Automatica, Vol. 23, No. 5, pp. 589–600, 1987
Printed in Great Britain.

0005–1098/87 $3.00 + 0.00
Pergamon Journals Ltd.

Restricted Exponential Forgetting in Real-time Identification*

RUDOLF KULHAVÝ†

A careful Bayesian formulation of the problem of forgetting obsolete information can substantially improve numerical reliability and quality of parameter tracking.

Key Words—Identification; parameter estimation; sampled data systems; time-varying systems; Bayes methods; adaptive control.

Abstract—Identification of time-varying stochastic systems is examined from the Bayesian viewpoint. The questions of which piece of posterior information should be forgotten and of how to forget it are discussed thoroughly, taking into account the main drawbacks of the commonly used technique of exponential forgetting. A novel procedure of suppressing obsolete information is suggested, endowed with two important features. Firstly, it modifies only that piece of so-far accumulated information which is being innovated by the currently measured data. Secondly, it makes it possible to respect partial prior knowledge about the location and/or evolution of parameters in a unified manner. The contributions of the procedure are illustrated in the special case of linear Gaussian regression-type models. A slight, but surprisingly effective, modification of the recursive least-squares method is derived which ensures a high numerical reliability of parameter tracking even for poorly excited systems. Its properties are demonstrated by a simulation example.

1. INTRODUCTION

IN MANY important practical problems (adaptive control, adaptive prediction etc.) the need for identification of time-variable systems arises although no explicit model of time variations of the system behaviour is available *a priori*. To ensure adaptivity of identification in such a case, a number of more or less *ad hoc* techniques have been designed. One of the most popular techniques, known as exponential forgetting (weighting, discounting, . . .), is based on a simple assumption that "information contained in a data item is the less reliable, the older the data item is".

However, this assumption is justified only if information carried by measured data is distributed over the whole parameter space and over the whole time horizon sufficiently "uniformly". If this condition is not fulfilled, i.e. the measured data do not carry sufficient information about all unknown parameters for a rather long time period, then the application of the exponential forgetting results in the loss of a piece of so-far accumulated information which is not compensated by the gain of new information. This case arises rather often in practice, e.g. due to linear feedback closed around the identified system, unsuitable parameterization including redundant and/or weakly identifiable parameters, rare changes of variables, input saturation etc.

The consequences are severe (Wittenmark and Åström, 1984). Any change of the character of system excitation may result (because of the previous loss of a piece of gathered information) in a sudden increase of the prediction error (the hitherto satisfactory model poorly describes the external system behaviour in new conditions). The adaptive controller (if used) may generate temporarily inappropriate control actions leading to a short-term instability of the closed loop ("bursting").

Different modifications of the standard exponential forgetting have been suggested in literature, e.g. a limitation of characteristic values of the parameter covariance matrix, an interruption of estimation whenever a sufficient quality of the output prediction is achieved, a data- and time-conditioned evaluation of the forgetting factor, often supplied with the possibility of detecting sudden changes of parameters etc. (several of these measures are discussed later in detail). Nevertheless, some shortcomings are still felt in practice in

* Received 10 June 1986; revised 6 January 1987; revised 26 March 1987. The original version of this paper was presented at the 7th IFAC/IFORS Symposium on Identification and System Parameter Estimation of York, U.K., during July 1985. The Published Proceedings of this IFAC Meeting may be ordered from Pergamon Books Ltd, Headington Hill Hall, Oxford OX3 0BW, U.K. This paper was recommended for publication in revised form by Associate Editor G. Goodwin under the direction of Editor P. C. Parks.

† Institute of Information Theory and Automation, Czechoslovak Academy of Sciences, 182 08 Prague, Czechoslovakia.

connection with parameter tracking techniques, especially insufficient robustness with respect to noninformative data, the great care that has to be devoted to the adjustment of "tuning knobs", and a difficult generalization to other than linear Gaussian regression-type models.

This paper attempts to give a rather general (Bayesian) insight into the problem of parameter tracking under lack of a model of parameter variations. Instead of purely heuristic constructions of suppressing obsolete information, an explicit model of updating the probability density of unknown parameters in time is built which fulfils intuitive demands on a rational forgetting. The resulting procedure has some novel features which overcome the main drawbacks of the standard exponential forgetting as well as the limitations of its known modifications.

2. PROBLEM STATEMENT IN BAYESIAN FRAMEWORK

The problem to be solved will be analysed in detail now using as a basis the Bayesian estimation scheme (Peterka, 1981).

Let us consider a stochastic system on which a finite data set is measured at discrete time instants $t = 1, 2, \ldots$. As usual, we denote the directly manipulated (possibly multivariate) input by $u(t)$ and the rest of the data forming the (m_y-dimensional) output by $y(t)$. Let the dependence of the system output on previous data be specified by the family of suitably parametrized probability densities

$$p(y(t) \mid t-1; u(t), k(t)), \quad t = 1, 2, \ldots \quad (1)$$

conditional on data measured up to and including the time $t-1$, the latest input $u(t)$, and the parameters $k(t)$.

In the following, the unknown parameters $k(t) \in K$ are taken as a multivariate random variable. Their prior uncertainty is described by the conditional (on all relevant information) probability density $p(k(1) \mid 0)$ defined with respect to a suitable measure κ.

Note that technical assumptions needed for a rigorous treatment of probability densities (in the following, simply called densities) are omitted below for the sake of simplicity (see e.g. Loève, 1960). For the same reason, no difference is made between the notation of a density as a whole and its value at a generic point.

Provided the input generator employs no other information about $k(t)$ than measured data (see the natural condition of control in Peterka, 1981), the *measurement updating* of the parameter density $p(k(t) \mid t-1) \to p(k(t) \mid t)$ is given by the Bayes theorem

$$p(k(t) \mid t) \propto p(y(t) \mid t-1; u(t), k(t)) p(k(t) \mid t-1) \quad (2)$$

where $\propto$ stands for proportionality (equality up to a normalizing constant).

For the recursion to be completed, the *time updating*

$$p(k(t) \mid t) \to p(k(t+1) \mid t) \quad (3)$$

has to be defined. This can be performed consistently within the margins of the Bayesian inference if complete prior knowledge of parameter variations is available. For instance, when describing the parameter evolution by a Markovian model $p(k(t+1) \mid t; k(t))$, the time updating is given by the integration

$$p(k(t+1) \mid t) = \int p(k(t+1) \mid t; k(t)) p(k(t) \mid t) \, \mathrm{d}\kappa(k(t)). \quad (4)$$

If such a piece of information is missing, the gap must be filled in another way, e.g. by using the technique of exponential forgetting.

Unfortunately, the standard exponential forgetting has proved to be too simple a model of discarding obsolete information. It suppresses all accumulated information regardless of the character of measured data and it does not admit the incorporation of available information about parameter variations other than through the choice of the forgetting factor. Some of the problems by which we pay for this simplicity have been mentioned in the Introduction. Apparently, it is worth re-considering the design of a rationally-based time updating (3) assuming at most partial (incomplete) information about parameter variations. This is the aim of this paper.

