16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

AR ORDER SELECTION WITH INFORMATION THEORETIC CRITERIA BASED ON LOCALIZED ESTIMATORS

Ciprian Doru Giurcăneanu and Seyed Alireza Razavi

Department of Signal Processing,
Tampere University of Technology, Finland
ciprian.giurcaneanu@tut.fi, alireza.razavi@tut.fi

ABSTRACT

As the Information Theoretic Criteria (ITC) for AR order selection are derived under the strong hypothesis of stationarity of the measured signals, it is not straightforward to utilize them in conjunction with the forgetting factor least-squares algorithms. In the previous literature, the attempts for solving the problem were focused on the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) and the Predictive Least Squares (PLS). This study provides a variant of the Predictive Densities Criterion (PDC) that it is compatible with the forgetting factor least-squares algorithms. We also introduce a modified version of the very new Sequentially Normalized Maximum Likelihood (SNML) criterion. Additionally, we give rigorous proofs for results concerning PLS and SNML.

1. INTRODUCTION

In most of the practical applications, the coefficients of the autoregressive (AR) models are estimated by algorithms that rely mainly on the recent observations and "forget" the past. Due to their design, the estimators are dubbed *localized*, and they have been intensively researched during the last two decades in the context of adaptive control and signal processing [1, 2].

As the Information Theoretic Criteria (ITC) are derived under the strong hypothesis of stationarity of the measured signals, they cannot be utilized in conjunction with the localized estimators (LE). Therefore, it is necessary to modify the structure selection criteria. Few attempts for solving the problem are mentioned in the previous literature: the most in-depth approach is the one from [3], where the Akaike Information Criterion (AIC) was re-designed for the LE case. The celebrated Bayesian Information Criterion (BIC) [4], which is equivalent with a crude variant of the Minimum Description Length (MDL) [5] was modified in [6] such that to be compatible with LE. We note that the expressions of BIC and AIC based on LE have been already applied in on-line spectral estimation for EEG signals [7] and in tracking of the fast varying systems [8]. The reference [6] contains also some heuristics on the LE-based formulae of the Predictive Least Squares (PLS) [9] and the Predictive Minimum Description Length criteria.

The previous studies do not discuss how the Predictive Densities Criterion (PDC) can be made compatible with the LE. Note that PDC was derived in [10] by resorting to Bayesian predictive densities, and its form coincides with another criterion introduced by Rissanen in [11].

The Sequentially Normalized Maximum Likelihood (SNML) was proposed very recently as a new model selection rule [12]. The major advantage of the SNML is given by its normalizing coefficient that can be computed easier than for the ordinary NML whose evaluation for AR and ARMA models is discussed in [13].

The aim of this study is to provide versions of the PDC and SNML criteria that can be employed in combination with the forgetting factor least-squares algorithms. Additionally, we give rigorous proofs for results concerning PLS and SNML.

This work was supported by the Academy of Finland, project No. 113572, 118355 and 213462.

The rest of the paper is organized as follows. The most important ITC that have been designed for stationary AR models are briefly revisited in Section 2. The definitions and notations concerning the forgetting factor least-squares algorithms are given in Section 3. They are further used in Section 4 to modify the ITC and to investigate their main properties. The order selection performances of the modified ITC are demonstrated in Section 5 for a piecewise AR process.

2. ORDER SELECTION CRITERIA FOR STATIONARY AR MODELS

We consider the order-k AR model,

$$y_t + a_1 y_{t-1} + \cdots + a_k y_{t-k} = \varepsilon_t, \tag{1}$$

where ε_t is zero-mean white gaussian noise of variance σ^2. We employ the notation $\mathbf{a} = [a_1, \ldots, a_k]^\top$ for the coefficients of the model, and the symbol $^\top$ denotes transposition.

The available measurements are $y_1, \ldots, y_n$, and we choose an integer m such that $k < m \ll n$. Let $m' = m - (k+1)$ and $t \in \{m, \ldots, n\}$. Next we define $\bar{\mathbf{y}}_t = [y_t, \ldots, y_{m'+1}]^\top$ and $\bar{\mathbf{x}}_t = [y_{t-1}, \ldots, y_{t-k}]^\top$, with the convention that $y_i = 0$ for $i < 1$. Additionally $\mathbf{X}_t = [\bar{\mathbf{x}}_t, \ldots, \bar{\mathbf{x}}_{m'+1}]$. For all possible values of t, the number of columns of $\mathbf{X}_t$ is larger than k.

Given $y_1, \ldots, y_t$, we estimate the parameters of the AR model by minimizing the least squares criterion

$$\sum_{i=m'+1}^{t} \left(y_i + \mathbf{a}^\top \bar{\mathbf{x}}_i \right)^2, \tag{2}$$

and consequently

$$\hat{\mathbf{a}}_t = -\mathbf{V}_t \mathbf{X}_t \bar{\mathbf{y}}_t, \tag{3}$$

with $\mathbf{V}_t = \left(\mathbf{X}_t \mathbf{X}_t^\top \right)^{-1}$. Moreover, $R_t \triangleq \bar{\mathbf{y}}_t^\top \left(\mathbf{I} - \mathbf{X}_t^\top \mathbf{V}_t \mathbf{X}_t \right) \bar{\mathbf{y}}_t$, and $\mathbf{I}$ is the identity matrix. The equations above are equivalent with the *prewindow method* for $m = k+1$, and with the *covariance method* for $m = 2k+1$ [2]. We denote $c_t = \bar{\mathbf{x}}_t^\top \mathbf{V}_{t-1} \bar{\mathbf{x}}_t$, and because $\mathbf{V}_{t-1}$ is positive definite we have $c_t > 0$. Lemma 2(i) from [14] leads to the identity

