Postprint of: Niedzwiecki M., Locally Adaptive Cooperative Kalman Smoothing and Its Application to Identification of Nonstationary Stochastic Systems, IEEE Transactions on Signal Processing, Vol. 60, Iss. 1 (2012), pp. 48-59, DOI: 10.1109/TSP.2011.2172432

Locally-adaptive Cooperative Kalman Smoothing and Its Application to Identification of Nonstationary Stochastic Systems

Maciej Niedźwiecki, *Member, IEEE*

***Abstract*—One of the central problems of the stochastic approximation theory is the proper adjustment of the smoothing algorithm to the unknown, and possibly time-varying, rate and mode of variation of the estimated signals/parameters. In this paper we propose a novel locally adaptive parallel estimation scheme which can be used to solve the problem of fixed-interval Kalman smoothing in the presence of model uncertainty. The proposed solution is based on the idea of cooperative smoothing – the Bayesian extension of the leave-one-out cross-validation approach to model selection. Within this approach the smoothed estimates are evaluated as a convex combination of the estimates provided by several competing smoothers. We derive computationally attractive algorithms allowing for cooperative Kalman smoothing and show how the proposed approach can be applied to identification of nonstationary stochastic systems.**

***Index Terms*—Parallel estimation schemes, system identification, Kalman smoothing.**

I. Introduction

In this paper we will consider the problem of noncausal estimation (smoothing) of an n-dimensional signal $\mathbf{s}(t) = [s_1(t), \ldots, s_n(t)]^{\mathrm{T}}$, based on noisy measurements

$$\mathbf{y}(t) = \mathbf{s}(t) + \mathbf{v}(t), \quad t = \ldots, -1, 0, 1, \ldots \tag{1}$$

where t denotes normalized discrete time and $\{\mathbf{v}(t)\}$ denotes n-dimensional white Gaussian noise. We will assume that the entire measurement history $\mathbf{Y}(N) = \{\mathbf{y}(1), \ldots, \mathbf{y}(N)\}$ is available, so at any instant $t \in [1, N]$ estimation of $\mathbf{s}(t)$ can be based on "past" ($i < t$), "present" ($i = t$) and "future" ($i > t$) measurements. We will also assume that the signal $\mathbf{s}(t)$ is generated by a state space model

$$\begin{aligned} \mathbf{x}(t+1) &= \mathbf{F}(t, \boldsymbol{\theta})\mathbf{x}(t) + \mathbf{G}(t, \boldsymbol{\theta})\mathbf{w}(t) \\ \mathbf{s}(t) &= \mathbf{H}(t, \boldsymbol{\theta})\mathbf{x}(t) \end{aligned} \tag{2}$$

where $\mathbf{x}(t)$ is a r-dimensional state vector, and $\{\mathbf{w}(t)\}$ denotes the l-dimensional driving noise – a white Gaussian sequence independent of $\{\mathbf{v}(t)\}$.

Finally, we will assume that the matrices $\mathbf{F}(t, \boldsymbol{\theta})_{r\times r}$, $\mathbf{G}(t, \boldsymbol{\theta})_{r\times l}$, $\mathbf{H}(t, \boldsymbol{\theta})_{n\times r}$ that appear in the state space description (2), as well as the covariance matrices of the noise sources $\mathrm{cov}[\mathbf{v}(t)] = \mathbf{V}(t, \boldsymbol{\theta})_{n\times n}$, $\mathrm{cov}[\mathbf{w}(t)] = \mathbf{W}(t, \boldsymbol{\theta})_{l\times l}$ are functions of a vector of unknown design parameters $\boldsymbol{\theta}$.

The structured information about the estimated signal, encapsulated in the state space model (2), is available in many passive (e.g. GPS - Global Positioning System) and active (e.g. radar-based) localization and navigation applications – see detailed studies in [1] and [2], among many others.

The time-varying nature of the matrices $\mathbf{F}$, $\mathbf{G}$, $\mathbf{H}$, $\mathbf{V}$ and $\mathbf{W}$ is often a consequence of the fact that the model (1) - (2) is a result of discretization, change of spatial coordinates and linearization around a time-dependent "setpoint", of a nonlinear continuous-time signal description

$$\begin{aligned} \dot{\mathbf{x}}_c &= \mathbf{f}[\mathbf{x}_c] + \mathbf{w}_c \\ \mathbf{y}_c &= \mathbf{h}[\mathbf{x}_c] + \mathbf{v}_c. \end{aligned} \tag{3}$$

The model uncertainty, represented by the vector $\boldsymbol{\theta}$, usually originates from an incomplete knowledge of the signal source or unknown/changing environmental conditions.

When the vector $\boldsymbol{\theta}$ is known, the optimal, in the mean-squared sense, smoothed estimate of $\mathbf{s}(t)$ is given in the form [3]

$$\widehat{\mathbf{s}}(t|N) = \mathrm{E}[\mathbf{s}(t)|\mathbf{Y}(N)]$$

and can be evaluated using the algorithm known as a fixed-interval Kalman smoother. Using the intuitively appealing framework proposed by Mayne [4] and Fraser [5], smoothing can be viewed as a result of combining the estimates yielded by two Kalman filters/predictors: the causal one, running forward in time, and the anticausal one [based on the backwards Markovian representation of $\mathbf{s}(t)$], running backward in time. When design parameters $\boldsymbol{\theta}$ are not known, the problem can be solved in two different ways. The first approach is based on sequential estimation of $\boldsymbol{\theta}$ and results in algorithms known as adaptive Kalman filters/smoothers [6]. In the second approach, several Kalman filters/smoothers, designed for different hypothetical values of $\boldsymbol{\theta}$, are run in parallel and the obtained results are merged in a statistically meaningful way. Such parallel estimation, or multiple-model, schemes, which can be traced back to Magill [7], are increasingly popular in modern navigation and tracking applications – see e.g. [8] and [9].

The approach that proved particularly useful in target tracking applications is called interactive multiple models (IMM). The IMM algorithm, originally proposed in [10], was designed for systems that can switch between different modes of behavior, characterized by different state space descriptions. The mode switching process is modelled as a Markov chain.

This work was supported in part by Foundation for Polish Science. The author is with the Faculty of Electronics, Telecommunications and Computer Science, Department of Automatic Control, Gdańsk University of Technology, ul. Narutowicza 11/12, Gdańsk, Poland (e-mail: maciekn@eti.pg.gda.pl).

The algorithm consists of a bank of Kalman filters (corresponding to different maneuvering hypotheses) and a linear combiner which computes the final state estimate as a convex combination of the component estimates. The term "interacting models" refers to the fact that, at the beginning of each cycle of model-conditioned Kalman filtering, the state vectors and error covariance matrices of component filters are mixed appropriately. The mixing step is equivalent to "hypothesis merging", and makes the filter bank behave more consistently compared to the no-mixing case. Over the past two decades the IMM approach has been perfected and extended in many important directions, such as handling the data origin uncertainty [the probabilistic data association (PDA) technique], for example. See e.g. [11] for a recent survey of the available solutions. The fixed-interval IMM smoothing algorithms can be found in [12] and [13].

In this paper we propose a novel locally adaptive parallel estimation scheme which can be used to solve the problem of Kalman smoothing in the presence of model uncertainty. The proposed solution is based on the idea of cooperative smoothing presented in [14]. Cooperative smoothing is a general framework which shows how credibility of constituent smoothers can be evaluated based on the errors ("matching errors") yielded by the appropriately modified ("holey") smoothers. Since this approach is based on the local statistics, it can deal with different forms of signal/system nonstationarity.

The structure of the cooperative smoother is the same as the structure of the IMM smoother, except that the hypothesis merging step is not employed. The main advantage of the new approach, besides very low computational complexity, is its universal character – unlike IMM schemes, cooperative smoothers can be used to combine, in a statistically meaningful way, results yielded by *any* smoothing algorithms, not necessarily those based on the Kalman theory. Hence, from the qualitative viewpoint, the cooperative approach complements the IMM approach, rather than competes with it.

Cooperative smoothing can be particularly useful in all applications, such as identification of nonstationary stochastic systems, where our prior knowledge about different modes of system variation is rather vague (if any), i.e., the incorporated state space models are to a greater extent instrumental than factual (physically motivated) [15]. In cases like this, one may be interested in combining results provided by smoothing algorithms based on different smoothing principles, e.g. combining Kalman smoothers with kernel smoothers, such as the algorithms proposed recently in [16]. Using the results presented in [16] and in this paper, one can easily design such "mixed" parallel estimation schemes.

It should be stressed that the general rules presented in [14] are usually not cost effective, i.e., for a particular class of combined smoothing algorithms they may be rather cumbersome to apply. Further work is usually needed to turn a general idea into an efficient computational algorithm. Such is the case with Kalman smoothers.

The contribution of this paper is threefold. First, we show how the idea of cooperative smoothing, originally proposed for univariate signals, can be extended to multivariate signals corrupted by Gaussian noise. Second, we show that matching errors, needed to compute credibility coefficients, can be expressed in terms of (easily computable) residual errors, i.e., that they can be evaluated without actually implementing the corresponding holey smoothers. This significantly simplifies the cooperative Kalman scheme, making it computationally attractive. Finally, we show how cooperative Kalman smoothing can be applied to identification of nonstationary stochastic systems.

II. Cooperative smoothing of multivariate signals

Similar to [14], we will start from considering a simplified smoothing scheme referred to as Bayesian pattern matching. The obtained results will be next applied, in an appropriately modified form, to solve more realistic smoothing problems.

A. Bayesian pattern matching

Denote by $T_{\rm a}(t)=[t-m,t+m]$ the local analysis window[1] of width $M=2m+1$, $M\geq n$, and let $\{\mathbf{s}_1(i),\ldots,\mathbf{s}_K(i), i\in T_{\rm a}(t)\}$ be a set of fixed (data-independent) signal patterns or waveforms. We will consider K equiprobable hypotheses $\mathcal{H}_k$, $\pi(\mathcal{H}_k)=1/K$, $k=1,\ldots,K$, of the form

$$\mathcal{H}_k:\ \mathbf{s}(i)=\mathbf{s}_k(i),\ \mathbf{v}(i)\sim\mathcal{N}(\mathbf{0},\mathbf{V}),\ \mathbf{V}>\mathbf{O},\quad i\in T_{\rm a}(t)$$

where $\mathbf{0}$ denotes the n-dimensional vector of zeros and $\mathbf{O}$ is the $n\times n$ null matrix.