3. PHILOSOPHY OF DISCARDING OBSOLETE INFORMATION

The initial situation can be summarized as follows. Owing to the lack of a probabilistic description of parameter variations, the Bayesian inference, more specifically, the time updating (3) of the posterior parameter density, cannot be directly completed. Instead, the density $p(k(t+1) \mid t)$ has to be chosen in each step of recursive estimation so as to best reflect the actual evolution of estimated parameters. To formulate the decision problem arisen, a set of alternatives (alternative parameter densities) must be known beforehand.

We shall assume information about possible changes of parameters at each time instant t in the form of a set of alternative parameter

densities

$$\{p_i(k(t+1)\mid t),\quad i=0,1,\ldots,N\}. \qquad (5)$$

The "zero" alternative density will correspond to the case of no parameter changes [the Dirac function $\delta(\cdot)$ model in (3)], i.e. coincide with the posterior density

$$p_0(k(t+1)\mid t)$$

$$=\int \delta(k(t+1)-k(t))p(k(t)\mid t)\,\mathrm{d}\kappa(k(t)); \qquad (6)$$

other alternative densities will correspond (in a way discussed more thoroughly in Section 5) to the worst supposed cases of parameter changes. While the posterior alternative $i=0$ describes all the so-far accumulated information about unknown parameters, the reference alternatives $i=1,\ldots,N$ are included to compensate ignorance of a complete model of parameter variations.

Intuitively, a reasonable model of the time updating (3) should fulfil the following requirements.

($\mathcal{R}1$) Only the evidentially obsolete piece of posterior knowledge is to be updated.

($\mathcal{R}2$) Posterior and reference knowledge are to be composed in a rationally based, but conceptually feasible way.

To make the model maximally simple, but at the same time safe from extremely unfavourable cases of system excitation, we take as obsolete just the piece of posterior information innovated by the current data.

4. STATISTICAL PRELIMINARIES

Before formalizing the above intuitive requirements, we shall recall two classical notions of mathematical statistics.

Statistical decision problem

Let us suppose that we are to choose a single parameter density $p^*(k)$ on the basis of a set of alternative parameter densities $\{p_i(k), i=0,1,\ldots,N\}$. This task can be treated as a statistical decision problem [see DeGroot (1970) for a general definition] fully described by the quadruple (H, D, L, α) where

H denotes the (super)parameter space composed of all alternative densities $p_i(k)$, $i=0,1,\ldots,N$,

D denotes the decision space including all probability densities $p(k)$ (with respect to a given measure κ),

L denotes the loss functional defined on $H\times D$ which is to measure the "unlikeness" or "discrepancy" of parameter densities from H and D,

α denotes the vector $(\alpha_0, \alpha_1, \ldots, \alpha_N)$ of the probabilities of the particular elements of the space H.

To find a simply realizable solution for a rather rich family of probability densities, we shall introduce the loss functional by means of the I-divergence (Csiszár, 1967), known also as Kullback–Leibler information or generalized entropy

$$L(p_i(k), p(k)) = \int p(k)\ln\frac{p(k)}{p_i(k)}\,\mathrm{d}\kappa(k). \qquad (7)$$

From different viewpoints it is reasonable to take as the optimal solution of the decision problem the Bayesian solution [cf. the min–max approach in Perez (1984)] minimizing the expected value of the loss functional (7)

$$p^*(k) = \arg\min_{p(k)\in D}\sum_{i=0}^{N}\alpha_i L(p_i(k), p(k)). \qquad (8)$$

Lemma 1. If the product of the densities $p_j(k)\in H$ for all j such that $\alpha_j>0$ is not zero a.e., then the optimization task (8) has a unique solution (up to the equivalency a.e.), namely

$$p(k)\propto\prod_{i=0}^{N}[p_i(k)]^{\alpha_i}. \qquad (9)$$

Proof. The proposition can be proved directly by substituting the density (9) into the minimized functional and using the basic property of the I-divergence functional (7) that $L(p(k), \bar{p}(k)) = 0$ iff $p(k)=\bar{p}(k)$ a.e. □

Note that if the assumption of Lemma 1 is not fulfilled, the task does not have a unique solution and the decision must be specified on the basis of additional requirements.

Sufficient mapping of parameters

Within the Bayesian framework the full description of available information about unknown parameters k is given by the appropriate parameter density $p(k)$. The extraction of some piece of initial information can be achieved using a (measurable) mapping $T: K\to K'$ from the original into a new parameter space. We shall need to find such a parameter mapping which makes it possible, loosely speaking, to distinguish two given parameter densities as exactly as the original parameters do. There is a mathematical formalization of the mapping with this property (Kullback and Leibler, 1951).

Definition. We say that the mapping T is *sufficient* with respect to the pair of parameter densities $\{p(k), \bar{p}(k)\}$ if nonnegative measur-

able functions g, $\tilde{g}$, h exist such that the equalities

$$p(k) = g(T(k))h(k)$$
$$\tilde{p}(k) = \tilde{g}(T(k))h(k) \tag{10}$$

hold a.e. We call such a mapping *minimal sufficient* if it is a function of any other sufficient mapping.

In the following, we shall be interested in a special type of this mapping, namely the mapping sufficient with respect to the prior and posterior densities $\{p(k(t) \mid t-1),\ p(k(t) \mid t)\}$.

Lemma 2. The minimal sufficient mapping of parameters with respect to $\{p(k(t) \mid t-1),\ p(k(t) \mid t)\}$ can be specified as follows

$$T_t(k(t)) = \begin{cases} p(y(t) \mid t-1; u(t), k(t)) \\ \quad \text{if } p(k(t) \mid t-1) > 0 \\ -1 \quad \text{if } p(k(t) \mid t-1) = 0. \end{cases} \tag{11}$$

Proof. The proposition follows from Lemma 2.3 in Csiszár (1967) and the relation between the posterior and prior densities given by the Bayes rule (2). □

Notice that the minimal sufficient mapping is given directly by the density (1) taken as a function of unknown parameters.

5. NEW METHOD OF FORGETTING

Now we are ready to give a mathematical formulation of a time-update model fulfilling our requirements.

Theorem 1. Let us assume that

(A1) a parameter mapping $T_t: K \to K'$ sufficient with respect to the prior and posterior densities $\{p(k(t) \mid t-1),\ p(k(t) \mid t)\}$ is specified, and

(A2) a set of alternative densities of unknown parameters $\{p_i(k(t+1) \mid t),\ i = 0, 1, \ldots, N\}$ is selected with the density $p_0(k(t+1) \mid t)$ given by (6).

Let us define the optimal-time updated density of unknown parameters $p(k(t+1) \mid t)$ by the following requirements:

(R1) the mapping T_t is sufficient with respect to the posterior and time-updated densities $\{p_0(k(t+1) \mid t),\ p(k(t+1) \mid t)\}$;

(R2) the marginal density $p(k'(t+1) \mid t)$ of $k'(t+1) = T_t(k(t+1))$ is given by solving the optimization task

$$\min_{p(k'(t+1))} \sum_{i=0}^{N} \alpha_i(t+1 \mid t) L(p_i(k'(t+1) \mid t), p(k'(t+1))) \tag{12}$$

where

(A3) the vector $\alpha(t+1 \mid t) = (\alpha_0(t+1 \mid t), \alpha_1(t+1 \mid t), \ldots, \alpha_N(t+1 \mid t))$ of the probabilities of the marginal alternative densities $p_i(k'(t+1) \mid t)$ is supposed to be known.

