Fully Probabilistic Knowledge Expression and Incorporation

Miroslav Kárný, Josef Andrýsek, Antonella Bodini,
Tatiana V. Guy, *Senior Member, IEEE,* Jan Kracík, Petr Nedoma,
and Fabrizio Ruggeri

Abstract

Exploitation of prior knowledge in parameter estimation is vital whenever data is not informative enough. Elicitation and quantification of prior knowledge is a well-elaborated art in societal and medical applications but not in the engineering ones. Frequently required involvement of a *facilitator* is mostly unrealistic due to either facilitator's high costs or the high complexity of modelled relationships that cannot be grasped by the human. This paper provides a facilitator-free approach exploiting a methodology of knowledge sharing.

The considered task assumes prospective models be indexed by an unknown finite-dimensional parameter. The parameter is estimated using: (i) observed data; (ii) a prior probability density function (pdf); and (iii) uncertain expert's information on the modelled data. The parametric model specifies pdf of the system's output conditioned on realised data and parameter. Data is assumed to enter the time-invariant model only via a finite-dimensional regression vector. The adopted methodology deals with expert's knowledge expressed as a collection of pdfs on the space of data trajectories. Instead of sampling from these pdfs and applying Bayes rule to the samples, the proposed approach uses the asymptotic formulae arisen from gedanken experiment relevant to the knowledge considered.

Manuscript received June 21, 2006; revised August 15, 2008. This work was supported by grants MŠMT ČR 1M0572, MŠMT ČR 2C06001, GAČR 102/08/0567 and Czech-Italian cooperation.

M. Kárný, J. Andrýsek, T.V. Guy, J. Kracík, and P. Nedoma are with the Adaptive Systems Department, Institute of Information Theory and Automation, Prague, Czech Republic, Email: { school, andrysek, guy, kracik, nedoma}@utia.cas.cz

A. Bodini, and F. Ruggeri are with Istituto di Matematica Applicata e Tecnologie Informatiche, Milano, Italy, Email: { anto, fabrizio}@mi.imati.cnr.it

The paper specifies knowledge-expressing pdfs for commonly accessible types of knowledge, applies the methodology to a normal controlled autoregressive model, and illustrates positive contributions of this knowledge processing.

Index Terms

Bayesian estimation, prior knowledge, automatised knowledge elicitation, just-in-time modelling

I. Introduction

Efficient use of prior knowledge influences quality of subsequent decision making (DM) relying on estimated models. The considered Bayesian DM paradigm [1] combines data with expert's knowledge quantified by a prior probability density function (pdf). Its construction dominates the activity known as knowledge elicitation, e.g. [2]. It is addressed repeatedly and a range of techniques has been developed, for review, see [3]. The majority of them, however, relies on a facilitator, who guides an expert providing knowledge and quantifies knowledge gathered. This way is expensive and can cope only with relatively simple cases.

Adaptive controllers and predictors based on recursive estimation [4], [6] are examples of important DM systems. Selection of their structure and their transient behavior depend, sometimes critically, on the properly quantified prior knowledge. Limitations of the available knowledge elicitation methodologies are even more strict for them, at least due to their assumed extensive use. This state has motivated search for elicitation techniques weakly dependent on the facilitator.

The papers [7], [8], [9] elaborate the desired technique for models within the exponential family (EF), [17] which dominates in applied adaptive systems as it enables exact recursive estimation on extending data sets. The referred technique transforms processed knowledge into so-called fictitious data, i.e., data that could be observed on the modelled system, and uses this data for estimation as real ones. Duality of some parts of this transformation has motivated attempts to treat the automatic elicitation as an optimisation under knowledge-reflecting constraints [10], [11]. This way is applicable only to specific cases, similarly as the other rare attempts [12].

Despite the increasing interest for the automatic knowledge elicitation, identification for control purposes has not employed its great potential yet. For examples transients are (at most) handled via specific control strategies, e.g. [13], [14].

This paper provides a solution that suppresses drawbacks and inconsistencies of the discussed predecessors, covers a wider range of types of knowledge and objectively determines relative impact of respective knowledge pieces. The improvement uses results [15], [16] that justify *inclusion of probabilistically expressed knowledge about possible data into the parameter estimation.*

For readers, this paper may serve as: (i) a review of improved automatic knowledge elicitation approach; and (ii) a justified methodology which allows a specific, case-dependent knowledge be relatively straightforwardly utilised for parameter estimation. The elaborated cases of prior knowledge are wide-spread in practice and thus may guide in active use of the proposed methodology.

Section II recalls the basic result [16], we rely on, summarises Bayesian estimation of the normal autoregressive model with exogenous inputs (ARX). Section III proposes processing of prevalent types of prior knowledge, like information about data ranges, rise time, frequency response, and simulated as well as obsolete data. Section IV proposes the way how to choose weights controlling the overall impact of the processed knowledge. Section V provides illustrative examples. Section VI summarises the results obtained.

II. PRELIMINARIES

The following basic notation is adopted. $\mathsf{X} = \{x_1, x_2, \ldots\}$ stands for a finite set. The cardinality of the set X is denoted $|\mathsf{X}|$ and provides the number of elements of X.

If x is a vector, x^{ℓ} denotes number of vector elements. Lower and upper bounds on $x \in \mathsf{X}$ are $\underline{x}$ and $\overline{x}$, respectively. They are meant entry-wise for vectors. If x is a matrix, $x \geq \underline{x}$ if and only if $(x - \underline{x})$ is positive semi-definite. x' stands for the transpose of x.

f is the pdf of a random variable, whose identifier is in its argument. $\tilde{x}$ means a specific value of the random variable x (typically, inserted into the condition). Subscript $t \in \mathsf{T}$ labels discrete time moments of observing/recording data d_t. $d(t)$ means the sequence of data records $d_1, d_2, \ldots, d_t$. Subscript $\tau \in \mathcal{T}$ refers either to the τth source/piece of prior knowledge or to time index of "fictitious" data, i.e. data that could be observed on the modelled system". The concrete meaning of τ is emphasised, if needed.

A. *Probabilistic description of knowledge*

A closed loop formed of system/process of interest and DM system is considered. The data record observed at time t, d_t, consists of the system output y_t and input u_t, i.e. $d_t \equiv (y_t, u_t)'$.

The addressed parameter estimation task concerns the *time-invariant parametric model* of the system. This model specifies the probability density function (pdf) of the scalar system output y_t conditioned by a finite-dimensional regression vector $\psi_t \equiv [u_t, d(t-1)]'$ and a finite-dimensional unknown parameter Θ:

$$M(\Psi_t, \Theta) = f(y_t | \underbrace{u_t, d(t-1)}_{\psi_t}, \Theta) = f(y_t|\psi_t, \Theta). \qquad (1)$$

The predicted system output y_t and the regression vector ψ_t form so-called *data vector*, $\Psi_t \equiv [y_t, \psi_t']'$.

Knowledge about the unknown parameter Θ is initially described by proper prior pdf $f(\Theta)$. Besides, available expert's knowledge (possibly imprecise and incomplete) of some system's characteristic often indirectly inform about the parameter and as such should be used in the parameter estimation. Each of the processed knowledge pieces is indexed by $\tau \in \mathcal{T}$ where $\mathcal{T}$ is a finite set of all sources/pieces of prior knowledge available. To describe this knowledge in order to exploit it, the following approach is considered.

- A gedanken experiment, reflecting the considered τth expert's knowledge, of the underlying system's characteristic, is assumed.
- The *possible* outcomes of this gedanken experiment are described by a pdf $f_\tau(\Psi)$, where Ψ is a data vector composed of so-called "fictitious data", i.e. data would be observed on the system if this experiment was performed in reality.
- The obtained knowledge-expressing pdf can be further used by the parameter estimation.

Note: τ refers to the τth gedanken experiment, i.e. to the τth source of the expert's knowledge.

The expert's knowledge provided by gedanken experiments is formally described by a set of pdfs $\mathsf{F}_{\mathcal{T}} \equiv \{f_\tau(\Psi)\}_{\tau\in\mathcal{T}}$, on the space of data trajectories, where cardinality $|\mathcal{T}|$ of the set $\mathcal{T}$, equals the total number of gedanken experiments performed. The paper [16] proposes the following definition of the prior pdf *denoted* by $f(\Theta|\mathsf{F}_{\mathcal{T}})$ which incorporates the knowledge provided by $\mathsf{F}_{\mathcal{T}}$:

$$f(\Theta|\mathsf{F}_{\mathcal{T}}) = \frac{f(\Theta)\exp\left(|\mathcal{T}|\,\Omega_{\mathcal{T}}(\Theta)\right)}{\int f(\Theta)\exp\left(|\mathcal{T}|\,\Omega_{\mathcal{T}}(\Theta)\right) d\Theta}, \text{ with} \tag{2}$$

$$\Omega_{\mathcal{T}}(\Theta) \equiv \frac{1}{|\mathcal{T}|}\sum_{\tau\in\mathcal{T}}\int f_\tau(\Psi)\ln[M(\Psi,\Theta)]\,d\Psi. \tag{3}$$

The function $\Omega_{\mathcal{T}}(\Theta)$ in (3) can be interpreted as an expectation of the logarithm of the parametric model with respect to the average pdf $\hat{f}(\Psi)$ representing the set $\mathsf{F}_{\mathcal{T}} \equiv \{f_\tau(\Psi)\}_{\tau\in\mathcal{T}}$:

$$\hat{f}(\Psi) \equiv \frac{1}{|\mathcal{T}|}\sum_{\tau\in\mathcal{T}} f_\tau(\Psi). \tag{4}$$

The definition (2) extends an applicability of Bayes rule [18] to the processing of non-random values of data vectors. Easy to see that, the data vectors $\tilde{\Psi}_\tau$, $\tau \in \mathcal{T}$, define the set

$$\mathsf{F}_{\mathcal{T}} \equiv \left\{f_\tau : f_\tau(\Psi) = \delta(\Psi - \tilde{\Psi}_\tau) \equiv \text{Dirac delta on } \tilde{\Psi}_\tau\right\}_{\tau\in\mathcal{T}}, \tag{5}$$

for which formula (2) reduces to the ordinary Bayes rule.

