Can We Preserve the Structure of Recursive Bayesian Estimation in a Limited-Dimensional Implementation?

Rudolf Kulhavý

Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
P.O. Box 18, 182 08 Prague
kulhavy@utia.cas.cz

1. Introduction

The Bayesian scheme of parameter estimation is a well-developed paradigm of mathematical statistics (DE FINETTI 1990, SAVAGE 1972) which essentially propagates the probability distribution of unknown quantitities conditional on observed data.

To make our treatment maximally transparent and free of technical subtleties, we shall consider the case of independent and identically distributed data $y_1, \ldots, y_k \in Y$ with a common probability function $m(y;\theta)$ parametrized by unknown parameters $\theta \in \Theta$ where both the data space Y and the parameter space Θ are finite.

Applying elementary rules of probability theory, we derive that the conditional probability function $p(\theta \,|\, y_1, \ldots, y_k)$ of θ given observed data $y_1, \ldots, y_k$ gets the form

$$p(\theta \,|\, y_1, \ldots, y_k) = c_{\rm n}\, p(\theta) \prod_{\kappa=1}^{k} m(y_\kappa;\theta)$$

where $c_{\rm n}$ is the normalizing factor.

With a simpler notation $p_k(\theta) = p(\theta \,|\, y_1, \ldots, y_k)$, $p_0(\theta) = p(\theta)$, the *recursive* formula for propagation of the posterior probability function reads

$$p_k(\theta) = c_{\rm n}\, p_{k-1}(\theta)\, m(y_k;\theta)\,. \tag{1}$$

The recursion (1) is conceptually very simple but its practical implementation is feasible only in special cases when its dimensionality can be reduced enough, namely in the case when the probability functions $m(y;\theta)$ admit the form

$$m(y;\theta) = c_{\rm n} \exp \sum_{i=1}^{n} f^i(\theta)\, g_i(y)\,. \tag{2}$$

In a generic case, we face extreme requirements on computational memory and time.

The problem thus reads how to approximate — in a theoretically well-grounded way — the optimal but unfeasible propagation of the conditional probability within a given memory and time.

The present paper illustrates the role of differential geometry in solving this challenging problem. Differential geometry has already proved to be an effective tool in statistical inference, in analysis and design of nonlinear control systems as well as in nonlinear filtering. Here we indicate that differential geometry is capable to address also the key issues of approximate Bayesian estimation — reduction of dimensionality and approximation of the theoretical inference scheme.

2. Reduction of Dimensionality

Practical implementation of recursive Bayesian estimation means primarily to find a limited-dimensional alternative of the theoretical scheme. Tens of different ways of reducing dimensionality of the problem have been reported in the literature. This has raised the natural question of existence of an "optimal" solution. We believe that if such a solution exists, it must preserve maximum of the "mathematical structure" of the theoretically optimal solution. The present Section analyses what the structure of Bayesian estimation is and what part of it can be preserved throughout approximation.

2.1. Probabilities and Translations

We start by describing the elementary geometry of a probability space determined by a special subgroup of its transformations related to (1). Our presentation follows essentially Chentsov (1972, §10); see also Kulhavý (1992a). Differential-geometric concepts are introduced rather briefly, the interested reader may find more background in textbooks like Boothby (1986) or Isham (1989).

Manifold of probabilities of unknown parameters. We consider a space P of mutually absolutely continuous probabilities on a finite measurable space $(\Theta, \mathcal{T})$ where Θ is a finite set with, say, $N+1$ elements and $\mathcal{T}$ is a finite algebra with atoms given by one-point sets $\{\theta\}$, $\theta \in \Theta$. Any probability measure is uniquely specified by the probability function

$$p := p(\theta) = \Pr(\{\theta\}) .$$

Because $p(\theta) > 0$ for any $\theta \in \Theta$ and $\sum_{\theta\in\Theta} p(\theta) = 1$, P is homeomorphic to an open simplex in $\mathbb{R}^N$ and probabilities of any N elements of Θ (as well as their diffeomorphic transformations) can serve as coordinates of p. Thus, P forms an N-dimensional *differentiable manifold.*

Note that from the statistical viewpoint the probability manifold P is both an *exponential* and *mixture* family of dimension N. Indeed, P is closed under the convex operations $c_{\mathrm{n}}\,(p)^{\alpha}\,(p')^{1-\alpha}$ and $\beta\,p + (1-\beta)\,p'$ where $\alpha,\ \beta \in [0,1]$ and c_{n} is the normalizing factor.

Group of unnormalized likelihoods. The set of all transformations of the manifold P forms a group. Our interest is in a subgroup of transformations with the structure of Eq. (1).

Let L be a space of unnormalized likelihoods

$$l := \{\, c\,l(\theta) \mid c > 0 \,\}$$

where $l(\theta) > 0$ for any $\theta \in \Theta$. L is a group with the group operation $l\,l' := \{\, c\,l(\theta)\,l'(\theta) \mid c > 0 \,\}$, the unit element $e := \{\, c\,1(\theta) \mid c > 0 \,\}$ and the inverse element $l^{-1} := \{\, c/l(\theta) \mid c > 0 \,\}$ where $l(\theta)$ and $l'(\theta)$ are arbitrary representatives of l and l', respectively. Note that the group operation is commutative, $l\,l' = l'l$, thus the group is Abelian. L can be regarded as a quotient group of the group of positive real-valued functions with the group operation of pointwise multiplication modulo its subgroup of constant positive functions $c\,1(\theta)$. L is homeomorphic to a set of rays in a positive cone in $\mathbb{R}^{N+1}$ (each ray corresponds to an equivalence class $l \in L$) and forms an N-dimensional differentiable manifold. Summing up the above properties with the fact that both the operation $(l, l') \mapsto l\,l'$ and $l \mapsto l^{-1}$ are smooth, we can conclude that L forms a *Lie group*.

Group action. We attach now to any $l \in L$ a specific transformation $\lambda_l : P \to P$ defined through $p \mapsto \bar{p} = \lambda_l(p) = p\,l$

$$\bar{p}(\theta) = \frac{p(\theta)\,l(\theta)}{\sum_{\theta\in\Theta} p(\theta)\,l(\theta)}$$

where $l(\theta)$ is any representative of l. Following the terminology of Chentsov (1972), we can call the above transformation *translation*. In fact, in a logarithmic coordinate system, we have $\log p'(\theta) = \log p(\theta) + \log l(\theta) + (\log c)1(\theta)$.

The map $l \mapsto \lambda_l$ is a homomorphism from L into the group of diffeomorphisms of P, i.e. it holds $p\,e = p$ for every $p \in P$ and $(p\,l)\,l' = p\,(l\,l')$ for every $p \in P$ and $l,\ l' \in L$. The map $(p, l) \mapsto p\,l$ from the product manifold $P \times L$ into the manifold P is clearly smooth. Thus, *the Lie group L acts smoothly on the differentiable manifold P* (see Fig. 1).

