Approximate Bayesian recursive estimation

Miroslav Kárny*

Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 4, 182 08 Prague, Czech Republic

ABSTRACT

Bayesian learning provides a firm theoretical basis of the design and exploitation of algorithms in data-streams processing (preprocessing, change detection, hypothesis testing, clustering, etc.). Primarily, it relies on a recursive parameter estimation of a firmly bounded complexity. As a rule, it has to approximate the exact posterior probability density (pd), which comprises unreduced information about the estimated parameter. In the recursive treatment of the data stream, the latest approximate pd is usually updated using the treated parametric model and the newest data and then approximated. The fact that approximation errors may accumulate over time course is mostly neglected in the estimator design and, at most, checked ex post. The paper inspects the estimator design with respect to the error accumulation and concludes that a sort of forgetting (pd flattening) is an indispensable part of a reliable approximate recursive estimation. The conclusion results from a Bayesian problem formulation complemented by the minimum Kullback–Leibler divergence principle. Claims of the paper are supported by a straightforward analysis, by elaboration of the proposed estimator to widely applicable parametric models and illustrated numerically.

1. Introduction

Data-streams processing [2,19] faces many challenges connected with data preprocessing, change detection, hypothesis testing, clustering, prediction, etc. These classical statistical topics [12] are instances of dynamic decision making under uncertainty and incomplete knowledge well-covered by Bayesian paradigm [7]. Its routine use is inhibited by the fact that the available formal solutions neglect the inherent need for the recursive (sequential) treatment. The paper counteracts this neglect with respect to parameter estimation, which forms the core of solutions of the mentioned problems.

The recursive estimation is rarely feasible without an information loss. Mostly, each data updating of estimates only approximates the lossless estimation [9]. Without a care, approximation errors may accumulate to the extent damaging the estimation quality. Stochastic approximations [5] dominate the analysis inspecting whether a specific estimator suffers from this problem or not. The design of estimators avoiding the accumulation is less developed and mostly relies on stochastic stability theory [28] limited by a non-trivial choice of an appropriate Lyapunov function.

Both the analysis and design predominantly focus on a point estimation. However, the recursive estimation serving to dynamic decision making is to provide a fuller information about the estimated parameter. The Bayesian estimation provides its most complete expression, namely, the posterior probability density of the unknown parameter (pd, Radon–Nikodým derivative with respect to a dominating measure, denoted d*. [33]). This explains the focus of the paper on the Bayesian estimation.

* Tel.: +420 266052274.
E-mail address: school@utia.cas.cz

http://dx.doi.org/10.1016/j.ins.2014.01.048
0020-0255/© 2014 Elsevier Inc. All rights reserved.
The inspection of the approximation-errors influence has been neglected within the Bayesian framework. Papers [20–24] represent a significant exception. They characterise the Bayesian approximate recursive estimation without an approximation-errors accumulation. They show that the accumulation is completely avoided if and only if a finite collection of fixed linear functionals acting on logarithm of the posterior pds are used as a (non-sufficient) statistic. The values of this statistic can be recursively updated by data and serve as information-bearing constraints for the design of the approximate posterior pds. This favourable class of statistics is, however, too narrow and excludes too many cases of practical interest. Thus, it is desirable to inspect an approximate recursive Bayesian estimation allowing non-zero errors caused by the recursive treatment while counteracting their accumulation. The paper proposes such an estimator. The proposed solution respects that the recursively stored information about the exact posterior pd (quantifying fully the available information) is inevitably partial. Then, the minimum Kullback–Leibler divergence (KLD) [27] principle [17,35] is to be used for its completion. Under general conditions, the completion adds forgetting to a common “naive” approximate recursive estimation, which takes the approximate posterior pd as an exact prior pd for the data updating.

The paper primarily aims to attract the research attention to the problem practically faced by any approximate recursive learning. This determines the relatively abstract presentation way. The excellent anonymous reviewers have served as an encouraging sample of readers who confirmed the presentation efficiency. The suppression of multitude features and technical details of an overall data-streams handling has allowed them to grasp well the essence of the addressed problem and of its solution. The focus on the problem core also determines the level of proofs’ details. The paper is not fully self-containing in this respect and relies on availability of the complementary information in referred papers. Technically, Section 2 formulates the addressed problem. Section 3 provides its solution and indicates that the accumulation of approximation errors is counteracted. It also guides how to choose the decisive data-dependent forgetting factor. Section 4 specialises the solution to an important class of parametric models and the corresponding feasible approximate posterior pds. An example illustrating general results is in Section 5. Section 6 contains closing remarks.

2. Addressed problem

A parametric model \( m_t = m_t(\Theta) \) describes a (modelled) output \( y_t \in \mathbb{Y}^+ \) stimulated by an (external) input \( u_t \in \mathbb{U}^+ \) at discrete-time moments labelled by \( t \in \mathbb{T} = \{1, 2, \ldots\} \). Data records \( d_t = (y_t, u_t) \) are processed sequentially. The parametric model \( m_t \) is a pd of the output \( y_t \), conditioned on the prior information, on the current input \( u_t \), on the past data records \( d_{t-1}, \ldots, d_1 \), and on an unknown parameter \( \Theta \in \Theta^+ \). The parameter is also unknown to the input generator. It means that \( u_t \) and \( \Theta \) are independent when conditioned on \( d_{t-1}, \ldots, d_1, i.e. \) natural conditions of control [32] are met.

Full information about the parameter \( \Theta \) at time \( t-1 \) is expressed by the exact posterior pd \( f_{t-1} = f_{t-1}(\Theta) = \{\Theta(u_t, d_{t-1}, \ldots, d_1) \} \) (quantifying fully the available information). The Bayes rule \( B_t \) updates this pd by the data record \( d_t \). The exact posterior pd evolves as follows

\[
\begin{align*}
 f_t &= B_t[f_{t-1}] \iff f_t(\Theta) = \frac{m_t(\Theta)f_{t-1}(\Theta)}{g_t(y_t)} \propto m_t(\Theta)f_{t-1}(\Theta), \quad \forall \Theta \in \Theta^+,
 g_t(y_t) &= \int_{\Theta^+} m_t(\Theta)f_{t-1}(\Theta) \, d\Theta,
\end{align*}
\]

where \( \propto \) denotes equality up to normalisation. The predictive pd \( g_t(y_t) \) is determined by (2) with the fixed condition \( u_t, d_{t-1}, \ldots, d_1 \) and an arbitrary output \( y \in \mathbb{Y}^+ \). The parametric model in (1) is treated as likelihood, i.e. as a function of \( \Theta \) for a fixed inserted data \( d_t, d_{t-1}, \ldots, d_1 \). The recursion (1) is initiated by a designer-supplied prior pd \( f_0 = f_0(\Theta) \) describing the available prior information. The updating (1) requires knowledge of the pd \( f_{t-1} \) and information that allows the evaluation of the likelihood \( m_t(\Theta) \). \( \forall \Theta \in \Theta^+ \). A \( K \)-dimensional statistic \( \psi_t \) (called regression vector, \( K < \infty \)), which can be updated recursively, is assumed to comprise such an information.