According to the hypothesis $\mathcal{H}_k$, the signal coincides, within the analysis window $T_{\rm a}(t)$, with the k-th pattern. The unknown covariance matrix $\mathbf{V}$ of the measurement noise, assumed constant over $T_{\rm a}(t)$, will be regarded as a nuisance parameter with an assigned noninformative prior distribution. Using the Jeffreys formula [17], the noninformative distribution for $\mathbf{V}$ can be obtained in the following improper form [18]

$$\pi(\mathbf{V}|\mathcal{H}_k)=\pi(\mathbf{V})\propto[\det(\mathbf{V})]^{-(n+1)/2} \tag{4}$$

where $\propto$ denotes proportionality.

Assuming that one and only one of the hypotheses is true, the optimal, in the mean-squared sense, estimate of $\mathbf{s}(t)$ can be obtained in the form [3]

$$\widehat{\mathbf{s}}(t)=\sum_{k=1}^{K}\mu_k(t)\mathbf{s}_k(t) \tag{5}$$

where $\mu_k(t), k=1,\ldots,K$, denote credibility coefficients – the posterior probabilities of different signal patterns given the set of local observations $\mathbf{Y}_T(t)=\{\mathbf{y}(t-m),\ldots,\mathbf{y}(t+m)\}$:

$$\begin{aligned}\mu_k(t)&=P(\mathcal{H}_k|\mathbf{Y}_T(t))\\&\propto\int_{\mathbf{V}>\mathbf{O}}p(\mathbf{Y}_T(t)|\mathbf{V},\mathcal{H}_k)\pi(\mathbf{V}|\mathcal{H}_k)\pi(\mathcal{H}_k)d\mathbf{V}.\end{aligned} \tag{6}$$

Note that

$$\begin{aligned}p(\mathbf{Y}_T(t)|\mathbf{V},\mathcal{H}_k)&=[\det(2\pi\mathbf{V})]^{-M/2}\\&\quad\times\exp\left\{-\frac{1}{2}\,\mathrm{tr}\left[\mathbf{V}^{-1}\mathbf{D}_k(t)\right]\right\}\end{aligned} \tag{7}$$

[1] We will assume, for simplicity, that $m<t\leq N-m$.

where

$$\mathbf{D}_k(t) = \sum_{i \in T_{\mathrm{a}}(t)} \mathbf{e}_k(i)\mathbf{e}_k^{\mathrm{T}}(i)$$

and $\mathbf{e}_k(i) = \mathbf{y}(i) - \mathbf{s}_k(i)$ denotes residual error. Therefore, under noninformative priors, one obtains

$$\begin{aligned}\mu_k(t) &\propto \int_{\mathbf{V}>\mathbf{O}} [\det(\mathbf{V})]^{-(M+n+1)/2} \\ &\quad \times \exp\left\{-\frac{1}{2}\,\mathrm{tr}\left[\mathbf{V}^{-1}\mathbf{D}_k(t)\right]\right\} d\mathbf{V} \\ &= 2^{\frac{nM}{2}}\,\Gamma_n\left(\tfrac{M}{2}\right)\{\,\det[\mathbf{D}_k(t)]\,\}^{-\frac{M}{2}} \end{aligned} \tag{8}$$

where $\Gamma_n(\cdot)$ denotes the multivariate gamma function. This result follows immediately from the well-known properties of the inverse Wishart distribution – see Appendix 1.

Using (8), credibility coefficients can be expressed in the following form

$$\mu_k(t) = \frac{\eta_k(t)}{\sum_{k=1}^{K}\eta_k(t)} \tag{9}$$

where

$$\eta_k(t) = \{\,\det[\mathbf{D}_k(t)]\,\}^{-M/2}. \tag{10}$$

B. Cooperative smoothing

Consider now the situation where, instead of fixed patterns, one uses data-dependent patterns locally adapted to the signal and given by

$$\widehat{\mathbf{s}}_k(t) = f_k[\mathbf{Y}(N)] \tag{11}$$

where $f_k[\cdot]$ is an arbitrary smoothing procedure. Depending on the smoothing approach, the function $f_k[\cdot]$ may take many different forms, both linear (Kalman smoother, kernel smoother [19], local polynomial approximation smoother [20]) and nonlinear (order statistical smoothers [21]).

Following [14], the cooperative smoother based on $\widehat{\mathbf{s}}_1(t), \ldots, \widehat{\mathbf{s}}_K(t)$ will be defined in the following form

$$\widehat{\mathbf{s}}(t) = \sum_{k=1}^{K} \mu_k^{\circ}(t)\widehat{\mathbf{s}}_k(t) \tag{12}$$

where

$$\begin{aligned}\mu_k^{\circ}(t) &= \frac{\eta_k^{\circ}(t)}{\sum_{k=1}^{K}\eta_k^{\circ}(t)} \\ \eta_k^{\circ}(t) &= \{\,\det[\mathbf{D}_k^{\circ}(t)]\,\}^{-M/2} \\ \mathbf{D}_k^{\circ}(t) &= \sum_{i \in T_{\mathrm{a}}(t)} \mathbf{e}_k^{\circ}(i)[\mathbf{e}_k^{\circ}(i)]^{\mathrm{T}} \end{aligned} \tag{13}$$

and

$$\mathbf{e}_k^{\circ}(t) = \mathbf{y}(t) - \widehat{\mathbf{s}}_k^{\circ}(t)$$

denotes the quantity which will be further referred to as matching error[2] – residual error yielded by the holey smoother $\widehat{\mathbf{s}}_k^{\circ}(t)$ associated with $\widehat{\mathbf{s}}_k(t)$

$$\widehat{\mathbf{s}}_k^{\circ}(t) = f_k[\mathbf{Y}^{\circ}(t,N)],\ \ \mathbf{Y}^{\circ}(t,N) = \mathbf{Y}(N) - \{\mathbf{y}(t)\}. \tag{14}$$

[2]In the classical regression analysis *normalized* matching errors are called deleted residuals.

According to (14), holey smoother $\widehat{\mathbf{s}}_k^{\circ}(t)$ is based on the same smoothing principle as $\widehat{\mathbf{s}}_k(t)$, except that it excludes $\mathbf{y}(t)$ from the set of measurements used for estimation of $\mathbf{s}(t)$. Since matching errors $\mathbf{e}_k^{\circ}(t)$ are pointwise independent of the measurement noise $\mathbf{v}(t)$, they allow for approximately unbiased evaluation of the local performance of $\widehat{\mathbf{s}}_k(t)$. In particular, the modified Bayesian-like combination rule, obtained when credibility coefficients are evaluated for matching errors, will not favor smoothers that "underestimate" the influence of measurement noise on the observed data.

Cooperative smoothing can be regarded as a Bayesian extension of the leave-one-out cross-validation approach to model selection [22].

On the qualitative level, cooperative smoothing closely resembles the prediction technique, known as Bayesian model averaging (BMA) – for more details see e.g. [23]. BMA is an intuitively attractive solution to the problem of accounting for model uncertainty – the final prediction is obtained by averaging predictions based on many competing models. It can be shown that averaging over all the models provides better average predictive ability, as measured by a logarithmic scoring rule, than using any single model [24]. The multiple-model approach to prediction, which can be traced back to the paper of Bates and Granger [25], has been successfully utilized in many fields, such as econometrics, environmental science and biology.

C. Computational hints

Some of the quantities involved in computation of credibility coefficients may take very large or very small values. The following modified expression, mathematically equivalent to (13), allows one to avoid numerical problems (such as numerical overflow) caused by improper scaling

$$\mu_k^{\circ}(t) = \frac{\exp\{\chi_k(t)\}}{\sum_{k=1}^{K}\exp\{\chi_k(t)\}}$$

where

$$\begin{aligned}\chi_k(t) &= \zeta_k(t) - \zeta_{\max}(t) \\ \zeta_k(t) &= \log\eta_k^{\circ}(t) = -\,(M/2)\log\{\,\det[\mathbf{D}_k^{\circ}(t)]\,\} \\ \zeta_{\max}(t) &= \max_{1\le k\le K}\zeta_k(t)\,. \end{aligned}$$

III. Cooperative Kalman smoothing

Suppose that the vector $\boldsymbol{\theta}$ in (2) can take K different values $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$, leading to K hypothetical models of the analyzed signal

$$\begin{aligned}\mathcal{M}_k:\quad \mathbf{x}(t+1) &= \mathbf{F}_k(t)\mathbf{x}(t) + \mathbf{G}_k(t)\mathbf{w}_k(t) \\ \mathbf{s}(t) &= \mathbf{H}_k(t)\mathbf{x}(t) \\ \mathbf{y}(t) &= \mathbf{s}(t) + \mathbf{v}_k(t) \end{aligned} \tag{15}$$

$$\mathbf{w}_k(t) \sim \mathcal{N}(\mathbf{0}, \mathbf{W}_k(t)),\quad \mathbf{v}_k(t) \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_k(t))$$

where $\mathbf{F}_k(t) = \mathbf{F}(t, \boldsymbol{\theta}_k)$, $\mathbf{G}_k(t) = \mathbf{G}(t, \boldsymbol{\theta}_k)$, etc.

A. Two-filter smoothing formula

For a given signal model $\mathcal{M}_k$, the optimal noncausal estimate of $\mathbf{x}(t)$ is given by

$$\widehat{\mathbf{x}}_k(t|N) = \mathrm{E}[\mathbf{x}(t)|\mathbf{Y}(N), \mathcal{M}_k]$$

and can be written down as a linear combination of the estimates provided by two Kalman filters/predictors operated forward in time and backward in time, respectively. The forward Kalman filter/predictor has the form

$$\begin{aligned}
\boldsymbol{\varepsilon}_k(t) &= \mathbf{y}(t) - \mathbf{H}_k(t)\widehat{\mathbf{x}}_k(t|t-1) \\
\mathbf{Q}_k(t) &= \mathbf{H}_k(t)\mathbf{P}_k(t|t-1)\mathbf{H}_k^{\mathrm{T}}(t) + \mathbf{V}_k(t) \\
\mathbf{K}_k(t) &= \mathbf{P}_k(t|t-1)\mathbf{H}_k^{\mathrm{T}}(t)\mathbf{Q}_k^{-1}(t) \\
\widehat{\mathbf{x}}_k(t|t) &= \widehat{\mathbf{x}}_k(t|t-1) + \mathbf{K}_k(t)\boldsymbol{\varepsilon}_k(t) \\
\mathbf{P}_k(t|t) &= \mathbf{P}_k(t|t-1) - \mathbf{K}_k(t)\mathbf{Q}_k(t)\mathbf{K}_k^{\mathrm{T}}(t) \\
\widehat{\mathbf{x}}_k(t+1|t) &= \mathbf{F}_k(t)\widehat{\mathbf{x}}_k(t|t) \\
\mathbf{P}_k(t+1|t) &= \mathbf{F}_k(t)\mathbf{P}_k(t|t)\mathbf{F}_k^{\mathrm{T}}(t) + \mathbf{G}_k(t)\mathbf{W}_k(t)\mathbf{G}_k^{\mathrm{T}}(t) \\
t &= 1, \ldots, N
\end{aligned} \tag{16}$$

where

$$\begin{aligned}
\widehat{\mathbf{x}}_k(t|t) &= \mathrm{E}[\mathbf{x}(t)|\mathbf{Y}(t), \mathcal{M}_k] \\
\widehat{\mathbf{x}}_k(t+1|t) &= \mathrm{E}[\mathbf{x}(t+1)|\mathbf{Y}(t), \mathcal{M}_k]
\end{aligned}$$

and $\mathbf{P}_k(t|t) = \mathrm{cov}[\widehat{\mathbf{x}}_k(t|t)]$, $\mathbf{P}_k(t+1|t) = \mathrm{cov}[\widehat{\mathbf{x}}_k(t+1|t)]$ denote the corresponding covariance matrices.