Note that use of the equation (2) avoids the artificial and costly processing via sampling of the pdf $f_\tau(\Psi)$ and subsequent applying Bayes rule to the samples obtained. The evaluation of (3) and *mapping of available domain-specific knowledge pieces* on the set $\mathsf{F}_{\mathcal{T}}$ become the most demanding tasks needed to be solved.

The function $\Omega_{\mathcal{T}}(\Theta)$, (3), has a simple form when the parametric model (1) belongs to the exponential family (EF) [17]

$$M(\Psi,\Theta) = A(\Theta)\exp\left\langle B(\Psi), C(\Theta)\right\rangle, \tag{6}$$

where $A(\Theta)$ is a non-negative scalar function of Θ. $B(\Psi)$ and $C(\Theta)$ are multivariate functions of compatible dimensions and the functional $\langle\cdot,\cdot\rangle$ is linear in the first argument.

For the models in EF, the pdf (2) has the form

$$f(\Theta|\mathsf{F}_{\mathcal{T}}) \propto f(\Theta)A^{|\mathcal{T}|}(\Theta)\exp\left\langle |\mathcal{T}|V, C(\Theta)\right\rangle, \text{ with} \tag{7}$$

$$V \equiv \frac{1}{|\mathcal{T}|}\sum_{\tau\in\mathcal{T}}\Upsilon_\tau,\ \Upsilon_\tau \equiv \int B(\Psi)f_\tau(\Psi)\,d\Psi,\ \tau\in\mathcal{T}.$$

The array V is the expectation of $B(\Psi)$ with respect to the average pdf $\hat{f}(\Psi)$ (4).

The pdfs $f_\tau(\Psi)$, $\tau \in \mathcal{T}$ which cause an identical modification of the prior pdf $f(\Theta)$, i.e provide the same increment Υ_τ in (7), are equivalent for the task concerned.

Selecting a specific representative of this equivalence class makes the mapping of domain-specific knowledge pieces on the set $\mathsf{F}_{\mathcal{T}}$ more transparent. Since the considered knowledge pieces express information of the type: *values of data vectors* Ψ_τ *are highly expected to be in the set* Ψ_τ, the uniform pdf with a support on Ψ_τ can be selected as such representative

$$f_\tau(\Psi) = \mathcal{U}_\Psi(\Psi_\tau) \equiv \text{ uniform pdf on the set } \Psi_\tau,\ \tau \in \mathcal{T}. \tag{8}$$

Restricting ourselves to parametric models in EF, it is natural to consider the conjugate prior pdf

$$\begin{aligned} f(\Theta) &= \frac{A^\nu(\Theta)\exp\langle V, C(\Theta)\rangle}{\mathcal{I}(V,\nu)}, \\ \mathcal{I}(V,\nu) &\equiv \int A^\nu(\Theta)\exp\langle V, C(\Theta)\rangle\, d\Theta, \end{aligned} \tag{9}$$

given by $V = \underline{V}$ and $\nu = \underline{\nu}$ which guarantee properness, i.e., $\mathcal{I}(V,\nu) < \infty$. The posterior pdf $f(\Theta|\tilde{d}(t), \mathsf{F}_{\mathcal{T}})$, obtained after processing the historical data observed up to time t, $\tilde{d}(t) = (\tilde{d}_1, \ldots, \tilde{d}_t)$, preserves this form ($t$ denotes discrete time moments of data recording). Arrays V_t and scalars ν_t determining the posterior pdf evolve recursively with V given by (7)

$$\begin{aligned} V_t &= V_{t-1} + B(\tilde{\Psi}_t),\ V_0 \equiv \underline{V} + |\mathcal{T}|V, \\ \nu_t &= \nu_{t-1} + 1,\ \nu_0 \equiv \underline{\nu} + |\mathcal{T}|,\ t = 1, 2, \ldots \end{aligned} \tag{10}$$

Thus, prior knowledge given by $\mathsf{F}_{\mathcal{T}}$ is expressed via an average pdf $\hat{f}(\Psi)$, (4), and total number of gedanken experiments $|\mathcal{T}|$. The prior knowledge modifies initial conditions in (10) from $\underline{V}$ and $\underline{\nu}$ to V_0 and ν_0, respectively.

Remarks

1) The presentation convenience has motivated an arbitrary choice of the uniform pdf as a representative of the equivalence class.
2) The choice of a nonuniform pdf $f_\tau(\Psi)$, which takes data vectors from a set $\Psi_\tau, \tau \in \mathcal{T}$, as highly expected, defines another equivalence class than the uniform pdf on Ψ_τ. Thus, the use of the uniform pdf (8) predetermines the mapping of domain-specific knowledge pieces on the set $\mathsf{F}_{\mathcal{T}}$. Influence of this choice should be studied within the framework of robust Bayesian estimation, [19].

3) Often, the τth piece of knowledge concerns data vectors Ψ with a non-random part, typically a part of the regression vector ψ. Let us decompose

$$\Psi' \equiv \left[{}^{U}\Psi',\, {}^{N}\Psi'\right] \tag{11}$$

so that ${}^{U}\Psi$ and ${}^{N}\Psi$ contain uncertain and non-random entries, respectively. Applying the chain rule to the pdf $f_\tau(\Psi) \equiv f_\tau\left({}^{U}\Psi,\, {}^{N}\Psi\right)$, we get $f_\tau(\Psi) = f_\tau\left({}^{U}\Psi \middle|\, {}^{N}\Psi\right) \delta\left({}^{N}\Psi - {}^{N}\tilde{\Psi}_\tau\right)$, where ${}^{N}\tilde{\Psi}_\tau$ is the considered value of the non-random part of the data vector. This factorisation implies

$$\Upsilon_\tau = \int B\left({}^{U}\Psi,\, {}^{N}\tilde{\Psi}_\tau\right) f_\tau\left({}^{U}\Psi \middle|\, {}^{N}\tilde{\Psi}_\tau\right) d\,{}^{U}\Psi. \tag{12}$$

4) Uncertain domain-specific knowledge can be practically provided in many forms which require a specific mapping of a knowledge piece on pdfs $f_\tau(\Psi), \tau \in \mathcal{T}$.

A number of mappings frequently met in practice is developed in Section III. However they do not cover a full range of possibilities. The yet common cases not elaborated here can be treated similarly as follows:

- Knowledge of the type "*if the regression vector ψ yields the value $\tilde{\psi}_\tau$, then the output $y \in \mathrm{Y}_\tau$*" can be expressed by the pdf $f_\tau(\Psi) = \mathcal{U}_y(\mathrm{Y}_\tau)\delta(\psi - \tilde{\psi}_\tau)$.

- Knowledge of the type "*if the regression vector yields the values $\left[{}^{U}\psi',\, {}^{N}\tilde{\psi}'_\tau\right]'$ with ${}^{U}\psi \in {}^{U}\Psi_\tau$, then the output $y \in \mathrm{Y}_\tau$*" can be expressed by $f_\tau(\Psi) = \mathcal{U}_{y,{}^{U}\psi}\left(\mathrm{Y}_\tau,\, {}^{U}\Psi_\tau\right) \delta({}^{N}\psi - {}^{N}\tilde{\psi}_\tau)$, i.e., the uniform pdf is used for uncertain part of data vector, ${}^{U}\Psi = \left[y,\, {}^{U}\psi'\right]'$.

- Fuzzy rules can be treated in the same way as in the previous case. It is sufficient to interpret the involved membership functions as conditional pdfs defining the pdf $f_\tau(\Psi)$ via the chain rule. In general, this is possible after a suitable normalisation.

B. Bayesian estimation of normal ARX model

The developed processing of prior knowledge can be applied to the autoregressive model with exogenous inputs (ARX model). Use of the chain rule for pdfs allows to consider the ARX model with the single output only. Its Bayesian estimation, exploiting data records $\tilde{d}(t) = (\tilde{d}_1, \ldots, \tilde{d}_t)$

observed up to the discrete time t, is recalled here. The normal ARX model belonging to EF (6) reads

$$
\begin{aligned}
M(\Psi,\Theta)=\mathcal{N}_y(\theta'\psi,r) &\equiv \underbrace{\frac{1}{\sqrt{2\pi r}}}_{A(\Theta)}\exp\left[-\frac{(y-\theta'\psi)^2}{2r}\right] \\
&= A(\Theta)\exp\left\{-\overbrace{\mathrm{tr}\left(\underbrace{\Psi\Psi'}_{B(\Psi)}\underbrace{\frac{[-1,\theta']'[-1,\theta']}{2r}}_{C(\Theta)}\right)}^{\langle B(\Psi),C(\Theta)\rangle}\right\}, \qquad (13)
\end{aligned}
$$

