THE BEST ORDER OF LONG AUTOREGRESSIVE MODELS FOR MOVING AVERAGE ESTIMATION

P.M.T. Broersen
Department of Applied Physics, Delft University of Technology
P.O.Box 5046, 2600 GA Delft, The Netherlands
phone + 31 15 278 6419, fax + 31 15 278 4263, email broersen@tn.tudelft.nl

ABSTRACT

Durbin's method for Moving Average (MA) estimation uses the estimated parameters of a long AutoRegressive (AR) model to compute the desired MA parameters. A theoretical order for that long AR model is ∞, but very high AR orders lead to inaccurate MA models in the finite sample practice. A new theoretical argument is presented to derive an expression for the best finite long AR order for a known MA process and a given sample size. Intermediate AR models of precisely that order produce the most accurate MA models. This new order differs from the best AR order to be used for prediction. An algorithm is presented that enables use of the theory for the best long AR order in known processes to data of an unknown process.

I. INTRODUCTION

In looking for a safe, robust and practical solution for the MA estimation problem, Durbin's method [1] is promising. A non-linear estimation problem is replaced by two stages of linear estimation. Firstly, the parameters of a long autoregressive model are estimated from the data. Afterwards, a second procedure uses those AR parameters as input. This method is based on the asymptotical equivalence of AR(∞) and MA(q) processes. Practice and simulations, however, have shown that the best AR order is finite and depends on the true process parameters and on the number of observations [2,3]. Non-linear algorithms for MA estimation are not considered in this paper, because it has been demonstrated that their performance in finite and small samples is poor [2]. Especially the lack of convergence or the convergence to non-invertable models prohibits general use of those methods in the routine analysis of time series [2,3]. However, an other two-stage method, mostly used in the estimation of ARMA models [4,5], can also be applied to MA estimation. The long AR model is used then to reconstruct residuals for the process that become regressors in an ordinary least squares solution.

The search is for a theoretical concept to find a best order for the long AR model as intermediate stage in the estimation of a MA model. The usual aim is the best order for AR *prediction*. However, the best order turned out to be a different one in simulations [3] where the *parameters* of a long AR model have been used to estimate MA parameters. Of course, the best AR order is defined here as the AR order yielding the final MA model with the highest prediction accuracy. The subject of order selection is extensively treated in linear regression with non-stochastic regressors. The best order for an estimated model depends on the intended use, which shortly can be either prediction or parameter accuracy [6]. Prediction as purpose gives a good fit of the model to the mean response of the given data; otherwise the smallest mean square error of the estimated parameters will be the criterion of quality.

In this paper, the theory for linear regression will be applied to three different AR order selection problems. The first type is characterised by using the AR model itself for prediction, the second type uses the reconstructed residuals of the AR model to find a MA model and the third type uses the AR model parameters to compute the MA model. The theory explains that the usual optimal AR order for prediction governs the first two types; the third type depends on the new long AR order for parameter accuracy. The computation of the two optimal orders is described for *known* MA parameters. Simulations show that the theoretical orders are also the best in practice. The theoretical knowledge about the best AR order can be utilised in a practical algorithm by adding a third stage.

II. LINEAR REGRESSION THEORY FOR MODEL ORDER SELECTION

In linear regression, the subject of order selection is known as subset selection or as the selection of variables. The size of the best estimated subset depends on the intended use of that subset model [6], which can be divided into two classes:

- prediction, or a good fit of the model to the mean response for the given range of the input regressor variables
- a small mean square error of the estimated parameters, or an accurate output for one specified input vector.

Both purposes will give different optimal numbers of variables for estimated models, depending on the true values of the parameters and their influence on the residual sum of squares. The theory is based on the reduction in the residual sum of squares that is found by including more parameters in the model. The requirements can be formulated *for true or for estimated parameters* of a given process.

- To be included in the best regression model for **prediction** [6], each group of r *true* process parameters should at least give a reduction of $r\sigma_\varepsilon^2$ in the residual sum of squares, where σ_ε^2 is the variance of the statistically independent residuals in the data. For *estimated* parameters, an additional reduction of σ_ε^2

is added for the variance of each parameter, so the total reduction must be greater than $2r\sigma_\varepsilon^2$. The factor 2 is the magic penalty factor in Mallows's C_p [6] and in Akaike's AIC criterion for autoregressive order selection [5].