If

(C1) the product of the marginal alternative densities $p_j(k'(t+1) \mid t)$ for all j such that $\alpha_j(t+1 \mid t) > 0$ is not zero a.e., and

(C2) $p_0(k'(t+1) \mid t) = 0$ implies $p(k'(t+1) \mid t) = 0$ a.e.,

then there exists a unique parameter density (up to the equivalency a.e.) fulfilling the requirements (R1) and (R2), namely

$$p(k(t+1) \mid t) \propto \begin{cases} \prod_{i=0}^{N} [p_i(T_t(k(t+1)) \mid t)]^{\alpha_i(t+1 \mid t)} \\ \qquad \times \dfrac{p_0(k(t+1) \mid t)}{p_0(T_t(k(t+1)) \mid t)} \\ \quad \text{if } p_0(T_t(k(t+1)) \mid t) > 0 \\ 0 \text{ if } p_0(T_t(k(t+1)) \mid t) = 0 \end{cases} \tag{13}$$

where we have denoted

$$p_i(T_t(k(t+1)) \mid t) = p_i(k'(t+1) \mid t) \quad \text{for} \quad k'(t+1) = T_t(k(t+1)).$$

Proof. The proposition follows from Lemma 1 given above and Lemma 2.2 in Csiszár (1967). □

The requirements (R1) and (R2) formalize the intuitive requirements ($\mathcal{R}1$) and ($\mathcal{R}2$) in terms of the sufficient mapping and statistical decision problem discussed in Section 4. Notice the multiplicative mixture way of combining the marginal alternative densities in (13) which is a result of the loss functional definition (7). This makes the application of Theorem 1 feasible for often met exponential-type parameter densities.

Assuming $T_t(k(t+1)) = k(t+1)$, $p_1(k(t+1) \mid t) \propto 1$, $N = 1$, we get a probabilistic definition of the standard exponential forgetting heuristically designed by Peterka (1981)

$$p(k(t+1) \mid t) \propto [p_0(k(t+1) \mid t)]^{\alpha_0(t+1 \mid t)}. \tag{14}$$

This relationship together with the fact that the application of any nontrivial (different from the identity) sufficient mapping of unknown parameters results in a restriction of forgetting account for the name "restricted exponential forgetting", used for referring to the new method below.

Discussion of assumptions

Before applying Theorem 1 to a concrete estimation problem, we must fulfil its basic assumptions, i.e.

(A1) choose a sufficient parameter mapping,
(A2) construct a set of alternative densities,
(A3) estimate the probabilities of marginal alternative densities.

Let us return to particular assumptions in detail.

(A1) The choice of a concrete sufficient mapping determines the level of restriction of forgetting. Thus, the use of the identity mapping implies no restriction, and the use of the minimal sufficient mapping implies maximal restriction. Obviously, the more closely we approach the maximal restriction, the more robust the parameter tracking is with respect to noninformative data. However, the same conclusion does not hold generally for the quality of output prediction. This can be improved in cases of good system excitation by utilizing a less strict restriction which ensures a faster reaction of estimator to future parameter changes. To sum up, the parameter mapping should be chosen in accordance with the expected innovating effect of future data (the level of system excitation) and with respect to the fact that a loss of robustness may be much more dangerous than a partial deterioration of the prediction quality.

(A2) The selection of suitable reference densities $p_i(k(t+1)\,|\,t)$, $i=1, 2, \ldots, N$ enables the user to specify at particular time instants areas of probable appearance of unknown parameter values. Utilizing all available information about parameter variations, these areas can often be made relatively small. In such a way the parameter tracking may be considerably improved; of course, at the cost of some additional computations associated with the evaluation of reference densities and/or a more complicated time-updating operation (13). There are no conceptual constraints on specification of reference densities—possibly except the conditions (C1) and (C2) which ensure the uniqueness of the resulting solution. They can be constructed as fixed, recursively updated (in accordance with a prespecified model of parameter variations), derived from the posterior density etc. If little prior information is available, or extreme computational demands must be satisfied, a noninformative reference density of unknown parameters (e.g. corresponding to a uniform distribution of probability) can be used. Specification of reference densities is dealt in detail in Kulhavý (1986).

(A3) The probabilities $\alpha_i(t+1\,|\,t)$ should be stated *a priori* according to the (subjective) belief in the marginal alternatives $p_i(k'(t+1)\,|\,t)$. Unfortunately, these probabilities generically depend (through the performed restriction) on measured data. The use of "average" probabilities may often be unsuitable, especially if the reference densities are extremely simple (the situation can be naturally solved by a careful specification of reference densities—cf. Kulhavý, 1986). The subjective determination of the optimal probabilities may be further complicated by their time variations. Hence, the need for a more sophisticated determination of the probabilities $\alpha_i(t+1\,|\,t)$ is quite appealing.

Theorem 2. Let us suppose that assumptions (A1) and (A2) from Theorem 1 are fulfilled.

Let us define the optimal probability vector $\alpha(t+1\,|\,t)$ so that

(R3) $\alpha(t+1\,|\,t)$ is a solution of the optimization task

$$\max_{\alpha(t+1)} \min_{p(k'(t+1))} M(\alpha(t+1), p(k'(t+1))) \quad (15)$$

with M defined by

$$M(\alpha(t+1), p(k'(t+1))) = \sum_{i=0}^{N} \alpha_i(t+1)[L(p_i(k'(t+1)\,|\,t), p(k'(t+1))) - \hat{L}_i(t+1\,|\,t)] \quad (16)$$

and

$$\hat{L}_0(t+1\,|\,t) = (1+\rho)L(p(k'(t)\,|\,t-1), p(k'(t)\,|\,t)) \quad (17)$$

$$\hat{L}_i(t+1\,|\,t) = L(p_i(k'(t)\,|\,t-1), p(k'(t)\,|\,t)), \; i=1, \ldots, N \quad (18)$$

where

(A3′) a scalar $\rho \geqslant 0$ is assumed prespecified.

If

(C1′) the product of the marginal alternative densities $p_i(k'(t+1)\,|\,t)$ for $i=0, 1, \ldots, N$ is not zero a.e., and

(C3) the values $\hat{L}_i(t+1\,|\,t)$, $i=0, 1, \ldots, N$ are finite,

then there exist at least one vector $\alpha(t+1\,|\,t)$ fulfilling the requirement (R3) and just one density (up to the equivalency a.e.)

$$p(k'(t+1)\,|\,t) \propto \prod_{i=0}^{N} [p_i(k'(t+1)\,|\,t)]^{\alpha_i(t+1|t)}. \quad (19)$$

The necessary and sufficient conditions for the optimal value $\alpha(t+1\,|\,t)$ are formed by the relations

$$L(p_i(k'(t+1)\,|\,t), p(k'(t+1)\,|\,t)) - \hat{L}_i(t+1\,|\,t) = \mu \quad (20a)$$

for all j such that $\alpha_j(t+1 \mid t) > 0$, and

$$L(p_i(k'(t+1) \mid t), p(k'(t+1) \mid t)) - \hat{L}_i(t+1 \mid t) \leqslant \mu \quad (20b)$$

for all j such that $\alpha_j(t+1 \mid t) = 0$, where μ is a real constant.

