Post-print of: Niedźwiecki M., Ciołek M., Kajikawa Y.: On adaptive covariance and spectrum estimation of locally stationary multivariate processes. AUTOMATICA. Vol. 82, (2017), p. 1-12. DOI: 10.1016/j.automatica.2017.04.033

On adaptive covariance and spectrum estimation of locally stationary multivariate processes☆

Maciej Niedźwiecki [a], Marcin Ciołek [a] Yoshinobu Kajikawa [b]

[a] Faculty of Electronics, Telecommunications and Computer Science, Department of Automatic Control, Gdańsk University of Technology, Narutowicza 11/12, 80-233 Gdańsk, Poland

[b] Department of Electrical and Electronic Engineering, Faculty of Engineering Science, Kansai University, Suita-shi, Osaka 564-8680, Japan

abstract

When estimating the correlation/spectral structure of a locally stationary process, one has to make two important decisions. First, one should choose the so-called estimation bandwidth, inversely proportional to the effective width of the local analysis window, in the way that complies with the degree of signal nonstationarity. Too small bandwidth may result in an excessive estimation bias, while too large bandwidth may cause excessive estimation variance. Second, but equally important, one should choose the appropriate order of the spectral representation of the signal so as to correctly model its resonant structure – when the order is too small, the estimated spectrum may not reveal some important signal components (resonances), and when it is too high, it may indicate the presence of some nonexistent components. When the analyzed signal is not stationary, with a possibly time-varying degree of nonstationarity, both the bandwidth and order parameters should be adjusted in an adaptive fashion. The paper presents and compares three approaches allowing for unified treatment of the problem of adaptive bandwidth and order selection for the purpose of identification of nonstationary vector autoregressive processes: the cross-validation approach, the full cross-validation approach, and the approach that incorporates the multivariate version of the generalized Akaike's final prediction error criterion. It is shown that the latter solution yields the best results and, at the same time, is very attractive from the computational viewpoint.

Keywords: Identification of nonstationary processes, Determination of estimation bandwidth, Model order selection

1. Introduction

Estimation of the correlation structure of multivariate time series is one of the fundamental techniques allowing one to "understand" experimental data, by revealing their internal relationships, in many research areas such as telecommunications, econometrics, biology, medicine, geophysics, etc. Since in a majority of cases the investigated signals are nonstationary, evaluation of the corresponding autocovariance functions is usually carried out using the local estimation approach, i.e., based on analysis of a short data segment extracted from the entire dataset by a sliding window of a certain width (Dahlhaus, 2012). Under the local stationarity assumptions the revealed signal correlation structure can be further investigated in the frequency domain using the concept of a time-varying signal spectrum (Dahlhaus, 2012).

One of the important decisions that must be taken when performing correlation and/or spectral analysis of a nonstationary signal is the choice of the size of the local analysis interval, which is inversely proportional to the so-called estimation bandwidth, i.e., the frequency range in which parameter changes can be tracked "successfully". Bandwidth optimization allows one to reach a compromise between the bias and variance of the corresponding estimates—large bandwidth results in covariance estimates with large variance but small bias, and small bandwidth causes the opposite effect. When the rate of signal nonstationarity changes over time, estimation bandwidth should be chosen in an adaptive way.

Another important parameter, which must be determined when spectral analysis is carried out, is the number of quantities

☆ This work was 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. The material in this paper was partially presented at the 41st IEEE International Conference on Acoustics Speech and Signal Processing, March, 20–25, 2016, Shanghai, China. This paper was recommended for publication in revised form by Associate Editor Juan I. Yuz under the direction of Editor Torsten Söderström.

http://dx.doi.org/10.1016/j.automatica.2017.04.033

that should be incorporated in the signal description to obtain the most adequate spectrum estimates, quantities such as the number of signal covariance matrices corresponding to different lags (in the nonparametric, i.e., data-driven approach), or the number of signal model parameters (in the parametric, i.e., model-based approach). This will be further referred to as the problem of selection of the order of spectral representation. When the selected order is too small, the estimated spectrum may not reveal some important signal components (resonances), while selecting too high order may result in spectral estimates that indicate the presence of nonexistent (spurious) signal components. From the qualitative viewpoint both alternatives are unsatisfactory. Similar to bandwidth selection, for nonstationary signals the order should be adjusted in an adaptive fashion.

For stationary signals order estimation is a well-explored statistical problem, which can be solved in many different ways. The most popular solutions are those based on the Akaike information criterion (AIC) (Akaike, 1974), Schwarz criterion, frequently referred to as the Bayesian information criterion (BIC) (Schwarz, 1978), and Rissanen's minimum description length (MDL) criterion (Rissanen, 1978). Generalized versions of the AIC and BIC criteria, applicable to local estimation schemes, were proposed in Niedźwiecki (1984, 1985), respectively.

Selection of the estimation bandwidth for the purpose of covariance/spectral analysis of nonstationary signals is a far less investigated topic. The solution that has gained a considerable attention in recent years, proposed in Goldenshluger and Nemirovski (1997) and further developed in Katkovnik (1999) and Stanković (2004), is based on the analysis of the intersection of the confidence intervals (ICI). The ICI approach, developed originally for the purpose of polynomial signal smoothing, was recently applied to covariance estimation in Fu, Chan, Di, Biswal, and Zhang (2014).

When the rate of signal nonstationarity is unknown, and possibly time-varying, several identification algorithms, with different estimation bandwidth settings, can be run in parallel and compared based on their interpolation or predictive capabilities. At each time instant the best-matching VAR model and the corresponding maximum entropy like spectrum estimator can be chosen by means of minimization, over the set of all models, the local performance index.

In this paper we present three approaches allowing for unified treatment of the order and bandwidth selection. The first approach, based on minimization of the local cross-validatory performance measure, was originally used for signal smoothing (Niedźwiecki, 2010). Later on, it was extended to the problem of noncausal identification of nonstationary finite impulse response (FIR) systems using the Kalman filter approach (Niedźwiecki, 2012) and the basis function approach (Niedźwiecki & Gackowski, 2013). Even though derived from the same general modeling principles, none of the solutions presented in the abovementioned papers is directly applicable to the problem of covariance/spectrum estimation. The second approach, based on the concept of full cross-validatory analysis, is a refinement of the first one. Finally, the third approach is based on assessment of predictive capabilities of models obtained for different bandwidth/order choices via the Akaike's final prediction error criterion.

2. Basic facts about the vector autoregressive representation

Consider a discrete stationary m-dimensional random signal $\{\mathbf{y}(t), t = \cdots, -1, 0, 1, \ldots\}$, $\mathbf{y}(t) = [y_1(t), \ldots, y_m(t)]^{\mathrm{T}}$, where t denotes the normalized (dimensionless) discrete time. Suppose that the first $n+1$ autocovariance matrices of $\mathbf{y}(t)$ are known, namely

$$\mathrm{E}[\mathbf{y}(t)\mathbf{y}^{\mathrm{T}}(t-l)] = \mathbf{R}_l, \quad l = 0, \ldots, n. \tag{1}$$

It is well-known from the Burg's work (Burg, 1967, 1975) that the maximum entropy (i.e., the most unpredictable) stationary process subject to the constraints (1) is the Gaussian vector autoregressive (VAR) process of order n satisfying the equation

$$\mathbf{y}(t) + \sum_{i=1}^{n} \mathbf{A}_i \mathbf{y}(t-i) = \boldsymbol{\epsilon}(t), \qquad \mathrm{cov}[\boldsymbol{\epsilon}(t)] = \boldsymbol{\rho} \tag{2}$$

where $\{\boldsymbol{\epsilon}(t)\}$ denotes m-dimensional white noise sequence with covariance matrix $\boldsymbol{\rho}$, and

$$\mathbf{A}_i = \begin{bmatrix} a_{11,i} & \cdots & a_{1m,i} \\ \vdots & & \vdots \\ a_{m1,i} & \cdots & a_{mm,i} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha}_{1i} \\ \vdots \\ \boldsymbol{\alpha}_{mi} \end{bmatrix}, \quad i = 1, \ldots, n$$

are the $m \times m$ matrices of autoregressive coefficients. The relationship between the autocovariance matrices (1) and parameters of the VAR model, known as the Yule–Walker (YW) equations, takes the form

$$[\mathbf{I}, \mathbf{A}_1, \ldots, \mathbf{A}_n]\, \mathscr{R} = [\boldsymbol{\rho}, \mathbf{O}, \ldots, \mathbf{O}] \tag{3}$$

where $\mathbf{I}$ and $\mathbf{O}$ denote the $m \times m$ identity and null matrices, respectively, and $\mathscr{R}$ is the block Toeplitz matrix of the form

$$\mathscr{R} = \begin{bmatrix} \mathbf{R}_0 & \cdots & \mathbf{R}_n \\ \vdots & \ddots & \vdots \\ \mathbf{R}_n^{\mathrm{T}} & \cdots & \mathbf{R}_0 \end{bmatrix}.$$

The maximum entropy (ME) extension of the autocovariance sequence (1) $\widehat{\mathbf{R}}_l = -\sum_{i=1}^{n} \mathbf{A}_i \widehat{\mathbf{R}}_{l-i}$, $l > n$, where $\widehat{\mathbf{R}}_i = \mathbf{R}_i$ for $0 \le i \le n$, which stems from the VAR signal model (2), leads to the following definition of the maximum entropy spectrum estimate

$$\widehat{\mathbf{S}}(\omega) = \sum_{i=-\infty}^{\infty} \widehat{\mathbf{R}}_i e^{-j\omega i} = \mathscr{A}^{-1}(e^{j\omega})\, \boldsymbol{\rho}\, \mathscr{A}^{-\mathrm{T}}(e^{-j\omega}) \tag{4}$$

where $j = \sqrt{-1}$, $\omega \in [0, \pi]$ denotes the normalized angular frequency, and $\mathscr{A}(z^{-1}) = \mathbf{I} + \sum_{i=1}^{n} \mathbf{A}_i z^{-i}$. Since the sequence of autocovariance matrices $\{\widehat{\mathbf{R}}_i, i = \cdots, -1, 0, 1, \ldots\}$, $\widehat{\mathbf{R}}_{-i} = \widehat{\mathbf{R}}_i^{\mathrm{T}}$, is by construction nonnegative definite, the corresponding spectral density matrix is also nonnegative definite $\widehat{\mathbf{S}}(\omega) \ge \mathbf{O}$, $\forall \omega \in [0, \pi]$. The off-diagonal elements of $\widehat{\mathbf{S}}(\omega)$, which can be interpreted as cross-spectral densities of different pairs of components of $\mathbf{y}(t)$, are in general complex-valued.

Two of our bandwidth/order selection procedures will be based on the results of signal interpolation. To derive the interpolation formula for the signal governed by the VAR model (2), suppose that all signal samples $\{\mathbf{y}(i),\ i = -\infty < i < \infty\}$ are known except for $\mathbf{y}(t)$. The least squares estimate of $\mathbf{y}(t)$ can be obtained from

$$\begin{aligned} \widehat{\mathbf{y}}(t) &= \arg\min_{\mathbf{y}(t)} \sum_{s=-\infty}^{\infty} \Big\| \mathbf{y}(s) + \sum_{i=1}^{n} \mathbf{A}_i \mathbf{y}(s-i) \Big\|^2 \\ &= \arg\min_{\mathbf{y}(t)} \sum_{s=t}^{t+n} \Big\| \mathbf{y}(s) + \sum_{i=1}^{n} \mathbf{A}_i \mathbf{y}(s-i) \Big\|^2 \\ &= \arg\min_{\mathbf{y}(t)} \mathbf{z}^{\mathrm{T}}(t)\, \mathscr{C}^{\mathrm{T}} \mathscr{C}\, \mathbf{z}(t) \end{aligned} \tag{5}$$

where $\mathbf{z}(t) = [\mathbf{y}^{\mathrm{T}}(t-n), \ldots, \mathbf{y}^{\mathrm{T}}(t+n)]^{\mathrm{T}}$,

$$\mathscr{C} = \begin{bmatrix} \mathbf{A}_n & \mathbf{A}_{n-1} & \cdots & \mathbf{A}_0 & \mathbf{O} & \cdots & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{A}_n & \cdots & \mathbf{A}_1 & \mathbf{A}_0 & \cdots & \mathbf{O} & \mathbf{O} \\ \vdots & & \ddots & & \ddots & \ddots & & \\ \mathbf{O} & \mathbf{O} & & \mathbf{A}_n & \cdots & & \mathbf{A}_1 & \mathbf{A}_0 \end{bmatrix}$$

and $\mathbf{A}_0 = \mathbf{I}$. Note that $\mathscr{C}$ is a $(n+1) \times (2n+1)$ block matrix made up of $m \times m$ dimensional blocks.