In (13) θ is a (column) vector of regression coefficients, ψ is the corresponding regression vector, r is noise variance, and tr denotes trace. This model is determined by the unknown parameter $\Theta=(\theta,r)$. Its conjugate prior pdf is a normal-inverse-gamma pdf, [10], (ψ^{ℓ} denotes number of elements of vector ψ)

$$
\begin{aligned}
f(\Theta|V,\nu) &= \mathcal{N}i\mathcal{G}_{\theta,r}(\hat{\theta},P,\hat{r},\nu) \equiv \mathcal{N}_{\theta}(\hat{\theta},rP)i\mathcal{G}_r(\hat{r},\nu) \\
&\equiv \frac{\exp\left\{-\frac{(\theta-\hat{\theta})'P^{-1}(\theta-\hat{\theta})+(\nu-2)\hat{r}}{2r}\right\}}{\mathcal{I}(V,\nu)\,r^{0.5(\nu+\psi^{\ell}+2)}} \qquad (14)\\
V &\equiv \begin{bmatrix} (\nu-2)\hat{r} & \hat{\theta}'P' \\ P\hat{\theta} & P^{-1} \end{bmatrix} \\
\mathcal{I}(V,\nu) &= \left[\frac{2}{(\nu-2)\hat{r}}\right]^{\frac{\nu}{2}} |2\pi P|^{0.5}\,\Gamma\left(\frac{\nu}{2}\right).
\end{aligned}
$$

This pdf is determined by a symmetric positive definite extended information matrix V – a direct counterpart of the array V in (7) – and by the scalar $\nu > 0$ interpreted as the number of degrees of freedom. The posterior pdf is also normal-inverse-gamma. The extended information matrix V_t and degree of freedom ν_t determining it are updated according to the recursive version of (7)

$$
V_t \equiv V_{t-1} + \underbrace{\tilde{\Psi}_t\tilde{\Psi}_t'}_{B(\tilde{\Psi}_t)}, \quad \nu_t = \nu_{t-1}+1, \qquad (15)
$$

where $\tilde{\Psi}_t$ is a data vector available at time t.

It can be shown [18] that $\hat{\theta}$, $\hat{r}$ and P are quantities well-known in connection with the recursive least squares (RLS). The recursion (15) is algebraically equivalent to RLS with initial values determining the prior pdf (14). The following correspondence holds (E and cov denote conditional expectation and covariance, respectively):

$$\begin{aligned}\hat{\theta}_t &\equiv \mathsf{E}[\theta|\tilde{d}(t)] \equiv \text{ RLS estimate of } \theta,\\ \hat{r}_t &\equiv \mathsf{E}[r|\tilde{d}(t)] = \frac{\text{RLS remainder}}{\nu_t - 2}, \quad \hat{r}P_t \equiv \mathsf{cov}[\theta|\tilde{d}(t)].\end{aligned}$$

This correspondence explains the "standard" choice of the prior pdf $f(\Theta)$, given by $\underline{V}$ and $\underline{\nu}$ and specified by $\underline{\hat{\theta}} = 0$, $\underline{P} =$ diagonal matrix with a large diagonal, $\underline{\hat{r}}$ and $\underline{\nu}$ are small positive numbers, [5]. This choice quantifies the assumption that θ and r are finite but knowledge of their values and relationships is very vague.

The technique developed here enriches this commonly acceptable practice by elaborating available expert knowledge expressed via a set of pdfs $\mathsf{F}_{\mathcal{T}}$ (5). Practically, the proposed approach provides better initial conditions of (10), i.e., better initial conditions of RLS.

III. PROCESSING OF COMMON TYPES OF PRIOR KNOWLEDGE

The processing presented below deals with the commonly available types of domain-specific knowledge. It constructs the mapping of knowledge piece on pdfs f_τ, $\tau \in \mathcal{T}$. It demonstrates typical ways of processing the knowledge and provides examples when equation (2) can be applied.

Section III-A deals with the prior knowledge of *ranges of data trajectories*, i.e., of ranges of data sequences *ordered* according to the "fictitious" time of the gedanken experiment considered. Knowledge of ranges of data trajectories is expressed via a uniform pdf on the set of highly expected data vectors. Then, equation (2) is applied.

Generally speaking a high number of commonly used types of prior knowledge about the system can be expressed in the above-mentioned way. A specific example of the approach is *static gain* quantification given in Section III-B. Section III-C describes a more complex example of prior knowledge processing, which concerns *rise time* and *dynamic time delay*. An exploitation of *obsolete, analogous and simulated data* is discussed in Section III-D. It reveals the need to prevent over-fitting of prior knowledge. The necessary balance between prior knowledge and observed data can be reached by using just-in-time-modelling methodology, Section III-D1. Alternatively, the weight $|\mathcal{T}|$ of the function $\Omega_{\mathcal{T}}(\Theta)$ in the exponent of (2) can be modified to $w \neq |\mathcal{T}|$. This solution is prepared in Section III-D2 and finalised in Section IV.

Quantification of *response's smoothness*, Section III-E, provides an example of widely accessible type of knowledge, whose processing requires Monte-Carlo-type evaluation.

The knowledge of *cut-off frequency*, Section III-F, and a *point on frequency response*, Section III-G, represent the cases allowing direct construction of the mapping of domain-specific knowledge pieces to pdfs $f_\tau, \tau \in \mathcal{T}$.

A. Ranges of data trajectories

Often, ranges of data trajectories are known from: (i) the system design phase; (ii) series of past experiments intended for estimation of particular system's characteristics (for example, step response) of the modelled system.

The ranges of data trajectories mean *ordered sequence of knowledge pieces* about data ranges. The data ranges induce ranges of data vectors Ψ_τ, $\tau \in \mathcal{T}$ (τ labels "fictitious time" of the gedanken experiment), with entries of data vector indexed by $i = 1, \ldots, \Psi^{\ell}$,

$$\Psi_\tau \in \mathrm{\Psi}_\tau \equiv \left[\underline{\Psi}_\tau, \overline{\Psi}_\tau\right] \quad \Leftrightarrow \quad \Psi_{\tau;i} \in \left[\underline{\Psi}_{\tau;i}, \overline{\Psi}_{\tau;i}\right]. \tag{16}$$

The vectors' ranges are determined by the lower $\underline{\Psi}_\tau$ and upper $\overline{\Psi}_\tau$, $\tau \in \mathcal{T}$, boundary values as follows:

$$\begin{aligned} \underline{\Psi}_\tau &= [\underline{\Psi}_{\tau;1}, \ldots, \underline{\Psi}_{\tau;\Psi^{\ell}}]', \\ \overline{\Psi}_\tau &= [\overline{\Psi}_{\tau;1}, \ldots, \overline{\Psi}_{\tau;\Psi^{\ell}}]', \end{aligned}$$

where the entries $\underline{\Psi}_{\tau;i}$ and $\overline{\Psi}_{\tau;i}$ are finite. The knowledge of the ranges is expressed via the uniform pdf on domain of Ψ

$$f_\tau(\Psi) = \mathcal{U}_\Psi\left(\mathrm{\Psi}_\tau\right) \equiv \prod_{i=1}^{\Psi^{\ell}} \frac{\chi_{[\underline{\Psi}_{\tau;i}, \overline{\Psi}_{\tau;i}]}(\Psi_i)}{\overline{\Psi}_{\tau;i} - \underline{\Psi}_{\tau;i}} \equiv \frac{\chi_{\mathrm{\Psi}_\tau}(\Psi)}{\int_{\mathrm{\Psi}_\tau} d\Psi},$$

with the indicator function $\chi_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$

For EF, the key quantity Υ_τ (7) becomes

$$\Upsilon_\tau = \int B(\Psi)\,\mathcal{U}_\Psi\left(\mathrm{\Psi}_\tau\right)\, d\Psi, \ \tau \in \mathcal{T}. \tag{17}$$

For the normal ARX model (13), the increment Υ_τ (17) of the extended information matrix reads

$$\Upsilon_\tau = \int \Psi\Psi' \mathcal{U}_\Psi\left(\Psi_\tau\right) d\Psi = \frac{1}{4}\left(\overline{\Psi}_\tau + \underline{\Psi}_\tau\right)\left(\overline{\Psi}_\tau + \underline{\Psi}_\tau\right)'$$
$$+\frac{1}{12}\mathsf{diag}\left[\left(\overline{\Psi}_{\tau;1} - \underline{\Psi}_{\tau;1}\right)^2, \ldots, \left(\overline{\Psi}_{\tau;\Psi^\ell} - \underline{\Psi}_{\tau;\Psi^\ell}\right)^2\right]. \tag{18}$$

B. Static gain

Static gain, rise time and dynamic delay characterise a system's response to a change from an equilibrium. The *static gain* of a system is the difference between the former value of system's output and its terminal value when the system's response reaches a new steady state after the system's input had been changed.

Static gain is typical information available in different application domains and its quantification was repeatedly addressed [7], [8]. The knowledge of the static gain's ranges $g \in \left[\underline{g}, \overline{g}\right]$ can be interpreted as a result of the following gedanken experiment: (i) let the inspected scalar system's output and system's input be at their initial constant levels y_1 and u_1, respectively; (ii) a step change Δ is applied to the system's input; (iii) when the system's response reaches a new steady state, the system's output stabilises within the interval $\left[y_1 + \underline{g}, y_1 + \overline{g}\right]$, i.e., the static gain $g \in \left[\underline{g}, \overline{g}\right]$.

This knowledge can be expressed via ranges of data vectors as follows. The initial data vector Ψ_1 expresses non-random knowledge of the initial constant value of the output y_1 and the input u_1 corresponding to an equilibrium state. Values of the input $u_1 + \Delta$ together with values of the stabilised output $y \in \left[y_1 + \underline{g}, y_1 + \overline{g}\right]$ determine the range of the terminal data vector $\Psi_2 \in \left[\underline{\Psi}_2, \overline{\Psi}_2\right]$. The pdfs expressing this knowledge are

$$f_1(\Psi) = \delta(\Psi - \Psi_1) \text{ and } f_2(\Psi) = \mathcal{U}_\Psi\left(\left[\underline{\Psi}_2, \overline{\Psi}_2\right]\right).$$

The processing coincides with that of data ranges.

