MERGING OF DATA KNOWLEDGE IN BAYESIAN ESTIMATION

Jan Kracík
Institute of Information Theory and Automation
P.O.box 18, 182 08 Praha 8, Czech Republic

Miroslav Kárný
Institute of Information Theory and Automation
P.O.box 18, 182 08 Praha 8, Czech Republic

Keywords: Bayesian estimation, prior information, multiple-participant decision making.

Abstract: Efficient multiple participant decision-making relies on cooperation of participants. Partially, it is reached by sharing knowledge. A specific but important case of this type is addressed here. Essentially, a participant passes to its partner distribution on common data and partner uses it for correcting its Bayesian parameter estimate.

1 INTRODUCTION

Decision making (DM) is the ultimate purpose of any cognitive system serving at various scales and domains: international, state or local-community levels; particular technical, medical and societal organizations; individual human beings etc. Attempts to optimize centrally the overall performance of a collection of mutually interacting participants reach soon the communication and evaluation complexity barriers. Use of distributed DM methodologies is then the only viable way towards desirable efficiency. Existing solutions overcome the complexity barrier by exploiting specificity of their application domains. Their transfer to different domains is, however, expensive in skilled manpower. None of them is able to serve as a common domain-independent pattern and thus the real need for applicable theory of distributed DM persists.

Careful inspection of DM (Savage, 1954; Berger, 1985) identifies the Bayesian theory as a prime candidate. Practical consequence, relevant to this paper, is that different subjects of distributed DM (participants) share a probabilistic information when cooperating. Existing approaches to a combination of low dimensional pdfs suffer from a significant ambiguity, e.g. (Jiroušek, 2003; Meneguzzo and Vecchiato, 2004). Furthermore, these approaches can be hardly integrated into the Bayesian framework. This motivated a research whose part is presented in this paper. Solution of a partial but important task – using of probabilistically described knowledge of data provided by another participant for improving Bayesian parameter estimation – is presented.

2 PROBLEM FORMULATION

A participant estimates an unknown finite-dimensional parameter Θ determining the parameterized model $m(\Psi_t, \Theta) \equiv f(y_t|\psi_t, \Theta) \equiv f(y_t|u_t, d(t-1), \Theta)$, where $f(\cdot|\cdot)$ is a conditional probability density function (pdf). 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 data vector Ψ_t is coupling of the modelled output y_t and of the corresponding regression vector ψ_t. Prior information, labelled d_0, is attached to the observed sequence $d_1, \ldots, d_{t-1}$. The participant estimates Θ in Bayesian way, i.e., evaluates the posterior pdf

$$f(\Theta|d(t)) \propto f(\Theta) \prod_{\tau=1}^{t} m(\Psi_\tau, \Theta). \qquad (1)$$

The symbol $\propto$ expresses equality without writing the normalizing data-dependent proportionality factor. The prior pdf $f(\Theta) \equiv f(\Theta|d_0)$ is related to the posterior pdf by the above version of Bayes rule iff the parameter Θ is unknown to the input generator, i.e.,
$f(u_t|d(t-1), \Theta) = f(u_t|d(t-1))$ (Peterka, 1981).

Kracík J. and Kárný M. (2005).
MERGING OF DATA KNOWLEDGE IN BAYESIAN ESTIMATION.
In *Proceedings of the Second International Conference on Informatics in Control, Automation and Robotics - Signal Processing, Systems Modeling and Control*, pages 229-232
DOI: 10.5220/0001174502290232

Another participant is assumed to deal with physically the same data $d(t)$ (possibly different realizations) and generate their joint pdf $f(d(t)) = \prod_{\tau=1}^{t} f(d_t|d(t-1))$ and evaluate marginal pdfs $M(\Psi_\tau)$ of data vectors. For simplicity of presentation, we assume that this function is time invariant. The pdfs $f(d_t|d(t-1))$ can be, for instance, output predictors obtained via Bayesian estimation and prediction of a model, which differs from $m(\Psi_t, \Theta)$. This participant provides its knowledge of $M(\Psi_t)$ to the former one. Another possibility is to interpret $M(\Psi_t)$ 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 SUFFICIENT STATISTIC FOR ANY PARAMETERIZED MODEL