Note that the L-action is free and transitive, i.e. any pair of points in P is connected by just one element of L. The appropriate translation is essentially given by the Radon-Nikodým derivative of the corresponding probability measures — see Lemma 10.2 and its Corollary in Chentsov (1972).

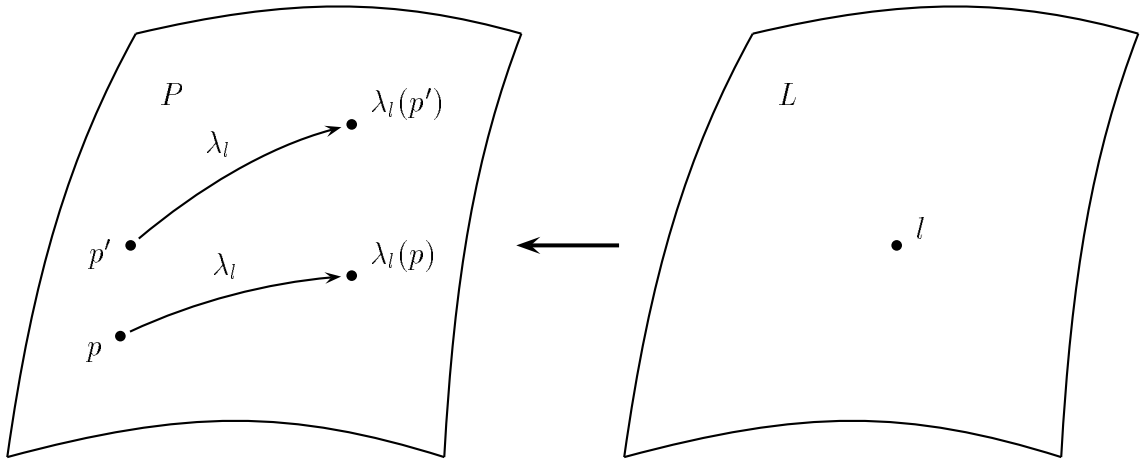

Figure 1: The action of the Lie group L of unnormalized likelihoods l on the differentiable manifold P of probabilities p of unknown parameters.

2.2. Translation-Closed Equivalence

To identify uniquely an arbitrary point p of the probability manifold P, we need to save a complete N-dimensional vector of its coordinates in a used coordinate chart. This may be in conflict with a limited computational memory available to us. Then we are forced to use a reduced description which makes the true point p uncertain and ambiguous.

We describe a natural way of reducing dimensionality which preserves the structure of the underlying probability manifold P under the L-action. Specifically, we look for a partition of the probability manifold P (i.e. an equivalence relation on P) that would be closed under the action of the Lie group L.

Principal fibre bundle. Let $\tilde{L} \subset L$ be a closed Lie subgroup of unnormalized likelihoods. Clearly, any subgroup $\tilde{L}$ of L acts smoothly again on the differentiable manifold P. The $\tilde{L}$-action is free (for any pair of points p, $p' \in P$ there exists at most one element $\tilde{l} \in \tilde{L}$ such that $p' = p\,\tilde{l}\,q$) but not transitive (there exist pairs of points p, $p' \in P$ that cannot be "connected" by any element of $\tilde{L}$).

The orbit $[\,p\,]$ of the $\tilde{L}$-action through $p \in P$ is the set of all points in P that can be reached from p

$$[\,p\,] := \{\, p' \in P \mid p' = p\,\tilde{l},\ \tilde{l} \in \tilde{L}\,\}\,. \tag{3}$$

This induces an equivalence relation on P: $p' \sim p$ if $p' \in [\,p\,]$. The probability manifold P is partitioned into a system of equivalence classes which is called the orbit space of the $\tilde{L}$-action on P and denoted $P/\tilde{L}$. It can be shown that $P/\tilde{L}$ is a differentiable manifold again (a quotient manifold of P relative to the equivalence relation $\sim$) and

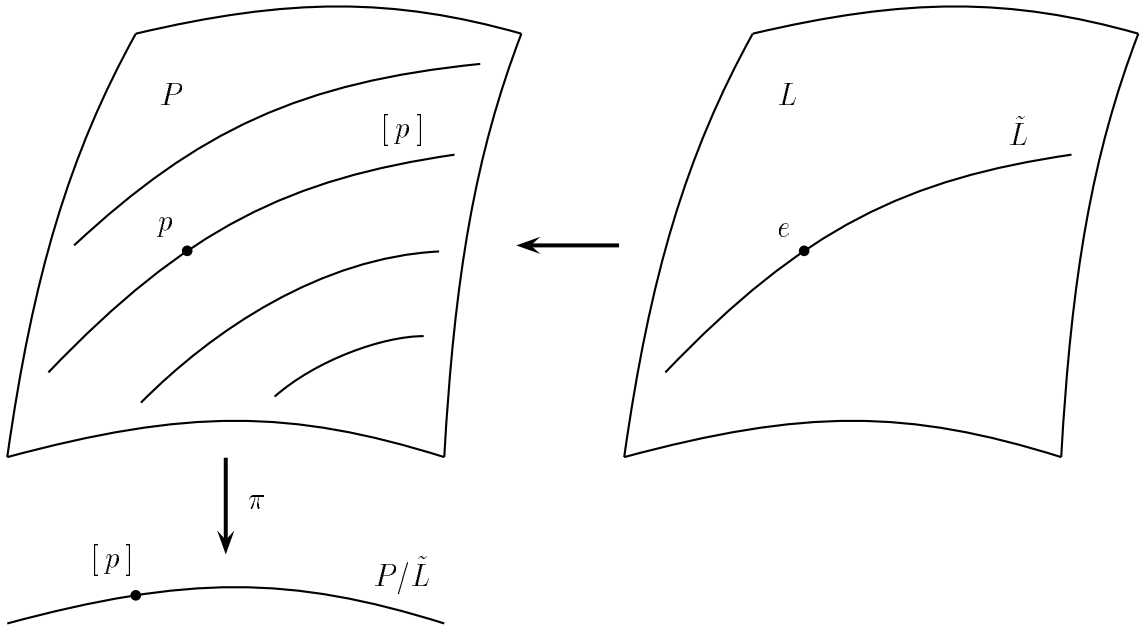

Figure 2: The principal fibre bundle produced by the action of a Lie subgroup $\tilde{L} \subset L$.

that the canonical projection map $\pi : P \to P/\tilde{L}$ defined through $p \mapsto [p]$ is smooth.

The inverse image $\pi^{-1}(x)$, $x \in P/\tilde{L}$ is called a fibre over x. In our case, the fibres of the bundle coincide with the orbits of the $\tilde{L}$-action on P. The freedom of the $\tilde{L}$-action on P implies that each orbit (fibre) is homeomorphic to $\tilde{L}$. Therefore, the action of $\tilde{L}$ on P produces a *principal fibre bundle*

$$(P, \pi, P/\tilde{L})$$

where P is a total space, $P/\tilde{L}$ is a base space and $\tilde{L}$ is a structure group of the bundle (see Fig. 2).

