Preprints of the
15th IFAC Symposium on System Identification
Saint-Malo, France, July 6-8, 2009

Parameter Estimation With Partial Forgetting Method

Kamil Dedecius * **Ivan Nagy** * **Miroslav Kárný** *
Lenka Pavelková *

* *Department of Adaptive Systems, Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Prague, Czech Republic; (e-mail: dedecius@utia.cas.cz)*

Abstract: The paper proposes a new estimating algorithm for linear parameter varying systems with slowly time-varying parameters when the rate of change of individual parameters is different. It introduces a true probability density function, describing ideally the behaviour of parameters. However, as it is unknown, we search for its best approximation. A convex combination of point estimates, defined by individual hypotheses about the true probability density function, is then approximated by a single density. That serves as the best available description of parameters' behaviour and it is therefore suitable e.g. for prediction purposes.

Keywords: Autoregressive models; Model; Parameter estimation; Prediction; Regression.

1. INTRODUCTION

Tracking of slowly varying parameters is an important task in the theory of adaptive systems. Majority of prediction and control algorithms, employing regression models like autoregression model (AR), autoregression model with exogenous inputs (ARX), autoregression model with moving average (ARMA) etc., assume a carefully defined model structure and correctly estimated parameters. Problems arise, when the model parameters vary in time. The problems of slowly time-varying model parameters were given a thorough attention. The exponential forgetting method, motivated by the idea of flattening the posterior probability density function [Peterka, 1981] or by time-weighted least squares (LS) [Jazwinski, 1970] dominates the group of solutions. Various modifications of this method were developed to solve the problem of information loss, when non-informative data are coming, e.g. the controlled forgetting, directional forgetting etc. (Kulhavý and Kárný [1984] and Cao and Schwartz [2000]). Other techniques employ the state-space model to describe the parameter changes. A typical example is the Kalman filter, estimating the parameters of a linear model with normal noise (see Kalman and Bucy [1961] and Kalman [1960]) and its modification like H_∞ filter, extended Kalman filters or particle filtering [Simon, 2006].

Many improvements of the exponential forgetting method solved its common drawback, but in contrast to the state-space based models, they lack the ability to appropriately track multiple parameters which vary each with different rates. This paper proposes a partial forgetting method, allowing to track the parameters even in this case.

The proposed forgetting technique employs some important facts from the field of the Bayesian modelling, summarized in Kárný et al. [2005]. As far as the authors are aware, the proposed concept is completely new.

The specific notation: $'$ denotes transposition, $\equiv$ is equivalence by definition, $\propto$ is proportionality, i.e. equivalence up to a constant factor. θ^* denotes a set of θ-values, $f(x)$ is probability density function where the random variable is determined by its argument x. The time is discrete, starting from 0.

2. PROBLEM STATEMENT

2.1 System model

Consider a discrete stochastic system observed at time instants $t = 1, 2, \ldots$ Let this system have directly manipulated input u_t, which affects the single system output y_t. The couples of inputs and outputs in each time instant t form the data vector $d_t = (u_t, y_t)$; the sequence $d(t) = (d_1, d_2, \ldots, d_t)$ describes the evolution of the system behaviour in time, i.e. from the beginning time instant 1 until the estimation time t.

Generally, the model output y_t depends on the previous data $d(t-1)$ and the current input u_t. This dependence is modelled by a conditional probability density function (pdf), which has the form

$$f(y_t|u_t, d(t-1), \theta_t) = f(y_t|\psi_t, \theta_t) \tag{1}$$

where θ_t stands for a model parameter (possibly multivariate column vector) and ψ_t is a column regression vector containing all data that have an influence to the output y_t.

2.2 Parameter estimation

According to the Bayesian approach, the unknown model parameter θ is a random variable. Then, it is possible to describe it by a probability density function, conditioned by the data available at the current time instant t, i.e. $f(\theta|d(t))$. If we apply the natural conditions of control [Peterka, 1981] saying

$$f(\theta_t|u_t, d(t-1)) = f(\theta|d(t-1)) \tag{2}$$

then the Bayes rule for recurrent parameter estimation reads

$$f(\theta_t|d(t)) \propto f(y_t|\psi_t,\theta_t) f(\theta_t|d(t-1)) \tag{3}$$

This relation can be viewed as the *data update*, as the new information carried by the data is incorporated into the parameter estimate.

