INTERNATIONAL JOURNAL OF ADAPTIVE CONTROL AND SIGNAL PROCESSING
*Int. J. Adapt. Control Signal Process.* 2005; **18**:1–15 *Prepared using* acsauth.cls *[Version: 2002/11/11 v1.00]*

# How to Exploit External Model of Data for Parameter Estimation?

Miroslav Kárný, Josef Andrýsek, *Antonella Bodini, Tatiana V. Guy, Jan Kracík, *Fabrizio Ruggeri

*Adaptive Systems Department*
*Institute of Information Theory and Automation*
*Academy of Sciences of the Czech Republic*
*P. O. Box 18, 182 08 Prague, Czech Republic*

**Instituto di Matematica Applicata e Tecnologie Informatiche*
*Consiglio Nazionale delle Ricerche*
*Via E. Bassini, 15, I-20133 Milano Italy*

## SUMMARY

Any cooperation in multiple-participant decision making (DM) relies on an exchange of individual knowledge pieces and aims. A general methodology of their rational exploitation without calling for an objective mediator is still missing. This paper proposes such a methodology in an important particular case in which a participant performs Bayesian parameter estimation and it is offered a model relating the observable data to their past history. The proposed solution is based on so called fully probabilistic design (FPD) of DM strategies. The result reduces to an "ordinary" Bayesian estimation if the offered model is the sample probability density function (pdf), i.e., if it provides additional observations. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: Bayesian estimation, decision making, fully probabilistic design, Kullback-Leibler divergence

## 1. INTRODUCTION

Parameter estimation is a basic technique of adapting a model from a suitable class to the modelled environment. The Bayesian estimation [1, 2] is a well established methodology harmonized with decision making (DM) under uncertainty, which is always the ultimate

[*] Correspondence to: Adaptive Systems Department
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
P. O. Box 18, 182 08 Prague, Czech Republic

Contract/grant sponsor: the common project of UTIA CR and CNR Italy – Advanced Techniques of Bayesian Decision Making in Complex Systems, AV ČR 1ET 100 750 404, 1ET 100 750 404, GAČR 102/03/P010, and MŠMT 1M0572

*Received July 2005*

*Accepted*

modelling aim. This estimation is rather straightforward evaluation of the posterior distribution, usually probability density function (pdf), of the unknown parameter given the observed data and available prior knowledge. The most important problems inherent to estimation are (i) modelling, i.e., design of the parameterized model, e.g. [3] (ii) knowledge elicitation, i.e., construction of the prior pdf e.g. [2, 4] (iii) design of evaluation algorithms for specific classes of parameterized models e.g. [5, 6, 7] (iv) analysis of the sensitivity and reliability of the results, e.g. [16, 8].

Here, a specific problem of the knowledge elicitation is addressed that arises in cooperative DM with multiple Bayesian participants. Among numerous variants, we focus on the cooperation scenarios summarized in the paper [9]. For the current paper, the following restrictions of the problem formulation are relevant (i) the respective participants use very different models for describing the same observed data; this prevents them to share prior knowledge on unknown parameters directly (ii) they decide on offering the knowledge to other participants who, in turn, decide on the way and degree of its exploiting; no mediator is supposed to perform this task.

Under these conditions, a participant can offer at most probabilistic distribution describing the common data and the receiving participant has to decide whether and how to exploit them for parameter estimation. This paper provides a justified algorithmic guideline how to do it using the idea presented in [10]. The cited paper exploits the fact that data enter the parameter estimation via a statistics in the form of a sample pdf and takes the knowledge offered in the form of a pdf on data as an initial value in recursive evaluation of this statistics. Its heuristic justification is supported here and generalized by applying so called fully probabilistic design (FPD) of DM strategies [11, 12, 13] to the addressed problem.

