INTERNATIONAL JOURNAL OF ADAPTIVE CONTROL AND SIGNAL PROCESSING
Int. J. Adapt. Control Signal Process. (2011)
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/acs.1270

Parameter tracking with partial forgetting method

K. Dedecius*,†, I. Nagy and M. Kárný

Department of Adaptive Systems, Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, P.O. Box 18, 182 08 Prague 8, Czech Republic

SUMMARY

This paper concerns the Bayesian tracking of slowly varying parameters of a linear stochastic regression model. The modelled and predicted system output is assumed to possess time-varying mean value, whereas its dynamics are relatively stable. The proposed estimation method models the system output mean value by time-varying offset. It formulates three extreme hypotheses on model parameters' variability: (i) no parameter varies; (ii) all parameters vary; and (iii) the offset varies. The Bayesian paradigm then provides a mixture as posterior distribution, which is appropriately projected to a feasible class. Exponential forgetting at 'second' hypotheses level allows tracking of slow variations of respective hypotheses.

The developed technique is an example of a general procedure called partial forgetting. Focus on a simple example allows to demonstrate essence of the approach. Moreover, it is important per se as it corresponds with a varying load of otherwise (almost) time-invariant dynamic system. Copyright © 2011 John Wiley & Sons, Ltd.

Received 23 November 2010; Revised 12 June 2011; Accepted 15 June 2011

KEY WORDS: regression models; model; parameter estimation; parameter tracking

1. INTRODUCTION

The modelling of dynamical systems is mostly complemented by estimation of their parameters based on observed data [1]. Often, the parameters (slowly) vary and have to be tracked. The conceptually correct Bayesian solutions explicitly model the parameters' variations and convert tracking into filtering. Specific variation models and evaluation techniques lead to the Kalman filter [2] and its modifications, such as H_∞ filter, extended Kalman filter, particle filtering [3], Markov Chain Monte Carlo based techniques [4] and so on. Filtering is naturally extended to smoothing [5]. Some alternative approaches comprise, for example, employment of reproducing kernel Hilbert spaces [6]. Various properties of filtering have been widely studied; papers [7–10] represent just samples of this rich area.

Still, there is a lot of cases in which detailed models of parameter variations are unavailable and they are substituted in various ways. An unknown-but-bounded types technique, for example [11], represents an intermediate stage between a detailed stochastic modelling and a group of tracking techniques still dominated by exponential forgetting (EF). The EF is viewed as time-weighted recursive least squares (RLS) [12] or as Bayesian flattening of the posterior distribution [13]. The Bayesian linear forgetting is its dual version [14]. Their basic variants suffer various drawbacks. For instance, EF method has a blow-up tendency: the gain of the estimation algorithm could grow without bounds for nonpersistently exciting signals [15]. This leads to many modifications, for example, RLS with constant forgetting factor in EF is modified to keep the covariance matrix bounded [16].

*Correspondence to: Kamil Dedecius, Department of Adaptive Systems, Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, P.O. Box 18, 182 08 Prague 8, Czech Republic.
†E-mail: dedecius@utia.cas.cz

Directional [17, 18], variable forgetting [19] and variable memory-length parameter estimation [20] represent various successful techniques preventing the EF blow-up both at RLS and Bayesian level.

Parameter entries may change at different rates. This calls for a sort of vector forgetting. For instance, the variable forgetting [19] has its vector counterpart in [21]. Surprisingly, this RLS related technique has no Bayesian counterpart. Such counterpart is highly desirable as it can be immediately combined with the consistent and versatile Bayesian framework, which allows, for instance, the inclusion of prior information [22], systematic model selection [13, 23], model averaging etc. [24]. Even more importantly, general Bayesian solutions are not confined to linear-in-parameters Gaussian autoregressive, regressive and possibly controlled models (briefly referred as regression models) for which they are algorithmically equivalent to RLS. For instance, the Bayesian estimation with EF and its extensions are immediately applicable to exponential family [25] that among others includes controlled Markov chains with unknown transition probabilities [26]. Besides the generality, a Bayesian algorithm being supplied with a valuable prior information is advantaged to the nonBayesian ones. However, if the prior is noninformative, the solutions are often identical, compare, for example, the EF [12] and the Bayesian view on it in [13]. This indicates importance of filling the discussed gap.

The proposed solution is based on the Bayesian treatment of appropriate hypotheses on parameter variations with the Bayesian EF used at the second hypotheses level. The basic steps are presented at general level independent of a particular form of parametric model. Then, they are elaborated for one practically important case of SISO regression model. Parameters describing its basic dynamics are assumed to vary slowly at most, but its offset may change substantially faster. The elaboration (i) is useful per se as the considered case models well the changes of the modelled system load (like the road traffic intensity); (ii) illustrates the general theory; and (iii) indicates relatively low computational demands of the proposed technique.

The organization of the paper is as follows: in Section 2, we describe the Bayesian parameter estimation methodology; in Section 3, the general idea of partial forgetting is presented; Section 4 applies the results to the popular Gaussian regression model; in Section 5, we provide illustrative experiments using artificial and real traffic data. Finally, Section 6 presents the conclusions.

Throughout: $'$ denotes transposition, $\equiv$ is defining equality, $\propto$ means proportionality. $x^\star \subset \mathbb{R}^{\dim(x)}$ is a set of x values. Most variables are viewed as random, thus, for the sake of better readability, we adopt the following conventions: The vectors $\boldsymbol{d}_t$ include observations at time t. $f(x)$ is a PDF of a (multivariate) random variable distinguished by the argument x. The conditional PDF $f_{t|\tau}(x|z) \equiv f(x_t = x | z_t = z, \boldsymbol{d}(\tau))$, where $t \geqslant \tau \in \{1, 2, 3, \ldots\}$ label discrete time instants and $\boldsymbol{d}(\tau) \equiv \{\text{prior knowledge}, \boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_\tau\}$. The variables $x, z \in \{y, u, \boldsymbol{\Theta}, \boldsymbol{\psi}, \ldots\}$. For instance, the notation $f_{t|t-1}(y|\boldsymbol{\psi}, \boldsymbol{\Theta})$ is identical to $f(y_t|\boldsymbol{\psi}_t, \boldsymbol{\Theta}_t, \boldsymbol{d}(t-1))$. For the sake of convenience, we do not distinguish among random variable, its realization and a corresponding argument of a PDF.

2. BAYESIAN PARAMETER ESTIMATION

Consider a stochastic SISO system observed at discrete time instants $t = 1, 2, \ldots$. Its directly manipulated input $u_t \in u^\star \subset \mathbb{R}$ affects the output $y_t \in y^\star \subset \mathbb{R}$. The input–output pairs $\boldsymbol{d}_t = [u_t, y_t]'$ observed at each time instant t form the data record.