$$|\mathbf{V}_t| / |\mathbf{V}_{t-1}| = 1/(1+c_t), \tag{4}$$

where the notation $|\cdot|$ is used for the matrix determinant. We utilize the following representations of the data

$$y_t + \hat{\mathbf{a}}_{t-1}^\top \bar{\mathbf{x}}_t = e_t, \tag{5}$$

$$y_t + \hat{\mathbf{a}}_t^\top \bar{\mathbf{x}}_t = \hat{e}_t. \tag{6}$$

Remark in the definitions above that R_t is the usual residual sum of squares, e_t is the forward a priori prediction error, and $\hat{e}_t$ is the forward a posteriori prediction error [2].

The well-known BIC has the expression [4],

$$\mathrm{BIC}(k) = \frac{n}{2}\ln\frac{R_n}{n} + \frac{k+1}{2}\ln n, \quad (7)$$

and PLS [9] is given by

$$\mathrm{PLS}(k) = \sum_{i=m+1}^{n} e_i^2. \quad (8)$$

Next we elaborate on the PDC formula [10] as a preparatory step for the results of the next Sections:

$$\begin{aligned}\mathrm{PDC}(k) &= -\ln\prod_{i=m+1}^{n}\left(\frac{1}{\sqrt{2\pi}}\frac{|\mathbf{V}_{i-1}^{-1}|^{1/2}}{|\mathbf{V}_{i}^{-1}|^{1/2}}\frac{\Gamma\left(\frac{i-m+1}{2}\right)}{\Gamma\left(\frac{i-m}{2}\right)}\right)\\ &\quad -\ln\prod_{i=m+1}^{n}\frac{(R_{i-1}/2)^{(i-m)/2}}{(R_i/2)^{(i-m+1)/2}} \quad (9)\\ &= -\ln\left(\frac{1}{\pi^{(n-m)/2}}\frac{\Gamma\left(\frac{n-m+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)}\frac{R_m^{1/2}}{R_n^{(n-m+1)/2}}\right)\\ &\quad +\ln\prod_{i=m+1}^{n}(1+c_i)^{1/2} \quad (10)\\ &\approx \frac{n}{2}\ln\frac{R_n}{n} + \frac{1}{2}\sum_{i=m+1}^{n}\ln(1+c_i) + \frac{1}{2}\ln n. \quad (11)\end{aligned}$$

The equation (9) is obtained by utilizing the formula (7) from [10] and by taking $m = 2k+1$. After using the identity (4) together with some simple manipulations we get (10). By applying the Stirling approximation for the *Gamma* function [15], we have

$$\ln\Gamma\left(\frac{n-m+1}{2}\right) \approx \frac{1}{2}\ln(2\pi) + \frac{n-m}{2}\ln\frac{n-m+1}{2} - \frac{n-m+1}{2}.$$

Next we drop the terms that do not depend on n, even if they depend on k. Since $m \ll n$ we employ the approximation $1/(n-m) \approx 1/n$. In (11), we neglect also the term $\frac{n}{2}\ln(2\pi\exp(1))$.

We consider the SNML formula from [12, 16], and for writing it more compactly we ignore the term $\frac{n}{2}\ln(2\pi\exp(1))$:

$$\begin{aligned}\mathrm{SNML}(k) &= \frac{n}{2}\ln\left(\frac{1}{n}\sum_{i=m+1}^{n}\hat{e}_i^2\right)\\ &\quad + \sum_{i=m+1}^{n}\ln(1+c_i) + \frac{1}{2}\ln n. \quad (12)\end{aligned}$$

The asymptotic analysis reveals the relationship between the four criteria. For example, it was shown in [6] that

$$\mathrm{PLS}(k) = R_n + \sigma^2 k \ln n\,(1+o(1)), \quad (13)$$

and the asymptotic equivalence between PLS and BIC was proven in [17]. In [16], it was verified the asymptotic equivalence between SNML and BIC, and the following limit was obtained as part of the proof: $\lim_{n\to\infty}\frac{\sum_{i=m+1}^{n}\ln(1+c_i)}{\ln n} = k$. The last result together with (11) lead to the equivalence between PDC and BIC for n large.

3. NON-STATIONARY CASE

When the hypothesis of stationarity is not verified, the loss function (2) is replaced by [6]

$$\sum_{i=1}^{t}\lambda^{t-i}\left(y_i + \mathbf{a}^\top\bar{\mathbf{x}}_i\right)^2. \quad (14)$$

The forgetting factor λ is positive and less than one, and the criterion (14) is minimized by

$$\hat{\mathbf{a}}_{\lambda,t} = -\mathbf{V}_{\lambda,t}\sum_{i=1}^{t}\lambda^{t-i}\bar{\mathbf{x}}_i y_i, \quad (15)$$

where $\mathbf{V}_{\lambda,t} = \left(\sum_{i=1}^{t}\lambda^{t-i}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^\top\right)^{-1}$.