The inspected problem arises when the exact posterior pd \( f_t = f_t(\Theta) \) is too complex and has to be replaced by an approximate pd \( p_t = p_t(\Theta) \). The pd \( p_t \) is a projection of \( f_t \) on a designer-selected set of feasible pds \( \mathbb{P}^+ \). In [8], it was shown that the pd \( \text{arg} \min_{p_t \in \mathbb{P}^+} \text{optimally} \) the exact posterior pd \( f_t \) is to minimise the KLD \( D(f_t||p) \) [27].

\[
\begin{align*}
 0_{\mathbb{P}^+} \in \arg \min_{p_t \in \mathbb{P}^+} D(f_t||p) &= \arg \min_{p_t \in \mathbb{P}^+} \int_{\Theta^+} f_t(\Theta) \ln \left( \frac{f_t(\Theta)}{p(\Theta)} \right) \, d\Theta.
\end{align*}
\]

Since a direct use of (3) with the exact pd \( f_t \) evolving according to (1) is prevented by the problem definition, the recursive evaluation without an additional error should evolve the optimal pd \( 0_{\mathbb{P}^+}(3) \), i.e. to update recursively the optimal approximation \( 0_{\mathbb{P}^+} \) of the exact posterior pd \( f_{t-1} \)

\[
(0_{\mathbb{P}^+}; m_t) \rightarrow 0_{\mathbb{P}^+}.
\]
The papers [20–24] mentioned in Introduction have shown that such a construction is possible if and only if the set \( p^* \) is delimited by values of a finite collection of linear, time and data invariant functionals \( \mathcal{F}_k(\ln(f_t)) \) fulfilling \( \mathcal{F}_k(1) = 0, k = 1, \ldots, K < \infty \). The exclusion of errors attributed to the recursive treatment excludes commonly used statistics, for instance, (i) the mean and covariance values in unscented approximation [15]; (ii) the likelihood values on a variable, say Monte Carlo generated, grid [11]; (iii) statistics determining finite Gaussian mixtures with a fixed number of components [3]; (iv) statistics yielded by variational Bayes [37]; etc. In other words, the non-commutativity of the Bayes rule (1) and the projection (3) to such classes of \( p^* \) makes the optimal recursive approximation (4) impossible. Thus, only a non-optimal approximate pd \( p_t \in p^* \) can be evolved instead of \( p_t. \) The ideal recursion (4) is to be replaced by the map recursively updating the pd \( p_{t-1} \) non-optimally approximating the exact posterior pd \( f_{t-1} \)

\[
(p_{t-1}, m_t) \rightarrow p_t.
\]

Such a feasible recursion is mostly constructed in the next naive way, \( t \in t^* \).

\[
\begin{align*}
\text{Evaluate} & \quad \tilde{f}_t = B_t[p_{t-1}] \\
\text{Find} & \quad \hat{p}_t \in \arg\min_{p_t \in p^{*}} D(\tilde{f}_t||p) \\
\text{Set} & \quad p_t = \hat{p}_t.
\end{align*}
\]

Often, other proximity measures than the KLD are employed but the use of an approximate \( p_{t-1} \) in the role of the exact pd \( f_{t-1} \) is the common flaw.

In summary, a question arises how to construct the map (5) respecting \( p_{t-1} \not= f_{t-1} \), or, in other words, how to modify the recursive approximate estimator (6) so that the approximation-errors accumulation is counteracted.

The next auxiliary proposition shows that a difference in prior pds has a tendency to diminish during the data updating.

**Proposition 1** (Contractive Nature of the Bayes Rule). Let \( \forall t \in t^* \), \( f_t \) be the exact posterior pds corresponding to the exact prior pd \( f_0 \) and \( p_t \) be the posterior pds initiated by another prior pd \( p_0 \) such that \( D_0 = D(f_0||p_0) < \infty \).

Then, the KLD values \( D_t = D(f_t||p_t) \) form super-martingale with respect to \( \sigma \)-algebras generated the processed data \( u_{t+1}, d_{t+1}, \ldots, d_t \). Consequently, \( D_t \) converges almost surely to a finite value \( D_{\infty} \in [0, D_0] \) for \( t \rightarrow \infty \).

**Proof.** Defining the predictive pds \( g_t(y) = \int \sigma \* m_{t}(\theta)f_{t-1}(\theta)d\theta \) and \( h_t(y) = \int \sigma \* m_{t}(\theta)p_{t-1}(\theta)d\theta \), the key super-martingale inequality is verified via the following evaluations of the conditional expectation

\[
E[D_t|u_t, d_{t-1}, \ldots, d_t] = \int j \sigma \* \left[ \int \sigma \* f_t(\theta) \ln \left( \frac{f_t(\theta)}{p_t(\theta)} \right) d\theta \right] dy
\]

\[
= \int j \sigma \* \int \sigma \* \left[ m_t(\theta)f_{t-1}(\theta) \ln \left( \frac{f_{t-1}(\theta)}{p_{t-1}(\theta)} \right) d\theta \right] dy - \int j \sigma \* \left[ g_t(y) \ln \left( \frac{g_t(y)}{h_t(y)} \right) \right] dy = D_{t-1} - D(g_t||h_t) \leq D_{t-1}.
\]

They use the Bayes rule (1), Fubini theorem [33], the equality \( \int j \sigma \* m_{t}(\theta)dy = 1 \), and non-negativity of the KLD [27]. The final claim is the property of any bounded super-martingale [30]. \( \square \)

A finite number of differences in evaluating of intermediate posterior pds has also tendency diminish. However, the approximations applied during potentially infinitely repeated data updating may completely spoil the super-martingale property. Then, the accumulated effect can cause the divergence of the KLD \( D_t \) for \( t \rightarrow \infty \), i.e. the compared pds can become asymptotically singular. The possibility of this behaviour formalises the faced problem.

3. Solution

The solution relies on a construction of a sequence of pds' sets containing the exact posterior pds.

