Postprint of: Niedźwiecki M., Ciołek M., Identification of nonstationary multivariate autoregressive processes – Comparison of competitive and collaborative strategies for joint selection of estimation bandwidth and model order, Digital Signal Processing, Vol. 78 (2018), pp. 72-81, DOI: 10.1016/j.dsp.2018.02.013

Identification of nonstationary multivariate autoregressive processes – comparison of competitive and collaborative strategies for joint selection of estimation bandwidth and model order

Maciej Niedźwiecki and Marcin Ciołek

***Abstract*—The problem of identification of multivariate autoregressive processes (systems or signals) with unknown and possibly time-varying model order and time-varying rate of parameter variation is considered and solved using parallel estimation approach. Under this approach, several local estimation algorithms, with different order and bandwidth settings, are run simultaneously and compared based on their predictive performance. First, the competitive decision schemes are considered. It is shown that the best parameter tracking results can be obtained when the order is selected based on minimization of the appropriately modified Akaike's final prediction error statistic, and the bandwidth is chosen using the localized version of the Rissanen's predictive least squares statistic. Next, it is shown that estimation results can be further improved if a collaborative decision is made by means of applying the Bayesian model averaging technique.**

***Index Terms*—Identification of nonstationary processes, determination of estimation bandwidth, model order selection.**

I. Introduction

AUTOREGRESSIVE models have found a large number of applications in many research areas such as signal prediction [1], adaptive control [2], equalization of telecommunication channels [3], biomedical signal analysis [4]–[7], elimination of impulsive disturbances from archive audio signals [8], and spectrum estimation [9], among many others. Such models are rarely based on physical insights and therefore their coefficients usually have no physical interpretation. However, they have some obvious advantages: they are easy to build using statistical inference and, more importantly, they allow mathematically tractable formulations and then solutions for problems arising in applications mentioned earlier.

When the analyzed process (system or signal) is nonstationary, identification of its autoregressive model can be carried out using local estimation techniques. In such a case two important decisions must be taken: selection of the estimation bandwidth (inversely proportional to the size of the local analysis window), i.e., the frequency range in which process parameters can be tracked "successfully" [10], and selection of the model order.

In the system identification case, estimation bandwidth should be chosen in accordance with the degree of system nonstationarity (quantifying how fast the statistics of the underlying process vary in time), so as to trade off the bias and variance components of the mean squared parameter tracking error. Selection of the model order should be made in a way that allows one to capture the dominant system dynamics under the information content constraints imposed by the limited estimation bandwidth. Both overfitted and underfitted models suffer from quantitative and qualitative drawbacks such as lower predictive capabilities, neglected or nonexistent dynamics, etc.

The appropriate choice of model order and estimation bandwidth is equally important in signal analysis applications, where autoregressive modeling is often used for the purpose of parametric spectrum estimation [11]. Misspecified order and/or bandwidth may result in the incorrect resonant structure of the estimated spectrum (existence or nonexistence of spectral peaks may lead to wrong qualitative interpretation of the spectrum), as well as in increased estimation errors.

Finally, we note that the problems of bandwidth and order selection are mutually coupled since, according to the principle of parsimony [10], smaller bandwidth allows one to estimate a larger number of model parameters and *vice versa*.

The problem of joint bandwidth and order adaptation for the purpose of *noncausal* identification of *autoregressive signals* was considered for the first time in our earlier papers [12] (for weighted Yule-Walker algorithms) and [13] (for doubly exponentially weighted lattice algorithms). In the current paper we will study the analogous problem for multivariate *autoregressive systems* with exogenous inputs, identified using *causal* exponentially weighted least squares algorithms. We will propose and compare several decision rules based on the modified (localized) Akaike's final prediction error (FPE) statistic [14], [15] and on the predictive least squares (PLS) principle [16], [17], [18]. Finally, we will show how the "hard selection" (competitive) scheme can be extended to a "soft selection" (collaborative) one, based on the Akaike's concept of Bayesian model averaging [19], [20]. Interestingly, some of the qualitative conclusions reached for causal estimators differ from those reported in [12] for noncausal ones. The paper extends preliminary results presented in [21].

Maciej Niedźwiecki and Marcin Ciołek are with the Department of Automatic Control, Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, Narutowicza 11/12, 80-233 Gdańsk, Poland, Tel: + 48 58 3472519; fax: +48 58 3415821.

II. System description

Consider the time-invariant multivariate system governed by the ARX (autoregressive with exogenous input) equation

$$\begin{gathered}\mathbf{y}(t)=\sum_{i=1}^{n}\mathbf{A}_i(t)\mathbf{y}(t-i)+\sum_{i=1}^{n}\mathbf{B}_i(t)\mathbf{u}(t-i)+\mathbf{e}(t)\\ \mathrm{cov}[\mathbf{e}(t)]=\boldsymbol{\rho}(t)\end{gathered}\tag{1}$$

where $t=1,2,\ldots$ denotes normalized (dimensionless) time, $\mathbf{y}(t)=[y_1(t),\ldots,y_{m_y}(t)]^{\mathrm{T}}$ denotes the m_y-dimensional output signal, $\mathbf{u}(t)=[u_1(t),\ldots,u_{m_u}(t)]^{\mathrm{T}}$ denotes the m_u-dimensional observable input signal, $\mathbf{e}(t)$ denotes zero-mean white input noise, and $\mathbf{A}_i(t)$, $\mathbf{B}_i(t)$ are the $m_y\times m_y$- and $m_y\times m_u$-dimensional matrices of time-varying autoregressive and input coefficients, respectively:

$$\mathbf{A}_i(t)=\begin{bmatrix}\boldsymbol{\alpha}_{1i}^{\mathrm{T}}(t)\\ \vdots\\ \boldsymbol{\alpha}_{m_y i}^{\mathrm{T}}(t)\end{bmatrix},\qquad \mathbf{B}_i(t)=\begin{bmatrix}\boldsymbol{\beta}_{1i}^{\mathrm{T}}(t)\\ \vdots\\ \boldsymbol{\beta}_{m_y i}^{\mathrm{T}}(t)\end{bmatrix}\tag{2}$$

$$i=1,\ldots,n\qquad\qquad i=1,\ldots,n,$$

where $\boldsymbol{\alpha}_{li}(t)=[a_{l1,i}(t),\ldots,a_{lm_y,i}(t)]^{\mathrm{T}}$ and $\boldsymbol{\beta}_{li}(t)=[b_{l1,i}(t),\ldots,b_{lm_u,i}(t)]^{\mathrm{T}}$ for $l=1,\ldots,m_y$.

Denote by $\boldsymbol{\theta}_n^j(t)=[\boldsymbol{\alpha}_{j1}^{\mathrm{T}}(t),\ldots,\boldsymbol{\alpha}_{jn}^{\mathrm{T}}(t),\boldsymbol{\beta}_{j1}^{\mathrm{T}}(t),\ldots,\boldsymbol{\beta}_{jn}^{\mathrm{T}}(t)]^{\mathrm{T}}$ the d_n-dimensional, $d_n=n(m_y+m_u)$, vector of parameters characterizing the j-th output, called also the j-th channel, of the ARX system, and by $\boldsymbol{\varphi}_n(t)=[\mathbf{y}^{\mathrm{T}}(t-1),\ldots,\mathbf{y}^{\mathrm{T}}(t-n),\mathbf{u}^{\mathrm{T}}(t-1),\ldots,\mathbf{u}^{\mathrm{T}}(t-n)]^{\mathrm{T}}$ – the corresponding regression vector (the same for all channels) of the same dimension. Finally, denote by $\boldsymbol{\theta}_n(t)=[(\boldsymbol{\theta}_n^1(t))^{\mathrm{T}},\ldots,(\boldsymbol{\theta}_n^{m_y}(t))^{\mathrm{T}}]^{\mathrm{T}}=\mathrm{vec}\{[\mathbf{A}_1(t),\ldots,\mathbf{A}_n(t),\mathbf{B}_1(t),\ldots,\mathbf{B}_n(t)]^{\mathrm{T}}\}$ the vector made up of all $D_n=m_y d_n$ system parameters and let $\boldsymbol{\Psi}_n(t)=\mathbf{I}_{m_y}\otimes\boldsymbol{\varphi}_n(t)=\mathrm{diag}\{\boldsymbol{\varphi}_n(t),\ldots,\boldsymbol{\varphi}_n(t)\}$ where the symbol $\otimes$ denotes Kronecker product of two matrices/vectors. Using this shorthand notation, (1) can be rewritten in the form

$$\mathbf{y}(t)=\boldsymbol{\Psi}_n^{\mathrm{T}}(t)\boldsymbol{\theta}_n(t)+\mathbf{e}(t)\ .\tag{3}$$

When system parameters vary slowly with time they can be estimated using a localized least squares (LS) algorithm, such as the one based on the well-known method of exponentially weighted least squares (EWLS). To achieve the effect of forgetting 'old' data, the sum of squares minimized in the method of least squares is replaced with the exponentially weighted sum of squares, resulting in the following EWLS estimator [22]

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_{n|k}(t)&=\arg\min_{\boldsymbol{\theta}_n}\sum_{i=0}^{t-1}\lambda_k^i\parallel\mathbf{y}(t-i)-\boldsymbol{\Psi}_n^{\mathrm{T}}(t-i)\boldsymbol{\theta}_n\parallel^2\\ \widehat{\boldsymbol{\rho}}_{n|k}(t)&=\frac{1}{L_k(t)}\sum_{i=0}^{t-1}\lambda_k^i[\mathbf{y}(t-i)-\boldsymbol{\Psi}_n^{\mathrm{T}}(t-i)\widehat{\boldsymbol{\theta}}_{n|k}(t)]\\ &\quad\times[\mathbf{y}(t-i)-\boldsymbol{\Psi}_n^{\mathrm{T}}(t-i)\widehat{\boldsymbol{\theta}}_{n|k}(t)]^{\mathrm{T}}\end{aligned}\tag{4}$$

where λ_k, $0<\lambda_k<1$ denotes the so-called forgetting constant (the subscript k is needed to differentiate between several candidate forgetting factors as discussed in Section III) and

$$L_k(t)=\sum_{i=0}^{t-1}\lambda_k^i=\frac{1-\lambda_k^t}{1-\lambda_k}\tag{5}$$

is the effective width of the exponential window quantifying the estimation memory of the EWLS tracker. In steady state, i.e., for large values of t, the effective window width converges to a constant value $L_k(\infty)=1/(1-\lambda_k)$.

