INTERNATIONAL JOURNAL OF ADAPTIVE CONTROL AND SIGNAL PROCESSING, VOL. 9, 525–546 (1995)

# RECURSIVE APPROXIMATION BY ARX MODEL: A TOOL FOR GREY BOX MODELLING

MIROSLAV KÁRNÝ, ALENA HALOUSKOVÁ AND PETR NEDOMA

*Department of Adaptive Systems, Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, POB 18, 182 08 Prague 8, Czech Republic*

## SUMMARY

The presented procedure computes approximate probabilistic models of complex dynamic phenomena recursively with respect to an increasing amount of observed evidence. Measured, fictitious as well as simulated data can be used in combination for obtaining a reasonably conservative approximate model. Thus information from a number of sources can be systematically merged using a refinement of the recently proposed method of Bayesian pooling of imprecise opinions from a variety of experts. It can be applied recursively as the number of treated items grows.

The procedure provides (i) a new tool needed for grey as well as black box modelling, (ii) a novel adaptation of probabilistic models and (iii) an approximation of a given model by a simpler one.

The general procedure is applied to the autoregressive model with exogenous variables (ARX). This example illustrates the adopted approach and contributes to the solution of the following tasks: (i) estimation of an appropriate model structure; (ii) incorporation of prior knowledge into the initial conditions of recursive least squares; (iii) construction of a reference for an advanced forgetting technique; (iv) approximation of a complex analytic/simulation model by an ARX model.

The behaviour of the procedure is illustrated on typical examples.

KEY WORDS mismodelling; prior information; ARX model; Bayesian statistics

## 1. INTRODUCTION

Modelling of complex dynamic phenomena is a well-developed art. The use of complete (white box) models may be unfeasible owing to their complexity and/or the excessive cost of obtaining relevant quantitative information. This has led to the extensive use of universal (black box) descriptions such as ARX models or neural nets and to advocating free-of-prior-information approaches. Gradually, grey box models have been found to be an adequate compromise: simplified models suitable for decision making, prediction or control are used while additional effort is spent in exploiting available information sources. The success of this compromise depends heavily on a solution of the following complex problem: can partially incompatible, uncertain pieces of information of a different nature be merged into the approximate model?

In this paper a widely applicable solution to this problem is advocated. The basic idea of *identifying a simplified model using all informational sources* is not new. This technique is, however, rarely used as there is a serious danger of 'overfitting'. The model produced by this technique may be excellent on a subspace of particular system behaviours but poor outside of it. This inconsistency is caused by incorrectly extracting information from a variety of

CCC 0890–6327/95/060525–22

heterogenous sources. Our solution is based on information pooling, suitable for repetitive and/or not fully compatible pieces of information.[1,2]

The addressed problem, estimation with an approximate model, has no widely accepted solution. The approach presented here is related to an approximation of a joint distribution using its low-dimensional marginals.[3,4] This methodology, however, does not admit errors (incompatibilities) in the given marginals. Moreover, it deals mostly with discrete-valued data.

Another approach which is well-elaborated for recursive estimation is presented in Reference 5. It searches for models from a class of restricted complexity, when available information is compressed into a recursively feasible non-sufficient statistic. By more deeply analysing the problem and employing the theory of large deviations,[6] an adequate model may be viewed as a minimizer of the Kullback–Leibler distance of the empirical data distribution and that provided by the model. The necessity to know the type of data dependence is the price paid for the gained depth. Moreover, the justification of the approach is of an asymptotic nature. This is not fully in accord with the Bayesian methodology that is desired because of its compatibility with decision making under uncertainty.

At this discussion level, the presented approach can be understood as an attempt to find a hypermodel which relates the ideal likelihood function to the uncertain local models. Then a Bayesian estimate of this likelihood is found. The non-reliance on independence assumptions is compensated by the heuristic nature of the model and by the less definite theoretical support. In that sense the presented theory is still in its experimental phase of searching for promising candidates. In this sense the presented work falls in the range of attempts represented by Reference 7.

The theory is elaborated for ARX models, which are known to describe a wide range of dynamic systems. It provides a novel method for:

(i) the improved estimation of the model structure
(ii) the incorporation of expert knowledge into initial conditions of recursive least squares
(iii) the construction of a reference for an advanced forgetting technique[8]
(iv) the approximation of a complex analytic/simulation model by an ARX model.

## OVERVIEW OF PAPER

The *standard learning problem*, i.e. collecting information about an unknown quantity $\Theta$ contained in the observed data sequence $D(t)$, $t \in t^* \equiv \{1, 2, \ldots\}$,† is considered. As opposed to standard estimation tasks, data can be gained from a variety of sources. They may differ in precision and reliability and may be partially incompatible and/or repeat the same piece of information. Even the basic parametrized model is allowed to be an approximation of the dynamic system to be described. Under such conditions, the suitable learning procedure has to be *conservative*: the obtained estimate has to be equipped with an adequate description of uncertainty.

The *formalization and solution* proceed as follows:

1. The probabilistic model $g_{\Theta;\tau}$, relating the unknown quantity $\Theta \in \Theta^*$ to the data $D(\tau)$, $\tau \in \tau^* = \{1, \ldots, t\}$ $t \in t^*$, is introduced. Then a widely applicable probabilistic model of mismodelling is introduced in Section 3. By the model of mismodelling a probabilistic relationship of models $g_{\Theta;\tau}$ to an ideal likelihood $l_\Theta(t)$ is understood. If the models are

† $A^*$ denotes always the set of possible $A$-values.

correct, then the Bayes rule provides such a relationship. The model of mismodelling is needed when conditions on the Bayes rule applicability are violated, i.e. when models $g_{\Theta;\tau}$ are imprecise. The violation of independence type assumptions is a typical example of such an imprecision.

2. The mismodelling model is parameterized by an unknown correct likelihood function $l_\Theta(t)$. Also additional nuisance (hyper)parameters $w$ and $R$ are introduced which enhance the flexibility of the model. The multivariate (hyper)parameter $(l, w, R)$ is assigned a flat (hyper)prior probability density function (PDF) and a standard Bayesian estimation is applied at the 'upper' (hyper)likelihood level. The maximum *a posterior* probability (MAP) point estimate $\hat{l}_\Theta(t)$ of the correct $l_\Theta(t)$ is then found in Section 4.
3. The sequence of the estimates $\hat{l}_\Theta(t)$ is inspected in Section 5. Models $g_{\Theta;\tau}$ from a subset of the exponential family are considered. For them the approximate estimation is recursively implementable. The theory is applied to an ARX model by inserting specific functions for the general symbols and by performing the necessary analytic computations, see Section 6. The resulting algorithm is summarized in Section 7 and typical illustrative examples are in Section 8.

## 3. MODELLING AND MISMODELLING

### 3.1. *Model of uncertain dynamic systems*

Here, a system model is described and the notation fixed.

The input–output behaviour of a dynamic system is inspected. The sequence of its inputs $U(t) = (u_1, \ldots, u_t)$, $t \in t^*$, consists of the manipulated data. The remaining measured data from $D(t)$ form the output sequence $Y(t) = (y_1, \ldots, y_t)$, i.e. $D(t) = (U(t), Y(t))$.

The *parametrized model* of the input–output relationship $(Y^*(\tau - 1),\ U^*(\tau)) \rightarrow y_\tau^*$ is the sequence of probability density functions (PDFs) of $y_\tau$ conditioned on $D(\tau - 1)$, $u_\tau$ and on an unknown $i_\Theta$-dimensional parameter $\Theta \in \Theta^*$:

$$g_{\Theta;\tau}^{\text{true}} \equiv f(y_\tau | D(\tau - 1),\ u_\tau,\ \Theta), \quad \tau \in \tau^*$$

In order to avoid the difficult area of stochastic processes a finite set $\Theta^* = \{\Theta^{[1]}, \ldots, \Theta^{[m]}\}$, $m > \infty$, is considered. The obtained results are extended to the continuous case of real-valued $\Theta$ by taking a formal limit $m \rightarrow \infty$.

In the Bayesian approach adopted, the set $\Theta^*$ of possible $\Theta$-values is equipped with a (pre)prior probability. It is described by the probability function (PF) $P(\Theta^{[i]})$ which distributes a degree of belief to various $\Theta \in \Theta^*$. In the continuous-space case this probability is specified by the (pre)prior PDF $p(\Theta)$ through the formula $P(\Theta^{[i]}) = \int_{\Theta^{[*i]}} p(\Theta)\, \mathrm{d}\Theta$. The sets $\Theta^{[*i]} \subset \Theta^*$ are neighbourhoods of the representative points $\Theta^{[i]}$.

The symbol is reserved for the (pre)prior PDF $p(\cdot)$, otherwise the letter $f(\cdot)$ is used for PDFs and PFs. The expectation given by $p$ is denoted $\mathscr{E}_p$.

### 3.2. *Model of mismodelling*

This subsection models the approximate nature of the system description. In order to stress the specific nature of the addressed problem, we call it the model of mismodelling. As in any modelling, the selection of the adequate description contains a substantial amount of heuristics.

Let us suppose that at any time instant $\tau \in \tau^*$ only an approximate description $g_{\Theta;\tau}$ of the system is available, i.e. $g_{\Theta;\tau} \approx g_{\Theta;\tau}^{\text{true}}$. The approximation is assumed to be good locally but insufficient for modelling global relationships within the possible data sequences $D(t)$.