3.1. Circumscribing set

At time \( t - 1 \), the approximate pd \( p_{t-1} \) represents the available information about the exact posterior pd \( f_{t-1} \). It differs both from the optimally approximating pd \( \hat{p}_{t-1} \) and from the unknown exact posterior pd \( f_{t-1} \). The already cited result [8] implies that a pd \( p_{t-1} \in p^* \) is an acceptable approximation of the pd \( f_{t-1} \) if there is a finite, ideally small, \( \beta_{t-1} \geq 0 \) such that \( D(f_{t-1}||p_{t-1}) \leq \beta_{t-1} < \infty \). Convexity of the functional \( D(f||p) \) with respect to the pd \( f \) implies convexity of the following set

\[
\{ f : D(f||p_{t-1}) \leq \beta_{t-1} < \infty \}
\]

within which the exact pd \( f_{t-1} \) stays. The expanded notation \( \hat{p}_{t-1} \) of \( f_{t-1} \) stresses its dependence on the given ”centre” \( p_{t-1} \) and ”radius” \( \beta_{t-1} \).

The following simple proposition shows that the Bayes rule maps the “ball” \( \hat{p}_{t-1} \) on a ball-like set.

Please cite this article in press as: M. Kárný, Approximate Bayesian recursive estimation, Inform. Sci. (2014), http://dx.doi.org/10.1016/j.ins.2014.01.048
Proposition 2 (The Bayes Rule Preserves Convexity). Let a set $\mathcal{F}$ of pd's on $\Theta^*$ be convex such that for any of its elements $f = f(\Theta)$ and the considered likelihood $m = m(\Theta)$ it holds $\int_{\Theta^*} m(\Theta)f(\Theta)d\Theta > 0$. Then, the image $B[f^*]$ of $f^*$ by the Bayes rule $B(1)$ is convex.

Proof. Let $f_k \in f^*$, $k = 1, 2$. The convexity of $f^*$ means that for any $z \in [0, 1]$ the pd $f_z = zf_1 + (1-z)f_2 \in f^*$. Its Bayes image is

$$B[f_z] = m(zf_1 + (1-z)f_2)$$
$$= \frac{m(zf_1 + (1-z)f_2)}{z \int_{\Theta^*} m(\Theta)f_1(\Theta)d\Theta + (1-z) \int_{\Theta^*} m(\Theta)f_2(\Theta)d\Theta}$$
$$= \frac{\epsilon B[f_1] + (1-\epsilon)B[f_2]}{\epsilon}$$

where the coefficient $\epsilon$ belongs to the interval $[0, 1]$ and the mapping $\epsilon \mapsto z$ is bijection. \(\square\)

The updated exact pd $f_t = B_t[f_{t-1}]$ as well as the pd $\tilde{f}_t = B_t[\tilde{f}_{t-1}]$ belong to the convex set $B_t[f^*_{t-1}] = \{ \tilde{f} = B_t[f], f \in f^*_{t-1} = f^*_{t-1}\} \}$, which is, however, quite complex in the considered generic case. Thus, it is necessary to circumscribe $B_t[f^*_{t-1}]$ by the set $f^*_{t} = f^*_{n_f}$, i.e. by the set of the form (7) for the increased time index.

Proposition 3 (Existence of $f^*_{n_f}$). Let the values of the predictive pd's $g_t(y) = \int_{\Theta^*} m_t(\Theta)f_{t-1}(\Theta)d\Theta, h_t(y) = \int_{\Theta^*} m_t(\Theta)p_{t-1}(\Theta)d\Theta$ be positive and finite for $y = y_t$. Let the likelihood $m_t$ as well as the ratio

$$\tilde{f}_t \equiv \frac{B_t[\tilde{f}_{t-1}]}{p_t}$$

be essentially bounded with respect to the measure $d\Theta$. Then a finite $\beta_t$ exists such that $f^*_{t} = f^*_{n_f}$ circumscribes $B_t[f^*_{t-1}]$.

Proof. The Bayes rule (1), the definition of the KLD (4), the definition of $\tilde{f}_t = B_t(p_{t-1})$ (6), the normalisation of $f_t(\Theta)$ and simple manipulations imply

$$D(f_t||p_t) = \int_{\Theta^*} m_t(\Theta)f_{t-1}(\Theta)ln\left(\frac{m_t(\Theta)f_{t-1}(\Theta)}{g_t(y_t)}\right)d\Theta$$
$$= \frac{1}{g_t(y_t)} \int_{\Theta^*} m_t(\Theta)f_{t-1}(\Theta)ln\left(\frac{f_{t-1}(\Theta)}{p_{t-1}(\Theta)}\right)d\Theta + \int_{\Theta^*} m_t(\Theta)f_{t-1}(\Theta)ln\left(\frac{\tilde{f}_t(\Theta)}{p_t(\Theta)}\right)d\Theta + \int_{\Theta^*} m_t(\Theta)f_{t-1}(\Theta)ln\left(\frac{h_t(y_t)}{g_t(y_t)}\right)$$

After the second equality, the first integral (multiplied by the finite factor $1/g_t(y_t)$) is bounded as it integrates the product of the essentially bounded $m_t(\Theta)$ and of the integrable function $f_{t-1}(\Theta)ln\left(\frac{f_{t-1}(\Theta)}{p_{t-1}(\Theta)}\right)$. The second summand is bounded by the finite value $ln\left(\text{esssup}_{\Theta^*} \frac{\tilde{f}_t(\Theta)}{p_t(\Theta)}\right)$. The third summand is bounded for positive and finite values of the involved predictive pd's. \(\square\)

3.2. Algorithmic construction of the circumscribing set

The construction of the circumscribing set $f^*_{t} = f^*_{n_f}$ coincides with the recursive choice of the centre $p_t$ and the radius $\beta_t$. The Bayes image $f_t = B_t[p_{t-1}]$ of the centre $p_{t-1}$ is an obvious candidate for the updated centre. It is generally out of the set $p^*$ of pd's of feasible forms so that it cannot be practically used. In accordance with [8], the optimal projection of the pd $f_t = B_t[p_{t-1}]$ on the set $p^*$ of feasible pd's offers as a new centre

$$\hat{p}_t \in \arg\min_{p \in p^*} D(f_t||p).$$

Then, it seemingly remains to select the radius $\beta_t$ so that $B_t[p_{t-1}][f^*_{n_f}] \subset f^*_{n_f}$. A permanent prolongation of this procedure is impossible as the centre would be updated in the naive way (6) and the error accumulation could cause an unbounded increase of the radius $\beta_t$ of the constructed circumscribing sets for $t \to \infty$. Thus, a centre $p_t$, which exploits the available information about the exact pd $f_t$ in a better way, is to be searched for.