The backward Kalman filter/predictor, which evaluates estimates based on the "future" data samples $\mathbf{Y}^{\mathrm{B}}(t) = \{\mathbf{y}(t), \ldots, \mathbf{y}(N)\}$

$$\begin{aligned}
\widehat{\mathbf{x}}_k^{\mathrm{B}}(t|t) &= \mathrm{E}[\mathbf{x}(t)|\mathbf{Y}^{\mathrm{B}}(t), \mathcal{M}_k^{\mathrm{B}}] \\
\widehat{\mathbf{x}}_k^{\mathrm{B}}(t-1|t) &= \mathrm{E}[\mathbf{x}(t-1)|\mathbf{Y}^{\mathrm{B}}(t), \mathcal{M}_k^{\mathrm{B}}]
\end{aligned}$$

has the same form as the forward algorithm (16), except that it is based on the backwards Markovian signal model

$$\begin{aligned}
\mathcal{M}_k^{\mathrm{B}}: \quad \mathbf{x}(t) &= \mathbf{F}_k^{\mathrm{B}}(t)\mathbf{x}(t+1) + \mathbf{G}_k^{\mathrm{B}}(t)\mathbf{w}_k^{\mathrm{B}}(t+1) \\
\mathbf{y}(t) &= \mathbf{H}_k(t)\mathbf{x}(t) + \mathbf{v}_k(t)
\end{aligned}$$

and is run backward in time.

When $\mathbf{P}_k^{-1}(1|0) = \mathbf{O}$ (noninformative prior) and when the state transition matrix $\mathbf{F}_k(t)$ is invertible for all t, parameters of the backwards Markovian model can be specified as follows:

$$\begin{aligned}
\mathbf{F}_k^{\mathrm{B}}(t) = \mathbf{F}_k^{-1}(t), \;\; \mathbf{G}_k^{\mathrm{B}}(t) &= -\mathbf{F}_k^{-1}(t)\mathbf{G}_k(t) \\
\mathbf{w}_k^{\mathrm{B}}(t+1) &= \mathbf{w}_k(t).
\end{aligned}$$

When $\mathbf{P}_k^{-1}(1|0) \neq \mathbf{O}$, i.e., some prior knowledge about the initial value of the state vector is available, construction of the backwards Markovian model is less straightforward – see [26] for more details. This complication, however, will have no impact on the analysis carried out in this paper.

The Mayne-Fraser (MF) two-filter formula can be summarized as follows

$$\begin{aligned}
\widehat{\mathbf{x}}_k(t|N) = \mathbf{P}_k(t|N) \big\{ &[\mathbf{P}_k(t|t)]^{-1}\widehat{\mathbf{x}}(t|t) \\
&+ [\mathbf{P}_k^{\mathrm{B}}(t|t+1)]^{-1}\widehat{\mathbf{x}}_k^{\mathrm{B}}(t|t+1) \big\}
\end{aligned} \tag{17}$$

$$\begin{aligned}
\mathbf{P}_k(t|N) &= \mathrm{cov}[\widehat{\mathbf{x}}_k(t|N)] \\
&= \big\{ [\mathbf{P}_k(t|t)]^{-1} + [\mathbf{P}_k^{\mathrm{B}}(t|t+1)]^{-1} \big\}^{-1}
\end{aligned} \tag{18}$$

where $\mathbf{P}_k^{\mathrm{B}}(t|t+1)$ denotes the covariance matrix of the backward Kalman "predictor" $\widehat{\mathbf{x}}_k^{\mathrm{B}}(t|t+1)$. The initial conditions for the backward algorithm are $\widehat{\mathbf{x}}_k^{\mathrm{B}}(N|N+1) = \mathbf{0}$ and $[\mathbf{P}_k^{\mathrm{B}}(N|N+1)]^{-1} = \mathbf{O}$. Such an improper covariance initialization can be easily handled when the backward Kalman algorithm is realized as an *information filter* (inverse covariance filter) – for more details see [3].

Based on (17), one obtains the following smoothed estimate of $\mathbf{s}(t)$

$$\begin{aligned}
\widehat{\mathbf{s}}_k(t|N) &= \mathrm{E}[\mathbf{s}(t)|\mathbf{Y}(N), \mathcal{M}_k] = \mathbf{H}_k(t)\widehat{\mathbf{x}}_k(t|N) \\
\mathbf{S}_k(t|N) &= \mathrm{cov}[\widehat{\mathbf{s}}_k(t|N)] = \mathbf{H}_k(t)\mathbf{P}_k(t|N)\mathbf{H}_k^{\mathrm{T}}(t).
\end{aligned} \tag{19}$$

B. Evaluation of matching errors

Local performance of the smoothed estimate $\widehat{\mathbf{s}}_k(t|N)$ can be evaluated in terms of the associated matching errors

$$\mathbf{e}_k^{\circ}(t) = \mathbf{y}(t) - \widehat{\mathbf{s}}_k^{\circ}(t|N)$$

where

$$\widehat{\mathbf{s}}_k^{\circ}(t|N) = \mathrm{E}[\mathbf{s}(t)|\mathbf{Y}^{\circ}(t,N), \mathcal{M}_k] = \mathbf{H}_k(t)\widehat{\mathbf{x}}_k^{\circ}(t|N)$$

and $\widehat{\mathbf{x}}_k^{\circ}(t|N)$ denotes the estimate yielded by the holey Kalman smoother

$$\widehat{\mathbf{x}}_k^{\circ}(t|N) = \mathrm{E}[\mathbf{x}(t)|\mathbf{Y}^{\circ}(t,N), \mathcal{M}_k].$$

Using the generalized Millman's theorem on optimal combination of two independent estimates [3], one can express $\widehat{\mathbf{x}}_k^{\circ}(t|N)$ as a linear combination of forward and backward Kalman predictors [note similarity to (17) - (18)]

$$\begin{aligned}
\widehat{\mathbf{x}}_k^{\circ}(t|N) = \mathbf{P}_k^{\circ}(t|N) \big\{ &[\mathbf{P}_k(t|t-1)]^{-1}\widehat{\mathbf{x}}(t|t-1) \\
&+ [\mathbf{P}_k^{\mathrm{B}}(t|t+1)]^{-1}\widehat{\mathbf{x}}_k^{\mathrm{B}}(t|t+1) \big\}
\end{aligned} \tag{20}$$

$$\begin{aligned}
\mathbf{P}_k^{\circ}(t|N) &= \mathrm{cov}[\widehat{\mathbf{x}}_k^{\circ}(t|N)] \\
&= \big\{ [\mathbf{P}_k(t|t-1)]^{-1} + [\mathbf{P}_k^{\mathrm{B}}(t|t+1)]^{-1} \big\}^{-1}
\end{aligned} \tag{21}$$

Now we are ready to state the main result of this section – we will show that matching errors $\mathbf{e}_k^{\circ}(t)$ can be expressed in terms of residual errors $\mathbf{e}_k(t)$.

Theorem

It holds that

$$\mathbf{e}_k^{\circ}(t) = \boldsymbol{\Sigma}_k^{-1}(t)\mathbf{e}_k(t) \tag{22}$$

where

$$\begin{aligned}
\boldsymbol{\Sigma}_k(t) &= \mathbf{I} - \mathbf{S}_k(t|N)\mathbf{V}_k^{-1}(t) \\
\mathbf{e}_k(t) &= \mathbf{y}(t) - \widehat{\mathbf{s}}_k(t|N).
\end{aligned} \tag{23}$$

Proof – see Appendix 2.

Note that, according to (22), matching errors can be evaluated without actually implementing the corresponding holey smoother.

C. Cooperative smoothing formula

The cooperative smoothing formula, merging results yielded by K Kalman smoothers, can be summarized as follows

$$\widehat{\mathbf{s}}(t|N) = \sum_{k=1}^{K} \mu_k^{\circ}(t)\widehat{\mathbf{s}}_k(t|N) \tag{24}$$

where $\mu_k^{\circ}(t), k = 1, \ldots, K$, denote credibility coefficients given by (13). Matching errors $\mathbf{e}_k^{\circ}(t)$, needed to compute $\mu_k^{\circ}(t)$, can be obtained from (22).

When all state space models share the some output matrix, i.e.,

$$\mathbf{H}_1(t) = \ldots = \mathbf{H}_K(t) = \mathbf{H}(t), \quad \forall t$$

the formula (24) can be rewritten in the form

$$\widehat{\mathbf{s}}(t|N) = \mathbf{H}(t)\widehat{\mathbf{x}}(t|N) \tag{25}$$

where

$$\widehat{\mathbf{x}}(t|N) = \sum_{k=1}^{K} \mu_k^{\circ}(t)\widehat{\mathbf{x}}_k(t|N) \tag{26}$$

denotes the cooperative state estimate.

Since the models $\mathcal{M}_k, k = 1, \ldots, K$ are regarded as mutually exclusive hypothetical signal descriptions, the covariance matrices of smoothed estimates (24) and (26) can be approximately[3] expressed in the form

$$\begin{aligned} \mathbf{S}(t|N) &\cong \sum_{k=1}^{K} \mu_k^{\circ}(t)\mathbf{S}_k(t|N) \\ \mathbf{P}(t|N) &\cong \sum_{k=1}^{K} \mu_k^{\circ}(t)\mathbf{P}_k(t|N). \end{aligned} \tag{27}$$

Remark

All results presented above remain valid if the state equation in (15) includes the deterministic input signal $\mathbf{u}(t)$

$$\mathbf{x}(t+1) = \mathbf{F}_k(t)\mathbf{x}(t) + \mathbf{D}_k(t)\mathbf{u}(t) + \mathbf{G}_k(t)\mathbf{w}_k(t)$$

that is, if $\mathcal{M}_k$ is the *system* model rather than the signal model. The only change that needs to be introduced comes in the state prediction equation, which in such a more general case should read

$$\widehat{\mathbf{x}}_k(t+1|t) = \mathbf{F}_k(t)\widehat{\mathbf{x}}_k(t|t) + \mathbf{D}_k(t)\mathbf{u}(t).$$

D. Efficient computational procedures

Even though the two-filter formula (17)-(18) proved to be very useful for derivation of the relationship (22), from the computational viewpoint it is not the most efficient way of evaluating the smoothed estimates. Below we summarize the two other procedures, known as the Rauch-Tung-Striebel (RTS) algorithm [27] and Bryson-Frazier (BF) algorithm [28]. In both cases the smoothed estimates are obtained by means of backward-time processing of the results yielded by the forward Kalman filter (16). Both algorithms update the covariance matrix $\mathbf{P}_k(t|N)$ needed to compute $\mathbf{S}_k(t|N)$ [cf. (19)], and both are free of technical subtleties associated with determination of the backwards Markovian model, used in the two-filter approach.