For example a single-input, single-output ARX model with the state in the phase form has the regression vector

$$\psi_t' \equiv [y_{t-1}, \ldots, y_{t-n}, u_t, \ldots, u_{t-m}], \tag{19}$$

where $n \geq 0$ and $m \geq 0$; $n, m \in \mathbf{Z}$.

For this case data vectors Ψ_1, $\underline{\Psi}_2$ and $\overline{\Psi}_2$ have the form

$$\begin{aligned}
\Psi_1 &= [y_1 \quad \ldots \; y_1 \quad u_1 \quad \ldots \; u_1]' \\
\underline{\Psi}_2 &= [y_1 + \underline{g} \; \ldots \; y_1 + \underline{g} \quad u_1 + \Delta \; \ldots \; u_1 + \Delta]' \\
\overline{\Psi}_2 &= [\underbrace{y_1 + \overline{g} \; \ldots \; y_1 + \overline{g}}_{(n+1) \text{ times}} \quad \underbrace{u_1 + \Delta \; \ldots \; u_1 + \Delta}_{(m+1) \text{ times}}]'.
\end{aligned}$$

Equations (7), (18) give the following form of the extended information matrix reflecting the processed prior knowledge of the static gain ($|\mathcal{T}| = 2$ in (7))

$$V = \frac{1}{2}\left\{ \tilde{\Psi}_1 \tilde{\Psi}_1' + \frac{1}{4}\left(\overline{\Psi}_2 + \underline{\Psi}_2\right)\left(\overline{\Psi}_2 + \underline{\Psi}_2\right)' \right.$$

$$\left. + \frac{1}{12} \text{diag}\Big[\underbrace{\left(\overline{g} - \underline{g}\right)^2, \ldots, \left(\overline{g} - \underline{g}\right)^2}_{(n+1) \text{ times}}, \; \underbrace{0, \ldots, 0}_{(m+1) \text{ times}} \Big] \right\}.$$

C. Rise time and dynamic delay

Rise time, ${}^{r}\tau$, refers to the time required for a system's output to rise from a specified low value to a specified high value of the final steady-state value of the desired system's response. *Dynamic delay*, ${}^{d}\tau$, is the time required for a system's response to change from zero to a small non-zero value specified.

Uncertain knowledge of rise time and dynamic delay can be converted to knowledge of data ranges in the following way.

- Both rise time and dynamic delay are gained on the system when deterministic step is applied to its input. Thus, the input forms the non-random part of the constructed data vectors, cf. (11). The non-random values are used for the construction of the increment Υ in (7) according to (12).
- The absolute value of the system's output is negligible until time ${}^{d}\tau$. The negligibility means the output is highly expected to be smaller than k times ($k \approx 0.1$) the guess of the static gain $\hat{g}$ (for clarity $\hat{g} > 0$), i.e.

$$y \in [-k \times \hat{g}, k \times \hat{g}], \quad \text{for } \tau \leq {}^{d}\tau. \tag{20}$$

- The value of the output within a time interval characterised by system's rise time ${}^{r}\tau$ is determined by

$$y \in [\underline{y}_{\tau}, \overline{y}_{\tau}], \; \text{for } \tau \in \left({}^{d}\tau, {}^{r}\tau\right]. \tag{21}$$

- For the time span $\tau > {}^{r}\tau$, the output y_τ is expected to be in the ranges

$$y \in \left[\max\Big((1-k)\times\hat{g},\ \underline{y}_{\tau}\Big),\ \min\Big((1+k)\times\hat{g},\ \overline{y}_{\tau}\Big)\right]. \tag{22}$$

Expressions (20)–(22) (cf. (16)) specify the ranges of data trajectories and are determined by the system's properties known to an expert. This specification allows direct application of the results obtained in Section III-A.

D. Data-base knowledge

Available realisations $\tilde{\Psi}_\tau$ of data vectors Ψ_τ, $\tau \in \mathcal{T}$ (τ refers to the τth item in a data base) can often serve as prior knowledge. The probabilistic description of this knowledge is $f_\tau(\Psi) \equiv \delta(\Psi - \tilde{\Psi}_\tau)$ and use of equation (2) reduces to ordinary Bayes estimation. This is a correct solution, if the realisations are obtained on the modelled system and in ordinary operational mode. The situation differs, if the realisations are: (i) obsolete; (ii) observed on a similar system; (iii) observed under significantly different operation conditions; (iv) obtained via simulation. Then, this knowledge has to be used carefully as the prior pdf may *practically* shrink at a wrong set, and the real data observed will not be able to change the result.

The problem is not critical when the number of processed data vectors is small and real data is informative [18]. Then, equation (2), reduced to the Bayes rule, can be directly applied. If the conditions are violated, two approaches are considered here: i) real-time selection of the informative past data, which are closely related to the current system's state; and ii) real-time weighting of the data processed to control the influence of the knowledge incorporated. Detailed description of both approaches is proposed below.

1) Real-time selection of representative data: The methodology called (among others) just-in-time modelling, e.g. [20], [21] can counteract the dangerous shrinking. This methodology assumes the ability to store and inspect a large number of data vectors in real time $t \in \mathsf{T}$. With them, the local model is built "just-in-time" as follows.

- Current observations are put in the regression vector $\tilde{\psi}_t, t \in \mathsf{T}$.
- A small number $|\mathcal{T}|$ of data vectors $\{\tilde{\Psi}_\tau\}_{\tau\in\mathcal{T}}$ with the regression vectors $\{\tilde{\psi}_\tau\}_{\tau\in\mathcal{T}}$ "close" to the observed one $\tilde{\psi}_t$ are selected from a data base. Here the subscript t refers to the real time, while τ to the τth piece of data-base knowledge serving as prior knowledge at time t.

- A local model at time t is fitted to the data vectors $\{\tilde{\Psi}_\tau\}_{\tau\in\mathcal{T}}$ corresponding to the regression vectors $\{\tilde{\psi}_\tau\}_{\tau\in\mathcal{T}}$ selected . By other words, (2) is applied with pdfs $f_\tau \in \mathsf{F}_\mathcal{T}$ being Dirac functions placed on the data vectors $\{\tilde{\Psi}_\tau\}_{\tau\in\mathcal{T}}$.
- The resulting model, delivered just-in-time, is used for predicting unknown value of the output y_t, i.e., for computing characteristics of the predictive pdf :
$f(y_t|\tilde{\psi}_t, \mathsf{F}_\mathcal{T}) \equiv \int M([y_t, \tilde{\psi}_t']', \Theta) f(\Theta|\mathsf{F}_\mathcal{T})\, d\Theta$.

The outlined idea is quite powerful if the modelled relation is smooth. It may, however, be sensitive to the definition of the closeness of regression vectors $\tilde{\psi}_t$ and $\tilde{\psi}_\tau$. The advocated probabilistic interpretation offers the following systematic approach.

The approach considers acceptance of the *natural conditions of control*, [18], which postulate that knowledge of the regression vector without knowledge of the corresponding output does not enrich knowledge about Θ, i.e.,

$$f(\Theta|\tilde{\psi}_t) = f(\Theta|\tilde{\psi}_\tau) = f(\Theta). \tag{23}$$

The regression vectors $\tilde{\psi}_\tau$ and $\tilde{\psi}_t$ can be assumed sufficiently close for the given task if the joint pdf of unknown output y_t and unknown finite-dimensional parameter Θ, determined by the vector $\tilde{\psi}_\tau$ selected from the data base, is close to that determined by the vector $\tilde{\psi}_t$ observed at time t, i.e.

$$f(y_t, \Theta|\tilde{\psi}_t) \approx f(y_t, \Theta|\tilde{\psi}_\tau). \tag{24}$$

The joint pdf $f(y_t, \Theta|\tilde{\psi}_t)$ can be rewritten in the following way (the validity of the second equality sign is lent by (23)):

$$f(y_t, \Theta|\tilde{\psi}_t) \equiv M([y_t, \tilde{\psi}_t']', \Theta)\, f(\Theta|\tilde{\psi}_t) = M([y_t, \tilde{\psi}_t']', \Theta)\, f(\Theta).$$

Similarly, pdf conditioned by $\tilde{\psi}_\tau$ reads:

$$f(y_t, \Theta|\tilde{\psi}_\tau) \equiv M([y_t, \tilde{\psi}_\tau']', \Theta) f(\Theta|\tilde{\psi}_\tau) = M([y_t, \tilde{\psi}_\tau']', \Theta) f(\Theta).$$

Under weak conditions [22], the Kullback-Leibler divergence [23] and its affine modifications having the same minimiser can serve as adequate measures of the inspected proximity. Under (23), the desired divergence takes the form

$$\mathcal{D}_{t\tau} = \int M([y_t, \tilde{\psi}_t']', \Theta) f(\Theta) \ln \left(\frac{M([y_t, \tilde{\psi}_t']', \Theta)}{M([y_t, \tilde{\psi}_\tau']', \Theta)} \right) dy_t d\Theta. \tag{25}$$

