Statistics and Its Interface Volume 7 (2014) 503–515

Fully probabilistic knowledge expression and incorporation

Miroslav Kárný, Tatiana V. Guy, Jan Kracík, Petr Nedoma, Antonella Bodini, and Fabrizio Ruggeri*

An exploitation of prior knowledge in parameter estimation becomes vital whenever measured data is not informative enough. Elicitation of quantified 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 complexity of modelled relationships that cannot be grasped by humans. This paper provides a facilitator-free approach based on an advanced knowledge-sharing methodology. It presents the approach on commonly available types of knowledge and applies the methodology to a normal controlled autoregressive model.

AMS 2000 subject classifications: Primary 62F15, 62C10; secondary 62M10.
Keywords and phrases: Bayesian estimation, Automatised knowledge elicitation, Just-in-time modelling, Controlled autoregressive model.

1. INTRODUCTION

An efficient use of prior knowledge influences quality of the decision making (DM) relying on estimated models. The considered Bayesian DM paradigm [5] combines observed data with prior knowledge quantified by a prior probability density function (pdf). Knowledge elicitation, i.e. mapping of prior knowledge onto the prior pdf is supported by a range of techniques [28, 12]. They generally rely on a facilitator, who guides the knowledge provider and quantifies knowledge gathered. The developed techniques mostly deal with societal and medical applications and focus on a quantification of experts' mental models [27]. The facilitator-dependent knowledge elicitation is expensive and can cope only with relatively simple cases. Moreover, it does not support exploitation of knowledge sources like simulation models or extensive data bases.

The current paper elaborates an elicitation technique that enables exploitation of all knowledge sources and weakly depends on a facilitator. The need to improve adaptive controllers and predictors based on recursive estimation [3, 26] has motivated this. Experience shows that the estimation of their structure as well as the resulting transient behaviour depend, sometimes critically, on the properly exploited prior knowledge.

The papers [14, 18, 19] proposed the desired elicitation technique that transforms knowledge into fictitious data, i.e. data that could be observed on a modelled system, and uses this data for the estimation as the real one. A formal structure of this transformation has also led to a treatment of the automatic elicitation as an optimisation under knowledge-reflecting constraints [16, 7]. The promising solutions are insufficiently general similarly as other rare attempts [2].

The estimation for control has not employed the potential of the knowledge elicitation: wild initial behaviour of adaptive controllers is at most handled via specific control strategies [21, 13].

The solution proposed here suppresses drawbacks and inconsistencies of its predecessors and covers a wider range of types of the elicited knowledge. Moreover, it determines a relative impact of provided knowledge, which increases robustness with respect to misleading knowledge. The approach is elaborated for models within the exponential family (EF), [4], that are widely used in applied adaptive systems as they enable an exact recursive estimation on extending data sets. The solution uses results [22, 15] that justify inclusion of probabilistically expressed knowledge about possible data into the parameter estimation. The paper provides the essence of the proposed knowledge elicitation. The elaborated cases (useful per se) offer a guide on how to apply it.

Section 2 summarises necessary preliminaries, namely, the adopted formula for processing probabilistic knowledge [15] and Bayesian estimation of the normal controlled autoregressive model (ARX). Section 3 describes an elicitation of prevalent types of prior knowledge. It shows how to use the basic knowledge about data ranges and how it can be applied to specific cases like rise time. The discussed use of extensive simulated or obsolete data sets reveals the need for selection of relevant data or their weighting. Quantification of knowledge on the system's response smoothness illustrates the use of Monte-Carlo evaluations. Some knowledge types can be elicited analytically as shown on partial knowledge of frequency response. Section 4 describes exploitation of real data to control an overall impact of the processed

*Corresponding author.

knowledge. Section 5 provides illustrative examples and Section 6 comments on the results obtained. Sections 3, 4 form the core of the paper.

2. PRELIMINARIES

The following notation is adopted. $\boldsymbol{x} = \{x_1, x_2, \ldots\}$ stands for a set of cardinality $|\boldsymbol{x}|$. If x is a vector, ℓ_x denotes its length and x' its transpose. The lower and upper bounds on $x \in \boldsymbol{x}$ are $\underline{x}$ and $\overline{x}$, respectively, and apply entry-wise to vectors. $f(x)$ is the pdf of a random variable with possible values x. ${}^{\mathrm{c}}x$ means a certain (realised) value of the random variable x. $\propto$ is a proportionality symbol. A subscript $t \in \boldsymbol{t} = \{1, 2, \ldots\}$ labels discrete time moments. $d(t)$ means a sequence $d_1, d_2, \ldots, d_t$. A subscript $\kappa \in \boldsymbol{\kappa} = \{1, 2, \ldots, |\boldsymbol{\kappa}|\}, |\boldsymbol{\kappa}| < \infty$, refers to the κth piece/source of prior knowledge. It can be represented by a large amount of "fictitious" data, i.e. data that could be observed on the modelled system. They are indexed by $\tau \in \boldsymbol{\tau} = \{1, 2, \ldots, |\boldsymbol{\tau}|\}, |\boldsymbol{\tau}| \leq \infty$.

2.1 Probabilistic description of knowledge

A closed loop formed by the modelled system and the adaptive DM system is considered. The data record $d_t = (y_t, u_t)$ observed at time $t \in \boldsymbol{t}$ consists of the system's output y_t and input u_t. The addressed parameter estimation concerns a time-invariant parametric model of the system. This model specifies the pdf of the scalar system's output y_t conditioned on a column regression vector ψ_t (of a length ℓ_ψ), which is a known function of u_t and $d(t-1)$. The model parameterised by a finite-dimensional unknown parameter Θ has the form

$$(1)\quad \begin{aligned} M(\Psi_t, \Theta) &= f(y_t|u_t, d(t-1), \Theta) = f(y_t|\psi_t, \Theta) \\ \Psi_t &= [y_t, \psi_t']'. \end{aligned}$$

The predicted system's output y_t and the regression vector ψ_t form a *data vector*, Ψ_t.

Knowledge of the unknown parameter Θ is initially described by a flat proper prior pdf $f(\Theta)$. Besides, available prior knowledge (possibly imprecise and incomplete) of some system's characteristic informs, often indirectly, about Θ and should be used in its estimation. Each of the processed knowledge sources is indexed by $\kappa \in \boldsymbol{\kappa}$, $|\boldsymbol{\kappa}| < \infty$. The proposed processing of all pieces of prior knowledge performs a gedanken experiment reflecting an underlying system's characteristic. The *possible* outcomes of this gedanken experiment are described by a pdf $f_\kappa(\Psi)$, where Ψ is a data vector composed of fictitious data, i.e. data which would be observed on the system if this experiment was performed in reality. Many data vectors, indexed by $\tau \in \boldsymbol{\tau}_\kappa = \{1, 2, \ldots, |\boldsymbol{\tau}_\kappa|\}$, $|\boldsymbol{\tau}_\kappa| \leq \infty$, can be considered when constructing $f_\kappa(\Psi)$. The set of knowledge-expressing pdfs

$$(2)\qquad \mathcal{K} = \{f_\kappa(\Psi),\ \Psi \in \boldsymbol{\Psi}_\kappa\}_{\kappa \in \boldsymbol{\kappa}}$$

is used for a modification of the (flat) prior pdf $f(\Theta)$ to a pdf denoted $f(\Theta|\mathcal{K})$. The pdf $f(\Theta|\mathcal{K})$ reflects the knowledge provided by $\mathcal{K}$ and serves for the subsequent standard Bayesian estimation as the prior pdf. To update $f(\Theta)$ to $f(\Theta|\mathcal{K})$, the formula proposed in [15] is adopted

$$(3)\quad \begin{aligned} f(\Theta|\mathcal{K}) &= \frac{f(\Theta)\exp\{\beta\,\Omega_{\mathcal{K}}(\Theta)\}}{\int_{\boldsymbol{\Theta}} f(\Theta)\exp\{\beta\,\Omega_{\mathcal{K}}(\Theta)\}\,\mathrm{d}\Theta} \\ &\propto f(\Theta)\exp\{\beta\,\Omega_{\mathcal{K}}(\Theta)\},\ \text{with} \end{aligned}$$

$$(4)\quad \Omega_{\mathcal{K}}(\Theta) = \sum_{\kappa \in \boldsymbol{\kappa}} \alpha_\kappa \int_{\boldsymbol{\Psi}_\kappa} f_\kappa(\Psi)\ln[M(\Psi, \Theta)]\,\mathrm{d}\Psi,$$

$\beta \in (0, \infty)$, $\alpha_\kappa \geq 0$, $\sum_{\kappa \in \boldsymbol{\kappa}} \alpha_\kappa = 1$, and $M(\Psi, \Theta)$ is given by (1). The function $\Omega_{\mathcal{K}}(\Theta)$, (4), is an expectation of the logarithm of the parametric model (1) with respect to the weighted average pdf $\hat{f}(\Psi)$ representing the prior knowledge in the pdfs from $\mathcal{K}$ (2)

$$(5)\qquad \hat{f}(\Psi) = \sum_{\kappa \in \boldsymbol{\kappa}} \alpha_\kappa f_\kappa(\Psi).$$

The formulae (3), (4) can be interpreted as follows. Let us consider a collection $\mathcal{K}_{|\boldsymbol{\tau}|} = \{\{{}^{\mathrm{c}}\Psi_{\kappa\tau} \sim f_\kappa(\Psi)\}_{\tau \in \boldsymbol{\tau}_\kappa}\}_{\kappa \in \boldsymbol{\kappa}}$, $|\boldsymbol{\tau}| = \min_{\kappa \in \boldsymbol{\kappa}}(|\boldsymbol{\tau}_\kappa|) \leq \infty$, of independent realisations ${}^{\mathrm{c}}\Psi_{\kappa\tau}$ of data vectors $\Psi \in \boldsymbol{\Psi}_\kappa$. Then Bayes rule [29], applied to them, updates the flat prior pdf $f(\Theta)$ to the posterior pdf $f(\Theta|\mathcal{K}_{|\boldsymbol{\tau}|})$, which is rewritten into the form resembling (3),

$$\begin{aligned} f(\Theta|\mathcal{K}_{|\boldsymbol{\tau}|}) &\propto f(\Theta)\prod_{\kappa \in \boldsymbol{\kappa}}\prod_{\tau \in \boldsymbol{\tau}_\kappa} M({}^{\mathrm{c}}\Psi_{\kappa\tau}, \Theta) \\ &= f(\Theta)\exp\left\{\beta_{|\boldsymbol{\tau}|}\sum_{\kappa \in \boldsymbol{\kappa}} \alpha_{\kappa|\boldsymbol{\tau}|}\cdot\right. \\ &\quad \left.\cdot\int_{\boldsymbol{\Psi}_\kappa} f_{\kappa|\boldsymbol{\tau}|}(\Psi)\ln[M(\Psi, \Theta)]\,\mathrm{d}\Psi\right\} \\ f_{\kappa|\boldsymbol{\tau}|}(\Psi) &= \frac{1}{|\boldsymbol{\tau}_\kappa|}\sum_{\tau \in \boldsymbol{\tau}_\kappa} \delta(\Psi - {}^{\mathrm{c}}\Psi_{\kappa\tau}) \\ &\quad (\text{a sample pdf of } \Psi \text{ from } \kappa\text{th source}) \\ \delta(\Psi - {}^{\mathrm{c}}\Psi_{\kappa\tau}) &= \text{Dirac delta on } {}^{\mathrm{c}}\Psi_{\kappa\tau} \quad \text{for} \quad \Psi, {}^{\mathrm{c}}\Psi_{\kappa\tau} \in \boldsymbol{\Psi}_\kappa \\ \alpha_{\kappa|\boldsymbol{\tau}|} &= \frac{|\boldsymbol{\tau}_\kappa|}{\beta_{|\boldsymbol{\tau}|}},\ \beta_{|\boldsymbol{\tau}|} = \sum_{\kappa \in \boldsymbol{\kappa}} |\boldsymbol{\tau}_\kappa|. \end{aligned}$$

