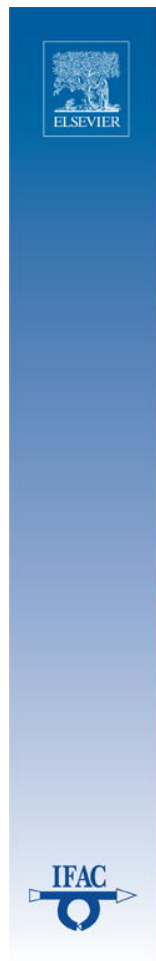


Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



ISSN 0005-1098

automatica

A Journal of IFAC The International
Federation of Automatic Control

www.elsevier.com/locate/automatica

This article was published in an Elsevier journal. The attached copy is furnished to the author for non-commercial research and education use, including for instruction at the author's institution, sharing with colleagues and providing to institution administration.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



Brief paper

On noncausal weighted least squares identification of nonstationary stochastic systems[☆]

Maciej Niedźwiecki¹, Szymon Gackowski

Faculty of Electronics, Telecommunications and Computer Science, Department of Automatic Control, Gdańsk University of Technology, Narutowicza 11/12, Gdańsk, Poland

ARTICLE INFO

Article history:

Received 3 November 2010

Received in revised form

15 February 2011

Accepted 8 April 2011

Available online 27 August 2011

Keywords:

System identification

Nonstationary processes

Noncausal estimation

ABSTRACT

In this paper, we consider the problem of noncausal identification of nonstationary, linear stochastic systems, i.e., identification based on prerecorded input/output data. We show how several competing weighted (windowed) least squares parameter smoothers, differing in memory settings, can be combined together to yield a better and more reliable smoothing algorithm. The resulting parallel estimation scheme automatically adjusts its smoothing bandwidth to the unknown, and possibly time-varying, rate of nonstationarity of the identified system. We optimize the window shape for a certain class of parameter variations and we derive computationally attractive recursive smoothing algorithms for such an optimized case.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

Consider the problem of identification of a discrete-time stochastic system governed by

$$y(t) = \boldsymbol{\varphi}^T(t)\boldsymbol{\theta}(t) + v(t) \quad (1)$$

where $t = \dots, -1, 0, 1, \dots$ denotes the normalized time, $y(t)$ denotes the system output, $\boldsymbol{\varphi}(t) = [\varphi_1(t), \dots, \varphi_n(t)]^T$ denotes the vector of input (regression) variables, $v(t)$ denotes measurement noise – a sequence of zero-mean independent random variables, and $\boldsymbol{\theta}(t) = [\theta_1(t), \dots, \theta_n(t)]^T$ is the vector of unknown, time-varying system coefficients.

Identification of nonstationary dynamic systems can be carried out using different frameworks, such as the local estimation approach, the basis function approach, or the approach based on Kalman filtering – to name only the most popular ones (Niedźwiecki, 2000).

In spite of methodological differences, the corresponding identification algorithms share one common feature – they all have finite estimation memory, i.e., they gradually “forget” information coming from the remote past as the new data becomes available. The appropriate choice of estimation memory is one

of the key issues in identification of nonstationary systems. The best results can be obtained if the estimation memory of the identification algorithm is selected so as to match the rate of nonstationarity of the analyzed system, trading off the variance and bias error components of the mean-squared parameter estimation error. When the rate of parameter changes varies with time, the estimation memory should be adjusted accordingly.

For causal estimation schemes (the corresponding identification algorithms are often referred to as parameter trackers), all aspects of identification of nonstationary systems, summarized above, have been pretty well worked out. While causality is an obvious requirement in all real-time prediction-oriented or control-oriented applications, there are many system identification tasks that can be performed off-line, based on the entire available observation record. Reconstruction of parameter trajectories of a time-varying communication channel, based on prerecorded input/output sequences (e.g. for simulation purposes), is a good example of such a problem. In cases like the one mentioned above, parameter estimates at any time instant can be based on both “past” and “future” measurements. Noncausal estimation schemes result in algorithms that are called parameter smoothers. Even though parameter smoothers yield better results than parameter trackers, their practical use is still very limited, mainly because of high computational requirements of the available procedures.

In this paper we present a new approach to the problem of parameter smoothing: (1) we design a parallel estimation scheme combining estimates yielded by several weighted least squares smoothers, characterized by different memory settings; (2) we show that statistics needed to design such a cooperative smoother can be expressed in terms of residual errors; (3) we prove

[☆] This work was supported by the Foundation for Polish Science. The material in this paper was presented at the 18th IFAC World Congress, August 28–September 2, 2011, Milano, Italy. This paper was recommended for publication in revised form by Associate Editor Er-Wei Bai under the direction of Editor Torsten Söderström.

E-mail addresses: maciekn@eti.pg.gda.pl (M. Niedźwiecki), szymon.gackowski@gmail.com (S. Gackowski).

¹ Tel.: +48 58 3472519; fax: +48 58 3415821.

that for parameter changes governed by the random-walk-plus-jumps model the optimal window shape is double exponential; (4) we present computationally attractive recursive smoothing algorithms that can be used in the double exponential weighting case.

2. Cooperative smoothing

Suppose that the observation history $\Omega(N) = \{y(N), \phi(N)\}$, $\mathcal{Y}(N) = \{y(1), \dots, y(N)\}$, $\Phi(N) = \{\phi(1), \dots, \phi(N)\}$, made up of N consecutive samples of the input/output variables, is available. We will look for the noncausal estimator $\hat{\theta}(t) = f[\Omega(N)]$ which, under certain constraints, minimizes the mean-squared estimation error $E[\|\hat{\theta}(t) - \theta(t)\|^2]$.