Problem formulation, Section 3, follows Section 2 summarizing the notation. The solution via FPD is in Section 4. The resulting generic algorithm and its stationary variant, Section 5, are followed by discussion interpreting it, Section 6, and by conclusions, Section 7

## 2. SOME NOTATIONS

All vector quantities are assumed to be columns and $'$ denotes transposition. Beside,

$\propto$ is the proportionality symbol;

$x^*$ stands for a set of all values of a quantity $x$;

$f(x)$ denotes probability density function (pdf) with names of the arguments referring to random variables they describe;

$d_t$ denotes $\mathring{d}$-dimensional data record at discrete time $t \in t^* \equiv \{1, 2, \ldots, \mathring{t}\}$; it is treated as a random variable;

$\underline{d}_t$ denotes realization of multi-dimensional random variable $d_t$;

$d(t) \equiv (d_t, d_{t-1}, \ldots, d_1)$ is time sequence of data records;

$\Theta$ stands for a finite-dimensional unknown parameter;

$\phi_{t-1}$ is a finite-dimensional state vector whose values are determined by a known deterministic mapping from $d(t-1)$ and the initial state $\phi_0$; for presentation simplicity, $\phi_0$ is assumed to be known and omitted in the conditions hereafter;

$\Psi_t \equiv [d_t', \phi_{t-1}']'$ is a data vector.

## 3. PROBLEM FORMULATION

A pair of DM units (*participants*) within a multiple-participant DM is considered. Each of participants solves own DM task, while participants' behavior have non-empty intersection. By participant's behavior, we understand all possible sequence of observed, decided and considered random variables. The participant's knowledge can be enriched by exploiting available knowledge of another participant concerning the common, potentially observable, part of their behaviors.

Let a common part of the behavior be $\mathring{d}$-dimensional data records $d$, observed at discrete time moments $\tau \in t^* \equiv \{1, 2, \ldots, \mathring{t}\}$. Formally speaking, at each time $\tau$, both participants observe realizations $\underline{d}_\tau$ of multi-dimensional random variable $d_\tau$ and design own models of its evolution. To formulate the problem, participants are distinguished by labels "first" and "second".

### First participant

The participant models evolution of the considered random variable by a finitely parameterized, time-invariant probability density function (pdf) with a finite memory

$$f_1(d_\tau|d(\tau-1), \phi_0, \Theta) = f_1(d_\tau|\phi_{\tau-1}, \Theta), \ \tau \in t^*. \tag{1}$$

A Bayesian type participant, which subjectively selects a prior pdf on parameters $f_1(\Theta)$, is considered. The joint pdf of data sequence up to the time $t \in t^*$ and the parameter, conditioned on $\phi_0$, is defined through the chain rule [1] that gets under the assumption (1) the form

$$f_1(d(t), \Theta) = \prod_{\tau=1}^{t} f_1(d_\tau|\phi_{\tau-1}, \Theta) \times f_1(\Theta), \ t \in t^*. \tag{2}$$

This participant plays an active role within the task considered. Its aim is to estimate unknown parameter $\Theta$ using *all knowledge available.*

### Second participant

The participant models the evolution of data by a conditional non-parameterized pdf

$$f_2(d_\tau|d(\tau-1)), \ \tau \in t^*. \tag{3}$$

The joint pdf of data records observed up to time $t \in t^*$ is determined by (3) via the chain rule as follows

$$f_2(d(t)) = \prod_{\tau=1}^{t} f_2(d_\tau|d(\tau-1)), \ t \in t^*. \tag{4}$$

### Problem addressed

Let us assume there exist a non-empty subsequence of data records, $d(\mathring{k}) \in d^*, \mathring{k} \le t$, on which the first participant takes the model (4) as a reliable information source. Then, the description

(4) can be exploited by the *first participant* to improve *its* knowledge on the common part of behavior. Let us emphasize that the parameterized model of the first participant is assumed to be fixed, thus the only part of participant's knowledge which can be changed is its prior pdf. Processing of the non-parameterized model (4), provided by the second participant, is expected to improve the guess of the first participant about the unknown parameter $\Theta$ of the model (1).

Note the following assumptions simplifying the explanations, but not influencing its generality, are adopted.

A1 Participants model sequences of data records starting at the time $\tau = 1$.

A2 Both models (2) and (4) describe the same data sequences, i.e., $\tau = 1, \ldots, t$. Generally, these sequences can be different, but they must have a non-void intersection.

A3 The subsequence of data records $d(\mathring{k})$, $\mathring{k} \leq t$, on which the first participant takes model (4) as reliable description of the reality, starts at time $\tau = 1$.

