This is a post-peer-review, pre-copyedit version of an article published in **2017 40th International Conference on Telecommunications and Signal Processing (TSP)**. The final authenticated version is available online at: https://doi.org/10.1109/TSP.2017.8075977

Akaike's Final Prediction Error Criterion Revisited

Maciej Niedźwiecki and Marcin Ciołek
Faculty of Electronics, Telecommunications and Informatics, Department of Automatic Control,
Gdańsk University of Technology
Narutowicza 11/12, 80-233 Gdańsk , Poland
Email: maciekn@eti.pg.gda.pl, marcin.ciolek@pg.gda.pl

***Abstract*—When local identification of a nonstationary ARX system is carried out, two important decisions must be taken. First, one should decide upon the number of estimated parameters, i.e., on the model order. Second, one should choose the appropriate estimation bandwidth, related to the (effective) number of input-output data samples that will be used for identification/tracking purposes. Failure to make the right decisions results in the model deterioration, both in the quantitative and qualitative sense. In this paper, we show that both problems can be solved using the suitably modified Akaike's final prediction error criterion. The proposed solution is next compared with another one, based on the Rissanen's predictive least squares principle.**

***Keywords*—ARX system identification; estimation bandwidth selection; generalized Akaike's criterion; model order selection**

I. Introduction

The final prediction error (FPE) criterion was the first of two tools proposed by Akaike for the purpose of model order selection [1], [2], [3]. Since the second one, presented several years later and known as the Akaike's information criterion (AIC) [4] has deeper statistical justification and wider range of applicability than FPE, it is much more frequently used and referred to. Both criteria were derived for time-invariant systems/signals operated under stationary conditions, and whenever both can be applied they asymptotically (for large data sets) yield the same results.

When characteristics of the analyzed system vary slowly with time, i.e., when the system may be regarded as locally stationary [5], its model can be continuously updated using the local estimation approach. In this case estimation of model parameters is based entirely, or primarily, on the most recent data samples.

When local identification techniques are used, two important decisions must be taken. First, similar to the stationary case, one should decide upon the model order, i.e., on the number of estimated parameters (or on the model structure, in a more general formulation). When the order is underestimated, the model may fail to describe some important parts of system dynamics (such as certain existing resonant modes). On the other hand, when the order is overestimated, i.e., when some nonexistent or insignificant model parameters are estimated, the model becomes less accurate (principle of parsimony) and less representative, e.g. it can suggest the presence of some nonexistent resonant modes.

The second decision that must be taken when the identified system is nonstationary concerns the estimation memory of the parameter tracking algorithm, i.e., the effective number of past input/output measurements taken into account during the estimation process. Short-memory algorithms are 'fast' (yield small tracking bias) but 'inaccurate' (yield large tracking variance) whereas the long-memory algorithms are 'slow' but 'accurate'. The best results are obtained when the estimation memory of the tracking algorithm, inversely proportional to its estimation bandwidth, is selected so as to match the degree of nonstationarity of the identified system, trading off the bias and variance error components.

We will show that both problems – selection of the model order and the most appropriate estimation bandwidth – can be successfully solved using the suitably modified FPE criterion. The FPE-based approach will be also compared with another solution to the problem of joint order and bandwidth selection, based on the predictive least squares (PLS) principle.

The potential applications of the proposed methods include, among many others, modeling and equalization of time-varying communication channels [6], [7].

II. Stationary case

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

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

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 $\mathbf{A}_i$, $\mathbf{B}_i$ are the $m_y \times m_y$- and $m_y \times m_u$-dimensional matrices of autoregressive and input coefficients, respectively:

$$\mathbf{A}_i = \begin{bmatrix} \boldsymbol{\alpha}_{1i} \\ \vdots \\ \boldsymbol{\alpha}_{m_y i} \end{bmatrix}, \quad i = 1, \ldots, n_y \qquad \mathbf{B}_i = \begin{bmatrix} \boldsymbol{\beta}_{1i} \\ \vdots \\ \boldsymbol{\beta}_{m_y i} \end{bmatrix}, \quad i = 1, \ldots, n_u,$$

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

where $\boldsymbol{\alpha}_{li} = [a_{l1,i}, \ldots, a_{lm_y,i}]$ and $\boldsymbol{\beta}_{li} = [b_{l1,i}, \ldots, b_{lm_u,i}]$ for $l = 1, \ldots, m$.