Thus, at time t, the desired data vectors $\Psi'_\tau = [\tilde{y}_\tau, \tilde{\psi}'_\tau]$ in the data base are those having the regression vectors $\tilde{\psi}_\tau$, $\tau \in \mathcal{T}$, which yield small values of the divergence $\mathcal{D}_{t\tau}$ (25). For EF and the conjugate prior pdf $f(\Theta)$, given by (9) with $V = \underline{V}$ and $\nu = \underline{\nu}$, $\mathcal{D}_{t\tau}$ becomes

$$\begin{aligned}\mathcal{D}_{t\tau} &= \int \frac{A^{\underline{\nu}+1}(\Theta)}{\mathcal{I}(\underline{V}, \underline{\nu})} \exp \left\langle \underline{V} + B([y_t, \tilde{\psi}'_t]'), C(\Theta) \right\rangle \\ &\times \left\langle B([y_t, \tilde{\psi}'_t]') - B([y_t, \tilde{\psi}'_\tau]'), C(\Theta) \right\rangle dy_t d\Theta.\end{aligned}$$

For a single-output normal ARX model (13), $\mathcal{D}_{t\tau}$ reads

$$\begin{aligned}\mathcal{D}_{t\tau} &= \int \frac{[\theta'(\tilde{\psi}_t - \tilde{\psi}_\tau)]^2}{2r} \mathcal{N}i\mathcal{G}_{\theta,r}(\underline{V}, \underline{\nu})\, d\theta dr \\ &= \frac{1}{2}\left[\frac{\underline{\nu}}{(\underline{\nu}-2)\hat{\underline{r}}}[\hat{\underline{\theta}}'(\tilde{\psi}_t - \tilde{\psi}_\tau)]^2 + (\tilde{\psi}_t - \tilde{\psi}_\tau)'\underline{P}(\tilde{\psi}_t - \tilde{\psi}_\tau)\right].\end{aligned} \tag{26}$$

RLS quantities $\hat{\underline{\theta}}$, $\hat{\underline{r}}$ and $\underline{P}$ are defined by (14) with $V = \underline{V}$ and $\nu = \underline{\nu}$. The result (26) follows from the basic properties of the normal and normal-inverse-gamma pdfs, see e.g. [10].

Let us consider (26). The first term in the square brackets is proportional to the normalised squared difference of outputs' predictions. Hence the values larger than one cannot be considered sufficiently small for it. The second term in the square brackets, (26), is proportional to the squared norm of $(\tilde{\psi}_t - \tilde{\psi}_\tau)$ weighted by a matrix $\underline{P}$. The equations (14) and (15) imply the matrix $\underline{P}$ can be interpreted as an inversion of the second moment of regression vectors divided by $\underline{\nu}$. Hence, the values larger than $\psi^\ell/\underline{\nu}$, (with ψ^ℓ is number of elements in ψ) cannot be taken as small. The discussion above indicates that a justified threshold, deciding on whether $\mathcal{D}_{t\tau}$ is small enough, i.e. whether $\tilde{\psi}_\tau$ is sufficiently close to $\tilde{\psi}_t$, can be found.

2) Weighting of knowledge-expressing data: Above, the regression vector ψ_t, around which the local model is built, is determined by real-time observations $\tilde{\psi}_t, t \in \mathrm{T}$. In other cases, a general rule for selecting "representative data" is missing and existing solutions, e.g. [24], are case-dependent. The problem is that a huge amount of data vectors has to be processed without a guide how to discard them individually. Consequently, processed prior knowledge can be "over-fitted" and influence of real observations can be diminished.

The problem applies to any data source, but it becomes especially important if the discussed data sample is generated by simulation models. Despite these models accumulate a substantial expert knowledge, their use in the subsequent choice of decision strategies is limited. The reason for that is required simplification of these models as the corresponding optimising design is often unfeasible.

Adaptive systems supported here rely on approximate models, too. They optimise decision strategy in real time by using a recursively estimated model from a tractable class of models. The approximation is constructed implicitly via Bayesian estimation, which guarantees that the asymptotically best approximation of a modelled system [10]. The learning transient can be substantially shortened, if the knowledge accumulated by the simulation model is projected onto the prior pdf of unknown parameter Θ. One possibility is to apply Bayes rule to simulated data vectors that is equivalent to (2) with the average pdf (4) equal to their sample pdf. However the result over-fits the projected knowledge and the data observed can hardly modify its inevitable flaws.

Therefore $\Omega_{\mathcal{T}}(\Theta)$ in (2) cannot be weighted by $|\mathcal{T}|$ - number of the processed simulated samples. The gained prior pdf should be flatten appropriately [10] and the weight should be $w \leq |\mathcal{T}|$.

For EF, the incorporation of the knowledge reflected in simulated data vectors implies the following processing.

- Collect the sample version of the *normalised array* $V = \frac{1}{|\mathcal{T}|}\sum_{\tau\in\mathcal{T}} \Upsilon_{\tau}$, see (7).
- Select the weight $w \in [0, |\mathcal{T}|)$, and create prior values V_0 and ν_0 according to (10) with w replacing $|\mathcal{T}|$, i.e.,

$$V_0 \equiv \underline{V} + wV, \;\; \nu_0 \equiv \underline{\nu} + w. \tag{27}$$

This weighting diminishes an influence of the factor modifying $f(\Theta)$ in (2) and controls the impact of the incorporated knowledge piece. An automatic choice of the weight $w \geq 0$ is proposed in Section IV.

E. Smoothness of system's response

Smoothness of the system's response is frequently available type of knowledge about the system. It can be expressed by a set of restrictions describing highly expected data trajectories gained in gedanken experiment ($\tau \in \mathcal{T}$ is time index of the fictitious time in the experiment)

$$d_{\tau} \in \mathsf{d}_{\tau|\tau-1} \equiv \left\{ d_{\tau} : |y_{\tau} - y_{\tau-1}| \leq q_{\tau} ||\psi_{\tau} - \psi_{\tau-1}|| \right\}. \tag{28}$$

The set (28) depends on $\Psi_{\tau-1}$ and it is parameterised by a continuity module $q_{\tau} > 0$ and the considered norm $|| \cdot ||$.

Similarly to the previous cases, this type of knowledge can be expressed via uniform pdfs on $\mathsf{d}_{\tau|\tau-1}$ restricted by (16). The pdfs together with the known deterministic mapping $(\Psi_{\tau-1}, d_\tau) \rightarrow \Psi_\tau$ determine the *conditional* pdfs $f_{\tau|\tau-1}(\Psi|\tilde{\Psi})$, describing the highly expected transitions $\tilde{\Psi}_{\tau-1} \rightarrow \tilde{\Psi}_\tau$, for $\tau \geq 2$. The pdf $f_1(\Psi)$ describing the expected initial data vector can be chosen by using, for instance, available knowledge of data ranges.

This construction represents cases when pdfs in the set $\mathsf{F}_\tau \equiv \{f_\tau(\Psi)\}_{\tau \in \mathcal{T}}$ are *given implicitly* as solutions of the equations

$$f_\tau(\Psi) = \int f_{\tau|\tau-1}(\Psi|\tilde{\Psi}) f_{\tau-1}(\tilde{\Psi})\, d\tilde{\Psi}, \ \tau = 2, \ldots, |\mathcal{T}|.$$

An explicit solution of these equations can hardly be obtained. The underlying conditional pdfs are, however, simple and Monte Carlo methodology can be applied. It draws random independent samples from $f_1(\Psi)$ and simulates realisations $\tilde{\Psi}_2, \ldots, \tilde{\Psi}_{|\mathcal{T}|}$ by drawing samples of data records uniformly distributed on $\mathsf{d}_{\tau|\tau-1}$. For EF, these realisations serve for evaluating the sample version of Υ_τ in (7).

The conceptual algorithm for EF is as follows.

Algorithm 1 (Evaluation of knowledge of the smoothness):
`Initial phase`

- Select the member of EF (6) determined by the functions $A(\Theta)$, $C(\Theta)$ and $B(\Psi)$.
- Choose a conjugate prior pdf (9) given by $\underline{V}$ and $\underline{\nu}$.
- Set the array $V = \underline{V}$.
- Specify ranges of data vectors Ψ_τ (16) and the continuity modules q_τ, $\tau \in \mathcal{T}$ as well as the norm $||\cdot||$, see (28).
- Select the number of runs $|k|$ and set $k = 0$.

`Iterative phase`

1) Set $k := k + 1$.
2) Generate the sample $\tilde{\Psi}_1 \sim f_1(\Psi)$.
3) Increment $V := \frac{k-1}{k} V + \frac{1}{k} B\left(\tilde{\Psi}_1\right)$.
4) Generate samples $\tilde{d}_\tau$ uniformly distributed on $\mathsf{d}_{\tau|\tau-1}$, determine sample $\tilde{\Psi}_\tau$ from $\tilde{\Psi}_{\tau-1}$ and $\tilde{d}_\tau$, and increment $V := V + \frac{1}{k} B\left(\tilde{\Psi}_\tau\right)$ for $\tau = 2, \ldots, |\mathcal{T}|$.
5) Stop, if $k \geq |k|$ or other stopping criterion is met. Otherwise, go to the beginning of Iterative phase.

6) Normalise the final $V := \frac{1}{|\mathcal{T}|}V$ and use it for the definition of the initial conditions in (10) with $|\mathcal{T}|$ replaced by the weight w according to (27).

F. Cut-off frequency

This section deals with an important case, when the result of the gedanken experiment can be evaluated analytically and the knowledge piece is expressed via directly constructed information matrix V (7). This case concerns, for example, knowledge of *cut-off frequency* $\omega_c \in (0, 2\pi)$. The term cut-off frequency refers to a boundary in a system's frequency response and represents the smallest frequency of the sinusoidal input at which the system's output begins to be almost zero. The presentation is made for the normal single-input, single-output ARX model with the phase-form regression vector (19).