Our design will incorporate a bank of K weighted least squares (WLS) smoothing algorithms (known also as kernel smoothers), working in parallel and yielding the estimates

$$\hat{\theta}_k(t) = \mathbf{R}_k^{-1}(t) \mathbf{r}_k(t), \quad k = 1, \dots, K \quad (2)$$

where $\mathbf{R}_k(t) = \sum_{i=1}^N w_k(t-i) \phi(i) \phi^T(i)$, $\mathbf{r}_k(t) = \sum_{i=1}^N w_k(t-i) y(i) \phi(i)$, and $w_k(i)$ denotes the symmetric, nonnegative and sidewise nonincreasing weighting (window) sequence

$$w_k(0) = 1 \geq w_k(1) \geq \dots \geq 0$$

$$w_k(-i) = w_k(i), \quad \forall i > 0, \quad \sum_{i=-\infty}^{\infty} w_k(i) < \infty. \quad (3)$$

Weighting is used to “localize” estimation, i.e., to make estimation results less sensitive to observations collected in the remote past and future. Note that when $w_k(i) \equiv 1$, $\hat{\theta}_k(t)$ reduces down to the ordinary least squares estimator.

The most important characteristic of a WLS smoother is its estimation memory. Suppose that

(A1) The sequence of regression vectors $\{\phi(t)\}$, independent of $\{v(t)\}$, is a stationary and ergodic process with positive definite correlation matrix $E[\phi(t) \phi^T(t)] = \Phi_0 > 0$.

Assumption (A1) is met, for example, by finite impulse response (FIR) systems subject to a stationary and persistent excitation, such as telecommunication channels. When the system is time-invariant, i.e., $\theta(t) = \theta_0$, $E[v^2(t)] = \sigma_v^2$, $\forall t$, one can show that (Niedźwiecki, 2000)

$$\text{cov}[\hat{\theta}_k(t)] \cong \frac{\sigma_v^2 \Phi_0^{-1}}{\zeta_k(t)}, \quad \zeta_k(t) = \frac{\left[\sum_{i=1}^N w_k(t-i) \right]^2}{\sum_{i=1}^N w_k^2(t-i)}. \quad (4)$$

The quantity $\zeta_k(t)$ denotes the so-called equivalent window width. According to (4), in the time-invariant case, all WLS smoothers characterized by the same value of $\zeta_k(t)$ are equivalent from the viewpoint of estimation accuracy (irrespective of the shape of the corresponding windows). Consequently, $\zeta_k(t)$ characterizes the amount of information about θ_0 which can be extracted from the input/output data as a result of applying the method of weighted least squares. The equivalent window width is a more adequate measure of estimation memory than the (better known) effective window width $\eta_k(t) = \sum_{i=1}^N w_k(t-i)$, especially when one wants to compare smoothing capabilities of WLS estimators characterized by different window shapes (Niedźwiecki, 2000).

Estimation memory should be matched to the rate of nonstationarity of the identified process, trading off the bias and variance error components. Short-memory smoothers are “flexible” (yield small estimation bias) but “inaccurate” (yield large estimation

variance), whereas long-memory smoothers are “rigid” but “accurate”. Whenever the identified system undergoes rapid changes, the memory of the smoothing algorithm (often referred to as its smoothing bandwidth) should be shortened, so as to allow for good trajectory matching; when the system changes are slow, the memory should be increased to make the parameter estimates more accurate. Based on these observations, our bank of WLS filters will be made up of algorithms with the same window shape, but different memory spans. Selection of the window shape will be discussed later in Section 3.

The problem of combining several smoothed estimates to obtain an overall estimate of improved quality, was recently considered in Niedźwiecki (2010). Assume that

(A2) The zero-mean white measurement noise $v(t)$ is distributed according to the generalized normal law² $v \sim \mathcal{G}_\beta \mathcal{N}(0, \alpha, \beta)$:

$$p(v; \chi, \alpha, \beta) = \frac{\beta}{2\alpha \Gamma(1/\beta)} \exp \left\{ - \left(\frac{|v - \chi|}{\alpha} \right)^\beta \right\}$$

where $\chi = 0$ is the location parameter, $\alpha > 0$ is the unknown scale parameter and $\beta \geq 1$ is the known shape parameter; $\Gamma(\cdot)$ denotes the Euler’s gamma function,

(A3) The process $\{\theta(t)\}$ is independent of $\{\phi(t)\}$ and $\{v(t)\}$.

Both assumptions hold true for majority of FIR systems, including nonstationary telecommunication channels.

Under (A2) and (A3) the combined estimate can be obtained in the following form

$$\hat{\theta}(t) = \sum_{k=1}^K \mu_k(t) \hat{\theta}_k(t), \quad \mu_k(t) = \frac{\xi_k(t)}{\sum_{k=1}^K \xi_k(t)} \quad (5)$$

where $\mu_k(t)$, $k = 1, \dots, K$, denote credibility coefficients (related to posterior probabilities of different trajectory “patterns”) and the quantities $\xi_k(t)$ are evaluated according to

$$\xi_k(t) = \left[\sum_{i \in T(t)} |\varepsilon_k^\circ(i)|^\beta \right]^{-M/\beta} \quad (6)$$

where $M = 2m + 1$ denotes the width of the local evaluation frame $T(t) = [t - m, t + m]$ centered at t . When $\beta \rightarrow \infty$ (the uniform noise case), one should set

