KYBERNETIKA — VOLUME 26 (1990), NUMBER 1

RECURSIVE BAYESIAN ESTIMATION UNDER MEMORY LIMITATION

RUDOLF KULHAVÝ

When the Bayes estimation scheme is to be applied *recursively*, a numerically feasible approximation of the ideal procedure is generically necessary. In this paper a conceptually novel approach to the approximation problem is elaborated. In contrast to known methods, the posterior probability density of an unknown parameter is interpreted as uncertain and its uncertainty is described by probabilistic tools again. It is shown that storing just a limited finite-dimensional description of posterior densities we are able to infer in a Bayes coherent way only in terms of transitions between equivalence classes of densities with the same description.

1. INTRODUCTION

Computational realization of the Bayesian approach to parameter estimation [1] is not at all trivial in nonstandard cases. In a *recursive* version, Bayes estimation immediately processes observations to update sequentially the probability distribution of unknown quantities. This way is inevitable for adaptive control or adaptive signal processing, for instance. It needs, however, to cope with computational complexity and especially memory demands due to storing a too large or continually growing statistic.

The first collection of papers concerning approximation of the ideal inference appeared in the late 1960s in connection with attempts to apply the Kalman filtering approach to systems which did not fulfil the assumptions of model linearity or normality of the underlying distribution. Many solutions were designed in the vein of the extended Kalman filter [2] or using more sophisticated expansions of the posterior density [3, 4]. An attempt to find a more global approach to density approximation resulted in the early 1970s in proposing several powerful schemes [5]–[10] (see [11] for a survey). Nevertheless, the problem has remained of interest and new approaches and methods have appeared since then – the list of references [12]–[19], far from being complete, can be mentioned at this place.

The approach elaborated conceptually in this paper is based on the observation that incompletely described posterior densities become in recursive estimation as uncertain as the parameter itself. By the way, this is the reason why recursive estimation is not analysable so easily as in the case when the true posterior is at disposal (see [19] for an interesting study of this case). In spite of this fact, most of the above mentioned techniques have tried to estimate the posterior density of the unknown parameter in a "point-estimation" way – without investigating or even discussing the influence of different approximation methods on uncertainty of the posterior density.

This is why we attempt to explore the problem of recursive Bayes inference under a limited description of posterior densities more thoroughly than usual. The fact that estimation and not filtering is dealt with is not accidental. On the one hand, the estimation case is more transparent because of its simpler structure. On the other hand, it turns out to be more critical and fundamental. The use of a model of time variations for unknown quantities in the filtering case usually prevents accumulation of errors and thus makes approximation less sensitive.

In our solution we strongly require that the results of estimation for full-description and limited-description cases are *coherent*. The concept of coherence is used here in a more restricted sense than e.g. by de Finetti [20]. Let us suppose two observers, one of which is able to accomplish the ideal Bayes procedure, but the other may store only a limited description of posteriors. Then we require that their conclusions are identical as regards the equivalence classes of probability densities induced by the used description. Incidentally, coherence in the above sense implies also the same results of one-shot and recursive estimation, a property very rare when known approximation methods are used.

Preliminary concepts of recursive Bayes estimation and of a limited finite-dimensional description of probability densities are introduced in Section 2. The key Section 3 poses and answers the question which posterior description is *Bayes-closed* in the sense that the description itself or, more generally, its probability distribution is computable in a recursive manner without complete knowledge of the true posterior density. The problem is analysed within a hierarchical Bayesian framework in the vein of Good [21, 22]. It is shown in Section 3.4 that the description closure requirement leaves surprisingly little freedom in the choice of a density description. Section 4 deals with approximation of the posterior density based on the Bayes-closed description. Using the derived results, a coherent framework for approximation is designed in Section 5. The concluding Section 6 summarizes the main results and indicates possible extensions of the approach.

2. PRELIMINARIES

2.1 Bayes parameter estimation

The standard Bayes estimation scheme [1] will be assumed throughout the paper.

Consider a stochastic system on which *data* are observed at discrete time instants labelled $t = 1, 2, \ldots$. Data include a directly manipulated input u_t and indirectly affected output y_t. The collection of all data items available up to time t is denoted by

$$x_t = (u_1, y_1, \ldots, u_t, y_t) .$$

To denote that no data are available, x_0 is used formally.

Assume the dependence of the system output on previous data described by a suitable parametric family of probability densities

$$p(y_t \mid x_{t-1}, u_t, \theta) = m(x_t; \theta) , \tag{1}$$

conditioned on the data observed up to the time $t - 1$, the latest input u_t and a constant (generally multivariate) *parameter* $\theta \in \Theta$. The shorthand notation $m(x_t; \theta)$ is preferred in the sequel when speaking about the likelihood function of θ for known data x_t. Assume that to evaluate the likelihood $m(x_t; \theta)$ needs to store only a finite limited amount of data.

Express prior uncertainty of θ through a probability density $p(\theta \mid x_0) = p(\theta)$ built on the basis of relevant initial information.

Then the updating of the density of the unknown parameter θ by data is given by the Bayes theorem, which in the case when the input generator employs no information about θ other than measured data (cf. the natural conditions of control in [1])

$$p(u_t \mid x_{t-1}, \theta) = p(u_t \mid x_{t-1}) \tag{2}$$

reduces to

$$p(\theta \mid x_t) \propto m(x_t; \theta)\, p(\theta \mid x_{t-1}) . \tag{3}$$

The symbol $\propto$ stands for proportionality (equality up to a normalizing factor).