Denote by $\boldsymbol{\theta}_n^j = [\boldsymbol{\alpha}_{j1}, \ldots, \boldsymbol{\alpha}_{jn_y}, \boldsymbol{\beta}_{j1}, \ldots, \boldsymbol{\beta}_{jn_u}]^{\mathrm{T}}$, where $n = \{n_y, n_u\}$, the $(n_y m_y + n_u m_u)$-dimensional vector of parameters characterizing the j-th 'channel' of the ARX system, and by $\boldsymbol{\varphi}_n(t) = [\mathbf{y}^{\mathrm{T}}(t-1), \ldots, \mathbf{y}^{\mathrm{T}}(t-n_y), \mathbf{u}^{\mathrm{T}}(t-1), \ldots, \mathbf{u}^{\mathrm{T}}(t-n_u)]^{\mathrm{T}}$ – the corresponding regression vector (the same for all channels). Using this notation, equation of the j-th channel can be put down in the form

$$y_j(t) = \boldsymbol{\varphi}_n^{\mathrm{T}}(t)\boldsymbol{\theta}_n^j + e_j(t) \tag{2}$$

and (1) can be written down more compactly as

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

where $\boldsymbol{\Psi}_n(t) = \mathbf{I} \otimes \boldsymbol{\varphi}_n(t) = \mathrm{diag}\{\boldsymbol{\varphi}_n(t), \ldots, \boldsymbol{\varphi}_n(t)\}$ (the symbol $\otimes$ denotes Kronecker product of two matrices/vectors), and $\boldsymbol{\theta}_n = [(\boldsymbol{\theta}_n^1)^{\mathrm{T}}, \ldots, (\boldsymbol{\theta}_n^{m_y})^{\mathrm{T}}]^{\mathrm{T}}$ is the vector combining all $(n_y m_y + n_u m_u)m_y$ system parameters.

Estimation of the vector $\boldsymbol{\theta}_n$ and the covariance matrix $\boldsymbol{\rho}$ can be carried out using the method of least squares (LS)

$$\widehat{\boldsymbol{\theta}}_n(t) = \arg\min_{\boldsymbol{\theta}_n} \sum_{i=1}^{t} \| \mathbf{y}(i) - \boldsymbol{\Psi}_n^{\mathrm{T}}(i)\boldsymbol{\theta}_n \|^2 \tag{4}$$

$$\widehat{\boldsymbol{\rho}}_n(t) = \frac{1}{t}\sum_{i=1}^{t}[\mathbf{y}(i) - \boldsymbol{\Psi}_n^{\mathrm{T}}(i)\widehat{\boldsymbol{\theta}}_n(t)][\mathbf{y}(i) - \boldsymbol{\Psi}_n^{\mathrm{T}}(i)\widehat{\boldsymbol{\theta}}_n(t)]^{\mathrm{T}}. \tag{5}$$

Both quantities can be computed recursively [8].

Suppose now that the model orders n_y and n_u are not known and should be also estimated. Denote by $\mathcal{N} = \{\{n_y^1, n_u^1\}, \ldots, \{n_y^N, n_u^N\}\}$ the set of N structural variants of the model (1). For example, one can set $\mathcal{N} = \{\{i, j\} : i = 1, \ldots, N_y, j = 1, \ldots, N_u\}$, $N = N_y N_u$. According to Akaike [3] the best fitting structural variant $n = \{n_y, n_u\}$ can be chosen from the set $\mathcal{N}$ by means of minimizing the following multivariate version of the FPE statistic

$$\widehat{n}(t) = \{\widehat{n}_y(t), \widehat{n}_u(t)\} = \arg\min_{n\in\mathcal{N}} \mathrm{MFPE}_n(t) \tag{6}$$

$$\mathrm{MFPE}_n(t) = \left[\frac{1 + \frac{d_n}{t}}{1 - \frac{d_n}{t}}\right]^{m_y} \det[\widehat{\boldsymbol{\rho}}_n(t)] \tag{7}$$

where $d_n = \dim[\boldsymbol{\varphi}_n(t)] = n_y m_y + n_u m_u$ denotes the number of coefficients estimated within each channel.

III. Nonstationary case

Suppose that the identified system slowly varies with time, namely, that it is governed by the following time-dependent ARX equation

$$\begin{gathered}\mathbf{y}(t) = \sum_{i=1}^{n_y} \mathbf{A}_i(t)\mathbf{y}(t-i) + \sum_{i=1}^{n_u} \mathbf{B}_i(t)\mathbf{u}(t-i) + \mathbf{e}(t) \\ \mathrm{cov}[\mathbf{e}(t)] = \boldsymbol{\rho}(t)\end{gathered} \tag{8}$$

or equivalently

$$\mathbf{y}(t) = \boldsymbol{\Psi}_n^{\mathrm{T}}(t)\boldsymbol{\theta}_n(t) + \mathbf{e}(t). \tag{9}$$

