Partial Forgetting in Autoregression Models

Kamil Dedecius
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 4
CZ-182 08 Prague, Czech Republic
dedecius@utia.cas.cz

***Abstract*— The assumption of constant parameters of the autoregression model sometimes fails, as the parameters may vary in time. If the parameters vary slowly, the problem is often solved using various forgetting methods like exponential forgetting, linear forgetting etc. However, most of them work on the model parameters probability density function with one common forgetting rate. In the case of different variability of individual parameters, these methods might fail. The developed partial forgetting method gives a new approach, which solves this problem. It releases individual parameters and allows them to change with different rates.**

I. INTRODUCTION

If a mathematical system model with fixed structure has to reflect the modeled reality in time evolution, it is often necessary to leave the assumption of constant model parameters and let them vary. The tracking of slowly varying parameters then employs various techniques of forgetting of the obsolete information. Apart from windowing estimation from a batch of recent data [1], the most popular methods are based on exponential [2][4] or linear [5] forgetting (e.g. the directional forgetting [6][7]). Some authors propose using the forgetting-factor least squares algorithm [8] and its extension – a finite-data-window least squares algorithm with a forgetting factor [9]. Another group of methods is based on the Kalman filter estimating the parameters of a linear model with normal noise [10][11] and its modifications like H_∞ filter or extended Kalman filters [12][13].

This paper tries to give a contribution to this field of interest. It introduces a partial forgetting method, allowing to estimate slowly varying parameters when they change each with a different rate.

II. MATHEMATICAL SYSTEM MODEL

Let us consider a discrete stochastic system observed at time instants $t = 1, 2, \ldots$ Let this system have directly manipulated inputs u_t affecting the system outputs y_t and introduce the data vector $d_t = (u_t, y_t)$. Then the time ordered sequence of these vectors $d(t) = (d_1, d_2, \ldots, d_t)$ describes the development of the system inputs and outputs in time, i.e. from the beginning time instant 1 until time t.

The model output y_t dependent on the previous data $d(t-1)$ and the current input u_t defines the conditional probability density function (pdf)

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

where $\theta \in \Theta$ stands for a model parameter (possibly multivariate); Θ is a vector of all parameters, in the case of model normality including the noise variance.

The mathematical system model has often the form of a regression model

$$y_t = \sum_{i=1}^{n} a_i y_{t-i} + \sum_{j=0}^{m} b_j u_{t-j} + c_t + e_t \tag{2}$$

where the regression parameters $a_i, b_j, c_t \in \Theta, i = 1, \ldots, n; j = 0, \ldots, m$. The term e_t denotes the white noise, i.e. normally distributed uncorrelated random variable with zero mean and constant variance, $e_t \sim \mathrm{N}(0, r)$. This variable causes the non-systematic dispersion of the (measured, predicted...) system output.

III. PARAMETER ESTIMATION

According to the Bayesian statistics the model parameter θ is a random variable, hence it is possible to describe it with a probability density function, conditioned by the data available at the current time instant t.

$$f(\theta|u_t, d(t-1)) \tag{3}$$

Under the natural conditions of control, when the information about the unknown parameter θ is derived only from past data and the variable u_t

cannot bring any additional information about it, the following simplification comes true [4]

$$f(\theta|u_t, d(t-1)) = f(\theta|d(t-1)) \tag{4}$$

Then, using the Bayes rule under natural conditions of control, the parameter is estimated as follows

$$f(\theta|d(t)) \propto f(y_t|\psi_t, \theta) f(\theta|d(t-1)) \tag{5}$$

or in the batch form

$$f(\theta|d(t)) \propto f(\theta|d(0)) \prod_{\tau=1}^{t} f(y_\tau|u_\tau, d(\tau-1), \theta) \tag{6}$$

where ψ_t is the regression vector, $f(\theta|d(0))$ is the initial knowledge about the parameter pdf, i.e. the probability distribution modelling the prior uncertainty about the parameter θ for $t = 0$ before the observed data $d(t)$ are incorporated. The product is the likelihood function $L_t(\theta, d(t)) = \prod_{\tau=1}^{t} f(y_\tau|u_\tau, d(\tau-1), \theta)$.

The last formula (6) has an important property – it is recursive. It means, that the parameter values can be estimated in a loop and the distribution is determined just by the history of the data.