A (generally iterative) search for a better centre simplifies substantially if the set of feasible pd's $p^*$ is closed with respect to taking geometric mean of its elements, if it is log-convex set. Further on, the log-convexity of $p^*$ is assumed without a substantial loss of applicability.

The axiomatically justified minimum KLD principle [35] recommends to replace the unknown, partially specified, exact pd $f_t$ by a pd $\tilde{f}_t$ with the smallest KLD on a pd representing the information before processing the information contained in
the pd \( \hat{p}_t \in p^* (8) \). The non-updated pd \( p_{t-1} \) is the available descriptor of such a (vague) information. The pd \( f_\gamma \) is searched in the trial circumscribing set \( f^* \) = \{f : D(f||\hat{p}_t) \leq \gamma \} parameterised by \( \gamma \in [0, \infty) \). The new centre \( \hat{p}_t \) is then found as the projection of \( f_\gamma \) on the set of feasible pds \( p^* \). Then, the trial radius \( \gamma \) is chosen so that the radius \( \hat{p}_t \) guaranteeing circumscription \( B_\gamma (f^* \gamma \) is the smallest one. Formally,

\[
\begin{align*}
  f_\gamma & \in \arg\min_{f : D(f||\hat{p}_t) \leq \gamma} D(f||p) \\
  p_{\gamma} & \in \arg\min_{p \in p^*} D(f_\gamma||p) \\
  (\hat{p}_t, \gamma) & \in \arg\min_{(p_t, \gamma) : \gamma \leq 0} D(f||p_t) \\
  p_t & = p_{\gamma}.
\end{align*}
\]

**Proposition 4 (The Optimal Centre).** Let the set \( p^* \) of approximate pds be log-convex. Then, the \( f_\gamma \) solving (9) is in \( p^* \). For a given trial radius \( \gamma \), the solution of (9) and (10) \( p_\gamma = f_\gamma \) has the form of stabilised forgetting [26] determined by a forgetting factor \( \lambda_\gamma \in [0, 1] \)

\[
\begin{align*}
  p_\gamma &= p_{\gamma} = \frac{\lambda_\gamma}{\lambda_\gamma + 1} \frac{1}{p_t^\gamma} p_{t-1}^\gamma = \frac{\lambda_\gamma}{\lambda_\gamma + 1} p_t^{-\lambda_\gamma} p_{t-1}^{\lambda_\gamma} \\
  L(\lambda_\gamma) &= \int_{\theta^*} p_t^\gamma p_{t-1}^{\lambda_\gamma} d\theta \\
  \lambda_\gamma &= 0 \text{ if } D(p_\gamma||\hat{p}_t) < \gamma \\
  \lambda_\gamma & \text{ solves } D(p_\gamma||\hat{p}_t) = \gamma \text{ otherwise.}
\end{align*}
\]

**Proof.** The Kuhn–Tucker optimality conditions [14] and the KLD property \( D(||g||d\theta) = 0 \iff f = g d\theta \)-almost surely provide the solution of the task (9). Its form and log-convexity of the set \( p^* \) imply that \( f_\gamma \in p^* \) and the nearest \( p_\gamma \in p^* (10) \) coincides with it. \( \square \)

The radiuses \( (\hat{p}_t, \gamma) \) guaranteeing that \( f^* \) circumscribes the set \( B_\gamma (f^* \gamma \), i.e. the solution of (11), depend on the chosen trial centre \( \hat{p}_t \), see (12). The found pd \( p_\gamma \) offers itself as a better centre. Under the adopted log-convexity of the set of feasible pds \( p^* \), the replacement of the centre \( \hat{p}_t \) by \( p_\gamma \) preserves the form (12) of \( p_\gamma \) and no improvement of \( \beta_\gamma \) can be gained. Thus, an iterative search is avoided.

To find the best radiuses \( (\hat{p}_t, \gamma) (11) \) either analytically or numerically is hard and the mapping of \( \gamma \) on the forgetting factor \( \lambda_\gamma \) is also complex. Instead, \( \lambda_\gamma \in [0, 1] \) approximatively minimising the expectation of the KLD \( \delta(f, \lambda) = D(f||p_\lambda) \), see (12), is searched for. The following proposition prepares the search and shows that the forgetting has indeed the potential to counteract the approximation-errors accumulation.

**Proposition 5 (Forgotten Probability Density Approximates Better).** To any pd \( f \) such that \( D(||\hat{p}_t|| < \infty, D(||p_{t-1}|| < \infty, \text{there is } \lambda_\gamma = \lambda(f) \in [0, 1] \text{ such that } D(||p_{\lambda}|| \leq D(||\hat{p}_t||. \)

**Proof.** The definitions of the KLD and the pd \( p_\lambda \) (12) give for any \( \lambda \in [0, 1] \)

\[
\begin{align*}
  \delta(f, \lambda) &= D(f||p_\lambda) = \lambda D(f||\hat{p}_t) + (1 - \lambda) D(f||p_{t-1}) + \ln(L(\lambda)).
\end{align*}
\]

The definition (12) of the pd \( p_\lambda \) also implies that \( p_{\lambda-1} = p_t \) and \( p_{\lambda-0} = p_{t-1} \). Denoting

\[
\begin{align*}
  R_\lambda &= \frac{p_{\lambda-1}}{p_{\lambda-0}} = \frac{p_t}{p_{t-1}},
\end{align*}
\]

derivatives of the function \( \delta(f, \lambda) \) get the forms

\[
\begin{align*}
  \frac{\partial \delta(f, \lambda)}{\partial \lambda} &= D(f||\hat{p}_t) - D(f||p_{t-1}) + \int_{\theta^*} p_\lambda(\theta) \ln(R_\lambda(\theta)) d\theta \\
  \frac{\partial^2 \delta(f, \lambda)}{\partial \lambda^2} &= \int_{\theta^*} p_\lambda(\theta) \ln^2(R_\lambda(\theta)) d\theta - \left[ \int_{\theta^*} p_\lambda(\theta) \ln(R_\lambda(\theta)) d\theta \right]^2.
\end{align*}
\]

The second derivative is the positive variance of \( \ln(R_\lambda(\theta)) \) with respect to the pd \( p_\lambda(\theta) \). Thus, \( \delta(f, \lambda) \) is a convex function of \( \lambda \in [0, 1] \) and reaches its minimum within this closed interval. \( \square \)

**Remarks.**

- The necessary condition \( \frac{\partial \delta(f, \lambda)}{\partial \lambda} = 0 \) for selecting the forgetting factor \( \lambda_\gamma = \lambda(f) \) minimising \( \delta(f, \lambda) = D(f||p_\lambda) \) has the appealing form, cf. (13) and (14),

Please cite this article in press as: M. Kárny, Approximate Bayesian recursive estimation, Inform. Sci. (2014), http://dx.doi.org/10.1016/ Lins.2014.01.048
\[
\int_{\Theta^*} f(\Theta) \ln(R_i(\Theta)) d\Theta = \int_{\Theta^*} p_\lambda(\Theta) \ln(R_i(\Theta)) d\Theta.
\]  
(15)

if the extreme is in (0, 1). Otherwise, \( \lambda = 0 \) if no \( \lambda \) solving (15) exists.

