A Brief Comparison of Selected Forgetting Methods

Kamil Dedecius

***Abstract*— This paper brings a comparison of three selected techniques for estimation of slowly varying parameters of input-output models. One of them is the exponential forgetting method, which is the most popular and simplest method, while another method – the alternative forgetting – is based on it. The third selected method is the partial forgetting, which presents a completely different approach to the slowly varying parameters. The comparison of these methods is based on a one-step ahead prediction of a predefined time series with models employing these forgetting methods. The prediction errors are then compared.**

I. INTRODUCTION

Tracking of parameters of a linear regressive models is a key factor in adaptive algorithms of all kinds [1]. However, if a process has to be modelled by a linear filter model, its stationarity is a limiting condition. A stochastic process $x(t)$ is said to be stationary if its statistics (distribution functions for each fixed t) are not affected by a time shift. It means, that two processes $x(t)$ and $x(t+\varepsilon)$ have the same statistics for any ε [2]. If the moments are finite, then the stationary process can be characterized with its first general and second central moments and the direction of time cannot be determined from them [3].

However, the condition of the stationarity cannot always be fulfilled, moreover, most real processes are in their nature not stationary (e.g. the longer-term economical series etc.). If a statistician has to model such processes, he needs to employ a scheme that allows changes in the moments like the mean value and variance. The forgetting-based estimation is then a popular option.

The goal of the paper is to demonstrate the prediction abilities of regression models employing parameter estimation with forgetting. The exponential [6][7] and alternative (sometimes called stabilized [5]) forgetting represent the more or less classical approach to slowly varying parameters, while the partial forgetting [12] is a newly developed method. For our purpose, just the noise free time series with characteristic roots in the right half-plane of the unit circle were modelled.

II. SYSTEM MODEL

Let's have a discrete stochastic system observed at time instants $t = 1, 2, \ldots$ This system can 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 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 on the output y_t.

Usually, the goal is to find such a probabilistic model that differs from the reality presented by an unknown true pdf as least as possible. As a measure of the difference between two pdf (or distributions in general), the Kullback-Leibler divergence defined as follows is used [4].

Definition 1 (Kullback-Leibler divergence): Let f and g be two continuous probability density functions of a random variable x. The Kullback-Leibler divergence, also known as relative entropy, is defined as

$$\mathsf{KL}\left(f(x)||g(x)\right) = \int f(x) \ln \frac{f(x)}{g(x)} \mathrm{d}x, \qquad x \in x^* \tag{2}$$

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.

III. NORMAL REGRESSION MODEL

If we assume normality of the linear regression model (1), we can consider the parameters to have Gauss-inverse-Wishart (GiW) distribution defined as follows [5]:

Proposition 1 (Gauss-inverse-Wishart pdf): The probability density functionof the Gauss-inverse-Wishart distribution has the form

$$GiW_\Theta(V, \nu) \equiv \frac{r^{-0.5(\nu+n+2)}}{I(V,\nu)} \exp\left\{\frac{-1}{2r}\begin{bmatrix}-1\\ \theta'\end{bmatrix}' V \begin{bmatrix}-1\\ \theta\end{bmatrix}\right\} \tag{3}$$

or

$$GiW_\Theta(L, D, \nu) \equiv \frac{r^{-0.5(\nu+n+2)}}{I(L,D,\nu)} \times \\ \times \exp\left\{\frac{-1}{2r}\left[(\theta-\hat{\theta})' C^{-1} (\theta-\hat{\theta}) + D_{LSR}\right]\right\} \tag{4}$$

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,

K. Dedecius is with the Department of Adaptive Systems, Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 4, 182 08 Prague, Czech Republic. dedecius@utia.cas.cz

V 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 coefficients
$\hat{\theta}$ is a least-squares (LS) estimate of θ
C is the covariance of LS estimate
D_{LSR} is the LS remainder
I stands for normalization integral

For the sake of generality, the time indices were omitted in the Proposition 1.