Knowledge of the cut-off frequency ω_c can be manifested in the gedanken experiment when the system's input is sinusoidal with frequency $\omega_c \leq \omega < 2\pi$ and the corresponding system's output is almost zero. Thus for the fixed frequency $\omega \in [\omega_c, 2\pi)$, the highly-expected fictitious data vectors at time $\tau \in \mathcal{T}$ are as follows ($m \geq 0$ and $n \geq 0$):

$$\tilde{\Psi}_{\omega,\tau} \equiv [\tilde{y}_\tau, \ldots, \tilde{y}_{\tau-n}, \sin(\omega\tau), \ldots, \sin(\omega(\tau - m))]'. \tag{29}$$

In (29) the involved outputs $(\tilde{y}_\tau, \ldots, \tilde{y}_{\tau-n})$ have zero mean, negligible correlations and a small variance r. Note the subscript ω indicates the considered fixed frequency $\omega \in [\omega_c, 2\pi)$, while τ refers to the time index of fictitious data, i.e. data would be observed if the experiment conducted in reality.

Generally the extended information matrix can be split into blocks in the following way:

$$V = \begin{bmatrix} R & T \\ T' & S \end{bmatrix}. \tag{30}$$

where R, T, S are matrices of dimensions $(n+1, n+1), (m+1, m+1)$ and $(n+1, m+1)$, respectively. Then the constructed extended information matrix V_ω, computed as the sample mean evaluated for the data vectors (29), can be written

$$V_\omega \equiv \lim_{|\mathcal{T}| \to \infty} \frac{1}{|\mathcal{T}|} \sum_{\tau=1}^{|\mathcal{T}|} \tilde{\Psi}_{\omega,\tau} \tilde{\Psi}'_{\omega,\tau} = \begin{bmatrix} rI_{n+1} & 0 \\ 0 & 0.5S_\omega \end{bmatrix}, \tag{31}$$

where I_n is the unit matrix of the order n and S_ω is the corresponding $(n+1, m+1)$-block of the decomposition (30) for the fixed frequency ω. Using the complex form of goniometric functions with j denoting imaginary unit, the (k,l)-element of the matrix S_ω with $k,l \in \{1,\dots,m+1\}$ can be written as follows

$$\begin{aligned}
S_\omega(k,l) &= 2 \lim_{|\mathcal{T}|\to\infty} \frac{1}{|\mathcal{T}|} \sum_{\tau=1}^{|\mathcal{T}|} \sin(\omega(\tau-k))\sin(\omega(\tau-l)) \\
&=- \lim_{|\mathcal{T}|\to\infty} \frac{1}{2|\mathcal{T}|} \sum_{\tau=1}^{|\mathcal{T}|} (\exp(j\omega(\tau-k)) - \exp(-j\omega(\tau-k))) \\
&\quad \times (\exp(j\omega(\tau-l)) - \exp(-j\omega(\tau-l))) \\
&= \frac{1}{2}\left[\exp(j\omega(k-l)) + \exp(-j\omega(k-l))\right] - \\
&\lim_{|\mathcal{T}|\to\infty} \frac{1}{2|\mathcal{T}|} \sum_{\tau=1}^{|\mathcal{T}|} [\exp(j\omega(2\tau-k-l)) + \exp(-j\omega(2\tau-k-l))] \\
&= \cos(\omega(k-l)). \qquad (32)
\end{aligned}$$

The last limit is zero, as it represents a bounded sum of the geometric sequences divided by $|\mathcal{T}|$.

The expressed knowledge is valid for any fixed frequency $\omega \in [\omega_c, 2\pi)$. Easy to see that all information matrices V_ω, indexed by $\omega \in [\omega_c, 2\pi)$ provide complementary knowledge pieces about the cut-off frequency ω_c. Hence the formal correspondence $\tau \leftrightarrow \omega$ and $\Upsilon_\tau \leftrightarrow V_\omega$ applied to equation (7) implies that an average of V_ω, $\omega \in [\omega_c, 2\pi)$ can adequately represent all these knowledge pieces. The desired extended information matrix V computed by averaging reads

$$V \equiv \int S_\omega \mathcal{U}_\omega([\omega_c, 2\pi])\, d\omega = \begin{bmatrix} rI_{n+1} & 0 \\ 0 & 0.5S \end{bmatrix}, \qquad (33)$$

$$S(k,l) = \begin{cases} 1 & \text{if } k=l, \\ -\frac{\sin(\omega_c|k-l|)}{|k-l|} & \text{if } k \neq l \end{cases} \quad k,l \in \{1,\dots,m+1\},$$

where $\mathcal{U}_\omega([\omega_c, 2\pi])$ is uniform pdf on $[\omega_c, 2\pi]$.

G. A point on frequency response

Knowledge of cut-off frequency is a special case of a partial knowledge of the system's frequency response. This knowledge can be available at least in connection with auto-tuners [25].

Recalling that the frequency response is the system's reaction to the sinusoidal input, let us assume a relevant gedanken experiment on the normal single-input, single-output ARX model with the regression vector in the phase form (19). Then the corresponding stationary system's response at time $\tau \in \mathcal{T}$ to the sinusoidal input with the frequency ω provides the fictitious data vectors ($m \geq 0$ and $n \geq 0$):

$$\tilde{\Psi}_{\omega,\tau} \equiv [a\sin(\omega\tau + \phi) + e_\tau, \ldots, a\sin(\omega(\tau - n) + \phi) + e_{\tau-n},$$
$$\sin(\omega\tau), \ldots, \sin(\omega(\tau - m))], \tag{34}$$

where the mutually uncorrelated noise elements e_τ have zero mean and the expert-specified variance r. In (34) the subscript ω indicates the considered frequency, while τ refers to the time index of fictitious data, i.e. data would be observed if the experiment conducted in reality.

The amplitude a represents the basic prior knowledge supplied. The phase shift $\phi \in [\underline{\phi}, \overline{\phi}] \subset [0, 2\pi]$ is another, usually more vague, part of this knowledge. For a fixed frequency ω and phase shift ϕ, the extended information matrix (7), denoted $V_{\omega\phi}$, coincides with the following sample moment evaluated for the data vectors (34)

$$\begin{aligned} V_{\omega\phi} &\equiv \lim_{|\mathcal{T}| \to \infty} \frac{1}{|\mathcal{T}|} \sum_{\tau=1}^{|\mathcal{T}|} \tilde{\Psi}_{\omega,\tau} \tilde{\Psi}'_{\omega,\tau} \\ &= \begin{bmatrix} rI_{n+1} + 0.5a^2 R_\omega & 0.5aT_{\omega\phi} \\ 0.5aT'_{\omega\phi} & 0.5S_\omega \end{bmatrix}, \end{aligned} \tag{35}$$

where matrices R_ω, S_ω and $T_{\omega\phi}$ are obtained via the decomposition (30). The entries of R_ω and S_ω are defined by (32) for the dimensions $(n+1, n+1)$ and $(m+1, m+1)$, respectively.

The (k, l)th entry of $(n+1, m+1)$-matrix $T_{\omega\phi}$ equals

$$T_{\omega\phi}(k, l) = \cos(\omega|k - l| + \phi), \tag{36}$$

with $k \in \{1, \ldots, n+1\}$ and $l \in \{1, \ldots, m+1\}$.

Similarly to cut-off frequency (see Section III-F), the final extended information matrix can be computed by averaging out $V_{\omega\phi}$ over the possible phase shifts $\phi \in [0, 2\pi)$. In the special, most uncertain, case when no knowledge of the phase shift $\phi \in [0, 2\pi)$ is available, the extended information matrix equals

$$V_\omega = \frac{1}{2\pi} \int_0^{2\pi} V_{\omega\phi}\, d\phi = \begin{bmatrix} rI_{n+1} + 0.5a^2 R_\omega & 0 \\ 0 & 0.5S_\omega \end{bmatrix}. \tag{37}$$

IV. WEIGHTS OF THE KNOWLEDGE PIECES

The influence of prior knowledge depends on the weight w with which the normalised array V, representing the processed pdfs $\mathsf{F}_{\mathcal{T}} = \{f_\tau(\Psi)\}_{\tau \in \mathcal{T}}$, is added to $\underline{V}$, see (27). The choice of w is critical issue for a balanced weighting of prior knowledge and information brought by data observed. It becomes even more critical, when several knowledge pieces, given by the collection of pdfs $\mathsf{F}_{\mathcal{T}_p} \equiv \{f_{\tau;p}(\Psi)\}_{\tau \in \mathcal{T}_p}$, $p \in \mathsf{P} \equiv \{1, \ldots, |\mathsf{P}|\}$, are to be combined. These pieces of knowledge may i) concern different aspects of the modelled system and be provided by one expert, or ii) reflect the same system's property but be provided by different experts. In order to solve this problem and reach the balanced combination of different knowledge pieces, the constructed functions $\Omega_{\mathcal{T}_p}$ (3) are weighted by $w_p \geq 0$, $p \in \mathsf{P}$, cf. (27). The weights are chosen under the following conditions.

- The parametric model belongs to EF, so that functions $\Omega_{\mathcal{T}_p}$ (3) are determined by arrays $V_p \equiv \frac{1}{|\mathcal{T}_p|} \sum_{\tau \in \mathcal{T}_p} \Upsilon_{\tau;p}$, $p \in \mathsf{P}$, see (7).
- The values of weights are chosen *after* observing a sufficient number t_s of real, informative learning data, i.e. $\tilde{d}(t_s) = (\tilde{d}_{t_1}, \ldots, \tilde{d}_{t_s})$.