In such a case system parameters can be estimated using a suitably localized LS algorithm, such as the one based on the well-known method of exponentially weighted least squares (EWLS). To achieve the effect of forgetting or discounting '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

$$\widehat{\boldsymbol{\theta}}_{n|k}(t) = \arg\min_{\boldsymbol{\theta}_n} \sum_{i=0}^{t-1} \lambda_k^i \| \mathbf{y}(t-i) - \boldsymbol{\Psi}_n^{\mathrm{T}}(t-i)\boldsymbol{\theta}_n \|^2 \tag{10}$$

$$\begin{aligned}\widehat{\boldsymbol{\rho}}_{n|k}(t) = \frac{1}{L_k(t)}\sum_{i=0}^{t-1} \lambda_k^i [\mathbf{y}(t-i) - \boldsymbol{\Psi}_n^{\mathrm{T}}(t-i)\widehat{\boldsymbol{\theta}}_{n|k}(t)] \\ \times [\mathbf{y}(t-i) - \boldsymbol{\Psi}_n^{\mathrm{T}}(t-i)\widehat{\boldsymbol{\theta}}_{n|k}(t)]^{\mathrm{T}}\end{aligned} \tag{11}$$

where λ_k, $0 < \lambda_k < 1$ denotes the so-called forgetting constant and

$$L_k(t) = \sum_{i=0}^{t-1} \lambda_k^i = \frac{1 - \lambda_k^t}{1 - \lambda_k} \tag{12}$$

is the effective width of the exponential window, quantifying estimation memory of the EWLS tracker.

Recursive algorithm for computation of the estimates $\widehat{\boldsymbol{\theta}}_{n|k}^j(t), j = 1, \ldots, m_y$ and $\widehat{\boldsymbol{\rho}}_{n|k}(t)$ given by (10) - (11) can be summarized as follows

$$\begin{aligned}\varepsilon_{n|k}^j(t) &= y_j(t) - \boldsymbol{\varphi}_n^{\mathrm{T}}(t)\widehat{\boldsymbol{\theta}}_{n|k}^j(t-1) \\ \widehat{\boldsymbol{\theta}}_{n|k}^j(t) &= \widehat{\boldsymbol{\theta}}_{n|k}^j(t-1) + \mathbf{g}_{n|k}(t)\varepsilon_{n|k}^j(t) \\ j &= 1, \ldots, m_y \\ \mathbf{g}_{n|k}(t) &= \frac{\mathbf{P}_{n|k}(t-1)\boldsymbol{\varphi}_n(t)}{\lambda_k + \boldsymbol{\varphi}_n^{\mathrm{T}}(t)\mathbf{P}_{n|k}(t-1)\boldsymbol{\varphi}_n(t)} \\ \mathbf{P}_{n|k}(t) &= \frac{1}{\lambda_k}[\mathbf{I} - \mathbf{g}_{n|k}(t)\boldsymbol{\varphi}_n^{\mathrm{T}}(t)]\mathbf{P}_{n|k}(t-1) \\ \mathbf{Q}_{n|k}(t) &= \lambda_k \mathbf{Q}_{n|k}(t-1) \\ &\quad + [1 - \mathbf{g}_{n|k}^{\mathrm{T}}(t)\boldsymbol{\varphi}_n(t)]\boldsymbol{\varepsilon}_{n|k}(t)\boldsymbol{\varepsilon}_{n|k}^{\mathrm{T}}(t) \\ L_k(t) &= \lambda_k L_k(t-1) + 1 \\ \widehat{\boldsymbol{\rho}}_{n|k}(t) &= \frac{1}{L_k(t)}\mathbf{Q}_{n|k}(t)\end{aligned} \tag{13}$$

where

$$\begin{aligned}\boldsymbol{\varepsilon}_{n|k}(t) &= \mathbf{y}(t) - \boldsymbol{\Psi}_n^{\mathrm{T}}(t)\widehat{\boldsymbol{\theta}}_{n|k}(t-1) \\ &= [\varepsilon_{n|k}^1(t), \ldots, \varepsilon_{n|k}^{m_y}(t)]^{\mathrm{T}} \\ \mathbf{P}_{n|k}(t) &= \left[\sum_{i=0}^{t-1} \lambda_k^i \boldsymbol{\varphi}_n(t-i)\boldsymbol{\varphi}_n^{\mathrm{T}}(t-i)\right]^{-1}\end{aligned}$$

and initial conditions should be set to $\widehat{\boldsymbol{\theta}}_{n|k}^j(0) = \mathbf{0}, j = 1, \ldots, m_y$, $\mathbf{Q}_{n|k}(0) = \mathbf{O}$, $L_k(0) = 0$ and $\mathbf{P}_{n|k}(0) = c\mathbf{I}$, where $c > 0$ denotes a large constant [8]. Note that the matrix $\mathbf{P}_{n|k}(0)$ is the same for all system channels, which is a consequence of the fact that all channels share the same regression vector $\boldsymbol{\varphi}_n(t)$.

IV. Selection of the estimation bandwidth

As already mentioned in Section 1, estimation bandwidth should be chosen in accordance with the degree of nonstationarity of the identified system, which is usually unknown and which itself may change over time. Instead of looking for the optimal value of λ_k, we will attempt to choose the best value of the forgetting constant from the set of K predefined values $\Lambda = \{\lambda_1, \ldots, \lambda_K\}$.

Assuming that the structure n of the model is fixed, consider K EWLS algorithms working in parallel and yielding the estimates $\widehat{\boldsymbol{\theta}}_{n|k}(t), \widehat{\boldsymbol{\rho}}_{n|k}(t)$, $k \in \mathcal{K} = \{1, \ldots, K\}$. Our task will be to select from $\mathcal{K}$ the most appropriate value of k at the instant t, i.e., the value that minimizes the judiciously chosen instantaneous measure of fit.