Dependence of the output y_t on the current input u_t and previous data $\boldsymbol{d}(t-1)$ is modelled by the conditional PDF

$$f_{t|t-1}(y|u, \boldsymbol{\Theta}) = f\left(y_t|u_t, \boldsymbol{\Theta}_t, \boldsymbol{d}_0, \ldots, \boldsymbol{d}_{t-1}\right) = f_{t|t-1}(y|\boldsymbol{\psi}, \boldsymbol{\Theta}) \tag{1}$$

where $\boldsymbol{\Theta} = [\Theta_1, \ldots, \Theta_N]$ is the value of an unknown multivariate model parameter $\boldsymbol{\Theta}_t$ and $\boldsymbol{\psi}$ is the value of a fixed length regression vector $\boldsymbol{\psi}_t$, fully determined by data u_t, $\boldsymbol{d}(t-1)$ influencing the output y_t. The prior knowledge is formed by initial data $\boldsymbol{d}(0)$ that is either chosen by an expert or observed before estimation.

The Bayesian approach treats the unknown model parameter $\boldsymbol{\Theta}_t$ as a random variable and describes it by a conditional PDF $f_{t|t-1}(\boldsymbol{\Theta})$. The Bayes' rule updates this PDF by a new data

DOI: 10.1002/acs

record $\boldsymbol{d}_t$ as follows [13],

$$f_{t|t}(\boldsymbol{\Theta}) \propto f_{t|t-1}(y|\boldsymbol{\psi}, \boldsymbol{\Theta}) f_{t|t-1}(\boldsymbol{\Theta}). \tag{2}$$

If the parameter is supposed to vary, the posterior PDF $f_{t|t}(\boldsymbol{\Theta})$ in (2) is to be updated to $f_{t+1|t}(\boldsymbol{\Theta})$ to complete the recursion. If not, $f_{t+1|t}(\boldsymbol{\Theta})$ is identical to $f_{t|t}(\boldsymbol{\Theta})$. Stochastic filtering [13] represents an exact solution of this time-update. It requires, however, an evolution model $\boldsymbol{\Theta}_t \rightarrow \boldsymbol{\Theta}_{t+1}$, which is often unknown or too complex to be treated. Such a situation calls for heuristic methods labelled as forgetting. The current paper contributes to this important field of interest. It follows the tracking scheme discussed in the Introduction, and in Section 3 it proposes a way to perform this update without an explicit evolution model, assuming that parameter entries vary slowly but at different rates.

The predictive PDF $f_{t+1|t}(y|\boldsymbol{\psi})$ provides the Bayesian output prediction. It employs the regression vector $\boldsymbol{\psi}_{t+1}$ depending on the considered input u_{t+1} and observed data $\boldsymbol{d}(t)$. It holds

$$f_{t+1|t}(y|\boldsymbol{\psi}) = \int_{\boldsymbol{\Theta}^\star} f_{t+1|t}(y|\boldsymbol{\psi}, \boldsymbol{\Theta}) f_{t|t}(\boldsymbol{\Theta}) \mathrm{d}\boldsymbol{\Theta}. \tag{3}$$

Note that point or interval estimates of $\boldsymbol{\Theta}_t$ and y_{t+1} are suitable characteristics of the PDFs $f_{t|t}(\boldsymbol{\Theta})$ and $f_{t+1|t}(y|\boldsymbol{\psi})$, respectively. The set $\boldsymbol{\Theta}^*$ is time-independent.

3. PARTIAL FORGETTING

Let us now be concern with time-varying parameters and assume that they vary with different rates. The way around the problem of ignorance of an explicit evolution model is to employ a heuristic forgetting method, evaluating the best available approximation of the posterior parameter PDF.

Although the desired updated posterior PDF $f_{t+1|t}(\boldsymbol{\Theta})$ is unknown, the simple structure of the problem allows us to characterize its expectations $\mathbb{E}$ within a small set of hypotheses $\{\mathrm{H}_{i;t}\}$. With respect to the announced application, three hypotheses are considered:

$$\begin{aligned}
\mathrm{H}_{0;t} &: \mathbb{E}\left[f_{t+1|t}(\boldsymbol{\Theta}) \middle| \boldsymbol{\Theta}, \mathrm{H}_{0;t}, \boldsymbol{d}(t)\right] = f_{t|t}(\boldsymbol{\Theta})\\
\mathrm{H}_{1;t} &: \mathbb{E}\left[f_{t+1|t}(\boldsymbol{\Theta}) \middle| \boldsymbol{\Theta}, \mathrm{H}_{1;t}, \boldsymbol{d}(t)\right] = g_{t|t}(\boldsymbol{\Theta})\\
\mathrm{H}_{2;t} &: \mathbb{E}\left[f_{t+1|t}(\boldsymbol{\Theta}) \middle| \boldsymbol{\Theta}, \mathrm{H}_{2;t}, \boldsymbol{d}(t)\right] = f_{t|t}\left(\Theta_2, \ldots, \Theta_N \middle| \Theta_1\right)\ h_{t|t}\left(\Theta_1\right)
\end{aligned} \tag{4}$$

Each of the hypotheses reflects our expectation on the true posterior PDF $f_{t+1|t}(\boldsymbol{\Theta})$ at each point $\boldsymbol{\Theta}$ under the knowledge of data $\boldsymbol{d}(t)$ and validity of the current hypothesis $\mathrm{H}_{i;t}$, $i = 0, 1, 2$. $\mathrm{H}_{0;t}$ corresponds to the expectation that no time evolution is needed. $\mathrm{H}_{1;t}$ expects that all parameter entries will evolve and should be described by an externally supplied alternative PDF $g_{t|t}(\boldsymbol{\Theta})$, for instance, a flattened version of $f_{t|t}(\boldsymbol{\Theta})$ (any other standard forgetting method can be used for its construction). $\mathrm{H}_{2;t}$ corresponds to the case that just a single parameter entry (the offset) is expected to change and may be described by a suitable externally supplied PDF $h_{t|t}(\Theta_1)$, whereas the conditional part $f_{t|t}(\Theta_2, \ldots, \Theta_N|\Theta_1)$ remains identical to that one in $\mathrm{H}_{0;t}$. The marginal part can be gained, for example, by flattening, or even the marginal of the prior PDF $f_{1|0}(\Theta_1)$ can be used. The other entries are expected to remain unchanged.

The number of similar hypotheses can be much larger. For instance, it is possible to formulate many hypotheses and to reduce their number by elimination of those that proved not to be valid through a longer modelling period. Another option is to set only a few hypotheses and to rely on the algorithm that it will do the best decision; that is, weights tunning. This is indeed a trade-off between the universality and the speed of computation. Anyway, it is important that even a small number of well formulated hypotheses will perform acceptably in most practical situations, because the Bayesian testing algorithm chooses the most probable information being at disposal. The ideal solution to this issue would consist of automated generating of hypotheses, which is, however,

DOI: 10.1002/acs

nontrivial. Although this issue was already given some effort, for example, [13, 26, 29], the remaining constraints still prevent its use in the presented method. Monte-Carlo type generating is possible as well, but has not been sufficiently elaborated yet.

The special choice (4) is the only sacrifice with respect to generality of the partial-forgetting description. It fits the application discussed in Section 4.