The term "sufficient number" formally means that at least one realisation $\tilde{\Psi}$ of data vector Ψ is available. Practically $\tilde{d}(t_s)$ must counteract poor robustness of the maximum likelihood estimates, see e.g. [26].

Under these weakly restrictive assumptions, the posterior pdf at any $t \geq t_s$ gets the form, cf. (6), (7), (27),

$$f\left(\Theta | \tilde{d}(t), \underline{\nu}, \underline{V}, V_1, \ldots, V_{|\mathsf{P}|}, w\right) \propto A(\Theta)^{\overbrace{\underline{\nu} + t}^{\underline{\nu}_t} + \sum_{p=1}^{|\mathsf{P}|} w_p}$$
$$\times \exp\left\langle \underbrace{\underline{V} + \sum_{t=1}^{t} B(\tilde{\Psi}_t)}_{\underline{V}_t} + \sum_{p=1}^{|\mathsf{P}|} w_p V_p, C(\Theta) \right\rangle. \qquad (38)$$

Note, that unlike τ refereing to either the τth piece of prior knowledge or time index of fictitious data, t refers to discrete time of *real* data observations.

In (38) the weight $w_p \geq 0, p \in \mathsf{P}$ determines the strength with which the pth knowledge item, expressed by V_p, is considered.

With the above notation, the addressed problem reduces to choice of the vector $w \equiv [w_1, \ldots, w_{|\mathsf{P}|}]'$ of non-negative numbers using the fixed knowledge $\underline{\nu}, \underline{V}, \underline{\nu}_t, \underline{V}_t, V_1, \ldots, V_{|\mathsf{P}|}$, $t \geq t_s$. For an instance of w, the predictive pdf, evaluated for the observed data $\tilde{d}(t)$, $t \geq t_s$ (t_s is a sufficient number of informative learning data observed), becomes, cf. (9), (38),

$$\begin{aligned} & f(\tilde{d}(t)|\underline{\nu}, \underline{V}, V_1, \ldots, V_{|\mathsf{P}|}, w) \\ & \quad = \frac{\mathcal{I}\left(\underline{V}_t + \sum_{p=1}^{|\mathsf{P}|} w_p V_p, \underline{\nu}_{|t|} + \sum_{p=1}^{|\mathsf{P}|} w_p\right)}{\mathcal{I}\left(\underline{V} + \sum_{p=1}^{|\mathsf{P}|} w_p V_p, \underline{\nu} + \sum_{p=1}^{|\mathsf{P}|} w_p\right)}. \end{aligned} \tag{39}$$

This predictive pdf is a likelihood function with respect to the unknown vector w. Rigorous Bayesian treatment would require assignment of a prior pdf over w and evaluation of the posterior pdf over their possible values. The related computational complexity makes us search for the maximum likelihood estimate of w for given $\underline{\nu}, \underline{V}, \underline{\nu}_t, \underline{V}_t, V_1, \ldots, V_{|\mathsf{P}|}, \tilde{d}(t)$, $t \geq t_s$ i.e., to maximise the predictive pdf (39) over $w_p \geq 0$, $p \in \mathsf{P}$. This choice respects the mentioned exceptional role of affine shifts of the Kullback-Leibler divergence [22]: the chosen maximum likelihood estimate minimises the Kerridge inaccuracy [27] of the empirical pdf – concentrated on the observed learning data $\tilde{d}(t)$, $t \geq t_s$ – on the optimised predictive pdf.

A rich set of optimisation procedures can be used for maximisation of the predictive pdf (39) as it is a nicely behaving function of the optimised weight. For instance, Hölder inequality implies that the logarithm of this predictive pdf is a difference of convex functions of w. Moreover, it has the ith derivative with respect to w, if the ith moment of $\ln(A(\Theta))$ and $C(\Theta)$, defining EF (6), exists for $w = 0$.

V. ILLUSTRATIVE EXAMPLES

The following normal ARX model (13) with scalar output y_t is considered, see Section II-B,

$$y_t = 1.81 y_{t-1} - 0.8187 y_{t-2} + 0.00468 u_t + 0.00438 u_{t-1} + e_t, \tag{40}$$

with white noise $e_t \sim \mathcal{N}_{e_t}(0, 10^{-4})$ and independent white exogenous scalar input $u_t \sim \mathcal{N}_{u_t}(0, 10^{-2})$. This is a discrete-time version of the continuous-time system with the transfer function $(1+s^2)^{-1}$ sampled with the period 0.1 sec.

The influence of incorporated prior knowledge is demonstrated by comparing the estimation results gained with and without use of prior knowledge. Each example contains the following steps:

- *Data generation* – a collection of $|t|$ learning data records are generated by model (40) for N realisations of noise e and input u.

- *Parameter estimation* – estimation, see Section II-B, is run twice: with and without prior knowledge. The runs without prior knowledge use the standard settings of the prior, Section II-B with diagonal of $\underline{P}$ equals to 10^6, $\underline{\hat{r}} = 10^{-4}$, $\underline{\nu} = 2$.
- *Evaluation of results* – the results achieved are judged according to time course of the regression-coefficients estimates and the prediction quality quantified by

$$Q \equiv \frac{\text{sample second moment of prediction errors}}{\text{variance of the noise } e_t \text{ in (40)}} \qquad (41)$$

  The quality is evaluated on validation data, generated after fixing ws in (27).

Example 1 illustrates influence of prior knowledge of a static gain $g \in [\underline{g}, \overline{g}] \equiv [0.9, 1.1]$ on the prediction. Processing steps:

- *Data generation* – a collection of $|t| = 200$ learning data records was generated by (40) for $N = 100$ different noise and input realisations.
- *Parameter estimation without prior knowledge* – the estimation run with the standard prior, see Section II-B.
- *Parameter estimation with prior knowledge* – the posterior pdf obtained from the learning data was combined with the prior knowledge of the static gain, represented by the second sum under the exponent in (38). The numerically computed optimal weight maximises the predictive pdf (39) evaluated for the learning data.
- *Evaluation of results* – an additional collection of 1000 validation data records was generated and used for evaluating the prediction quality (41).

The results are in Figure 1. The left subplot presents the optimal weights w computed for each of $N = 100$ noise and input realisations. The higher value of the weight, the more informative contribution and stronger influence of knowledge processed.

The right subplot of Figure 1 presents a histogram of the prediction quality differences(41) for the estimation with prior knowledge and without it. Therefore, the prediction with prior knowledge is worse if the difference presented is positive. The histogram confirms mainly positive influence of processed prior knowledge. Quantitatively, it is seen on a sample statistics of the prediction quality differences

```
 mean   median minimum maximum
-0.363 -0.240 -1.664  0.180
```

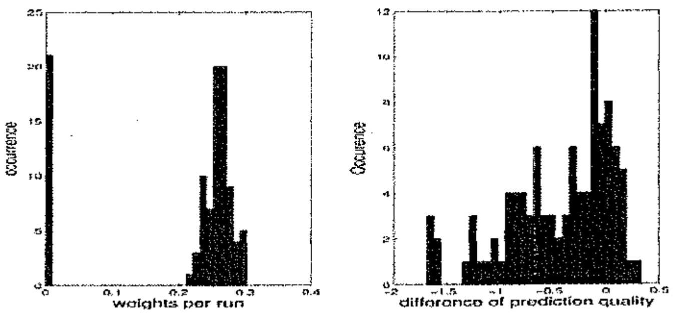

Fig. 1. Influence of prior knowledge of static gain: weights (27) (left) and histogram of differences of the prediction quality over realisations (right).

Example 2 illustrates influence of prior knowledge of a static gain $g \in [\underline{g}, \overline{g}] \equiv [0.9, 1.1]$ on point estimates of the regression coefficients in (40). Processing steps:

- *Data generation* – a collection of $|t| = 20$ learning input-output data was generated by (40) to initialise the estimation and to find the optimal weights. An additional collection of 150 validation data records was generated and used in the parameter estimation.
- *Parameter estimation without prior knowledge* – estimation run on 150 validation data records using the standard prior and learning data, see Section II-B.
- *Parameter estimation with prior knowledge* – estimation run on 150 validation data records using the prior pdf enriched by the knowledge of a static gain. The numerically computed optimal weight maximises the predictive pdf (39) evaluated for the learning data.
- *Evaluation of results* – the obtained time courses of the point estimates of the coefficient $b_0 = 0.00468$ at u_t (40) were recorded and compared for estimation: with and without prior knowledge.

The obtained results are in Figure 2. The left-hand subplot depicts the logarithm of the predictive pdf in learning data as a function of the optimised weight. The curve illustrates that the maximum can be simply found. The right-hand subplot, Figure 2, shows evolution of the b_0-estimates for both cases. The trajectory of b_0-estimates is smoother and closer to the true value of the regression coefficient with prior knowledge.

Example 3 illustrates incorporation of prior knowledge of data ranges and combination of several pieces of knowledge.

