Akademie věd České republiky
Ústav teorie informace a automatizace, v.v.i.

Academy of Sciences of the Czech Republic
Institute of Information Theory and Automation

RESEARCH REPORT

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

Partial Forgetting

A new method for tracking time-variant parameters

No. 2249 March 23, 2009

ÚTIA AV ČR, P.O.Box 18, 182 08 Prague, Czech Republic
Tel: +420 286892337, Fax: +420 266052068, Url: http://www.utia.cas.cz,
E-mail: *dedecius@utia.cas.cz

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 [1] or by time-weighted least squares (LS) [7] 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, restricted exponential forgetting etc. [2][3][17]. Some methods employ other approaches like linear forgetting [15]. Another group of 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 [4][5] and its modifications like H_∞ filter, extended Kalman filters [6] or particle filtering [6].

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 with different rates. This paper proposes a partial forgetting method, allowing to track the parameters even in this case.

The problem is stated in Section 2, where the system model and the theory of parameter estimation is recalled. Section 3 introduces the concept of partial forgetting method, hypotheses about parameter distribution (Section 3.1) and search for the approximation of the true parameter distribution (Section 3.2). In Section 4, the algorithm of parameter estimation with partial forgetting is described. The practical realization of the method is in Section 5, where the partial forgetting is derived for normal autoregression model. Finally, Section 6 brings tests on both artificially generated and real data and demonstrates the advantages of the method. Concluding remarks are in Section 7.

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 t is discrete.

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 time of estimation 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 on 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 [1] 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* – the new information carried by the data is incorporated into the parameter estimate.

In the case of time-variant parameters, the successive step after the *data update* is the *time update*, formally given

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

If the parameters were time-invariant ($\theta_{t+1} = \theta_t$), the time update would be 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 of the right-hand side. Unfortunately such model is not always available. Another approach is to modify the whole time-update and make it admit slow permanent changes of parameter estimates. Such an approach is called time weighting, time discounting or simply forgetting [8].

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 could be expressed as follows:

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 even if the parameters change with different rates, all of them are forgotten with the same rate. For instance, suppose use of the exponential forgetting on a two-dimensional parameter pdf. In addition, suppose that the individual parameters are slowly varying in time, each one with a different rate of change. In this example, the exponential forgetting can easily fail, because it cannot catch the individual parameter changes. Either it is tuned according to the more quickly changing parameter and the slower one is completely forgotten, or it is fitted to the slower one and the estimation cannot follow the quicker parameter. No matter how this simple case might seem marginal, it can occur in some data measured on real systems.

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 within the class of admissible pdfs. To this end we formulate 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, $\mathsf{E}\left[d({}^{T}f, \tilde{f})\right] \to \min$.

As the distance (or more correctly divergence) measure, we use the Kullback-Leibler divergence [9] in the form

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

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 $\mathsf{KL}\left(f||g\right) \neq \mathsf{KL}\left(g||f\right)$, nor the triangle inequality. Some interesting properties of the Kullback-Leibler divergence are

- By definition $\mathsf{KL}\left(f||g\right) \geq 0$
- iff $f(x) = g(x)$ almost everywhere on $x^* \Longrightarrow \mathsf{KL}\left(f||g\right) = 0$
- iff $g(x) = 0 \wedge f(x) > 0$ on a set of positive Lebesgue measure $\Longrightarrow \mathsf{KL}\left(f||g\right) = \infty$

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 hyper-distribution describing the pdf ${}^{T}f$, but it is too complicated and we will drop the idea to construct it. 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:

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

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

where f_A is an alternative probability density function (preferably flat, e.g. the prior one), expressing uncertainty arising from parameter changes.

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. The hypotheses $H_1 - H_n$ represent the cases when only one parameter varies and its marginal pdf is replaced with an alternative pdf. 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 an 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 thus they are treated as known. Hence the expectation is taken over all possible forms of ${}^{T}f$.

Each of these hypotheses is assigned its weight, characterized as a probability of becoming true during the time run. That is why they must fulfill $\lambda_i \in [0, 1], \quad i = 0, \ldots, 2^n - 1$ 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 expectation 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 $\mathcal{C} = \{\theta, d(t)\}$

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 mixture components. Under general conditions, as a 'measure' of dissimilarity between two distributions, it is convenient to use the Kullback-Leibler divergence [9]. Hence the approximative pdf could be selected as that one which minimizes the expected

divergence between the mixture and itself

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

Using the relation (8), we found 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.