3.1. *Evolution of hypotheses probabilities*

Each of the hypotheses is assigned prior probability $\lambda_{i;t|t-1}$ of becoming true at the particular time instant

$$\lambda_{i;t|t-1} = \mathsf{Probability}_{t|t-1}(H_i), \qquad \sum_{i=0}^{2} \lambda_{i;t|t-1} = 1. \tag{5}$$

Bayes' rule provides data update of these probabilities

$$\lambda_{i;t|t} = \mathsf{Probability}_{t|t}(H_i) \propto \lambda_{i;t|t-1} \mathfrak{F}_{i;t|t-1}, \tag{6}$$

where

$$\mathfrak{F}_{i;t|t-1} = \int_{\mathbf{\Theta}^*} f_{t|t-1}(y|\boldsymbol{\psi}, \mathbf{\Theta}) \mathbb{E}\left[f_{t|t-1}(\mathbf{\Theta}) \middle| \mathbf{\Theta}, H_{i;t}, \boldsymbol{d}(t) \right] \mathsf{d}\mathbf{\Theta}.$$

Notice that each probability is updated by the predictive PDF (3) relevant to the induced hypothetic PDF from (4).

The exact time update of $\lambda_{i;t|t} \rightarrow \lambda_{i;t+1|t}$ would require a model of the time evolution $\mathrm{H}_{i;t} \rightarrow \mathrm{H}_{i;t+1}$ of respective hypotheses. It is convenient to address this tracking problem at hypotheses level by a sort of forgetting. Bayesian flattening yields

$$\lambda_{i;t+1|t} \propto \lambda_{i;t|t}^{\alpha}, \tag{7}$$

with the 'second level' forgetting factor $\alpha \in (0, 1)$ that can be interpreted as probability that the hypothesis validity does not change with time. Experience shows that $\alpha \geqslant 0.95$ is suitable (c.f. [13]).

The next section demonstrates that the probabilities $\lambda_{i;t+1|t}$ weight the estimation results obtained within respective hypotheses. Their evolution (6), (7) is robustly driven by the optional α. The initial values $\lambda_{i;1|0}$ can be chosen by an expert or selected randomly within the appropriate probabilistic simplex.

3.2. *Approximate estimation of $f_{t+1|t}(\mathbf{\Theta})$*

It remains to construct a 'point' estimate of the unknown $f_{t+1|t}(\mathbf{\Theta})$ and use it in a subsequent recursive estimation. Expected value $\mathbb{E}\left[f_{t+1|t}(\mathbf{\Theta}) \middle| \mathbf{\Theta}, \boldsymbol{d}(t) \right]$ offers itself as such estimate. It is a convex combination of the PDFs induced by hypotheses and weighted by their probabilities

$$\begin{aligned} \mathbb{E}\left[f_{t+1|t}(\mathbf{\Theta}) \middle| \mathbf{\Theta}, \boldsymbol{d}(t) \right] &= \mathbb{E}\left[\mathbb{E}\left[f_{t+1|t}(\mathbf{\Theta}) \middle| \mathbf{\Theta}, \boldsymbol{d}(t), H_i \right] \middle| \mathbf{\Theta}, \boldsymbol{d}(t) \right] \\ &= \sum_{i=0}^{2} \lambda_{i;t+1|t} \mathbb{E}\left[f_{t+1|}(\mathbf{\Theta}) \middle| \mathbf{\Theta}, \boldsymbol{d}(t), H_i \right]. \end{aligned} \tag{8}$$

Though the mixture (8) expresses our expectation about the unknown PDF $f_{t+1|t}(\mathbf{\Theta})$, its use for modelling in (1) is impractical. Instead, we approximate it by the PDF $\tilde{f}_{t+1|t}(\mathbf{\Theta})$ from a set $f_{t+1|t}^{\star}$ of feasible PDFs, whose divergence from the original mixture is the smallest one. The paper [27] provides arguments for selecting the Kullback–Leibler (KL) divergence [28] defined as follows.

Definition 1 (Kullback–Leibler divergence)
Let f and g be two PDFs of a random variable X, acting on a common set $x^{\star}$. The Kullback–Leibler divergence is defined

$$\mathfrak{D}(f||g) = \int_{x^\star} f(x) \ln \frac{f(x)}{g(x)} \mathrm{d}x. \tag{9}$$

It can be shown that $\mathfrak{D}(f||g) \geqslant 0$ with equality for $f(x) = g(x)$ almost everywhere on $x^\star$.

Note that the minimization of the KL divergence of the expected PDF (8) on $\tilde{f}_{t+1|t}(\boldsymbol{\Theta})$ is equivalent to the minimization of the expected KL divergence $\mathbb{E}\left[\mathfrak{D}\left(f_{t+1|t}||\tilde{f}_{t+1|t}\right)\right]$ of unknown $f_{t+1|t}$ on $\tilde{f}_{t+1|t}$

$$\tilde{f}_{t+1|t}(\boldsymbol{\Theta}) = \underset{f \in f^\star_{t+1|t}}{\arg\min}\ \mathbb{E}\left[\mathfrak{D}\left(f_{t+1|t}||f\right)\Big|\boldsymbol{\Theta}, \boldsymbol{d}(t)\right], \tag{10}$$

hence, we use (8) in place of $f_{t+1|t}$. The resulting best approximation $f_{t+1|t}(\boldsymbol{\Theta}) \equiv \tilde{f}_{t+1|t}(\boldsymbol{\Theta})$, found within the set of feasible PDFs $f^\star_{t+1|t}$, is used as the prior PDF for the subsequent data-update and parameter tracking steps.

4. APPLICATION TO GAUSSIAN REGRESSION MODEL

This section demonstrates an application of the developed method to the popular Gaussian regression model. First, the standard theory of the Bayesian regression with Gaussian model and its conjugate prior Gauss-inverse-Wishart distribution is recalled in a necessary detail. On the base of this theory, summarized, for example, in [26], we derive an approximation of the unknown PDF $f_{t+1|t}(\boldsymbol{\Theta})$. The abstract formula (10) gets its application counterpart. The data update of probabilities of hypotheses follows.

To derive the partial forgetting method for the Gaussian regression model, let us follow the way it is described in Section 3. Suppose that a SISO system can be modelled by a regression model, whose output y_t is influenced by inputs $u_t, \dots, u_{t-q+1}$, previous outputs $y_{t-1}, \dots, y_{t-p}$ and offset, forming a regression vector $\boldsymbol{\psi}_t \in \mathbb{R}^n$, where $n = p + q + 1$; $p, q \in \mathbb{N}^0$. Then, the output is characterized by a Gaussian PDF with the mean $\boldsymbol{\psi}'_t\boldsymbol{\theta}_t$ and variance r,

$$f_{t|t-1}(y|\boldsymbol{\psi}, \boldsymbol{\Theta}) = \mathcal{N}(\boldsymbol{\psi}'_t\boldsymbol{\theta}_t, r) \equiv \frac{1}{\sqrt{2\pi r}} \exp\left\{\frac{\left(y_t - \boldsymbol{\psi}'_t\boldsymbol{\theta}_t\right)^2}{2r}\right\}, \qquad \boldsymbol{\Theta} = \{\boldsymbol{\theta}, r\} \tag{11}$$

where $\boldsymbol{\theta}_t$ is a real n-vector of regression coefficients. The number of unknown parameters $N = n + 1$. The recursive Bayesian estimation of (11) employs the Gauss-inverse-Wishart PDF, representing a conjugate (i.e., reproducing) prior PDF suitable for estimation of model parameters [13]. An exhaustive description of $\mathcal{G}i\mathcal{W}$ PDF and its use can be found in [26], pp. 251–260.