The global ('true') likelihood, i.e., $f^{\text{true}}(D(t)\,|\,\Theta)$ is taken as a function of $\Theta$ and contains complete uncertain-data-based information on the unknown parameter $\Theta$. The Bayesian set-up combines it with non-data-based information quantified by the (pre)prior PF $P(\Theta)$ into the entire-information-compressing posterior PF

$$f(\Theta\,|\,D(t)) \propto f^{\text{true}}(D(t)\,|\,\Theta)P(\Theta) \propto l_\Theta(t)P(\Theta).$$

Here, '$\propto$' means equality up to a $\Theta$-independent factor and $l_\Theta(t)$ denotes the likelihood, i.e. the PDF $f^{\text{true}}(D(t)\,|\,\Theta)$ possibly with a factor independent of $\Theta$ omitted. Under broadly met natural conditions of control[9] the likelihood is the product of (true) partial likelihoods $g_{\Theta;\tau}^{\text{true}}$

$$l_\Theta(t) \equiv \prod_{\tau=1}^{t} g_{\Theta;\tau}^{\text{true}} \tag{1}$$

The severity of the non-equality

$$l_\Theta(t) \neq \prod_{\tau=1}^{t} g_{\Theta;\tau}$$

that *we wish to respect* depends on the degree of non-equality of $g_{\Theta;\tau}$ and $g_{\Theta;\tau}^{\text{true}}$ for $\tau \in \tau^*$.

Let us assume that the ratios $\rho_{\Theta;\tau} = g_{\Theta;\tau}^{\text{true}}/g_{\Theta;\tau}$ are well-defined and positive on a common support included in the support of the (pre)prior distribution. Then, by (i) taking the logarithm of (1) and normalizing the result by the number of samples $t$, (ii) using the definition of the ratio $\rho$, (iii) adding the term $\ln(g_{\Theta;\tau})$ to both sides of the transformed identity, (iv) leaving only $\ln(g_{\Theta;\tau})$ on the left-hand side and (v) adding and subtracting the term $(w_\tau(t) - 1/t)\ln(l_\Theta(\Theta))$, with some scalar $w_\tau(t)$ to and from the resultant right-hand side, we obtain

$$\begin{aligned}\ln(g_{\Theta;\tau}) &= \frac{1}{t}\ln(l_\Theta(t)) + \left[\ln(g_{\Theta;\tau}) - \frac{1}{t}\sum_{i=1}^{1}\ln(g_{\Theta;i})\right]_{\mathrm{a}} + \left[\frac{1}{t}\sum_{i=1}^{1}\ln(\rho_{\Theta;i})\right]_{\mathrm{b}} \\ &\equiv w_\tau(t)\ln(l_\Theta(t)) + e_{\Theta,\tau}(t)\end{aligned}$$

where

$$e_{\Theta,\tau}(t) = \left[\ln(g_{\Theta;\tau}) - \frac{1}{t}\sum_{i=1}^{1}\ln(g_{\Theta;i})\right]_{\mathrm{a}} + \left[\frac{1}{t}\sum_{i=1}^{1}\ln(\rho_{\Theta;i})\right]_{\mathrm{b}} - \left(w_\tau(t) - \frac{1}{t}\right)\ln(l_\Theta(t))$$

The introduced noise $e_{\Theta,\tau}(t)$ is the sum of the following: $[\cdot]_{\mathrm{a}}$, the deviation of $\ln(g_{\Theta;\tau})$ from the sample mean of all $g_{\Theta;i}$, $i \in \tau^*$; $[\cdot]_{\mathrm{b}}$, the sample mean of the modelling errors; $(w_\tau(t) - 1/t)\ln(l_\Theta(t))$, a function proportional to the logarithm of the ideal likelihood. The derived relationship is obtained by deduction from the correct expression for the log likelihood of interest. Now it is extended by an inductive step. When doing this, an arbitrary $t \in t^*$ is fixed and suppressed in notation until Section 5. $\Theta$, and $\tilde{\Theta}$ denote arbitrary fixed points in $\Theta^*$.

The induction is based on *interpreting the introduced noise* $e_\Theta \equiv [e_{\Theta;1}, \ldots, e_{\Theta;t}]^{\mathrm{T}}$ (superscript T denotes transposition) *as a normal multivariate variable with zero mean and with the specific correlation structure*

$$\operatorname{cov}[e_\Theta e_{\tilde{\Theta}}^{\mathrm{T}}\,|\,l, R] = \frac{\Delta(\Theta, \tilde{\Theta})R}{P(\Theta)} \tag{2}$$

where $\Delta$ is Kronecker symbol. $R \geqslant 0$ (positive semidefinite) is an arbitrary $(t, t)$ covariance matrix.

Let us comment on this choice.

1. *Zero mean* can be achieved by subtracting a suitable portion of $\ln(l_\Theta)$ from $[\cdot]_a + [\cdot]_b$, by removing a systematic bias in the noise. In other words, the ideal log likelihood $\ln[l_\Theta]$ is assumed to be the only common factor (in terms of factor analysis) in the matrix of logarithmic local models $\ln(g_{\Theta,\tau})$, $\tau \in \tau^*$, $\Theta \in \Theta^*$. This matrix plays the role of observable data. Note that the smaller the bias-correcting term $(\mathrm{w}_\tau(t) - 1/t)\ln(l_\Theta(t)) \approx 0$ is, the closer the model $g_{\Theta;\tau}$ is to $g_{\Theta;\tau}^{\text{true}}$. Thus for good models

$$w_\tau \approx 1/t \tag{3}$$

2. *Covariance structure* postulates a lack of mutual correlations of $e_\Theta$ and $e_{\tilde{\Theta}}$ for $\Theta \neq \tilde{\Theta}$. It reflects the presumption that mutual relationships of $g_{\Theta;\tau}$ and $g_{\tilde{\Theta};\tilde{\tau}}$ are fully respected by the selected first moment. The inverse dependence of dispersions on the (pre)prior PF $P(\Theta)$ respects the fact that any reasonable modelling is more careful nearby *a priori* expected values of $\Theta$ and admits larger errors for less expected possibilities. In other words, the columns of the $(t, m)$ matrix $G$ of the weighted (hyper)data with entries

$$G_{\tau,i} \equiv p^{0\cdot 5}(\Theta^{[i]})\ln(g_{\Theta^{[i]};\tau}), \quad \tau \in \tau^*, \quad i = 1, \ldots, m \tag{4}$$

are assumed to have a common $\Theta$-invariant covariance $R$. The unrestricted form of the mutual correlations in the 'time-direction' makes the model flexible. All types of dependences between various time instants are admitted which conform with the addressed problem.
3. *Normality* could be justified by the maximum entropy principle: the most uncertain model with finite covariance is normal. It is, however, fair to say that the analytical feasibility is the decisive reason for selecting this model. It is known to be a good approximation of all distributions with a practically limited support. It fails in describing heavy-tail situations with possible outliers. From this viewpoint the assumption that all $g_{\Theta;\tau}$ are good (local) approximations of the corresponding $g_{\Theta;\tau}^{\text{true}}$ is merely reconfirmed.

In summary, the adopted assumptions specify a (hyper)parametrized model postulating that the (hyper)data $G$ in (4) conditioned on the (hyper)parameter consisting of $w \equiv [w_1, \ldots, w_t]^{\mathrm{T}}$, $L \equiv [P^{0\cdot 5}(\Theta^{[1]})\ln(l_{\Theta^{[1]}}), \ldots, P^{0\cdot 5}(\Theta^{[m]})\ln(l_{\Theta^{[m]}})]$ and $R \geqslant 0$ are described by the normal PDF

$$f(G \mid L, w, R) \equiv \mathcal{N}_G(wL^{\mathrm{T}}, I \otimes R), \tag{5}$$

where $\mathcal{N}_X(\hat{X}, \mathscr{X})$ denotes the normal PDF of $X$ given by the expected value $\hat{X}$ and covariance $\mathscr{X}$. $I$ is a unit matrix of appropriate dimension and $\otimes$ denotes the Kronecker product, which expresses formally the covariance structure of type (2).

### *3.3. Prior distribution of (hyper)parameter*

As outlined in Section 2, the posterior PDF of the (hyper)parameter $(L, R, w)$ conditioned on the (hyper)data $G$ is sought. For this the (hyper)parametrized model proposed in the previous subsection has to be complemented by a (hyper)prior PDF on $(L, R, w)$.

The (hyper)parametrized model reflects the commonly accepted prior knowledge. Thus it is relevant to choose *a priori* independent $L$, and $R$ with a very flat (hyper)prior PDF.

The key quantity $L$ is simply assigned the (improper) uniform prior distribution. The covariance matrix $R$ is also given the improper uniform prior distribution on $R \geqslant 0$. It is obtained, however, as a limit (for $\gamma \to 0^+$) of the flat inverse Wishart distribution

$$f(R) \propto \exp[-0{\cdot}5\gamma \operatorname{tr}(R^{-1})], \qquad \gamma > 0, \qquad \operatorname{tr} = \text{matrix trace}$$

The use of the regularizing parameter $\gamma > 0$, which gives a slight preference to independence of (hyper)noise entries, simplifies the necessary algebra. The asymptotic case with $\gamma \to 0^+$ has to be taken, however, as the generic one. It includes cases with a singular $R$ which corresponds to the exact modelling and covers repetitions of information pieces.

The choice of the prior PDF for $w$ is the most peculiar one. It serves as a method of dealing with the non-uniqueness of the adopted parametrization; namely, the expected value determining factor analysis model (5) is not unique. To demonstrate this, let $\bar{L}$ denote $L$ normalized to a unit norm, $\|\bar{L}\|^2 \equiv \bar{L}^{\mathrm{T}}\bar{L} = 1$, and $\alpha$ be any nonzero scalar. Then,

$$w\bar{L}^{\mathrm{T}} = [w/\alpha][\alpha\bar{L}^{\mathrm{T}}] = [w/\alpha][p^{0{\cdot}5}(\Theta^{[1]})\ln(\bar{l}^{\alpha}_{\Theta^{[1]}}), \ldots, p^{0{\cdot}5}(\Theta^{[m]})\ln(\bar{l}^{\alpha}_{\Theta^{[m]}})]$$

The required estimate $\hat{f}(\Theta \mid G)$, obtained from the ideally separated factor $[\alpha \ln(\bar{l}_{\Theta})]$, has the form

$$\hat{f}(\Theta \mid G) = \frac{\bar{l}^{\alpha}_{\Theta} P(\Theta)}{\sum_{\Theta \in \Theta^*} \bar{l}^{\alpha}_{\Theta} P(\Theta)} \tag{6}$$

which is very different for different $\alpha$s. At the same time the proportionality $l_{\Theta} \propto f^{\text{true}}(D(t) \mid \Theta)$ cannot be exploited due to the lack of knowledge of this function on the whole $D^*(t)$. This function is evaluated (and estimated) for the 'measured' data sample $D(t)$ only.