We choose m such that the inverse $\mathbf{V}_{\lambda,t}$ exists for $t = m$, and we show that such a selection guarantees the inverse $\mathbf{V}_{\lambda,t}$ to exist for all $t \geq m$. It is useful to denote $\mathbf{A}_{\lambda,t} = \sum_{i=1}^{t}\lambda^{t-i}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^\top$, where $t \in \{m,\dots,n\}$. According to the Theorem 8.1.8 from [18], there exists $\mu \in [0,1]$ such that the smallest eigenvalue of $\mathbf{A}_{\lambda,t}$, $m < t \leq n$, can be expressed as $\lambda\ell_{\lambda,t-1} + \mu\|\bar{\mathbf{x}}_t\|^2$, where $\ell_{\lambda,t-1}$ is an eigenvalue of $\mathbf{A}_{\lambda,t-1}$. The observation $\ell_{\lambda,t-1} > 0$ concludes the proof.

For $t \in \{m+1,\dots,n\}$, we consider the data representations that are similar with (5) and (6):

$$\begin{aligned} y_t + \hat{\mathbf{a}}_{\lambda,t-1}^\top\bar{\mathbf{x}}_t &= e_{\lambda,t}, \quad (16)\\ y_t + \hat{\mathbf{a}}_{\lambda,t}^\top\bar{\mathbf{x}}_t &= \hat{e}_{\lambda,t}. \quad (17)\end{aligned}$$

Let $R_{\lambda,t}$ be the value of the loss function (14) evaluated at $\mathbf{a} = \hat{\mathbf{a}}_{\lambda,t}$. Relying on results from [2], we can easily write the formulae:

$$\begin{aligned} R_{\lambda,t} &= \lambda R_{\lambda,t-1} + e_{\lambda,t}^2/(1+c_{\lambda,t}) \quad (18)\\ &= \lambda R_{\lambda,t-1} + e_{\lambda,t}^2(1-d_{\lambda,t}), \quad (19)\\ \frac{|\mathbf{V}_{\lambda,t}|}{|\mathbf{V}_{\lambda,t-1}|} &= \frac{1}{\lambda^k(1+c_{\lambda,t})} = \frac{1-d_{\lambda,t}}{\lambda^k}, \quad (20)\end{aligned}$$

where $c_{\lambda,t} = \lambda^{-1}\bar{\mathbf{x}}_t^\top\mathbf{V}_{\lambda,t-1}\bar{\mathbf{x}}_t$ and $d_{\lambda,t} = \bar{\mathbf{x}}_t^\top\mathbf{V}_{\lambda,t}\bar{\mathbf{x}}_t$. Since $\mathbf{V}_{\lambda,t}$ is positive definite, we get for all $t \geq m$,

$$0 < d_{\lambda,t} < 1. \quad (21)$$

The ITC given in (7),(8),(11),(12) are obtained under the hypothesis that the AR coefficients are estimated with (3). In the next Section we investigate how the ITC can be re-designed to use the estimation (15) instead of (3).

4. MODIFIED INFORMATION THEORETIC CRITERIA

The traditional way of modifying BIC is to replace in (7), R_n with $R_{\lambda,n}$, and n with the *effective number of samples* $n_{ef} = \sum_{i=0}^{n-1}\lambda^i$ [1]. Because $\lim_{n\to\infty} n_{ef} = n_{ef}^\infty = 1/(1-\lambda)$, the formula employed in most of the applications is [6, 8]

$$\mathrm{BIC}_\lambda(k) = \frac{n_{ef}^\infty}{2}\ln\frac{R_{\lambda,n}}{n_{ef}^\infty} + \frac{k+1}{2}\ln n_{ef}^\infty. \quad (22)$$

In [6], the PLS criterion (8) is altered such that

$$\mathrm{PLS}_\lambda(k) = \sum_{i=m+1}^{n}\lambda^{n-i}e_{\lambda,i}^2. \quad (23)$$

To gain more insight on (23), we resort to a practice that it is common for the analysis of the adaptive algorithms, namely to examine the behavior of the estimators under the time-invariant conditions. We assume:

(A1) $y_1,\dots,y_n$ are outcomes of the gaussian stationary AR process defined in (1), for which $E[\bar{\mathbf{x}}_t\bar{\mathbf{x}}_t^\top] = \mathbf{C} > 0$.

(A2) For λ close to one and $n\to\infty$, we have:

$$\begin{aligned}\sum_{i=1}^{n}\lambda^{n-i}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^\top &\approx \mathbf{G}, \quad (24)\\ \sum_{i=1}^{n}\lambda^{n-i}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^\top\varepsilon_i^2 &\approx \mathbf{H}, \quad (25)\end{aligned}$$

where $\mathbf{G} = \frac{1}{1-\lambda}\mathbf{C}$ and $\mathbf{H} = \frac{\sigma^2}{1-\lambda}\mathbf{C}$.

The assumption *(A2)* is used frequently in the analysis of the adaptive algorithms. The interested reader can find in [1] and the references therein the conditions for which *(A2)* is verified.