Denote by $J = \{1, \ldots, n, n+2, \ldots, 2n+1\}$ the set of indexes indicating positions of known samples within the vector $\mathbf{z}(t)$, and by $I = \{n+1\}$—the analogous set indicating position of the unknown sample. Denote by $\mathbf{C}_m = \mathcal{C}_{|J>}$ the $(n+1)\times 1$ block matrix obtained after removing from $\mathcal{C}$ its $2n$ block columns indicated by the set J, and by $\mathbf{C}_o = \mathcal{C}_{|I>}$—the $(n+1)\times 2n$ block matrix obtained in the analogous way using the set I. Finally, denote by $\mathbf{z}_o(t) = \mathbf{z}(t)_{|I>} = [\mathbf{y}^{\mathrm{T}}(t-n), \ldots, \mathbf{y}^{\mathrm{T}}(t-1), \mathbf{y}^{\mathrm{T}}(t+1), \ldots, \mathbf{y}^{\mathrm{T}}(t+n)]^{\mathrm{T}}$ the $2n\times 1$ block vector of known samples obtained after removing $\mathbf{y}(t)$ from $\mathbf{z}(t)$. Using this notation, the estimated value of $\mathbf{y}(t)$, given by (5), can be written down in the form Niedźwiecki (1993)

$$\widehat{\mathbf{y}}(t) = -[\mathbf{C}_m^{\mathrm{T}}\mathbf{C}_m]^{-1}\mathbf{C}_m^{\mathrm{T}}\mathbf{C}_o\mathbf{z}_o(t) \tag{6}$$

or, more explicitly, as

$$\widehat{\mathbf{y}}(t) = -\left[\sum_{i=0}^{n}\mathbf{A}_i^{\mathrm{T}}\mathbf{A}_i\right]^{-1}\sum_{i=0}^{n}\mathbf{A}_i^{\mathrm{T}}\mathbf{p}_i(t) \tag{7}$$

where

$$\mathbf{p}_i(t) = \sum_{\substack{l=0\\ l\neq i}}^{n}\mathbf{A}_l\mathbf{y}(t+i-l), \quad i = 0, \ldots, n. \tag{8}$$

Note that $\widehat{\mathbf{y}}(t)$ depends only on n samples preceding $\mathbf{y}(t)$ and n samples succeeding $\mathbf{y}(t)$, which is consistent with the fact that the signal $\{\mathbf{y}(t)\}$ governed by (2) is a Markov process of order n.

3. Local estimation technique

When the investigated process is nonstationary, but its characteristics vary slowly with time, the covariance/spectral analysis can be carried out under the "local stationarity" framework. An elegant theory of locally stationary processes, based on the concept of infill asymptotics (in which a fixed-length time interval is sampled over a finer and finer grid of points as the sample size increases) was worked out by Dahlhaus (1997, 2012). Without getting into mathematical details, we note that the probabilistic structure of such processes at a selected time instant t can be examined using local estimation techniques, e.g. by means of processing a fixed-length data segment $\{\mathbf{y}(t-k), \ldots, \mathbf{y}(t), \ldots, \mathbf{y}(t+k)\}$ "centered" at t. The integer number k, which controls the size of the local analysis interval $[t-k, t+k]$, will be further referred to as a bandwidth parameter.

3.1. Yule–Walker estimator

The local estimates of the autocovariance matrices (1) can be obtained using the formula

$$\widehat{\mathbf{R}}_{l,k}(t) = \frac{1}{L_k}\mathbf{P}_{l,k}(t), \quad l = 0, \ldots, n \tag{9}$$

where

$$\begin{aligned}\mathbf{P}_{l,k}(t) &= \sum_{i=-k+l}^{k} w_k(i)w_k(i-l)\mathbf{y}(t+i)\mathbf{y}^{\mathrm{T}}(t+i-l)\\ &= \sum_{i=-k+l}^{k}\mathbf{y}_k(t+i|t)\mathbf{y}_k^{\mathrm{T}}(t+i-l|t)\end{aligned} \tag{10}$$

and $\mathbf{y}_k(t-k|t), \ldots, \mathbf{y}_k(t+k|t)$ is the tapered data sequence $\mathbf{y}_k(t+i|t) = w_k(i)\mathbf{y}(t+i)$, $i = -k, \ldots, k$. The weights $w_k(i)$ are defined as $w_k(i) = h(i/k)$, where $h : [-1, 1] \to \mathbb{R}_+$ denotes a symmetric data taper function $h(x) = h(-x) \geq 0$ taking its largest value at 0 [for convenience we will assume that $h(0) = 1$] and smoothly decaying to 0 at the edges.

Finally, the normalizing constant in (9) takes the form $L_k = \sum_{i=-k}^{k} w_k^2(i) \cong k\int_{-1}^{1} h^2(x)dx$. Based on the set of covariance estimates (9), the local VAR signal model

$$\mathbf{y}(t) + \sum_{i=1}^{n}\widehat{\mathbf{A}}_{i,k}(t)\mathbf{y}(t-i) = \boldsymbol{\epsilon}(t), \qquad \mathrm{cov}[\boldsymbol{\epsilon}(t)] = \widehat{\boldsymbol{\rho}}_k(t) \tag{11}$$

can be obtained by solving for $\widehat{\mathbf{A}}_{1,k}(t), \ldots, \widehat{\mathbf{A}}_{n,k}(t)$ and $\widehat{\boldsymbol{\rho}}_k(t)$ the corresponding Yule–Walker equations

$$[\mathbf{I}, \widehat{\mathbf{A}}_{1,k}(t), \ldots, \widehat{\mathbf{A}}_{n,k}(t)]\,\widehat{\boldsymbol{\mathcal{R}}}_k(t) = [\widehat{\boldsymbol{\rho}}_k(t), \mathbf{O}, \ldots, \mathbf{O}] \tag{12}$$

where $\widehat{\boldsymbol{\mathcal{R}}}_k(t)$ is a block Toeplitz matrix obtained by replacing the true autocovariance matrices $\mathbf{R}_i$, appearing in $\boldsymbol{\mathcal{R}}$, with their local estimates $\widehat{\mathbf{R}}_{i,k}(t)$. An efficient procedure for solving (12) is known as the Whittle–Wiggins–Robinson (WWR) algorithm. WWR algorithm is a multivariate extension of the Levinson–Durbin algorithm—for the discussion of its basic properties see Complement C8.6 in Söderström and Stoica (1988).

3.2. Relation to the weighted least squares estimator

Denote by $\boldsymbol{\theta}_l = [\boldsymbol{\alpha}_{l1}, \ldots, \boldsymbol{\alpha}_{ln}]^{\mathrm{T}}$ the vector of parameters characterizing the lth "channel" of the VAR process, and by $\boldsymbol{\varphi}(t) = [\mathbf{y}^{\mathrm{T}}(t-1), \ldots, \mathbf{y}^{\mathrm{T}}(t-n)]^{\mathrm{T}}$ the corresponding regression vector. Using this notation, equation of the lth channel can be written down in the form

$$y_l(t) + \boldsymbol{\varphi}^{\mathrm{T}}(t)\boldsymbol{\theta}_l = \epsilon_l(t) \tag{13}$$

and (2) can be rewritten more compactly as

$$\mathbf{y}(t) + \boldsymbol{\Psi}^{\mathrm{T}}(t)\boldsymbol{\theta} = \boldsymbol{\epsilon}(t) \tag{14}$$

where $\boldsymbol{\Psi}(t) = \mathbf{I}\otimes\boldsymbol{\varphi}(t) = \mathrm{diag}\{\boldsymbol{\varphi}(t), \ldots, \boldsymbol{\varphi}(t)\}$ ($\otimes$ denotes Kronecker product of two matrices/vectors) and $\boldsymbol{\theta} = [\boldsymbol{\theta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_m^{\mathrm{T}}]^{\mathrm{T}} = \mathrm{vec}\{[\mathbf{A}_1|\ldots|\mathbf{A}_n]^{\mathrm{T}}\}$ is the column vector combining, in a row-wise manner, all autoregressive coefficients gathered in the $m\times mn$ matrix $[\mathbf{A}_1|\ldots|\mathbf{A}_n]$. Note that the regression vector in (13) is the same for all channels.

It is straightforward to check [see e.g. Section 3.4.2 in Stoica & Moses, 1997] that if n signal samples preceding and succeeding the frame $[t-k, t+k]$ are zeroed, namely $\mathbf{y}(t\pm k\pm 1) = \cdots = \mathbf{y}(t\pm k\pm n) = \mathbf{0}$ and a similar extension is applied to the data window $[w_k(k+1) = \cdots = w_k(k+n) = 0]$, the YW estimator $\widehat{\boldsymbol{\theta}}_k(t) = [\widehat{\boldsymbol{\theta}}_{1,k}^{\mathrm{T}}(t), \ldots, \widehat{\boldsymbol{\theta}}_{m,k}^{\mathrm{T}}(t)]^{\mathrm{T}} = \mathrm{vec}\{[\widehat{\mathbf{A}}_{1,k}(t)|\ldots|\widehat{\mathbf{A}}_{n,k}(t)]^{\mathrm{T}}\}$ is identical with the least squares estimator with data weighting (LSW) defined as

$$\widehat{\boldsymbol{\theta}}_k(t) = \arg\min_{\boldsymbol{\theta}}\sum_{i=-k}^{k+n}\|\mathbf{y}_k(t+i|t) + \boldsymbol{\Psi}_k^{\mathrm{T}}(t+i|t)\boldsymbol{\theta}\|^2 \tag{15}$$

where $\boldsymbol{\Psi}_k(t+i|t) = \mathbf{I}\otimes\boldsymbol{\varphi}_k(t+i|t)$ and $\boldsymbol{\varphi}_k(t+i|t) = [\mathbf{y}_k^{\mathrm{T}}(t+i-1|t), \ldots, \mathbf{y}_k^{\mathrm{T}}(t+i-n|t)]^{\mathrm{T}}$. Moreover, when $n \ll k$, one can use the approximation $\boldsymbol{\varphi}_k(t+i|t) \cong w_k(i)\boldsymbol{\varphi}(t+i)$, leading to

$$\widehat{\boldsymbol{\theta}}_k(t) \cong \arg\min_{\boldsymbol{\theta}}\sum_{i=-k}^{k} v_k(i)\|\mathbf{y}(t+i) + \boldsymbol{\Psi}^{\mathrm{T}}(t+i)\boldsymbol{\theta}\|^2 \tag{16}$$

where the squared data taper serves as the weighting sequence

$$v_k(i) = w_k^2(i), \quad i\in[-k, k]. \tag{17}$$

The estimation scheme (16), known as weighted least squares (WLS), is well understood from the statistical viewpoint which will be helpful when deriving some technical results below. It should be stressed that, unlike the YW scheme, the WLS scheme *does not* guarantee stability of the VAR model, which is a prerequisite for well-posed parametric spectrum estimation.

3.3. Covariance and spectrum estimation

Since in this paper we are primarily interested in analyzing the evolution of the instantaneous (local) autocovariance function $\{\mathbf{R}_i(t),\ i = \cdots, -1, 0, 1, \ldots\}$ of $\mathbf{y}(t)$, and its instantaneous spectral density function $\mathbf{S}(\omega, t)$, the time-varying VAR model (11) will be regarded – very much like in the maximum entropy approach – as a "meta-model", serving mainly both purposes mentioned above. According to Dahlhaus (2012), both $\mathbf{R}_i(t)$ and $\mathbf{S}(\omega, t)$ are well-defined quantities which can be interpreted as characteristics of a stationary process $\{\mathbf{y}_0(s)\}$ "tangent" to $\{\mathbf{y}(s)\}$ at the instant t.

The important property of the approximation (11) is that as long as the matrix $\widehat{\mathcal{R}}_k(t)$ is positive definite [which is always the case when the estimates (12) are incorporated—provided that the sequence $\{\mathbf{y}(t)\}$ is persistently exciting in some deterministic Stoica & Moses, 1997 or stochastic Niedźwiecki & Guo, 1991 sense] the obtained model is always stable in the sense that all zeros z_i of the characteristic polynomial $\det[\widehat{\mathcal{A}}_k(z^{-1}, t)]$, where

$$\widehat{\mathcal{A}}_k(z^{-1}, t) = \mathbf{I} + \sum_{i=1}^{n} \widehat{\mathbf{A}}_{i,k}(t) z^{-i} \tag{18}$$

lie inside the unit circle in the complex plane: $|z_i| < 1,\ i = 1, \ldots, mn$.