Remark 1. All probability densities mentioned above and introduced in the sequel are understood as Radon-Nikodým derivatives of corresponding probability distributions with respect to suitable dominating measures. To keep notation on a reasonably simple level, we use mostly the symbol $p(\cdot)$ for densities and $\lambda(\cdot)$ for dominating measures.

2.2 Limited description of posterior density

In actual estimation we are able to store only a limited finite-dimensional description of densities $p(\theta \mid x_t)$. Let us formalize the concept of a posterior description exactly.

To avoid problems with considering probabilities on infinitely-dimensional spaces, we shall assume that the density $p(\theta \mid x_t)$ can be interpreted, for any t, as a member

of a sufficiently rich family parametrized by an N-dimensional *hyperparameter* w (w could be viewed as a sufficient statistic too)

$$\mathscr{F} = \{p(\theta; w_t) \mid w_t \in W \subset \mathbb{R}^N\} . \tag{4}$$

A one-to-one correspondence between the posterior $p(\theta \mid x_t)$ and the hyperparameter w_t is assumed throughout the paper.

The above assumptions may seem artificial at the first sight. However, note that working with quantized data, the number of different likelihoods $m(x_t; \theta)$ is always finite (although very large in general). This makes our formulation quite realistic.

Now the posterior description represents an n-dimensional vector mapping ($n \leqq N$)

$$\chi: \mathscr{F} \to \mathbb{R}^n . \tag{5}$$

The mapping induces on the family $\mathscr{F}$ or, equivalently, the hyperparameter space W an obvious equivalence relation

$$p(\theta; w_t) \sim p(\theta; \tilde{w}_t) \Leftrightarrow w_t \sim \tilde{w}_t \Leftrightarrow \chi(p(\theta; w_t)) = \chi(p(\theta; \tilde{w}_t)) . \tag{6}$$

Mere knowledge of the description χ allows to distinguish just among the *equivalence classes* of densities

$$[p(\theta; w_t)] \triangleq \{p(\theta; \tilde{w}_t) \mid p(\theta; \tilde{w}_t) \sim p(\theta; w_t)\} \tag{7}$$

or, equivalently, the equivalence classes of hyperparameters

$$[w_t] \triangleq \{\tilde{w}_t \mid \tilde{w}_t \sim w_t\} . \tag{8}$$

With respect to this fact, it is useful to assume a more specific parametrization of the family $\mathscr{F}$

$$w_t = (w_t^+, w_t^-) \in W = W^+ \times W^- \tag{9}$$

with

$$w_t^+ \triangleq \chi(p(\theta; w_t)) . \tag{10}$$

Here the superscript w^+ suggests available while w^- missing information. Note that the description hyperparameter w_t^+ specifies only the equivalence class $[p(\theta; w_t)]$ in which $p(\theta; w_t)$ lies while additional knowledge of the complementary hyperparameter w_t^- makes it possible to determine the density $p(\theta; w_t)$ completely.

3. BAYES-CLOSED DESCRIPTION OF POSTERIOR

3.1 Problem formulation

In the standard estimation scheme, the evolution of $p(\theta \mid x_t)$ is fully specified by the system model defining the density (1) and by the prior density $p(\theta)$. Therefore, there is no uncertainty in determining $p(\theta \mid x_t)$.

In the case that just a limited description χ is at our disposal, the set of all possible densities matching the current value of the description must be considered for Bayes inference and the density $p(\theta \mid x_t)$ of the unknown parameter θ becomes

uncertain. Owing to the limited description, we are able to infer at most about the equivalence classes of equidescription densities. However, is it possible to infer in a Bayes-coherent way about equivalence classes without inferring at the same time about densities within these classes? It is not difficult to find descriptions which are not "closed" with respect to the Bayes rule. A classical example is the description by the first n moments of the unknown parameter θ. It is well known (cf. Example 2 below) that the posterior moments depend generically on all prior moments, i.e. their computation requires complete prior knowledge.

Therefore, the key problem of coherent approximation reads: *Which description χ, if any, allows to realize Bayes inference in a closed manner — making the diagram*

$$\begin{array}{ccc} p(\theta \mid x_{t-1}) & \xrightarrow{(3)} & p(\theta \mid x_t) \\ {\scriptstyle \chi(\cdot)}\downarrow & & \downarrow{\scriptstyle \chi(\cdot)} \\ [p(\theta \mid x_{t-1})] & \longrightarrow & [p(\theta \mid x_t)] \end{array} \tag{11}$$

commutative?

The problem is formulated exactly and solved in this section. Uncertainty of the posterior density $p(\theta \mid x_t) \equiv p(\theta; w_t)$ is described by the density $p(w_t \mid x_t)$ of the hyperparameter w_t. This density is assumed to be taken with respect to a suitable dominating product measure λ providing $\lambda(\mathrm{d}w_t) = \lambda(\mathrm{d}w_t^+)\,\lambda(\mathrm{d}w_t^-)$. We specify the evolution of $p(w_t \mid x_t)$ for the full-description case and of $p(w_t^+ \mid x_t)$ for the limited-description case. Analysing the results for the latter case, we formulate requirements on the Bayes-closed description. Finally, we present the basic result stating that, under certain assumptions, there exists a description with the needed properties.

3.2 Analysis in hierarchical Bayes setting

The following analysis has a preliminary character. It is aimed at finding conditions which the Bayes-closed description of posterior densities must satisfy.