$$\xi_k(t) = \left[\max_{i \in T(t)} |\varepsilon_k^\circ(i)| \right]^{-M}.$$

The matching errors $\varepsilon_k^\circ(i)$, used to determine credibility coefficients, are defined as $\varepsilon_k^\circ(i) = y(i) - \phi^T(i) \hat{\theta}_k^\circ(i)$ where $\hat{\theta}_k^\circ(i)$ denotes the *holey* smoother associated with $\hat{\theta}_k(i)$ – the estimation procedure that is identical with the original one, except that it excludes the “central” sample $y(t)$ from the set of measurements used for estimation of $\theta(t)$:

$$\hat{\theta}_k^\circ(t) = [\mathbf{R}_k^\circ(t)]^{-1} \mathbf{r}_k^\circ(t), \quad k = 1, \dots, K \quad (7)$$

where $\mathbf{R}_k^\circ(t) = \sum_{i=1, i \neq t}^N w_k(t-i) \phi(i) \phi^T(i)$ and $\mathbf{r}_k^\circ(t) = \sum_{i=1, i \neq t}^N w_k(t-i) y(i) \phi(i)$.

Unlike residual errors $\varepsilon_k(i) = y(i) - \phi^T(i) \hat{\theta}_k(i)$, matching errors are pointwise independent of the measurement noise $v(i)$, and hence they allow for approximately unbiased evaluation of the local performance of $\hat{\theta}_k(i)$.

² The generalized normal law incorporates such practically important distributions as Gaussian ($\beta = 2$), Laplace ($\beta = 1$), and uniform ($\beta \rightarrow \infty$).

The combination scheme (5)–(6) can be regarded as a Bayesian extension of the leave-one-out cross-validation approach to selection of smoothing bandwidth (Friedl & Stampfer, 2002). We will show that matching errors $\varepsilon_k^\circ(i)$ can be computed without actually implementing the corresponding holey smoothers.

Corollary 1. *Matching errors corresponding to WLS smoothers can be expressed in terms of residual errors*

$$\varepsilon_k^\circ(i) = \varepsilon_k(i)/[1 - q_k(t)] \tag{8}$$

where $q_k(t) = \boldsymbol{\varphi}^\top(t)\mathbf{R}_k^{-1}(t)\boldsymbol{\varphi}(t)$.

Proof. Note that $\hat{\boldsymbol{\theta}}_k^\circ(t) = [\mathbf{R}_k(t) - \boldsymbol{\varphi}(t)\boldsymbol{\varphi}^\top(t)]^{-1} \times [\mathbf{R}_k\hat{\boldsymbol{\theta}}_k(t) - y(t)\boldsymbol{\varphi}(t)]$. It is straightforward to check that (provided that all inverses below exist)

$$[\mathbf{R}_k(t) - \boldsymbol{\varphi}(t)\boldsymbol{\varphi}^\top(t)]^{-1} = \mathbf{R}_k^{-1}(t) + \frac{\mathbf{R}_k^{-1}(t)\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^\top(t)\mathbf{R}_k^{-1}(t)}{1 - \boldsymbol{\varphi}^\top(t)\mathbf{R}_k^{-1}(t)\boldsymbol{\varphi}(t)}$$

Combining both equations, one arrives at

$$\begin{aligned} \varepsilon_k^\circ(t) &= y(t) - \boldsymbol{\varphi}^\top(t)\hat{\boldsymbol{\theta}}_k^\circ(t) = y(t) - \boldsymbol{\varphi}^\top(t)\hat{\boldsymbol{\theta}}_k(t) \\ &\quad + q_k(t)y(t) - \frac{q_k(t)\boldsymbol{\varphi}^\top(t)\hat{\boldsymbol{\theta}}_k(t)}{1 - q_k(t)} + \frac{q_k^2(t)y(t)}{1 - q_k(t)} \\ &= [y(t) - \boldsymbol{\varphi}^\top(t)\hat{\boldsymbol{\theta}}_k(t)]/[1 - q_k(t)]. \quad \square \end{aligned}$$

3. Window choice and efficient computational procedures

3.1. Optimization of the window shape

Suppose that parameter trajectory can be modeled as a process with orthogonal increments

(A4) The process $\{\boldsymbol{\theta}(t)\}$ is governed by

$$\boldsymbol{\theta}(t) = \boldsymbol{\theta}(t - 1) + \mathbf{z}(t), \quad \mathbf{z}(t) = \mathbf{z}_1(t) + \delta(t)\mathbf{z}_2(t)$$

where $\{\mathbf{z}_1(t)\}$, $\{\mathbf{z}_2(t)\}$ and $\{\delta(t)\}$ denote sequences of mutually orthogonal and self-orthogonal random variables such that $\text{cov}[\mathbf{z}_1(t)] = \mathbf{Z}_1$, $\text{cov}[\mathbf{z}_2(t)] = \mathbf{Z}_2$, $\forall t$ and $\delta(t) = \{1 \text{ w.p. } p_0, 0 \text{ w.p. } (1 - p_0)\}$, where $p_0, 0 \leq p_0 \ll 1$, denotes the probability of a jump at any instant t .

Note that for “small” values of p_0 and \mathbf{Z}_1 , and “large” values of \mathbf{Z}_2 , the process $\{\boldsymbol{\theta}(t)\}$ described by (A4) has two distinct modes of variation: the slow mode, governed by the random-walk equation (\mathbf{z}_1 -components) and resulting in a random drift of parameter values, and the fast mode, described by the “sparse” random-walk equation (\mathbf{z}_2 -components), resulting in occasional jump changes of parameter values. Such a random-walk-plus-jumps (RWJ) model was originally proposed in Falconer and Gitlin (1978) to describe a class of time-varying telecommunication channels.