A. Estimation of slowly varying parameters

The case of slowly varying parameters supposes, that the "new" (in time) parameter value θ lays close to the previous value. This assumption is crucial to many forgetting methods, as they cannot catch the rapid changes of parameter value. The problematics of fast varying parameters is solved e.g. in [14][15].

There is a couple of ways how to cope with the slowly varying parameters. One possible approach is to alter the recursive parameter estimation relation (6), so that it admits slow permanent changes of parameter estimates. Such an approach is called time weighting, time discounting or simply forgetting. In this case, the parameter estimation (6) changes to the parameter tracking, which can be divided into three basic steps:

1) Collecting the newest data d_t.
2) Performing the data update of the parameter probability density function.
3) Performing the time update in the form of forgetting.

The data update step is equivalent to the relation (5). It is sometimes written in the following indexed form

$$\begin{aligned} f_{t-1|t}(\theta|d(t)) \propto\, & f(y_t|u_t, d(t-1), \theta) \times \\ & \times f_{t-1|t-1}(\theta|d(t-1)) \end{aligned} \tag{7}$$

where the multiindex $\cdot|\cdot$ describes in order the 'time index' of parameters separated with the $|$ sign from the 'time index' of data.

The time update works on the data-updated parameter pdf. In the case of constant parameters, it is only formal addition of 1 to the time index of appropriate variables, hence $(t-1)+1$. In the case of forgetting, the time update takes various forms, e.g. in exponential forgetting it is equivalent to flattening of the parameter pdf [4]

$$f_{t|t}(\theta|d(t)) = [f_{t-1|t}(\theta|d(t))]^\lambda, \quad \lambda \in [0, 1] \tag{8}$$

where the forgetting factor λ is usually not lower than 0.95, $\lambda \geq 0.95$.

The main problem of most forgetting methods consists in the fact, that all parameters are being forgotten with one common rate. For instance, the equation (8) applies factor λ on the whole parameter pdf of any dimension. If this pdf is two-dimensional ($\theta = (\theta_1, \theta_2)$) and one parameter varies quickly than the other, the choice of λ is complicated. Faster forgetting helps tracking of the faster changing parameter, while the information about the other one is being lost. Slower forgetting maintains information about the slower parameter, but the information about the other parameter gets inaccurate (outdated).

IV. Partial forgetting

The basic idea of the partial forgetting, allowing individual parameters tracking, is based on an unknown and random true (multidimensional) parameter probability density function $f_T(\theta|d(t))$ (T denotes 'true'), ideally describing the actual distribution of parameters. However it is unknown to us, but for our purpose it is sufficient to construct only its point estimates, given by hypotheses about the individual parameters behaviour. Each of these hypotheses describes which configuration of parameters vary and with which probability and induces one parameter pdf – a point estimate of the true pdf $f_T(\theta|d(t))$. These estimates produce a mixture of pdfs, describing the random true parameter pdf, and the goal is to find its best approximation $\tilde{f}(\theta|d(t))$. This approximant should minimize the expectation of distance between the mixture and itself $\mathsf{E}\left[d(f_T, \tilde{f})\right] \rightarrow \min$. For this purpose we use the Kullback-Leibler divergence in the form

$$\mathsf{D}\left(f_T(\theta)\middle|\middle|\tilde{f}(\theta)\right) = \int f_T(\theta) \ln \frac{f_T(\theta)}{\tilde{f}(\theta)} \mathrm{d}\theta \tag{9}$$

A. AR(1) model

For the sake of simplicity let's consider just a first order autoregression model AR(1). The transition to higher orders is due to computational efficiency still in development stage.

The first order autoregression model is a derivate of the regression model (2), formally describing a system with output influenced by the previous output, an absolute term and the model noise.

$$y_t = \theta_1 + \theta_2 y_{t-1} + e_t \tag{10}$$

The AR(1) Gaussian model has three parameters – the regression model parameters θ_1, θ_2 and the noise variance r, forming the model parameter vector Θ

$$\Theta = (\theta, r) = (\theta_1, \theta_2, r) \tag{11}$$

B. Distribution

The subject of partial forgetting method is based on an unknown true multivariate parameter probability density function $f_T(\theta_1, \theta_2, r|d(t))$. The assumption of model normality leads to Gauss-inverse-Wishart distribution of the parameters, $f_T \sim GiW_\Theta(V, \nu)$ given by the following definition [16].