The incorporation of prior information is the only possible remedy for this problem. We claim that the available information is well quantified by the PDF

$$f(w \mid L, R) = \mathcal{N}_w\left(w_0, \frac{R}{\varepsilon\alpha^2}\right)$$

where $w_0 = \mathbb{1}/\operatorname{rank}(R)$, $\mathbb{1}$ is a $t$-vector of units, $\varepsilon > 0$ is a small scalar and $\alpha^2 = \|L\|^2 = L^{\mathrm{T}}L$.

This choice has the following justification.

1. The *normal form* is chosen in order to preserve the necessary degree of numerical feasibility. Owing to the considered flatness ($\varepsilon$ is small) the assumption is not restrictive.
2. The *expected value* $w_0 = \mathbb{1}/\operatorname{rank}(R)$ respects that the restricted applicability or the 'neighbourhood' of formula (1) for the 'true' likelihood is modelled. According to (3), $w$ is expected to be close to $\mathbb{1}/t$. For any regular $R$, $\operatorname{rank}(R) = t$ and the discussed expected value can be equivalently expressed as $\mathbb{1}/\operatorname{rank}(R)$. For the generic singular case ('visible' when $\gamma \to 0^+$), the alternative expression becomes key method for dealing with the data repetition. It respects the fact that coincidence of the noise in several evidence pieces can be caused by repetitions only. It is sufficient to recognize that the (hyper)data are always continuous-valued random variables. Such data should be counted only once in the conservatism requiring situation with which we are dealing.
3. The *chosen covariance structure* says essentially that the stringency of the requirement (3) increases with increasing precision of local models and with increasing confidence in $L$ (increasing $\alpha$ in (6)).

### *3.4. Summary of model of mismodelling*

The introduced (hyper)parametrized model and the chosen prior distribution on the (hyper)parameter define, through the Bayes rule, the posterior distribution

$$f(\bar{L}, \alpha, w, R \mid G) \propto \mathcal{N}_G(\alpha w \bar{L}^{\mathrm{T}}, I \otimes R)\mathcal{N}_w\left(w_0, \frac{R}{\varepsilon\alpha^2}\right)\exp[-0{\cdot}5\gamma\,\mathrm{tr}(R^{-1})] \tag{7}$$

where $\bar{L}_i = P^{0\cdot5}(\Theta^{[i]})\ln(\bar{l}_{\Theta^{[i]}})$, $i = 1, \ldots, m$. $P(\Theta)$ denotes (pre)prior PDF of the unknown parameter $\Theta \in \Theta^* = \{\Theta^{[i]}\}_{i=1}^{m}$. $\bar{l}_{\Theta^{[i]}}$ is the key unknown quantity, namely the likelihood corresponding to the true (unknown) description of the modelled dynamic system. It is normalized so that $\|\bar{L}\|^2 = 1$. $\alpha$ is an unknown scalar determining the norm of the unknown 'true' log likelihood $L$ (corresponding to the true description of the modelled dynamic system), i.e. $L = \alpha\bar{L}$. $w$ is an unknown $t$-dimensional vector. It is a nuisance parameter defining the structure of the problem. $I$ is a unit matrix and $\otimes$ denotes the Kronecker product. $R$ is a positive semidefinite $(t, t)$ covariance matrix. It is nuisance parameter as well. $G_{\tau,i} = [P(\Theta^{[i]})^{0\cdot5}\ln(g_{\Theta^{[i]};\tau})]_{\tau\in\tau^*, i=1,\ldots,m}$ is the matrix containing (hyper)data available for estimating the unknown likelihood. It consists of weighted logarithms of local models. They are interpreted as an approximation of the 'true' description of the system with inputs $u$ and outputs $y$: $g_{\Theta,\tau} \approx f^{\mathrm{true}}(y_\tau / D(\tau - 1), u_\tau, \Theta)$, $D(t) = (y_1, u_1, \ldots, y_t, u_t)$. $w_0(t) = \mathbb{1}/\mathrm{rank}(R)$, where a $\mathbb{1}$ is a $t$-vector of units. $\gamma > 0$ is regularizing scalar which will be sent to zero. $\varepsilon > 0$ is a sufficiently small auxiliary scalar.

## 4. MAP ESTIMATE OF UNKNOWN LIKELIHOOD

The posterior PDF (7) compresses all available information on the (hyper)parameter $(\bar{L}, \alpha, w, R)$. Ideally the marginal (posterior) distribution on $L = \alpha\bar{L}$ should be computed and some point or interval estimate found. For computational reasons, only the maximizer $\hat{L}$ — the maximum *a posteriori* probability (MAP) estimate — is constructed here. This point estimate at (hyper)level is asymptotically $(m \to \infty)$ a function at the basic probabilistic level. Thus the MAP estimate of the entire likelihood and consequently of the posterior PDF on $\Theta$ is found. In this way, the problem of combining the assumed 'nasty' information sources is solved.

The MAP estimate is derived using the singular value decomposition[10] of the $(t, m)$ matrix $G$:

$$\begin{gathered} G = S[\mathfrak{D}, 0]V^{\mathrm{T}}, \qquad SS^{\mathrm{T}} = S^{\mathrm{T}}S = I, \qquad VV^{\mathrm{T}} = V^{\mathrm{T}}V = I, \qquad I = \text{unit matrix} \\ S = [S^{[1]}, \ldots, S^{[t]}], \qquad V = [V^{[1]}, \ldots, V^{[m]}] \\ \mathfrak{D} = \mathrm{diag}[\mathfrak{D}_1, \ldots, \mathfrak{D}_\delta, 0, \ldots, 0], \qquad \mathfrak{D}_1 \geqslant \mathfrak{D}_2 \cdots \geqslant \mathfrak{D}_\delta > 0, \qquad \delta = \mathrm{rank}(G) \leqslant \min(t, m) \end{gathered} \tag{8}$$

*Proposition 1* (*MAP estimate of L*)

For sufficiently small $\varepsilon$ and $\gamma \to 0^+$ the MAP estimate $\hat{L}$ of $L$ is $\hat{L} = G^{\mathrm{T}}\nu$, where $\nu = \alpha\bar{\nu} \equiv \alpha S^{[1]}\mathbb{1}$, $\|\bar{\nu}\| = 1$, called the *mixing vector*, is the eigenvector corresponding to the maximum eigenvalue of the *mixing matrix* $Q^{[m]} \equiv GG^{\mathrm{T}}$.

The length $\alpha$ of the mixing vector is $\alpha = \delta/(\nu^{\mathrm{T}}\mathbb{1})$.

*Proof.* First the maximization over $R$ is performed. For the regular $R$ considered now, the vector $w_0(R) \equiv \mathbb{1}/\mathrm{tr}(R) = \mathbb{1}/t$, i.e. it does not influence this optimization part. Its role becomes significant for $\gamma \to 0^+$.

Writing explicitly normal PDFs in (7), rotating the argument of the trace in the exponential and denoting

$$\tilde{C} \equiv \gamma I + (G - wL^{\mathrm{T}})(G - wL^{\mathrm{T}})^{\mathrm{T}} + \varepsilon(w - \mathbb{1}/t)(w - \mathbb{1}/t)^{\mathrm{T}} L^{\mathrm{T}} L > 0$$

we get $f(R, L, w \mid G) \propto |R|^{-0{\cdot}5(m+1)} \exp[-0{\cdot}5 \operatorname{tr}(R^{-1}\tilde{C})]$. This function is maximized by $\hat{R} = \tilde{C}/(m+1)$, which gives the value $f(\hat{R}, L, w \mid G) \propto |C|^{-0{\cdot}5(m+1)}$, where

$$C \equiv \gamma I + (G - wL^{\mathrm{T}})(G - wL^{\mathrm{T}})^{\mathrm{T}} + \varepsilon(w - \hat{w}_0)(w - \hat{w}_0)^{\mathrm{T}} q^{\mathrm{T}} q > 0, \qquad \hat{w}_0 = \mathbb{1}/\operatorname{rank}(C).$$

Notice the implicit form of this estimate.

The maximization over $w$ and $L$ reduces to the minimization of the determinant $|C|$. Using the singular value decomposition (8) and completion of squares, the minimized determinant can be written in the form

$$\begin{aligned}
|C| &= |\gamma I + ([\mathscr{D}, 0] - vx^{\mathrm{T}})([\mathscr{D}, \mathrm{o}] - vx^{\mathrm{T}})^{\mathrm{T}} + \varepsilon(v - v_0)(v - v_0)^{\mathrm{T}} x^{\mathrm{T}} x| \\
&= \left|\gamma I + \mathscr{D}^2 - \frac{[\mathscr{D}, 0]\bar{x}\bar{x}^{\mathrm{T}}[\mathscr{D}, 0]^{\mathrm{T}}}{1+\varepsilon} + \frac{\varepsilon}{1+\varepsilon}(\alpha^2 v_0 v_0^{\mathrm{T}} - v_0 x^{\mathrm{T}}[\mathscr{D}, 0]^{\mathrm{T}} - [\mathscr{D}, 0] x v_0^{\mathrm{T}})\right. \\
&\quad \left. + (1+\varepsilon)\alpha^2 \left(v - \frac{[\mathscr{D}, 0]\alpha\bar{x} + \varepsilon\alpha^2 v_0}{(1+\varepsilon)\alpha^2}\right)\left(v - \frac{[\mathscr{D}, 0]\alpha\bar{x} + \varepsilon\alpha^2 v_0}{(1+\varepsilon)\alpha^2}\right)^{\mathrm{T}}\right|
\end{aligned}$$

where $x \equiv VL \equiv \alpha\bar{x}$, $v = S^{\mathrm{T}} w$ and $v_0 = S^{\mathrm{T}} \hat{w}_0$.