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

$$V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} = \\ = L'DL = \begin{bmatrix} 1 & 0 \\ L_1 & L_2 \end{bmatrix}' \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ L_1 & L_2 \end{bmatrix} \quad (5)$$

Here, the left upper-corner elements of the V and D matrices are nonnegative scalars, $D_1, V_{11} \in \mathbb{R}^+$. Recalling Proposition 1, the least-square estimate of parameters $\hat{\theta} \equiv L_2^{-1} L_1$ has the covariance $C \equiv L_2^{-1} D_2^{-1} (L_2^{-1})'$ and the least-square reminder $D_{LSR} \equiv D_1$.

IV. PARAMETER ESTIMATION AND FORGETTING

According to the Bayesian approach, the unknown model parameter θ_t 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. $f(\theta_t|d(t))$. Under the natural conditions of control [6] the Bayes rule for recurrent parameter estimation reads

$$f(\theta_t|d(t)) \propto f(y_t|\psi_t, \theta_t) f(\theta_t|d(t-1)) \quad (6)$$

where $\propto$ denotes proportionality, i.e. equality up to a constant factor.

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

$$f(\theta_{t+1}|d(t)) = \int_{\theta^*} f(\theta_{t+1}|d(t), \theta_t) f(\theta_t|d(t)) \, \mathrm{d}\theta \quad (7)$$

There are multiple ways how to obtain the posterior pdf in (7). A popular approach is to employ some forgetting method, enabling discounting of a (potentially) outdated information. For purpose of this paper, the exponential, alternative and partial forgetting were selected.

A. Exponential Forgetting

The exponential forgetting, also known as time-weighted least squares [7] or flattening the posterior probability density function [6], dominates the methods of solution of slowly time-variant parameters issue. This approach introduces just one new parameter $\lambda \in (0, 1]$ called forgetting factor, which is usually not lower than 0.95. This factor causes the flattening of the pdf by exponentation of it

$$f(\theta_{t+1}|d(t)) = [f(\theta_t|d(t))]^\lambda \quad (8)$$

1) Exponential Forgetting in Normal Model: If we suppose using the exponential forgetting method in normal model, then the probability density function is powered by the forgetting factor λ, which leads to the following time-update scheme:

$$V_t = \lambda V_{t-1} \quad (9)$$

$$\nu_t = \lambda \nu_{t-1} \quad (10)$$

Indeed, the forgetting step is as simple as the multiplication of the extended information matrix and the counter (number of degrees of freedom) by the forgetting factor.

B. Alternative Forgetting

Sometimes, the exponential forgetting is viewed as an optimization problem of 'balancing' two probability density functions f_1 and f_2. The unknown true pdf $\hat{f}$ describing the distribution of model parameters then equals to f_1 with probability λ, making the probability of the second pdf f_2 to be $1 - \lambda$. Then, the following proposition can be made [5][8].

Proposition 2: Let an unknown true probability density function f, describing the distribution of unknown parameters, be equal to pdf f_1 with probability $\lambda \in [0, 1]$ and to pdf f_2 with probability $1 - \lambda$. Let the pdfs f_1 and f_2 be mutually non-orthogonal and have the same support θ^*. Then, the relation

$$\hat{f}(\theta) \propto [f_1(\theta)]^\lambda [f_2(\theta)]^{1-\lambda} \quad (11)$$

describes the estimate of f, obtained as a solution of the optimization problem

$$\min_f [\lambda \mathsf{KL}(f||f_1) + (1 - \lambda) \mathsf{KL}(f||f_2)] \quad (12)$$

The proof can be found in [8]

In practical use, the pdf f_1 is usually the filtered one, obtained after the data update (6), while the another one (f_2) is any appropriate (preferably flat, e.g. the prior) pdf. The relation (11) defines the time update step.

1) Alternative Forgetting in Normal Model: As it was already mentioned above, the alternative forgetting (also known as stabilized [5] assumes use of an alternative information. For the sake of convenience, such an information takes the same form (distribution) as the latest information available. In the case of a normal regressive model, the alternative information is represented by a normal probability density function with own information matrix V_A and counter (degrees of freedom) ν_A. The time update (i.e. forgetting) has the following form

$$V_t = \lambda V_{t-1} + (1 - \lambda) V_A \quad (13)$$

$$\nu_t = \lambda \nu_{t-1} + (1 - \lambda) \nu_A \quad (14)$$

Just like in the case of a simple exponential forgetting, the factor $\lambda \in [0, 1]$ denotes the weight (probability).