Closure under the group action. We verify that the orbit space $P/\tilde{L}$ is closed indeed under the L-action so that for any $p \in P$ and $l \in L$ it holds

$$[p]\, l = [p\, l]\,. \tag{4}$$

Denoting $[p] = p\,\tilde{L}$, we can rewrite the above relation as $p\,\tilde{L}\, l = p\, l\,\tilde{L}$. Because of the freedom of the $\tilde{L}$-action on P, the last relation is valid if and only if the left and right cosets of $\tilde{L}$ coincide: $l\,\tilde{L} = \tilde{L}\, l$. It means that to meet our requirement, $\tilde{L}$ need to be a *normal* subgroup of L. In our case, L is Abelian and so any subgroup $\tilde{L} \subset L$ is normal.

Subgroups of unnormalized likelihoods. Subgroups of L have a rather simple structure. Let V be an N-dimensional vector space composed of classes of real-valued

functions differing by at most an additive constant

$$v := \{\, v(\theta) + a1(\theta) \,|\, a \in \mathbb{R} \,\}\,.$$

V can be regarded as a quotient vector space of the real vector space of dimension $N+1$ modulo its subspace of constant functions $a1(\theta)$. V is homeomorphic to a set of parallel straight lines in $\mathbb{R}^{N+1}$.

Choosing any N linearly independent elements $v_1, \ldots, v_N \in V$ as a basis of V, we can express a generic element $l \in L$ in the form

$$l = \exp \sum_{j=1}^{N} \alpha^j v_j$$

where $\alpha^j \in \mathbb{R}$, $j = 1, \ldots, N$. A Lie subgroup $\tilde{L} \subset L$ of dimension $N - n > 0$ is then composed of elements

$$\tilde{l} = \exp \sum_{j=1}^{N-n} \alpha^j \tilde{v}_j$$

where $\alpha^j \in \mathbb{R}$, $j = 1, \ldots, N-n$ and $\tilde{v}_1, \ldots, \tilde{v}_{N-n}$ are linearly independent elements of an $(N-n)$-dimensional vector subspace $\tilde{V} \subset V$.

The fibre of the bundle $(P, \pi, P/\tilde{L})$ over $p \in P$ is thus composed of the points

$$p_\alpha = p \exp \sum_{j=1}^{N-n} \alpha^j \tilde{v}_j\,. \tag{5}$$

Each fibre $[\,p\,]$ is an $(N-n)$-dimensional differential submanifold of P and, from the statistical viewpoint, an *exponential* subfamily of P.

2.3. Orthogonal Projections

The concept of the principal fibre manifold provides an elegant general tool for dimensional reduction. The choice of a particular subgroup $\tilde{L}$ means simply to decide on probability translations $\lambda_{\tilde{l}}$ which will be indistinguishable in a reduced dimension.

A direct specification of $\tilde{L}$ by listing basis vectors $\tilde{v}_1, \ldots, \tilde{v}_{N-n}$ is, however, intractable in realistic cases when N is large. We have to look for a more "concrete" representation of the base space $P/\tilde{L}$ and the projection map π.

Construction of isomorphic bundles. Let us consider a bundle (P, π^*, Q) where Q is a differentiable submanifold of P. The bundle is isomorphic to $(P, \pi, P/\tilde{L})$ if there exist maps $f_1 : P/\tilde{L} \to Q$ and $f_2 : Q \to P/\tilde{L}$ such that $f_1 \circ f_2 = \mathrm{id}_Q$, $f_2 \circ f_1 = \mathrm{id}_{P/\tilde{L}}$ and $\pi^* = f_1 \circ \pi$. In fact, it is easy to find such maps. Let f_2 be the canonical projection map π restricted to Q, $\pi_{|Q}$. Let f_1 be a smooth map $\sigma : P/\tilde{L} \to P$ such that $\pi \circ \sigma = \mathrm{id}_{P/\tilde{L}}$, i.e.

let σ be a smooth *cross-section* of the bundle $(P, \pi, P/\tilde{L})$. Finally, let $Q = \sigma(P/\tilde{L})$ be the image of $P/\tilde{L}$ in the cross-section σ. Then the following diagram clearly commutes in both directions:

$$\begin{array}{ccc} & P & \\ \pi \swarrow & & \searrow \pi^* = \sigma \circ \pi \\ P/\tilde{L} & \underset{\pi_{|Q}}{\overset{\sigma}{\rightleftarrows}} & Q \end{array}$$

In other words, to construct a bundle (P, π^*, Q) isomorphic to the principal fibre bundle $(P, \pi, P/\tilde{L})$, we can choose an arbitrary n-dimensional differentiable submanifold Q of P intersecting each fibre $[\,p\,] \in P/\tilde{L}$ at just one point and then to define the projection map π^* as an assignment $p \mapsto q^p = [\,p\,] \cap Q$.

Submanifold orthogonal to fibres. We show one construction which has a particular geometric flavour. Namely, we construct a submanifold Q of P which is orthogonal (in a sense which will be made explicit in a moment) to all fibres of the bundle $(P, \pi, P/\tilde{L})$.

Let Q be a set of points

$$q_\beta = \sum_{i=1}^{n} \beta^i q_i + \Big(1 - \sum_{i=1}^{n} \beta^i\Big) q_0 \tag{6}$$

where q_0, q_1, ..., q_n are n fixed, linearly independent points of P and the coefficients $\beta^1, \ldots, \beta^n$ are any real scalars such that $q_\beta \in P$. It is easy to show that Q forms an n-dimensional differentiable submanifold of P and $(\beta^1, \ldots, \beta^n)$ represent specific coordinates of the points $q \in Q$. From the statistical viewpoint, Q is an n-dimensional *mixture* subfamily of P.

With regard to (5), the fibre $[\,q_\beta\,]$ of the bundle $(P, \pi, P/\tilde{L})$ over q_β is composed of the points

$$p_{\alpha,\beta} = q_\beta \, \exp \sum_{j=1}^{N-n} \alpha^j \tilde{v}_j \, . \tag{7}$$

We say that the vectors $\partial/\partial\alpha^j$ and $\partial/\partial\beta^i$ tangent to the submanifolds $[\,q_\beta\,]$ and Q, respectively, are mutually orthogonal at $p_{\alpha=0,\beta} = q_\beta$ if their inner products are zero

$$\left\langle \frac{\partial}{\partial\alpha^j}, \frac{\partial}{\partial\beta^i} \right\rangle = \sum_{\theta\in\Theta} \left(\frac{\partial}{\partial\alpha^j} \log p_{\alpha,\beta}(\theta) \right)_{\alpha=0} \left(\frac{\partial}{\partial\beta^i} \log p_{\alpha,\beta}(\theta) \right)_{\alpha=0} p_{\alpha=0,\beta}(\theta) = 0 \tag{8}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, N-n$. The inner-product definition (8) is strongly motivated by the fact that it is the only definition invariant under the change of parametrization θ as well as the coordinates α and β (CHENTSOV 1972).

