

# An efficient algorithm for selecting optimal configurations of AR-coefficients

R. Moddemeijer

*Abstract*—There exists an essential difference between the correct Auto Regressive (AR) model and the optimal AR-model. We try to find an optimal model balancing between flexibility, using many AR-parameters, and low variance, using only a few AR-parameters. We select an optimal AR-parameter configuration consisting of zero and non-zero parameters given a maximum AR-order. This optimal configuration will be selected using a Modified Information Criterion (MIC) which is closely related to Akaike's criterion (AIC). This MIC allows an a priori selection of the probability of estimating too many parameters.

We present the theoretical foundation of the method and verify this method by simulations. The method is based on pivoting the Hessian matrix by Gauß-Jordan pivots. As a result we can now select an optimal parameter configuration with an a priori probability of selecting a configuration with a too large number of parameters given an a priori selected maximum AR-order.

*Keywords*—AIC, Akaike criterion, AR, autoregressive processes, composite hypothesis, maximum likelihood, model order, system identification, time series analysis.

## I. INTRODUCTION

Hocking [1] showed that two completely different definitions of the best model order can be given. The usual one applies to the overall fit of the model and leads to selection criteria like the Akaike criterion. The other order is applicable to situations where the estimated Auto-Regressive (AR) parameters themselves are important because they are used as an intermediate to compute quantities like harmonic frequencies, spectra, Moving Average (MA) or Auto-regressive Moving Average (ARMA) model parameters.

We distinguish between the *correct* model and the *optimal* model. Our definition of the *correct* model coincides with the *true* signal model, the model which generated the signal. It can be a better strategy to estimate another model from an observed AR-process, the *optimal* model, which has a simpler structure and can more reliably be estimated. Due to the finite sample size not all aspects of the correct model can be estimated with a sufficient accuracy, so we balance between a minimum estimation error (variance) and a minimum modeling error (bias). This problem of balancing bias and variance is a statistical problem which leads to the concept of an optimal model in the mean square sense. This optimal model is slightly biased with respect to the correct model and performs better with respect to the correct model due to a reduced variance.

Assume we have a  $I^{th}$  order AR-model with  $I$  AR-parameters  $a_1, a_2, \dots, a_I$ . Some of these parameters will be large, some will be small and some are zero. Why is it

necessary to estimate all these parameters of an observed AR-process? Zero parameters can be put to zero, avoiding any estimation error. Small parameters can be negligible with respect to the estimation error. In this case these parameters are meaningless and can also be fixated at zero. We try to find a balance between flexibility, using many parameters, and low variance, using only a few parameters. To deal with this problem we use a large parameter space with only a few active (non-zero) parameters.

In practical situations neither the AR-order nor the number of negligible parameters is a priori known. Therefore we search for an efficient strategy to estimate the configuration of significant parameters. Our algorithm is based on an ARMA-parameter estimation algorithm [2] which was used to find alternative ARMA-parameter configurations providing a better fit to the time-series published in the book of Box & Jenkins [3]. This algorithm has been combined with the AR-order estimation algorithm based on the *Modified Information Criterion* (MIC) [4]. This algorithm is capable of estimating an AR-order given an a priori selected probability of selecting a too high order.

We apply the method of testing composite hypotheses [5, chapter 35], [6, pp. 86–96] to selection of the AR-parameter configuration. A composite hypothesis is a hypothesis which is specified except for a few parameters to be estimated. An example of two composite hypotheses is: *the average temperature on earth is constant*, with one unknown parameter, or *the average temperature increases linearly*, with two unknown parameters. This is one of the most difficult estimation problems because the model of the first hypothesis is contained in the model of the second hypothesis. To solve this problem many criteria are developed. We use MIC [4] which is similar to the Akaike criterion (AIC).

The Akaike criterion [7], [8], [9] integrates the method of Maximum Likelihood (ML) to estimate parameters and the likelihood-ratio test to discriminate the hypotheses. The Akaike criterion is widely accepted and is for example used to determine the order of AR- [10], [11], [12] and ARMA-models [13]. Other applications of the Akaike criterion are: maximum likelihood estimation [14], data-mining [15], EEG and EMG data processing [16], [17] and Geophysics [18]. There exist a number of competitive criteria [19], [20], [21].

In case of the estimation of a parameter configuration we will successively be adding parameters to an initially empty space of non-zero parameters. We are dealing with a null hypothesis, the current configuration of non-zero parameters is sufficient, and a set of alternative hypotheses corresponding with the parameters which can possibly be added

to this space of non-zero parameters. All these hypotheses are composite hypotheses because the relevant parameters should also be estimated.

## II. BALANCING BIAS AND VARIANCE

The principle of balancing bias and variance can easily be explained by the following statistical model [22], [23], [24]. Assume an unbiased statistic  $s$  with a given variance:

$$\begin{aligned} \text{BIAS} \{ \underline{s} \} &= 0 \\ \text{MSE} \{ \underline{s} \} &= \text{VAR} \{ \underline{s} \} \end{aligned} \quad (1)$$

where MSE means the mean square error. Can we construct a better statistic? In mean square sense it can be done. Assume the new statistic  $s' = \lambda s$ . This statistic has bias and mean square error:

$$\begin{aligned} \text{BIAS} \{ \underline{s}' \} &= (\lambda - 1) \text{E} \{ \underline{s} \} \\ \text{MSE} \{ \underline{s}' \} &= (\lambda - 1)^2 \text{E} \{ \underline{s} \}^2 + \lambda^2 \text{VAR} \{ \underline{s} \} \end{aligned} \quad (2)$$

We search for the minimum mean square error as a function of  $\lambda$  and find the *optimal* statistic  $s' = \lambda s$  where:

$$\lambda = \frac{\text{E} \{ \underline{s} \}^2}{\text{E} \{ \underline{s} \}^2 + \text{VAR} \{ \underline{s} \}} \quad (3)$$

Consequently  $0 \leq \lambda \leq 1$ . This simple example shows the principle of balancing bias and variance. If we accept a small bias, the decrease of the mean square error by the reduction of the variance dominates the increase by the bias. This means that we can construct better estimators by accepting some bias.