Proposition 4.1. *If (A1) and (A2) are satisfied, then*

$$\mathrm{PLS}_\lambda(k) = R_{\lambda,n} + \sigma^2 k(1+O(1)). \tag{26}$$

Proof. The most important ideas of the proof are inspired by [14, 17, 19], where the analysis is restricted to the case $\lambda = 1$. The case $\lambda \in (0,1)$ poses supplementary difficulties, and we give in the Appendix A.1 the results that lead to (26). More precisely, (26) is readily obtained from Lemma A.1, Lemma A.4 and Lemma A.5. □

Remark that *(A1)* guarantees the model to be the *correct* one. The key point in practical applications is to evaluate PLS_λ for various values of k, and to choose the order that minimizes the criterion. Under mild conditions, a result similar with Proposition 4.1 can be also obtained for *incorrect* models. As the proof is lengthy, we do not include it in this note. More importantly, from the equation (13) we know that $\ln n_{ef}^\infty$ should be a factor in the penalty term of PLS_λ. Unfortunately, the second term of (26) does not contain such a factor, which prevents us to conclude that PLS_λ and BIC_λ are asymptotically equivalent. Therefore we expect PLS_λ to have modest performances. This was noticed heuristically in [6], where the following ad-hoc criterion was proposed to replace PLS_λ:

$$\mathrm{SRM}_\lambda(k) = \sum_{i=m+1}^{n} \lambda^{n-i} e_{\lambda,i}^2 + k.$$

In the reference [6], it was also coined the name SRM of this criterion.

The preparatory results (9)-(11) suggest the following version of the PDC:

$$\begin{aligned}\mathrm{PDC}_\lambda(k) &= \frac{n_{ef}^\infty}{2}\ln\frac{R_{\lambda,n}}{n_{ef}^\infty} - \ln\prod_{i=m+1}^{n}\frac{|\mathbf{V}_{i-1,\lambda}^{-1}|^{1/2}}{|\mathbf{V}_{i,\lambda}^{-1}|^{1/2}} \\ &\quad + \frac{1}{2}\ln n_{ef}^\infty \qquad (27)\\ &= \frac{n_{ef}^\infty}{2}\ln\frac{R_{\lambda,n}}{n_{ef}^\infty} + \frac{1}{2}\sum_{i=m+1}^{n}\ln\left((1+c_{\lambda,i})\lambda^k\right) \\ &\quad + \frac{1}{2}\ln n_{ef}^\infty. \qquad (28)\end{aligned}$$

The expression (28) was derived from (27) by utilizing (20). For the asymptotic analysis, it is very convenient to use (27): the penalty term is given by $\frac{1}{2}\ln\frac{|\mathbf{V}_{n,\lambda}^{-1}|}{|\mathbf{V}_{m,\lambda}^{-1}|} + \frac{1}{2}\ln n_{ef}^\infty = \frac{k+1}{2}\ln n_{ef}^\infty + \frac{1}{2}\ln\frac{|\mathbf{C}|}{|\mathbf{V}_{m,\lambda}^{-1}|}$. The last equality can be easily verified by resorting to (24), and it shows that PDC_λ and BIC_λ are equivalent for n_{ef}^∞ large.

Based on (12), it is natural to define

$$\begin{aligned}\mathrm{SNML}_\lambda(k) &= \frac{n_{ef}^\infty}{2}\ln\left(\frac{1}{n_{ef}^\infty}\sum_{i=m+1}^{n}\lambda^{n-i}\hat{e}_{\lambda,i}^2\right) \\ &\quad + \sum_{i=m+1}^{n}\ln\left((1+c_i)\lambda^k\right) + \frac{1}{2}\ln n_{ef}^\infty, \qquad (29)\end{aligned}$$

which leads to

Proposition 4.2. *If (A1) and (A2) are satisfied, then*

$$\begin{aligned}\mathrm{SNML}_\lambda(k) &= \frac{n_{ef}^\infty}{2}\ln\frac{R_{\lambda,n}}{n_{ef}^\infty} - \frac{k}{2}(1+O(1)) \\ &\quad + k\ln n_{ef}^\infty(1+O(1)) + \frac{1}{2}\ln n_{ef}^\infty. \qquad (30)\end{aligned}$$

Proof. The result is a straightforward consequence of Lemma A.6 from the Appendix A.2, and the identity $\sum_{i=m+1}^{n}\ln\left((1+c_i)\lambda^k\right) = k\ln n_{ef}^\infty + \ln\frac{|\mathbf{C}|}{|\mathbf{V}_{m,\lambda}^{-1}|}$ that we have obtained in the analysis of the penalty term for PDC_λ. □

The "big-O" terms from (30) make difficult the comparison between the asymptotic result of Proposition 4.2 and the BIC formula (22). To gain more insight, the performances of the modified ITC are compared in the next Section by resorting to simulations.

5. EXPERIMENTAL RESULTS

To illustrate the time-varying case, we consider a piecewise AR process that was also used in [6]. The number of samples is $n = 4000$, and the break points are $n_{(j)} = 1000j$ for $j \in \{1,2,3\}$. We take conventionally $n_{(0)} = 0$ and $n_{(4)} = n$. Hence the outcomes y_t of the process are given by

$$y_t + a_{j1}y_{t-1} + \cdots + a_{j,k_{(j)}}y_{t-k_{(j)}} = \varepsilon_{jt}, \tag{31}$$

where $j \in \{1,2,3\}$ and $t \in \{n_{(j-1)}+1,\ldots,n_{(j)}\}$. The AR order is $k_{(j)}$ within the j-th frame, and the noise sequence ε_{jt} is white gaussian with mean zero and unitary variance. More precisely, $k_{(1)} = 0$, $k_{(2)} = 6$, $k_{(3)} = 8$ and $k_{(4)} = 0$. The coefficients of the order-6 AR process within the second frame are [-0.4397 -0.1316 0.0905 -0.1053 -0.2814 0.5120]$^\top$, and the coefficients of the order-8 AR process within the third frame are [-0.9896 0.8097 -0.8912 0.6736 -0.7575 0.5850 -0.6077 0.5220]$^\top$. The interested reader can find in [6] the spectra for the two AR models and some details on how they have been constructed to mimic the speech spectrum.