Definition 1: The probability density function of the Gauss-inverse-Wishart (GiW) distribution has the following form

$$GiW_\Theta(V, \nu) \equiv \frac{r^{-0.5(\nu+n+2)}}{I(V, \nu)} \times \\ \times \exp\left\{ \frac{-1}{2r} \begin{bmatrix} -1 \\ \theta' \end{bmatrix}' V \begin{bmatrix} -1 \\ \theta \end{bmatrix} \right\} \tag{12}$$

or

$$GiW_\Theta(L, D, \nu) \equiv \frac{r^{-0.5(\nu+n+2)}}{I(L, D, \nu)} \times \\ \times \exp\left\{ \frac{-1}{2r} \left[(\theta - \hat{\theta})' C^{-1} (\theta - \hat{\theta}) + D_{LSR} \right] \right\} \tag{13}$$

where the individual terms have the following meaning:

ν – degrees of freedom,
ψ – regression vector
θ – vector of regression parameters
n – length of the regression vector,
r – variance of model noise,
V_t – 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 $L'DL$ factorization the terms L and D are obtained. The D matrix upper-corner term is the least square reminder D_{LSR}.
C – the covariance matrix
I – normalization integral

In our particular case of the AR(1) model the pdf has the form

$$f(\theta_1, \theta_2, r|d(t)) \propto r^{-0.5(\nu+n+1)} \times \\ \times \exp\left\{ \frac{-1}{2r} \begin{bmatrix} -1 \\ \theta_1 \\ \theta_2 \end{bmatrix}' V_t \begin{bmatrix} -1 \\ \theta_1 \\ \theta_2 \end{bmatrix} \right\}, \quad t = 1, 2, \ldots \tag{14}$$

C. Hypotheses

The true parameter pdf $f_T(\theta_1, \theta_2, r|d(t))$ is unknown and random due to the variability of individual model parameters. It is possible to consider a distribution describing the pdf f_T, which is too complicated. For our purpose, it suffices to take into account the point estimates according to the individual hypotheses about the parameters' behaviour. These estimates are given by the expectations. In the presented case, we obtain the following four hypotheses:

$$\begin{aligned} H_1 &: \mathsf{E}\left[f_T(\theta_1, \theta_2, r|d(t))|\theta_1, \theta_2, r, d(t), H_1\right] = \\ &= f(\theta_1, \theta_2, r|d(t)) \\ H_2 &: \mathsf{E}\left[f_T(\theta_1, \theta_2, r|d(t))|\theta_1, \theta_2, r, d(t), H_3\right] = \\ &= f(\theta_2|\theta_1, r, d(t)) f_A(\theta_1, r) \\ H_3 &: \mathsf{E}\left[f_T(\theta_1, \theta_2, r|d(t))|\theta_1, \theta_2, r, d(t), H_2\right] = \\ &= f(\theta_1|\theta_2, r, d(t)) f_A(\theta_2, r) \\ H_4 &: \mathsf{E}\left[f_T(\theta_1, \theta_2, r|d(t))|\theta_1, \theta_2, r, d(t), H_4\right] = \\ &= f_A(\theta_1, \theta_2, r) \end{aligned} \tag{15}$$

where f comes from the filtration (6), while the alternative pdf f_A is any appropriate alternative, preferably flat pdf, e.g. the initial (prior) one. The hypotheses employ the alternative pdf if the parameter varies, otherwise the pdf for particular parameter (or parameters) stay unchanged.

All four hypothetic densities have the Gauss-inverse-Wishart distribution. The first hypothesis uses the density obtained after the data update as an expectation of the true pdf, while the last uses a completely alternative density. The hypotheses H_2 and H_3 employ the conditional parts from the data-updated pdf (7), but their 'marginal' parts describing the individual regression parameters (and noise variance) are changed with any appropriate alternative. The method for a joint GiW pdf decomposition to conditional and marginal pdfs can be found in [16].

Each of these hypotheses has assigned a weight, characterized as a probability of becoming true during the time run. Let these weights be $\lambda_i \in [0,1], \quad i = 1, \ldots, 4$ and $\sum_{i=1}^{4} \lambda_i = 1$. The true parameter pdf may be expressed as a convex combination of the hypothetic densities

$$f_T(\theta_1, \theta_2, r|d(t)) = \\ = \sum_{i=1}^{4} \lambda_i \mathsf{E}\left[f_T(\theta_1, \theta_2, r)|\theta_1, \theta_2, r, d(t), H_i\right] \quad (16)$$

Hence, we obtained a mixture of four GiW density functions, which describes the reality.