A. Final prediction error based approach

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.

As an instantaneous measure of fit we will adopt the following quantity

$$\boldsymbol{\delta}_{n|k}(t) = \mathrm{E}\left\{[\widetilde{\mathbf{y}}(t) - \widetilde{\boldsymbol{\Psi}}_n(t)\widehat{\boldsymbol{\theta}}_{n|k}(t)][\widetilde{\mathbf{y}}(t) - \widetilde{\boldsymbol{\Psi}}_n(t)\widehat{\boldsymbol{\theta}}_{n|k}(t)]^{\mathrm{T}}\right\} \tag{14}$$

where the expectation is carried out with respect to $\Xi(t)$ *and* $\widetilde{\Xi}(t)$. According to (14), 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, i.e., that it is governed by (3) and that the sequence of regression vectors $\{\boldsymbol{\varphi}_n(t)\}$ is zero-mean, stationary and ergodic with covariance matrix $\mathrm{cov}[\boldsymbol{\varphi}_n(t)\} = \boldsymbol{\Phi}_0$. 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 [15]. This means that in fact only local stationarity is required.

Assume that the true system order $n = \{n_y, n_u\}$ is not underdetermined (if the adopted values of n_y and/or n_u are greater than the true values, the corresponding entries in $\boldsymbol{\theta}_n$ should be set to zero). Under the (local) stationarity assumptions made above and some technical assumptions guaranteeing finite sample invertibility of the exponentially weighted regression matrix [see e.g. [13]], it can be shown that

$$\begin{aligned} \mathrm{E}[\widehat{\boldsymbol{\theta}}_{n|k}(t)] &\cong \boldsymbol{\theta}_n \\ \mathrm{cov}[\widehat{\boldsymbol{\theta}}_{n|k}(t)] &= \frac{\boldsymbol{\rho} \otimes \boldsymbol{\Phi}_0^{-1}}{N_k(t)} + o\left(\frac{1}{N_k(t)}\right) \end{aligned} \tag{15}$$

where

$$N_k(t) = \frac{[\sum_{i=0}^{t-1} \lambda_k^i]^2}{\sum_{i=0}^{t-1} \lambda_k^{2i}} = \frac{(1-\lambda_k^t)(1+\lambda_k)}{(1+\lambda_k^t)(1-\lambda_k)} \tag{16}$$

denotes the so-called equivalent width of the exponential window, quantifying the amount of information extracted from the input-output data by means of applying the EWLS scheme.

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

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

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

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

Finally, using (15), one arrives at

$$\begin{aligned} \boldsymbol{\delta}_{n|k}(t) &\cong \boldsymbol{\rho} + \frac{1}{N_k(t)}\,\mathrm{E}\left\{[\mathbf{I} \otimes \widetilde{\boldsymbol{\varphi}}^{\mathrm{T}}(t)][\boldsymbol{\rho} \otimes \boldsymbol{\Phi}_0^{-1}][\mathbf{I} \otimes \widetilde{\boldsymbol{\varphi}}(t)]\right\} \\ &= \boldsymbol{\rho} + \frac{1}{N_k(t)}\,\mathrm{E}\left\{\boldsymbol{\rho} \otimes [\widetilde{\boldsymbol{\varphi}}^{\mathrm{T}}(t)\boldsymbol{\Phi}_0^{-1}\widetilde{\boldsymbol{\varphi}}(t)]\right\} \\ &= \boldsymbol{\rho} + \frac{1}{N_k(t)}\,\boldsymbol{\rho}\,\mathrm{tr}[\boldsymbol{\Phi}_0^{-1}\mathrm{E}\{\widetilde{\boldsymbol{\varphi}}(t)\widetilde{\boldsymbol{\varphi}}^{\mathrm{T}}(t)\}] \\ &= \left[1 + \frac{d_n}{N_k(t)}\right]\boldsymbol{\rho} \end{aligned} \tag{17}$$

where the second transition stems from the following identity $(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) = (\mathbf{AC})\otimes(\mathbf{BD})$ which holds for Kronecker products.

In a similar way, one can show that [see [15] in the univariate case]

$$\mathrm{E}[\widehat{\boldsymbol{\rho}}_{n|k}(t)] \cong \left[1 - \frac{d_n}{N_k(t)}\right]\boldsymbol{\rho}. \tag{18}$$

Combining (17) with (18), one arrives at the following estimate of $\boldsymbol{\delta}_{n|k}(t)$ based on $\Xi(t)$

$$\widehat{\boldsymbol{\delta}}_{n|k}(t) = \frac{1 + \frac{d_n}{N_k(t)}}{1 - \frac{d_n}{N_k(t)}}\,\widehat{\boldsymbol{\rho}}_{n|k}(t) \tag{19}$$

and leads to the bandwidth selection rule

$$\widehat{k}(t) = \arg\min_{k \in \mathcal{K}} \mathrm{MFPE}_{n|k}(t) \tag{20}$$

$$\begin{aligned} \mathrm{MFPE}_{n|k}(t) &= \det[\widehat{\boldsymbol{\delta}}_{n|k}(t)] \\ &= \left[\frac{1 + \frac{d_n}{N_k(t)}}{1 - \frac{d_n}{N_k(t)}}\right]^{m_y} \det[\widehat{\boldsymbol{\rho}}_{n|k}(t)] \end{aligned} \tag{21}$$

which is in fact an extension of the rule proposed by Akaike [note that (21) reduces down to (7) when $\lambda_k = 1$].