The successive step after the *data update* is the *time update*, formally given

$$f(\theta_{t+1}|d(t)) = \int_{\theta^*} f(\theta_{t+1}|d(t),\theta_t) f(\theta_t|d(t))\, \mathsf{d}\theta \tag{4}$$

In the case of time-invariant parameters $\theta_{t+1} = \theta_t$, the time update is rather a formal step. However, a mathematical model with a fixed structure and constant parameters is not always suitable for modelling the reality and it is often necessary to admit that its parameters vary. There is a couple of methods how to obtain the posterior pdf in (4), one of them is to consider an explicit model of parameter changes on the right-hand side. Unfortunately such model is not always available. Another approach is to modify the whole time-update to make it admit slow permanent changes of parameter estimates. Such an approach is called time weighting, time discounting or simply forgetting.

Remark 1. In this paper, the case of slowly varying parameters is considered, which can be formally written as $\theta_t \approx \theta_{t-1}$. In regard to this proximity, we don't write the parameters with time index anymore.

The summary of estimation of slowly varying parameters with forgetting:

(1) Collect the newest data d_t.
(2) Perform the data update of the parameter probability density function (3).
(3) Perform the time update (4) in the form of forgetting

The main problem of the majority of forgetting methods consists in the fact, that all parameters are forgotten with the same rate.

3. PARTIAL FORGETTING

The basic idea of partial forgetting, allowing tracking of individual parameters, is based on the notion of unknown true parameter probability density function ${}^{T}f(\theta|d(t))$. This pdf describes ideally the actual behaviour of the model parameters. Our aim is to find its best approximation over all formulated hypotheses about the variability of individual parameter elements. The hypotheses specify whether and which configuration of parameters changes. Each hypothesis has its own probability with which it is supposed to be valid and induces a probability density function, which should be used on condition of the hypothesis validity. Division of the reality into several specific cases, according to the specified hypotheses, leads to the description of the true pdf in the form of a mixture of densities. The goal is to find the best approximation $\tilde{f}$ of this mixture, regardless on the knowledge which hypothesis is true at the moment.

This approximate pdf is constructed so that it would minimize expectation of a distance between the mixture and itself. As the distance (or more correctly divergence) measure, we use the Kullback-Leibler divergence [Kullback and Leibler, 1951] in the form

$$\mathsf{D}\left(f(x)||g(x)\right) = \int f(x) \ln \frac{f(x)}{g(x)} \mathsf{d}x, \qquad x \in x^* \tag{5}$$

It measures the divergence of a pair of pdfs f and g, acting on a set x^*. However, it cannot be considered as a distance measure, since it does not satisfy neither the symmetry $\mathsf{D}\left(f||g\right) \neq \mathsf{D}\left(g||f\right)$, nor the triangle inequality.

3.1 Hypotheses

As it has been mentioned, the method of partial forgetting is based on an unknown random true multivariate parameter pdf ${}^{T}f(\theta|d(t)) = {}^{T}f(\theta_1,\ldots,\theta_n|d(t))$, $n = 1, 2, \ldots$. The problem is, that such a pdf is not available to us, as we are not sure about the variability of individual parameters. Theoretically, it would be possible to consider a hyperdistribution describing the pdf ${}^{T}f$, however, it is too complicated and we will drop the idea. For our purposes, it is fully sufficient to take into account its point estimates constructed on the basis of the individual hypotheses about the parameters behaviour. These hypotheses are given by the expectations as follows:

$$\begin{aligned}
H_0 &: \mathsf{E}\left[{}^{T}f(\theta|d(t))|\theta,d(t),H_0\right] = f(\theta|d(t)) \\
H_1 &: \mathsf{E}\left[{}^{T}f(\theta|d(t))|\theta,d(t),H_1\right] = \\
&\quad = f(\theta_2,\ldots,\theta_n|\theta_1,d(t)) f_A(\theta_1) \\
H_2 &: \mathsf{E}\left[{}^{T}f(\theta|d(t))|\theta,d(t),H_2\right] = \\
&\quad = f(\theta_1,\theta_3,\ldots,\theta_n|\theta_2,d(t)) f_A(\theta_2) \\
&\cdots \\
H_n &: \mathsf{E}\left[{}^{T}f(\theta|d(t))|\theta,d(t),H_n\right] = \\
&\quad = f(\theta_1,\ldots,\theta_{n-1}|\theta_n,d(t)) f_A(\theta_n) \\
H_{n+1} &: \mathsf{E}\left[{}^{T}f(\theta|d(t))|\theta,d(t),H_{n+1}\right] = \\
&\quad = f(\theta_3,\ldots,\theta_n|\theta_1,\theta_2,d(t)) f_A(\theta_1,\theta_2) \\
H_{n+2} &: \mathsf{E}\left[{}^{T}f(\theta|d(t))|\theta,d(t),H_{n+2}\right] = \\
&\quad = f(\theta_2,\theta_4\ldots,\theta_n|\theta_1,\theta_3,d(t)) f_A(\theta_1,\theta_3) \\
&\cdots \\
H_{2^n-2} &: \mathsf{E}\left[{}^{T}f(\theta|d(t))|\theta,d(t),H_{2^n-2}\right] = \\
&\quad = f(\theta_n|\theta_1,\ldots,\theta_{n-1},d(t)) f_A(\theta_1,\ldots,\theta_{n-1}) \\
H_{2^n-1} &: \mathsf{E}\left[{}^{T}f(\theta|d(t))|d(t),H_{2^n-1}\right] = f_A(\theta)
\end{aligned} \tag{6}$$