As already mentioned, the time-varying VAR meta-model opens interesting analytical opportunities. First, it allows one to evaluate the ME-like extension of the autocovariance function for the lags $l > n$, i.e. beyond the range of estimation

$$\widehat{\mathbf{R}}_{l,k}(t) = -\sum_{i=1}^{n} \widehat{\mathbf{A}}_{i,k}(t) \widehat{\mathbf{R}}_{l-i,k}(t), \quad l > n. \tag{19}$$

Second, the VAR model can serve as a basis for evaluation of the instantaneous signal spectrum

$$\begin{aligned} \widehat{\mathbf{S}}_k(\omega, t) &= \sum_{i=-\infty}^{\infty} \widehat{\mathbf{R}}_{i,k}(t) e^{-j\omega i} \\ &= \widehat{\mathcal{A}}_k^{-1}(e^{j\omega}, t)\, \widehat{\boldsymbol{\rho}}_k(t)\, \widehat{\mathcal{A}}_k^{-\mathrm{T}}(e^{-j\omega}, t). \end{aligned} \tag{20}$$

We note that when the local stationarity assumptions, given in Dahlhaus (2012), are met, the time-varying spectral density function

$$\mathbf{S}(\omega, t) = \mathcal{A}^{-1}(e^{j\omega}, t)\ \boldsymbol{\rho}(t)\ \mathcal{A}^{-T}(e^{-j\omega}, t)$$

governed by a stable VAR model

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

is uniquely defined in the rescaled time domain. In the non-rescaled case, considered e.g. by Priestley in his work on evolutionary spectra (Priestley, 1965), such uniqueness is not guaranteed.

3.4. Window carpentry

The problem of selection of the shape of the window $\{w_k(i)\}$ can be discussed from several different perspectives.

First, since the analyzed process is nonstationary, it is reasonable to assign higher weights to measurements taken at instants close to t (which is our time-point of interest), and lower weights to measurements from instants far from t. Second, since unweighted YW estimates are identical to LS estimates obtained for the original data sequence extended with n zero samples at the segment beginning and at its end, data tapering allows one to smooth out signal discontinuities introduced by such a modification, and hence to reduce the associated estimation bias. Both observations suggest usage of symmetric bell-shaped windows.

Some additional insights into the problem of window selection can be gained from the frequency-domain analysis. It is known that in the univariate case the weighted YW estimators minimize the so-called Whittle likelihood—the Kullback–Leibler based "distance" between the parametric AR spectrum and nonparametric weighted periodogram one (Dahlhaus, 1997). In nonparametric spectrum estimation, weighting is applied to reach the desired bias–variance tradeoff. This can be achieved by choosing windows with the appropriate energy spectrum, namely the ones with reduced sidelobe structure (to minimize spectral leakage) and, at the same time, with a relatively narrow main lobe (to minimize spectrum "smearing").

The last, practically important aspect of window selection is computational complexity. The preferable form of the window is the one that allows for recursive computation of the quantities $\mathbf{P}_{l,k}(t)$ given by (10).

We will show that the cosinusoidal window

$$w_k(i) = \cos\frac{\pi i}{2(k+1)}, \quad i \in [-k, k] \tag{21}$$

fulfills all requirements mentioned above. First, the weights decay to zero at both ends of the analysis interval, which guarantees smooth transition from data to no-data. Second, when the LS reinterpretation of the YW scheme is applied, the squared data window, which from the qualitative viewpoint corresponds to the so-called lag window in the Blackman–Tukey correlogram analysis, has the form

$$v_k(i) = w_k^2(i) = \frac{1}{2}\left[1 + \cos\frac{\pi i}{k+1}\right] \tag{22}$$

which can be recognized as the Hann (raised cosine) window—one of the standard choices in classical nonparametric spectrum estimation, offering good bias–variance tradeoff.

To show that the window (21) allows for recursive computation of (10), note that $w_k(i)$ can be expressed in the form

$$w_k(i) = \frac{1}{2}\left[e^{\frac{j\pi i}{2(k+1)}} + e^{-\frac{j\pi i}{2(k+1)}}\right]$$

leading to

$$\begin{aligned} w_k(i) w_k(i-l) = {} & \frac{1}{4} e^{\frac{j\pi i}{k+1}} e^{-\frac{j\pi l}{2(k+1)}} + \frac{1}{4} e^{-\frac{j\pi i}{k+1}} e^{\frac{j\pi l}{2(k+1)}} \\ & + \frac{1}{2}\cos\frac{\pi l}{2(k+1)}. \end{aligned}$$

Note that

$$\mathbf{P}_{l,k}(t) = \sum_{i=-k+l}^{k} w_k(i) w_k(i-l) \mathbf{H}_l(t+i)$$

where $\mathbf{H}_l(t+i) = \mathbf{y}(t+i)\mathbf{y}^{\mathrm{T}}(t+i-l)$. Let

$$\mathbf{F}_{l,k}(t) = \sum_{i=-k+l}^{k} \mathbf{H}_l(t+i)$$

$$\mathbf{G}_{l,k}(t) = \sum_{i=-k+l}^{k} \mathbf{H}_l(t+i) e^{\frac{j\pi i}{k+1}}.$$

Observe that both quantities defined above are recursively computable

$$\begin{aligned} \mathbf{F}_{l,k}(t+1) &= \mathbf{F}_{l,k}(t) - \mathbf{H}_l(t-k+l) + \mathbf{H}_l(t+k+1) \\ \mathbf{G}_{l,k}(t+1) &= e^{-\frac{j\pi}{k+1}} \mathbf{G}_{l,k}(t) + e^{\frac{j\pi l}{k+1}} \mathbf{H}_l(t-k+l) \\ &\quad + e^{\frac{j\pi k}{k+1}} \mathbf{H}_l(t+k+1). \end{aligned} \tag{23}$$

The quantity $\mathbf{P}_{l,k}(t)$ can be obtained from

$$\mathbf{P}_{l,k}(t) = \frac{1}{2}\cos\frac{\pi l}{2(k+1)}\,\mathbf{F}_{l,k}(t) + \frac{1}{2}\mathrm{Re}\left[\mathbf{G}_{l,k}(t)e^{-\frac{j\pi l}{2(k+1)}}\right]. \tag{24}$$

A single time update using (23)–(24) requires, for a selected value of l, $11m^2$ real multiply–add operations. Note that the computational load does not depend on the size of the analysis interval. When the direct method (10) is used, the analogous count is $(2k+1-l)m^2$ per time update.

As all sliding window subtract–add algorithms, recursive algorithms (23) are prone to diverge due to unbounded accumulation of round-off errors. Therefore, to prevent this from happening, they should be periodically (e.g. every 1 million steps or so) reset by direct (nonrecursive) computation of the quantities $\mathbf{F}_{l,k}(t)$ and $\mathbf{G}_{l,k}(t)$.

4. Selection of the estimation bandwidth

So far we have assumed that the bandwidth parameter k is fixed prior to autocovariance/spectrum estimation. For a nonstationary process with constant-known "degree of nonstationarity" the optimal value of k, i.e., the one that minimizes the mean-squared estimation error, can be found analytically (Dahlhaus & Giraitis, 1998). Unfortunately, in practice such a prior knowledge is not available. Additionally, the degree of signal nonstationarity may itself change with time. On the qualitative level, it is known that the optimal value of the bandwidth parameter increases as the identified signal becomes more and more stationary, and conversely—when the degree of signal nonstationarity is high, short analysis windows may be required to guarantee the best tradeoff between the bias component of the mean-squared error (which grows with k) and its variance component (which decays with k).

Rather than trying to design a single estimation algorithm equipped with an adjustable bandwidth-controlling parameter, we will consider a parallel estimation scheme made up of K simultaneously working algorithms with different bandwidth settings: $k_i,\ i = 1,\ldots,K$. The results yielded by the competing algorithms will be combined in a way that takes into account their locally assessed performance.

4.1. Cross-validation based approach

As a local performance measure one can use the sum of "squared" leave-one-out interpolation errors $\mathbf{e}^{\circ}_k(t) = \mathbf{y}(t) - \widehat{\mathbf{y}}^{\circ}_k(t)$, where $\widehat{\mathbf{y}}^{\circ}_k(t)$ denotes the estimate of $\mathbf{y}(t)$ based *exclusively* on k samples preceding and k samples succeeding $\mathbf{y}(t)$. To derive the suitable interpolation formula, we will first define the "holey" counterpart of the VAR model (11)

$$\mathbf{y}(t) + \sum_{i=1}^{n}\widehat{\mathbf{A}}^{\circ}_{i,k}(t)\mathbf{y}(t-i) = \boldsymbol{\epsilon}^{\circ}(t), \qquad \mathrm{cov}[\boldsymbol{\epsilon}^{\circ}(t)] = \widehat{\boldsymbol{\rho}}^{\circ}_k(t) \tag{25}$$

obtained in an analogous way as (11), except that the central sample $\mathbf{y}(t)$ is excluded from the estimation process. The corresponding parameter estimates can be obtained by solving the modified set of Yule–Walker equations

$$[\mathbf{I}, \widehat{\mathbf{A}}^{\circ}_{1,k}(t), \ldots, \widehat{\mathbf{A}}^{\circ}_{n,k}(t)]\ \widehat{\boldsymbol{\mathcal{R}}}^{\circ}_k(t) = [\widehat{\boldsymbol{\rho}}^{\circ}_k(t), \mathbf{O}, \ldots, \mathbf{O}] \tag{26}$$

where the matrix $\widehat{\boldsymbol{\mathcal{R}}}^{\circ}_k(t)$ is made up of "holey" covariance estimates [note that, according to our earlier assumptions, $w_k(0) = 1$]

$$\widehat{\mathbf{R}}^{\circ}_{l,k}(t) = \frac{1}{L^{\circ}_k}\,\mathbf{P}^{\circ}_{l,k}(t) \tag{27}$$

$$\begin{aligned}\mathbf{P}^{\circ}_{l,k}(t) &= \mathbf{P}_{l,k}(t)\big|_{\mathbf{y}(t)=\mathbf{0}} \\ &= \begin{cases} \mathbf{P}_{l,k}(t) - \mathbf{y}(t)\mathbf{y}^{\mathrm{T}}(t) & l = 0 \\ \mathbf{P}_{l,k}(t) - \mathbf{y}(t)\mathbf{y}^{\mathrm{T}}_k(t-l|t) & \\ \quad -\mathbf{y}_k(t+l|t)\mathbf{y}^{\mathrm{T}}(t) & 1 \le l \le k+1 \\ \mathbf{P}_{l,k}(t) & l > k+1 \end{cases}\end{aligned} \tag{28}$$

$$L^{\circ}_k = \sum_{\substack{i=-k \\ i\neq 0}}^{k} w^2_k(i) = L_k - 1.$$

Based on (25), one arrives at the following interpolation formula borrowed from the theory of stationary VAR processes [cf. (7) and (8)]

$$\widehat{\mathbf{y}}^{\circ}_k(t) = -\left[\sum_{i=0}^{n}[\widehat{\mathbf{A}}^{\circ}_{i,k}(t)]^{\mathrm{T}}\widehat{\mathbf{A}}^{\circ}_{i,k}(t)\right]^{-1}\sum_{i=0}^{n}[\widehat{\mathbf{A}}^{\circ}_{i,k}(t)]^{\mathrm{T}}\mathbf{p}^{\circ}_{i,k}(t) \tag{29}$$

where

$$\mathbf{p}^{\circ}_{i,k}(t) = \sum_{\substack{l=0 \\ l\neq i}}^{n}\widehat{\mathbf{A}}^{\circ}_{l,k}(t)\mathbf{y}(t+i-l), \quad i = 0,\ldots,n. \tag{30}$$

Interpolation errors will be accumulated over a local evaluation window $T(t) = [t-d, t+d]$ of width $D = 2d+1 > m$, forming the matrix

$$\mathbf{Q}^{\circ}_k(t) = \sum_{s\in T(t)}\mathbf{e}^{\circ}_k(s)[\mathbf{e}^{\circ}_k(s)]^{\mathrm{T}}.$$

At each time instant t the bandwidth parameter will be chosen from the set $\mathcal{K} = \{k_i,\ i = 1,\ldots,K\}$ so as to "minimize" the matrix $\mathbf{Q}^{\circ}_k(t)$, namely

$$\widehat{k}(t) = \arg\min_{k\in\mathcal{K}}\det\left[\mathbf{Q}^{\circ}_k(t)\right]. \tag{31}$$

The corresponding spectral density estimate will take the form

$$\widehat{\mathbf{S}}(\omega, t) = \widehat{\mathbf{S}}_{\widehat{k}(t)}(\omega, t). \tag{32}$$

The procedure described above is based on the technique known in statistics as cross-validation. In this approach the quality of the model obtained for a given (training) dataset is judged by checking its ability to "explain", e.g. predict, data samples excluded from the estimation process (validation dataset) (Friedl & Stampfer, 2002). When only one sample is excluded at a time – as in the case considered – the procedure is known as a leave-one-out cross-validation.