[3]Credibility coefficient are approximate, not exact, posterior probabilities.

1) Rauch-Tung-Striebel smoothing formula: Let

$$\mathbf{A}_k(t) = \mathbf{P}_k(t|t)\mathbf{F}_k^{\mathrm{T}}\mathbf{P}_k^{-1}(t+1|t).$$

The RTS formula can be summarized as follows

$$\begin{aligned} \widehat{\mathbf{x}}_k(t|N) &= \widehat{\mathbf{x}}_k(t|t) + \mathbf{A}_k(t)\left[\widehat{\mathbf{x}}_k(t+1|N) - \widehat{\mathbf{x}}_k(t+1|t)\right] \\ \mathbf{P}_k(t|N) &= \mathbf{P}_k(t|t) \\ &\quad + \mathbf{A}_k(t)[\mathbf{P}_k(t+1|N) - \mathbf{P}_k(t+1|t)]\mathbf{A}_k^{\mathrm{T}}(t) \\ t &= N-1, \ldots, 1 \end{aligned} \tag{28}$$

The initial conditions $\widehat{\mathbf{x}}_k(N|N)$ and $\mathbf{P}_k(N|N)$ are provided by the forward Kalman filter.

2) Bryson-Frazier smoothing formula: Let

$$\mathbf{B}_k(t) = \mathbf{F}_k(t)[\mathbf{I} - \mathbf{K}(t)\mathbf{H}_k^{\mathrm{T}}(t)].$$

The BF formula is given in the form

$$\begin{aligned} \mathbf{r}_k(t-1) &= \mathbf{B}_k^{\mathrm{T}}(t)\mathbf{r}_k(t) + \mathbf{H}_k^{\mathrm{T}}(t)\mathbf{Q}_k^{-1}(t)\boldsymbol{\varepsilon}_k(t) \\ \mathbf{R}_k(t-1) &= \mathbf{B}_k^{\mathrm{T}}(t)\mathbf{R}_k(t)\mathbf{B}_k(t) + \mathbf{H}_k^{\mathrm{T}}(t)\mathbf{Q}_k^{-1}(t)\mathbf{H}_k(t) \\ \widehat{\mathbf{x}}_k(t|N) &= \widehat{\mathbf{x}}_k(t|t-1) + \mathbf{P}_k(t|t-1)\mathbf{r}_k(t-1) \\ \mathbf{P}_k(t|N) &= \mathbf{P}_k(t|t-1) \\ &\quad - \mathbf{P}_k(t|t-1)\mathbf{R}_k(t-1)\mathbf{P}_k(t|t-1) \\ t &= N-1, \ldots, 1 \end{aligned} \tag{29}$$

with initial conditions set to $\mathbf{r}_k(N) = \mathbf{0}$ and $\mathbf{R}_k(N) = \mathbf{O}$.

The formula similar to the BF algorithm (even though originally derived from the RTS algorithm), known as modified Bryson-Frazier (MBF) smoother, was proposed by Bierman [29], [30].

Remark

To improve numerical conditioning and stability of computations, the equation-based MF, RTS, BF and MBF algorithms can be replaced with their fast square-root versions [31], [32], [33]. Such factorized algorithms, which update the square-root $\mathbf{P}_k^{1/2}(t|N)$ of the matrix $\mathbf{P}_k(t|N)$ [$\mathbf{P}_k^{1/2}(t|N)\mathbf{P}_k^{\mathrm{T}/2}(t|N) = \mathbf{P}_k(t|N)$], are also amenable to parallel and systolic implementation.

IV. Application – identification of nonstationary stochastic systems

Consider the problem of identification, based on prerecorded input-output data, of a discrete-time stochastic system governed by

$$y(t) = \boldsymbol{\varphi}^{\mathrm{T}}(t)\boldsymbol{\theta}(t) + v(t) \tag{30}$$

where $\boldsymbol{\varphi}(t) = [\varphi_1(t), \ldots, \varphi_r(t)]^{\mathrm{T}}$ denotes the vector of input (regression) variables and $\boldsymbol{\theta}(t) = [\theta_1(t), \ldots, \theta_r(t)]^{\mathrm{T}}$ is the vector of unknown, time-varying system coefficients.

The locally-adaptive identification algorithm, developed below, is based on the concept of cooperative Kalman smoothing. In the context of system identification, similar solutions (convex combination of competing estimates) were proposed for parameter trackers, see e.g. [34] and [35]. However, to the best of our knowledge, no such results are yet available for non-causal estimation schemes combining parameter smoothers.

A. Nonprobabilistic problem formulation

The smoothed estimate $\widehat{\boldsymbol{\theta}}(t|N)$ of $\boldsymbol{\theta}(t)$ can be obtained by solving the following minimization problem

$$\widehat{\boldsymbol{\theta}}(\cdot|N) = \arg\min_{\boldsymbol{\theta}(\cdot)} \left\{ \sum_{t=1}^{N} \left[y(t) - \boldsymbol{\varphi}^{\mathrm{T}}(t)\boldsymbol{\theta}(t) \right]^2 + \xi \sum_{t=1}^{N} \| \nabla^p \boldsymbol{\theta}(t) \|^2 \right\} \tag{31}$$

where $\nabla^p \boldsymbol{\theta}(t)$ denotes the p-th order difference of $\boldsymbol{\theta}(t)$

$$\nabla^p \boldsymbol{\theta}(t) = (1 - q^{-1})^p \boldsymbol{\theta}(t) = \sum_{i=0}^{p} f_i \boldsymbol{\theta}(t-i)$$

$$f_i = (-1)^i \begin{pmatrix} p \\ i \end{pmatrix}, \; i = 0, \ldots, p$$

q^{-1} is the backward shift operator and $\xi > 0$ denotes the user-dependent smoothness tradeoff parameter.

Such a problem formulation can be traced back to Whittaker [36], who used it to work out a procedure for signal smoothing. In the context of system identification (estimation of an impulse response of a dynamic system), the same approach was later pursued by Shiller [37]. As Whittaker stated it, the estimates (31) balance a tradeoff between infidelity to the data, represented by the first sum on the right side of (31), and infidelity to a p-th order difference equation constraint, represented by the second sum in (31).

B. Stochastic embedding

The deterministic (as far as description of $\boldsymbol{\theta}(t)$ is concerned), regularized least squares problem formulated by Whittaker was given a probabilistic reinterpretation by Akaike [38]. Suppose that the following stochastic integrated random-walk (IRW) parameter model is adopted

$$\nabla^p \boldsymbol{\theta}(t) = \mathbf{w}(t), \quad \mathrm{cov}[\mathbf{w}(t)] = \mathbf{W} = \sigma_w^2 \mathbf{I} \tag{32}$$

where $\{\mathbf{w}(t)\}$ denotes a zero-mean i.i.d. sequence, independent of $\{v(t)\}$ and $\{\boldsymbol{\varphi}(t)\}$. Since $\nabla^p \boldsymbol{\theta}(t) = 0$ implies

$$\theta_i(t) = \sum_{j=1}^{p} a_{ij} t^{j-1}, \quad i = 1, \ldots, r$$

where a_{ij} denote arbitrary constants, the IRW model can be regarded as a local, or "perturbed", power series model of parameter variation. Generally, the larger the order p of the IRW model, the smoother the corresponding parameter trajectory. As noted in [38], when the sequences $\{v(t)\}$ and $\{\mathbf{w}(t)\}$ are Gaussian, and when ξ is adopted in the form

$$\xi = \frac{\sigma_w^2}{\sigma_v^2} \; ,$$

evaluation of the maximum likelihood estimates of parameters governed by (32) is equivalent to solving (31).

The final step was made by Norton [39] and Kitagawa and Gersch [40], who embedded smoothness constraints into a state-space system model

$$\begin{aligned} \mathbf{x}(t) &= \mathbf{F}\mathbf{x}(t-1) + \mathbf{G}\mathbf{w}(t) \\ y(t) &= \boldsymbol{\psi}^{\mathrm{T}}(t)\mathbf{x}(t) + v(t) \end{aligned} \tag{33}$$

where

$$\mathbf{x}(t) = \begin{bmatrix} \boldsymbol{\theta}(t) \\ \boldsymbol{\theta}(t-1) \\ \vdots \\ \boldsymbol{\theta}(t-p+1) \end{bmatrix}, \quad \boldsymbol{\psi}(t) = \begin{bmatrix} \boldsymbol{\varphi}(t) \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}$$

$$\mathbf{F} = \begin{bmatrix} -f_1\mathbf{I} & -f_2\mathbf{I} & \ldots & -f_{p-1}\mathbf{I} & -f_p\mathbf{I} \\ \mathbf{I} & \mathbf{O} & \ldots & \mathbf{O} & \mathbf{O} \\ & & \vdots & & \\ \mathbf{O} & \mathbf{O} & \ldots & \mathbf{I} & \mathbf{O} \end{bmatrix}, \mathbf{G} = \begin{bmatrix} \mathbf{I} \\ \mathbf{O} \\ \vdots \\ \mathbf{O} \end{bmatrix}$$

Note that

$$\boldsymbol{\theta}(t) = \mathbf{G}^{\mathrm{T}}\mathbf{x}(t) \tag{34}$$

which means that estimation of $\boldsymbol{\theta}(t)$ is equivalent to obtaining a smoothed estimate of the state vector $\mathbf{x}(t)$ based on $\{y(i), \boldsymbol{\varphi}(i), i = 1, \ldots, N\}$. Under Gaussian assumptions this can be achieved using Kalman smoother.

As argued in [15], in the context of system identification the coefficient ξ can be regarded as an instrumental variable, a sort of a user-dependent "knob" allowing one to tune the Kalman parameter tracker to the degree of nonstationarity of the identified process. In particular, when system parameters are assumed to vary according to the random-walk model ($p = 1$), one can show that the "estimation memory" of a Kalman parameter tracker based on (33) is inversely proportional to $\sqrt{\xi}$ [15].

