Prior information in structure estimation

M. Kárný, P. Nedoma, N. Khaylova and L. Pavelková

Abstract: The problem of a joint quantification of prior knowledge and structure estimation is solved within the dynamic exponential family of models. The result is elaborated for normal controlled autoregressive models and illustrated on a simulated example. The problem arose as a substantial ingredient of the automatic commission of adaptive controllers, described in a companion paper. From this perspective, work also serves as an illustration of technology used within this broader context.

Principal symbols

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

1 Introduction

Adaptive LQG controllers that minimise approximately quadratic criteria using recursively estimated linear gaussian models have become a standard in the academic environment [1–3]. The version with controlled autoregressive models (ARX) is among the most successful ones as confirmed by their full-scale applications [4–6]. At the same time, their potential is far from being adequately exploited. One of the main reasons for this is the requirement for expensive manpower-intensive commissioning of such controllers in many practical contexts. This fact, that also applies to other adaptive controllers like GPC, MUSMAR etc. [3, 7], stimulated an extensive project [8–10]. It aims at creating a complete computer support of the commission. At present, a full solution for single-input single-output systems is implemented in the software system Designer. It covers

(i) data preprocessing
(ii) quantification of prior information [11]
(iii) selection of the model structure [12]
(iv) offline estimation [13] that serves for initialisation of the online estimation as well as an alternative model needed for parameter tracking [14] by the stabilised forgetting
(v) estimation of the forgetting factor
(vi) tuning of kernels in the optimised quadratic loss so that user's aims and restrictions are met; this also provides offline prediction of closed-loop behaviour [15]

This paper proposes a correct combination of steps (ii) and (iii). The summary of steps (ii)–(iv) also serves the companion paper [16], which provides a solution of the most difficult step (vi). Thus, in addition to particular improvements, these two papers report on the technology used within Designer.

The paper is relatively self-containing. Readers less familiar with the bayesian treatment adopted are referred to [13].

2 Problem formulation and solution

The essence of the addressed problem and its conceptual solution are described after technical preliminaries. The ARX model used belongs to the exponential family. Its handling is very simple when its member-specific details are left aside; this explains why the basic ideas are presented in terms of this family.

2.1 Basic notation and operations

The following notation is used throughout the text.

PDFs are distinguished by the identifiers in their arguments. No formal distinction is made between a random

IEE Proceedings online no. 20030861
doi: 10.1049/ip-cta:20030861
Paper first received 7th March 2003 and in revised form 28th July 2003

The authors are with the Institute of Information Theory and Automation, Pod vodárenskou věží 4, 182 08 Prague 8, POB 18, Czech Republic

variable, its realisation and argument of PDFs. The correct meaning follows from the context.

The following elementary operations of PDFs are used [13]:

$$\begin{aligned}
\text{normalisation}\quad & \int f(a)\,da \equiv \int_{a\in a^*} f(a)\,da = 1 \\
\text{chain rule}\quad & f(a,b|c) = f(a|b,c)f(b|c) \\
\text{marginalisation}\quad & f(a|c) = \int f(a,b|c)\,db \\
\text{Bayes' rule}\quad & f(a|b,c) = \frac{f(b|a,c)f(a|c)}{\int f(b|a,c)f(a|c)\,da} \\
& \propto f(b|a,c)f(a|c)
\end{aligned} \tag{1}$$

The bayesian paradigm that 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 realisation of any variable that is at its disposal.

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

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

2.2 Estimation and prediction in exponential family

The parameterised 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 ith parameterised model belongs to the (dynamic) exponential family IFF it can be written in the form

$$\begin{aligned}
f(d_{i,t}|d_{i+1;t},\ldots,d_{\mathring{d};t},d(t-1),\Theta_{i\mathcal{S}}) &= f(d_{i;t}|\boldsymbol{\psi}_{i\mathcal{S};t},\Theta_{i\mathcal{S}}) \\
&= A(\Theta_{i\mathcal{S}})\exp[\langle B(\boldsymbol{\Psi}_{i\mathcal{S};t}), C(\Theta_{i\mathcal{S}})\rangle]
\end{aligned} \tag{3}$$

where $\Psi'_{i\mathcal{S};t} \equiv [d_{i;t}, \psi'_{i\mathcal{S};t}]$ is the 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 nonnegative scalar function defined on $\Theta^*_{i\mathcal{S}}$ $\langle\cdot,\cdot\rangle$ is a functional that is linear in the first argument and $B(\cdot)$, $C(\cdot)$ 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.

The practical significance of the exponential family becomes obvious when summarising the corresponding estimation and prediction [13], as follows.

Proposition 2.1 (Estimation and prediction in exponential family): Let the parameterised model have the form (3) and the parameters $\Theta_{\mathcal{S}} \equiv \Theta(\mathring{\mathcal{S}})$ be *a priori* independent, i.e. $f(\Theta_{\mathcal{S}}) = \Pi_{i=1}^{\mathring{d}} f(\Theta_{i\mathcal{S}})$. Moreover, let the conjugate prior PDFs $f(\Theta_{i\mathcal{S}})$ [17] be used

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

The conjugate PDFs have the parameterised-model-induced functional form $\mathcal{G}$. They are given 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

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

$$\begin{aligned}
\mathcal{I}(V_{i\mathcal{S};t},\nu_{i\mathcal{S};t}) \equiv \int & A^{\nu_{i\mathcal{S};t}}(\Theta_{i\mathcal{S}})\exp[\langle V_{i\mathcal{S};t}, C(\Theta_{i\mathcal{S}})\rangle] \\
& \times \chi_{\Theta^*_{i\mathcal{S}}}(\Theta_{i\mathcal{S}})\,d\Theta_{i\mathcal{S}}
\end{aligned} \tag{5}$$

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

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

The predictive PDF, modelling evolution of the ith data entry (ith channel), is given by the formula

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

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

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

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

Thus, the estimation and prediction can be performed channel-wise. Focus 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 integral $\mathcal{I}(V_{\mathcal{S}},\nu_{\mathcal{S}})$ has to be evaluated.

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

The influence of particular knowledge items $\mathcal{K} \in \mathcal{K}^* \equiv \{1,\ldots,\mathring{\mathcal{K}}\}$ on descriptive abilities of the adopted model is inspected. 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.