4 Algorithm of the partial forgetting

The algorithm of the partial forgetting method can be described as follows:

Initial mode, for $t = 0$

- Specify the appropriate hypotheses $H_i \in \mathcal{H}^*$, $i = 0, \ldots, 2^n - 1$ about expectation of the true parameter pdf $\mathsf{E}\left[{}^{T}f\right]$.
- Select the probabilities $\lambda_i \in [0, 1]$, $i = 0, \ldots, 2^n - 1$ of relevance of individual hypothesis $H_i \in \mathcal{H}^*$.
- Specify proper alternative pdfs for subset of hypotheses $\mathcal{H}^* \backslash H_0$ (e.g. using the prior pdf)

On-line mode, for $t > 0$

1. Collect the newest data d_t
2. Perform the data update (3)
3. Construct appropriate pdf for each hypothesis from $\mathcal{H}^*$ (6) with a proper alternative parameter behaviour, i.e.:
   (a) Compute the pdf after the data update;
   (b) Change the pdf describing the related parameter(s) with its alternative;
4. Compute the minimally divergent pdf
5. If $t \leq t_{end}$ (t_{end} is the ending time of the estimation), go to the step 1.

The probability density function from the step 4 forms the optimal estimate of the true parameter probability density function.

Remark 2 *The qualities like consistency and bias of the estimate are critically dependent on the choice of hypotheses and weights, which should be solved as an optimization problem, however it is computationally demanding. The preferred approach is to choose from the set of hypotheses $\mathcal{H}^*$ only those which can become significant during the time development. It means, that we can select only a subset of the hypotheses set $\mathcal{H}^*$ and consider the remaining possible hypotheses to have weights equal to zero.*

5 Derivation for normal regression model

If we assume normality of the regression model (1), we can consider the parameters to have Gauss-inverse-Wishart (GiW) distribution defined as follows [10]:

Proposition 1 (Gauss-inverse-Wishart pdf) *The probability density functionof 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

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

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*

The expression of individual terms (the normalization integral in particular) can be found in [10]. The important terms are given later in this paper.

The extended information matrix is symmetric and positively definite and therefore factorable to unique unit triangular matrix L and the unique unit diagonal matrix D as follows

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

Here, the left upper-corner elements of the V and D matrices 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 the 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).

Proposition 2 (Low-dimensional pdfs of GiW pdf) *Given a distribution $GiW_{[\theta', \theta']',r}(V, \nu)$. 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

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

The proof can be found in [10]

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 of the information matrix. The permutation algorithm is given in [10] 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 [10]:

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*

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

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 [10].

To find the best approximation of the mixture (7) of GiW densities, we need to find the minimum of the Kullback-Leibler divergence (Proposition 3) by taking derivatives with respect to $\hat{\tilde{\theta}}, \tilde{C}, \tilde{D}_{LSR}$ and $\tilde{\nu}$. Useful identities are $\frac{\partial}{\partial X}\ln|AXB| = (X^{-1})'$ and $\frac{\partial}{\partial X}\mathsf{Tr}(AX) = A'$).

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 parameters (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*

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

- *and the counter (degrees of freedom)*

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

where

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

Remark 3 *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. [11][12][13]).*

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.

6 Experiments

The partial forgetting method was tested on both artificially generated and real data and the results were compared to the exponential forgetting method, which is the most popular approach to time-variant parameters in linear stochastic systems.

The exponential forgetting is formally motivated by time-weighted least squares [7] or flattening the posterior pdf [1]. The time update has the following form

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

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

In both cases, the related systems were modeled 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 coefficients and e_t denotes the normally distributed white noise with zero mean and constant variance. y_t denotes the modelled system output.

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

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

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 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. The Matlab software was used for this purpose.

6.1 User-defined data

First, we try to predict the development of an artificially generated time series. This series has the following form

$$y_t = \left(0.9 - \frac{1}{t}\right) y_{t-1} + 2, \qquad t = 1, 2, \ldots; y(1) = 2$$

For testing purposes the data were noiseless, thus we expected almost precise prediction and a very little relative prediction error. Although the variable term is the multiplier of the previous output, the first order autoregressive model (22) was expected to catch the development of the data by both regression coefficients θ_1, θ_2. As the source of alternative information for the partial forgetting-based estimation, we used the prior obtained from the first 10 data samples. The following 50 samples were used directly for prediction.