The Bayesian parameter estimation is described by the Bayes rule (1). It can be rewritten as follows

$$f(\Theta|d(t)) \propto f(\Theta) \exp\left[\sum_{\tau=1}^{t} \ln(m(\Psi_\tau, \Theta))\right] =$$

$$= \exp\left[\int \sum_{\tau=1}^{t} \delta(\Psi - \Psi_\tau) \ln(m(\Psi, \Theta))\, d\Psi\right]. \quad (2)$$

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 on set Ψ^* of possible data vectors Ψ. Formally clean version is obtained by the correct interpretation of $\int \delta(\Psi - \Psi_\tau) g(\Psi)\, d\Psi$ as the linear functional assigning to a function $g(\Psi)$ its value in Ψ_τ. The quotation marks at the term empirical distribution stress that against the traditional assumptions the involved data vectors are statistically dependent.

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 other data records, say* $d_{t+1}, \ldots, d_{\bar{t}}$*, is equivalent to adding sufficient statistic corresponding to* $d_{t+1}, \ldots, d_{\bar{t}}$ *to the statistic* $\sum_{\tau=1}^{t} \delta(\Psi - \Psi_\tau)$.

4 MERGING DATA BASED KNOWLEDGE

The observations made in the previous section determine the way to incorporate knowledge expressed by $M(\Psi)$ into the parametric estimation connected with the model $m(\Psi, \Theta)$. Taking information $M(\Psi)$ as a pdf of, say ν, virtual observations, the sufficient statistic for the posterior pdf $f(\Theta|d(t), M, \nu)$ based on both real and virtual observations is determined by $t + \nu$ data records with the pdf

$$\frac{1}{t+\nu} \sum_{\tau=1}^{t} \delta(\Psi - \Psi_\tau) + \frac{\nu}{t+\nu} M(\Psi)$$

in the place of the empirical pdf. Note that the idea of virtual data is quite common, e.g. (Kárný et al., 2001). For instance, Bayesian estimation with a conjugate prior pdf is often interpreted as estimation with additional virtual data (determining the original prior) and a uniform prior pdf.

Contrary to $M(\Psi)$, the weight ν assigned to the information $M(\Psi)$ is not supposed to be given. Generally, it is subjectively assigned by the the participant making the parametric estimation, and expresses the weight it gives to the participant serving as an information source.

Using this way, we get the parameter estimate that respects both knowledge sources

$$f(\Theta|d(t), M, \nu) \propto f(\Theta) \exp\{ \quad (3)$$

$$\int \left[\sum_{\tau=1}^{t} \delta(\Psi - \Psi_\tau) + \nu M(\Psi)\right] \ln(m(\Psi, \Theta))\, d\Psi\}$$

$$\propto f(\Theta|d(t)) \exp\left[\nu \int M(\Psi) \ln(m(\Psi, \Theta))\, d\Psi\right].$$

Remarks

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

$$f(\Theta|M, \nu) = f(\Theta|d(\nu)).$$

2. An intuitive way to use information $M(\Psi)$ as ν data records is to generate ν random samples from $M(\Psi)$ and evaluate the posterior pdf with these samples. For sufficiently large ν such posterior pdf is expected to be close to the posterior $f(\Theta|M, \nu)$ 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 $f(\Theta|M, \nu)$ is not influenced by any randomness.

3. The "merging" weights are controlled by the optional scalar $\nu > 0$.
4. It is worth stressing that the function $M(\Psi_t)$ is to be joint pdf of the output y_t and the regression vector ψ_t similarly as in the case of independent Ψs.

5 EXAMPLES IN EXPONENTIAL FAMILY

Let us consider a parameterized model in the exponential family (Barndorff-Nielsen, 1978)

$$m(\Psi, \Theta) = A(\Theta) \exp\left\langle B(\Psi), C(\Theta)\right\rangle, \qquad (4)$$

where the functions A, B, C are known functions of respective arguments. $A(\Theta) \geq 0$ is the scalar one, B, C are vectorial functions of compatible dimensions and $\langle B(\Psi), C(\Theta)\rangle$ is a functional linear in the first argument.