For $|\boldsymbol{\tau}| \to \infty$, the sample pdfs $f_{\kappa|\boldsymbol{\tau}|}(\Psi)$ converge generally to $f_\kappa(\Psi)$ and the weights $\alpha_{\kappa|\boldsymbol{\tau}|}$ to $\alpha_\kappa \geq 0$, $\sum_{\kappa \in \boldsymbol{\kappa}} \alpha_\kappa = 1$. The posterior pdf $f(\Theta|\mathcal{K}_{|\boldsymbol{\tau}|})$, however, converges to a pdf that over-fits the processed data-vectors samples as $\beta_{|\boldsymbol{\tau}|} \geq |\boldsymbol{\tau}|$. Knowing that the processed knowledge is imprecise, we have to make this posterior pdf more flat. The simplest flattening [16] shrinks $\beta_{|\boldsymbol{\tau}|}$ by a factor $\zeta_{|\boldsymbol{\tau}|} \in (0, 1)$ such that $\beta_{|\boldsymbol{\tau}|}\zeta_{|\boldsymbol{\tau}|} \to \beta < \infty$. Thus, the formulae (3), (4) can be interpreted as a flattened version of Bayes rule applied to infinite amounts of independent samples from $\mathcal{K}$. Their use

i) avoids demanding sampling, ii) excludes over-fitting of imprecise prior knowledge and iii) respects relevance of knowledge pieces. Section 4 shows how to optimise the weights

$$w_\kappa = \beta\alpha_\kappa \tag{6}$$

which control influence of the respective knowledge pieces $f_\kappa(\Psi)$ on $f(\Theta|\mathcal{K})$, see (3). The discussed formulae serve us here as a tool only. A reader interested in alternative views is referred to [15].

Note that the support (the set of its positivity) of the flat prior pdf $f(\Theta)$ includes the support of the pdf $f(\Theta|\mathcal{K})$ and the support of any posterior pdf using it as prior pdf. Thus, a constrained support of $f(\Theta)$ quantifies well hard constraints on the parameter range and we can focus on quantification of "soft" constraints expressing the processed knowledge.

2.2 Knowledge description in exponential family

The considered parametric models (1) from the exponential family (EF) [4] have the form

$$M(\Psi,\Theta) = A(\Theta)\exp\left\langle B(\Psi), C(\Theta)\right\rangle, \tag{7}$$

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 its first argument.

The evaluation of the function $\Omega_{\mathcal{K}}(\Theta)$ (4) for the model (7) is simple, as the pdf (3) becomes

$$\begin{aligned} f(\Theta|\mathcal{K}) &\propto f(\Theta)A^\beta(\Theta)\exp\left\langle \beta V, C(\Theta)\right\rangle, \text{ with} \\ V &= \sum_{\kappa\in\boldsymbol{\kappa}} \alpha_\kappa\Lambda_\kappa, \\ \text{where } \Lambda_\kappa &= \int_{\boldsymbol{\Psi}_\kappa} B(\Psi) f_\kappa(\Psi)\,\mathrm{d}\Psi,\ \kappa\in\boldsymbol{\kappa}. \end{aligned} \tag{8}$$

The array V is an expectation of $B(\Psi)$ (7) with respect to the average pdf $\hat{f}(\Psi)$ (5). Its increments Λ_κ, weighted by α_κ, are expectations of $B(\Psi)$ with respect to $f_\kappa(\Psi)$.

Thus, application of (3) and (4) to the model (7) reduces knowledge elicitation to a *mapping of available domain-specific knowledge pieces* on the set $\mathcal{K}$ (2). This mapping is constructed by employing maximum entropy principle [31], which assigns to the κth piece of knowledge the pdf $f_\kappa(\Psi)$ compatible with the processed piece and having the highest entropy.

We consider knowledge expressed in the form: *realisations of data vector Ψ are highly expected to be in a set $\boldsymbol{\Psi}_\kappa$*. To simplify the processing, we neglect a possible occurrence of Ψ out of this set. Then, the maximum entropy principle provides

$$f_\kappa(\Psi) = \mathcal{U}_\Psi\left(\boldsymbol{\Psi}_\kappa\right), \tag{9}$$

where $\mathcal{U}_\Psi\left(\boldsymbol{\Psi}_\kappa\right)$ denotes the pdf of uniform random data vectors on the set $\boldsymbol{\Psi}_\kappa$, $\kappa\in\boldsymbol{\kappa}$. Having (9), it remains to specify the flat prior pdf $f(\Theta)$. For the parametric models in EF (7), the following conjugate prior pdf is considered

$$\begin{aligned} f(\Theta) &= \frac{A^\nu(\Theta)\exp\left\langle V, C(\Theta)\right\rangle}{\mathcal{I}(V,\nu)}, \\ \mathcal{I}(V,\nu) &= \int_{\boldsymbol{\Theta}} A^\nu(\Theta)\exp\left\langle V, C(\Theta)\right\rangle \mathrm{d}\Theta, \end{aligned} \tag{10}$$

with $V=\underline{V}$ and $\nu=\underline{\nu}$, making pdf $f(\Theta)$ flat and guaranteeing $\mathcal{I}(\underline{V},\underline{\nu})<\infty$.

Then, the posterior pdf $f(\Theta|\,{}^{\mathrm{c}}d(t),\mathcal{K})$, reflecting prior knowledge $\mathcal{K}$ (2) and data vectors ${}^{\mathrm{c}}\Psi(t)$ made of the measured data ${}^{\mathrm{c}}d(t)$, preserves the conjugate form (10) i.e. $f(\Theta|\,{}^{\mathrm{c}}d(t),\mathcal{K}) = A^{\nu_t}(\Theta)\exp\left\langle V_t, C(\Theta)\right\rangle/\mathcal{I}(V_t,\nu_t)$. The arrays V_t and scalars ν_t evolve recursively

$$\begin{array}{rclrcl} V_t &=& V_{t-1}+B({}^{\mathrm{c}}\Psi_t), & V_0 &=& \underline{V}+\beta V, \\ \nu_t &=& \nu_{t-1}+1, & \nu_0 &=& \underline{\nu}+\beta, \\ t &\in& \boldsymbol{t}=\{1,\ldots,|\boldsymbol{t}|\}, & & & \end{array} \tag{11}$$

with V defined in (8). Prior knowledge (2) influences initial conditions in (11) by adding the term βV to $\underline{V}$. Section 3 shows its influence generally and Example 2 of Section 5 numerically.

Remarks

• The indicator function of a constrained support can be included into the flat prior pdf. This preserves the recursion (11) but makes evaluation of normalisation factor and moments of the pdf $f(\Theta|\mathcal{K})$ numerically demanding.

• The maximum entropy principle generally provides non-zero pdf $f_\kappa(\Psi)$ out of the highly expected set $\boldsymbol{\Psi}_\kappa, \kappa\in\boldsymbol{\kappa}$. An influence of its approximation by the uniform pdf having the support on $\boldsymbol{\Psi}_\kappa$ is expected to be negligible as the optimised weighting by w_κ (6), discussed in Section 4, diminishes influence of the neglected tails of $f_\kappa(\Psi)$ out of $\boldsymbol{\Psi}_\kappa$.

This hypothesis can be studied within the framework of robust Bayesian estimation, [30].

• The κth piece of knowledge often concerns data vectors Ψ with a certain (non-random) part, typically a part of the regression vector ψ. Let us decompose

$$\Psi' = \left[{}^{\mathrm{r}}\Psi',\ {}^{\mathrm{c}}\Psi'_\kappa\right], \tag{12}$$

where ${}^{\mathrm{r}}\Psi$ contains uncertain (random) entries and ${}^{\mathrm{c}}\Psi_\kappa$ certain (realised) entries. Applying the chain rule to the pdf $f_\kappa(\Psi)=f_\kappa\left({}^{\mathrm{r}}\Psi,\,{}^{\mathrm{c}}\Psi_\kappa\right)$, we get, see (8),

$$\Lambda_\kappa = \int_{{}^{\mathrm{r}}\boldsymbol{\Psi}_\kappa} B\left(\left[{}^{\mathrm{r}}\Psi',\,{}^{\mathrm{c}}\Psi'_\kappa\right]'\right) f_\kappa\left({}^{\mathrm{r}}\Psi\right|{}^{\mathrm{c}}\Psi_\kappa)\ \mathrm{d}\,{}^{\mathrm{r}}\Psi. \tag{13}$$

• Uncertain, domain-specific, knowledge can be provided in many forms. Each requires a ready (algorithmic) mapping of this knowledge form on the pdfs in $\mathcal{K}$ (2). A number of such mappings, frequently met in practice, is designed in

Section 3. They do not cover a full range of possibilities. Many cases can be treated similarly as in Section 3 or as follows:

⋆ Knowledge "*if the regression vector* ψ *yields the value* ${}^{c}\psi_\kappa$, *then the output* $y \in \boldsymbol{y}_\kappa$" quantifies by

$$\Lambda_\kappa = \int_{\boldsymbol{y}_\kappa} B\left([y,\, {}^{c}\psi_\kappa']'\right) \mathcal{U}_y(\boldsymbol{y}_\kappa)\, \mathrm{d}y.$$

⋆ Knowledge "*if the regression vector yields the values* $[{}^{r}\psi',\, {}^{c}\psi_\kappa']'$ *with* ${}^{r}\psi \in {}^{r}\boldsymbol{\psi}_\kappa$, *then the output* $y \in \boldsymbol{y}_\kappa$" quantifies similarly: the averaging with respect to uniform pdf is applied to the whole uncertain part ${}^{r}\Psi = [y,\, {}^{r}\psi']'$ of the data vector Ψ.