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

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

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

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

This proposition serves for tracking of slow parameter changes using stabilised forgetting [14]. 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 forgetting is applied. The externally supplied alternative PDF ${}^a f(\Theta)$ describes possible changes of estimated parameters before measuring next data. The weight λ, called the 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 ${}^a f(\Theta) \propto 1$. Use of the preprior PDF $\bar{f}(\Theta) \propto \mathcal{G}_\Theta(\bar{V}, \bar{\nu})$ as the alternative PDF ${}^a f(\Theta)$ is wiser. Even a flat PDF $\bar{f}(\Theta)$ respects finite Θ values. It conservatively guarantees that the support of the forgotten posterior PDF does not move on the area containing infinite values of Θ.

The geometric mean of PDFs serves also for finding a representative $f(\Theta \,|\, d(\mathring{t}), \mathcal{K}(\mathring{\mathcal{K}}))$ 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 ascribed to each of them coincides with the posterior probability $f(\mathcal{K} \,|\, d(\mathring{t})) \propto \mathcal{L}(d(\mathring{t}), \mathcal{K})$. The representative $f(\Theta \,|\, d(\mathring{t}), \mathcal{K}(\mathring{\mathcal{K}}))$, called the merger and motivated in [11], is chosen as the weighted geometric mean

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

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 entries allowed in 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 $2^{\mathring{\psi}_\mathcal{R}}$ choices. This number is usually excessive and precludes straightforward bayesian structure estimation via comparison of 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 $V_{\mathcal{S};\mathring{t}}$ constitutes the main computational burden related to structure estimation. The current implementation in the system Designer [18] relies on nesting of competitive structures $\mathcal{S}$ within the richest one $\mathcal{R}$. The model, given by the data vectors $\Psi_{\mathcal{S};t}$, is said to be 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 parameterised model with the richest structure belong to the exponential family (3) $f(d_t \,|\, \psi_{\mathcal{R};t}, \Theta_\mathcal{R}) = A(\Theta_\mathcal{R}) \times \exp[\langle B(\Psi_{\mathcal{R};t}), C(\Theta_\mathcal{R}) \rangle]$. 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 $B(\Psi_{\mathcal{S};t}) = N_\mathcal{S}[B(\Psi_{\mathcal{R};t})]$. Assume that

$$V_{\mathcal{S};0} = N_\mathcal{S}[V_{\mathcal{R};0}] \tag{12}$$

Then the V-statistics of the posterior PDFs of both models are nested, i.e. they 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

$$f(\mathcal{S} \,|\, d(t)) \propto \frac{\mathcal{I}(N_\mathcal{S}[V_{\mathcal{R};t}], \nu_{\mathcal{S};t})}{\mathcal{I}(N_\mathcal{S}[V_{\mathcal{R};0}], \nu_{\mathcal{S};0})} f(\mathcal{S}) \; \forall t \in t^* \tag{13}$$

Thus, for the nested models and nested prior statistics, it is sufficient to collect the V-statistic for the richest structure. This helps but only partially. Full evaluation of the PF values (13) on the complete set of $2^{\mathring{\psi}_\mathcal{R}}$ competing structures is still prohibitive. Thus, the maximum *a posteriori* probability (MAP) estimate of $\mathcal{S}$ has to be searched for. The nonnormalised values of the PF (13) evaluated during the search provides useful partial information on highly probable structures. The following conceptual search algorithm is used [12].

Algorithm 2.1 (Structure estimation with nested priors): Initial phase

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

$${}^\Delta V_{\mathcal{R};\mathring{t}} \equiv \sum_{t=1}^{\mathring{t}} B(\Psi_{\mathcal{R};t}) \tag{14}$$

- Select the prior statistic $V_{\mathcal{R};0}$ so that $V_{\mathcal{S};0}$ are nested in it $\forall \mathcal{S} \in \mathcal{S}^*$ (12).
- Specify prior PF $f(\mathcal{S})$ of competitive structures, typically uniform.

Search phase

This is run until the prespecified number of restarts (needed for global maximisation) is reached.

1. Initialise the current guess of the structure.
   Empty, richest and user-specified structures of regression vectors are used. These options are complemented by guesses 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 neighbourhood of the current guess of the structure, i.e. maximise within it the posterior partial likelihood

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

   The neighbourhood 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 defined in the last two categories.

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

end of Do

2.4 Quantification of prior knowledge

The quality of parameter and structure estimation is sensitive to the amount of information actually available in the learning data. Thus, any available prior knowledge must be used. In the bayesian set-up, it is fed through the prior PDF. Here, how the latter can be constructed is outlined. The following circumstances are specific to technological applications:

⊕ Groups of widely accessible knowledge types exist.
⊕ Experimental data $d(\mathring{t})$ measured on the modelled system are available.
⊕ Prior PDFs are restricted to being conjugate as so the prior knowledge is to be translated into values of the prior statistics V, ν.
⊖ The person feeding the prior knowledge does not care about the probabilistic tool-set used.
⊖ No supervisor for knowledge elicitation and judgement of expert competence is available.
⊖ Knowledge items processed are expected to be repetitive, not fully consistent and differing in precision and nature. Mutual dependence between knowledge items is undefined.

The following quantification algorithm respects these conditions [11].

Algorithm 2.2 (Quantification of prior knowledge): Initialisation phase

- Select functional form of the *i*th parameterised model (2) of a fixed structure $\mathcal{S}$ in the exponential family under consideration.
- Collect the real-data-dependent increment ${}^{\Delta}V_{\mathcal{S};\mathring{t}}$ of the V-statistics according to (14) for $\mathcal{R} = \mathcal{S}$.
- Split the existing knowledge into internally consistent knowledge items (see Section 3.1).
- Select the preprior 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 (preprior) knowledge available.
- Initialise the normalisation factor $s = 0$.

Quantification phase runs for 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 ${}^{\Delta}V_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}}, {}^{\Delta}\nu_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}}$ of the preprior statistics $\bar{V}_{\mathcal{S}}$, $\bar{\nu}_{\mathcal{S}}$ so that $V_{\mathcal{SK};0} = {}^{\Delta}V_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}} + \bar{V}_{\mathcal{S}}$ and $\nu_{\mathcal{SK};0} = {}^{\Delta}\nu_{\mathcal{S};\mathring{\tau}\mathcal{K}} + \bar{\nu}_{\mathcal{S}}$ describe the considered 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 normalisation factor s

$$\mathcal{L}(d(\mathring{t}), \mathcal{S}, \mathcal{K}) = \frac{\mathcal{I}\left({}^{\Delta}V_{\mathcal{SK};\mathring{t}} + V_{\mathcal{SK};0,\mathring{t}} + \nu_{\mathcal{SK};0}\right)}{\mathcal{I}(V_{\mathcal{SK};0}, \nu_{\mathcal{SK};0})},$$
$$s = s + \mathcal{L}(d(\mathring{t}), \mathcal{S}, \mathcal{K}). \tag{15}$$