Definition 2 (Gauss-inverse-Wishart PDF)
The Gauss-inverse-Wishart PDF has the form

$$\mathcal{G}i\mathcal{W}(\boldsymbol{V}, \nu) \equiv \frac{r^{-0.5(\nu+n+2)}}{\mathcal{I}(\boldsymbol{V}, \nu)} \exp\left\{\frac{-1}{2r}\begin{bmatrix}-1\\ \theta\end{bmatrix}' V \begin{bmatrix}-1\\ \theta\end{bmatrix}\right\}$$

or in terms of the decomposition $\boldsymbol{V} = \boldsymbol{L}'\boldsymbol{D}\boldsymbol{L}$, where $\boldsymbol{L}$ is a unit lower triangular matrix and $\boldsymbol{D}$ is a diagonal matrix

$$\mathcal{G}i\mathcal{W}(\boldsymbol{L}, \boldsymbol{D}, \nu) \equiv \frac{r^{-0.5(\nu+n+2)}}{\mathcal{I}(\boldsymbol{L}, \boldsymbol{D}, \nu)} \exp\left\{\frac{-1}{2r}\left[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\boldsymbol{C}^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + D_y\right]\right\}.$$

The individual terms have the following meaning:

$\boldsymbol{V} \in \mathbb{R}^{N \times N}$, $N = n + 1$, is the extended information matrix, that is, symmetric square positive definite matrix, which carries the information about the past data,
$\nu \in \mathbb{R}^+$ stands for the degrees of freedom,
n denotes the length of the regression vector $\boldsymbol{\psi}$,
$r \in \mathbb{R}^+$ is the noise variance.

DOI: 10.1002/acs

With the split of $\boldsymbol{L}$ and $\boldsymbol{D}$ to block matrices of corresponding dimensions (D_y scalar)

$$\boldsymbol{L} = \begin{bmatrix} 1 & \\ \boldsymbol{L}_{y\psi} & \boldsymbol{L}_{\psi} \end{bmatrix}, \quad \boldsymbol{D} = \begin{bmatrix} D_y & \\ & \boldsymbol{D}_{\psi} \end{bmatrix}$$

$\hat{\boldsymbol{\theta}} \equiv \boldsymbol{L}_{\psi}^{-1} \boldsymbol{L}_{y\psi}$ is the least-squares (LS) estimate of $\boldsymbol{\theta}$,
$\boldsymbol{C} \equiv \boldsymbol{L}_{\psi}^{-1} \boldsymbol{D}_{\psi}^{-1} (\boldsymbol{L}_{\psi}^{-1})' \in \mathbb{R}^{n \times n}$ is the LS covariance of $\hat{\boldsymbol{\theta}}$,
$D_y \in \mathbb{R}^{+}$ is the LS remainder,
$\mathcal{I}$ stands for the normalization integral

$$\mathcal{I}(\boldsymbol{L}, \boldsymbol{D}, \nu) \equiv \Gamma(0.5\nu) \sqrt{\frac{2^{\nu}(2\pi)^{n}}{D_y^{\nu} |\boldsymbol{D}_{\psi}|}}. \tag{12}$$

As proved in [26], the used $\boldsymbol{L}'\boldsymbol{D}\boldsymbol{L}$ decomposition allows simple evaluation of marginal PDF of $\mathcal{G}i\mathcal{W}$ PDF.

Theorem 1 (Marginal PDF of $\mathcal{G}i\mathcal{W}$ PDF)
Given a PDF $f(\boldsymbol{\Theta}) = f(\boldsymbol{\theta}_{\alpha}, \boldsymbol{\theta}_{\beta}, r) = \mathcal{G}i\mathcal{W}(\boldsymbol{V}, \nu)$. Let $\boldsymbol{L}'\boldsymbol{D}\boldsymbol{L}$ be the corresponding decomposition of the extended information matrix $\boldsymbol{V}$ of its PDF as follows:

$$\boldsymbol{L} \equiv \begin{bmatrix} 1 & & \\ \boldsymbol{L}_{y\alpha} & \boldsymbol{L}_{\alpha} & \\ \boldsymbol{L}_{y\beta} & \boldsymbol{L}_{\alpha\beta} & \boldsymbol{L}_{\beta} \end{bmatrix}, \quad \boldsymbol{D} \equiv \begin{bmatrix} D_y & & \\ & \boldsymbol{D}_{\alpha} & \\ & & \boldsymbol{D}_{\beta} \end{bmatrix}. \tag{13}$$

Then, the marginal PDF for (θ_{α}, r) can be extracted:

$$f(\boldsymbol{\theta}_{\alpha}, r) \equiv \mathcal{G}i\mathcal{W}\left(\begin{bmatrix} 1 & \\ \boldsymbol{L}_{y\alpha} & \boldsymbol{L}_{\alpha} \end{bmatrix}, \begin{bmatrix} D_y & \\ & \boldsymbol{D}_{\alpha} \end{bmatrix}, \nu \right). \tag{14}$$

4.1. Parameter estimation

Parameter estimation in the Gaussian regression model, as given in (2), evolves the statistics $\boldsymbol{V}$ and ν by the data update

$$\boldsymbol{V}_{t|t} = \boldsymbol{V}_{t|t-1} + \boldsymbol{\Psi}_t \boldsymbol{\Psi}_t', \qquad \nu_{t|t} = \nu_{t|t-1} + 1, \tag{15}$$

where $\boldsymbol{\Psi}_t \equiv [y_t, \boldsymbol{\psi}_t']'$ is an extended regression vector. The recursion (15) expressed in terms of the LS representation coincides with the RLS [13].