Let us stress that the problem is formulated and treated regarding the first participant. The second participant plays a passive role of the information source and does not participate in the processing. The further explanation thus mainly concerns the first participant and if there is no explicit indication, the first participant is considered.

## 4. FULLY PROBABILISTIC TREATMENT

The participant is provided with an external non-parametric data model $f_2(d(\mathring{k}))$ describing a part of data $d(\mathring{k}), \mathring{k} \in t^*$ observed by the participant. It intends to exploit this additional knowledge for estimation of parameter $\Theta$ of the own parameterized model $f_1(d(t)|\Theta)$, which describes the whole data history, including the part $d(\mathring{k})$ modelled by an external model. The discrepancy in the modelled collections of random variables is the key obstacle in problem formalization.

The proper knowledge incorporation should result in the model, which preserves the original knowledge about data and parameter, while be enriched on the additional knowledge provided by the outer model.

The proposed solution rests on three steps.

*Step 1* *To remove discrepancy between modelled collections of random variables*
The outer description $f_2(d(\mathring{k}))$, $d(\mathring{k}) \in d^*(\mathring{k})$ is extended to "two"-dimensional space $(d^*(\mathring{k}), \Theta^*)$ such that the common marginal pdf of the resulting $f_2(d(\mathring{k}), \Theta)$ coincides with $f_2(d(\mathring{k}))$.

*Step 2* *To ensure preserving the knowledge provided the participant's model.*
The model closest to the two-dimensional $f_1(d(\mathring{k}), \Theta)$ is found within the set of models resulting from *Step* 1.

*Step 3* *To ensure incorporating knowledge from the outer model*
The original prior pdf $f_1(\Theta)$ of the parameter $\Theta$ is replaced by the pdf, which contains the knowledge provided by the extended external model.

*Step 1 – Extension of the outer model*

Meaningful extensions of the outer data model $f_2(d(\mathring{k}))$ to a pdf $f_2(d(\mathring{k}), \Theta)$ describing two-dimensional space $(d^*(\mathring{k}), \Theta^*)$ has to combine the non-parametric data description at disposal (4) and an arbitrary prior pdf $f(\Theta)$ expressing prior knowledge on the parameter $\Theta$.

The chain rule applied to the most general possible extension

$$f_2(d(\mathring{k}), \Theta) = \prod_{k=1}^{\mathring{k}} f_2(d_k|d(k-1), \Theta) \times f(\Theta), \mathring{k} \leq t$$

would require the second participant to relate data not only to the past observed history but also to the parameter $\Theta$ unused by it. This implies the need to restrict the possible extensions by the following realistic assumption

$$f_2(d_k|d(k-1), \Theta) = f_2(d_k|d(k-1)), \ \ k = 1, \ldots, \mathring{k}, \tag{5}$$

expressing the conditional independence of predictions made by the second participant on the parameter unused for data modelling.

Then, the resulting joint descriptions, for the given $f_2(d_k|d(k-1))$ and an arbitrary $f(\Theta)$, read

$$f(d(\mathring{k}), \Theta) = \prod_{k=1}^{\mathring{k}} \underbrace{f_2(d_k|d(k-1))}_{\text{outer model}} \times \underbrace{f(\Theta)}_{\text{guess on }\Theta} = f_2(d(\mathring{k})) \times f(\Theta). \tag{6}$$

*Step 2 – Knowledge preservation*

The optional pdf $f(\Theta)$ determines the possible joint pdf (6) for the given $f_2(d_k|d(k-1))$. The best extension of the outer model should properly describe data subsequence $d(\mathring{k})$, while respecting available knowledge on $\Theta$. Within the task considered, the proper selection of the pdf $f(\Theta)$ is expected to provide the joint pdf, which is (6) as close as possible to the joint pdf (2) for all $d(\mathring{k})$. In other words, the pdf $f(\Theta)$ that minimizes a "distance" between the pdfs (2) and (6) is searched for while the pdfs $f_1(d_k|\phi_{k-1}, \Theta)$, $f_1(\Theta)$ and $f_2(d(\mathring{k}))$ are given.