The verbal expression of the given hypotheses is the following: H_0 assumes that no parameter varies, hence the data-updated pdf is used in (3) directly as the time updated one. The hypotheses $H_1,\ldots,H_n$ represent cases when only one parameter varies, thus its marginal pdf is replaced with a suitable alternative. The following hypotheses present cases when a specific subset of parameters vary. The last hypothesis H_{2^n-1} expresses the case when all parameters vary. Here, the whole data updated pdf is substituted by a suitable (preferably flat) alternative.

Notice that in the hypotheses definition the random element is the whole pdf ${}^{T}f$. All other variables like parameters and data occur in the condition and are treated as known, hence the expectation is taken over all possible forms of ${}^{T}f$.

Each hypothesis H_i is assigned its weight λ_i, characterized as a probability of becoming true during the time run. Hence $\lambda_i \in [0,1], \quad i = 0, \ldots, 2^n - 1$ and $\sum_{i=1}^{2^n-1} \lambda_i = 1$.

3.2 Approximative pdf

The convex combination of the probability density functions according to individual hypotheses produces the expression of the true parameter probability density function.

$$\begin{aligned} \mathsf{E}\left[{}^{T}f(\theta|d(t))|\mathcal{C}\right] &= \mathsf{E}\left[\mathsf{E}\left[{}^{T}f(\theta|d(t))|\mathcal{C}, H_i\right]|\mathcal{C}\right] = \\ &= \sum_{i=0}^{2^n-1} \lambda_i \mathsf{E}\left[{}^{T}f(\theta|d(t))|\mathcal{C}, H_i\right] \end{aligned} \tag{7}$$

where the condition $\mathcal{C}$ comprises all what is supposed to be known.

We search for an approximative pdf $\tilde{f}(\theta|d(t))$ of the mixture (7) that belongs to the same family of distributions as the true pdf ${}^{T}f$. Under general conditions, as a 'measure' of dissimilarity between two distributions, it is convenient to use the Kullback-Leibler divergence [Bernardo, 1979]. Hence the approximative pdf could be selected as that which minimizes the expected divergence between the mixture and itself

$$\begin{aligned} &\arg \min_{f^*(\theta|d(t))} \mathsf{E}\left[\mathsf{D}\left({}^{T}f \middle|\middle| \tilde{f}\right)|\mathcal{C}\right] = \\ &= \arg \min_{f^*(\theta|d(t))} \mathsf{E}\left[\int_{\theta^*} {}^{T}f(\theta|d(t)) \ln \frac{{}^{T}f(\theta|d(t))}{\tilde{f}(\theta|d(t))} \,\mathrm{d}\theta|\mathcal{C}\right] = \\ &= \arg \min_{f^*(\theta|d(t))} \int_{\theta^*} \mathsf{E}\left[{}^{T}f(\theta|d(t))|\mathcal{C}\right] \ln \frac{1}{\tilde{f}(\theta|d(t))} \,\mathrm{d}\theta = \\ &= \arg \min_{f^*(\theta|d(t))} \int_{\theta^*} \sum_{i=0}^{2^n-1} \lambda_i \mathsf{E}\left[{}^{T}f(\theta|d(t))|\mathcal{C}, H_i\right] \ln \frac{1}{\tilde{f}(\theta|d(t))} \,\mathrm{d}\theta \end{aligned} \tag{8}$$

where the condition $\mathcal{C}$ again comprises all what is supposed to be known.

Using the relation (8), we can find the best approximation of the true parameter probability density function $\tilde{f}(\theta|d(t))$. This pdf ideally approximates the probabilistic description of the real behaviour of model parameters, i.e whether any of their subset possibly varies in time or not.

