Use of Prior Information in Structure Estimation

Miroslav Kárný, Petr Nedoma, Natalia Khaylova, Lenka Pavelková
Institute of Information Theory and Automation
Pod vodárenskou věží 4, 182 08 Prague 8, POB 18, Czech Republic
school@utia.cas.cz tel. +420-2-66052274,
fax +420-2-66052068

Key words: adaptive control, Bayesian estimation, prior information, ARX model

1 Introduction

Adaptive LQG controllers that minimize approximately quadratic criterion while using recursively estimated linear Gaussian models become a standard in academic environment [1, 2, 3]. The version with controlled autoregressive models (ARX) belongs among the most successful ones as confirmed by their full scale applications [4, 5, 6]. At the same time, their potential is far from being adequately exploited. Man-power expensive commission is one of the main reasons of this undesirable state. This fact – that applies also to other adaptive controllers like GPC, MUSMAR etc. [3, 7] – stimulated an extensive project [8, 9, 10]. It aims at creating a complete computer support of the commission. At present, a full solution for single-input single-outputs systems is implemented in software system DESIGNER. It covers (i) data-preprocessing; (ii) quantification of prior information [**?**]; (iii) selection of the model structure [11]; (iv) off-line estimation [12] that serves for initialization of the on-line estimation as well as an alternative in the stabilized forgetting used for parameter tracking [13]; (v) estimation of the forgetting factor; (vi) tuning of kernels in the optimized quadratic loss so that user's aims and restrictions are met; this also provides off-line prediction of closed-loop behavior [14].

This paper proposes a correct combination of steps (ii) and (iii). The summary of steps (ii) - (iv) serves also to the companion paper [15], that provides a novel solution of the most difficult step (vi). Thus, in addition to particular improvements, the paper pair informs on the "technology used" within the DESIGNER.

Review of necessary results is given in Section 2 leading to the addressed problem formulation and its conceptual solution. Mapping of processed

knowledge on so called fictitious data is contained in Section 3. The results are elaborated to ARX model in Section 4, illustrated in Section 5 and complemented by remarks of Section 6.

The paper is relatively self-containing. The readers less familiar with the adopted Bayesian set up are referred to [12].

2 Problem formulation and solution

Here, the essence of the addressed problem and its conceptual solution are described after technical preliminaries. The used ARX model belongs to the exponential family. Its formal handling is very simple when its member-specific details are left aside. This motivates us to present all elements within it.

2.1 Basic notation and operations

The following notation is used throughout the text.

Symbol	Meaning
$\equiv$	equality by definition
x^*	a set of x-values
$\mathring{x}$	the number of elements in the vector or sequence x
$f(\cdot\|\cdot)$	conditional probability (density) functions (p(d)f)
$\mathcal{E}$, var, cov	expectation, variance and covariance, respectively
$d(t)$	the sequence $(d_1, \ldots, d_t)$
$\mathcal{S}$, $\mathcal{K}$	model structure and knowledge item, respectively
t	discrete-time, always the last subscript after $_;$
$\tau_{\mathcal{K}}$	discrete time of fictitious data expressing knowledge $\mathcal{K}$
i	the subscript of the entry $d_{i;t}$ of a data item d_t
I, tr, $'$	unit matrix, trace and transposition, respectively
$\lfloor^{\psi}L$	a non-numerical index ψ of a variable L
V, ν	statistics describing conjugate pdf to exponential family
$\bar{}$	bar distinguishing (flat) pre-prior pdf
$\lfloor^{\Delta}$	mark of increments of statistics obtained from data only

Note that pdfs are distinguished by the identifiers in their arguments. No formal distinction is made between random variable, its realization and argument of pdfs. The correct meaning follows from the context.

The following elementary operations of non-negative pdfs are used [12]

$$\begin{aligned}
\text{Normalization} \quad & \int f(a)\,da \equiv \int_{a\in a^*} f(a)\,da = 1, \qquad (1)\\
\text{Chain rule} \quad & f(a,b|c) = f(a|b,c)f(b|c),\\
\text{Marginalization} \quad & f(a|c) = \int f(a,b|c)\,db,\\
\text{Bayes rule} \quad & f(a|b,c) = \frac{f(b|a,c)f(a|c)}{\int f(b|a,c)f(a|c)\,da} \propto f(b|a,c)f(a|c).
\end{aligned}$$

Bayesian paradigm we exploit operates on the joint pdf of all uncertain variables encountered. It composes this pdf from its elements and derives its particular marginal or conditional versions using (1). It inserts in them the measured realization of any variable that is at disposal.

Here, sequence of multi-variate data $d(\mathring{t}) = (d_1, \ldots, d_{\mathring{t}})$, unknown, finite-dimensional parameters $\Theta_{\mathcal{S}}$ and unknown structures $\mathcal{S}$ of an adequate model are considered. The joint pdf is composed as follows

$$\begin{aligned}
&\overbrace{\text{joint pdf}}\ \overbrace{\text{prior pdf}\,|\,\mathcal{S}}\ \overbrace{\text{prior on } \mathcal{S}}\ \overbrace{\qquad\qquad\qquad\qquad\text{parameterized model}\qquad\qquad\qquad\qquad}\\
&f(d(\mathring{t}),\Theta_{\mathcal{S}},\mathcal{S}) = f(\Theta_{\mathcal{S}}|\mathcal{S})\ \times f(\mathcal{S}) \times \prod_{t=1}^{\mathring{t}}\prod_{i=1}^{\mathring{d}} f(d_{i;t}|d_{i+1;t},\ldots,d_{\mathring{d};t},d(t-1),\Theta_{i\mathcal{S}})\\
&\Theta_{\mathcal{S}} \equiv (\Theta_{1\mathcal{S}},\ldots,\Theta_{\mathring{d}\mathcal{S}}). \qquad (2)
\end{aligned}$$

2.2 Estimation and prediction in exponential family

The parameterized model in (2) of a fixed structure $\mathcal{S}$ is the central modelling element. Within the control context, when the amount of observed data is permanently increasing, the following models are predominantly used.

Agreement 2.1 (Exponential family) *The i-th parameterized model belongs to the* (dynamic) exponential family *iff it can be written in the form*

$$\begin{aligned}
f(d_{i;t}|d_{i+1;t},\ldots,d_{\mathring{d};t},d(t-1),\Theta_{i\mathcal{S}}) &= f(d_{i;t}|\psi_{i\mathcal{S};t},\Theta_{i\mathcal{S}}) \qquad (3)\\
&= A(\Theta_{i\mathcal{S}})\exp\left[\langle B(\Psi_{i\mathcal{S};t}),C(\Theta_{i\mathcal{S}})\rangle\right],
\end{aligned}$$

where $\Psi'_{i\mathcal{S};t} = [d_{i;t},\psi'_{i\mathcal{S};t}]$ *is* data vector*, given by a finite dimensional* regression vector $\psi_{i\mathcal{S};t}$*, depending on* $d_{i+1;t},\ldots,d_{\mathring{d};t}$ *and on* $d(t-1)$*; it is assumed that the values of all data vectors* $\Psi_{i\mathcal{S};t-1}$*,* $i = 1,\ldots,\mathring{d}$*, can be recursively updated using the newest data item* d_t *only,*
$A(\cdot)$ *is a non-negative scalar function defined on* $\Theta^*_{i\mathcal{S}}$*,*

$\langle\cdot,\cdot\rangle$ *is a functional that is linear in the first argument,*
$B(\cdot)$, $C(\cdot)$ *are either vector or matrix functions of compatible, finite dimensions. They are defined on* $\Psi^*_{i\mathcal{S};t}$ *and* $\Theta^*_{i\mathcal{S}}$*, respectively.*

Practical significance of the exponential family becomes obvious when summarizing the corresponding estimation and prediction, [12].

Proposition 2.1 (Estimation and prediction in exponential family) *Let the parameterized model have the form (3) and the parameters* $\Theta_{\mathcal{S}} = \Theta(\mathring{\mathcal{S}})$ *be a priori independent, i.e.* $f(\Theta_{\mathcal{S}}) = \prod_{i=1}^{\mathring{d}} f(\Theta_{i\mathcal{S}})$*. Let, moreover, the* conjugate prior *pdfs* $f(\Theta_{i\mathcal{S}})$ [16]

$$f(\Theta_{i\mathcal{S}}) \propto A^{\nu_{i\mathcal{S};0}}(\Theta_{i\mathcal{S}}) \exp[\langle V_{i\mathcal{S};0}, C(\Theta_{i\mathcal{S}})\rangle] \chi_{\Theta^*_{i\mathcal{S}}}(\Theta_{i\mathcal{S}}) \equiv \mathcal{G}_{\Theta_{i\mathcal{S}}}(V_{i\mathcal{S};0}, \nu_{i\mathcal{S};0}) \quad (4)$$

are used. The conjugate pdfs have the parameterized-model-induced functional form $\mathcal{G}$*. They are determined by the* prior finite-dimensional statistic $V_{i\mathcal{S};0}$*, by the* prior sample counter $\nu_{i\mathcal{S};0}$ *and* indicator $\chi_{\Theta^*_{i\mathcal{S}}}(\Theta_{i\mathcal{S}})$ *of the set* $\Theta^*_{i\mathcal{S}}$*.*

Then, the parameters $\Theta_{i\mathcal{S}}$ *are independent a posteriori and the respective* posterior pdfs $f(\Theta_{i\mathcal{S}}|d(t))$ *preserve the functional form of the prior pdfs*

$$f(\Theta_{i\mathcal{S}}|d(t)) = \frac{A^{\nu_{i\mathcal{S};t}}(\Theta_{i\mathcal{S}}) \exp[\langle V_{i\mathcal{S};t}, C(\Theta_{i\mathcal{S}})\rangle] \chi_{\Theta^*_{i\mathcal{S}}}(\Theta_{i\mathcal{S}})}{\mathcal{I}(V_{i\mathcal{S};t}, \nu_{i\mathcal{S};t})} = \frac{\mathcal{G}_{\Theta_{i\mathcal{S}}}(V_{i\mathcal{S};t}, \nu_{i\mathcal{S};t})}{\mathcal{I}(V_{i\mathcal{S};t}, \nu_{i\mathcal{S};t})}$$

$$\mathcal{I}(V_{i\mathcal{S};t}, \nu_{i\mathcal{S};t}) = \int A^{\nu_{i\mathcal{S};t}}(\Theta_{i\mathcal{S}}) \exp[\langle V_{i\mathcal{S};t}, C(\Theta_{i\mathcal{S}})\rangle] \chi_{\Theta^*_{i\mathcal{S}}}(\Theta_{i\mathcal{S}})\, d\Theta_{i\mathcal{S}}. \quad (5)$$

The involved statistics $V_{i\mathcal{S};t}$, $\nu_{i\mathcal{S};t}$ *can be updated recursively*

$$V_{i\mathcal{S};t} = V_{i\mathcal{S};t-1} + B(\Psi_{i\mathcal{S};t}),\ \nu_{i\mathcal{S};t} = \nu_{i\mathcal{S};t-1} + 1 \text{ with a priori chosen } V_{i\mathcal{S};0},\ \nu_{i\mathcal{S};0}. \quad (6)$$

The predictive pdf*, modelling evolution of the* i*-th data entry (*i-th channel*), is given by the formula*

$$f(d_{i;t}|d_{i+1;t}, \ldots, d_{\mathring{d};t}, d(t-1), \mathcal{S}) = \frac{\mathcal{I}(V_{i\mathcal{S};t-1} + B(\Psi_{i\mathcal{S};t}), \nu_{i\mathcal{S};t-1} + 1)}{\mathcal{I}(V_{i\mathcal{S};t-1}, \nu_{i\mathcal{S};t-1})}. \quad (7)$$

The overall predictive pdf, given by the structure $\mathcal{S}$*, is product of pdfs (7) over* i*. The joint pdf of data conditioned by the structure* $\mathcal{S}$ *is*

$$f(d(\mathring{t})|\mathcal{S}) \quad = \quad \prod_{i=1}^{\mathring{d}} \mathcal{L}_i(d(\mathring{t}), \mathcal{S}) \text{ with } \mathcal{L}_i(d(\mathring{t}), \mathcal{S}) = \frac{\mathcal{I}(V_{i\mathcal{S};\mathring{t}}, \nu_{i\mathcal{S};\mathring{t}})}{\mathcal{I}(V_{i\mathcal{S};0}, \nu_{i\mathcal{S};0})} \quad (8)$$

called partial likelihood*.* $\mathcal{L}_i(d(\mathring{t}), \mathcal{S})$ *expresses descriptive abilities of the model having the structure* $\mathcal{S}$ *judged with respect to* i*-th channel.*

Thus, the estimation and prediction can be performed for respective is, channel-wise. We can focus the attention on a fixed channel and *drop the index* i remembering that a scalar variable is predicted.