Under assumptions (A1)–(A4) it is possible to show that among all windows of the same equivalent width, the one-sided exponential window guarantees the best tracking performance (Niedźwiecki, 1986). The exponentially weighted least squares (EWLS) trackers have the form

$$\hat{\boldsymbol{\theta}}_k(t) = \tilde{\mathbf{R}}_k^{-1}(t)\tilde{\mathbf{r}}_k(t), \quad k = 1, \dots, K \tag{9}$$

where $\tilde{\mathbf{R}}_k(t) = \sum_{i=1}^t \lambda_k^{t-i} \boldsymbol{\varphi}(i)\boldsymbol{\varphi}^\top(i)$, $\tilde{\mathbf{r}}_k(t) = \sum_{i=1}^t \lambda_k^{t-i} y(i)\boldsymbol{\varphi}(i)$ and $\lambda_k, 0 < \lambda_k < 1, k = 1, \dots, K$, denote the so-called forgetting constants. This corresponds to choosing $w_k(i) = \lambda_k^i, i \geq 0$. The recursive algorithm for evaluation of $\hat{\boldsymbol{\theta}}_k(t)$ is undoubtedly the best-known and the most frequently used procedure for causal identification (tracking) of nonstationary systems.

Similar analysis can be carried out for WLS smoothers. It leads to the following.

Table 1
Exact smoothing algorithm.

Forward-time run:
 $[\tilde{\mathbf{R}}_k(0) = 0, \tilde{\mathbf{r}}_k(0) = 0]$
 $\tilde{\mathbf{R}}_k(t) = \lambda_k \tilde{\mathbf{R}}_k(t - 1) + \boldsymbol{\varphi}(t)\boldsymbol{\varphi}^\top(t)$
 $\tilde{\mathbf{r}}_k(t) = \lambda_k \tilde{\mathbf{r}}_k(t - 1) + y(t)\boldsymbol{\varphi}(t)$
 $t = 1, \dots, N$
Backward-time run:
 $[\mathbf{R}_k(N) = \tilde{\mathbf{R}}_k(N), \mathbf{r}_k(N) = \tilde{\mathbf{r}}_k(N)]$
 $\mathbf{R}_k(t) = \lambda_k \mathbf{R}_k(t + 1) + (1 - \lambda_k^2)\tilde{\mathbf{R}}_k(t)$
 $\mathbf{r}_k(t) = \lambda_k \mathbf{r}_k(t + 1) + (1 - \lambda_k^2)\tilde{\mathbf{r}}_k(t)$
 $\hat{\boldsymbol{\theta}}_k(t) = \mathbf{R}_k^{-1}(t)\mathbf{r}_k(t)$
 $t = N - 1, \dots, 1$

Corollary 2. *Under assumptions (A1)–(A4) the best smoothing results can be obtained when the window is double exponential:*

$$w_k(i) = \lambda_k^{|i|}, \quad \forall i.$$

Proof. see Appendix.

The corresponding EWLS smoothers are given by

$$\hat{\boldsymbol{\theta}}_k(t) = \mathbf{R}_k^{-1}(t)\mathbf{r}_k(t), \quad k = 1, \dots, K \tag{10}$$

where $\mathbf{R}_k(t) = \sum_{i=1}^N \lambda_k^{|t-i|} \boldsymbol{\varphi}(i)\boldsymbol{\varphi}^\top(i)$ and $\mathbf{r}_k(t) = \sum_{i=1}^N \lambda_k^{|t-i|} y(i)\boldsymbol{\varphi}(i)$. Below we will work out recursive procedures for computation of $\hat{\boldsymbol{\theta}}_k(t)$ and $\mu_k(t)$. \square

3.2. Exact smoothing algorithm

The smoothed estimate (10) can be computed recursively. The procedure is two-step. The first, forward-time run, is needed to compute the quantities $\tilde{\mathbf{R}}_k(t)$ and $\tilde{\mathbf{r}}_k(t)$ [cf. (9)]. The quantities $\mathbf{R}_k(t)$, $\mathbf{r}_k(t)$ and $\hat{\boldsymbol{\theta}}_k(t)$ are computed from $\tilde{\mathbf{R}}_k(t)$ and $\tilde{\mathbf{r}}_k(t)$ during the second, backward-time run. The entire procedure was summarized in Table 1.

The exact algorithm has two drawbacks. First, prior to computation of $\hat{\boldsymbol{\theta}}_k(t)$, all quantities $\tilde{\mathbf{R}}_k(t)$, $\tilde{\mathbf{r}}_k(t)$, $t = 1, \dots, N$ have to be stored in the computer memory. Second, since – unlike $\tilde{\mathbf{R}}_k^{-1}(t)$ – the matrix $\mathbf{R}_k^{-1}(t)$ cannot be computed recursively, a direct inversion of the $n \times n$ matrix $\mathbf{R}_k(t)$ must be performed at each step of the backward-time algorithm.

