ARTICLE IN PRESS

Information Sciences xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Approximate Bayesian recursive estimation

Miroslav Kárný *

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

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Approximate parameter estimation
Bayesian recursive estimation
Kullback–Leibler divergence
Forgetting

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 $\mathsf{m}_t = \mathsf{m}_t(\Theta)$ describes a (modelled) output $y_t \in y^\star$[1] stimulated by an (external) input $u_t \in u^\star$ at discrete-time moments labelled by $t \in t^\star = \{1, 2, \ldots\}$. 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 $\Theta \in \Theta^\star$. The parameter is also unknown to the input generator. It means that u_t and Θ 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 Θ at time $t-1$ is expressed by the *exact* posterior pd $\mathsf{f}_{t-1} = \mathsf{f}_{t-1}(\Theta) = \mathsf{f}(\Theta|u_t, d_{t-1}, \ldots, d_1) = \mathsf{f}(\Theta|d_{t-1}, \ldots, d_1)$ (quantifying fully the available information). The Bayes rule $\mathcal{B}_t$ updates this pd by the data record d_t. The exact posterior pd evolves as follows

$$\mathsf{f}_t = \mathcal{B}_t[\mathsf{f}_{t-1}] \iff \mathsf{f}_t(\Theta) = \frac{\mathsf{m}_t(\Theta)\mathsf{f}_{t-1}(\Theta)}{\mathsf{g}_t(y_t)} \propto \mathsf{m}_t(\Theta)\mathsf{f}_{t-1}(\Theta), \quad \forall\, \Theta \in \Theta^\star, \tag{1}$$

$$\mathsf{g}_t(y_t) = \int_{\Theta^\star} \mathsf{m}_t(\Theta)\mathsf{f}_{t-1}(\Theta)\mathrm{d}\Theta, \tag{2}$$

where $\propto$ denotes equality up to normalisation. The *predictive pd* $\mathsf{g}_t(y)$ is determined by (2) with the fixed condition $u_t, d_{t-1}, \ldots, d_1$ and an arbitrary output $y \in y^\star$. The parametric model in (1) is treated as *likelihood*, i.e. as a function of Θ for a fixed inserted data $d_t, d_{t-1}, \ldots, d_1$. The recursion (1) is initiated by a designer-supplied prior pd $\mathsf{f}_0 = \mathsf{f}_0(\Theta)$ 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 $\mathsf{m}_t(\Theta), \quad \forall\, \Theta \in \Theta^\star$. A κ-dimensional statistic ψ_t (called *regression vector*, $\kappa < \infty$), which can be updated recursively, is assumed to comprise such an information.

The inspected problem arises when the exact posterior pd $\mathsf{f}_t = \mathsf{f}_t(\Theta)$ is too complex and has to be replaced by an *approximate* pd $\mathsf{p}_t = \mathsf{p}_t(\Theta)$. The pd p_t is a projection of f_t on a designer-selected *set of feasible pds* $\mathsf{p}^\star$. In [8], it was shown that the pd ${}^{O}\mathsf{p}_t \in \mathsf{p}^\star$ approximating *optimally* the exact posterior pd f_t is to minimise the KLD $\mathsf{D}(\mathsf{f}_t||\mathsf{p})$ [27].[2]

$${}^{O}\mathsf{p}_t \in \arg\min_{\mathsf{p}\in\mathsf{p}^\star} \mathsf{D}(\mathsf{f}_t||\mathsf{p}) = \arg\min_{\mathsf{p}\in\mathsf{p}^\star} \int_{\Theta^\star} \mathsf{f}_t(\Theta) \ln\left(\frac{\mathsf{f}_t(\Theta)}{\mathsf{p}(\Theta)}\right)\mathrm{d}\Theta. \tag{3}$$

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 ${}^{O}\mathsf{p}_t$ (3), i.e. to update recursively the *optimal* approximation ${}^{O}\mathsf{p}_{t-1}$ of the exact posterior pd f_{t-1}

$$\left({}^{O}\mathsf{p}_{t-1}, \mathsf{m}_t\right) \rightarrow {}^{O}\mathsf{p}_t. \tag{4}$$

[1] $x^\star$ denotes a set of xs. It is either a non-empty subset of a finite-dimensional real space or a subset of pds acting on the set of unknown parameters. The scalar-valued output is considered without a loss of generality as the multivariate case can always be treated entry-wise [16].

[2] The KLD, defined in (3) by the integral expression after equality, is conditioned on the data $d_t, \ldots, d_1$. The adopted simplified notation does not mark the condition explicitly. The KLD has many properties of distance between pds in its argument like non-negativity, equality to zero for almost surely equal arguments, etc. It is, however, asymmetric and does not meet triangle inequality.

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

The papers [20–24] mentioned in Introduction have shown that such a construction is possible if and only if the set $\mathsf{p}^\star$ is delimited by values of a finite collection of linear, time and data invariant, functionals $\mathcal{F}_k(\ln(\mathsf{f}_t))$ 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 $\mathsf{p}^\star$ makes the optimal recursive approximation (4) impossible. Thus, only a *non-optimal* approximate pd $\mathsf{p}_t \in \mathsf{p}^\star$ can be evolved instead of ${}^{o}\mathsf{p}_t$. 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}

$$(\mathsf{p}_{t-1}, \mathsf{m}_t) \rightarrow \mathsf{p}_t. \tag{5}$$

Such a feasible recursion is mostly constructed in the next *naive* way, $t \in t^\star$,

$$\begin{array}{lll} \text{Evaluate} & \tilde{\mathsf{f}}_t = & \mathcal{B}_t[\mathsf{p}_{t-1}] \\ \text{Find} & \tilde{\mathsf{p}}_t \in & \arg\min\limits_{\mathsf{p}\in\mathsf{p}^\star} \mathsf{D}(\tilde{\mathsf{f}}_t||\mathsf{p}) \\ \text{Set} & \mathsf{p}_t = & \tilde{\mathsf{p}}_t. \end{array} \tag{6}$$

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 $\mathsf{p}_{t-1} \neq \mathsf{f}_{t-1}$ 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* $\forall\, t \in t^\star, \mathsf{f}_t$ *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* $\mathsf{D}_0 = \mathsf{D}(\mathsf{f}_0||\mathsf{p}_0) < \infty$.

Then, the KLD values $\mathsf{D}_t = \mathsf{D}(\mathsf{f}_t||\mathsf{p}_t)$ *form super-martingale with respect to* σ*-algebras generated the processed data* $u_{t+1}, d_t, \ldots, d_1$. *Consequently,* D_t *converges almost surely to a finite value* $\mathsf{D}_\infty \in [0, \mathsf{D}_0]$ *for* $t \rightarrow \infty$.

