Signal Processing 90 (2010) 451–466

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

AR order selection in the case when the model parameters are estimated by forgetting factor least-squares algorithms ☆

Ciprian Doru Giurcăneanu *, Seyed Alireza Razavi

Department of Signal Processing, Tampere University of Technology, P.O. Box 553, FIN-33101 Tampere, Finland

ARTICLE INFO

Article history:
Received 27 March 2009
Received in revised form
2 July 2009
Accepted 10 July 2009
Available online 17 July 2009

Keywords:
Non-stationary signals
Information theoretic criteria
Autoregressive models
Order selection
Forgetting factor least-squares algorithms

ABSTRACT

During the last decades, the use of information theoretic criteria (ITC) for selecting the order of autoregressive (AR) models has increased constantly. Because the ITC are derived under the strong assumption that the measured signals are stationary, it is not straightforward to employ them in combination 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). In connection with PLS, an ad hoc criterion called SRM was also introduced. In this paper, we modify the predictive densities criterion (PDC) and the sequentially normalized maximum likelihood (SNML) criterion such that to be compatible with the forgetting factor least-squares algorithms. Additionally, we provide rigorous proofs concerning the asymptotic approximations of four modified ITC, namely PLS, SRM, PDC and SNML. Then, the four criteria are compared by simulations with the modified variants of BIC and AIC.

1. Introduction

Autoregressive (AR) modeling is widely used for stationary time series because it yields high resolution power spectral density estimates. However, in most of the practical applications, the signals are non-stationary, and they are approximated by piecewise AR processes, which are generally called segments. The number of segments, their locations, and the AR order for each segment are not assumed to be a priori known. Therefore, finding the "best" segmentation reduces to solve an optimization problem which is difficult because the search space is huge [3]. Various sub-optimal solutions have been proposed in the past: for example, in the dyadic approach, the length of each segment is assumed to be a power of two, and the signal is divided into blocks, in a dyadic manner, up to a pre-specified scale [19]. It is beyond the scope of this paper to investigate exhaustively the vast literature on segmentation.

More importantly, the methods mentioned above have the disadvantage of a high computational burden, and they are not suitable when the spectrum of non-stationary signals must be computed on-line [9]. In such applications, the coefficients of the AR models are estimated by algorithms that use 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 [12,18].

Accuracy of spectral estimation depends crucially on the AR order selection, which makes it necessary to employ information theoretic criteria (ITC). Because the ITC are derived under the strong hypothesis that the

☆ The article extends the results of the paper "AR order selection with information theoretic criteria based on localized estimators", authored by C.D. Giurcăneanu and S.A. Razavi, and presented at EUSIPCO 2008, Lausanne, Switzerland.

* Corresponding author. Tel.: +358 3 3115 3832; fax: +358 3 3115 4989.
E-mail addresses: ciprian.giurcaneanu@tut.fi (C.D. Giurcăneanu), alireza.razavi@tut.fi (S.A. Razavi).

0165-1684/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2009.07.011

measured signals are stationary, they cannot be used in conjunction with the localized estimators. Therefore, it is mandatory to modify the ITC, and the previous literature contains few attempts: a pioneer approach is the one from [16], where the Akaike information criterion (AIC) [2] was re-designed for the case of localized estimators. The celebrated Bayesian information criterion (BIC) [27], which is equivalent with a crude variant of the minimum description length (MDL) selection rule [21] was modified in [11,17] such that to be compatible with the localized estimators. We mention that the modified expressions of BIC and AIC have been applied in on-line spectral estimation for EEG signals [9] and in tracking of the fast varying systems [32]. Ref. [11] contains some heuristics on the localized estimators-based formula for the predictive least squares (PLS) [22], and introduces also an ad hoc criterion, which is dubbed SRM (the significance of the acronym SRM is not given in [11]).

The previous studies do not discuss how the predictive densities criterion (PDC) can be made compatible with the localized estimators. Note that PDC was derived in [4] by using Bayesian predictive densities, and it is equivalent with another criterion introduced by Rissanen [23].

The sequentially normalized maximum likelihood (SNML) was proposed recently as a new model selection rule [25,26]. The major advantage of the SNML is given by its normalizing coefficient that can be computed much easier than for the ordinary NML whose evaluation for AR and autoregressive-moving-average (ARMA) models is discussed in [8]. The acronym SNLS (sequentially normalized least squares) is employed sometimes instead of SNML, but hereafter we prefer to use SNML.

The rest of the paper is organized as follows. The most important ITC designed for stationary AR models are briefly revisited in Section 2. The definitions and notations concerning the modified ITC are outlined in Section 3, where we also introduce variants of the PDC and SNML criteria that can be employed in combination with the forgetting factor least-squares algorithms. The principal result within Section 4 is Proposition 4.1, which is devoted to the asymptotic approximations for the modified ITC. Some of the asymptotic formulas have been included, without complete proofs, in [7]. In this paper, we provide rigorous proofs under two main assumptions, called ($\mathbb{A}1$) and ($\mathbb{A}2$). A novel aspect is the study on how to select the forgetting factor such that ($\mathbb{A}1$) and ($\mathbb{A}2$) are satisfied. The performance of the modified ITC is demonstrated with simulated data, in Section 5, by computing two figures of merit at each sampling point: the empirical probability of estimating the true order and the average spectrum estimation error. This extends the findings from [7], where only the first figure of merit was employed in evaluation. For the experiments in [7], the data have been produced by a piecewise AR model, which was taken from [11]. In our simulations, we use again the model from [7] for three different experimental settings, together with other two models from [3,19]: a piecewise AR process with dyadic structure and a slowly varying AR process. We do not restrain the experiments to the modified variants of BIC, PLS, PDC and SNML as it was done in [7], and we consider additionally a form of AIC which is suitable to be used in conjunction with localized estimators.

2. Order selection criteria for stationary AR models

We consider the stationary AR model with order k,

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

where ε_t is zero-mean white Gaussian noise of variance σ^2. With the convention that the symbol $(\cdot)^\top$ denotes transposition, we employ the notation $\mathbf{a} = [a_1, \ldots, a_k]^\top$ for the coefficients of the model.

Suppose the measurements $y_1, \ldots, y_n$ are available. We choose an integer m such that $k < m \ll n$. Let $m' = m - (k+1)$ and $t \in \{m, \ldots, n\}$. Then we define $\bar{\mathbf{y}}_t = [y_t, \ldots, y_{m'+1}]^\top$ and $\bar{\mathbf{x}}_t = [y_{t-1}, \ldots, y_{t-k}]^\top$, where $y_i = 0$ for $i < 1$. Additionally, we have $\mathbf{X}_t = [\bar{\mathbf{x}}_t, \ldots, \bar{\mathbf{x}}_{m'+1}]$. Remark that the number of columns of $\mathbf{X}_t$ is larger than k, for all $t \in \{m, \ldots, n\}$. It is useful to denote $\mathbf{V}_t = (\mathbf{X}_t \mathbf{X}_t^\top)^{-1}$.

Given $y_1, \ldots, y_t$, we estimate the parameters of the AR model in (1) by minimizing the least-squares criterion

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

which leads to

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

and the residual sum of squares $R_t = \bar{\mathbf{y}}_t^\top (\mathbf{I} - \mathbf{X}_t^\top \mathbf{V}_t \mathbf{X}_t) \bar{\mathbf{y}}_t$, where $\mathbf{I}$ is the identity matrix. The equations above are equivalent with the *prewindow method* for $m = k+1$, and with the *covariance method* for $m = 2k+1$ [12]. 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] implies

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

where the operator $|\cdot|$ is used for the determinant of the matrix in the argument. Based on the definitions from [12], the forward a priori prediction error is given by

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

while the forward a posteriori prediction error has the expression

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

The well-known BIC is computed by using [27]

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

and PLS [22] is evaluated with the formula

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

We elaborate on the PDC [4] as a preparatory step for the results included in the next sections:

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

Eq. (9) is obtained by using the formula (7) from [4] and by taking $m=2k+1$. The identity in (4) together with some simple manipulations yield (10). Then, we proceed like in [7] to get (11). We consider the SNML formula from [26], and we employ the approximation $(n-m)/2\approx n/2$ because $n\gg m$. After ignoring the term $(n/2)\ln(2\pi\exp(1))$, we get

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

The asymptotic analysis reveals the relationship between the four criteria. For example, it was shown in [31] that PLS and BIC are asymptotically equivalent by proving that

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

if

$$\lim_{n\to\infty}\frac{1}{n}R_n=\sigma^2.$$

A similar result was obtained previously in [11]. In [26], the asymptotic equivalence between SNML and BIC was verified, 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.

All the ITC discussed above share a common feature, namely they can be derived by applying the minimum description length (MDL) principle [24]. It is known that the MDL-based criteria are *consistent*: if the true model is among the candidates, then the probability that this model is selected goes to one as the sample size increases [10]. The fact that AIC does not have the same property was traditionally considered to be a drawback. But AIC is asymptotically *efficient*: it selects the candidate model that minimizes the one-step mean squared error of prediction [28]. We refer to [20] for a very lucid analysis of BIC and AIC.

In time-varying environments, the main concern is not the consistent estimation of the model order, but choosing the model which has the best performance in terms of prediction, spectrum estimation, or adaptive control. This is why we include in our study the Akaike criterion [2]:

$$\mathrm{AIC}(k)=\frac{n}{2}\ln\frac{R_n}{n}+k+1. \qquad (14)$$

Its expression for the time-varying case, along with the modified formulas of all other ITC that we investigate, will be given in the next section.

3. Non-stationary case

When the hypothesis of stationarity is not satisfied, the loss function (2) is replaced by [11]

$$\sum_{i=1}^{t}\lambda^{t-i}(y_i+\mathbf{a}^{\top}\bar{\mathbf{x}}_i)^2. \qquad (15)$$

The forgetting factor λ is positive and less than one, and the criterion (15) is minimized by

$$\hat{\mathbf{a}}_{\lambda,t}=-\mathbf{V}_{\lambda,t}\sum_{i=1}^{t}\lambda^{t-i}\bar{\mathbf{x}}_iy_i, \qquad (16)$$

where $\mathbf{V}_{\lambda,t}=(\sum_{i=1}^{t}\lambda^{t-i}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^{\top})^{-1}$. We choose m such that the inverse $\mathbf{V}_{\lambda,t}$ exists for $t=m$. It was proven in [7] that such a selection guarantees the inverse $\mathbf{V}_{\lambda,t}$ to exist for all $t\geq m$. Similarly with (5) and (6), we define for $t\in\{m+1,\ldots,n\}$:

$$e_{\lambda,t}=y_t+\hat{\mathbf{a}}_{\lambda,t-1}^{\top}\bar{\mathbf{x}}_t, \qquad (17)$$

$$\hat{e}_{\lambda,t}=y_t+\hat{\mathbf{a}}_{\lambda,t}^{\top}\bar{\mathbf{x}}_t. \qquad (18)$$

Let $R_{\lambda,t}$ be the value of the loss function (15) evaluated at $\mathbf{a}=\hat{\mathbf{a}}_{\lambda,t}$. Relying on results from [12], we can easily write the identities:

$$\begin{aligned}R_{\lambda,t}&=\lambda R_{\lambda,t-1}+e_{\lambda,t}^2/(1+c_{\lambda,t}) \qquad (19)\\ &=\lambda R_{\lambda,t-1}+e_{\lambda,t}^2(1-d_{\lambda,t}), \qquad (20)\end{aligned}$$

$$\frac{|\mathbf{V}_{\lambda,t}|}{|\mathbf{V}_{\lambda,t-1}|}=\frac{1}{\lambda^k(1+c_{\lambda,t})}=\frac{1-d_{\lambda,t}}{\lambda^k}, \qquad (21)$$

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

$$0<d_{\lambda,t}<1,\quad\forall t\in\{m+1,\ldots,n\}. \qquad (22)$$

The ITC given in (7)–(8), (11)–(12) and (14) are obtained under the hypothesis that the AR coefficients are estimated by (3). We show next how the ITC can be redesigned to use the estimation (16) instead of (3).

The traditional way of modifying BIC is to operate in (7) the following changes: R_n is replaced by $R_{\lambda,n}$, and n is replaced by the *effective number of samples*, $n_{\mathrm{ef}}=\sum_{i=0}^{n-1}\lambda^i$ [17]. This leads to

$$\mathrm{BIC}_{\lambda}(k)=\frac{n_{\mathrm{ef}}}{2}\ln\frac{R_{\lambda,n}}{n_{\mathrm{ef}}}+\frac{k+1}{2}\ln n_{\mathrm{ef}}. \qquad (23)$$

In [7,11], the formula above was further modified by employing instead of n_{ef} its asymptotic value, $\lim_{n\to\infty} n_{\mathrm{ef}} = n_{\mathrm{ef}}^{\infty} = 1/(1-\lambda)$. In this paper, we prefer to use the formula in (23) because it has better capabilities for estimating the structure when the sample size, n, is small.