The estimation and prediction reduce to algebraic operations with the *finite-dimensional statistic* $V_{\mathcal{S};t}$ and of the *sample counter* $\nu_{\mathcal{S};t}$. Moreover, a single type of the integral $\mathcal{I}(V_{\mathcal{S}}, \nu_{\mathcal{S}})$ has to be evaluated.

The need to get the *complete recursion* explains the requirement for the possibility to update data vector $\Psi_{\mathcal{S};t}$ recursively. Note that this requirement excludes use of models with unknown moving average noise.

We inspect influence of particular knowledge items $\mathcal{K} \in \mathcal{K}^* = \{1, \ldots, \mathring{\mathcal{K}}\}$ on descriptive abilities of the adopted model. The notation $\mathcal{L}(d(\mathring{t}), \mathcal{S}, \mathcal{K})$ stresses the use of the prior pdf $f(\Theta_{\mathcal{S}}|\mathcal{K})$. Similarly, $\mathcal{L}(d(\mathring{t}), \mathcal{S}, \mathcal{K}(\mathring{\mathcal{K}}))$ denotes the joint predictive pdf obtained when using the prior pdf $f(\Theta_{\mathcal{S}}|\mathcal{K}(\mathring{\mathcal{K}}))$ that includes all knowledge items available.

We need the following proposition (the fixed index $\mathcal{S}$ is dropped).

Proposition 2.2 (Weighted geometric mean of conjugate pdfs) *Let* $f(\Theta)$, ${}^{\lfloor a}f(\Theta)$ *be a pair of pdfs conjugated to the parameterized model in the exponential family, i.e.* $f(\Theta) \propto \mathcal{G}_\Theta(V, \nu)$ *and* ${}^{\lfloor a}f(\Theta) \propto \mathcal{G}_\Theta\left({}^{\lfloor a}V, {}^{\lfloor a}\nu\right)$.

Then, their geometric mean $f_\lambda \propto f^\lambda\, {}^{\lfloor a}f^{1-\lambda}$, *weighted by the factor* $\lambda \in [0, 1]$, *is the conjugated pdf* $f_\lambda(\Theta) \propto \mathcal{G}_\Theta(V_\lambda, \nu_\lambda)$ *whose statistics are*

$$V_\lambda = \lambda V + (1-\lambda)\, {}^{\lfloor a}V,\ \nu_\lambda = \lambda\nu + (1-\lambda)\, {}^{\lfloor a}\nu. \tag{9}$$

This proposition serves for tracking of slow parameter changes using the *stabilized forgetting* [13]. There, $f(\Theta) \equiv f(\Theta|d(t))$ is the posterior pdf of Θ based on data $d(t)$ measured up to the moment t when the forgetting is applied. The externally supplied *alternative pdf* ${}^{\lfloor a}f(\Theta)$ describes possible changes of estimated parameters before measuring next data. The weight λ called *forgetting factor* is interpreted as the probability that the parameters do not change. The usual *exponential forgetting* is obtained by taking the completely flat alternative ${}^{\lfloor a}f(\Theta) \propto 1$. Use of the pre-prior pdf $\bar{f}(\Theta) \propto \mathcal{G}_\Theta(\bar{V}, \bar{\nu})$ as the alternative pdf ${}^{\lfloor a}f(\Theta)$ is more wise.

Mostly, $\bar{f}(\Theta)$ is a flat pdf that respects just finiteness of Θ values. It conservatively guarantees that the forgotten posterior pdf does not move on the area of improbable infinite values of Θ.

The geometric mean of pdfs serves us also for finding a representant $f(\Theta|d(\mathring{t}), \mathcal{K}(\mathring{\mathcal{K}}))$ of several pdfs $f(\Theta|d(\mathring{t}), \mathcal{K}),\ \mathcal{K} \in \mathcal{K}^*$, each including a piece of prior knowledge $\mathcal{K}$ about Θ. For the measured data $d(\mathring{t})$ and no prejudice, the degree of belief to each of them coincides with the posterior probability

$f(\mathcal{K}|d(\mathring{t})) \propto \mathcal{L}(d(\mathring{t}),\mathcal{K})$. The representant $f(\Theta|d(\mathring{t}),\mathcal{K}(\mathring{\mathcal{K}}))$, called *merger* and motivated in [**?**], is chosen as the weighted geometric mean

$$\begin{aligned} f(\Theta|d(\mathring{t}),\mathcal{K}(\mathring{\mathcal{K}})) &\propto \prod_{\mathcal{K}=1}^{\mathring{\mathcal{K}}} \left[f(\Theta|d(\mathring{t}),\mathcal{K})\right]^{f(\mathcal{K}|d(\mathring{t}))} \\ &= \mathcal{G}_\Theta\left(\sum_{\mathcal{K}=1}^{\mathring{\mathcal{K}}} f(\mathcal{K}|d(\mathring{t}))V_{\mathcal{K};\mathring{t}}, \sum_{\mathcal{K}=1}^{\mathring{\mathcal{K}}} f(\mathcal{K}|d(\mathring{t}))\nu_{\mathcal{K};\mathring{t}}\right). \end{aligned} \tag{10}$$

2.3 Structure estimation in nested exponential family

For a fixed functional form of the exponential family, the *model structure* $\mathcal{S}$ is determined by the allowed entries of the regression vector. By collecting the potential entries into the *richest regression vector* $\psi_{\mathcal{R};t}$, the estimation of the model structure can be formulated as a selection of indices in it. They mark those entries that should be used in the proper regression vector $\psi_{\mathcal{S};t}$. There are $2^{\mathring{\psi}_\mathcal{R}}$ of such options. This number is usually excessive one and prevents the straightforward Bayesian structure estimation through judging the posterior probabilities of competing structures $f(\mathcal{S}|d(\mathring{t})) \propto \mathcal{L}(d(\mathring{t}),\mathcal{S})f(\mathcal{S})$ These posterior probabilities qualify a posteriori the discrete pointers $\mathcal{S}$ that have the prior pf $f(\mathcal{S})$.

The accumulation of $V_{\mathcal{S};\mathring{t}}$ makes the main computational burden related to the structure estimation. The current implementation in the system DESIGNER [17] relies on nesting of competitive structures $\mathcal{S}$ in the richest one $\mathcal{R}$. The model, given by the data vectors $\Psi_{\mathcal{S};t}$, is called nested in the richest one $\Psi_{\mathcal{R};t}$ if there is a *linear nesting mapping* $N_\mathcal{S}$ such that, cf. (3),

$$B(\Psi_{\mathcal{S};t}) = N_\mathcal{S}[B(\Psi_{\mathcal{R};t})]. \tag{11}$$

This notion and Proposition 2.1 imply the following statement.

Proposition 2.3 (Nesting in exponential family) *Let the parameterized model with the* richest structure *belong to the exponential family (3)* $f(d_t|\psi_{\mathcal{R};t},\Theta_\mathcal{R}) = A(\Theta_\mathcal{R})\exp[\langle B(\Psi_{\mathcal{R};t}), C(\Theta_\mathcal{R})\rangle]$. *Let us consider another model* $f(d_t|\psi_{\mathcal{S};t},\Theta_\mathcal{S}) = A(\Theta_\mathcal{S})\exp[\langle B(\Psi_{\mathcal{S};t}), C(\Theta_\mathcal{S})\rangle]$. *Let* $N_\mathcal{S}$ *be a time-invariant,* linear nesting mapping *such that* $B(\Psi_{\mathcal{S};t}) = N_\mathcal{S}[B(\Psi_{\mathcal{R};t})]$. *Let us assume that* $\boxed{V_{\mathcal{S};0} = N_\mathcal{S}[V_{\mathcal{R};0}]}$.

Then, the V-statistics of the posterior pdfs of both models are related by the nesting mapping $V_{\mathcal{S};\mathring{t}} = N_\mathcal{S}[V_{\mathcal{R};\mathring{t}}]$ *and the posterior probability on the*

structure $\mathcal{S}$ *is given by the formula*

$$f(\mathcal{S}|d(\mathring{t})) \propto \frac{\mathcal{I}(N_{\mathcal{S}}[V_{\mathcal{R};\mathring{t}}], \nu_{\mathcal{S};\mathring{t}})}{\mathcal{I}(N_{\mathcal{S}}[V_{\mathcal{R};0}], \nu_{\mathcal{S};0})} f(\mathcal{S}). \tag{12}$$

Thus, for the nested models and *nested prior statistics*, it is sufficient to collect the V-statistic for the richest structure. It helps but only partially. Full evaluation of the pf values (12) on the complete set of $2^{\mathring{\psi}_{\mathcal{R}}}$ competitive structures is still impossible. Thus, *maximum a posteriori probability* (MAP) estimate of $\mathcal{S}$ has to be searched for. The non-normalized values of the pf (12) evaluated during the search provide useful partial information on highly probable structures. The following conceptual search algorithm is used [11].

Algorithm 2.1 (SEN: MAP structure estimate of nested model)
Initial phase

- *Collect the real-data-dependent increment* ${}^{\lfloor\Delta}V_{\mathcal{R};\mathring{t}}$ *of the* V*-statistic corresponding to the richest structure of the data vector* $\Psi_{\mathcal{R};t}$

$${}^{\lfloor\Delta}V_{\mathcal{R};\mathring{t}} = \sum_{t=1}^{\mathring{t}} B(\Psi_{\mathcal{R};t}). \tag{13}$$

- *Select the prior statistic* $V_{\mathcal{R};0}$ *so that* $V_{\mathcal{S};0}$ *are nested in it* $\forall \mathcal{S} \in \mathcal{S}^*$.
- *Specify prior pf* $f(\mathcal{S})$ *of competitive structures, often as uniform one.*

Search phase *is run until the pre-specified number of restarts is reached.*

1. *Initialize the current guess of the structure.*

   Empty, richest and user-specified structures of regression vectors are used. These options are complemented by structures selected randomly among a priori possible ones.

2. *Do while the value of the posterior partial likelihood increases.*

   (a) *Make full search for the best structure within a* neighborhood *of the current guess of the structure, i.e. maximize within it the posterior partial likelihood*

$$\mathcal{L}(d(\mathring{t}), \mathcal{S}) f(\mathcal{S}) = \frac{\mathcal{I}\left(N_{\mathcal{S}}\left[{}^{\lfloor\Delta}V_{\mathcal{R};\mathring{t}} + V_{\mathcal{R};0}\right], \mathring{t} + \nu_{\mathcal{R};0}\right)}{\mathcal{I}\left(N_{\mathcal{S}}[V_{\mathcal{R};0}], \nu_{\mathcal{R};0}\right)} f(\mathcal{S}).$$

   The neighborhood consists of all structures gained by:

- *adding a single entry to the current guess of the structure,*
- *removing a single entry from the current guess of the structure,*
- *considering structures nested in those named above.*

(b) Take the maximizer as a new current guess of the structure.