- In contrast, if the primary concern is **accurate estimates of the parameters**, the required reduction of the residual sum of squares for any group of *true* parameters is only σ_ε^2, independent of the number of parameters involved to achieve that reduction [6]. This means that also smaller parameter values have to be included in the best subset for accurate parameters. Unfortunately, no good selection criterion based on *estimated* parameters is available for this class of applications.

The selection theory for regression will now be applied to time series. A common restriction to AR models is that they are hierarchical, so an AR(p) model contains exactly the first p parameters and subset models are not considered. The first class, prediction, is used in the derivation of criteria for AR order selection. The second theoretical class, parameter accuracy, will be applied to AR models of a true MA(q) process, with given MA parameters. The true AR model has order ∞ and true AR reflection coefficients κ_i can be found by computing all q+1 non-zero covariances of the MA(q) process, adding zeros for the covariances of greater time shifts and use the Levinson recursion [5] to compute as many AR parameters as desired. Define a MA(q) process as:

$$y_n = \varepsilon_n + b_1\varepsilon_{n-1} + \cdots + b_q\varepsilon_{n-q} \quad \text{or} \quad y_n = B_q(z)\varepsilon_n \qquad (1)$$

where ε_n is a series of independent identically distributed random variables with mean zero and variance σ_ε^2. The variance of the process y_n can be expressed in the MA parameters b_i or the AR reflection coefficients κ_i as:

$$\sigma_y^2 = \sigma_\varepsilon^2\left(1+\sum_{i=1}^{q} b_i^2\right) = \sigma_\varepsilon^2 / \prod_{i=1}^{\infty} 1-\kappa_i^2 \,. \qquad (2)$$

The expectation of the residual sum of squares of all AR models from AR(1) to AR(∞) can be determined. The residual sum of squares found by including the theoretical values of the reflection coefficients κ_i of an AR(p) model of N observations of a MA(q) process is denoted RSS_p. It becomes:

$$RSS_p = \sum_{n=1}^{N}\sigma_y^2\prod_{i=1}^{p}\left(1-\kappa_i^2\right) = RSS_{p-1}\left(1-\kappa_p^2\right). \qquad (3)$$

The asymptotical expectation of an estimated reflection coefficient k_i^2 is approximately 1/N for high orders where all k_i of still higher orders are small, and exactly 1/N for all reflection coefficients above the true AR process order. The best AR order, K, for prediction is found to be that specific order for which all models of orders K-r have a value for RSS_{K-r} exceeding RSS_K with at least $r\sigma_\varepsilon^2$, while all models of orders K+r have a value for RSS_{K+r} that is less than $r\sigma_\varepsilon^2$ smaller than RSS_K, for all r. Unfortunately, that order cannot be determined easily, because the equation (3) for RSS_p is multiplicative. The correspondence between Mallows's C_p [6] and Akaike's AIC criterion is used to find an asymptotical approximation to that order K by using the generalisation of AIC denoted GIC(p,α), [7], with penalty factor 1 for α:

$$GIC(p,\alpha) = \ln\left(RSS_p / N\right) + \alpha p / N \,. \qquad (4)$$

So K is the order with minimum GIC(p,1), p=0,1,..,∞.

The order with the best parameter accuracy M is found as the order M for which the residual sum of squares, RSS_M, is σ_ε^2 greater than RSS_∞ where all AR parameters are included. RSS_∞ is equal to $N\sigma_\varepsilon^2$ for N observations, as follows from (2) and (3). So the expectation for RSS_M is $(N+1)\sigma_\varepsilon^2$. This means that the residual variance equals $(1+1/N)\sigma_\varepsilon^2$. If the best orders concerned are greater than 0.1N, it is better to replace the asymptotical quantity 1/N by its method dependent finite sample value v_i [7,8].