4. DERIVATION FOR NORMAL REGRESSION MODEL

If we assume normality of the regression model (1), we can consider the parameters to have Gauss-inverse-Wishart distribution, ${}^{T}f \sim GiW_{\Theta}(V, \nu)$ defined as follows [Kárný et al., 2005]:

Proposition 1. (Gauss-inverse-Wishart pdf). The probability density function of the Gauss-inverse-Wishart distribution has the form

$$GiW_{\Theta}(V,\nu) \equiv \frac{r^{-0.5(\nu+n+2)}}{I(V,\nu)} \exp\left\{\frac{-1}{2r}\begin{bmatrix}-1\\ \theta'\end{bmatrix}' V \begin{bmatrix}-1\\ \theta\end{bmatrix}\right\} \tag{9}$$

or

$$\begin{aligned} GiW_{\Theta}(L,D,\nu) \equiv &\frac{r^{-0.5(\nu+n+2)}}{I(L,D,\nu)} \cdot \\ &\cdot \exp\left\{\frac{-1}{2r}\left[(\theta-\hat{\theta})' C^{-1} (\theta-\hat{\theta}) + D_{LSR}\right]\right\} \end{aligned} \tag{10}$$

where the individual terms have the following meaning:

- ν stands for degrees of freedom,
- n denotes length of the regression vector $[-1, \theta']'$,
- r is the variance of model noise,
- V_t is the extended information matrix, i.e. symmetric square $n \times n$ dimensional non-zero positive definite matrix, which carries the information about the past data. By its $L'DL$ decomposition, the terms L and D are obtained.
- θ is a vector of regression parameters
- $\hat{\theta}$ is a least-squares (LS) estimate of θ
- I stands for normalization integral
- C is the covariance of LS estimate
- D_{LSR} is the LS reminder
- $\Theta \equiv (\theta', r)'$ collects the unknown model parameters

The expression of individual terms (the normalization integral in particular) can be found in Kárný et al. [2005]. The important terms are given later in this paper.

A key property of the extended information matrix V is its $L'DL$ factorability to unique unit triangular matrix L and the unique unit diagonal matrix D

$$\begin{aligned} V &= \begin{bmatrix} {}^{\lfloor d}V & {}^{\lfloor d\psi}V' \\ {}^{\lfloor d\psi}V & {}^{\lfloor \psi}V \end{bmatrix} = \\ &= L'DL = \begin{bmatrix} 1 & 0 \\ {}^{\lfloor d\psi}L & {}^{\lfloor \psi}L \end{bmatrix}' \begin{bmatrix} {}^{\lfloor d}D & 0 \\ 0 & {}^{\lfloor \psi}D \end{bmatrix} \begin{bmatrix} 1 & 0 \\ {}^{\lfloor d\psi}L & {}^{\lfloor \psi}L \end{bmatrix} \end{aligned} \tag{11}$$

Apparently, the left upper-corner V and D matrix elements are scalars, ${}^{\lfloor d}D$, ${}^{\lfloor d}V \in \mathbb{R}$. Recalling Proposition 1, the least-square estimate of parameters $\hat{\theta} \equiv {}^{\lfloor \psi}L^{-1}\, {}^{\lfloor d\psi}L$ has the covariance $C \equiv {}^{\lfloor \psi}L^{-1}\, {}^{\lfloor \psi}D^{-1} ({}^{\lfloor \psi}L^{-1})'$ and the least-square reminder $D_{LSR} \equiv {}^{\lfloor d}D$.

Suppose, that the GiW pdf given above represents a density obtained by the data-update step (3) and the next logical step to be determined is the time update in the form of forgetting. First, we have to construct appropriate hypotheses about the individual regression parameters' behaviour (6). Note, that the variance r is varied together with regression parameters from θ^*, when the GiW pdf is decomposed [Kárný et al., 2005] and the hypothetical pdfs constructed.

Proposition 2. (Low-dimensional pdfs of GiW pdf). Given a distribution $GiW_{[{}^{a}\theta', {}^{b}\theta']', r}(V, \nu)$. Let $L'DL$ be the decomposition of the extended information matrix V of its probability density function as follows:

$$L \equiv \begin{bmatrix} 1 & & \\ {}^{\lfloor da}L & {}^{\lfloor a}L & \\ {}^{\lfloor db}L & {}^{\lfloor ab}L & {}^{\lfloor b}L \end{bmatrix}, D \equiv \begin{bmatrix} {}^{\lfloor d}D & & \\ & {}^{\lfloor a}D & \\ & & {}^{\lfloor b}D \end{bmatrix} \tag{12}$$