C. Partial Forgetting

The above mentioned methods lack the 'explicit' ability to track the system with parameters which vary each with a different rate. To solve this case, the partial forgetting method [12] was developed.

The method of partial forgetting is based on an unknown random true multivariate parameter pdf ${}^{T}f(\theta|d(t)) = {}^{T}f(\theta_1,\dots,\theta_n|d(t))$, $n \in \mathbb{N}$. As this pdf is unknown and unavailable, it would theoretically be possible to consider a hyper-distribution describing it. However, such a distribution would be too complicated and unconvenient for our purpose, but it would be fully sufficient to take into account only its point estimates constructed on the basis of the hypotheses about the individual parameters behaviour. These hypotheses are given by the expectations as follows:

$$
\begin{aligned}
H_0 &: \mathsf{E}\left[{}^{T}f(\theta|d(t))|\theta,d(t),H_0\right] = f(\theta|d(t))\\
H_1 &: \mathsf{E}\left[{}^{T}f(\theta|d(t))|\theta,d(t),H_1\right] =\\
&\quad = f(\theta_2,\dots,\theta_n|\theta_1,d(t))f_A(\theta_1)\\
H_2 &: \mathsf{E}\left[{}^{T}f(\theta|d(t))|\theta,d(t),H_2\right] =\\
&\quad = f(\theta_1,\theta_3,\dots,\theta_n|\theta_2,d(t))f_A(\theta_2)\\
&\dots\\
H_n &: \mathsf{E}\left[{}^{T}f(\theta|d(t))|\theta,d(t),H_n\right] =\\
&\quad = f(\theta_1,\dots,\theta_{n-1}|\theta_n,d(t))f_A(\theta_n)\\
H_{n+1} &: \mathsf{E}\left[{}^{T}f(\theta|d(t))|\theta,d(t),H_{n+1}\right] =\\
&\quad = f(\theta_3,\dots,\theta_n|\theta_1,\theta_2,d(t))f_A(\theta_1,\theta_2)\\
H_{n+2} &: \mathsf{E}\left[{}^{T}f(\theta|d(t))|\theta,d(t),H_{n+2}\right] =\\
&\quad = f(\theta_2,\theta_4\dots,\theta_n|\theta_1,\theta_3,d(t))f_A(\theta_1,\theta_3)\\
&\dots\\
H_{2^n-2} &: \mathsf{E}\left[{}^{T}f(\theta|d(t))|\theta,d(t),H_{2^n-2}\right] =\\
&\quad = f(\theta_n|\theta_1,\dots,\theta_{n-1},d(t))f_A(\theta_1,\dots,\theta_{n-1})\\
H_{2^n-1} &: \mathsf{E}\left[{}^{T}f(\theta|d(t))|d(t),H_{2^n-1}\right] = f_A(\theta) \qquad (15)
\end{aligned}
$$

The pdf $f(\cdot)$ denotes the probability density function obtained after the data update (6), while the pdf f_A is any appropriate alternative. We use it to explicitly declare that one or more parameters have different distribution (e.g. because they vary). Usually we keep the same distribution and only change its moments. In the case of the normal distribution we can employ a flat pdf, causing the release of the parameters' values. One of possible sources of the flat pdf is the prior information.

Each of the hypotheses mentioned above has assigned a weight (probability) $\lambda_i, i = 1,\dots,2^n-1$ of becoming true during the time run. As each hypothesis represents an atomic random event, they altogether compose the whole probability space and the sum of their weights must be equal to one, $\sum_i \lambda_i = 1$ and $\lambda_i \in [0,1]$.