The Kullback-Leibler (KL) divergence $\mathcal{D}(f_2||f_1)$ [14] is known to be a good measure of proximity of the pair of pdfs $f_2 \equiv f_2(d(\mathring{k}), \Theta)$, $f_1 \equiv f_1(d(\mathring{k}), \Theta)$. It is defined by the formula

$$\mathcal{D}(f_2||f_1) \equiv \int f_2(d(\mathring{k}), \Theta) \ln \left( \frac{f_2(d(\mathring{k}), \Theta)}{f_1(d(\mathring{k}), \Theta)} \right) dd(\mathring{k}) d\Theta. \tag{7}$$

The choice of this version of the (asymmetric) KL divergence is heuristically motivated and *ex post* justified by reasonable properties of the obtained solution, see Section 6.

Note that the multiple definite integration over the domain of its argument is denoted in the simplified way by $\int \cdot \, dd(\mathring{k}) d\Theta$. Several integration signs are used whenever Fubini theorem on multiple integration is exploited.

The KL divergence (7) can be re-written as follows

$$
\begin{aligned}
\mathcal{D}(f_2||f_1) &= \int f(\Theta)\left[\ln\left(\frac{f(\Theta)}{f_1(\Theta)}\right) + \int f_2(d(\mathring{k}))\sum_{k=1}^{\mathring{k}} \ln\left(\frac{f_2(d_k|d(k-1))}{f_1(d_k|\phi_{k-1},\Theta)}\right) dd(\mathring{k})\right] d\Theta \\
&= \int f(\Theta)\left[\ln\left(\frac{f(\Theta)}{f_1(\Theta)}\right) + \sum_{k=1}^{\mathring{k}} \int f_2(d(k)) \ln\left(\frac{f_2(d_k|d(k-1))}{f_1(d_k|\phi_{k-1},\Theta)}\right) dd(k)\right] d\Theta \\
&= \int f(\Theta)\left[\ln\left(\frac{f(\Theta)}{f_1(\Theta)}\right) - \underbrace{\sum_{k=1}^{\mathring{k}} \int f_2(d_k,\phi_{k-1}) \ln(f_1(d_k|\phi_{k-1},\Theta))\, dd(k)}_{\Omega_{\mathring{k}}(\Theta)}\right] d\Theta \\
&+ \underbrace{\sum_{k=1}^{\mathring{k}} \int f_2(d(k)) \ln(f_2(d_k|d(k-1)))\, dd(k)}_{c(\mathring{k})} \qquad (8) \\
&= \mathcal{D}\left(f(\Theta)\left|\left|\frac{f_1(\Theta)\exp(\Omega_{\mathring{k}}(\Theta))}{\int f_1(\Theta)\exp(\Omega_{\mathring{k}}(\Theta))\, d\Theta}\right.\right.\right) + \underbrace{c(\mathring{k}) - \ln\left(\int f_1(\Theta)\exp(\Omega_{\mathring{k}}(\Theta))\, d\Theta\right)}_{\text{the term independent of } f(\Theta)}.
\end{aligned}
$$

The basic properties of the KL divergence, stating that

$$\mathcal{D}(f_2||f_1) \geq 0 \text{ and } \mathcal{D}(f_2||f_1) = 0 \text{ iff } f_2 = f_1 \text{ almost everywhere}, \qquad (9)$$

imply that the pdf minimizing the KL divergence (8), denoted symbolically $f(\Theta|f_1)$, is given by the following explicit formula

$$f(\Theta|f_1) = \frac{f_1(\Theta)\exp(\Omega_{\mathring{k}}(\Theta))}{\int f_1(\Theta)\exp(\Omega_{\mathring{k}}(\Theta))\, d\Theta}. \qquad (10)$$

Note that the needed joint pdfs $f_2(\Psi_k) \equiv f_2(d_k,\phi_{k-1})$ are gained from the pdf $f_2(d(k))$, $k \leq \mathring{k}$. For instance, if the state is in so-called phase form $\phi_{k-1} = [d'_{k-1}, d'_{k-2}, \ldots, d'_{k-n}]'$ with a finite order $n \geq 1$, the pdf $f_2(\Psi_k)$ is the marginal pdf $f_2(d_k,\ldots,d_{k-n}) = \int f_2(d(k))\, dd(k-n-1)$.