Then, the GiW probability density function can be decomposed to the low-dimensional marginal pdf

$$f({}^{\lfloor a}\theta, r) \sim GiW_{{}^{\lfloor a}\theta, r}\left(\begin{bmatrix} 1 & \\ {}^{\lfloor da}L & {}^{\lfloor a}L \end{bmatrix}, \begin{bmatrix} {}^{\lfloor d}D & \\ & {}^{\lfloor a}D \end{bmatrix}, \nu\right) \tag{13}$$

and the low-dimensional conditional pdf

$$f(\,{}^{\lfloor b}\theta|\,{}^{\lfloor a}\theta, r) \sim$$
$$\sim N_{{}^{\lfloor b}\theta}\left({}^{\lfloor b}L^{-1}\left({}^{\lfloor db}L - {}^{\lfloor ab}L\,{}^{\lfloor a}\theta\right), r\left({}^{\lfloor b}L'\,{}^{\lfloor b}D\,{}^{\lfloor b}L\right)^{-1}\right) \tag{14}$$

The proof can be found in Kárný et al. [2005]

This proposition allows us to select and change the marginal pdf for parameter ${}^{\lfloor a}\theta$ by replacing the proper rows in the $L'DL$-factorized information matrix with suitable alternative. To change the marginal pdf inherent to parameter ${}^{\lfloor b}\theta$, it is necessary to permute the proper rows, in the particular case. The permutation algorithm is given in Kárný et al. [2005] as well.

As given in Section 3.2, the convex combination of the hypothetic pdfs with weights λ_i leads to a mixture of densities approximating the true parameter probability density function. To approximate this mixture with a single GiW density we search for the minimally divergent (in the Kullback-Leibler divergence sense) pdf as given in (8). The Kullback-Leibler divergence introduced by (5) of two GiW distributions is given by the following proposition [Kárný et al., 2005]:

Proposition 3. (KL divergence of two GiW pdfs). Given two distributions with probability density functions f and $\tilde{f}$. The Kullback-Leibler divergence of these two functions has the following form

$$\mathsf{D}\left(f\middle\|\tilde{f}\right) = \ln\frac{\Gamma(0.5\tilde{\nu})}{\Gamma(0.5\nu)} - 0.5\ln|C\tilde{C}^{-1}| + 0.5\tilde{\nu}\ln\frac{D_{LSR}}{\tilde{D}_{LSR}}$$
$$+ 0.5(\nu-\tilde{\nu})\psi_0(0.5\nu) - 0.5n - 0.5\nu + 0.5\mathrm{Tr}\left(C\tilde{C}^{-1}\right)$$
$$+ 0.5\frac{\nu}{D_{LSR}}\left[\left(\hat{\theta}-\hat{\tilde{\theta}}\right)'\tilde{C}^{-1}\left(\hat{\theta}-\hat{\tilde{\theta}}\right) + \tilde{D}_{LSR}\right] \tag{15}$$

where $\psi_0(\cdot)$ denotes the digamma function, i.e. the first logarithmic derivative of the gamma function $\Gamma(\cdot)$.

The proof is not trivial and is given in [Kárný et al., 2005].

To find the best approximation of mixture (7) made from GiW densities, we need to find the minimum of the Kullback-Leibler divergence in Proposition 3 by taking derivatives with respect to $\hat{\tilde{\theta}}, \tilde{C}, \tilde{D}_{LSR}$ and $\tilde{\nu}$. The results give the following proposition.

Proposition 4. Given a convex combination (mixture) of n Gauss-inverse-Wishart pdfs. Its best approximation in the sense of the minimizer of the Kullback-Leibler divergence, holding the GiW distribution, is given by the following statistics

- $\tilde{\theta}$ – the regression coefficients

$$\hat{\tilde{\theta}} = \left(\sum_{i=1}^{n}\lambda_i\frac{\nu_i}{D_{LSR,i}}\right)^{-1}\cdot\left(\sum_{i=1}^{n}\lambda_i\frac{\nu_i}{D_{LSR,i}}\hat{\theta}_i\right) \tag{16}$$

- $\tilde{D}_{LSR}$ – the least-squares reminder

$$\tilde{D}_{LSR} = \tilde{\nu}\cdot\left(\sum_{i=1}^{n}\lambda_i\frac{\nu_i}{D_{LSR,i}}\right)^{-1} \tag{17}$$

- $\tilde{C}$ – the least-square covariance matrix

$$\tilde{C} = \sum_{i=1}^{n}\lambda_i C_i + \text{`}$$
$$+ \sum_{i=1}^{n}\lambda_i\frac{\nu_i}{D_{LSR;i}}\left[\left(\hat{\theta}_i-\hat{\tilde{\theta}}\right)\left(\hat{\theta}_i-\hat{\tilde{\theta}}\right)'\right] \tag{18}$$

- and the counter (degrees of freedom)

$$\tilde{\nu} = \frac{1+\sqrt{1+\frac{4}{3}(A-\ln 2)}}{2(A-\ln 2)} \tag{19}$$

where

$$A = \ln\left(\sum_{i=1}^{n}\lambda_i\frac{\nu_i}{D_{LSR,i}}\right) + \sum_{i=1}^{n}\lambda_i\ln D_{LSR,i} -$$
$$- \sum_{i=1}^{n}\lambda_i\,\psi_0(0.5\nu_i) \tag{20}$$

The Proposition can be proved by minimizing the differentiated Kullback-Leibler divergence of two Gauss-inverse-Wishart pdfs (15).