Proof. Defining the predictive pds $\mathsf{g}_t(y) = \int_{\Theta^\star} \mathsf{m}_t(\Theta)\mathsf{f}_{t-1}(\Theta)\mathrm{d}\Theta$ and $\mathsf{h}_t(y) = \int_{\Theta^\star} \mathsf{m}_t(\Theta)\mathsf{p}_{t-1}(\Theta)\mathrm{d}\Theta$, the key super-martingale inequality is verified via the following evaluations of the conditional expectation

$$\begin{aligned} \mathsf{E}[\mathsf{D}_t|u_t, d_{t-1}, \ldots, d_1] &= \int_{y^\star} \mathsf{g}_t(y)\left[\int_{\Theta^\star} \mathsf{f}_t(\Theta)\ln\left(\frac{\mathsf{f}_t(\Theta)}{\mathsf{p}_t(\Theta)}\right)\mathrm{d}\Theta\right]\mathrm{d}y \\ &= \int_{y^\star}\int_{\Theta^\star} \mathsf{m}_t(\Theta)\mathsf{f}_{t-1}(\Theta)\ln\left(\frac{\mathsf{f}_{t-1}(\Theta)}{\mathsf{p}_{t-1}(\Theta)}\right)\mathrm{d}\Theta\mathrm{d}y - \int_{y^\star} \mathsf{g}_t(y)\ln\left(\frac{\mathsf{g}_t(y)}{\mathsf{h}_t(y)}\right)\mathrm{d}y = \mathsf{D}_{t-1} - \mathsf{D}(\mathsf{g}_t||\mathsf{h}_t) \leqslant \mathsf{D}_{t-1}. \end{aligned}$$

They use the Bayes rule (1), Fubini theorem [33], the equality $\int_{y^\star} \mathsf{m}_t(\Theta)\mathrm{d}y = 1$, and non-negativity of the KLD [27]. The final claim is the property of any bounded super-martingale [30]. □

A finite number of differences in evaluating of intermediate posterior pds have 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 ${}^{o}\mathsf{p}_{t-1}$ and from the unknown exact posterior pd f_{t-1}. The already cited result [8] implies that a pd $\mathsf{p}_{t-1} \in \mathsf{p}^\star$ is an acceptable approximation of the pd f_{t-1} if there is a *finite*, ideally small, $\beta_{t-1} \geqslant 0$ such that $\mathsf{D}(\mathsf{f}_{t-1}||\mathsf{p}_{t-1}) \leqslant \beta_{t-1} < \infty$. Convexity of the functional $\mathsf{D}(\mathsf{f}||\mathsf{p})$ with respect to the pd f implies convexity of the following set $\mathsf{f}^\star_{t-1} = \mathsf{f}^\star_{\mathsf{p}_{t-1}\beta_{t-1}}$ of pds on $\Theta^\star$

$$\mathsf{f}^\star_{t-1} = \mathsf{f}^\star_{\mathsf{p}_{t-1}\beta_{t-1}} = \{\mathsf{f} : \mathsf{D}(\mathsf{f}||\mathsf{p}_{t-1}) \leqslant \beta_{t-1} < \infty\} \tag{7}$$

within which the exact pd f_{t-1} stays. The expanded notation $\mathsf{f}^\star_{\mathsf{p}_{t-1}\beta_{t-1}}$ of $\mathsf{f}^\star_{t-1}$ stresses its dependence on the given "*centre*" p_{t-1} and "*radius*" β_{t-1}.

The following simple proposition shows that the Bayes rule maps the "ball" $\mathsf{f}^\star_{t-1}$ 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* $\mathsf{f}^\star$ *of pds on* $\Theta^\star$ *be convex such that for any of its elements* $\mathsf{f} = \mathsf{f}(\Theta)$ *and the considered likelihood* $\mathsf{m} = \mathsf{m}(\Theta)$ *it holds* $\int_{\Theta^\star} \mathsf{m}(\Theta)\mathsf{f}(\Theta)\mathrm{d}\Theta > 0$. *Then, the image* $\mathcal{B}[\mathsf{f}^\star]$ *of* $\mathsf{f}^\star$ *by the Bayes rule* $\mathcal{B}$ (1) *is convex.*

Proof. Let $\mathsf{f}_k \in \mathsf{f}^\star, k = 1, 2$. The convexity of $\mathsf{f}^\star$ means that for any $\alpha \in [0, 1]$ the pd $\mathsf{f}_\alpha = \alpha\mathsf{f}_1 + (1-\alpha)\mathsf{f}_2 \in \mathsf{f}^\star$. Its Bayes image is

$$\begin{aligned}\mathcal{B}[\mathsf{f}_\alpha] &= \frac{\mathsf{m}(\alpha\mathsf{f}_1 + (1-\alpha)\mathsf{f}_2)}{\alpha\int_{\Theta^\star}\mathsf{m}(\Theta)\mathsf{f}_1(\Theta)\mathrm{d}\Theta + (1-\alpha)\int_{\Theta^\star}\mathsf{m}(\Theta)\mathsf{f}_2(\Theta)\mathrm{d}\Theta}\\ &= \underbrace{\frac{\alpha\int_{\Theta^\star}\mathsf{m}(\Theta)\mathsf{f}_1(\Theta)\mathrm{d}\Theta}{\alpha\int_{\Theta^\star}\mathsf{m}(\Theta)\mathsf{f}_1(\Theta)\mathrm{d}\Theta + (1-\alpha)\int_{\Theta^\star}\mathsf{m}(\Theta)\mathsf{f}_2(\Theta)\mathrm{d}\Theta}}_{\epsilon}\mathcal{B}(\mathsf{f}_1) + \underbrace{\frac{(1-\alpha)\int_{\Theta^\star}\mathsf{m}(\Theta)\mathsf{f}_2(\Theta)\mathrm{d}\Theta}{\alpha\int_{\Theta^\star}\mathsf{m}(\Theta)\mathsf{f}_1(\Theta)\mathrm{d}\Theta + (1-\alpha)\int_{\Theta^\star}\mathsf{m}(\Theta)\mathsf{f}_2(\Theta)\mathrm{d}\Theta}}_{1-\epsilon}\mathcal{B}(\mathsf{f}_2)\\ &= \epsilon\mathcal{B}[\mathsf{f}_1] + (1-\epsilon)\mathcal{B}[\mathsf{f}_2],\end{aligned}$$

where the coefficient ϵ belongs to the interval $[0, 1]$ and the mapping $\epsilon \leftrightarrow \alpha$ is bijection. □

The updated exact pd $\mathsf{f}_t = \mathcal{B}_t[\mathsf{f}_{t-1}]$ as well as the pd $\tilde{\mathsf{f}}_t = \mathcal{B}_t[\mathsf{p}_{t-1}]$ belong to the convex set $\mathcal{B}_t[\mathsf{f}^\star_{t-1}] = \{\tilde{\mathsf{f}} : \tilde{\mathsf{f}} = \mathcal{B}_t[\mathsf{f}], \mathsf{f} \in \mathsf{f}^\star_{t-1} = \mathsf{f}^\star_{\mathsf{p}_{t-1}\beta_{t-1}}\}$, which is, however, quite complex in the considered generic case. Thus, it is necessary to circumscribe $\mathcal{B}_t[\mathsf{f}^\star_{t-1}]$ by the set $\mathsf{f}^\star_t = \mathsf{f}^\star_{\mathsf{p}_t\beta_t} = \{\mathsf{f} : \mathsf{D}(\mathsf{f}||\mathsf{p}_t) \leqslant \beta_t < \infty\}$, i.e. by the set of the form (7) for the increased time index.