To reduce the estimation bias caused by the fact that the "central" sample $\mathbf{y}(t)$ is zeroed in (28), after calculating the leave-one-out signal interpolation $\mathbf{y}^{\circ}_k(t)$, one can recompute the covariance estimates setting $\mathbf{y}(t)$ to $\widehat{\mathbf{y}}^{\circ}_k(t)$ instead of $\mathbf{0}$:

$$\widehat{\mathbf{R}}^{\bullet}_{l,k}(t) = \frac{1}{L_k}\,\mathbf{P}^{\bullet}_{l,k}(t) \tag{33}$$

$$\begin{aligned}\mathbf{P}^{\bullet}_{l,k}(t) &= \mathbf{P}_{l,k}(t)\big|_{\mathbf{y}(t)=\widehat{\mathbf{y}}^{\circ}_k(t)} \\ &= \begin{cases} \mathbf{P}^{\circ}_{l,k}(t) + \widehat{\mathbf{y}}^{\circ}_k(t)[\widehat{\mathbf{y}}^{\circ}_k(t)]^{\mathrm{T}} & l = 0 \\ \mathbf{P}^{\circ}_{l,k}(t) + \widehat{\mathbf{y}}^{\circ}_k(t)\mathbf{y}^{\mathrm{T}}_k(t-l|t) & \\ \quad +\mathbf{y}_k(t+l|t)[\widehat{\mathbf{y}}^{\circ}_k(t)]^{\mathrm{T}} & 1 \le l \le k+1 \\ \mathbf{P}^{\circ}_{l,k}(t) & l > k+1. \end{cases}\end{aligned} \tag{34}$$

The corresponding VAR model can be obtained by solving

$$[\mathbf{I}, \widehat{\mathbf{A}}^{\bullet}_{1,k}(t), \ldots, \widehat{\mathbf{A}}^{\bullet}_{n,k}(t)]\ \widehat{\boldsymbol{\mathcal{R}}}^{\bullet}_k(t) = [\widehat{\boldsymbol{\rho}}^{\bullet}_k(t), \mathbf{O}, \ldots, \mathbf{O}] \tag{35}$$

where the block Toeplitz matrix $\widehat{\boldsymbol{\mathcal{R}}}^{\bullet}_k(t)$, made up of the estimates (33), has the same structure as $\widehat{\boldsymbol{\mathcal{R}}}^{\circ}_k(t)$. Note that, similar to the model resulting from (26), the corrected model is also "holey" in the sense that its parameters do not depend on the central sample

$\mathbf{y}(t)$. The idea of the correction described above goes back to Bunke, Droge, and Polzehl (1999) where it was the cornerstone of the so-called full cross-validatory analysis.

Using the corrected model, one can compute [in the same way as described before—see (29) and (30)] the corrected signal interpolation $\widehat{\mathbf{y}}_k^{\bullet}(t)$ and the associated interpolation error $\mathbf{e}_k^{\bullet}(t) = \mathbf{y}(t) - \widehat{\mathbf{y}}_k^{\bullet}(t)$. Then, using the errors $\mathbf{e}_k^{\bullet}(t)$ in lieu of $\mathbf{e}_k^{\circ}(t)$, one can select k by means of minimizing the determinant of the matrix

$$\mathbf{Q}_k^{\bullet}(t) = \sum_{s\in T(t)} \mathbf{e}_k^{\bullet}(s)[\mathbf{e}_k^{\bullet}(s)]^{\mathrm{T}}.$$

4.2. Final prediction error based approach

Another approach to selection of the estimation bandwidth is based on evaluation of predictive, rather than interpolation, capabilities of the compared models. Denote by $\widetilde{\mathcal{Y}}_k(t) = \{\widetilde{\mathbf{y}}(t-k), \ldots, \widetilde{\mathbf{y}}(t+k)\}$ another realization of the analyzed data sequence, independent of the sequence $\mathcal{Y}_k(t) = \{\mathbf{y}(t-k), \ldots, \mathbf{y}(t+k)\}$ used for identification purposes. As an alternative measure of fit, one can adopt determinant of the following matrix of mean squared prediction errors

$$\delta_k(t) = \mathrm{E}\left\{[\widetilde{\mathbf{y}}(t) + \widetilde{\boldsymbol{\Psi}}^{\mathrm{T}}(t)\widehat{\boldsymbol{\theta}}_k(t)][\widetilde{\mathbf{y}}(t) + \widetilde{\boldsymbol{\Psi}}^{\mathrm{T}}(t)\widehat{\boldsymbol{\theta}}_k(t)]^{\mathrm{T}}\right\} \tag{36}$$

where the expectation is taken with respect to $\widetilde{\mathcal{Y}}_k(t)$ *and* $\mathcal{Y}_k(t)$. Note that the matrix $\delta_k(t)$ reflects prediction accuracy observed when the model is verified using an independent dataset.

We will work out the estimate of $\delta_k(t)$ in the case where $\widehat{\boldsymbol{\theta}}_k(t)$ is the WLS estimate given by (16), and the process $\{\mathbf{y}(t)\}$ can be regarded as stationary in the analysis interval $[t-k, t+k]$, i.e., it is governed by (14).

First, under the assumptions made above and some additional regression matrix invertibility conditions, such as those given in Niedźwiecki and Guo (1991), one can show that (see Appendix A)

$$\mathrm{E}[\widehat{\boldsymbol{\theta}}_k(t)] \cong \boldsymbol{\theta}, \qquad \mathrm{cov}[\widehat{\boldsymbol{\theta}}_k(t)] = \frac{\boldsymbol{\rho}\otimes\boldsymbol{\Phi}_0^{-1}}{N_k} + o\left(\frac{1}{N_k}\right) \tag{37}$$

where $\boldsymbol{\Phi}_0 = \mathrm{E}[\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{T}}(t)]$ and

$$N_k = \frac{\left[\sum\limits_{i=-k}^{k} v_k(i)\right]^2}{\sum\limits_{i=-k}^{k} v_k^2(i)} \cong k\,\frac{\left[\int_{-1}^{1} h^2(x)dx\right]^2}{\int_{-1}^{1} h^4(x)dx} \tag{38}$$

denotes the equivalent width of the window $\{v_k(i)\}$, or the equivalent estimation memory of the WLS algorithm (Niedźwiecki, 2000) (which in the literature on nonparametric spectrum estimation is usually referred to as its equivalent noise bandwidth).

Denote by $\Delta\widehat{\boldsymbol{\theta}}_k(t) = \widehat{\boldsymbol{\theta}}_k(t) - \boldsymbol{\theta}$ the parameter estimation error. Based on (37), and on the fact that the quantities $\widetilde{\boldsymbol{\varepsilon}}(t)$ and $\widetilde{\boldsymbol{\Psi}}(t)$ are mutually independent and independent of $\widehat{\boldsymbol{\theta}}_k(t)$, one obtains

$$\begin{aligned}
\delta_k(t) &= \mathrm{E}\left\{[\widetilde{\boldsymbol{\varepsilon}}(t) + \widetilde{\boldsymbol{\Psi}}^{\mathrm{T}}(t)\Delta\widehat{\boldsymbol{\theta}}_k(t)][\widetilde{\boldsymbol{\varepsilon}}(t) + \widetilde{\boldsymbol{\Psi}}^{\mathrm{T}}(t)\Delta\widehat{\boldsymbol{\theta}}_k(t)]^{\mathrm{T}}\right\} \\
&= \boldsymbol{\rho} + \mathrm{E}\left\{\widetilde{\boldsymbol{\Psi}}^{\mathrm{T}}(t)\Delta\widehat{\boldsymbol{\theta}}_k(t)\Delta\widehat{\boldsymbol{\theta}}_k^{\mathrm{T}}(t)\widetilde{\boldsymbol{\Psi}}(t)\right\} \\
&= \boldsymbol{\rho} + \mathrm{E}\left\{\widetilde{\boldsymbol{\Psi}}^{\mathrm{T}}(t)\mathrm{cov}[\widehat{\boldsymbol{\theta}}_k(t)]\widetilde{\boldsymbol{\Psi}}(t)\right\} \\
&\cong \boldsymbol{\rho} + \frac{1}{N_k}\,\mathrm{E}\left\{[\mathbf{I}\otimes\widetilde{\boldsymbol{\varphi}}^{\mathrm{T}}(t)][\boldsymbol{\rho}\otimes\boldsymbol{\Phi}_0^{-1}][\mathbf{I}\otimes\widetilde{\boldsymbol{\varphi}}(t)]\right\} \\
&= \boldsymbol{\rho} + \frac{1}{N_k}\,\mathrm{E}\left\{\boldsymbol{\rho}\otimes[\widetilde{\boldsymbol{\varphi}}^{\mathrm{T}}(t)\boldsymbol{\Phi}_0^{-1}\widetilde{\boldsymbol{\varphi}}(t)]\right\}
\end{aligned}$$

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

Since

$$\begin{aligned}
\mathrm{E}\left\{\boldsymbol{\rho}\otimes[\widetilde{\boldsymbol{\varphi}}^{\mathrm{T}}(t)\boldsymbol{\Phi}_0^{-1}\widetilde{\boldsymbol{\varphi}}(t)]\right\} &= \boldsymbol{\rho}\,\mathrm{tr}[\boldsymbol{\Phi}_0^{-1}\mathrm{E}\{\widetilde{\boldsymbol{\varphi}}(t)\widetilde{\boldsymbol{\varphi}}^{\mathrm{T}}(t)\}] \\
&= \boldsymbol{\rho}\,\mathrm{tr}[\boldsymbol{\Phi}_0^{-1}\boldsymbol{\Phi}_0] = \boldsymbol{\rho}\,mn
\end{aligned}$$

one finally arrives at

$$\delta_k(t) \cong \left[1 + \frac{mn}{N_k}\right]\boldsymbol{\rho}. \tag{39}$$

Note that the term $(mn/N_k)\boldsymbol{\rho}$ in (39), which grows when the estimation bandwidth decreases, can be interpreted as the prediction "loss" due to inaccuracy of parameter estimates.

Another useful relationship can be obtained by examining the WLS estimate of the driving noise covariance [which under (17) is approximately equal to the YW estimate of the same quantity]

$$\begin{aligned}
\widehat{\boldsymbol{\rho}}_k(t) = \frac{1}{L_k}\sum_{i=-k}^{k} v_k(i)&[\mathbf{y}(t+i) + \boldsymbol{\Psi}^{\mathrm{T}}(t+i)\widehat{\boldsymbol{\theta}}_k(t)] \\
&\times [\mathbf{y}(t+i) + \boldsymbol{\Psi}^{\mathrm{T}}(t+i)\widehat{\boldsymbol{\theta}}_k(t)]^{\mathrm{T}}
\end{aligned} \tag{40}$$

where

$$L_k = \sum_{i=-k}^{k} v_k(i) \tag{41}$$

denotes the effective width of the window $\{v_k(i)\}$.

It can be shown that under stationary conditions it holds (see Appendix B)

$$\mathrm{E}[\widehat{\boldsymbol{\rho}}_k(t)] \cong \left[1 - \frac{mn}{N_k}\right]\boldsymbol{\rho}. \tag{42}$$

Combining (42) with (39), one arrives at the following estimate of $\delta_k(t)$

$$\widehat{\delta}_k(t) = \frac{1 + \frac{mn}{N_k}}{1 - \frac{mn}{N_k}}\,\widehat{\boldsymbol{\rho}}_k(t) \tag{43}$$

leading to

$$\begin{aligned}
\widehat{k}(t) &= \arg\min_{k\in\mathcal{K}} \det[\widehat{\delta}_k(t)] \\
&= \arg\min_{k\in\mathcal{K}} \left[\frac{1 + \frac{mn}{N_k}}{1 - \frac{mn}{N_k}}\right]^m \det[\widehat{\boldsymbol{\rho}}_k(t)]
\end{aligned} \tag{44}$$

which can be recognized as the generalized version of the Akaike's multivariate final prediction error (MFPE) criterion (Akaike, 1971) proposed in Niedźwiecki (1984) (with a slightly different justification) for model order selection.