⋆ Fuzzy rules can be treated similarly as above. It suffices to interpret the involved membership functions as non-normalised conditional pdfs used in expectation (13).

2.3 Bayesian estimation of normal controlled autoregressive model

The discussed processing of prior knowledge is applied to the normal controlled autoregressive model (ARX). Use of the chain rule allows us to consider the normal ARX model with a single output. Its Bayesian estimation, exploiting data records ${}^{c}d(t) = ({}^{c}d_1, \ldots, {}^{c}d_t)$ observed up to the discrete time $t \in \boldsymbol{t}$, is recalled here. The normal ARX model belongs to EF (7)

$$M(\Psi,\Theta) = \mathcal{N}_y(\theta'\psi, r) = \frac{1}{\sqrt{2\pi r}} \exp\left[-\frac{(y-\theta'\psi)^2}{2r}\right] \tag{14}$$
$$= A(\Theta) \exp\left\{-\mathrm{tr}\left(\Psi\Psi' \frac{[-1,\theta']'[-1,\theta']}{2r}\right)\right\},$$

$A(\Theta) = \frac{1}{\sqrt{2\pi r}}$, $B(\Psi) = \Psi\Psi'$, $C(\Theta) = \frac{[-1,\theta']'[-1,\theta']}{2r}$ and $\langle B, C\rangle = -\mathrm{tr}(BC)$.

In (14) 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 $\Theta = (\theta, r)$. The conjugate prior pdf is a normal-inverse-gamma pdf $\mathcal{N}i\mathcal{G}$, [16],

$$f(\Theta|V,\nu) = \mathcal{N}i\mathcal{G}_{\theta,r}(\hat\theta, P, \hat r, \nu)$$
$$= \frac{\exp\left\{-\frac{1}{2r}[(\theta-\hat\theta)'P^{-1}(\theta-\hat\theta)+(\nu-2)\hat r]\right\}}{\mathcal{I}(V,\nu)\, r^{0.5(\nu+\ell_\psi+2)}}$$

$$V = \begin{bmatrix} (\nu-2)\hat r & \hat\theta' P' \\ P\hat\theta & P^{-1} \end{bmatrix}, \tag{15}$$

$$\mathcal{I}(V,\nu) = \left[\frac{2}{(\nu-2)\hat r}\right]^{\frac{\nu}{2}} |2\pi P|^{0.5}\, \Gamma\left(\frac{\nu}{2}\right),$$

where ℓ_ψ denotes the length of ψ and Γ is the gamma function. The scalar $\hat r$, the column vector $\hat\theta$ and the matrix P are interpreted in (17) below. The pdf (15) is determined by a symmetric positive definite extended information matrix V – the array V in (8) specialised to the normal ARX model – and by the scalar $\nu > 0$ interpreted as the number of degrees of freedom. The posterior pdf is also normal-inverse-gamma with V_t and ν_t updated according to the following specialised version of the recursion (11)

$$V_t = V_{t-1} + \underbrace{{}^{c}\Psi_t\, {}^{c}\Psi_t'}_{B({}^{c}\Psi_t)}, \quad \nu_t = \nu_{t-1} + 1, \tag{16}$$

where ${}^{c}\Psi_t = [{}^{c}y_t,\, {}^{c}\psi_t']'$ is a data vector available at time $t \in \boldsymbol{t}$. It can be shown [29] that $\hat\theta$, $\hat r$ and P in (15) are quantities known in connection with the recursive least squares (RLS) algorithm, to which (16) is algebraically equivalent. The following correspondence holds ($\mathsf{E}_t[\cdot]$ and $\mathsf{cov}_t[\cdot]$ denote expectation and covariance conditioned on information processed up to the time $t \in \boldsymbol{t}$, respectively)

$$\hat\theta_t = \mathsf{E}_t[\theta] = \text{ RLS estimate of } \theta, \tag{17}$$
$$\hat r_t = \mathsf{E}_t[r] = \frac{\text{RLS remainder}}{\nu_t - 2}, \quad \hat r_t P_t = \mathsf{cov}_t[\theta].$$

This correspondence motivates the "standard" choice of the prior pdf $f(\Theta)$, given by $\underline{V}$ and $\underline{\nu}$ specified via (15) by $\underline{\hat\theta} = 0$, $\underline{P}$ = diagonal matrix with a large diagonal, $\underline{\hat r}$ and $\underline{\nu} - 2$ chosen as small positive numbers, [25]. This choice quantifies an assumption that θ and r are finite but knowledge of their values and relations is very vague. The technique developed here enriches this commonly accepted practice by including available prior knowledge. The proposed approach provides better initial conditions of (11), i.e. better initial conditions of RLS, equivalent to the recursion (16) [29],

$$\hat\theta_t = \hat\theta_{t-1} + G_t({}^{c}y_t - \hat\theta_{t-1}'\, {}^{c}\psi_t),$$
$$G_t = \frac{P_{t-1}\, {}^{c}\psi_t}{1 + {}^{c}\psi_t' P_{t-1}\, {}^{c}\psi_t},$$
$$P_t = P_{t-1} - G_t\, {}^{c}\psi_t' P_{t-1}.$$

The recursion updates values of the sufficient statistics determining the posterior pdf, expressed in RLS terms (15). It uses the observed realisations of the output ${}^{c}y_t$ and the regression vector ${}^{c}\psi_t$ entering the normal variant (14) of the parametric model (1).

The choice of V_0 determines the needed initial conditions via (15), (17). Importantly, the advocated knowledge elicitation influences the whole V_0 and thus influences both prior estimate of regression coefficients $\hat\theta_0$ *and* the gains G_t with which the prediction errors ${}^{c}y_t - \hat\theta_{t-1}'\, {}^{c}\psi_t$ modify the estimates $\hat\theta_t$, $t \geq 1$. Example 2 in Section 5 illustrates this influence numerically.

3. PROCESSING OF COMMON TYPES OF PRIOR KNOWLEDGE

The processing presented below deals with knowledge types commonly available in the control domain. It constructs typical mappings of knowledge pieces on pdfs in $\mathcal{K}$

(2) and applies (3) to models in EF, i.e. provides the increments Λ_κ in (8). The mappings are mostly specialised to the normal ARX model for which

$$\Lambda_\kappa = \int_{\boldsymbol{\Psi}_\kappa} \Psi\Psi' f_\kappa(\Psi)\,\mathrm{d}\Psi. \tag{18}$$

Section 3.1 deals with the prior knowledge of ranges of data trajectories, i.e. the ranges of data sequences indexed by the "fictitious" time of the gedanken experiment.

A high number of common types of prior knowledge about the system can be expressed in terms of data ranges. A simple example of this type is quantification of knowledge about *static gain*, see Section 3.2.1. Section 3.2.2 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 3.3. It makes explicit the need to counteract possible over-fitting of prior knowledge. The necessary balance between prior knowledge and data observed can be partially reached by using just-in-time-modelling methodology, Section 3.3.1. Generally, the balance can be and is to be controlled by the weights w_κ (6). Their choice is briefly discussed in Section 3.3.2 and finalised in Section 4. Quantification of *response's smoothness*, Section 3.4, provides an example of a widely-spread knowledge type, whose processing requires Monte-Carlo-type evaluation. The knowledge of *cut-off frequency*, Section 3.5.1, and of a *point on frequency response*, Section 3.5.2, represent the domain-specific knowledge analytically mappable on $\mathcal{K}$ (2).

3.1 Basic quantification: ranges of data trajectories

Ranges of data trajectories are often known from: i) the system design phase, ii) series of past experiments performed for estimation of particular characteristics of the modelled system, e.g. step response. Ranges of data trajectories mean knowledge pieces constructed from *ordered sequence of data ranges*. Data ranges induce ranges of data vectors

$$\Psi \in \boldsymbol{\Psi}_\kappa = \left[\underline{\Psi}_\kappa, \overline{\Psi}_\kappa\right], \text{ which is a shorthand notation for } \Psi_i \in \left[\underline{\Psi}_{\kappa i}, \overline{\Psi}_{\kappa i}\right], i = 1, \ldots, \ell_\Psi,\ \kappa \in \boldsymbol{\kappa}. \tag{19}$$

They are determined by the lower $\underline{\Psi}_\kappa$ and upper $\overline{\Psi}_\kappa$ boundary values with finite entries $\underline{\Psi}_{\kappa i}$ and $\overline{\Psi}_{\kappa i}$.

The respective ranges are treated individually (thus indexed by $\kappa \in \boldsymbol{\kappa}$) and expressed via the uniform pdfs $f_\kappa(\Psi) = \mathcal{U}_\Psi([\underline{\Psi}_\kappa, \overline{\Psi}_\kappa])$ on the intervals (19) in accordance with the adopted maximum entropy. For EF, the increment Λ_κ (8) becomes

$$\Lambda_\kappa = \int_{\left[\underline{\Psi}_\kappa, \overline{\Psi}_\kappa\right]} B(\Psi)\,\mathcal{U}_\Psi\left(\left[\underline{\Psi}_\kappa, \overline{\Psi}_\kappa\right]\right)\,\mathrm{d}\Psi. \tag{20}$$

For the normal ARX model, the increment Λ_κ (20) of the extended information matrix V reads

$$\begin{aligned}\Lambda_\kappa &= \int_{\left[\underline{\Psi}_\kappa, \overline{\Psi}_\kappa\right]} \Psi\Psi'\mathcal{U}_\Psi\left(\left[\underline{\Psi}_\kappa, \overline{\Psi}_\kappa\right]\right)\,\mathrm{d}\Psi \\ &= \frac{1}{4}\left(\overline{\Psi}_\kappa + \underline{\Psi}_\kappa\right)\left(\overline{\Psi}_\kappa + \underline{\Psi}_\kappa\right)' \\ &+ \frac{1}{12}\mathsf{diag}\left[\left(\overline{\Psi}_{\kappa 1} - \underline{\Psi}_{\kappa 1}\right)^2, \ldots, \left(\overline{\Psi}_{\kappa \ell_\Psi} - \underline{\Psi}_{\kappa \ell_\Psi}\right)^2\right].\end{aligned} \tag{21}$$

3.2 Exploitation of the basic quantification

Static gain, rise time and dynamic delay characterise a system's response to a change from an equilibrium. They are examples of traditional characteristics of standardised experiments with inspected real systems. All of them suit for gedanken experiments.