At every sample point, the ITC must be computed for each order between $K_{min} = 0$ and $K_{max} = 15$. We choose $m = 2K_{max}$, and we resort to the fast implementation of the forgetting factor least-squares algorithm that it is based on predictive lattice filters [20]. We take $\lambda = 0.99$, which is equivalent with $n_{ef}^\infty = 100$. The value of the forgetting factor is the same as in [6]. For all the ITC that have been discussed in the previous Sections, we plot in Figure 1 and Figure 2 the percentage of correctly estimating the true order of the piecewise AR process (31). The number of runs is 5000.

The results shown in Figure 1 are in perfect agreement with those reported in [6]. The performances of PLS_λ are modest, hence the empirical evidence supports the claim of the Proposition 4.1. The capabilities of SRM_λ are superior to those of PLS_λ, but SRM_λ compares favorably with BIC_λ only in the second frame. We also noticed during the experiments that the number of correct order estimations by SRM_λ and by PLS_λ can become almost equal if the variance of the driven noise for the piecewise AR process is not unitary. Among the investigated ITC, SRM_λ is the only one affected by the variance of the driven noise.

Note in Figure 2 that SNML_λ and PDC_λ respond rapidly to an increase in order, but slowly when the order decreases. BIC_λ is very good in estimating the structure for the zero-order model.

For the Figures 3 and 4, the experimental settings are the same like in Figures 1 and 2, except the forgetting factor that is taken $\lambda = 0.995$ ($n_{ef}^\infty = 200$) instead of $\lambda = 0.99$ ($n_{ef}^\infty = 100$). Since the memory is longer than in the previous case, SNML_λ, PDC_λ and BIC_λ improve their accuracy during the frames when the model does not change, but they are less sensitive to parameter changes. This observation is in line with the principle of uncertainty [1].

6. CONCLUSION

Transforming the ITC to become compatible with the forgetting factor least-squares algorithms is not a trivial task, especially for criteria that do not involve explicitly the residual sum of weighted squares $R_{\lambda,n}$. In our study, we resorted to asymptotic analysis for decomposing each criterion into the goodness-of-fit term and the penalty term. The performances of various ITC have been illustrated by simulations with a piecewise AR model.

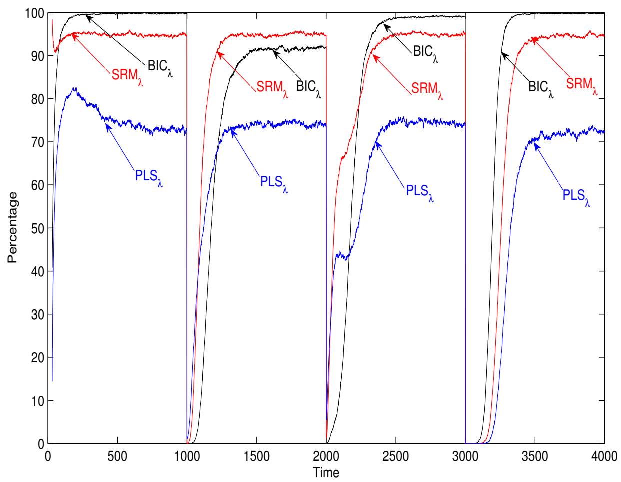

Figure 1: Percentages of the true order estimations for BIC_λ (black), PLS_λ (blue) and SRM_λ (red). The forgetting factor is $\lambda = 0.99$.

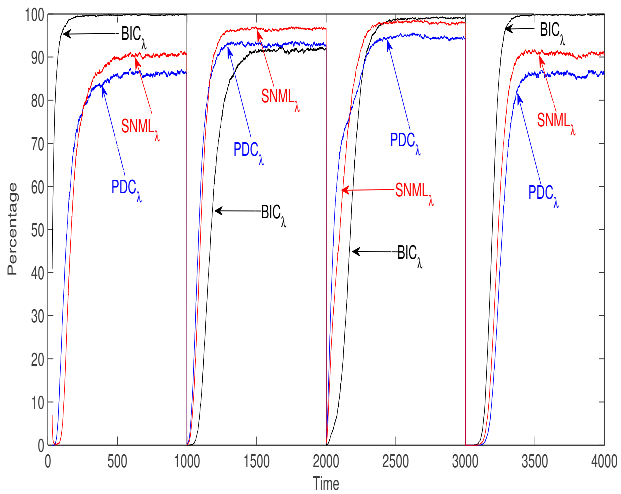

Figure 2: Percentages of the true order estimations for BIC_λ (black), SNML_λ (red) and PDC_λ (blue). The forgetting factor is $\lambda = 0.99$.

A. APPENDIX

A.1 Auxiliary results for Proposition 4.1

Lemma A.1. *The following identity is verified:*

$$PLS_\lambda(k) = \sum_{i=m+1}^{n} \lambda^{n-i} e_{\lambda,i}^2 = R_{\lambda,n} + \sum_{j=1}^{3} \mathcal{S}_j, \tag{32}$$

where $\mathcal{S}_1 \triangleq \sum_{i=m+1}^n \lambda^{n-i} d_{\lambda,i}\varepsilon_i^2$, $\mathcal{S}_2 \triangleq \sum_{i=m+1}^n \lambda^{n-i} d_{\lambda,i}\left[\left(\hat{\mathbf{a}}_{\lambda,i-1}-\mathbf{a}\right)^\top \bar{\mathbf{x}}_i\right]^2$ *and* $\mathcal{S}_3 \triangleq 2\sum_{i=m+1}^n \lambda^{n-i} d_{\lambda,i}\left[\left(\hat{\mathbf{a}}_{\lambda,i-1}-\mathbf{a}\right)^\top \bar{\mathbf{x}}_i\right]\varepsilon_i$.