Simulations have been carried out to validate those theoretically based applications of linear regression orders to time series. The accuracy of the finally resulting estimated MA models is reported as the average of the Selection Error SE [3]. That is defined as a scaled expectation of the prediction error. It is computed by applying the parameters of an MA(q') model estimated from y_n to new and independent observations x_n of the same process as given in (1):

$$\eta_n = \frac{x_n}{\hat{B}(z)} = \frac{B(z)}{\hat{B}(z)}\varepsilon_n \,. \qquad (5)$$

The Selection Error SE(q') of a MA(q') model of a MA(q) process is now defined as [3]:

$$SE(q') = N\left(\sigma_\eta^2 / \sigma_\varepsilon^2 - 1\right) \qquad (6)$$

if all zeros of the equation $\hat{B}(z) = 0$ are inside the unit circle. For estimated zeros exactly on the unit circle, the Selection Error is infinite unless the true process B(z) had a zero at exactly the same location. Moreover, the Selection Error approaches infinity for zeros close to the unit circle.

Fig.1 gives an example of the results. The theoretically computed optimal model orders are 12 and 20 for prediction and parameter accuracy, respectively. The minima in Fig.1 are

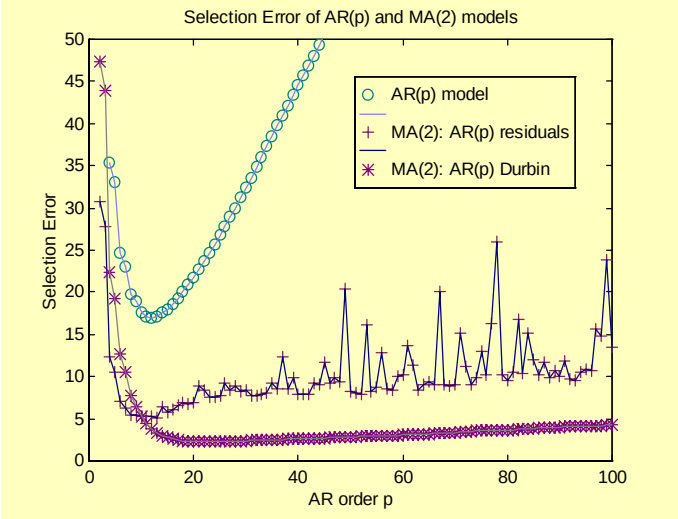

Fig.1 Average Selection Error of an AR(p) and two MA(2) models; 1000 simulation runs; N=400, $b_1 = -0.24$, $b_2 = -0.6$

at order 12 for AR models and for MA(2) models, estimated from the residuals of AR(p) models. The minimum SE for the MA(2) model, estimated from the AR parameters with Durbin's method is at order 23. The minimum is rather flat. So the theoretical result of 20 for the best long AR model order was a much better approximation than the order 12 with best AR predictions. Simulations with several other MA processes have shown that always the calculated AR order K with minimum GIC(K,1) is the best order for AR models used for prediction. Likewise, the long AR model order M gives the highest accuracy of the MA model, if the AR(M) parameters are used to compute the MA model with Durbin's method.

The algorithm that uses the *residuals* of the long AR model to compute the MA parameters causes problems in practice. The zeros can lay outside the unit circle or close to it. For small sample sizes many runs give useless non-invertible results. The best AR order in successful simulation runs is K with minimum GIC(K,1): the best order for AR prediction. Fig.1 shows an irregular pattern for the residual based MA(2) Selection Error. This is caused by a small number of runs with zero's close to the unit circle. Remark that the minimum Selection Error is 5.12; the minimum is 2.24 for Durbin's method and the asymptotical minimum value in maximum likelihood theory is 2, equal to the number of estimated parameters So the use of residuals of long AR models is not advisable in MA estimation. Durbin's method always gives invertible results, and also with better quality because the Selection Error is smaller. Use of AR residuals in ARMA estimation has the same problems with non-invertible MA models.

III. SLIDING WINDOW TECHNIQUE

So far, the best AR order has been determined for a *given* MA process. The order can be computed by determining the reduction of the residual sum of squares RSS_M for the true values of the reflection coefficients in (3). Unfortunately, no AR order selection criterion can select the desired best theoretical order M from data. Existing selection criteria are directly or indirectly based on a transformation of the estimated residual sum of squares as a function of the model order. An estimated k_p gives a multiplicative reduction with $1-k_p^2$ in the residual variance. The problem is to detect a bias contribution to the residual variance of about 1/N for r combined parameters where the variance of each parameter has a contribution of about 1/N. This would require a penalty of magnitude 1+1/r , instead of the penalty factor 2 in AIC. The difficulty is that a penalty less than 2 gives enormous costs of overfit [7]. If the best order is higher than the order with minimum prediction error, no existing order selection criterion is appropriate to select that order from given observations.