In [11], the PLS criterion (8) was altered such that

$$\mathrm{PLS}_{\lambda}(k) = \sum_{i=m+1}^{n} \lambda^{n-i} e_{\lambda,i}^{2}, \tag{24}$$

and the following ad hoc criterion was introduced as an improvement of PLS_{λ}:

$$\mathrm{SRM}_{\lambda}(k) = \sum_{i=m+1}^{n} \lambda^{n-i} e_{\lambda,i}^{2} + k. \tag{25}$$

The preparatory results (9)–(11) suggest to modify PDC as follows:

$$\mathrm{PDC}_{\lambda}(k) = \frac{n_{\mathrm{ef}}}{2}\ln\frac{R_{\lambda,n}}{n_{\mathrm{ef}}} - \ln\prod_{i=m+1}^{n}\frac{|\mathbf{V}_{i-1,\lambda}^{-1}|^{1/2}}{|\mathbf{V}_{i,\lambda}^{-1}|^{1/2}} + \frac{1}{2}\ln n_{\mathrm{ef}} \tag{26}$$

$$= \frac{n_{\mathrm{ef}}}{2}\ln\frac{R_{\lambda,n}}{n_{\mathrm{ef}}} + \frac{1}{2}\sum_{i=m+1}^{n}\ln[(1+c_{\lambda,i})\lambda^{k}] + \frac{1}{2}\ln n_{\mathrm{ef}}. \tag{27}$$

Note that (27) was derived from (26) by applying (21). Based on (12), it is natural to define

$$\mathrm{SNML}_{\lambda}(k) = \frac{n_{\mathrm{ef}}}{2}\ln\left(\frac{1}{n_{\mathrm{ef}}}\sum_{i=m+1}^{n}\lambda^{n-i}\hat{e}_{\lambda,i}^{2}\right) + \sum_{i=m+1}^{n}\ln[(1+c_{\lambda,i})\lambda^{k}] + \frac{1}{2}\ln n_{\mathrm{ef}}. \tag{28}$$

We apply to AIC-formula in (14) the same changes that have been previously used to transform the expression of BIC from (7) to the BIC_{λ}-formula in (23), and we readily obtain

$$\mathrm{AIC}_{\lambda}(k) = \frac{n_{\mathrm{ef}}}{2}\ln\frac{R_{\lambda,n}}{n_{\mathrm{ef}}} + k + 1. \tag{29}$$

Sometimes the practitioners prefer to employ the criterion above after replacing n_{ef} by n_{ef}^{∞} (see, for example, [9]). More interestingly, Ref. [16] gives theoretical grounds for replacing in (29) the effective number of samples (n_{ef}) by the *equivalent number of samples* (n_{eq}), where $n_{\mathrm{eq}} = (\sum_{i=0}^{n-1}\lambda^{i})^{2}/(\sum_{i=0}^{n-1}\lambda^{2i})$. The effect of this change can be understood better by considering the following result from [18]: $\lim_{n\to\infty}(n_{\mathrm{eq}}/n_{\mathrm{ef}}) = 1+\lambda \approx 2$ when λ is close to one. The interested reader can find more on the significance of n_{eq} in [18, Chapter 4]. In our experiments, we prefer to use the criterion (29). The reason for our choice is twofold: (i) the comparison with BIC_{λ}-formula from (23) is fair if we employ n_{ef} and not n_{ef}^{∞}; (ii) using n_{eq} instead of n_{ef} is not a very common option for the practitioners.

In the next section, we learn more about the modified ITC by evaluating them under time-invariant conditions.

4. Analysis of the modified information theoretic criteria

To investigate the behavior of PLS_{λ}, SRM_{λ}, PDC_{λ} and SNML_{λ} under time-invariant conditions, we assume:

($\mathbb{A}1$) $y_1,\ldots,y_n$ are outcomes of the Gaussian stationary AR process defined in (1), for which $E[\bar{\mathbf{x}}_t\bar{\mathbf{x}}_t^{\top}] = \mathbf{C}$.

($\mathbb{A}2$) For λ close to one and $n\to\infty$, we have:

$$\sum_{i=1}^{n}\lambda^{n-i}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^{\top} \approx \mathbf{G}_{\lambda}, \tag{30}$$

$$\sum_{i=1}^{n}\lambda^{n-i}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^{\top}\varepsilon_i^{2} \approx \mathbf{H}_{\lambda}, \tag{31}$$

where

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

In ($\mathbb{A}1$), $E[\cdot]$ is the expectation operator, and the matrix $\mathbf{C}$ is supposed to be positive definite. Remark that ($\mathbb{A}1$) guarantees the model to be the *correct* one. To circumvent some technical difficulties, we do not consider the case of *incorrect* models. We mention for completeness that the incorrect model case was omitted also in [26].

The approximation (30) is used frequently in the analysis of the adaptive algorithms (see, for example, [18] and the references therein). Let us note that $\lim_{n\to\infty} E[\mathbf{G}_{\lambda,n}] = \mathbf{G}_{\lambda}$, where $\mathbf{G}_{\lambda,n} = \sum_{i=1}^{n}\lambda^{n-i}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^{\top}$. Relying on this property, Eleftheriou and Falconer proposed in [5] to decompose $\mathbf{G}_{\lambda,n}$ into two terms: $\mathbf{G}_{\lambda,n} = \mathbf{G}_{\lambda} + \tilde{\mathbf{G}}_{\lambda,n}$. The perturbation matrix $\tilde{\mathbf{G}}_{\lambda,n}$ is assumed to have the following properties: (i) is symmetric; (ii) its entries are zero-mean random variables and they are statistically independent from the random vector $\bar{\mathbf{x}}_n$. Hence, the approximation in (30) is equivalent to ignoring the contribution of $\tilde{\mathbf{G}}_{\lambda,n}$ when $\mathbf{G}_{\lambda,n}$ is evaluated. According to heuristics from [5], this can be done if, for $n\to\infty$, $\mathbf{G}_{\lambda,n}$ fluctuates slowly around its mean. A more solid approach is the one from [15], where the following condition is used to find out when theapproximation (30) can be applied:

$$\mathrm{tr}(E[(\mathbf{G}_{\lambda,n} - E[\mathbf{G}_{\lambda,n}])^{2}]) \ll \mathrm{tr}((E[\mathbf{G}_{\lambda,n}])^{2}) \quad \text{for } n\to\infty. \tag{32}$$

The operator $\mathrm{tr}(\cdot)$ denotes the trace of the matrix in the argument. Conventionally, the matrix $\mathbf{G}_{\lambda,n}$ is called *quasi-deterministic* whenever (32) is satisfied [15].

In Appendix A.1, we discuss how the forgetting factor λ must be selected such that the matrices $\sum_{i=1}^{n}\lambda^{n-i}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^{\top}$ and $\sum_{i=1}^{n}\lambda^{n-i}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^{\top}\varepsilon_i^{2}$ are quasi-deterministic, in the sense of the definition above. We find the condition $(1+\lambda)/(1-\lambda) \gg 8k$, and we compare it with the results from [1,5]. The derivations within Appendix A.1 give also a hint on the accuracy of the approximations (30) and (31) when λ satisfies $(1+\lambda)/(1-\lambda) \gg 8k$. It was already pointed out in [12, see p. 648] that, in the previous literature, $\sum_{i=1}^{n}\lambda^{n-i}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^{\top}$ is proven to be quasi-deterministic by assuming the vectors $\bar{\mathbf{x}}_1,\ldots,\bar{\mathbf{x}}_n$ are statistically independent. Our approach appears to be novel because we do not use such an assumption for $\bar{\mathbf{x}}_1,\ldots,\bar{\mathbf{x}}_n$.

After these preliminaries, we are prepared to formalize the principal result.

Proposition 4.1. *If* ($\mathbb{A}1$) *and* ($\mathbb{A}2$) *are satisfied, then*

$$\mathrm{PLS}_\lambda(k) = R_{\lambda,n} + O(\sigma^2 k), \tag{33}$$

$$\mathrm{SRM}_\lambda(k) = R_{\lambda,n} + k + O(\sigma^2 k), \tag{34}$$

$$\begin{aligned}\mathrm{PDC}_\lambda(k) = {}& \frac{n_{\mathrm{ef}}^\infty}{2} \ln \frac{R_{\lambda,n}}{n_{\mathrm{ef}}^\infty} + \frac{k+1}{2} \ln n_{\mathrm{ef}}^\infty \\ & + \frac{1}{2} \ln \frac{|\mathbf{C}|}{|\mathbf{V}_{m,\lambda}^{-1}|} + o(1),\end{aligned} \tag{35}$$

$$\begin{aligned}\mathrm{SNML}_\lambda(k) = {}& \frac{n_{\mathrm{ef}}^\infty}{2} \ln \frac{R_{\lambda,n}}{n_{\mathrm{ef}}^\infty} + \frac{2k+1}{2} \ln n_{\mathrm{ef}}^\infty \\ & + \ln \frac{|\mathbf{C}|}{|\mathbf{V}_{m,\lambda}^{-1}|} - O(k).\end{aligned} \tag{36}$$

The proof is deferred to Appendix A.2.

Discussion: The equality in (35) shows that PDC_λ and BIC_λ are equivalent for n_{ef}^∞ large. Because $\ln n_{\mathrm{ef}}^\infty$ does not appear explicitly as a factor in the penalty term of PLS_λ, we cannot conclude that PLS_λ and BIC_λ are asymptotically equivalent. The approximation in (33) is in line with the heuristic result given by Eq. (1.11) from [11], and it makes us to expect modest estimation performance for PLS_λ. To gain more insight, we investigate in Appendix A.3 the relationship between the asymptotic approximations (13) and (33). The "big-O" term in (36) poses troubles when one wants to check if SNML_λ is asymptotically equivalent to BIC_λ.

In the next section, the performance of the modified ITC is evaluated by simulations.

5. Experimental results

Computational aspects and performance evaluation: Assuming the observations $y_1, \ldots, y_n$ are available, the ITC whose performance we want to evaluate must be computed at each time moment for AR orders between K_{min} and K_{max}. Because in our simulations the true AR order takes values between zero and eight, we choose $K_{min} = 0$ and $K_{max} = 15$. To reduce the computational burden, we use predictive lattice filters for the implementation of the forgetting factor least-squares estimator [29]. Note that all variables involved in (23)–(25) and (27)–(29) are byproducts of the algorithm from [29]. We mention for completeness that $m = 2K_{max}$ in our settings.

When the forgetting factor is $\lambda \in (0,1)$, at each time moment $t \in \{m+1, \ldots, n\}$, the order $\hat{k} \in \{K_{min}, \ldots, K_{max}\}$ selected by an information theoretic rule, ITC_λ, is the one which minimizes the criterion. Let us assume that the data are simulated such that the true AR order at instant t is k, and let us consider N_r independent realizations of $y_1, \ldots, y_n$. Then we count in $N_c(\mathrm{ITC}_\lambda)$ how many times, after observing the first t samples, the AR order estimated by ITC_λ, $\hat{k}$, coincides with k. Hence, the empirical probability of correctly estimating the true order after observing t samples is $P_c(\mathrm{ITC}_\lambda) = N_c(\mathrm{ITC}_\lambda)/N_r$. This is the first figure of merit that we use to compare the performance of various selection criteria. Remark that $P_c(\mathrm{ITC}_\lambda)$ depends also on t, and not only on ITC_λ, but we drop t for having a simpler notation. In all the examples discussed next, we calculate the empirical probability of correctly estimating the true AR order from $N_r = 5000$ simulation runs.

If the true AR process at instant t is the one from Eq. (1), then its spectrum is given by the well-known formula [9,19],

$$\phi(f) = \frac{\sigma^2}{|1 + a_1 \exp(-i2\pi f) + \cdots + a_k \exp(-i2\pi f k)|^2}, \tag{37}$$

where $0 \le f \le 0.5$ and $i = \sqrt{-1}$. Let us suppose that, after observing the first t samples, ITC_λ selects the AR order $\hat{k}$, and the parameter estimates are $\hat{a}_{r,1}, \ldots, \hat{a}_{r,\hat{k}}, \hat{\sigma}_r^2$. Then the estimated AR spectrum is

$$\hat{\phi}_{r,\mathrm{ITC}_\lambda}(f) = \frac{\hat{\sigma}_r^2}{|1 + \hat{a}_{r,1} \exp(-i2\pi f) + \cdots + \hat{a}_{r,\hat{k}} \exp(-i2\pi f \hat{k})|^2}. \tag{38}$$

To keep the formula as simple as possible, we do not emphasize in the equation above the dependency of the parameter estimates on t, λ and $\hat{k}$. The second figure of merit that we use in the performance evaluation of the ITC is the average spectrum estimation error measured for $N_r = 5000$ independent realizations of the process,

$$\Delta_\phi(\mathrm{ITC}_\lambda) = \frac{1}{N_r} \frac{1}{M_j/2 + 1} \sum_{r=1}^{N_r} \sum_{j=0}^{M_j/2} \left[\ln \frac{\hat{\phi}_{r,\mathrm{ITC}_\lambda}(j/M_j)}{\phi(j/M_j)} \right]^2, \tag{39}$$

where $\phi(\cdot)$ and $\hat{\phi}_{r,\mathrm{ITC}_\lambda}(\cdot)$ were defined in (37) and (38), respectively. Performance measures that are similar with (39) have been employed in [3,16,17,19]. Like in [16], we take $M_j = 200$, and the reason for our choice will be evident from the description of the examples below. Remark that for the computation of $\Delta_\phi(\mathrm{ITC}_\lambda)$, it is necessary to calculate the AR model coefficients at each instant $t \in \{m+1, \ldots, n\}$. This can be done efficiently by using the lattice parameters that have been computed as part of the procedure for the selection of the optimum order $\hat{k}$ (the details of the algorithm can be found in [6,12]).

We consider three different examples to illustrate the capabilities of BIC_λ, PLS_λ, SRM_λ, PDC_λ, SNML_λ and AIC_λ.