Let us suppose that the function $M(\Psi)$ defines well the expectation

$$V \equiv \int M(\Psi) B(\Psi)\, d\Psi. \qquad (5)$$

Then, the factor modifying the prior pdf has the conjugated form

$$g(\Theta, \nu, V) \equiv A(\Theta)^{\nu} \exp\left\langle \nu V, C(\Theta)\right\rangle. \qquad (6)$$

If the prior pdf is also chosen conjugated one

$$f(\Theta) = \frac{g(\Theta, \bar{\nu}, \bar{V})}{\mathcal{I}(\bar{\nu}, \bar{V})},\ \mathcal{I}(\nu, V) = \int g(\Theta, \nu, V)\, d\Theta, \qquad (7)$$

then the posterior pdfs have the same fixed functional form given by $g(\Theta, \nu_t, V_t)$ with the statistics ν_t, V_t evolving as follows

$$\begin{aligned} \nu_t &= \nu_{t-1} + 1,\ V_t = V_{t-1} + B(\Psi_t), \qquad (8)\\ \nu_0 &= \bar{\nu} + \nu,\ V_0 = \bar{V} + \nu V. \end{aligned}$$

Thus, the externally supplied pdf $M(\Psi)$ adds ν and V to the initial values of the statistics selected by the participant that runs the parameter estimation.

If the DM task allows us to wait for collecting the statistics $\bar{V}_t = \sum_{\tau=1}^{t} B(\Psi_\tau) + \bar{V}$ and $\bar{\nu}_t = t + \bar{\nu}$ for some realization of data vectors, it is possible to select the optimal weight ν_o by maximizing the corresponding posterior likelihood function

$$\nu_o = \arg\max_{\nu} \frac{\mathcal{I}(\nu + \bar{\nu}_t, \bar{V}_t + \nu V)}{\mathcal{I}(\nu, \nu V)}, \qquad (9)$$

where $\bar{\nu}_t = t + \bar{\nu}$ and $\bar{V}_t = \sum_{\tau=1}^{t} B(\Psi_\tau) + \bar{V}$.

If we cannot wait, several competitive values of ν have to be chosen and the corresponding posterior likelihoods compared in recursive mode.

Normal ARX model is the most prominent example of a dynamic model in the exponential family. It is described by the parameterized model

$$\begin{aligned} & m(\Psi, \Theta) \equiv \mathcal{N}_{y_t}(\theta' \psi_t, r) \qquad (10)\\ = & \underbrace{\frac{1}{\sqrt{2\pi r}}}_{A(\Theta)} \exp \underbrace{\mathrm{tr}\left(\underbrace{\Psi_t \Psi_t'}_{B(\Psi_t)} \underbrace{\frac{-1}{2r}[-1, \theta'][-1, \theta']'}_{C(\Theta)} \right)}_{\langle \cdot, \cdot \rangle} \end{aligned}$$

where $\mathcal{N}_y(\mu, \rho)$ is normal pdf with mean μ and variance ρ; the regression coefficients θ and variance form the unknown parameter Θ, $\mathrm{tr}(N)$ is the trace of matrix N, and $'$ denotes transposition.

The marked correspondence with exponential family shows that the moments needed in connection with $M(\Psi)$ are the non-central second moments of the data vector Ψ

$$V = \int M(\Psi) \Psi \Psi'\, d\Psi. \qquad (11)$$

The updating (8), describing completely the posterior pdfs in the conjugate Gauss-inverse-Wishart form, can be shown to be algebraically equivalent to recursive least-squares algorithm (Peterka, 1981). The information from the second participant simply modifies its initial conditions. Their careful choice is known to influence substantially the transient behavior of the algorithm. Often, it is vital, especially in closed decision-making (control) loop.