Proposition 3 (*Existence of* $\mathsf{f}^\star_{\mathsf{p}_t\beta_t}$). *Let the values of the predictive pds* $\mathsf{g}_t(y) = \int_{\Theta^\star}\mathsf{m}_t(\Theta)\mathsf{f}_{t-1}(\Theta)\mathrm{d}\Theta, \mathsf{h}_t(y) = \int_{\Theta^\star}\mathsf{m}_t(\Theta)\mathsf{p}_{t-1}(\Theta)\mathrm{d}\Theta$ *be positive and finite for* $y = y_t$. *Let the likelihood* m_t *as well as the ratio*

$$\frac{\tilde{\mathsf{f}}_t}{\mathsf{p}_t} = \frac{\mathcal{B}_t(\mathsf{p}_{t-1})}{\textit{a new chosen centre in } \mathsf{p}^\star}$$

be essentially bounded with respect to the measure $\mathrm{d}\Theta$. *Then a finite* β_t *exists such that* $\mathsf{f}^\star_t = \mathsf{f}^\star_{\mathsf{p}_t\beta_t}$ *circumscribes* $\mathcal{B}_t[\mathsf{f}^\star_{t-1}]$.

Proof. The Bayes rule (1), the definition of the KLD (4), the definition of $\tilde{\mathsf{f}}_t = \mathcal{B}_t(\mathsf{p}_{t-1})$ (6), the normalisation of $\mathsf{f}_t(\Theta)$ and simple manipulations imply

$$\begin{aligned}\mathsf{D}(\mathsf{f}_t||\mathsf{p}_t) &= \int_{\Theta^\star}\frac{\mathsf{m}_t(\Theta)\mathsf{f}_{t-1}(\Theta)}{\mathsf{g}_t(y_t)}\ln\left(\frac{\mathsf{m}_t(\Theta)\mathsf{f}_{t-1}(\Theta)}{\mathsf{g}_t(y_t)\mathsf{p}_t(\Theta)}\right)\mathrm{d}\Theta\\ &= \frac{1}{\mathsf{g}_t(y_t)}\int_{\Theta^\star}\mathsf{m}_t(\Theta)\mathsf{f}_{t-1}(\Theta)\ln\left(\frac{\mathsf{f}_{t-1}(\Theta)}{\mathsf{p}_{t-1}(\Theta)}\right)\mathrm{d}\Theta + \int_{\Theta^\star}\frac{\mathsf{m}_t(\Theta)\mathsf{f}_{t-1}(\Theta)}{\mathsf{g}_t(y_t)}\ln\left(\frac{\tilde{\mathsf{f}}_t(\Theta)}{\mathsf{p}_t(\Theta)}\right)\mathrm{d}\Theta + \ln\left(\frac{\mathsf{h}_t(y_t)}{\mathsf{g}_t(y_t)}\right).\end{aligned}$$

After the second equality, the first integral (multiplied by the finite factor $1/\mathsf{g}_t(y_t)$) is bounded as it integrates the product of the essentially bounded $\mathsf{m}_t(\Theta)$ and of the integrable function $\mathsf{f}_{t-1}(\Theta)\ln\left(\frac{\mathsf{f}_{t-1}(\Theta)}{\mathsf{p}_{t-1}(\Theta)}\right)$. The second summand is bounded by the finite value $\ln\left(\mathrm{essup}_{\Theta^\star}\frac{\tilde{\mathsf{f}}_t(\Theta)}{\mathsf{p}_t(\Theta)}\right)$. The third summand is bounded for positive and finite values of the involved predictive pds. □

The adopted assumptions are probably stronger than necessary but they are verifiable as the realisation of the treated data record implies positivity of $\mathsf{g}_t(y_t)$ and other conditions can be inspected analytically.

3.2. Algorithmic construction of the circumscribing set

The construction of the circumscribing set $\mathsf{f}^\star_t = \mathsf{f}^\star_{\mathsf{p}_t\beta_t}$ coincides with the recursive choice of the centre p_t and the radius β_t. The Bayes image $\tilde{\mathsf{f}}_t = \mathcal{B}_t[\mathsf{p}_{t-1}]$ of the centre p_{t-1} is an obvious candidate for the updated centre. It is generally out of the set $\mathsf{p}^\star$ of pds of feasible forms so that it cannot be practically used. In accordance with [8], the optimal projection of the pd $\tilde{\mathsf{f}}_t = \mathcal{B}_t[\mathsf{p}_{t-1}]$ on the set $\mathsf{p}^\star$ of feasible pds offers as a new centre

$$\tilde{\mathsf{p}}_t \in \arg\min_{\mathsf{p}\in\mathsf{p}^\star}\mathsf{D}(\tilde{\mathsf{f}}_t||\mathsf{p}). \tag{8}$$

Then, it seemingly remains to select the radius β_t so that $\mathcal{B}_t[\mathsf{f}^\star_{\mathsf{p}_{t-1}\beta_{t-1}}] \subset \mathsf{f}^\star_{\tilde{\mathsf{p}}_t\beta_t}$. 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 radiuses* β_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 pds $\mathsf{p}^\star$ is closed with respect to taking geometric mean of its elements, if it is *log-convex set*. Further on, the log-convexity of $\mathsf{p}^\star$ 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 f_γ with the smallest KLD on a pd representing the information before processing the information contained in

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

the pd $\tilde{\mathsf{p}}_t \in \mathsf{p}^\star$ (8). The non-updated pd p_{t-1} is the available descriptor of such a (vague) information. The pd f_γ is searched in the *trial* circumscribing set $\mathsf{f}^\star_{\tilde{\mathsf{p}}_t\gamma} = \{\mathsf{f} : \mathsf{D}(\mathsf{f}||\tilde{\mathsf{p}}_t) \leqslant \gamma\}$ parameterised by $\gamma \in [0, \infty)$. The new centre p_γ is then found as the projection of f_γ on the set of feasible pds $\mathsf{p}^\star$. Then, the trial radius γ is chosen so that the radius β_t guaranteeing circumscription $\mathcal{B}_t(\mathsf{f}^\star_{t-1}) \subset \mathsf{f}^\star_{\mathsf{p}_\gamma\beta_t}$ is the smallest one. Formally,

$$\mathsf{f}_\gamma \in \arg\min_{\{\mathsf{f}:\mathsf{D}(\mathsf{f}||\tilde{\mathsf{p}}_t)\leqslant\gamma\}} \mathsf{D}(\mathsf{f}||\mathsf{p}_{t-1}) \tag{9}$$

$$\mathsf{p}_\gamma \in \arg\min_{\mathsf{p}\in\mathsf{p}^\star} \mathsf{D}(\mathsf{f}_\gamma||\mathsf{p}) \tag{10}$$