To select ranges properly, two independent data sets, of the length $|\mathcal{T}_p| = 50$, $p = 1, 2$ were

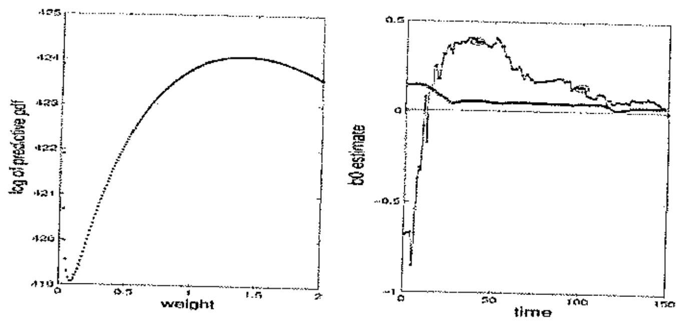

Fig. 2. Influence of prior knowledge of static gain: the logarithm of predictive pdf (39) as a function of the weight w (left) and time courses of b_0 estimate (right). The time course without prior knowledge is marked by circles. The straight line corresponds to the simulated value of the coefficient b_0.

generated by (40) for twenty noise and input realisations. The realistic ranges of data vectors $[\underline{\Psi}_{\tau;p}, \overline{\Psi}_{\tau;p}]$, $\tau \in \mathcal{T}_p$, $p = 1, 2$, were determined as envelopes of these simulated data sets. The respective arrays V_p were evaluated according to (7) and (18). Processing steps:

- *Data generation* – a collection of $|t| = 300$ learning data records was generated by (40) for $N = 100$ different noise and input realisations.
- *Parameter estimation without prior knowledge* – estimation run using the standard prior, see Section II-B.
- *Parameter estimation with prior knowledge* – estimation run using the standard prior combined with processed knowledge items and learning data. The numerically found optimal weights $[w_1, w_2]$, maximising (39), fix the impact of these knowledge pieces in (38).
- *Evaluation of results* – an additional collection of 1000 validation data records was generated and used for evaluating the prediction quality (41) .

The left subplot of Figure 3 shows the logarithm of the predictive pdf in a two-dimensional space of the weights $(w_1, w_2)'$. The maximum is marked by circle. The plot corresponds to the last noise and input realisations. The right subplot presents histogram of the differences of the prediction quality (41) for the estimation with prior knowledge and without it. Positive difference indicates the processed prior knowledge has worsened the prediction. The histogram confirms predominantly positive influence of the prior knowledge processed. Sample statistics of differences of the prediction quality is

```
mean     median minimum maximum
```

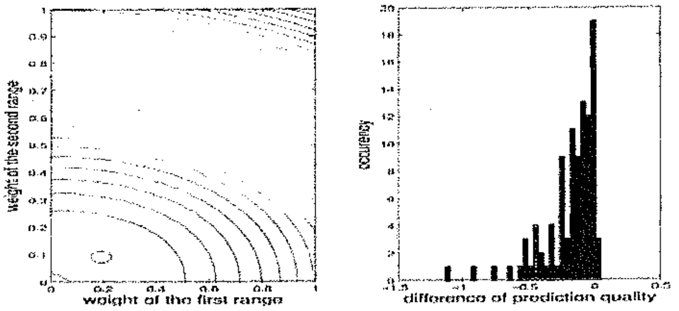

Fig. 3. Influence of the combined knowledge of data ranges: the contours the predictive-pdf logarithm (39) as a function of two-dimensional weight (left) and histogram of differences of the prediction quality over realisations (right).

```
-0.188  -0.121 -1.138  0.047
```

VI. CONCLUDING REMARKS

The paper concerns elicitation and quantification of prior knowledge frequently met in engineering domain. The adopted methodology works with prior knowledge expressed as a collection of pdfs on the space of data trajectories. Illustrative examples indicate a strong improvement of the estimation and prediction results implied by the proposed approach. The experiments confirm that it improves estimation of model structure as well as quality of adaptive control.

The presented methodology presents a further step towards the ultimate aim: facilitator-free incorporation of domain-specific knowledge into the prior pdf utilised by Bayesian estimation. A unified elicitation methodology, based on incorporating knowledge of data ranges, and objective way of mixing various knowledge pieces represent main progress.

The paper provides an approach, which i) covers a wide range of various knowledge types; ii) removes drawbacks and inconsistencies of the predecessors; iii) objectively determines a relative impact of the knowledge piece processed. Specifically, the treatment of data ranges does not rely on artificial models mimic to the estimated one so that the arbitrariness connected with them is removed. This applies to smoothness [7], rise time and dynamic delay [8].

An explicit introduction of the initial-level-fixing data vector $\tilde{\Psi}_1$ in knowledge quantification allows us to respect at least partially non-linear nature of the modelled system.

Quantification of the knowledge of the gain respects uncertainty of outputs in the regression vector unlike the old solution [8].

Treatment of cut-off frequency respects that the output diminishes for all frequencies behind it, unlike [8]. Concerning frequency response, uncertainty about the phase characteristic was not covered before. Moreover, if amplitudes are highly expected to be a given interval for a range of frequencies, it suffices to average V_ω (37) over this frequency range.

The foreseen open problems include: (i) the elicitation of knowledge provided by (possibly fuzzy) production rules; (ii) relaxation of the restrictive assumption on the uniform pdfs used; (iii) robustness analysis; and (iv) extensive real-life testing.

These technical steps will enhance the achieved conceptual and algorithmic improvements. The major progress expected is however in facilitator-free quantification of domain-specific *control aims*. It is achievable by applying the presented methodology to so-called ideal pdf, i.e pdf expressing control aims within the fully probabilistic design of control strategies [28].

REFERENCES

[1] J.O. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York, 1985.

[2] K.M. Osei-Bryson, "Supporting knowledge elicitation and consensus building for Dempster-Shafer decision models", *International Journal of Intelligent Systems*, vol. 18, no. 1, pp. 129–148, 2003.

[3] P.H. Garthwaite, J.B. Kadane, and A. O'Hagan, "Statistical methods for eliciting probability distributions", *Journal of the American Statistical Association*, vol. 100, no. 470, pp. 680–700, Jun 2005.

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

[5] L. Ljung, *System Identification: Theory for the User*, Prentice-Hall, London, 1987.

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

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

[8] M. Kárný, N. Khailova, P. Nedoma, and J. Böhm, "Quantification of prior information revised", *International Journal of Adaptive Control and Signal Processing*, vol. 15, no. 1, pp. 65–84, 2001.

[9] M. Kárný, P. Nedoma, N. Khailova, and L. Pavelková, "Prior information in structure estimation", *IEE Proceedings — Control Theory and Applications*, vol. 150, no. 6, pp. 643–653, 2003.

[10] M. Kárný, J. Böhm, T. V. Guy, L. Jirsa, I. Nagy, P. Nedoma, and L. Tesař, *Optimized Bayesian Dynamic Advising: Theory and Algorithms*, Springer, London, 2005.

[11] B. Betró and A. Guglielmi, "Methods for global prior robustness under generalized moment conditions", in *Robust Bayesian Analysis*, D. Rios-Insua and F. Rugerri, Eds., pp. 273 –294. Springer Verlag, New York, 2000.

[12] Abonyi J., Babuska R., Verbruggen H.B., and Szeifert F., "Incorporating prior knowledge in fuzzy model identification", *International Journal of Systems Science*, vol. 5, no. 31, pp. 657–667, 2000.

[13] Kosut R.L., "Iterative adaptive control: Windsurfing with confidence", in *Model Identification and Adaptive Control - From Windsurfing to Telecommunications*, Goodwin G.C., Ed. Springer-Verlag, London, UK, 2001.

[14] Hang C.C., Astrom K.J., and Wang Q.G, "Relay feedback auto-tuning of process controllers a tutorial review", *Journal of Process Control*, vol. 12, pp. 143–162, 2002.

[15] J. Kracík and M. Kárný, "Merging of data knowledge in Bayesian estimation", in *Proceedings of the Second International Conference on Informatics in Control, Automation and Robotics*, J. Filipe, J. A. Cetto, and J. L. Ferrier, Eds., Barcelona, September 2005, pp. 229–232, INSTICC.

[16] M. Kárný, J. Andrýsek, A. Bodini, T. V. Guy, J. Kracík, and F. Ruggeri, "How to exploit external model of data for parameter estimation?", *International Journal of Adaptive Control and Signal Processing*, vol. 20, no. 1, pp. 41–50, 2006.

[17] O. Barndorff-Nielsen, *Information and exponential families in statistical theory*, Wiley, New York, 1978.

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

[19] D. Rios-Insua and F. Rugerri, *Robust Bayesian Analysis*, Springer Verlag, New York, 2000.

[20] G. Bontempi, M. Birattari, and H. Bersini, "Lazy learning for local modelling and control design", *International Journal of Control*, vol. 72, no. 7–8, pp. 643–658, 1999.

[21] J.Y. Li, G.Z. Dong, K. Ramamohanarao, and L.S. Wong, "Deeps: A new instance-based lazy discovery and classification system", *Machine Learning*, vol. 54, no. 2, pp. 99–124, 2004.

[22] J. M. Bernardo, "Expected information as expected utility", *The Annals of Statistics*, vol. 7, no. 3, pp. 686–690, 1979.

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

[24] J.S. Edwards, T. Alifantis, R.D. Hurrion, J. Ladbrook, S. Robinson, and A. Waller, "Using a simulation model for knowledge elicitation and knowledge management", *Simulation Modelling Practice and Theory*, vol. 12, no. 7–8, pp. 527–540, 2004.

[25] D.W. Clarke, "Pretuning and adaptation of PI controllers", *IEE Proceedings-Control Theory and Applications*, vol. 150, no. 6, pp. 585–598, 2003.

[26] V.G. da Fonseca and N.R.J. Fieller, "Distortion in statistical inference: the distinction between data contamination and model deviation", *Metrika*, vol. 63, no. 2, pp. 169–190, 2006.

[27] D.F. Kerridge, "Inaccuracy and inference", *Journal of Royal Statistical Society*, vol. B 23, pp. 284–294, 1961.

[28] M. Kárný and T. V. Guy, "Fully probabilistic control design", *Systems & Control Letters*, vol. 55, no. 4, pp. 259–265, 2006.