- The inequality claimed by Proposition 5 cannot be improved. It is trivially met for \( \lambda = 1 \) corresponding to the “naive” approximate learning step. The illustrative example, Section 5, indicates that it often—but not always—happens. Asymptotically, when the approximate posterior concentrates around the optimal projection of the almost surely converging exact posterior pd, the optimality of \( \lambda = 1 \) is even desirable.

The exact posterior pd \( f_t \) is unknown so that the optimal \( \lambda = \lambda(f_t) \) minimising \( \delta(f_t, \lambda) \) cannot be constructed. Proposition 1 shows that the pd \( \hat{f}_t = B_t(p_{t-1}) \) is expected to be nearer to the unknown exact pd \( f_t \) than the approximate pd \( p_{t-1} \) to the exact pd \( f_{t-1} \). Also, Proposition 3 guarantees that \( D(f||p_t) < \infty \), \( D(f||p_{t-1}) < \infty \) for \( f \in \{ f_t, f_{t-1} \} \). As the left-hand side of (15) depends linearly on the considered pd \( f_t \), it can be expected that the use \( f = \hat{f}_t \) instead of \( f = f_t \) will lead to \( \lambda_t(\hat{f}_t) \), which makes \( D(f_t||p_{t-1}) = \delta(f_t, \lambda_t(\hat{f}_t)) < \delta(f_t, \lambda = 1) = D(f_t||p_t) \). Thus, it is reasonable to use \( \lambda_t = \lambda_t(\hat{f}_t) \) instead of the inaccessible \( \lambda_t(f_t) \). This is the only heuristic step in designing the final recursive estimation applicable \( \forall t \in t^* \).

**Algorithm 1.** Approximate Recursive Bayesian Estimation

| Evaluate | \( \hat{f}_t = B_t[p_{t-1}] \propto m_{t|t-1} \) |
|----------|-------------------------------------------------|
| Find     | \( \hat{p}_t \in \text{argmin}_{p \in \Theta^*} D(\hat{f}_t||p) \) |
| Define   | \( p_t \propto \hat{p}_t^{\hat{f}_t} \) for \( \lambda \in [0, 1] \), and \( R_t = \frac{\hat{f}_t}{p_{t-1}} \) |
| Find     | \( \lambda_t \in [0, 1] \) solving |

\[
\int_{\Theta^*} \hat{f}_t(\Theta) \ln(R_i(\Theta)) d\Theta = \int_{\Theta^*} p_\lambda(\Theta) \ln(R_i(\Theta)) d\Theta \quad (\bullet)
\]

if no solution of (\( \bullet \)) exists set \( \lambda_t = \begin{cases} 0 & \text{if } \int_{\Theta^*} \hat{f}_t(\Theta) \ln(R_i(\Theta)) d\Theta \leq 0 \\ 1 & \text{otherwise} \end{cases} \)

Forget \( p_t = p_\lambda \).

**Remarks.**

- Algorithm 1 rectifies the estimator (6) by complementing it by a stabilised forgetting applied to the naive approximate pd \( \hat{p}_t \), while using the pd \( p_{t-1} \) as the needed alternative.
- Algorithm 1 reduces to the naive estimator (6) for \( \lambda_t = 1 \). It happens whenever no projection is needed after data updating, whenever \( f_t = \hat{p}_t \in \Theta^* \). Thus, it does not spoil the feasible optimal recursion. This smooth transition to the exact recursive estimation is an intuitively desirable property.
- Similar “flattening” techniques counteracting errors’ accumulation are already used, for instance, within applications of Monte Carlo methods for a recursive parameter estimation [31]. The proposed treatment formally supports the need for a technique of this type and recommends its form (12) without adding a new “tuning knob”.

4. Dynamic regression model with an arbitrary noise

The widely applicable class of dynamic regression parametric models with an external inputs (briefly regression models) with an arbitrary noise is considered here. It: (i) is useful per se; (ii) illustrates the respective steps of the general estimation algorithm; (iii) indicates a mild increase of the computational complexity needed for the proposed counteracting of the error accumulation. Algorithm 1 is tailored to this common parametric model, which arises from the following functional expansion of the conditional expectation

\[
y_t = E[y_t|u_t, d_{t-1}, \ldots, d_1, \Theta] + \sqrt{\tau} e_t = \theta^\prime \psi_t + \sqrt{\tau} e_t, \quad (\theta)\text{' is transposition.}
\]  
(17)

Such an expansion is often possible at least locally. The noise \( \sqrt{\tau} e_t \) as the difference between the output and its conditional expectation has the conditional expectation equal to zero. As such, it is uncorrelated over time and uncorrelated with data in the condition. The \( K \)-dimensional regression vector \( \psi_t, K < \infty \), is supposed to be recursively updatable. This makes the recursive updating of the data vector \( \Phi_t \) consisting of the output \( y_t \) and of the regression vector \( \psi_t \) possible. The unknown parameter \( \Theta = (\theta, r) \) includes the regression coefficients \( \theta \) and a scaling parameter \( r > 0 \) used in a pd \( \frac{1}{\sqrt{\tau}} \varphi(e, r) \) defining the distribution of the scaled noise \( e \). With it, the parametric regression model gets the form

Please cite this article in press as: M. Kárný, Approximate Bayesian recursive estimation, Inform. Sci. (2014), http://dx.doi.org/10.1016/
\[ m_t(\Theta) = \frac{1}{\sqrt{t}} g\left( \frac{|1, -\theta|}{\sqrt{t}}, r \right) = \frac{1}{\sqrt{t}} g(\eta^t\Phi_t, r). \eta^t = \eta^t(\theta, r) = \frac{|1, -\theta|}{\sqrt{t}}. \]  