This principle of balancing bias and variance can also be applied to AR-modeling. Assume we estimate an AR-parameter, with the statistic  $s$  or the alternative constant statistic  $s' = 0$ . The constant statistic  $s'$  performs better if the mean square error is smaller:  $\text{MSE} \{ \underline{s}' \} < \text{MSE} \{ \underline{s} \}$ . This condition is fulfilled if  $\text{E} \{ \underline{s} \}^2 < \text{VAR} \{ \underline{s} \}$ . Assume the correct AR-model has a small parameter. The value of this parameter can be negligible with respect to the variance; in this case it's better to omit the parameter in the optimal model. Omitting a parameter introduces some bias but also reduces the variance. This leads to a new concept of AR-modeling. We use a  $J^{\text{th}}$  order AR-model with  $J$  independently adjustable AR-parameters. Of every parameter we decide during the estimation process whether the parameter is essential for modeling or not. We only estimate the essential parameters of the AR-model and keep the other parameters equal to zero. In the next sections we develop the theory and verify the results using this concept.

## III. THEORY

Assume the following discrete time AR-models:

1.  $x_n = \epsilon_n$
2.  $x_n = 0.5 x_{n-1} + 0.25 x_{n-2} + 0.125 x_{n-3} + 0.0625 x_{n-4} + 0.03125 x_{n-5} + 0.015625 x_{n-6} + \epsilon_n$
3.  $x_n = 0.55 x_{n-1} + 0.05 x_{n-2} + \epsilon_n$
4.  $x_n = 0.75 x_{n-1} - 0.50 x_{n-2} + \epsilon_n$
5.  $x_n = 0.75 x_{n-1} - 0.50 x_{n-4} + \epsilon_n$
6.  $x_n = 0.50 x_{n-1} - 0.25 x_{n-4} + \epsilon_n$

where  $\epsilon$  is normally distributed white noise with zero mean and fixed variance  $\sigma^2$ .

$$f_\epsilon(\epsilon; \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\epsilon^2}{2\sigma^2}} \quad (4)$$

Assume the stochastic signal  $\underline{x}$ , generated by one of these AR-models, and a sequence of  $N$  observations  $x_1, x_2, \dots, x_N$ . Estimate the generation model of  $\underline{x}$ ; i.e. estimate the configuration and the values of the non-zero coefficients of the AR-model.

Assume the  $J^{\text{th}}$  order AR-model and its conditional probability density function:

$$f_x(x_n | x_{n-1}, \dots, x_{n-J}; \mathbf{a}, \sigma) = f_\epsilon(\epsilon_n(\mathbf{a}); \sigma) \quad (5)$$

where  $\mathbf{a} = (a_1, \dots, a_J)$  and where

$$\epsilon_n(\mathbf{a}) = x_n - \sum_{j=1}^J a_j x_{n-j} \quad (6)$$

This AR-model has a parameter vector  $\mathbf{p} = (a_1, a_2, \dots, a_J, \sigma)$  of  $J + 1$  independently adjustable parameters. If we fixate some AR-parameters at zero, the number of independently adjustable parameters will be less than  $J + 1$ .

The generation models with different configurations of zero and non-zero AR-coefficients are called the hypotheses. The dimension  $\dim \mathbf{p}$  of the parameter vector  $\mathbf{p}$  for a given parameter configuration is the number of independently adjustable parameters, i.e. the number of non-zero parameters. Given a sequence of observations we determine the parameter configuration by selecting the *optimal* hypothesis.

The case of testing two composite hypotheses is elaborately discussed in our earlier publication [4]. Given the *Maximum of the Average Log Likelihood* (MALL), i.e. the *Average Log Likelihood* (ALL) function at the ML parameter estimate  $\hat{\mathbf{p}}$ , defined by

$$\hat{\text{E}} \{ \underline{\mathcal{L}}(\hat{\mathbf{p}}) \} = \frac{1}{N} \sum_{n=1}^N \log f_x(x_n | \hat{\mathbf{p}}) \quad (7)$$

where  $\hat{\text{E}} \{ \underline{\mathcal{L}}(\hat{\mathbf{p}}) \}$  denotes the average, the statistic to estimate the mean, of the stochastic variable  $\underline{\mathcal{L}}(\hat{\mathbf{p}}) = \log f(\underline{x} | \hat{\mathbf{p}})$ . The quantity  $N \hat{\text{E}} \{ \underline{\mathcal{L}}(\hat{\mathbf{p}}) \}$  equals the well-known maximum of the log likelihood function in the method of maximum likelihood [26, chapter 7][27, section 11.5].

The *Generalized Information Criterion* (GIC) [28] is used to select an optimal hypothesis from a set of composite hypotheses:

$$\text{GIC}(\lambda) = -\hat{\text{E}} \{ \underline{\mathcal{L}}(\hat{\mathbf{p}}) \} + \frac{\lambda \dim \hat{\mathbf{p}}}{N} \quad (8)$$

where  $\lambda$  is an a priori defined constant that depends on the structure of  $\hat{\mathbf{p}}$ . The optimal hypothesis is considered to be [25] the hypothesis having a minimum GIC( $\lambda$ ). Different values [10] for  $\lambda$  are used: Bhansali ( $\lambda = \frac{1}{2}$ ) [25], [29], Akaike ( $\lambda = 1$ , [4] so AIC = GIC(1)) [7], Broersen ( $\lambda = \frac{3}{2}$ ) [30] and Åström ( $\lambda = 2$ ) [31]. Even considerably larger values of  $\lambda$ , even

depending on  $N$ , have been used [20], [19]. The concept of testing a set of composite hypotheses using a criterion which is derived to test a pair of composite hypotheses is *wrong!*

In earlier work [4] we have argued that the problem of AR-order estimation can be solved by subsequently testing pairs of hypotheses using the ALL-ratio at the ML-estimates (MALL-ratio) as a criterion:

$$\widehat{E} \{ \Delta \mathcal{L}(\widehat{\mathbf{p}}_0, \widehat{\mathbf{p}}_1) \} = \widehat{E} \{ \mathcal{L}(\widehat{\mathbf{p}}_1) \} - \widehat{E} \{ \mathcal{L}(\widehat{\mathbf{p}}_0) \} > \eta_{high} \quad (9)$$

We accept the alternative hypothesis with parameter vector  $\widehat{\mathbf{p}}_1$  if the MALL-ratio is larger than  $\eta_{high}$ ; in the other case we accept the null hypothesis with parameter vector  $\widehat{\mathbf{p}}_0$ . The value of  $\eta_{high}$  depends on the difference in number of independently adjustable parameters ( $\dim \widehat{\mathbf{p}}_1 - \dim \widehat{\mathbf{p}}_0$ ) and the a priori selected *False Alarm Probability* (FAP)  $\alpha$ ; i.e. the probability on selecting a too high order (see table VI in [4]).