The EWLS estimates can be computed recursively which allows for real-time applications. Moreover, the computations can be arranged in an order-recursive way, which means that all lower-order models $\widehat{\boldsymbol{\theta}}_{n|k}(t)$, $\widehat{\boldsymbol{\rho}}_{n|k}(t)$, $n=1,\ldots,N-1$, can be obtained in the course of estimation of the highest-order model $\widehat{\boldsymbol{\theta}}_{N|k}(t)$, $\widehat{\boldsymbol{\rho}}_{N|k}(t)$ [23].

When the identified system is nonstationary and its identification is carried out using the EWLS approach, two important design decisions must be taken. First, the system order(s) should be chosen appropriately. If the number of estimated coefficients is too small, i.e., the order is underestimated, the obtained system model may fail to correctly describe system dynamics. If the order is overestimated, i.e., if some superfluous coefficients are estimated, the descriptive (e.g. predictive) capabilities of the model also deteriorate - the fact well known in statistics [22]. Second, the estimation memory of the parameter tracking algorithm should be chosen so as to match the degree of system nonstationarity[1], trading off the bias and variance components of the mean squared parameter tracking error. The effective memory $L_k(t)$ should be large when parameters vary slowly with time, and small in the presence of fast parameter changes [10], [24]. If the degree of system nonstationarity changes over time, estimation memory should be selected in an adaptive fashion. This problem is often referred to as adaptive bandwidth scheduling.

III. Competitive order and bandwidth scheduling

Our approach is based on parallel estimation. Consider K time- and order-recursive EWLS algorithms, with different forgetting factors $\lambda_k, k=1,\ldots,K$, working in parallel and yielding at each time instant KN estimates: $\widehat{\boldsymbol{\theta}}_{n|k}(t)$, $\widehat{\boldsymbol{\rho}}_{n|k}(t)$, $n=1,\ldots,N$, $k=1,\ldots,K$. Within the competitive framework, one looks for the best-local values of n and k. The model adopted at the instant t has the form

$$\widehat{\boldsymbol{\theta}}_{\widehat{n}(t)|\widehat{k}(t)}(t),\quad \widehat{\boldsymbol{\rho}}_{\widehat{n}(t)|\widehat{k}(t)}(t).\tag{6}$$

When system identification/tracking is carried out using the EWLS approach, instead of the LS approach, the local model order and estimation bandwidth selection can be performed by minimizing over n and k the following generalized version of

[1]For a nonstationary autoregressive proces, nonstationarity degree can be defined in terms of the local rate of change, with respect to time, of its time-varying autocorrelation function $\mathbf{R}_{\mathrm{y}}(t,\tau)=\mathrm{E}[\mathbf{y}(t)\mathbf{y}^{\mathrm{T}}(t-\tau)]$, see [25].

the MFPE statistic [originally developed in [26] for estimation of the model order only] – see Appendix

$$\begin{aligned} \mathrm{MFPE}_{n|k}(t) &= \det \mathbf{P}_{n|k}(t) \\ \mathbf{P}_{n|k}(t) &= \frac{1+\frac{d_n}{N_k(t)}}{1-\frac{d_n}{N_k(t)}}\,\widehat{\boldsymbol{\rho}}_{n|k}(t) \end{aligned} \tag{7}$$

where $N_k(t)$ denotes the so-called equivalent width of the exponential weighting sequence [10]

$$N_k(t) = \frac{\left(\sum_{i=0}^{t-1}\lambda_k^i\right)^2}{\sum_{i=0}^{t-1}\lambda_k^i} = \frac{(1-\lambda_k^t)(1+\lambda_k)}{(1+\lambda_k^t)(1-\lambda_k)}\,. \tag{8}$$

Note that in the steady state $N_k(\infty) = (1+\lambda_k)/(1-\lambda_k)$.

Let $\mathcal{K} = \{1,\ldots,K\}$ and $\mathcal{N} = \{1,\ldots,N\}$. The combined order-bandwidth optimization can be performed using the following rule

$$\{\widehat{n}(t), \widehat{k}(t)\} = \arg\min_{\substack{n\in\mathcal{N}\\ k\in\mathcal{K}}} \mathrm{MFPE}_{n|k}(t). \tag{9}$$

As to the model order selection, the same (asymptotically) results can be obtained by minimization, over n, of the generalized AIC statistic [26]

$$\begin{aligned} \mathrm{AIC}_{n|k}(t) &= L_k(t)\log\det\widehat{\boldsymbol{\rho}}_{n|k}(t) + 2\frac{L_k(t)}{N_k(t)}D_n \\ &\cong L_k(t)\log \mathrm{MFPE}_{n|k}(t). \end{aligned} \tag{10}$$

where the last transition follows from the fact that for $N_k(t) \gg d_n$ it holds that $\log[1 \pm d_n/N_k(t)] \cong \pm d_n/N_k(t)$.

As an alternative way of choosing the model order and estimation bandwidth, one can consider minimization of the localized version of the PLS statistic [27], [28]

$$\begin{aligned} \mathrm{PLS}_{n|k}(t) &= \det \mathbf{Q}_{n|k}(t) \\ \mathbf{Q}_{n|k}(t) &= \sum_{i=0}^{L-1} \boldsymbol{\varepsilon}_{n|k}(t-i)\boldsymbol{\varepsilon}_{n|k}^{\mathrm{T}}(t-i) \end{aligned} \tag{11}$$

where $\boldsymbol{\varepsilon}_{n|k}(t) = \mathbf{y}(t) - \boldsymbol{\Psi}_n^{\mathrm{T}}(t)\widehat{\boldsymbol{\theta}}_{n|k}(t-1)$ denotes the one-step-ahead prediction error, and $L \in [20, 50]$ is the width of the local decision window $T(t) = [t-L+1, t]$. This corresponds to choosing the model that shows the best-local predictive performance.

The PLS-based decision rule takes the form

$$\{\widehat{n}(t), \widehat{k}(t)\} = \arg\min_{\substack{n\in\mathcal{N}\\ k\in\mathcal{K}}} \mathrm{PLS}_{n|k}(t). \tag{12}$$

In addition to selection strategies based entirely on the MFPE or PLS statistics, one can consider mixed strategies, which exploit MFPE for model order selection and (subsequently) PLS for estimation bandwidth selection or *vice versa*. Such mixed strategies are two-stage procedures. When MFPE is used for order selection, and PLS for bandwidth selection, one can first look for the best value of n for each value of k, and then search for the best value of k (mixed strategy A)

$$\begin{aligned} \widehat{n}_k(t) &= \arg\min_{n\in\mathcal{N}} \mathrm{MFPE}_{n|k}(t) \\ \widehat{k}(t) &= \arg\min_{k\in\mathcal{K}} \mathrm{PLS}_{\widehat{n}_k(t)|k}(t) \end{aligned} \tag{13}$$

TABLE I
DESCRIPTION OF FOUR VARIANTS OF MIXED SELECTION

Variant	A	B	C	D
Step 1	MFPE(n)	PLS(k)	PLS(n)	MFPE(k)
Step 2	PLS(k)	MFPE(n)	MFPE(k)	PLS(n)

or first search for k and then for n (mixed strategy B)

$$\begin{aligned} \widehat{k}_n(t) &= \arg\min_{k\in\mathcal{K}} \mathrm{PLS}_{n|k}(t) \\ \widehat{n}(t) &= \arg\min_{n\in\mathcal{N}} \mathrm{MFPE}_{n|\widehat{k}_n(t)}(t). \end{aligned} \tag{14}$$

The remaining two mixed strategies (C, D), specified in Table I, correspond to the situation where order decisions are based on PLS and bandwidth decisions on MFPE.

Remark 1

Since the pioneering work of Akaike [14], [29], a large number of alternative/modified model order selection criteria have been proposed based on statistical arguments and/or on empirical evidence, see e.g. [30], [31] and references therein. Our focus on MFPE/AIC and PLS has two reasons. First, when appropriately modified both approaches can be used for joint model order and estimation bandwidth selection for the purpose of identification of nonstationary autoregressive processes. Second, both AIC and PLS have their Bayesian reinterpretations, which will be exploited for further improvement of identification results.

Remark 2

The problem of selection of bandwidth parameters for a bank of competing algorithms was studied in [12]. As shown there, to maximize robustness of the parallel estimation scheme, the effective memory spans of the competing algorithms should form a geometric progression, i.e., it should hold that

$$N_k(\infty) = \gamma N_{k-1}(\infty), \quad k = 2,\ldots,K \tag{15}$$

where $\gamma > 1$. The value of γ can be determined analytically for the assumed acceptable level of "insignificant increase" (compared with the ideal switching scheme) of the mean squared parameter estimation error in the case where the optimal bandwidth is unknown but constant, and for the assumed degree of smoothness of parameter trajectory. For the acceptable relative performance degradation equal to 10%, the values of γ range from 1.57 (for smooth deterministic parameter trajectories) to 2.43 (for random-walk type trajectories). If nothing is known about parameter variation, our recommended default value is $\gamma = 2$ (memory doubling technique).