(18)

When delimiting known properties of the noise, its distribution \( g \) is obtained by the minimum KLD principle. For instance, assuming that \( r \) is a finite noise variance and reducing the minimum KLD principle to maximum entropy principle, the allocated noise distribution is Gaussian. In the Gaussian case with the pd \( g(e, r) = (2\pi)^{-0.5} \exp \left[-\frac{e^2}{2}\right] \), the regression model has the Gauss-inverse-Wishart (GiW, [32]) as the conjugated prior

\[ p_t(\Theta) = p_t, \eta(\theta, r) = \frac{r^{-0.5(\nu_0+\kappa+2)} \exp\left[-0.5\eta^tL_t\eta\right]}{J(L_t, v_t)}, \Theta = (\theta, r), \]  

(19)

where \( L_t \) is lower triangular matrix with positive diagonal (symbolically, \( L_t > 0 \); it is Choleski square root of the extended information matrix), scalar \( v_t > 0 \) and \( \kappa \) is the length of the regression vector \( \psi_t \), which equals to the number of regression coefficients.

For the Gaussian regression model, this pd preserves its form during estimation and its use converts the data updating of the posterior pds into the algebraic updating

\[ L_t' = L_{t-1} + \Phi_t\Phi_t', \quad v_t = v_{t-1} + 1. \]  

(20)

The recursion starts from the chosen prior values \( L_0 > 0 \) and \( v_0 > 0 \).

The GiW pds have a chance to approximate well the exact posterior pds even for non-Gaussian regression models. This follows from the fact that the exact posterior pds quickly concentrate on a narrow support in the posterior pds into the algebraic updating from Section 3.2.1.

4.1. Specialisation of Algorithm 1

The specialisation of Algorithm 1 exploits the following elementary properties of the GiW pd, proved for instance in [16]. They are expressed in terms of the following splitting of the extended information matrix \( V_t = L_t' L_t \)

\[ V_t = \begin{bmatrix} V_{ty} & V_{ty} \\ V_{ty} & V_{t}\psi \end{bmatrix} = \begin{bmatrix} L_t^2 + L_{t0}L_0 & L_{t0}L_0' \\ L_{t0}L_0' & L_0L_0' \end{bmatrix} > 0. \]  

(21)

There, \( V_{ty}, V_{\psi} \) are scalars. The time index, separated by semicolon, is dropped whenever its occurrence brings no information.

**Proposition 6 (Some Properties of GiW Probability Density).** Let us consider GiW pds \( p_{t-1, v_{t-1}}, p_{t, v_t} \) with \( J(L_{t-1}, v_{t-1}) < \infty \) and \( J(L_t, v_t) < \infty \). Then, the pd \( p_t, \theta \propto p_t, \eta^{-1} \), given by \( \lambda \in [0, 1] \), is the pd \( p_{t, \eta} \) with

\[ L_t' = \lambda \left[ L_t', \sqrt{\lambda^t - 1L_{t-1}'} \right], \quad v_t = v_{t-1} + (1 - \lambda)v_{t-1}. \]  

(22)

For a GiW pd given by statistic values \( L_t > 0, v_t > 0 \) the normalisation factor \( J(L_t, v_t) \) in (19) is finite. Consequently, \( J(L_t, v_t) < \infty \) for \( L_{t-1} > 0, v_{t-1} > 0 \). The set \( \theta^* \) of GiW pds (19), serving as approximate feasible posterior pds, is log-convex set and the construction from Section 3.2 is directly applicable.

It holds

\[ J(L, v) = \pi^{0.5C} \Gamma(0.5v)(0.5A)^{-0.5v} C^{0.5}, \quad E[\theta|L, v] = \hat{\theta}, \quad E[r^{-1}|L, v] = \frac{1}{r} = \frac{\gamma}{A} \]

\[ E[\ln(r)|L, v] = \ln(0.5A) - \Psi(0.5v), \quad E[\eta^t|L, v] = \frac{1}{r} \left[ \begin{array}{c} \mathbf{1} \\ \hat{\eta} \\ \hat{\eta}^t \end{array} \right]. \]

There psi function \( \Psi(x) = \frac{\partial}{\partial x}, \Gamma(x) = \int_0^\infty z^{x-1} \exp(-z)dz < \infty \) for \( x > 0 \), [1], and entities known from least squares estimation method are used

\[ C = V_{\psi}^{-1} = L_{\psi}^{-1} \left(L_{\psi}^{-1}\right)', \quad A = V_y - V_{\psi'}V_{\psi}^{-1}V_{\psi y} = L_y^2, \quad \hat{\theta} = L_{\psi y}^{-1}L_{\psi y}. \]  

(23)

Recalling that the GiW pds (19) form the assumed set \( \theta^* \) of feasible approximate pds, the specialisation of Algorithm 1 starts with the application of the Bayes rule to the pd \( p_{t-1} = p_{t-1, v_{t-1}} \) and likelihood of the form (18)

\[ f_t(\theta, r) = g(\eta^t\Phi_t, r)p_{t-1, v_{t-1}+1}(\theta, r). \]  

(24)

The association of the factor \( \frac{1}{r} \) with the prior pd \( p_{t-1, v_{t-1}} \) makes the considered likelihood essentially bounded in generic case.

The KLD definition, the form of the approximating GiW pd and (24) imply that the approximate pd \( p_t = p_{t, v_t} \) minimising the KLD \( D(f_t||p) \) is determined by the statistic values \( L_t, v_t \).