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

$$
\begin{aligned}
\mathsf{E}\left[{}^{T}f(\theta|d(t))|\mathcal{C}\right] &= \mathsf{E}\left[\mathsf{E}\left[{}^{T}f(\theta|d(t))|\mathcal{C},H_i\right]|\mathcal{C}\right] =\\
&= \sum_{i=0}^{2^n-1}\lambda_i \mathsf{E}\left[{}^{T}f(\theta|d(t))|\mathcal{C},H_i\right] \qquad (16)
\end{aligned}
$$

where $\mathcal{C} = \{\theta, d(t)\}$

We search for an approximative pdf $\tilde{f}(\theta|d(t))$ of the mixture (16) that belongs to the same family of distributions as the mixture components. Under general conditions, as a 'measure' of dissimilarity between two distributions, we use the Kullback-Leibler divergence described in the former part of the paper. Hence the approximative pdf could be selected as that one which minimizes the expected divergence between the mixture and itself

$$
\begin{aligned}
&\underset{\tilde{f}\in\tilde{f}^*(\theta|d(t))}{\arg\min}\ \mathsf{E}\left[\mathsf{KL}\left({}^{T}f\middle\|\tilde{f}\right)|\mathcal{C}\right] =\\
&= \underset{\tilde{f}\in\tilde{f}^*(\theta|d(t))}{\arg\min}\ \mathsf{E}\left[\int_{\theta^*} {}^{T}f(\theta|d(t))\ln\frac{{}^{T}f(\theta|d(t))}{\tilde{f}(\theta|d(t))}\,\mathrm{d}\theta|\mathcal{C}\right] =\\
&= \underset{\tilde{f}\in\tilde{f}^*(\theta|d(t))}{\arg\min}\int_{\theta^*}\mathsf{E}\left[{}^{T}f(\theta|d(t))|\mathcal{C},H_i\right]\ln\frac{1}{\tilde{f}(\theta|d(t))}\,\mathrm{d}\theta =\\
&= \underset{\tilde{f}\in\tilde{f}^*(\theta|d(t))}{\arg\min}\int_{\theta^*}\sum_{i=0}^{2^n-1}\lambda_i\mathsf{E}\left[{}^{T}f(\theta|d(t))|\mathcal{C},H_i\right]\times\\
&\qquad\times\ln\frac{1}{\tilde{f}(\theta|d(t))}\,\mathrm{d}\theta \qquad (17)
\end{aligned}
$$

Using the relation (17), 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.

1) Partial Forgetting in Normal Model: Here, with regard to the theoretical approach described above, the situation gets more complicated than in the case of the previous two forgetting methods. The derivation of the forgetting was thoroughly described in [12], to avoid the informational overhead in this paper, let's summarize just the result. First, note that we need to enumerate the hypotheses (15) and the related probability density functions. The successive step consists in evaluation of a mixture of these pdfs as given in (16) and its approximation (17) with respect to the distribution parameters. The results of this procedure is given by the following proposition.

Proposition 3: 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

$$
\hat{\tilde{\theta}} = \left(\sum_{i=1}^{n}\lambda_i\frac{\nu_i}{D_{LSR,i}}\right)^{-1}\cdot\left(\sum_{i=1}^{n}\lambda_i\frac{\nu_i}{D_{LSR,i}}\hat{\theta}_i\right) \qquad (18)
$$

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

$$\tilde{D}_{LSR} = \tilde{\nu} \cdot \left(\sum_{i=1}^{n} \lambda_i \frac{\nu_i}{D_{LSR,i}} \right)^{-1} \tag{19}$$

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

$$\tilde{C} = \sum_{i=1}^{n} \lambda_i C_i + \sum_{i=1}^{n} \lambda_i \frac{\nu_i}{D_{LSR;i}} \left[\left(\hat{\theta}_i - \hat{\tilde{\theta}} \right) \left(\hat{\theta}_i - \hat{\tilde{\theta}} \right)' \right] \tag{20}$$

- and the counter (degrees of freedom)

$$\tilde{\nu} = \frac{1 + \sqrt{1 + \frac{4}{3}(A - \ln 2)}}{2(A - \ln 2)} \tag{21}$$

where

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

$$- \sum_{i=1}^{n} \lambda_i \, \psi_0(0.5\nu_i) \tag{22}$$

The proof is omitted as it would be necessary to include the derivation of the Kullback-Leibler divergence of normal probability density functions, its motivation is given in [12]. 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. [9][10][11]).