4.2. Partial forgetting

In this part of the paper, we apply the theory described in Section 3 on the Gaussian regression model (11) with the $\mathcal{G}i\mathcal{W}$ conjugate prior PDF. Recall from (4) that we need to specify three hypotheses about parameter variability. Here, the PDFs induced by the hypotheses (4) will differ in terms of statistics $\boldsymbol{V}$ and ν:

$$\begin{aligned} \mathrm{H}_{0;t} &: \mathbb{E}\left[f_{t+1|t}(\boldsymbol{\Theta}) \middle| \boldsymbol{\Theta}, \boldsymbol{d}(t), \mathrm{H}_{0;t} \right] = \mathcal{G}i\mathcal{W}\left(\boldsymbol{V}_{t|t}^{f}, \nu_{t|t}^{f} \right) = \mathfrak{G}_{0;t|t} \\ \mathrm{H}_{1;t} &: \mathbb{E}\left[f_{t+1|t}(\boldsymbol{\Theta}) \middle| \boldsymbol{\Theta}, \boldsymbol{d}(t), \mathrm{H}_{1;t} \right] = \mathcal{G}i\mathcal{W}\left(\boldsymbol{V}_{t|t}^{g}, \nu_{t|t}^{g} \right) = \mathfrak{G}_{1;t|t} \\ \mathrm{H}_{2;t} &: \mathbb{E}\left[f_{t+1|t}(\boldsymbol{\Theta}) \middle| \boldsymbol{\Theta}, \boldsymbol{d}(t), \mathrm{H}_{2;t} \right] = \mathcal{G}i\mathcal{W}\left(\boldsymbol{V}_{t|t}^{f|h}, \nu_{t|t}^{h} \right) = \mathfrak{G}_{2;t|t} \end{aligned} \tag{16}$$

where to stress connection with (4), the upper indices denote the correspondence to particular PDFs f, g, h. The hypothesis $\mathrm{H}_{2;t}$ in (4) assumes the exchange of marginal PDF of related parameter

DOI: 10.1002/acs

with a suitable alternative h. For the Gaussian regression model, we may proceed with the help of Theorem 1. It allows us to select the statistics of the marginal PDF to be exchanged with alternative ones, for example, from $\mathrm{H}_{1;t}$.

Suppose, that we are given the probabilities $\lambda_{i;t+1|t}$ for $i = 0, 1, 2$ according to (5). Then, we can express the expectation of the true PDF $f_{t+1|t}(\boldsymbol{\Theta})$ as the mixture (8),

$$\mathbb{E}\left[f_{t+1|t}(\boldsymbol{\Theta}) \middle| \boldsymbol{\Theta}, \boldsymbol{d}(t) \right] = \sum_{i=0}^{2} \lambda_{i;t+1|t} \mathfrak{G}_{i;t|t}. \tag{17}$$

According to Section 3, the mixture (17) has to be approximated by a single PDF from the same family. In order to do this, we use the Kullback–Leibler divergence of $\mathcal{G}i\mathcal{W}$ PDFs as follows:

Theorem 2 (KL divergence of two $\mathcal{G}i\mathcal{W}$ PDFs)
Given two $\mathcal{G}i\mathcal{W}$ distributions with PDFs f and $\tilde{f}$. The Kullback–Leibler divergence of these two functions has the following form

$$\begin{aligned}
\mathfrak{D}\left(f \middle|\middle| \tilde{f}\right) = &\ln \frac{\Gamma(0.5\tilde{\nu})}{\Gamma(0.5\nu)} - 0.5 \ln |\boldsymbol{C}\tilde{\boldsymbol{C}}^{-1}| + 0.5\tilde{\nu} \ln \frac{D_y}{\tilde{D}_y} \\
&+ 0.5(\nu - \tilde{\nu})\Upsilon(0.5\nu) - 0.5n - 0.5\nu + 0.5\mathsf{Tr}\left(\boldsymbol{C}\tilde{\boldsymbol{C}}^{-1}\right) \\
&+ 0.5\frac{\nu}{D_y}\left[\left(\hat{\boldsymbol{\theta}} - \hat{\tilde{\boldsymbol{\theta}}}\right)' \tilde{\boldsymbol{C}}^{-1} \left(\hat{\boldsymbol{\theta}} - \hat{\tilde{\boldsymbol{\theta}}}\right) + \tilde{D}_y\right],
\end{aligned} \tag{18}$$

where $\Upsilon(\cdot)$ denotes the digamma function, that is, the first logarithmic derivative of the gamma function $\Gamma(\cdot)$.

The proof is nontrivial and is given in [26].

In our case, we substitute the expectation of $f_{t+1|t}$ for f and search for its best approximation $\tilde{f}_{t+1|t}$ that minimizes the expected KL divergence. For this purpose, we take its derivatives with respect to $\hat{\tilde{\boldsymbol{\theta}}}, \tilde{\boldsymbol{C}}, \tilde{D}_y$ and $\tilde{\nu}$. The result of the minimization is summarized in the following proposition.

Proposition 1
Given a convex combination (mixture) of Gauss-inverse-Wishart PDFs (17). Its best approximation minimizing the Kullback–Leibler divergence within the set of the $\mathcal{G}i\mathcal{W}$ distributions is given by the following parameters (statistics):

$$\begin{aligned}
\hat{\tilde{\boldsymbol{\theta}}}_{t+1|t} &= \left(\sum_{i=0}^{2} \lambda_{i;t+1|t} \frac{\nu_{i;t|t}}{D_{yi;t|t}}\right)^{-1} \left(\sum_{i=0}^{2} \lambda_{i;t+1|t} \frac{\nu_{i;t|t}}{D_{yi;t|t}} \hat{\boldsymbol{\theta}}_{i;t|t}\right) \\
\tilde{D}_{y;t+1|t} &= \tilde{\nu}_{i;t+1|t} \left(\sum_{i=0}^{2} \lambda_{i;t+1|t} \frac{\nu_{i;t|t}}{D_{yi;t|t}}\right)^{-1} \\
\tilde{\boldsymbol{C}}_{t+1|t} &= \sum_{i=0}^{2} \lambda_{i;t+1|t} \left[\frac{\nu_{i;t|t}}{D_{yi;t|t}} \left(\hat{\boldsymbol{\theta}}_{i;t|t} - \hat{\tilde{\boldsymbol{\theta}}}_{i;t+1|t}\right) \left(\hat{\boldsymbol{\theta}}_{i;t|t} - \hat{\tilde{\boldsymbol{\theta}}}_{i;t+1|t}\right)' + \boldsymbol{C}_{i;t|t}\right] \\
\tilde{\nu}_{t+1|t} &= \frac{1 + \sqrt{1 + \frac{4}{3}(A - \ln 2)}}{2(A - \ln 2)} \\
A &= \ln\left(\sum_{i=0}^{2} \lambda_{i;t+1|t} \frac{\nu_{i;t|t}}{D_{yi;t|t}}\right) + \sum_{i=0}^{2} \lambda_{i;t+1|t} \left[\ln D_{yi;t|t} - \Upsilon(0.5\nu_{i;t|t})\right].
\end{aligned}$$

$\Upsilon(\cdot)$ is the digamma function. The index i refers to the i-th hypothesis; that is, to the PDF $\mathfrak{G}_{i;t|t}$.