As to selection of the number of algorithms working in parallel, the typical choices are $K = 2$ (short-memory algorithm and long-memory algorithm) and $K = 3$ (short-memory algorithm, nominal-memory algorithm, and long-memory algorithm). Computational load of the parallel estimation scheme grows linearly with K. We note, however, that it can be significantly reduced (almost to the load of a single algorithm) using the postfiltering technique described in [32].

IV. Collaborative order and bandwidth scheduling

Collaborative strategy is based on Bayesian reasoning. In this framework all unknown parameters (such as n and/or k) are regarded as random variables with assigned *a priori* distributions. Consequently, Bayesian estimates of quantities that depend on such parameters often take the form of a weighted sum of conditional estimates with weights equal to appropriately defined posterior probabilities.

Consider the situation where the bandwidth parameter k is fixed and the model order n is unknown. The optimal, in the mean square sense, Bayesian predictor can be expressed in the form

$$\widehat{\mathbf{y}}_{\mathcal{N}|k}(t+1|t) = \sum_{n\in\mathcal{N}} \mu_{n|k}(t)\widehat{\mathbf{y}}_{n|k}(t+1|t) \tag{16}$$

where

$$\begin{aligned}\widehat{\mathbf{y}}_{n|k}(t+1|t) &= \mathbf{\Psi}_n^{\mathrm{T}}(t+1)\widehat{\boldsymbol{\theta}}_{n|k}(t)\\ &= \sum_{i=1}^{n} \widehat{\mathbf{A}}_{i,n|k}(t)\mathbf{y}(t-i+1) + \sum_{i=1}^{n} \widehat{\mathbf{B}}_{i,n|k}(t)\mathbf{u}(t-i+1)\end{aligned} \tag{17}$$

denotes predictor corresponding to the model of order n, and $\mu_{n|k}(t) \geq 0, \quad n \in \mathcal{N}, \quad \sum_{n\in\mathcal{N}} \mu_{n|k}(t) = 1$, denote *a posteriori* probabilities of different model orders given the data set $\mathcal{Y}(t) = \{\mathbf{y}(i), i \leq t\}$ available at the instant t.

According to (16), the optimal predictor is the weighted linear (convex) combination of competitive predictors with the weights reflecting our confidence in that n is the right choice of the model order. Unlike the competitive, "hard selection" case discussed in the previous section, where the predictor is seeked in the form $\widehat{\mathbf{y}}_{\widehat{n}(t)|k}(t+1|t)$, the Bayesian solution is collaborative, i.e., "soft". Since it takes into account uncertainty embedded in the selection process (especially when several competing models demonstrate similar predictive capabilities), some performance improvements over the model switching strategy can be expected.

Note that the Bayesian formula (16) can be rewritten as

$$\begin{aligned}\widehat{\mathbf{y}}_{\mathcal{N}|k}(t+1|t) &= \mathbf{\Psi}_N^{\mathrm{T}}(t+1)\bar{\boldsymbol{\theta}}_{\mathcal{N}|k}(t)\\ &= \sum_{i=1}^{N} \bar{\mathbf{A}}_{i,N|k}(t)\mathbf{y}(t-i+1) + \sum_{i=1}^{N} \bar{\mathbf{B}}_{i,N|k}(t)\mathbf{u}(t-i+1)\end{aligned} \tag{18}$$

where $\bar{\boldsymbol{\theta}}_{\mathcal{N}|k}(t) = \mathrm{vec}\{[\bar{\mathbf{A}}_{1,N|k}(t), \ldots, \bar{\mathbf{A}}_{N,N|k}(t), \bar{\mathbf{B}}_{1,N|k}(t), \ldots, \bar{\mathbf{B}}_{N,N|k}(t)]^{\mathrm{T}}\}$ denotes the D_N-dimensional vector of parameters of the averaged system model

$$\begin{aligned}\bar{\mathbf{A}}_{i,N|k}(t) &= \sum_{n=i}^{N} \mu_{n|k}(t)\widehat{\mathbf{A}}_{i,n|k}(t)\\ \bar{\mathbf{B}}_{i,N|k}(t) &= \sum_{n=i}^{N} \mu_{n|k}(t)\widehat{\mathbf{B}}_{i,n|k}(t)\\ i &= 1, \ldots, N.\end{aligned} \tag{19}$$

The corresponding estimate of $\boldsymbol{\rho}(t)$ can be obtained from

$$\bar{\boldsymbol{\rho}}_{\mathcal{N}|k}(t) = \sum_{n=1}^{N} \mu_{n|k}(t)\widehat{\boldsymbol{\rho}}_{n|k}(t). \tag{20}$$

As argued by Akaike in his Bayesian extension of the AIC criterion [19], [20], in the stationary case and under noninformative [33] *a priori* distribution $\pi(\cdot)$ imposed on n [$\pi(n) = 1/N, n \in \mathcal{N}$], one can adopt

$$\mu_n(t) \propto \exp\left[-\frac{1}{2}\mathrm{AIC}_n(t)\right] \tag{21}$$

where $\mathrm{AIC}_n(t)$ denotes Akaike's statistic [which can be obtained from (10) after setting $\lambda_k = 1$] and the symbol $\propto$ denotes proportionality (with constant of proportionality that does not depend on n). Akaike called such posteriors model likelihoods. In the nonstationary case model likelihoods can be defined as

$$\mu_{n|k}(t) \propto \exp\left[-\frac{1}{2}\mathrm{AIC}_{n|k}(t)\right] \cong \left[\det \mathbf{P}_{n|k}(t)\right]^{-L_k(t)/2} \tag{22}$$

where the last transition follows from (7) and (10).

Remark 1

The averaged model has always the maximum order N, which means that when $N > n$, it is overparametrized. In spite of this, the predictor $\widehat{\mathbf{y}}_{\mathcal{N}|k}(t+1|t)$ based on such a model (which differs from the predictor $\widehat{\mathbf{y}}_{N|k}(t+1|t)$, based exclusively on the maximum-order model) performs better, in the mean square sense, than the predictor $\widehat{\mathbf{y}}_{\widehat{n}(t)|k}(t+1|t)$ which relies on the point estimate of n. This seeming contradiction shows clearly that the problems of model order estimation and model parameter estimation should be considered and solved jointly.

The estimation bandwidth uncertainty can be dealt with in an analogous way as model order uncertainty – rather than selecting "the most appropriate" model $\bar{\boldsymbol{\theta}}_{\mathcal{N}|\widehat{k}(t)}(t)$ from the set $\bar{\boldsymbol{\theta}}_{\mathcal{N}|1}(t), \ldots, \bar{\boldsymbol{\theta}}_{\mathcal{N}|K}(t)$, one can apply once more the model averaging approach. Note that the optimal Bayesian predictor can be expressed in the form

$$\widehat{\mathbf{y}}_{\mathcal{N}|\mathcal{K}}(t+1|t) = \sum_{k\in\mathcal{K}} \mu_k(t)\widehat{\mathbf{y}}_{\mathcal{N}|k}(t+1|t) \tag{23}$$

where $\mu_k(t), k \in \mathcal{K}$, $\sum_{k\in\mathcal{K}} \mu_k(t) = 1$, denote *a posteriori* probabilities of k given $\mathcal{Y}(t)$. The optimal predictor can be rewritten in terms of the averaged vector of parameter estimates

$$\widehat{\mathbf{y}}_{\mathcal{N}|\mathcal{K}}(t+1|t) = \mathbf{\Psi}_N^{\mathrm{T}}(t+1)\bar{\boldsymbol{\theta}}_{\mathcal{N}|\mathcal{K}}(t) \tag{24}$$

where

$$\bar{\boldsymbol{\theta}}_{\mathcal{N}|\mathcal{K}}(t) = \sum_{k\in\mathcal{K}} \mu_k(t)\bar{\boldsymbol{\theta}}_{\mathcal{N}|k}(t). \tag{25}$$

Likewise, the estimate of the covariance matrix of the input noise can be obtained from

$$\bar{\boldsymbol{\rho}}_{\mathcal{N}|\mathcal{K}}(t) = \sum_{k\in\mathcal{K}} \mu_k(t)\bar{\boldsymbol{\rho}}_{\mathcal{N}|k}(t). \tag{26}$$

Let

$$\mathbf{Q}_{\mathcal{N}|k}(t) = \sum_{i=0}^{L-1} \boldsymbol{\varepsilon}_{\mathcal{N}|k}(t-i)\boldsymbol{\varepsilon}_{\mathcal{N}|k}^{\mathrm{T}}(t-i) \tag{27}$$

where $\boldsymbol{\varepsilon}_{\mathcal{N}|k}(t) = \mathbf{y}(t) - \boldsymbol{\Psi}_N^{\mathrm{T}}(t)\bar{\boldsymbol{\theta}}_{\mathcal{N}|k}(t-1)$ denotes the one-step-ahead prediction error yielded by the corresponding order-averaged model. As shown in [28], which presents a Bayesian reinterpretation of the PLS criterion (based on the concept of prequential likelihood [34]), under noninformative priors imposed on k [$\pi(k) = 1/K, k = 1, \ldots, K$] and $\boldsymbol{\rho}$ [$\pi(\boldsymbol{\rho}) \propto (\det \boldsymbol{\rho})^{-(m_y+1)/2}$], the weights $\mu_k(t)$, called model credibility coefficients, can be obtained in the form

$$\mu_k(t) \propto \left[\det \mathbf{Q}_{\mathcal{N}|k}(t)\right]^{-L/2} \tag{28}$$

which resembles the Akaike's formula (22).