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

$$f(\Theta_{\mathcal{S}}\,|\,d(\mathring{t}), \mathcal{K}(\mathring{\mathcal{K}})) \propto f(d(\mathring{t})\,|\,\Theta_{\mathcal{S}}) \times \prod_{\mathcal{K}=1}^{\mathring{\mathcal{K}}} [f(\Theta_{\mathcal{S}}\,|\,\mathcal{K})]^{f(\mathcal{K}|d(\mathring{t}),\mathcal{S})} \tag{16}$$

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

it gives

$$f(\Theta_{\mathcal{S}}|d(\mathring{t}),\mathcal{K}(\mathring{\mathcal{K}})) \propto \mathcal{G}_{\Theta_{\mathcal{S}}} \times \underbrace{\left({}^{\Delta}V_{\mathcal{S};\mathring{t}} + \underbrace{\sum_{\mathcal{K}=1}^{\mathring{\mathcal{K}}} f(\mathcal{K}|d(\mathring{t}),\mathcal{S}) V_{\mathcal{SK};0,\mathring{t}}}_{V_{\mathcal{SK}(\mathring{\mathcal{K}});0}} + \underbrace{\sum_{\mathcal{K}=1}^{\mathring{\mathcal{K}}} f(\mathcal{K}|d(\mathring{t}),\mathcal{S}) \nu_{\mathcal{SK};0}}_{\nu_{\mathcal{SK}(\mathring{\mathcal{K}});0}} \right)}_{V_{\mathcal{SK}(\mathring{\mathcal{K}});\mathring{t}}} \underbrace{\qquad}_{\nu_{\mathcal{SK}(\mathring{\mathcal{K}});\mathring{t}}} \tag{17}$$

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 addressed problem stems from the fact that usually the prior 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 algorithm 2.1 cannot be directly used if prior knowledge is to be exploited. This fact was overlooked in the former implementations [19] and caused worse estimation results than we hoped for. Recognising the problem, the remedy is straightforward. The following conceptual algorithm is used.

Algorithm 2.3 (Structure estimation with prior knowledge): Initial phase

- Select the parameterised model in the exponential family and the richest structure of the underlying data vector $\Psi_{\mathcal{R}}$.
- Select the preprior 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 expresses the common knowledge available while requiring that the same knowledge be expressed for all nested structures of interest.

Structure estimation with the nested prior statistics

- Apply algorithm 2.1 to get a preselected number $\mathring{\mathcal{S}}$ of structures $\mathcal{S} \in \mathcal{S}^* \equiv \{1, 2, \ldots, \mathring{\mathcal{S}}\}$ 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.* (17).
2. Evaluate descriptive abilities of the best merger, within the given structure $\mathcal{S}$, *cf.* (16),

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

3. Provide $\hat{\mathcal{S}}\in \text{Arg}\max_{\mathcal{S}^*}\mathcal{L}(d(\mathring{t}),\mathcal{S})f(\mathcal{S})$ 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 online estimation and as the alternative PDF needed in stabilised forgetting.

3 Fictitious data

Here, the construction of the common information basis, i.e. fictitious data, is recalled and refined [11]. It allows one 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 $\mathcal{K}$ the 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, e.g. 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 $f(\Theta_{\mathcal{S}}|\mathcal{K})$ is equal to a flattened version of the posterior PDF $f(\Theta_{\mathcal{S}}|d(\mathring{\tau}_{\mathcal{K}}))\propto\mathcal{G}_{\Theta_{\mathcal{S}}}(V_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}},\nu_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}})$. The handling of such a knowledge item is described in detail.

The posterior PDF is obtained by Bayer's rule applied to the preprior PDF $\bar{f}(\Theta_{\mathcal{S}})\propto\mathcal{G}_{\Theta_{\mathcal{S}}}(\bar{V}_{\mathcal{S}},\bar{\nu}_{\mathcal{S}})$ with stabilised forgetting. Forgetting is used to counteract mismodelling. The preprior PDF is used as the alternative PDF. Thus, only an appropriate forgetting factor λ needs to be chosen. A comparison of partial likelihoods obtained for various forgetting factors serves this purpose. This 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 λs (forming a representative grid defined on λ-domain) as different knowledge pieces. This is done from now on and reference to λ is suppressed.

The nature of the fictitious data blocks implies that the PDFs

$$\begin{aligned}f(\Theta_{\mathcal{S}}|d(\mathring{\tau}_{\mathcal{K}}))&=\mathcal{G}_{\Theta_{\mathcal{S}}}(V_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}},\nu_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}})\\&\equiv\mathcal{G}_{\Theta_{\mathcal{S}}}\left({}^{\Delta}V_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}}+\bar{V}_{\mathcal{S}},{}^{\Delta}\nu_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}}+\bar{\nu}_{\mathcal{S}}\right)\end{aligned}\qquad(18)$$

reflect system properties approximately. Consequently, these PDFs have to be flattened adequately before merging such a piece of knowledge. This is of great importance since, for instance, simulators may provide a huge amount of data that may over-fit the posterior PDFs at wrong positions. The geometric mean could serve for flattening these PDFs. Its use is, however, unnecessary since an appropriate weighting is applied during merging of all knowledge items anyway.

3.2 Construction of fictitious data

Here, the focus is 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 prespecified stimulus. Static gain, a point on the frequency response and time-constants all serve as prominent examples. Such a knowledge item $\mathcal{K}$ is always uncertain to some degree. It can be interpreted as a collection of partial chracterisations of several predictors. Each of them is expressed in terms of its prior PDF $f(\Theta_{\mathcal{S}}|\tau_{\mathcal{K}})$

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

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

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

$h(\psi_{\mathcal{S};\tau_{\mathcal{K}}})$ and $H(\Psi_{\mathcal{S};\tau_{\mathcal{K}}})=H(d_{\tau_{\mathcal{K}}},\psi_{\mathcal{S};\tau_{\mathcal{K}}})$ are known vector functions of the indicated arguments. When there is no PDF $f(\Theta_{\mathcal{S}}|\tau_{\mathcal{K}})$ fulfilling (19) and (20), this information source is inconsistent and has to be split into several, internally consistent, knowledge sources. Then the restrictions (19) and (20) do not determine fully the constructed PDF $f(\Theta_{\mathcal{S}}|\tau_{\mathcal{K}})$. Pragmatic reasons encourage one 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) and (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 preprior PDF $\bar{f}(\Theta_{\mathcal{S}})$. The choice is made among those meeting (19) and (20). The Kullback–Leibler distance [20] $\mathcal{D}(f\|\bar{f})=\int f(\Theta_{\mathcal{S}}|\tau_{\mathcal{K}})\ln[f(\Theta_{\mathcal{S}}|\tau_{\mathcal{K}})/\bar{f}(\Theta_{\mathcal{S}})]\,d\Theta_{\mathcal{S}}$ is used as an adequate proximity measure.