Remark 2. The given expression of counter employs an approximation of the digamma function $\psi_0(\tilde{\nu})$. The approximation was done on base of the Bernoulli numbers, however multiple methods can be used (see e.g. Bernardo [1976] or Spouge [1994] or Cody et al. [1973]).

A Gauss-inverse-Wishart probability density function (10) constructed with the found terms (16), (17), (18) and (19) can be used as the best approximation of the parameters' reality and hence used e.g. for prediction purposes.

5. EXPERIMENT

This experiment demonstrates the effect of the partial forgetting-based estimation on a prediction with an autoregressive model. The results are compared to the prediction with exponential forgetting method, which is the most popular approach to estimation of time-variant parameters in linear stochastic systems.

The exponential forgetting is formally motivated by time-weighted least squares [Jazwinski, 1970] or flattening the posterior pdf [Peterka, 1981]. The time update has the following form

$$[f(\theta|d(t))]^{\lambda}, \qquad \lambda\in(0,1] \tag{21}$$

where pdf $f(\theta|d(t))$ is the data-updated pdf from (3) and λ is the forgetting factor, usually not lower than 0.95.

5.1 Transportation data

For the practical testing purposes, the real traffic data were used. The data sample consists of traffic intensities measured in Prague, Czech Republic, with the sampling period equal to five minutes. For our purposes, the data window of 300 samples was used (see Fig. 1). The initial 10 samples were used as a source of alternative information.

The traffic system was modelled with a first-order autoregression model AR(1) in the form

$$y_t = \theta_1 + \theta_2 y_{t-1} + e_t, \quad t = 1,2,\ldots \tag{22}$$

where $\theta = (\theta_1, \theta_2)'$ are regression parameters and e_t denotes the normally distributed white noise with zero mean and constant variance. y_t denotes the modelled traffic intensity.

According to the model, the appropriate four hypotheses about the true pdf equivalent to those given in (6) were constructed as follows

$$\begin{aligned} H_0 &: \mathsf{E}\left[f_T(\theta_1,\theta_2,r|d(t))|\theta_1,\theta_2,r,d(t),H_0\right] = \\ &= f(\theta_1,\theta_2,r|d(t)) \\ H_1 &: \mathsf{E}\left[f_T(\theta_1,\theta_2,r|d(t))|\theta_1,\theta_2,r,d(t),H_1\right] = \\ &= f(\theta_2|\theta_1,r,d(t))f_A(\theta_1,r) \\ H_2 &: \mathsf{E}\left[f_T(\theta_1,\theta_2,r|d(t))|\theta_1,\theta_2,r,d(t),H_2\right] = \\ &= f(\theta_1|\theta_2,r,d(t))f_A(\theta_2,r) \\ H_3 &: \mathsf{E}\left[f_T(\theta_1,\theta_2,r|d(t))|\theta_1,\theta_2,r,d(t),H_3\right] = \\ &= f_A(\theta_1,\theta_2,r) \end{aligned} \tag{23}$$

The optimization problem consisted in the search for optimal weights $\lambda = [\lambda_0, \lambda_1, \lambda_2, \lambda_3]$ of hypotheses H_0, H_1, H_2 H_3. The quality of estimation was evaluated by the prediction ability. As a criterion of the prediction quality, the relative prediction error RPE defined as follows was considered

$$RPE = \frac{1}{s}\sqrt{\frac{\sum_{i=1}^{T}(y_{p;i} - y_i)^2}{T}} \tag{24}$$

where y_i denotes the real system output, $y_{p;i}$ is the predicted output and s is the sample standard deviation of data on horizon T.

Characteristics	Partial forg.	Exp. forg.
Hyp./Forg. weight(s)	$\lambda = [0.9, 0.1, 0, 0]$	$\lambda = 0.985$
Rel. pred. error	0.0422	0.0989
Pred. error – minimum	–1.0930	–2.3060
Pred. error – maximum	3.1240	3.8140
Pred. error – average	0.0934	0.7709
Pred. error – st. deviation	0.6215	1.2530

Table 1. Elementary characteristics of AR(1) model with partial and exponential forgetting.