The course of the modelled (i.e. generated) data shows Figure 1.

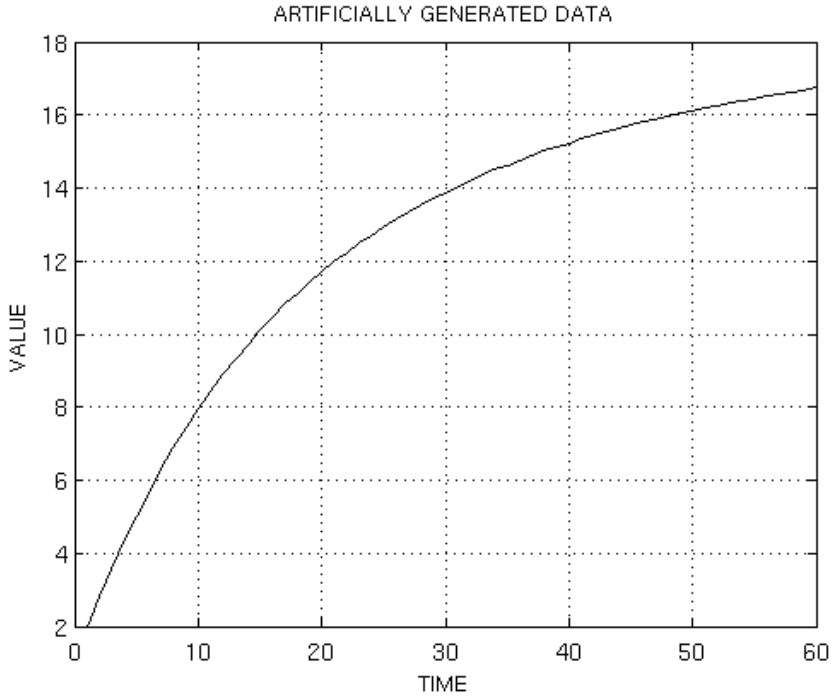

Figure 1: Artificially generated data.

The test showed that the partial-based estimation of the regression coefficients θ_1, θ_2 led to a better prediction than the prediction with exponential forgetting. The optimal forgetting weight of the exponential forgetting was 0.95, which led to the relative prediction error $RPE = 0.0045$. The partial forgetting minimizes the error with weights $\lambda = [0.89, 0, 0.05, 0.06]$ leading to $RPE = 0.0017$.

In the Figures 2 and 3, the course of parameters estimates of the autoregressive model with partial and exponential forgetting is shown. The absolute prediction errors are depicted in the Figures 4 and 5, respectively.

The Table 1 summarizes a few interresting characteristics of the prediction, namely the relative prediction error RPE and statistics of the absolute prediction errors (residues) – the maximum and minimum error, the mean and the standard deviation. Apparently, the prediction errors are more symmetric around zero in the case of the prediction with the partial forgetting-based parameter estimation.

Characteristics	Partial forg.	Exp. forg.
Rel. pred. error	0.0017	0.1576
Pred. error – minimum	–0.0037	–0.0140
Pred. error – maximum	0.0113	0.0113
Pred. error – average	–0.0009	–0.0077
Pred. error – st. deviation	0.0040	0.0072

Table 1: Artificially generated data: Elementary characteristics of AR(1) model with partial forgetting with $\lambda = [0.89, 0, 0.05, 0.06]$ and exponential forgetting with $\lambda = 0.95$.