$$(\beta_t, \gamma_t) \in \arg\min_{(\beta,\gamma)\text{such that}\{\mathsf{f}^\star_{\mathsf{p}_\gamma\beta}\supset\mathcal{B}_t(\mathsf{f}^\star_{t-1})\}} \beta \tag{11}$$

$$\mathsf{p}_t = \mathsf{p}_{\gamma_t}.$$

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

$$\begin{aligned} \mathsf{p}_\gamma &= \mathsf{p}_{\lambda_\gamma} = \frac{\tilde{\mathsf{p}}_t^{\lambda_\gamma}\mathsf{p}_{t-1}^{1-\lambda_\gamma}}{\int_{\Theta^\star}\tilde{\mathsf{p}}_t^{\lambda_\gamma}\mathsf{p}_{t-1}^{1-\lambda_\gamma}\,d\Theta} = \frac{\tilde{\mathsf{p}}_t^{\lambda_\gamma}\mathsf{p}_{t-1}^{1-\lambda_\gamma}}{\mathsf{L}(\lambda_\gamma)} \\ \lambda_\gamma &= 0 \quad \text{if} \quad \mathsf{D}(\mathsf{p}_{t-1}||\tilde{\mathsf{p}}_t) < \gamma \\ \lambda_\gamma &\; \textit{solves} \quad \mathsf{D}(\mathsf{p}_\gamma||\tilde{\mathsf{p}}_t) = \gamma \; \textit{otherwise}. \end{aligned} \tag{12}$$

Proof. The Kuhn–Tucker optimality conditions [14] and the KLD property $\mathsf{D}(\mathsf{f}||\mathsf{g}) = 0 \iff \mathsf{f} = \mathsf{g}$ $d\Theta$-almost surely provide the solution of the task (9). Its form and log-convexity of the set $\mathsf{p}^\star$ imply that $\mathsf{f}_\gamma \in \mathsf{p}^\star$ and the nearest $\mathsf{p}_\gamma \in \mathsf{p}^\star$ (10) coincides with it. □

The radiuses (β_t, γ_t) guaranteeing that $\mathsf{f}^\star_{\mathsf{p}_{\gamma_t}\beta_t}$ circumscribes the set $\mathcal{B}_t(\mathsf{f}^\star_{t-1})$, i.e. the solution of (11), depend on the chosen trial centre $\tilde{\mathsf{p}}_t$ (8), see (12). The found pd p_{γ_t} offers itself as a better centre. Under the adopted log-convexity of the set of feasible pds $\mathsf{p}^\star$, the replacement of the centre $\tilde{\mathsf{p}}_t$ by p_{γ_t} preserves the form (12) of p_{γ_t} and no improvement of β_t can be gained. Thus, an iterative search is avoided.

To find the best radiuses (β_t, γ_t) (11) either analytically or numerically is hard and the mapping of γ_t on the forgetting factor λ_{γ_t} is also complex. Instead, $\lambda_t \in [0, 1]$ approximately minimising the expectation of the KLD $\delta(\mathsf{f}_t, \lambda) = \mathsf{D}(\mathsf{f}_t||\mathsf{p}_\lambda)$, 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* $\mathsf{D}(\mathsf{f}||\tilde{\mathsf{p}}_t) < \infty, \mathsf{D}(\mathsf{f}||\mathsf{p}_{t-1}) < \infty$, *there is* $\lambda_t = \lambda_t(\mathsf{f}) \in [0, 1]$ *such that* $\mathsf{D}(\mathsf{f}||\mathsf{p}_{\lambda_t}) \leqslant \mathsf{D}(\mathsf{f}||\tilde{\mathsf{p}}_t)$.

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

$$\delta(\mathsf{f}, \lambda) = \mathsf{D}(\mathsf{f}||\mathsf{p}_\lambda) = \lambda\mathsf{D}(\mathsf{f}||\tilde{\mathsf{p}}_t) + (1-\lambda)\mathsf{D}(\mathsf{f}||\mathsf{p}_{t-1}) + \ln(\mathsf{L}(\lambda)).$$

The definition (12) of the pd p_λ also implies that $\mathsf{p}_{\lambda=1} = \tilde{\mathsf{p}}_t$ and $\mathsf{p}_{\lambda=0} = \mathsf{p}_{t-1}$. Denoting

$$\mathsf{R}_t = \frac{\mathsf{p}_{\lambda=1}}{\mathsf{p}_{\lambda=0}} = \frac{\tilde{\mathsf{p}}_t}{\mathsf{p}_{t-1}}, \tag{13}$$

derivatives of the function $\delta(\mathsf{f}, \lambda)$ get the forms

$$\begin{aligned} \frac{\partial\delta(\mathsf{f}, \lambda)}{\partial\lambda} &= \mathsf{D}(\mathsf{f}||\tilde{\mathsf{p}}_t) - \mathsf{D}(\mathsf{f}||\mathsf{p}_{t-1}) + \int_{\Theta^\star}\mathsf{p}_\lambda(\Theta)\ln(\mathsf{R}_t(\Theta))d\Theta \\ \frac{\partial^2\delta(\mathsf{f}, \lambda)}{\partial\lambda^2} &= \int_{\Theta^\star}\mathsf{p}_\lambda(\Theta)\ln^2(\mathsf{R}_t(\Theta))d\Theta - \left[\int_{\Theta^\star}\mathsf{p}_\lambda(\Theta)\ln(\mathsf{R}_t(\Theta))d\Theta\right]^2. \end{aligned} \tag{14}$$

The second derivative is the positive variance of $\ln(\mathsf{R}_t(\Theta))$ with respect to the pd $\mathsf{p}_\lambda(\Theta)$. Thus, $\delta(\mathsf{f}, \lambda)$ is a convex function of $\lambda \in [0, 1]$ and reaches its minimum within this closed interval. □

Remarks.

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

$$\int_{\Theta^\star} \mathsf{f}(\Theta)\ln(\mathsf{R}_t(\Theta))\mathrm{d}\Theta = \int_{\Theta^\star} \mathsf{p}_{\lambda_t}(\Theta)\ln(\mathsf{R}_t(\Theta))\mathrm{d}\Theta, \tag{15}$$

if the extreme is in $(0,1)$. Otherwise, $\lambda_t \in \{0,1\}$ if no λ_t solving (15) exists.
- The inequality claimed by Proposition 5 cannot be improved. It is trivially met for $\lambda_t = 1$ 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 $\lambda_t = 1$ is even desirable.