Please cite this article in press as: M. Kárný, Approximate Bayesian recursive estimation, Inform. Sci. (2014), http://dx.doi.org/10.1016/j.ins.2014.01.048
\[ \bar{L}_t, \quad \tilde{v}_t \in \text{arg min}\{ -2\ln|L_t| + 2\ln(\Gamma(0.5\nu)) - \nu\ln(0.5L_t^2) + (\nu + \kappa + 2)\alpha_t + \text{tr}[B_{ty}(L_t^2 + L_{ty}L_{yy}^2) + 2B_{tyy}L_{tyy}^2 + \text{tr}[B_{tyy}L_{yy}^2]] \}
\]

\[ a_t = \int_{(0,\infty)^2} \ln(r) g(\eta^T\Phi r)_{p_{k1-1,1+1}(0,r)} d\eta dr \]

\[ B_t = \int_{(0,\infty)^2} \eta^T g(\eta^T\Phi r)_{p_{k1-1,1+1}(0,r)} d\eta dr \]

\[ \bar{V}_{t,ty} = B_{ty} \quad B_{tyy} \]

\[ \bar{V}_{t,y} = \bar{v}_{t,ty} \quad \bar{V}_{t,yy} = -B_{ty}^{-1} \quad \bar{V}_{t,tyy} = -U_{ty}^{-1} U_{t,yy}^{-1} \]

\[ \ln(0.5\bar{v}_t) - \Psi(0.5\bar{v}_t) = a_t + \ln(B_{yy}) \quad B_{ty} = U_{ty}^2 \]

\[ \bar{V}_t = \bar{L}_t \bar{L}_t \]

**Proposition 7** (The Best Projection on a GiW Probability Density). The GiW minimiser \( p_{k1-1,1+1}(0,r) \) of \( D(\bar{v}_t|\rho_{\bar{v},t}) \) for \( \bar{v}_t(0,r) \), see (24), is given by

\[ \bar{V}_{t,ty} = B_{ty} \quad B_{tyy} = -B_{ty}^{-1} \quad \bar{V}_{t,yy} = -U_{ty}^{-1} U_{t,yy}^{-1} \]

\[ \bar{V}_{t,y} = \bar{v}_{t,ty} \quad \bar{V}_{t,yy} = -B_{ty}^{-1} U_{t,yy} \quad \bar{V}_{t,tyy} = \bar{V}_{t,y} U_{t,yy} \]

\[ \ln(0.5\bar{v}_t) - \Psi(0.5\bar{v}_t) = a_t + \ln(B_{yy}) \quad B_{ty} = U_{ty}^2 \]

where \( \Psi \) and the least-squares entities (23), for \( \bar{V}_t = \bar{L}_t \bar{L}_t \), are used.

**Proof.** The proof exploits the necessary conditions for extreme, the matrix identities \( \frac{\partial \ln(C)}{\partial x} = C^{-1} \quad \frac{\partial \ln(C)}{\partial x} = D' \) and simple manipulations. □

The Eq. (26) for the positive scalar \( \tau = 0.5\bar{v}_t \) is the only one to be solved iteratively. The following proposition shows that the solution exists. It indicates that no problems should be encountered in a numerical solution.

**Proposition 8** (The Solution of (26)). The solution of (26) exists.

**Proof.** The right-hand side of the inspected equation has the form \( a_t + \ln(B_{yy}) = -\ln(\tau^r) + \ln(\tau^s) \geq 0 \), cf. (25). The non-negativity of this expression follows from Jensen inequality [33] applied either to the exact expectation or its Monte Carlo approximation. The same inequality also allows the bounding of the left-hand side

\[ \ln(\tau) - \Psi(\tau) = \ln(\tau) - \frac{\partial \ln(\int_0^\infty z^{-1}\exp(-z)dz)}{\partial \tau} = \ln(\tau) - \int_0^\infty \ln(z) z^{-1}\exp(-z)dz > \ln(\tau) - \ln\left(\frac{\Gamma(\tau+1)}{\Gamma(\tau)}\right) = 0 \]

where the equality \( \Gamma(\tau+1) = \tau \Gamma(\tau) \), [1], is used.

Moreover, \( -\Psi(\tau) = \bar{v}_t + \zeta(\tau) \), for \( \tau \rightarrow 0^+ \), where \( \zeta(\tau) \) is a bounded function [1]. Thus, the left-hand side \( \ln(\tau) - \Psi(\tau) = \frac{1}{\bar{v}_t}(\ln(\tau+1) + \zeta(\tau)) \) increases to infinity for \( \tau \rightarrow 0^+ \). This together with the continuous dependence of this function on \( \tau \) implies the claim. □

Having the projection \( p_{k1,1+1} \), it remains to find the forgetting factor \( \lambda_t \) meeting (●) in (16).

**Proposition 9** (Forgetting factor). Let us consider the approximate GiW pd (19) and denote

\[ \Delta V = V_{t-1} - \tilde{v}_t \quad \Delta V = V_{t-1} - \bar{V}_t \]

Then, the forgetting factor \( \lambda_t \) is

\[ (i) \text{ either a unique solution of the equation } \Delta v_t + \text{tr}(U_t U_t \Delta V) \]

\[ = \Delta v \ln(0.5\lambda_t) - \Psi(0.5v_t) + \text{tr}\left[ \frac{1}{\bar{V}_t} \begin{bmatrix} 1 & \hat{\theta}_t' & \hat{\theta}_t \hat{\theta}_t' + \bar{r}_t C_t \end{bmatrix} \Delta V \right] \]

Please cite this article in press as: M. Kárný, Approximate Bayesian recursive estimation, Inform. Sci. (2014), http://dx.doi.org/10.1016/j.ins.2014.01.048
\[
\Delta v = [\ln(0.5L_0^2) - \Psi(0.5v_0)] + \text{tr}\left(L_0^{-1}\left(L_0^{-1}\right)^\top \Delta V\right),
\]

(29)

where \(a_1, B_t = U'_t U_t\) are defined in (25). The used least squares entities correspond to the extended information matrix \(V_t = L_t^\top L_t\), see (22) also defining \(v_t\).

(ii) or equals 0 for (cf. (23))

\[
\Delta v_t + \text{tr}(U'_t U_t \Delta V) + 2 \ln \left(\frac{J(L_{t-1}, v_{t-1})}{J(L_t, v_t)}\right) \leq 0
\]

(30)

(iii) otherwise equals 1.

**Proof.** The statement is specialisation of the general results and the equation (\(\star\)) in (16). In the considered case,

\[
2 \ln |R_t(\theta, r)| = \Delta v + \text{tr}(U'_t U_t \Delta V) + \omega
\]

The left-hand side (\(l_{ih}\)) of the inspected equation (\(\star\)) gets the form

\[
l_{ih} = \Delta v_t + \text{tr}(U'_t U_t \Delta V) + \omega
\]

The right-hand side (\(r_{ih}\)) gets the analytical form

\[
r_{ih} = \Delta v E_t[\ln(r)] + \text{tr}(E_t[\eta|\eta'] \Delta V) + \omega,
\]

where \(E_t\) is expectation with respect to the inspected \(p_{t,y_t}\). The explicit formulae for the needed moments are in Proposition 6 and then it remains to set \(l_{ih} = r_{ih}\). If the solution does not exist then \(\lambda = 0\) is minimiser if \(D(f_t||p_{t-1}) \leq D(f_t||p_t)\), which translates into the condition (30). \(\square\)

This concludes specialisation of Algorithm 1 for the regression model (18) and the GiW class of approximate psds (19).