Both $f(\Theta_{\mathcal{S}}|\tau_{\mathcal{K}})$, built in this way, and preprior PDF $\bar{f}(\Theta_{\mathcal{S}})$ belong to the same conjugate class. Thus, their ratio can be interpreted as a product of the parameterised model at some fictitious data vectors, leading to the statistics increments ${}^{\Delta}V_{\mathcal{S};\tau_{\mathcal{K}}}$, ${}^{\Delta}\nu_{\mathcal{S};\tau_{\mathcal{K}}}$. 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 Bayes's rule. Thus, $f(\Theta_{\mathcal{S}}|\mathcal{K})$ is determined by statistics

$${}^{\Delta}V_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}}=\sum_{\tau_{\mathcal{K}}=1}^{\mathring{\tau}_{\mathcal{K}}}{}^{\Delta}V_{\mathcal{S};\tau_{\mathcal{K}}},\;{}^{\Delta}\nu_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}}=\sum_{\tau_{\mathcal{K}}=1}^{\mathring{\tau}_{\mathcal{K}}}{}^{\Delta}\nu_{\mathcal{S};\tau_{\mathcal{K}}}\qquad(21)$$

The values ${}^{\Delta}V_{\mathcal{S};\tau_{\mathcal{K}}}$, ${}^{\Delta}\nu_{\mathcal{S};\tau_{\mathcal{K}}}$ are chosen so that the PDFs

$$\mathcal{G}_{\Theta_{\mathcal{S}}}\left({}^{\Delta}V_{\mathcal{S};\tau_{\mathcal{K}}}+\bar{V}_{\mathcal{S}},{}^{\Delta}\nu_{\mathcal{S};\tau_{\mathcal{K}}}+\bar{\nu}_{\mathcal{S}}\right)$$

minimise the Kullback-Leibler distance to $\mathcal{G}_{\Theta_{\mathcal{S}}}(\bar{V}_{\mathcal{S}},\bar{\nu}_{\mathcal{S}})$ under restrictions (19) and (20).

The optimisation is elaborated for the ARX model in the following Section.

4 Application to normal ARX model

The normal ARX model and its variants are predominantly used in practice. This determined that the system Designer is oriented to it and encourages one to specialise 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):

$$\begin{aligned}f(d|\psi,\Theta)&=\mathcal{N}_d(\theta'\psi,r)\\&=A(\Theta)\exp[\langle B(\Psi),C(\Theta)\rangle]\;\text{with}\end{aligned}\qquad(22)$$

$\Theta=[\theta,r]=[\text{regression coefficients, noise variance}]$, $A(\Theta)=(2\pi r)^{-0.5}\langle B,C\rangle=\text{tr}[B'C]$, $B(\Psi)=\Psi\Psi'$, $C(\Theta)=(2r)^{-1}[-1,\theta']'[-1,\theta']$. This correspondence determines the conjugate prior (4) in the form known as the Gauss-inverse-Wishart (GiW) PDF [21]

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

The $(\mathring{\Psi},\mathring{\Psi})$-dimensional extended information matrix V can be chosen symmetric and must be positive definite as the function $\mathcal{G}_\Theta(V,\nu)$ is to be normalised to a PDF. Consequently, there is the numerically advantageous L′DL decomposition of this matrix [22]

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

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

$$\mathcal{G}_\Theta(L,D,\nu) = \frac{r^{-0.5(\nu+\mathring{\psi}+2)}}{\mathcal{I}(L,D,\nu)} \times \exp\left\{-\frac{1}{2r}\left[(\theta-\hat{\theta})'C^{-1}(\theta-\hat{\theta}) + {}^{d}D\right]\right\} \tag{25}$$

where $\hat{\theta} = {}^{\psi}L^{-1\,d\psi}L =$ least-squares (LS) estimate of θ, $C = {}^{\psi}L^{-1}\psi D^{-1}({}^{\psi}L')^{-1} =$ LS covariance factor, and ${}^{d}D =$ LS remainder. The normalisation integral is

$$\mathcal{I}(L,D,\nu) = \Gamma(0.5\nu)\ {}^{d}D^{-0.5\nu} \prod_{j=1}^{\mathring{\psi}} {}^{\psi}D_{jj}^{-0.5} 2^{0.5\nu} (2\pi)^{0.5\,\mathring{\psi}} \text{ with} \tag{26}$$

$$\Gamma(x) = \int_0^\infty z^{x-1}\exp(-z)\,dz < \infty \quad \text{for } x > 0$$

The gamma function is finite IFF $\nu > 0$ and V is positive definite $\Leftrightarrow D_{jj} > 0,\ j = 1,\ldots,\mathring{\Psi}$. Under this condition, the normalisation integral $\mathcal{L}$ is finite.

The GiW PDF has the following moments

$$\mathcal{E}[r\,|\,L,D,\nu] = \frac{{}^{d}D}{\nu-2} = \hat{r},\ \mathrm{var}[r\,|\,L,D,\nu] = \frac{2\hat{r}^2}{\nu-4}$$
$$\mathcal{E}[\theta\,|\,L,D,\nu] = {}^{\psi}L^{-1\,d\psi}L = \hat{\theta}$$
$$\mathrm{cov}[\theta\,|\,L,D,\nu] = \frac{{}^{d}D}{\nu-2}\,{}^{\psi}L^{-1}\,{}^{\psi}D^{-1}({}^{\psi}L')^{-1} = \hat{r}C \tag{27}$$

Proposition 2.1 specialises to the following normal variant.

Proposition 4.2 (Estimation and 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 $\mathcal{G}_\Theta({}^{a}L,{}^{a}D,{}^{a}\nu)$ in stabilised forgetting with 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

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

The rectangular matrix $\mathbf{L}'$ is mapped on $[L_t',0]$ by the regular matrix $\mathcal{T}$ that diagonalises 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 to the LS quantities, proposition 4.1, implies that its updating is equivalent to recursive least squares [13]. It is equipped with tracking ability through stabilised forgetting. Numerically, its $L'DL$ decomposition is evaluated by using an efficient, rank-one updating [22] that creates the mapping $\mathcal{T}$ (see proposition 4.2).