The exact posterior pd f_t is unknown so that the optimal $\lambda_t = \lambda_t(\mathsf{f}_t)$ minimising $\delta(\mathsf{f}_t, \lambda)$ cannot be constructed. Proposition 1 shows that the pd $\tilde{\mathsf{f}}_t = \mathcal{B}_t(\mathsf{p}_{t-1})$ 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 $\mathsf{D}(\mathsf{f}||\tilde{\mathsf{p}}_t) < \infty, \mathsf{D}(\mathsf{f}||\mathsf{p}_{t-1}) < \infty$ for $\mathsf{f} \in \{\tilde{\mathsf{f}}_t, \mathsf{f}_t\}$. As the left-hand side of (15) depends linearly on the considered pd f, it can be expected that the use $\mathsf{f} = \tilde{\mathsf{f}}_t$ instead of $\mathsf{f} = \mathsf{f}_t$ will lead to $\lambda_t(\tilde{\mathsf{f}}_t)$, which makes $\mathsf{D}(\mathsf{f}_t||\mathsf{p}_{\lambda_t(\tilde{\mathsf{f}}_t)}) = \delta(\mathsf{f}_t, \lambda_t(\tilde{\mathsf{f}}_t)) < \delta(\mathsf{f}_t, \lambda = 1) = \mathsf{D}(\mathsf{f}_t||\tilde{\mathsf{p}}_t)$. Thus, it is reasonable to use $\lambda_t = \lambda_t(\tilde{\mathsf{f}}_t)$ instead of the inaccessible $\lambda_t(\mathsf{f}_t)$. This is the only heuristic step in designing the final recursive estimation applicable $\forall t \in t^\star$.

Algorithm 1. Approximate Recursive Bayesian Estimation

$$\begin{aligned}
&\textit{Evaluate}\quad \tilde{\mathsf{f}}_t = \mathcal{B}_t[\mathsf{p}_{t-1}] \propto \mathsf{m}_t\mathsf{p}_{t-1}\\
&\textit{Find}\quad \tilde{\mathsf{p}}_t \in \operatorname*{argmin}_{\mathsf{p}\in\mathsf{p}^\star} \mathsf{D}(\tilde{\mathsf{f}}_t||\mathsf{p})\\
&\textit{Define}\quad \mathsf{p}_\lambda \propto \tilde{\mathsf{p}}_t^{\lambda}\mathsf{p}_{t-1}^{1-\lambda},\ \textit{for}\ \lambda \in [0,1], \textit{and}\ \mathsf{R}_t = \frac{\tilde{\mathsf{p}}_t}{\mathsf{p}_{t-1}}\\
&\textit{Find}\quad \lambda_t \in [0,1] \textit{solving}
\end{aligned} \tag{16}$$

$$\int_{\Theta^\star} \tilde{\mathsf{f}}_t(\Theta)\ln(\mathsf{R}_t(\Theta))\mathrm{d}\Theta = \int_{\Theta^\star} \mathsf{p}_{\lambda_t}(\Theta)\ln(\mathsf{R}_t(\Theta))\mathrm{d}\Theta\ \ (\bullet)$$

$$\textit{if no solution of } (\bullet) \textit{ exists set } \lambda_t = \begin{cases} 0 & \textit{if} \quad \int_{\Theta^\star} \tilde{\mathsf{f}}_t(\Theta)\ln(\mathsf{R}_t(\Theta))\mathrm{d}\Theta \le 0 \\ 1 & \textit{otherwise} \end{cases}$$

$$\textit{Forget}\quad \mathsf{p}_t = \mathsf{p}_{\lambda_t}.$$

Remarks.

- Algorithm 1 rectifies the estimator (6) by complementing it by a stabilised forgetting applied to the naive approximate pd $\tilde{\mathsf{p}}_t$, 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 $\tilde{\mathsf{f}}_t = \tilde{\mathsf{p}}_t \in \mathsf{p}^\star$. 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

$$y_t = \mathsf{E}[y_t|u_t, d_{t-1}, \ldots, d_1, \Theta] + \sqrt{r}e_t = \theta'\psi_t + \sqrt{r}e_t, ' \text{ is transposition}. \tag{17}$$

Such an expansion is often possible at least locally. The noise $\sqrt{r}e_t$ 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 κ-dimensional regression vector $\psi_t, \kappa < \infty$, is supposed to be recursively updatable. This makes the recursive updating of the *data vector* Φ_t consisting of the output y_t and of the regression vector ψ_t possible. The unknown parameter $\Theta = (\theta, r)$ includes the *regression coefficients* θ and a *scaling parameter* $r > 0$ used in a pd $\frac{1}{\sqrt{r}}\mathsf{g}(e, r)$ defining the distribution of the scaled noise e. With it, the parametric regression model gets the form

$$\mathsf{m}_t(\Theta) = \frac{1}{\sqrt{r}}\mathsf{g}\left(\frac{[1,-\theta']}{\sqrt{r}}\Phi_t, r\right) = \frac{1}{\sqrt{r}}\mathsf{g}(\eta'\Phi_t, r),\ \eta' = \eta'(\theta, r) = \frac{[1,-\theta']}{\sqrt{r}}. \tag{18}$$

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 $\mathsf{g}(e, r) = (2\pi)^{-0.5}\exp\left[-\frac{e^2}{2}\right]$, the regression model has the Gauss-inverse-Wishart (GiW, [32]) as the conjugated prior

$$\mathsf{p}_t(\Theta) = \mathsf{p}_{L_t,\nu_t}(\theta, r) = \frac{r^{-0.5(\nu_t+\kappa+2)}\exp[-0.5\eta' L_t' L_t \eta]}{\mathsf{J}(L_t, \nu_t)},\ \Theta = (\theta, r), \tag{19}$$

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 $\nu_t > 0$ and κ is the length of the regression vector ψ_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

$$L_t' L_t = L_{t-1}' L_{t-1} + \Phi_t \Phi_t',\ \nu_t = \nu_{t-1} + 1. \tag{20}$$

The recursion starts from the chosen prior values $L_0 > 0$ and $\nu_0 > 0$.

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 $\Theta^\star = (\theta^\star, r^\star)$. It happens under general conditions [6] allowing the extension of the classical large-deviation theorem [34] to externally stimulated dynamic systems.

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 $V_t = L_t' L_t$

$$V_t = \begin{bmatrix} V_{t;y} & V_{t;y\psi}' \\ V_{t;y\psi} & V_{t;\psi} \end{bmatrix} = \begin{bmatrix} L_y^2 + L_{y\psi}' L_{y\psi} & L_{y\psi}' L\psi \\ L_\psi' L_{y\psi} & L_\psi' L_\psi \end{bmatrix} > 0. \tag{21}$$

There, $V_{t;y}, L_{t;y}$ 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* $\mathsf{p}_{L_{t-1},\nu_{t-1}}, \mathsf{p}_{\widetilde{L}_t,\tilde{\nu}_t}$ *with* $\mathsf{J}(L_{t-1}, \nu_{t-1}) < \infty$ *and* $\mathsf{J}(\widetilde{L}_t, \tilde{\nu}_t) < \infty$. *Then, the pd* $\mathsf{p}_\lambda \propto \tilde{\mathsf{p}}_t^\lambda \mathsf{p}_{t-1}^{1-\lambda}$, *given by* $\lambda \in [0, 1]$, *is the pd* $\mathsf{p}_{L_\lambda,\nu_\lambda}$ *with*

$$L_\lambda' L_\lambda = \lambda\left[\widetilde{L}_t', \sqrt{\lambda^{-1} - 1}L_{t-1}'\right]\left[\widetilde{L}_t', \sqrt{\lambda^{-1} - 1}L_{t-1}'\right]',\ \nu_\lambda = \lambda\tilde{\nu}_t + (1-\lambda)\nu_{t-1}. \tag{22}$$