Proof. The existence result can be proved using Proposition 1.2 from Chapter II in Ekeland and Temam (1976) taking into account the properties of the optimized functional. The necessary and sufficient conditions for $\alpha(t+1 \mid t)$ follow from Theorem 4.4.1 in Gallager (1968). □

The requirement (R3) formalizes the fact that we seek the probability vector $\alpha(t+1 \mid t)$ as the least favourable between all probability vectors $\alpha(t+1)$. To compensate the pessimistic character of max–min decision-making, we use the modification (17)–(18) for incorporating partial information about the "success" of particular alternatives in describing the parameter evolution (cf. Savage, 1954). Notice that the loss estimates are given by the discrepancies of the posterior density with respect to the prior density and the reference densities at the preceding time instant. Due to parameter variations at the time $t+1$ the loss for the posterior alternative is expected to be higher than the evaluated discrepancy. The expected increase is incorporated into (17) as simply as possible—through a single, prespecified factor $\rho \geqslant 0$.

Note that to be consistent we should take the probabilities $\alpha_i(t+1 \mid t)$ as unknown parameters of the time-update model and recursively update them within the Bayesian framework. Unfortunately, the conceptual troubles caused by the restriction and computational complexity make this procedure so far unfeasible.

Discussion of assumptions

Theorem 2 is supposed to complement Theorem 1. Notice that in this connection assumptions (A1) and (A2) remain valid, but assumption (A3) is replaced by a weaker assumption (A3′). Hence, instead of the vector of the probabilities $\alpha_i(t+1 \mid t)$ now only the scalar ρ is supposed prespecified. It can be shown [by applying the expectation $E(\cdot \mid t-1; u(t))$ to the conditions (20)] that this factor represents the steady-state ratio [or its upper bound if $\alpha_0(t+1 \mid t) = 0$] of the expected amount of "forgotten" information $L(p_0(k'(t+1) \mid t), p(k'(t+1) \mid t))$ to the expected amount of information contained in the latest data $L(p(k'(t) \mid t-1), p(k'(t) \mid t)) = L(p(k(t) \mid t-1), p(k(t) \mid t))$. Thus, the needed probabilities are evaluated automatically depending on information in measured data and the user is to adjust the gain of this information "feedback" through ρ. Naturally, parameter tracking is less sensitive to the value of ρ than to the values of $\alpha_i(t+1 \mid t)$.

Notice that for the combination of Theorem 1 and 2 the uniqueness conditions (C1′), (C2), (C3) are stricter than the original (C1), (C2).

6. APPLICATION TO REGRESSION-TYPE MODELS

To illustrate the contributions of the restricted exponential forgetting, we shall describe two interesting practical applications of linear Gaussian regression-type models. More about the practical implications of the above theory can be found in Kulhavý and Kárný (1984), Kárný *et al.* (1985), Kulhavý (1986), and Kárný and Kulhavý (1987).

Single output model with known noise dispersion

Let the identified system be described by the univariate Gaussian regression-type model generating the density (1) in the form

$$p(y(t) \mid t-1; u(t), P(t)) = (2\pi)^{-1/2}\sigma^{-1} \exp\{-(y(t) - P^{\mathrm{T}}(t)z(t))^2/2\sigma^2\}. \quad (21)$$

The vector $z(t)$ is a known function of the latest input $u(t)$ and previous data measured up to the time $t-1$. Provided the dispersion σ^2 is known, just the vector $P(t)$ represents the unknown parameters $k(t)$.

Let us suppose that the prior density of $P(t)$ is Gaussian

$$p(P(t) \mid t-1) \propto \exp\{-(P(t) - \hat{P}(t \mid t-1))^{\mathrm{T}} \times C^{-1}(t \mid t-1)(P(t) - \hat{P}(t \mid t-1))/2\sigma^2\}. \quad (22)$$

Using the Bayes rule (2) we easily derive that the posterior density $p(P(t) \mid t)$ is of the same form, but with the statistics $\hat{P}(t \mid t)$, $C(t \mid t)$ given by the well-known recursive least-squares formulae

$$\hat{P}(t \mid t) = \hat{P}(t \mid t-1) + \frac{C(t \mid t-1)z(t)}{1 + \zeta(t \mid t-1)}\hat{\varepsilon}(t \mid t-1) \quad (23)$$

$$C^{-1}(t \mid t) = C^{-1}(t \mid t-1) + z(t)z^{\mathrm{T}}(t) \quad (24)$$

where

$$\hat{\varepsilon}(t \mid t-1) = y(t) - \hat{P}^{\mathrm{T}}(t \mid t-1)z(t) \quad (25)$$

$$\zeta(t \mid t-1) = z^{\mathrm{T}}(t)C(t \mid t-1)z(t). \quad (26)$$

Specialization of a forgetting scheme. To suppose as little as possible, we consider a single

reference density of unknown parameters $p_1(P(t+1)\,|\,t) \propto 1$ modelling the case of maximal parameter uncertainty (all parameter changes are admitted). Following the usual terminology the forgetting factor $\varphi(t+1\,|\,t)$ is introduced explicitly instead of the probability $\alpha_0(t+1\,|\,t)$. The forgetting factor is assumed known at first.

To illustrate the effect of a forgetting restriction, we compare three forms of the parameter mapping T_t:

Case 1:

$$T_t(P) = P \tag{27}$$

Case 2:

$$T_t(P) = y(t) - P^{\mathrm{T}} z(t) \tag{28}$$

Case 3:

$$T_t(P) = (y(t) - P^{\mathrm{T}} z(t))^2. \tag{29}$$

According to Lemma 2 the mapping (29) is minimal sufficient with respect to the pair of prior and posterior densities [it is a one-to-one transformation of the mapping (11)]. Clearly, the mappings (27) and (28) are then sufficient too, but generically not minimal sufficient.

Case 1—*no restriction.* Using the data-update formulae (23)–(24) and Theorem 1 under the above assumptions with the mapping (27), it can be proved that the Gaussian form (22) of the prior parameter density reproduces throughout the whole estimation and the statistics $\hat{P}$, C evolve as follows:

$$\hat{P}(t+1\,|\,t) = \hat{P}(t\,|\,t-1) + \frac{C(t\,|\,t-1)z(t)}{1+\zeta(t\,|\,t-1)}\,\hat{\varepsilon}(t\,|\,t-1) \tag{30}$$

$$C^{-1}(t+1\,|\,t) = \varphi(t+1\,|\,t)[C^{-1}(t\,|\,t-1) + z(t)z^{\mathrm{T}}(t)]. \tag{31}$$

Case 2—*restriction to a linear projection of parameters.* Using the same procedure as in Case 1, only with the mapping (28) instead of (27), it can be proved that the Gaussian form (22) of the prior parameter density reproduces again throughout the whole estimation, but the statistics $\hat{P}$, C evolve in a slightly different way.