Within our concept we have the same null hypotheses but a set of alternative hypotheses instead of only one alternative hypothesis. We want to extend the successful concept of our earlier work to this situation.

#### IV. MODEL SELECTION BY GAUSS-JORDAN PIVOTS

A Gauß-Jordan pivot is an operation on a matrix  $V$  resulting in the pivoted matrix  $\tilde{V}$ . Performing a Gauß-Jordan pivot with pivot element  $V_{ij}$  exchanges the elements  $a_i$  and  $b_j$  in the matrix equation

$$\begin{bmatrix} a_0 \\ \vdots \\ a_{i-1} \\ a_i \\ a_{i+1} \\ \vdots \\ a_J \end{bmatrix} = V \begin{bmatrix} b_0 \\ \vdots \\ b_{j-1} \\ b_j \\ a_{j+1} \\ \vdots \\ b_J \end{bmatrix} \quad \text{so} \quad \begin{bmatrix} a_0 \\ \vdots \\ a_{i-1} \\ b_j \\ a_{i+1} \\ \vdots \\ a_J \end{bmatrix} = \tilde{V} \begin{bmatrix} b_0 \\ \vdots \\ b_{j-1} \\ a_i \\ a_{j+1} \\ \vdots \\ b_J \end{bmatrix} \quad (10)$$

Performing Gauß-Jordan pivot using all main diagonal elements of a matrix as pivot element is equivalent to inversion of the matrix. Using a selected subset of main diagonal element leads to a *partial inverse* of the matrix. The Gauß-Jordan pivot of the matrix  $V$  with pivot element  $V_{ij}$  resulting in the matrix  $\tilde{V}$  can be computed by:

$$\begin{aligned} \tilde{V}_{mn} &= V_{mn} - V_{mj}V_{in}/V_{ij} \quad \text{where } m \neq i \text{ and } n \neq j \\ \tilde{V}_{mi} &= V_{mj}/V_{ij} \quad \text{where } n \neq j \\ \tilde{V}_{in} &= V_{in}/V_{ij} \quad \text{where } m \neq i \\ \tilde{V}_{ij} &= 1/V_{ij} \end{aligned} \quad (11)$$

We determine the ALL-function given the parameter vector  $\mathbf{p}$  by substitution of (4) into (7):

$$\widehat{E} \{ \mathcal{L}(\mathbf{p}) \} = \frac{1}{N} \sum_{n=1}^N \left( -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{\epsilon_n^2(\mathbf{a})}{2\sigma^2} \right) \quad (12)$$

Searching the maximum of the ALL-function as a function of the parameter vector  $\mathbf{p}$  leads to the ML estimate  $\widehat{\mathbf{p}}$ . We

can separate the estimation of  $\sigma$  and of  $\mathbf{a} = (a_1, a_2, \dots, a_J)$ . Consequently we find:

$$\hat{\sigma}^2 = v(\widehat{\mathbf{a}}) \quad \text{where } v(\mathbf{a}) = \frac{1}{N} \sum_{n=1}^N \epsilon_n^2(\mathbf{a}) \quad (13)$$

and

$$\begin{aligned} \widehat{\mathbf{a}} &= -v_{\mathbf{aa}}^{-1}(\mathbf{0})v_{\mathbf{a}}(\mathbf{0}) \\ v(\widehat{\mathbf{a}}) &= v(\mathbf{0}) - \frac{1}{2}v_{\mathbf{a}}(\mathbf{0})v_{\mathbf{aa}}^{-1}(\mathbf{0})v_{\mathbf{a}}(\mathbf{0}) \end{aligned} \quad (14)$$

where

$$v_{\mathbf{a}}(\mathbf{0}) = \left. \frac{\partial}{\partial \mathbf{a}} v(\mathbf{a}) \right|_{\mathbf{a}=\mathbf{0}} \quad v_{\mathbf{aa}}(\mathbf{0}) = \left. \frac{\partial}{\partial \mathbf{a}} \frac{\partial}{\partial \mathbf{a}}^T v(\mathbf{a}) \right|_{\mathbf{a}=\mathbf{0}} \quad (15)$$

are the vector of first derivatives and the matrix of second derivatives (Hessian matrix) with respect to the parameters. By  $\mathbf{0}$  we mean a vector with all elements equal zero.

We construct the matrix  $V$ :

$$V = \begin{bmatrix} v(\mathbf{0}) & -\frac{1}{2}v_{\mathbf{a}}^T(\mathbf{0}) \\ -\frac{1}{2}v_{\mathbf{a}}(\mathbf{0}) & \frac{1}{2}v_{\mathbf{aa}}(\mathbf{0}) \end{bmatrix} \quad (16)$$

with elements  $V_{ij}$  ( $0 \leq i \leq J$  and  $0 \leq j \leq J$ ). These elements can be computed by:

$$V_{ij} = \sum_{n=1}^N x_{n-i}x_{n-j} \quad (17)$$

Performing  $J$  Gauß-Jordan pivots on this matrix subsequently using  $V_{jj}$  ( $1 \leq j \leq J$ ) as pivot element leads to:

$$\begin{aligned} &\begin{bmatrix} v(\widehat{\mathbf{a}}) & \widehat{\mathbf{a}}^T \\ -\widehat{\mathbf{a}} & -2v_{\mathbf{aa}}(\mathbf{0})^{-1} \end{bmatrix} \\ &= \begin{bmatrix} v(\mathbf{0}) - \frac{1}{2}v_{\mathbf{a}}(\mathbf{0})v_{\mathbf{aa}}^{-1}(\mathbf{0})v_{\mathbf{a}}(\mathbf{0}) & -v_{\mathbf{a}}^T(\mathbf{0})v_{\mathbf{aa}}^{-1}(\mathbf{0}) \\ v_{\mathbf{aa}}^{-1}(\mathbf{0})v_{\mathbf{a}}(\mathbf{0}) & 2v_{\mathbf{aa}}(\mathbf{0})^{-1} \end{bmatrix} \end{aligned} \quad (18)$$

Notice that  $V_{00}$  is the only element of the main diagonal that is not used as pivot element. Similarly we can determine partial inverses of the matrix representing a ML-estimate by using only a subset of the full parameter space. E.g. we find  $v(\widehat{\mathbf{a}})$  where  $\widehat{\mathbf{a}} = (\widehat{a}_1, \widehat{a}_2, 0, \dots, 0)$  by only using  $V_{11}$  and  $V_{22}$  as pivot element.