DOI: 10.1002/acs

Proof
Except on ν, all results are obtained directly using the derivative rules $\frac{\partial}{\partial \boldsymbol{X}} \ln |\boldsymbol{AXB}| = (\boldsymbol{X}^{-1})'$ and $\frac{\partial}{\partial \boldsymbol{X}} \mathsf{Tr}(\boldsymbol{AX}) = \boldsymbol{A}'$. Differentiation of (18) with f being the mixture (8) with respect to $\tilde{\nu}_{t+1|t}$ yields

$$\frac{\partial \mathfrak{D}\left(f_{t+1|t} \middle\| \tilde{f}_{t+1|t}\right)}{\partial \tilde{\nu}_{t+1|t}} = \Upsilon(0.5\tilde{\nu}_{t+1|t}) - \ln \tilde{\nu}_{t+1|t} + A = 0,$$

where A is given above. The digamma function $\Upsilon(0.5\tilde{\nu}_{t+1|t})$ is approximated [30]

$$\Upsilon(x) = \ln x - \frac{1}{2x} - \frac{1}{12x^2} + \underbrace{O\left(\frac{1}{x^4}\right)}_{\to 0},$$

hence, for $x = 0.5\tilde{\nu}_{t+1|t}$ and back substitution, we get a quadratic equation, which has the claimed form. □

4.3. Data update of hypotheses probabilities

In Section 3.1, we introduced a method for tuning the weights of Hypotheses (4). The data update (6) consists of recursive update of probabilities (5) by the predictive PDFs. Expressing the predictive PDF in term of a ratio of normalization integrals (12) allows us to specify the corresponding rule for the Gaussian regression model; that is,

$$\lambda_{i;t|t} \propto \lambda_{i;t|t-1} \frac{\mathcal{I}_{i;t|t}(\boldsymbol{L}, \boldsymbol{D}, \nu)}{\mathcal{I}_{i;t|t-1}(\boldsymbol{L}, \boldsymbol{D}, \nu)}, \tag{19}$$

where $\mathcal{I}(\cdot)$ is the normalization integral (12).

Algorithm 1: Summary of estimation with partial forgetting

```
Initialization:
Set prior statistics $V_{1|0}$ and $\nu_{1|0}$ of $\mathcal{G}i\mathcal{W}$ pdf $f_{1|0}$.
Set hypotheses $\{H_{i,0}\}$ and weights $\lambda_{i,1|0}, i = \{0, 1, 2\}$.
Set second-level forgetting factor $\alpha$.
Online steps:
for $t = 1, 2, \ldots$ do
    Input: $d_t = [y_t, \psi_t']'$
    Calculate data update of statistics $V_{t|t-1} \to V_{t|t}$ and $\nu_{t|t-1} \to \nu_{t|t}$ – Eq. (15)
    if Evaluate point prediction is true then
        Input: $\psi_{t+1}$
        return $\hat{y}_{t+1} = \psi_{t+1}'\hat{\theta}_{t|t}$
    end
    forall $H_{i,t},\ i = \{0, 1, 2\}$ do
        Set pdf $\mathfrak{G}_{i;t|t}$ – Eq. (16)
        Calculate data update of probability $\lambda_{i,t|t-1} \to \lambda_{i,t|t}$ – Eq. (19)
        Calculate forgetting $\lambda_{i,t|t} \to \lambda_{i,t+1|t}$ – Eq. (7)
    end
    Calculate statistics $\hat{\tilde{\theta}}_{t+1|t}, \tilde{D}_{y;t+1|t}, \tilde{C}_{t+1|t}, \tilde{\nu}_{t+1|t}$ of approximate pdf $\tilde{f}_{t+1|t}$ – Proposition 1.
    Set $f_{t+1|t} \equiv \tilde{f}_{t+1|t}$ for the next time step.
end
```

5. EXAMPLES

In this section, we demonstrate the method on two examples, one being a simulation example and the other one being based on a real traffic data.

5.1. *Simulation example*

This example shows the ability of the method to track simulated coefficients of a first-order Gaussian autoregressive model with $\boldsymbol{\psi}_t = [y_{t-1}, 1]'$. The data was generated with $y_0 = 0.5$ and coefficients $\theta_{1,t} = 0.95$ for the whole time run and $\theta_{2,t} = 1.5$ for $t = 1, \dots, 100$ and $\theta_{2,t} = 0.5 + x_t^3$ for $t = 101, \dots, 200$, where x_t are 100 consecutive evenly spaced numbers from the interval $[-1, 0.5]$, respectively. The additive Gaussian white noise had zero mean and for this special purpose, a small standard deviation $\sigma = 0.1$. The data evolution is depicted in Figure 1. The prior information was gathered from the first 20 samples; the following 180 samples were used for estimation. The split is depicted in Figure 1 by a vertical dashed line. The prior probabilities $\lambda_{i,1|0} = 1/3$ for $i = \{0, 1, 2\}$, that is, no hypothesis was preferred by the user. Their flattening factor $\alpha = 0.99$. To demonstrate the feasibility of the method, any use of an externally supplied expert information in hypotheses was avoided. Instead, we flattened the posterior PDF by factor 0.85 and used it as an information for H_1 and a source of the marginal PDF of $\theta_{2,t}$ for H_2.

Figure 2 shows the evolution of point estimates of the regression coefficients. Apparently, after the faster change at $t = 100$, the model lacked any useful information, but the estimation quickly stabilizes and the coefficients are estimated well during their evolution. The mean squared errors were 0.0022 for θ_1 and 0.8815 for θ_2, respectively.

The evolution of probabilities of the three hypotheses is depicted in Figure 3. During the first linear part, the weights of H_1 and H_2 were continually suppressed. After $t = 100$, their probabilities rapidly grew to reflect the change and as the data became almost linear, they slowly decreased.

5.2. *Traffic data example*

This example demonstrates the ability of the method to track a time-varying mean value of a stochastic process. The data batch represents working day traffic intensities measured in Prague, Czech Republic – Figure 4 (thin line).

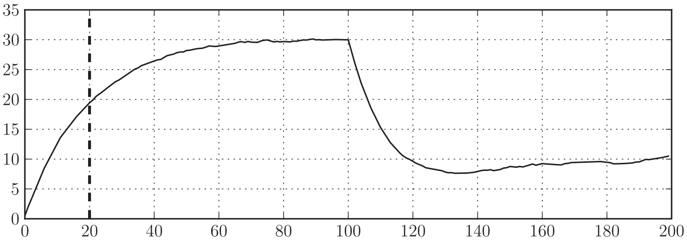

Figure 1. Artificial data series.

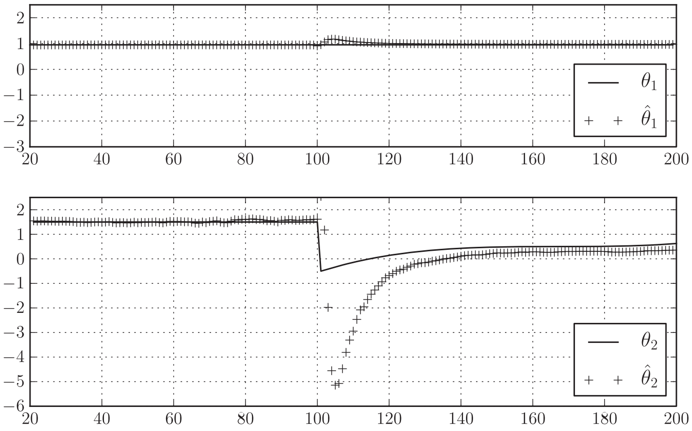

Figure 2. True and estimated coefficients.