Example 1 (*Piecewise AR process*). The following model was originally proposed in [11]:

$$y_t = \begin{cases} \varepsilon_t, & 1 \le t \le 1000, \\ 0.4397 y_{t-1} + 0.1316 y_{t-2} \\ \quad -0.0905 y_{t-3} \\ \quad +0.1053 y_{t-4} + 0.2814 y_{t-5} \\ \quad -0.5120 y_{t-6} + \varepsilon_t, & 1001 \le t \le 2000, \\ 0.9896 y_{t-1} - 0.8097 y_{t-2} \\ \quad +0.8912 y_{t-3} \\ \quad -0.6736 y_{t-4} + 0.7575 y_{t-5} \\ \quad -0.5850 y_{t-6} \\ \quad +0.6077 y_{t-7} - 0.5220 y_{t-8} + \varepsilon_t, & 2001 \le t \le 3000, \\ \varepsilon_t, & 3001 \le t \le 4000, \end{cases} \tag{40}$$

where the noise sequence $\varepsilon_1, \ldots, \varepsilon_{4000}$ is white Gaussian with zero-mean and variance $\sigma^2 = 1$. The interested reader can find in [11] the spectra for the AR models within the second and the third frame, and also some details on how they have been designed to mimic the speech spectrum.

In our experiments, the value of the forgetting factor is the same as in [11], namely $\lambda = 0.99$, which is equivalent to $n_{\text{ef}}^{\infty} = 100$. Because in the previous literature, BIC$_\lambda$ is the mostly used selection rule, we show $P_c(\text{BIC}_\lambda)$ and $\Delta_\phi(\text{BIC}_\lambda)$ in the top plots of Fig. 1. For comparison, we take BIC$_\lambda$ as a reference, and we plot in the same figure the differences $P_c(\text{ITC}_\lambda) - P_c(\text{BIC}_\lambda)$ and $\Delta_\phi(\text{ITC}_\lambda) - \Delta_\phi(\text{BIC}_\lambda)$ for all other ITC whose capabilities are evaluated in our simulations.

Remark in Fig. 1 that PLS$_\lambda$, PDC$_\lambda$ and SNML$_\lambda$ are less effective than BIC$_\lambda$ in estimating correctly the structure for the zero-order model within the first frame, especially when the number of observations is small.

For PDC$_\lambda$, this drawback can be explained by observing in (26) that the penalty term is $\frac{1}{2}\ln|\sum_{i=1}^{n}\lambda^{n-i}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^\top| - \frac{1}{2}\ln|\sum_{i=1}^{m}\lambda^{n-i}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^\top| + \frac{1}{2}\ln n_{\text{ef}}$. When $\lambda = 1$, $|\sigma^{-2}\sum_{i=1}^{n}\lambda^{n-i}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^\top|$ is the Fisher information (FI) for the AR parameters $a_1, \ldots, a_k$. It is known that $\frac{1}{2}\ln(\text{FI})$ has desirable properties and asymptotically it becomes equal to $(k/2)\ln n$, which is the penalty term for BIC. A more comprehensive discussion on this issue can be found in Section 5 of [31]. Therefore, subtracting $\frac{1}{2}\ln|\sum_{i=1}^{m}\lambda^{n-i}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^\top|$ from $\frac{1}{2}\ln(\text{FI})$ leads to a penalty term which is likely to favor the higher order models, especially when n is not much larger than m. The undesirable effects described above can be limited by using the method proposed in [4] for the initialization of ITC.

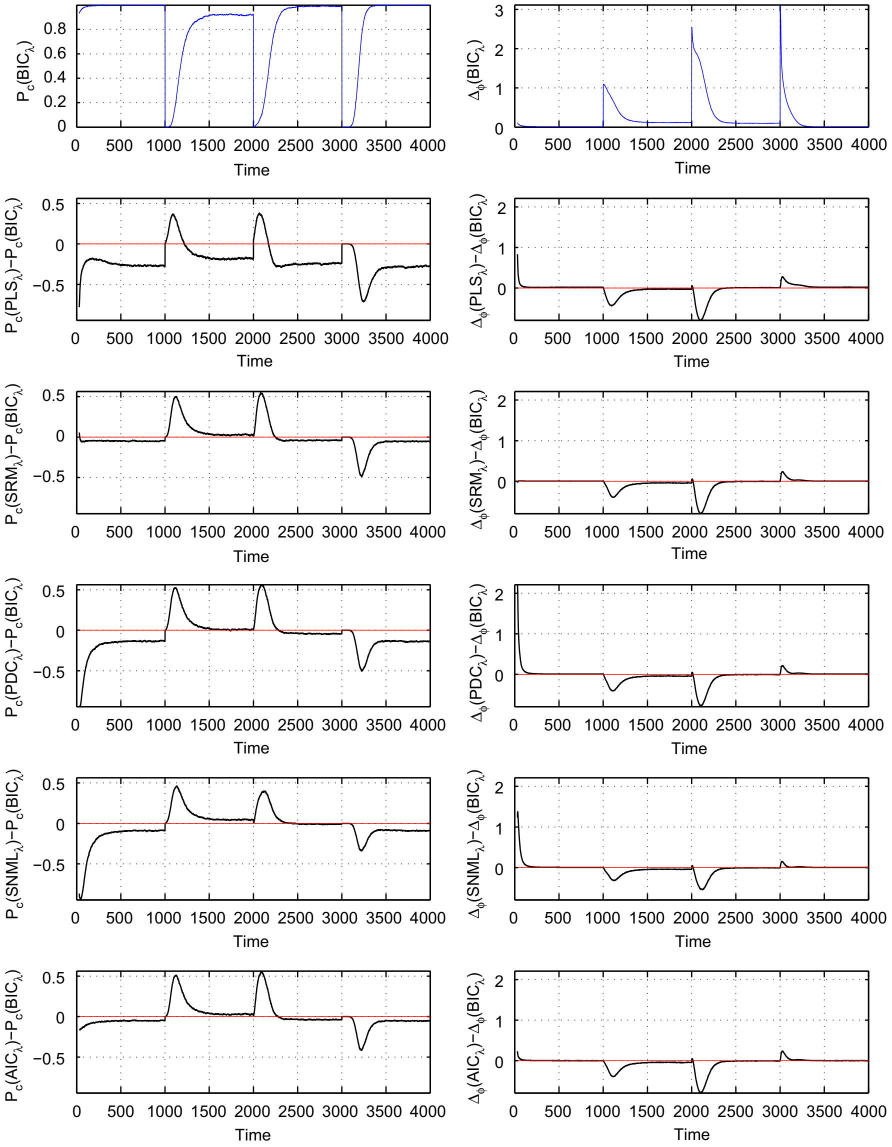

Fig. 1. Example 1—the true model is given in (40), the driven noise variance is $\sigma^2 = 1$, and the forgetting factor is $\lambda = 0.99$. The empirical probability of correctly estimating the true order, $P_c(\text{BIC}_\lambda)$, is shown in the first plot, while the second plot shows the average spectrum estimation error, $\Delta_\phi(\text{BIC}_\lambda)$. The other plots represent the differences $P_c(\text{ITC}_\lambda) - P_c(\text{BIC}_\lambda)$ and $\Delta_\phi(\text{ITC}_\lambda) - \Delta_\phi(\text{BIC}_\lambda)$ for $\text{ITC}_\lambda \in \{\text{PLS}_\lambda, \text{SRM}_\lambda, \text{PDC}_\lambda, \text{SNML}_\lambda, \text{AIC}_\lambda\}$.

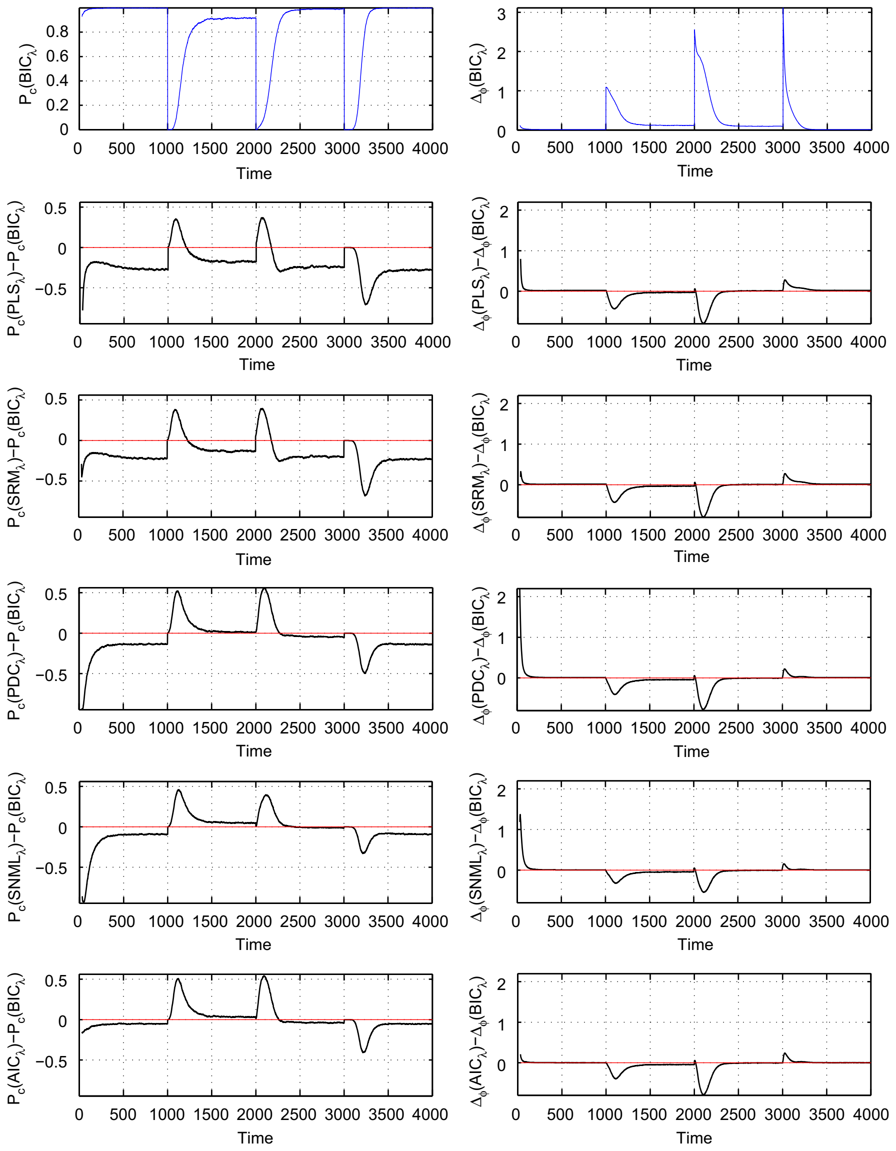

Fig. 2. Example 1—the true model is given in (40), the driven noise variance is $\sigma^2 = 10$, and the forgetting factor is $\lambda = 0.99$. All graphical conventions are the same like in Fig. 1.

The same heuristics can be extended from the case $\lambda = 1$ to $\lambda \in (0,1)$. The findings on PDC$_\lambda$ can be applied also to SNML$_\lambda$ because $\sum_{i=m+1}^{n} \ln[(1+c_{\lambda,i})\lambda^k] = \ln|\sum_{i=1}^{n} \lambda^{n-i}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^\top| - \ln|\sum_{i=1}^{m} \lambda^{n-i}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^\top|$ is part of its penalty term.

We can also remark in Fig. 1 that the performance of PLS$_\lambda$ is modest in all frames, and not only in the first one. Hence, the empirical evidence supports the asymptotic result in (33), where it is easy to see that the goodness-of-fit term dominates the penalty term. This is why PLS$_\lambda$ tends to over-estimate the model order, as was already noticed in [11]. The capabilities of SRM$_\lambda$ are superior to those of PLS$_\lambda$, but SRM$_\lambda$ compares favorably with BIC$_\lambda$ only in the second frame. Moreover, SRM$_\lambda$ performs very similarly to AIC$_\lambda$ during the first three frames, and is slightly inferior to AIC$_\lambda$ in the fourth frame.

By focusing on the breaks at the time instants 1001, 2001 and 3001, we observe in Fig. 1 that all criteria are faster than BIC$_\lambda$ when responding to an increase of the AR order, but they are slower than BIC$_\lambda$ when the AR order decreases.