2.4 Quantification of prior knowledge

Parameter and structure estimation are sensitive to the information content of the learning data. As a rule, the available data are poorly informative. Then, prior knowledge has to be used. In the Bayesian set up, it is feeded through the prior pdf. Here, we outline the way how it can be constructed. The following conditions are specific for technological applications.

$\oplus$ Groups of widely accessible knowledge types exist.
$\oplus$ Experimental data $d(\mathring{t})$ measured on the modelled system are available.
$\oplus$ Admissible prior pdfs are restricted to conjugate ones, thus the prior knowledge is to be translated into values of the prior statistics V, ν.
$\ominus$ The person feeding the prior knowledge does not care about the probabilistic tool set exploited.
$\ominus$ No supervisor for knowledge elicitation and judgement of the expert competence is at disposal.
$\ominus$ Knowledge items processed are expected to be repetitive, not fully consistent and differing in precision and nature. Mutual dependencies of knowledge items are unknown.

The following quantification algorithm respects these conditions [**?**].

Algorithm 2.2 (Quantification of prior knowledge)
Initiation phase

- *Select the i-th parameterized model (2) of a fixed structure $\mathcal{S}$ in the exponential family you deal with.*
- *Collect the real-data-dependent increment ${}^{\lfloor\Delta}V_{\mathcal{S};\mathring{t}}$ of the V-statistics according to (13) for $\mathcal{R} = \mathcal{S}$.*
- *Split the existing knowledge into internally consistent knowledge items.*
- *Select the* pre-prior pdf $\bar{f}(\Theta_{\mathcal{S}}) \propto \mathcal{G}_{\Theta_{\mathcal{S}}}(\bar{V}_{\mathcal{S}}, \bar{\nu}_{\mathcal{S}})$ *on unknown parameters $\Theta_{\mathcal{S}}$ that expresses the common knowledge available.*
- *Initialize the auxiliary scalar normalization factor $s = 0$.*

Quantification phase *runs for the internally consistent knowledge items $\mathcal{K} \in \mathcal{K}^*$.*

- *Translate the knowledge item $\mathcal{K}$ into the fictitious-data $d(\mathring{\tau}_{\mathcal{K}})$ dependent increments ${}^{\lfloor\Delta}V_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}}$, ${}^{\lfloor\Delta}\nu_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}}$ of the pre-prior statistics $\bar{V}_{\mathcal{S}}$, $\bar{\nu}_{\mathcal{S}}$ so that $V_{\mathcal{SK};0} = {}^{\lfloor\Delta}V_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}} + \bar{V}_{\mathcal{S}}$ and $\nu_{\mathcal{SK};0} = {}^{\lfloor\Delta}\nu_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}} + \bar{\nu}_{\mathcal{S}}$ reflect the processed knowledge item, i.e. $f(\Theta_{\mathcal{S}}|\mathcal{K}) \propto \mathcal{G}_{\Theta_{\mathcal{S}}}(V_{\mathcal{SK};0}, \nu_{\mathcal{SK};0})$.*

- *Evaluate the descriptive abilities gained by exploiting this knowledge on real data $d(\mathring{t})$ and update the normalization factor s*

$$\mathcal{L}(d(\mathring{t}),\mathcal{S},\mathcal{K}) = \frac{\mathcal{I}\left({}^{\lfloor\Delta}V_{\mathcal{SK};\mathring{t}} + V_{\mathcal{SK};0}, \mathring{t} + \nu_{\mathcal{SK};0}\right)}{\mathcal{I}(V_{\mathcal{SK};0}, \nu_{\mathcal{SK};0})},\ s = s + \mathcal{L}(d(\mathring{t}),\mathcal{S},\mathcal{K}). \quad (14)$$

Merging phase *combines particular knowledge items into the merger (10)*

$$f(\Theta_{\mathcal{S}}|d(\mathring{t}),\mathcal{K}(\mathring{\mathcal{K}})) \ \propto\ f(d(\mathring{t})|\Theta_{\mathcal{S}}) \prod_{\mathcal{K}=1}^{\mathring{\mathcal{K}}} [f(\Theta_{\mathcal{S}}|\mathcal{K})]^{f(\mathcal{K}|d(\mathring{t}),\mathcal{S})} \quad (15)$$

$$f(\mathcal{K}|d(\mathring{t}),\mathcal{S}) \ =\ \frac{\mathcal{L}(d(\mathring{t}),\mathcal{S},\mathcal{K})}{s}.$$

It gives $f(\Theta_{\mathcal{S}}|d(\mathring{t}),\mathcal{K}(\mathring{\mathcal{K}})) \propto$

$$\mathcal{G}_{\Theta_{\mathcal{S}}}\Bigg(\underbrace{\underbrace{{}^{\lfloor\Delta}V_{\mathcal{S};\mathring{t}} + \sum_{\mathcal{K}=1}^{\mathring{\mathcal{K}}} f(\mathcal{K}|d(\mathring{t}),\mathcal{S})V_{\mathcal{SK};0}}_{V_{\mathcal{SK}(\mathring{\mathcal{K}});0}}, \mathring{t} + \underbrace{\sum_{\mathcal{K}=1}^{\mathring{\mathcal{K}}} f(\mathcal{K}|d(\mathring{t}),\mathcal{S})\nu_{\mathcal{SK};0}}_{\nu_{\mathcal{SK}(\mathring{\mathcal{K}});0}}}_{V_{\mathcal{SK}(\mathring{\mathcal{K}});\mathring{t}}\qquad\qquad\qquad\qquad\qquad\qquad \nu_{\mathcal{SK}(\mathring{\mathcal{K}});\mathring{t}}}\Bigg).$$

It remains to specify the meaning of internally consistent knowledge item and to show how to construct the increments of statistics on fictitious data. These aspects are covered in Section 3.

2.5 Addressed problem and its conceptual solution

The problem, we are facing, stems from the fact that usually the *prior sufficient statistic* $V_{\mathcal{SK};0}$ expressing the piece of knowledge $\mathcal{K}$ within the structure $\mathcal{S}$ *is not nested* in the statistics corresponding to that with the richest data vector. In other words, the *efficient nested structure estimation* Algorithm 2.1 *cannot be directly used whenever a non-trivial prior knowledge is to be exploited*. This fact was overlooked in our former implementations and publications and caused worse estimation results than expected.

Knowing the problem, the remedy is rather straightforward. The following conceptual algorithm is used.

Algorithm 2.3 (Structure estimation with prior knowledge)
Initial phase

- *Select the parameterized model in the exponential family and the richest structure of the underlying regression vector* $\psi_{\mathcal{R}}$.
- *Select the pre-prior pdf* $\bar{f}(\Theta_{\mathcal{R}}) \propto \mathcal{G}_{\Theta_{\mathcal{R}}}(\bar{V}_{\mathcal{R}}, \bar{\nu}_{\mathcal{R}})$ *on unknown parameters* $\Theta_{\mathcal{R}}$ *that express the common knowledge available while requiring that the* same knowledge is expressed for all nested structures *of interest.*

Structure estimation with nested prior knowledge

- *Apply the algorithm SEN 2.1 in order to find a pre-selected number* $\mathring{\mathcal{S}}$ *of structures* $\mathcal{S} \in \mathcal{S}^* = \{1, 2, \ldots, \mathring{\mathcal{S}}\}$ *having the highest posterior probabilities when using the restricted prior knowledge described by the pdf* $\bar{f}(\Theta_{\mathcal{R}})$.

Inclusion of prior knowledge for promising structures $\mathcal{S} \in \mathcal{S}^*$

1. *Apply Algorithm 2.2 for the fixed structure* $\mathcal{S}$ *to get the statistics of the best merger* $V_{\mathcal{SK}(\mathring{\mathcal{K}});\tau}$, $\nu_{\mathcal{SK}(\mathring{\mathcal{K}});\tau}$, $\tau \in \{0, \mathring{t}\}$, *cf. (15).*
2. *Evaluate descriptive abilities of the best merger, within the given structure* $\mathcal{S}$, *cf. (15),*

$$\mathcal{L}(d(\mathring{t}), \mathcal{S}) = \frac{\mathcal{I}\left(V_{\mathcal{SK}(\mathring{\mathcal{K}});\mathring{t}}, \nu_{\mathcal{SK}(\mathring{\mathcal{K}});\mathring{t}}\right)}{\mathcal{I}\left(V_{\mathcal{SK}(\mathring{\mathcal{K}});0}, \nu_{\mathcal{SK}(\mathring{\mathcal{K}});0}\right)}.$$

3. *Provide* $\hat{\mathcal{S}} \in \operatorname{Arg\,max}_{\mathcal{S}^*} \mathcal{L}(d(\mathring{t}), \mathcal{S}) f(\mathcal{S})$ *as the structure estimate and its statistics* $V_{\hat{\mathcal{S}}\mathcal{K}(\mathring{\mathcal{K}};\mathring{t}}$, $\nu_{\hat{\mathcal{S}}\mathcal{K}(\mathring{\mathcal{K}};\mathring{t}}$ *as initial conditions for the subsequent on-line estimation and as the alternative pdf needed in the stabilized forgetting.*

3 Fictitious data

Here, the construction of the common information basis, i.e. fictitious data, is recalled and refined [**?**]. It allows us to cope with knowledge items of a different nature in a unified way.

3.1 Internally consistent fictitious data blocks

Some information sources provide knowledge pieces $\mathcal{K}$ *in form of data blocks* $d(\mathring{\tau}_{\mathcal{K}})$. They include obsolete or interpolated data measured on the system in question or data measured on a similar system, data from identification experiments violating usual working conditions – like measurement of step response – and data gained from a realistic simulation.

The data block $d(\mathring{\tau}_{\mathcal{K}})$ expressing the knowledge piece $\mathcal{K}$ is called internally consistent iff $f(\Theta_{\mathcal{S}}|\mathcal{K})$ coincides with a flattened version of the posterior pdf $f(\Theta_{\mathcal{S}}|d(\mathring{\tau}_{\mathcal{K}})) \propto \mathcal{G}_{\Theta_{\mathcal{S}}}(V_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}}, \nu_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}})$. Let us describe it in a detail.

The posterior pdf is obtained by the Bayes rule applied to the pre-prior pdf $\bar{f}(\Theta_{\mathcal{S}}) \propto \mathcal{G}_{\Theta_{\mathcal{S}}}(\bar{V}_{\mathcal{S}}, \bar{\nu}_{\mathcal{S}})$ with the stabilized forgetting. The *forgetting is used in order to counteract mismodelling*. The pre-prior pdf is used as the alternative pdf. Thus, the adequate forgetting factor λ is to be chosen only. A comparison of descriptive abilities of the posterior pdfs obtained for various forgetting factors serves well to this purpose. It is done anyway during merging of individual knowledge pieces. Thus, it suffices to take pdfs $f(\Theta_{\mathcal{S}}|d(\mathring{\tau}_{\mathcal{K}})), \lambda) \propto \mathcal{G}_{\Theta_{\mathcal{S}}}(V_{\mathcal{S}\lambda;\mathring{\tau}_{\mathcal{K}}}, \nu_{\mathcal{S}\lambda;\mathring{\tau}_{\mathcal{K}}})$ gained for different λ as different knowledge pieces. It is done from now on and the reference to λ is suppressed.