The determinant is minimized by $v = [\mathscr{D}, 0]\bar{x}/\alpha + \varepsilon v_0/(1+\varepsilon)$ and the corresponding value has the form

$$|C| = \left|\gamma I + \mathscr{D}^2 - [\mathscr{D}, 0]\bar{x}\bar{x}^{\mathrm{T}}[\mathscr{D}, 0]^{\mathrm{T}} + \frac{\varepsilon}{1+\varepsilon}(\alpha v_0 - [\mathscr{D}, 0]\bar{x})(\alpha v_0 - [\mathscr{D}, 0]\bar{x})^{\mathrm{T}}\right|$$

Denoting $z = (\gamma I + \mathscr{D}^2)^{-0{\cdot}5}[\mathscr{D}, 0]\bar{x}$ and $\mu = (\gamma I + \mathscr{D}^2)^{-0{\cdot}5} v_0$, the minimized function becomes

$$\begin{aligned}
|C| &\propto \left| I - zz^{\mathrm{T}} + \frac{\varepsilon}{1+\varepsilon}(\alpha\mu - z)(\alpha\mu - z)^{\mathrm{T}} \right| \\
&= 1 - z^{\mathrm{T}} z + \frac{\varepsilon}{1+\varepsilon}[(1 - z^{\mathrm{T}} z)(\alpha\mu - z)^{\mathrm{T}}(\alpha\mu - z) + (\alpha z^{\mathrm{T}}\mu - z^{\mathrm{T}} z)^2]
\end{aligned}$$

This quadratic function of $\alpha$ is minimized by

$$\alpha = \frac{\mu^{\mathrm{T}} z}{\mu^{\mathrm{T}}\mu(1 - z^{\mathrm{T}} z) + (z^{\mathrm{T}}\mu)^2}$$

giving

$$|C| \propto \frac{1 - z^{\mathrm{T}} z}{1+\varepsilon}\left(1 + \varepsilon \frac{\mu^{\mathrm{T}}\mu}{\mu^{\mathrm{T}}\mu(1 - z^{\mathrm{T}} z) + (z^{\mathrm{T}}\mu)^2}\right)$$

It remains to find $z$ minimizing this expression. This determines the direction $\hat{\bar{L}}$. The discussion is restricted to the generic case of simple positive singular values $\mathscr{D}_1 > \mathscr{D}_2 > \cdots > \mathscr{D}_\delta > 0$.

Let $z^{[i]}$ correspond to $x^{[i]}$ = vector having unity in the $i$th position as the only non-zero entry. Such $z^{[i]}$ correspond to the unit-length vectors $\bar{L}^{[i]} = V^{[i]}$ for $i = 1, \ldots, \delta$. For them

$\lim_{\gamma\to 0^+}(1-(z^{[i]})^{\mathrm{T}}z^{[i]})=0$ and this limit is positive for any other possible $z$. Thus the minimizer valid for $\gamma\to 0^+$ lies within the finite set of $z^{[i]}$ only. For sufficiently small $\varepsilon$, different singular values and any $\gamma>0$ a direct inspection shows that $z^{[1]}$ gives the smallest value.

In order to obtain the final results for $\gamma\to 0^+$, it remains to notice that, for the chosen $v$, $\alpha$, $\bar{x}$ and $\gamma\to 0$, $\mathrm{rank}(C)=\mathrm{rank}(G)$, to express $\hat{L}=\alpha V^{[1]}=\alpha/\mathfrak{D}_1 G^{\mathrm{T}}S^{[1]}$ and to include $1/\mathfrak{D}_1$ in $\alpha$. □

The chosen form of the resultant estimate allows a simple formal transition to the continuous counterpart of Proposition 1.

*Proposition 2 (MAP estimate of $f(\Theta\,|\,D(t))$)*

Let $g_{\Theta;\tau}$ be positive on a common support $\Theta^*$, sufficiently smooth in $\Theta\in\Theta^*$, where $\Theta^*$ is an $i_\Theta$-dimensional real space. Let also $\mathscr{E}_p[\ln^2(g_{\cdot;\tau})]=\int\ln^2(g_{\Theta;\tau})p(\Theta)\,\mathrm{d}\Theta<\infty$ for all $\tau\in\tau^*$.

Then the MAP estimate $\hat{f}(\Theta\,|\,D(t))$ of $f(\Theta\,|\,D(t))$ is

$$\hat{f}(\Theta\,|\,D(t))=\frac{\prod_{\tau=1}^{t} g_{\Theta;\tau}^{\nu_\tau}p(\Theta)}{\int\prod_{\tau=1}^{t} g_{\Theta;\tau}^{\nu_\tau}p(\Theta)\,\mathrm{d}\Theta}$$

where $\nu=[\nu_1,\ldots,\nu_t]^{\mathrm{T}}$ is the eigenvector corresponding to the maximum eigenvalue of the symmetric positive semidefinite $(t,t)$ matrix $Q$, with entries

$$Q_{\tau\tilde{\tau}}=\mathscr{E}_p[\ln g_{\cdot;\tau}\ln g_{\cdot;\tilde{\tau}}],\quad \tau,\tilde{\tau}\in\tau^* \tag{9}$$

The length $\alpha$ of $\nu=\alpha\bar{\nu}$, $\|\bar{\nu}\|=1$, is given by the formula

$$\alpha=\frac{\mathrm{rank}(Q)}{\nu^{\mathrm{T}}\mathbf{1}} \tag{10}$$

*Proof.* This is a direct consequence of Proposition 1 for $m\to\infty$ and $\max P(\Theta^{[i]})\to 0$, as under the adopted integrability conditions $Q^{[m]}\to\Theta$ in (9). Then the formula for the discrete version of the MAP estimate of $L$, definitions of the involved quantities and invariance of the MAP estimate to regular transformations (here the Bayes rule) are used. □

## 5. RECURSIVE IMPLEMENTATION

Recursive learning, estimation, etc. deal with a sequence of problems by exploiting the previous results to obtain a 'cheap' solution of the current task.

Here the sequence of MAP estimates $\hat{f}(\Theta\,|\,D(t))$ of the posterior PDFs $f(\Theta\,|\,D(t))$, $t\in t^*$, is linked. In a generic case the entire $(t,t)$ matrix $Q(t)$ in (9) is needed for this purpose. This means that the dimension of the sufficient statistic grows with $t$. Thus, an approximation based on a reduced (not sufficient) statistic is needed. The choice of a suitable recursively feasible approximation is obvious here: the data $\ln(g_{\Theta;\tau})$, $\tau\in\tau^*$, have to be projected onto a fixed-finite dimensional subspace embedded into the space of possible estimates $\hat{f}(\Theta\,|\,D(t))$. A general description of this approach, closely related to Reference 11, can be found in Reference 2. For the treated exponential subfamily no such projection is needed.

### 5.1. *Recursive exponential family*

The models $g_{\Theta;\tau}$ are said to belong to the *recursive exponential family* if

(i) a known transformation, say $\tilde{T}$, exists mapping the *data vector* $\Phi_\tau = [y_\tau, \phi_\tau]$ and the parameter $\Theta$ on a data-dependent $i_\psi$-vector $\psi_\tau$ and an $i_\psi$-vector function $\tilde{\Theta}$ of $\Theta$, respectively, so that

$$\ln(g_{\Theta;\tau}) = -0{\cdot}5\psi_\tau^{\mathrm{T}}\tilde{\Theta} \tag{11}$$

(ii) the models are determined by the *regression vector*, a known $i_\phi$-dimensional function $\phi_\tau$ of $D(\tau - 1)$, $u_\tau$ which can be recursively updated, i.e. $\phi_{\tau+1}$ is a known function of $\phi_\tau$ and $y_\tau$, $u_{\tau+1}$

(iii) the elements of the mixing matrix are finite.

An explicit construction leading to the *canonical form* (11) is given in Section 6 for the ARX model. The recursive feasibility is the key advantage of this model class, which is rich enough to approximate a wide class of realistic systems. Its known feasibility applies also to the inspected approximation problem, as is demonstrated in the subsequent text.

### 5.2. *Mixing matrix and vector*

With the vectors $\psi_\tau^{\mathrm{T}}$, $\tau \in \tau^*$, taken as data-dependent rows of the $(t, i_\psi)$ matrix $\Psi(t)$,

$$\Psi^{\mathrm{T}}(t) = [\psi_1, \ldots, \psi_t] \tag{12}$$

the mixing $(t, t)$ matrix $Q(t)$ in (9) becomes

$$Q(t) = \Psi(t)\mathcal{A}\Psi^{\mathrm{T}}(t) \tag{13}$$

The fixed $(i_\psi, i_\psi)$ matrix $\mathcal{A}$ is defined by

$$\mathcal{A} = \mathscr{E}_p[\tilde{\Theta}\tilde{\Theta}^{\mathrm{T}}]$$

Note that the omitted factor of 0·25 leaves the eigenstructure unaffected.

The MAP estimate of the likelihood takes the form

$$\ln(\hat{l}_\Theta(t)) = -0{\cdot}5\nu^{\mathrm{T}}(t)\Psi(t)\tilde{\Theta} \tag{14}$$

where $\nu(t)$ is the mixing $t$-vector, i.e.

$$\rho\nu(t) = Q(t)\nu(t) = \Psi(t)\mathcal{A}\Psi^{\mathrm{T}}(t)\nu(t) \tag{15}$$

and $\rho$ is the largest eigenvalue of the mixing $(t, t)$ matrix $Q(t)$ in (13).

### 5.3. *The recursive form of $\hat{l}_\Theta(t)$*

The likelihood estimate $\hat{l}_\Theta(t)$ can be evaluated recursively within a limited memory as shown below.