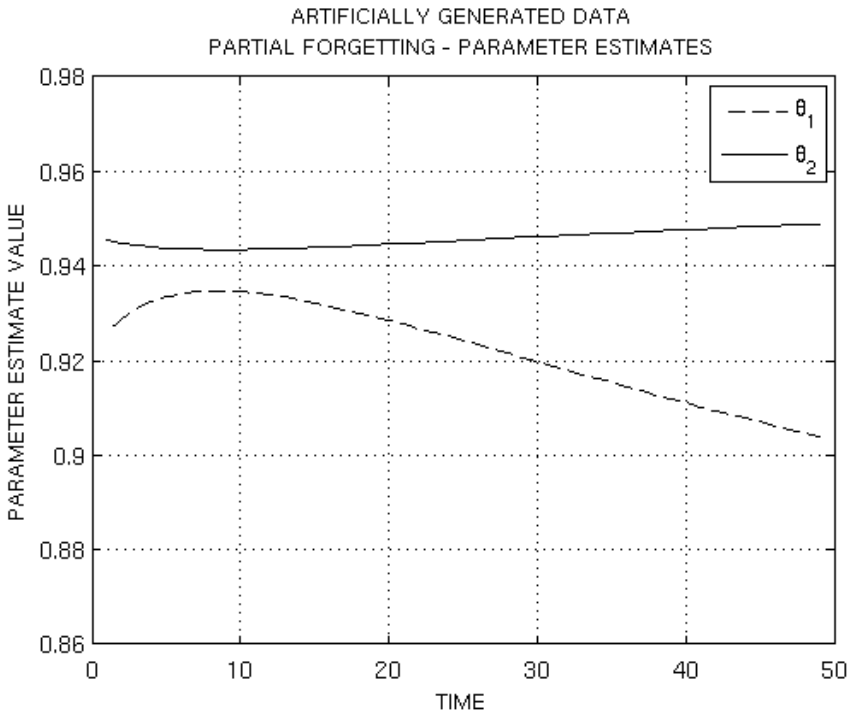

Figure 2: AR(1) with partial forgetting: Evolution of model parameters estimates

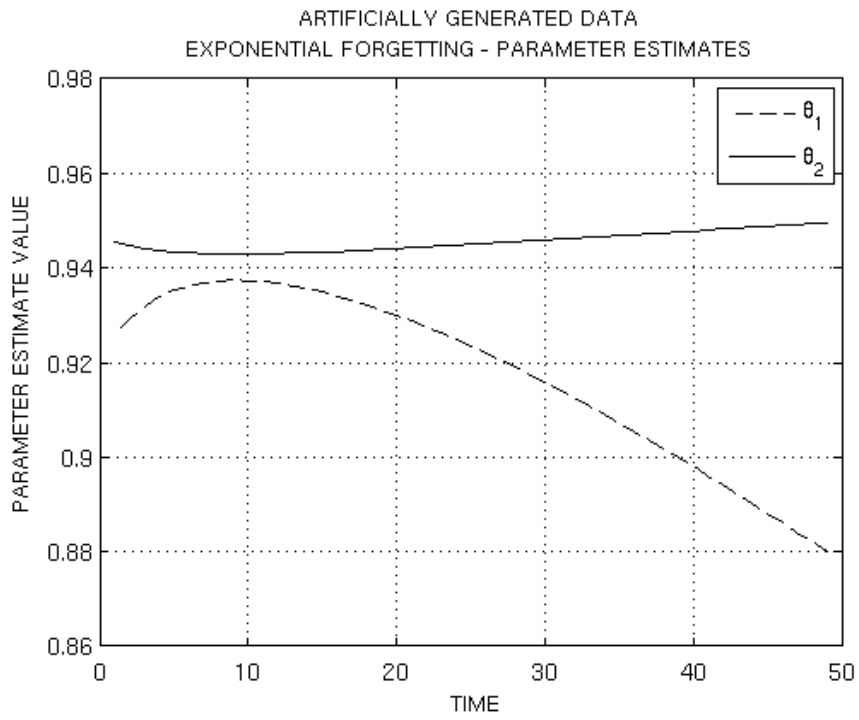

Figure 3: AR(1) with exponential forgetting: Evolution of model parameters estimates