4.3 Construction of fictitious data

Here, processing of the most common case of prior knowledge is presented. 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), by

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

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

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

where $\hat{\theta}'_{\tau_\mathcal{K}}, \hat{\mathbf{r}}_{\tau_\mathcal{K}}, C_{\tau_\mathcal{K}}$ are LS equivalents of the statistics $V_{\tau_\mathcal{K}}$ resulting from the minimisation of the KL distance to the preprior PDF.

Typically, the expert provides the range $[\underline{d}_{\tau_\mathcal{K}},\bar{d}_{\tau_\mathcal{K}}]$ 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 $\hat{d}_{\tau_\mathcal{K}} = 0.5(\underline{d}_{\tau_\mathcal{K}} + \bar{d}_{\tau_\mathcal{K}})$ and ${}^{d}r_{\tau_\mathcal{K}} = [0.5(\bar{d}_{\tau_\mathcal{K}} - \underline{d}_{\tau_\mathcal{K}})]^2$.

The preprior PDF used in the minimisation is assumed to be of the form

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

It is given by positive scalars ${}^{d}\varepsilon, \varepsilon, \bar{\nu}$. 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 preprior PDF (31), the optimised Kullback–Leibler distance is (the subscript $\tau_\mathcal{K}$ is also suppressed whenever possible)

$$2\mathcal{D}(f\|\bar{f}) = \omega(\nu) + \varepsilon\mathrm{tr}|C| - \ln|C| + \bar{\nu}\ln(\hat{r}) + \frac{\nu}{(\nu-2)\hat{r}}\left[\varepsilon\hat{\theta}'\hat{\theta} + {}^{d}\varepsilon\right] \quad (32)$$

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

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

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

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

$$q = \frac{\nu_{\tau_{\mathcal{K}}}}{\nu_{\tau_{\mathcal{K}}} - 2}(\alpha\rho + \beta), \quad \alpha = \frac{\hat{d}^2}{d_r}, \quad \beta = \frac{{}^{d}\varepsilon}{d_r} > 0$$
$$b = \rho + 1 - q + \bar{\nu}, \quad c = \rho(-\bar{\nu} + q) + q,$$
$$x = 0.5\left(-b + \sqrt{b^2 + 4c}\right) \quad (34)$$

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

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

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

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

$$\hat{\theta} = \frac{\hat{d}\psi}{\psi'\psi} \quad (36)$$

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

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

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

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

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 $x = \frac{-\bar{\nu}(\rho+x)}{\rho+x+1} + q$, which converts to the following quadratic equation in x, $x^2 + \underbrace{[\rho + 1 - q + \bar{\nu}]}_{b}x - \underbrace{[\rho(-\bar{\nu}+q)+q]}_{c} = 0$. ν sufficiently close to 2 (from the right) gives $c > 0$ and the equation has the unique positive (meaningful) solution $x = 0.5(-b + \sqrt{b^2 + 4c})$.

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

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

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

is taken. Inserting the relationship between x and q from (37), one can express $\nu_{\tau_{\mathcal{K}}} - 2$ term in the equation as a function of x. It gives a third-order algebraic equation for x with a real solution guaranteed. Standard but lengthy analysis establishes the uniqueness of the solution in the meaningful domain. □

4.4 Practical examples of prior knowledge

Here, we list common prior pieces of available knowledge and ways to construct the data vectors $\Psi'_i = [\hat{d}_i, \psi'_i]$ (fixed subscripts $\tau_{\mathcal{K}}, \mathcal{S}$ associated are again suppressed). Multivariate data items d_t and the common case where the state is in 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 the ith 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 express the fact that the data in the jth channel $d_{j;t-\delta}$ are in the constructed regression vector ψ_i.

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

- Knowledge of the static gain $\hat{d}_i = \hat{g}$ of the ith channel on a stimulus from the jth 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 [23] that the knowledge of a point of the frequency response stimulated by a periodic signal on the jth channel, given by the magnitude $\hat{a}(\omega)$ and phase shift $\phi(\omega)$ 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 $a(\omega) = [a(\omega), \bar{a}(\omega)]$ 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 (16) deals with 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 excludes an isolated pass through the zero level. Again, it generates several competitive knowledge items balanced by the merging procedure.
- Knowledge of the range of the dominant time constant is implemented by modelling the lower and upper envelope of 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 the difference of envelopes determines the variance d_r.
- The envelope of measured data, obtained from a periodic measurement, is handled in the same way as dominant time constant.
- The smoothness of the step response [24] 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 Bayes' rule, would be violated.

4.5 Overall algorithm for normal ARX model

Here, we put together the algorithmic elements for the normal ARX model. The recommended options correspond to preprocessed data $d(\mathring{t})$ with outliers and measurement noise suppressed and with data scaled so that their means are approximately zero and variances are about one. The evaluations attempt to conserve computational resources as much as possible.

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

Algorithm 4.1 (Structure estimation with prior knowledge): Initial phase

- Select 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 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 $\bar{L}_{\mathcal{R}} = I$, $\bar{D}_{\mathcal{R}} = \mathrm{diag}[{}^d\varepsilon, \varepsilon\overbrace{[1, \ldots, 1]}^{\mathring{\psi}_{\mathcal{R}}}]$ and $\bar{\nu}$ determining the flat preprior PDF on the richest possible parameterisation. Requirements on finiteness of the *a priori* assigned expectation of r, and the need for a flat PDF $\bar{f}(r)$, lead to the use of $\bar{\nu} = 3$, see (27). For this choice, ${}^d\varepsilon$ is the variance of the unpredictable part of the modelled data. It is sufficient to consider a few categories of noise-to-signal ratio. For instance, the values (0.1%, 1%, 5%, 10%, 50%) correspond to ${}^d\varepsilon = (10^{-6}, 10^{-4}, 0.0025, 0.01, 0.25)$. For stable systems, that are predominantly treated, the autoregression coefficients do not excess 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 $\varepsilon \approx 25\,{}^d\varepsilon/\gamma^2$ is an appropriate conservative option.
- Select a 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 the L′DL decomposition of the increment of the extended information matrix corresponding to the richest data vectors $\Psi_{\mathcal{R};t}$, see (14),

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

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

- 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 preprior PDF

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

- 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 statistics

- Select the factors of the preprior 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 ${}^{\Delta}L_{i\mathcal{R};\mathring{t}}$, ${}^{\Delta}D_{i\mathcal{R};\mathring{t}}$, corresponding to the ith 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 ${}^{\Delta}L_{\mathcal{R};\mathring{t}}$, ${}^{\Delta}D_{\mathcal{R};\mathring{t}}$. The L′DL decompositions destroyed by this selection have to be recovered using rank-one corrections [22].
- Apply algorithm 2.1 giving L′DL factors of the preprior extended information matrices $\bar{L}_{i\mathcal{S}}$, $\bar{D}_{i\mathcal{S}}$ and their data-dependent increments ${}^{\Delta}L_{i\mathcal{S};\mathring{t}}$, ${}^{\Delta}D_{i\mathcal{S};\mathring{t}}$. They correspond to the most probable structures $\mathcal{S} \in \mathring{\mathcal{S}}^*$ found when just the nested preprior knowledge is used.
- Select $\mathcal{S} \in \mathcal{S}^*$.