Especially simple and appealing variant is obtained when data vectors $\Psi_k = [d'_k, \phi'_{k-1}]'$ form a stationary process. Note that this process is defined by the initial state $\phi_0$, pdf $f_2(d_k|d(k-1))$ and by the definition of the state $d(k-1) \rightarrow \phi_{k-1}$. In the stationary case, the pdfs $f_2(\Psi_k) = f_2(d_k,\phi_{k-1})$ do not depend on $k$ and

$$f(\Theta|f_1) \propto f_1(\Theta)\exp\left(\mathring{k}\int f_2(\Psi)\ln(f_1(d|\phi,\Theta))\, d\Psi\right), \quad \Psi \equiv [d',\phi']'. \qquad (11)$$

In summary, the pdf $f(\Theta|f_1)$ (10) (or its special version (11)) defines the extension (6), which is the nearest to the pdf (2) on the time interval $k \in k^* \equiv \{1,\ldots,\mathring{k}\}$.

*Step 3 – Knowledge exploiting*

Now the participant should modify its former prior pdf $f_1(\Theta)$ to a pdf $f(\Theta)$ that includes information obtained from the outer model (4). Sticking at the selected quantification of the extension quality (7), we search for the minimizer, symbolically denoted $f(\Theta|f_2)$,

$$f(\Theta|f_2) = \arg\min_{f(\Theta)} \mathcal{D}\left( f_2(d(\mathring{k}))f(\Theta|f_1) \left|\left| \prod_{k=1}^{\mathring{k}} f_1(d_k|\phi_{k-1},\Theta)f(\Theta) \right.\right.\right), \tag{12}$$

where the pdf $f(\Theta|f_1)$ is determined by (10). The first identity in (8) shows that

$$\mathcal{D}\left( f_2(d(\mathring{k}))f(\Theta|f_1) \left|\left| \prod_{k=1}^{\mathring{k}} f_1(d_k|\phi_{k-1},\Theta)f(\Theta) \right.\right.\right) = \mathcal{D}(f(\Theta|f_1)||f(\Theta)) + \text{term independent of } f(\Theta).$$

This form and the cited basic properties of the KL divergence (9) imply that the minimizing argument in (12) is

$$f(\Theta|f_2) = f(\Theta|f_1) = \frac{f_1(\Theta)\exp[\Omega_{\mathring{k}}(\Theta)]}{\int f_1(\Theta)\exp[\Omega_{\mathring{k}}(\Theta)]\, d\Theta}, \tag{13}$$

or, similarly to (11), for the stationary case

$$f(\Theta|f_2) = f(\Theta|f_1) \propto f_1(\Theta)\exp\left(\mathring{k}\int f_2(\Psi)\ln(f_1(d|\phi,\Theta))\, d\Psi\right). \tag{14}$$

Definitely, the proposed construction leading to the final formula (13) is not the only possible way of incorporating the knowledge offered by the second participant. Unlike others, possibly more straightforward alternatives, this variant has highly desirable properties, see Section 6.

## 5. ALGORITHMIC SUMMARY

Recalling the problem of knowledge exploitation in the "two-participant" setup (Section 3), the following algorithm can be advised.

The first participant aims to use the non-parametric model, provided by the second participant, to improve the parameter estimation of the own model describing its environment.

**Step 1** The second participant offers the non-parametric description of a common part of data sequences to the first participant.

**Step 2** The **first participant** selects the length $\mathring{k}$ of the subsequence $d(\mathring{k})$, on which it takes the second participant's model as a reliable description of reality.

Now the first participant has at disposal: its own model to be estimated $f_1(d(\tau)|\Theta)$, $\tau \leq t$; its prior guess $f_1(\Theta)$ on the unknown parameter $\Theta$ and the offered data model $f_2(d(\mathring{k}))$ it trusts. Its aim is to modify the prior pdf $f_1(\Theta)$ so that it will reflect knowledge provided by the model offered.

**Step 3** The **first participant**

- extends one-dimensional data model $f_2(d(\mathring{k}))$ to a set of "two"-dimensional pdfs $f^*(d(\mathring{k}), \Theta)$ to combine own ideas about the unknown parameter $\Theta$ and the second participant's knowledge about the data $d(\mathring{k})$;
- selects the model $f(d(\mathring{k}), \Theta|f_1)$ within the created set $f^*(d(\mathring{k}), \Theta)$, which is the closest to the own "two"-dimensional model $f_1(d(\mathring{k}), \Theta)$;
- modifies the original prior pdf $f_1(\Theta)$ to the prior pdf $f(\Theta|f_2)$ so that the "two"-dimensional pdf $f(d(\mathring{k}), \Theta|f_2) \equiv f_1(d(\mathring{k})|\Theta) f(\Theta|f_2)$ is closest to $f(d(\mathring{k}), \Theta|f_1)$. Recall that the parameterized model $f_1(d(\mathring{k})|\Theta)$ is taken as the given one.