While the quantity $\det[\widehat{\boldsymbol{\rho}}_k(t)]$ in (44) tends to decrease when the bandwidth parameter k decreases (reflecting decrease in bias errors), the bandwidth-dependent multiplier behaves in the opposite way (reflecting increase in variance errors). Hence, the choice based on (44) allows one to balance the bias–variance tradeoff. Note that this mechanism is similar to that observed in the order selection case, where the residual noise variance decreases and the multiplier increases with growing n.

4.3. Determinant or trace?

When designing the bandwidth/order selection criteria, one can consider different scalar measures of the 'magnitude' of the final prediction error matrix (43), determinant and trace being the two obvious possibilities. To guarantee congruence between the cross-validation approach and MFPE, we proposed to minimize

determinant of the matrix $\mathbf{Q}_k^\circ(t)$. However, in principle, one can also consider the following trace variant of (31)

$$\widehat{k}(t) = \arg\min_{k\in\mathcal{K}} \operatorname{tr}\left[\mathbf{Q}_k^\circ(t)\right] = \arg\min_{k\in\mathcal{K}} \sum_{s\in T(t)} \|\mathbf{e}_k^\circ(s)\|^2. \tag{45}$$

4.4. Optimization of the bandwidth selection procedures

In this section we will try to find how one should choose the bandwidth parameters k_i, $i = 1, \ldots, K$ in order to maximize robustness of the parallel estimation schemes described above. Our considerations will be based on a hypothetical model of dependence of the local mean-squared parameter estimation error on the bandwidth parameter k. For univariate autoregressive processes such dependence was rigorously analyzed in Dahlhaus and Giraitis (1998). Assuming that the AR model is uniformly stable (i.e., the roots of the characteristic polynomial are uniformly bounded away from the unit circle), and that parameter trajectories are sufficiently smooth (uniformly bounded first, second and third derivatives), it was shown that

$$\mathrm{E}\left[\|\widehat{\boldsymbol{\theta}}_k(t) - \boldsymbol{\theta}(t)\|^2\right] \cong \frac{b_1(t)}{k} + b_2(t)k^4 \tag{46}$$

where the first term on the right hand side of (46) corresponds to the variance component of the MSE, and the second term is its (squared) bias component. The positive constants $b_1(t)$ and $b_2(t)$ depend on the shape of the data taper function $h(\cdot)$ and on the rate of signal nonstationarity measured by the second derivative of $\boldsymbol{\theta}(t)$ with respect to time (interestingly, but not surprisingly, the bias error is zero if $d\boldsymbol{\theta}(t)/dt \neq 0$ but $d^2\boldsymbol{\theta}(t)/dt^2 = 0$, i.e., if signal parameters vary linearly with time). The optimal instantaneous value k_{opt}, i.e., the one that minimizes (46) with respect to k, is given by[3]

$$k_{\mathrm{opt}}(t) = \left[\frac{b_1(t)}{4b_2(t)}\right]^{\frac{1}{5}}. \tag{47}$$

Of course, when the quantities $b_1(t)$ and $b_2(t)$ are unknown, which is almost always the case in practice, one cannot use the analytical formula (47).

Under the assumptions made in Dahlhaus and Giraitis (1998), the optimal data window is given by

$$h(x) = \sqrt{1-x^2}, \quad x \in [-1, 1] \tag{48}$$

which can be recognized as the square root of the Epanechnikov kernel (Epanechnikov, 1969), widely used in nonparametric probability density estimation.

It should be stressed that the exact form of the bias–variance tradeoff depends on the assumed degree of smoothness of parameter changes. As shown in Niedźwiecki and Gackowski (2011), when parameter trajectory can be modeled as a random process with orthogonal increments (such as random walk), the mean squared parameter estimation error can be expressed in the form

$$\mathrm{E}\left[\|\widehat{\boldsymbol{\theta}}_k(t) - \boldsymbol{\theta}(t)\|^2\right] \cong \frac{c_1(t)}{k} + c_2(t)k \tag{49}$$

leading to $k_{\mathrm{opt}}(t) = [c_1(t)/c_2(t)]^{1/2}$, and the optimal window shape is two-sided exponential. Even though this result was obtained for a different estimation problem – identification of a FIR system using the method of weighted least squares – its qualitative implications seem to be valid also in the currently discussed case.

The analysis carried out below will be based on the following more general model of the bias–variance tradeoff considered in Stanković (2004)

$$\mathrm{E}\left[\|\widehat{\boldsymbol{\theta}}_k(t) - \boldsymbol{\theta}(t)\|^2\right] = \frac{d_1(t)}{k^p} + d_2(t)k^r$$

where $p, r \geq 1$ are integer numbers.

Assuming that $d_1(t)$ and $d_2(t)$ vary slowly with time, so that in the analysis interval $T(t)$ both quantities can be regarded as (unknown) constants $[d_1(s) = d_1, d_2(s) = d_2, s \in T(t)]$, our local performance measure can be approximately written down in the form

$$I(k, d_1, d_2) = \frac{d_1}{k^p} + d_2 k^r. \tag{50}$$

For given values of d_1 and d_2 the minimum of $I(k, d_1, d_2)$ is attained for $k_{\mathrm{opt}} = (pd_1/rd_2)^{1/(p+r)}$, leading to

$$I(k_{\mathrm{opt}}, d_1, d_2) = d_1^{\frac{r}{p+r}} d_2^{\frac{p}{p+r}} \left(\left[\frac{p}{r}\right]^{\frac{r}{p+r}} + \left[\frac{r}{p}\right]^{\frac{p}{p+r}}\right).$$

Denote by $\delta = 1 + \varepsilon$, where ε is a small positive constant, the multiplier which allows one to specify what is meant by an "insignificant increase" of the performance measure. For example, if insignificant changes are defined as those not exceeding 10% of the optimal value, one should set $\delta = 1.1$. To determine the range of the values of $k \in \Omega_k = [\underline{k}, \overline{k}]$ for which the performance of the identification algorithm is "suboptimal" for fixed values of d_1 and d_2, one should solve for k the inequality

$$I(k, d_1, d_2) \leq \delta \min_k I(k, d_1, d_2) = \delta I(k_{\mathrm{opt}}, d_1, d_2). \tag{51}$$

The insensitivity zone Ω_k can be widened if instead of a single estimation algorithm, one uses K algorithms with different bandwidth parameters $k_1 < k_2 < \cdots k_K$. If such a parallel estimation scheme is equipped with an ideal switching rule, i.e., the one that always selects the best performance, the corresponding outcome is

$$I(\mathcal{K}, d_1, d_2) = \min\{I(k_i, d_1, d_2), k_i \in \mathcal{K}\}$$

and the insensitivity zone takes the form $\Omega_{\mathcal{K}} = \bigcup_{i=1}^K [\underline{k}_i, \overline{k}_i]$.

Consider the case where $K = 2$ (two algorithms working in parallel) and $k_2 = \gamma k_1$, $\gamma > 1$. Denote by k_* the coordinate of the intersection point of the characteristics $I(k_1, d_1, d_2)$ and $I(k_2, d_1, d_2)$. To maximize the size of the insensitivity zone while preserving its compactness, the value of γ should be chosen so as to guarantee that $\underline{k}_1 = \overline{k}_2 = k_*$ and (see Fig. 1)

$$I(k_*, d_1, d_2) = I(\gamma k_*, d_1, d_2) = \delta I(k_{\mathrm{opt}}, d_1, d_2). \tag{52}$$

Solving (52), one obtains

$$k_* = \left[\frac{d_1(\gamma^p - 1)}{d_2\gamma^p(\gamma^r - 1)}\right]^{\frac{1}{p+r}}$$

and

$$\delta = \frac{\left[\frac{\gamma^p(\gamma^r-1)}{\gamma^p-1}\right]^{\frac{p}{p+r}} + \left[\frac{\gamma^p-1}{\gamma^p(\gamma^r-1)}\right]^{\frac{r}{p+r}}}{\left[\frac{r}{p}\right]^{\frac{p}{p+r}} + \left[\frac{p}{r}\right]^{\frac{r}{p+r}}}. \tag{53}$$

Note that the relationship between the insensitivity multiplier δ and the bandwidth scaling coefficient γ, established above, *does not* depend on d_1 and d_2 and hence it holds true for *all* values of these constants (which are usually unknown).

[3] To avoid unnecessary complications, from this point on, k will be regarded as a real number; the approximate integer solution can be obtained by rounding k_{opt} to the nearest integer number.

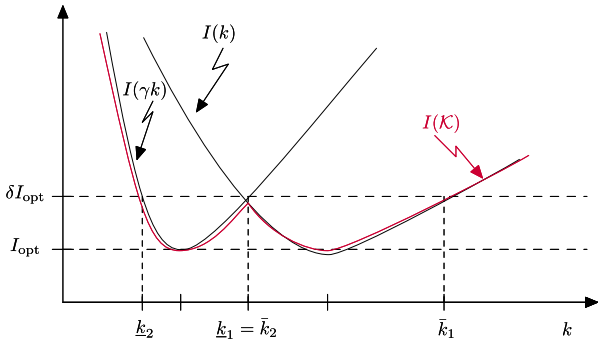

Fig. 1. Error characteristics corresponding to the "ideal" switching rule ($K = 2$).

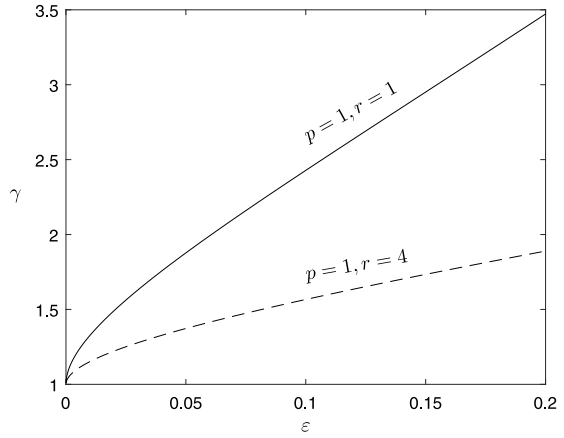

Fig. 2. Dependence of γ on ε in two cases discussed in the paper.

From the discussion carried out above it is clear that for $K > 2$ one should set $\bar{k}_{i+1} = \underline{k}_i$, $i = 1, \ldots, K-1$, which results in $k_{i+1} = \gamma^i k_1$, $i = 1, \ldots, K-1$ and $\Omega_{\mathcal{K}} = [\underline{k}_K, \bar{k}_1]$, i.e., to maximize robustness of the parallel estimation scheme, the consecutive bandwidth parameters k_i should form a geometric progression.

When γ obeys (53), the boundaries of the insensitivity zone, determined by (52), can be expressed in the form

$$\underline{k} = \left[\frac{r(\gamma^p - 1)}{p\gamma^p(\gamma^r - 1)}\right]^{\frac{1}{p+r}} k_{\mathrm{opt}}, \qquad \bar{k} = \left[\frac{r\gamma^p(\gamma^r - 1)}{p(\gamma^p - 1)}\right]^{\frac{1}{p+r}} k_{\mathrm{opt}}.$$

Fig. 2 shows dependence of γ on ε in two cases discussed above. For $\varepsilon = 0.1$ the corresponding values of γ obtained from (53) are equal to 2.43 for $p = r = 1$, and 1.57 for $p = 1, r = 4$. We note that for $p = r = 1$ such an inverse relationship can be established analytically: $\gamma = \left[\sqrt{\varepsilon^2 + 2\varepsilon} + \varepsilon + 1\right]^2$.

5. Joint bandwidth and order selection

So far we have been assuming that the number of estimated autocovariance matrices n, i.e., the order of the VAR model, is fixed prior to estimation. We will show that the proposed approaches can be easily extended to joint bandwidth and order selection.

In the context of maximum entropy spectrum estimation, selection of the model order is an important decision that must be taken (Stoica & Moses, 1997). When the order is too small, i.e., the VAR model is underfitted, some important information about the resonant structure of the analyzed process may be lost. If the order is too large, i.e., the corresponding model is overfitted, some nonexistent spectral resonances may be detected leading to false qualitative conclusions.