Inclusion of prior knowledge for promising structure

- Select a knowledge item $\mathcal{K}$ in the list $\mathcal{K}_i^* = \{1, \ldots, \mathring{\mathcal{K}}_i\}$ available for the ith channel.
- Set the normalisation factor needed in merging, i.e. $s = 0$.

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 $\hat{d}_{i\mathcal{S};\tau_{\mathcal{K}}}$, ${}^d r_{i\mathcal{S};\tau_{\mathcal{K}}}$ 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

$$ L'_{i\mathcal{S};\tau_{\mathcal{K}}}D_{i;\mathcal{S}\tau_{\mathcal{K}}}L_{i\mathcal{S};\tau_{\mathcal{K}}} = \begin{bmatrix} L_{i\mathcal{S};\tau_{\mathcal{K}}-1} \\ \Psi'_{i\mathcal{S};\tau_{\mathcal{K}}} \end{bmatrix}' \mathrm{diag}\begin{bmatrix} D_{i\mathcal{S};\tau_{\mathcal{K}}-1} \\ 1 \end{bmatrix} \begin{bmatrix} L_{i\mathcal{S}\tau_{\mathcal{K}}-1} \\ \Psi'_{i\mathcal{S};\tau_{\mathcal{K}}} \end{bmatrix} $$

$$ \nu_{i\mathcal{S};\tau_{\mathcal{K}}} = \nu_{i\mathcal{S};\tau_{\mathcal{K}}-1} + {}^{\Delta}\nu_{i\mathcal{S};\tau_{\mathcal{K}}}. $$

- Run the estimation with stabilised 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 in ${}^{\Delta}L_{i\mathcal{S};\mathring{\tau}_{\mathcal{K}}}$, ${}^{\Delta}D_{i\mathcal{S};\mathring{\tau}_{\mathcal{K}}}$ and ${}^{\Delta}\nu_{\mathcal{S};\mathring{\tau}_{\mathcal{K}}}$

- Evaluate the partial likelihood

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

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

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

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

- Take a new knowledge item $\mathcal{K}$, if the list $\mathcal{SK}_i^*$ is not exhausted, and go to *Processing of knowledge items*. Otherwise continue.
- Set $L_{i\mathcal{S}\mathcal{K}(\overset{\circ}{\mathcal{K}})} = I,\ D_{i\mathcal{S}\mathcal{K}(\overset{\circ}{\mathcal{K}})} = \bar{D}_{i\mathcal{S}},\ \nu_{i\mathcal{S}\mathcal{K}(\overset{\circ}{\mathcal{K}})} = \bar{\nu}_{i\mathcal{S}}$.
- Select $\mathcal{K} \in \mathcal{K}_i^*$.

Evaluation of the merger within the structure $\mathcal{S}$

- Update

$$f_i(\mathcal{K} \mid d(\overset{\circ}{t}),\mathcal{S}) = \frac{\mathcal{L}_i(d(\overset{\circ}{t}),\mathcal{S},\mathcal{K})}{s},\ L'_{i\mathcal{S}\mathcal{K}(\overset{\circ}{\mathcal{K}})}D_{i\mathcal{S}\mathcal{K}(\overset{\circ}{\mathcal{K}})}L_{i\mathcal{S}\mathcal{K}(\overset{\circ}{\mathcal{K}})}$$

$$= L'_{i\mathcal{S}\mathcal{K}(\overset{\circ}{\mathcal{K}})}D_{i\mathcal{S}\mathcal{K}(\overset{\circ}{\mathcal{K}})}L_{i\mathcal{S}\mathcal{K}(\overset{\circ}{\mathcal{K}})} + f_i(\mathcal{K} \mid d(\overset{\circ}{t}),\mathcal{S}) \times {}^{\Delta}L'_{i\mathcal{S};;\overset{\circ}{\tau}_{\mathcal{K}}}\,{}^{\Delta}D_{i\mathcal{S};\overset{\circ}{\tau}_{\mathcal{K}}}\,{}^{\Delta}L_{i\mathcal{S};\overset{\circ}{\tau}_{\mathcal{K}}}$$

$$\nu_{i\mathcal{S}\mathcal{K}(\overset{\circ}{\mathcal{K}})} = \nu_{i\mathcal{S}\mathcal{K}(\overset{\circ}{\mathcal{K}})} + f_i(\mathcal{K} \mid d(\overset{\circ}{t}),\mathcal{S})\,{}^{\Delta}\nu_{i\mathcal{S};\overset{\circ}{\tau}_{\mathcal{K}}}.$$

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

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

- Set $\hat{\mathcal{S}} = \mathcal{S},\ \hat{\mathcal{L}}_i = \mathcal{L}_i(d(\overset{\circ}{t}),\mathcal{S})$ and store the statistics corresponding to the posterior PDF ${}^{\Delta}L'_{i\hat{\mathcal{S}};\overset{\circ}{t}}\,{}^{\Delta}D_{i\hat{\mathcal{S}};\overset{\circ}{t}}\,{}^{\Delta}L_{i\hat{\mathcal{S}};\overset{\circ}{t}} + L'_{i\hat{\mathcal{S}}\mathcal{K}(\overset{\circ}{\mathcal{K}})}D_{i\hat{\mathcal{S}}\mathcal{K}(\overset{\circ}{\mathcal{K}})}L_{i\hat{\mathcal{S}}\mathcal{K}(\overset{\circ}{\mathcal{K}})}$ and $\overset{\circ}{t}+\nu_{i\hat{\mathcal{S}}}$ if $\mathcal{L}_i(d(\overset{\circ}{t}),\mathcal{S}) > \hat{\mathcal{L}}_i$.
- Select a new structure $\mathcal{S}$ if the list $\mathcal{S}^*$ of most probable cases is not exhausted and go to the *Inclusion of prior knowledge for promising structure*. Otherwise continue.
- Offer the structure $\hat{\mathcal{S}}$ as the recommended one for the ith channel with the corresponding stored posterior statistics.
- Increment i and go to *Cycle over indices i of modelled data entries* if $i \leq \overset{\circ}{d}$. Otherwise stop.