Remark

For sufficiently large values of t, the quantity $N_k(t)$ in (21) can be replaced with its steady state value $N_k(\infty) = (1 + \lambda_k)/(1 - \lambda_k)$. Note that for the typical values of λ_k, which are close to 1, it holds that $N_k(\infty) \cong 2L_k(\infty) = 2/(1 - \lambda_k)$.

To understand the bandwidth selection mechanism behind (20) - (21), note that when the identified system is nonstationary, increasing estimation memory of the EWLS algorithm (picking the value of λ_k that is closer to 1) will usually result in increased value of $\widehat{\boldsymbol{\rho}}_{n|k}(t)$, which is the consequence of increasing the bias component of parameter estimation errors. At the same time, the multiplier in (21) will decrease, reflecting decrease in the estimation variance. The value of k selected according to (20) is therefore a result of a trade-off between estimation bias and estimation variance.

B. Predictive least squares based approach

The predictive least squares (PLS) approach was originally proposed by Rissanen as a tool for estimation of an order of a stationary AR process [16], [17]. Later on the local (sliding window) PLS statistic emerged as a limiting case of the Bayesian bandwidth estimation procedure [12], [14] based on prequential analysis [9] - for nonstationary ARX processes. The corresponding decision rule has the form

$$\widehat{k}(t) = \arg\min_{k\in\mathcal{K}} \mathrm{PLS}_{n|k}(t) \tag{22}$$

$$\mathrm{PLS}_{n|k}(t) = \det\left[\sum_{i=0}^{L-1} \boldsymbol{\varepsilon}_{n|k}(t-i)\boldsymbol{\varepsilon}_{n|k}^{\mathrm{T}}(t-i)\right] \tag{23}$$

where $L \in [20, 50]$ is the width of the local decision window $T(t) = [t - L + 1, t]$. According to (22) - (23), the bandwidth selected at the instant t corresponds to the model with the best-recent predictive capabilities, namely the one which minimizes the prediction error statistic accumulated over the recent past.

V. Joint order and bandwidth selection

The bandwidth selection statistic (21) is identical with the statistic proposed in [11] for the purpose of model order selection (based on an alternative quality measure). The selection mechanism is similar to the one described above. For a fixed bandwidth parameter k, increasing the model order will increase the multiplier in (21), but, at the same time, will decrease residual errors [along with $\widehat{\boldsymbol{\rho}}_{n|k}(t)$], which is another manifestation of the bias-variance compromise. Based on the observations made above, we propose the following joint order-and-bandwidth selection rule

$$\begin{aligned}\{\widehat{n}(t), \widehat{k}(t)\} &= \{\widehat{n}_y(t), \widehat{n}_u(t), \widehat{k}(t)\} \\ &= \arg\min_{\substack{n\in\mathcal{N}\\ k\in\mathcal{K}}} \mathrm{MFPE}_{n|k}(t).\end{aligned} \tag{24}$$

As an alternative, one can consider the analogous decision rule based on the predictive least squares principle

$$\begin{aligned}\{\widehat{n}(t), \widehat{k}(t)\} &= \{\widehat{n}_y(t), \widehat{n}_u(t), \widehat{k}(t)\} \\ &= \arg\min_{\substack{n\in\mathcal{N}\\ k\in\mathcal{K}}} \mathrm{PLS}_{n|k}(t).\end{aligned} \tag{25}$$

Remark

Suppose that the family of competing models of different orders is restricted to the case where $n_y = n_u$, i.e., where the number of autoregressive coefficients is the same as the number of input coefficients. Denote by $\widetilde{\boldsymbol{\varphi}}_n(t) = [\mathbf{y}^{\mathrm{T}}(t-1), \mathbf{u}^{\mathrm{T}}(t-1), \ldots, \mathbf{y}^{\mathrm{T}}(t-n_y), \mathbf{u}^{\mathrm{T}}(t-n_y)]^{\mathrm{T}}$ the rearranged vector of regression variables, and let $\widetilde{\boldsymbol{\Psi}}_n(t) = \mathbf{I} \otimes \widetilde{\boldsymbol{\varphi}}_n(t)$. Similarly, let $\widetilde{\boldsymbol{\theta}}_n = [(\widetilde{\boldsymbol{\theta}}_n^1)^{\mathrm{T}}, \ldots, (\widetilde{\boldsymbol{\theta}}_n^{m_y})^{\mathrm{T}}]^{\mathrm{T}}$ where $\widetilde{\boldsymbol{\theta}}_n^j = [\boldsymbol{\alpha}_{j1}, \boldsymbol{\beta}_{j1}, \ldots, \boldsymbol{\alpha}_{jn_y}, \boldsymbol{\beta}_{jn_y}]^{\mathrm{T}}$. The model (9) can be rewritten in the following equivalent form

