PROCESSING OF EXPERT INFORMATION IN BAYESIAN PARAMETER ESTIMATION

Jan Kracík

Department of Adaptive Systems
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Prague, CZECH REPUBLIC
E-mail: `kracik@utia.cas.cz`

Abstrakt: In Bayesian statistics an expert information is typically represented by a prior probability density function of a parameter. However, often an expert is able to provide an information in form of probability distribution of data only. In this paper we propose a method for utilization of an expert information of this type.

Keywords: Bayesian estimation, multiple participant decision making, expert information

1. INTRODUCTION

Efficient multiple-participant decision making (Andrýsek *et al.*, 2004) relies on a cooperation of participants. Basic element of such cooperation is a mutual exchange of individual knowledge pieces, which are, in case of Bayesian decision makers (Berger, 1985), represented by probability density functions. As individual participants may use completely different parameterized models, the knowledge can be exchanged only by sharing probability density functions (pdfs) of quantities describing data which are in common of communicating participants. The pdf provided by another participant represents, in fact, an expert information described by probability distribution of data quantity.

In this paper, a rationale and technically feasible framework which allows a Bayesian decision maker to utilize an expert information in form of a probabilistic model of data is proposed in a simple form. The presented method is consistent with Bayesian learning from data – learning from an external probabilistic model is equivalent to the ordinary Bayesian learning, if the external model is described by an empirical probability density function.

2. PROBLEM FORMULATION

A participant, as a Bayesian decision-maker, is supposed to model a system (i.e. a part of the world of his interest) by a conditional pdf

$$f(y_t|\psi_t, \Theta) \equiv f(y_t|u_t, d(t-1), \Theta). \quad (1)$$

In it, the modelled system output y_t depends on a system input u_t and past data history $d(t-1) \equiv (d_0, d_1, \ldots, d_{t-1})$, $d_\tau = (y_\tau, u_\tau)$, via a finite-dimensional regression vector ψ_t only.

The model 1 is supposed to be known up to an unknown parameter $\Theta \in \Theta^*$. For the purpose of a decision making, the participant has to collect information about the parameter Θ, which is performed by evaluating the posterior pdf

$$f(\Theta|d(t)) \propto f(\Theta) \prod_{\tau=1}^{t} f(y_t|\psi_t, \Theta), \tag{2}$$

where $\propto$ means equality up to a uniquely determined normalizing factor, which is independent of Θ. $f(\Theta) \equiv f(\Theta|d(0))$ represents an initial knowledge about parameter Θ. Bayes rule 2 holds under a naturally fulfilled assumption (Peterka, 1981) that inputs u_t depends only on the observed data, i.e., $f(u_t|d(t-1), \Theta) = f(u_t|d(t-1))$.

Another participant is assumed to deal with physically the same data $d(\cdot)$ (possibly different observations) and generate a joint pdf $h(\Psi) \equiv h(\Psi_\tau)$ of a data vector $\Psi_\tau \equiv (y_\tau, \psi_\tau)$ for some τ. $h(\Psi)$ can be, for instance, an output predictor obtained via Bayesian estimation and prediction of a model, which differs from $f(\Psi_t, \Theta)$. This participant provides its knowledge of $h(\Psi)$ to the former one. Another possibility is to interpret $h(\Psi)$ as an additional information provided by an expert. Question arises how this information can be used for correcting the posterior pdf of Θ. An answer to this question is the problem addressed within the paper.

3. MERGING DATA BASED KNOWLEDGE

The Bayes rule (3) can be rewritten as follows

$$\begin{aligned} f(\Theta|d(t)) &\propto f(\Theta) \exp\left[\sum_{\tilde{t}=1}^{t} \ln(f(y_{\tilde{t}}|\psi_{\tilde{t}}, \Theta))\right] = \\ &= f(\Theta) \exp\left[\int \sum_{\tilde{t}=1}^{t} \delta(\Psi - \Psi_{\tilde{t}}) \ln(f(y|\psi, \Theta)) d\Psi\right]. \end{aligned} \tag{3}$$

The expression $\sum_{\tau=1}^{t} \delta(\Psi - \Psi_\tau)$, determined by the Dirac delta function, can be interpreted as t-multiple of the empirical pdf f_t on Ψ.