When a full description of the posterior density is at disposal, Bayes estimation is characterized by the following lemma which provides a "meta" — formulation of the standard scheme.

Lemma 1. Let us assume that a prior density $p(w_0 \mid x_0) = p(w_0)$ is prespecified and the natural conditions of control are satisfied for both the parameter θ and hyperparameter w_{t-1}

$$p(u_t \mid x_{t-1}, \theta, w_{t-1}) = p(u_t \mid x_{t-1})\,. \tag{12}$$

Then the evolution of densities

$$p(w_{t-1} \mid x_{t-1}) \to p(w_t \mid x_t)$$

is described by the recursive functional equation

$$p(w_t \mid x_t) \propto \int p(w_t \mid x_t, w_{t-1})\, m(x_t; w_{t-1})\, p(w_{t-1} \mid x_{t-1})\, \lambda(\mathrm{d}w_{t-1})\,. \tag{13}$$

Here

$$m(x_t; w_{t-1}) = \int m(x_t; \theta)\, p(\theta; w_{t-1})\, \lambda(\mathrm{d}\theta) \tag{14}$$

and

$$p(w_t \mid x_t, w_{t-1}) = \delta(w_t - \mathrm{B}(x_t, w_{t-1})) \tag{15}$$

where $\delta(\cdot)$ denotes the Dirac function and the operator $\mathrm{B}(x_t, \cdot)$: $W \to W$ is defined by the Bayes rule (3) (the posteriors $p(\theta \mid x_t)$ are identified with the points w_t).

Proof. The density $p(w_t \mid x_t)$ can be evaluated from the joint density $p(w_t, w_{t-1} \mid x_t)$ by integration

$$p(w_t \mid x_t) = \int p(w_t, w_{t-1} \mid x_t)\, \lambda(\mathrm{d}w_{t-1})\,.$$

Using the probability "chain rule", we get

$$p(w_t \mid x_t) = \int p(w_t \mid x_t, w_{t-1})\, p(w_{t-1} \mid x_t)\, \lambda(\mathrm{d}w_{t-1})\,.$$

Owing to the meaning of the hyperparameter w_t the translation density $p(w_t \mid x_t, w_{t-1})$ reduces to (15) where $\mathrm{B}(x_t, \cdot)$ is specified under the natural conditions of control (12) by the Bayes relation (3).

The density $p(w_{t-1} \mid x_t)$ can be evaluated by conditioning from the joint density

$$p(w_{t-1} \mid x_t) \propto p(w_{t-1}, u_t, y_t \mid x_{t-1})\,.$$

By applying the "chain rule" and taking into account the natural conditions of control (12) again, we derive

$$p(w_{t-1} \mid x_t) \propto m(x_t; w_{t-1})\, p(w_{t-1} \mid x_{t-1})$$

with $m(x_t; w_{t-1}) = p(y_t \mid x_{t-1}, u_t, w_{t-1})$ for known data x_t. As $m(x_t; \theta)$ is known and the density $p(\theta \mid x_{t-1})$ is determined unambiguously by w_{t-1}, we get immediately

$$m(x_t; w_{t-1}) = \int m(x_t; \theta)\, p(\theta; w_{t-1})\, \lambda(\mathrm{d}\theta)$$

which gives the desired result. □

In the above way the standard Bayes estimation is nested into a more general framework with both the parameter θ and the parameter density $p(\theta \mid x_t)$ uncertain. Where we need below to distinguish both levels in terminology, we use the concept of density types introduced by Good [21, 22]. The adjective *type-I* is reserved for densities $p(\theta; w_t)$ of the unknown parameter θ whereas the adjective *type-II* is applied to densities $p(w_t \mid x_t)$ of the unknown hyperparameter w_t.

Notice that taking the type-II prior density (hyperprior) in the form

$$p(w_0) = \delta(w_0 - \bar{w}_0)\,,$$

we get

$$p(w_t \mid x_t) = \delta(w_t - \bar{w}_t)$$

with

$$\bar{w}_t = \mathrm{B}(x_t, \bar{w}_{t-1})\,.$$

Thus, if all uncertainty in the choice of the type-I prior $p(\theta; w_0)$ is eliminated, the type-II scheme is specialized to the standard Bayes estimation.

Example 1. The notions introduced above can be visualized using a trivial example when the parameter θ has only 3 possible values $\theta \in \{1, 2, 3\}$. Then the type-I density $p(\theta \mid x_t)$ is fully described by the triple

$$w_t = (p(\theta = 1 \mid x_t),\ p(\theta = 2 \mid x_t),\ p(\theta = 3 \mid x_t))$$

and may be identified with the appropriate point of a probability simplex. Uncertainty of this point can be described by a type-II density $p(w_t \mid x_t)$. The whole situation is illustrated in Figure 1. This example will repeatedly serve us to make explanations more transparent.