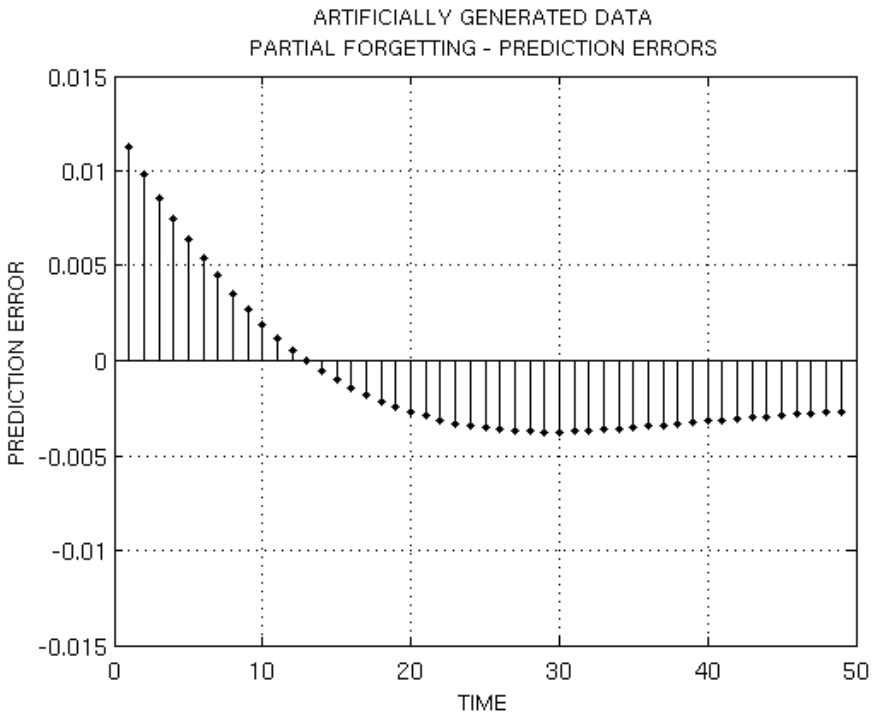

Figure 4: AR(1) with partial forgetting: Prediction errors

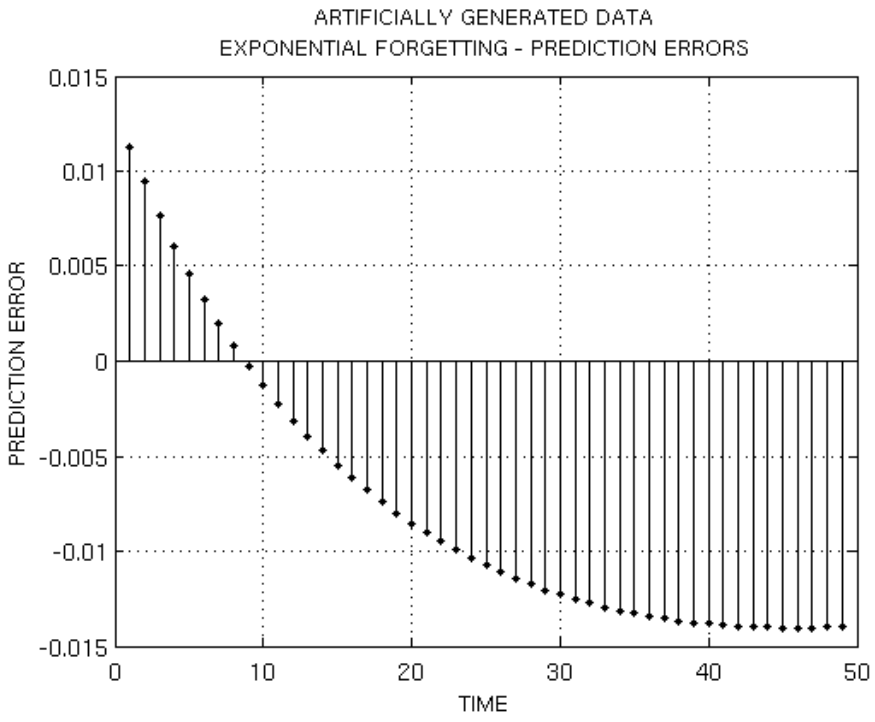

Figure 5: AR(1) with exponential forgetting: Prediction errors

6.2 Transportation data

The next test used the real traffic data. The data sample consisted of traffic intensities measured in Prague, Czech Republic, with the sampling period equal to five minutes. For testing purposes, a data window of 300 samples was used (see Fig. 6). The initial 10 samples were used as a source of alternative information.

Again, the normal first order autoregression model AR(1) defined by Equation (22) was used.