Let $T(t)$ be an orthogonal data-dependent $(t, t)$-matrix such that

$$T(t)\Psi(t) = \begin{bmatrix} \mathcal{B}(t) \\ 0 \end{bmatrix} \tag{16}$$

where $\mathcal{B}(t)$ is a full-rank $(i_{\mathcal{B}(t)}, i_\Psi)$-matrix. Such a transformation exists[10] and

$$i_{\mathcal{B}(t)} \leq \min(t, i_\Psi) \tag{17}$$

The Equation (14), orthogonality of $T(t)$ and the identity (16) imply

$$-2\ln(\hat{l}_\Theta(t)) = \nu^{\mathrm{T}}(t)\Psi(t)\tilde{\Theta} = \nu^{\mathrm{T}}(t)T^{\mathrm{T}}(t)T(t)\Psi(t)\tilde{\Theta} = \nu^{\mathrm{T}}(t)T^{\mathrm{T}}(t)\begin{bmatrix}\mathcal{B}(t)\\ 0\end{bmatrix}\tilde{\Theta}$$

The equation determining the $t$-vector $T(t)\nu(t)$ is obtained by multiplying (15) by $T(t)$ and using (16):

$$\rho T(t)\nu(t) = T(t)Q(t)\nu(t) = T(t)\Psi(t)\mathcal{A}\Psi^{\mathrm{T}}(t)T^{\mathrm{T}}(t)T(t)\nu(t) = \begin{bmatrix}\mathcal{B}(t)\\ 0\end{bmatrix}\mathcal{A}[\mathcal{B}^{\mathrm{T}}(t), 0]T(t)\nu(t)$$

This implies that

$$T(t)\nu(t) = \begin{bmatrix}z(t)\\ 0\end{bmatrix}$$

where the dimension of the non-zero part $z(t)$ does not exceed the bounded value $i_{\mathcal{B}(t)}$ in (17); $z(t)$ is the eigenvector corresponding to the largest eigenvalue of the non-zero matrix $W(t)$,

$$W(t) = \mathcal{B}(t)\mathcal{A}\mathcal{B}^{\mathrm{T}}(t) \tag{18}$$

It provides the estimate of the likelihood in the form

$$\ln(\hat{l}_\Theta(t)) = -0{\cdot}5z^{\mathrm{T}}(t)\mathcal{B}(t)\tilde{\Theta} \tag{19}$$

For achieving the overall recursivity of the needed evaluations within a bounded memory space, it is important that $\mathcal{B}(t+1)$ can be constructed directly from $\mathcal{B}(t)$ and $\psi^{\mathrm{T}}_{t+1}$. The evaluation is based on the simple observation (see (12) and (13), that

$$\begin{bmatrix}T(t) & 0\\ 0 & 1\end{bmatrix}\Psi(t+1) = \begin{bmatrix}T(t) & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}\Psi(t)\\ \psi^{\mathrm{T}}_{t+1}\end{bmatrix} = \begin{bmatrix}\mathcal{B}(t)\\ 0\\ \psi^{\mathrm{T}}_{t+1}\end{bmatrix} \tag{20}$$

Thus, as the product of orthogonal matrices is orthogonal, it is sufficient (for constructing $T(t+1)$) to find the orthogonal transformation 'zeroing' the new row $\psi^{\mathrm{T}}_{t+1}$ with respect to $\mathcal{B}(t)$. Boundedness of the $\mathcal{B}(t+1)$ dimensions is guaranteed by (17).

It remains to evaluate the length (10) of the mixing vector. It is determined by:

(i) $\mathrm{rank}(Q(t))$ which coincides with $i_{\mathcal{B}}$ when assuming the generic case of regular A.

(ii) The product $\bar{\nu}^{\mathrm{T}}(t)\mathbb{1} = \bar{z}^{\mathrm{T}}(t)b(t)$, where $b(t)$ consists of $i_{\mathcal{B}}$ leading entries of the $t$-vector $\mathbb{1}$ rotated in the same way as $\Psi(t)$; thus, the transformation zeroing the new row $\psi^{\mathrm{T}}_{t+1}$ when transforming $\mathcal{B}(t) \rightarrow \mathcal{B}(t+1)$ has to be applied to $[b^{\mathrm{T}}(t), 1]^{\mathrm{T}}$ in order to get $[b^{\mathrm{T}}(t+1), \bullet]^{\mathrm{T}}$ where '$\bullet$' stands for a scalar.

Let us summarize.

*Proposition 3 (recursive MAP estimate of $\hat{l}_\Theta(t)$)*

For the recursive exponential family, the MAP estimate $\hat{l}_\Theta(t)$ of the unknown likelihood $l_\Theta(t)$ has the form (19), where $z(t)$ is the eigenvector corresponding to the largest eigenvalue of the symmetric positive semidefinite $(i_{\mathcal{B}}, i_{\mathcal{B}})$ matrix $W(t)$ in (18) with the bounded dimension (17). The matrix $\mathcal{B}(t)$ determining $W(t)$ can be evaluated recursively by the orthogonal transformation $\mathcal{T}(t+1)$ for which

$$\mathcal{T}(t+1)\begin{bmatrix}\mathcal{B}(t)\\ \psi_{t+1}^{\mathrm{T}}\end{bmatrix}=\begin{bmatrix}\mathcal{B}(t+1)\\ 0\end{bmatrix}$$

The length of the vector $z(t)$ is

$$\alpha(t)=\frac{i_{\mathcal{B}(t)}}{\bar{z}^{\mathrm{T}}(t)b(t)}$$

The transformation $\mathcal{T}(t+1)$ defines recursion for the vector $b(t)$:

$$\mathcal{T}(t+1)\begin{bmatrix}b(t)\\ 1\end{bmatrix}=\begin{bmatrix}b(t+1)\\ \bullet\end{bmatrix}$$

where '●' marks an unimportant scalar.

*Proof.* This is implied by Proposition 2 and the above algebra. □

Notice that the (pre)prior PDF $p(\cdot)$ and the form of $\psi$ and $\tilde{\Theta}$ are the case-specific options which determine the case-specific evaluation of the matrix $\mathcal{A}$.

## 6. APPLICATION TO ARX MODEL

The main *practical contribution* contained in this section results from a straightforward application of the previous theory. The results are directly and widely applicable and serve as an illustration of the general approach.

### 6.1. *ARX model*

Without loss of generality[12] the single-output case is considered. The Gaussian ARX model is specified by the PDF

$$g_{[\theta,r];\tau}=\mathcal{N}_{y_\tau}(\theta^{\mathrm{T}}\phi_\tau,r)\equiv(2\pi r)^{-0\cdot5}\exp[-0\cdot5r^{-1}([-1,\theta^{\mathrm{T}}]\Phi_\tau)^2] \tag{21}$$

where $\theta$ is the $i_\phi$-vector of unknown real *regression coefficients*; $\phi_\tau$ is the *regressor*, a known function of $D(\tau-1)$, $u_\tau$; $\Phi_\tau=[y_\tau,\phi_\tau]^{\mathrm{T}}$ is the corresponding *data vector* with dimension $i_\Phi=i_\phi+1$; $r$ is an unknown *dispersion*; and $\Theta=[\theta,r]$ is the unknown parameter defining the input–output model of the considered dynamic uncertain system.

This popular model is positive on a data-independent support and its logarithm is spanned over a finite fixed number of functions of $\Theta$. For a wide set of (pre)prior PDFs it belongs to the recursive exponential family.

### 6.2. *Canonical form of ARX model*

Proposition 3 is formulated in terms of the canonical form. The ARX model is transformed here to it for a specific choice of the (pre)prior PDF $p(\Theta)$.

The (pre)prior PDF $p(\Theta)$ is expected to be flat and its analytical form should not influence the results significantly. For computational reasons a self-reproducing Gauss–inverse–Wishart

(GiW) PDF[13] is selected

$$p(\theta, r) = \mathrm{GiW}_{\theta,r}(\mathscr{L}^{\mathrm{T}}\mathscr{L}, v) \propto r^{-0\cdot5(v+i_{\phi}+2)} \exp\left(-\frac{[-1, \theta^{\mathrm{T}}]\mathscr{L}^{\mathrm{T}}\mathscr{L}[-1, \theta^{\mathrm{T}}]^{\mathrm{T}}}{2r}\right) \tag{22}$$

The normalizing constant in (22) is finite, i.e. the PDF $p(\theta, r)$ is well-defined, if the $(i_{\Phi}, i_{\Phi})$ matrix $\mathscr{L}$ is regular and the scalar $v$ positive. The matrix $\mathscr{L}$ can always be and will be selected as lower triangular with a positive diagonal (Cholesky factorization):

$$\mathscr{L} = \begin{bmatrix} \mathscr{L}_y & 0 \\ \mathscr{L}_{y\phi} & \mathscr{L}_{\phi} \end{bmatrix}$$

$\mathscr{L}_y$ is a positive scalar in this decomposition.

The unknown quantity $X$ is transformed to the canonical form

$$X^{\mathrm{T}} \equiv [-\rho, x^{\mathrm{T}}] \equiv [-r^{-0\cdot5}, r^{-0\cdot5}\theta^{\mathrm{T}}]\mathscr{L}^{\mathrm{T}}$$

i.e. $\rho$ is a positive scalar and $x$ is an $i_{\phi}$ vector. For this variable

$$P(X) = \frac{\rho^{v-1}}{2^{(v-2)/2}\Gamma(v/2)} \exp\left(-\frac{q^2}{2}\right)(2\pi)^{-i_{\phi}/2} \exp\left(-\frac{x^{\mathrm{T}}x}{2}\right) = \mathscr{Q}_{\rho}\left(-\frac{v}{2}\right)\mathscr{N}_x(0, I) \tag{23}$$

The function $\mathscr{Q}_{\rho}(v/2)$, denoting the factor depending on $\rho > 0$, could be converted into the gamma PDF with $v/2 > 0$ degrees of freedom by the additional substitution $0\cdot5\rho^2 \rightarrow \tilde{\rho}$.