Mismodelling or obsolete nature of the fictitious data blocks imply that the pdfs

$$f(\Theta_{\mathcal{S}}|d(\mathring{\tau}_{\mathcal{K}})) = \mathcal{G}_{\Theta_{\mathcal{S}}}(V_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}}, \nu_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}}) \equiv \mathcal{G}_{\Theta_{\mathcal{S}}}\left({}^{\lfloor\Delta}V_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}} + \bar{V}_{\mathcal{S}},\ {}^{\lfloor\Delta}\nu_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}} + \bar{\nu}_{\mathcal{S}}\right) \tag{16}$$

reflect system properties approximately only. Consequently, these pdfs have to be adequately flattened before merging such a piece of knowledge. This is of an extreme importance as, for instance, simulators may provide a huge amount of data that may over-fit the posterior pdfs at wrong positions. The geometric mean serves well for flattening of these pdfs, cf. (16),

$$\begin{aligned} f(\Theta_{\mathcal{S}}|\mathcal{K}) &\propto \left[f(\Theta_{\mathcal{S}}|d(\mathring{\tau}_{\mathcal{K}}))\right]^{\Lambda}_{\mathcal{S}\mathcal{K}} \left[\bar{f}(\Theta_{\mathcal{S}})\right]^{1-\Lambda_{\mathcal{S}\mathcal{K}}} \\ &\propto \mathcal{G}_{\Theta_{\mathcal{S}}}\left(\Lambda_{\mathcal{S}\mathcal{K}}V_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}} + (1-\Lambda_{\mathcal{S}\mathcal{K}})\bar{V}_{\mathcal{S}}, \Lambda_{\mathcal{S}\mathcal{K}}\nu_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}} + (1-\Lambda_{\mathcal{S}\mathcal{K}})\bar{\nu}_{\mathcal{S}}\right) \\ &= \mathcal{G}_{\Theta_{\mathcal{S}}}\left(\Lambda_{\mathcal{S}\mathcal{K}}\,{}^{\lfloor\Delta}V_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}} + \bar{V}_{\mathcal{S}}, \Lambda_{\mathcal{S}\mathcal{K}}\,{}^{\lfloor\Delta}\nu_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}} + \bar{\nu}_{\mathcal{S}}\right). \end{aligned} \tag{17}$$

The scalar $\Lambda_{\mathcal{S}\mathcal{K}} \in [0,1]$, called *flattening factor*, is selected to maximize the descriptive ability, i.e.

$$\Lambda_{\mathcal{S}\mathcal{K}} = \arg\max_{\Lambda\in[0,1]} \frac{\mathcal{I}\left(\,{}^{\lfloor\Delta}V_{\mathcal{S};\mathring{t}} + \Lambda\,{}^{\lfloor\Delta}V_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}} + \bar{V}_{\mathcal{S}}; \mathring{t} + \Lambda\,{}^{\lfloor\Delta}\nu_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}} + \bar{\nu}_{\mathcal{S}}\right)}{\mathcal{I}\left(\Lambda\,{}^{\lfloor\Delta}V_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}} + \bar{V}_{\mathcal{S}}; \Lambda\,{}^{\lfloor\Delta}\nu_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}} + \bar{\nu}_{\mathcal{S}},\right)}. \tag{18}$$

The values of the flattening factor $\Lambda_{\mathcal{SK}}$ are structure-dependent. Thus, the optimized flattening spoils possible nesting of ${}^{\lfloor\Delta}V_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}}$.

3.2 Construction of fictitious data

Here, we focus on those prior knowledge items that *do not have directly the form of data blocks* but can be interpreted as the *expected system response on a pre-specified stimulus*. Static gain, a point on frequency response, time-constants serve as their prominent examples. Such a knowledge item $\mathcal{K}$ is *always uncertain* to some degree. It can be interpreted as a collection of partial characterizations of the several predictors. Each of them is expressed in terms of its prior pdf $f(\Theta_{\mathcal{S}}|\tau_{\mathcal{K}})$

$$f(d_{\tau_{\mathcal{K}}}|\psi_{\mathcal{S};\tau_{\mathcal{K}}}) = \int f(d_{\tau_{\mathcal{K}}}|\psi_{\mathcal{S};\tau_{\mathcal{K}}}, \Theta_{\mathcal{S}}) f(\Theta_{\mathcal{S}}|\tau_{\mathcal{K}})\, d\Theta_{\mathcal{S}} \tag{19}$$

for respective regression vectors $\psi_{\mathcal{S};\tau_{\mathcal{K}}}$, $\tau_{\mathcal{K}} = 1, \ldots, \mathring{\tau}_{\mathcal{K}}$. Mostly, the $\tau_{\mathcal{K}}$-th part of the knowledge piece $\mathcal{K}$ can be expressed as initial moments of the pdf (19). Formally,

$$h(\psi_{\mathcal{S};\tau_{\mathcal{K}}}) = \int H(d_{\tau_{\mathcal{K}}}, \psi_{\mathcal{S};\tau_{\mathcal{K}}}) f(d_{\tau_{\mathcal{K}}}|\psi_{\mathcal{S};\tau_{\mathcal{K}}})\, dd_{\tau_{\mathcal{K}}}. \tag{20}$$

$h(\psi_{\mathcal{S};\tau_{\mathcal{K}}})$ and $H(\Psi_{\mathcal{S};\tau_{\mathcal{K}}}) = H(d_{i;\tau_{\mathcal{K}}}, \psi_{\mathcal{S};\tau_{\mathcal{K}}})$ are known vector functions of the indicated arguments. When there is no pdf $f(\Theta_{\mathcal{S}}|\tau_{\mathcal{K}})$ fulfilling (19), (20), then this information source is inconsistent and has to be split into several, internally consistent, knowledge sources. Then, the restrictions (19), (20) do not determine fully the constructed pdf $f(\Theta_{\mathcal{S}}|\tau_{\mathcal{K}})$. Pragmatic reasons make us to search within the class of conjugate pdfs. Moreover, it is reasonable to construct such a pdf $f(\Theta_{\mathcal{S}}|\tau_{\mathcal{K}})$ that reflects *just* the knowledge item expressed by (19), (20). Thus, it makes sense to choose such a pdf $f(\Theta_{\mathcal{S}}|\tau_{\mathcal{K}})$ that is the nearest one to the flat pre-prior pdf $\bar{f}(\Theta_{\mathcal{S}})$. The choice is made among those meeting (19), (20). The Kullback-Leibler distance [18] $\mathcal{D}(f||\bar{f}) = \int f(\Theta_{\mathcal{S}}|\tau_{\mathcal{K}}) \ln\left[f(\Theta_{\mathcal{S}}|\tau_{\mathcal{K}})/\bar{f}(\Theta_{\mathcal{S}})\right]\, d\Theta_{\mathcal{S}}$ is used as an adequate proximity measure.

Both the found $f(\Theta_{\mathcal{S}}|\tau_{\mathcal{K}})$ and pre-prior pdf $\bar{f}(\Theta_{\mathcal{S}})$ belong to the same conjugate class. Thus, their ratio can be interpreted as a product of the parameterized model at some fictitious data vectors, leading to the statistics increments ${}^{\lfloor\Delta}V_{\mathcal{S};\tau_{\mathcal{K}}}$, ${}^{\lfloor\Delta}\nu_{\mathcal{S};\tau_{\mathcal{K}}}$. The knowledge item $\mathcal{K}$ is supposed to be internally consistent. Consequently, fictitious data vectors obtained for various $\tau_{\mathcal{K}}$ should be processed by using the Bayes rule. The result is then scaled

in the same way as the consistent data block, Section 3.1. Thus, $f(\Theta_{\mathcal{S}}|\mathcal{K})$ has the form (17) with statistics

$$\lfloor^{\Delta}V_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}} = \sum_{\tau_{\mathcal{K}}=1}^{\mathring{\tau}_{\mathcal{K}}} \lfloor^{\Delta}V_{\mathcal{S};\tau_{\mathcal{K}}}, \quad \lfloor^{\Delta}\nu_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}} = \sum_{\tau_{\mathcal{K}}=1}^{\mathring{\tau}_{\mathcal{K}}} \lfloor^{\Delta}\nu_{\mathcal{S};\tau_{\mathcal{K}}}, \text{ where} \tag{21}$$

$\mathcal{G}_{\Theta_{\mathcal{S}}} = \left(\lfloor^{\Delta}V_{\mathcal{S};\tau_{\mathcal{K}}} + \bar{V}_{\mathcal{S}}, \lfloor^{\Delta}\nu_{\mathcal{S};\tau_{\mathcal{K}}} + \bar{\nu}_{\mathcal{S}}\right)$ are pdfs minimizing the Kullback-Leibler distance to $\mathcal{G}_{\Theta_{\mathcal{S}}}\left(\bar{V}_{\mathcal{S}}, \bar{\nu}_{\mathcal{S}}\right)$ under restrictions (19), (20). The factor $\Lambda_{\mathcal{S}\mathcal{K}}$ results from the maximization (18).

Specific optimization steps are elaborated for the ARX model in the next section.

4 Application to normal ARX model

The system DESIGNER, in which the presented results are predominantly used, relies on the normal ARX model and its variants. This makes us to specialize the general solution to this case.

4.1 Estimation and prediction with normal ARX model

The normal ARX model belongs to the exponential family with the following correspondence to its general form (3)

$$\begin{aligned} f(d|\psi,\Theta) = \mathcal{N}_d(\theta'\psi, r) = A(\Theta)\exp[\langle B(\Psi), C(\Theta)\rangle] \quad \text{with} \\ \Theta = [\theta, r] = [\text{regression coefficients, noise variance}],\ A(\Theta) = (2\pi r)^{-0.5} \\ \langle B, C\rangle = \mathrm{tr}[B'C],\ B(\Psi) = \Psi\Psi',\ C(\Theta) = (2r)^{-1}[-1,\theta']'[-1,\theta']. \end{aligned} \tag{22}$$

This correspondence determines the conjugate prior (4) in the form known as the Gauss-inverse-Wishart (GiW) pdf [19]

$$\mathcal{G}_{\Theta}(V,\nu) = r^{-0.5(\nu+\mathring{\psi}+2)}\exp\left\{-\frac{1}{2r}\mathrm{tr}\left(V\left[-1,\theta'\right]'\left[-1,\theta'\right]\right)\right\}. \tag{23}$$

The $(\mathring{\Psi},\mathring{\Psi})$-dimensional *extended information matrix* V can be chosen symmetric and must be *positive definite*. Otherwise, the function $\mathcal{G}_{\Theta}(V,\nu)$ cannot be normalized to a pdf. Thus, we can use the numerically advantageous $L'DL$ decomposition of this matrix [20]

$$\begin{aligned} V &= L'DL,\ L = \text{lower triangular with unit diagonal},\ D = \text{diagonal} \\ L &= \begin{bmatrix} 1 & 0 \\ \lfloor^{d\psi}L & \lfloor^{\psi}L \end{bmatrix},\ D = \mathrm{diag}\left[\lfloor^{d}D, \lfloor^{\psi}D\right],\ \lfloor^{d}D = \text{scalar}. \end{aligned} \tag{24}$$

Proposition 4.1 (Basic properties and moments of the GiW pdf) *The conjugate GiW pdf has the following alternative expression*

$$\begin{aligned}
\mathcal{G}_\Theta(L,D,\nu) &= \frac{r^{-0.5(\nu+\mathring{\psi}+2)}}{\mathcal{I}(L,D,\nu)}\exp\left\{-\frac{1}{2r}\left[\left(\theta-\hat{\theta}\right)' C^{-1}\left(\theta-\hat{\theta}\right)+ {}^{\lfloor d}D\right]\right\}\\
\hat{\theta} &= {}^{\lfloor\psi}L^{-1}\,{}^{\lfloor d\psi}L = \textit{ least-squares (LS) estimate of } \theta \qquad (25)\\
C &= {}^{\lfloor\psi}L^{-1}\,{}^{\lfloor\psi}D^{-1}\left({}^{\lfloor\psi}L'\right)^{-1} = \textit{LS covariance factor}\\
{}^{\lfloor d}D &= \textit{LS remainder.}
\end{aligned}$$

The normalization integral is

$$\begin{aligned}
\mathcal{I}(L,D,\nu) &= \Gamma(0.5\nu)\,{}^{\lfloor d}D^{-0.5\nu}\prod_{j=1}^{\mathring{\psi}}{}^{\lfloor\psi}D_{jj}^{-0.5}2^{0.5\nu}(2\pi)^{0.5\mathring{\psi}} \quad \textit{with} \qquad (26)\\
\Gamma(x) &= \int_0^\infty z^{x-1}\exp(-z)\,dz<\infty \quad \textit{for } x>0.
\end{aligned}$$

Thus, it is finite iff $\nu>0$ and V is positive definite $\Leftrightarrow D_{jj}>0,\ j=1,\ldots,\mathring{\Psi}$. Under this condition, the normalization integral $\mathcal{I}$ is positive.