For a GiW pd given by statistic values $L_t > 0, \nu_t > 0$ *the normalisation factor* $\mathsf{J}(L_t, \nu_t)$ *in* (19) *is finite. Consequently,* $\mathsf{J}(L_\lambda, \nu_\lambda) < \infty$ *for* $L_{t-1} > 0, \nu_{t-1} > 0$. *The set* $\mathsf{p}^\star$ *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

$$\mathsf{J}(L, \nu) = \pi^{0.5\kappa}\Gamma(0.5\nu)(0.5\Lambda)^{-0.5\nu}|C|^{0.5},\ \mathsf{E}[\theta|L, \nu] = \hat{\theta},\ \mathsf{E}[r^{-1}|L, \nu] = \frac{1}{\hat{r}} = \frac{\nu}{\Lambda}$$

$$\mathsf{E}[\ln(r)|L, \nu] = \ln(0.5\Lambda) - \Psi(0.5\nu),\ \mathsf{E}[\eta\eta'|L, \nu] = \frac{1}{\hat{r}}\begin{bmatrix} 1 & \hat{\theta}' \\ \hat{\theta} & \hat{\theta}\hat{\theta}' + \hat{r}C \end{bmatrix}.$$

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

$$C = V_\psi^{-1} = L_\psi^{-1}\left(L_\psi^{-1}\right)',\ \Lambda = V_y - V_{y\psi}' V_\psi^{-1} V_{y\psi} = L_y^2,\ \hat{\theta} = L_\psi^{-1} L_{y\psi}. \tag{23}$$

Recalling that the GiW pds (19) form the assumed set $\mathsf{p}^\star$ of feasible approximate pds, the specialisation of Algorithm 1 starts with the application of the Bayes rule to the pd $\mathsf{p}_{t-1} = \mathsf{p}_{L_{t-1},\nu_{t-1}}$ and likelihood of the form (18)

$$\tilde{\mathsf{f}}_t(\theta, r) \propto \mathsf{g}(\eta'\Phi_t, r)\mathsf{p}_{L_{t-1},\nu_{t-1}+1}(\theta, r). \tag{24}$$

The association of the factor $\frac{1}{\sqrt{r}}$ with the prior pd $\mathsf{p}_{L_{t-1},\nu_{t-1}}$ 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 $\tilde{\mathsf{p}}_t = \mathsf{p}_{\widetilde{L}_t,\tilde{\nu}_t}$ minimising the KLD $\mathsf{D}(\tilde{\mathsf{f}}_t||\mathsf{p})$ is determined by the statistic values $\widetilde{L}_t, \tilde{\nu}_t$

$$\widetilde{L}_t,\quad \tilde{v}_t \in \arg\min_{L,v}\{-2\ln|L_\psi| + 2\ln(\Gamma(0.5v)) - v\ln(0.5L_y^2) + (v+\kappa+2)a_t + \mathrm{tr}[B_{t;y}(L_y^2 + L'_{y\psi}L_{y\psi}) + 2B'_{t;y\psi}L'_\psi L_{y\psi}] + \mathrm{tr}[B_{t;\psi}L'_\psi L_\psi]\}$$

$$a_t = \int_{(\theta^\star,r^\star)} \ln(r)\frac{\mathrm{g}(\eta'\Phi_t,r)\mathrm{p}_{L_{t-1},v_{t-1}+1}(\theta,r)}{\int_{(\theta^\star,r^\star)}\mathrm{g}(\underline{\eta}'\Phi_t,\underline{r})\mathrm{p}_{L_{t-1},v_{t-1}+1}(\underline{\theta},\underline{r})\mathrm{d}\underline{\theta}\mathrm{d}\underline{r}}\mathrm{d}\theta\mathrm{d}r$$

$$B_t = \int_{(\theta^\star,r^\star)} \eta\eta'\frac{\mathrm{g}(\eta'\Phi_t,r)\mathrm{p}_{L_{t-1},v_{t-1}+1}(\theta,r)}{\int_{(\theta^\star,r^\star)}\int_{(\theta^\star,r^\star)}\mathrm{g}(\underline{\eta}'\Phi_t,\underline{r})\mathrm{p}_{L_{t-1},v_{t-1}+1}(\underline{\theta},\underline{r})\mathrm{d}\underline{\theta}\mathrm{d}\underline{r}}\mathrm{d}\theta\mathrm{d}r = \begin{bmatrix} B_{t;y} & B'_{t;y\psi} \\ B_{t;y\psi} & B_{t;\psi}\end{bmatrix} = \begin{bmatrix} U_{t;y} & 0 \\ U_{t;y\psi} & U'_{t;\psi}\end{bmatrix}\begin{bmatrix} U_{t;y} & U'_{t;y\psi} \\ 0 & U_{t;\psi}\end{bmatrix}. \tag{25}$$

The scalar a_t and the positive definite matrix B_t can be found efficiently by Monte Carlo sampling from the GiW pd $\mathrm{p}_{L_{t-1},v_{t-1}+1}(\theta,r)$. The upper-triangular Choleski factors U_t of B_t splits similarly as V_t (21).

Proposition 7 (*The Best Projection on a GiW Probability Density*). *The GiW minimiser $\mathrm{p}_{\widetilde{L}_t,\tilde{v}_t}(\theta,r)$ of $\mathrm{D}(\tilde{\mathrm{f}}_t||\mathrm{p}_{L,v})$ for $\tilde{\mathrm{f}}_t(\theta,r)$, see (24), is given by*

$$\begin{aligned}
&\widetilde{V}^{-1}_{t;\psi} = \widetilde{C}_t = B_{t;\psi} \iff \widetilde{L}_{t;\psi} = \left(U^{-1}_{t;\psi}\right)'\\
&\widetilde{V}_{t;y\psi} = -B^{-1}_{t;y}\widetilde{V}_{t;\psi}B_{t;y\psi} \iff \widetilde{L}_{t;y\psi} = -U^{-1}_{t;y}\left(U^{-1}_{t;\psi}\right)'U_{t;y\psi} \iff \hat{\tilde{\theta}}_t = -U^{-1}_{t;y}U_{t;y\psi}\\
&\widetilde{V}_{t;y} = \frac{\tilde{v}_t}{B_{t;y}} + \widetilde{V}'_{t;y\psi}\widetilde{V}^{-1}_{t;\psi}\widetilde{V}_{t;y\psi} \iff \tilde{\Lambda}_t = \frac{\tilde{v}_t}{B_{t;y}} = \frac{1}{\hat{\tilde{r}}_t} \iff \widetilde{L}_{t;y} = \frac{\sqrt{\tilde{v}_t}}{U_{t;y}}\\
&\ln(0.5\tilde{v}_t) - \Psi(0.5\tilde{v}_t) = a_t + \ln(B_{t;y}),\ B_{t;y} = U^2_{t;y},
\end{aligned} \tag{26}$$

where Ψ function and the least-squares entities (23), for $\widetilde{V}_t = \widetilde{L}'_t\widetilde{L}_t$, are used.