DOI: 10.1002/acs

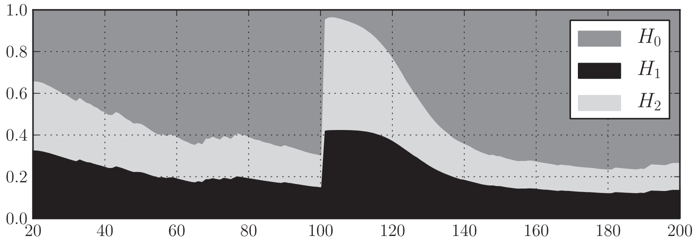

Figure 3. Evolution of probabilities of hypotheses H_i.

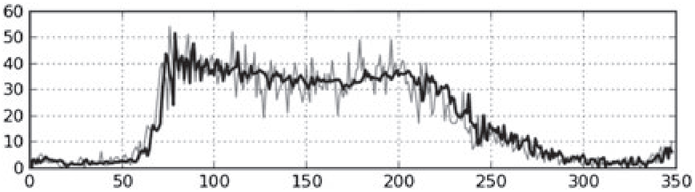

Figure 4. True data (thin line) and predictions (thick line).

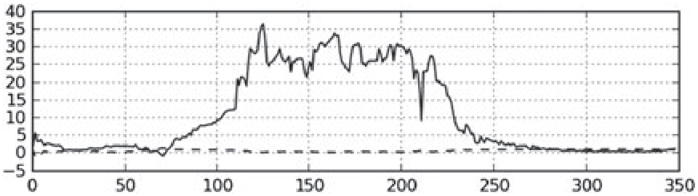

Figure 5. Parameter estimates – dynamics (dashed line) and the offset (solid line).

A first-order autoregression model with $\boldsymbol{\psi}_t = [1, y_{t-1}]'$ was used for a step-ahead prediction of the data. The initial $\mathcal{G}i\mathcal{W}$ statistics were $\boldsymbol{V}_{1|0}$ with diagonal $[0.1, 0.01, 0.01]$ and zeros elsewhere, and $\nu_{1|0} = 5$. The alternative information for H_1 was created each time step by multiplication of both statistics by 0.85, which is identical to EF. An alternative marginal PDF for the offset in H_2 was made the same way. Weights flattening factor $\alpha = 0.99$. Both factors were chosen directly without optimization. The estimation started with uniformly distributed weights $\lambda_{i;1|0} = 1/3$.

The predictions are depicted in Figure 4 (thick line). The model followed the course of the true data quite well. Root mean squared error (RMSE) of prediction was 5.328. To compare, the EF with factor 0.98 led to RMSE $= 6.039$. Except for the noise variance estimation, the GiW regression model is equivalent to RLS. A counterpart of the proposed method is then the vector forgetting [21], providing qualitatively similar results with RMSE $= 5.2937$ using forgetting factors 0.85 for θ_1 (offset) and 0.99 for θ_2. However, compared with the proposed method, this combination was obtained by intensive optimization and the method is strictly connected with RLS.

The evolution of model parameters – the offset θ_1 and dynamics θ_2 – is depicted in Figure 5. Evidently, the model preferred to track the change in the mean value with the offset.

6. CONCLUSION

The problem of estimation of a linear time varying model with different speed of parameter variations has been faced within the parametric Bayesian framework with an approximation minimizing the Kullback–Leibler divergence. It has been shown that a varying mean value of the system output can be tracked with the offset of the model. The presented approach allows to formulate hypotheses about the multivariate parameter variability and recursively tune their probabilities. Furthermore, the user can insert available expert information connected with alternative hypotheses H_1 and H_2

into the estimation process. The risk of discrepancy between the true and the estimated parameter values is suppressed by the recursive weighting as well.

APPENDIX A: SIMULATION EXAMPLE – COMMENTED APPLICATION OF ALGORITHM 1

```
 1 Initialization:
 2 Set prior statistics $\boldsymbol{V}_{1|0}$ and $\nu_{1|0}$ of $\mathcal{G}i\mathcal{W}$ pdf $f_{1|0}$.
   We use the prior pdf with initial matrix with diagonal $[0.1, 0.01, 0.01]$ updated by data vectors
   in the form $d_t = [y_t, y_{t-1}, 1]', t = 1, \cdots, 20$, see (15).
 3 Set hypotheses $\{H_{i,0}\}$ and weights $\lambda_{i,1|0}, i = \{0, 1, 2\}$.
   The scheme for future generating of three hypotheses – Eq. (16) – is set. In the particular case,
   the prior $\mathcal{G}i\mathcal{W}$ pdf stands for $H_{0,t}$, its flattened version obtained by multiplication of its
   statistics by factor 0.85 for $H_{1,t}$ and the prior pdf with flattened marginal of $\theta_2$ selected
   according to Eq. (14) for $H_{2,t}$. The probabilities of hypotheses are set to be equal, i.e., 1/3.
 4 Set second-level forgetting factor $\alpha$.
   We set $\alpha = 0.99$.
 5 Online steps:
 6 for $t = 1, 2, \ldots$ do
       Since the initial 20 data were used as a source of the prior pdf, we start estimation with
       $t = 21$. This is an example-specific case without any loss of generality.
 7     Input: $\boldsymbol{d}_t = [y_t, \boldsymbol{\psi}_t']'$
       In the particular case, we have $\boldsymbol{d}_t = [y_t, y_{t-1}, 1]'$.
 8     Calculate data update of statistics $\boldsymbol{V}_{t|t-1} \to \boldsymbol{V}_{t|t}$ and $\nu_{t|t-1} \to \nu_{t|t}$ – Eq. (15)
       We update the statistics $\mathbf{V}_{t|t-1}$ and $\nu_{t|t-1}$ of the prior pdf using the data vector
       $\boldsymbol{d}_t = [y_t, y_{t-1}, 1]'$ according to (15). The output is the data-updated pdf with statistics $\boldsymbol{V}_{t|t}$
       and $\nu_{t|t}$. This ordinary estimation procedure will be later followed by forgetting.
 9     if Evaluate point prediction is true then
           The prediction is calculated in the ordinary statistical way. In the particular case,
           $\boldsymbol{\psi}_{t+1} = [y_t, 1]'$ and estimated $\hat{\boldsymbol{\theta}}_{t|t}$ follows from Definition 2.
10         Input: $\boldsymbol{\psi}_{t+1}$
11         return $\hat{y}_{t+1} = \boldsymbol{\psi}_{t+1}'\hat{\boldsymbol{\theta}}_{t|t}$
12     end
       The forgetting procedure begins here, when new data $\mathbf{d}_{t+1}$ is available. In the cycle,
       hypotheses (16) about the true pdf will be constructed and tested.
13     forall $H_{i,t},\ i = \{0, 1, 2\}$ do
14         Set pdf $\mathfrak{G}_{i;t|t}$ – Eq. (16)
           The updated pdf (line 8) and the scheme (line 3) is used to set pdf $\mathfrak{G}_{i;t|t}$ of $H_{i,t}$.
15         Calculate data update of probability $\lambda_{i,t|t-1} \to \lambda_{i,t|t}$ – Eq. (19)
           The pdf $\mathfrak{G}_{i;t|t}$ is updated by $\mathbf{d}_{t+1}$ and the ratio of normalization integrals after and
           before this update is calculated. It serves as a factor updating $\lambda_{i,t|t-1}$ to $\lambda_{i,t|t}$, Eq. (19).
16         Calculate forgetting $\lambda_{i,t|t} \to \lambda_{i,t+1|t}$ – Eq. (7)
           The hypotheses probabilities are forgotten by exponentiation by $\alpha = 0.99$, set on line 4.
17     end
       Now, we have three hypotheses $H_{0,t}, H_{1,t}$ and $H_{2,t}$, i.e., three pdfs, and their probabilities
       $\lambda_{0,t+1|t}, \lambda_{1,t+1|t}$ and $\lambda_{2,t+1|t}$. Hence the approximate pdf can be calculated.
18     Calculate statistics $\tilde{\hat{\boldsymbol{\theta}}}_{t+1|t}, \tilde{D}_{y;t+1|t}, \tilde{\boldsymbol{C}}_{t+1|t}, \tilde{\nu}_{t+1|t}$ of approximate pdf $\tilde{f}_{t+1|t}$ – Proposition 1.
       This is a direct use of statistics of the hypothetic pdfs and Proposition 1. The forgetting
       procedure ends here and the ordinary estimation continues with the next time step.
19     Set $f_{t+1|t} \equiv \tilde{f}_{t+1|t}$ for the next time step.
20 end
```