V. EXPERIMENTS

This section brings tests based on one-step ahead predictions of the system output using the first-order autoregression model AR(1) modelling the system in the form

$$y_{t+1} = \theta_1 + \theta_2 y_t \tag{23}$$

where θ_1 is the absolute term and θ_2 is the dynamics, y_t is the system output in time t. The usually present normal uncorrelated white noise with zero mean was omitted. Because of the character of the AR(1) process model, only the non-oscillating processes with poles in the right halfplane were included to the test.

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.

If there was a need for an alternative, the prior information build up from the first few data was used. The exponential- and alternative-based estimators were run as defined, the partial forgetting hypotheses were selected the following way: the hypothesis H_0, employing the plain filtered pdf, was taken as the most probable one, the hypothesis H_3 consisting of the completely alternative pdf was decided to be very unlikely and assigned with weight of zero. The two remaining hypotheses weights' were searched with the genetic algorithms (GAs) and are shown in the tables below. The weight of H_0 can be calculated as a complement to unity.

A. Constant parameters

First, try to model the time series which is stationary and stable. This series was generated by the equation

$$y_{t+1} = 0.9 y_t - 0.2, \quad t = 1, 2, \ldots, 300 \tag{25}$$

with initial value $y_1 = 0$. This time series is shown in the Fig. 1.

Such a time series can be modelled both by a models with and without forgetting. In a very short time aspect, this series changes its statistics, however in a longer time it is stable. The forgetting methods only lead to faster "stabilization" of the parameter estimation. The prediction quality gives Table I. The alternative information for partial and alternative forgetting was built up from the first five data.

The prediction error is the least in the case of partial forgetting, which leads to the conclusion that the short-time process is better modelled by a model with released parameter (absolute term).

B. Time-varying dynamics

Now, we will try to model the series generated by the relation

$$y_{t+1} = (0.9 - 1/t) y_t + 2, \quad t = 1, 2, \ldots, 300 \tag{26}$$

initialized with $y_1 = 2$. Such a process has for each time instant t characteristic root inside the unit circle. However,

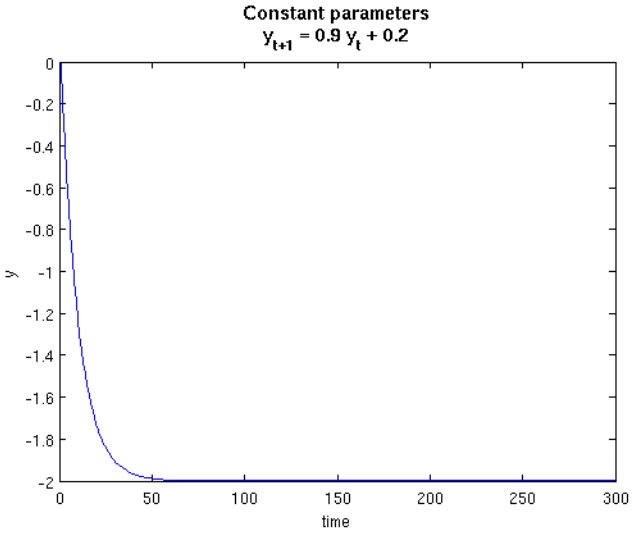

Fig. 1. Constant parameters.

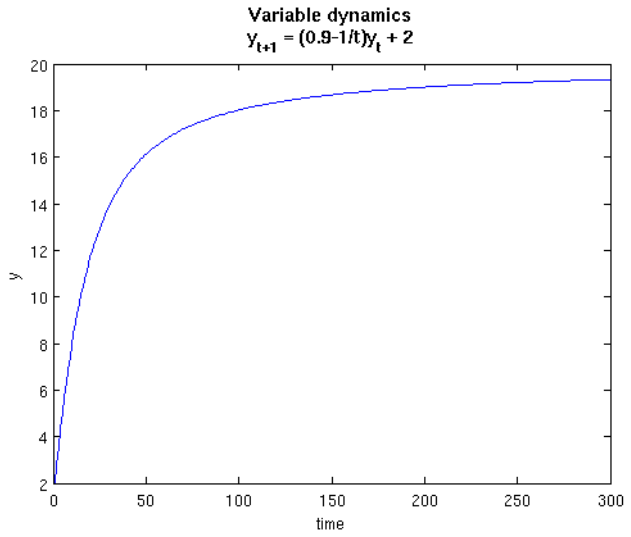

Fig. 2. Time-varying dynamics.