Remark 2

The collaborative, model averaging scheme (24), with the weights assigned according to (22) and (28), can be regarded a Bayesian extension of the mixed competitive selection scheme A. While the model credibility coefficients are applicable to *any* set of models (irrespective of the identification principles and estimation bandwidth settings), the model likelihoods are well-defined *only* for a family of EWLS-based models obtained using the same bandwidth settings. For this reason Bayesian extension of the mixed competitive schemes B, C and D is not possible along the presented lines. Since, as we will demonstrate in Section V, performance of scheme A is comparable with that of scheme B, and much better than performance of schemes C and D, the above-mentioned limitation does not seem to be serious.

V. Parametric spectrum estimation

Setting $\mathbf{u}(t) \equiv 0$ in (1), one obtains the following model

$$\mathbf{y}(t) = \sum_{i=1}^{n} \mathbf{A}_i(t)\mathbf{y}(t-i) + \mathbf{e}(t), \quad \mathrm{cov}[\mathbf{e}(t)] = \boldsymbol{\rho}(t) \tag{29}$$

which describes the time-varying autoregressive (AR) signal. According to Dahlhaus [36], [35], [25], when AR coefficients vary smoothly with time and the model (29) is uniformly stable, the signal $\{\mathbf{y}(t)\}$ belongs to the class of locally stationary processes with well-defined evolutionary spectral representation. The corresponding instantaneous spectral density function has the form

$$\mathbf{S}(\omega, t) = \boldsymbol{\mathcal{A}}^{-1}[e^{-j\omega}, \boldsymbol{\alpha}(t)]\ \boldsymbol{\rho}(t)\ \boldsymbol{\mathcal{A}}^{-\mathrm{T}}[e^{j\omega}, \boldsymbol{\alpha}(t)] \tag{30}$$

where $j = \sqrt{-1}$, $\omega \in (-\pi, \pi]$ denotes the normalized angular frequency, $\boldsymbol{\alpha}(t) = \mathrm{vec}\{[\mathbf{A}_1(t), \ldots, \mathbf{A}_n(t)]^{\mathrm{T}}\}$ denotes the vector of all AR coefficients and

$$\boldsymbol{\mathcal{A}}[e^{-j\omega}, \boldsymbol{\alpha}(t)] = \mathbf{I} - \sum_{i=1}^{n} \mathbf{A}_i(t)e^{-ij\omega}. \tag{31}$$

The theory of locally stationary processes developed by Dahlhaus, based on the concept of infill asymptotics[2], is statistically more consistent than the earlier attempt made by Priestley [37]. In particular, the time-varying spectral density function (30) is uniquely defined in the rescaled time domain, the property that does not hold in the non-rescaled case considered by Priestley.

[2]In this framework a fixed-length time interval is sampled over a finer and finer grid of points as the sample size increases, which results in a triangular array of increasingly stationary processes.

Based on (29), the competitive parametric spectrum estimate can be expressed in the form

$$\begin{aligned}\widehat{\mathbf{S}}(\omega, t) = \boldsymbol{\mathcal{A}}^{-1}[e^{-j\omega}, &\widehat{\boldsymbol{\alpha}}_{\widehat{n}(t)|\widehat{k}(t)}(t)]\ \widehat{\boldsymbol{\rho}}_{\widehat{n}(t)|\widehat{k}(t)}(t) \\ &\times \boldsymbol{\mathcal{A}}^{-\mathrm{T}}[e^{j\omega}, \widehat{\boldsymbol{\alpha}}_{\widehat{n}(t)|\widehat{k}(t)}(t)]\end{aligned} \tag{32}$$

where $\widehat{\boldsymbol{\alpha}}_{n|k}(t) = \mathrm{vec}\{[\widehat{\mathbf{A}}_{1,n|k}(t), \ldots, \widehat{\mathbf{A}}_{n,n|k}(t)]^{\mathrm{T}}\}$, and the collaborative one takes the form

$$\begin{aligned}\bar{\mathbf{S}}(\omega, t) = \boldsymbol{\mathcal{A}}^{-1}[e^{-j\omega}, &\bar{\boldsymbol{\alpha}}_{\mathcal{N}|\mathcal{K}}(t)]\ \bar{\boldsymbol{\rho}}_{\mathcal{N}|\mathcal{K}}(t) \\ &\times \boldsymbol{\mathcal{A}}^{-\mathrm{T}}[e^{j\omega}, \bar{\boldsymbol{\alpha}}_{\mathcal{N}|\mathcal{K}}(t)]\end{aligned} \tag{33}$$

where $\widehat{\boldsymbol{\rho}}_{\mathcal{N}|\mathcal{K}}(t)$ is given by (26) and [cf. (25)]

$$\begin{aligned}\bar{\boldsymbol{\alpha}}_{\mathcal{N}|\mathcal{K}}(t) &= \mathrm{vec}\{[\bar{\mathbf{A}}_{1,N|\mathcal{K}}(t), \ldots, \bar{\mathbf{A}}_{N,N|\mathcal{K}}(t)]^{\mathrm{T}}\} \\ \bar{\mathbf{A}}_{i,N|\mathcal{K}}(t) &= \sum_{k\in\mathcal{K}} \mu_k(t)\bar{\mathbf{A}}_{i,N|k}(t), \quad i = 1, \ldots, N.\end{aligned} \tag{34}$$

Remark

As shown in [13], minimization of the FPE statistic is equivalent to minimization of the mean-square log spectral distortion measure evaluated for gain normalized spectra. This is another good reason to use MFPE in the spectral estimation case.

VI. Computer simulations

A. System identification

Performance of the proposed joint order and bandwidth selection methods was checked by means of computer simulation. Dynamics of the simulated three-input three-output ($m_y = m_u = 3$) ARX system was based on three stable time-invariant "anchor" models M_1, M_2, and M_3 of orders 1, 2 and 3, respectively. For M_3 we adopted the model of a rotary cement kiln (established experimentally) presented in [38]. This model is characterized by 3 matrices of autoregressive coefficients $\mathbf{A}_1^0, \mathbf{A}_2^0, \mathbf{A}_3^0$ and 3 matrices of input coefficients $\mathbf{B}_1^0, \mathbf{B}_2^0, \mathbf{B}_3^0$, specified in Table II. The anchor model M_2 was obtained from M_3 by zeroing its highest-order coefficients ($\mathbf{A}_3^0 = \mathbf{B}_3^0 = 0$), and the anchor model M_1 – by retaining only the first-order coefficients ($\mathbf{A}_3^0 = \mathbf{B}_3^0 = \mathbf{A}_2^0 = \mathbf{B}_2^0 = 0$). All anchor models specified above are stable.

The time-varying ARX model was obtained by morphing anchor model M_1 into M_2 and anchor model M_2 into M_3. Transition from the model M_1, valid at the instant t_1, to the model M_2, valid at the instant t_2, was realized using the following transformations

$$\mathbf{A}_2(t) = \mu(t)\mathbf{A}_2^0, \quad \mathbf{B}_2(t) = \mu(t)\mathbf{B}_2^0, \quad t \in [t_1, t_2] \tag{35}$$

where $\mu(t) = (t - t_1)/(t_2 - t_1)$. The remaining parameters were kept constant: $\mathbf{A}_1(t) = \mathbf{A}_1^0$, $\mathbf{B}_1(t) = \mathbf{B}_1^0$, $t \in [t_1, t_2]$. Transition from the model M_2, valid at the instant t_3, to the model M_3, valid at the instant t_4, was realized in an analogous way, namely

$$\mathbf{A}_3(t) = \eta(t)\mathbf{A}_3^0, \quad \mathbf{B}_3(t) = \eta(t)\mathbf{B}_3^0, \quad t \in [t_3, t_4] \tag{36}$$

where $\eta(t) = (t - t_3)/(t_4 - t_3)$, and $\mathbf{A}_1(t) = \mathbf{A}_1^0$, $\mathbf{B}_1(t) = \mathbf{B}_1^0$, $\mathbf{A}_2(t) = \mathbf{A}_2^0$, $\mathbf{B}_2(t) = \mathbf{B}_2^0$, $t \in [t_3, t_4]$.

TABLE II
COEFFICIENTS OF THE ANCHOR MODEL M_3

$\mathbf{A}_1^0$			$\mathbf{B}_1^0$		
0.692	0.002	-0.009	0.056	-0.609	1.577
0.256	0.295	-0.001	0.042	-0.844	0.110
-0.193	-0.021	0.666	0.057	2.850	-0.502
$\mathbf{A}_2^0$			$\mathbf{B}_2^0$		
0.260	0.005	0.001	0.032	1.044	-1.128
-0.654	0.147	0.084	-0.655	11.988	-0.266
0.275	-0.002	-0.019	0.100	-3.541	0.757
$\mathbf{A}_3^0$			$\mathbf{B}_3^0$		
0.039	0.005	-0.012	-0.122	-0.524	-0.511
0.282	0.029	-0.036	0.315	7.169	0.035
0.120	0.000	0.128	-0.048	1.234	-0.669

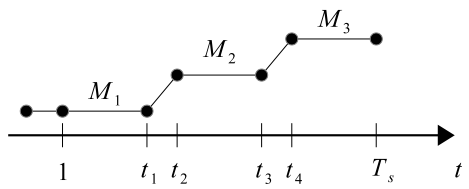

Fig. 1. Morphing scenario used in simulation tests

The applied morphing scenario is symbolically depicted in Fig. 1. The identified system, analyzed in the interval $[1, T_s]$, had 3 periods of time-invariance (M_1–M_1, M_2–M_2, M_3–M_3), each of length $l_1 = 2000$, interleaved with 2 periods of nonstationary behavior (M_1–M_2, M_2–M_3), each of length $l_2 = 500$ ($T_s = 3l_1 + 2l_2 = 7000$). The coordinates of the breakpoints, marked by bullets in Fig. 1, were equal to: $t_1 = 2000$, $t_2 = 2500$, $t_3 = 4500$ and $t_4 = 5000$. Data generation was started 1000 instants prior to $t = 1$ so that for all competing algorithms estimation and evaluation of the results could be started at the instant $t = 1$.