3.3. Simplified smoothing algorithm

Under (A1), for sufficiently large values of the effective window width, it holds that (Niedźwiecki, 2000) $\mathbf{R}_k^{-1}(t) \cong \boldsymbol{\Phi}_0^{-1}/\eta_k(t)$. Combining this result with (10), one obtains the following approximation $\mathbf{r}_k(t) \cong \eta_k(t)\boldsymbol{\Phi}_0\hat{\boldsymbol{\theta}}_k(t)$. In an analogous way, one arrives at $\tilde{\mathbf{r}}_k(t) \cong \tilde{\eta}_k(t)\boldsymbol{\Phi}_0\hat{\boldsymbol{\theta}}_k(t)$, where $\tilde{\eta}_k(t) = \sum_{i=1}^t \lambda_k^{t-i} = (1 - \lambda_k^t)/(1 - \lambda_k)$ denotes the effective width of the one-sided exponential window. Substituting both approximations into the recursive equation derived for $\mathbf{r}_k(t)$ (see Table 1), and multiplying both sides of this equation by $\boldsymbol{\Phi}_0^{-1}$, one obtains $\hat{\boldsymbol{\gamma}}_k(t) = \lambda_k \hat{\boldsymbol{\gamma}}_k(t + 1) + (1 - \lambda_k^2)\tilde{\boldsymbol{\gamma}}_k(t)$ where $\hat{\boldsymbol{\gamma}}_k(t) = \eta_k(t)\hat{\boldsymbol{\theta}}_k(t)$ and $\tilde{\boldsymbol{\gamma}}_k(t) = \tilde{\eta}_k(t)\hat{\boldsymbol{\theta}}_k(t)$.

Finally, note that the effective window widths $\tilde{\eta}_k(t)$ and $\eta_k(t)$ can be computed recursively using the following equations: $\tilde{\eta}_k(t) = \lambda_k \tilde{\eta}_k(t - 1) + 1, t = 1, \dots, N$, and $\eta_k(t) = \lambda_k \eta_k(t + 1) + (1 - \lambda_k^2)\tilde{\eta}_k(t), t = N - 1, \dots, 1$, respectively. The simplified EWLS smoother was summarized in Table 2.

The core of the forward-time procedure is constituted by the well-known recursive EWLS algorithm – the recursive version of (9). The matrix $\mathbf{P}_k(t)$ updated by this algorithm is equal to $\mathbf{R}_k^{-1}(t)$, i.e., inversion of the regression matrix is carried out recursively. Note that the only quantities that should be memorized during the forward-time run are the $n \times 1$ vectors $\tilde{\boldsymbol{\gamma}}_k(i), i = 1, \dots, N$. Note also that no matrix computations are involved during the backward-time run – the smoothed estimates are obtained by means of backward-time filtering of the scaled EWLS estimates.

Table 2
Simplified smoothing algorithm.

Forward-time run:	
$\hat{\boldsymbol{\theta}}_k(0) = \mathbf{0}$, $\tilde{\eta}_k(0) = 0$, $\tilde{\mathbf{P}}_k(0) = c\mathbf{I}$, $c \gg 1$	
$\boldsymbol{\theta}_k(t) = \hat{\boldsymbol{\theta}}_k(t-1) + \mathbf{h}_k(t)e_k(t)$	
$e_k(t) = y(t) - \boldsymbol{\varphi}^T(t)\hat{\boldsymbol{\theta}}_k(t-1)$	
$\mathbf{h}_k(t) = \frac{\tilde{\mathbf{P}}_k(t-1)\boldsymbol{\varphi}(t)}{\lambda_k + \boldsymbol{\varphi}^T(t)\tilde{\mathbf{P}}_k(t-1)\boldsymbol{\varphi}(t)}$	
$\tilde{\mathbf{P}}_k(t) = [\mathbf{I} - \mathbf{h}_k(t)\boldsymbol{\varphi}^T(t)]\tilde{\mathbf{P}}_k(t-1)/\lambda_k$	
$\tilde{\eta}_k(t) = \lambda_k\tilde{\eta}_k(t-1) + 1$	
$\tilde{\mathbf{Y}}_k(t) = \tilde{\eta}_k(t)\hat{\boldsymbol{\theta}}_k(t)$	
$t = 1, \dots, N$	
Backward-time run:	
$[\hat{\boldsymbol{\theta}}_k(N) = \tilde{\boldsymbol{\theta}}_k(N), \eta_k(N) = \tilde{\eta}_k(N)]$	
$\eta_k(t) = \lambda_k\eta_k(t+1) + (1 - \lambda_k^2)\tilde{\eta}_k(t)$	
$\hat{\mathbf{Y}}_k(t) = \lambda_k\hat{\mathbf{Y}}_k(t+1) + (1 - \lambda_k^2)\tilde{\mathbf{Y}}_k(t)$	
$\boldsymbol{\theta}_k(t) = \hat{\mathbf{Y}}_k(t)/\eta_k(t)$	
$t = N-1, \dots, 1$	

3.4. Evaluation of credibility coefficients

In order to compute credibility coefficients, one should evaluate the quantities $q_k(t)$, $k = 1, \dots, K$. While this is straightforward when the exact algorithm is used (since the matrix $\mathbf{R}_k^{-1}(t)$ is directly available), in the case of the simplified algorithm the situation is not clear as neither the matrix $\mathbf{R}_k^{-1}(t)$ nor the matrix $\mathbf{R}_k(t)$ is updated by this algorithm.

To work out an approximation of $q_k(t)$, we will again use the relationships $\mathbf{R}_k(t) \cong \eta_k(t)\boldsymbol{\Phi}_0$ and $\mathbf{R}_k(t) \cong \tilde{\eta}_k(t)\boldsymbol{\Phi}_0$ which hold under the assumption (A1). Using these relationships, one obtains $\mathbf{R}_k(t) \cong \tilde{\eta}_k(t)\tilde{\mathbf{R}}_k(t)/\tilde{\eta}_k(t)$, and consequently $\hat{q}_k(t) = [\tilde{\eta}_k(t)/\eta_k(t)]\boldsymbol{\varphi}^T(t)\tilde{\mathbf{P}}_k(t)\boldsymbol{\varphi}(t)$.