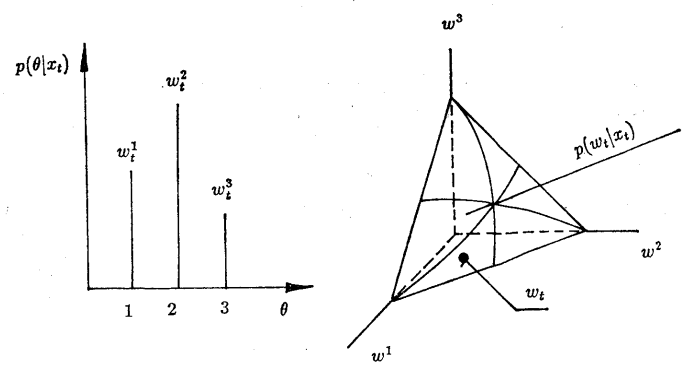

Fig. 1. An illustration of the hierarchical scheme of Bayes estimation in the case the unknown parameter θ has only 3 possible values $\theta \in \{1, 2, 3\}$. The type-II density $p(w_t \mid x_t)$ "measures" uncertainty in determining the type-I density $p(\theta \mid x_t)$.

Naturally, estimation is restricted radically when only a limited description χ is available.

Lemma 2. Let us assume a prior density $p(w_0^+ \mid x_0) = p(w_0^+)$ prespecified and the natural conditions of control (12) satisfied. Then the evolution of the type-II density

$$p(w_{t-1}^+ \mid x_{t-1}) \to p(w_t^+ \mid x_t)$$

is described by the recursive functional equation

$$p(w_t^+ \mid x_t) \propto \int\{\int[\int p(w_t \mid x_t, w_{t-1})\, \lambda(\mathrm{d}w_t^-)]\, m(x_t; w_{t-1}) \cdot$$
$$\cdot\, p(w_{t-1}^- \mid x_{t-1}, w_{t-1}^+)\, \lambda(\mathrm{d}w_{t-1}^-)\}\, p(w_{t-1}^+ \mid x_{t-1})\, \lambda(\mathrm{d}w_{t-1}^+) \qquad (16)$$

where $p(w_t \mid x_t, w_{t-1})$ and $m(x_t; w_{t-1})$ are defined by (15) and (14) respectively and w stands for the pair (w^+, w^-).

Proof. By integrating both sides of the relation (13) over the values of w_t^- (using Fubini's theorem for changing the order of integration on the right-hand side), we get

$$p(w_t^+ \mid x_t) \propto \int[\int p(w_t \mid x_t, w_{t-1})\, \lambda(\mathrm{d}w_t^-)]\, m(x_t; w_{t-1})\, p(w_{t-1} \mid x_{t-1})\, \lambda(\mathrm{d}w_{t-1}) .$$

Using the "chain rule", we factorize

$$p(w_{t-1} \mid x_{t-1}) = p(w_{t-1}^- \mid x_{t-1}, w_{t-1}^+)\, p(w_{t-1}^+ \mid x_{t-1}) .$$

Then, integrating now over the values of both w_{t-1}^+ and w_{t-1}^-, we derive

$$p(w_t^+ \mid x_t) \propto \int\{\int[\int p(w_t \mid x_t, w_{t-1})\, \lambda(\mathrm{d}w_t^-)]\, m(x_t; w_{t-1})\,.$$
$$.\, p(w_{t-1}^- \mid x_{t-1}, w_{t-1}^+)\, p(w_{t-1}^+ \mid x_{t-1})\, \lambda(\mathrm{d}w_{t-1}^-)\}\, \lambda(\mathrm{d}w_{t-1}^+)$$

However, as $p(w_{t-1}^+ \mid x_{t-1})$ does not depend on w_{t-1}^-, we can extract it from the integral over the values of w_{t-1}^- and get the relation (16). □

Notice that to evaluate the marginal density $p(w_t^+ \mid x_t)$ in (16), we need generically to know the conditional density $p(w_{t-1}^- \mid x_{t-1}, w_{t-1}^+)$, i.e. the joint density $p(w_{t-1} \mid x_{t-1})$ in fact. However, this is in contradiction with the possibility of storing only a limited description χ.

Example 2. The example with the finite parameter space illustrates the problems vividly. If we consider the density of θ described by its first moment

$$\chi(p(\theta; w_t)) = \int \theta\, p(\theta; w_t)\, \lambda(\mathrm{d}\theta)\,,$$

we simply verify that the posterior densities derived from a single equivalence class

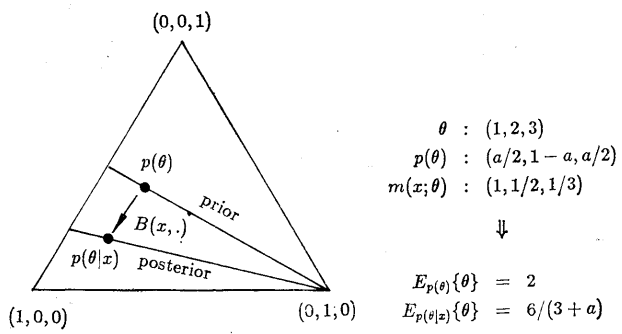

Fig. 2. The first moment equivalence class of prior densities $[p(\theta)] = \{\tilde{p}(\theta) \mid \mathsf{E}_{\tilde{p}(\theta)}\{\theta\} = \mathsf{E}_{p(\theta)}\{\theta\}\}$ for $\theta \in \{1, 2, 3\}$ is formed by a straight line. Note that after the Bayes udpate by the given $m(x; \theta)$ the posterior densities $p(\theta \mid x)$ provide different mean values of θ dependent on the prior density $p(\theta)$ (parametrized by $a > 0$). Thus, posterior densities lie in *different* equivalence classes.

$[p(\theta; w_{t-1})]$ may lie in different equivalence classes $[p(\theta; w_t)]$ (cf. Fig. 2). This inconvenient fact is known in the general setting as a *moment closure problem* [5].