The ARX model (21) parametrized by the transformed variables $X$, becomes

$$g_{X;\tau} = \rho \exp[-0\cdot5([\rho, x^{\mathrm{T}}]\tilde{\bar{\Phi}}_{\tau})^2] = \exp\{-0\cdot5[-2\ln(\rho) + ([\rho, x^{\mathrm{T}}]\tilde{\bar{\Phi}}_{\tau})^2]\}$$

where

$$\bar{\varepsilon}_{\tau} = -\mathscr{L}_y^{-1}[y_{\tau} + (\mathscr{L}_f^{-1}\mathscr{L}_{y\phi})^{\mathrm{T}}\phi_{\tau}], \qquad \tilde{\phi}_{\tau} = (\mathscr{L}_{\phi}^{\mathrm{T}})^{-1}\phi_{\tau}$$

The canonical form is finally obtained by mapping $\tilde{\bar{\Phi}}^{\mathrm{T}} = [\bar{\varepsilon}, \bar{\phi}^{\mathrm{T}}]$ to the $i_{\psi}$-vector

$$\psi^{\mathrm{T}} = [1, \tilde{\bar{\Phi}}_1^2, \ldots, \tilde{\bar{\Phi}}_{i\Phi}^2, \tilde{\bar{\Phi}}_1\tilde{\phi}^{\mathrm{T}}, \tilde{\phi}_1[\tilde{\phi}_2, \ldots, \tilde{\phi}_{i_{\phi}}], \ldots, \tilde{\phi}_{i_{\phi}-1}\tilde{\phi}_{i_{\phi}}] \tag{24}$$

with $i_{\psi} = 1 + 0\cdot5 i_{\Phi}(i_{\Phi} + 1)$.

Similarly, an $i_{\psi}$-vector $\tilde{\Theta}$ is assigned to the transformed unknown parameter $X^{\mathrm{T}} = (\rho, x^{\mathrm{T}})$:

$$\tilde{\Theta}^{\mathrm{T}} = [-2\ln(\rho), \rho^2, x_1^2, \ldots, x_{i_{\phi}}^2, 2\rho x^{\mathrm{T}}, 2x_1[x_2, \ldots, x_{i_{\phi}}], \ldots, 2x_{i_{\phi}-1}x_{i_{\phi}}] \tag{25}$$

With this notation, the logarithm of the local model (21) takes the form (11).

### 6.3. *Evaluation of mixing quantities*

The matrix $\mathscr{A}$ is the second moment of the vector $\tilde{\Theta}$ in (25) with respect to the (pre)prior PDF (23). It can be expressed in terms of logarithmic derivatives $\mathscr{S}(a) \equiv \partial \ln[\Gamma(a)]/\partial a$ of the Euler gamma function $\Gamma(a)$, $a > 0$. Its approximate value can be based on the formula[14]

$$\mathscr{S}(a) \equiv \frac{\partial \ln(\Gamma(a))}{\partial a} \approx \ln(a) - \frac{1}{2a} - \frac{1}{12a^2} + \frac{1}{120a^4}$$

It holds that

$$\mathscr{A} = \begin{bmatrix} \mu_2 & \mu_3 & \mu_1 \mathbb{1}^{\mathrm{T}} & 0 & 0 \\ \mu_3 & \mu_4 & \mu_5 \mathbb{1}^{\mathrm{T}} & 0 & 0 \\ \mu_1 \mathbb{1} & \mu_5 \mathbb{1} & \mu_6 I + \mathbb{1}\mathbb{1}^{\mathrm{T}} & 0 & 0 \\ 0 & 0 & 0 & 4\mu_5 I & 0 \\ 0 & 0 & 0 & 0 & 4I \end{bmatrix}$$

where $\mathbb{1}$ is now and $i_\Theta$ dimensional vector of units. The unit matrices at $\mu_6$ and $\mu_5$ are of the $i_\Theta$-type and the unit matrix with the coefficient '4' completes the dimension of $\mathscr{A}$ to the square $i_\psi$-type. The scalars involved are defined by

$$\mu_1 \equiv -2\mathscr{E}_p[\ln(\rho)] = -\ln(2) - \mathscr{S}(0{\cdot}5v)$$

$$\mu_2 \equiv 4\mathscr{E}_p[\ln^2(\rho)] = \mu_1^2 + \frac{\partial \mathscr{S}(0{\cdot}5v)}{\partial(0{\cdot}5v)}$$

$$\mu_3 \equiv -2\mathscr{E}_p[\rho^2 \ln(\rho)] = -(v+2)(\ln(2) + \mathscr{S}(0{\cdot}5v+1))$$

$$\mu_4 \equiv \mathscr{E}_p[\rho^4] = (v+4)(v+2)$$

$$\mu_5 \equiv \mathscr{E}_p[\rho^4] = v+2$$

$$\mu_6 \equiv \int_{-\infty}^{\infty} h^4 \mathscr{N}_h(0,1)\, dh - 1 = 2$$

The structure of $\mathscr{A}$ leads to the following lower triangular square root $A$ of $\mathscr{A} = AA^{\mathrm{T}}$ needed in the implementation:

$$A = \begin{bmatrix} M & & \\ 0 & 2\surd(\mu_5) I & \\ 0 & 0 & 2I \end{bmatrix} \tag{26}$$

where the $(i_\phi, i_\phi)$ matrix $M$ is a triangular Choleski square root of the top-left submatrix of $\mathscr{A}$:

$$MM^{\mathrm{T}} = \begin{bmatrix} 1 & \mu_1 & \mu_5 & \mathbb{1}^{\mathrm{T}} \\ \mu_1 & \mu_2 & \mu_3 & \mu_1 \\ \mu_5 & \mu_3 & \mu_4 & \mu_5 \mathbb{1}^{\mathrm{T}} \\ \mathbb{1} & \mu_1 \mathbb{1} & \mu_5 \mathbb{1} & \mu_6 I + \mathbb{1}\mathbb{1}^{\mathrm{T}} \end{bmatrix} \tag{27}$$

This square root can be precomputed and the matrix $W(t)$ in (18) expressed as

$$W(t) = B(t)B^{\mathrm{T}}(t), \quad \text{with } B(t) = \mathscr{B}(t)A$$

A standard QR algorithm[10] can be used for updating the matrix $\mathscr{B}(t)$. As the vectors $\psi_\tau$ have the first entry equal to unity, the vector $b(t)$ needed in evaluating the length $\alpha$ of the mixing vector in (10) coincides with the first column of $\mathscr{B}(t)$.

There are many efficient procedures for determining the maximum eigenvector of a positive semidefinite (mixing) matrix. Iterative versions starting from the mixing vector obtained in the previous time step seem to be preferable.

It is worth noticing that the mixing matrix often has positive entries (this is always true when the data $D(t)$ are discrete). Then the Perron theory[15] implies that the mixing vector has non-negative entries.

### 6.4. *Final form of estimate* $\hat{f}(\Theta \mid D(t))$

With the mixing vector $z(t)$ and the matrix $\mathcal{B}(t)$, the data-dependent vector in the exponent of the likelihood estimate (19) is just the product

$$H^{\mathrm{T}}(t) \equiv z^{\mathrm{T}}(t)\mathcal{B}(t) \equiv [\beta(t), H_{11}(t), \ldots, H_{i_\Phi i_\Phi}, H_{12}, \ldots, H_{1 i_\Phi}(t), \ldots, H_{i_\Phi - 1 i_\Phi}(t)]$$

The names given to the particular entries help us to write the estimate in a final compact form.

Defining the $(i_\Phi, i_\Phi)$ matrix $\mathcal{H}(t)$ with entries $\mathcal{H}_{ij} = \mathcal{H}_{ji} = H_{ij}(t)$ for $i \leqslant j$, the estimate $\hat{f}(X \mid D(t))$ of the posterior PDF $f(X \mid D(t))$ becomes

$$\hat{f}(X \mid D(t)) \propto X_1^{\upsilon - 1 + \beta(t)} \exp[-0{\cdot}5(X^{\mathrm{T}} V_H(t) X)] \quad \text{with } V_H(t) \equiv \mathcal{H}(t) + I \tag{28}$$

By reverting back from $X = [\rho, x^{\mathrm{T}}]^{\mathrm{T}}$ to the original variables $(\theta, r)$, the final estimate of the posterior PDF is found to have the GiW form

$$\hat{f}(r, \theta / D(t)) = \mathrm{GiW}_{r,\theta}(\mathcal{L}^{\mathrm{T}} V_H(t) \mathcal{L}, \upsilon + \beta(t)) \tag{29}$$

## 7. ALGORITHMIC SUMMARY

*Off-line phase*

1. *Select (pre)prior statistics* $\mathcal{L}, \upsilon$, respecting moments of (22), $\hat{r} \equiv \mathcal{E}_p[r] = \mathcal{L}_y / \upsilon$, $\mathrm{cov}(\theta) = \hat{r} \mathcal{L}_\phi^{-1} (\mathcal{L}_\phi^{-1})^{\mathrm{T}}$ (see Section 8).
2. *Compute the matrix* $A$ according to (26) and (27).
3. *Set the initial condition* $\mathcal{B}(0) = 0$, $t = 0$ in (16) and fill $\phi(0)$ in (21).

*On-line phase*

4. *Set* $t = t + 1$.
5. *Measure* new data $u_t$, $y_t$ and construct the vector $\psi_t$ in (24).
6. *Update* the matrix $\mathcal{B}(t-1)$ to $\mathcal{B}(t)$ by the 'orthogonal' zeroing of $\psi_t^{\mathrm{T}}$ (20).
7. *Compute* the mixing vector $\nu(t)$ of the matrix $\mathcal{B}(t) A A^{\mathrm{T}} \mathcal{B}(t)$, including its length $\alpha = i_{\mathcal{B}(t)} / (\nu^{\mathrm{T}}(t) b(t))$ where $b(t)$ is the first column of $B(t)$.
8. *Evaluate* the statistic $H(t)$ determining the constructed estimate, i.e. the scalar $\beta(t)$ and the matrix $V_H$ in (28).
9. *Evaluate* desired characteristics of the approximate PDF (29) and go to step 4.

## 8. ILLUSTRATIVE EXAMPLES

The original incentive for this research was provided by attempts to ensure the automatic commission of adaptive controllers. Effective use of prior information proved to be very useful in this respect. Even such simple and often available information as the approximate static gain and/or the supposed (not precise) form of the step response can improve the start of the adaptive control.