4. Simulation results

The simulated two-tap FIR system, inspired by the channel estimation applications, was governed by

$$y(t) = \theta_1(t)u(t-1) + \theta_2(t)u(t-2) + v(t) \quad (11)$$

where $u(t) = \pm 1$, $\sigma_u^2 = 1$, denotes the pseudo-random binary signal (PRBS) – the sequence transmitted over a telecommunication channel – and $v(t)$ denotes a zero-mean white noise. Note that the assumptions (A1)–(A4) are straightforward for this kind of application.

Two variants of parameter changes were considered: (A) discontinuous (piecewise-constant), and (B) continuous (chirp-like); see Fig. 1.

The forgetting constants λ_k were chosen so as to make the steady state equivalent width of the one-sided exponential window $\zeta_k = (1 + \lambda_k)/(1 - \lambda_k)$ and double exponential window $\zeta_k = (1 + \lambda_k)^3 / [(1 - \lambda_k)(1 + \lambda_k^2)]$ identical. The measurement noise was Gaussian ($\beta = 2$): $v(t) \sim \mathcal{N}(0, \sigma_v^2)$.

Table 3 presents comparison of the steady-state accumulated mean-squared parameter estimation errors $J_\theta = E_v\{\sum_t \|\hat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}(t)\|^2\}$ obtained for different variants of EWLS trackers and EWLS smoothers. Cooperative tracker (CT) was obtained using the formulas (5)–(6) after replacing the matching errors with the one-step-ahead prediction errors – see Niedźwiecki (1992) for more details. The equivalent widths of the competing EWLS trackers/smoothers were set equal to $\zeta_1 = 13$, $\zeta_2 = 37$ and $\zeta_3 = 109$. The width of the evaluation frame was equal to $M = 21$. To eliminate transient effects, the summation in J_θ was restricted to the interval [101, 4900]. Ensemble averaging $E_v(\cdot)$ was performed over 100 realizations (always the same) of the measurement noise $\{v(t)\}$.

The advantages of smoothing are clear after comparing results presented in Table 3. Note also that both cooperative smoothers –

Table 3

Comparison of parameter estimation errors obtained for 3 EWLS trackers (T_1, T_2, T_3), 3 EWLS smoothers (S_1, S_2, S_3), the cooperative tracker (CT) and two variants of cooperative smoothers – exact (CS) and approximate (CS*). Simulations were performed for 2 types of parameter changes (A, B), and several noise intensities (Gaussian noise).

P	σ_v	T_1	T_2	T_3	CT	S_1	S_2	S_3	CS	CS*
A	0.05	61.22	192.94	559.20	60.45	18.39	51.23	159.12	17.63	15.08
	0.10	65.54	194.30	559.49	62.30	22.80	52.81	159.62	19.59	17.48
	0.15	72.72	196.76	560.13	65.31	30.12	55.24	160.16	22.76	21.36
	0.20	83.17	200.48	561.61	69.62	40.34	58.98	161.71	26.84	26.54
	0.25	96.21	205.01	562.95	74.66	53.75	63.67	163.42	32.43	33.44
B	0.30	112.47	210.30	564.61	81.10	70.01	69.34	164.98	38.96	41.49
	0.05	14.86	103.19	567.83	14.85	1.04	1.44	36.94	0.76	0.87
	0.10	17.36	103.95	567.91	17.20	3.61	2.32	37.10	2.14	2.45
	0.15	21.54	105.29	568.03	20.86	7.82	3.82	37.92	4.07	4.82
	0.20	27.76	107.95	569.73	26.16	13.84	5.87	38.38	6.52	7.96
0.25	35.32	110.23	569.90	32.21	21.64	8.70	39.37	9.67	11.93	
0.30	44.99	114.21	572.31	39.56	31.05	11.94	40.47	13.11	16.39	

the exact algorithm (CS) and its simplified version (CS*) – yield either better results (for piecewise-constant parameter trajectories), or only slightly worse results (for continuous parameter trajectories) than the best constituent smoothers (S_1, S_2, S_3).

5. Conclusion

The problem of noncausal identification of linear stochastic systems was considered and a new parameter smoothing procedure, incorporating a bank of weighted least squares estimation algorithms, was described. At each time step the estimates yielded by the competing algorithms are combined using the recently proposed rules of cooperative smoothing. The resulting parallel estimation scheme automatically adjusts its smoothing bandwidth to the unknown, and possibly time-varying, rate of nonstationarity of the identified system. It can also account for the distribution of measurement noise.

Appendix. (Optimization of the window shape)

Consider the WLS smoother with an infinite past/future data support. Such a steady state algorithm can be rewritten in the following normalized form

$$\hat{\boldsymbol{\theta}}(t) = \mathbf{R}^{-1}(t)\mathbf{r}(t)$$

where $\mathbf{R}(t) = \sum_{i=-\infty}^{\infty} \tilde{w}(i)\boldsymbol{\varphi}(t-i)\boldsymbol{\varphi}^T(t-i)$, $\mathbf{r}(t) = \sum_{i=-\infty}^{\infty} \tilde{w}(i)y(t-i)\boldsymbol{\varphi}(t-i)$ and $\tilde{w}(i) = w(i)/\sum_{i=-\infty}^{\infty} w(i)$. Note that $\sum_{i=-\infty}^{\infty} \tilde{w}(i) = 1$ and $\sum_{i=-\infty}^{\infty} \tilde{w}^2(i) = 1/\zeta$, where ζ denotes the equivalent window width; cf. (4). Given these constraints, we will look for the optimal window shape for a class of parameter trajectories described by the RWJ model (A4), under assumptions (A1)–(A3).