ACKNOWLEDGEMENT

This paper is supported by Grant Agency of the Czech Republic, grant number 102/08/0567.

REFERENCES

1. Söderström T, Stoica P. *System Identification*. Prentice Hall: New York, 1989.
2. Kalman RE, Bucy RS. New results in linear filtering and prediction theory. *Journal Of Basic Engineering* 1961; **83**(1):95–108.
3. Andrieu C, Doucet A, Singh SS, Tadic VB. Particle methods for change detection, system identification, and control. *Proceedings of the IEEE* 2004; **90**(3):423–438.
4. Min C. A Gibbs sampling approach to estimation and prediction of time-varying-parameter models. *Computational Statistics and Data Analysis* 1998; **27**(2):171–194.
5. Young P. Time Variable Parameter Estimation. *Proceedings of the 15th IFAC Symposium on System Identification (SYSID 2009)*, Saint-Malo, France, 2009; 432–437. IFAC PapersOnline, ISSN: 1474–6670.
6. Pillonetto G. Identification of time-varying systems in reproducing kernel hilbert spaces. *IEEE Transactions on Automatic Control* 2008; **53**(9):2202–2209.
7. Guo L. Estimating time-varying parameters by the kalman filter based algorithm: stability and convergence. *IEEE Transactions on Automatic Control* 1990; **35**(2):141–147.
8. Guo L, Ljung L. Performance analysis of general tracking algorithms. *IEEE Transactions on Automatic Control* 1995; **40**(8):1388–1402.
9. Kumar PR. Convergence of adaptive control schemes using least-squares parameter estimates. *IEEE Transactions on Automatic Control* 1990; **35**(4):416–424.
10. Ravikanth R, Meyn SP. Bounds on achievable performance in the identification and adaptive control of time varying systems. *IEEE Transactions on Automatic Control* 1999; **44**(4):670–682.
11. Layton KJ, Weyer E, Campi M. Online Algorithms for the Construction of Guaranteed Confidence Sets for the Parameters of Time-Varying Systems. *Proceedings of the 15th IFAC Symposium on System Identification (SYSID 2009)*, Saint-Malo, France, 2009; 426–431. IFAC PapersOnline, ISSN: 1474–6670.
12. Jazwinski AH. *Stochastic Processes and Filtering Theory*. Academic Press: New York, 1970.
13. Peterka V. Bayesian approach to system identification. In *Trends and Progress in System Identification*, Ekhoff P (ed.). Pergamon Press: Oxford, 1981; 239–304.
14. Kulhavý R, Kraus FJ. On duality of regularized exponential and linear forgetting. *Automatica* 1996; **32**(10): 1403–1415.
15. Milek JT. *Stabilized Adaptive Forgetting in Recursive Parameter Estimation*. Verlag der Fachvereine Hochschulverlag AG an der ETH Zurich: Zurych, 1995.
16. Bittanti S, Campi M. Bounded error identification of time-varying parameters by rls techniques. *IEEE Transactions on Automatic Control* 1994; **39**(5):1106–1110.
17. Kulhavý R, Kárný M. Tracking of slowly varying parameters by directional forgetting. In *Preprints of the 9th IFAC World Congress,* Budapest, Hungary, vol. X, 1984; 78–83.
18. Cao L, Schwartz H. Directional forgetting algorithm based on the decomposition of the information matrix. *Automatica* 2000; **36**(11):1725–1731.
19. Fortescue TR, Kershenbaum LS, Ydstie BE. Implementation of self-tuning regulators with variable forgetting factors. *Automatica* 1981; **17**(6):831–835.
20. Rojas AJ, Goodwin GC, Renton C. Short or Long Memory Estimators? *Proceedings of the 15th IFAC Symposium on System Identification (SYSID 2009)*, Saint-Malo, France, 2009; 1038–1043. IFAC PapersOnline, ISSN: 1474-6670.
21. Saelid S, Foss B. Adaptive Controllers with a Vector Variable Forgetting Factor. *The 22nd IEEE Conference on Decision and Control,* San Antonio, TX. vol. **22**, 1983; 1488–1494.
22. Kadane JB, Dickey JM, Winkler RL, Smith WS, Peters SC. Interactive elicitation of opinions for a normal linear model. *Journal of the American Statistical Association* 1980; **75**(372):845–854.
23. Kárný M, Kulhavý R. Structure determination of regression-type models for adaptive prediction and control. In *Bayesian Analysis of Time Series and Dynamic Models* Chapter 12. Marcel Dekker: New York, 1988; 313–345.
24. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial. *Statistical science* 1999; **14**(4):382–401.
25. Barndorff-Nielsen O. *Information and Exponential Families in Statistical Theory*. Wiley: New York, 1978.
26. Kárný M. *Optimized Bayesian Dynamic Advising*. Springer: London, 2005.
27. Bernardo JM. Expected information as expected utility. *Annals of Statistics* 1979; **7**(3):686–690.
28. Kullback S, Leibler RA. On Information and sufficiency. *Annals of Mathematical Statistics* 1951; **22**(1):79–86.
29. Cox J, Ingemar J. Modeling a dynamic environment using a bayesian multiple hypothesis approach. *Artificial Intelligence* 1994; **66**(2):311–344.
30. Bernardo JM. Algorithm AS 103 Psi (Digamma) Function. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 1976; **25**(3):315–317.

DOI: 10.1002/acs