Similarly as in [38], the pseudo-random binary type sequence with magnitude $|u_1(t)| = |u_2(t)| = |u_3(t)| = u_0, \forall t$, $u_0 = 0.1$, and covariance matrix $\mathrm{cov}[\mathbf{u}(t)] = u_0^2\mathbf{I}_3$ (the same in all experiments) was used as an observable input signal. The unobservable noise sequence $\{\mathbf{e}(t)\}$, white and independent of $\{\mathbf{u}(t)\}$, was Gaussian: $\mathbf{e}(t) \sim \mathcal{N}(0, \sigma_e^2\mathbf{I}_3)$, $\sigma_e = 0.01$.

As a performance measure, quantifying the tracking capabilities of different estimation algorithms, the squared parameter tracking error $d_{\mathrm{PAR}}(t) = ||\widehat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}(t)||^2$ was used. Evaluation was based on comparison of the mean scores obtained after combined time and ensemble averaging of $d_{\mathrm{PAR}}(t)$ (over $t \in [1, T_s]$ and 100 independent realizations of $\{\mathbf{e}(t)\}$).

Table III shows the mean scores yielded by 3 EWLS algorithms ($\lambda_1 = 0.98, \lambda_2 = 0.99, \lambda_3 = 0.995$) run for models of different orders ($n = 1, \ldots, 10$) and by 7 adaptive order-and-bandwidth selection schemes (PLS, MFPE, A, B, C, D, Bayesian). Additionally, Figs. 2 and 3 show the evolution of the bandwidth and order selection statistics based on PLS and MFPE.

First of all, note that all adaptive bandwidth-and-order selection schemes yield much better results than their fixed-bandwidth fixed-order counterparts. Actually, note that when the model order is not underestimated ($n, N \geq 3$), performance of the best nonadaptive algorithm, observed for $\lambda_2 = 0.99$ and $n = 3$ ($\bar{d}_{\mathrm{PAR}} = 2.98$), is worse than performance of any of the proposed adaptive schemes, irrespective of the choice of the maximum order N. Note also that among all adaptive solutions the Bayesian scheme provides results that are uniformly the best.

According to Fig. 2, which shows the locally time averaged histograms of the results of bandwidth selection based on PLS and MFPE (each time bin covers 250 samples), shorter-memory algorithms are preferred in the presence of parameter variation, i.e., in the intervals $[t_1, t_2]$ and $[t_3, t_4]$; they are switched back to the longer-memory ones when system dynamics becomes time-invariant again. Fig. 3 summarizes the order selection capabilities of both adaptive schemes. Both order selection criteria appropriately react to the change of system order, which takes place at the instants t_1 (from 1 to 2) and t_3 (from 2 to 3).

Note that the bandwidth selection capabilities of the PLS-based approach are slightly better than those of the MFPE-based approach, and that the order selection capabilities of the MFPE-based approach are much better than the analogous capabilities of the PLS-based approach. This is exactly the reason why the mixed strategies A and B yield better parameter tracking results than the competitive strategies based exclusively on PLS or MFPE. The very best results are obtained when the collaborative Bayesian strategy is adopted.

Finally, we note that the conclusions reached differ from those reported in our earlier paper [12], devoted to *noncausal* identification of nonstationary AR signals, where the decision rule based on the cross-validation statistic, similar to PLS, failed to correctly identify both order and bandwidth parameters. This effect is most likely caused by the fact that, unlike prediction errors used for evaluation of causal estimation algorithms, the sequence of leave-one-out interpolation errors exploited in [12] may be strongly internally correlated.

B. Parametric spectrum estimation

Unlike the system identification experiment, the stable AR anchor models M_1, M_2 and M_3, of orders 2, 4 and 6, respectively, were specified in the lattice form. The model M_3 was obtained by means of identifying – using the Whittle-Wiggins-Robinson algorithm [22] – a fragment of a stereo audio recording ($m_y = 2$). It had the form $M_3 = \{\boldsymbol{\Delta}_1^0, \ldots, \boldsymbol{\Delta}_6^0, \mathbf{R}_0^0\}$, where $\boldsymbol{\Delta}_i^0, i = 1, \ldots, 6$ denote the matrices of normalized reflection (partial autocorrelation) coefficients and $\mathbf{R}_0^0 = \mathrm{cov}[\mathbf{y}(t)]$ – see Table IV. The remaining two models were obtained by zeroing the higher-order reflection coefficients in M_3, namely $M_2 = \{\boldsymbol{\Delta}_1^0, \ldots, \boldsymbol{\Delta}_4^0, \mathbf{R}_0^0\}$ and $M_1 = \{\boldsymbol{\Delta}_1^0, \boldsymbol{\Delta}_2^0, \mathbf{R}_0^0\}$. We note that every lattice representation $\{\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_n, \mathbf{R}_0\}$ can be uniquely transformed into the direct representation $\{\mathbf{A}_1, \ldots, \mathbf{A}_n, \boldsymbol{\rho}\}$, and *vice versa* [22].

TABLE III

COMPARISON OF PARAMETER TRACKING ERRORS OBTAINED FOR 3 FIXED-ORDER ($n = 1, \ldots, 10$) EWLS ALGORITHMS WITH DIFFERENT FORGETTING CONSTANTS $\lambda_1, \lambda_2, \lambda_3$, WITH THE RESULTS YIELDED BY 2 ORDER-AND-BANDWIDTH-ADAPTIVE PARALLEL ESTIMATION SCHEMES BASED ON THE PLS STATISTIC ($L = 30$) AND THE MFPE STATISTIC, 4 MIXED COMPETITIVE PLS/MFPE ADAPTATION SCHEMES (A, B, C, D), AND THE COLLABORATIVE BAYESIAN SCHEME BASED ON MODEL AVERAGING. THE FIRST COLUMN SHOWS THE MODEL ORDER n (FOR FIXED-ORDER SOLUTIONS) OR THE MAXIMUM MODEL ORDER N (FOR ORDER-ADAPTIVE SOLUTIONS). FOR $n, N \geq 3$ THE BEST RESULTS IN EACH ROW ARE SHOWN IN BOLDFACE.

ARX modeling – PAR measure

n/N	λ_1	λ_2	λ_3	PLS	MFPE	A	B	C	D	Bayesian
1	133.13	132.02	131.22	132.50	132.38	132.50	132.50	132.38	132.38	132.42
2	25.29	25.06	24.83	25.14	25.17	25.13	25.13	25.18	25.19	25.12
3	3.28	2.98	4.47	1.04	0.90	0.87	0.90	1.11	1.16	**0.85**
4	8.52	6.64	7.81	1.40	0.99	0.90	0.97	1.62	1.72	**0.86**
5	14.18	10.71	11.98	1.57	1.05	0.90	1.04	1.84	2.00	**0.86**
6	20.14	14.83	15.88	1.67	1.10	0.90	1.09	1.95	2.17	**0.86**
7	26.32	18.92	19.71	1.74	1.13	0.90	1.15	2.00	2.28	**0.86**
8	32.86	23.21	23.66	1.80	1.15	0.90	1.20	2.02	2.37	**0.86**
9	39.62	27.61	27.78	1.84	1.17	0.90	1.25	2.04	2.43	**0.86**
10	46.84	32.24	31.99	1.87	1.18	0.90	1.31	2.05	2.47	**0.86**

‘

TABLE V

COMPARISON OF PARAMETER TRACKING ERRORS OBTAINED FOR 3 FIXED-ORDER ($n = 1, \ldots, 10$) EWLS ALGORITHMS WITH DIFFERENT FORGETTING CONSTANTS $\lambda_1, \lambda_2, \lambda_3$, WITH THE RESULTS YIELDED BY 2 ORDER-AND-BANDWIDTH-ADAPTIVE PARALLEL ESTIMATION SCHEMES BASED ON THE PLS STATISTIC ($L = 30$) AND THE MFPE STATISTIC, 4 MIXED COMPETITIVE PLS/MFPE ADAPTATION SCHEMES (A, B, C, D), AND THE COLLABORATIVE BAYESIAN SCHEME BASED ON MODEL AVERAGING. THE FIRST COLUMN SHOWS THE MODEL ORDER n (FOR FIXED-ORDER SOLUTIONS) OR THE MAXIMUM MODEL ORDER N (FOR ORDER-ADAPTIVE SOLUTIONS). FOR $n, N \geq 6$ THE BEST RESULTS IN EACH ROW ARE SHOWN IN BOLDFACE.