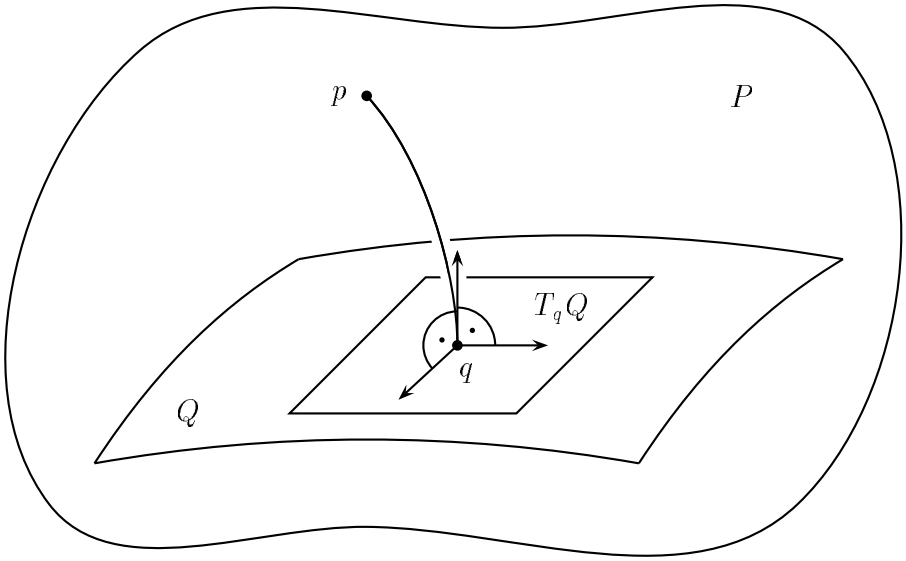

Figure 3: The orthogonal projection of a point $p \in P$ onto a submanifold $Q \subset P$ (T_qQ is the tangent space of Q at q).

After substituting from (7) into (8), we get a system of conditions

$$\sum_{\theta\in\Theta} [q_i(\theta) - q_0(\theta)]\, \tilde{v}_j(\theta) = 0 \tag{9}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, N-n$ where $\tilde{v}_j(\theta)$ stands for any representative of $\tilde{v}_j \in \tilde{V}$. Clearly, for a given basis $\tilde{v}_1, \ldots, \tilde{v}_{N-n} \in \tilde{V}$, it is always possible to find points $q_0, q_1, \ldots, q_n$ so to fulfil the conditions.

Note that (9) is equivalent to the identity

$$\sum_{\theta\in\Theta} [q(\theta) - q'(\theta)]\, [\log p(\theta) - \log p'(\theta)] = 0 \tag{10}$$

valid for every pair q, $q' \in Q$ and every pair $p \sim p'$. We say that $q^p \in Q$ is an orthogonal projection of $p \in P$ along the fibre $[p]$ if it holds

$$\sum_{\theta\in\Theta} [q_i(\theta) - q_0(\theta)]\, [\log p(\theta) - \log q^p(\theta)] = 0$$

for $i = 1, \ldots, n$ (see Fig. 3).

The fact that Q intersects each fibre at just one point can be proved in more ways, e.g. using a sort of Pythagorean theorem valid for probabilities (Chentsov 1972, Csiszár 1975, Amari 1985).

We can now turn the above construction around and choose first an n-dimensional submanifold $Q \subset P$ and then define an equivalence relation on P through the orthogonal projection $\pi^* : p \mapsto q^p$. More precisely, two points p and p' are said to be equivalent, $p \sim p'$, if the orthogonal projections of them onto Q coincide, $q^p = q^{p'}$. In this way we get a principal fibre bundle (P, π^*, Q) with the same effect of dimensional reduction.

2.4. Repetitive Structure of Estimation

Only some of the translations $\lambda_l : P \to P$ may actually appear as a result of the conditioning (1) for a given model. Specifically, it is the translations produced by the likelihoods in the form

$$l_k = \exp k v_k$$

where $k = 1, 2, \ldots$ and $v_k \in V$ is represented by

$$v_k(\theta) = \frac{1}{k} \sum_{\kappa=1}^{k} \log m(y_\kappa; \theta) \,.$$

The form of v_k reflects the natural repetitive structure of estimation (Lauritzen 1988). With the increasing time k, the variation of v_k gets typically smaller and this motivates us to pay special attention to subgroups of L composed of the points

$$l_t = \exp t v^r$$

with $t \in \mathbb{R}$ and $v^r \in V$ represented by

$$v^r(\theta) = \sum_{y \in Y} r(y) \, \log m(y; \theta) \tag{11}$$

where $r(y)$ is a probability function of y. It is easy to see that for the empirical probability function

$$r_k(y) = \frac{1}{k} \sum_{\kappa=1}^{k} \delta(y - y_\kappa) \,, \tag{12}$$

where $\delta(x) = 1$ for $x = 0$ and 1 elsewhere, we have $v^{r_k} = v_k$. It shows a rather close relationship between the empirical probability function $r_k \in R$ and the posterior probability function $p_k \in P$ which we analyse and employ further.

Manifold of probabilities of observed data. Quite analogously as in Section 2.1, we can introduce a space R of mutually absolutely continuous probabilities on a finite measurable space $(Y, \mathcal{Y})$ where Y is a finite set of, say, $M + 1$ elements and $\mathcal{Y}$

is a finite algebra with atoms given by one-point sets $\{y\}$, $y \in Y$. Any probability measure is uniquely specified by the probability function

$$r := r(y) = \Pr(\{y\}) .$$

P is homeomorphic to an open simplex in $\mathbb{R}^M$ and probabilities of any M elements of Y can serve as coordinates of r. R forms an M-dimensional differentiable manifold.

Obviously, R comprises model probability functions $m(y;\theta)$ for every $\theta \in \Theta$. To include even empirical probability functions $r_k(y)$ for all data sequences $y_1, y_2, \ldots, y_k$ and every k, we have to consider the boundary of the simplex R along with the inside.

One-parameter subgroups of likelihoods. For any element $v \in V$ the map $t \mapsto \exp tv$ is a smooth homomorphism from the additive group of the real line $\mathbb{R}$ into L

$$\exp(t_1 + t_2)v = \exp(t_1 v)\, \exp(t_2 v)$$

and thus the set $\{\exp tv \,|\, t \in \mathbb{R}\}$ forms a *one-parameter Lie subgroup* of L. It is a classical result of differential geometry that every one-parameter subgroup of the Lie group L is an integral curve of an invariant vector field X^A on L generated by a tangent vector $A \in T_e L$ to L at the unit element $e \in L$ (note that because L is Abelian, we do not need to distinguish between the left- and right-invariant vector fields). Another fundamental result of differential geometry states an isomorphism of the tangent space $T_e L$ and the Lie algebra $\mathcal{L}(L)$ of all invariant vector fields on L, regarded as a real vector space.