3.2.1 Simple case

The *static gain* g of a system is a (negative) difference between the initial value of the system's output ${}^{\mathrm{c}}y_1$ and its steady-state value ${}^{\mathrm{c}}y_1 + g$ reached after the system's input change from the initial value ${}^{\mathrm{c}}u_1$ to the value ${}^{\mathrm{c}}u_1 + 1$. The knowledge $g \in [\underline{g}, \overline{g}]$ is often available and its quantification was repeatedly addressed [14, 18]. Its definition can be interpreted as the gedanken experiment:

- the inspected scalar system's output and input are at their initial constant levels ${}^{\mathrm{c}}y_1$, ${}^{\mathrm{c}}u_1$,
- a unit step change is applied to the system's input,
- the system's output reaches a new steady state in the interval $[{}^{\mathrm{c}}y_1 + \underline{g}, {}^{\mathrm{c}}y_1 + \overline{g}]$.

The initial input-output values determine the realisation of the data vector ${}^{\mathrm{c}}\Psi_1$. The steady-state data vector Ψ_2 contains the certain part ${}^{\mathrm{c}}\Psi_2$ made of ${}^{\mathrm{c}}u_1 + 1$ and uncertain one ${}^{\mathrm{r}}\Psi_2$ formed by the stabilised system's output $y \in [{}^{\mathrm{c}}y_1 + \underline{g}, {}^{\mathrm{c}}y_1 + \overline{g}]$ determining the range of the terminal data vector ${}^{\mathrm{r}}\Psi_2 \in [{}^{\mathrm{r}}\underline{\Psi}_2, {}^{\mathrm{r}}\overline{\Psi}_2]$. A detailed exploitation of this knowledge is well visible on a single-input, single-output normal ARX model with the state in the phase form. It has the regression vector

$$\psi_t' = [y_{t-1}, \ldots, y_{t-n}, u_t, \ldots, u_{t-m}],\ n, m \geq 0. \tag{22}$$

The processed fictitious data vectors in initial ($\tau = 1$) and terminal ($\tau = 2$) steady state are ${}^{\mathrm{c}}\Psi_1 = {}^{\mathrm{c}}\Psi_{\tau=1} = [{}^{\mathrm{c}}y_1, \ldots, {}^{\mathrm{c}}y_1, {}^{\mathrm{c}}u_1, \ldots, {}^{\mathrm{c}}u_1]'$, ${}^{\mathrm{c}}\Psi_{\tau=2} = [{}^{\mathrm{c}}u_1 + 1, \ldots, {}^{\mathrm{c}}u_1 + 1]'$, ${}^{\mathrm{r}}\underline{\Psi}_{\tau=2} = [{}^{\mathrm{c}}y_1 + \underline{g}, \ldots, {}^{\mathrm{c}}y_1 + \underline{g}]'$, ${}^{\mathrm{r}}\overline{\Psi}_{\tau=2} = [{}^{\mathrm{c}}y_1 + \overline{g}, \ldots, {}^{\mathrm{c}}y_1 + \overline{g}]'$ with $(n+1)$ and $(m+1)$ entries in the parts related to the system's output and input, respectively. Equations (8), (21) and the definition ${}^{\mathrm{c}}\Psi_2 = {}^{\mathrm{c}}\Psi_1 + [0.5(\overline{g} + \underline{g}), \ldots, 0.5(\overline{g} + \underline{g}), 1, \ldots, 1]'$ give the increment Λ (18) of the extended information matrix V (15)

$$\Lambda = \frac{1}{2}\Big\{ {}^{c}\Psi_1\,{}^{c}\Psi_1' + {}^{c}\Psi_2\,{}^{c}\Psi_2' + \\ + \frac{1}{12}\mathrm{diag}\Big[\underbrace{(\overline{g}-\underline{g})^2,\ldots,(\overline{g}-\underline{g})^2}_{(n+1)\text{ times}},\ \underbrace{0,\ldots,0}_{(m+1)\text{ times}}\Big]\Big\}.$$

3.2.2 More complex case

Rise time, ρ, refers to the time required for the system's output to rise from a specified initial value, say zero, to a specified final steady-state value. *Dynamic delay*, Δ, is the time required for the system's output to change from zero to a small non-zero value. Both characteristics can be converted into the data-ranges type of knowledge as follows.

- The system is assumed to be in a steady state characterised by a constant system's input ${}^{c}u_1$ and the corresponding constant system's output ${}^{c}y_1 = 0$.
- The inspected characteristics are gained when the unit step ${}^{c}u_1 \to {}^{c}u_1 + 1$ is applied. Thus, the system's input forms a certain part of the constructed data vectors, see (12). The certain values define the increment Λ in (8) according to (13).
- The system's output is negligible until time Δ, i.e., its absolute value is highly expected to be smaller than ε times (say, $\varepsilon \approx 0.1$) the guess of the static gain, $\hat{g}$ (for instance, $\hat{g} = 0.5(\overline{g}+\underline{g})$, for clarity $\hat{g} > 0$), i.e., the expected system's output range is

$$\boldsymbol{y}_\tau = [-\varepsilon \times \hat{g}, \varepsilon \times \hat{g}], \quad \text{for } \tau \le \Delta. \tag{23}$$

- Given bounds $\underline{y}_\tau$, $\overline{y}_\tau$ determine the expected range of the output before the rise time ρ

$$\boldsymbol{y}_\tau = [\underline{y}_\tau, \overline{y}_\tau], \ \text{for } \tau \in (\Delta, \rho]\,. \tag{24}$$

- For the (fictitious) time span $\tau > \rho$, the system's output y_τ is expected to be in the intervals

$$\boldsymbol{y}_\tau = \Big[\max\Big((1-\varepsilon)\times\hat{g},\ \underline{y}_\tau\Big),\ \min\Big((1+\varepsilon)\times\hat{g},\ \overline{y}_\tau\Big)\Big]\,. \tag{25}$$

(23)–(25) specify the ranges of data records and consequently of data-vector trajectories, see (19), and allow a direct application of the results obtained in Section 3.1.

3.3 Relevant prior knowledge

Available realisations ${}^{c}\Psi_\kappa$ of data vector Ψ, with $\kappa \in \boldsymbol{\kappa}$ referring to the κth item in an extensive data source (typically, data base), often serve as prior knowledge. The probabilistic description of this knowledge is $f_\kappa(\Psi) = \delta(\Psi - {}^{c}\Psi_\kappa)$ and the use of equation (3) reduces to ordinary Bayes estimation if $w_\kappa = \beta\alpha_\kappa = 1$. This is a correct solution, if the realisations are obtained on the modelled system and in an 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 so much that the real data observed will not be able to change this. The problem is not critical if a number of processed data vectors is small and real data is informative [29]. Then, equation (3), reduced to Bayes rule, can be directly applied. If these conditions are violated, two approaches are used (see below): i) a real-time selection of the relevant realisations, which are closely related to the current system's state, ii) non-unit weighting of the processed data that can control the influence of the knowledge incorporated.

3.3.1 Real-time selection of the relevant data

The methodology called (among others) just-in-time modelling, e.g. [8, 24], can counteract the mentioned shrinking of the pdf $f(\Theta|\mathcal{K})$. This methodology assumes the ability to store a large number of data vectors and to inspect them in real time. The local model is built "just-in-time" relying on the following selection of relevant data.

- Current observations made at real time $t \in \boldsymbol{t}$ are put in the regression vector ${}^{c}\psi_t$.
- A small number $|\boldsymbol{\kappa}|$ of stored data vectors $\{{}^{c}\Psi_\kappa\}_{\kappa\in\boldsymbol{\kappa}}$ with the regression vectors $\{{}^{c}\psi_\kappa\}_{\kappa\in\boldsymbol{\kappa}}$ "close" to the currently observed ${}^{c}\psi_t$ are selected. Here, the subscript κ refers to the stored record serving as the κth prior-knowledge piece exploited at $t \in \boldsymbol{t}$.

This choice relies on the hypothesis that similar causes, represented by close regression vectors, lead to similar consequences reflected in outputs.

- Parameter Θ of a (local) model $M(\Psi_t, \Theta)$ (1) is estimated at time $t \in \boldsymbol{t}$ by applying Bayes rule to the data vectors $\{{}^{c}\Psi_\kappa\}_{\kappa\in\boldsymbol{\kappa}}$ corresponding to the selected regression vectors $\{{}^{c}\psi_\kappa\}_{\kappa\in\boldsymbol{\kappa}}$. It means that (3) is applied with pdfs $f_\kappa \in \mathcal{K}$ being Dirac functions placed on the data vectors $\{{}^{c}\Psi_\kappa\}_{\kappa\in\boldsymbol{\kappa}}$ and the weights $w_\kappa = 1$, (6).
- The pdf $f(\Theta|\mathcal{K})$ is used for predicting an unknown value of the output y_t modelled by the predictive pdf $f(y_t|\,{}^{c}\psi_t, \mathcal{K}) = \int_{\boldsymbol{\Theta}} M([y_t, {}^{c}\psi_t']', \Theta) f(\Theta|\mathcal{K})\,\mathrm{d}\Theta$.

The outlined idea is quite powerful if the modelled relation of y and ψ is smooth. It may, however, be sensitive to the definition of the closeness of regression vectors ${}^{c}\psi_t$ and ${}^{c}\psi_\kappa$. The probabilistic treatment offers the following systematic approach, which considers acceptance of the *natural conditions of control*, [29]. They postulate that knowledge of the regression vector without the corresponding output says nothing about the parameter Θ, i.e.

$$f(\Theta|\,{}^{c}\psi_t) = f(\Theta|\,{}^{c}\psi_\kappa) = f(\Theta). \tag{26}$$

The regression vectors ${}^{c}\psi_\kappa$ and ${}^{c}\psi_t$ can be assumed sufficiently close if the joint pdf of yet unobserved system's output y_t and unknown finite-dimensional parameter Θ, given by the vector ${}^{c}\psi_\kappa$ selected from the data base, is close to that given by ${}^{c}\psi_t$ observed at time $t \in \boldsymbol{t}$, i.e.

$$f(y_t, \Theta|\,{}^{c}\psi_t) \approx f(y_t, \Theta|\,{}^{c}\psi_\kappa). \tag{27}$$

Under (26), the joint pdf $f(y_t, \Theta|\,{}^{c}\psi_t)$ can be rewritten in the following way

$$\begin{aligned} f(y_t, \Theta|\,{}^{c}\psi_t) &= M([y_t,\, {}^{c}\psi_t']', \Theta) f(\Theta|\,{}^{c}\psi_t) \\ &= M([y_t,\, {}^{c}\psi_t']', \Theta) f(\Theta). \end{aligned}$$