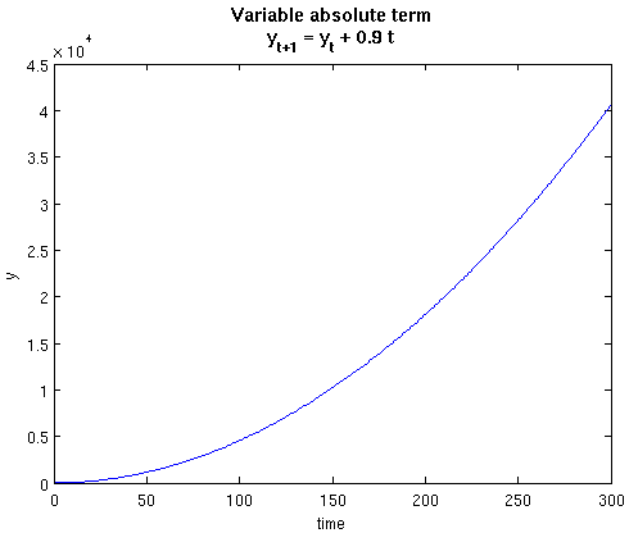

Fig. 3. Time-varying absolute term.

in a short term, the change of the nominator of the term y_t can be viewed to be quite significant and therefore making this particular example suitable for our purpose.

The course of the modelled data is depicted in the Fig. 2

Apparently, in this case the partial forgetting method led to a better prediction than the other two methods, thus its aim to track the parameters with different rate of change was fulfilled. According to the intuition the exponential forgetting based estimation produced the worst results. The results of the latter two methods could yet be improved if there were a more suitable alternative information. The alternative was built from the first five data.

C. Time-varying absolute term

$$y_{t+1} = y_t + 0.9t, \quad t = 1, 2, \ldots, 300 \tag{27}$$

with $y_1 = 2$. This process is a unit root process [13], as its single root is equal to one, hence it is nonstationary.

TABLE I

CONSTANT PARAMETERS: ONE STEP-AHEAD PREDICTION OF TIME SERIES.

Method	Weight(s)	RPE
Exponential	0.95	0.0501
Alternative	0.7569	0.0021
Partial	[0.0063, 0.5059]	0.0001

TABLE II

TIME-VARYING DYNAMICS: ONE STEP-AHEAD PREDICTION OF TIME SERIES.

Method	Weight(s)	RPE
Exponential	0.95	0.00336
Alternative	0.4078	0.00085
Partial	[0.1435, 0.6122]	0.00061

TABLE III

TIME-VARYING ABSOLUTE TERM: ONE STEP-AHEAD PREDICTION OF TIME SERIES.

Method	Weight(s)	RPE
Exponential	0.95	19.564e-05
Alternative	0.001	7.0712e-05
Partial	[0.0086, 0.6973]	6.436e-05

Its character is explosive, which in combination with the absolute term change makes it suitable for prediction with forgetting-based parameter estimation.

The course of the series is shown in the Fig. 3. The results of the prediction for all three methods shows Table III.

In this case the partial forgetting led again to the best prediction, however the difference between it and the alternative forgetting were relatively not very big. The alternative information was built from the first 10 data samples. The exponential forgetting led to the worst prediction quality (in comparison to the other two methods) as one would intuitively expect.

D. Time-varying absolute term and dynamics

The last test demonstrates the prediction ability for a system model with varying both regression coefficients. The system is simulated with the following time series

$$y_{t+1} = (1 + 10^{-4}t)y_t + 10^{-3}t, \quad t = 1, 2, \ldots, 300 \tag{28}$$

initialized with $y_1 = 0$. The single root of the process lays outside the unit circle in the right halfplane, hence the process is nonstationary. It is depicted in the Fig. 4.