3.3 Description closure requirements

It is possible to find by analysing Lemma 2 that, without having the density $p(w_{t-1}^- \mid x_{t-1}, w_{t-1}^+)$ in hand, we can perform the evolution (16) only if

1. we know a priori the equivalence class in which $p(\theta \mid x_{t-1})$ lies, and
2. the Bayes operator $\mathrm{B}(x_t, \cdot)$ maps the class $[w_{t-1}]$ on to a single equivalence class $[w_t]$ again.

In such a case the evolution (16) of the marginal density $p(w_{t-1}^{+} \mid x_{t-1})$, being a Dirac function pointing to the true point w_{t-1}^{+}, does bot depend on the conditional density $p(w_{t-1}^{-} \mid x_{t-1}, w_{t-1}^{+})$. The marginal density $p(w_t^{+} \mid x_t)$ will be a Dirac function again, pointing to the true point w_t^{+}.

Hence, the description χ is sufficient for realizing Bayes inference about respective equivalence classes if the following two requirements are satisfied.

Requirement 1 (on prior knowledge)

Let the value $\bar{w}_0$ of w_0 be such that $p(\theta; \bar{w}_0) = p(\theta)$.
Then $p(w_0^{+}) = \delta(w_0^{+} - \bar{w}_0^{+})$ must be set.

Requirement 2 (on density description)

If $p(\theta; w_{t-1}) \sim p(\theta \mid x_{t-1})$,
then $p(\theta \mid \mathrm{B}(x_t, w_{t-1})) \sim p(\theta \mid x_t)$ must hold
for any data x_t in the Bayes operator $\mathrm{B}(x_t, \cdot)$.

To sum up both requirements, estimation is to be reduced to the recursive determination of the equivalence class in which the true (type-I) posterior density lies. While to satisfy Requirement 1 is easy, Requirement 2 is not trivial and will be analysed in detail in the next section.

Remark 2. Note that Requirement 2 implies a special property of the induced equivalence. Let us interpret the family of probability densities of θ as a one-object category [23] whose morphisms are real, positive almost everywhere (for simplicity), measurable functions $m(\theta)$ defining the Bayes-rule translations

$$p(\theta) \to \frac{m(\theta)\, p(\theta)}{\int m(\theta)\, p(\theta)\, \lambda(\mathrm{d}\theta)}$$

and the composition of morphisms is defined by their multiplication. Then instead of Requirement 2 we could require the equivalence $\sim$, defined as a kernel of the description χ, to be a *congruence* on this category, i.e. to satisfy

$$\text{if} \quad m_1 \sim m_1' \quad \text{and} \quad m_2 \sim m_2', \quad \text{then} \quad m_2 m_1 \sim m_2' m_1' .$$

3.4 Construction of Bayes-closed description

The following result shows that our Bayes-coherent formulation of the approximation problem has a non-trivial solution.

Result 1. Assume that for any possible $w_t \in W$ the functions $\ln p(\theta; w_t)$ are from a Hilbert space $\mathrm{L}_2(\Theta, \mathscr{T}, r(\theta))$ with $\mathscr{T}$ standing for a σ-algebra of subsets of Θ and $r(\theta)$ representing a density of θ with respect to a dominating measure λ on $(\Theta, \mathscr{T})$.

Let the inner product be defined for any $f(\theta), g(\theta) \in L_2(\Theta, \mathscr{T}, r(\theta))$ by

$$\langle f(\theta), g(\theta) \rangle = \mathsf{E}_{r(\theta)}\{f(\theta)\, g(\theta)\} = \int f(\theta)\, g(\theta)\, r(\theta)\, \lambda(\mathrm{d}\theta) \tag{17}$$

Then there exist continuous Fréchet differentiable descriptions of densities $p(\theta; w_t)$ which are Bayes-closed in the sense of Requirement 2. Their entries are defined by

$$\chi_i(p(\theta; w_t)) = \langle h_i(\theta), \ln p(\theta \cdot w_t) \rangle\,, \quad i = 1, \ldots, n \tag{18}$$

where $h_i(\theta)$, $i = 1, \ldots, n$ are arbitrary but fixed functions from $L_2(\Theta, \mathscr{T}, r(\theta))$ satisfying

$$\langle h_i(\theta), 1 \rangle = 0\,, \quad i = 1, \ldots, n \tag{19}$$

All other Bayes-closed descriptions are related to the above ones through continuous regular transformations.

Proof. Owing to the assumptions, the problem can be formulated using logarithmic functions of densities and likelihoods and results in searching for a linear functional of a log-density. The proposition follows by application of the Riesz representation theorem [24]. For more detail see [25]. □

The assumption of Result 1 implies restrictions on possible families (4) and on the choice of the density $r(\theta)$.

Example 3. To satisfy the assumption of Result 1 in our finite parameter space example, we shall consider only probabilities from the interior of the probability simplex (another possibility of taking probabilities from the interior of some simplex

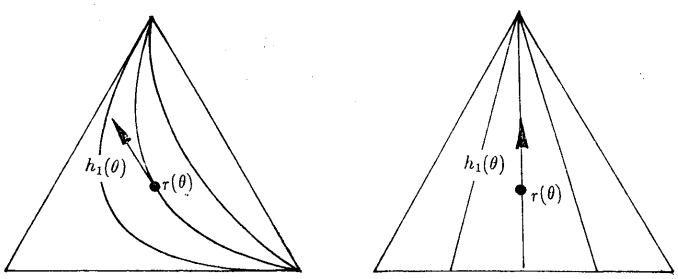

Fig. 3. An illustration of equivalence classes for $\theta \in \{1, 2, 3\}$ and various density descriptions. The vector $h_1(\theta)$ is tangential to the equivalence class at $r(\theta)$.

edge would be little illustrative). Figure 3 demonstrates the "shape" of equivalence classes for a fixed density $r(\theta)$ but various functions $h_1(\theta)$.