For $\|z(t)\| > 0$:

$$\hat{P}(t+1\,|\,t) = \hat{P}(t\,|\,t-1) + \frac{C(t\,|\,t-1)z(t)}{1+\zeta(t\,|\,t-1)}\,\hat{\varepsilon}(t\,|\,t-1) \tag{32a}$$

$$C^{-1}(t+1\,|\,t) = C^{-1}(t\,|\,t-1) + \gamma(t)z(t)z^{\mathrm{T}}(t) \tag{33a}$$

where

$$\gamma(t) = \varphi(t+1\,|\,t) - (1-\varphi(t+1\,|\,t))\zeta^{-1}(t\,|\,t-1). \tag{34}$$

For $\|z(t)\| = 0$:

$$\hat{P}(t+1\,|\,t) = \hat{P}(t\,|\,t-1) \tag{32b}$$

$$C(t+1\,|\,t) = C(t\,|\,t-1). \tag{33b}$$

It should be emphasized that equation (33a) does not imply a simple data weighting because the "weight" $\gamma(t)$ can be negative too. An alternative form of (33a) gives a deeper insight:

$$C^{-1}(t+1\,|\,t) = [C^{-1}(t\,|\,t-1) - \zeta^{-1}(t\,|\,t-1)z(t)z^{\mathrm{T}}(t)] + \varphi(t+1\,|\,t)[(\zeta^{-1}(t\,|\,t-1)+1)z(t)z^{\mathrm{T}}(t)]. \tag{33a}$$

Here we have decomposed the regular matrix $C^{-1}(t\,|\,t)$ into two singular matrices in square brackets. Obviously, exponential forgetting is applied only to the second matrix of rank 1 which extracts from $C^{-1}(t\,|\,t)$ "all data dyads" of the form $z(t)z^{\mathrm{T}}(t)$.

Notice that the algorithm has two branches according to the Euclidean norm $\|z(t)\| = \sqrt{z^{\mathrm{T}}(t)z(t)}$ of the regressor.

Comparison of Cases 1 *and* 2. Let us describe the Gaussian parameter densities considered in the illustrative example by the corresponding ellipsoids of concentration

$$(P-\hat{P})^{\mathrm{T}} C^{-1} (P-\hat{P}) = \text{const.} \tag{35}$$

It is simple to verify by comparing the formulae (24), (31), and (33) that

- —the data updating decreases only the diameter of the ellipsoid of concentration in the direction of the current Kalman gain vector $C(t\,|\,t)z(t)$—while this diameter is multiplied by the factor $1/\sqrt{1+\zeta(t\,|\,t-1)}$, the geometrically conjugate diameters, in the directions orthogonal to the current regressor vector $z(t)$, remain unchanged;
- —the forgetting in Case 1 (the standard exponential forgetting) increases all conjugate diameters of the ellipsoid of concentration, in Case 2 only the diameter in the Kalman gain direction $C(t\,|\,t)z(t)$—the corresponding diameters are multiplied by the factor $1/\sqrt{\varphi(t+1\,|\,t)}$.

Different effects of forgetting in Cases 1 and 2 are illustrated for a two-dimensional vector of regression coefficients in Figs 1 and 2.

The contribution of the restriction in Case 2 becomes clear e.g. if the regressor vector is composed of the latest input and the last output $z^{\mathrm{T}}(t) = [u(t), y(t-1)]$ firmly related by the linear feedback $u(t) = -cy(t-1)$. Then the regressor vector $z^{\mathrm{T}}(t) = [-c, 1]y(t-1)$ does not

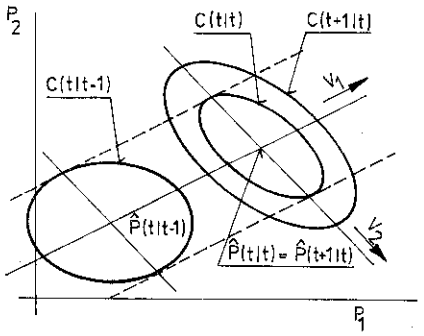

FIG. 1. Evolution of the ellipse of concentration in the case of the standard exponential forgetting [the direction v_1 is parallel with $C(t \mid t)z(t)$, v_2 is orthogonal to $z(t)$].

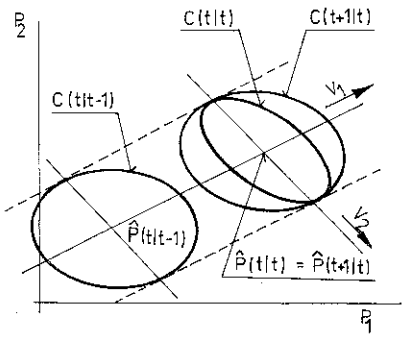

FIG. 2. Evolution of the ellipse of concentration in the case of the restricted exponential forgetting (the same notation as in Fig. 1 is used).

change its direction and data may carry new information at most about the linear combination $-cP_1(t) + P_2(t)$. In Case 1, the diameter of the ellipse of concentration orthogonal to the regressor direction permanently increases—possibly up to a numerical breakdown of identification. In Case 2, the possibility of the estimator wind-up is conceptually prevented—identification becomes considerably robust with respect to linearly dependent regressors.

Case 3—*maximal restriction.* Using the Bayes rule (2) and Theorem 1 under the above assumptions with the mapping (29), it can be proved that the resulting time-updated density of unknown parameters is not purely Gaussian

$$p(P(t+1) \mid t) \propto \frac{\exp\{-(P(t+1) - \hat{P}(t+1 \mid t))^{\mathrm{T}} C^{-1}(t+1 \mid t) \times (P(t+1) - \hat{P}(t+1 \mid t))/2\sigma^2\}}{[1 + \exp\{-2\hat{\varepsilon}(t \mid t-1) \times (y(t) - P^{\mathrm{T}}(t+1)z(t))/\zeta(t \mid t-1)\}]^{1-\varphi(t+1|t)}}. \quad (36)$$

Here the statistics $\hat{P}(t+1 \mid t)$, $C(t+1 \mid t)$ coincide with those derived in Case 2. This means that the probability originally normally distributed (as in Case 2) is re-distributed in favour of the parameter values $P(t+1)$ giving the same sign of the output error $y(t) - P^{\mathrm{T}}(t+1)z(t)$ as the prediction error $\hat{\varepsilon}(t \mid t-1)$ has. After applying Theorem 1 in the next steps the result is much more complicated, but the essential effect remains unchanged: parameter estimation becomes less sensitive with respect to stochastic influences (due to the asymmetry of parameter densities) so that a smaller value of the forgetting factor can be used.

To sum up, Case 2 seems to be a reasonable compromise between robustness and feasibility of parameter estimation. This is why just its adaptive version is discussed below.