The alternative information for the alternative and partial forgetting was made up from the first five data. The partial forgetting led to a bit better prediction than the alternative one.

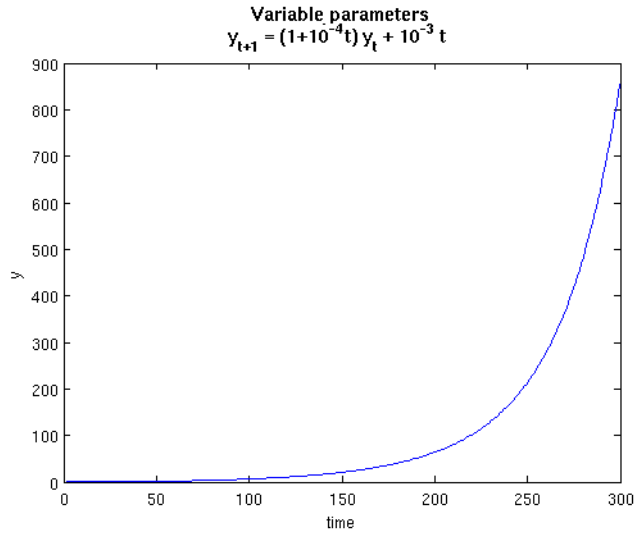

Fig. 4. Time-varying both parameters.

TABLE IV

TIME-VARYING BOTH PARAMETERS: ONE STEP-AHEAD PREDICTION OF TIME SERIES.

Method	**Weight(s)**	**RPE**
Exponential	0.95	33.478e-05
Alternative	0.001	9.789e-05
Partial	[0.0020,0.2490]	9.216e-05

VI. CONCLUSION

The paper brought a brief description and comparison of three selected forgetting techniques, that can be used in estimation of slowly varying parameters of nonstationary stochastic systems. In the first part, they were shortly introduced and described with references to the literature. The second part of the paper compares their use in some scenarios.

According to the tests given above, the best prediction quality can be obtained with the partial forgetting, which is more or less comparable to the alternative forgetting. The reasoning leads to the conclusion that the exponential forgetting, which is very fast, is usefull in the realtime systems, when the time cost is the key element.

The main drawback of the partial forgetting consists in the computational cost. While the exponential and alternative forgettings are based on a simple exponentation of the probability density function, which in the case of normality leads to a simple multiplication of the appropriate terms, the partial forgetting method needs a construction of hypotheses and nontrivial approximation of a mixture of the hypothetical pdfs. Moreover, the search for optima represents a multidimensional problem and beside the search for optimal weights we can yet search for optimal alternative.

VII. FUTURE WORK

This paper brought a brief comparison of three selected forgetting techniques in the first-order autoregressive models, used for one-step ahead predictions of stochastic processes with a non-oscilating course. The future work will consist in a deeper analysis of abilities of the forgetting methods and development of a methodology for a selection of the optimal weights and alternative information for the partial forgetting. Currently, the underlaying environment for practical testing and developement is formed by the traffic situation in Prague.

REFERENCES

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

[2] Najim, K., Ikonen, E., Daoud, AK. *Stochastic processes: estimation, optimization and analysis*. Butterworth-Heinemann,l 2004.

[3] Pollock, DSG, *A Handbook of Time-series Analysis, Signal Processing and Dynamics*. Academic Press, 1999.

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

[5] Kárný, M. et al. *Optimized Bayesian Dynamic Advising*, Springer, 2005.

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

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

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

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

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

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

[12] Dedecius, K. et al. *Partial Forgetting. A new method for tracking time-variant parameters*, UTIA AV ČR. Research Report 2249, [http://library.utia.cas.cz/separaty/2009/AS/dedecius-partial%20forgetting.%20a%20new%20method%20for%20tracking%20time-variant%20parameters.pdf], 2009.

[13] Hamilton, J.D. *Time series analysis*. Princeton University Press, 1994.