The GiW pdf has the following moments

$$\begin{aligned}
\mathcal{E}[r|L,D,\nu] &= \frac{{}^{\lfloor d}D}{\nu-2}=\hat{r},\ \mathrm{var}[r|L,D,\nu]=\frac{2\hat{r}^2}{\nu-4}\\
\mathcal{E}[\theta|L,D,\nu] &= {}^{\lfloor\psi}L^{-1}\,{}^{\lfloor d\psi}L=\hat{\theta} \qquad (27)\\
\mathrm{cov}[\theta|L,D,\nu] &= \frac{{}^{\lfloor d}D}{\nu-2}\,{}^{\lfloor\psi}L^{-1}\,{}^{\lfloor\psi}D^{-1}\left({}^{\lfloor\psi}L'\right)^{-1}=\hat{r}C.
\end{aligned}$$

Proposition 2.1 specializes to the following normal variant.

Proposition 4.2 (Estimation & prediction with ARX model) *Let the normal ARX model (22) be considered, together with the conjugate GiW prior pdf $\mathcal{G}_\Theta(L_0,D_0,\nu_0)$ (25) and the alternative pdf $\mathcal{G}_\Theta\left({}^{\lfloor a}L,\,{}^{\lfloor a}D,\,{}^{\lfloor a}\nu\right)$ in the stabilized forgetting with the forgetting factor $\lambda\in[0,1]$.*

Then, the posterior pdf is the GiW pdf $\mathcal{G}_\Theta(L_t,D_t,\nu_t)$ and its sufficient statistics evolve according to the recursions

$$\begin{aligned}
L_t'D_tL_t &= \underbrace{\begin{bmatrix}L_{t-1}\\ \Psi_t'\\ {}^{\lfloor a}L\end{bmatrix}'}_{\mathbf{L'}}\underbrace{\mathrm{diag}\begin{bmatrix}\lambda D_{t-1}\\ \lambda\\ (1-\lambda)\,{}^{\lfloor a}D\end{bmatrix}}_{\mathbf{D}}\underbrace{\begin{bmatrix}L_{t-1}\\ \Psi_t'\\ {}^{\lfloor a}L\end{bmatrix}}_{\mathbf{L}} \qquad (28)\\
\nu_t &= \lambda(\nu_{t-1}+1)+(1-\lambda)\,{}^{\lfloor a}\nu,\quad L_0,\,D_0,\,\nu_0 \textit{ given a priori.}
\end{aligned}$$

The rectangular matrix $\mathbf{L}'$ *is mapped on* $[L'_t, 0]$ *by the regular matrix* $\mathcal{T}$ *that diagonalizes the* $(\mathring{\Psi}, \mathring{\Psi})$*-left upper corner in* $\mathcal{T}^{-1}\mathrm{diag}[D, 1](\mathcal{T}')^{-1}$.

The predictive pdf is known as the Student pdf. *For any data vector* $\Psi = [d, \psi']'$*, its values can be found numerically as the ratio (7).*

4.2 Internally consistent fictitious data blocks

Processing of internally consistent data blocks coincides with Bayesian estimation of the ARX normal model. The one-to-one mapping of the extended information matrix V on the LS quantities, Proposition 4.1, implies that its updating is equivalent to recursive least squares [12]. It is equipped with tracking ability through the stabilized forgetting. Numerically, its $L'DL$ decomposition is evaluated by using an efficient, rank-one updating [20] that creates the mapping $\mathcal{T}$, see Proposition 4.2.

For the assumed positive definite $\bar{V}_{\mathcal{S}}$ and $\bar{\nu}_{\mathcal{S}} > 0$, Proposition 4.1 implies that the partial likelihood is continuous and bounded function for $\Lambda \in [0, 1]$. Thus, the maximizer (18) exist. It can be simply found by using, for instance, golden-section based maximization.

4.3 Construction of fictitious data

The most common case of prior knowledge type is solved here as the practically used example of its quantification. Specifically, initial moments of the predictive pdfs $f(d_{\tau_{\mathcal{K}}}|\psi_{\tau_{\mathcal{K}}})$ are assumed to be given for a fixed regression vector $\psi_{\tau_{\mathcal{K}}}$ (index of the fixed structure $\mathcal{S}$ is suppressed)

$$\begin{aligned} \hat{d}_{\tau_{\mathcal{K}}} &= \int d_{\tau_{\mathcal{K}}} f(d_{\tau_{\mathcal{K}}}|\psi_{\tau_{\mathcal{K}}})\, dd_{\tau_{\mathcal{K}}} \\ {}^{\lfloor d}r_{\tau_{\mathcal{K}}} &= \int (d_{\tau_{\mathcal{K}}} - \hat{d}_{\tau_{\mathcal{K}}})^2 f(d_{\tau_{\mathcal{K}}}|\psi_{\tau_{\mathcal{K}}})\, dd_{\tau_{\mathcal{K}}}. \end{aligned} \tag{29}$$

It corresponds to the knowledge $h(\psi_{\tau_{\mathcal{K}}}) = \left[\hat{d}_{\tau_{\mathcal{K}}},\, {}^{\lfloor d}r_{\tau_{\mathcal{K}}}\right]$, $H(d_{\tau_{\mathcal{K}}}, \psi_{\tau_{\mathcal{K}}}) = \left[d_{\tau_{\mathcal{K}}}, (d_{\tau_{\mathcal{K}}} - \hat{d}_{\tau_{\mathcal{K}}})^2\right]$ in (20). For the ARX model, the restriction on the constructed prior pdf (29) becomes, Proposition 4.1,

$$\hat{d}_{\tau_{\mathcal{K}}} = \hat{\theta}'_{\tau_{\mathcal{K}}}\psi_{\tau_{\mathcal{K}}},\ {}^{\lfloor d}r_{\tau_{\mathcal{K}}} = \hat{r}_{\tau_{\mathcal{K}}}(1 + \zeta_{\tau_{\mathcal{K}}}),\ \zeta_{\tau_{\mathcal{K}}} = \psi'_{\tau_{\mathcal{K}}} C_{\tau_{\mathcal{K}}} \psi_{\tau_{\mathcal{K}}}, \tag{30}$$

where $\hat{\theta}'_{\tau_{\mathcal{K}}}, \hat{r}_{\tau k}, C_{\tau_{\mathcal{K}}}$ are LS equivalents of the statistics $V_{\tau_{\mathcal{K}}}$ resulting from the minimization of the KL distance to the pre-prior pdf.

Typically, the expert provides the range $\left[\underline{d}_{\tau_{\mathcal{K}}}, \overline{d}_{\tau_{\mathcal{K}}}\right]$ of the response $d_{\tau_{\mathcal{K}}}$ on the stimulus coded in $\psi_{\mathcal{S};\tau_{\mathcal{K}}}$. Then, neglecting a small asymmetry of the Student pdf, we choose $\hat{d}_{\tau_{\mathcal{K}}} = 0.5\left(\underline{d}_{\tau_{\mathcal{K}}} + \overline{d}_{\tau_{\mathcal{K}}}\right)$ and ${}^{\lfloor d}r_{\tau_{\mathcal{K}}} = \left[0.5\left(\overline{d}_{\tau_{\mathcal{K}}} - \underline{d}_{\tau_{\mathcal{K}}}\right)\right]^2$.

The pre-prior pdf used in the minimization is assumed in the form

$$\bar{f}(\Theta) = \mathcal{G}_\Theta(I, \mathrm{diag}[\,{}^{\lfloor d}\varepsilon, \varepsilon \underbrace{[1, \ldots, 1]}_{\mathring{\psi}}], \bar{\nu}). \tag{31}$$

It is given by positive scalars ${}^{\lfloor d}\varepsilon, \varepsilon, \bar{\nu}$. Such a pdf expresses the common knowledge that the parameters are finite and, importantly, *this knowledge is preserved for all nested structures.*

Using Proposition 4.1 for the pre-prior pdf (31), the optimized Kullback-Leibler distance is (the *subscript* $\tau_{\mathcal{K}}$ *is also temporarily suppressed* whenever possible)

$$\begin{aligned} 2\mathcal{D}(f||\bar{f}) &= \omega(\nu) + \varepsilon \mathrm{tr}\,[C] - \ln|C| - \bar{\nu}\ln(1+\zeta) + \bar{\nu}\ln(\hat{r}) \\ &+ \frac{\nu}{(\nu-2)\hat{r}}\left[\varepsilon\hat{\theta}'\hat{\theta} + {}^{\lfloor d}\varepsilon\right], \end{aligned} \tag{32}$$

where $\omega(\nu)$ includes all terms depending on ν only. The optimization of this function with respect to ν is rather involved and, importantly, its results do not have an intuitive support. This makes us to minimize the function (32) with respect to the remaining arguments only and interpret the results as updating of $\bar{V}$ by a rank-one ${}^{\lfloor\Delta}V$ matrix defined by a *fictitious data vector* Ψ. Then, the restrictions (30) determine even ${}^{\lfloor\Delta}\nu$ uniquely.

Proposition 4.3 (Optimal fictitious data vector) *The LS quantities* $\hat{\theta}$, $C, \hat{r}$ *minimizing the function (32) under the restrictions (30) are obtained by the least-squares-type updating of the pre-prior statistics (31) using the fictitious data vector*

$$\Psi'_{\tau_{\mathcal{K}}} = \left[\hat{d}\left(\frac{\rho}{\sqrt{x}} + \sqrt{x}\right), \sqrt{x}\psi'\right], \ \rho = \frac{\varepsilon}{\psi'\psi}. \tag{33}$$

The weight $x > 0$ *is determined by the following formulas*

$$\begin{aligned} q &= \frac{\nu_{\tau_{\mathcal{K}}}}{\nu_{\tau_{\mathcal{K}}} - 2}(\alpha\rho + \beta), \quad \alpha = \frac{\hat{d}^2}{{}^{\lfloor d}r}, \quad \beta = \frac{{}^{\lfloor d}\varepsilon}{{}^{\lfloor d}r_{\tau_{\mathcal{K}}}} > 0 \\ b &= \rho + 1 - q + \bar{\nu}, \quad c = \rho(-\bar{\nu} + q) + q, \quad x = 0.5\left(-b + \sqrt{b^2 + 4c}\right). \end{aligned} \tag{34}$$

The corresponding $\nu_{\tau_{\mathcal{K}}}$ *is specified by the implicit formula*

$$\nu_{\tau_{\mathcal{K}}} = 2 + (1 + \rho + x)\left(\frac{\beta}{\rho + x} + \frac{\alpha\rho}{x}\right). \tag{35}$$

The coupled equations (34) (35) have a unique solution in the meaningful domain $x > 0$, $\nu_{\tau_{\mathcal{K}}} > 2$.