D. Mixture approximation

As written above, the mixture (16) obtained as a convex combination of GiW pdfs should be approximated by a single pdf $\tilde{f}$, coming from the same distribution. This approximation is done by minimization of the Kullback-Leibler divergence between it and $\tilde{f}$.

The Kullback-Leibler divergence of two GiW distributions is given by the following lemma [16]:

Lemma 1: Let's have two distributions with probability density functions f and $\tilde{f}$. The Kullback-Leibler divergence of these two functions has the following form

$$\mathsf{D}\left(f \middle|\middle| \tilde{f}\right) = \ln \frac{\Gamma(0.5\tilde{\nu})}{\Gamma(0.5\nu)} - 0.5 \ln |C\tilde{C}^{-1}| + \\ + 0.5\tilde{\nu} \ln \frac{D_{LSR}}{\tilde{D}_{LSR}} + 0.5(\nu - \tilde{\nu})\psi_0(0.5\nu) - 0.5n - \\ - 0.5\nu + 0.5\mathrm{Tr}\left(C\tilde{C}^{-1}\right) + 0.5\frac{\nu}{D_{LSR}} \times \\ \times \left[\left(\hat{\theta} - \hat{\tilde{\theta}}\right)' \tilde{C}^{-1} \left(\hat{\theta} - \hat{\tilde{\theta}}\right) + \tilde{D}_{LSR}\right] \quad (17)$$

where $\psi_0(\cdot)$ denotes the digamma function, i.e. the first logarithmic derivative of the gamma function. The proof is not trivial and can be found in [16].

This lemma, applied directly on f_T and unknown optimal estimate $\tilde{f}$, helps us to find the minimally divergent GiW distribution defined by parameters $\tilde{\theta}$, the least-squares reminder $\tilde{D}_{LSR}$, the covariance matrix $\tilde{C}$ of the least-square estimate of parameters $\tilde{\theta}$ and the counter describing the degrees of freedom $\tilde{\nu}$.

The found parameters of the minimizing Gauss-inverse-Wishart pdf are:

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

$$\hat{\tilde{\theta}} = \left(\sum_{i=1}^{4} \lambda_i \frac{\nu_i}{D_{LSR,i}}\right)^{-1} \times \\ \times \left(\sum_{i=1}^{4} \lambda_i \frac{\nu_i}{D_{LSR,i}} \hat{\theta}_i\right) \quad (18)$$

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

$$\tilde{D}_{LSR} = \tilde{\nu} \cdot \left(\sum_{i=1}^{4} \lambda_i \frac{\nu_i}{D_{LSR,i}}\right)^{-1} \quad (19)$$

- $\tilde{C}$ – the covariance matrix

$$\tilde{C} = \sum_{i=1}^{4} \lambda_i C_i + \\ + 2\left(\frac{\sum_{i=1}^{4} \lambda_i \cdot \frac{\nu_i}{D_{LSR,i}} \hat{\theta}_i}{\sum_{i=1}^{4} \lambda_i \cdot \frac{\nu_i}{D_{LSR,i}}}\right) \times \\ \times \left(\frac{\sum_{i=1}^{4} \lambda_i \cdot \frac{\nu_i}{D_{LSR,i}} \hat{\theta}_i}{\sum_{i=1}^{4} \lambda_i \cdot \frac{\nu_i}{D_{LSR,i}}}\right)' \times \\ \times \sum_{i=1}^{4} \lambda_i \frac{\nu_i}{D_{LSR,i}} + \\ + \sum_{i=1}^{4} \hat{\theta}_i \hat{\theta}_i' \cdot \lambda_i \frac{\nu_i}{D_{LSR,i}} \quad (20)$$

- and the counter (degrees of freedom)

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

where

$$A = \ln\left(\sum_{i=1}^{4} \lambda_i \frac{\nu_i}{D_{LSR,i}}\right) + \\ + \sum_{i=1}^{4} \lambda_i \ln D_{LSR,i} - \sum_{i=1}^{4} \lambda_i \ \psi_0(0.5\nu_i) \quad (22)$$

The obtained parameters $\tilde{\theta}, \tilde{D}_{LSR}, \tilde{C}$ and $\tilde{\nu}$ might be used to construct the minimally divergent Gauss-inverse-Wishartian probability density function $\tilde{f}$ (13), approximating the mixture (16). There arises only minor need of approximation, related to the digamma function ψ_0 in the expression of the counter $\tilde{\nu}$ (21) or (22) respectively. A couple of approximation methods and algorithms exists, see e.g. [18][19][20].