Some interesting results and statistics are summarized in the Table 2. Again, it compares the AR(1) models with parameter estimation with partial and exponential forgetting methods and shows the relative prediction error and a few interesting statistics of the absolute prediction errors. Apparently, the partial forgetting based estimation with weights $\lambda = [0.9, 0.1, 0, 0]$ led to smaller relative prediction error $RPE = 0.0422$, while the exponential forgetting worked best with weight $\lambda = 1.0$ (i.e. no forgetting) leading to $RPE = 0.0989$. The absolute prediction errors were smaller and less biased in the case of the partial forgetting method.

Figures 7 and 8 show the evolution 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 9 and 10 respectively show the course of prediction errors for both forgetting methods. The prediction with partial forgetting led to smaller and more symmetrical (around zero) errors than the exponential forgetting.

Characteristics	Partial forg.	Exp. forg.
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 2: Elementary characteristics of AR(1) model with partial forgetting with $\lambda = [0.9, 0.1, 0, 0]$ and exponential forgetting with $\lambda = 1.0$.

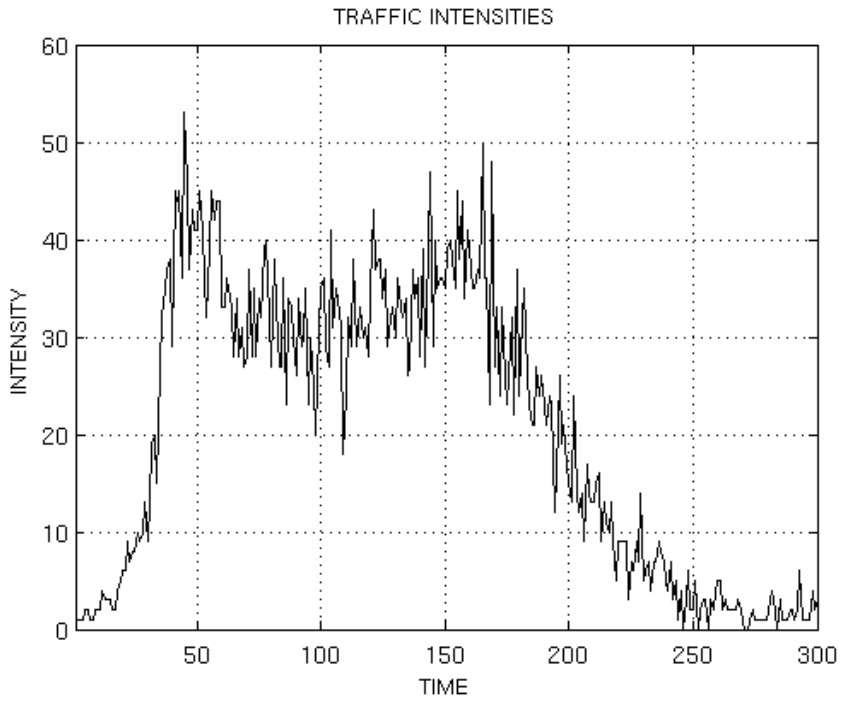

Figure 6: Real course of traffic intensities.