An algorithm has been developed that uses a preliminary estimate of the MA model, as if it were the true process, to determine the order of the long AR model. Firstly, the preliminary MA model is estimated with Durbin's method with the sliding window technique [2,3], using a long AR model with length greater than twice the selected AR order K. In Fig.1, it is demonstrated that poor MA models can be found if the long AR order is too low, too high AR orders are less dangerous. Moreover, the theoretical requirement for the best order for parameter accuracy showed that this order is the same or higher as the best order for prediction. Therefore, a first guess is made by taking the number of MA parameters plus two times the selected AR order for prediction as the preliminary long AR model. The stages of the algorithm are:

First stage: AR

Estimate AR models of orders 0 to N/2 from the data with the Burg AR estimation method. The Yule-Walker method for AR estimation gives sometimes slightly better results, but it will become much worse for zeros close to the unit circle [3]. So for unknown data, Burg's method is to be preferred. Select the AR order K, preferably with FSIC [8] which is the only criterion that can deal with AR models of order N/2. In using other criteria than FSIC, the result in simulations is worse and the maximum AR order has to be limited to N/3 or even to N/4 to evade selection peculiarities.

Second stage: SW

Use Durbin's method to estimate MA(q) models of order 1 to a maximum chosen candidate order L; each MA(q) model is estimated from a long AR model with 2K+q parameters, a sliding window that depends on the selected AR order K and on the number of MA parameters q that is computed; if 2K+q is greater than N/2, this latter order will be used. Compute the residual variances (with backforecasting [3]) by substituting the estimated MA(0) to MA(L) parameters in the data models. Use these residual variances to select the MA order Q with minimum GIC(Q,3), with α=3 in (4), or a similar criterion; GIC(p,3) is similar to AIC but has a better balance between underfitting and overfitting risks [7]. The asymptotical derivation of properties of GIC(p,α) for AR models is also valid for MA models.

Third stage: SW+

Treat the preliminary model MA(Q) as if it is the true process and determine the best long AR order M for that 'given' MA(Q) process and recalculate the MA(Q) parameters from this final AR(M) model.

Third stage NL:

Compute the non-linear MA model of order Q of the second stage with the least squares algorithm using backforecasted residuals. This model is only accepted if all zeros are in the invertible region; otherwise, the model found in the second stage is kept. This non-linear third stage is introduced to investigate the possibility of selecting the structure with some basic algorithm and to improve the preliminary estimate with an approximate maximum likelihood estimator. In this way, it is possible to compare the behaviour of the best asymptotical estimator with the SW and SW+ algorithms in finite sample simulations.

IV. SIMULATION RESULTS

Theoretical derivations have an asymptotical validity, but the finite sample performance can only be determined in practice

or calibrated in simulations. Many different processes have been simulated, for various process orders and sample sizes. Table 1 presents the results for 50 observations on MA(4) processes. Those processes have been generated by using 'reflection coefficients' [1 β −β β −β] and compute the MA(4) parameters with the Levinson recursion [5]. By taking β less than 1 in absolute value, this ensures that all true processes are invertible. The average results of AR, SW, SW+ and NL have been investigated. For a comparison, the best theoretical long AR order has been used in the simulations to compute the MA(4) model of the true order, so neither AR nor MA order selection took place; this result is given in the Tables in the row 'theory'.

Table 1: Selection Error of estimated and selected MA models as a function of β, with various algorithms. N=50.