Computation of the forgetting factor in Case 2. Evaluating the value of the forgetting factor $\varphi(t+1 \mid t)$ according to Theorem 2 we get

$$\varphi^{-1}(t+1 \mid t) = 1 + (1+\rho)\left[\ln(1 + \zeta(t \mid t-1)) + \frac{\zeta(t \mid t-1)}{1 + \zeta(t \mid t-1)}(\hat{\varepsilon}_N^2(t \mid t-1) - 1)\right] \quad (37)$$

where the normalized square of the prediction error is

$$\hat{\varepsilon}_N^2(t \mid t-1) = \frac{\hat{\varepsilon}^2(t \mid t-1)}{\sigma^2(1 + \zeta(t \mid t-1))}. \quad (38)$$

For rather small values of $\zeta(t \mid t-1)$ the natural logarithm in (37) may be approximated by

$$\ln(1 + \zeta(t \mid t-1)) \approx \frac{\zeta(t \mid t-1)}{1 + (\zeta(t \mid t-1)/2)}. \quad (39)$$

The expected value of the forgetting factor for $\hat{\varepsilon}_N^2(t \mid t-1) = 1$ is

$$\hat{\varphi}^{-1}(t+1 \mid t) = 1 + (1+\rho)\ln(1 + \zeta(t \mid t-1)). \quad (40)$$

In the steady state of parameter estimation when

$$z^{\mathrm{T}}(t)C(t+1 \mid t)z(t) \approx z^{\mathrm{T}}(t)C(t \mid t-1)z(t) \quad (41)$$

the expected value of the forgetting factor is roughly

$$\hat{\varphi}^{-1} \approx 1 + 2\rho \quad (42)$$

as we easily derive after substituting (33), (40) and (39) into (41). The last relation makes it possible to parameterize the expression (37) directly by means of the value $\hat{\varphi}$ which may be easier to specify.

Notice that the gain of the normalized prediction error square in (37) depends on the uncertainty of the regression coefficients $P(t)$ [through $\zeta(t \mid t-1)$]. This implies a high degree of adaptivity of the algorithm, e.g. a roughly exponential increase of $\varphi(t+1 \mid t)$ in the initial phase of estimation as well as in re-tuning after important parameter changes.

Multiple output model with unknown noise covariance matrix

Let us consider the linear Gaussian regression-type model in the case where the system output

may be multivariate ($m_y \geq 1$), and the covariance matrix $R(t)$ of the unpredictable component of the system output is unknown. Since for technical reasons it is somewhat more advantageous to work with the inversion $W(t) = R^{-1}(t)$, called the precision matrix, we shall assume that the model results in the density

$$p(y(t) \mid t-1; u(t), W(t), P(t)) = (2\pi)^{-m_y/2} |W(t)|^{1/2} \exp\{-(y(t) - P^{\mathrm{T}}(t)z(t))^{\mathrm{T}} W^{-1}(t)(y(t) - P^{\mathrm{T}}(t)z(t))/2\} \quad (43)$$

where now the unknown parameters are formed by the (m_z, m_y)-matrix of regression coefficients $P(t)$ and the (m_y, m_y)-matrix of precision $W(t)$. It is easy to verify that if the prior density of unknown parameters at time t is of the Gauss–Wishart form

$$p(W(t), P(t) \mid t-1) \propto |W(t)|^{(\nu(t|t-1)+m_z-m_y-1)/2} \times \exp\{-\mathrm{tr}[W(t)\Lambda(t \mid t-1)]/2\} \times \exp\{-\mathrm{tr}[W(t)(P(t) - \hat{P}(t \mid t-1))^{\mathrm{T}} C^{-1}(t \mid t-1) \times (P(t) - \hat{P}(t \mid t-1))]/2\}, \quad (44)$$

then the posterior density is of the same form too. The density is fully specified by a scalar $\nu(t \mid t-1) > m_y - 1$ and matrices $\hat{P}(t \mid t-1)$, $C(t \mid t-1) > 0$, $\Lambda(t \mid t-1) > 0$.

Specialization of a forgetting scheme. To suppose as little as possible, we consider again a single "noninformative" reference density of unknown parameters—now taken as a limit case of the parameter density in the form (44) for $C^{-1}(t+1 \mid t) \to 0$, $\Lambda(t+1 \mid t) \to 0$, $\nu(t+1 \mid t) \to 0$, i.e.

$$p_1(W(t+1), P(t+1) \mid t) \propto |W(t+1)|^{(m_z-m_y-1)/2}. \quad (45)$$

To keep feasibility of parameter estimation, we choose the linear parameter mapping

$$T_t(W, P) = (W, y(t) - P^{\mathrm{T}} z(t)) \quad (46)$$

which is probably the minimal sufficient mapping saving the Gauss–Wishart form of the parameter density.

Results. Using the Bayes rule (2) and Theorem 1 under the above assumptions, it can be proved that the Gauss–Wishart form (44) of the prior density of parameters reproduces throughout the whole estimation and the statistics $\hat{P}$, C, Λ, ν evolve according to the following formulae. Recall that the weight $\gamma(t)$ has been introduced above by (34).

For $\|z(t)\| > 0$:

$$\hat{P}(t+1 \mid t) = \hat{P}(t \mid t-1) + \frac{C(t \mid t-1)z(t)}{1 + \zeta(t \mid t-1)} \hat{\varepsilon}^{\mathrm{T}}(t \mid t-1) \quad (47a)$$

$$C^{-1}(t+1 \mid t) = C^{-1}(t \mid t-1) + \gamma(t)z(t)z^{\mathrm{T}}(t). \quad (48a)$$

For $\|z(t)\| = 0$:

$$\hat{P}(t+1 \mid t) = \hat{P}(t \mid t-1) \quad (47b)$$

$$C(t+1 \mid t) = C(t \mid t-1). \quad (48b)$$

Independently of $\|z(t)\|$:

$$\Lambda(t+1 \mid t) = \varphi(t+1 \mid t)\left[\Lambda(t \mid t-1) + \frac{\hat{\varepsilon}(t \mid t-1)\hat{\varepsilon}^{\mathrm{T}}(t \mid t-1)}{1 + \zeta(t \mid t-1)}\right] \quad (49)$$

$$\nu(t+1 \mid t) = \varphi(t+1 \mid t)[\nu(t \mid t-1) + 1]. \quad (50)$$

Evaluating the optimal value of the forgetting factor $\varphi(t+1 \mid t)$ according to Theorem 2, we get a pair of transcedent equations containing rational, logarithmic, gamma and psi functions of the statistics $\nu(t \mid t-1)$, $\zeta(t \mid t-1)$ [defined by (26)], and $\eta(t \mid t-1)$ introduced as follows

$$\eta(t \mid t-1) = \hat{\varepsilon}^{\mathrm{T}}(t \mid t-1)\Lambda^{-1}(t \mid t-1)\hat{\varepsilon}(t \mid t-1). \quad (51)$$

These equations are too complicated to be presented here, but they can be well approximated by the explicit formula (37) with the normalized square of the prediction error

$$\hat{\varepsilon}_N^2(t \mid t-1) = \frac{1}{m_y} \frac{[\nu(t \mid t-1) + 1]\eta(t \mid t-1)}{1 + \zeta(t \mid t-1) + \eta(t \mid t-1)}. \quad (52)$$

Note that all comments made above concerning the formula (37) are then valid too.