Controlled Markov chain is another example of the model describing well dynamic systems. It models discrete-valued outputs that depend on discrete-valued regression vector by the table

$$\begin{aligned} & f(y_t | u_t, d(t-1), \Theta) = m(\Psi_t, \Theta) \equiv \qquad (12)\\ \equiv\ & \Theta_{y_t|\psi_t} = \exp \underbrace{\left[\sum_{\Psi \in \Psi^*} \underbrace{\delta(\Psi - \Psi_t)}_{B_\Psi(\Psi_t)} \underbrace{\ln(\Theta_{y|\psi})}_{C_\Psi(\Theta)} \right]}_{\langle \cdot, \cdot \rangle} \end{aligned}$$

whose entries $\Theta_{y|\psi}$ form the unknown parameter Θ. The parameter belongs to a subset (determined possibly by some additional information) of the convex set $\Theta^* \equiv \left\{ \Theta_{y|\psi} : \Theta_{y|\psi} \geq 0,\ \sum_{y \in y^*} \Theta_{y|\psi} = 1 \right\}$.

The externally supplied model $M(\Psi)$ simply assigns probabilities to various possible values of $\Psi \in \Psi^*$ and the factor modifying the prior pdf has the form

$$\exp\left(\nu \sum_{\Psi \in \Psi^*} M(\Psi) \ln(\Theta_{y|\psi}) \right) = \prod_{\Psi \in \Psi^*} \Theta_{y|\psi}^{\nu M(\Psi)}. \qquad (13)$$

This expression is proportional to the conjugate Dirichlet pdf determined by the table $\nu M(\Psi)$ that

can be interpreted as the number of occurrences of the data vector Ψ. Choosing the prior pdf $f(\Theta)$ in the Dirichlet form $\propto \prod_{\Psi\in\Psi^*} \Theta_{y|\psi}^{\bar{V}_{y|\psi}-1}$, the externally supplied information increases it to the initial value $V_0 = \bar{V} + \nu M$. The posterior pdf is also Dirichlet one given by the occurrence table V_t. It evolves starting from the initial value V_0. The updating by the observed data $V_t = V_{t-1} + B(\Psi_t)$ adds the number of occurrences of the values $\Psi_\tau = \Psi$, $\tau \leq t$, to the Ψth entry of the table V_0.

Again, importance of the prior knowledge can be hardly over-stressed: the estimation of controlled Markov chains is formally extremely simple but the dimension of the occurrence table V grows exponentially with the cardinality of the set Ψ^*. Consequently, there is a lack of data in majority practical cases and, moreover, their information content is as a rule insufficient.

6 CONCLUSIONS

The simple presented result is a quite powerful and practical tool. Considering the parameterized model $m(\Theta, \Psi)$ from the exponential family and a conjugate prior pdf, the posterior pdf $f(\Theta|M, \nu)$ remains in the conjugate form as it is in "proper" Bayesian estimation. Evaluation of $\int M(\Psi)\ln(m(\Psi,\Theta))\,d\Psi$ 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 $\int M(\Psi)\ln(m(\Psi,\Theta))\,d\Psi$. In this case, the use of $M(\Psi)$ 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).

Physically motivated models, black box models of much higher order than the estimated one, relationships described by production rules stimulated by real data in past etc. may serve as data-vectors sources. In this way, model simplification and translation between various knowledge domains are addressed in a justified, purposeful and simple way.

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

Similarly, the case when information about some entries of Ψ_t is only offered is unsolved. It is expected that assignment the extension of M to the set Ψ^* by a very flat marginal on "non-reported" entries of Ψ will solve this problem.

Even with this open problems pending, the proposed "technology" is straightforward to pass uncertain knowledge from one participant to another one and thus to combine very different knowledge sources.

ACKNOWLEDGEMENTS

This research has been partially supported by GAČR grant 102/03/0049, AV ČR project BADDYR 1ET100750401, MŠMT grant 1M6798555601 DAR, and AV ČR project 1ET100750404.

REFERENCES

Barndorff-Nielsen, O. (1978). *Information and exponential families in statistical theory*. Wiley, New York.

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

Jiroušek, R. (2003). On experimental system for multidimensional model development MUDIN. *Neural Network World*, (5):513–520.

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

Kárný, M., Khailova, N., Nedoma, P., and Böhm, J. (2001). Quantification of prior information revised. *International Journal of Adaptive Control and Signal Processing*, 15(1):65–84.

Meneguzzo, D. and Vecchiato, W. (2004). Copula sensitivity in collateralized debt obligations and basket default swaps. *Journal of Futures Markets*, 24(1):37–70.

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

Savage, L. (1954). *Foundations of Statistics*. Wiley, New York.