The presented form of the posterior pdf has an important consequence: *the number of data records together with the empirical pdf of data vectors form a sufficient statistic for estimation of any parameterized model that deals with the data vectors* $\{\Psi_t\}$. Furthermore, *updating posterior pdf* $f(\Theta|d(t))$ *by another data records, say* $d_{t+1}, \ldots, d_{\tilde{t}}$*, is equivalent to adding sufficient statistic corresponding to* $d_{t+1}, \ldots, d_{\tilde{t}}$ *to the statistic* $\sum_{\tau=1}^{t} \delta(\Psi - \Psi_\tau)$.

These facts lead us to the following idea. Having a knowledge expressed by $h(\Psi)$ instead of a set of samples we use in 3 the pdf $h(\Psi)$ directly in place of the empirical pdf. Contrary to $h(\Psi)$, the weight ν assigned to the information $h(\Psi)$ is not supposed to be given. Generally, it is subjectively assigned by the the participant making the parametric estimation, and expresses the weight (corresponding to the number of "virtual" observations) it gives to the participant serving as information source.

Finally, the resulting "posterior" pdf combining the prior pdf $f(\Theta)$, data $d(t)$, and an expert information $h(\Psi)$ with weight ν has a form

$$f(\Theta|d(t), h, \nu) \propto f(\Theta) \exp\left\{\int \left[\sum_{\tilde{t}=1}^{t} \delta(\Psi - \Psi_{\tilde{t}}) + \nu h(\Psi)\right] \ln(f(y|\psi, \Theta))\, d\Psi\right\} =$$

$$= \quad f(\Theta|d(t)) \exp\left[\nu \int h(\Psi) \ln(f(y|\psi,\Theta))\, d\Psi\right]. \tag{4}$$

Remarks

1. Evidently, $f(\Theta|d(t), h, \nu)$ is neither a real posterior pdf nor a conditional pdf as $f(h, \nu)$ is not defined. It should be taken rather as a symbol with a clear meaning.

2. In the proposed method the information $h(\Psi)$ is processed "data-like" in the following sense. Suppose that $h(\Psi)$ is an empirical density from ν data records, i.e., $h(\Psi) = \frac{1}{\nu}\sum_{\tilde{t}=1}^{\nu} \delta(\Psi, \Psi_{\tilde{t}})$, and data vectors $\Psi_1, \ldots, \Psi_\nu$ arise from a sequence of data $d(\nu)$. Then
$$f(\Theta|h, \nu) = f(\Theta|d(\nu)).$$

3. An intuitive way how to utilize an information $h(\Psi)$ as ν data records is to generate ν random samples from $h(\Psi)$ and evaluate the posterior pdf with these samples. For sufficiently large ν such a posterior pdf is to be close to the posterior $f(\Theta|h, \nu)$ as the empirical distribution converges to the real one. However, for small ν the posterior pdf based on the random samples strongly depends on their realization while $f(\Theta|h, \nu)$ is not influenced by any randomness.

4. CONCLUSIONS

The simple presented result is a quite powerful and practical tool. Considering the parameterized model $f(y|\psi, \Theta)$ from an exponential family and a conjugate prior pdf, the posterior pdf $f(\Theta|h, \nu)$ remains in the conjugate form as it is in "proper" Bayesian estimation. Evaluation of $\int h(\Psi) \ln(f(y|\psi, \Theta))\, d\Psi$ often reduces into evaluation of moments of Ψ. Moreover, a simulation model of a quite different nature than the estimated one can be used for estimating $\int h(\Psi) \ln(f(y|\psi, \Theta))\, d\Psi$. In this case, it is often reduced into evaluation of sample moments of Ψ.

More complex models – probabilistic mixtures – can be estimated by the proposed method using, e.g., a slightly modified Quasi-Bayes algorithm (Kárný *et al.*, 2005).

The choice of the weight ν is an open problem, whose solution is, however, predictable: Bayesian hypothesis testing and real data observed by the participant should provide flexible universal solution.

In summary, the proposed method serves not only for communication of participants in multiple participant decision making, but more generally for exploiting any expert information in form of pdf of data to the Bayesian estimation.

ACKNOWLEDGEMENTS

This work has been supported by AV ČR grants 1ET 100750401, 100750404, S1075351, GA ČR grant no. 102/03/0049, MŠMT 1M0572 DAR, and Czech-Slovenian project MŠMT 8-2006-06 "Data-driven modelling for decision making support and process monitoring".

REFERENCE

Andrýsek, J., Kárný, M. and Kracík, J.., Eds. (2004). *Multiple Participant Decision Making*. Advanced Knowledge International. Adelaide.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag. New York.

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

Peterka, V. (1981). Bayesian system identification. In: *Trends and Progress in System Identification* (P. Eykhoff, Ed.). pp. 239–304. Pergamon Press. Oxford.