C. Whittaker scheme revisited

The optimal-global values of the smoothness hyperparameters p and ξ can be obtained by minimizing some statistical loss function $Q(p, \xi)$, such as the Akaike's AIC statistics [40] or the generalized cross-validation index [22]. This can be achieved by a discrete search[4] over the parameters p and ξ, which is equivalent to running several algorithms corresponding to different values of p and ξ, and choosing the estimates providing the smallest value of $Q(p, \xi)$. By optimal-global we mean such constant parameter values that minimize the adopted loss function for the entire available data record. This should be contrasted with the optimal-local approach, pursued in this article, where the best-fitting parameters are searched independently for each location t, within the entire data set, of a short analysis window $T_{\mathrm{a}}(t)$. Finally, an intermediary approach is also possible, where the analyzed process is segmented and optimization is carried out independently for each data segment. Such a semi-local approach was described in [40].

Consider the situation where K Kalman smoothing algorithms, obtained for different values of design parameters $(p, \xi)_k, k = 1, \ldots, K$, and yielding the estimates $\widehat{\boldsymbol{\theta}}_k(t|N)$, are run in parallel.

[4]Even though a numerical search for the optimal values of smoothness hyperparameters is possible [41], it is usually computationally prohibitive.

The proposed locally-adaptive smoothed estimate of $\boldsymbol{\theta}(t)$ has the form

$$\begin{aligned}\widehat{\boldsymbol{\theta}}(t|N) &= \sum_{k=1}^{K} \mu_k^{\circ}(t)\widehat{\boldsymbol{\theta}}_k(t|N) \\ \widehat{\boldsymbol{\theta}}_k(t|N) &= \mathbf{G}_k^{\mathrm{T}}\widehat{\mathbf{x}}_k(t|N), \quad k=1,\ldots,K \end{aligned} \tag{35}$$

where $\widehat{\mathbf{x}}_k(t|N)$ denotes the estimate yielded by the k-th Kalman smoother.

For identification purposes it is recommended to use the following *normalized* version of the forward Kalman filter/predictor

$$\begin{aligned}\varepsilon_k(t) &= y(t) - \boldsymbol{\psi}_k^{\mathrm{T}}(t)\widehat{\mathbf{x}}_k(t|t-1) \\ \widetilde{q}_k(t) &= \boldsymbol{\psi}_k^{\mathrm{T}}(t)\widetilde{\mathbf{P}}_k(t|t-1)\boldsymbol{\psi}_k(t) + 1 \\ \mathbf{k}_k(t) &= \widetilde{\mathbf{P}}_k(t|t-1)\boldsymbol{\psi}_k(t)/\widetilde{q}_k(t) \\ \widehat{\mathbf{x}}_k(t|t) &= \widehat{\mathbf{x}}_k(t|t-1) + \mathbf{k}_k(t)\boldsymbol{\varepsilon}_k(t) \\ \widetilde{\mathbf{P}}_k(t|t) &= \widetilde{\mathbf{P}}_k(t|t-1) - \mathbf{k}_k(t)\mathbf{k}_k^{\mathrm{T}}(t)\widetilde{q}_k(t) \\ \widehat{\mathbf{x}}_k(t+1|t) &= \mathbf{F}_k\widehat{\mathbf{x}}_k(t|t) \\ \widetilde{\mathbf{P}}_k(t+1|t) &= \mathbf{F}_k\widetilde{\mathbf{P}}_k(t|t)\mathbf{F}_k^{\mathrm{T}} + \mathbf{G}_k\mathbf{G}_k^{\mathrm{T}}\xi_k \\ t &= 1,\ldots,N \end{aligned} \tag{36}$$

where $\widetilde{\mathbf{P}}_k(t+1|t) = \mathbf{P}_k(t+1|t)/(\sigma_v^2)_k$, $\widetilde{\mathbf{P}}_k(t|t) = \mathbf{P}_k(t|t)/(\sigma_v^2)_k$ denote normalized covariance matrices and $\widetilde{q}_k(t)$ is the normalized version of $q_k(t)$ – the scalar counterpart of the matrix $\mathbf{Q}_k(t)$ updated in (16). The normalized algorithm yields the same estimates as its unnormalized version, but it is equipped with only one design parameter ξ_k instead of two parameters $(\sigma_w^2, \sigma_v^2)_k$. This feature of the normalized algorithm confirms the well-known fact that parameter estimation properties of the Kalman-filter-based identification algorithm depend on the assumed ratio of variances $(\sigma_w^2/\sigma_v^2)_k$ rather than on each of these variances alone [15].

The corresponding MF, RTS and MBF smoothing formulas remain unchanged – all that one has to do is replace the quantities $\mathbf{P}_k(t|t)$, $\mathbf{P}_k(t+1|t)$ and $q_k(t)$ in (17)-(18), (28) and (29) with their normalized counterparts.

In the special case considered in this section the credibility coefficients can be obtained from

$$\mu_k^{\circ}(t) \propto \left\{ \sum_{i \in T_{\mathrm{a}}(t)} [e_k^{\circ}(i)]^2 \right\}^{-M/2} \tag{37}$$

where

$$e_k^{\circ}(i) = \frac{e_k(i)}{\widetilde{\sigma}_k(t)} = \frac{y(i) - \boldsymbol{\psi}_k^{\mathrm{T}}(i)\widehat{\mathbf{x}}_k(i|N)}{1 - \boldsymbol{\psi}_k^{\mathrm{T}}(i)\widetilde{\mathbf{P}}_k(i|N)\boldsymbol{\psi}_k(i)}$$

and $\widetilde{\mathbf{P}}_k(i|N)$ denotes the normalized covariance matrix of the estimate $\widehat{\mathbf{x}}_k(i|N)$ provided by Kalman smoother.

D. Non-Gaussian Extensions

In the scalar measurement case, considered in this section, the cooperative smoothing formula (35) can be extended to a wider class of measurement noise distributions, known as generalized normal [42]. The generalized normal law incorporates such practically important distributions as Gaussian, Laplace and uniform. For example, when noise distribution is Laplace (heavy-tailed), credibility coefficients should be evaluated according to [14]

$$\mu_k^{\circ}(t) \propto \left\{ \sum_{i \in T_{\mathrm{a}}(t)} |e_k^{\circ}(i)| \right\}^{-M}. \tag{38}$$

In the uniform (light-tailed) noise case, one should set

$$\mu_k^{\circ}(t) \propto \left[\max_{i \in T_{\mathrm{a}}(t)} |e_k^{\circ}(i)| \right]^{-M}. \tag{39}$$

The smoothing formula based on (38) is more robust (e.g. to outliers) than the one based on (37). Unfortunately, due to technical difficulties associated with the form of the multivariate Laplace distribution [43], a straightforward extension of (38) to multivariate signals does not seem to be possible.

For white but non-Gaussian noise sources the estimates yielded by the constituent Kalman smoothers cannot be claimed optimal any more, but they still have a nice statistical interpretation: they can be viewed as an orthogonal projection of the state vector $\mathbf{x}(t)$ onto a linear space spanned by the available measurements. This means that they are minimum mean square *linear* estimates of $\mathbf{x}(t)$, optimal in the sense of Whittaker (31).

E. Computational Complexity

In the scalar measurement case ($n = 1$), analyzed in this section, the MF algorithm (17) - (18) is computationally the least attractive one, as it requires inversion of 3 matrices of dimension $r \times r$ at each time step (only 2 inversions are needed if the information filter is used for the backward sweep), and the BF and MBF algorithms are the most attractive ones, as they do not require matrix inversion at all. The RTS algorithm comes in the middle since it requires 1 inversion per (reverse) time update.

Since the computational load associated with evaluation of credibility coefficients $\mu_k^{\circ}(t)$ is negligible compared to the cost of evaluating the corresponding estimates $\widehat{\mathbf{x}}_k(i|N)$, the computational burden of the cooperative smoother is not much higher than the cost of running K standard Kalman smoothing algorithms. In contrast with this, the IMM smoother described in [13] consists of K^2 MF-type smoothing algorithms operating in parallel. Hence, for systems with Markovian switching parameters, cooperative smoother can be regarded as a computationally attractive alternative to the IMM smoother (of course, this computational advantage comes at the cost of decreased smoothing efficiency).

V. Simulation results

To check performance of the proposed parameter smoothing algorithm, the following two-tap FIR system (inspired by channel equalization applications) was simulated

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

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.

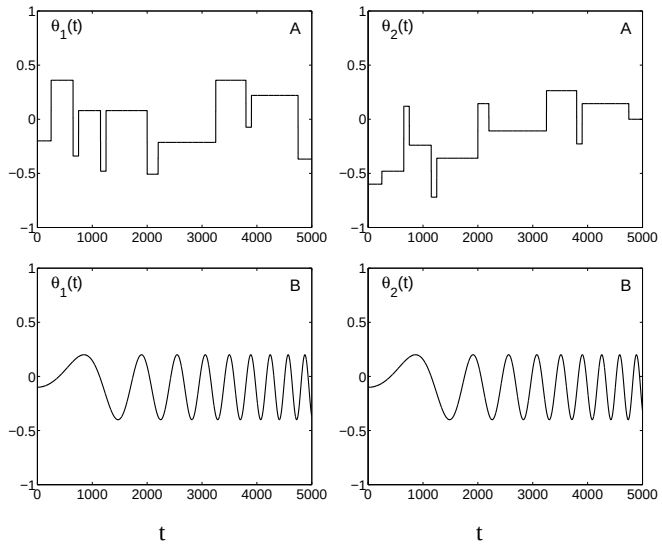

Fig. 1. Two variants of parameter changes used in computer simulations: discontinuous (two upper plots) and continuous (two lower plots).

Two different scenarios of parameter changes were considered: A) discontinuous, step-like and B) continuous, chirp-like – see Fig. 1. Due to appropriate scaling, all parameter trajectories have the same L_2 norm.

For each of the compared algorithms the steady state accumulated mean-squared parameter estimation errors were computed

$$\mathrm{E}_v\{\textstyle\sum_{t=101}^{4900} \parallel \widehat{\boldsymbol{\theta}}(t|5000) - \boldsymbol{\theta}(t) \parallel^2\}. \qquad (40)$$

To eliminate transient effects, the summation in (40) was restricted to the interval $[101, 4900]$. Ensemble averaging $\mathrm{E}_v(\cdot)$ was performed over 100 realizations of the measurement noise $\{v(t)\}$ – the same in all experiments. The procedure was repeated for each of 6 noise intensities ranging from $\sigma_v = 0.05$ (SNR=26 dB) to $\sigma_v = 0.30$ (SNR=10.5 dB). The width of the evaluation frame was equal to $M = 21$. The distribution of noise was either Gaussian or Laplacian.