To investigate further the ranking of the five selection criteria, we simulate another $N_r = 5000$ realizations of the process in (40). In this set of experiments, the variance of the driven noise is chosen to be $\sigma^2 = 10$ and not $\sigma^2 = 1$ as it was for the previous runs. The estimation results obtained for $\lambda = 0.99$ are plotted in Fig. 2, and by comparing them with those from Fig. 1, we note immediately that SRM$_\lambda$ is the only criterion affected by the increase of σ^2. More precisely, the empirical probability of correctly estimating the true order, $P_c(\text{SRM}_\lambda)$, becomes almost equal to $P_c(\text{PLS}_\lambda)$

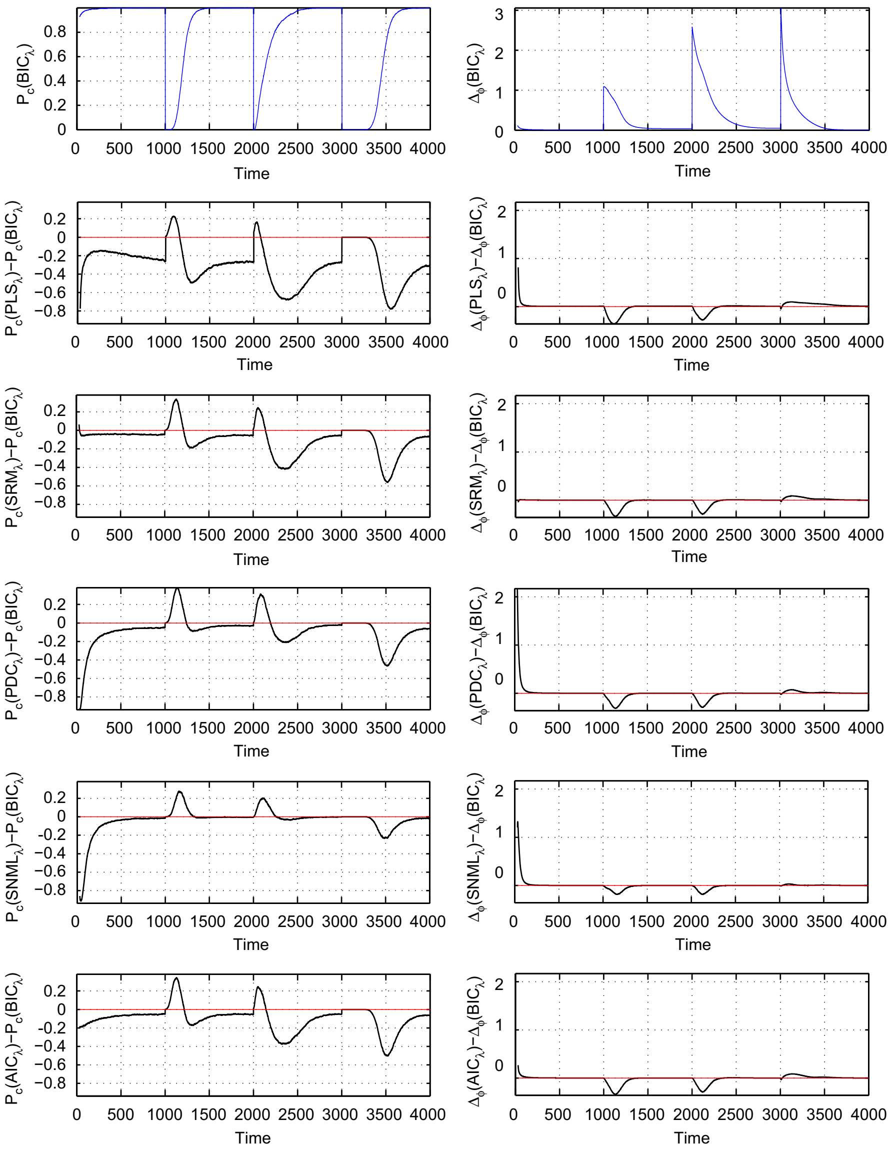

Fig. 3. Example 1—the true model is given in (40), the driven noise variance is $\sigma^2=1$, and the forgetting factor is $\lambda=0.995$. All graphical conventions are the same like in Fig. 1.

when $\sigma^2=10$. This is easily explained by the results of Proposition 4.1. The asymptotic approximations (33) and (34) lead to the following outcomes: (i) the goodness-of-fit term is the same for both PLS$_\lambda$ and SRM$_\lambda$; (ii) when $\sigma^2 \gg 1$, we have also $\sigma^2 k \gg k$, and the penalty terms of the two criteria are almost the same.

Next we analyze the influence of λ on ITC$_\lambda$. To this end, we consider the same experimental settings like those used to produce Fig. 1, except the forgetting factor that now is taken $\lambda=0.995$ ($n_{\text{ef}}^{\infty}=200$) instead of $\lambda=0.99$ ($n_{\text{ef}}^{\infty}=100$). The results are plotted in Fig. 3. Let us concentrate on the behavior of BIC$_\lambda$ in the second frame. In Fig. 1, $P_c(\text{BIC}_\lambda)<1$ for all time points between 1001 and 2000, whereas in Fig. 3, $P_c(\text{BIC}_\lambda)$ attains value one approximately at the time moment 1500 and remains at this level until the end of the frame. Then we focus on the fourth frame: in Fig. 1, $P_c(\text{BIC}_\lambda)$ equals one shortly after the time moment 3000, whereas in Fig. 3 $P_c(\text{BIC}_\lambda)$ attains the same level only after the time moment 3500. Thus, the effect of using longer memory is that BIC$_\lambda$ improves its accuracy during the frames when the model does not change, but it is less sensitive to parameter changes. The same is true for PDC$_\lambda$ and SNML$_\lambda$. This behavior is in line with the principle of uncertainty which is outlined in [18].

We can also see in Fig. 3 that PLS$_\lambda$ yields modest results, while the capabilities of SRM$_\lambda$ are moderate. Switching from $\lambda=0.99$ to 0.995 has a negative effect on the estimation accuracy of AIC$_\lambda$, which makes it to be inferior to BIC$_\lambda$ at all sampling points, except the beginning of the second and the

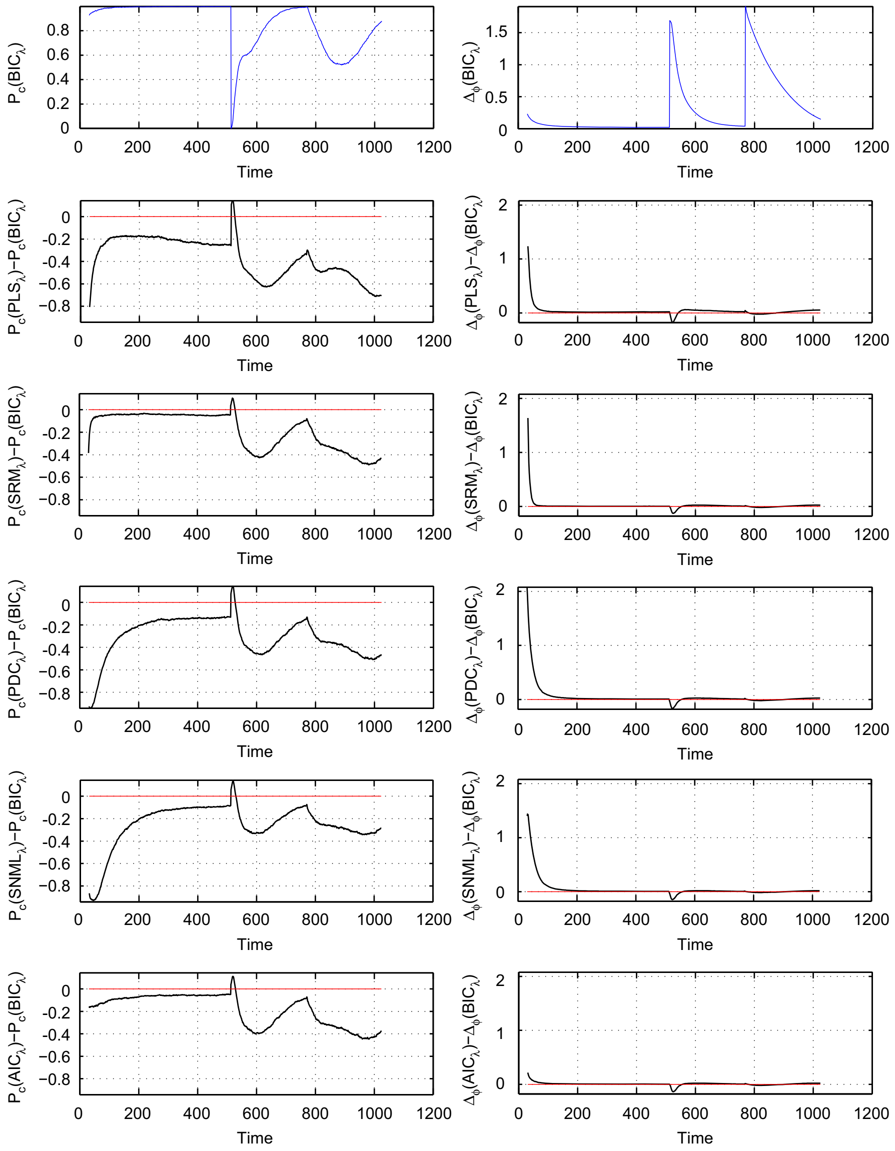

Fig. 4. Example 2—the true model is given in (41), the driven noise variance is $\sigma^2=1$, and the forgetting factor is $\lambda=0.99$. All graphical conventions are the same like in Fig. 1.

third frame. This confirms that the increase of n_{ef} affects AIC$_\lambda$ like the increase of the sample size affects AIC.

We conclude the discussion related to Example 1 by mentioning that, in Figs. 1–3, the average spectrum estimation errors are smaller for a particular ITC$_\lambda$ than for BIC$_\lambda$ only when $P_c(\text{ITC}_\lambda)>P_c(\text{BIC}_\lambda)$.

Example 2 (*Piecewise AR process with dyadic structure*). The true model for this example is taken from [3,19]:

$$y_t=\begin{cases}0.9y_{t-1}+\varepsilon_t, & 1\le t\le 512,\\ 1.69y_{t-1}-0.81y_{t-2}+\varepsilon_t, & 513\le t\le 768,\\ 1.32y_{t-1}-0.81y_{t-2}+\varepsilon_t, & 769\le t\le 1024,\end{cases}\qquad(41)$$

where ε_t is zero-mean white Gaussian noise with unitary variance. Because the lengths of all stationary frames are a power of two, the model is ideal for the segmentation algorithms that divide the time series into blocks, in a dyadic manner, up to a pre-specified scale.

For an easier comparison with the previous example, we choose the forgetting factor $\lambda=0.99$, and we plot the estimation results in Fig. 4 by using the same conventions like in Figs. 1–3. Remark in the first plot of Fig. 4 that BIC$_\lambda$ is very good in estimating the structure for the order-1 AR process within the first frame. In the same frame, SRM$_\lambda$ and AIC$_\lambda$ approach the performance of BIC$_\lambda$ much faster than PDC$_\lambda$ and SNML$_\lambda$. The difference $P_c(\text{PLS}_\lambda)-P_c(\text{BIC}_\lambda)$ is smaller than -0.2 for almost all time points within the first frame, hence the performance of PLS$_\lambda$ is modest.

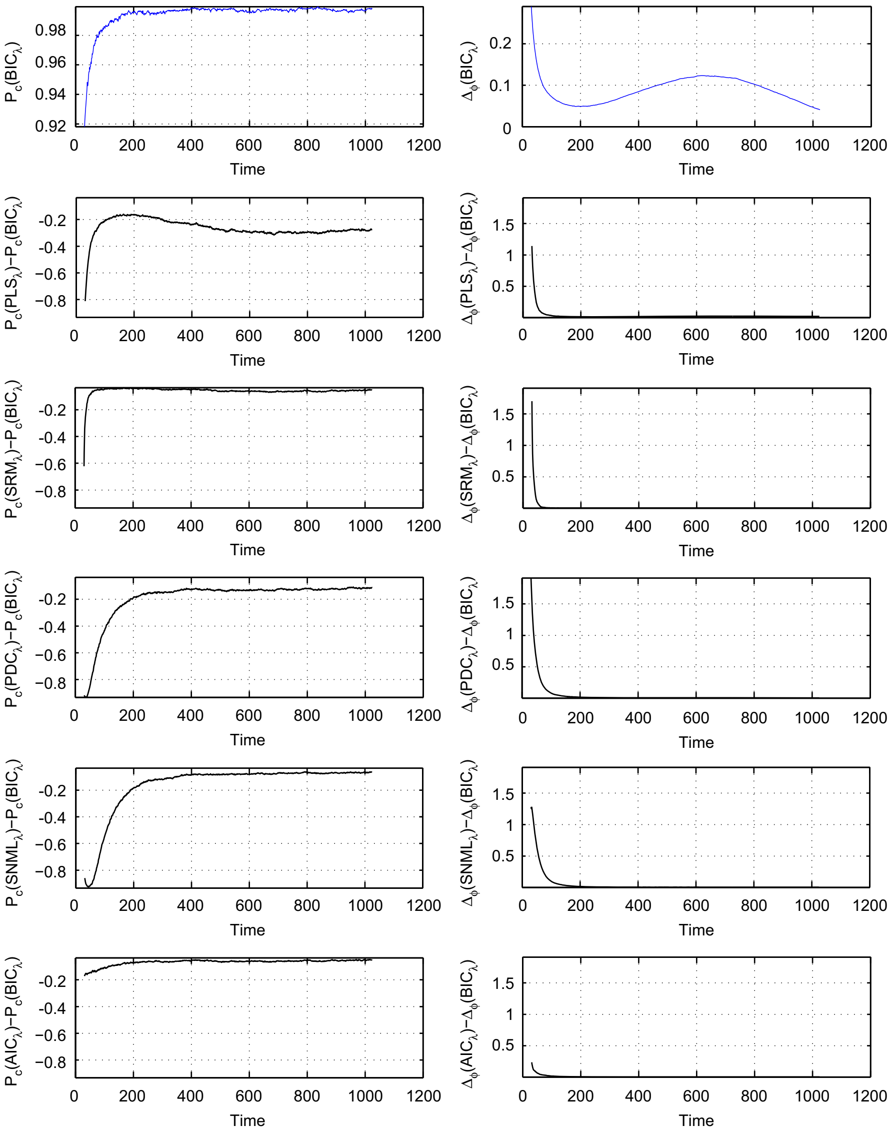

Fig. 5. Example 3—the true model is given in (42), the driven noise variance is $\sigma^2 = 1$, and the forgetting factor is $\lambda = 0.99$. All graphical conventions are the same like in Fig. 1.