Similarly, the joint pdf of y_t and Θ conditioned on ${}^{c}\psi_\kappa$ reads

$$\begin{aligned} f(y_t, \Theta|\,{}^{c}\psi_\kappa) &= M([y_t,\, {}^{c}\psi_\kappa']', \Theta) f(\Theta|\,{}^{c}\psi_\kappa) \\ &= M([y_t,\, {}^{c}\psi_\kappa']', \Theta) f(\Theta). \end{aligned}$$

Under weak conditions [6], the Kullback-Leibler divergence [23] of $f(y_t, \Theta|\,{}^{c}\psi_t)$ on $f(y_t, \Theta|\,{}^{c}\psi_\kappa)$ is an adequate measure of the inspected proximity. Under (26), the divergence reads

$$\begin{aligned} &(28)\ \mathcal{D}_{t\kappa} \\ &= \int_{(\boldsymbol{y}_t, \boldsymbol{\Theta})} M\left([y,\, {}^{c}\psi_t']'\right) f(\Theta) \ln\left(\frac{M\left([y,\, {}^{c}\psi_t']', \Theta\right)}{M\left([y,\, {}^{c}\psi_\kappa']', \Theta\right)}\right) \mathrm{d}y\mathrm{d}\Theta. \end{aligned}$$

Thus, at time $t \in \boldsymbol{t}$, the relevant data vectors ${}^{c}\Psi_\kappa' = [{}^{c}y_\kappa,\, {}^{c}\psi_\kappa']$ have the regression vectors ${}^{c}\psi_\kappa$, $\kappa \in \boldsymbol{\kappa}$, which yield small values of the divergence $\mathcal{D}_{t\kappa}$ (28). The next discussion shows that the term "small" can be well quantified.

For EF and the conjugate prior pdf $f(\Theta)$, given by (10) with $V = \underline{V}$ and $\nu = \underline{\nu}$, $\mathcal{D}_{t\kappa}$ becomes

$$\begin{aligned} \mathcal{D}_{t\kappa} = \int_{(\boldsymbol{y}_t, \boldsymbol{\Theta})} &\langle B([y,\, {}^{c}\psi_t']') - B([y,\, {}^{c}\psi_\kappa']'), C(\Theta)\rangle \frac{A^{\underline{\nu}+1}(\Theta)}{\mathcal{I}(\underline{V}, \underline{\nu})} \\ &\times \exp\langle \underline{V} + B([y,\, {}^{c}\psi_t']'), C(\Theta)\rangle \,\mathrm{d}y\mathrm{d}\Theta. \end{aligned}$$

For a single-output normal ARX model (14), $\mathcal{D}_{t\kappa}$ reads

$$\begin{aligned} \mathcal{D}_{t\kappa} &= \int_{\boldsymbol{\theta}, r\geq 0} \frac{[\theta'({}^{c}\psi_t - {}^{c}\psi_\kappa)]^2}{2r} \mathcal{N}i\mathcal{G}_{\theta,r}(\underline{V}, \underline{\nu})\,\mathrm{d}\theta\mathrm{d}r \\ &= \frac{1}{2}\left[\frac{\underline{\nu}}{(\underline{\nu}-2)\underline{\hat{r}}}[\underline{\hat{\theta}}'({}^{c}\psi_t - {}^{c}\psi_\kappa)]^2 + \right. \\ &\quad \left. + ({}^{c}\psi_t - {}^{c}\psi_\kappa)'\underline{P}({}^{c}\psi_t - {}^{c}\psi_\kappa)\right]. \end{aligned} \tag{29}$$

The quantities $\underline{\hat{\theta}}$, $\underline{\hat{r}}$ and $\underline{P}$ are defined by (15) with $V = \underline{V}$ and $\nu = \underline{\nu}$. The result (29) follows from the basic properties of the normal and normal-inverse-gamma pdfs, see for example [16]. The first summand in the square brackets above is proportional to the normalised squared difference of the outputs' predictions based on $\underline{\hat{\theta}}$, $\underline{\hat{r}}$ and the regression vectors ${}^{c}\psi_t$, ${}^{c}\psi_\kappa$. Hence the values of $\mathcal{D}_{t\kappa}$ much larger than one cannot be considered small. The second summand is proportional to the squared Euclidean norm of $({}^{c}\psi_t - {}^{c}\psi_\kappa)$ weighted by the matrix $\underline{P}$. The equations (15) and (16) imply that 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 $\ell_\psi/\underline{\nu}$ cannot be taken as small. This indicates that the values $\mathcal{D}_{t\kappa} \leq \bar{\mathcal{D}} < 0.5(1 + \ell_\psi/\underline{\nu})$ are small.

3.3.2 On knowledge weighting

The regression vector ${}^{c}\psi_t$, around which the local model is built, determines the selection of the relevant prior data while the threshold $\bar{\mathcal{D}}$ controls the degree of relevance and the amount of processed data. Without such "reference", typically in off-line processing mode, just case-dependent choices of the relevant data, e.g. [11], are at disposal. Consequently, a huge amount of data vectors of varying relevance has to be often processed. Then, the prior knowledge can be "over-fitted" and an influence of real observations diminished.

The problem applies to any data-rich source, but it becomes especially important if the processed data sample is generated by simulation models. Despite these models accumulating a substantial prior knowledge, their use in the subsequent design of decision strategies is limited as the optimising design is often unfeasible without their simplification.

Adaptive systems supported here rely on approximate models, too. They optimise a decision strategy in real time by using a recursively estimated approximate model from a tractable class of parametric models, typically from EF. The approximation is constructed implicitly via Bayesian estimation, which guarantees the asymptotically best approximation of the modelled system [16].

The learning transient can be substantially shortened, if the knowledge accumulated in a simulation model is projected onto the prior pdf $f(\Theta|\mathcal{K})$. Application of Bayes rule to simulated data vectors ${}^{c}\Psi_{\kappa\tau}$, $\tau \in \boldsymbol{\tau} = \{1, \ldots, |\boldsymbol{\tau}_\kappa|\}$ does the job due to the mentioned approximation ability. The formula (3) counteracts the danger of over-fitting as the function $\Omega_{\mathcal{K}}(\Theta)$ weights the individual sample pdf $f_\kappa(\Psi) = \frac{1}{|\boldsymbol{\tau}_\kappa|}\sum_{\tau\in\boldsymbol{\tau}_\kappa} \delta(\Psi - {}^{c}\Psi_{\kappa\tau})$ representing $\{{}^{c}\Psi_{\kappa\tau}\}_{\tau\in\boldsymbol{\tau}_\kappa}$. The weighting can be interpreted as a use of a flattened version of the pdf, [16], obtained after standard Bayesian estimation from the sample $\{{}^{c}\Psi_{\kappa\tau}\}_{\tau\in\boldsymbol{\tau}_\kappa}$.

Altogether, the incorporation of the κth knowledge piece provided by a large amount $|\boldsymbol{\tau}_\kappa|$, say simulated, data vectors reduces (for EF) to a collection of the sample version of the *normalised increment*

$$\Lambda_\kappa = \frac{1}{|\boldsymbol{\tau}_\kappa|}\sum_{\tau=1}^{|\boldsymbol{\tau}_\kappa|} \Lambda_{\kappa\tau} \tag{30}$$

and its weighted inclusion into (8). The choice of the weights (6), controlling the impact of the incorporated knowledge pieces, is to be subjective without additional information. Section 4 provides an automatic choice of the weights w_κ, when real data records are at disposal.

3.4 Monte-Carlo quantification

Smoothness of the system's response to standardised system's inputs ${}^{c}u_\kappa$, $\kappa \in \boldsymbol{\kappa}$, is a frequently available type of knowledge about the system. It can be expressed by a set of

restrictions describing the highly expected output trajectories gained in the gedanken experiment

$$y_\kappa \in \boldsymbol{y}_\kappa = \{y_\kappa : |y_\kappa - {}^{c}y_{\kappa-1}| \le q_\kappa || {}^{c}\psi_\kappa - {}^{c}\psi_{\kappa-1}||\}, \tag{31}$$

for $\kappa \in \boldsymbol{\kappa}$. Each set (31) depends on ${}^{c}u_\kappa$, ${}^{c}\Psi_{\kappa-1}$, a supplied continuity module $q_\kappa > 0$ and the norm $||\cdot||$.

Similar to the previous cases, this type of knowledge can be expressed via uniform pdfs on $\boldsymbol{y}_\kappa$. They, together with the known deterministic mapping $({}^{c}\Psi_{\kappa-1}, {}^{c}u_\kappa, y_\kappa) \to \Psi_\kappa$, determine the *conditional* pdfs $f_{\kappa|\kappa-1}(\Psi|\,{}^{c}u, {}^{c}\Psi)$, expressing the highly expected transitions $({}^{c}\Psi_{\kappa-1}, {}^{c}u_\kappa) \to \Psi_\kappa$, for $\kappa \ge 2$. The pdf $f_1(\Psi)$, describing the expected initial data vectors, can be chosen by using, for instance, available knowledge of data ranges. This construction represents the case when pdfs in the set $\mathcal{K} = \{f_\kappa(\Psi)\}_{\kappa\in\boldsymbol{\kappa}}$ are *given implicitly* as solutions of the equations

$$f_\kappa(\Psi) = \int_{\boldsymbol{\Psi}_{\kappa-1}} f_{\kappa|\kappa-1}(\Psi|\,{}^{c}u_\kappa, \Psi_{\kappa-1}) f_{\kappa-1}(\Psi_{\kappa-1})\,\mathrm{d}\Psi_{\kappa-1},$$

for $\kappa = 2, \dots, |\boldsymbol{\kappa}|$. An explicit solution of these equations can hardly be obtained. The underlying conditional pdfs $f_{\kappa|\kappa-1}(\Psi|\,{}^{c}u_\kappa, {}^{c}\Psi_{\kappa-1})$ are, however, simple and Monte Carlo methodology can be applied. It draws random independent samples ${}^{c}\Psi_{1\tau} \sim f_1(\Psi)$, $\tau \in \boldsymbol{\tau} = \{1, \dots, |\boldsymbol{\tau}|\}$, and simulates realisations ${}^{c}\Psi_{2\tau}, \dots, {}^{c}\Psi_{|\boldsymbol{\kappa}|\tau}$ by using the considered ${}^{c}u_\kappa$ and drawing the system's output samples ${}^{c}y_\kappa$ from the uniform pdf on $\boldsymbol{y}_\kappa$ (31). For EF, these realisations serve for evaluating a sample version of the increment $\Lambda_\kappa = \frac{1}{|\boldsymbol{\tau}|}\sum_{\tau=1}^{|\boldsymbol{\tau}|} B({}^{c}\Psi_{\kappa\tau})$ (30), which is then used in (8).