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

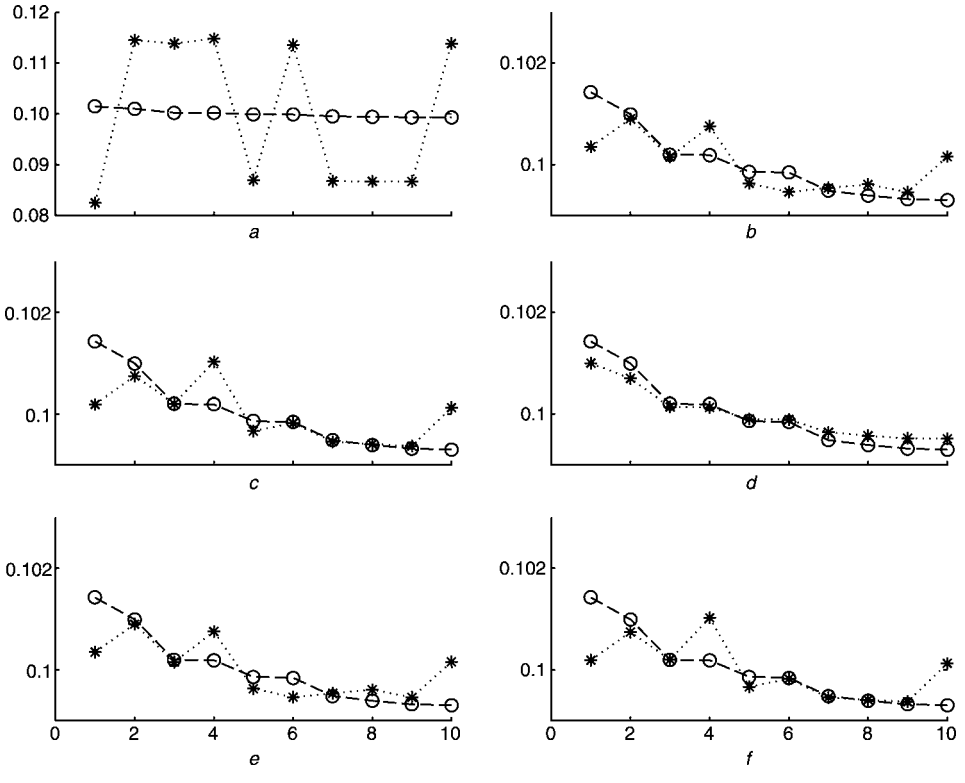

Fig. 1 *Structure estimation with prior knowledge*

Number $\mathcal{S}$ on x-axis coincide with those in Table 1. ○ probabilities obtained with weak nested prior knowledge. ∗ probabilities resulted from use of prior knowledge described in title of subplot as follows:
a Gain
b Data
c Data envelope
d Time constants
e Gain + data
f Gain + data envelope

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

5 Example

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

$$\begin{aligned} y_t &= {}^o a_1 y_{t-1} + {}^o a_2 y_{t-2} + {}^o b_0 u_t + {}^o b_1 u_{t-1} + \sqrt{{}^o r}\, e_t \\ &= \underbrace{[1.81, 0.8187, 0.0438, 0.00468]}_{{}^o\theta'} \underbrace{[y_{t-1}, y_{t-2}, u_t, u_{t-1}]'}_{\psi_t} \\ &\quad + \underbrace{\sqrt{0.0001}}_{\sqrt{{}^o r}} e_t \\ &\Leftrightarrow f(y_t \mid u_t, d(t-1), {}^o\Theta) = f(y_t \mid \psi_t, {}^o\Theta) \\ &\quad = \mathcal{N}_{y_t}({}^o\theta'\psi_t, {}^o r), \end{aligned}$$

The 'real' data $d(\mathring{t}) = d(300)$ were 'measured' on this system.

Algorithm 2.1 was applied with the richest structure of the phase form given by the order $\mathring{\delta}_{\mathcal{R}} = 6$ and the recommended nested preprior PDF was 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.

The following internally consistent knowledge items were considered.

The influence of the knowledge items and their combinations described in Table 2 on the posterior probabilities within the set of structures given in Table 1 were inspected. They are shown in Figs. 1 and 2.

Observe that

- the correct structure is in fourth position with weak prior knowledge
- static gain, data envelope and data increase probability of the true structure significantly (not necessarily sufficiently)
- time constant offers no benefit in this case

Table 1: Most probable structures for weak nested prior knowledge

S	y_{t-1}	y_{t-2}	y_{t-3}	u_t	u_{t-1}	other significant regressors	f(S | d(300))
1	*	*					0.828000
2	*	*			*		0.154000
3	*	*		*			0.006850
4	*	*		*	*		0.006560
5	*	*	*				0.001840
6	*	*				u_{t-6}	0.001710
7	*	*				y_{t-4}	0.000422
8	*	*				y_{t-5}	0.000290
9	*	*				y_{t-6}	0.000218
10	*	*	*		*		0.000204

Significant regressors are marked by $*$.

Table 2: Description of tested prior knowledge items

$\mathcal{K}$	Meaning	Specified by	Figure title
1	static gain	gain in [0.99,1.01]	gain
2	time constant	constant in [0.82,0.84] sampling period 0.1 [s]	time constants
3	data	100 samples of the model that arises from the "true" system by setting $b_0 = 0$	data
4	data envelope	envelope of 20 step responses	fictitious data