Table 1 summarizes results obtained for 6 Kalman smoothers ($\mathrm{K}_1, \ldots, \mathrm{K}_6$) and 2 cooperative smoothers (K_{1-6}, K_{1-3}). The first 3 Kalman algorithms (K_1, K_2, K_3) were based on the first-order IRW model ($p = 1$, $\xi_1 = 0.002$, $\xi_2 = 0.018$, $\xi_3 = 0.16$), and the remaining 3 algorithms (K_4, K_5, K_6) – on the second-order IRW model ($p = 2$, $\xi_4 = 6 \cdot 10^{-8}$, $\xi_5 = 5 \cdot 10^{-6}$, $\xi_6 = 4 \cdot 10^{-4}$). Cooperative smoothers K_{1-6} and K_{1-3} combined all 6 Kalman algorithms and the first 3 Kalman algorithms, respectively. The design parameters ξ_k of constituent smoothers were not optimized in any way – the corresponding values were chosen so that within each group of algorithms $\{\mathrm{K}_1, \mathrm{K}_4\}$, $\{\mathrm{K}_2, \mathrm{K}_5\}$ and $\{\mathrm{K}_3, \mathrm{K}_6\}$, the equivalent memory spans [15] were the same and equal to 90 samples, 30 samples and 10 samples, respectively.

Note that the smoother K_{1-6}, which combines algorithms based on both first-order and second-order IRW models, is uniformly better than K_{1-3}, i.e., it provides the smallest estimation errors in all cases considered. Note also that it yields either better results (for the system with step-like parameter changes) or only slightly worse results (for the system with chirp-like parameter changes) than the best smoothers among $\mathrm{K}_1, \ldots, \mathrm{K}_6$. An obvious advantage of all cooperative smoothers is their increased robustness to unknown and possibly time-varying degree of nonstationarity of the identified system.

To get better insight into the quality of reconstructions provided by the cooperative smoothing scheme, look at the plots displayed in Figs. 2 (for step-like parameter changes) and 3 (for chirp-like parameter changes). Each figure shows the best-fitting trajectory, the mid-score trajectory (50th score) and the worst-fitting trajectory from among 100 estimated parameter trajectories yielded, for different realizations of Gaussian measurement noise ($\sigma_v = 0.15$), by the K_{1-6} smoother. Note that the main source of quality degradation, when one moves from the best result to the worst result, is due to sporadic short-lived switching artifacts. Switching artifact occurs when, owing to a specific local noise pattern, one of less efficient smoothers temporarily dominates the entire smoother bank. Most of these artifacts can be easily eliminated by means of median postfiltering of the sequence $\{\widehat{\boldsymbol{\theta}}(t|N)\}$. For example, in the case illustrated in Figs. 2 and 3, 11-point median smoothing allows one to reduce the average estimation loss (40) by aproximately 10%: from 3.98 to 3.75 for step-like parameter changes, and from 0.63 to 0.57 for chirp-like parameter changes. It also makes estimation results more homogeneous across different realizations.

Median filter is just the simplest representative of a whole family of nonlinear order statistical filters with noteworthy properties: they are capable of removing outliers without blurring step-like signal features [21]. Exploration of the potential for nonlinear postfiltering[5] seems to be an interesting topic for future research.

VI. Conclusion

We have considered the problem of Kalman smoothing in the presence of model uncertainty and we have proposed a novel parallel estimation (multiple-model) scheme which combines results yielded by several competing Kalman smoothers. The solution was based on the recently proposed concept of cooperative smoothing. After extending the rules of cooperative smoothing to multivariate signals, we have shown that credibility coefficients – which play the role of mixing coefficients in the cooperative smoothing formula – can be easily evaluated using quantities provided by the constituent Kalman algorithms. Finally, we have shown how cooperative Kalman smoothing can be applied to identification of nonstationary stochastic systems and we have indicated possible directions for future research (nonlinear postfiltering).

Appendix 1 [derivation of (8)]

Consider a $n \times n$ random matrix $\mathbf{X} > \mathbf{O}$ which obeys the inverse Wishart distribution with an $n \times n$ inverse scale matrix $\mathbf{Y} > \mathbf{O}$ and $p \geq n$ degrees of freedom: $\mathbf{X} \sim \mathcal{W}^{-1}(\mathbf{Y}, p)$. The

[5] Linear postfiltering schemes were proposed earlier in [44].

Table 1
Comparison of parameter estimation errors obtained for 3 Kalman smoothers based on the first-order IRW model (K_1, K_2, K_3), 3 Kalman smoothers based on the second-order IRW model (K_4, K_5, K_6), and 2 cooperative smoothers (K_{1-6}, K_{1-3}). Simulations were performed for 2 variants of parameter changes (A,B) and 6 noise intensities.

Gaussian noise:

P	σ_v	K_1	K_2	K_3	K_4	K_5	K_6	K_{1-6}	K_{1-3}
A	0.05	26.30	8.40	**3.00**	8.26	25.21	91.50	**2.80**	**2.80**
	0.10	26.42	8.76	**4.11**	8.68	25.37	91.50	**3.22**	3.25
	0.15	26.04	9.42	**6.04**	9.45	25.62	91.69	**3.98**	4.10
	0.20	26.89	10.25	**8.63**	10.46	25.92	91.71	**4.82**	5.08
	0.25	27.24	**11.31**	11.86	11.67	26.32	91.86	**5.83**	6.27
	0.30	27.66	**12.91**	16.07	13.53	26.93	91.70	**7.20**	7.80
B	0.05	3.95	0.15	0.22	**0.08**	0.12	65.70	**0.08**	0.13
	0.10	4.01	0.35	0.86	0.32	**0.20**	65.71	**0.27**	0.41
	0.15	4.20	0.75	1.97	0.77	**0.35**	65.83	**0.63**	0.90
	0.20	4.47	1.29	3.51	1.38	**0.56**	66.15	**1.07**	1.51
	0.25	4.68	1.88	5.41	2.07	**0.78**	65.14	**1.53**	2.16
	0.30	4.80	2.76	7.86	3.11	**1.10**	66.03	**2.15**	3.00

Laplacian noise:

P	σ_v	K_1	K_2	K_3	K_4	K_5	K_6	K_{1-6}	K_{1-3}
A	0.05	26.26	8.35	**2.98**	8.24	25.20	91.47	**2.74**	2.75
	0.10	26.40	8.79	**4.14**	8.74	25.34	91.52	**3.19**	3.23
	0.15	26.63	9.40	**5.97**	9.46	25.60	91.68	**3.80**	3.91
	0.20	27.00	10.38	**8.62**	10.54	26.00	91.82	**4.54**	4.98
	0.25	27.20	**11.28**	11.64	11.59	26.18	91.95	**5.28**	5.66
	0.30	27.71	**12.71**	15.76	13.23	26.89	91.97	**6.47**	7.04
B	0.05	3.97	0.15	0.22	**0.09**	0.13	65.77	**0.06**	0.12
	0.10	4.06	0.38	0.87	0.35	**0.23**	65.78	**0.24**	0.39
	0.15	4.15	0.75	1.94	0.77	**0.36**	65.74	**0.50**	0.77
	0.20	4.36	1.31	3.49	1.42	**0.60**	65.84	**0.84**	1.29
	0.25	4.43	1.84	5.26	2.06	**0.76**	65.71	**1.14**	2.76
	0.30	4.75	2.64	7.61	2.98	**1.06**	66.03	**1.62**	2.45

probability density function of $\mathbf{X}$ takes the form [18]

$$p(\mathbf{X}) = \frac{[\det(\mathbf{X})]^{-(p+n+1)/2}[\det(\mathbf{Y})]^{p/2}}{2^{pn/2}\Gamma_n(\frac{p}{2})} \times \exp\left\{-\frac{1}{2}\,\mathrm{tr}[\mathbf{X}^{-1}\mathbf{Y}]\right\}$$

and since it must integrate to 1 when $\mathbf{X}$ spans the entire space of $n \times n$ positive definite matrices, one obtains the following integral formula (with no probabilistic connotation)

$$\int_{\mathbf{X}>\mathbf{O}} [\,\det(\mathbf{X})\,]^{-(p+n+1)/2} \exp\left\{-\frac{1}{2}\,\mathrm{tr}\left[\mathbf{X}^{-1}\mathbf{Y}\right]\right\} d\mathbf{X} = 2^{\frac{pn}{2}}\Gamma_n\left(\frac{p}{2}\right)[\,\det(\mathbf{Y})\,]^{-\frac{p}{2}}. \quad (41)$$

The result (8) follows immediately from (41) after setting $\mathbf{X} = \mathbf{V}$, $\mathbf{Y} = \mathbf{D}_k(t)$ and $p = M$.

Appendix 2 [derivation of (22) - (23)]

The proof will exploit the well-known matrix inversion lemma [45], given below in a slightly more general ($\pm$) form

Lemma

Provided that all inverses below exist:

$$[\mathbf{A} \pm \mathbf{BCD}]^{-1} = \mathbf{A}^{-1} \mp \mathbf{A}^{-1}\mathbf{B}\left[\mathbf{C}^{-1} \pm \mathbf{DA}^{-1}\mathbf{B}\right]^{-1}\mathbf{DA}^{-1}.$$

■

Applying the matrix inversion lemma ($+$) to the fifth recursion of (16), one arrives at

$$\mathbf{P}_k(t|t) = \left[\mathbf{P}_k^{-1}(t|t-1) + \mathbf{H}_k^{\mathrm{T}}(t)\mathbf{V}_k^{-1}(t)\mathbf{H}_k(t)\right]^{-1} \quad (42)$$

which, when combined with (18), leads to

$$\mathbf{P}_k(t|N) = \left\{[\mathbf{P}_k(t|t-1)]^{-1} + \mathbf{H}_k^{\mathrm{T}}(t)\mathbf{V}_k^{-1}(t)\mathbf{H}_k(t) + [\mathbf{P}_k^{\mathrm{B}}(t|t-1)]^{-1}\right\}^{-1}. \quad (43)$$

Combining the first and the fourth recursion of (16), one obtains

$$\widehat{\mathbf{x}}_k(t|t) = \mathbf{K}_k(t)\mathbf{y}(t) + [\mathbf{I} - \mathbf{K}_k(t)\mathbf{H}_k(t)]\widehat{\mathbf{x}}_k(t|t-1) \quad (44)$$

Multiplying both sides of the fifth recursion of (16) with $\mathbf{P}_k^{-1}(t|t-1)$, and incorporating the third recursion, one arrives at

$$\begin{aligned}\mathbf{P}_k(t|t)\mathbf{P}_k^{-1}(t|t-1) &= \mathbf{I} - \mathbf{K}_k(t)\mathbf{Q}_k(t)\mathbf{K}_k^{\mathrm{T}}(t)\mathbf{P}_k^{-1}(t|t-1)\\ &= \mathbf{I} - \mathbf{P}_k(t|t-1)\mathbf{H}_k^{\mathrm{T}}(t)\mathbf{Q}_k^{-1}(t)\mathbf{H}_k^{\mathrm{T}}(t)\\ &= \mathbf{I} - \mathbf{K}_k(t)\mathbf{H}_k(t). \end{aligned} \quad (45)$$