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

Consider the case where $\mathcal{N} = \{\{n_y, n_y\}, n_y = 1, \ldots, N\}$. Then, for a growing order n_y, the corresponding regression vectors form a nested family, namely $\boldsymbol{\varphi}_{\{n_y,n_y\}}(t) \prec \boldsymbol{\varphi}_{\{n_y+1,n_y+1\}}(t)$, $n_y = 1, \ldots, N-1$, where $\mathbf{x} \prec \mathbf{y}$ means that $\mathbf{x}$ is a subvector of $\mathbf{y}$. Using this property, one can easily derive order-recursive algorithms for evaluation of $\widehat{\widetilde{\boldsymbol{\theta}}}_{n|k}(t)$ [which is a permuted version of $\widehat{\boldsymbol{\theta}}_{n|k}(t)$], such as the algorithm proposed in [10].

VI. Computer simulations

Performance of the proposed joint order and bandwidth selection methods was checked by means of computer simulation. Dynamics of the simulated two-input two-output ($m_y = m_u = 2$) ARX system was based on two stable time-invariant "anchor" models: the second-order model M_2 ($n_y = n_u = 2$):

$$\boldsymbol{\Delta}_1^0 = \begin{bmatrix} -0.9135 & -0.0816 \\ 0.0561 & -0.9363 \end{bmatrix}, \quad \boldsymbol{\Delta}_2^0 = \begin{bmatrix} 0.8815 & 0.0437 \\ -0.0768 & 0.9564 \end{bmatrix}$$

$$\mathbf{B}_1^0 = \begin{bmatrix} 3.0082 & 0.0289 \\ 0.0884 & 2.8099 \end{bmatrix}, \quad \mathbf{B}_2^0 = \begin{bmatrix} -3.7434 & -0.0681 \\ -0.1713 & -3.2701 \end{bmatrix}$$

and the fourth-order model M_4 ($n_y = n_u = 4$):

$$\boldsymbol{\Delta}_1^0 = \begin{bmatrix} -0.9135 & -0.0816 \\ 0.0561 & -0.9363 \end{bmatrix}, \quad \boldsymbol{\Delta}_2^0 = \begin{bmatrix} 0.8815 & 0.0437 \\ -0.0768 & 0.9564 \end{bmatrix}$$

$$\boldsymbol{\Delta}_3^0 = \begin{bmatrix} -0.6134 & -0.3783 \\ 0.1237 & -0.7222 \end{bmatrix}, \quad \boldsymbol{\Delta}_4^0 = \begin{bmatrix} 0.6854 & 0.1603 \\ -0.0671 & 0.7490 \end{bmatrix}$$

$$\mathbf{B}_1^0 = \begin{bmatrix} 3.0082 & 0.0289 \\ 0.0884 & 2.8099 \end{bmatrix}, \quad \mathbf{B}_2^0 = \begin{bmatrix} -3.7434 & -0.0681 \\ -0.1713 & -3.2701 \end{bmatrix}$$

$$\mathbf{B}_3^0 = \begin{bmatrix} 2.3177 & 0.0517 \\ 0.1069 & 1.9134 \end{bmatrix}, \quad \mathbf{B}_4^0 = \begin{bmatrix} -0.6278 & -0.0136 \\ -0.0210 & -0.4893 \end{bmatrix}$$

where $\boldsymbol{\Delta}_1^0, \ldots, \boldsymbol{\Delta}_4^0$ denote the so-called matrices of normalized reflection coefficients which uniquely define the autoregressive part of the ARX model and can be "translated" to matrices of autoregressive coefficients appearing in (1). Note that the matrices $\boldsymbol{\Delta}_1^0$, $\boldsymbol{\Delta}_2^0$, $\mathbf{B}_1^0$ and $\mathbf{B}_2^0$ are the same in both anchor models.

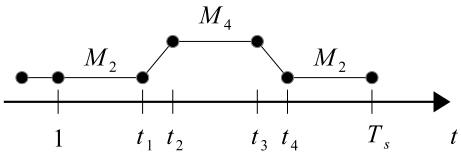

Fig. 1. Morphing scenario used in simulation tests

The time-varying ARX model was obtained by morphing anchor model M_2 into M_4 and *vice versa*. Transition from the model M_2, valid at the instant t_1, to the model M_4, valid at the instant t_2, was realized using the folowing transformations

$$\mathbf{\Delta}_i(t) = \mu(t)\mathbf{\Delta}_i^0, \quad \mathbf{B}_i(t) = \mu(t)\mathbf{B}_i^0$$
$$i = 3, 4 \quad t \in [t_1, t_2]$$