This section illustrates the application of the proposed procedure as well as its contribution to the key tasks met in adaptive prediction/control design.

### 8.1. *Coding of prior information*

Variety of information sources induces variety of forms in which the information is provided (opinions, pictures, analytic expressions, data file, simulation model). For efficient processing a unified coding is almost a necessity. The use of so-called 'fictitious data' proved to be a suitable tool for handling the problem. This technique is an integral part of the identification area folklore and fits in with the presented algorithm. Essentially the information source is asked to provide the information in the form of fictitious data. The reply should have the following form.

> If the system is stimulated with the input sequence $U_f(t_f)$ the output sequence is expected in the range $[\underline{Y}_f(t_f), \bar{Y}_f(t_f)]$ with a high probability.

The scaled data sequence

$$y_\tau = \frac{\mathscr{L}_y}{v}\,\frac{\bar{y}_{\mathrm{f}\tau} + \underline{y}_{\mathrm{f}\tau}}{\bar{y}_{\mathrm{f}\tau} - \underline{y}_{\mathrm{f}\tau}}, \qquad u_\tau = \frac{\mathscr{L}_y}{v}\,\frac{2u_{\mathrm{f}\tau}}{\bar{y}_\tau - \underline{y}_\tau}, \qquad \tau \in \tau_{\mathrm{f}}^{*} = \{1, \dots, t_{\mathrm{f}}\}$$

can then be interpreted as the data sequence measured on the inspected linear ARX model. The 'fictitious noise' dispersion is equal to the expected value $\hat{r}$ of the model noise dispersion $r$. Use of such information reduces to the estimation with these data. The estimation has to be conservative as the fictitious noise is not white.

The required form of the fictitious data is believed to be 'natural' for process experts and sometimes may result from analytic/simulation sources too.

If just a single input–output trajectory is available, then scaling of inputs and outputs is the only possibility for expressing its precision. The guessed range of the innovations in the scaled fictitious data has to be close to the output dispersion predicted by the (pre)prior PDF, namely to $\hat{r}(1 + \| \phi_\tau^{\mathrm{T}} \mathscr{L}^{-1} \|^2)$. This rough rule is mostly sufficient, especially for the usual case of diagonal (pre)prior (!) $\mathscr{L}$.

### 8.2. *Tasks supported by theory*

In the computer-aided preliminary design of demanding adaptive systems, any piece of prior information available should be used for possible performance improvement. The decisive first design step, the *system structure estimation*,[16] gives more adequate results when prior information is respected.

The quality of the initial learning period of the adaptive predictor/controller is substantially influenced by its initial state — in the case of the ARX model by the recursive *least squares* (RLS) initialization. The appropriate values are just statistics of the GiW PDF (29) obtained after merging all information pieces.

Objects which have to be modelled with varying parameters (because of the system nature or because of the approximate modelling) represent the main application area of adaptive systems. They can rarely work without excitation-robust forgetting. A very general solution which fits our probabilistic language is proposed in Reference 8. It needs an *alternative PDF of the unknown parameter* $\Theta$ describing *a priori* possible parameter uncertainty. This is directly provided by the advocated algorithm.

### 8.3. *Experimental conditions*

The influence of prior information in the listed tasks is illustrated on simple simulation examples. A single-output/single-input continuous system with the transfer function $1/(1 + s^2)$ sampled with a period of 0·1 s is considered. The corresponding discrete-time simulation system is

$$y_t = 1{\cdot}81 y_{\tau-1} - 0{\cdot}8187 y_{\tau-2} + 0{\cdot}00468 u_\tau + 0.00438 u_{\tau-1} + \sigma e_\tau \tag{30}$$

The sequence $e$ is white normal zero-mean noise with unit dispersion. In open loop experiments the same type of independent noise serves as input. The noise standard deviation $\sigma$ varies in particular cases.

Notice the small values of coefficients at input. It is non-trivial to recognize their significance from noisy data.

In the on-line phase when the adaptive system (predictor or controller) is in action, a simplified system model of the first order ARX model is estimated as

$$g_{a,b_1,b_2,k,r;\tau} = \mathcal{N}_{y_\tau}(a y_{\tau-1} + b_0 u_\tau + b_1 u_{\tau-1} + k, r) \tag{31}$$

with the unknown parameter $\Theta \equiv [\theta^{\mathrm{T}}, r] \equiv [a, b_0, b_1, k, r]$.

The following types of prior information pieces are used:

(i) An approximate static *gain* $\approx 1$ expressed by input–output data equal to a constant $S$, processed as a single $\psi$ (note that in accordance with the developed theory an arbitrary repetition leads to identical results).

(ii) An approximate *step* response expressed by the fictitious data $(u_{\mathrm{f}}, y_{\mathrm{f}})$.

$$y_{\mathrm{f},\tau} = 0{\cdot}957 y_{\mathrm{f},\tau-1} + 0{\cdot}023 u_{\mathrm{f},\tau} + 0{\cdot}024 u_{\mathrm{f},\tau-1}$$
$$u_{\mathrm{f},\tau} = y_{\mathrm{f},\tau} = 0 \quad \text{for } \tau \leqslant 0, \qquad u_{\mathrm{f},\tau} = 1 \quad \text{for } \tau \geqslant 1$$

which are amplified by a factor $S$.

(iii) A short *data* sample (of length 20) generated by the 'true' system.

The statistics determining the (pre)prior PDF are $\mathcal{L} = 0{\cdot}1I$, $v = 1$. It means that the magnitudes of the highly expected regression coefficients are of zero order. The amplitude of the white noise is of the order $-1$. The used scalings of data $S = 10$ or $S = 1$ imply that the precision assigned to information sources is of the order $-2$ or $-1$, respectively.

The information is processed by the proposed algorithm (Section 7), recorded in files and used in the experiments.

The particular experiments vary significantly with realizations of the system noise. To suppress this influence, the experiments are repeated 50 times with different noise realizations and results are appropriately summarized.

### 8.4. Prior information in structure estimation

Given the 'true' data sample, the maximum *a posteriori* probability estimate[16] of the best model structure is found in a sufficiently broad space of possible models.

All the highly probable structures contained the second order autoregression $\phi_\tau = [y_{\tau-1}, y_{\tau-2}, \ldots]$. The results of 100 runs with different extents of prior information (very *vague*, static *gain*, *step* response, real *data* sample) and with different system noise levels $\sigma$ are recorded in Table I. Whenever an input is recognized as significant, the result is counted as successful. The scaling factor is $S = 10$.

In order to show that the proposed algorithm processes data appropriately, prior information is also generated from the *data* sample by the standard *RLS* algorithm.

The results are almost self-explanatory. The true model structure is easy to find when the system noise level ($\sigma$) is below $0{\cdot}02$ (Table I, row 1). When the standard deviation of the system noise exceeds $0{\cdot}06$, it is almost impossible to find the true structure. Within this range

Table I. Number of successes in structure estimation

| | Prior information | | | | |
|---|---|---|---|---|---|
| Noise $\sigma$ | Vague | Gain | Step | Data | RLS |
| 0·02 | 97 | 98 | 99 | 96 | 96 |
| 0·03 | 66 | 79 | 88 | 69 | 57 |
| 0·04 | 30 | 48 | 66 | 36 | 24 |
| 0·06 | 6 | 8 | 31 | 10 | 4 |

the use of prior information has substantial positive influence. The best results are obtained for the *step* response information, then the *gain* information and the *data* sample. The *RLS* processing has only a minor influence.

### 8.5. *Initialization of adaptive predictors/controllers*

This subsection demonstrates the positive influence of incorporating prior information in the initial phase of recursive estimation and control.

The system (30) with the dispersion $\sigma = 0{\cdot}03$, the model (31) and the scaling factor $S = 1$ are considered.

The open loop results are quantified by the (average) absolute prediction error in Figure 1. The influence of the *gain* information is quite similar to that of the *step* response information, so only the former case (broken curve) is presented. The full curve corresponds to the *vague* prior information and the chain curve to the *data* sample.

The better identification gives chance for an improved start of an adaptive controller based on it. To demonstrate this, the open loop is closed by a standard receding horizon adaptive

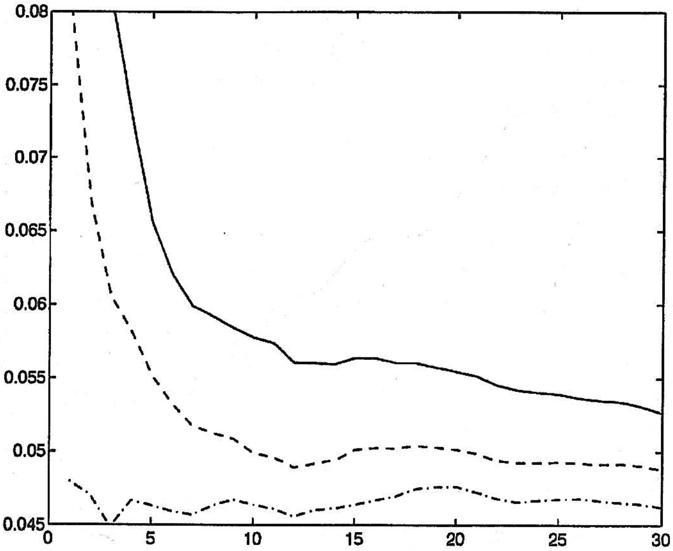

Figure 1. Prediction error

linear–quadratic–Gaussian (LQG) controller based on the model (31). The average absolute control error, evaluated for a step change in setpoint, is depicted in Figure 2 (the notation of the curves is the same as in the open loop case). The non-linear nature of the controller makes the influence of the prior information on the control quality non-monotonic. After a few steps, however, the positive influence of the prior information is clearly demonstrated.