Proof. For each $t \in \{m+1,\dots,n\}$, we consider the equation (19) and we multiply it by λ^{n-t}. We sum together all the resulting equalities, and the identity

$$\sum_{i=m+1}^{n} \lambda^{n-i} e_{\lambda,i}^2 = R_{\lambda,n} - \lambda^{n-m} R_{\lambda,m} + \sum_{i=m+1}^{n} \lambda^{n-i} d_{\lambda,i} e_{\lambda,i}^2 \tag{33}$$

is obtained. As $\lambda^{n-m}R_{\lambda,m} \approx 0$ asymptotically, we ignore this term from the identity above. This observation together with (1) and (16) lead to (32). □

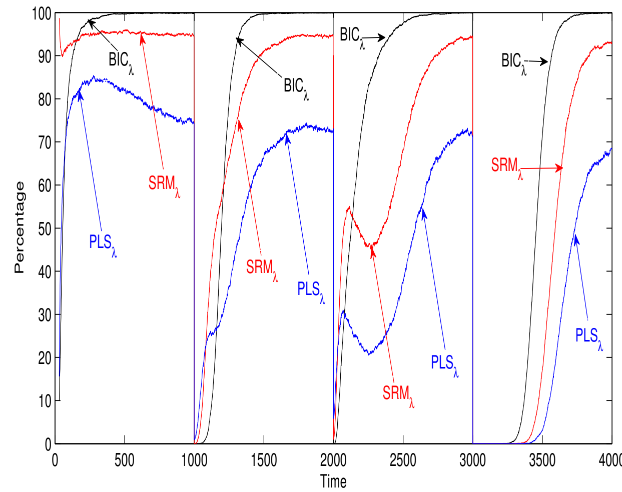

Figure 3: Percentages of the true order estimations for BIC_λ (black), PLS_λ (blue) and SRM_λ (red). The forgetting factor is $\lambda = 0.995$.

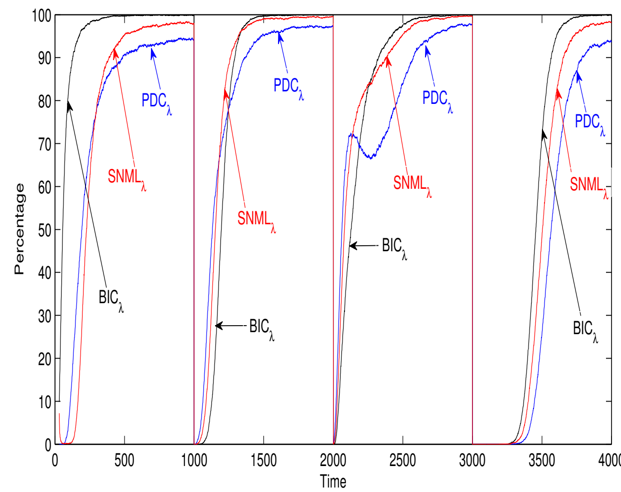

Figure 4: Percentages of the true order estimations for BIC_λ (black), SNML_λ (red) and PDC_λ (blue). The forgetting factor is $\lambda = 0.995$.

Lemma A.2. *For each $i > m$,*

$$\mathbf{V}_{\lambda,i} = \frac{1}{\lambda}\mathbf{V}_{\lambda,i-1} - \frac{1}{\lambda^2}\frac{\mathbf{V}_{\lambda,i-1}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^\top\mathbf{V}_{\lambda,i-1}}{1+c_{\lambda,i}}, \tag{34}$$

$$\bar{\mathbf{x}}_i^\top\mathbf{V}_{\lambda,i} = \frac{\bar{\mathbf{x}}_i^\top\mathbf{V}_{\lambda,i-1}}{\lambda(1+c_{\lambda,i})}. \tag{35}$$

Proof. Both identities are straightforward applications of the matrix inversion lemma [2]. □

Lemma A.3. *We have the following results:*

$$\lim_{n\to\infty}\sum_{i=m+1}^{n}\lambda^{n-i}d_{\lambda,i} < \infty, \tag{36}$$

$$\lim_{n\to\infty}\mathcal{S}_1 = \lim_{n\to\infty}\sum_{i=m+1}^{n}\lambda^{n-i}d_{\lambda,i}\varepsilon_i^2 < \infty \ a.s., \tag{37}$$

$$\lim_{n\to\infty}\sum_{i=m+1}^{n}\lambda^{n-i}\left(\lambda^{n-i}\bar{\mathbf{x}}_i^\top\mathbf{V}_{\lambda,n}\bar{\mathbf{x}}_i\right)\varepsilon_i^2 < \infty \ a.s. \tag{38}$$