In **Step 2** the following method can be used to select a non-subjective length $\mathring{k}$.

**Step 2a** The **first participant** may select $\mathring{k}$, expressing its trust in the information offered by the second participant, using observed *realizations* of data $\underline{d}(t)$. It inserts them into likelihood function on possible compared lengths $\mathring{k}_\iota,\ \iota = 1, \ldots, \mathring{\iota}$,

$$\begin{aligned} L(\underline{d}(t), \mathring{k}_\iota) &= \prod_{\tau=1}^{t} f(\underline{d}_t|\underline{d}(t-1), \mathring{k}_\iota), \quad \text{with} \qquad (15) \\ f(d_t|d(t-1), \mathring{k}_\iota) &\propto \int f_1(d_t|\phi_{t-1}, \Theta) f_1(d(t-1)|\Theta) f_1(\Theta|f_2, \mathring{k}_\iota)\, d\Theta \text{ for } \mathring{k} = \mathring{k}_\iota, \end{aligned}$$

where $f_1(\Theta|f_2, \mathring{k}_\iota)$ are modifications of the original prior pdf $f_1(\Theta)$ by pdfs $f_2(d(\mathring{k}_\iota))$, $\iota = 1, \ldots, \mathring{\iota}$.

Typically, the first participant will select $\mathring{k} = \arg\max_{\mathring{k}_\iota, \iota=1,\ldots,\mathring{\iota}} L(\underline{d}(t), \mathring{k}_\iota)$.

## 6. VERIFICATION OF REASONABILITY OF $f(\Theta|f_2)$

The above construction contains heuristic steps like use of the extension step or the choice of the KL-divergence version. Thus, it is important to check whether the result has desirable properties in particular, practically important, cases.

### *6.1. Singularity of data models*

Let us inspect behavior of the constructed pdf $f(\Theta|f_2)$ when the data models are (partially) singular.

Let for some $k \in k^*$ and a given value $\Theta$ the $f_1(d_k|\phi_{k-1}, \Theta) = 0$ on a set $\Psi^*_{0;k} \subset \Psi^*$ for which $\int_{\Psi^*_{0;k}} f_2(\Psi_k)\, d\Psi_k > 0$. Then, it is easy to find that $f(\Theta|f_2) = 0$ for the considered $\Theta$ for the considered $\Theta$, see (13).

Thus, the parameter values, which lead to the essential discrepancy of predictions of both participants, get zero probability. This property seems to be intuitively desirable. The incorporation of the information contained in $f_2$ fails completely if it happens for all $\Theta \in \Theta^*$. This failure is good indicator of absolute incompatibility of both considered predictors. Of course, it is responsibility of the first participant whether it takes such a situation seriously or not.

### *6.2. Case of measured data*

Let us assume that the second participant observes realizations $\underline{d}_k$ of data records $d_k$, $k \in k^* \equiv \{1, \ldots, \mathring{k}\}$ and thus it is able to construct the realizations $\underline{\Psi}_k$ of data vectors $\Psi_k$. The realization $\underline{\Psi}_k$ of each data vector $\Psi_k$ is equivalent to the pdf $f_2(d_k, \phi_{k-1}) = f_2(\Psi_k) \equiv \delta(\Psi_k - \underline{\Psi}_k)$. The used Dirac function $\delta(\Psi_k - \underline{\Psi}_k)$ is the formal pdf of the probabilistic measure concentrated on the realization $\underline{\Psi}_k$. Thus, passing pdfs $f_2(\Psi_k) \equiv \delta(\Psi_k - \underline{\Psi}_k)$, $k \in k^*$, from the second participant to the first one leads to the same result as observing the realized data vectors $\underline{\Psi}(\mathring{k})$ by the first participant. Inserting the discussed pdfs into the general formula for the pdf $f(\Theta|f_2)$, we get