**Algorithm 2.** Approximate Recursive Bayesian Estimator for (18 and 19)

---

**Initial phase**

- Specify the regression model by specifying the modelled output \(y \in y^*\), structure of the regression vector \(\psi\) and the pd \(g\) of the noise.
- Set \(t = 0\) and specify the values of the prior statistic \(L_t, v_t\).
- Select parameters controlling evaluations, i.e. the number of Monte Carlo samples of \(\Theta = (\theta, r)\) and a precision used in solving equations for \(v_t\) and \(\lambda_t\).

**Recursive phase**, running for \(t \in t^*\),

1. Increase time counter.
2. Evaluate \(a_t, U_t\) in (25) for the processed data vector \(\Phi'_t = [y_t, \psi'_t]\).
3. Evaluate the statistic values \(L_t, v_t\) according to (26).
4. Determine the forgetting factor \(\lambda_t\) as specified in Proposition 9.
5. Set \(L_t L_t = \lambda_t L_t L_t + (1 - \lambda_t)L'_t L_{t-1}, v_t = \lambda_t v_t + (1 - \lambda_t)v_{t-1}\).

---

In a generic case, Step 2 of Recursive phase is performed by Monte Carlo with samples from GiW pd \(p_{t-1, v_{t-1}}\) (19). The iterative solution in Steps 3, 4 can employ any simple standard method. Step 4 brings the only computational overheads with respect to the naive recursive estimation.

Note that updating of Choleski square roots \(L_t, U_t\) can be efficiently performed by using standard rotation-based algorithms, see for instance [16].

---

5. Illustrative example

The presented example illustrates behaviour of the proposed estimator and indicates that the naive estimator (6), even when complemented by a fixed forgetting, exhibits the divergence the paper is fighting with.

The simple regression model with Cauchy noise was estimated

\[
m_t(\Theta) = \frac{\pi \sqrt{r}}{r} \left(1 + \frac{(y_t - \theta \psi_t)^2}{r}\right)^{-1}, \quad \Theta = ([\theta_1, \theta_2]', r), \quad \psi_t = [u_t, y_{t-1}]',
\]

(32)
where $\theta_1$, $\theta_2$ are real scalars, $r > 0$ and scalar real inputs $u_t$, $t \in \mathbb{Z}$, are realisations of a sequence of independent normal random variables with zero mean and unit variance. The simulated system is described by the same pd with $\Theta_s = (\theta_s', r_s') = ([0.8, 0.9], 1)$.

The specialisation of the theory as described in Section 4 was applied with the following options:

- The prior GiW distribution (19) given by unit $L_0$ and $v_0$ was used.
- The generalised moments $a_t, b_t$ (25) were evaluated by straightforward Monte Carlo methods using 600 samples from the GiW pd $p_{\theta_t, v_t, t+1}$. The unnecessarily high number of samples was chosen in order to see clearly the influence of approximation-errors accumulation.
- The equation for $v_t$ (26) was solved by a simple secant method and the equation for $\lambda_t$, Algorithm 2, was solved by brute force on a regular grid of forgetting factors with the step 0.005. The used factorised version of the estimator, Algorithm 2, indeed suppressed otherwise occurring numerical troubles.
- The simulation length of 2000 data pairs was sufficient to reach stationary phase of the estimation.
- The stabilised forgetting (12) with the constant $k_c = 0.95$ was used for comparison.

The simulated data are in Fig. 1 (left), typical behaviour of the proposed forgetting factor $\lambda_t$ is in Fig. 1 (right). Fig. 2 (left) shows time course of relative errors

$$E[\theta[L_t, v_t] - \theta_t]/\theta_t \text{ (meant entry-wise and applicable as } \theta_t \neq 0)$$

(33)

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{simulated_data.png}
\caption{Left: Simulated data, i.e. the modelled output (blue, large values) and the external input (green, small values). Right: Time course of the used forgetting factor $\lambda_t$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{relative_estimation_errors.png}
\caption{Left: Relative estimation errors (33) for the proposed estimator. Right: Relative estimation errors when the forgetting factor is $\lambda = 0.95$. The blue and green lines concern the first and second parameter entry, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)}
\end{figure}
for the proposed estimator and Fig. 2 (right) shows the same quantities for the naive estimator combined with a fixed forgetting.

A range of experiments differing in noise distributions, their realisations and prior pd was performed. Their outcomes looked quite similar. They confirmed that the presented results are typical and weakly dependent on the specific options made. The high occurrence of the extreme forgetting factors in \(0,1\) is characteristic to the very heavy-tailed Cauchy parametric model. The models with lighter tails are less demanding and infrequently require \(\lambda = 0\): they rarely drop the updating and projection pair.

6. Concluding remarks

Adaptive control [4] has been the author's original research domain. There, the recursive treatment and adaptation to changes dominate and a lot have been done in this area. Comparing to general cases, the processed data streams are there definitely simpler. The adaptive control, however, demonstrated that the common problems ranging from data preprocessing, detection of abrupt changing, adaptation to slow changes [25,26], dynamic clustering [16], etc. are systematically solvable within a single axiomatised framework [18]. The current paper fills a significant gap of the referred theory by designing a recursive Bayesian estimator, which counteracts the accumulation of approximation errors caused by the recursive treatment.

The proposed estimator is useful on its own. More importantly, it touches of the common problem of widely used recursive techniques like sequential Monte Carlo parameter estimation, variational Bayes, unscented-transformation based estimation, etc. The problem is especially urgent in the considered parameter estimation, where the errors are not damped by a stable state evolution. Respecting these observations in the future research promises a non-trivial improvements of established estimators and filters. This hypothesis is supported by recent attempts of this type, e.g. [29].

A finer structuring of the circumscribing set is a promising direction of the future research. Ideas connected with directional [25] and partial [10] forgetting are surely applicable.

Treatment of general data streams requires a systematic effort and corresponding expertise to get general algorithmic solutions available now only in the simpler adaptive-control context. This paper intends to stimulate a wider research interest in this respect. Recent papers like [13], dealing with drifts in data streams, confirm the need for it. The same observation applies to handling of crowd-wisdom-based classifiers [36], which obviously have to gradually switch from the batch to adaptive data-stream processing.

Acknowledgements

The research reflected in this paper has been supported by GAČR 13-13502S. Dr. Kamil Dedecius helped significantly in initial phase of the research that indicated we are on a proper track.

References