Partial Forgetting in Estimation of Regression Models

Ivan Nagy, Lenka Pavelková, Kamil Dedecius

October 6, 2007

1 Introduction

Exponential forgetting is a well known tool for coping with slowly varying parameters during estimation [1]. It avoids modeling of the time development of the parameters, which use to be difficult, and it substitutes modeling by simple gradual suppression of the the importance of older data items to the values of estimated parameters. Thus, changing values of the parameter estimates can break through.

The technique of exponential forgetting is a big help in estimation of unknown model parameters. However, to use it, one has to be careful. If the identified system is well excited, so that the measured data which are used for estimation currently bring information about the system, the estimation can run with a relatively small forgetting coefficient and everything is OK. However, if some modes of the system give, from some reasons, data which are not informative, then the information previously obtained is forgotten, new one does not come and the process of estimation breaks down. That is why, a careful dealing with the forgetting coefficient is recommended.

In addition to what has been spoken above, the situation especially in real applications is still more complicated. It can be seen, that some parameters change and some do not or they change too, but negligibly. So, if we desire good following of the parameters values on condition of a careful dealing with the forgetting coefficient, we find that it is necessary to use several forgetting factors and to forget only those parameters that really change.

The problem of differentiated forgetting for systems, whose parameters partially slowly change and partially do not change is the topic of this paper.

Notation

In the context of system theory and automatic control we introduce the following notation:

If x is the studied signal (variable), then

x	is a sequence of random variables indexed by discrete time,
x_t	is a value of x at discrete time instant t,
$x_{t:\tau}$	denotes items from time t to τ (for $t \leq \tau$ is empty),
$x(t)$	is a subsequence of x involving items from the beginning up to the index t.

Problem solved

On the background of our problem is estimation of a dynamic regression model with a vector of unknown parameters Θ on the basis of data evidence whose item measured at time instant t is denoted by d_t. If the parameter Θ is changed, we have to consider it a state.

The problem of state estimation (for state space model with known parameters) is analytically solved in running time by the well known Kalman filter [3]. This filter evolves the parameter probability density function (pdf) from some prior one by embedding information extracted from currently measured data. It has two steps: (i) data update - filtration and (ii) time update - prediction. If the state is represented by slowly varying parameters, the prediction step is substituted by forgetting. In this paper we will consider just the "prediction step" in which the forgetting is performed.

Problem formulation

Let us have a vector of parameters $\Theta = [\Theta_1, \Theta_2, \cdots, \Theta_n]$. Some items of this vector change in time and the remaining stay constant. We have two descriptions of the parameter

- $f_1(\Theta)$ which is valid if the parameter is constant (it. is e.g. the pdf from the previous step of estimation), and
- $f_2(\Theta)$ which holds for the case that Θ is changing (which is some flat pdf e.g. the prior one).

Our task is to construct a description $\hat{f}(\Theta)$ which will describe the parameter Θ if we do not exactly know which items of Θ change and which are constant.

2 Preliminaries

The goal of this paper is to introduce a new forgetting scheme that can be used in an estimation of non-stationary dynamic systems structurally described by normal regression model with slowly varying parameters.

Basic principle of the Bayesian estimation

Let us consider a dynamic system with scalar valued output y_t and suppose it can be described by a regression type model of the form

$$f\left(y_t|\psi_t,\theta\right) = N\left(\theta'\psi_t,\sigma^2\right) \tag{1}$$

where

- ψ_t is a regression vector (containing data on which y_t depends),
- θ is a vector of regression parameters, corresponding to ψ_t,
- σ^2 is a variance of the noise e_t affecting y_t.

As we suppose just slow changes of the model parameters which cannot be modeled as a stochastic process in time, we do not assign them the time index.