The increase in AR order from the first frame to the second one is detected by all criteria, except BIC$_\lambda$. In spite of modest performance in the beginning of the second frame, BIC$_\lambda$ is the best in both the second and the third frames. For the time moments $t \geq 513$, SNML$_\lambda$ is the second-best and is followed closely by AIC$_\lambda$.

Example 3 (*Slowly varying AR process with order 2*). The true model is also taken from [3,19], and it is an order-2 AR process for which the first coefficient changes slowly over time, while the second coefficient is constant. This makes the true spectrum to change from one time point to another, as it can be seen in Fig. 5 from [3], or Fig. 12 from [19]. Thus, we simulate $N_r = 5000$ realizations of the model:

$$y_t = -a_{1,t}y_{t-1} - 0.81y_{t-2} + \varepsilon_t, \quad 1 \leq t \leq 1024, \tag{42}$$

where $a_{1,t} = -0.8[1 - 0.5\cos(\pi t/1024)]$, and ε_t is zero-mean white Gaussian noise with unitary variance.

The values of $P_c(\cdot)$ and $\Delta_\phi(\cdot)$ obtained when $\lambda = 0.99$ are plotted in Fig. 5. Similarly with what we have seen in the previous examples, the accuracy of PDC$_\lambda$ and SNML$_\lambda$ improves slower than the accuracy of the other criteria when the sample size increases. However, after the time moment 200, PLS$_\lambda$ is the only selection rule which has difficulties in correctly estimating the order of the model. It was already shown experimentally in [26] that BIC is very accurate in estimating the structure of stationary AR models whose order is at most two. Our results for Examples 2 and 3 confirm that the same is true for BIC$_\lambda$ applied to non-stationary AR models.

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. In this paper, we focused on five ITC which can be seen as embodiments of the MDL principle. Additionally, a modified variant of AIC was considered. For decomposing each MDL-based criterion into the goodness-of-fit term and the penalty term, we resorted to an asymptotic analysis.

Both the theoretical and the experimental results lead to the following outcomes: (a) BIC_λ estimates accurately the structure of non-stationary AR models whose order is at most two, which extends the similar result that was previously obtained for BIC in the stationary case; (b) PLS_λ has modest performance; (c) SRM_λ is superior to PLS_λ, but the accuracies of the two criteria become almost equal when the variance of the driven noise increases; (d) PDC_λ is faster than BIC_λ in detecting the increase of the AR order, but it is slower than BIC_λ in detecting when the AR order decreases; (e) SNML_λ and PDC_λ have similar behaviors, with the supplementary remark that SNML_λ is slightly superior to PDC_λ; (f) AIC_λ is superior to SNML_λ when the true AR order is at most two and the sample size is small, but it becomes inferior to SNML_λ when λ is chosen such that the equivalent number of samples (n_{ef}) is large enough.

The investigation can be further extended to the case of variable forgetting factor, which is known to account better for the non-stationarity of the signal. A more advanced option is to select the forgetting factor such that to minimize the description length. The difficult part is the theoretical analysis which will be more complicated than that outlined in this paper.

Acknowledgment

This work was supported by the Academy of Finland, Project nos. 113572, 118355 and 213462.

Appendix A

A.1. On the assumption ($\mathbb{A}\mathbf{2}$)

In this appendix, we investigate the conditions for which the approximations in (30) and (31) are sharp. To perform the analysis, we consider like in [16] that the covariance function of the Gaussian process defined by $\bar{\mathbf{z}}_t = [\bar{\mathbf{x}}_t^\top \ \varepsilon_t]^\top$ is exponentially decaying, or equivalently, there exist $\eta > 0$ and $\zeta \in (0,1)$ such that for all integers v, the magnitudes of the entries of $E[\bar{\mathbf{z}}_t \bar{\mathbf{z}}_{t+v}^\top]$ do not exceed $\eta\zeta^{|v|}$. For an arbitrary integer v, we denote $c(v) = E[y_t y_{t+v}]$ and $s(v) = E[\varepsilon_t y_{t+v}]$. Therefore, we have

$$|c(v)| \le \eta\zeta^{k-1+|v|} \quad \forall v \in \{\ldots,-1,0,1,\ldots\}, \tag{43}$$

$$|s(v)| \le \eta\zeta^{k+v} \quad \forall v \in \{0,1,\ldots\}. \tag{44}$$

Additionally, $s(v)=0$ for $v<0$ and $\eta \ge \sigma^2$.

We focus on the selection of the forgetting factor λ such that (32) holds true. Simple calculations lead to

$$E[(\mathbf{G}_{\lambda,n} - E[\mathbf{G}_{\lambda,n}])^2] = \left(\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda^{2n-i-j} E[\bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^\top \bar{\mathbf{x}}_j \bar{\mathbf{x}}_j^\top]\right) - \left(\sum_{i=1}^{n} \lambda^{n-i}\right)^2 \mathbf{C}^2, \tag{45}$$

and by applying the well-known formula (see, for example, [13]) that gives the expectation of the product of four jointly Gaussian random variables in terms of their first- and second-order moments, we obtain for all $i,j \in \{1,\ldots,n\}$:

$$\begin{aligned}
&\operatorname{tr}(E[\bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^\top \bar{\mathbf{x}}_j \bar{\mathbf{x}}_j^\top]) \\
&= \sum_{p=1}^{k}\sum_{q=1}^{k} E[y_{i-p}y_{i-q}y_{j-q}y_{j-p}] = \sum_{p=1}^{k}\sum_{q=1}^{k} c(|p-q|)^2 && (46)\\
&\quad + \sum_{p=1}^{k}\sum_{q=1}^{k} c(|i-j-(p-q)|)c(|i-j+(p-q)|) && (47)\\
&\quad + \sum_{p=1}^{k}\sum_{q=1}^{k} c(|i-j|)^2 && (48)\\
&= kc(0)^2 + 2\sum_{r=1}^{k-1}(k-r)c(r)^2 && (49)\\
&\quad + kc(|i-j|)^2 + 2\sum_{r=1}^{k-1}(k-r)c(|i-j-r|)c(|i-j+r|) && (50)\\
&\quad + k^2 c(|i-j|)^2. && (51)
\end{aligned}$$

We note that Eqs. (46)–(48) are a particular case of Eqs. (2.8)–(2.12) from [13]. Remark also that the matrix $\mathbf{C}$ is Toeplitz and symmetric, thus it is fully defined by its first row $[c(0)\ c(1)\cdots c(k-1)]$. By using (45) and (49)–(51), we get

$$\begin{aligned}
&\operatorname{tr}(E[(\mathbf{G}_{\lambda,n} - E[\mathbf{G}_{\lambda,n}])^2]) \\
&= \left(\sum_{i=1}^{n}\lambda^{n-i}\right)^2 \left(kc(0)^2 + 2\sum_{r=1}^{k-1}(k-r)c(r)^2\right) \\
&\quad + \sum_{i=1}^{n}\lambda^{2n-2i}\left(kc(0)^2 + 2\sum_{r=1}^{k-1}(k-r)c(r)^2\right) \\
&\quad +2\sum_{\ell=1}^{n-1}\sum_{i=\ell+1}^{n}\lambda^{2n-2i+\ell} \\
&\quad \times\left(kc(\ell)^2 + 2\sum_{r=1}^{k-1}(k-r)c(|\ell-r|)c(|\ell+r|)\right) \\
&\quad + \sum_{i=1}^{n}\lambda^{2n-2i}k^2c(0)^2 \\
&\quad +2\sum_{\ell=1}^{n-1}\sum_{i=\ell+1}^{n}\lambda^{2n-2i+\ell}k^2c(\ell)^2 \\
&\quad -\left(\sum_{i=1}^{n}\lambda^{n-i}\right)^2\left(kc(0)^2 + 2\sum_{r=1}^{k-1}(k-r)c(r)^2\right) \\
&= c(0)^2\left(\frac{1-\lambda^{2n}}{1-\lambda^2}(k^2+k)\right) && (52)\\
&\quad + \sum_{r=1}^{k-1}c(r)^2\left(\frac{1-\lambda^{2n}}{1-\lambda^2}2(k-r) + \frac{\lambda^r-\lambda^{2n-r}}{1-\lambda^2}2(k^2+k)\right) && (53)\\
&\quad + \sum_{\ell=k}^{n-1}c(\ell)^2\left(\frac{\lambda^\ell-\lambda^{2n-\ell}}{1-\lambda^2}2(k^2+k)\right) && (54)\\
&\quad + 4\sum_{\ell=1}^{n-1}\sum_{i=\ell+1}^{n}\lambda^{2n-2i+\ell}\sum_{r=1}^{k-1}(k-r)c(|\ell-r|)c(|\ell+r|). && (55)
\end{aligned}$$

We apply (43) to the term in (54),

$$\begin{aligned}\sum_{\ell=k}^{n-1} c(\ell)^2 \left(\frac{\lambda^{\ell}-\lambda^{2n-\ell}}{1-\lambda^2} 2(k^2+k)\right) &\leq \frac{2\eta^2\zeta^{2k-2}(k^2+k)}{1-\lambda^2}\sum_{\ell=k}^{n-1}(\lambda^{\ell}-\lambda^{2n-\ell})\zeta^{2\ell} \\ &= \frac{2\eta^2\zeta^{2k-2}(k^2+k)}{1-\lambda^2}\left(\frac{\lambda^k\zeta^{2k}-\lambda^n\zeta^{2n}}{1-\lambda\zeta^2}-\frac{\lambda^{2n-k}\zeta^{2k}-\lambda^n\zeta^{2n}}{1-\lambda^{-1}\zeta^2}\right),\end{aligned}$$

and for $n \to \infty$, we get

$$\sum_{\ell=k}^{n-1} c(\ell)^2 \left(\frac{\lambda^{\ell}-\lambda^{2n-\ell}}{1-\lambda^2} 2(k^2+k)\right) \leq 2\frac{k^2+k}{1-\lambda^2}\frac{\eta^2}{1-\lambda\zeta^2}\lambda^k\zeta^{4k-2}. \tag{56}$$

Next we consider (43) and the term in (55):

$$\begin{aligned}&4\sum_{\ell=1}^{n-1}\sum_{i=\ell+1}^{n}\lambda^{2n-2i+\ell}\sum_{r=1}^{k-1}(k-r)c(|\ell-r|)c(|\ell+r|) \\ &\leq 4\sum_{\ell=1}^{n-1}\sum_{i=\ell+1}^{n}\lambda^{2n-2i+\ell}\sum_{r=1}^{k-1}(k-r)\eta^2\zeta^{2k-2}\zeta^{|\ell-r|+|\ell+r|} \\ &\leq 4\eta^2\zeta^{2k-2}\sum_{\ell=1}^{n-1}\sum_{i=\ell+1}^{n}\lambda^{2n-2i+\ell}\sum_{r=1}^{k-1}(k-r)\zeta^{2\ell} \qquad (57) \\ &= 4\eta^2\zeta^{2k-2}\sum_{\ell=1}^{n-1}\sum_{i=\ell+1}^{n}\lambda^{2n-2i+\ell}\zeta^{2\ell}\sum_{r=1}^{k-1}(k-r) \\ &= \frac{2\eta^2\zeta^{2k-2}(k^2-k)}{1-\lambda^2}\sum_{\ell=1}^{n-1}(\lambda^{\ell}-\lambda^{2n-\ell})\zeta^{2\ell} \\ &= \frac{2\eta^2\zeta^{2k-2}(k^2-k)}{1-\lambda^2}\left(\frac{\lambda\zeta^2-\lambda^n\zeta^{2n}}{1-\lambda\zeta^2}-\frac{\lambda^{2n-1}\zeta^2-\lambda^n\zeta^{2n}}{1-\lambda^{-1}\zeta^2}\right). \qquad (58)\end{aligned}$$

All the derivations above are straightforward. The inequality in (57) is a consequence of the following result: for all $\ell \in \{1,\dots,n-1\}$ and $r \in \{1,\dots,k-1\}$, $2\ell \leq |\ell-r|+|\ell+r|$, with equality if and only if $r \leq \ell$. By taking $n \to \infty$ in (58), we obtain

$$\begin{aligned}&4\sum_{\ell=1}^{n-1}\sum_{i=\ell+1}^{n}\lambda^{2n-2i+\ell}\sum_{r=1}^{k-1}(k-r)c(|\ell-r|)c(|\ell+r|) \\ &\leq 2\frac{k^2-k}{1-\lambda^2}\frac{\eta^2}{1-\lambda\zeta^2}\lambda\zeta^{2k}. \qquad (59)\end{aligned}$$