Suppose that, instead of a fixed-order model, for each bandwidth parameter k, one considers a family of VAR models of different orders $n \in \mathcal{N} = \{1, \ldots, N\}$ obtained by solving the Yule–Walker equations of the form

$$[\mathbf{I}, \widehat{\mathbf{A}}_{1,k|n}(t), \ldots, \widehat{\mathbf{A}}_{n,k|n}(t)]\, \widehat{\mathcal{R}}_{k|n}(t) = [\widehat{\boldsymbol{\rho}}_{k|n}(t), \mathbf{O}, \ldots, \mathbf{O}], \quad n = 1, \ldots, N. \tag{54}$$

The symbol $\widehat{\mathbf{A}}_{i,k|n}(t)$ denotes the estimate of the matrix $\mathbf{A}_i$ obtained for the model of order n and bandwidth k, and $\widehat{\boldsymbol{\rho}}_{k|n}(t)$ is the corresponding estimate of the covariance matrix $\boldsymbol{\rho}$—the additional subscript n was introduced to distinguish between models of different orders. Since the matrices $\widehat{\mathcal{R}}_{k|n}(t)$, $n = 1, \ldots, N$, made up of the covariance estimates (9), are nested, i.e., $\widehat{\mathcal{R}}_{k|n}(t) \prec \widehat{\mathcal{R}}_{k|n+1}(t)$, $n = 1, \ldots, N-1$ ($\mathbf{A} \prec \mathbf{B}$ means that $\mathbf{A}$ is the principal submatrix of $\mathbf{B}$), signal identification can be carried out in an order-recursive manner, i.e., for a given value of k, all VAR models of orders $n = 1, \ldots, N$ can be obtained during a single run of the WWR algorithm.

Denote by $\mathbf{e}^{\circ}_{k|n}(t)$ the leave-one-out signal interpolation error obtained (in the way described in the previous section) for the model of order n and bandwidth k, and let

$$\mathbf{Q}^{\circ}_{k|n}(t) = \sum_{s \in T(t)} \mathbf{e}^{\circ}_{k|n}(s)[\mathbf{e}^{\circ}_{k|n}(s)]^{\mathrm{T}}.$$

The spectral density estimate can be obtained from

$$\widehat{\mathbf{S}}(\omega, t) = \widehat{\mathbf{S}}_{\widehat{k}(t)|\widehat{n}(t)}(\omega, t) \tag{55}$$

where

$$\{\widehat{k}(t), \widehat{n}(t)\} = \arg\min_{\substack{k \in \mathcal{K} \\ n \in \mathcal{N}}} \det\left[\mathbf{Q}^{\circ}_{k|n}(t)\right]. \tag{56}$$

When the full cross-validatory analysis is applied, the matrix $\mathbf{Q}^{\circ}_{k|n}(t)$ in (56) should be replaced with $\mathbf{Q}^{\bullet}_{k|n}(t)$. Finally, when the prediction-oriented approach is used, the joint order/bandwidth selection rule becomes

$$\{\widehat{k}(t), \widehat{n}(t)\} = \arg\min_{\substack{k \in \mathcal{K} \\ n \in \mathcal{N}}} \left[\frac{1 + \frac{mn}{N_k}}{1 - \frac{mn}{N_k}}\right]^m \det[\widehat{\boldsymbol{\rho}}_{k|n}(t)]. \tag{57}$$

6. Simulation results

To evaluate different approaches to bandwidth and order selection, one needs information about evolution of true parameters and true instantaneous spectrum of the analyzed process. This precludes using real-world processes as the "ground truth" model behind such nonstationary data is usually not known. To generate artificial VAR process that has some practical relevance and yet fulfills the above requirement, we used the "morphing" technique. First, 3 time-invariant "anchor" VAR models (A, B, C), of order $n = 4$, were obtained by performing local identification of a stereo ($m = 2$) audio signal. The identified fragments differed in their resonance structures – see Fig. 3. Anchor models were specified in the lattice form $\{\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_n, \mathbf{R}_0\}$, where $\boldsymbol{\Delta}_i$, $i = 1, \ldots, n$ denote matrices of normalized reflection coefficients (partial correlation coefficients) which can be obtained as a byproduct of the WWR algorithm – see Complement C8.6 in Söderström and Stoica (1988). The lattice representation of a stable VAR model is unique and can be uniquely transformed into the direct representation $\{\mathbf{A}_1, \ldots \mathbf{A}_n, \boldsymbol{\rho}\}$. It is known that for a stable VAR model (such as the one obtained using the WWR algorithm) it holds that $\sigma_{\max}(\boldsymbol{\Delta}_i) < 1$, $i = 1, \ldots, n$, where $\sigma_{\max}(\boldsymbol{\Delta}_i)$ denotes the largest singular value of the matrix $\boldsymbol{\Delta}_i$, i.e., its spectral norm.

The time-variant model used to generate artificial VAR data was obtained by morphing one anchor model into another one. For

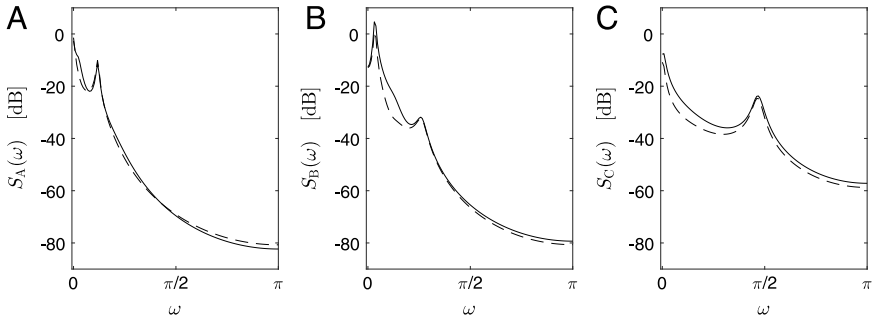

Fig. 3. Spectral density functions (autospectra) corresponding to 3 anchor models.

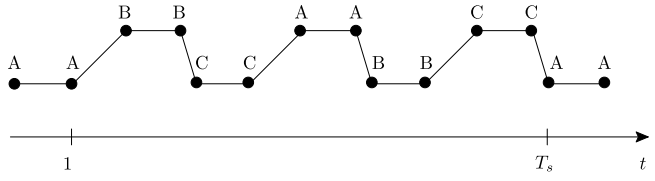

Fig. 4. Morphing scenario used in simulation tests.

example, the transition from the model $\{\boldsymbol{\Delta}_1^{\mathrm{A}}, \ldots, \boldsymbol{\Delta}_4^{\mathrm{A}}, \mathbf{R}_0^{\mathrm{A}}\}$, valid at the instant t_1, to the model $\{\boldsymbol{\Delta}_1^{\mathrm{B}}, \ldots, \boldsymbol{\Delta}_4^{\mathrm{B}}, \mathbf{R}_0^{\mathrm{B}}\}$, valid at the instant t_2, was realized using the following transformations

$$\mathbf{R}_0(t) = \mu(t)\mathbf{R}_0^{\mathrm{A}} + [1-\mu(t)]\mathbf{R}_0^{\mathrm{B}}$$
$$\boldsymbol{\Delta}_i(t) = \mu(t)\boldsymbol{\Delta}_i^{\mathrm{A}} + [1-\mu(t)]\boldsymbol{\Delta}_i^{\mathrm{B}}$$
$$i = 1, \ldots, 4, \; t \in [t_1, t_2]$$

where $\mu(t) = (t_2 - t)/(t_2 - t_1)$.

Using the triangle inequality, which holds for all matrix norms, one can easily show that $\sigma_{\max}[\boldsymbol{\Delta}_i(t)] < 1$, $i = 1, \ldots, 4$, $t \in [t_1, t_2]$, which means that the resulting time-variant model is at all times stable. It should be stressed that model stability *is not* guaranteed if the morphing technique is used to merge direct VAR representations.

The applied morphing scenario is symbolically depicted in Fig. 4. The generated signal $\{\mathbf{y}(t), \; t = 1, \ldots, T_s\}$ has periods of stationarity (A–A, B–B, C–C) interleaved with periods of nonstationary behavior, both slower (A–B, C–A, B–C) and faster (B–C, A–B, C–A). Data generation starts 1000 instants prior to $t = 1$ and continues for 1000 instants after $t = T_s$ (in both cases using the model A) so that no matter what bandwidth and model order, the estimation process and evaluation of its results can be in all cases started at the instant 1 and ended at the instant T_s.

To check performance of the compared algorithms under different rates of signal nonstationarity, 5 different values of T_s were adopted (2100, 4200, 8400, 16 800 and 33 600) corresponding to 5 different speeds of parameter variation, further denoted by S_1, S_2, S_3, S_4 and S_5, respectively (S_1 corresponds to the highest rate of nonstationarity and S_5-to the lowest rate).

Two instantaneous performance measures were used to evaluate simulation results: the squared parameter estimation error

$$d_{\mathrm{PAR}}(t) = \|\widehat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}(t)\|^2$$

which quantifies discrepancy between the estimated model and the true model in the time domain, and the relative entropy rate (RER) (Ferrante, Masiero, & Pavon, 2012)

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

Table 1
Estimation results obtained for 5 windows of the same equivalent width (Epanechnikov, Hann, rectangular, Bartlett, Gauss) under 5 speeds of parameter variation (S_1, S_2, S_3, S_4, S_5). The upper table shows mean relative entropy rates (RER) and the lower one—mean squared parameter estimation errors (PAR). The best results in each column are shown in boldface.

$\{v(t)\}$	S_1	S_2	S_3	S_4	S_5
RER					
Epan.	0.647	0.444	0.375	0.366	0.380
Hann	**0.432**	**0.169**	**0.077**	**0.051**	**0.050**
rect.	1.830	1.690	1.642	1.650	1.681
Bartlett	0.586	0.368	0.299	0.287	0.301
Gauss	0.764	0.576	0.512	0.503	0.519
PAR					
Epan.	5.584	6.299	7.172	7.903	9.074
Hann	**1.513**	**0.594**	**0.325**	**0.269**	**0.332**
rect.	30.145	33.215	34.751	35.075	35.047
Bartlett	4.151	4.404	5.193	5.777	6.659
Gauss	8.110	9.800	11.285	12.286	13.316

which is a multivariate extension of the Itakura–Saito spectral distortion measure.

Final evaluation was based on comparison of the mean scores obtained after combined time and ensemble averaging of $d_{\mathrm{PAR}}(t)$ and $d_{\mathrm{RER}}(t)$ (over $t \in [1, T_s]$ and 100 independent realizations of $\{\mathbf{y}(t)\}$).

Experiment 1

The aim of this experiment was to check the influence of the window shape on estimation accuracy. Five different lag windows $v(t)$ (Epanechnikov, Hann, rectangular, Bartlett and Gauss) were applied with the same equivalent width equal to 301 samples. The corresponding data tapers had the form $w(t) = \sqrt{v(t)}$. Table 1 summarizes results obtained for different rates of nonstationarity. It can be easily seen that estimation results depend quite strongly on the window shape. When the window is rectangular, i.e., no taper is applied, modeling errors are much higher than those obtained for bell-shaped windows. The Hann window (i.e., cosinusoidal taper) consistently yielded the best results, also in other simulation experiments, not reported here.

Interestingly, the Hann window systematically yielded better results than the Epanechnikov window recommended in Dahlhaus and Giraitis (1998), which is most probably caused by the fact that, unlike the first case, in the second case the derivative of the window prototyping function $h^2(x)$ is not equal to zero at both ends of the analysis interval. Based on this experience, in all further experiments the cosinusoidal data taper was used.