Proof: The minimization of the function (32) with respect to $\hat{\theta}$ gives directly

$$\hat{\theta} = \frac{\hat{d}\psi}{\psi'\psi} \tag{36}$$

irrespective of other variables. Inserting this $\hat{\theta}$ into the optimized function (32) and using the second restriction in (30) for expressing $\hat{r} = {}^{\lfloor d}r/(1+\zeta)$, we get the following function minimized with respect to C

$$\begin{aligned} 2\mathcal{D}(f||\bar{f}) &= \omega(\nu) + \varepsilon \mathrm{tr}\,[C] - \ln|C| - \bar{\nu}\ln(1+\zeta) + (1+\zeta)q \\ q &= \frac{\nu}{\nu-2}\left[\underbrace{\rho\frac{\hat{d}^2}{{}^{\lfloor d}r}}_{\alpha} + \underbrace{\frac{{}^{\lfloor d}\varepsilon}{{}^{\lfloor d}r}}_{\beta}\right], \quad \rho = \frac{\varepsilon}{\psi'\psi}. \end{aligned}$$

Taking its derivatives with respect to C and using the identity $\frac{\partial \ln|C|}{\partial C} = C^{-1}$, we get the necessary condition for minimum

$$C^{-1} = \varepsilon I + x\psi\psi' \text{ with } x = -\frac{\bar{\nu}}{1+\zeta} + q. \tag{37}$$

This implicit definition, $\zeta = \psi' C \psi$, is resolved using the formula $(\varepsilon I + x\psi\psi')^{-1} = \varepsilon^{-1}[I - \frac{x\psi\psi'}{\varepsilon + x\psi'\psi}]$. It gives the equation $x = \frac{-\bar{\nu}(\rho+x)}{\rho+x+1} + q$, that converts into the quadratic one for x, $x^2 + \underbrace{[\rho + 1 - q + \bar{\nu}]}_{b} x - \underbrace{[\rho\,(-\bar{\nu}+q) + q]}_{c} = 0$. For ν sufficiently close to 2 (from right), we get $c > 0$ and the equation has the unique positive (meaningful) solution $x = 0.5\left(-b + \sqrt{b^2 + 4c}\right)$.

The found form of the updating of the pre-prior covariance factor $\bar{C} = \varepsilon^{-1} I$ (37) implies that the fictitious regression vector corresponding to the $\tau_{\mathcal{K}}$-th part of the knowledge item is simply $\psi_{\tau_{\mathcal{K}}} = \sqrt{x}\psi$. The derived formula for $\hat{\theta}$ (36) is obtained if we take fictitious output $d_{\tau_{\mathcal{K}}} = \frac{\hat{d}}{\sqrt{x}}(\rho + x)$.

The least-squares remainder, Proposition 4.1, that corresponds to this updating has the value ${}^{\lfloor d}D = {}^{\lfloor d}\varepsilon + \frac{\hat{d}^2\rho(\rho+x)}{x}$. At the same time, the estimate of the noise variance meeting the given restrictions has the form $\hat{r} = \frac{{}^{\lfloor d}D}{\nu-2} = \frac{{}^{\lfloor d}r(\rho+x)}{1+\rho+x}$. Thus, the found results can be interpreted as updating by the fictitious data vector $\Psi_{\tau_{\mathcal{K}}} = [d_{\tau_{\mathcal{K}}}, \psi'_{\tau_{\mathcal{K}}}]'$ iff we take

$$\nu_{\tau_{\mathcal{K}}} - 2 = (1 + \rho + x)\left(\frac{\beta}{\rho + x} + \frac{\alpha\rho}{x}\right).$$

Inserting the relationship between x and q from (37), we can express the searched $\nu_{\tau_{\mathcal{K}}} - 2$ occurring in the above equation as a function of x. It gives a

third order algebraic equation for x. It has always real solution. Standard but lengthy analysis shows uniqueness of the solution in the meaningful domain. □

4.4 Practical examples of prior knowledge

Here, we list common prior pieces knowledge available and ways of the data vectors $\Psi'_i = \left[\hat{d}_i, \psi'_i\right]$ construction (fixed subscripts $\tau_{\mathcal{K}}, \mathcal{S}$ are still suppressed). The multi-variate data items d_t and the common case of the state in the phase form $[d'_{t-1}, \ldots, d'_{t-\mathring{\delta}}, 1]$ of the order $\mathring{\delta}$ are considered. The structure of the data vector is described by the index i, pointing to i-th output channel, and by the list l_i of indices (j, δ) with $j \in \{1, \ldots, \mathring{d}\}$ and $\delta \in \{0, \ldots, \mathring{\delta}\}$. The indices say that the data on the j-th channel $d_{j;t-\delta}$ are in the constructed regression vector ψ_i.

In all cases discussed below, the *entries of* ψ_i that are *not explicitly mentioned are set to zero.*

The knowledge of the *static gain* $\hat{d}_i = \hat{g}$ of the i-th channel on stimulus from the j-th channel is expressed by setting $d_{i;t-\delta} = \hat{g}$ and $d_{j;t-\delta} = 1$ for all delays δ in the list l_i.

It is shown in [21] that the knowledge of a point of the *frequency response* on the j-th channel – given by the module $\hat{a}(\omega)$ and phase $\phi(\omega)$ shift at frequency ω – is determined by a pair of data vectors with $d_{i;\delta} = \hat{a}(\omega)\sin(\phi(\omega)+\delta\omega)$, $d_{j;\delta} = \sin(\delta\omega)$ and $d_{i;\delta} = \hat{a}(\omega)\cos(\phi(\omega)+\delta\omega)$, $d_{j;\delta} = \sin(\delta\omega)$. The range of $a(\omega) = [\underline{a}(\omega), \overline{a}(\omega)]$ can often be well specified. The uncertainty in the phase $\phi(\omega)$ is simply reflected by considering a relatively coarse grid within the uncertainty range and processing each case as an individual data item. The subsequent merging (15) cares about the proper weighting.

Knowledge of *cut-off frequency* is a special case of frequency knowledge with practically zero amplitude, i.e. the amplitude range is given by the point estimate of the standard deviation of the noise. Introduction of this knowledge for several frequencies higher than the cut-off one excludes isolated pass through the zero level. Again, it generates several competitive knowledge items balanced by the merging.

The knowledge of a range of the *dominant time constant* is implemented by modelling lower and upper envelope on the impulse response generated by the first order models with time constants equal to the specified bounds on the time constant. Data are filled from the average trajectory into Ψ_i while distance of envelopes determine the variance ${}^{\lfloor d}r$.

Envelope of measured data, obtained from a periodic measurement, is handled in the same way.

Smoothness of the step response [22] can be respected by enforcing its second order difference to be close to zero.

Note, that the lengths of the samples in "simulated"responses have to be limited so that stationary values are not repeated too much. Otherwise, the assumption on internal consistency, i.e. applicability of the Bayes rule, would be violated.

4.5 Overall algorithm for normal ARX model

Here, we put the algorithmic element together for the normal ARX model. The recommended options correspond with the pre-processed data $d(\mathring{t})$ without outliers, suppressed measurement noise and *data scaled* so that their means are approximately zero and variances are about one. The evaluation is organized so that computation resources are preserved as much as possible.

The explicit reference is made here to the channel treated (subscript $_i$).

Algorithm 4.1 (Structure estimation with prior pdf)
Initial phase

- *Select the grids on:*
  forgetting factors $\{\lambda\}$*, used for processing of internally consistent data blocks,*
  phases $\{\phi(\omega)\}$ *that complete definitions of the frequency response,*
  frequencies $\{\omega_c\}$ *that guarantee that frequency response is close to zero above the cut-off frequency.*

- *Select the number of repetitive starts in the nested structure estimation algorithm SEN 2.1.*

- *Select the order* $\mathring{\delta}_{\mathcal{R}}$ *of the richest data vector* $\Psi_{\mathcal{R};t} = [d'_t, \ldots, d'_{t-\mathring{\delta}_{\mathcal{R}}}, 1]'$ *that includes all potential entries when predicting all modelled entries* $d_{i;t},\ i = 1, \ldots, \mathring{d}$*, of the data item* d_t*.*

- *Specify statistics* $\bar{L}_{\mathcal{R}} = I,\ \bar{D}_{\mathcal{R}} = \mathrm{diag}[\,\lfloor^d\varepsilon, \varepsilon\overbrace{[1, \ldots, 1]}^{\mathring{\psi}_{\mathcal{R}}}]$ *and* $\bar{\nu}$ *determining the flat pre-prior pdf on the richest possible parameterization.*

  Requirements on finiteness of the a priori assigned expectation of r *and need for a flat pdf* $\bar{f}(r)$ *hint to use* $\bar{\nu} = 3$*, see (27). For this choice,* $\lfloor^d\varepsilon$ *is the variance of the unpredictable part of the modelled data. It is sufficient to consider a few categories of the noise-to-signal ratio.*

For instance, the values (0.1%, 1%, 5%, 10%, 50%) correspond with $\lfloor^{d}\varepsilon = (1e-6, 1e-4, 2.5e-3, 1e-2, 0.25)$.

For stable systems, that are predominantly treated, the auto-regression coefficients do not cross the value $\gamma = \begin{pmatrix} \mathring{\delta}_{\mathcal{R}} \\ 0.5\mathring{\delta}_{\mathcal{R}} \end{pmatrix}$. *The regression coefficients are, as a rule, much smaller. Properties of the GiW pdf imply that* $\varepsilon \approx 25\, \lfloor^{d}\varepsilon/\gamma^2$ *is an appropriate conservative option.*

- *Select the number* $\mathring{\mathcal{S}}$*, say several tens, of competitive structures to be refined by using prior information.*

- *Use the available real data* $d_t, t = 1, \ldots, \mathring{t}$ *to update* $L'DL$ *decomposition of the* increment *of the extended information matrix corresponding to the richest data vectors* $\Psi_{\mathcal{R};t}$*, see (13),*

$$\lfloor^{\Delta}L'_{\mathcal{R};t}\, \lfloor^{\Delta}D_{\mathcal{R};t}\, \lfloor^{\Delta}L_{\mathcal{R};t} = \begin{bmatrix} \lfloor^{\Delta}L_{\mathcal{R};t-1} \\ \Psi'_{\mathcal{R};t} \end{bmatrix}' \mathrm{diag} \begin{bmatrix} \lfloor^{\Delta}D_{\mathcal{R};t-1} \\ 1 \end{bmatrix} \begin{bmatrix} \lfloor^{\Delta}L_{\mathcal{R};t-1} \\ \Psi'_{\mathcal{R};t} \end{bmatrix}$$

$$\lfloor^{\Delta}\nu_{\mathcal{R};t} = \lfloor^{\Delta}\nu_{\mathcal{R};t-1} + 1, \quad \textit{with} \quad \lfloor^{\Delta}L_{\mathcal{R};0} = I,\ \lfloor^{\Delta}D_{\mathcal{R};0} = 0,\ \lfloor^{\Delta}\nu_{\mathcal{R};0} = 0.$$

- *Evaluate the* $L'DL$ *decomposition of the extended information matrix corresponding to the richest data vectors* $\Psi_{\mathcal{R};t}$*, i.e. add the statistics of the pre-prior pdf*

$$L'_{\mathcal{R};\mathring{t}} D_{\mathcal{R};\mathring{t}} L_{\mathcal{R};\mathring{t}} = \begin{bmatrix} \lfloor^{\Delta}L'_{\mathcal{R};\mathring{t}} \\ I \end{bmatrix}' \mathrm{diag} \begin{bmatrix} \lfloor^{\Delta}D_{\mathcal{R};\mathring{t}} \\ \bar{D}_{\mathcal{R}} \end{bmatrix} \begin{bmatrix} \lfloor^{\Delta}L_{\mathcal{R};\mathring{t}} \\ I \end{bmatrix}.$$

- *Set the channel index* $i = 1$.