As in our case the tangent space $T_e(L)$ of L at the unit element $e \in L$ can be identified in a natural way with the real vector space V, we get an important correspondence

$$v \in V \;\leftrightarrow\; A \in T_e L \;\leftrightarrow\; X \in \mathcal{L}(L)$$

among vectors (equivalence classes of real-valued functions) $v := \{\, v(\theta) + a1(\theta) \,|\, a \in \mathbb{R}\,\}$, tangent vectors A to L at e and invariant vector fields X on L.

One-parameter subgroup orbits. Let us assume that $v \neq \{\, a1(\theta) \,|\, a \in \mathbb{R}\,\}$. Then the orbit of the action of a one-parameter subgroup $t \mapsto \exp tv$ through $p \in P$ is the curve

$$t \mapsto \phi_t^v(p) = p \exp tv .$$

Clearly, when $v = v^{r_k}$ where r_k is the empirical probability function (12) at time k, the orbit goes through the posterior probability function at time k, $\phi_k^v(p_0) = p_k$.

We can sum up all the connections as follows (cf. Fig. 4):

$$\begin{aligned}
r \in R &\;\mapsto\; v^r \in V \;\; (A^{v^r} \in T_e L,\; X^{A^{v^r}} \in \mathcal{L}(L)) \\
t \in \mathbb{R} &\;\mapsto\; \exp tv^r \in L \\
p \in P &\;\mapsto\; p \exp tv^r \in P
\end{aligned}$$

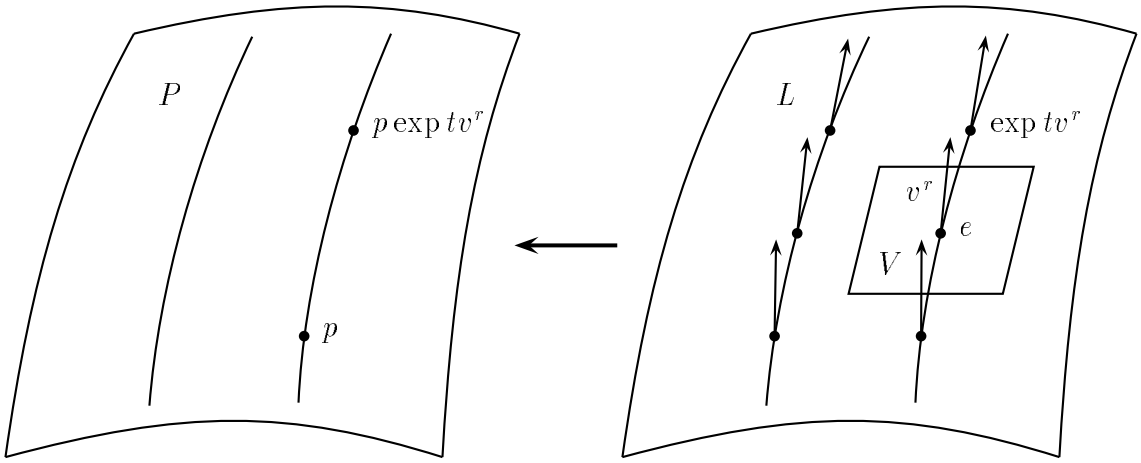

Figure 4: The action of a one-parameter subgroup $\exp tv^r \subset L$ on P. The tangent space T_eL of L at e is identified with the vector space V.

Note that it is the dimension of the subspace

$$V^m = \text{span}\,\{\, v^r \in V \,|\, r \in R \,\} \subset V$$

that essentially decides on the dimensionality of Bayesian estimation. In the special case when the probability functions $m(y;\theta)$ admit the form (2), V^m is spanned by the vectors $f_i := \{\, f_i(\theta) + a1(\theta) \,|\, a \in \mathbb{R} \,\}$ for $i = 1, \ldots, n$.

2.5. Induced Equivalence

In a limited dimension, the action of a one-parameter subgroup passes naturally to the equivalence classes $[\,p_t\,] \in P/\tilde{L}$. Specifically, with the definition $\tilde{\phi}_t^v([\,p\,]) = [\,p\,] \exp tv$, the following diagram commutes

$$\begin{array}{ccc} P & \xrightarrow{\phi_t^v} & P \\ \pi \downarrow & & \downarrow \pi \\ P/\tilde{L} & \xrightarrow{\tilde{\phi}_t^v} & P/\tilde{L} \end{array}$$

owing to the closure property (4).

Equivalent tangent vectors. The action of $\tilde{L}$ on L itself induces an equivalence relation on the vector space V (and correspondingly on the tangent space T_eL and the Lie algebra $\mathcal{L}(L)$): $v \sim v'$ if $v - v' = \tilde{v} \in \tilde{V}$, i.e., owing to (9), if it holds

$$\sum_{\theta \in \Theta} [q_i(\theta) - q_0(\theta)]\,[v(\theta) - v'(\theta)] = 0 \tag{13}$$

for $i = 1, \ldots, n$ where $v(\theta)$ and $v'(\theta)$ are arbitrary representatives of v and v', respectively.

The action produces a vector bundle $(V, \nu, V/\tilde{V})$ where each fibre $[\,v\,] \in V/\tilde{V}$ possesses the structure of an $(N-n)$-dimensional real vector space and $\nu : V \to V/\tilde{V}$ is the canonical projection map. Analogous vector bundles are induced on the tangent space $T_e L$ and the Lie algebra of invariant vector fields $\mathcal{L}(L)$ (regarded as real vector space).

Equivalent probabilities of observed data. The equivalence on V is "pulled back" through the map $r \in R \mapsto v^r \in V$ (11) on the manifold R. Two vectors v^r and $v^{r'} \in V$ are equivalent if they meet the conditions (13)

$$\sum_{\theta \in \Theta} [q_i(\theta) - q_0(\theta)]\,[v^r(\theta) - v^{r'}(\theta)] = 0\,.$$

After substituting for $v^r(\theta)$ and $v^{r'}(\theta)$ from (11), we get first

$$\sum_{\theta \in \Theta} [q_i(\theta) - q_0(\theta)] \sum_{y \in Y} [r(y) - r'(y)] \log m(y;\theta) = 0$$

and after changing the order of summation

$$\sum_{y \in Y} [r(y) - r'(y)] \sum_{\theta \in \Theta} [q_i(\theta) - q_0(\theta)] \log m(y;\theta) = 0\,.$$

Introducing the functions of observed data

$$\xi_i(y) = \sum_{\theta \in \Theta} [q_i(\theta) - q_0(\theta)] \log m(y;\theta)\,, \tag{14}$$

we get the conditions

$$\sum_{y \in Y} r(y)\,\xi_i(y) = \sum_{y \in Y} r'(y)\,\xi_i(y) \tag{15}$$

which clearly define a partition of R.