Obviously, by a specific choice of the density $r(\theta)$ and the functions $h_i(\theta)$, $i = 1, \ldots, n$ we influence the properties of the appropriate equivalence classes. One appealing possibility how to proceed is elaborated in [25].

4. APPROXIMATION OF POSTERIOR

Analysing recursive Bayes estimation with only a limited description of the true posterior stored, we have found a specific description which can be updated in a Bayes-closed manner. The problem of *approximation* arises in the case that not a class of densities matching the current description, but one representative of this class is required for inference or decision-making. The problem is discussed in this section.

4.1 Uncertainty of posterior density

Uncertainty in determining the posterior $p(\theta \mid x_t) \equiv p(\theta; w_t)$ is completely specified by the density

$$p(w_t \mid x_t) = p(w_t^+ \mid x_t)\, p(w_t^- \mid x_t, w_t^+)\,. \tag{20}$$

There is a fundamental difference between the factors on the right-hand side of (20). The marginal density $p(w_t^+ \mid x_t)$ specifies a probability distribution on the collection of equivalence classes. If we choose the prior $p(w_0^+)$ in accordance with Requirement 1 and use the Bayes-closed description (18) satisfying Requirement 2, we get

$$p(w_t^+ \mid x_t) = \delta(w_t^+ - \bar{w}_t^+) \tag{21}$$

pointing to the point

$$(\bar{w}_t^+, \cdot) = \mathrm{B}(x_t, (\bar{w}_{t-1}^+, \cdot))\,. \tag{22}$$

Thus, the value $\bar{w}_t^+$ of the description hyperparameter w_t^+ can be determined exactly regardless of the value of the complementary hyperparameter w_t^-.

While information about w_t^+ contained in observations can be successively accumulated as indicated above, information about w_t^- must be supplied either *a priori*, when only the description (18) is evaluated recursively, or *outside the Bayesian framework*, when the above coherent approach is combined with some non-Bayes estimation technique. In any case, the true conditional density $p(w_t^- \mid x_t, w_t^+)$ describing a probability distribution within particular equivalence classes must be substituted by its, more or less subjective, assessment $\hat{p}(w_t^- \mid x_t, w_t^+)$.

Remark 3. Through the choice of $\hat{p}(w_t^- \mid x_t, w_t^+)$ we can (and must in many cases) assign a positive probability only to densities from a subset of $\mathscr{F}$. In such a way we avoid considering too complicated parametric representations. However, the true posterior $p(\theta \mid x_t)$ need not then be a member of the selected parametric family.

4.2 Decision-theoretic framework of approximation

The consequences of a specific choice of $\hat{p}(w_t^- \mid x_t, w_t^+)$ can be evaluated in a right way only in terms of final decisions. We shall assume a classical structure of the statis-

tical decision problem [26] when our task is to select an optimal decision from the space of possible decisions, say $a \in A$. The loss caused by taking a specific decision for a specific value of the parameter θ is measured by the loss function

$$L: \Theta \times A \to \langle 0, \infty) \,. \tag{23}$$

From the Bayes point of view, optimal decision-making reduces to minimization of the expected value of the loss function

$$\inf_{a \in A} \int L(\theta, a)\, p(\theta; w_t)\, \lambda(\mathrm{d}\theta) \,. \tag{24}$$

If we do not know the density $p(\theta; w_t)$, which is our case, we have to eliminate its uncertainty using the type-II density $p(w_t \mid x_t)$:

$$\inf_{a \in A} \int \{ \int L(\theta, a)\, p(\theta; w_t)\, \lambda(\mathrm{d}\theta) \}\, p(w_t \mid x_t)\, \lambda(\mathrm{d}w_t) \tag{25}$$

By a trivial re-arrangement of integration in (25) (applying Fubini's theorem), we get an alternative form

$$\inf_{a \in A} \int L(\theta, a)\, \hat{p}(\theta \mid x_t)\, \lambda(\mathrm{d}\theta) \tag{26}$$

with an explicit posterior representative

$$\hat{p}(\theta \mid x_t) = \int \{ \int p(\theta; w_t)\, p(w_t^- \mid x_t, w_t^+)\, \lambda(\mathrm{d}w_t^-) \}\, p(w_t^+ \mid x_t)\, \lambda(\mathrm{d}w_t^+) \tag{27}$$

where we used the factorized form (20) of $p(w_t \mid x_t)$. The density (27) represents the optimal result of inference with respect to the Bayes decision principle.

When $p(w_0 \mid x_0) = \delta(w_0 - \bar{w}_0)$ and a full description of $p(\theta \mid x_t)$ is saved, $\hat{p}(\theta \mid x_t)$ coincides with $p(\theta \mid x_t)$. When a limited but Bayes-closed description is stored, the marginal density $p(w_t^+ \mid x_t)$ reduces to (21) whereas the conditional density $p(w_t^- \mid x_t, w_t^+)$ must be substituted by an "estimate" $\hat{p}(w_t^- \mid x_t, w_t^+)$. In this case $\hat{p}(\theta \mid x_t)$ is an approximation of the true posterior $p(\theta \mid x_t)$.