where

$$\mu(t) = \frac{t - t_1}{t_2 - t_1}.$$

The remaining parameters were kept constant: $\mathbf{\Delta}_i(t) = \mathbf{\Delta}_i^0$, $\mathbf{B}_i(t) = \mathbf{B}_i^0$, $t \in [t_1, t_2]$, $i = 1, 2$. Such a morphing technique guarantees stability of of the resulting time-variant model at all times as long as both anchor models are stable (stability is not guaranteed if morphing is applied directly to the matrices of autoregressive coefficients). Transition from the model M_4, valid at the instant t_3, to the model M_2, valid at the instant t_4, was realized in an analogous way, namely

$$\mathbf{\Delta}_i(t) = \eta(t)\mathbf{\Delta}_i^0, \quad \mathbf{B}_i(t) = \eta(t)\mathbf{B}_i^0$$
$$i = 3, 4 \quad t \in [t_3, t_4]$$

where

$$\eta(t) = \frac{t_4 - t}{t_4 - t_3}.$$

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_2–M_2, M_4–M_4, M_2–M_2), each of length l_1, interleaved with 2 periods of nonstationary behavior (M_2–M_4, M_4–M_2), each of length l_2 ($T_s = 3l_1 + 2l_2$). Data generation was started 1000 instants prior to $t = 1$ so that, no matter what bandwidth and model order, the estimation process and evaluation of its results could be started at the instant $t = 1$.

To check performance of the compared algorithms under different rates of nonstationarity, 3 cases were considered corresponding to: fast parameter changes ($T_s = 14000$, $l_1 = 4000$, $l_2 = 1000$), nominal parameter changes ($T_s = 28000$, $l_1 = 8000$, $l_2 = 2000$) and slow parameter changes ($T_s = 56000$, $l_1 = 16000$, $l_2 = 4000$).

The pseudo-random binary type sequence with magnitude $|u_1(t)| = |u_2(t)| = u_0, \forall t$, $u_0 = 0.1$, and covariance matrix $\mathrm{cov}[\mathbf{u}(t)] = u_0^2\mathbf{I}$ (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: $\mathbf{e}(t) \sim \mathcal{N}(\mathbf{0}, \sigma_e^2\mathbf{I})$, $\sigma_e = 0.01$.

TABLE I. COMPARISON OF ESTIMATION RESULTS 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, RESPECTIVETY.

Fast parameter changes

n/N_y	λ_1	λ_2	λ_3	PLS	MFPE
1	47.43	47.06	46.85	47.15	47.16
2	23.53	23.30	23.05	23.39	23.27
3	12.63	12.36	12.34	11.97	12.01
4	1.58	1.54	2.18	0.47	0.45
5	4.12	3.53	4.63	1.67	1.58
6	5.32	4.21	4.91	1.88	1.66
7	6.49	4.81	5.27	1.94	1.65
8	7.78	5.42	5.61	2.02	1.67
9	9.51	6.29	6.06	2.08	1.70
10	11.03	7.02	6.42	2.13	1.70

Nominal parameter changes

n/N_y	λ_1	λ_2	λ_3	PLS	MFPE
1	47.35	47.00	46.83	47.07	47.10
2	23.82	23.63	23.48	23.67	23.56
3	12.63	12.29	12.27	12.03	11.99
4	1.36	0.91	1.20	0.25	0.13
5	3.32	2.54	2.69	0.81	0.58
6	4.51	3.09	2.97	0.95	0.59
7	5.91	3.79	3.42	1.04	0.59
8	7.27	4.45	3.75	1.12	0.60
9	8.79	5.20	4.17	1.20	0.60
10	10.24	5.86	4.52	1.26	0.60

Slow parameter changes

n/N_y	λ_1	λ_2	λ_3	PLS	MFPE
1	47.36	47.01	46.83	47.10	47.10
2	23.61	23.44	23.33	23.47	23.38
3	12.59	12.20	12.08	11.95	11.92
4	1.33	0.71	0.61	0.22	0.08
5	3.02	1.91	1.88	0.47	0.16
6	4.40	2.60	2.26	0.63	0.17
7	5.91	3.34	2.63	0.75	0.17
8	7.47	4.08	3.03	0.85	0.17
9	8.97	4.77	3.39	0.92	0.17
10	10.43	5.48	3.76	0.98	0.17

As a performance measure, quantifying the tracking capabilities of different estimation algorithms, the squared parameter estimation error $d(t) = ||\widehat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}(t)||^2$ was used. Evaluation was based on comparison of mean scores obtained after combined time and ensemble averaging of $d(t)$ (over $t \in [1, T_s]$ and 20 independent realizations of $\{\mathbf{e}(t)\}$).

Table 1 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 ($n_y = n_u = 1, \ldots, 10$) and by 2 adaptive order-and-bandwidth selection schemes (MFPE, PLS).