We say that two probability functions $r(y)$ and $r'(y)$ are equivalent, $r \sim r'$, if they fulfil the conditions (15) for $i = 1, \ldots, n$. The above derivation implies that if $r \sim r'$, then $v^r \sim v^{r'}$. The equivalence relation $\sim$ on R produces a bundle $(R, \rho, R/\!\sim)$ where the quotient manifold $R/\!\sim$ is composed of the equivalence classes

$$[\,r\,] := \{\, r' \in R \,|\, r' \sim r \,\}$$

and $\rho : R \to R/\!\sim$ is the canonical projection map. Note that from the statistical viewpoint the fibres (equivalence classes) of the bundle form *mixture* subfamilies of R. Their dimension is equal to $M - n$ provided the functions $\xi_1(y), \ldots, \xi_n(y)$ are linearly independent.

Morphism between bundles. There exists a *morphism* between the bundles $(R, \rho, R/\!\sim)$ and $(V, \nu, V/\tilde{V})$, i.e. a pair of maps $f_1 : R \to V$ and $f_2 : R/\!\sim \to V/\tilde{V}$ such that the following diagram commutes

$$\begin{array}{ccc} R & \xrightarrow{f_1} & V \\ \rho \downarrow & & \downarrow \nu \\ R/\!\sim & \xrightarrow{f_2} & V/\tilde{V} \end{array} .$$

Specifically, the map f_1 is given by the assignment $r \mapsto v^r$ (11) and f_2 is well defined through the assignment $[\, r \,] \mapsto [\, v^{[r]} \,]$ because the derivation of (15) implies

$$v^{[r]} \subset [\, v^r \,] .$$

Therefore, the pair of maps (f_1, f_2) is fibre-preserving. Note that the recognition of the repetitive structure of estimation in (11) has brought a new (typically quite substantial) piece of information about the true vector v^r.

2.6. Orthogonal projections II

Analogously as in Section 2.3 we construct a differentiable submanifold of the manifold R orthogonal to all fibres of the bundle $(R, \rho, R/\!\sim)$.

Submanifold orthogonal to fibres. Let S be a set of points

$$s_\alpha = s_0 \exp \sum_{i=1}^{n} \alpha^i \log \frac{s_i}{s_0} \tag{16}$$

where $s_i \in S$, $i = 0, 1, \ldots, n$ are related to $q_i \in Q$ as follows

$$s_i(y) = c_{\mathbf{n}} \exp \sum_{\theta \in \Theta} q_i(\theta) \log m(y; \theta) .$$

Note that from the statistical viewpoint S is an exponential subfamily of R generated by the functions $\xi_i(y)$ (14), $i = 1, \ldots, n$

$$s_\alpha(y) = c_{\mathbf{n}}\, s_0(y) \exp \sum_{i=1}^{n} \alpha^i \xi_i(y) .$$

The dimension of S is equal to n provided the points s_0, s_1, ..., s_n are in a "general position", i.e. the functions $\xi_1(y)$, ..., $\xi_n(y)$ are linearly independent.

For a fixed point $s_\alpha \in S$, the fibre $[\, s_\alpha \,]$ of the bundle $(R, \rho, R/\!\sim)$ over s_α is composed of the points

$$r_{\alpha,\beta} = \sum_{j=1}^{M-n} \beta^j r_j + \Big(1 - \sum_{j=1}^{M-n} \beta^j\Big) s_\alpha \tag{17}$$

where s_α, r_1, ..., r_{M-n} are $M-n+1$ linearly independent points of $[\, s_\alpha \,]$ and β^1, ..., β^{M-n} are any real scalars such that $r_{\alpha,\beta} \in R$.

We say that the vectors $\partial/\partial\alpha^i$ and $\partial/\partial\beta^j$ tangent to the submanifolds S and $[\, s_\alpha \,]$, respectively, are mutually orthogonal at $r_{\alpha,\beta=0} = s_\alpha$ if their inner products are zero

$$\left\langle \frac{\partial}{\partial\alpha^i}, \frac{\partial}{\partial\beta^j} \right\rangle = \sum_{y\in Y} \left(\frac{\partial}{\partial\alpha^i} \log r_{\alpha,\beta}(y) \right)_{\beta=0} \left(\frac{\partial}{\partial\beta^j} \log r_{\alpha,\beta}(y) \right)_{\beta=0} r_{\alpha,\beta=0}(y) = 0 \qquad (18)$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, M-n$. The inner-product definition (18) is essentially identical to (8) and is motivated again by its invariance properties.

After substituting from (17) and (16) into (18) we get the conditions

$$\sum_{y\in Y} [r_j(y) - s_\alpha(y)]\, \xi_i(y) = 0$$

for $i = 1, \ldots, n$ which are identical to the identity

$$\sum_{y\in Y} [r(y) - r'(y)]\, [\log s(y) - \log s'(y)] = 0 \qquad (19)$$

valid for every pair s, $s' \in S$ and every pair $r \sim r'$. The conditions are certainly met owing to (15). We say that $s^r \in S$ is an orthogonal projection of the point $r \in R$ along the fibre $[\, r \,] \in R/\!\sim$ onto S if

$$\sum_{y\in Y} [r(y) - s^r(y)]\, \xi_i(y) = 0$$

for $i = 1, \ldots, n$. It can be shown that the bundle (R, ρ^*, S) where $\rho^* : R \to S$ is the orthogonal projection map $r \mapsto s^r$ is isomorphic to $(R, \rho, R/\!\sim)$.

Dual geometry. The bundles $(P, \pi, P/\tilde{L})$ and $(R, \rho, R/\!\sim)$ have significant dual features (see Fig. 5).

- The fibres $[\, p \,] \in P/\tilde{L}$ and $[\, r \,] \in R/\!\sim$ form exponential and mixture subfamilies of P and R, respectively.
- The submanifolds $Q \subset P$ and $S \subset R$ orthogonal to all fibres of the bundles $(P, \pi, P/\tilde{L})$ and $(R, \rho, R/\!\sim)$ form mixture and exponential subfamilies of P and R, respectively.
- The maps $q_i \mapsto s_i$ and $r_k \mapsto p_k$ have essentially the same exponential structure.

This fact is not accidental, it is just manifestation of the existence of dual connections (namely, exponential and mixture ones) on probability manifolds — see Dawid (1975), Amari (1985), cf. Kulhavý (1990).