Therefore, (44) can be written down in the form

$$\widehat{\mathbf{x}}_k(t|t) = \mathbf{K}_k(t)\mathbf{y}(t) + \mathbf{P}_k(t|t)\mathbf{P}_k^{-1}(t|t-1)\widehat{\mathbf{x}}_k(t|t-1) \quad (46)$$

leading to

$$\mathbf{P}_k^{-1}(t|t)\widehat{\mathbf{x}}_k(t|t) = \mathbf{P}_k^{-1}(t|t)\mathbf{K}_k(t)\mathbf{y}(t) + \mathbf{P}_k^{-1}(t|t-1)\widehat{\mathbf{x}}_k(t|t-1). \quad (47)$$

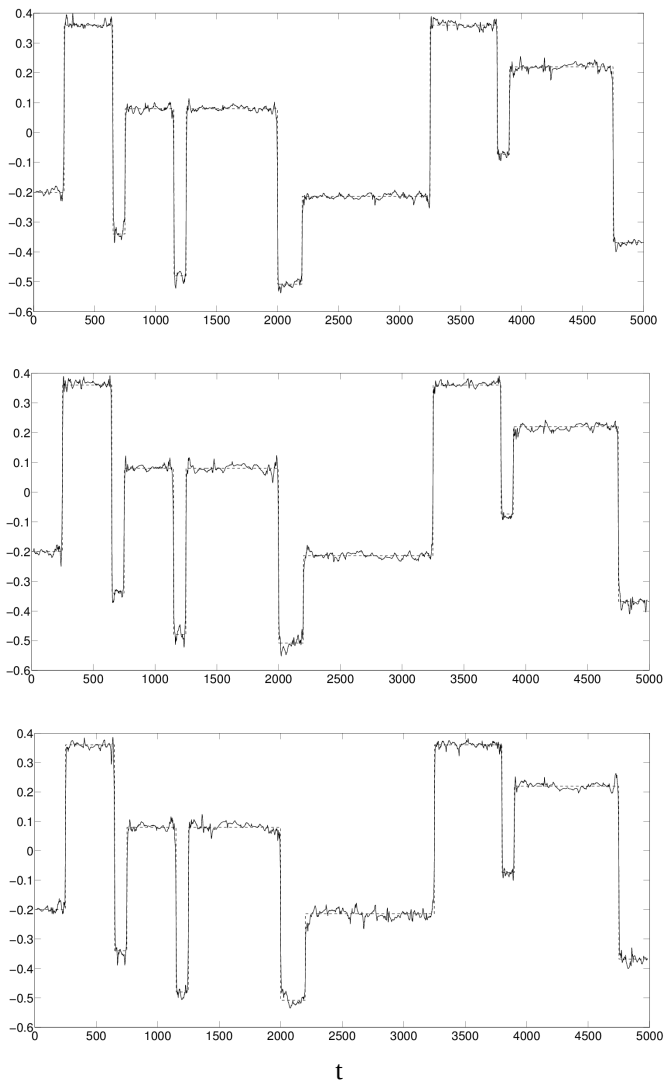

Fig. 2. The best-fitting (top plot), mid-score (middle plot) and the worst-fitting (bottom plot) parameter trajectories from among 100 trajectories generated, for different realizations of Gaussian measurement noise ($\sigma_v = 0.15$), by the K_{1-6} smoother. In each case the plot of the estimated parameter trajectory $\{\widehat{\theta}_1(t|5000)\}$ is superimposed on the plot of the true (step-like) parameter trajectory $\{\theta_1(t|5000)\}$.

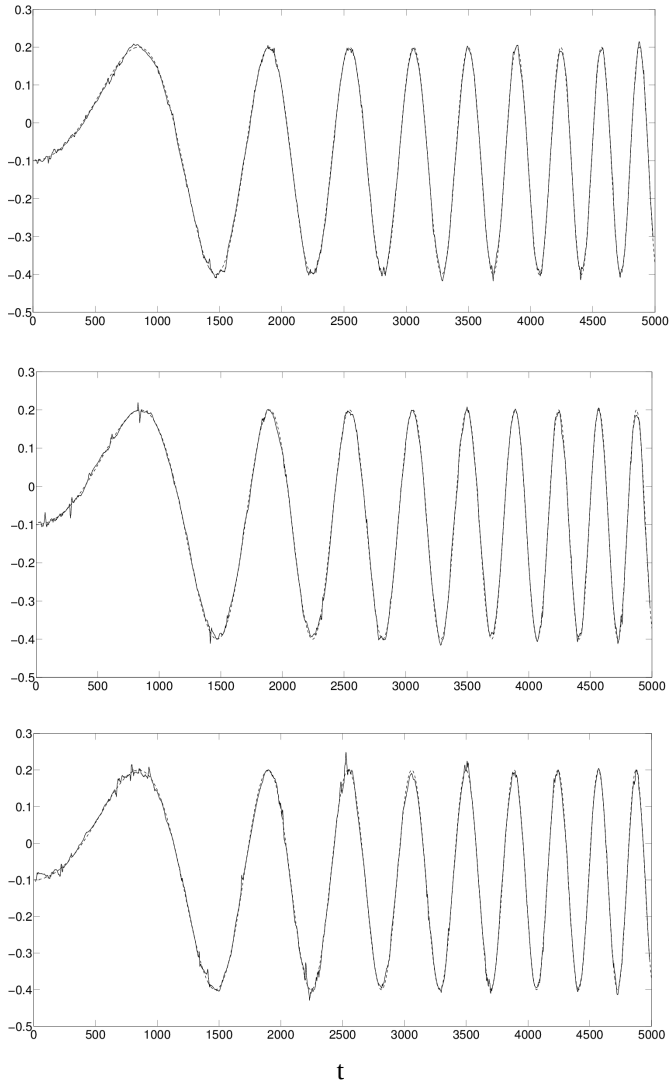

Fig. 3. The best-fitting (top plot), mid-score (middle plot) and the worst-fitting (bottom plot) parameter trajectories from among 100 trajectories generated, for different realizations of Gaussian measurement noise ($\sigma_v = 0.15$), by the K_{1-6} smoother. In each case the plot of the estimated parameter trajectory $\{\widehat{\theta}_1(t|5000)\}$ is superimposed on the plot of the true (chirp-like) parameter trajectory $\{\theta_1(t|5000)\}$.

Using (42) and the third recursion of (16), one obtains

$$\begin{aligned}\mathbf{P}_k^{-1}(t|t)\mathbf{K}_k(t) &= \left[\mathbf{P}_k^{-1}(t|t-1) + \mathbf{H}_k^{\mathrm{T}}(t)\mathbf{V}_k^{-1}(t)\mathbf{H}_k(t)\right] \\ &\quad \times \mathbf{P}_k(t|t-1)\mathbf{H}_k^{\mathrm{T}}(t)\mathbf{Q}_k^{-1}(t) \\ &= \mathbf{H}_k^{\mathrm{T}}(t)[\mathbf{I} + \mathbf{V}_k^{-1}(t)\mathbf{H}_k(t)\mathbf{P}_k^{-1}(t|t-1)\mathbf{H}_k^{\mathrm{T}}(t)]\mathbf{Q}_k^{-1}(t).\end{aligned} \tag{48}$$

Since the second recursion of (16) can be rewritten in the form

$$\mathbf{V}_k^{-1}(t)\mathbf{Q}_k(t) = \mathbf{I} + \mathbf{V}_k^{-1}(t)\mathbf{H}_k(t)\mathbf{P}_k^{-1}(t|t-1)\mathbf{H}_k^{\mathrm{T}}(t) \tag{49}$$

one arrives at

$$\mathbf{P}_k^{-1}(t|t)\mathbf{K}_k(t) = \mathbf{H}_k^{\mathrm{T}}(t)\mathbf{V}_k^{-1}(t) \tag{50}$$

and [cf. (47)]

$$\begin{aligned}\mathbf{P}_k^{-1}(t|t)\widehat{\mathbf{x}}_k(t|t) &= \mathbf{H}_k^{\mathrm{T}}(t)\mathbf{V}_k^{-1}(t)\mathbf{y}(t) \\ &\quad + \mathbf{P}_k^{-1}(t|t-1)\widehat{\mathbf{x}}_k(t|t-1).\end{aligned} \tag{51}$$

Note that, using (17), (18), (43) and (51), one can rewrite (20) in the form

$$\begin{aligned}\widehat{\mathbf{x}}_k^{\circ}(t|N) &= \left[\mathbf{P}_k^{-1}(t|N) - \mathbf{H}_k^{\mathrm{T}}(t)\mathbf{V}_k^{-1}(t)\mathbf{H}_k(t)\right]^{-1} \\ &\quad \times \left[\mathbf{P}_k^{-1}(t|N)\widehat{\mathbf{x}}_k(t|N) - \mathbf{H}_k^{\mathrm{T}}(t)\mathbf{V}_k^{-1}(t)\mathbf{y}(t)\right]\end{aligned} \tag{52}$$

Using the matrix inversion lemma again $(-)$, one obtains

$$\begin{aligned}&\left[\mathbf{P}_k^{-1}(t|N) - \mathbf{H}_k^{\mathrm{T}}(t)\mathbf{V}_k^{-1}(t)\mathbf{H}_k(t)\right]^{-1} \\ &= \mathbf{P}_k(t|N) + \mathbf{P}_k(t|N)\mathbf{H}_k^{\mathrm{T}}(t)\mathbf{Z}_k^{-1}(t)\mathbf{H}_k(t)\mathbf{P}_k(t|N)\end{aligned} \tag{53}$$

where

$$\begin{aligned}\mathbf{Z}_k(t) &= \mathbf{V}_k(t) - \mathbf{H}_k(t)\mathbf{P}_k(t|N)\mathbf{H}_k^{\mathrm{T}}(t) \\ &= \mathbf{V}_k(t) - \mathbf{S}_k(t|N).\end{aligned} \tag{54}$$