E. Tests

The derived method of partial forgetting was tested on traffic data representing road intensities measured in Prague, Czech republic. The sampling period of the measurement was 5 minutes and the data window contains 600 samples. The course of the selected intensities is shown in the figure 1.

The traffic system was modelled with a first order autoregression model AR(1)

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

The hypotheses about the parameter distribution are equivalent to those given in (15) with probabilities (weights) $\lambda_1, \lambda_2, \lambda_3, \lambda_4$. As a source of alternative pdf(s) the prior pdf obtained from preceding 10 data samples was used. Using the relations (18) – (22), the Gauss-inverse-Wishart distribution parameters were calculated and the best weights searched. The optimization criterion was the minimization of the prediction error.

Definition 2 (relative prediction error): Let y be the true random vector and y_p the predicted random vector, both with length n. Let us denote s the standard deviation of y. We define the relative prediction error in the form

$$RPE = \frac{1}{s}\sqrt{\frac{\sum(y-y_p)^2}{n}} \tag{24}$$

The best hypotheses' weights found with a MATLAB software were

- $\lambda_1 = 0.99$
- $\lambda_2 = 0$
- $\lambda_3 = 0.01$
- $\lambda_4 = 0$

giving the prediction error RPE = 0.0525.

To have a comparison, the same data were predicted with a first order autoregression model AR(1) with the most popular forgetting method – the exponential forgetting (for the time update formula see equation (8)). In this case, the best forgetting rate found was $\lambda = 0.985$ (RPE = 0.1576).

Figures 2 and 3 show the course of prediction residuals $y - y_p$ for partial and exponential forgetting method, respectively. The partial forgetting leads to smaller and less biased residuals than the exponential forgetting.

Figure 4 shows the course of both regression parameters θ_1, θ_2. The real traffic intensity course could be best predicted with variable absolute term – parameter θ_1.

V. CONCLUSIONS AND FUTURE WORKS

The partial forgetting method is designed for tracking of slowly varying parameters of linear stochastic processes when the individual parameters vary with different rates. It is based on hypotheses about the individual parameters variability, introducing the point estimates of the true parameter probability density functions. A convex

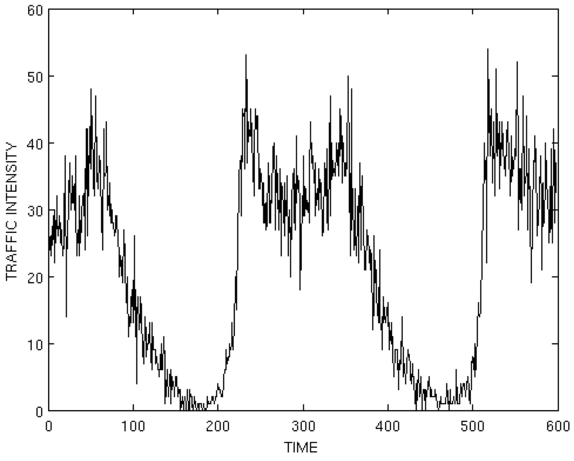

Fig. 1. Real data course

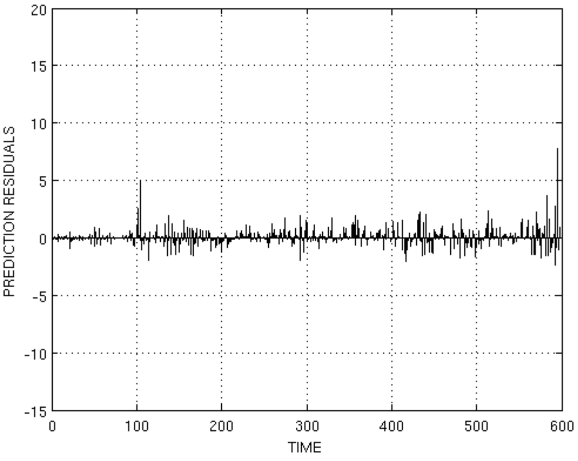

Fig. 2. Prediction residuals (partial forgetting)