Information loss. Let us denote the set of all model probability functions of observed data as $M = \{\, m(y;\theta) \,|\, \theta \in \Theta \,\} \subset R$. We introduce a space of probability

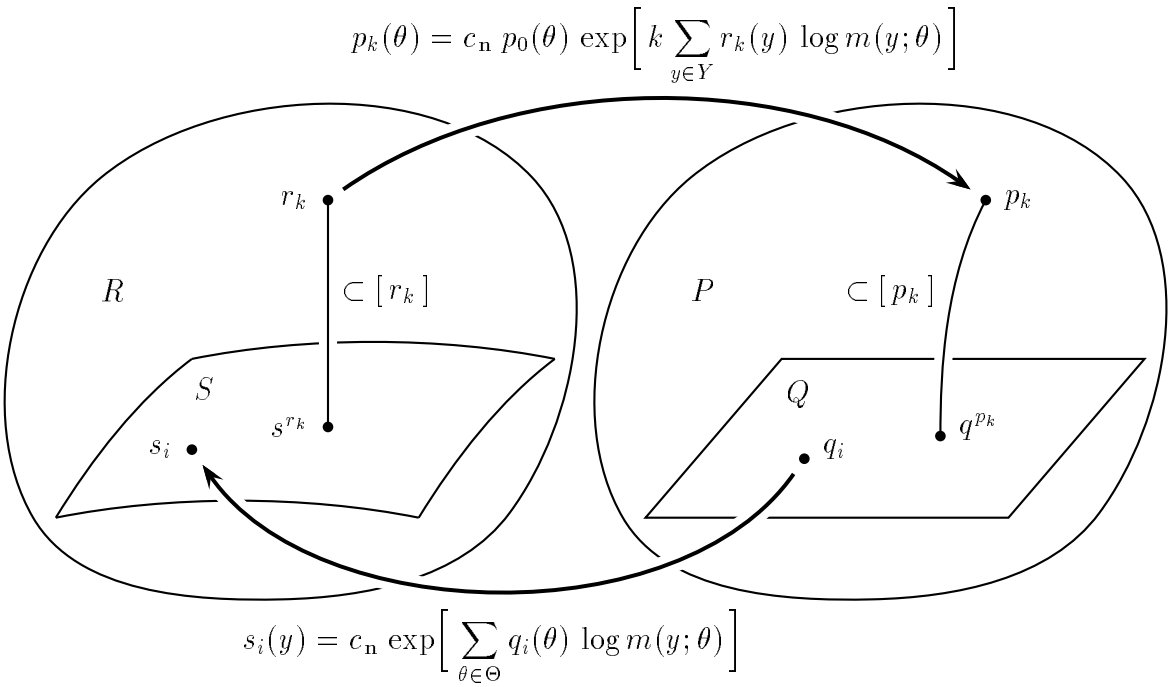

Figure 5: The dual features of geometry of empirical and posterior probabilities (for the sake of an easier comparison, mixture and exponential subfamilies are envisaged as straight and curved objects, respectively).

functions of observed data $M^* \subset R$ as the minimal smooth submanifold of R which envelops M and is closed under the convex operation $(p,p') \mapsto c_{\rm n}\, p^{\alpha}\, p'^{1-\alpha}$, $\alpha \in \mathbb{R}$. Thus M^* is the minimal exponential subfamily of R containing $m(y;\theta)$ for every $\theta \in \Theta$.

Note that S is necessarily a smooth submanifold (exponential subfamily) of M^*. An important special case comes when $S = M^*$. Then $m(y;\theta) \in S$ for every $\theta \in \Theta$ and the relation (19) implies

$$\sum_{y\in Y} [r(y) - r'(y)]\, \log m(y;\theta) = a1(\theta),\ a \in \mathbb{R}\,.$$

which combined with (11) gives the identity $v^r = v^{r'}$ valid for every pair of equivalent points $r \sim r'$.

Thus, if $S = M^*$, the ambiguity of the empirical probability function r_k due to the equivalence on R does not result in any loss of information necessary for evaluating the true posterior probability function $p_k = p_0\, \exp kv^{r_k}$ (Kulhavý 1993).

On the contrary, when the dimension of M^* is so big that we have to specify S as a proper submanifold of M^*, knowledge of the equivalence class $[\, r_k\,]$ is not sufficient

to determine the true posterior p_k (KULHAVÝ 1993). All we know about p_k is that it lies within the set

$$\{\, p_0 \, \exp k v^{\tilde{r}} \,|\, \tilde{r} \in [\, r_k \,] \,\} \subset [\, p_k \,] \,.$$

3. Extensions to More General Models

The differential-geometric view of Bayesian estimation presented above for the case of independent and identically distributed data is general enough to be applied to more complex model situations.

Infinite parameter space. When the parameter space Θ is a subset of Euclidean space, the probability manifold P is infinitely-dimensional. When Y is still finite, we can avoid working in infinite dimension by restricting to the space $\bar{P} \subset P$ of all possible posterior densities which is finite-dimensional in this case; see KULHAVÝ (1992a) for more details.

Infinite data space. When even the data space Y is a subset of Euclidean space, the above approach is not effective because the space of all possible posterior densities is generically infinitely-dimensional. Moreover, the manifold R of all densities of Y is infinitely-dimensional. This makes building of differential geometry on the manifolds a more delicate matter, e.g. all fibres in P and R become infinitely-dimensional submanifolds. The basic features of the above geometric picture of the problem seem, however, preserved.

Dependent data. When observed data $y_1, \ldots, y_k$ are dependent, modelled by a regression-type model $m(y_k \,|\, z_{k-1}; \theta)$ where z_{k-1} is a known function of data up to time $k-1$, the conditioning operation (1) modifies as follows

$$p_k(\theta) = c_{\mathrm{n}} \, p_{k-1}(\theta) \, m(y_k \,|\, z_{k-1}; \theta) \,. \tag{20}$$

Let us stress that the multiplicative structure of (1) and (20) is essentially identical. The repetitive structure of estimation is, however, more complex in the present case. We have to introduce the manifold R as a space of densities $r(y,z)$ of (y,z) and define the empirical density at time k as

$$r_k(y,z) = \frac{1}{k} \sum_{\kappa=1}^{k} \delta((y,z) - (y_\kappa, z_{\kappa-1}))$$

where $\delta(\cdot)$ denotes Dirac function. The posterior density p_k is then related to the empirical density r_k through the formula

$$p_k(\theta) = c_{\mathrm{n}} \; p_0(\theta) \; \exp\Big[\, k \int r_k(y,z) \; \log m(y \,|\, z; \theta) \, \mathrm{d}y \, \mathrm{d}z \,\Big] \,.$$

Note that the model densities $m(y \mid z;\theta)$ are not elements of the manifold R but they can be "lifted" by an arbitrary density $w(z)$ of z so that $m(y \mid z;\theta)\, w(z) \in R$. With this setting, the differential geometry of Bayesian estimation of dynamic models comes very close to what we have described in this paper; cf. KULHAVÝ (1992b).

4. Which Approximation?

We have shown that much of the structure of recursive Bayesian estimation can be preserved even in a limited dimension. This fact is important in its own right, but it should also contribute to the quality of a subsequent approximation of Bayesian inference.