The main contribution of the presented algorithm lies in the increased numerical reliability of identification. Owing to the restriction of forgetting through the linear mapping (46) only information about the precision matrix $W(t+1)$ and the current projection of regression coefficients $y(t) - P^{\mathrm{T}}(t+1)z(t)$ is suppressed, but it is suppressed uniformly regardless of possibly different rates of their evolution. The discrepancy between the uniform forgetting and different variations of parameters determined by the mapping (46) can be solved by respecting prior information about parameter variations through a nontrivial reference density (Kulhavý, 1986).

Relationship to other algorithms

The method of the restricted exponential forgetting formulated in Theorem 1 is a direct continuation of the idea of the directional forgetting proposed in Kulhavý and Kárný

(1984). In the cited paper, the parameter mapping was required to be independent of the "rest" of the unknown parameters and the exponential forgetting was introduced in a purely heuristic way. In this respect, the present paper offers a precise formulation and further extension of the original idea. The resulting algorithms for the linear Gaussian regression-type models are very close—a slight modification is only in formula (50) which is now unified for both algorithm branches.

A similar algorithm, as concerns the structure of formula (48), was suggested as a direct modification of the recursive least-squares by Hägglund (1983, 1985). Instead of using an explicit forgetting factor, he required the covariance matrix of the regression coefficients $P(t+1)$ to converge to a prespecified diagonal matrix.

Naturally, there are other possibilities of eliminating the estimator wind-up, but these attain robustness of parameter estimation at the cost of additional, in its essence artificial, constraints, e.g. by keeping the trace $\mathrm{tr}\, C(t+1 \mid t)$ (Landau and Lozano, 1979) or the diagonal entries $C_{ii}(t+1 \mid t)$ (Saelid *et al.*, 1985) constant. Note that these constraints are closely tied up with the recursive least-squares scheme.

Although the method of determining the forgetting factor proposed in Theorem 2 cannot be considered as foolproof, formula (37) surprisingly covers some previous results. The idea of Fortescue *et al.* (1981) to keep a constant desired amount of accumulated information results (applied in accordance with our approach, i.e. with the forgetting applied after the data updating) in the formula

$$\varphi^{-1}(t+1 \mid t) = 1 + \frac{1}{N_0}\,\hat{\varepsilon}^2_N(t \mid t-1) \qquad (53)$$

where the factor N_0 ("asymptotic memory length") is supposed to be *a priori* chosen. Wellstead and Sanoff (1981) proposed to improve the initial tuning by combining the forgetting factor (53) with an exponential growing to unity. Bertin *et al.* (1986) recommended that the influence of large prediction errors in (53) is limited by prefiltering $\hat{\varepsilon}^2(t \mid t-1)$. Such features can be attained to a certain degree automatically if the "memory length" is adjusted dynamically as $N_0 = (1+\zeta^{-1}(t \mid t-1))/(1-\rho)$. However, the resulting formula is then nothing but a rough approximation of the relation (37).

Note to implementation

A numerically reliable implementation of the above algorithms on a computer requires the positive (semi)definiteness of the matrices $C(t+1 \mid t)$, $\Lambda(t+1 \mid t)$ to be saved throughout the computation. This can be achieved by evolving the factors of the Choleski square-root or $L-D$ factorization of the matrices (Kulhavý and Kárný, 1984; Kárný *et al.* 1985) instead of the matrices themselves. By a suitable organization of the computation all statistics needed for evaluating the forgetting factor can be determined at the same time.The algorithms can be used even in the case of a singular matrix $C(t \mid t-1)$ if the singular branch of the computation defined originally by $\|z(t)\| = 0$ is entered whenever $\zeta(t \mid t-1) = 0$ [or $\zeta(t \mid t-1) <$ "machine zero"]. The computational demands are comparable with the recursive least-squares algorithm using the standard exponential forgetting.

7. SIMULATION EXAMPLE

A single input single output system was simulated by the linear Gaussian regression-type model generating the density (43) with

$$z^{\mathrm{T}}(t) = [u(t), y(t-1), u(t-1), \ldots, y(t-3), u(t-3), v(t)]$$
$$P^{\mathrm{T}} = [b_0, a_1, b_1, \ldots, a_3, b_3, 1],\ W = 1.$$

The parameters a_i, b_i corresponded to the discretized transfer function $1/(1+Ts)^3$ sampled with the period $1.25T$. The disturbance $v(t)$ giving a variable output level was modelled as the random walk with $N(0, 0.0025)$-distributed increments. The system was controlled to the zero setpoint by the self-tuning controller supposing the reduced structure of the system model (Böhm *et al.*, 1984)

$$z^{\mathrm{T}}(t) = [u(t), y(t-1), u(t-1), 1]$$
$$P^{\mathrm{T}} = [b_0, a_1, b_1, c].$$

The penalty $\omega = 0.01$ on the input increments was chosen. It should be emphasized that the disturbance $v(t)$ was not supposed measurable.

Simulation runs were performed with the estimation algorithm (47)–(50) using the forgetting factor (37) and (52) as well as with the standard exponentially forgotten least squares using the forgetting factor (53) and (52) [the procedure of Fortescue *et al.* (1981) with a recursively estimated dispersion σ^2]. Estimation started from $\hat{P}(1 \mid 0) = 0$, $C(1 \mid 0) = I$, $\Lambda(1 \mid 0) = 1$, $\nu(1 \mid 0) = 1$.

The novel algorithm improved the results obtained with the standard exponential forgetting in several respects (some of the following observations are demonstrated in Figs 3–5):

—the average value of the forgetting factor and, consequently, the quality of control were in a

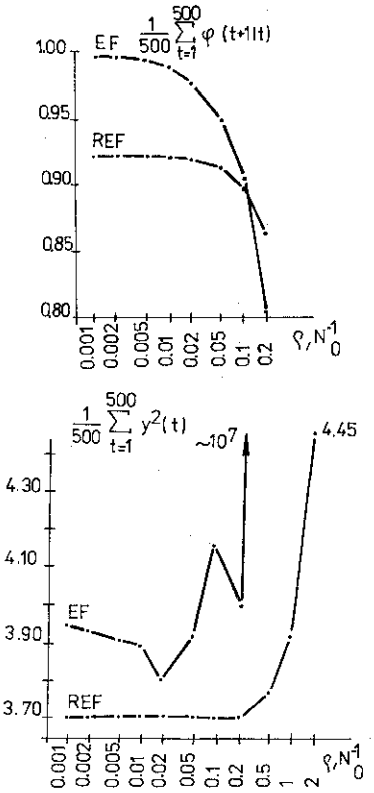

FIG. 3. Sensitivity of the average forgetting factor and the quality of control to the value of the heuristic factor ρ or $1/N_0$. The standard exponential forgetting (EF) using the factor (53) (Fortescue *et al.*, 1981) and the restricted exponential forgetting (REF) using the factor (37) are compared.

wide range ($\rho > 0.2$, $N_0 > 5$) much less sensitive to the value of the appropriate heuristic factor;

—the quality of control achievable by a suitable choice of the heuristic factor (ρ or N_0) was slightly better;

—numerically reliable estimation and adaptive control were ensured even for rather small average values of the forgetting factor;

—the small values of the forgetting factor at the first steps of estimation accelerated the initial convergence of parameter estimates;

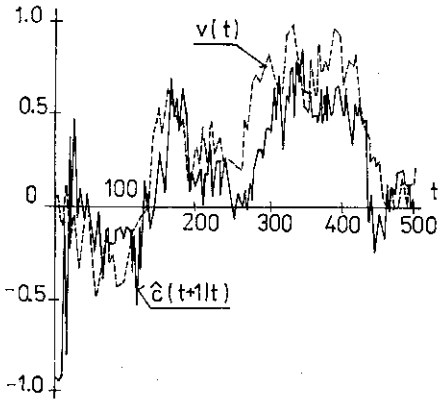

FIG. 4. Tracking of the unmeasurable external disturbance $v(t)$ by the absolute term estimate $\hat{c}(t+1 \mid t)$ using the restricted exponential forgetting with $\rho = 0.2$.