If we have  $J$  AR-parameters, there exist  $2^J$  partial inverses representing ML-estimates with 0, 1, 2 to  $J$  AR-parameters. If  $J$  becomes large, the selection of the optimal partial inverse becomes rather computationally intensive. To reduce the computational effort, we only use the main diagonal element  $V_{kk}$  which lead to a maximum reduction of  $V_{00}$  as pivot element. This reduction can be predicted by evaluation of  $V_{00} - \tilde{V}_{00}$ :

$$V_{00} - \tilde{V}_{00} = V_{0k}V_{k0}/V_{kk} \quad (19)$$

During the partial inversion process we use the element  $V_{kk}$  which has a maximum  $V_{0k}V_{k0}/V_{kk}$  as the next pivot element. Now the decision remains whether the inclusion of the parameter  $a_k$  to the set of non-zero parameters leads to a better model. Due to the properties of  $V$  any main diagonal element  $V_{kk}$  ( $1 \leq k \leq J$ ), which is used for the first

time as pivot element during the partial inversion process, leads to a decrease of  $V_{00}$  because  $V_{0k}V_{k0}/V_{kk} \geq 0$ . This is not surprising, because any additional parameter leads to a model with smaller  $\hat{\sigma}$ . We are searching for the *optimal* model, and not for the model which leads to the smallest  $\hat{\sigma}$ . We use MIC to construct a threshold to determine whether an additional parameter leads to a better model.

We use the ALL-function as a criterion. Substitution of the ML parameter estimate into the ALL-function leads to the Maximum Average Log Likelihood (MALL) of the hypothesis. Substituting (12) into (13) we find:

$$\widehat{E} \{ \mathcal{L}(\widehat{\mathbf{p}}) \} = -\frac{1}{2} \ln (2\pi\hat{\sigma}^2) - \frac{1}{2} \quad (20)$$

We determine the increase of the MALL in case of the introduction of one additional parameter. Assume we have an estimated parameter vector  $\widehat{\mathbf{p}}_0$ , given the parameter configuration of the null hypothesis, and similar estimated parameter vectors  $\widehat{\mathbf{p}}_k$ , given parameter configurations of the alternative hypotheses, which all have one additional AR-parameter with respect to the null hypothesis. The MALL-ratio depends on  $k$  and equals:

$$\begin{aligned} \widehat{E} \{ \Delta \mathcal{L}(\widehat{\mathbf{p}}_0, \widehat{\mathbf{p}}_k) \} &= \widehat{E} \{ \mathcal{L}(\widehat{\mathbf{p}}_k) \} - \widehat{E} \{ \mathcal{L}(\widehat{\mathbf{p}}_0) \} = \ln \left( \frac{\hat{\sigma}_0}{\hat{\sigma}_k} \right) \\ &= -\frac{1}{2} \ln \left( 1 - \frac{V_{0k}V_{k0}}{V_{00}V_{kk}} \right) \end{aligned} \quad (21)$$

Compare this equation with (9). If the additional parameter is superfluous ( $\hat{\sigma}_0 \approx \hat{\sigma}_k$ ) this statistic is according to earlier publications [4], [32], [33] chi-squared distributed with one degree of freedom. If the additional parameter is a significant contribution to the model ( $\hat{\sigma}_0 \gg \hat{\sigma}_k$ ) the expectation of (21) differs significantly from zero, the statistic is normally distributed [4]. We are only interested in the first case because the second case is a typical indication that the algorithm should proceed in adding more parameters.

Assume there are  $K$  different parameters which can possibly be added tot the space of non-zero parameters. If  $K = 1$  we have a situation which equals the situation with AR-order estimation where you possibly add one pre-selected parameter in every iteration step. The threshold for the FAP  $\alpha$  in case of one additional parameter can be solved from the equation [4]:

$$\alpha = \int_{2N\eta_{high}}^{\infty} f_{\chi^2}(x) dx = 1 - \text{erf} \left( \sqrt{N\eta_{high}} \right) \quad (22)$$

In general we want to find the threshold  $\eta_{high}$  given  $\alpha$ ; therefore we determine the inverse function:

$$\eta_{high} = \frac{(\text{erf}^{-1}(1 - \alpha))^2}{N} \quad (23)$$

If  $K > 1$  the threshold should be higher because there are two or more possibilities of adding one parameter to the space of non-zero parameters.

Assume that the MALL-ratios,  $\widehat{E} \{ \Delta \mathcal{L}(\widehat{\mathbf{p}}_0, \widehat{\mathbf{p}}_k) \}$  for different  $k$ , in case of adding different superfluous parameters are statistically independent amongst each other (see also

	1%	2%	5%	10%	20%
1	6.6349	5.41189	3.84146	2.70554	1.64237
2	7.8749	6.62592	5.00183	3.79791	2.61927
3	8.6093	7.34857	5.70129	4.46923	3.24407
4	9.13371	7.86577	6.20466	4.95627	3.70473
5	9.54216	8.26917	6.59854	5.33915	4.07015
6	9.87691	8.60009	6.92236	5.65489	4.37324
7	10.1606	8.88074	7.19742	5.92367	4.63233
8	10.4068	9.12445	7.43657	6.15775	4.85867
9	10.6244	9.33986	7.43657	6.36513	5.05967
10	10.8192	9.53288	7.64815	6.55130	5.24048

TABLE I

THE VALUE  $2N\eta_{high}$  AS A FUNCTION OF THE FALSE ALARM PROBABILITY  $\alpha$  COMPUTED USING THE PI-DISTRIBUTION AS A FUNCTION OF  $K$ .

figure 1). At this moment there is no theoretical evidence for this assumption. Empirically we obtain good results, so this assumption seems to be reasonable.

Given this assumption, the maximum of  $\widehat{E} \{ \Delta \mathcal{L}(\widehat{\mathbf{p}}_0, \widehat{\mathbf{p}}_k) \}$  is pi-distributed, where the pi-distribution is defined in appendix A. Instead of (9) we use as a criterion:

$$\max \widehat{E} \{ \Delta \mathcal{L}(\widehat{\mathbf{p}}_0, \widehat{\mathbf{p}}_k) \} > \eta_{high} \quad (24)$$

where we search the maximum as a function of the  $K$  alternative hypotheses. Similar to (22) we determine the threshold as a function of the FAP by solving  $\eta_{high}$  from the equation:

$$\alpha = \int_{2N\eta_{high}}^{\infty} f_{\Pi}(x) dx = 1 - (\text{erf}(\sqrt{N\eta_{high}}))^K \quad (25)$$

where  $f_{\Pi}$  is the pi-distribution. We solve  $\eta_{high}$ :

$$\eta_{high} = \frac{(\text{erf}^{-1}(\frac{\sqrt{1-\alpha}}{N}))^2}{N} \quad (26)$$

The corresponding values of  $2N\eta_{high}$  can be found in table I. If the statistical dependence of the MALL-ratios plays a significant role, we should use a lower threshold  $\eta_{high}$ , because the effective number of parameters decreases. Keeping  $\eta_{high}$  unchanged will reduce the FAP.