AR modeling – PAR measure

n/N	λ_1	λ_2	λ_3	PLS	MFPE	A	B	C	D	Bayesian
1	38.76	38.76	38.76	38.75	38.76	38.75	38.75	38.76	38.76	38.75
2	25.68	25.58	25.53	25.60	25.60	25.60	25.60	25.60	25.60	25.59
3	17.11	16.77	16.66	16.76	16.75	16.73	16.73	16.74	16.76	16.69
4	10.59	10.02	9.84	10.01	9.89	9.81	9.82	10.00	10.04	9.79
5	4.37	3.38	3.23	3.05	3.02	2.92	2.91	3.03	3.06	2.86
6	3.46	1.97	1.63	1.36	1.30	1.07	1.10	1.38	1.41	**1.02**
7	4.92	2.69	2.00	1.56	1.36	1.17	1.19	1.58	1.59	**1.08**
8	6.41	3.45	2.40	1.72	1.41	1.23	1.26	1.73	1.75	**1.13**
9	7.91	4.20	2.77	1.87	1.44	1.26	1.32	1.85	1.89	**1.16**
10	9.47	4.97	3.14	2.00	1.46	1.28	1.36	1.95	2.02	**1.18**

‘

TABLE VI

COMPARISON OF SPECTRAL ESTIMATION ERRORS OBTAINED FOR 3 FIXED-ORDER ($n = 1, \ldots, 10$) EWLS ALGORITHMS WITH DIFFERENT FORGETTING CONSTANTS $\lambda_1, \lambda_2, \lambda_3$, WITH THE RESULTS YIELDED BY 2 ORDER-AND-BANDWIDTH-ADAPTIVE PARALLEL ESTIMATION SCHEMES BASED ON THE PLS STATISTIC ($L = 30$) AND THE MFPE STATISTIC, 4 MIXED COMPETITIVE PLS/MFPE ADAPTATION SCHEMES (A, B, C, D), AND THE COLLABORATIVE BAYESIAN SCHEME BASED ON MODEL AVERAGING. THE FIRST COLUMN SHOWS THE MODEL ORDER n (FOR FIXED-ORDER SOLUTIONS) OR THE MAXIMUM MODEL ORDER N (FOR ORDER-ADAPTIVE SOLUTIONS). FOR $n, N \geq 6$ THE BEST RESULTS IN EACH ROW ARE SHOWN IN BOLDFACE.

AR modeling – RER measure

n/N	λ_1	λ_2	λ_3	PLS	MFPE	A	B	C	D	Bayesian
1	1.835	1.761	1.730	1.800	1.809	1.800	1.800	1.809	1.809	1.784
2	0.682	0.614	0.586	0.632	0.643	0.631	0.631	0.643	0.643	0.621
3	0.442	0.365	0.337	0.374	0.391	0.373	0.372	0.390	0.390	0.365
4	0.268	0.183	0.158	0.180	0.197	0.176	0.174	0.197	0.197	0.172
5	0.205	0.109	0.083	0.093	0.114	0.091	0.088	0.112	0.110	0.088
6	0.205	0.094	0.064	0.069	0.087	0.064	**0.061**	0.088	0.083	**0.061**
7	0.238	0.107	0.070	0.072	0.089	0.066	0.063	0.091	0.084	**0.062**
8	0.274	0.121	0.077	0.075	0.090	0.067	0.064	0.093	0.085	**0.063**
9	0.312	0.136	0.083	0.077	0.090	0.067	0.066	0.095	0.085	**0.063**
10	0.352	0.150	0.090	0.079	0.091	0.068	0.067	0.097	0.086	**0.064**

‘

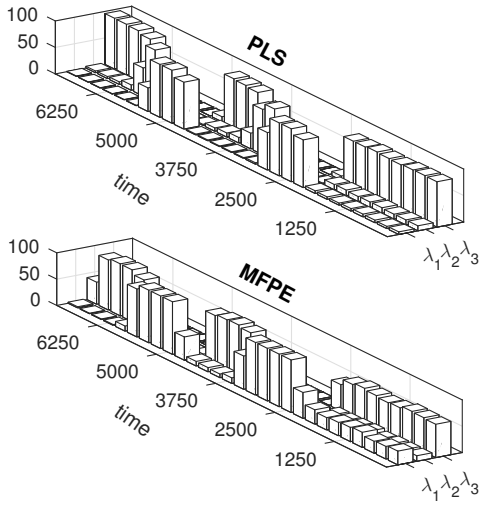

Fig. 2. Histograms of the results of estimation bandwidth selection, obtained for 100 process realizations ($T_s = 7000$).

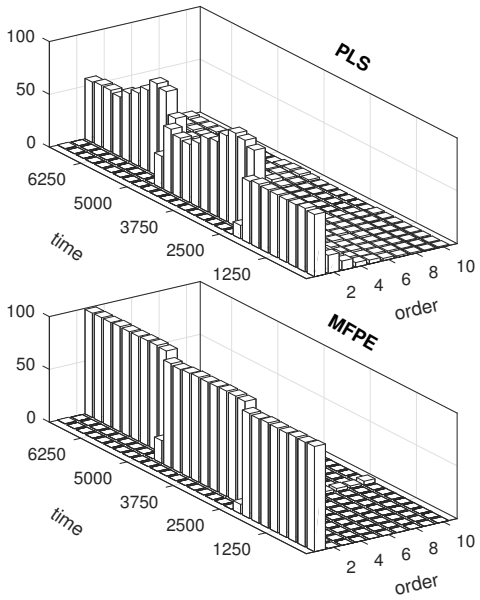

Fig. 3. Histograms of model order estimates, obtained for 100 process realizations ($T_s = 7000$).

The identified process, analyzed in the interval $[1, T_s]$, had 5 periods of time-invariance (M_1–M_1, M_2–M_2, M_3–M_3, M_2–M_2, M_1–M_1), each of length $l_1 = 2000$, interleaved with 2 periods of nonstationary behavior (M_1–M_2, M_2–M_3), each of length $l_2 = 500$ ($T_s = 5l_1 + 2l_2 = 11000$), and with 2 abrupt changes (M_3–M_2, M_2–M_1). The coordinates of the breakpoints, marked by bullets in Fig. 4, were equal to: $t_1 = 2000$, $t_2 = 2500$, $t_3 = 4500$, $t_4 = 5000$, $t_5 = 7000$, and $t_6 = 9000$.

Similarly as in the system identification experiment, transition between different models was realized by morphing, with a constant speed, one set of reflection coefficients into another one. For example, the transition from the model M_2, valid at

TABLE IV
LATTICE FORM OF THE ANCHOR MODEL M_3

$\mathbf{R}_0^0$

0.006	0.007
0.007	0.011

$\boldsymbol{\Delta}_1^0$

0.987	-0.040
0.044	0.991

$\boldsymbol{\Delta}_2^0$

-0.740	-0.353
0.158	-0.768

$\boldsymbol{\Delta}_3^0$

0.529	0.329
-0.243	0.496

$\boldsymbol{\Delta}_4^0$

-0.590	0.013
0.022	-0.430

$\boldsymbol{\Delta}_5^0$

0.553	0.361
-0.322	0.385

$\boldsymbol{\Delta}_6^0$

-0.322	-0.093
-0.048	-0.378

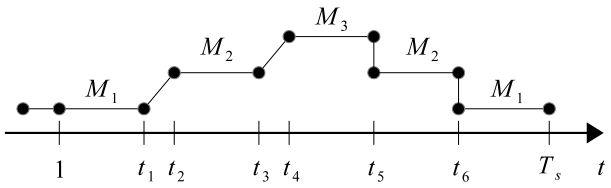

Fig. 4. Morphing scenario used in simulation tests.

the instant t_3, to the model M_3, valid at the instant t_4, was realized using the following transformations

$$\begin{aligned} \mathbf{R}_0(t) &= \mathbf{R}_0^0 \\ \boldsymbol{\Delta}_i(t) &= \boldsymbol{\Delta}_i^0 \\ & \quad i = 1, \ldots, 4, \\ \boldsymbol{\Delta}_i(t) &= \mu(t)\boldsymbol{\Delta}_i^0 \\ & \quad i = 5, \ldots, 6, \qquad t \in [t_3, t_4] \end{aligned} \tag{37}$$

where $\mu(t) = (t - t_3)/(t_4 - t_3)$. The resulting time-varying reflection coefficients $\boldsymbol{\Delta}_1(t), \ldots, \boldsymbol{\Delta}_6(t)$ were then transformed into autoregressive coefficients $\mathbf{A}_1(t), \ldots, \mathbf{A}_6(t)$ and the driving noise covariance $\boldsymbol{\rho}(t)$. As explained in [5], the time-variant model obtained in such a way is uniformly stable, the property that generally does not hold true if morphing is applied directly to the matrices of autoregressive coefficients. The evolutionary spectrum of the simulated nonstationary AR process is shown in Fig. 5.

Two performance measures were used to evaluate simulation results: the parameter tracking error $d_{\mathrm{PAR}}(t)$, defined earlier, and the relative entropy rate

$$\begin{aligned} d_{\mathrm{RER}}(t) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \Big\{ & \mathrm{tr}\left[\left(\mathbf{S}(\omega, t) - \widehat{\mathbf{S}}(\omega, t) \right) \widehat{\mathbf{S}}^{-1}(\omega, t) \right] \\ & - \log\det\left[\mathbf{S}(\omega, t) \widehat{\mathbf{S}}^{-1}(\omega, t) \right] \Big\} \, d\omega \end{aligned} \tag{38}$$

which is an extension, to the multivariate case, of the well-known Itakura-Saito spectral distortion measure. The results of comparison of different approaches, gathered in Table V (PAR) and Table VI (RER), stay in agreement with those obtained

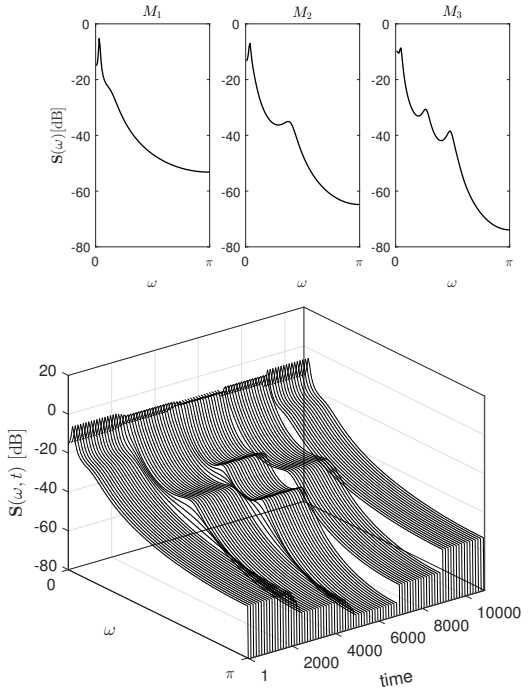

Fig. 5. Single channel spectra of stationary anchor processes (three upper plots) and single channel evolutionary spectrum of the simulated nonstationary autoregressive process (lower plot).