When the maximum model order $N_y = N_u$ is not underfitted, i.e., when $N_y \geq 4$, both adaptive schemes outperform the non-adaptive fixed-order-fixed-bandwidth algorithms. The reason for this is that both adaptive algorithms detect and adequately react to system nonstationarity by appropriately adjusting model order and estimation bandwidth. According to Fig. 2, which shows the locally time averaged histograms of the results of bandwidth selection (each time bin covers 500 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

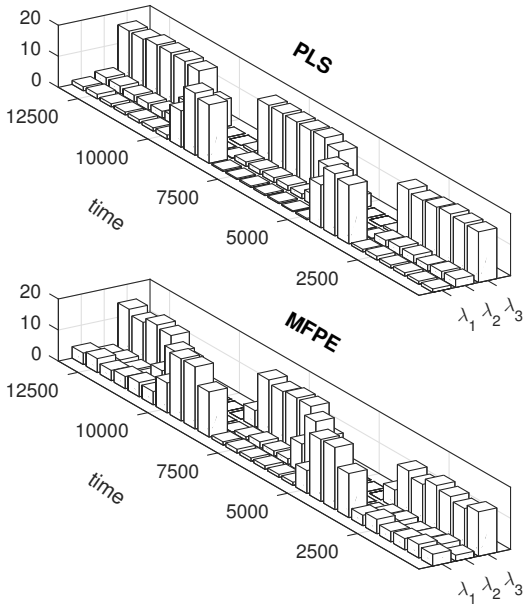

Fig. 2. Histograms of the results of estimation bandwidth selection, obtained for 20 process realizations ($T_s = 14000$).

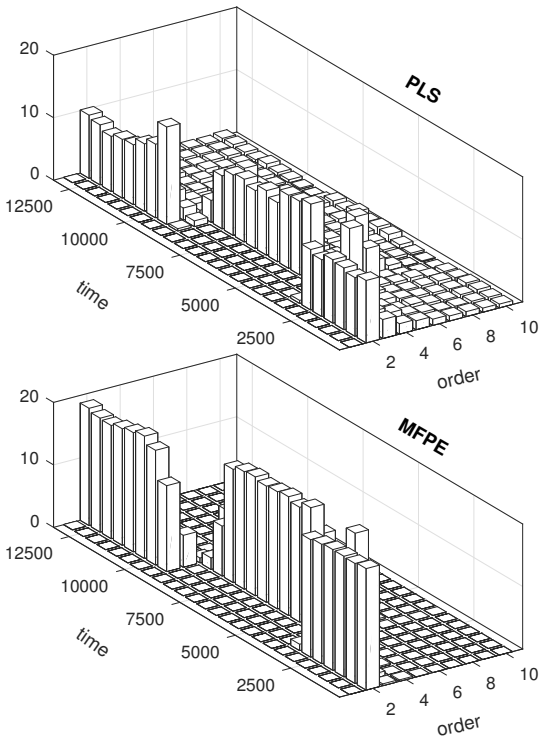

Fig. 3. Histograms of model order estimates, obtained for 20 process realizations ($T_s = 14000$).

ones when system dynamics becomes time-invariant again. The results shown in Fig. 2 were obtained for the high speed of parameter variation. In this case $T_s = 14000$, $t_1 = 4000$, $t_2 = 5000$, $t_3 = 9000$ ans $t_4 = 10000$. Fig. 3 summarizes the order selection capabilities of both adaptive schemes. Both order selection criteria appropriately react to the temporary change of system order from 2 to 4 which takes place in the interval $[t_2, t_3]$. Although the bandwidth selection capabilities of the MFPE-based approach seem to be slightly worse than those of the PLS-based approach, its order selection capabilities are much better, which results in a considerably better overall performance for all rates of system nonstationarity.

VII. Conclusion

The problem of adaptive selection of model order and estimation bandwidth for the purpose of identification of a nonstationary ARX system was considered. It was shown that when suitably modified, the Akaike's final prediction error (FPE) criterion, originally developed for selection of the model order only, can be successfully used for joint order and bandwidth selection. Finally, it was shown that the FPE-based approach favorably compares with the approach based on the Rissanen's predictive least squares principle.

References

[1] H. Akaike, "Fitting autoregressive models for prediction," *Ann. Inst. Statist. Math.*, vol. 21, pp. 243–247, 1969.

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

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

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

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

[6] K.E. Baddour and N.C. Beaulieu, "Autoregressive modeling for fading channel simulation, " *IEEE Trans. Wireless Conmm.*, vol. 4, pp. 1650–1662, 2005.

[7] R.O. Adeogun, "Channel prediction for mobile MIMO wireless communication systems, "Ph.D. Thesis, Victoria University of Wellington, 2015.

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

[9] 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.

[10] M. Kárný, "Bayesian estimation of model order," *Problems of Control and Information Theory*, vol. 9, pp. 33–46, 1980.

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

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

[13] M. Niedźwiecki and L. Guo, "Nonasymptotic results for finite-memory WLS filters," *IEEE Trans. Automat. Contr.*, vol. 36, pp. 515–522, 1991.

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

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

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

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