Denote by $\bar{\boldsymbol{\theta}}(t) = E[\hat{\boldsymbol{\theta}}(t)|\Theta]$ the mean path of parameter estimates for a given (true) parameter trajectory $\Theta = \{\boldsymbol{\theta}(i), -\infty < i < \infty\}$. The mean-squared parameter tracking error can be decomposed into the variance and bias components, respectively: $\sigma(t) = E[\|\hat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}(t)\|^2] = \sigma_v(t) + \sigma_b(t)$, where averaging extends to all realizations of $\mathcal{Z} = \{\mathbf{z}(i), -\infty < i < \infty\}$, $\boldsymbol{\varphi} = \{\boldsymbol{\varphi}(i), -\infty < i < \infty\}$ and $\mathcal{V} = \{v(i), -\infty < i < \infty\}$.

Under (A1)–(A3) it holds that (Niedźwiecki, 2000) $\bar{\boldsymbol{\theta}}(t) \cong \sum_{i=-\infty}^{\infty} \tilde{w}(i)\boldsymbol{\theta}(t-i)$, i.e., the mean path of parameter estimates $\{\bar{\boldsymbol{\theta}}(t)\}$ can be approximately viewed as an output of a linear noncausal filter with the impulse response $\{\tilde{w}(t)\}$, excited by the process $\{\boldsymbol{\theta}(t)\}$. Using this result, the bias component of the mean-squared parameter estimation error can be expressed in the form

$$\sigma_b(t) = E[\|\bar{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}(t)\|^2] \cong \text{tr} \left\{ \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \tilde{w}(i)\tilde{w}(j)E[\Delta\boldsymbol{\theta}(t, i)\Delta\boldsymbol{\theta}(t, j)] \right\} \quad (12)$$

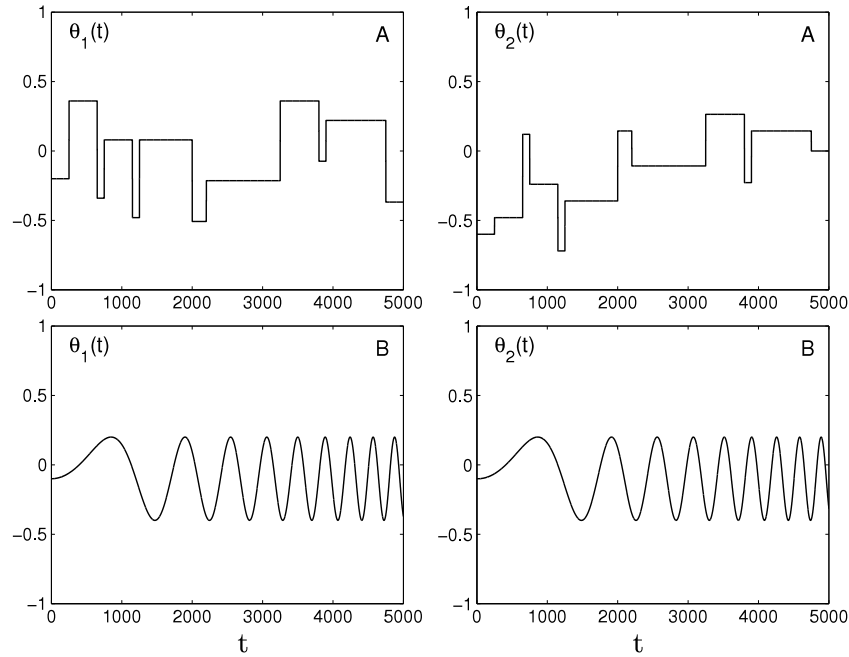


Fig. 1. Two variants of parameter changes used in computer simulations: discontinuous (two upper plots) and continuous (two lower plots).

where $\Delta\theta(t, i) = \theta(t) - \theta(t - i)$. Under (A4) it holds that $E[\Delta\theta(t, i)\Delta\theta(t, j)] = \{\mathbf{Z}f(i, j) \text{ if } i, j \geq 0, \mathbf{Z}f(|i|, |j|) \text{ if } i, j < 0, 0 \text{ elsewhere}\}$, where $f(i, j) = \min(i, j)$ and $\mathbf{Z} = \mathbf{Z}_1 + p_0\mathbf{Z}_2$.

Since the window is assumed to be symmetric, one arrives at

$$\sigma_b(t) = 2 \operatorname{tr}\{\mathbf{Z}\} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \tilde{w}(i)\tilde{w}(j)f(i, j). \quad (13)$$

Analyzing the variance component of the mean-squared parameter estimation error, one arrives at $\sigma_v(t) = E[\|\hat{\theta}(t) - \theta(t)\|^2] = \operatorname{tr}\{\mathbf{S}\} + \sigma_n(t)$ where $\mathbf{S} = \sigma_v^2\Phi_0^{-1}/\zeta$ is the covariance matrix that describes statistical variability of parameter estimates under stationary conditions [cf. (4)] and $\sigma_n(t) > 0$ denotes the extra variability term that can be attributed to parameter changes (for a time-invariant system $\sigma_n(t) = 0$). Similar to Niedźwiecki (1986), when the autocovariance function of the input process $\{\varphi(t)\}$ can be upper-bounded by the exponentially decaying sequence, one can show that the term $\sigma_n(t)$ is dominated by $\sigma_b(t)$. Therefore, to minimize $\sigma(t)$ for a given value of ζ , one should find such a window shape that minimizes (13). Our optimization problem can be stated as follows: find $\{\tilde{w}(i)\}$ such that

$$\sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \tilde{w}(i)\tilde{w}(j)f(i, j) \longrightarrow \min \quad (14)$$

subject to the following constraints: $\sum_{i=-\infty}^{\infty} \tilde{w}(i) = 1$, $\sum_{i=-\infty}^{\infty} \tilde{w}^2(i) = 1/\zeta$, and $\tilde{w}(i) = \tilde{w}(-i) \geq 0$, $\forall i$.