$$\begin{aligned} f(\Theta|f_2) &\propto f_1(\Theta) \exp\left( \sum_{k \in k^*} \int \delta(\Psi_k - \underline{\Psi}_k) \ln(f_1(d_k|\phi_{k-1}, \Theta))\, d\Psi_k \right) \\ &= f_1(\Theta) \prod_{k=1}^{\mathring{k}} f_1(\underline{d}_k|\underline{\phi}_{k-1}, \Theta) \propto f_1(\Theta|\underline{d}(\mathring{k})). \end{aligned} \tag{16}$$

In other words, $f(\Theta|f_2)$ reduces to the "ordinary" posterior pdf conditioned by the observed data if the pdf $f_2(\Psi_k)$ is fully concentrated on the observations.

### *6.3. Interplay of parametric and non-parametric estimation*

The expression for the posterior pdf used in (16) can be re-written in the following slightly modified form

$$\begin{aligned} f_1(\Theta|\underline{d}(t)) &\propto f_1(\Theta) \exp\left( \sum_{k=1}^{t} \int \delta(\Psi_k - \underline{\Psi}_k) \ln(f_1(d_k|\phi_{k-1}, \Theta))\, d\Psi_k \right) \\ &= f_1(\Theta) \exp\left( t \int \frac{1}{t} \sum_{k=1}^{t} \delta(\Psi - \underline{\Psi}_k) \ln(f_1(d|\phi, \Theta))\, d\Psi \right) \\ &\equiv f_1(\Theta) \exp\left( t \int s(\Psi|\underline{\Psi}(t)) \ln(f_1(d|\phi, \Theta))\, d\Psi \right). \end{aligned} \tag{17}$$

In the formula (17) all realized data vectors enter via the (formal) sample pdf

$$s(\Psi|\underline{\Psi}(t)) = \frac{1}{t} \sum_{k \in k^*} \delta(\Psi - \underline{\Psi}_k). \tag{18}$$

This non-parametric estimate of the distribution of data vector $\Psi$ enters *any* Bayesian parametric estimation whose parameterized model depends on the discussed data vector $\Psi$. Thus, the nonparametric estimation "precedes" the parametric one. As the second participant offers information in the space of data vectors only, it is natural to search for established ways of incorporating prior knowledge into the non-parametric estimation. The methodology based on Dirichlet processes is available to this purpose. It is well developed for independent observations [15] so that we restrict ourselves to this case in this paragraph. It makes no harm as we take the subsequent notes as an additional check only.

Let $\Psi_k$, $k \in k^*$, be mutually independent with a common, time-invariant, *unknown pdf* $f_1(\Psi)$. Within the cited framework, the prior distribution of this infinite-dimensional parameter is modelled as a Dirichlet process and it is determined by an expected value, say $g(\Psi)$, and by a scalar precision parameter, say $\nu_0 > 0$. Let us take the pdf $f_2(\Psi)$ offered by the second participant as the parameter of the prior Dirichlet process $g(\Psi) \equiv f_2(\Psi)$. The conjugation of Dirichlet process implies that the $f_1(\Psi)$ remains to be Dirichlet process also a posteriori. The posterior expected value and precision parameter $\nu_t$ assigned after observing $\underline{\Psi}(t)$ becomes

$$\begin{aligned} f_1(\Psi|\underline{\Psi}(t), f_2) &\equiv \frac{\nu_t - \nu_0}{\nu_t} s(\Psi|\underline{\Psi}(t)) + \frac{\nu_0}{\nu_t} f_2(\Psi) \\ \nu_t &\equiv t + \nu_0. \end{aligned} \tag{19}$$

The product $t \times s(\Psi|\underline{\Psi}(t))$, defining the posterior pdf (17), can be interpreted the expectation of the unknown pdf $f(\Psi)$ multiplied by $\nu_t$, i.e., $\nu_t \times f_1(\Psi|\underline{\Psi}(\mathring{t}), f_2)$, in the special, non-informative case obtained for $\nu_0 \to 0$. By replacing this estimate in (17) by the informative estimate (19), we get

$$\begin{aligned} & f(\Theta|\underline{d}(\mathring{t}), f_2) \propto f_1(\Theta) \\ & \times \quad \exp\left( \sum_{k=1}^{t} \int \delta(\Psi_k - \underline{\Psi}_k) \ln(f_1(y_k|\psi_k, \Theta))\, d\Psi_k + \nu_0 \int f_2(\Psi) \ln(f_1(y|\psi, \Theta))\, d\Psi \right), \end{aligned}$$

which, for $\nu_0 = \mathring{k}$, coincides with the ordinary Bayesian learning that exploits the recommended prior pdf $f(\Theta|f_2)$ (13).