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

The Eq. (26) for the positive scalar $\tau = 0.5\tilde{v}_t$ 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 $a_t + \ln(B_{t;y}) = -\mathrm{E}[\ln(r^{-1})] + \ln(\mathrm{E}[r^{-1}]) \geqslant 0$, 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

$$\ln(\tau) - \Psi(\tau) = \ln(\tau) - \frac{\partial\ln\left(\int_0^\infty z^{\tau-1}\exp(-z)\mathrm{d}z\right)}{\partial\tau} = \ln(\tau) - \int_0^\infty \ln(z)\frac{z^{\tau-1}\exp(-z)}{\int_0^\infty \underline{z}^{\tau-1}\exp(-\underline{z})\mathrm{d}\underline{z}}\mathrm{d}z \geqslant \ln(\tau) - \ln\left(\frac{\Gamma(\tau+1)}{\Gamma(\tau)}\right) = 0,$$

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

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

Having the projection $\mathrm{p}_{\tilde{L}_t,\tilde{v}_t}$, it remains to find the forgetting factor λ_t meeting (•) in (16).

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

$$\Delta v = v_{t-1} - \tilde{v}_t,\quad \Delta V = V_{t-1} - \widetilde{V}_t. \tag{27}$$

Then, the forgetting factor $\lambda_t \in [0,1]$ minimising $\delta(\tilde{\mathrm{f}}_t,\lambda) = \mathrm{D}(\tilde{\mathrm{f}}_t||\mathrm{p}_{L_\lambda,v_\lambda})$ is

(i) either a unique solution of the equation

$$\Delta v a_t + \mathrm{tr}\left(U'_t U_t \Delta V\right) \tag{28}$$

$$= \Delta v[\ln(0.5\Lambda_\lambda) - \Psi(0.5v_\lambda)] + \mathrm{tr}\left(\frac{1}{\hat{r}_\lambda}\begin{bmatrix} 1 & \hat{\theta}'_\lambda \\ \hat{\theta}_\lambda & \hat{\theta}_\lambda\hat{\theta}'_\lambda + \hat{r}_\lambda C_\lambda\end{bmatrix}\Delta V\right)$$

$$= \Delta\nu[\ln(0.5L^2_{y\lambda}) - \Psi(0.5\nu_\lambda)] + \mathrm{tr}\left(L_\lambda^{-1}\left(L_\lambda^{-1}\right)'\Delta V\right), \tag{29}$$

where $a_t, B_t = U_t'U_t$ are defined in (25). The used least squares entities correspond to the extended information matrix $V_\lambda = L_\lambda' L_\lambda$, see (22) also defining ν_λ,

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

$$\Delta\nu a_t + \mathrm{tr}\left(U_t'U_t\Delta V\right) + 2\ln\left(\frac{\mathsf{J}(L_{t-1},\nu_{t-1})}{\mathsf{J}(\widetilde{L}_t,\tilde{\nu}_t)}\right) \leqslant 0 \tag{30}$$

(iii) otherwise equals 1.

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

$$2\ln(\mathsf{R}_t(\theta,r)) = \Delta\nu\ln(r) + \mathrm{tr}(\eta\eta'\Delta V) + \underbrace{2\ln\left(\frac{\mathsf{J}(L_{t-1},\nu_{t-1})}{\mathsf{J}(\widetilde{L}_t,\tilde{\nu}_t)}\right)}_{\omega} \tag{31}$$

The left-hand side (lh_t) of the inspected equation (•) gets the form

$$lh_t = \Delta\nu a_t + \mathrm{tr}\left(U_t'U_t\Delta V\right) + \omega$$

The right-hand side (rh_t) gets the analytical form

$$rh_t = \Delta\nu\mathsf{E}_\lambda[\ln(r)] + \mathrm{tr}(\mathsf{E}_\lambda[\eta\eta']\Delta V) + \omega,$$

where E_λ is expectation with respect to the inspected $\mathsf{p}_{L_\lambda,\nu_\lambda}$. The explicit formulae for the needed moments are in Proposition 6 and then it remains to set $lh_t = rh_t$. If the solution does not exist then $\lambda = 0$ is minimiser if $\mathsf{D}(\tilde{\mathsf{f}}_t||\mathsf{p}_{t-1}) \leqslant \mathsf{D}(\tilde{\mathsf{f}}_t||\tilde{\mathsf{p}}_t)$, which translates into the condition (30). □

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

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

Initial phase

- Specify the regression model by specifying the modelled output $y \in y^\star$, structure of the regression vector ψ and the pd g of the noise.
- Set $t = 0$ and specify the values of the prior statistic L_t, ν_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 ν_t and λ_t.

Recursive phase, running for $t \in t^\star$,

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 $\widetilde{L}_t, \tilde{\nu}_t$ according to (26).
4. Determine the forgetting factor λ_t as specified in Proposition 9.
5. Set $L_t'L_t = \lambda_t\widetilde{L}_t'\widetilde{L}_t + (1-\lambda_t)L_{t-1}'L_{t-1}, \nu_t = \lambda_t\tilde{\nu}_t + (1-\lambda_t)\nu_{t-1}$.

In a generic case, Step 2 of Recursive phase is performed by Monte Carlo with samples from GiW pd $\mathsf{p}_{L_{t-1},\nu_{t-1}+1}$ (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

$$\mathsf{m}_t(\Theta) = \left[\pi\sqrt{r}\left(1 + \frac{(y_t - \theta'\psi_t)^2}{r}\right)\right]^{-1}, \quad \Theta = ([\theta_1,\theta_2]', r), \quad \psi_t = [u_t, y_{t-1}]', \tag{32}$$

where θ_1, θ_2 are real scalars, $r > 0$ and scalar real inputs $u_t, t \in t^\star$, 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 ν_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 $\mathsf{p}_{L_{t-1},\nu_{t-1}+1}$. The unnecessarily high number of samples was chosen in order to see clearly the influence of approximation-errors accumulation.
- The equation for $\tilde{\nu}_t$ (26) was solved by a simple secant method and the equation for λ_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 $\lambda_\gamma = 0.95$ was used for comparison.