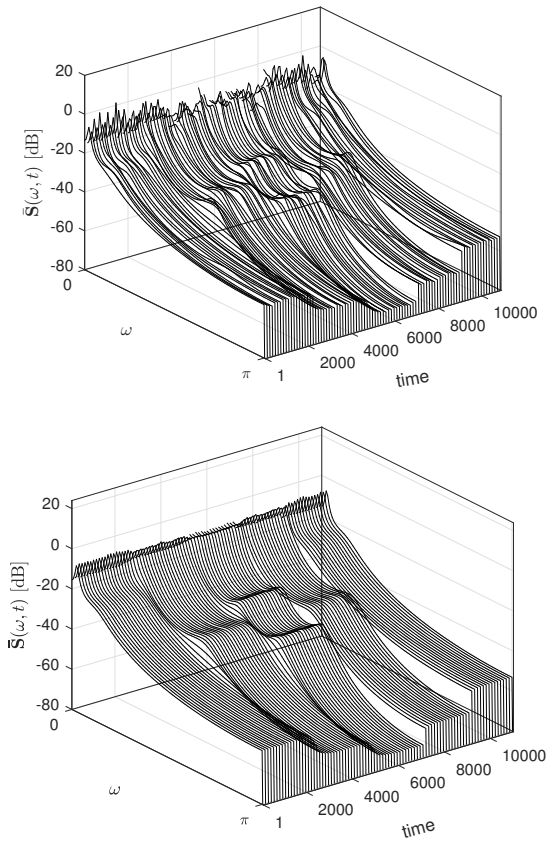

Fig. 6. Estimated evolutionary spectrum of the simulated nonstationary autoregressive process obtained for a single process realization (upper figure) and averaged over 100 realizations (lower figure).

earlier in the system identification case: if the order is not underestimated ($N \geq 6$) the Bayesian scheme provides results that are uniformly the best according to both performance measures.

Estimates of the evolutionary spectrum, obtained for a typical simulation run of the Bayesian approach, and averaged over 100 process realizations, are shown in Fig. 6.

VII. Conclusion

The problem of causal identification of nonstationary multivariate autoregressive processes, with unknown order and unknown rate of parameter variation, was considered and solved by combining results yielded by several parameter trackers with different order and bandwidth settings working in parallel. Two decision approaches were considered, based on the localized versions of Akaike's final prediction error (FPE) and the predictive least squares (PLS) statistics, respectively. It was shown that the best results can be obtained when both approaches are combined, namely, when FPE is used for model order selection, and PLS – for estimation bandwidth selection. On a qualitative level these findings differ from the conclusions reached in our earlier papers devoted to noncausal identification of nonstationary autoregressive signals. Finally, it was shown that even better results can be obtained if the competitive estimation strategy is replaced with the collaborative one.

The proposed identification schemes usually outperform the fixed-order fixed-bandwidth algorithms they are made up of and can be realized in a computationally affordable way.

VIII. Acknowledgements

This work was partially supported by the National Science Center under the agreement UMO-2015/17/B/ST7/03772. Computer simulations were carried out at the Academic Computer Centre in Gdańsk.

References

[1] A. Elnagar and K. Gupta, "Motion prediction of moving objects based on autoregressive model," *IEEE Trans. Syst. Man Cyb.*, Part A, vol. 28, pp. 803-810, 1998.

[2] A.M. Souza, R.W. Samohyl, and C.O. Malavé, "Multivariate feed back control: an application in productive processes," *Computers & Industrial Eng.*, vol. 46, pp. 837-850, 2004.

[3] K.E. Baddour, N.C. Beaulieu, "Autoregressive models for fading channel simulation," IEEE Trans. Wireless Comm., vol. 4, pp. 1650–1662, 2005.

[4] T. Wada, M. Jinnouchi and Y. Matsumura, "Application of autoregressive modelling for the analysis of clinical and other biological data," *Ann. Inst. Statist. Math.*, vol. 40, pp. 211-227, 1998.

[5] S.G. Fabri, K.P. Camilleri, and T. Cassar, "Parametric modelling of EEG data for the identification of mental tasks," *Biomed. Eng. Trends in Electron., Commun. & Software*, A. Laskovski Ed., pp. 367-386, 2011.

[6] A. Schlögl, *The Electroencephalogram and the Adaptive Autoregressive Model: Theory and Applications*. Aachen, Germany: Shaker Verlag, 2000.

[7] A. A. Zaman, M. Ferdjallah, and A. Khamayseh, "Muscle fatigue analysis for healthy adults using TVAR model with instantaneous frequency estimation, in *Proc. IEEE 38th Southeast. Symp. Syst. Theory (SSST)*, Cookeville, TN, USA, pp. 244-247, 2006.

[8] M. Niedźwiecki, M. Ciołek, and K. Cisowski, "Elimination of impulsive disturbances from stereo audio recordings using vector autoregressive modeling and variable-order Kalman filtering," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 23, pp. 970-981, 2015.

[9] P. Lesage, F. Glangeaud and J. Mars, "Applications of autoregressive models and time-frequency analysis to the study of volcanic tremor and long-period events," *J. Volc. Geotherm. Res.*, vol. 114, pp. 391417, 2002.

[10] M. Niedźwiecki, *Identification of Time-varying Processes*. Wiley, 2000.

[11] S. Kay, *Modern Spectrum Estimation: Theory and Application*, New Jersey: Prentice-Hall, 1999.

[12] M. Niedźwiecki, M. Ciołek, and Y. Kajikawa, "On adaptive covariance and spectrum estimation of locally stationary multivariate processes," *Automatica*, vol. 82, pp. 1–12, 2017.

[13] M. Niedźwiecki, M. Meller, and D. Chojnacki, "Lattice filter based autoregressive spectrum estimation with joint model order and estimation bandwidth adaptation," *Proc. 56th IEEE Conference on Decision and Control*, Melbourne, Australia, pp. 4618–4625, 2017.

[14] H. Akaike, "Statistical predictor identification," *Ann. Inst. Statist. Math.*, vol. 22, pp. 203–217, 1970.

[15] H. Akaike, "Autoregressive model fitting for control," *Ann. Inst. Statist. Math.*, vol. 23, pp. 163–180, 1971.

[16] J. Rissanen and V. Wertz, "Structure estimation by accumulated prediction error criterion," *Proc. 7th IFAC Symposium on Identification and System Parameter Estimation*, York, U.K., pp. 757–759, 1985.

[17] J. Rissanen, "A predictive least squares principle," *IMA J. Math. Control Inform.*, vol. 3, pp. 211-222, 1986.

[18] J. Rissanen, "Order estimation by accumulated prediction errors," *J. App. Probab.*, vol. 23, pp. 55-61, 1986.

[19] H. Akaike, "On the likelihood of a time series model," *J. Roy. Statist. Soc. D*, vol. 27, pp. 217–235, 1978.

[20] H. Akaike, "A Bayesian extension of the minimum AIC procedure of autoregressive model fitting," *Biometrika*, vol. 66, pp. 237–242, 1979.

[21] M. Niedźwiecki and M. Ciołek, "Akaike's final prediction error criterion revisited," *Proc. 40th International Conference on Telecommunications and Signal Processing*, Barcelona, Spain, pp. 237–242, 2017.

[22] T. Söderström and P. Stoica, *System Identification*, Englewood Cliffs NJ: Prentice-Hall, 1988.

[23] S. Haykin, *Adaptive Filter Theory*, Upper Saddle River NJ: Prentice-Hall, 1996.

[24] R. Dahlhaus and L. Giraitis, "On the optimal segment length for parameter estimates for locally stationary time series," J. Time Series Anal., vol. 19, pp. 629-655, 1998.

[25] R. Dahlhaus, "Locally stationary processes," *Handbook Statist.*, vol. 25, pp. 1–37, 2012.

[26] M. Niedźwiecki, "On the localized estimators and generalized Akaike's criteria," *IEEE Trans. Automat. Contr.*, vol. 29, pp. 970–983, 1984.

[27] M. Niedźwiecki, "Identification of nonstationary stochastic systems using parallel estimation schemes," *IEEE Trans. Automat. Contr.*, vol. 35, pp. 329–334, 1990.

[28] M. Niedźwiecki, "Multiple model approach to adaptive filtering," *IEEE Trans. Signal Process.*, vol. 40, pp. 470–474, 1992.

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

[30] P. Stoica and Y. Selén, "Model-order selection: A review of information criterion rules," *IEEE Sign. Process. Mag.*, vol. 21, pp. 36-47, 2004.

[31] K.P. Burnham and D.R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-theoretic Approach*, New York: Springer-Verlag, 2002.

[32] M. Niedźwiecki, "Identification of time-varying systems using combined parameter estimation and filtering", *IEEE Trans. Acoust., Speech and Sign. Process.*, vol. 38, pp. 679–686, 1990.

[33] V. Peterka, "Bayesian system identification," *Automatica*, vol. 17, pp.41-53, 1981.