Assume we have  $J$  AR-parameters and an empty space of non-zero parameters; consequently all parameters are zero. In the  $j^{th}$  iteration ( $1 \leq j \leq J$ ) of the estimation process we can add one parameter to the space of non-zero parameters by performing a Gauß-Jordan pivot, so this space contains exactly  $j - 1$  different AR-parameters. This means that we can choose between  $K = J - j + 1$  different parameters (pivot-elements) to be added. We select the parameters with the largest  $V_{0k}V_{k0}/V_{kk}$ ; this is the parameter which leads to a maximum increase of the MALL. We compare with (24) the  $K$  results of (21) with  $\eta_{high}$  of (26); if this MALL-ratio is larger than  $\eta_{high}$  we add the parameter to the space of non-zero parameters, by pivoting the matrix  $V$  with pivot element  $V_{kk}$ , and continue the iteration. If the MALL-ratio is smaller we stop the iteration and the optimal model is reached.

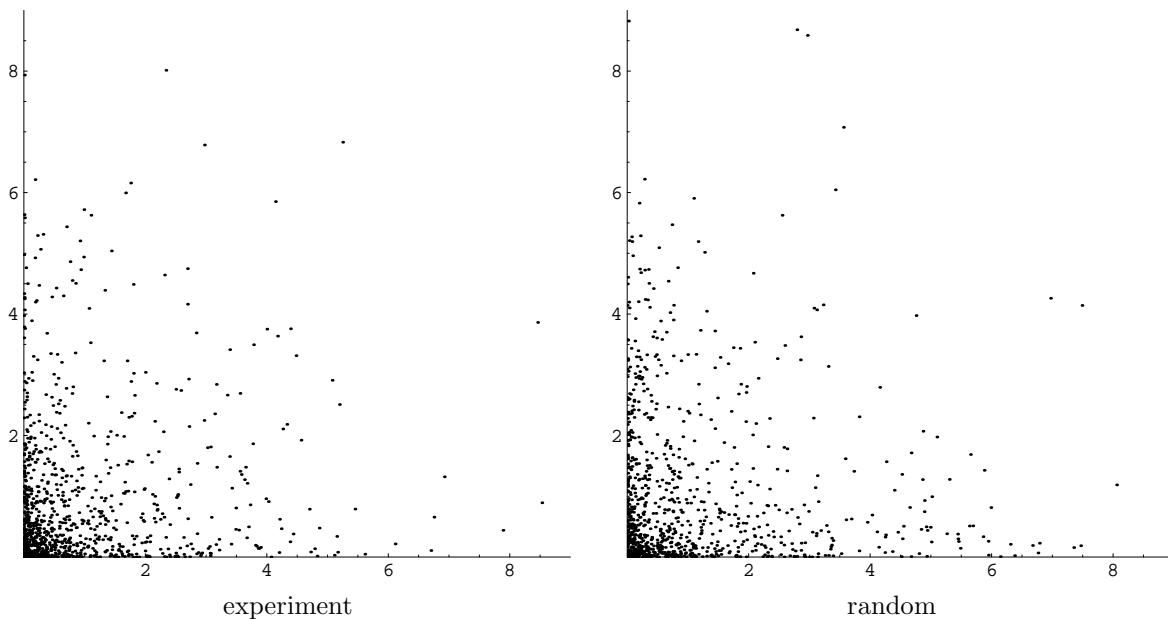


Fig. 1. Assume the AR-model  $x_n = 0.50x_{n-1} - 0.25x_{n-4} + \epsilon_n$ . In the left-hand figure (experiment) 1000 MALL-ratios multiplied by  $2N$ ; i.e.  $2N\hat{E}\{\Delta_L(\hat{\mathbf{p}}_0, \hat{\mathbf{p}}_k)\}$ , with sample size  $N = 1000$  of the correct null-hypothesis  $\mathbf{p}_0 = (a_1, 0, 0, a_4, \sigma)$  and the two alternative hypotheses  $\mathbf{p}_2 = (a_1, a_2, 0, a_4, \sigma)$  and  $\mathbf{p}_3 = (a_1, 0, a_3, a_4, \sigma)$  are plotted. In the right-hand figure 1000 random data-points (random) of two independent chi-squared distributions with one degree of freedom are plotted. Both plots are similar so we may conclude that the MALL-ratios are sufficiently independent amongst each other.

V. COMPLEXITY

The computational effort determines whether an algorithm can easily be applied or not. Our algorithm is computationally rather efficient. The computational effort will mainly be determined by the number of samples  $N$ , the number of evaluated parameters  $J$  and the actually estimated number of parameters  $M$ . The construction of the matrix  $V$ , which is initially symmetric and approximately Toeplitz costs  $\mathcal{O}\{N \times J\}$  units of time. The pivot search, which costs  $\mathcal{O}\{J\}$  time, is repeated  $M$  times and is negligible with respect to  $M$  times pivoting the matrix, which is an  $\mathcal{O}\{J^2\}$  operation. If the number of samples is large, the most costly operation is the construction of  $V$  and the complexity behaves like  $\mathcal{O}\{N \times J\}$ . So, for large  $N$  the algorithm has a linear complexity. The pivoting process, which costs  $\mathcal{O}\{M \times J^2\}$  units of time, dominates for small  $N$ .

VI. RESULTS

First we have tested the behavior of our method in case of no AR-model; i.e. the number of AR-parameters and also the model order is zero. In this case we are fitting AR-models to white noise. We expect to estimate with probability  $1 - \alpha$  the model with no AR-parameters. These results are confirmed by table II. There is a good agreement between the theoretically predicted probability (95%) and the observed probability of the model with no AR-parameters.