3.5 Analytical quantification

This section deals with important types of knowledge for which the gedanken experiment can be evaluated analytically. The presentation is made for the normal single-input, single-output ARX model with the phase-form regression vector, given by $m, n \ge 0$, see (22). It means that the knowledge piece is expressed via analytically constructed increment Λ (18) of V (15).

3.5.1 Cut-off frequency

The term cut-off frequency refers to the smallest frequency $\omega_c \in (0, 2\pi)$ of the sinusoidal system's input that leaves the system's output almost uninfluenced: the system's output stays around the initial, say zero, value. Thus, considering the sinusoidal system's input of a fixed frequency $\omega \in [\omega_c, 2\pi)$, the highly-expected fictitious data vectors are as follows

$$\Psi_{\tau\omega} = [y_\tau, \dots, y_{\tau-n}, \sin(\tau\omega), \dots, \sin(\omega(\tau-m))]', \tag{32}$$

for $\tau \in \boldsymbol{\tau} = \{1, \dots, |\boldsymbol{\tau}|\}$, $|\boldsymbol{\tau}| \to \infty$.

In (32) the involved system's outputs $(y_\tau, \dots, y_{\tau-n})$ have zero mean, negligible correlations and a small variance σ^2, typically $\sigma^2 = \hat{\underline{r}}$, see Section 2.3.

In correspondence with the output-input structure of the regression vector (22), the increments Λ of (18) the extended information matrix V (15) can be split into the blocks

$$\Lambda = \begin{bmatrix} R & T \\ T' & S \end{bmatrix}, \tag{33}$$

where R, S, T are matrices of dimensions $(n+1, n+1)$, $(m+1, m+1)$, $(n+1, m+1)$, respectively. In accordance with (30), the increment Λ_ω of the extended information matrix V is the sample mean of $\Psi_{\tau\omega}\Psi'_{\tau\omega}$ for the data vectors (32)

$$\begin{aligned} \Lambda_\omega &= \lim_{|\boldsymbol{\tau}|\to\infty} \frac{1}{|\boldsymbol{\tau}|}\sum_{\tau=1}^{|\boldsymbol{\tau}|} \Psi_{\tau\omega}\Psi'_{\tau\omega} = \begin{bmatrix} R_\omega & 0 \\ 0 & 0.5S_\omega \end{bmatrix} \\ &= \begin{bmatrix} \sigma^2 I_{n+1} & 0 \\ 0 & 0.5S_\omega \end{bmatrix}, \end{aligned} \tag{34}$$

where I_{n+1} is the unit matrix of the order $n+1$ and S_ω is the corresponding $(m+1, m+1)$-block of the decomposition (33) for the fixed frequency ω. Using the complex form of goniometric functions with j denoting imaginary unit, the (k,l)-entry $S_\omega(k,l)$ of the matrix S_ω with $k, l \in \{1, \dots, m+1\}$ can be written as follows

$$\begin{aligned} S_\omega(k,l) &= 2\lim_{|\boldsymbol{\tau}|\to\infty} \frac{1}{|\boldsymbol{\tau}|}\sum_{\tau=1}^{|\boldsymbol{\tau}|} \sin(\omega(\tau-k))\sin(\omega(\tau-l)) \\ &= -\lim_{|\boldsymbol{\tau}|\to\infty} \frac{1}{2|\boldsymbol{\tau}|}\sum_{\tau=1}^{|\boldsymbol{\tau}|} [\exp(j\omega(\tau-k)) - \exp(-j\omega(\tau-k))] \\ &\quad \times [\exp(j\omega(\tau-l)) - \exp(-j\omega(\tau-l))] \\ &= \frac{1}{2}[\exp(j\omega(\tau-l)) + \exp(-j\omega(\tau-l))] \\ &\quad - \lim_{|\boldsymbol{\tau}|\to\infty} \frac{1}{2|\boldsymbol{\tau}|}\sum_{\tau=1}^{|\boldsymbol{\tau}|} [\exp(j\omega(2\tau-k-l)) + \exp(-j\omega(2\tau-k-l))] \\ &= \cos(\omega(k-l)). \end{aligned} \tag{35}$$

The last limit is zero, as it is a bounded sum of the geometric sequences divided by $|\boldsymbol{\tau}| \to \infty$.

The expressed knowledge is valid for any fixed frequency $\omega \in [\omega_c, 2\pi)$. The knowledge of the cut-off frequency ω_c is expressed by the collection of increments Λ_ω of the extended information matrices V for all $\omega \in [\omega_c, 2\pi)$. The formal correspondence between κ and ω and between Λ_κ and Λ_ω, respectively, applied to equation (30) reveals that an average of Λ_ω over $\omega \in [\omega_c, 2\pi)$ adequately represents all these knowledge pieces. An overall increment Λ of V, computed by averaging, reads

$$\Lambda = \int_{\omega_c}^{2\pi} \Lambda_\omega \mathcal{U}_\omega([\omega_c, 2\pi])\,\mathrm{d}\omega = \begin{bmatrix} \sigma^2 I_{n+1} & 0 \\ 0 & 0.5S \end{bmatrix}, \tag{36}$$

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

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

3.5.2 A point on frequency response

Knowledge of cut-off frequency is a special case of a partial knowledge of the system's frequency response available at least in connection with auto-tuners [9]. Recalling that the frequency response is the system's reaction to the sinusoidal system's input, a relevant gedanken experiment is described by the data vectors

$$\begin{aligned}\Psi_{\tau\omega\phi} &= [a\sin(\tau\omega+\phi)+e_\tau,\dots,a\sin(\omega(\tau-n)+\phi)+\\ &\quad +e_{\tau-n},\sin(\tau\omega),\dots,\sin(\omega(\tau-m))]' \\ &= {}^{c}\Psi_{\tau\omega\phi}+[e_\tau,\dots,e_{\tau-n},0,\dots,0]',\ \tau\in\boldsymbol{\tau},\end{aligned} \tag{37}$$

where a is the supplied estimate of the amplitude. Uncertainty of this knowledge is modelled by the mutually uncorrelated noise elements e_τ with zero mean and a priori specified variance σ^2, typically $\sigma^2=\hat{\underline{r}}$, see Section 2.3. In (37) the subscript $\omega\phi$ indicates the considered frequency and the phase shift. 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 a fixed phase shift ϕ, the increment $\Lambda_{\omega\phi}$ (18) of the extended information matrix V (15) coincides with the following sample moment evaluated for the data vectors (37)

$$\begin{aligned}\Lambda_{\omega\phi} &= \lim_{|\boldsymbol{\tau}|\to\infty}\frac{1}{|\boldsymbol{\tau}|}\sum_{\tau=1}^{|\boldsymbol{\tau}|}\Psi_{\tau\omega\phi}\Psi'_{\tau\omega\phi}\\ &= \begin{bmatrix}\sigma^2 I_{n+1}+a^2R_\omega & 0.5aT_{\omega\phi}\\ 0.5aT'_{\omega\phi} & 0.5S_\omega\end{bmatrix},\end{aligned}$$

where matrices R_ω, S_ω and $T_{\omega\phi}$ are obtained via the decomposition (33). The entries of R_ω and S_ω are defined by (34), (35). The (k,l)th entry $T_{\omega\phi}(k,l)$ of the $(n+1,m+1)$-matrix $T_{\omega\phi}$ equals

$$T_{\omega\phi}(k,l)=\cos(\omega|k-l|+\phi). \tag{38}$$

Similarly to cut-off frequency, see Section 3.5.1, the final increment Λ_ω of the extended information matrix can be computed by averaging $\Lambda_{\omega\phi}$ over the possible phase shifts $\phi\in[\underline{\phi},\overline{\phi}]$. In the most uncertain case when no knowledge of the phase shift $\phi\in[0,2\pi]$ is available, it holds

$$\Lambda_\omega=\frac{1}{2\pi}\int_0^{2\pi}\Lambda_{\omega\phi}\,\mathrm{d}\phi=\begin{bmatrix}\left(\sigma^2+a^2\right)I_{n+1} & 0\\ 0 & 0.5S_\omega\end{bmatrix}. \tag{39}$$

4. DATA-BASED KNOWLEDGE WEIGHTING

The influence of prior knowledge depends on the weights w_κ (6) with which the increments Λ_κ, representing the processed pdfs in $\mathcal{K}=\{f_\kappa(\Psi)\}_{\kappa\in\boldsymbol{\kappa}}$, enter $f(\Theta|\mathcal{K})$, see (8).

The choice of w_κ is critical issue for a balanced weighting of prior knowledge and information brought by observed data. It becomes even more critical, when the combined knowledge pieces: i) concern different aspects of the modelled system, or ii) reflect the same system's property but offered by different knowledge sources. The weights can be chosen automatically *after* observing a sufficient number $\upsilon\in\boldsymbol{t}$ of real, informative data ${}^{c}d(\upsilon)=({}^{c}d_1,\dots,{}^{c}d_\upsilon)$. The term "sufficient number" formally means that at least one observed realisation ${}^{c}\Psi$ of data vector Ψ is available. In practice ${}^{c}d(\upsilon)$ must counteract a poor robustness of the maximum likelihood estimates, see [10].