Proof. Based on (21), we get immediately $\sum_{i=m+1}^n \lambda^{n-i} d_{\lambda,i} < \sum_{i=m+1}^n \lambda^{n-i} = \frac{\lambda^{m+1}-\lambda^{n+1}}{1-\lambda}$, and (36) is obtained by applying the comparison test for convergence. The result (37) is a direct consequence of (36) and Lemma 2(iii) from [14]. Equation (34) implies $\lambda^{n-i}\bar{\mathbf{x}}_i^\top \mathbf{V}_{\lambda,n}\bar{\mathbf{x}}_i \leq \bar{\mathbf{x}}_i^\top \mathbf{V}_{\lambda,i}\bar{\mathbf{x}}_i$ for all $i \in \{m+1,\ldots,n\}$, hence $\sum_{i=m+1}^{\infty}\lambda^{n-i}\left(\lambda^{n-i}\bar{\mathbf{x}}_i^\top \mathbf{V}_{\lambda,n}\bar{\mathbf{x}}_i\right)\varepsilon_i^2 \leq \sum_{i=m+1}^{\infty}\lambda^{n-i}d_{\lambda,i}\varepsilon_i^2 < \infty$ *a.s.* □

Lemma A.4. *For n large,* $\mathcal{S}_2+\mathcal{S}_3 = O(\mathcal{S}_1)$.

Proof. From (1) and (15) we have $\mathcal{S}_2 = \sum_{i=m+1}^{n}\lambda^{n-i}d_{\lambda,i}\left[\bar{\mathbf{x}}_i^\top \mathbf{V}_{\lambda,i-1}\sum_{j=1}^{i-1}\lambda^{i-1-j}\bar{\mathbf{x}}_j\varepsilon_j\right]^2$, and from (21) we get $0 \leq \mathcal{S}_2 \leq \mathcal{S}_4$, where $\mathcal{S}_4 \triangleq \sum_{i=m+1}^{n}\lambda^{n-i}\left[\bar{\mathbf{x}}_i^\top \mathbf{V}_{\lambda,i-1}\sum_{j=1}^{i-1}\lambda^{i-1-j}\bar{\mathbf{x}}_j\varepsilon_j\right]^2$. Next we consider $D_{\lambda,i} \triangleq \left[\sum_{j=1}^{i}\lambda^{i-j}\bar{\mathbf{x}}_j^\top\varepsilon_j\right]\mathbf{V}_{\lambda,i}\left[\sum_{j=1}^{i}\lambda^{i-j}\bar{\mathbf{x}}_j\varepsilon_j\right]$. Let $\lim_{n\to\infty}d_{\lambda,n} = d_\lambda$. Note that $d_\lambda \in (0,1)$ (see also (20)). After some simple manipulations that use Lemma A.2, we apply Lemma 2(iii) from [14] to obtain $(1-d_\lambda)\mathcal{S}_4(1+o(1)) = -D_{\lambda,n}+\lambda^{n-m-1}D_m+\mathcal{S}_1$. As a consequence of (38), $\lim_{n\to\infty}D_{\lambda,n} < \infty$, and the equation above leads to $\mathcal{S}_2 = O(\mathcal{S}_1)$. We utilize the Cauchy-Schwarz inequality to get $[\mathcal{S}_3/(2\mathcal{S}_1)]^2 \leq \mathcal{S}_2/\mathcal{S}_1$, which concludes the proof. □

Lemma A.5. *For n large,* $\mathcal{S}_1 = k\sigma^2+O(1)$.

Proof. (a similar reasoning can be found at p.7 in [17]) Equation (24) guarantees for any $\delta > 0$ there exists i_0 such that for all $i > i_0$ the following holds: $\|\mathbf{V}_{\lambda,i}-\mathbf{G}^{-1}\| \leq \delta/\|\mathbf{G}\|$. As $\mathbf{G}$ is positive definite, we also have for all $\bar{\mathbf{x}} \in \Re^{k\times 1}$, $\|\bar{\mathbf{x}}\|^2/\|\mathbf{G}\| \leq \bar{\mathbf{x}}^\top\mathbf{G}^{-1}\bar{\mathbf{x}}$. Note that $\|\cdot\|$ denotes the 2-norm. Simple calculations lead to $(1-\delta)\mathcal{S}_5 \leq \sum_{i=i_0}^{n}\lambda^{n-i}d_{\lambda,i}\varepsilon_i^2 \leq (1+\delta)\mathcal{S}_5$, where $\mathcal{S}_5 \triangleq \sum_{i=i_0}^{n}\lambda^{n-i}\bar{\mathbf{x}}_i^\top\mathbf{G}^{-1}\bar{\mathbf{x}}_i\varepsilon_i^2$. Tacking $\delta \to 0$, we get

$$\begin{aligned}\mathcal{S}_1 &= tr\left(\mathbf{G}^{-1}\sum_{i=m+1}^{\infty}\lambda^{n-i}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^\top\varepsilon_i^2\right)+O(1)\\ &= tr(\mathbf{G}^{-1}\mathbf{H})+O(1)\\ &= k\sigma^2+O(1),\end{aligned}$$

where $tr(\cdot)$ denotes the trace of the matrix in the argument. □

A.2 Auxiliary result for Proposition 4.2

Lemma A.6. *The following results hold:*

$$\sum_{i=m+1}^{n}\lambda^{n-i}\hat{e}_{\lambda,i}^2 = R_{\lambda,n}-\mathcal{S}_1(1+O(1)), \tag{39}$$

where $\mathcal{S}_1$ *is defined in Lemma A.1.*

$$\ln\left(\frac{1}{n_{ef}^{\infty}}\sum_{i=m+1}^{n}\lambda^{n-i}\hat{e}_{\lambda,i}^2\right) = \ln\frac{R_{\lambda,n}}{n_{ef}^{\infty}}-\frac{k}{n_{ef}^{\infty}}(1+O(1)). \tag{40}$$