### 8.6. *Prior information in stabilized forgetting*

Forgetting prevents the influence of obsolete information and thus makes adaptation possible. In connection with recursive least squares, exponential forgetting is the simplest and very widespread (RLS-EF), but it suffers from the well-known lack of robustness with respect to data informativity. This danger can be removed by combining the current PDF f(Θ) with a prespecified alternative $f_a(\Theta)$.[17] For ARX models (GiW distributions) this technique reduces to linear additive stabilization of the information matrix as proposed by Kraus and analysed in Reference 18. It is called the stabilized recursive least squares with invariant alternative (RLS-SI). In this method, mixing of the information matrix with some alternative takes place; as no closer specification of the alternative is given, a diagonal matrix is usually supposed. The mathematical analysis presented is valid for the full matrix alternative too, but hints are missing for its appropriate choice. The tested algorithm provides such an alternative. As shown in the experiments, the proper choice of an adequate matrix alternative covariance (using the available prior knowledge) can bring performance improvement.

The following experiment gives an example of the use of prior information (*step* response knowledge) in RLS-SI identification algorithms with the full matrix alternative (full curves). The results are compared with the usual exponential forgetting (RLS-EF, broken curves) and also with the RLS-SI with the simple diagonal alternative covariance constant *I* (dotted curves). The used constant predicts time variations in the range of percentages of the prior uncertainty range.

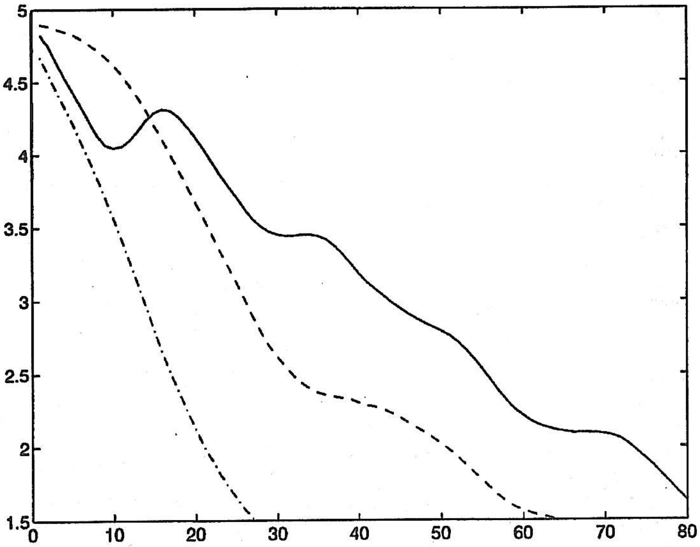

Figure 2. Output reference

The system (30) with $\sigma = 0{\cdot}06$ is identified as the first order ARX model (31). The influence of undermodelling is amplified by using white noise input with standard deviation equal to three up to time 50. Then input changes are switched off in order to test the behaviour of the estimation under strong non-informativity conditions.

Figure 3 shows the average results of repeated experiments. The prediction of RLS-EF is almost identical with that of RLS-SI with the full matrix alternative and both are visibly better than that provided by RLS-SI with the diagonal alternative.

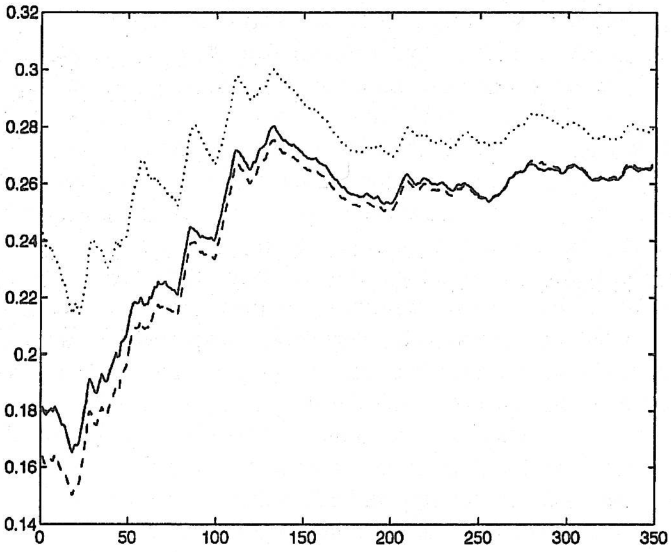

Figure 3. Prediction error

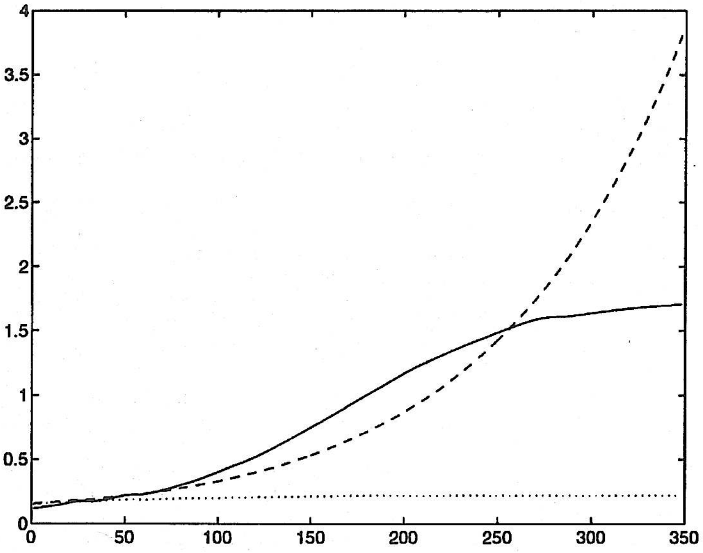

Figure 4. Covariance norm

These results should be compared with the covariance behaviour displayed in Figure 4. For RLS-EF the covariance wind-up is visible, RLS-SI has the expected stabilizing effect. The overdamping by the selected diagonal is paid for by the deterioration of the prediction.

## 9. CONCLUSIONS

An *automatic* recursive procedure has been proposed for creating an approximate description of a stochastic dynamic system that uses uncertain, not fully compatible and repetitive pieces of information originating from (i) expert knowledge, (ii) preliminary data and (iii) a complex analytic/simulation model. The procedure can merge an increasing amount of evidence and provides a reasonably conservative probabilistic quantification of the system description. It has been designed and tested for widely applicable ARX models. In this way the typical representation method of black box models comes closer to the more desirable grey box models while preserving simple applicability in control/prediction tasks.

At the practical level the algorithm allows the incorporation of prior knowledge into structure estimation as well as into the initial conditions of recursive least squares. It serves for construction of a reference for an advanced forgetting technique. These applications have been demonstrated in the paper but the procedure may serve in other areas as well. It may be used as an alternative forgetting technique if increased evaluation complexity is allowed. When treating data from a complex source, the algorithm generates the approximant of the system as a byproduct. It serves as a construction tool for merging simple adaptive systems into a more complex one. It offers a lot but (as yet?) lacks deeper analysis.

## ACKNOWLEDGEMENTS

This work has been partially supported by GA ČR, Grant 102/93/0228 and by GA AV ČR, Grants 275109 and 275110.

## REFERENCES

1. Kárný, M., A. Halousková and L. Zörnigová, 'On pooling expert opinions', in Blanke, M. and T. Söderström (eds), *Prepr. SYSID'94*, Vol. 2, Danish Automation Society, Copenhagen, 1994, pp. 477–482.
2. Kárný, M., D. Hajmán and A. Halousková, 'On conservative recursive learning', in L. Kulhavá, M. Kárný and K. Warwick (eds), *Prepr. Eur. IEEE Workshop CMP'94*, ÚTIA AV ČR, Prague, 1994, pp. 77–82.
3. Perez, A., and R. Jiroušek, 'Constructing an intensional expert system (INES)', in *Medical Decision Making*, Elsevier, Amsterdam, 1986, pp. 307–313.
4. Jiroušek, R., 'Solution of the marginal problem and decomposable distributions', *Kybernetika*, **27**, 403–412 (1991).
5. Kulhavý, R., 'Bayesian estimation, large deviations, and incomplete data', *Prepr. IEEE Conf. on Decision and Control*, **1**, 755–756. Orlando, FL, IEEE, New York, (1994).
6. Sanov, I. N., 'On probability of large deviations of random variables', *Mat. Sborn.*, **42**, *11–44 (1957) (in Russian); Engl. transl. Select. Transl. Math. Stat. Probab.*, **I**, 213–244 (1961).
7. Ninness, B. M., and G. C. Goodwin, 'Rapprochment between bounded-error and stochastic estimation theory', *Int. j. adapt. control signal process.*, *9*, 107–132 (1995).
8. Kulhavý, R., 'Restricted exponential foregetting in real-time identification', *Automatica*, **23**, 589–600 (1987).
9. Peterka, V., 'Bayesian system identification', in Eykhoff, P. (ed), *Trends and Progress in System Identification*, Pergamon, Oxford, 1981, pp. 239–304.
10. Golub, G. H., and C. F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 1989.
11. Kulhavý, R., I. Nagy and J. Spousta, 'Towards real-time implementation of Bayesian parameter estimation', *Prepr. 4th IFAC Symp. ACASP'92*, Vol. 2, Academic, Grenoble, 1992, pp. 125–247.
12. Kárný, M., 'Parametrization of multi-output multi-input autoregressive–regressive models for self-tuning control', *Kybernetika*, **28**, 402–412 (1992).
13. Anderson, T. W. *An Introduction to Multivariate Statistical Analysis*, Wiley, New York, 1958.

14. Abramowitz, M., and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1972.
15. Fiedler, M., *Special Matrices and Their Use in Numerical Mathematics*, SNT, Prague, 1981.
16. Kárný, M., and R. Kulhavý, 'Structure determination of regression-type models for adaptive prediction and control', in Spall, J. C. (ed.), *Bayesian Analysis of Time Series and Dynamic Models*, Marcel Dekker, New York, 1988, Chap. 12.
17. Kulhavý, R., and M. B. Zarrop, 'On general concept of forgetting', *Int. J. Control*, **58**, 905–924 (1993).
18. Milek, J. J., and F. J. Kraus, 'Stabilized least squares estimators', *Technical Report 91-02*, Automatic Control Laboratory, ETH Zürich, 1991.