Substituting (53) into (52), one gets

$$\begin{aligned}\widehat{\mathbf{x}}_k^{\circ}(t|N) &= \widehat{\mathbf{x}}_k(t|N) + \mathbf{P}_k(t|N)\mathbf{H}_k^{\mathrm{T}}(t)\mathbf{Z}_k^{-1}(t)\mathbf{H}_k(t)\widehat{\mathbf{x}}_k(t|N)\\ &- \mathbf{P}_k(t|N)\mathbf{H}_k^{\mathrm{T}}(t)\mathbf{Z}_k^{-1}(t)\mathbf{H}_k(t)\mathbf{P}_k(t|N)\mathbf{H}_k^{\mathrm{T}}(t)\mathbf{V}_k^{-1}(t)\mathbf{y}(t)\\ &- \mathbf{P}_k(t|N)\mathbf{H}_k^{\mathrm{T}}(t)\mathbf{V}_k^{-1}(t)\mathbf{y}(t) \qquad (55)\end{aligned}$$

Using (55), one arrives at

$$\begin{aligned}&\mathbf{e}_k^{\circ}(t) = \mathbf{y}(t) - \mathbf{H}_k(t)\widehat{\mathbf{x}}_k^{\circ}(t|N) = \mathbf{y}(t) - \mathbf{H}_k(t)\widehat{\mathbf{x}}_k(t|N)\\ &- \mathbf{H}_k(t)\mathbf{P}_k(t|N)\mathbf{H}_k^{\mathrm{T}}(t)\mathbf{Z}_k^{-1}(t)\mathbf{H}_k(t)\widehat{\mathbf{x}}_k(t|N)\\ &+ \mathbf{H}_k(t)\mathbf{P}_k(t|N)\mathbf{H}_k^{\mathrm{T}}(t)\mathbf{Z}_k^{-1}(t)\mathbf{H}_k(t)\mathbf{P}_k(t|N)\mathbf{H}_k^{\mathrm{T}}(t)\\ &\times \mathbf{V}_k^{-1}(t)\mathbf{y}(t) + \mathbf{H}_k(t)\mathbf{P}_k(t|N)\mathbf{H}_k^{\mathrm{T}}(t)\mathbf{V}_k^{-1}(t)\mathbf{y}(t)\\ &= \left[\mathbf{I} + \mathbf{S}_k(t|N)\mathbf{V}_k^{-1}(t) + \mathbf{S}_k(t|N)\mathbf{Z}_k^{-1}(t)\mathbf{S}_k(t|N)\mathbf{V}_k^{-1}(t)\right]\\ &\times \mathbf{y}(t) - \left[\mathbf{I} + \mathbf{S}_k(t|N)\mathbf{Z}_k^{-1}(t)\right]\mathbf{H}_k(t)\widehat{\mathbf{x}}_k(t|N). \qquad (56)\end{aligned}$$

According to (54) it holds that

$$\mathbf{S}_k(t|N) = \mathbf{V}_k(t) - \mathbf{Z}_k(t). \qquad (57)$$

Substituting (57) into (56), one gets (after elementary calculations)

$$\begin{aligned}\mathbf{e}_k^{\circ}(t) &= \mathbf{V}_k(t)\mathbf{Z}_k^{-1}(t)[\mathbf{y}(t) - \mathbf{H}_k(t)\widehat{\mathbf{x}}_k(t|N)]\\ &= \mathbf{V}_k(t)\mathbf{Z}_k^{-1}(t)\mathbf{e}_k(t) \qquad (58)\end{aligned}$$

which is identical with (22) – (23).

REFERENCES

[1] R.G. Brown and P.Y.C. Hwang, *Introduction to Random Signals and Applied Kalman Filtering*. Wiley, 1992.

[2] F. Gustafsson, *Adaptive Filtering and Change Detection*. Wiley, 2000.

[3] F. Lewis, *Optimal Estimation*. Wiley, 1986.

[4] D.Q. Mayne, "A solution of the smoothing problem for linear dynamic systems," *Automatica*, vol. 4, pp. 73–92, 1966.

[5] D.C. Fraser, "A new technique for the optimal smoothing of data," Sc.D. Dissertation, MIT, Cambridge, MA., 1967 (see also *IEEE Trans. Automat Contr.* vol. 14, pp. 387-390, 1969).

[6] J.S. Meditch, *Stochastic Optimal Linear Estimatio and Control*. McGraw-Hill, 1969.

[7] D.T. Magill, "Optimal adaptive estimation of sampled stochastic processes," *IEEE Trans. Automat Contr.*, vol. 10, pp. 434–439, 1965.

[8] Y. Bar-Shalom, Y. Li and T. Kirubarajan *Estimation with applications to Tracking and Navigation*. Wiley, 2001.

[9] C. Hide, T. Moore and M. Smith "Adaptive Kalman filtering for low-coct INS/GPS," *The Journal of Navigation*, vol. 56, pp. 143–152, 2003.

[10] H.A.P. Blom and Y. Bar-Shalom, "The interacting multiple-model algorithm for systems with Markovian switching coefficients," *IEEE Trans. Automat. Contr.*, vol. 33, pp. 780–783 , 1988.

[11] X.R. Li and V.P. Jilkov, "Survey of maneuvering target tracking. Part V: Multiple-model methods," *IEEE Trans. Aerospace and Electron. Syst.*, vol.41, pp. 1255–1321, 2005.

[12] H.A.P. Blom and Y. Bar-Shalom, "Time-reversion of a hybrid state stochastic difference system with a jump-linear smoothing applications," *IEEE Trans. Inform. Theory*, vol. 36, pp. 836–847 , 1990.

[13] R.E. Helmick, W.D. Blair and S.A. Hoffman, " Fixed-interval smoothing for Markovian switching systems," *IEEE Trans. Inform. Theory*, vol. 41, pp. 1845–1855 , 1995.

[14] M. Niedźwiecki, "Easy recipes for cooperative smoothing," *Automatica*, vol. 46, pp. 716–720, 2010.

[15] M. Niedźwiecki, *Identification of Time-varying Processes*. Wiley, 2001.

[16] M. Niedźwiecki and S. Gackowski "On noncausal weighted least squares identification of nonstationary stochastic systems," *Automatica*, vol. 47, pp. 2239–2245, 2011.

[17] H. Jeffreys, "On invariant form for the prior probability in estimation problems," *Proc. of the Royal Society of London, Series A, Mathematical and Physical Sciences*, vol. 186, pp. 453–461, 1946.

[18] G. Box and G. Tiao *Bayesian Inference in Statistical Analysis*. Addison-Wesley, 1973.

[19] W. Härdle, W. *Applied Nonparametric Regression*. Cambridge University Press, 1990.

[20] V. Katkovnik, K. Egiazarian and J. Astola. *Local Approximation Techniques in Signal and Image Processing*. SPIE Press, 2006.

[21] A.C. Bovik, T. S. Huang and D. C. Munson, "A generalization of median filtering using linear combinations of order statistics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 31, 1342–1350, 1983.

[22] H. Friedl and E. Stampfer, "Cross-validation," in *Encyclopedia of Environmetrics*, A.H. El-Shaarawi & W.W. Piegorsch, Eds., vol. 1, pp. 452–460, Wiley, 2002.

[23] J.A. Hoeting, D. Madigan, A.E. Raftery and C.T. Volinsky, "Bayesian model averaging: A tutorial," *Statistical Science*, vol. 14, pp. 382–417, 1999 [the corrected version of this paper, free of typesetting errors, can be reached from the BMA home page: http://www2.research.att.com/ volinsky/bma.html].

[24] D. Madigan and A.E. Raftery, "Model selection and accounting for model uncertainty in graphical models using Occam's window," *J. Amer. Statist. Assoc.*, vol. 89, pp. 1535–1546, 1973.

[25] J.M. Bates and C.W.J. Granger, "The combination of forecasts," *Operational Research Quarterly*, vol. 20, pp. 451–468, 1969.

[26] G. Verghese and T. Kailath, "A further note on backwards Markovian models," *IEEE Trans. Inf. Theory*, vol. 25, pp. 121–124, 1979.

[27] H.E. Rauch, F. Tung and C.T. Striebel, "Maximum likelihood estimates of linear dynamic systems," *AIAA J.*, vol. 3, pp. 1445–1450, 1965.

[28] A.E. Bryson and M. Frazier, "Smoothing for linear and nonlinear systems," *Aero Syst. Dvi.*, Wright-Patterson Air Force Base, OH, TDR 63-119, pp. 353–364, 1963.

[29] G.J. Bierman, "Fixed interval smoothing with discrete measurements," *Int. J. Control*, vol. 18, pp. 65–75, 1973.

[30] G.J. Bierman, *Factorization Methods for Discrete Sequential Estimation*. Academic Press, 1977.

[31] P. Park and T. Kailath, "Square-root Bryson-Frazier smoothing algorithms," *IEEE Trans. Automat. Contr.*, vol. 40, pp. 761–766, 1995.

[32] P. Park and T. Kailath, "New square-root smoothing algorithms," *IEEE Trans. Automat. Contr.*, vol. 41, pp. 727–732, 1996.

[33] R.G. Gibbs, "Square root modified Bryson-Frazier smoother," *IEEE Trans. Automat. Contr.*, vol. 56, pp. 452–456, 2011.

[34] M. Niedźwiecki, "Identification of nonstationary stochastic systems using parallel estimation schemes," *IEEE Trans. Automat. Contr.*, vol. 35, pp. 329–334, 1992.

[35] T.M. Silva and V.H. Nascimento, "Improving the tracking capability of adaptive filters via convex combination," *IEEE Trans. Signal Process.*, vol. 15, pp. 3137–3149, 2008.

[36] E.T. Whittaker, "On a new method of graduation," *Proc. Edinborough Math. Soc.*, vol. 41, pp. 63–75, 1923.

[37] R. Shiller, "A distributed lag estimator derived from smoothness priors," *Econometrica*, vol. 41, pp. 775–778, 1973.

[38] H. Akaike, "Likelihood and the Bayes procedure," in *Bayesian Statistics*, J.M. Bernado, M.H. DeGroot, D.V. Lindley & A.F.M. Smith, Eds. Valencia, Spain, University Press, pp. 141–166, 1980.

[39] J.P. Norton, "Optimal smoothing in the identification of linear time-varying systens," *Proc. IEE*, vol. 122, 663–668, 1975.

[40] G. Kitagawa and W. Gersch, "A smoothness priors time-varying AR coefficient modeling of nonstationary covariance time series," *IEEE Trans. Automat. Contr.*, vol. 30, pp. 48–56, 1985.

[41] S.N. Wood *Generalized Additive Models: An Introduction with R. Chapman and Hall*. CRC Press, 2006.

[42] N. Saralees, "A generalized normal distribution," *J. Appl. Stat.*, vol. 32, pp. 685–694, 2005.

[43] T. Eltoft, T. Kim and T-W. Lee, "On the multivariate Laplace distribution," *IEEE Trans. Signal Process.*, vol. 54, pp. 300–303, 2006.

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

[45] T. Söderstrm̈ and P. Stoica *System Identification*. Prentice Hall, 1989.

Maciej Niedźwiecki was born in Poznań, Poland in 1953. He 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 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.