We must admit, however, that there is still much ambiguity on what the "right" approximation scheme is and even, which requirements such a scheme should meet. We thus finish by a critical survey of several approaches suggested by the above "structural insight".

Orthogonal projection of the posterior probability function. The simplest idea of approximation is to utilize the existence of a submanifold $Q \subset P$ orthogonal to all fibres of the bundle $(P, \pi, P/\tilde{L})$ and construct an approximate posterior probability function $\hat{p}_k(\theta)$ through the orthogonal projection of the true posterior p_k along the fibre $[\, p_k \,]$ onto Q

$$\hat{p}_k = q^{p_k} \in [\, p_k \,]\,. \tag{21}$$

Such a solution has attractive properties both locally (due to the orthogonal projection) and globally (it minimizes Kullback-Leibler distance between p_k and q_k); see KULHAVÝ (1990), KULHAVÝ (1992a) for details. The negative feature is its poor asymptotic behaviour. We have seen that with k increasing, the posterior probability functions p_k get typically closer and closer to the orbit of a one-parameter subgroup action $p_t = p_0 \exp tv$ for some $v \in V$. The projections of p_k onto a mixture subfamily Q of P do not possess this property. In fact, no *fixed* mixture family can catch a typical successive concentration of the posterior probability on a single point of P.

Orthogonal projection of the empirical probability function. The drawback of the solution (21) can be avoided if we recognize and respect the repetitive structure of estimation, specifically, if we look for an approximate posterior probability function $\hat{p}_k$ only within the set of probability functions that can be reached with different distributions of data $\tilde{r} \in [\, r_k \,]$

$$\hat{p}_k \in p_0 \exp\, kv^{[\, r_k \,]} \subset [\, p_k \,]\,. \tag{22}$$

The set is typically much smaller that $[\, p_k \,]$. In such a way we eliminate from possible

candidates on $\hat{p}_k$ those probability functions which cannot be produced by the recursion (1).

With the above restriction the approximate solution $\hat{p}_k$ can be regarded as a posterior probability function

$$\hat{p}_k = p_0 \exp kv^{\hat{r}_k} \tag{23}$$

for some approximation $\hat{r}_k \in [\, r_k \,]$ of the true empirical probability function r_k. An intuitively attractive construction of $\hat{r}_k$ is through the orthogonal projection of r_k along the fibre $[\, r_k \,]$ onto S

$$\hat{r}_k = s^{r_k} \in [\, r_k \,]\,. \tag{24}$$

Note that in this case we attempt to approximate the orbit of a one-parameter subgroup action $p_t = p_0 \exp tv^r$ (for the current distribution r of observed data) rather than isolated posterior points p_k. This should result in better asymptotic properties of the approximation. Our (limited) simulational experience has confirmed the expectation: we have achieved a very good agreement of the true and approximate posteriors, improving with $k \to \infty$.

The orthogonal projection $\hat{r}_k$ can be justified in more ways. The point $\hat{r}_k$ achieves the minimum of Kullback-Leibler distance between r_k and S, the minimum of dual Kullback-Leibler distance between an arbitrary fixed point $s \in S$ and $[\, r_k \,]$, the maximum of (relative) entropy along $[\, r_k \,]$; see KULHAVÝ (1992b), KULHAVÝ (1993) for more details.

Propagation of posterior uncertainty. The solution (23)–(24) aims at a reasonable "point estimate" of the true posterior probability function. Uncertainty of p_k is simply neglected. A more appropriate Bayesian solution may, however, be considered.

First note that knowledge of the equivalence class $[\, p_k \,]$ or $[\, r_k \,]$ is equivalent to knowledge of the value of a specific data statistic suggested by (15)

$$\bar{\xi}_{i,k} = \sum_{y \in Y} r_k(y)\, \xi_i(y),\ \ i = 1, \ldots, n\,.$$

This statistic is generically *non-sufficient*. It can be shown (KULHAVÝ 1992b) that the probability function of θ conditional on the reduced data statistic (rather than all data) is a *prior* expectation of the unknown posterior probability function within $[\, p_k \,]$. This fact suggests that the uncertainty of θ should depend on the loss of information in the projection map $\pi : P \to P/\tilde{L}$.

If we want to mimic the feature, it appears reasonable to leave the constraint (22) and construct an approximate probability function as

$$\hat{p}_k = p_0 \exp k\hat{v}_k \tag{25}$$

where the vector $\hat{v}_k \in V$ is related in some way to the set $\{ v^{\tilde{r}} \in V \,|\, \tilde{r} \in [\, r_k \,] \}$. For instance, in KULHAVÝ (1993) the function $\hat{v}_k(\theta)$ has been defined as Kullback-Leibler distance between the equivalence class $[\, r_k \,]$ and particular model points $m_\theta \in M$.

Even from the above brief remarks, it is clear that approximation of recursive Bayesian estimation in a limited dimension presents a number of non-trivial conceptual questions which still wait for a deeper theoretical analysis. The same is true about numerical implementation of resulting approximate estimators which we have not discussed here at all.

5. References

AMARI, S. (1985). *Differential-Geometrical Methods in Statistics.* Springer-Verlag, New York.

BOOTHBY, W.M. (1986). *An Introduction to Differentiable Manifolds and Riemannian Geometry*, 2nd ed. Academic Press, London.

CHENTSOV, N.N. (1972). *Statistical Decision Rules and Optimal Inference* (in Russian). Nauka, Moscow. English translation (1982), Amer. Math. Soc., Providence, RI.

CSISZÁR, I. (1975). I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.*, **3**, 146–158.

DAWID, A.P. (1975). Discussion to the paper by B. Efron: Defining the curvature of a statistical problem (with application to second order efficiency). *Ann. Statist.*, **3**, 1231–1234.

DE FINETTI, B. (1990). *Theory of Probability*, Vol. 1 and 2, Wiley Classics Library Edition. Wiley, Chichester.

ISHAM, C.J. (1989). *Modern Differential Geometry for Physicists.* World Scientific, Singapore.

KULHAVÝ, R. (1990). Recursive nonlinear estimation: a geometric approach. *Automatica*, **26**, 545–555.

KULHAVÝ, R. (1992a). Recursive nonlinear estimation: geometry of a space of posterior densities. *Automatica*, **28**, 313–323.

KULHAVÝ, R. (1992b). On design of approximate finite-dimensional estimators: the Bayesian view. In *Preprints of the IFAC Workshop on Mutual Impact of Computing Power and Control Theory*, Prague, Czechoslovakia. pp. 13–24.

KULHAVÝ, R. (1993). Can approximate Bayesian estimation be consistent with the ideal solution? In *Preprints of the 12th IFAC World Congress*, Sydney, Australia. Vol. 4, pp. 225–228.

LAURITZEN, S.L. (1988). *Extremal Families and Systems of Sufficient Statistics.* Springer-Verlag, Berlin.

SAVAGE, L.J. (1972). *The Foundations of Statistics*, 2nd edition. Dover Publ., New York.