Experiment 2

The purpose of the second experiment was to check joint bandwidth and order selection properties of the proposed approaches. The parallel estimation scheme was made up of 5 algorithms with bandwidth parameters set to $k_1 = 100$, $k_2 = 150$, $k_3 = 225$, $k_4 = 337$ and $k_5 = 505$ ($\gamma = 1.5$). For cosinusoidal

Table 2
Comparison of mean RER scores obtained for 3 approaches to joint bandwidth and order selection (cross-validation—$CV_\circ$, full cross-validation—$CV_\bullet$, final prediction error—MFPE) under 5 speeds of parameter variation ($S_1, \ldots, S_5$); $k_1, \ldots, k_5$ denote different bandwidths of fixed-bandwidth–fixed-order algorithms working in parallel (the best scores are shown in boldface) and $\mathcal{K}_1 = \{k_1, k_2\}$, $\mathcal{K}_2 = \{k_1, k_2, k_3\}$, $\mathcal{K}_3 = \{k_1, k_2, k_3, k_4\}$, $\mathcal{K}_4 = \{k_1, k_2, k_3, k_4, k_5\}$ denote different configurations of bandwidth-order selection algorithms. The best results among $CV_\circ$, $CV_\bullet$ and MFPE (for each configuration) are shown in boldface. GT denotes ground truth results, i.e., results obtained for the true model order ($n = 4$) under optimal switching.

	k	S_1	S_2	S_3	S_4	S_5
Joint bandwidth and order selection (RER measure)						
$n = 4$	k_1	**0.261**	0.180	0.159	0.155	0.159
	k_2	0.291	**0.138**	0.094	0.083	0.087
	k_3	0.432	0.169	**0.077**	0.051	0.050
	k_4	0.705	0.269	0.097	**0.044**	0.032
	k_5	1.093	0.460	0.164	0.056	**0.026**
$n = 10$	k_1	0.378	0.311	0.289	0.283	0.285
	k_2	**0.340**	0.208	0.170	0.159	0.162
	k_3	0.432	**0.200**	0.121	0.098	0.097
	k_4	0.654	0.269	**0.120**	**0.072**	0.062
	k_5	1.001	0.424	0.166	**0.072**	**0.045**
GT	$\mathcal{K}_1$	0.218	0.119	0.090	0.082	0.087
	$\mathcal{K}_2$	0.207	0.099	0.061	0.048	0.050
	$\mathcal{K}_3$	0.206	0.093	0.050	0.033	0.030
	$\mathcal{K}_4$	0.206	0.093	0.047	0.027	0.020
$CV_\circ$	$\mathcal{K}_1$	0.780	0.665	0.628	0.622	0.666
	$\mathcal{K}_2$	0.859	0.655	0.587	0.563	0.597
	$\mathcal{K}_3$	1.117	0.776	0.644	0.601	0.614
	$\mathcal{K}_4$	1.446	0.953	0.738	0.664	0.652
$CV_\bullet$	$\mathcal{K}_1$	0.415	0.289	0.258	0.252	0.252
	$\mathcal{K}_2$	0.429	0.230	0.170	0.155	0.153
	$\mathcal{K}_3$	0.552	0.253	0.142	0.107	0.102
	$\mathcal{K}_4$	0.715	0.343	0.160	0.095	0.078
MFPE	$\mathcal{K}_1$	**0.246**	**0.149**	**0.116**	**0.104**	**0.106**
	$\mathcal{K}_2$	**0.245**	**0.136**	**0.095**	**0.077**	**0.074**
	$\mathcal{K}_3$	**0.255**	**0.134**	**0.089**	**0.066**	**0.059**
	$\mathcal{K}_4$	**0.286**	**0.136**	**0.088**	**0.062**	**0.052**

Table 3
Comparison of mean PAR scores obtained for 3 approaches to joint bandwidth and order selection (cross-validation—$CV_\circ$, full cross-validation—$CV_\bullet$, final prediction error—MFPE) under 5 speeds of parameter variation ($S_1, \ldots, S_5$); $k_1, \ldots, k_5$ denote different bandwidths of fixed-bandwidth–fixed-order algorithms working in parallel (the best scores are shown in boldface) and $\mathcal{K}_1 = \{k_1, k_2\}$, $\mathcal{K}_2 = \{k_1, k_2, k_3\}$, $\mathcal{K}_3 = \{k_1, k_2, k_3, k_4\}$, $\mathcal{K}_4 = \{k_1, k_2, k_3, k_4, k_5\}$ denote different configurations of bandwidth-order selection algorithms. The best results among $CV_\circ$, $CV_\bullet$ and MFPE (for each configuration) are shown in boldface. GT denotes ground truth results, i.e., results obtained for the true model order ($n = 4$) under optimal switching.

	k	S_1	S_2	S_3	S_4	S_5
Joint bandwidth and order selection (PAR measure)						
$n = 4$	k_1	**0.994**	0.827	0.896	0.985	1.286
	k_2	1.034	**0.539**	0.463	0.471	0.612
	k_3	1.513	0.594	**0.325**	0.269	0.332
	k_4	2.486	0.926	0.340	**0.200**	0.190
	k_5	4.389	1.510	0.536	0.203	**0.132**
$n = 10$	k_1	9.108	8.250	8.022	8.007	8.503
	k_2	7.531	5.955	5.526	5.356	5.564
	k_3	**7.056**	4.667	3.924	3.623	3.673
	k_4	7.683	**4.217**	2.974	2.511	2.445
	k_5	9.446	4.445	**2.599**	**1.860**	**1.665**
GT	$\mathcal{K}_1$	0.720	0.453	0.429	0.449	0.589
	$\mathcal{K}_2$	0.658	0.339	0.259	0.238	0.307
	$\mathcal{K}_3$	0.640	0.301	0.190	0.153	0.162
	$\mathcal{K}_4$	0.629	0.291	0.163	0.113	0.099
$CV_\circ$	$\mathcal{K}_1$	9.298	9.432	9.545	9.695	10.488
	$\mathcal{K}_2$	10.362	9.913	9.679	9.613	10.213
	$\mathcal{K}_3$	12.451	11.439	10.874	10.636	10.984
	$\mathcal{K}_4$	14.384	12.805	12.225	11.907	11.987
$CV_\bullet$	$\mathcal{K}_1$	4.286	3.872	4.012	4.263	4.460
	$\mathcal{K}_2$	3.415	2.575	2.515	2.597	2.759
	$\mathcal{K}_3$	3.929	2.283	1.815	1.704	1.831
	$\mathcal{K}_4$	4.956	2.666	1.707	1.355	1.387
MFPE	$\mathcal{K}_1$	**2.888**	**1.603**	**1.228**	**1.156**	**1.449**
	$\mathcal{K}_2$	**2.920**	**1.477**	**0.943**	**0.759**	**0.854**
	$\mathcal{K}_3$	**3.049**	**1.474**	**0.874**	**0.620**	**0.603**
	$\mathcal{K}_4$	**3.516**	**1.499**	**0.863**	**0.581**	**0.489**

taper the corresponding equivalent window widths are given by $N_k = 4(k+1)/3$. For both cross-validation approaches the width of the evaluation window was set to $D = 2d + 1 = 41$; this choice is by no means critical as almost identical results were obtained for all values of D from the interval [31, 51].

Tables 2 and 3 show the mean RER and PAR scores obtained for 4 different configurations of the compared bandwidth selection algorithms ($\mathcal{K}_1 = \{k_1, k_2\}$, $\mathcal{K}_2 = \{k_1, k_2, k_3\}$, $\mathcal{K}_3 = \{k_1, k_2, k_3, k_4\}$, $\mathcal{K}_4 = \{k_1, k_2, k_3, k_4, k_5\}$) and 5 speeds of parameter variation. Additionally, they show the scores obtained when the order is set to its maximum value $n = 10$ and to the true value $n = 4$, as well as the ground truth scores corresponding to the best switching scenario (determined experimentally). The average relative standard deviations of the results shown in Tables 2 and 3 are equal to 8.6% and 16.4%, respectively.

For all rates of parameter variation ($S_1, \ldots, S_5$) and all configurations of the parallel estimation scheme ($\mathcal{K}_1, \ldots, \mathcal{K}_4$), the best results are provided by the MFPE-based approach. Note that in majority of cases these results are better, or at least comparable, with those yielded by non-adaptive, fixed-bandwidth–fixed-order algorithms incorporated in the parallel scheme.

Experiment 3

The last experiment aimed at examining the evolution of joint bandwidth and order selection decisions in the borderline situation where both the coefficients and the order of the underlying signal model change abruptly [we note that the case of isolated parameter jumps is covered by the theory of locally stationary processes—see Dahlhaus, 2009]. According to the simulation scenario, depicted in Fig. 5, at the instant $t = 501$ the second-order forming filter (D) was switched to the fourth-order filter (E). Data generation was started 500 instants prior to $t = 1$ and continued for 500 instants after $t = 1000$. Evolution of the autospectrum of one of the signal channels is shown in Fig. 5. The parallel estimation scheme was made up of 5 algorithms with bandwidth parameters set to $k_1 = 44$, $k_2 = 66$, $k_3 = 100$, $k_4 = 150$ and $k_5 = 225$. Fig. 6 shows the locally time averaged histograms of the results of bandwidth and order selection (each time bin covers 20 consecutive time instants) obtained for 100 process realizations.

The results obtained in the MFPE case are satisfactory. Exactly as one would expect, the estimation bandwidth parameter is gradually decreased prior to the jump and gradually increased after the jump. Similarly, most of the time the correct model order is selected. Note that in the close vicinity of the jump, for $t \in [460, 540]$, MFPE selects high-order models. Since in this time interval even the shortest analyzed data frames, i.e., those corresponding to $k_1 = 44$, have a mixed spectral content, such behavior is fully understandable.

When the full cross-validation approach is used, both bandwidth and order selection statistics are less satisfactory than those observed for the MFPE approach.

Finally, when the standard cross-validation approach is applied, the results are unsatisfactory in both aspects—the estimation bandwidth parameter is overestimated and the model order is quite often underestimated.

Summary of simulation results

The main conclusions that can be drawn from our simulation study can be summarized as follows:

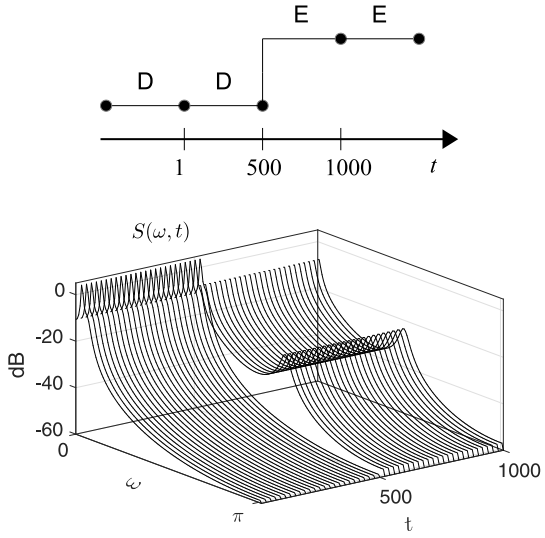

Fig. 5. Simulation scenario in the case of abrupt model change (top figure) and the corresponding time-varying autospectrum of one of the signal channels (bottom figure).

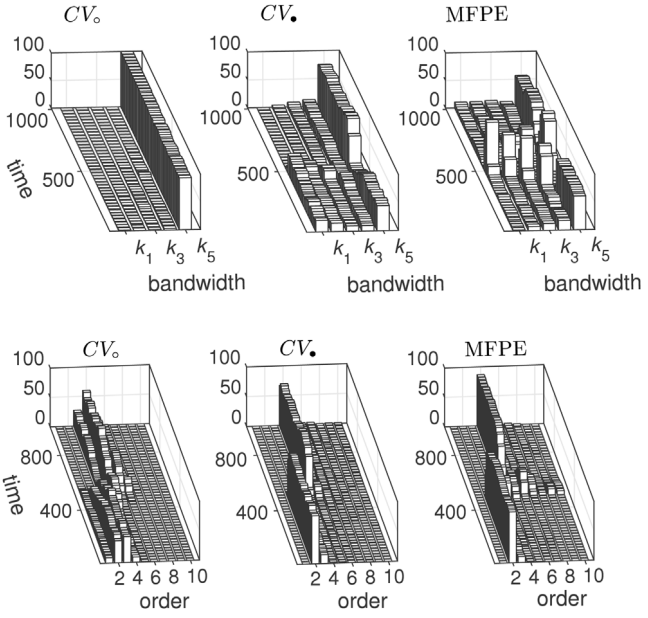

Fig. 6. Locally time averaged histograms of the results of bandwidth selection (upper figures) and order selection (lower figures).

(1) When VAR model is identified using the WWR algorithm, data tapering is an important and highly recommended preprocessing step. The proposed cosinusoidal taper not only allows for recursive computation of covariance estimates needed to run the WWR algorithm, but it also offers very good estimation bias–variance tradeoff compared with other windows of the same equivalent width.
(2) The multivariate version of the generalized Akaike's final prediction error (MFPE) criterion, originally proposed as a tool for model order selection only, yields very good results when applied to joint bandwidth and order selection. It is also attractive from the computational viewpoint as all quantities needed to evaluate the MFPE statistics are provided (at no additional computational cost) by the corresponding fixed-bandwidth WWR algorithms incorporated in the parallel estimation scheme.
(3) Both cross-validation approaches discussed in the paper yield worse results than the MFPE-based approach. Additionally, they are computationally much more involved. The full cross-validation approach performs better than its standard version.
(4) When applied to joint bandwidth and order selection, the MFPE-based parallel estimation schemes usually outperform the non-adaptive fixed-bandwidth–fixed-order algorithms.
(5) Adoption of the trace variants of the bandwidth/order selection criteria yielded results that were slightly inferior to those obtained using (56) and (57).