Denote by $w_g(s)$ the widow-generating function of a real argument $s \in [0, \infty)$, obeying the condition $w(i) = w_g(i)$, $i \geq 0$, and let $\tilde{w}_g(s) = w_g(s)/\int_0^{\infty} w_g(s)ds$. For sufficiently large values of ζ , the discrete-time problem (14) can be solved approximately by solving the corresponding continuous-time problem: find $\{\tilde{w}_g(s)\}$ such that

$$\int_0^{\infty} \int_0^{\infty} \tilde{w}_g(s_1)\tilde{w}_g(s_2)f(s_1, s_2)ds_1ds_2 \longrightarrow \min \quad (15)$$

under the constraints: $\int_0^{\infty} \tilde{w}_g(s)ds = 0.5$, $\int_0^{\infty} \tilde{w}_g^2(s)ds = 1/(2\zeta)$, and $\tilde{w}_g(s) \geq 0$, $\forall s \geq 0$.

Note that the double integral in (15) can be rewritten as $2 \int_0^{\infty} s\tilde{w}_g(s)[1 - \int_0^s \tilde{w}_g(\tau)d\tau]ds$ and let $x(s) = \int_0^s \tilde{w}_g(\tau)d\tau$. Using these expressions, one can convert (15) into the so-called isoperimetric problem of the calculus of variations (Gelfand & Fomin, 2000): find $x(s)$ such that:

$$\int_0^{\infty} s\dot{x}(s)[1 - x(s)]ds \longrightarrow \min \quad (16)$$

under the constraints: $x(0) = 0$, $x(\infty) = 0.5$, $\int_0^{\infty} [\dot{x}(s)]^2 ds = 1/(2\zeta)$, and $\dot{x}(s) \geq 0$, $\forall s \geq 0$.

The basis function for (16) takes the form $H(s, x, \dot{x}) = s\dot{x} - s\dot{x}x + c\dot{x}^2$, where c denotes a constant parameter. To find the optimal solution, one has to solve the Euler–Lagrange equation

$$\frac{\partial H}{\partial x} - \frac{d}{ds} \left(\frac{\partial H}{\partial \dot{x}} \right) = 2c\ddot{x} - x + 1 = 0. \quad (17)$$

The only admissible solution of (17) has the form $x_{\text{opt}}(s) = [1 - e^{-\alpha s}]/2$, $\alpha = 4/\zeta$, leading to $\tilde{w}_g^{\text{opt}}(s) = \alpha e^{-\alpha s}/2$, $s \geq 0$. This means that the optimal steady-state window for the considered class of parameter changes is double exponential, i.e., it has the form $w_{\text{opt}}(i) = \lambda^{|i|}$, $\forall i$, where $\lambda = e^{-\alpha} < 1$.

References

- Falconer, D. D., & Gitlin, R. D. (1978). Optimum reception of digital data signals in the presence of timing-phase hits. *Bell System Technical Journal*, 57, 3181–3208.
- Friedl, H., & Stampfer, E. (2002). Cross-validation. In A. H. El-Shaarawi, & W. W. Piegorsch (Eds.), *Encyclopedia of environmetrics: Vol. 1* (pp. 452–460). New York: Wiley.
- Gelfand, I. M., & Fomin, S. V. (2000). *Calculus of variations*. Dover Publ.
- Niedźwiecki, M. (1986). Optimization of the window shape in weighted least squares identification of a class of nonstationary systems. In *Proc. 7th int. conf. on anal. and opt. of syst.* (pp. 889–901). Antibes, France: Springer-Verlag.
- Niedźwiecki, M. (1992). Identification of nonstationary stochastic systems using parallel estimation schemes. *IEEE Transactions on Automatic Control*, 35, 329–334.
- Niedźwiecki, M. (2000). *Identification of time-varying processes*. New York: Wiley.
- Niedźwiecki, M. (2010). Easy recipes for cooperative smoothing. *Automatica*, 46, 716–720.



Maciej Niedźwiecki was born in Poznań, Poland in 1953. He received the M.Sc. and Ph.D. degrees from the Gdańsk University of Technology, 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. During 1990–1993, he served as a Vice Chairman of Technical Committee on Theory of the International Federation of Automatic Control (IFAC).

He is currently Associate Editor for IEEE Transactions on Signal Processing, 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 is the author of the book *Identification of Time-varying Processes* (Wiley, 2000).

He works as a Professor and Head of the Department of Automatic Control, Faculty of Electronics, Telecommunications and Computer Science, Gdańsk University of Technology. His main areas of research interests include system identification, statistical signal processing and adaptive systems.



Szymon Gackowski received the M.Sc. degree in automatic control from the Gdańsk University of Technology, Gdańsk, Poland, in 2008. Since 2009 he is a Ph.D. student in the Department of Automatic Control, Faculty of Electronics, Telecommunications and Computer Science, Gdańsk University of Technology. His professional interests include signal processing with emphasis on image processing.