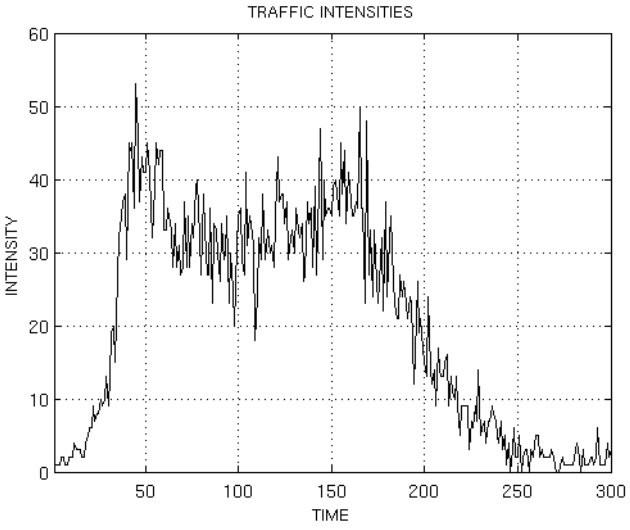

Fig. 1. Real course of traffic intensities.

Some interesting results and statistics are shown in the Table 1. It compares the AR(1) models with parameter estimation with partial and exponential forgetting methods, respectively. The first row shows the weights of particular hypotheses $\lambda = [\lambda_0, \ldots, \lambda_3]$ for partial forgetting and the forgetting factor λ of exponential forgetting. Other table rows show the relative prediction error and a few interesting statistics of the prediction errors. Apparently, the partial forgetting based estimation leads to smaller relative prediction error. The (absolute) prediction errors (differences between real and predicted values for specific time instants) are smaller and show that the prediction of y_t is less biased.

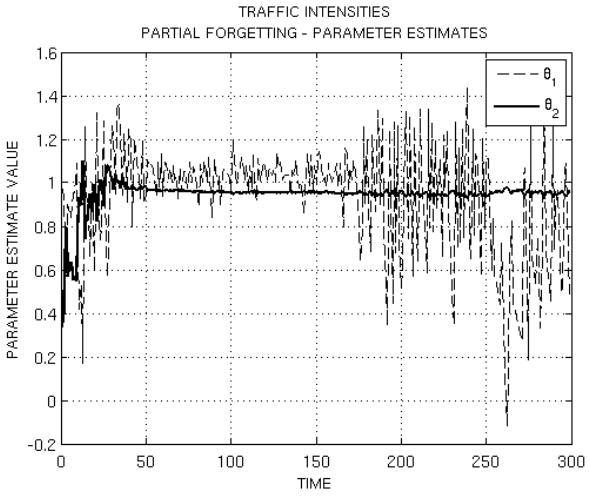

Fig. 2. AR(1) with partial forgetting: Evolution of model parameters estimates

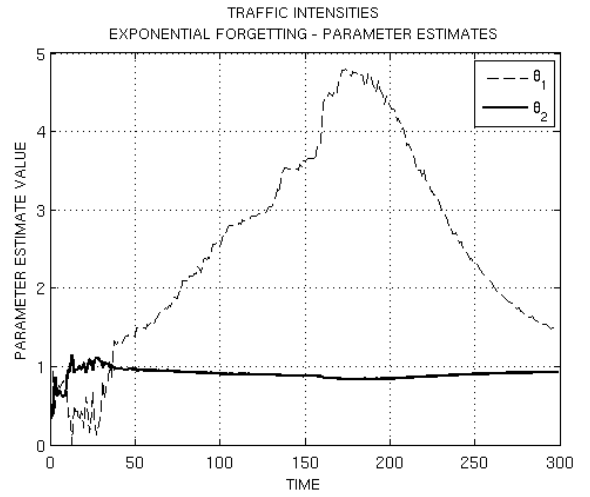

Fig. 3. AR(1) with exponential forgetting: Evolution of model parameters estimates

Figures 2 and 3 show the evolution course of model parameter estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ during the estimation for both forgetting methods. Apparently the changes are caught by the absolute term in both cases, as one would intuitively expect.

Figures 4 and 5 respectively show the course of prediction errors for both forgetting methods. The prediction with partial forgetting leads to smaller and more symmetrical (around zero) errors than the exponential forgetting.