Cycle over indices i of the modelled entries in data records

- *Set the auxiliary description of the structure* $\hat{\mathcal{S}} = \emptyset,\ \hat{\mathcal{L}}_i = -\infty$ *needed for the MAP estimation.*

Structure estimation with nested prior knowledge

- *Select the factors of the pre-prior and posterior extended information matrices,* $\bar{L}_{i\mathcal{R}}$*,* $\bar{D}_{i\mathcal{R}}$*,* $L_{i\mathcal{R};\mathring{t}}$*,* $D_{i\mathcal{R};\mathring{t}}$*, as well as of the increment* $\lfloor^{\Delta}L_{i\mathcal{R};\mathring{t}}$*,* $\lfloor^{\Delta}D_{i\mathcal{R};\mathring{t}}$*, corresponding to the* i*-th predicted data entry* $d_{i;t}$ *and the richest regression vector* $\psi_{i\mathcal{R};t}$*.*

  They are nested in $L_{\mathcal{R};\mathring{t}}$*,* $D_{\mathcal{R};\mathring{t}}$*,* $\bar{L}$*,* $\bar{D}$*, and* $\lfloor^{\Delta}L_{\mathcal{R};\mathring{t}}$*,* $\lfloor^{\Delta}D_{\mathcal{R};\mathring{t}}$*. The* $L'DL$ *decompositions spoiled by this selection have to be recovered using the rank-one corrections [20].*

- *Apply the algorithm SEN (2.1) giving $L'DL$ factors of the pre-prior extended information matrices $\bar{L}_{i\mathcal{S}}$, $\bar{D}_{i\mathcal{S}}$ and their data-dependent increments ${}^{\lfloor\Delta}L_{i\mathcal{S};\mathring{t}}$, ${}^{\lfloor\Delta}D_{\mathcal{S};\mathring{t}}$. They correspond to the most probable structures $\mathcal{S} \in \mathcal{S}^*$ found when just the nested pre-prior knowledge is used.*

- *Select $\mathcal{S} \in \mathcal{S}^*$.*

Inclusion of prior knowledge for promising structures

- *Select a knowledge item $\mathcal{K}$ in the list $\mathcal{K}_i^* = \{1, \ldots, \mathring{\mathcal{K}}_i\}$ available for the i-th channel.*

- *Set the normalization factor needed in merging $s = 0$.*

  Processing of knowledge items

- *Do* if *the individual knowledge item $\mathcal{K}$* has to be converted into fictitious data vectors

  - *Set $L_{i\mathcal{S};0} = \bar{L}_{i\mathcal{S}}$, $D_{i\mathcal{S};0} = \bar{D}_{i\mathcal{S}}$, $\nu_{i\mathcal{S};0} = \bar{\nu}$.*
  - *For $\tau_{\mathcal{K}} = 1, \ldots, \mathring{\tau}_{\mathcal{K}}$*
    - *Generate data reflecting $\tau_{\mathcal{K}}$-th part of the knowledge item $\mathcal{K}$ given by $\hat{d}_{i\mathcal{S};\tau_{\mathcal{K}}}$, ${}^{\lfloor d}r_{i\mathcal{S};\tau_{\mathcal{K}}}$ and $\psi_{i\mathcal{S};\tau_{\mathcal{K}}}$, cf. (29).*
    - *Evaluate fictitious data vectors $\Psi_{i\mathcal{S};\tau_{\mathcal{K}}}$ and its $\nu_{i\mathcal{S};\tau_{\mathcal{K}}}$, cf. (33),*

$$\Psi'_{i\mathcal{S};\tau_{\mathcal{K}}} = \left[\hat{d}_{i\mathcal{S};\tau_{\mathcal{K}}}\left(\frac{\rho_{i\mathcal{S};\tau_{\mathcal{K}}}}{\sqrt{x_{i\mathcal{S};\tau_{\mathcal{K}}}}} + \sqrt{x_{i\mathcal{S};\tau_{\mathcal{K}}}}\right), \sqrt{x_{i\mathcal{S};\tau_{\mathcal{K}}}}\psi'_{i\mathcal{S};\tau_{\mathcal{K}}}\right].$$

    - *Update*

$$\begin{aligned} L'_{i\mathcal{S};\tau_{\mathcal{K}}} D_{i\mathcal{S};\tau_{\mathcal{K}}} L_{i\mathcal{S};\tau_{\mathcal{K}}} &= \begin{bmatrix} L_{i\mathcal{S};\tau_{\mathcal{K}}-1} \\ \Psi'_{i\mathcal{S};\tau_{\mathcal{K}}} \end{bmatrix}' \operatorname{diag}\begin{bmatrix} D_{i\mathcal{S};\tau_{\mathcal{K}}-1} \\ 1 \end{bmatrix} \begin{bmatrix} L_{i\mathcal{S}\tau_{\mathcal{K}}-1} \\ \Psi'_{i\mathcal{S};\tau_{\mathcal{K}}} \end{bmatrix} \\ \nu_{i\mathcal{S};\tau_{\mathcal{K}}} &= \nu_{\mathcal{S};\tau_{\mathcal{K}}-1} + {}^{\lfloor\Delta}\nu_{i\mathcal{S};\tau_{\mathcal{K}}}. \end{aligned}$$

- *Run the estimation with the stabilized forgetting for the selected forgetting factors λ and with the alternative pdf given by $\bar{L}_{i\mathcal{S}}$, $\bar{D}_{i\mathcal{S}}$, $\bar{\nu}$* if *the knowledge item $\mathcal{K}$ is formed by an* internally consistent data block.

  Store the results into ${}^{\lfloor\Delta}L_{i\mathcal{S};\mathring{\tau}_{\mathcal{K}}}$, ${}^{\lfloor\Delta}D_{i\mathcal{S};\mathring{\tau}_{\mathcal{K}}}$ and ${}^{\lfloor\Delta}\nu_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}}$

- *Select the flattening factor $\Lambda_{i\mathcal{SK}}$ maximizing the partial likelihood*

$$\mathcal{L}_i(d(\mathring{t}),\mathcal{S},\mathcal{K}) = \frac{\mathcal{I}\left(V_{i\mathcal{SK};\mathring{t}},\mathring{t}+\Lambda_{i\mathcal{SK}}\,{}^{\lfloor\Delta}\nu_{i\mathcal{S};\mathring{\tau}_\mathcal{K}}+\bar{\nu}_{i\mathcal{S}}\right)}{\mathcal{I}\left(\Lambda_{i\mathcal{SK}}\,{}^{\lfloor\Delta}L'_{i\mathcal{S};\mathring{\tau}_\mathcal{K}}\,{}^{\lfloor\Delta}D_{i\mathcal{S};\mathring{\tau}_\mathcal{K}}\,{}^{\lfloor\Delta}L_{i\mathcal{S};\mathring{\tau}_\mathcal{K}}+\bar{L}'_{i\mathcal{S}}\bar{D}_{i\mathcal{S}}\bar{L}_{i\mathcal{S}},\Lambda_{i\mathcal{SK}}\,{}^{\lfloor\Delta}\nu_{i\mathcal{S};\mathring{\tau}_\mathcal{K}}+\bar{\nu}_{i\mathcal{S}}\right)}$$

$$V_{i\mathcal{SK};\mathring{t}} = {}^{\lfloor\Delta}L'_{i\mathcal{S};\mathring{t}}\,{}^{\lfloor\Delta}D_{i\mathcal{S};\mathring{t}}\,{}^{\lfloor\Delta}L_{i\mathcal{S};\mathring{t}}+\Lambda_{i\mathcal{SK}}\,{}^{\lfloor\Delta}L'_{i\mathcal{S};\mathring{\tau}_\mathcal{K}}\,{}^{\lfloor\Delta}D_{i\mathcal{S};\mathring{\tau}_\mathcal{K}}\,{}^{\lfloor\Delta}L_{i\mathcal{S};\mathring{\tau}_\mathcal{K}}+\bar{L}'_{i\mathcal{S}}\bar{D}_{i\mathcal{S}}\bar{L}_{i\mathcal{S}}$$

$$s = s+\mathcal{L}_i(d(\mathring{t}),\mathcal{S},\mathcal{K})$$

Notice that ${}^{\lfloor\Delta}L'_{i\mathcal{S};\mathring{t}}$, ${}^{\lfloor\Delta}D_{i\mathcal{S};\mathring{t}}$ and $\bar{L}'_{i\mathcal{S}}, \bar{D}_{i\mathcal{S}}, \bar{\nu}_{i\mathcal{S}}$ depend only on the structural indices $i,\mathcal{S}$ and not on $\mathcal{K}$.

- *Take a new knowledge item $\mathcal{K}$, if the list $S\mathcal{K}_i$ is not exhausted, and go to* Processing of knowledge items. *Otherwise continue.*

- *Set $L_{i\mathcal{SK}(\mathring{\mathcal{K}})} = I,\ D_{i\mathcal{SK}(\mathring{\mathcal{K}})} = \bar{D}_{i\mathcal{S}},\ \nu_{i\mathcal{SK}(\mathring{\mathcal{K}})} = \bar{\nu}_{i\mathcal{S}}$.*

- *Select $\mathcal{K}\in\mathcal{K}_i^*$.* Evaluation of the merger within the structure $\mathcal{S}$

- *Update*

$$\begin{aligned} f_i(\mathcal{K}|d(\mathring{t}),\mathcal{S}) &= \frac{\mathcal{L}_i(d(\mathring{t}),\mathcal{S},\mathcal{K})}{s},\quad L'_{i\mathcal{SK}(\mathring{\mathcal{K}})}D_{i\mathcal{SK}(\mathring{\mathcal{K}})}L_{i\mathcal{SK}(\mathring{\mathcal{K}})} \\ &= L'_{i\mathcal{SK}(\mathring{\mathcal{K}})}D_{i\mathcal{SK}(\mathring{\mathcal{K}})}L_{i\mathcal{SK}(\mathring{\mathcal{K}})}+f_i(\mathcal{K}|d(\mathring{t}),\mathcal{S})\,{}^{\lfloor\Delta}L'_{i\mathcal{S};;\mathring{\tau}_\mathcal{K}}\,{}^{\lfloor\Delta}D_{i\mathcal{S};\mathring{\tau}_\mathcal{K}}\,{}^{\lfloor\Delta}L_{i\mathcal{S};\mathring{\tau}_\mathcal{K}} \\ \nu_{i\mathcal{SK}(\mathring{\mathcal{K}})} &= \nu_{i\mathcal{SK}(\mathring{\mathcal{K}})}+f_i(\mathcal{K}|d(\mathring{t}),\mathcal{S})\,{}^{\lfloor\Delta}\nu_{i\mathcal{S};\mathring{\tau}_\mathcal{K}}. \end{aligned}$$

- *Select a new $\mathcal{K}$ if their list $\{1,\ldots,\mathring{\mathcal{K}}_i\}$ is not exhausted and go to* Evaluation of the merger within the structure $\mathcal{S}$. *Otherwise continue.*