In table III we test our method in case of the AR-model  $x_n = 0.5x_{n-1} + 0.25x_{n-2} + 0.125x_{n-3} + 0.0625x_{n-4} + 0.03125x_{n-4} + 0.015625x_{n-6} + \epsilon_n$  where the importance of the parameters  $a_i$  decreases with  $i$ . We expect an un-

	0	1	2	3	4	5	6	7	8	corr.
$N = 100$										
2	9513	475	12							9513
4	9500	487	13	0	0					9500
6	9480	512	8	0	0	0	0			9480
8	9508	481	11	0	0	0	0	0	0	9508
10	9468	521	10	1	0	0	0	0	0	9468
$N = 1000$										
2	9487	497	16							9487
4	9500	481	19	0	0					9500
6	9488	498	14	0	0	0	0			9488
8	9465	524	11	0	0	0	0	0	0	9465
10	9395	586	19	0	0	0	0	0	0	9395
$N = 10000$										
2	9523	459	18							9523
4	9484	501	14	1	0					9484
6	9455	529	16	0	0	0	0			9455
8	9461	532	7	0	0	0	0	0	0	9461
10	9403	584	13	0	0	0	0	0	0	9403

TABLE II

THE OBSERVED NUMBER OF PARAMETER CONFIGURATIONS AS A FUNCTION OF THE NUMBER OF AR-PARAMETERS (HORIZONTAL) AND AS A FUNCTION OF THE SIZE OF THE AR-PARAMETER SPACE  $J$  (VERTICAL) GIVEN 10.000 SEQUENCES OF  $N$  OBSERVATIONS GENERATED ACCORDING TO THE AR-MODEL  $x_n = \epsilon_n$ . THE PROBABILITY ON A TOO HIGH NUMBER OF AR-PARAMETERS  $\alpha$  IS 5%. IN THE LAST COLUMN WE PRESENT THE NUMBER OF TIMES THE MODEL HAS CORRECTLY BEEN IDENTIFIED.

derestimation of the number of parameters which decreases with increasing sample size  $N$ . For  $N = 10000$  the five AR-parameter model seems to be optimal. As expected the number of estimated models with 6 parameters decreases with increasing  $J$  due to the increasing threshold.

	0	1	2	3	4	5	6	7	8	corr.
$N = 100$										
2	0	257	9743							0
4	0	197	8052	1719	32					0
6	0	263	8450	1273	14	0	0			0
8	0	340	8645	1003	12	0	0	0	0	0
10	0	379	8758	852	11	0	0	0	0	0
$N = 1000$										
2	0	0	10000							0
4	0	0	0	2572	7428					0
6	0	0	0	3004	6777	217	2			0
8	0	0	0	3737	6138	122	3	0	0	0
10	0	0	0	4245	5660	92	3	0	0	0
$N = 10000$										
2	0	0	10000							0
4	0	0	0	0	10000					0
6	0	0	0	0	256	7532	2212			2212
8	0	0	0	0	431	8523	1014	30	2	725
10	0	0	0	0	572	8685	714	28	1	431

TABLE III

AS TABLE II USING THE AR-MODEL  $x_n = 0.5x_{n-1} + 0.25x_{n-2} + 0.125x_{n-3} + 0.0625x_{n-4} + 0.03125x_{n-5} + 0.015625x_{n-6} + \epsilon_n$  AND  $\alpha = 5\%$ .

	0	1	2	3	4	5	6	7	8	corr.
$N = 100$										
2	1	9286	713							713
4	1	9448	535	16	0					235
6	2	9494	493	11	0	0	0			144
8	5	9505	479	10	1	0	0	0	0	103
10	5	9470	513	11	1	0	0	0	0	94
$N = 1000$										
2	0	6570	3430							3430
4	0	7618	2311	69	2					1869
6	0	8056	1881	62	1	0	0			1444
8	0	8274	1657	66	3	0	0	0	0	1205
10	0	8308	1628	62	2	0	0	0	0	1131
$N = 10000$										
2	0	9	9991							9991
4	0	50	9509	428	13					9473
6	0	80	9495	407	17	1	0			9449
8	0	103	9473	414	10	0	0	0	0	9421
10	0	122	9460	405	13	0	0	0	0	9407

TABLE IV

AS TABLE II USING THE AR-MODEL  $x_n = 0.55x_{n-1} + 0.05x_{n-2} + \epsilon_n$  AND  $\alpha = 5\%$ .

A model with a possibly irrelevant parameter  $a_2 = 0.05$  is presented in table IV. We expect that for small  $N$  the models with and without the parameter  $a_2$  are indistinguishable. This is confirmed by the clear preference for the one AR-parameter model for  $N \leq 1000$ . For a large sample size  $N = 10000$  we are capable of correctly identifying the correct model. Our claim, that in less than 5% of the case a model with too many parameters has been identified, is confirmed by the simulation results, although in a significant number of cases a too simple model has been found.

In table V we observe for the AR-model  $x_n = 0.75x_{n-1} - 0.50x_{n-2} + \epsilon_n$  an excellent correspondence between the theory and the simulation. Only for  $N = 100$  a significant (13%) number of models with too many parameters have

	0	1	2	3	4	5	6	7	8	corr.
$N = 100$										
2	0	8	9992							9992
4	0	29	8598	1302	71					8541
6	0	69	8551	1329	51	0	0			8480
8	0	97	8551	1303	48	1	0	0	0	8471
10	0	109	8541	1308	40	2	0	0	0	8451
$N = 1000$										
2	0	0	10000							10000
4	0	0	9537	450	13					9537
6	0	0	9560	436	4	0	0			9560
8	0	0	9567	427	6	0	0	0	0	9567
10	0	0	9541	454	5	0	0	0	0	9541
$N = 10000$										
2	0	0	10000							10000
4	0	0	9517	472	11					9517
6	0	0	9511	479	10	0	0			9511
8	0	0	9551	442	7	0	0	0	0	9551
10	0	0	9549	443	8	0	0	0	0	9549

TABLE V

AS TABLE II USING THE AR-MODEL  $x_n = 0.75x_{n-1} - 0.50x_{n-2} + \epsilon_n$  AND  $\alpha = 5\%$ .

been estimated. In about 10% of these cases the model with the non-zero parameters  $a_1, a_2$  and  $a_3$  has been identified. We also observe an increased preference for the one AR-parameter model, which is caused by the relative large threshold necessary too deal with the statistical fluctuations.

In fact the number of observations is too small to reliably distinguish between a one and a two AR-parameter model. By comparing the third column, the number of two AR-parameter models, with the last column, the number of correctly identified models, we can determine the fraction of incorrectly identified two AR-parameter models. In case of  $N = 100$  a large number of incorrect two parameter models have been estimated; this problem disappears for  $N \geq 1000$ . The results are in good agreement with our theoretical claim: in less than 5% of the cases a three or more parameter model has been estimated.