Example 4. Consider a specific equivalence class $[p(\theta; w_t)]$ for a fixed value

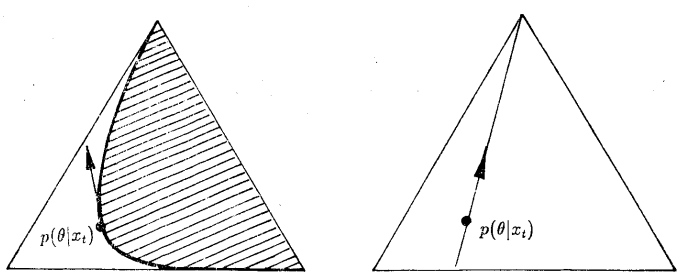

Fig. 4. Convex hulls of equivalence classes induced for $\theta \in \{1, 2, 3\}$ by various density descriptions. These sets are composed of possible Bayes estimates of the true density. When the equivalence class is closed with respect to the "mixture" operation — being a straight line, the set of possible alternatives is radically restricted.

$\bar{w}_t^+$ of the description hyperparameter w_t^+. The set of all possible approximations $\hat{p}(\theta \mid x_t)$ (for all $\hat{p}(w_t^- \mid x_t, \bar{w}_t^+)$) forms a convex hull of the equivalence class. Figure 4 illustrates a dramatic change in these sets for slightly different descriptions. Notice a remarkable fact that when the equivalence class is "closed" with respect to (27) (being a straight line in our case), the convex hull coincides with the equivalence class itself.

5. COMPLETE APPROXIMATION SCHEME

Summarizing the above results, we get a complete framework for recursive Bayes estimation using only a limited finite-dimensional but *Bayes-closed* description of type-I densities. In order that the effect of the limited density description may become more transparent, both procedures – ideal and approximated – are recalled in parallel using the type-II terms.

Full Description	*Limited Description*
(1) choice of prior density:	
$p(w_0) = \delta(w_0 - \bar{w}_0)$	$p(w_0^+) = \delta(w_0^+ - \bar{w}_0^+)$
(2) Bayes evolution:	
$p(w_t \mid x_t) = \delta(w_t - \bar{w}_t)$	$p(w_t^+ \mid x_t) = \delta(w_t^+ - \bar{w}_t^+)$
where	where
$\bar{w}_t = \mathrm{B}(x_t, \bar{w}_{t-1})$	$(\bar{w}_t^+, \cdot) = \mathrm{B}(x_t, (\bar{w}_{t-1}^+, \cdot))$
(3) Bayes-optimal parameter density:	
$p(\theta \mid x_t) = p(\theta; \bar{w}_t)$	$\hat{p}(\theta \mid x_t) = \int p(\theta; \bar{w}_t^+, w_t^-) \cdot \hat{p}(w_t^- \mid x_t, \bar{w}_t^+)\, \lambda(\mathrm{d}w_t^-)$

Note the structure of the scheme for a limited description. While the first two steps describe an exact evolution of the description hyperparameter w_t^+, the third step represents *approximation* itself. Thus, two features are strictly separated in the above scheme:

1. Using the Bayes-closed description of posterior densities, we are able to accomplish the ideal Bayes inference on corresponding equivalence classes.
2. With only a limited description at disposal, we are not able to accumulate information about densities within equivalence classes in a Bayes-coherent way. All information about them, supplied through $\hat{p}(w_t^- \mid x_t, w_t^+)$, must be specified either a priori or outside the Bayes scheme.

Commonly in literature posterior descriptions different from the Bayes-closed ones are used. Then, of course, we lose the above insight and control over estimation.

The first two steps of the suggested scheme are made explicit by the next result.

Result 2. The Bayes inference about the equivalence classes $[p(\theta \mid x_t)]$ induced

by the description (18) is fully specified by the evaluation of the prior description

$$\chi_i(p(\theta)) = \mathsf{E}_{r(\theta)}\{h_i(\theta) \ln p(\theta)\}, \quad i = 1, \ldots, n \tag{28}$$

and by the recursive updating of the posterior description

$$\chi_i(p(\theta \mid x_t)) = \chi_i(p(\theta \mid x_{t-1})) + \mathsf{E}_{r(\theta)}\{h_i(\theta) \ln m(x_t; \theta)\}, \quad i = 1, \ldots, n \tag{29}$$

with the expectation $\mathsf{E}_{r(\theta)}\{\cdot\}$ defined in Result 1.

Proof. The above equations follow directly from the description (18) and the Bayes rule (3).

Remark 4. The fact that the entries $\chi_i(\cdot)$ are updated independently of each other makes *parallel processing* possible. The overall time needed to compute a new description vector can be substantially reduced in such a way.

6. CONCLUDING REMARKS

Recursive Bayes estimation has been studied in the special case when only a limited description of the posterior density has been assumed stored. The approach adopted to solve the problem is characterized by several novel features:

- the density of an unknown parameter is interpreted as *uncertain*:
- all densities with the same description, forming an *equivalence class*, are considered together as the basic unit for inference;
- the approximation results are required to be *coherent* (not contradictory) with the ideal Bayes inference;
- the posterior description is required to be *Bayes-closed*, i.e. computable in a recursive manner without complete knowledge of the true posterior density.