- *Evaluate partial likelihood expressing assigned to the structure $\mathcal{S}$*

$$\mathcal{L}_i(d(\mathring{t}),\mathcal{S}) = \frac{\mathcal{I}\left({}^{\lfloor\Delta}L'_{i\mathcal{S};\mathring{t}}\,{}^{\lfloor\Delta}D_{i\mathcal{S};\mathring{t}}\,{}^{\lfloor\Delta}L_{i\mathcal{S};\mathring{t}}+L'_{i\mathcal{SK}(\mathring{\mathcal{K}})}D_{i\mathcal{SK}(\mathring{\mathcal{K}})}L_{i\mathcal{SK}(\mathring{\mathcal{K}})},\mathring{t}+\nu_{i\mathcal{SK}(\mathring{\mathcal{K}})}\right)}{\mathcal{I}\left(L'_{i\mathcal{SK}(\mathring{\mathcal{K}})}D_{i\mathcal{SK}(\mathring{\mathcal{K}})}L_{i\mathcal{SK}(\mathring{\mathcal{K}})},\nu_{i\mathcal{SK}(\mathring{\mathcal{K}})}\right)}.$$

- *Set $\hat{\mathcal{S}}=\mathcal{S},\ \hat{\mathcal{L}}_i=\mathcal{L}_i(d(\mathring{t}),\mathcal{S})$ and store the statistics corresponding to the posterior pdf ${}^{\lfloor\Delta}L'_{i\hat{\mathcal{S}};\mathring{t}}\,{}^{\lfloor\Delta}D_{i\hat{\mathcal{S}};\mathring{t}}\,{}^{\lfloor\Delta}L_{i\hat{\mathcal{S}};\mathring{t}}+L'_{i\hat{\mathcal{S}}\mathcal{K}(\mathring{\mathcal{K}})}D_{i\hat{\mathcal{S}}\mathcal{K}(\mathring{\mathcal{K}})}L_{i\hat{\mathcal{S}}\mathcal{K}(\mathring{\mathcal{K}})}$ and $\mathring{t}+\nu_{i\hat{\mathcal{S}}}$ if $\mathcal{L}_i(d(\mathring{t}),\mathcal{S})>\hat{\mathcal{L}}_i$.*

- *Select a new structure $\mathcal{S}$ if the list $\mathcal{S}^*$ of the most probable ones is not exhausted and go to the* Inclusion of prior knowledge for promising structures. *Otherwise continue.*

- *Offer the structure $\hat{\mathcal{S}}$ as the recommended one for i-th channel with the corresponding stored posterior statistics.*

- *Increase i and go to* Cycle over indices i of modelled data entries *if $i \leq \mathring{d}$. Otherwise stop.*

The algorithm provides also, till unavailable, estimate of the best order in factorization (2). It must, however, be complemented by a check against incorrect dependence loops.

It is important that the algorithm can cope with problems having $\mathring{d}$ of the order several tens.

5 Illustrative example

Contribution of prior knowledge to structure estimation results is illustrated on single-input single-output simulated system. It corresponds to two-dimensional $d_t = [y_t, u_t]'$. The system input u_t is generated as white normal noise with variance 0.3. The modelled system output y_t is *simulated* by the ARX model determined by the "objective parameter" ${}^{\lfloor o}\Theta = \left[{}^{\lfloor o}\theta,\, {}^{\lfloor o}r \right]$. It is usually written in form of the difference equation driven by the normalized white normal noise $e_t \sim \mathcal{N}_{e_t}(0,1)$

$$\begin{aligned} y_t &= {}^{\lfloor o}a_1 y_{t-1} + {}^{\lfloor o}a_2 y_{t-2} + {}^{\lfloor o}b_0 u_t + {}^{\lfloor o}b_1 u_{t-1} + \sqrt{{}^{\lfloor o}r}\, e_t \\ &= \underbrace{[1.81, 0.8187, 0.0438, 0.00468]}_{{}^{\lfloor o}\theta'} \underbrace{[y_{t-1}, y_{t-2}, u_t, u_{t-1}]'}_{\psi_t} + \underbrace{\sqrt{0.001}}_{\sqrt{{}^{\lfloor o}r}}\, e_t \\ &\Leftrightarrow \quad f\left(y_t | u_t, d(t-1), {}^{\lfloor o}\Theta\right) = f\left(y_t | \psi_t, {}^{\lfloor o}\Theta\right) = \mathcal{N}_{y_t}\left({}^{\lfloor o}\theta'\psi_t, {}^{\lfloor o}r\right). \end{aligned}$$

The "real" data $d(\mathring{t}) = d(300)$ were "measured" on this system (MATLAB simulation with the seed of normal generator equal to 1).

The SEN algorithm 2.1 was applied with the richest structure of the phase form given by the order $\mathring{\delta}_{\mathcal{R}}$ and the recommended nested pre-prior pdf were used. The number of restarts was 10 and $\mathring{\mathcal{S}} = 10$ of the best structures were stored giving the significant entries marked by * and the related posterior pfs

S	a_1	a_2	a_3	b_0	b_1	b_2	$f(\mathcal{S}\|d(300))$
1	*	*					0.6935
2	*	*			*		0.2572
3	*	*	*				0.0185
4	*	*		*			0.0111
5	*	*		*	*		0.0077
⋮			⋮				⋮

The correct structure was assigned the 5th largest probability. Inclusion of the prior knowledge $\mathcal{K} = 1 \equiv$ static gain is in range [1.0310 1.0518] made the largest pf $f(\mathcal{S} = 2|d(300), \mathcal{K} = 1)$ corresponding to the second order model with non-zero b_1.

... ...

The following analysis is based on data sample only and on the prior knowledge about the system:

1. the system static gain is in the range

2. the dominant time constants in ranges:
   $\lambda_1 = [0.8254 \quad 0.8421]; \quad \lambda_2 = [1.1237 \quad 1.3734]$

The conclusion are:

- the SEN algorithm of nested structure estimation does not discover the dependency of output on input;

- the "true" structure (appears among "best" structures with a low probability (the last in the table above).

In this case, dependency of output on input is discovered. The prior knowledge of static gain carries the information about dependency but not about dynamics - it can explain the absence of the b_0 in the result.

Now, the knowledge of the time constants is applied. The prior knowledge carries the information about dependency and system dynamics. This time, the "true" structure a_1, a_2, b_0, b_1 is discovered.

A range of prior knowledge and their combinations were tested in the task of the structure estimation. The experiments show that use of prior pdf is not a magic tool that can solve any problem – it only slightly enlarges the range of successful solutions.

CCC>> MK: information on some improvement to former version, some information on weights <<CCC

6 Concluding remarks

Significance of the inclusion of prior knowledge into black-box models is still underestimated. The theory and algorithms presented in this paper solve this problem to a significant extent. The available practical experience confirms that the use of even vague knowledge may decide on the success or failure of structure estimation and consequently on the success or failure of the controller design.

A wider and more precise use of the prior knowledge is especially important in the context of the prior design of the advanced controller with incomplete knowledge [15].

The theory and algorithms have been elaborated for the LQG-type-design set up. The same problems are met out of this class and the adopted methodology may serve to them, too.

The problem of the joint processing of the prior knowledge and structure estimation is solved here for the exponential family but elaborated for the ARX models only. It determines a direction of a further development: the controlled-Markov-chain case should be elaborated, too. It would strengthen dynamic modelling of systems with dynamic discrete data.

Acknowledgements

This research has been partially supported by grants GA AVČR S1075102, S1075351 and GAČR 102/03/0049.

References

[1] K.J. Astrom and B. Wittenmark, *Adaptive Control*, Addison-Wesley Publishing Company, Reading, Massachusetts, 1989.

[2] P.E. Wellstead and M.B. Zarrop, *Self-tuning Systems*, John Wiley & Sons, Chichester, 1991.

[3] E. Mosca, *Optimal, Predictive, and Adaptive Control*, Prentice Hall, 1994.

[4] M. Kárný, A. Halousková, J. Böhm, R. Kulhavý, and P. Nedoma, "Design of linear quadratic adaptive control: Theory and algorithms for practice", *Kybernetika*, vol. 21, 1985, Supplement to No. 3, 4 ,5, 6.

[5] A. Johnson, "LQG applications in the process industries", *Chemical Engineering Practice*, vol. 48, no. 16, pp. 2829–2838, 1993.

[6] M. Elbelkacemi, A. Lachhab, M. Limouri, D. Dahhou, and A. Essaid, "Adaptive control of a water supply system", *Control Engineering Practice*, vol. 9, no. 3, pp. 343–349, 2001.

[7] D.W. Clarke, *Advances in Model-Based Predictive Control*, Oxford University Press, Oxford, 1994.

[8] M. Kárný and A. Halousková, "Implementing LQG adaptive control: a CAD approach", in *Preprints of the 9th IFAC/IFORS Symposium on Identification and System Parameter Estimation*, pp. 1585–1590. AKA PRINT Nyomdaipari, Budapest, 1991.

[9] M. Kárný and A. Halousková, "Pre-tuning of self-tuners", in *Advances in Model-Based Predictive Control*, D. Clarke, Ed., pp. 333–343. Oxford University Press, Oxford, 1994.

[10] J. Bůcha, M. Kárný, P. Nedoma, J. Böhm, and J. Rojíček, "Designer 2000 project", in *International Conference on Control '98*, London, September 1998, pp. 1450–1455, IEE.

[11] M. Kárný and R. Kulhavý, "Structure determination of regression-type models for adaptive prediction and control", in *Bayesian Analysis of Time Series and Dynamic Models*, J.C. Spall, Ed. Marcel Dekker, New York, 1988, chapter 12.

[12] V. Peterka, "Bayesian system identification", in *Trends and Progress in System Identification*, P. Eykhoff, Ed., pp. 239–304. Pergamon Press, Oxford, 1981.

[13] R. Kulhavý and M. B. Zarrop, "On general concept of forgetting", *International Journal of Control*, vol. 58, no. 4, pp. 905–924, 1993.

[14] M. Kárný, T. Jeníček, and W. Ottenheimer, "Contribution to prior tuning of LQG selftuners", *Kybernetika*, vol. 26, no. 2, pp. 107–121, 1990.

[15] M. Novák and J. Böhm, "Automated multivariable adaptive controller design", *IEE Trans. on Control theory and its applications*, 2003, This issue.

[16] J.M. Bernardo and A.F.M. Smith, *Bayesian theory*, John Wiley & Sons, Chichester, New York, Brisbane, Toronto, Singapore, 1997, 2nd edition.

[17] P. Nedoma, M. Kárný, and J. Böhm, "Software tools for use of prior knowledge in design of LQG adaptive controllers", in *Preprints of the IFAC Workshop on Adaptive Systems in Control and Signal Processing*, Glasgow, August 1998, pp. 425–429, IFAC.

[18] S. Kullback and R. Leibler, "On information and sufficiency", *Annals of Mathematical Statistics*, vol. 22, pp. 79–87, 1951.

[19] A. Zellner, *An Introduction to Bayesian Inference in Econometrics*, J. Wiley, New York, 1976.

[20] G.J. Bierman, *Factorization Methods for Discrete Sequential Estimation*, Academic Press, New York, 1977.

[21] N. Khaylova, "Exploitation of prior knowledge in adaptive control design", Tech. Rep., FAV ZČU, University of West Bohemia, Pilsen, Czech Republic, 2001, PhD Thesis.

[22] M. Kárný, "Quantification of prior knowledge about global characteristics of linear normal model", *Kybernetika*, vol. 20, no. 5, pp. 376–385, 1984.