By collecting the results from (52), (53), (56) and (59), the following asymptotic inequality is proven:

$$\begin{aligned}&\frac{(1-\lambda)^2}{k}\mathrm{tr}(E[(\mathbf{G}_{\lambda,n}-E[\mathbf{G}_{\lambda,n}])^2]) \\ &\leq \left(2\frac{1-\lambda}{1+\lambda}k\right)c(0)^2\frac{k+1}{2k} \qquad (60) \\ &+\left(2\frac{1-\lambda}{1+\lambda}k\right)\sum_{r=1}^{k-1}c(r)^2\left(\frac{k-r}{k^2}+\frac{k+1}{k}\lambda^r\right) \qquad (61) \\ &+\left(2\frac{1-\lambda}{1+\lambda}k\right)\left(\frac{k+1}{k}\frac{\eta^2}{1/(\lambda\zeta^2)-1}(\lambda\zeta^4)^{k-1}\right) \qquad (62) \\ &+\left(2\frac{1-\lambda}{1+\lambda}k\right)\left(\frac{k-1}{k}\frac{\eta^2}{1/(\lambda\zeta^2)-1}\zeta^{2(k-1)}\right). \qquad (63)\end{aligned}$$

Moreover, we have the identity:

$$\frac{(1-\lambda)^2}{k}\mathrm{tr}((E[\mathbf{G}_{\lambda,n}])^2) = c(0)^2 + \sum_{r=1}^{k-1}c(r)^2\frac{2(k-r)}{k}. \tag{64}$$

Because the coefficient of $c(0)^2$ in (64) is one, we have forced in (60)–(63) the common factor $2((1-\lambda)/(1+\lambda))k$ such that the coefficient of $c(0)^2$ in (60), namely $(k+1)/(2k)$, does not exceed one; it is obvious that $(k+1)/(2k) \leq 1$ for $k \geq 1$. By comparing (60)–(63) with (64), we remark that (32) holds under the condition $2((1-\lambda)/(1+\lambda))k \ll 1$ if the covariance function of the AR process is rapidly decreasing. To be more precise, let us compare the coefficient of $c(r)^2$ from (61) with the one from (64). It is evident that $k > 1$ and $r \in \{1,\dots,k-1\}$. Because λ is close to one, the effect of λ^r is marginal, and we ignore it. Then, it is elementary to observe that

$$\frac{2(k-r)}{k}-\left(\frac{k-r}{k^2}+\frac{k+1}{k}\right) \geq -1+\left(\frac{1}{k}-\frac{1}{k^2}\right) > -1.$$

Hence, for $2((1-\lambda)/(1+\lambda))k \ll 1$ we have

$$\left(2\frac{1-\lambda}{1+\lambda}k\right)\left(\frac{k-r}{k^2}+\frac{k+1}{k}\lambda^r\right) \ll \frac{2(k-r)}{k}.$$

In (62) and (63), the smaller is ζ, the smaller are the factors

$$\left(\frac{k+1}{k}\frac{\eta^2}{1/(\lambda\zeta^2)-1}(\lambda\zeta^4)^{k-1}\right)$$

and

$$\left(\frac{k-1}{k}\frac{\eta^2}{1/(\lambda\zeta^2)-1}\zeta^{2(k-1)}\right).$$

Therefore, the approximation in (30) becomes sharper when the covariance function of the AR process decreases rapidly. We note also that, because λ is close to one, we have $2/(1+\lambda) \approx 1$. Thus, $2((1-\lambda)/(1+\lambda))k \ll 1$ is almost the same like the usual condition $1/(1-\lambda) \gg k$ [1,5].

Next we focus on (31) and we find conditions for λ such that the matrix $\mathbf{H}_{\lambda,n} = \sum_{i=1}^{n}\lambda^{n-i}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^{\top}\varepsilon_i^2$ is quasi-deterministic at steady-state. Remark that

$$\begin{aligned}&E[(\mathbf{H}_{\lambda,n}-E[\mathbf{H}_{\lambda,n}])^2] \\ &\quad= \left(\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda^{2n-i-j}E[\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^{\top}\varepsilon_i^2\bar{\mathbf{x}}_j\bar{\mathbf{x}}_j^{\top}\varepsilon_j^2]\right)-\sigma^4\left(\sum_{i=1}^{n}\lambda^{n-i}\right)^2\mathbf{C}^2,\end{aligned}$$

$$\begin{aligned}&\mathrm{tr}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda^{2n-i-j}E[\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^{\top}\varepsilon_i^2\bar{\mathbf{x}}_j\bar{\mathbf{x}}_j^{\top}\varepsilon_j^2]\right) \\ &\quad= \left[\sum_{\substack{i,j=1\\ i=j}}^{n}+\sum_{\substack{i,j=1\\ i\neq j}}^{n}\right]\times[\lambda^{2n-i-j}\mathrm{tr}(E[\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^{\top}\varepsilon_i^2\bar{\mathbf{x}}_j\bar{\mathbf{x}}_j^{\top}\varepsilon_j^2])].\end{aligned}$$

With some simple calculations, we obtain

$$\begin{aligned}
&\sum_{\substack{i,j=1\\ i=j}}^{n} \lambda^{2n-i-j}\mathrm{tr}(E[\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^\top \varepsilon_i^2\bar{\mathbf{x}}_j\bar{\mathbf{x}}_j^\top \varepsilon_j^2])\\
&\quad=\sum_{i=1}^{n}\lambda^{2n-2i}E[\mathrm{tr}((\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^\top \varepsilon_i)^2)]\\
&\quad=\sum_{i=1}^{n}\lambda^{2n-2i}E\left[\sum_{p=1}^{k}\sum_{q=1}^{k}y_{i-p}^2y_{i-q}^2\varepsilon_i^4\right]\\
&\quad=\sum_{i=1}^{n}\lambda^{2n-2i}E[\varepsilon_i^4]\sum_{p=1}^{k}\sum_{q=1}^{k}E[y_{i-p}^2y_{i-q}^2]\\
&\quad=3\sigma^4\sum_{i=1}^{n}\lambda^{2n-2i}\sum_{p=1}^{k}\sum_{q=1}^{k}(c(0)^2+2c(|p-q|)^2)\\
&\quad=3\sigma^4\frac{1-\lambda^{2n}}{1-\lambda^2}\left(k^2c(0)^2+2\left(kc(0)^2+2\sum_{r=1}^{k-1}(k-r)c(r)^2\right)\right)\\
&\quad=3\sigma^4\frac{1-\lambda^{2n}}{1-\lambda^2}\left((k^2+2k)c(0)^2+4\sum_{r=1}^{k-1}(k-r)c(r)^2\right).
\end{aligned}\tag{65}$$

For $i\neq j$ and $p,q\in\{1,\ldots,k\}$, the expectation formula for the product of Gaussian random variables [16] leads to

$$\begin{aligned}
&E[y_{i-p}y_{i-q}\varepsilon_i^2y_{j-q}y_{j-p}\varepsilon_j^2]\\
&\quad=\sigma^4c(|p-q|)^2\\
&\qquad+\sigma^4c(|i-j-(p-q)|)c(|i-j+(p-q)|)\\
&\qquad+\sigma^4c(|i-j|)^2+2\sigma^2\\
&\qquad\times\underbrace{c(|p-q|)s(|i-j|-p)s(|i-j|-q)U(|i-j|-\max(p,q))}_{T(i,j,p,q)},
\end{aligned}$$

where $U(\cdot)$ has value one whenever the argument is nonnegative, and otherwise it takes value zero. By using the equation above together with (65) and some previous results, we get

$$\begin{aligned}
&\mathrm{tr}(E[(\mathbf{H}_{\lambda,n}-E[\mathbf{H}_{\lambda,n}])^2])\\
&\quad=3\sigma^4\frac{1-\lambda^{2n}}{1-\lambda^2}\left((k^2+2k)c(0)^2+4\sum_{r=1}^{k-1}(k-r)c(r)^2\right)\\
&\qquad+\sigma^4\left(\left(\sum_{i=1}^{n}\lambda^{n-i}\right)^2-\frac{1-\lambda^{2n}}{1-\lambda^2}\right)\\
&\qquad\times\left(kc(0)^2+2\sum_{r=1}^{k-1}(k-r)c(r)^2\right)\\
&\qquad+2\sigma^4\sum_{\ell=1}^{n-1}\sum_{i=\ell+1}^{n}\lambda^{2n-2i+\ell}\\
&\qquad\times\left(kc(\ell)^2+2\sum_{r=1}^{k-1}(k-r)c(|\ell-r|)c(|\ell+r|)\right)\\
&\qquad+2\sigma^4\sum_{\ell=1}^{n-1}\sum_{i=\ell+1}^{n}\lambda^{2n-2i+\ell}k^2c(\ell)^2\\
&\qquad+4\sigma^2\sum_{\ell=1}^{n-1}\sum_{i=\ell+1}^{n}\lambda^{2n-2i+\ell}\sum_{p=1}^{k}\sum_{q=1}^{k}T(i,i-\ell,p,q)\\
&\qquad-\sigma^4\left(\sum_{i=1}^{n}\lambda^{n-i}\right)^2\left(kc(0)^2+2\sum_{r=1}^{k-1}(k-r)c(r)^2\right).
\end{aligned}$$

From (43) and (44), we have $T(i,i-\ell,p,q)\le\eta^3\zeta^{3k-1}\zeta^{2\ell}\zeta^{|p-q|-(p+q)}$, thus

$$\begin{aligned}
&4\sigma^2\sum_{\ell=1}^{n-1}\sum_{i=\ell+1}^{n}\lambda^{2n-2i+\ell}\sum_{p=1}^{k}\sum_{q=1}^{k}T(i,i-\ell,p,q)\\
&\quad\le4\sigma^2\eta^3\zeta^{3k-1}\sum_{\ell=1}^{n-1}\sum_{i=\ell+1}^{n}\lambda^{2n-2i+\ell}\zeta^{2\ell}\sum_{p=1}^{k}\sum_{q=1}^{k}\zeta^{|p-q|-(p+q)}\\
&\quad\le\frac{4\sigma^2\eta^3\zeta^{k-1}k^2}{1-\lambda^2}\left(\frac{\lambda\zeta^2-\lambda^n\zeta^{2n}}{1-\lambda\zeta^2}-\frac{\lambda^{2n-1}\zeta^2-\lambda^n\zeta^{2n}}{1-\lambda^{-1}\zeta^2}\right),
\end{aligned}$$

where the last inequality is obtained by observing that $-2k\le|p-q|-(p+q)<0$ for all $p,q\in\{1,\ldots,k\}$. By taking $n\to\infty$, we get

$$\begin{aligned}
&4\sigma^2\sum_{\ell=1}^{n-1}\sum_{i=\ell+1}^{n}\lambda^{2n-2i+\ell}\sum_{p=1}^{k}\sum_{q=1}^{k}T(i,i-\ell,p,q)\\
&\quad\le4\sigma^2\frac{k^2}{1-\lambda^2}\frac{\eta^3}{1-\lambda\zeta^2}\lambda\zeta^{k+1},
\end{aligned}$$

and we readily obtain

$$\frac{(1-\lambda)^2}{k\sigma^4}\mathrm{tr}(E[(\mathbf{H}_{\lambda,n}-E[\mathbf{H}_{\lambda,n}])^2])$$

$$\le\left(8\frac{1-\lambda}{1+\lambda}k\right)c(0)^2\frac{3k+5}{8k}\tag{66}$$

$$+\left(8\frac{1-\lambda}{1+\lambda}k\right)\sum_{r=1}^{k-1}c(r)^2\left(\frac{5(k-r)}{4k^2}+\frac{k+1}{4k}\lambda^r\right)\tag{67}$$

$$+\left(8\frac{1-\lambda}{1+\lambda}k\right)\left(\frac{k+1}{4k}\frac{\eta^2}{1/(\lambda\zeta^2)-1}(\lambda\zeta^4)^{k-1}\right)\tag{68}$$

$$+\left(8\frac{1-\lambda}{1+\lambda}k\right)\left(\frac{k-1}{4k}\frac{\eta^2}{1/(\lambda\zeta^2)-1}\zeta^{2(k-1)}\right)\tag{69}$$

$$+\left(8\frac{1-\lambda}{1+\lambda}k\right)\left(\frac{\eta}{2\sigma^2}\frac{\eta^2}{1/(\lambda\zeta^2)-1}\zeta^{k-1}\right).\tag{70}$$

Similarly with (64), we have

$$\frac{(1-\lambda)^2}{k\sigma^4}\mathrm{tr}((E[\mathbf{H}_{\lambda,n}])^2)=c(0)^2+\sum_{r=1}^{k-1}c(r)^2\frac{2(k-r)}{k}.\tag{71}$$

In (66)–(70), the common factor $8((1-\lambda)/(1+\lambda))k$ was chosen such that the coefficient $(3k+5)/(8k)$ that multiplies $c(0)^2$ in (66) does not exceed one, the coefficient of $c(0)^2$ in (71). By comparing (66)–(70) with (71), we decide that the inequality $8((1-\lambda)/(1+\lambda))k\ll1$ must be satisfied such that the matrix $\mathbf{H}_{\lambda,n}$ is quasi-deterministic. Remark that the condition is slightly stronger than the previously found condition for $\mathbf{G}_{\lambda,n}$ to be quasi-deterministic.

A.2. Proof of Proposition 4.1

The most important ideas of the proof are inspired by [14,30,31], where the analysis is restricted to the case $\lambda = 1$. Because the case $\lambda \in (0,1)$ poses supplementary difficulties, we first demonstrate some auxiliary results that will be instrumental for the main proof given in Section A.2.2.