In the considered case, the posterior pdf at time $\upsilon\in\boldsymbol{t}$ gets the form, cf. (7), (8),

$$\begin{aligned}&f\left(\Theta\left|{}^{c}d(\upsilon),\underline{\nu},\underline{V},\Lambda_1,\dots,\Lambda_{|\boldsymbol{\kappa}|},w\right.\right)\propto A(\Theta)^{\underline{\nu}_\upsilon+\sum_{\kappa=1}^{|\boldsymbol{\kappa}|}w_\kappa}\\ &\times\exp\left\langle\underline{V}_\upsilon+\sum_{\kappa=1}^{|\boldsymbol{\kappa}|}w_\kappa\Lambda_\kappa,C(\Theta)\right\rangle\\ &\underline{V}_\upsilon=\underline{V}+\sum_{t=1}^{\upsilon}B({}^{c}\Psi_t),\quad \underline{\nu}_\upsilon=\underline{\nu}+\upsilon.\end{aligned} \tag{40}$$

Let us stress, that unlike κ referring to the κth piece of prior knowledge, t and υ refer to discrete time of *real* data observations. In (40) the weight $w_\kappa=\beta\alpha_\kappa\geq 0$ determines the strength of the κth knowledge piece. The choice of the vector $w\in\boldsymbol{w}=\{w=[w_1,\dots,w_{|\boldsymbol{\kappa}|}]',\ w_\kappa\geq 0\}$ is based on the fixed knowledge of $\underline{\nu},\underline{V},\underline{\nu}_\upsilon,\underline{V}_\upsilon,\Lambda_1,\dots,\Lambda_{|\boldsymbol{\kappa}|}$. For an instance of w, the predictive pdf, evaluated for the observed data ${}^{c}d(\upsilon)$, reads, cf. (10), (40),

$$\begin{aligned}&f({}^{c}d(\upsilon)|\underline{\nu},\underline{V},\Lambda_1,\dots,\Lambda_{|\boldsymbol{\kappa}|},w)\\ &=\frac{\mathcal{I}\left(\underline{V}_\upsilon+\sum_{\kappa=1}^{|\boldsymbol{\kappa}|}w_\kappa\Lambda_\kappa,\underline{\nu}_\upsilon+\sum_{\kappa=1}^{|\boldsymbol{\kappa}|}w_\kappa\right)}{\mathcal{I}\left(\underline{V}+\sum_{\kappa=1}^{|\boldsymbol{\kappa}|}w_\kappa\Lambda_\kappa,\underline{\nu}+\sum_{\kappa=1}^{|\boldsymbol{\kappa}|}w_\kappa\right)}.\end{aligned} \tag{41}$$

It is the likelihood function of the unknown $w\in\boldsymbol{w}$. The rigorous Bayesian treatment would require assignment of a prior pdf over $\boldsymbol{w}$ and the evaluation of the posterior pdf on $\boldsymbol{w}$. The related computational complexity motivates the search for the maximum likelihood estimate of w for the given $\underline{\nu},\underline{V},\underline{\nu}_\upsilon,\underline{V}_\upsilon,\Lambda_1,\dots,\Lambda_{|\boldsymbol{\kappa}|}$, i.e. the maximiser of (41) on $\boldsymbol{w}$. This choice respects the mentioned exceptional role of the Kullback-Leibler divergence [6] as the maximum likelihood estimate minimises its affine transformation, namely, the Kerridge inaccuracy [20] of the sample pdf of the observed data ${}^{c}d(\upsilon)$ on the optimised predictive pdf.

Hölder inequality implies that the logarithm of the likelihood function (41) is a difference of convex functions of $w\in\boldsymbol{w}$. Moreover, it has the finite ith derivative with respect to w, if the ith moments of $\ln(A(\Theta))$ and $C(\Theta)$, defining EF (7), exist. Consequently, a rich set of optimisation algorithms is available for its maximisation.

5. ILLUSTRATIVE EXAMPLES

The examples demonstrate: i) discarding of irrelevant prior knowledge via weighting, Example 1, ii) positive in-

fluence of prior knowledge on parameter estimates, Example 2, iii) combination of knowledge pieces about the same system's property, Example 3.

Extensive numerical studies of other combinations, comparisons with previous processing versions as well as a study illustrating just-in-time modelling methodology will be published elsewhere.

In all examples, the following normal ARX model (14) with single system's output y_t is considered, see Section 2.3,

(42)
$$y_t = 1.81y_{t-1} - 0.8187y_{t-2} + 0.00468u_t + 0.00438u_{t-1} + e_t,$$

with white normal noise $e_t \sim \mathcal{N}_{e_t}(0, 10^{-4})$ and the independent white exogenous system's 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 used noise variance makes the simulation realistic as the autoregressive part amplifies the noise roughly by hundred times. The chosen variance of the simulated input, which is attenuated by the corresponding coefficient, makes its influence on the output similar to that of the noise. The relation between both influences determines the ease and rate of the estimation and the chosen case is relatively hard. The simulation with well stimulating input u_t in the open controlled loop is favourable for data-based estimation [25]. This set up makes it harder to demonstrate positive effects of the prior-knowledge incorporation. When successful, the demonstration guarantees that much stronger positive effects can be expected under much harder conditions within a closed control loop, e.g. [18].

The influence of incorporated prior knowledge is demonstrated by comparing the estimation results gained with and without use of prior knowledge. Each example has the following steps repeated for N realisations of the noise e and system's input u.

• *Data generation* – a collection of υ, see (40), learning data records are generated by model (42).

• *Parameter estimation* – estimation, see Section 2.3, is run twice on learning data: with and without prior knowledge. The runs without prior knowledge use the standard settings of the prior, Section 2.3, with diagonal of $\underline{P}$ equals to 10^6, $\underline{\hat{r}} = 10^{-4}$, $\underline{\nu} = 2 + 10^{-6}$.

• *Evaluation of results* – the results are judged according to the prediction quality quantified by

$$(43) \qquad Q = \frac{\text{sample second moment of prediction errors}}{\text{variance of the noise } e_t \text{ in (42)}},$$

which is evaluated on validation data, generated after fixing w_κ in (6).

Figures, like the time course of the regression-coefficients estimates, provide qualitative insight.

Example 1 illustrates influence of prior knowledge of a static gain $g \in [\underline{g}, \overline{g}] = [0.9, 1.1]$ on the prediction. The processing steps were run for $N = 100$ noise and system's input realisations:

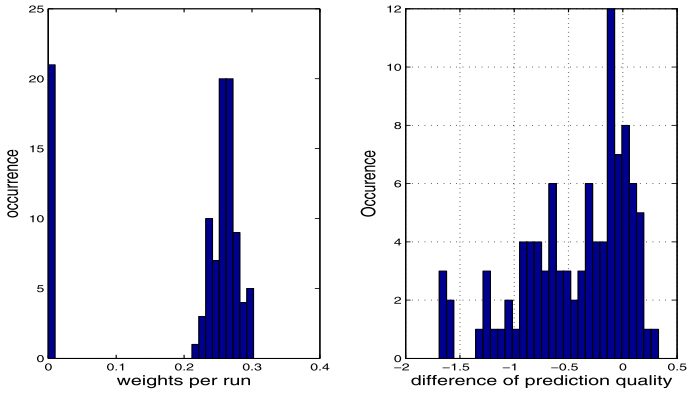

Figure 1. Influence of prior knowledge of static gain: histogram of weights (8) over realisations (left) and histogram of differences of the prediction quality over realisations (right).

• *Data generation* – a collection of $\upsilon = 200$ learning data records was generated by (42).

• *Parameter estimation without prior knowledge* – the estimation uses the standard prior, see Section 2.3.

• *Parameter estimation with prior knowledge* – the posterior pdf obtained from the learning data was combined with single prior knowledge of the static gain, Section 3.2.1. The numerically computed weight $w = w_1$ maximises the predictive pdf (41) evaluated for the learning data $^{\upsilon}d(\upsilon)$.

• *Evaluation of results* – 1,000 additional validation data records were used for evaluating the prediction quality (43).

The results are in Figure 1. The left subplot presents the histogram of the optimal weights w computed for each of $N = 100$ noise and system's input realisations. The higher value of the weight, the more informative contribution and stronger influence of the knowledge processed. It is worth noticing that more than 20% realisations led to zero weight: the processed knowledge is perceived as irrelevant. The right subplot of Figure 1 presents a histogram of the prediction quality differences (43) for the estimation with the prior knowledge and without it. Therefore, the prediction with the prior knowledge is worse if the difference presented is positive. The histogram confirms predominantly positive influence of the processed prior knowledge. Quantitatively, it is seen on elementary statistics of the prediction-quality differences *evaluated on validation* data: mean $= -0.363$, median $= -0.240$, minimum $= -1.664$, maximum $= 0.180$.

Remarks

• The occurrences of (almost) zero weights, suppressing adverse effect of the processed knowledge, cumulate into a single column in the left subplot of Figure 1. This and the discrete nature of occurrence counts caused the observed gap in the graph.

• Realisations' randomness causes deviations from the predominantly positive effects of the included prior knowledge.

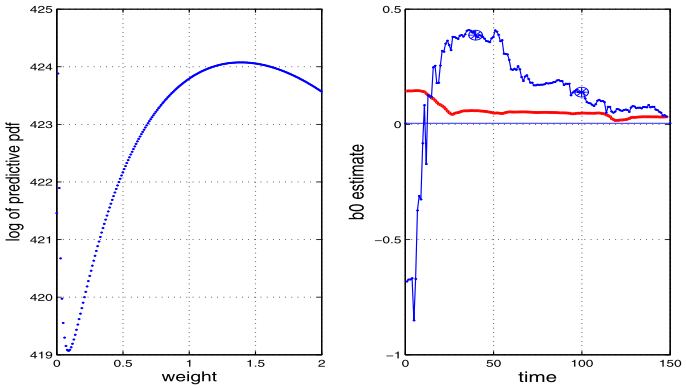

Figure 2. Influence of prior knowledge of static gain: the logarithm of predictive pdf (41) as a function of the weight w (left) and time courses of b_0 estimate (right). Circles mark the time course without prior knowledge. The straight line marks the simulated coefficient b_0.

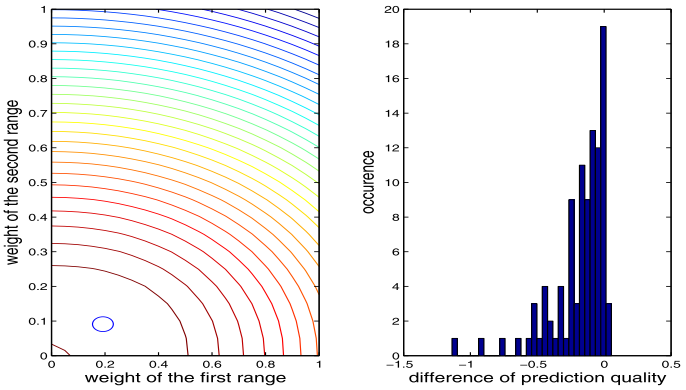

Figure 3. Influence of the combined knowledge of data ranges: the contours the predictive-pdf logarithm (41) as a function of two-dimensional weight (left) and histogram of differences of the prediction quality over realisations (right).

Sometimes it masks the deterministic part of the simulated relations. This explains the observed down-weighting of knowledge constructed from the data measured on the true system.