7. Conclusion

The problem of identification of nonstationary multivariate autoregressive process was considered and solved using the parallel estimation technique. It was shown that the most important task of parallel estimation – adaptive selection of the estimation bandwidth and order of the local autoregressive model – can be accomplished using the multivariate version of the generalized Akaike's final prediction error criterion. The resulting estimation scheme usually outperforms the non-adaptive fixed-bandwidth–fixed-order algorithms it is made up of. It is computationally attractive and does not rely on any subjectively determined quantities such as decision thresholds, confidence levels, etc.

The two alternative approaches presented in the paper, based on the concept of cross-validation, yield worse results than the final prediction error based criterion.

Appendix A. Outline of derivation of (37)

Straightforward calculations show that under (14) the solution of (16) obeys $\widehat{\boldsymbol{\theta}}_k(t) = \boldsymbol{\theta} - [\mathbf{I} \otimes \boldsymbol{\Phi}_k^{-1}(t)]\boldsymbol{\xi}_k(t)$, where

$$\boldsymbol{\Phi}_k(t) = \frac{1}{L_k} \sum_{i=-k}^{k} v_k(i)\boldsymbol{\varphi}(t+i)\boldsymbol{\varphi}^{\mathrm{T}}(t+i)$$

$$\boldsymbol{\xi}_k(t) = \frac{1}{L_k} \sum_{i=-k}^{k} v_k(i)\boldsymbol{\epsilon}(t+i) \otimes \boldsymbol{\varphi}(t+i).$$

Using one of the generalized versions of the strong law of large numbers for weighted sums of random variables [see e.g. Taylor, 1978], one can show that $\lim_{k\to\infty} \boldsymbol{\Phi}_k(t) = \mathrm{E}[\boldsymbol{\Phi}_k(t)] = \boldsymbol{\Phi}_0$, and hence $\lim_{k\to\infty}^{\text{a.s.}} \boldsymbol{\Phi}_k^{-1}(t) = \boldsymbol{\Phi}_0^{-1}$, which justifies the following approximation (valid as long as the effective width of the window is sufficiently large): $\widehat{\boldsymbol{\theta}}_k(t) - \boldsymbol{\theta} \cong -[\mathbf{I} \otimes \boldsymbol{\Phi}_0^{-1}]\boldsymbol{\xi}_k(t)$. Based on this approximation, one obtains $\mathrm{E}[\widehat{\boldsymbol{\theta}}_k(t)] \cong \boldsymbol{\theta}$ and

$$\mathrm{cov}[\widehat{\boldsymbol{\theta}}_k(t)] \cong [\mathbf{I} \otimes \boldsymbol{\Phi}_0^{-1}]\boldsymbol{\Gamma}_k(t)[\mathbf{I} \otimes \boldsymbol{\Phi}_0^{-1}] \tag{58}$$

where

$$\begin{aligned}\boldsymbol{\Gamma}_k(t) = \mathrm{E}\{\boldsymbol{\xi}_k(t)\boldsymbol{\xi}_k^{\mathrm{T}}(t)\} &= \frac{1}{L_k^2} \sum_{i_1=-k}^{k} \sum_{i_2=-k}^{k} v_k(i_1)v_k(i_2) \\ &\quad \times \mathrm{E}\{[\boldsymbol{\epsilon}(t+i_1)\boldsymbol{\epsilon}^{\mathrm{T}}(t+i_2)] \otimes [\boldsymbol{\varphi}(t+i_1)\boldsymbol{\varphi}^{\mathrm{T}}(t+i_2)]\} \\ &= \frac{1}{L_k^2} \sum_{i=-k}^{k} v_k^2(i)[\mathrm{E}\{\boldsymbol{\epsilon}(t+i)\boldsymbol{\epsilon}^{\mathrm{T}}(t+i)\}] \\ &\quad \otimes [\mathrm{E}\{\boldsymbol{\varphi}(t+i)\boldsymbol{\varphi}^{\mathrm{T}}(t+i)\}] = \frac{1}{N_k}\,\boldsymbol{\rho} \otimes \boldsymbol{\Phi}_0. \end{aligned} \tag{59}$$

The third transition in (59) stems from the fact that $\mathrm{E}[\boldsymbol{\epsilon}(t)] = \mathbf{0}\ \forall t$, $\boldsymbol{\epsilon}(t+i_1)$ and $\boldsymbol{\epsilon}(t+i_2)$ are mutually independent for $i_1 \neq i_2$, and $\boldsymbol{\epsilon}(t+i_1)$ is independent of $\boldsymbol{\varphi}(t+i_2)$ for $i_1 \geq i_2$.

Combining (58) with (59), one obtains

$$\mathrm{cov}[\widehat{\boldsymbol{\theta}}_k(t)] \cong \frac{1}{N_k}\,\boldsymbol{\rho} \otimes \boldsymbol{\Phi}_0^{-1}. \tag{60}$$

To show that the approximation in (60) holds up to terms of order $o(1/N_k)$, some additional technical assumptions must be made guaranteeing stochastic invertibility of the matrix $\mathbf{\Phi}_k(t)$ for finite values of k.

Appendix B. Outline of derivation of (37)

Let $\widehat{\rho}_{ij,k}(t) = [\widehat{\boldsymbol{\rho}}_k(t)]_{ij}$. Straightforward but tedious calculations lead to

$$\widehat{\rho}_{ij,k}(t) = \frac{1}{L_k}\sum_{l=-k}^{k} v_k(l)\epsilon_i(t+l)\epsilon_j(t+l) - \boldsymbol{\xi}_{j,k}^{\mathrm{T}}(t)\mathbf{\Phi}_k^{-1}(t)\boldsymbol{\xi}_{i,k}(t) = J_1(t) + J_2(t)$$

where $\boldsymbol{\xi}_{i,k}(t) = \frac{1}{L_k}\sum_{l=-k}^{k} v_k(l)\epsilon_i(t+l)\boldsymbol{\varphi}(t+l)$. Note that $\mathrm{E}[J_1(t)] = \rho_{ij}$. Furthermore, using the approximation $\mathbf{\Phi}_k^{-1}(t) \cong \mathbf{\Phi}_0^{-1}$, one obtains $\mathrm{E}[J_2(t)] \cong -\mathrm{tr}[\mathbf{\Phi}_0^{-1}\mathrm{E}\{\boldsymbol{\xi}_{i,k}(t)\boldsymbol{\xi}_{j,k}^{\mathrm{T}}(t)\}]$. According to (59), it holds that $\mathrm{E}\{\boldsymbol{\xi}_{i,k}(t)\boldsymbol{\xi}_{j,k}^{\mathrm{T}}(t)\} = [\mathbf{\Gamma}_k(t)]_{ij} = \frac{\rho_{ij}}{N_k}\,\mathbf{\Phi}_0$, leading to

$$\mathrm{E}[\widehat{\rho}_{ij,k}(t)] \cong \left[1 - \frac{mn}{N_k}\right]\rho_{ij}, \quad i,j = 1,\ldots,m$$

which is equivalent to (37).

References

Akaike, H. (1971). Autoregressive model fitting for control. *Annals of the Institute of Statistical Mathematics*, *23*, 163–180.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Bunke, O., Droge, B., & Polzehl, J. (1999). Model selection, transformations and variance estimation in nonlinear regression. *Statistics*, *33*, 197–240.

Burg, J.P. (1967). Maximum entropy spectral analysis. In *Proc. 37th meet. society of exploration geophysicists*.

Burg, J. P. (1975). *Maximum entropy spectral analysis*. (Ph.D. Dissertation), Stanford, CA: Dept. of Geophysics, Stanford University.

Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *The Annals of Statistics*, *30*, 351–413.

Dahlhaus, R. (2009). Local inference for locally stationary time series based on the empirical spectral measure. *Journal of Econometrics*, *151*, 101–112.

Dahlhaus, R. (2012). Locally stationary processes. *Handbook of Statistics*, *25*, 1–37.

Dahlhaus, R., & Giraitis, L. (1998). On the optimal segment length for parameter estimates for locally stationary time series. *Journal of Time Series Analysis*, *19*, 629–655.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability and its Applications*, *14*, 153–158.

Ferrante, A., Masiero, C., & Pavon, M. (2012). Time and spectral domain relative entropy: A new approach to multivariate spectral estimation. *IEEE Transactions on Automatic Control*, *57*, 2561–2575.

Friedl, H., & Stampfer, E. (2002). Cross-validation. In A. H. El-Shaarawi, & W. W. Piegorsch (Eds.), *Encyclopedia of environmetrics. Vol. 1* (pp. 452–460). Wiley.

Fu, Z., Chan, S.-C., Di, X., Biswal, B., & Zhang, Z. (2014). Adaptive covariance estimation of non-stationary processes and its application to infer dynamic connectivity from fMRI. *IEEE Transactions on Biomedical Circuits and Systems*, *8*, 228–239.

Goldenshluger, A., & Nemirovski, A. (1997). On spatial adaptive estimation of nonparametric regression. *Mathematical Methods of Statistics*, *6*, 135–170.

Katkovnik, V. (1999). A new method for varying adaptive bandwidth selection. *IEEE Transactions on Signal Processing*, *47*, 2567–2571.

Niedźwiecki, M. (1984). On the localized estimators and generalized Akaike's criteria. *IEEE Transactions on Automatic Control*, *29*, 970–983.

Niedźwiecki, M. (1985). Bayesian-like autoregressive spectrum estimation in the case of unknown process order. *IEEE Transactions on Automatic Control*, *30*, 950–961.

Niedźwiecki, M. (1993). Statistical reconstruction of multivariate time series. *IEEE Transactions on Signal Processing*, *41*, 451–457.

Niedźwiecki, M. (2000). *Identification of time-varying processes*. Wiley, 2000.

Niedźwiecki, M. (2010). Easy recipes for cooperative smoothing. *Automatica*, *46*, 716–720.

Niedźwiecki, M. (2012). Locally adaptive cooperative Kalman smoothing and its application to identification of nonstationary stochastic systems. *IEEE Transactions on Signal Processing*, *60*, 48–59.

Niedźwiecki, M., & Gackowski, S. (2011). On noncausal weighted least squares identification of nonstationary stochastic systems. *Automatica*, *47*, 2239–2245.

Niedźwiecki, M., & Gackowski, S. (2013). New approach to noncausal identification of nonstationary stochastic FIR systems subject to both smooth and abrupt parameter changes. *IEEE Transactions on Automatic Control*, *58*, 1847–1853.

Niedźwiecki, M., & Guo, L. (1991). Nonasymptotic results for finite-memory WLS filters. *IEEE Transactions on Automatic Control*, *36*, 515–522.

Priestley, M. B. (1965). Evolutionary spectra and non-stationary processes. *Journal of the Royal Statistical Society, Series B*, *27*, 204–237.

Rissanen, J. (1978). Modeling by shortest data descriptiona. *Automatica*, *14*, 465–471.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Söderström, T., & Stoica, P. (1988). *System identification*. Englewood Cliffs NJ: Prentice-Hall.

Stanković, L. (2004). Performance analysis of the adaptive algorithm for bias-to-variance tradeoff. *IEEE Transactions on Signal Processing*, *52*, 1228–1234.

Stoica, P., & Moses, R. L. (1997). *Introduction to spectral analysis*. Prentice Hall.

Taylor, R. L. (1978). Stochastic convergence of weighted sums of random elements in linear spaces. *Lecture Notes in Mathematics*, *672*.

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.

Yoshinobu Kajikawa received the B.Eng. and M.Eng. degrees in electrical engineering from Kansai University, Osaka, Japan, and the D.E. degree in communication engineering from Osaka University in 1991, 1993, and 1997, respectively. He joined Fujitsu Ltd., Kawasaki, Japan, in 1993 and engaged in research on active noise control. In 1994, he joined Kansai University, where he is now a professor. His current research interests lie in the area of signal processing for acoustic systems.

Dr. Kajikawa is a senior member of the Institute of Electronics, Information and Communication Engineers (IEICE) and the Institute of Electrical and Electronics Engineers (IEEE), and a member of the European Association for Signal Processing (EURASIP), the Acoustical Society of Japan (ASJ), and the Asia and Pacific Signal and Information Processing Association (APSIPA). He is currently serving as an associate editor for the Journal of the ASJ and IET Signal Processing. He is a member-at-large of BoG in APSIPA. He is the author or coauthor of more than 180 articles in journals and conference proceedings and has been responsible for 7 patents. He received the 2012 Sato Prize Paper Award from the ASJ, and the Best Paper Award in APCCAS 2014.