β	–0.8	–0.4	0.0	0.4	0.8
SW	12.00	8.25	1.56	12.00	12.24
SW+	12.52	7.98	1.19	11.86	12.66
NL	44.20	97.70	93.17	105.26	31.45
theory	12.22	5.24	0.00	7.69	13.44
AR	21.88	10.02	2.14	13.82	23.32

It is remarkable that the third stage SW+ is hardly an improvement in comparison with the second stage SW; but it would be if K or 3K had been used for the long AR order instead of 2K+q. For β= – 0.8 and 0.8, the selection error of SW was even better than that of the row 'theory'. The explanation is that the maximum considered AR order is N/2, so 25 in Table 1 and for those values of β the best AR order is greater than 25. So even 'theory' is not based on the best AR order then; one may say that 50 observations is not enough for a reliable estimation in those examples.

This can be demonstrated in Table 2, where the MA(4) example with β= – 0.8 has been simulated for different sample sizes. Again SW+ with third stage gives no useful improvement of the SE, so the third stage can be omitted. NL is almost as good as SW for N=1000. Eventually, for still greater N, it may become the same or even slightly better. The theoretical asymptotical value for the selection error is 4 if the correct MA(4) model would be selected in all simulation runs. SW is already quite close to that limiting value. In most examples, NL performs poorly with often estimated zeros outside the unit circle or close to it, giving a non-invertible model or a high value for the SE. Only for large samples, say N>1000 and for true zeros close to the unit circle, some examples have been simulated where NL was somewhat better than SW, but even then the risk of non-invertible models remained.

Table 2: Selection Error of estimated and selected MA models as a function of N, with various algorithms. β= - 0.8.

N	20	50	100	200	500	1000
SW	16.84	12.00	9.93	8.65	6.52	6.12
SW+	13.88	12.52	10.09	8.46	6.70	6.24
NL	127.38	44.20	15.23	13.63	11.10	6.54
theory	13.76	12.22	9.33	7.22	5.15	5.60
AR	19.71	21.88	29.58	35.90	41.88	52.60

A study of selection losses for AR modelling has been carried out [7]. In comparison with AR modelling, the accuracy of the finally selected MA models with the sliding window technique for Durbin is remarkably good: the Selection Error for estimated MA models is close to the theoretical minimum value 4 for MA(4) processes. This might be caused by the fact that MA models are computed from AR parameters, without considering the residual variance of the MA model. That variance is only computed afterwards for an order selection criterion. Not using the residual sum of squares twice is an advantage in MA selection in comparison with AR. As a consequence, MA order selection gives a smaller average contribution to the SE of the best fixed order model than AR order selection.

V. CONCLUDING REMARKS

A theoretical value for the best order of a long AR model in Durbin's method of MA estimation has been derived. It is characterised as the order yielding AR parameter estimates with the smallest mean square error. Simulations show that MA models calculated from the *parameters* of that long AR model have the smallest error of prediction. However, if the *residuals* of a long AR model are used, the best AR order is the order with the best AR prediction accuracy, which is lower.

The sliding window technique performs very well for all sample sizes where enough information is available. If the true correlation is not yet damped out for N/2, no method is very accurate but the performance of SW remains reasonable.

REFERENCES

[1] Durbin, J., "Efficient estimation of parameters in moving average models", *Biometrika*, 46, p 306-316, 1959.

[2] Broersen, P.M.T. and H.E. Wensink, *"Small sample MA estimation with Durbin's method"*, Selected papers from the 9-th IFAC-IFORS Symposium on Identification and System Parameter Estimation, Budapest 1991, Oxford: Pergamon, p 983-98, 1992.

[3] Broersen, P.M.T. and H.E. Wensink, "Practical aspects of Moving Average Estimation", *Proc. 15th GRETSI conference*, p 201-204, 1995.

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

[5] Kay, S.M. and S.L. Marple, "Spectrum analysis, a modern perspective", *Proc. IEEE*, vol 69, p 1380-1419, 1981.

[6] Hocking, R.R.,"The analysis and selection of variables in linear regression", *Biometrics*, vol.32, p 1-49, 1976.

[7] Broersen, P.M.T. and H.E. Wensink, "On the penalty factor for autoregressive order selection in finite samples", *IEEE Trans. on Signal Processing*, march 1996.

[8] Wensink, H.E. and P.M.T. Broersen, "*Estimating the Kullback-Leibler information for autoregressive model order selection in finite samples*", Signal Processing VII: Theories and Applications, Proc. Eusipco Conf., Edinburgh, p 1847-1850, 1994.