According to the Bayesian approach to the estimation, see [2], to be able to compute the (posterior) parameter description at time t, we need to have the prior one in a form of conditional probability density function (pdf)

$$f\left(\Theta|d\left(t-1\right)\right), \tag{2}$$

where

- Θ is a set of all model parameters, i.e. $\left\{\theta,\sigma^2\right\}$ and
- $d\left(t-1\right)$ denotes all data $d_\tau = \left[y_\tau,u_\tau\right]'$ from the beginning of estimation up to the time $t-1$.

The Bayes rule reads

$$f\left(\Theta|d\left(t\right)\right) \propto f\left(y_t|\psi_t,\theta\right) f\left(\Theta|d\left(t-1\right)\right) \tag{3}$$

starting with a very prior pdf $f\left(\Theta|d\left(0\right)\right)$, where $d\left(0\right)$ denotes prior information acquired before the start of the estimation.

Remark: *The Bayes rule respects the natural conditions of control, introduced in [cit]*

Estimation with forgetting

A classical "exponential" forgetting can be obtained by the indicated modification of the Bayes rule

$$f\left(\Theta|d\left(t+1\right)\right) \propto f\left(y_{t+1}|\psi_{t+1},\theta\right) f\left(\Theta|d\left(t\right)\right)^{\lambda} \tag{4}$$

where $\lambda \in (0,1)$ is the forgetting rate.

Remark: *The exponentiation of the prior pdf means flattening of the posterior pdf which represents forgetting the information brought by historical data. As the Bayes rule is recursive in time, the older is the information the more is forgotten.*

3 Partial forgetting

In this section we will approach the main topic of this paper which is "time update" of estimated parameter in case when some of its items change and the rest of them does not.

Let us denote $\Theta = [\Theta_1, \Theta_2, \ldots, \Theta_n]$ parameters of a regression model (involving regression coefficients θ and noise variance σ^2).

Further, we suppose, we are in the process of estimation and we already performed the filtration step, in which the information brought by data has been used for data update of the pdf of parameters. We obtained the filtered parameter pdf - let us denote it $f_T(\Theta)$. Besides this pdf, we have some flat (e.g. prior) one, which carries very little information - we denote if $f_A(\Theta)$.

Now, we suppose, that some items of Θ change and we would like to use individual forgetting rate for each parameter item. That is why, we formulate the hypotheses $H_i,\ i = 1, 2, \ldots, 2^n$ (n is the number of parameter items). The i-th hypothesis H_i claims

> H_i: The i-th combination of parameters changes, the rest of them stays unchanged during the prediction step of the estimation.

The i th combination $c^{n;i}$ of parameters is defined as n-vector of items of a binary number whose value is equal to i and whose coefficients that are equal to one point to the parameter items that change.

Example: For a model with 4 parameter items we have:

$$\Theta = [\Theta_1, \Theta_2, \Theta_3, \Theta_4]',$$

the hypothesis H_5 claims that Θ_2 and Θ_4 change, because in binary form the number 5 is represented by $c^{4;5} = [0, 1, 0, 1]$.

As we do not know which of the hypothesis is true, we assign them probabilities. Thus the hypotheses H_i is supposed to be true with the probability λ_i.

To be able to deal with individual parameter items, we decompose the parameter description which is the joint pdf of the unknown parameter into a product of individual conditional ones

$$f(\Theta) = f(\Theta_n|\Theta_{1:n-1})\, f(\Theta_{n-1}|\Theta_{1:n-2}) \cdots f(\Theta_1) \tag{5}$$

In this way, also the pdfs f_T and f_A are decomposed.