The AR-model  $x_n = 0.75x_{n-1} - 0.50x_{n-4} + \epsilon_n$  (see table VI) seems to be difficult to identify. For  $N = 10000$  in about 95% of the cases the model with the non-zero parameters  $a_1, a_4$  and  $a_5$  has been identified. Furthermore in about 3.5% of the cases the four parameter model  $a_1, a_4, a_5$  and  $a_6$  has been found. Given the number of observations these models are obviously indistinguishable from the correct model. Because there is a strong coupling between  $x_n$  and  $x_{n-1}$  it is not surprising to find a non-zero coefficient  $a_5$ , caused by the corresponding coupling between  $x_{n-4}$  and  $x_{n-5}$ . For  $J = 2$ , in the first row, no correct AR-models has been identified because the AR-order of a second order model is too low to model a fourth order signal.

We have added the model  $x_n = 0.50x_{n-1} - 0.25x_{n-4} + \epsilon_n$  of table VII because of the problems with the identification of the model in table VI. The coupling in this model is less pronounced resulting in excellent results. Our claim of less than 5% wrong identifications with irrelevant additional parameters has never been violated. For  $N \geq 1000$  all

	0	1	2	3	4	5	6	7	8	corr.
$N = 100$										
2	0	132	9868							0
4	0	0	9554	439	7					9533
6	0	0	2976	4845	2101	78	0			2947
8	0	0	2988	4840	2073	99	0	0	0	2954
10	0	0	2988	4818	2114	80	0	0	0	2966
$N = 1000$										
2	0	0	10000							0
4	0	0	9529	451	20					9529
6	0	0	783	6559	2494	160	4			783
8	0	0	789	6578	2488	142	3	0	0	789
10	0	0	794	6564	2522	117	3	0	0	794
$N = 10000$										
2	0	0	10000							0
4	0	0	9531	457	12					9531
6	0	0	0	9238	732	27	3			0
8	0	0	0	9282	692	26	0	0	0	0
10	0	0	0	9291	684	24	1	0	0	0

TABLE VI

AS TABLE II USING THE AR-MODEL  $x_n = 0.75x_{n-1} - 0.50x_{n-4} + \epsilon_n$  AND  $\alpha = 5\%$ .

	0	1	2	3	4	5	6	7	8	corr.
$N = 100$										
2	23	8747	1230							0
4	26	2737	7006	219	12					6487
6	32	3261	6460	241	6	0	0			5682
8	47	3645	6076	228	4	0	0	0	0	5293
10	45	3764	5969	218	4	0	0	0	0	5143
$N = 1000$										
2	0	3597	6403							0
4	0	0	9514	477	9					9514
6	0	0	9528	466	6	0	0			9528
8	0	0	9532	458	10	0	0	0	0	9532
10	0	0	9546	446	8	0	0	0	0	9546
$N = 10000$										
2	0	0	10000							0
4	0	0	9503	487	10					9503
6	0	0	9499	483	17	1	0			9499
8	0	0	9529	456	13	2	0	0	0	9529
10	0	0	9505	480	15	0	0	0	0	9505

TABLE VII

AS TABLE II USING THE AR-MODEL  $x_n = 0.50x_{n-1} - 0.25x_{n-4} + \epsilon_n$  AND  $\alpha = 5\%$ .

identified two AR-parameter models correspond with the correct model.

Finally we have used the model of table VII in table VIII to verify the behavior of our method for other values of  $\alpha$ . In this table we also find an excellent agreement between the simulations and the theory.

VII. CONCLUSIONS

Due to (20) the Maximum Average Log Likelihood (MALL) is a decreasing function of the variance of the residuals  $\hat{\sigma}^2$ . So a larger MALL leads to smaller residuals. Therefore the MALL is a good measure of the overall model fit. With respect to the residual variance, the *correct* model is not always the *optimal* model. It is a better strategy to estimate the AR-parameter configuration instead of the AR-order, accepting that the AR-parameters

	0	1	2	3	4	5	6	7	8	corr.
$\alpha = 1\%$										
2	0	6089	3911							0
4	0	0	9912	88	0					9912
6	0	0	9907	93	0	0	0			9907
8	0	0	9907	93	0	0	0	0	0	9907
10	0	0	9918	82	0	0	0	0	0	9918
$\alpha = 2\%$										
2	0	5108	4892							0
4	0	0	9812	187	1					9812
6	0	0	9807	191	2	0	0			9807
8	0	0	9811	187	2	0	0	0	0	9811
10	0	0	9822	178	0	0	0	0	0	9822
$\alpha = 10\%$										
2	0	2447	7553							0
4	0	0	9060	889	51					9060
6	0	0	9022	934	43	1	0			9022
8	0	0	9070	892	38	0	0	0	0	9070
10	0	0	9112	854	33	1	0	0	0	9112
$\alpha = 20\%$										
2	0	1377	8623							0
4	0	0	8118	1664	218					8118
6	0	0	8114	1669	199	17	1			8114
8	0	0	8155	1639	196	8	2	0	0	8155
10	0	0	8199	1613	175	13	0	0	0	8199

TABLE VIII

AS TABLE II USING THE AR-MODEL  $x_n = 0.50x_{n-1} - 0.25x_{n-4} + \epsilon_n$ . INSTEAD OF THE SAMPLE SIZE WE MODIFY  $\alpha$  FOR A GIVEN SAMPLE SIZE  $N = 1000$

do not form a consecutive row of non-zero parameters up to the AR-order. From a statistical point of view, parameters with a value negligible with respect to the estimation error, can better be fixated at zero. This leads to accepting some bias in the optimal model in order to decrease the estimation error.

The proposed algorithm for estimating AR-parameter configurations performs satisfactory. In most applications the claim of an a priori selected probability of estimating a model with too many parameters is fulfilled. The only theoretical weakness is the assumption of the statistical independence of the MALL-ratios (21). Although this assumption seems to be confirmed by simulations, a better theoretical foundation is necessary.

In some cases our method fails systematically (see table VI); we believe that this failure is not caused by a wrong estimation method, but has a theoretical foundation. Some models are given a certain number of observations  $N$  indistinguishable. The problem of indistinguishable models has been neglected in the literature.

APPENDIX

I. THE PI-DISTRIBUTION

Assume  $K$  independent stochastic variables distributed according to a chi-squared distribution with one degree of freedom.

$$f_{\chi^2}(x_k) = \frac{e^{-\frac{1}{2}x_k}}{\sqrt{2\pi x_k}} \tag{27}$$

We derive the distribution of the maximum of these  $K$  variables and we call this distribution the pi-distribution.