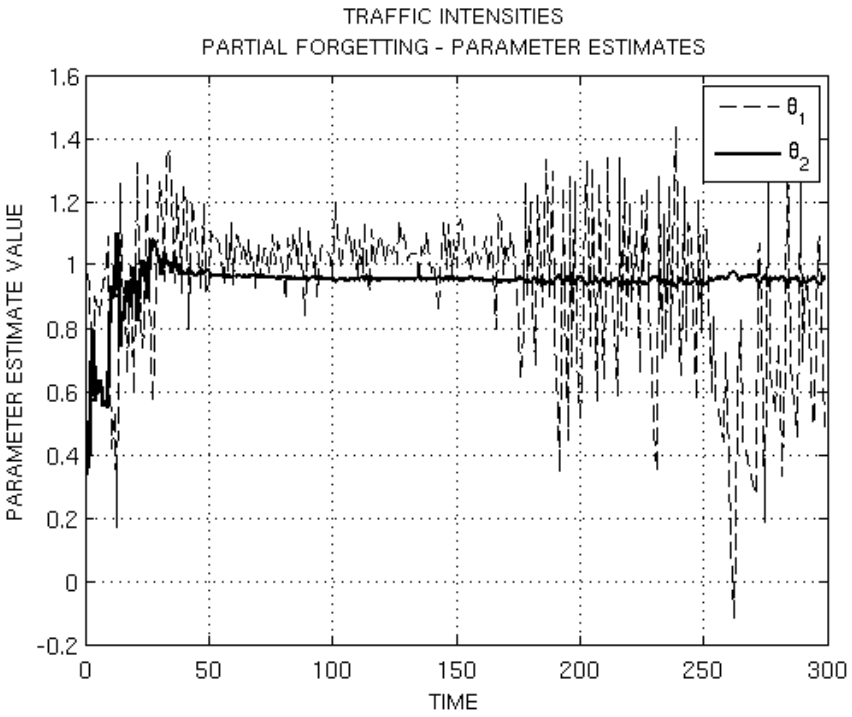

Figure 7: AR(1) with partial forgetting: Evolution of model parameters estimates

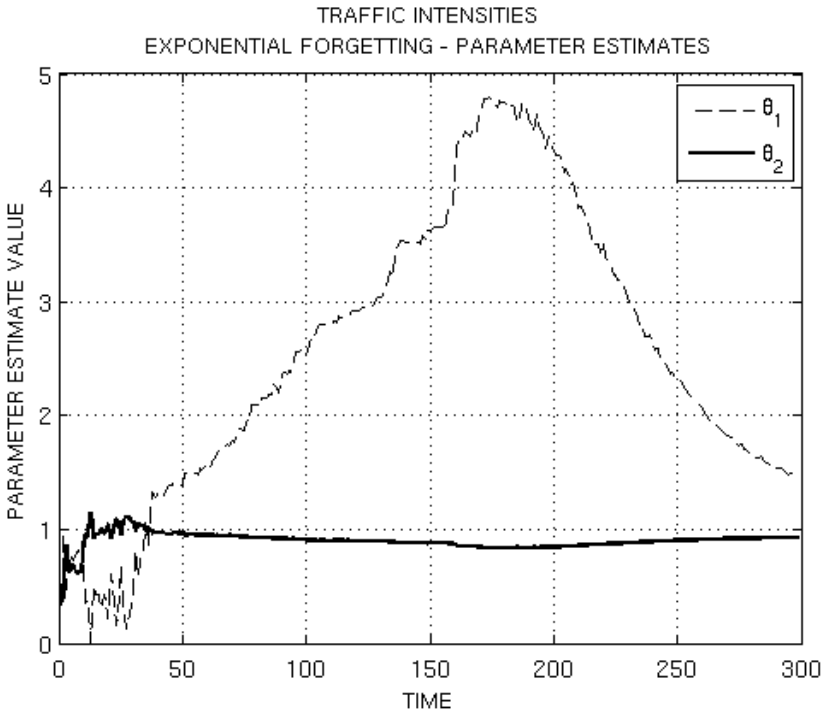

Figure 8: AR(1) with exponential forgetting: Evolution of model parameters estimates