Now, if we knew that the hypothesis H_i is true, then in the prediction step of identification the parameter pdf would be constructed such that for its decomposition (5) it holds:

$$\begin{aligned} &f(\Theta_i|\Theta_{1:i-1}) = \\ &\quad = f_A(\Theta_i|\Theta_{1:i-1}), \quad \text{if } \Theta_i \text{ changes;} \\ &\quad = f_T(\Theta_i|\Theta_{1:i-1}), \quad \text{if } \Theta_i \text{ does not change.} \end{aligned}$$

So, if the hypothesis H_i with the combination $c^{n,i}$ of changing parameter items is valid then the true parameter pdf is

$$f^{(i)} = \prod_{k-1}^{n} f_{c,k}^{(i)},$$

where

$$\begin{aligned} f_{c,k}^{(i)} &= f_A(\Theta_k|\Theta_{1:k-1}) \quad \text{for } c = c_k^{n,i} = 1 \text{ (for changing parameter items), and} \\ f_{c,k}^{(i)} &= f_T(\Theta_k|\Theta_{1:k-1}) \quad \text{for } c = c_k^{n,i} = 0 \text{ (for constant parameter items).} \end{aligned}$$

Thus for example for four parameters and hypothesis H_5,for which it is $c^{4,5} = [0, 1, 0, 1]$, we have (see the Example)

$$f^{(5)} = f_T(\Theta_4|\Theta_{1:3})\, f_A(\Theta_3|\Theta_{1:2})\, f_T(\Theta_2|\Theta_1)\, f(\Theta_1).$$

Having 2^n hypotheses we can construct the same number of such possible parameter pdfs. Each of them corresponds to one hypothesis, i.e. to one combination of changing and not changing parameter items.

If the knowledge about the true hypothesis is missing, we have to consider the parameter pdf random with realizations $f^{(i)}$ and their probabilities equal to those of the hypotheses $\lambda_i = \Pr(H_i) = \Pr\left(f^{(i)}\right)$. Let us denote this random pdf by $f_H(\Theta)$.

The resulting parameter pdf of parameters as a result of the prediction step of estimation we introduce as such pdf $f_R(\Theta)$ that minimizes expectation of Kullback-Liebler distance from the pdf $f_H(\Theta)$

$$\hat{f}_R(\Theta) = \arg\min_{f_H} E\left[f_R(\Theta) \ln \frac{f_R(\Theta)}{f_H(\Theta)}\right]$$

where the expectation is taken over the random f_H.

The result of the optimization is simple and instructive

$$\hat{f}_R(\Theta) = \prod_{i=1}^{n} \left(f^{(i)}(\Theta)\right)^{\lambda_i}. \tag{6}$$

Instead of taking one true parameter pdf we take all possible and weight them according to their probabilities.

Summary

Let us summarize the results:

For

$f_T(\Theta)$	parameter pdf for constant parameters (e.g. from the last step of estimation),
$f_A(\Theta)$	parameter pdf for changing parameters (e.g. prior pdf),
H_i	hypothesis that a specific subset of parameters change while the rest stays constant,
λ_i	probability that H_i is true,
$f^{(i)}(\Theta)$	parameter pdf of parameters corresponding to H_i, i.e. composed of those conditional factors (see (5)) of $f_T(\Theta)$ that correspond to constant parameters and those factors of $f_A(\Theta)$ that correspond to changing parameters,

the resulting parameter pdf $f_R(\Theta)$, modeling constant parameters according to $f_T(\Theta)$ and variable ones according to $f_A(\Theta)$ with given probabilities whether specific parameters vary or not is given by the relation (6).

4 Conclusions

An extension of of the famous exponential forgetting used for identification of model parameters has been presented. The extension consists in a possibility to respect the fact, that some of the estimated parameters slowly change, and thus they need some forgetting coefficient smaller than one, and the rest of the coefficients do not change, and should have the forgetting coefficient equal to one. At the same time, it is shown, that the forgetting problem can be formulated as a problem of testing hypotheses.

References

[1] R. Kulhavý and M. B. Zarrop. On a general concept of forgetting. *International Journal of Control*, 58(4):905–924, 1993.

[2] V. Peterka. Bayesian approach to system identification. In P. Eykhoff, editor, *Trends and Progress in System Identification*, pages 239–304. Pergamon Press, Oxford, 1981.

[3] Greg Welch and Gary Bishop. An Introduction to the Kalman Filter. Technical Report 95-041, UNC-CH Computer Science, 1995.