In spite of a lot of open questions, the approach used and first results found in this part are believed to give a new insight into the challenging problem of approximating recursive estimation. Although the problem has been studied within the Bayesian framework, many results may easily be adapted to maximum likelihood techniques.

Two promising ways of continuing the line sketched in the paper offer. First, the equivalence approach seems to be applicable immediately in the case of filtering too (problems known as parameter tracking, state estimation, etc.) if the joint densities of all uncertain quantities are characterized by a finite-dimensional description.

Second, the equivalence approach turns out to be only a special case of a more general, say "*invariance approach*" viewing the approximation of Bayes estimation as building invariants with respect to the Bayes rule translation. Other invariants in addition to the density description derived in this part are probability distributions which dominate or are dominated by the posterior distribution. Such distributions

can be updated recursively in a suitable parametric form ([17] illustrates this possibility). The important question in this respect is whether other types of invariants can be constructed (the results of Chentsov [23] may be useful here).

(Received August 11, 1989.)

REFERENCES

[1] V. Peterka: Bayesian approach to system identification. In: Trends and Progress in System Identification (P. Eykhoff, ed.), Pergamon Press, Oxford 1981.

[2] A. H. Jazwinski: Stochastic Processes and Filtering Theory. Academic Press, New York 1970.

[3] H. W. Sorenson and A. R. Stubberud: Non-linear filtering by approximation of the a posteriori density. Internat. J. Control *8* (1968), 33—51.

[4] K. Srinivasan: State estimation by orthogonal expansion of probability distributions. IEEE Trans. Automat. Control *AC-15* (1970), 3—10.

[5] R. S. Bucy and K. D. Senne: Digital synthesis of non-linear filters. Automatica *7* (1971), 287—298.

[6] H. W. Sorenson and D. L. Alspach: Recursive Bayesian estimation using Gaussian sums. Automatica *7* (1971), 465—479.

[7] D. L. Alspach: Gaussian sum approximations in nonlinear filtering and control. In: Estimation Theory (D. G. Lainiotis, ed.), American Elsevier, New York 1974.

[8] J. L. Center: Practical nonlinear filtering of discrete observations by generalized least-squares approximation of the conditional probability distribution. In: Proc. of 2nd Symp. on Nonlinear Estimation, San Diego 1971.

[9] R. J. P. de Figueiredo and J. G. Jan: Spline filters. In: Proc. of 2nd Symp. on Nonlinear Estimation, San Diego 1971.

[10] D. G. Lainiotis and J. G. Deshpande: Parameter estimation using splines. In: Estimation Theory (D. G. Lainiotis, ed.), American Elsevier, New York 1974.

[11] H. W. Sorenson: On the development of practical nonlinear filters. In: Estimation Theory (D. G. Lainiotis, ed.), American Elsevier, New York 1974.

[12] A. H. Wang and R. L. Klein: Implementation of nonlinear estimators using monospline. In: Proc. of 13th IEEE Conf. on Decision and Control, 1976.

[13] D. V. Lindley: Approximate Bayesian methods. In: Bayesian Statistics (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), University Press, Valencia 1980.

[14] O. L. R. Jacobs: Recursive estimation for non-linear Wiener systems by on-line implementation of Bayes' rule. Trans. Inst. M. C. *7* (1985), 245—250.

[15] M. Kárný and K. M. Hangos: Approximation of the Bayes rule. In: Proc. of 7th IFAC/IFORS Symp. on Identification and System Parameter Estimation, York 1985.

[16] A. R. Stubberud and G. H. Xia: A fixed complexity nonlinear estimation technique. In: Proc. of 25th Conf. on Decision and Control, Athens 1986.

[17] M. Kárný and K. M. Hangos: One-sided approximation of Bayes rule: theoretical background. In: Proc. of 10th IFAC Congress, Munich 1987.

[18] S. C. Kramer and H. W. Sorenson: Bayesian parameter estimation. In: Proc. of 1987 Amer. Control. Conf., Minneapolis 1987.

[19] J. M. Bernardo: Approximations in statistics from a decision theoretical viewpoint. In: Probability and Bayesian Statistics (R. Viertl, ed.), Plenum Press, New York 1987.

[20] B. de Finetti: Theory of Probability: A Critical Introductory Treatment. Wiley, New York 1970 (Vol. 1), Chichester 1972 (Vol. 2).

[21] I. J. Good: The Estimation of Probabilities: An Essay on Modern Bayesian Methods. MIT Press, Cambridge 1965.

[22] I. J. Good: Some history of the hierarchical Bayesian methodology. In: Bayesian Statistics (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), University Press, Valencia 1980.
[23] N. N. Chentsov: Statistical Decision Rules and Optimal Inference (in Russian). Nauka, Moscow 1972. English translation: Translation of Mathematical Monographs, *53*, AMS, Rhode Island 1982.
[24] R. Larsen: Functional Analysis: An Introduction. Dekker, New York 1973.
[25] R. Kulhavý: A Bayes-closed approximation of recursive nonlinear estimation. Internat. J. Adaptive Control and Signal Processing (submitted).
[26] L. J. Savage: The Foundations of Statistics. Wiley, New York 1954.

Ing. Rudolf Kulhavý, CSc., Ústav teorie informace a automatizace ČSAV (Institute of Information — Czechoslovak Academy of Sciences), Pod vodárenskou věží 4, 182 08 Praha 8. Czechoslovakia.