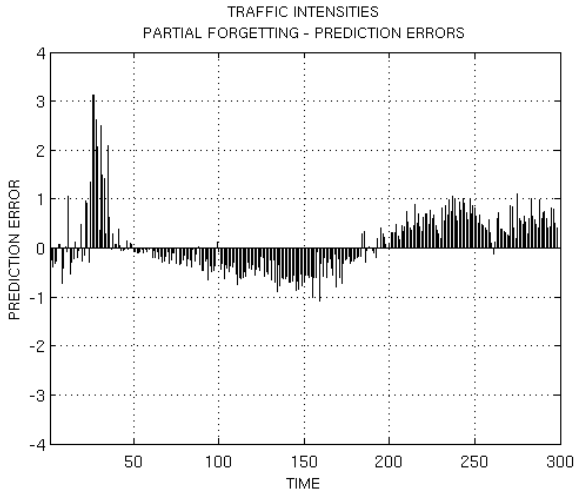

Fig. 4. AR(1) with partial forgetting: Prediction errors

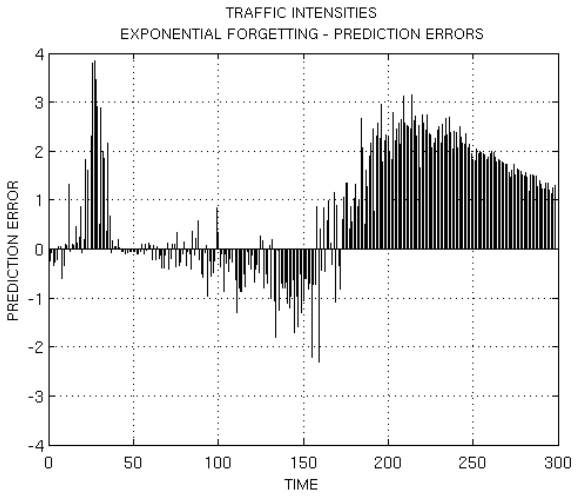

Fig. 5. AR(1) with exponen. forgetting: Prediction errors

6. CONCLUSIONS

The paper describes a new method suitable for tracking of slowly time-varying parameters of a linear stochastic model with parameters that vary in time with different rates. It is based on an unknown true probability density function, describing the real behaviour of parameters. To find its approximation, we define hypotheses about this pdf, introducing its point estimates. Their convex combination is approximated to find the minimally divergent (in the Kullback-Leibler divergence sense) pdf, well describing the parameters and therefore convenient e.g. for prediction purposes.

The challenge is to find a method for selecting significant hypotheses from the set of all possible hypotheses, as well as the choice of their weights. Any theoretical concept would be welcome.

REFERENCES

J.M. Bernardo (1976). *Algorithm AS 103: Psi (digamma) function*, Applied Statistics, Vol. 25, No. 3 (1976), pp. 315-317.

J.M. Bernardo (1979). *Expected information as expected utility*. The Annals of Statistics, Vol. 7, No. 3, pp. 686-690.

L. Cao, H. Schwartz (2000). *Directional forgetting algorithm based on the decomposition of the information matrix*, Automatica, vol. 36, no. 11, pp. 1725–1731.

W.J. Cody, A.J. Strecok, H.C. Thacher (1973). *Chebyshev approximations for the psi function*, Mathematics of Computation, Vol. 27, No. 121 (1973), pp. 123-127

L. Guo, L. Ljung. (1994) *Performance analysis of general tracking algorithms*, in Proceedings of the 33rd Conference on Decision and Control, pp.2851-2855.

A.H. Jazwinski (1970). *Stochastic processes and filtering theory*. New York: Academic Press.

R.E. Kalman, R.S. Bucy (1961). *New results in linear filtering and prediction theory*.

R.E. Kalman. (1960) *A new approach to linear filtering and prediction problems*. Journal of Basic Engineering 82 (1), pp. 35-45.

M. Kárný et al. (2005) *Optimized Bayesian dynamic advising*, Springer.

S. Kullback, R.A. Leibler (1951). *On information and sufficiency*. Annals of Mathematical Statistics vol. 22, no. 1, pp. 79–86.

R. Kulhavý, M. Kárný (1984). *Tracking of slowly varying parameters by directional forgetting*, In Preprints of the 9th IFAC World Congress, Budapest, Vol. X, pp. 78-83.

R. Kulhavý (1987). *Restricted exponential forgetting in real-time identification*, Automatica, vol. 23, no. 5, pp. 589–600.

R. Kulhavý, F.J. Kraus (1996). *On duality of regularized exponential and linear forgetting*, Automatica, vol. 32/10, pp. 1403–1415.

L. Ljung (1999). *System identification: Theory for the user*. Prentince-Hall, Englewood Cliffs, N.J.

V. Peterka (1981). *Bayesian approach to system identification*, in *Trends and Progress in System Identification*, P. Ekhoff, Ed., pp. 239-304. Pergamon Press, Oxford.

D. Simon (2006). *Optimal state estimation: Kalman, H Infinity, and nonlinear approaches*. Wiley-Interscience.

J.L. Spouge (1994). *Computation of the gamma, digamma, and trigamma functions*, SIAM Journal on Numerical Analysis, Vol. 31, No. 3 (1994), pp. 931-944