Example 2 illustrates influence of prior knowledge of a static gain $g \in [\underline{g}, \overline{g}] = [0.9, 1.1]$ on point estimates of the regression coefficients in (42). The processing steps were run once:

- *Data generation* – a collection of $\upsilon = 20$ learning data records was generated by (42) to initialise the estimation and to find the optimal weights.
- *Parameter estimation without prior knowledge* – estimation run on 150 additional data records using the standard prior, see Section 2.3.
- *Parameter estimation with prior knowledge* – estimation run on the additional data records using the prior pdf enriched by the knowledge of a static gain. The numerically computed weight maximises the predictive pdf (41) evaluated for the learning data ${}^{c}d(\upsilon)$.
- *Evaluation of results* – the obtained time courses of the point estimates (the time course of an appropriate entry of $\hat{\theta}_t$, see (17)) of the coefficient $b_0 = 0.00468$ at u_t (42) 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 likelihood as a function of the optimised weight w. The curve illustrates smoothness of the maximised function as well as existence of non-trivial maximum. 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 an incorporation of prior knowledge of data ranges and combination of several pieces of knowledge. To select ranges properly, two independent data sets, $|\boldsymbol{\kappa}| = 2$, each of the length 50 were generated by (42). The realistic ranges of data vectors $[\underline{\Psi}_\kappa, \overline{\Psi}_\kappa]$ were determined as envelopes of these simulated data sets. The processing was run for $N = 100$:

- *Data generation* – a collection of $\upsilon = 100$ learning data records was generated by (42).
- *Parameter estimation without prior knowledge* – estimation run using the standard prior.
- *Parameter estimation with prior knowledge* – estimation run using the standard prior combined with the processed knowledge items and learning data. The numerically found weights $w = [\hat{w}_1, \hat{w}_2]'$, maximising (41), fixed the impact of knowledge pieces in (40).
- *Evaluation of results* – an additional collection of 1,000 validation data records was generated and used for evaluating the prediction quality (43).

The left subplot of Figure 3 shows the logarithm of the likelihood in a two-dimensional space of the weights $w = [w_1, w_2]'$. The maximum is marked by a circle. The plot corresponds to the last noise and system's input realisations. The right subplot presents histogram of the differences of the prediction quality (43) for the estimation with the prior knowledge and without it. A positive difference indicates the used prior knowledge has worsened the prediction. The histogram confirms positive influence of the prior knowledge processed. Statistics of the prediction-quality differences on validation data are: $\mathsf{mean} = -0.188$, $\mathsf{median} = -0.121$, $\mathsf{minimum} = -1.138$, $\mathsf{maximum} = 0.047$.

6. CONCLUDING REMARKS

The paper concerns elicitation and quantification of prior knowledge frequently met in the engineering domain. The adopted methodology works with prior knowledge expressed as a collection of pdfs on the space of data trajectories. The illustrative examples indicate a visible improvement of

the estimation and prediction results implied by the proposed approach. The experience confirms [19] that the proposed inclusion of the prior knowledge improves the model structure estimation as well as the quality of adaptive control.

The reported methodology represents a significant step towards the facilitator-free incorporation of domain-specific knowledge into the prior pdf utilised by the Bayesian estimation. A unified elicitation methodology, based on the incorporating knowledge of data ranges, and data-based merging of various knowledge pieces represent the 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 [18], iii) objectively determines a relative impact of a knowledge piece processed. The last item is extremely important as it increases the robustness and the quality of knowledge elicitation: the knowledge piece that does not improve prediction based on vague prior knowledge gets a negligible weight as illustrated by Example 1, Section 5.

The paper focuses on normal controlled autoregressive model. This stimulates questions, concerning its practical impact and applicability width, which are briefly commented here.

• ARX model is useful per se. It serves to adaptive controllers applied in controlling technological processes (primarily in chemical and energy industries) or in metal production or transportation [16]. A recent survey and many references are in [32]). Cheaper commission of such controllers, suppression of wild adaptation transients and increased robustness are our main contributions to them.

• The gedanken experiment unifies treatment of expert knowledge, knowledge arising from the controlled system design, from preliminary data (even obsolete or collected only on a similar system or simulation model). This methodology is model-independent.

• The approach is immediately applicable to other members of the exponential family and its mixtures, providing a powerful modelling tool of complex processes. Their estimation is sensitive to initialisation [1, 16] and can be substantially improved by the proposed methodology.

The foreseen open problems include elicitation of knowledge provided by, possibly fuzzy, production rules and robustness analysis (see Remarks in Section 2.2). These technical steps will enhance the achieved conceptual and algorithmic improvements. The main progress is, however, expected in elaborating facilitator-free quantification of domain-specific *decision making preferences* (control aims). It can be achieved by applying the methodology to the pdf expressing control aims within the fully probabilistic design of the control strategies [17].

ACKNOWLEDGEMENTS

The authors gratefully acknowledge support of Bilateral project CNR Italy and AVČR "Advanced techniques of Bayesian decision making in complex systems" and GAČR 13-13502S.

Received 16 October 2013

REFERENCES

[1] Ari, C., Aksoy S., Arikan O., 2012. Maximum likelihood estimation of Gaussian mixture models using stochastic search. Pattern Recognition 45 (7), 2804–2816.

[2] Abonyi, J., Babuska, R., Verbruggen, H., Szeifert, F., 2000. Incorporating prior knowledge in fuzzy model identification. Int. J. of Systems Science 5 (31), 657–667.

[3] Astrom, K., Wittenmark, B., 1989. Adaptive Control. Addison-Wesley, Massachusetts.

[4] Barndorff-Nielsen, O., 1978. Information and exponential families in statistical theory. New York. MR0489333

[5] Berger, J., 1985. Statistical Decision Theory and Bayesian Analysis. Springer, New York. MR0804611

[6] Bernardo, J. M., 1979. Expected information as expected utility. The Annals of Statistics 7 (3), 686–690. MR0527503

[7] Betró, B., Guglielmi, A., 2000. Methods for global prior robustness under generalized moment conditions. In: Rios-Insua, D., Ruggeri, F. (Eds.), Robust Bayesian Analysis. Springer Verlag, New York, pp. 273–294. MR1795221

[8] Bontempi, G., Birattari, M., Bersini, H., 1999. Lazy learning for local modelling and control design. Int. J. of Control 72 (7–8), 643–658. MR1679876

[9] Clarke, D., 2003. Pretuning and adaptation of PI controllers. IEE Proceedings-Control Theory and Applications 150 (6), 585–598.

[10] da Fonseca, V., Fieller, N., 2006. Distortion in statistical inference: The distinction between data contamination and model deviation. Metrika 63, 169–190. MR2242538

[11] Edwards, J., Alifantis, T., Hurrion, R., Ladbrook, J., Robinson, S., Waller, A., 2004. Using a simulation model for knowledge elicitation and knowledge management. Simulation Modelling Practice and Theory 12 (7–8), 527–540.

[12] Garthwaite, P., Kadane, J., O'Hagan, A., Jun 2005. Statistical methods for eliciting probability distributions. J. of the American Statistical Association 100 (470), 680–700. MR2170464

[13] Hang, C. C., Astrom, K. J., Wang, Q. G., 2002. Relay feedback auto-tuning of process controllers a tutorial review. J. of Process Control 12, 143–162.

[14] Kárný, M., 1984. Quantification of prior knowledge about global characteristics of linear normal model. Kybernetika 20 (5), 376–385. MR0776327

[15] Kárný, M., Andrýsek, J., Bodini, A., Guy, T. V., Kracík, J., Ruggeri, F., 2006. How to exploit external model of data for parameter estimation? Int. J. of Adaptive Control and Signal Processing 20 (1), 41–50. MR2199130

[16] Kárný, M., Böhm, J., Guy, T. V., Jirsa, L., Nagy, I., Nedoma, P., Tesař, L., 2006. Optimized Bayesian Dynamic Advising: Theory and Algorithms. Springer, London.

[17] Kárný, M., Guy, T. V., 2006. Fully probabilistic control design. Systems & Control Letters 55 (4), 259–265. MR2202563

[18] Kárný, M., Khailova, N., Nedoma, P., Böhm, J., 2001. Quantification of prior information revised. Int. J. of Adaptive Control and Signal Processing 15 (1), 65–84.

[19] Kárný, M., Nedoma, P., Khailova, N., Pavelková, L., 2003. Prior information in structure estimation. IEE Proc. – Control Theory and Applications 150 (6), 643–653.

[20] Kerridge, D., 1961. Inaccuracy and inference. J. of Royal Statistical Society B 23, 284–294. MR0123375

[21] Kosut, R. L., 2001. Iterative adaptive control: Windsurfing with confidence. In: G.C., G. (Ed.), Model Identification and Adaptive Control – From Windsurfing to Telecommunications. Springer-Verlag, London, UK.

[22] Kracík, J., Kárný, M., 2005. Merging of data knowledge in Bayesian estimation. In: Filipe, J., Cetto, J. A., Ferrier, J. L. (Eds.), Proc. of the Second Int. Conference on Informatics in Control, Automation and Robotics. INSTICC, Barcelona, pp. 229–232.
[23] Kullback, S., Leibler, R., 1951. On information and sufficiency. Annals of Mathematical Statistics 22, 79–87. MR0039968
[24] Li, J., Dong, G., Ramamohanarao, K., Wong, L., 2004. Deeps: A new instance-based lazy discovery and classification system. Machine Learning 54 (2), 99–124.
[25] Ljung, L., 1987. System Identification: Theory for the User. Prentice-Hall, London.
[26] Mosca, E., 1994. Optimal, Predictive, and Adaptive Control. Prentice Hall.
[27] O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J., Rakow, T., 2006. Uncertain Judgement: Eliciting Experts' Probabilities. John Wiley & Sons.
[28] Osei-Bryson, K., 2003. Supporting knowledge elicitation and consensus building for Dempster-Shafer decision models. Int. J. of Intelligent Systems 18 (e:1), 129–148.
[29] Peterka, V., 1981. Bayesian system identification. In: Eykhoff, P. (Ed.), Trends and Progress in System Identification. Pergamon Press, Oxford, pp. 239–304. MR0746139
[30] Rios-Insua, D., Ruggeri, F., 2000. Robust Bayesian Analysis. Springer Verlag, New York. MR1795206
[31] Shore, J., Johnson, R., 1980. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. IEEE Transactions on Information Theory 26 (1), 26–37. MR0560389
[32] Sánches, J. M., Lemos, J. M., Rodellar, J., 2013. Survey of industrial optimized adaptive control. Int. J. of Adaptive Control and Signal Processing 26 (10), 881–918. MR2983741

Miroslav Kárný
UTIA AVCR
Prague
Czech Republic
E-mail address: school@utia.cas.cz
url: www.utia.cz/AS

Tatiana V. Guy
UTIA AVCR
Prague
Czech Republic
E-mail address: guy@utia.cas.cz
url: www.utia.cz/AS

Jan Kracík
Dept. Applied Mathematics
VSB-Technical University of Ostrava
Czech Republic
E-mail address: jan.kracik@vsb.cz

Petr Nedoma
UTIA AVCR
Prague
Czech Republic
E-mail address: guy@utia.cas.cz

Antonella Bodini
CNR IMATI
Milano
Italy
E-mail address: anto@mi.imati.cnr.it

Fabrizio Ruggeri
CNR IMATI
Via Bassini 15
I-20133 Milano
Italy
E-mail address: fabrizio@mi.imati.cnr.it
url: www.mi.imati.cnr.it/fabrizio