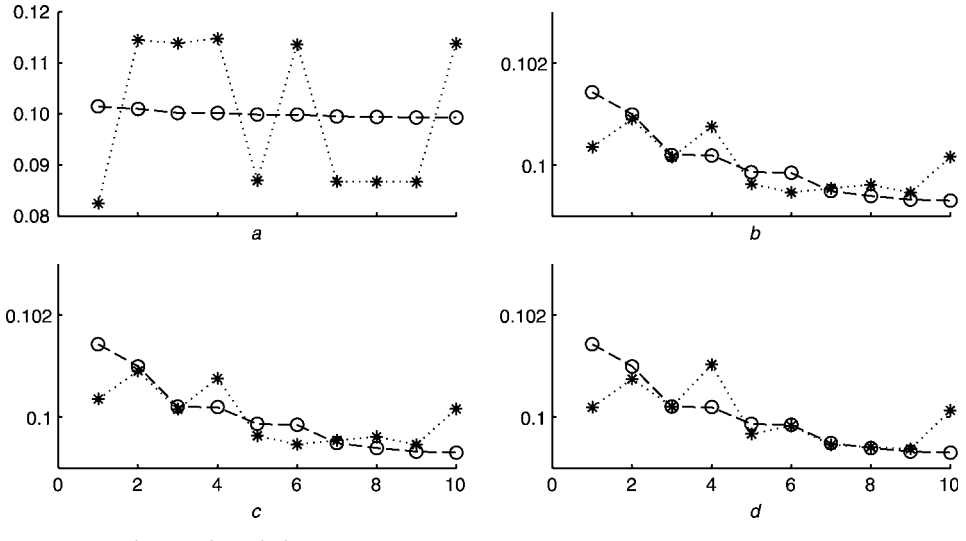

Fig. 2 *Structure estimation with prior knowledge*

Number $\mathcal{S}$ on x-axis coincide with those in Table 1. ○ probabilities obtained with weak nested prior knowledge, $*$ probabilities resulted from use of prior knowledge described in title of subplot as follows;

a Gain + time constant
b Data + data envelope
c Data + time constant
d Time constants
e Gain + data
f Gain + data envelope

- combinations of knowledge items do not guarantee significant improvement
- the combination of good and bad knowledge items does not rescue the result

These observations are typical. Prior knowledge is not a fail-safe means of overcoming the lack of information in real data but it helps. Quantification of individual knowledge pieces and their merging seem to perform well.

6 Concluding remarks

The significance of including prior knowledge in black-box models is still underestimated. The theory and algorithms presented treat this problem under circumstances met regularly in technological applications, *cf.* Section 2.4. Available practical experience supports the view that the use of even vague knowledge may decide the success or failure of structure estimation, and consequently the quality of the controller design.

A wider and more precise use of prior knowledge is especially important in the context of prior design of an advanced controller with incomplete knowledge [16].

The theory and algorithms have been elaborated for the LQG-type design set-up. The same problems are met beyond this class, and the adopted methodology can be applied there also.

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

7 Acknowledgments

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

8 References

1 Astrom, K.J., and Wittenmark, B.: 'Adaptive control' (Addison-Wesley, Reading, MA, 1989)
2 Wellstead, P.E., and Zarrop, M.B.: 'Self-tuning systems' (Wiley, Chichester, 1991)
3 Mosca, E.: 'Optimal, predictive and adaptive control' (Prentice Hall, New Jersey, 1994)
4 Kárný, M., Halousková, A., Böhm, J., Kulhavý, R., and Nedoma, P.: 'Design of linear quadratic adaptive control: Theory and algorithms for practice', *Kybernetika*, 1985, **21**, Supplement to No. 3, 4, 5, 6, pp. 1–99
5 Johnson, A.: 'LQG applications in the process industries', *Chem. Eng. Prog.*, 1993, **48**, (16), pp. 2829–2838
6 Elbelkacemi, M., Lachhab, A., Limouri, M., Dahhou, D., and Essaid, A.: 'Adaptive control of a water supply system', *Control Eng. Pract.*, 2001, **9**, (3), pp. 343–349
7 Clarke, D.W.: 'Advances in model-based predictive control' (Oxford University Press, Oxford, 1994)
8 Kárný, M., and Halousková, A.: 'Implementing LQG adaptive control: a CAD approach'. Proc. 9th IFAC/IFORS Symp. on Identification and system parameter estimation, AKA PRINT Nyomdaipari, Budapest, 1991, pp. 1585–1590
9 Kárný, M., and Halousková, A.: 'Pretuning of selftuners' in Clarke, D. (Ed.): 'Advances in model-based predictive control' (Oxford University Press, Oxford, 1994), pp. 333–343
10 Bůcha, J., Kárný, M., Nedoma, P., Böhm, J., and Rojiček, J.: 'Designer 2000 project'. Proc. Inf. Conf. on Control IEE, London, September 1998, pp. 1450–1455
11 Kárný, M., Khailova, N., Böhm, J., and Nedoma, P.: 'Quantification of prior information revised', *Int. J. Adapt. Control Signal Process.*, 2001, **15**, (1), pp. 65–84
12 Kárný, M., and Kulhavý, R.: '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
13 Peterka, V.: 'Bayesian system identification' in Eykhoff, P. (Ed.): 'Trends and progress in system identification' (Pergamon, Oxford, 1981), pp. 239–304
14 Kulhavý, R., and Zarrop, M.B.: 'On general concept of forgetting', *Int. J. Control*, 1993, **58**, (4), pp. 905–924
15 Kárný, T., Jeniček, T., and Ottenheimer, W.: 'Contribution to prior tuning of LQG selftuners', *Kybernetika*, 1990, **26**, (2), pp. 107–121
16 Novák, M., and Böhm, J.: 'Automated multivariable adaptive controller design', *IEE Proc. Control Theory Appl.*, 2003
17 Bernardo, J.M., and Smith, A.F.M.: 'Bayesian theory' (Wiley, Chichester, 1997, 2nd edn.)
18 Nedoma, P., Kárný M., and Böhm, J.: 'Software tools for use of prior knowledge in design of LQG adaptive controllers', in Proc. IFAC Workshop on Adaptive systems in control and signal processing, Glasgow, August 1998, pp. 425–429
19 Nedoma, P., Kárný, M., Böhm, J., Rojíček, J., and Berec, L.: 'ABET98. Adaptive Bayesian Estimation Toolbox for MATLAB'. Technical Report 1937, ÚTIA AV ČR, Praha, 1998
20 Kullback, S., and Leibler, R.: 'On information and sufficiency', *Annals Math. Stat.*, 1951, **22**, pp. 79–87
21 Zellner, A.: 'Introduction to bayesian inference in econometrics' (Wiley, New York, 1976)
22 Bierman, G.J.: 'Factorization methods for discrete sequential estimation' (Academic Press, New York, 1977)
23 Khaylova, N.: 'Exploitation of prior knowledge in adaptive control design', PhD thesis, FAV ZČU, University of West Bohemia, Pilsen, Czech Republic, 2001
24 Kárný, M.: 'Quantification of prior knowledge about global characteristics of linear normal model', *Kybernetika*, 1984, **20**, (5), pp. 376–385