[34] A.P. Dawid, "Present position and potential developments: some personal view, statistical theory the prequential approach," *J. Roy. Statist. Soc. A*, vol. 147, pp. 278–292, 1984.

[35] R. Dahlhaus, "Fitting time series models to nonstationary processes," *Ann. Statist.*, vol. 25, pp. 1–37, 1997.

[36] R. Dahlhaus, "Local inference for locally stationary time series based on the empirical spectral measure," *J. Econometrics*, vol. 151, pp. 101–112, 2009.

[37] M.B. Priestley, "Evolutionary spectra and non-stationary processes," *J. R. Stat. Soc.*, Series B, vol. 27, pp. 204–237, 1965.

[38] M.A. Doustari, H. Nakamine, N. Sannomiya, K. Watanabe, and K. Shinmura, "Modeling of rotary cement kiln," *Proc. 23rd ISCIE Symposium on Stochastic Systems Theory and its Application*, Osaka, Japan, pp. 185–188, 1991.

Appendix [outline of derivation of (7)]

Denote by $\Xi(t) = \{\xi(1), \ldots, \xi(t)\}$, $\xi(i) = \{\mathbf{y}(i), \mathbf{u}(i)\}$, the data set available at the instant t, and by $\widetilde{\Xi}(t) = \{\widetilde{\xi}(1), \ldots, \widetilde{\xi}(t)\}$, $\widetilde{\xi}(i) = \{\widetilde{\mathbf{y}}(i), \widetilde{\mathbf{u}}(i)\}$ – another, independent realization of $\Xi(t)$ obtained from the analyzed system under the same experimental conditions. This means that the corresponding excitation signals $\{\widetilde{\mathbf{u}}(t)\}$ (observable) and $\{\widetilde{\mathbf{e}}(t)\}$ (unobservable) are independent realizations of $\{\mathbf{u}(t)\}$ and $\{\mathbf{e}(t)\}$, respectively.

Following [15], as an instantaneous measure of fit we will adopt the following final prediction error statistic

$$\begin{aligned}&\boldsymbol{\delta}_{n|k}(t)\\&= \mathrm{E}\left\{[\widetilde{\mathbf{y}}(t) - \widetilde{\boldsymbol{\Psi}}_n^{\mathrm{T}}(t)\widehat{\boldsymbol{\theta}}_{n|k}(t)][\widetilde{\mathbf{y}}(t) - \widetilde{\boldsymbol{\Psi}}_n^{\mathrm{T}}(t)\widehat{\boldsymbol{\theta}}_{n|k}(t)]^{\mathrm{T}}\right\}\end{aligned}$$

where the expectation is carried out with respect to $\Xi(t)$ and $\widetilde{\Xi}(t)$. According to this measure, the quality of the model is checked on an independent data set, different from that used for identification purposes.

We will derive a stationary approximation of $\boldsymbol{\delta}_{n|k}(t)$. Suppose that the analyzed system is stationary and that the sequence of regression vectors $\{\boldsymbol{\varphi}_n(t)\}$ is zero-mean, stationary and ergodic with covariance matrix $\mathrm{cov}[\boldsymbol{\varphi}_n(t)] = \boldsymbol{\Phi}_n$. Since identification is carried out using exponential forgetting, estimation results practically do not depend on very 'old' data samples, namely on samples collected $2L_k(\infty)$ time instants prior to t, or earlier [10]. This means that in fact only the local stationarity is required.

Denote by $\Delta\widehat{\boldsymbol{\theta}}_{n|k}(t) = \widehat{\boldsymbol{\theta}}_{n|k}(t) - \boldsymbol{\theta}_n$ the parameter estimation error. Observe that

$$\widetilde{\mathbf{y}}(t) - \widetilde{\boldsymbol{\Psi}}_n^{\mathrm{T}}(t)\widehat{\boldsymbol{\theta}}_{n|k}(t) = \widetilde{\mathbf{e}}(t) - \widetilde{\boldsymbol{\Psi}}_n^{\mathrm{T}}(t)\Delta\widehat{\boldsymbol{\theta}}_{n|k}(t).$$

Furthermore, since the quantities $\widetilde{\mathbf{e}}(t)$ and $\widetilde{\boldsymbol{\Psi}}_n(t)$ are mutually independent and independent of $\Delta\widehat{\boldsymbol{\theta}}_{n|k}(t)$, it holds that

$$\begin{aligned}&\boldsymbol{\delta}_{n|k}(t) =\\&= \mathrm{E}\left\{[\widetilde{\mathbf{e}}(t) - \widetilde{\boldsymbol{\Psi}}_n^{\mathrm{T}}(t)\Delta\widehat{\boldsymbol{\theta}}_{n|k}(t)][\widetilde{\mathbf{e}}(t) - \widetilde{\boldsymbol{\Psi}}_n^{\mathrm{T}}(t)\Delta\widehat{\boldsymbol{\theta}}_{n|k}(t)]^{\mathrm{T}}\right\}\\&= \boldsymbol{\rho} + \mathrm{E}\left\{\widetilde{\boldsymbol{\Psi}}_n^{\mathrm{T}}(t)\Delta\widehat{\boldsymbol{\theta}}_{n|k}(t)\Delta\widehat{\boldsymbol{\theta}}_{n|k}^{\mathrm{T}}(t)\widetilde{\boldsymbol{\Psi}}_n(t)\right\}\\&= \boldsymbol{\rho} + \mathrm{E}\left\{\widetilde{\boldsymbol{\Psi}}_n^{\mathrm{T}}(t)\mathrm{cov}[\widehat{\boldsymbol{\theta}}_{n|k}(t)]\widetilde{\boldsymbol{\Psi}}_n(t)\right\}\end{aligned}$$

where $\boldsymbol{\rho} = \mathrm{cov}[\widetilde{\mathbf{e}}(t)] = \mathrm{cov}[\mathbf{e}(t)]$. If the true system order n is not underdetermined, it can be shown that the estimator $\widehat{\boldsymbol{\theta}}_{n|k}(t)$ is (approximately) unbiased and [26]

$$\mathrm{cov}[\widehat{\boldsymbol{\theta}}_{n|k}(t)] = \frac{\boldsymbol{\rho} \otimes \boldsymbol{\Phi}_n^{-1}}{N_k(t)} + o\left(\frac{1}{N_k(t)}\right)$$

where $N_k(t)$ is given by (8). Combining the last two results, one arrives at

$$\begin{aligned}&\boldsymbol{\delta}_{n|k}(t)\\&\cong \boldsymbol{\rho} + \frac{1}{N_k(t)}\,\mathrm{E}\left\{[\mathbf{I}_{m_y} \otimes \widetilde{\boldsymbol{\varphi}}_n^{\mathrm{T}}(t)][\boldsymbol{\rho} \otimes \boldsymbol{\Phi}_n^{-1}][\mathbf{I}_{m_y} \otimes \widetilde{\boldsymbol{\varphi}}_n(t)]\right\}\\&= \boldsymbol{\rho} + \frac{1}{N_k(t)}\,\mathrm{E}\left\{\boldsymbol{\rho} \otimes [\widetilde{\boldsymbol{\varphi}}_n^{\mathrm{T}}(t)\boldsymbol{\Phi}_n^{-1}\widetilde{\boldsymbol{\varphi}}_n(t)]\right\}\\&= \boldsymbol{\rho} + \frac{1}{N_k(t)}\,\boldsymbol{\rho}\,\mathrm{tr}[\boldsymbol{\Phi}_n^{-1}\mathrm{E}\{\widetilde{\boldsymbol{\varphi}}_n(t)\widetilde{\boldsymbol{\varphi}}_n^{\mathrm{T}}(t)\}] = \left[1 + \frac{d_n}{N_k(t)}\right]\boldsymbol{\rho}\end{aligned}$$

where the second transition follows from the identity $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD})$.

Since it holds that [12]

$$\mathrm{E}[\widehat{\boldsymbol{\rho}}_{n|k}(t)] \cong \left[1 - \frac{d_n}{N_k(t)}\right] \boldsymbol{\rho}$$

the unbiased estimate of $\boldsymbol{\delta}_{n|k}(t)$ can be obtained in the form

$$\widehat{\boldsymbol{\delta}}_{n|k}(t) = \frac{1 + \frac{d_n}{N_k(t)}}{1 - \frac{d_n}{N_k(t)}} \, \widehat{\boldsymbol{\rho}}_{n|k}(t) = \mathbf{P}_{n|k}(t) \; .$$

Maciej Niedźwiecki received the M.Sc. and Ph.D. degrees from the Technical University of Gdańsk, Gdańsk, Poland and the Dr.Hab. (D.Sc.) degree from the Technical University of Warsaw, Warsaw, Poland, in 1977, 1981 and 1991, respectively. He spent three years as a Research Fellow with the Department of Systems Engineering, Australian National University, 1986-1989. In 1990 - 1993 he served as a Vice Chairman of Technical Committee on Theory of the International Federation of Automatic Control (IFAC). He is the author of the book *Identification of Time-varying Processes* **(Wiley, 2000). His main areas of research interests include system identification, statistical signal processing and adaptive systems.**

Dr. Niedźwiecki is currently a member of the IFAC committees on Modeling, Identification and Signal Processing and on Large Scale Complex Systems, and a member of the Automatic Control and Robotics Committee of the Polish Academy of Sciences (PAN). He works as a Professor and Head of the Department of Automatic Control, Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology.

Marcin Ciołek received the M.Sc. and Ph.D. degrees from the Gdańsk University of Technology (GUT), Gdańsk, Poland, in 2010 and 2017, respectively. Since 2017, he has been working as an Adjunct Professor in the Department of Automatic Control, Faculty of Electronics, Telecommunications and Informatics, GUT. His professional interests include speech, music and biomedical signal processing.