A.2.1. Auxiliary results

Lemma A.1. *The following identity holds*:

$$\mathrm{PLS}_\lambda(k) = \sum_{i=m+1}^{n} \lambda^{n-i} e^2_{\lambda,i} = R_{\lambda,n} + \sum_{j=1}^{3} \mathscr{S}^{(j)}_{\lambda,n}, \tag{72}$$

where

$$\mathscr{S}^{(1)}_{\lambda,n} = \sum_{i=m+1}^{n} \lambda^{n-i} d_{\lambda,i} \varepsilon_i^2,$$

$$\mathscr{S}^{(2)}_{\lambda,n} = \sum_{i=m+1}^{n} \lambda^{n-i} d_{\lambda,i} [(\hat{\mathbf{a}}_{\lambda,i-1} - \mathbf{a})^\top \bar{\mathbf{x}}_i]^2,$$

$$\mathscr{S}^{(3)}_{\lambda,n} = 2\sum_{i=m+1}^{n} \lambda^{n-i} d_{\lambda,i} [(\hat{\mathbf{a}}_{\lambda,i-1} - \mathbf{a})^\top \bar{\mathbf{x}}_i] \varepsilon_i.$$

Proof. For each $t \in \{m+1,\ldots,n\}$, we consider Eq. (20) and we multiply it by λ^{n-t}. We sum together all the resulting equalities, and the identity

$$\sum_{i=m+1}^{n} \lambda^{n-i} e^2_{\lambda,i} = R_{\lambda,n} - \lambda^{n-m} R_{\lambda,m} + \sum_{i=m+1}^{n} \lambda^{n-i} d_{\lambda,i} e^2_{\lambda,i} \tag{73}$$

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 (17) lead to (72). □

Lemma A.2. *We have the following results*:

$$\lim_{n\to\infty} \sum_{i=m+1}^{n} \lambda^{n-i} d_{\lambda,i} < \infty, \tag{74}$$

$$\lim_{n\to\infty} \mathscr{S}^{(1)}_{\lambda,n} = \lim_{n\to\infty} \sum_{i=m+1}^{n} \lambda^{n-i} d_{\lambda,i} \varepsilon_i^2 < \infty \quad a.s. \tag{75}$$

Proof. Based on (22), 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 (74) is obtained by applying the comparison test for convergence. The result (75) is a direct consequence of (74) and Lemma 2(iii) from [14]. □

Lemma A.3. *For n large*, $\mathscr{S}^{(2)}_{\lambda,n} + \mathscr{S}^{(3)}_{\lambda,n} = O(\mathscr{S}^{(1)}_{\lambda,n})$ *a.s.*

Proof. From (1) and (16), we have

$$\begin{aligned}
&[(\hat{\mathbf{a}}_{\lambda,i-1} - \mathbf{a})^\top \bar{\mathbf{x}}_i]^2 \\
&= \left[\bar{\mathbf{x}}_i^\top \left(-\mathbf{a} - \mathbf{V}_{\lambda,i-1} \sum_{j=1}^{i-1} \lambda^{i-1-j} \bar{\mathbf{x}}_j (-\bar{\mathbf{x}}_j^\top \mathbf{a} + \varepsilon_j)\right)\right]^2 \\
&= \left[\bar{\mathbf{x}}_i^\top \left(-\mathbf{a} + \mathbf{V}_{\lambda,i-1} \left(\sum_{j=1}^{i-1} \lambda^{i-1-j} \bar{\mathbf{x}}_j \bar{\mathbf{x}}_j^\top\right) \mathbf{a} \right.\right. \\
&\qquad \left.\left. - \mathbf{V}_{\lambda,i-1} \sum_{j=1}^{i-1} \lambda^{i-1-j} \bar{\mathbf{x}}_j \varepsilon_j \right)\right]^2 \\
&= \left[\bar{\mathbf{x}}_i^\top \mathbf{V}_{\lambda,i-1} \sum_{j=1}^{i-1} \lambda^{i-1-j} \bar{\mathbf{x}}_j \varepsilon_j\right]^2.
\end{aligned} \tag{76}$$

Next we introduce $D_{\lambda,i}$, $i \in \{m \ldots, n\}$, and we obtain a recursive formula for it:

$$\begin{aligned}
D_{\lambda,i} &= \left[\sum_{j=1}^{i} \lambda^{i-j} \bar{\mathbf{x}}_j^\top \varepsilon_j\right] \mathbf{V}_{\lambda,i} \left[\sum_{j=1}^{i} \lambda^{i-j} \bar{\mathbf{x}}_j \varepsilon_j\right] \\
&= \left[\sum_{j=1}^{i-1} \lambda^{i-j} \bar{\mathbf{x}}_j^\top \varepsilon_j\right] \mathbf{V}_{\lambda,i} \left[\sum_{j=1}^{i-1} \lambda^{i-j} \bar{\mathbf{x}}_j \varepsilon_j\right] + \bar{\mathbf{x}}_i^\top \mathbf{V}_{\lambda,i} \bar{\mathbf{x}}_i \varepsilon_i^2 \\
&\quad + 2\bar{\mathbf{x}}_i^\top \mathbf{V}_{\lambda,i} \left[\sum_{j=1}^{i-1} \lambda^{i-j} \bar{\mathbf{x}}_j \varepsilon_j\right] \varepsilon_i \\
&= \lambda D_{\lambda,i-1} - \left[\sum_{j=1}^{i-1} \lambda^{i-j} \bar{\mathbf{x}}_j^\top \varepsilon_j\right] \frac{1}{\lambda^2} \frac{\mathbf{V}_{\lambda,i-1} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^\top \mathbf{V}_{\lambda,i-1}}{1 + c_{\lambda,i}} \\
&\quad \times \left[\sum_{j=1}^{i-1} \lambda^{i-j} \bar{\mathbf{x}}_j \varepsilon_j\right] && (77) \\
&\quad + \bar{\mathbf{x}}_i^\top \mathbf{V}_{\lambda,i} \bar{\mathbf{x}}_i \varepsilon_i^2 + 2 \frac{\bar{\mathbf{x}}_i^\top \mathbf{V}_{\lambda,i-1}}{\lambda(1 + c_{\lambda,i})} \left[\sum_{j=1}^{i-1} \lambda^{i-j} \bar{\mathbf{x}}_j \varepsilon_j\right] \varepsilon_i && (78) \\
&= \lambda D_{\lambda,i-1} - \frac{\left[\bar{\mathbf{x}}_i^\top \mathbf{V}_{\lambda,i-1} \sum_{j=1}^{i-1} \lambda^{i-1-j} \bar{\mathbf{x}}_j \varepsilon_j\right]^2}{1 + c_{\lambda,i}} + d_{\lambda,i} \varepsilon_i^2 \\
&\quad + \frac{2\bar{\mathbf{x}}_i^\top \mathbf{V}_{\lambda,i-1} \left[\sum_{j=1}^{i-1} \lambda^{i-1-j} \bar{\mathbf{x}}_j \varepsilon_j\right] \varepsilon_i}{1 + c_{\lambda,i}}. && (79)
\end{aligned}$$

In (77) and (78), we applied the following two identities which are consequences of the matrix inversion lemma [12]: for all $i > m$,

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

and

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

We multiply (79) by λ^{n-i} for each $i \in \{m+1,\ldots,n\}$, and after summing together all the resulting equalities, we get

$$\begin{aligned}
&D_{\lambda,n} - \lambda^{n-m} D_{\lambda,m} + \sum_{i=m+1}^{n} \frac{\lambda^{n-i} [\bar{\mathbf{x}}_i^\top \mathbf{V}_{\lambda,i-1} \sum_{j=1}^{i-1} \lambda^{i-1-j} \bar{\mathbf{x}}_j \varepsilon_j]^2}{1 + c_{\lambda,i}} \\
&\quad = \sum_{i=m+1}^{n} \lambda^{n-i} d_{\lambda,i} \varepsilon_i^2 + 2 \sum_{i=m+1}^{n} \frac{\lambda^{n-i} \bar{\mathbf{x}}_i^\top \mathbf{V}_{\lambda,i-1} [\sum_{j=1}^{i-1} \lambda^{i-1-j} \bar{\mathbf{x}}_j \varepsilon_j] \varepsilon_i}{1 + c_{\lambda,i}}.
\end{aligned} \tag{80}$$

Lemma 2(iii) from [14] leads to $\sum_{i=m+1}^{n} u_i\varepsilon_i = o(\sum_{i=m+1}^{n} u_i^2)+O(1)$ a.s., where

$$u_i = \frac{\lambda^{n-i}\bar{\mathbf{x}}_i^\top \mathbf{V}_{\lambda,i-1}[\sum_{j=1}^{i-1}\lambda^{i-1-j}\bar{\mathbf{x}}_j\varepsilon_j]}{1+c_{\lambda,i}}$$

when $i>m$, and conventionally $u_i=0$ for $i\le m$. Because $(1+c_{\lambda,i})/\lambda^{n-i}\ge 1$ for all $i>m$, we also have

$$\sum_{i=m+1}^{n} u_i\varepsilon_i = o\left(\sum_{i=m+1}^{n}\frac{1+c_{\lambda,i}}{\lambda^{n-i}}u_i^2\right)+O(1)\text{ a.s.} \tag{81}$$

Based on (75), (80) and (81), we note that

$$\lim_{n\to\infty}\sum_{i=m+1}^{n}\frac{\lambda^{n-i}[\bar{\mathbf{x}}_i^\top\mathbf{V}_{\lambda,i-1}\sum_{j=1}^{i-1}\lambda^{i-1-j}\bar{\mathbf{x}}_j\varepsilon_j]^2}{1+c_{\lambda,i}}<\infty \quad \text{a.s.}$$

Therefore, by using (76) and the identity $1/(1+c_{\lambda,i}) = 1-d_{\lambda,i}$, Eqs. (80) and (81) can be re-written as

$$\left\{\sum_{i=m+1}^{n}\lambda^{n-i}(1-d_{\lambda,i})[(\hat{\mathbf{a}}_{\lambda,i-1}-\mathbf{a})^\top\bar{\mathbf{x}}_i]^2\right\}(1+o(1)) = \mathscr{S}^{(1)}_{\lambda,n}+O(1)\text{ a.s.,}$$

and by resorting to the definition of $\mathscr{S}^{(2)}_{\lambda,n}$, we readily obtain $\mathscr{S}^{(2)}_{\lambda,n}=O(\mathscr{S}^{(1)}_{\lambda,n})$ a.s. We apply the Cauchy–Schwarz inequality to get $[\mathscr{S}^{(3)}_{\lambda,n}/(2\mathscr{S}^{(1)}_{\lambda,n})]^2\le\mathscr{S}^{(2)}_{\lambda,n}/\mathscr{S}^{(1)}_{\lambda,n}$, which concludes the proof. □

Lemma A.4. *For n large,* $\mathscr{S}^{(1)}_{\lambda,n}=k\sigma^2+O(1)$.

Proof. Let us consider the usual norms for vectors and matrices. Thus, $\|\bar{\mathbf{x}}\|^2=\bar{\mathbf{x}}^\top\bar{\mathbf{x}}$ for an arbitrary $\bar{\mathbf{x}}\in\mathfrak{R}^{k\times 1}$. For $\mathbf{M}\in\mathfrak{R}^{k\times k}$, $\|\mathbf{M}\|^2$ equals the largest eigenvalue of $\mathbf{M}^\top\mathbf{M}$. Eq. (30) guarantees for any $\delta>0$ there exists i_0 such that for all $i>i_0$ the following inequality is verified: $\|\mathbf{V}_{\lambda,i}-\mathbf{G}_\lambda^{-1}\|\le\delta/\|\mathbf{G}_\lambda\|$. As $\mathbf{G}_\lambda$ is positive definite, we also have for all $\bar{\mathbf{x}}\in\mathfrak{R}^{k\times 1}$, $\|\bar{\mathbf{x}}\|^2/\|\mathbf{G}_\lambda\|\le\bar{\mathbf{x}}^\top\mathbf{G}_\lambda^{-1}\bar{\mathbf{x}}$. Simple calculations lead to $(1-\delta)\mathscr{S}^{(4)}_{\lambda,n}\le\sum_{i=i_0+1}^{n}\lambda^{n-i}d_{\lambda,i}\varepsilon_i^2\le(1+\delta)\mathscr{S}^{(4)}_{\lambda,n}$, where $\mathscr{S}^{(4)}_{\lambda,n}=\sum_{i=i_0+1}^{n}\lambda^{n-i}\bar{\mathbf{x}}_i^\top\mathbf{G}_\lambda^{-1}\bar{\mathbf{x}}_i\varepsilon_i^2$. Taking $\delta\to 0$, we get $\sum_{i=i_0+1}^{n}\lambda^{n-i}d_{\lambda,i}\varepsilon_i^2=\mathscr{S}^{(4)}_{\lambda,n}$. Without loss of generality, we assume $i_0>m$, and by applying (31) and the identity above, we obtain

$$\begin{aligned}
\lim_{n\to\infty}\mathscr{S}^{(1)}_{\lambda,n} &= \lim_{n\to\infty}\left[\sum_{i=m+1}^{i_0}+\sum_{i=i_0+1}^{n}\right]\left(\lambda^{n-i}d_{\lambda,i}\varepsilon_i^2\right)\\
&= \lim_{n\to\infty}\sum_{i=i_0+1}^{n}\left(\lambda^{n-i}\bar{\mathbf{x}}_i^\top\mathbf{G}_\lambda^{-1}\bar{\mathbf{x}}_i\varepsilon_i^2\right)+O(1)\\
&= \lim_{n\to\infty}\left[\sum_{i=1}^{n}-\sum_{i=1}^{i_0}\right](\lambda^{n-i}\bar{\mathbf{x}}_i^\top\mathbf{G}_\lambda^{-1}\bar{\mathbf{x}}_i\varepsilon_i^2)+O(1)\\
&= \lim_{n\to\infty}\sum_{i=1}^{n}(\lambda^{n-i}\bar{\mathbf{x}}_i^\top\mathbf{G}_\lambda^{-1}\bar{\mathbf{x}}_i\varepsilon_i^2)+O(1)\\
&= \lim_{n\to\infty}\operatorname{tr}(\mathbf{G}_\lambda^{-1}\sum_{i=1}^{n}\lambda^{n-i}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^\top\varepsilon_i^2)+O(1)\\
&= \operatorname{tr}(\mathbf{G}_\lambda^{-1}\mathbf{H}_\lambda)+O(1)\\
&= k\sigma^2+O(1). \quad \square
\end{aligned}$$