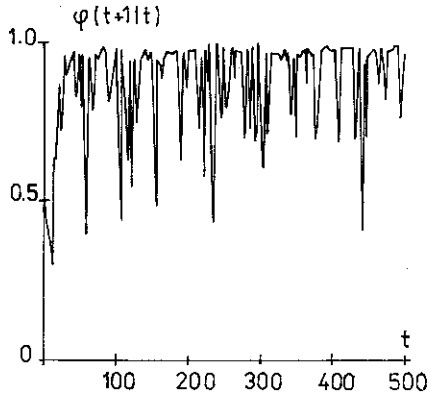

FIG. 5. Time course of the forgetting factor $\varphi(t+1 \mid t)$ in the case of the restricted exponential forgetting with $\rho = 0.2$.

—the variances of all the estimated regression coefficients with the exception of the absolute term c were significantly smaller.

8. CONCLUSIONS

A rationally based way of supressing obsolete information has been suggested in the paper with two goals pursued—on the one hand, to ensure a numerical reliability of identification even in the situations where little information is contained in measured data, on the other hand, to keep feasibility of the computation for a rather broad class of system models.

The core of the paper consists of general theoretical results. Theorem 1 presents in statistical terms a new formulation of a rational information forgetting. This generalizes the standard understanding of the exponential forgetting and enlightens the relationship between the pieces of information gathered from data and lost by a forgetting. Moreover, a practically important possibility of incorporating available (though incomplete) information about parameter variations follows from the general formulation. Theorem 2 offers a conceptually feasible way in which the forgetting may be made adaptive.

These results make it possible to design a "forgetting" procedure adequately to the chosen model of system as well as to available knowledge of its parameter variations. As an illustration, the restricted exponential forgetting is elaborated in the paper for linear Gaussian regression-type models. Simple modifications of the recursive least-squares algorithm are derived which (as extensive simulation experience as well as first practical experience indicate) overcome the essential drawbacks of the standard exponential forgetting in a qualitative way. A further improvement of parameter tracking can be achieved by an algorithm utilizing prior information about parameter variations (through

the reference density of a general form) which is derived in the self-contained paper by Kulhavý (1986). Applications for other types of system models, e.g. finite Markov chains, are elaborated.

The results achieved indicate that the Bayesian statistics are a suitable tool for solving complex engineering problems. Nevertheless, feasibility requirements have forced us to avoid e.g. the Bayesian evaluation of the forgetting factor or the use of the minimal sufficient parameter mapping. Progress in building a systematic approach to the approximation of the Bayesian inference could help to overcome these limitations.

REFERENCES

Bertin, D., S. Bittanti and P. Bolzern (1986). Tracking of nonstationary systems by means of different prediction-error directional forgetting techniques. *Preprints 2nd IFAC Workshop on Adaptive Systems in Control and Signal Processing,* Lund, pp. 91–96.

Böhm, J., A. Halousková, M. Kárný and V. Peterka (1984). Simple LQ self-tuning controllers. *Preprints 9th IFAC Congress,* Vol. XII, Budapest, pp. 171–176.

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungarica,* **2,** 299–318.

DeGroot, M. H. (1970). *Optimal Statistical Decisions.* McGraw-Hill, New York.

Ekeland, I. and R. Temam (1976). *Convex Analysis and Variational Problems.* North-Holland, Amsterdam.

Fortescue, T. R., L. S. Kershenbaum and B. E. Ydstie (1981). Implementation of self-tuning regulators with variable forgetting factors. *Automatica,* **17,** 831–835.

Gallager, R. G. (1968). *Information Theory and Reliable Communication.* Wiley, New York.

Hägglund, T. (1983). The problem of forgetting old data in recursive estimation. *Proc. IFAC Workshop on Adaptive System in Control and Signal Processing,* San Francisco.

Hägglund, T. (1985). Recursive estimation of slowly time-varying parameters. *Preprints 7th IFAC/IFORS Symp. Identification and System Parameter Estimation,* Vol. 2, pp. 1137–1142. University of York.

Kárný, M., A. Halousková, J. Böhm, R. Kulhavý and P. Nedoma (1985). Design of linear quadratic adaptive control: theory and algorithms for practice. Supplement to the journal *Kybernetika,* **21,** Nos 3, 4, 5, 6.

Kárný, M. and R. Kulhavý (1987). Structure determination of regression-type models for adaptive prediction and control. To appear in Spall, J. C. (Ed.), *Bayesian Analysis of Time Series and Dynamic Models.* Marcel Dekker, New York.

Kulhavý, R. (1986). Directional tracking of regression-type model parameters. *Preprints 2nd IFAC Workshop on Adaptive Systems in Control and Signal Processing,* Lund, pp. 97–102.

Kulhavý, R. and M. Kárný (1984). Tracking of slowly varying parameters by directional forgetting. *Preprints 9th IFAC Congress,* Vol. X, Budapest, pp. 178–183.

Kullback, S. and R. Leibler (1951). On information and sufficiency. *Ann. Math. Statist.,* **22,** 79–87.

Landau, I. D. and R. Lozano (1979). Unification of discrete time explicit model reference adaptive control designs. *Automatica,* **17,** 593–611.

Loève, M. (1960). *Probability Theory.* Van Nostrand, Princeton.

Perez, A. (1984). "Barycenter" of a set of probability measures and its application in statistical decision. *Proc. 6th Symp. Computational Statistics,* Prague, pp. 154–159.

Peterka, V. (1981). Bayesian approach to system identification. In Eykhoff, P. (Ed.), *Trends and Progress in System Identification,* Chap. 8, pp. 239–304. Pergamon Press, Oxford.

Saelid, S., O. Egeland and B. Foss (1985). A solution to the blow-up problem in adaptive controllers. *Modeling Ident. Control,* **6,** 39–56.

Savage, L. J. (1954). *The Foundations of Statistics.* Wiley, New York.

Wellstead, P. E. and S. P. Sanoff (1981). Extended self-tuning algorithm. *Int. J. Control,* **34,** 433–455.

Wittenmark, B. and K. J. Åström (1984). Practical issues in the implementation of self-tuning control. *Automatica,* **20,** 595–605.