We have chosen for this name because it sounds like chi and it is derived by a product.

The cumulative distribution function of a chi-squared distribution with one degree of freedom equals:

$$F_{\chi^2}(x_k) = \text{erf}\left(\sqrt{x_k/2}\right) \quad (28)$$

The cumulative distribution function of the maximum  $y = \max(x_1, x_2, \dots, x_K)$  equals:

$$F_{\Pi}(y) = \left(\text{erf}\left(\sqrt{y/2}\right)\right)^K \quad (29)$$

This is the cumulative distribution of the pi-distribution. Differentiating the result leads to the probability density function of the pi-distribution.

$$f_{\Pi}(y) = K \frac{\left(\text{erf}\left(\sqrt{y/2}\right)\right)^{K-1}}{e^{y/2} \sqrt{2\pi y}} \quad (30)$$

The pi-distribution with  $K = 1$  corresponds with the chi-squared distribution with one degree of freedom.

REFERENCES

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

[2] R. Moddemeijer, "An ARMA model identification algorithm," in *Seventh Symposium on Information Theory in the Benelux*, D. E. Boeke, Ed., Noordwijkerhout (NL), May 22-23 1986, pp. 151–159, Werkgemeenschap Informatie- en Communicatietheorie, Enschede (NL).

[3] G. E. P. Box and G. M. Jenkins, *Time series analysis: forecasting and control*, Holden-Day series in time series analysis. Holden-Day, San Francisco, 1970.

[4] R. Moddemeijer, "Testing composite hypotheses applied to AR order estimation; the Akaike-criterion revised," Submitted to IEEE Transactions on Signal Processing, 1997.

[5] H. Cramér, *Mathematical methods of statistics*, Princeton, Princeton Univ. Press, 1945.

[6] H. L. van Trees, *Detection, estimation, and modulation theory part I*, John Wiley & Sons, Inc., New York, 1968.

[7] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proc. 2nd Int. Symp. on Information Theory*, P. N. Petrov and F. Csaki, Eds., Budapest (H), 1973, pp. 267–281, Akademia Kiado.

[8] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. on Information Theory*, vol. 19, no. 6, pp. 716–723, 1974.

[9] Y. Sakamoto, *Akaike information criterion statistics*, Reidel Publ. Comp., Dordrecht (NL), 1986.

[10] P. M. T. Broersen, "Selecting the order of autoregressive models from small samples," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 33, no. 4, pp. 874–879, 1985.

[11] D. Burshtein and E. Weinstein, "Some relations between the various criteria for autoregressive model order determination," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 33, no. 4, pp. 1017–1019, 1985.

[12] R. Shibata, "Selection of the order of an autoregressive model by Akaike's information criterion," *Biometrika*, vol. 63, no. 1, pp. 117–126, 1976.

[13] T. Westerlund, "A digital quality control system for an industrial dry process rotary cement kiln," *IEEE Trans. on Automatic Control*, vol. 26, no. 4, pp. 885–890, 1981.

[14] F. Gustafsson and H. Hjalmarsson, "Twenty-one ML estimators for model selection," *Automatica*, vol. 31, no. 10, pp. 1377–1392, 1995.

[15] C. Glymour et. al., "Statistical interference and data mining," *Communications of the ACM*, vol. 39, no. 11, pp. 35–41, 1996.

[16] H. Akaike, "Recent development of statistical methods for spectral estimation," in *Recent Advances in EEG and EMG*

*Data Processing*, N. Yamaguchi and K. Fujisawa, Eds., pp. 63–78. Elsevier, Amsterdam (NL), 1981.

[17] H. W. Steinberg et. al., "Fitting autoregressive models to EEG time series: An empirical comparison of estimates of the order," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 33, no. 1, pp. 143–150, 1985.

[18] K. W. Hipel, "Geophysical model discrimination using the Akaike information criterion," *IEEE Trans. on Automatic Control*, vol. 26, pp. 358–378, 1981.

[19] J. Rissanen, "Modelling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[20] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *J. R. Statist. Soc. Ser. B*, vol. 41, pp. 190–195, 1979.

[21] P. M. T. Broersen, "On orders for long AR models for ARMA estimation," in *The First European Conf. on Signal Analysis and Prediction*, Prague, Czech Republic, June 24-27 1997.

[22] C. M. Stein, "Multiple regression," in *Essays in Honor of Harald Hotelling*, chapter 37. Stanford Univ. Press, 1960.

[23] W. James and C. M. Stein, "Estimation with quadratic loss," in *Proc. Fourth Berkeley Symp. 1*, 1961, pp. 361–379.

[24] C. M. Stein, "Confidence sets for the mean of a multivariate distribution," *Journal of the Royal Statistical Society, Series B*, vol. 24, pp. 265–296, 1962.

[25] R. J. Bhansali, "A Monte Carlo comparison of the regression method and the spectral methods of regression," *Journal American Statistical Association*, vol. 68, no. 343, pp. 621–625, 1973.

[26] S. Brandt, *Statistical and Computational Methods in Data Analysis*, North-Holland Publ. Comp., Amsterdam (NL), 2<sup>nd</sup> edition, 1976.

[27] E. Kreyszig, *Introductory Mathematical Statistics*, John Wiley & Sons, Inc., New York, 1970.

[28] P. M. T. Broersen and H. E. Wensink, "On the penalty factor for autoregressive order selection in finite samples," *IEEE Trans. on Signal Processing*, vol. 44, no. 3, pp. 748–752, 1996.

[29] G. A. Miller, "Note on the bias of information estimates," in *Information Theory in Psychology*, H. Quastler, Ed., pp. 94–100. Glencoe, Ill., 1955.

[30] P. M. T. Broersen, "The ABC of autoregressive order selection criteria," in *11<sup>th</sup> IFAC Symp. on System Identification, SYSID '97*, Kitakyushu, Fukuoka, Japan, July 8-11 1997, vol. 1, pp. 231–236, Society of Instrument and Control Engineers (SICE).

[31] K. J. Åström, "Maximum likelihood and prediction error methods," *Automatica*, vol. 16, pp. 551–574, 1980.

[32] V. K. Rohatgi, *An introduction to probability theory and mathematical statistics*, John Wiley & Sons, Inc., New York, 1976.

[33] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *Ann. Math. Stat.*, vol. 9, pp. 60–62, 1938.