Lemma A.5. *The following results hold*:

$$\sum_{i=m+1}^{n}\lambda^{n-i}\hat{e}^2_{\lambda,i}=R_{\lambda,n}-\mathscr{S}^{(1)}_{\lambda,n}(1+O(1)), \tag{82}$$

$$\ln\left(\frac{1}{n^\infty_{\mathrm{ef}}}\sum_{i=m+1}^{n}\lambda^{n-i}\hat{e}^2_{\lambda,i}\right)=\ln\frac{R_{\lambda,n}}{n^\infty_{\mathrm{ef}}}-\frac{k}{n^\infty_{\mathrm{ef}}}(1+O(1)). \tag{83}$$

Proof. Because $\hat{e}^2_{\lambda,i}=(1-d_{\lambda,i})^2e^2_{\lambda,i}$ for all $i\in\{m+1,\dots,n\}$ [12], Eq. (73) implies $\sum_{i=m+1}^{n}\lambda^{n-i}\hat{e}^2_{\lambda,i}=R_{\lambda,n}-\lambda^{n-m}R_{\lambda,m}-\sum_{i=m+1}^{n}\lambda^{n-i}d_{\lambda,i}e^2_{\lambda,i}+\mathscr{S}^{(5)}_{\lambda,n}$, where $\mathscr{S}^{(5)}_{\lambda,n}=\sum_{i=m+1}^{n}\lambda^{n-i}d^2_{\lambda,i}e^2_{\lambda,i}$. We know from Lemma A.1 that $\sum_{i=m+1}^{n}\lambda^{n-i}d_{\lambda,i}e^2_{\lambda,i}=\sum_{j=1}^{3}\mathscr{S}^{(j)}_{\lambda,n}$, and from Lemma A.3 we have $\mathscr{S}^{(2)}_{\lambda,n}+\mathscr{S}^{(3)}_{\lambda,n}=O(\mathscr{S}^{(1)}_{\lambda,n})$. The inequality in (22) leads to $0<\mathscr{S}^{(5)}_{\lambda,n}<\sum_{i=m+1}^{n}\lambda^{n-i}d_{\lambda,i}e^2_{\lambda,i}$. Additionally $\lambda^{n-m}R_{\lambda,m}\approx 0$, and the result in (82) is proven. Then

$$\ln\left(\frac{1}{n^\infty_{\mathrm{ef}}}\sum_{i=m+1}^{n}\lambda^{n-i}\hat{e}^2_{\lambda,i}\right)=\ln\left(\frac{R_{\lambda,n}}{n^\infty_{\mathrm{ef}}}\right)+\ln\left(1-\frac{k}{n^\infty_{\mathrm{ef}}}\frac{\sigma^2}{\frac{R_{\lambda,n}}{n^\infty_{\mathrm{ef}}}}(1+O(1))\right),$$

which is a consequence of (82) and Lemma A.4. To get (83), we use $R_{\lambda,n}/n^\infty_{\mathrm{ef}}\approx\sigma^2$ and $\ln(1-\xi)\approx-\xi$ for $|\xi|$ close to zero. □

A.2.2. Main results

Remark in the proofs above that the assumption ($\mathbb{A}1$) was needed for all lemmas, while ($\mathbb{A}2$) was used only for Lemmas A.4 and A.5. Next we explain how the identities (33)–(36) are derived:

- Eq. (33): is readily obtained from Lemmas A.1, A.3 and A.4.
- Eq. (34): is a straightforward consequence of (24), (25) and (33).
- Eq. (35): by using (26) and (30), we get asymptotically the following expression for the penalty term of PDC$_\lambda$:

$$\begin{aligned}
&-\ln\prod_{i=m+1}^{n}\frac{|\mathbf{V}^{-1}_{i-1,\lambda}|^{1/2}}{|\mathbf{V}^{-1}_{i,\lambda}|^{1/2}}+\frac{1}{2}\ln n^\infty_{\mathrm{ef}}\\
&=\frac{1}{2}\ln\frac{|\mathbf{V}^{-1}_{n,\lambda}|}{|\mathbf{V}^{-1}_{m,\lambda}|}+\frac{1}{2}\ln n^\infty_{\mathrm{ef}}\\
&=\frac{k+1}{2}\ln n^\infty_{\mathrm{ef}}+\frac{1}{2}\ln\frac{|\mathbf{C}|}{|\mathbf{V}^{-1}_{m,\lambda}|}+o(1),
\end{aligned} \tag{84}$$

  which leads to (35).
- Eq. (36): is obtained by applying (83) to the first term in (28), and by using (84) to re-write the other two terms in (28).

A.3. On the relationship between the asymptotic approximations (13) and (33)

Based on the definitions from Lemma A.1, we note that $\mathscr{S}^{(1)}_{\lambda,n}$ becomes $\mathscr{S}^{(1)}_{n}=\sum_{i=m+1}^{n}d_i\varepsilon_i^2$ when $\lambda=1$. The crucial difference between $\mathscr{S}^{(1)}_{\lambda,n}$ and $\mathscr{S}^{(1)}_{n}$ is that $\lim_{n\to\infty}\mathscr{S}^{(1)}_{\lambda,n}<\infty$ a.s., $\forall\lambda\in(0,1)$, as we know from Lemma A.2, whereas $\lim_{n\to\infty}\mathscr{S}^{(1)}_{n}=\infty$ [31]. Moreover, we have shown in Lemmas A.1–A.3 that $\mathrm{PLS}_\lambda(k)=R_{\lambda,n}+\mathscr{S}^{(1)}_{\lambda,n}(1+O(1))$, whereas in Theorem 2.2 from [31], it is proven under mild conditions that $\mathrm{PLS}(k) = R_n + \mathscr{S}_n^{(1)}(1+o(1))$. To continue the brief comparison between the two cases, we re-write Eqs. (30) and (31) as

$$\frac{1}{n_{\mathrm{ef}}^{\infty}}\sum_{i=1}^{n}\lambda^{n-i}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^{\top} \approx \mathbf{C}$$

and

$$\frac{1}{n_{\mathrm{ef}}^{\infty}}\sum_{i=1}^{n}\lambda^{n-i}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^{\top}\varepsilon_i^2 \approx \sigma^2\mathbf{C},$$

respectively. Observe in both equations that the normalization factor, n_{ef}^{∞}, depends on λ and it is independent of n, the number of terms in the two summations. According to Lemma A.4, this property leads to $\lim_{n\to\infty}\mathscr{S}_{\lambda,n}^{(1)} = k\sigma^2 + O(1)$. On the other hand, the ergodicity guarantees that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^{\top} = \mathbf{C}$$

and

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^{\top}\varepsilon_i^2 = \sigma^2\mathbf{C},$$

which implies

$$\lim_{n\to\infty}\frac{1}{\ln n}\mathscr{S}_n^{(1)} = \sigma^2 k.$$

This result was previously obtained (see [31, p. 7]) by applying the same methodology as the one used in Lemma A.4, with the major difference that the normalization factor in both

$$\frac{1}{n}\sum_{i=1}^{n}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^{\top}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^{\top}\varepsilon_i^2$$

is given by the number of terms within each summation, and not by n_{ef}^{∞} as it is the case for (30)–(31). This explains the presence of the factor $1/\ln n$ in

$$\lim_{n\to\infty}\frac{1}{\ln n}\mathscr{S}_n^{(1)} = \sigma^2 k.$$

References

[1] T. Adali, S.H. Ardalan, On the effect of input signal correlation on weight misadjustment in the RLS algorithm, IEEE Trans. Signal Process. 43 (4) (1995) 988–991.

[2] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control AC-19 (December 1974) 716–723.

[3] R.A. Davis, T.C.M. Lee, G.A. Rodriguez-Yam, Structural break estimation for nonstationary time series models, J. Am. Stat. Assoc. 101 (2006) 223–239.

[4] P.M. Djuric, S.M. Kay, Order selection of autoregressive models, IEEE Trans. Signal Process. 40 (1992) 2829–2833.

[5] E. Eleftheriou, D. Falconer, Tracking properties and steady-state performance of RLS adaptive filter algorithms, IEEE Trans. Acoust. Speech Signal Process. 34 (5) (1986) 1097–1110.

[6] B. Friedlander, Lattice filters for adaptive processing, Proc. IEEE 70 (1982) 829–868.

[7] C.D. Giurcăneanu, S.A. Razavi, AR order selection with information theoretic criteria based on localized estimators, in: Proceedings of the 16th European Signal Processing Conference (Eusipco 2008), Lausanne, Switzerland, August 25–29, 2008 ⟨http://www.eurasip.org/Proceedings/Eusipco/Eusipco2008/papers/1569102210.pdf⟩.

[8] C.D. Giurcăneanu, J. Rissanen, Estimation of AR and ARMA models by stochastic complexity, in: H.-C. Ho, C.-K. Ing, T.L. Lai (Eds.), Time Series and Related Topics, Institute of Mathematical Statistics Lecture Notes-Monograph Series, vol. 52, 2006, pp. 48–59 ⟨http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.lnms/1196285965⟩.

[9] S. Goto, M. Nakamura, 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. Signal Process. 43 (1995) 1519–1522.

[10] P.D. Grünwald, The Minimum Description Length Principle, MIT Press, Cambridge, MA, 2007.

[11] E.J. Hannan, A.J. McDougall, D.S. Poskitt, Recursive estimation of autoregressions, J. R. Stat. Soc. B 51 (1989) 217–233.

[12] S. Haykin, Adaptive Filter Theory, Prentice-Hall, Englewood Cliffs, NJ, 2002.

[13] P.H.M. Janssen, P. Stoica, On the expectation of the product of four matrix-valued Gaussian random variables, IEEE Trans. Autom. Control 33 (9) (1988) 867–870.

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

[15] O.M. Macchi, N.J. Bershad, Adaptive recovery of a chirped sinusoid in noise, Part I: performance of the RLS algorithm, IEEE Trans. Signal Process. 39 (3) (1991) 583–594.

[16] M. Niedzwiecki, On the localized estimators and generalized Akaike's criteria, IEEE Trans. Autom. Control 29 (11) (1984) 970–983.

[17] M. Niedzwiecki, Bayesian-like autoregressive spectrum estimation in the case of unknown process order, IEEE Trans. Autom. Control 30 (10) (1985) 950–961.

[18] M. Niedzwiecki, Identification of Time-varying Processes, Wiley, New York, 2000.

[19] H.C. Ombao, J.A. Raz, R. von Sachs, B.A. Malow, Automatic statistical analysis of bivariate nonstationary time series. In memory of Jonathan A. Raz, J. Am. Stat. Assoc. 96 (454) (2001) 543–560.

[20] D.S. Poskitt, On the selection of irregular, misspecified regression models: a comment on folklore, J. Japan Stat. Soc. 38 (1) (2008) 75–86.

[21] J. Rissanen, Modeling by shortest data description, Automatica 14 (1978) 465–471.

[22] J. Rissanen, Order estimation by accumulated prediction errors, J. Appl. Prob. 23A (1986) 55–61.

[23] J. Rissanen, Stochastic complexity, J. R. Stat. Soc. B 49 (1987) 252–265.

[24] J. Rissanen, Information and Complexity in Statistical Modeling, Springer, Berlin, 2007.

[25] J. Rissanen, T. Roos, Conditional NML universal models, in: Proceedings of the Information Theory and Applications Workshop, University California, San Diego, USA (ITA-07), January 29–February 2, 2007, pp. 337–341 ⟨http://ita.ucsd.edu/workshop/07/files/paper/paper_212.pdf⟩.

[26] T. Roos, J. Rissanen On sequentially normalized maximum likelihood models, in: Proceedings of the Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-08), Tampere, Finland, August 18–20, 2008 ⟨http://sp.cs.tut.fi/WITMSE08/Proceedings/⟩.

[27] G. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (1978) 461–464.

[28] R. Shibata, Asymptotically efficient selection of the order of the model for estimating parameters of a linear process, Ann. Stat. 8 (1) (1980) 147–164.

[29] M. Wax, Order selection for AR models by predictive least squares, IEEE Trans. Acoust. Speech Signal Process. 36 (1988) 581–588.

[30] C.Z. Wei, Adaptive prediction by least squares predictors in stochastic regression models with applications to time series, Ann. Stat. (1987) 1667–1682.

[31] C.Z. Wei, On predictive least squares principles, Ann. Stat. 20 (1992) 1–42.

[32] Y. Zheng, Z. Lin, Recursive adaptive algorithms for fast and rapidly time-varying systems, IEEE Trans. Circuits Syst. II 50 (2003) 602–614.