Proof. Because $\hat{e}_{\lambda,i}^2 = (1-d_{\lambda,i})^2e_{\lambda,i}^2$ for all $i \in \{m+1,\ldots,n\}$ [2], the equation (33) implies $\sum_{i=m+1}^{n}\lambda^{n-i}\hat{e}_{\lambda,i}^2 = R_{\lambda,n}-\lambda^{n-m}R_{\lambda,m}-\sum_{i=m+1}^{n}\lambda^{n-i}d_{\lambda,i}e_{\lambda,i}^2+\mathcal{S}_6$, where $\mathcal{S}_6 \triangleq \sum_{i=m+1}^{n}\lambda^{n-i}d_{\lambda,i}^2e_{\lambda,i}^2$. We know from Lemma A.1 that $\sum_{i=m+1}^{n}\lambda^{n-i}d_{\lambda,i}e_{\lambda,i}^2 = \sum_{j=1}^{3}\mathcal{S}_j$, and from Lemma A.4 we have $\mathcal{S}_2+\mathcal{S}_3 = O(\mathcal{S}_1)$. The inequality (21) leads to $0 < \mathcal{S}_6 < \sum_{i=m+1}^{n}\lambda^{n-i}d_{\lambda,i}e_{\lambda,i}^2$. Additionally $\lambda^{n-m}R_{\lambda,m} \approx 0$, and the result (39) is readily obtained. Then $\ln\left(\frac{1}{n_{ef}^{\infty}}\sum_{i=m+1}^{n}\lambda^{n-i}\hat{e}_{\lambda,i}^2\right) = \ln\left(\frac{R_{\lambda,n}}{n_{ef}^{\infty}}\right) + \ln\left(1-\frac{k}{n_{ef}^{\infty}}\frac{\sigma^2}{\frac{R_{\lambda,n}}{n_{ef}^{\infty}}}(1+O(1))\right)$, which is a consequence of (39) and Lemma A.5. To get (40), we use $\frac{R_{\lambda,n}}{n_{ef}^{\infty}} \approx \sigma^2$ and $\ln(1-\xi) \approx -\xi$ for $|\xi|$ close to zero. □

REFERENCES

[1] M. Niedzwiecki, *Identification of time-varying processes*, John Wiley, 2000.

[2] S. Haykin, *Adaptive filter theory*, Prentice Hall, 1996.

[3] M. Niedzwiecki, “On the localized estimators and generalized Akaike’s criteria,” *IEEE Trans. on Automatic Control*, vol. 29, pp. 970–983, 1984.

[4] G. Schwarz, “Estimating the dimension of a model,” *Ann. Stat.*, vol. 6, pp. 461–464, 1978.

[5] J. Rissanen, “Modeling by shortest data description,” *Automatica*, vol. 14, pp. 465–471, 1978.

[6] E.J. Hannan, A.J. McDougall, and D. S. Poskitt, “Recursive estimation of autoregressions,” *J. Roy. Stat. Soc. B*, vol. 51, pp. 217–233, 1989.

[7] S. Goto, M. Nakamura, and K. Uosaki, “On-line spectral estimation of nonstationary time series based on AR model parameter estimation and order selection with a forgetting factor,” *IEEE Trans. on Signal Processing*, vol. 43, pp. 1519–1522, 1995.

[8] Y. Zheng and Z. Lin, “Recursive adaptive algorithms for fast and rapidly time-varying systems,” *IEEE Trans. on Circuits and Systems II*, vol. 50, pp. 602–614, 2003.

[9] J. Rissanen, “Order estimation by acumulated prediction errors,” *J. Appl. Prob.*, vol. 23A, pp. 55–61, 1986.

[10] P.M. Djuric and S.M. Kay, “Order selection of autoregressive models,” *IEEE Trans. on Signal Processing*, vol. 40, pp. 2829–2833, 1992.

[11] J. Rissanen, “Stochastic complexity,” *J. Roy. Stat. Soc. B*, vol. 49, pp. 252–265, 1987.

[12] J. Rissanen and T.Roos, “Conditional NML universal models,” *Proc. Information Theory and Applications Workshop, Univ. California, San Diego, USA (ITA-07)*, pp. 337–341, 29 Jan. - 2 Feb. 2007.

[13] C.D. Giurcaneanu and J. Rissanen, “Estimation of AR and ARMA models by stochastic complexity,” in *Time series and related topics*, Hwai-Chung Ho, Ching-Kang Ing, and Tze Leung Lai, Eds., vol. 52, pp. 48–59. Institute of Mathematical Statistics Lecture Notes-Monograph Series, 2006.

[14] T.L. Lai and C.Z. Wei, “Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems,” *Ann. Stat.*, vol. 10, pp. 154–166, 1982.

[15] E. Artin, *The Gamma function*, Holt, Rinehart and Winston, Inc., 1964.

[16] J. Rissanen and T.Roos, “Sequentially normalized ML universal models,” personal communication (21 pages), May 2007.

[17] C.Z. Wei, “On predictive least squares principles,” *Ann. Stat.*, vol. 20, pp. 1–42, 1982.

[18] G.H. Golub and C.F. Van Loan, *Matrix computations*, Johns Hopkins Univ. Press, 1996.

[19] C.Z. Wei, “Adaptive prediction by least squares predictors in stochastic regression models with applications to time series,” *Ann. Stat.*, pp. 1667–1682, 1987.

[20] M. Wax, “Order selection for AR models by predictive least squares,” *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 36, pp. 581–588, 1988.