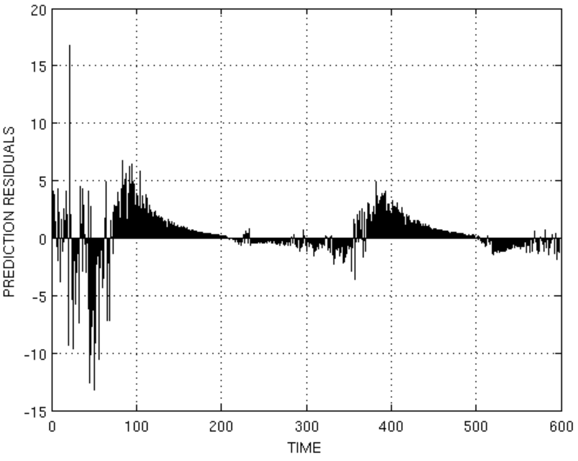

Fig. 3. Prediction residuals (exponential forgetting)

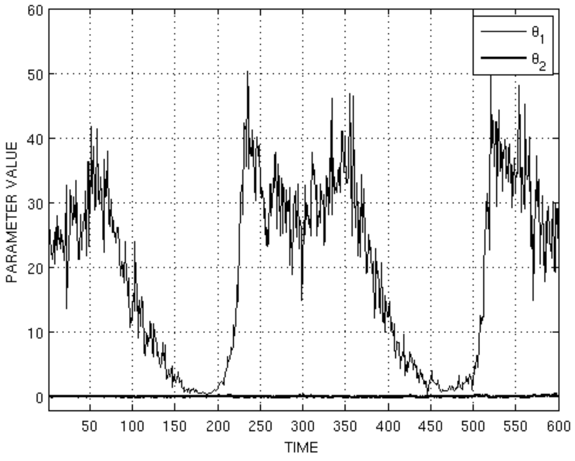

Fig. 4. Time evolution of parameters (partial forgetting)

combination of these pdfs gives a mixture of densities, which is approximated by the minimally divergent (in the Kullback-Leibler divergence sense) pdf from the same distribution as the true parameter pdf. The resulting pdf represents the best available description of the regression parameters distribution and is convenient e.g. for prediction purposes.

As the exponential forgetting method is the most popular approach to slowly varying parameters (many other methods are derived from it), the partial forgetting was tested and compared to it. The test on real data sample has shown that the developed method gives better results. However, it has some drawbacks, consisting in computationally more demanding optimization (search for rates) and complications with regression models of higher order. Both these problems are 'solving in progress'.

VI. ACKNOWLEDGEMENT

The author thanks to Dr. M. Kárný and I. Nagy from Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Prague, Czech republic, for their kind guidance and supervision during the research.

The research was partly supported by the Czech-Slovenian project KONTAKT MEB090607.

REFERENCES

[1] Middleton, R.H. et al., *Desing issues in adaptive control*, IEEE Trans. Automatic Control, vol. 33, pp. 50-58, 1988

[2] Jazwinski, A.H., *Stochastic Processes and Filtering Theory*. Academic Press, New York, 1970.

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

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

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

[6] Cao, L., Schwartz, H., *Directional forgetting algorithm based on the decomposition of the information matrix*, Automatica, vol. 36, no. 11, pp. 17251731, 2000.

[7] Kulhavy, R., Kárný, M., *Tracking of slowly varying parameters by directional forgetting*, In Preprints of the 9th IFAC World Congress, Budapest, Vol. X, pp. 78-83.

[8] Ding, F., Chen, T., *Performance bounds of forgetting factor least squares alogrithm for time-varying systems with finite measurement data*, IEEE Transactions on Circuits and Systems-I: Regular Papers 52 (3), pp. 555-566, 2005.

[9] Ding, F., Xiao, Y., *A finite-data-window least squares algorithm with a forgetting factor for dynamical modeling*, Applied Mathematics and Computation, vol. 1 (43), pp. 184–192, 2007.

[10] Kalman, R.E., Bucy, R.S. *New Results in Linear Filtering and Prediction Theory*. 1961

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

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

[13] Haykin, S., *Adaptive Filter Theory (3rd ed.)*, Prentice Hall, 1996.

[14] Li, Z., *Parameter tracking for stochastic time-varying systems*, in Proceedings of the 3rd World Congress on Intelligent Control and Automation, vol 3, pp.2248-2253. 2000

[15] Li, Z., *Discrete-time adaptive control for linear fast time-varying systems*, IEEE Trans. Aut, Contr. AC-32, 1987, pp.444-447.

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

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

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

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

[20] 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