## 7. CONCLUSIONS

The elaborated methodology of passing external knowledge in terms of data distribution into a parametric estimation seems to posses a wide range of desirable properties, including simplicity. The application range is even wider: it includes not only the motivating cooperation of participants but also quantification of prior knowledge, re-sampling of continuous time signals, approximation of complex models by simpler ones etc. These applications together with encouraging experiments will be described in independent papers which are under preparations. The attempt to focus on the problem, its solution and basic interpretation made us to omit experiments of this type here. We however want to stress that the derived methodology is very practical and vital for applications with a few data like those met in nuclear medicine [17, 18].

## REFERENCES

1. V. Peterka, "Bayesian system identification", in *Trends and Progress in System Identification*, P. Eykhoff, Ed., pp. 239–304. Pergamon Press, Oxford, 1981.
2. J.O. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York, 1985.
3. T. Bohlin, *Interactive System Identification: Prospects and Pitfalls*, Springer-Verlag, New York, 1991.
4. 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.
5. D.M. Titterington, A.F.M. Smith, and U.E. Makov, *Statistical Analysis of Finite Mixtures*, John Wiley, New York, 1985.
6. R. Kulhavý, *Recursive Nonlinear Estimation: A Geometric Approach*, vol. 216 of *Lecture Notes in Control and Information Sciences*, Springer-Verlag, London, 1996.
7. 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, ISBN 1-85233-928-4, pp. 552.

8. M. Kárný, P. Nedoma, and V. Šmídl, "Cross-validation of controlled dynamic models: Bayesian approach", in *IFAC World Congress, Preprints*, IFAC, Ed. IFAC, Prague, 2005.
9. M. Kárný and T.V. Guy, "On dynamic decision-making scenarios with multiple participants", in *Multiple Participant Decision Making*, J. Andrýsek, M. Kárný, and J. Kracík, Eds., Adelaide, May 2004, pp. 17–28, Advanced Knowledge International.
10. J. Kracík and Kárný M, "Merging of data knowledge in Bayesian estimation", in *ICINCO*, Barcelona, September 2005, accepted, IFAC.
11. M. Kárný, "Towards fully probabilistic control design", *Automatica*, vol. 32, no. 12, pp. 1719–1722, 1996.
12. M. Kárný and T.V. Guy, "Fully probabilistic control design", *Systems & Control Letters*, 2004, accepted.
13. M. Kárný and T.V. Guy, "Stationary fully probabilistic control design", in *ICINCO*, Barcelona, September 2005, accepted, IFAC.
14. S. Kullback and R. Leibler, "On information and sufficiency", *Annals of Mathematical Statistics*, vol. 22, pp. 79–87, 1951.
15. T.S. Ferguson, "A Bayesian analysis of some nonparametric problems", *The Annals of Statistics*, vol. 1, pp. 209–230, 1973.
16. D.R. Insua and F. Ruggeri Eds., *Robust Bayesian Analysis*, Lecture Notes in Statistics. Springer Verlag, New York, 2000.
17. P. Gebouský, M. Kárný, and A. Quinn, "Lymphoscintigraphy of upper limbs: A Bayesian framework", in *Bayesian Statistics 7*, J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. David, D. Heckerman, and A.F.M. Smith, Eds., Oxford, 2003, pp. 543–552, Clarendon Press, Proceedings of the Seventh Valencia International Meeting, June 2–6, 2002.
18. J. Heřmanská, M. Kárný, J. Zimák, L. Jirsa, and M. Šámal, "Improved prediction of therapeutic absorbed doses of radioiodine in the treatment of thyroid carcinoma", *The Journal of Nuclear Medicine*, vol. 42, no. 7, pp. 1084–1090, 2001.