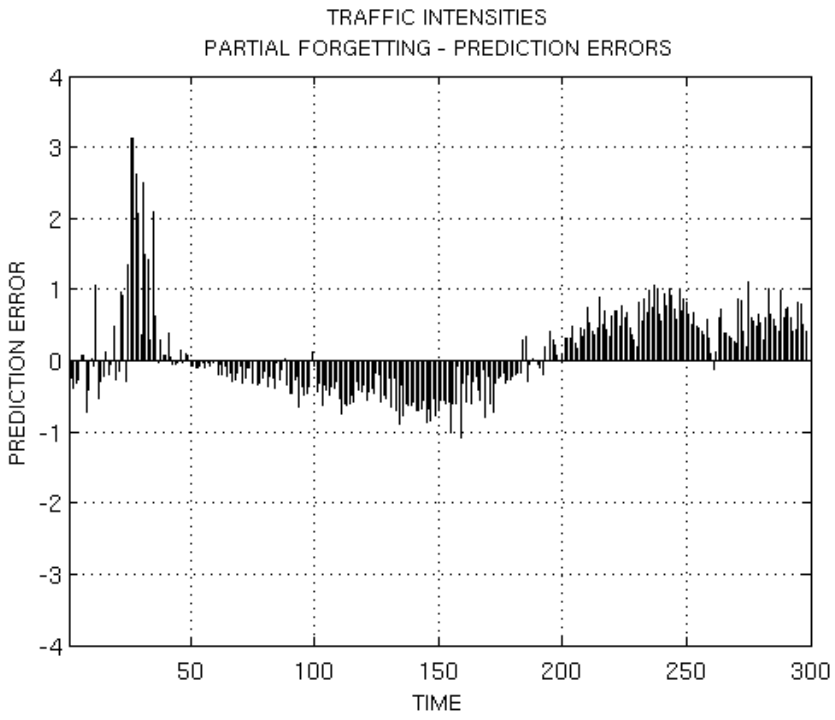

Figure 9: AR(1) with partial forgetting: Prediction errors

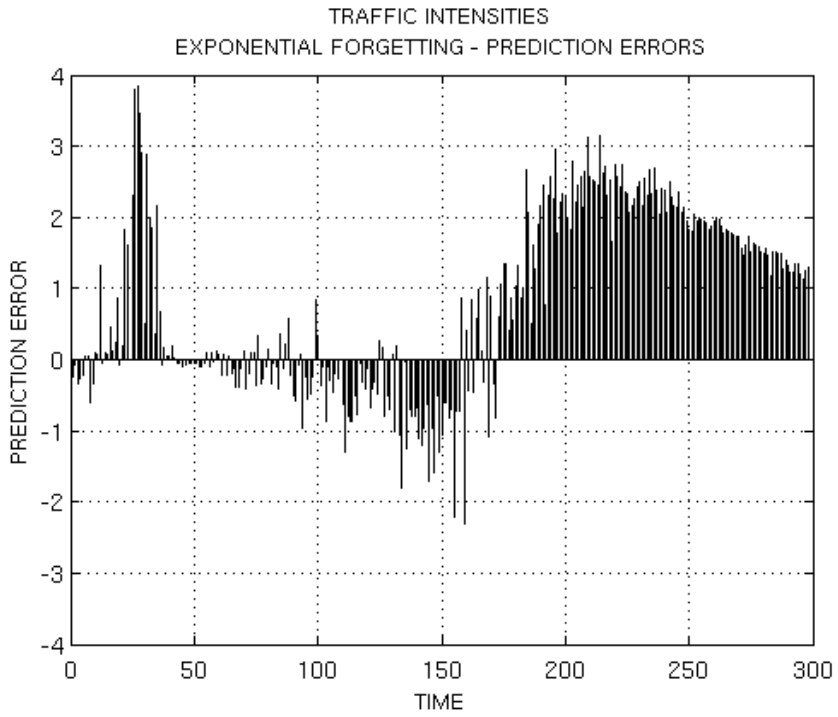

Figure 10: AR(1) with exponential forgetting: Prediction errors

7 Conclusions

The paper described 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 tests on both artificially generated and real traffic data demonstrate, that this approach to slowly time-variant model parameters is suitable and the obtained results show the improvement of the prediction quality in comparison to the exponential forgetting.

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. Also, there may be multiple approaches to the problematics of the suitable alternative(s). Any theoretical concept would be welcome.

References

[1] Peterka, V. (1981). *Bayesian Approach to System Identification*, in *Trends and Progress in System Identification*, P. Ekhoff, Ed., pp. 239–304. Pergamon Press, Oxford.

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

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

[4] Kalman, R.E. & Bucy, R.S. (1961). *New Results in Linear Filtering and Prediction Theory.*

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

[6] Simon, D. (2006). *Optimal State Estimation: Kalman, H Infinity, and Nonlinear Approaches.* Wiley-Interscience.

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

[8] Söderström, T. & Stoica, P. (1989). *System Identification.* New York: Prentice Hall.

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

[10] Kárný, M. et al. (2005). *Optimized Bayesian Dynamic Advising*, Springer.

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

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

[13] Cody, W.J., Strecok, A.J. & Thacher, H.C. (1973). *Chebyshev Approximations for the Psi Function*, Mathematics of Computation, Vol. 27, No. 121 (1973), pp. 123–127.

[14] Guo, L. & Ljung, L. (1994). *Performance Analysis of General Tracking Algorithms*, in Proceedings of the 33rd Conference on Decision and Control, pp.2851–2855.

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

[16] Ljung, L. (1987). *System Identification: Theory for the User.* Prentince-Hall, Englewood Cliffs, N.J.

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