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

$$(\mathsf{E}[\theta|L_t, \nu_t] - \theta_s)/\theta_s \text{ (meant entry-wise and applicable as } \theta_s \neq 0) \tag{33}$$

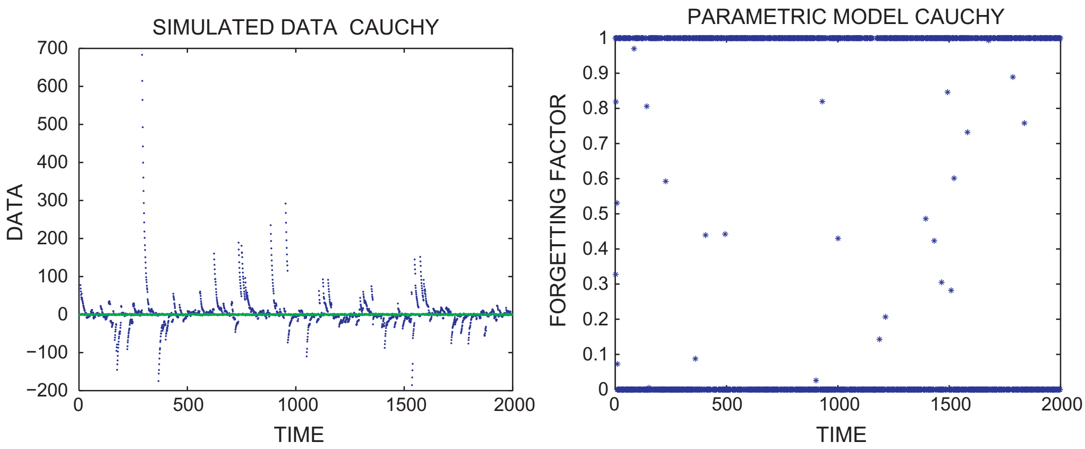

Fig. 1. Left: Simulated data, i.e. the modelled output (blue, large values) and the external input (green, small values). Right: Time course of the used forgetting factor λ_t. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

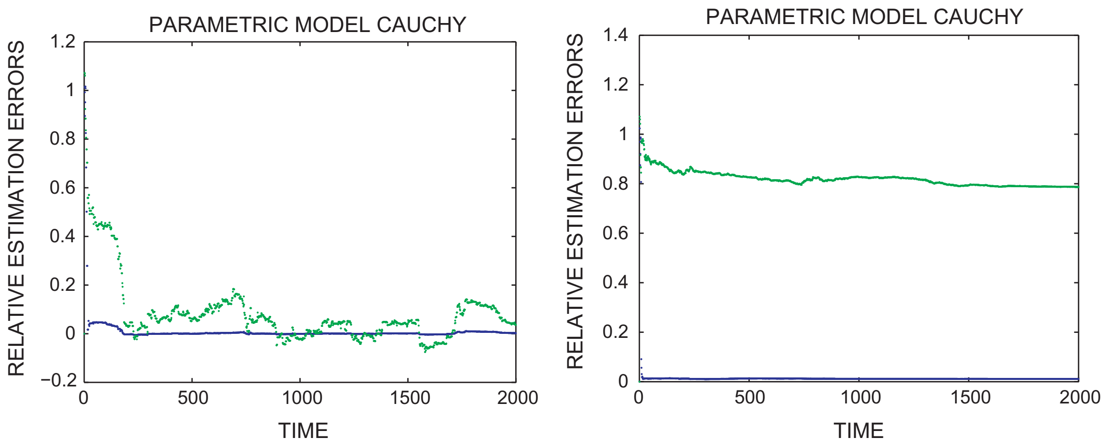

Fig. 2. Left: Relative estimation errors (33) for the proposed estimator. Right: Relative estimation errors when the forgetting factor is $\lambda = 0.95$. The blue and green lines concern the first and second parameter entry, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

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

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

[1] M. Abramowitz, I.A. Stegun, Handbook of Mathematical Functions, Dover Publications, New York, 1972.
[2] C. Aggarwal (Ed.), Data Streams: Models and Algorithms, Springer, 2007.
[3] D.L. Alspach, H.W. Sorenson, Nonlinear Bayesian estimation using Gaussian sum approximation, IEEE Trans. Autom. Control 17 (4) (1972) 439–448.
[4] K.J. Astrom, B. Wittenmark, Adaptive Control, Addison-Wesley, Massachusetts, 1989.
[5] A. Benveniste, M. Métivier, P. Priouret, Adaptive Algorithms and Stochastic Approximations, Springer, Berlin, 1990.
[6] L. Berec, M. Kárný, Identification of reality in Bayesian context, in: K. Warwick, M. Kárný (Eds.), Computer-Intensive Methods in Control and Signal Processing, Birkhäuser, 1997, pp. 181–193.
[7] J.O. Berger, Statistical Decision Theory and Bayesian Analysis, Springer, New York, 1985.
[8] J.M. Bernardo, Expected information as expected utility, Ann. Stat. 7 (3) (1979) 686–690.
[9] F. Daum, Nonlinear filters: beyond the Kalman filter, IEEE Aero. Electron. Syst. Mag. 20 (8) (2005) 57–69.
[10] K. Dedecius, I. Nagy, M. Kárný, Parameter tracking with partial forgetting method, Int. J. Adapt. Control Signal Process. 26 (1) (2012) 1–12.
[11] A. Doucet, V.B. Tadić, Parameter estimation in general state-space models using particle methods, Ann. Inst. Stat. Math. 55 (2) (2003) 409–422.
[12] M.M. Gaber, A. Zaslavsky, S. Krishnaswamy, Mining data streams: a review, SIGMOD Record 34 (2) (2005) 1–26.
[13] J. Gama, I. Ziobailte, A. Bifet, ACM computing surveys, J. Math. Econ. 1 (1) (2013).
[14] R. Horst, H. Tuy, Global Optimization, Springer, 1996. 727 pp..
[15] S.J. Julier, J.K. Uhlmann, H.F. Durrant-Whyte, A new approach for the nonlinear transformation of means and covariances in linear filters, IEEE Trans. Autom. Control 5 (3) (2000) 477–482.
[16] M. Kárný, J. Böhm, T.V. Guy, L. Jirsa, I. Nagy, P. Nedoma, L. Tesař, Optimized Bayesian Dynamic Advising: Theory and Algorithms, Springer, 2006.
[17] M. Kárný, T.V. Guy, On support of imperfect Bayesian participants, in: T.V. Guy, M. Kárný, D.H. Wolpert (Eds.), Decision Making with Imperfect Decision Makers, vol. 28, Springer, Berlin, 2012. Intelligent Systems Reference Library.
[18] M. Kárný, T. Kroupa, Axiomatisation of fully probabilistic design, Inform. Sci. 186 (1) (2012) 105–113.
[19] L. Khan, Data stream mining: challenges and techniques, in: Proc. 22nd IEEE International Conference on Tools with Artificial Intelligence, 2010.
[20] R. Kulhavý, A Bayes-closed approximation of recursive nonlinear estimation, Int. J. Adapt. Control Signal Process. 4 (1990) 271–285.
[21] R. Kulhavý, Recursive Bayesian estimation under memory limitations, Kybernetika 26 (1990) 1–20.
[22] R. Kulhavý, Recursive nonlinear estimation: a geometric approach, Automatica 26 (3) (1990) 545–555.
[23] R. Kulhavý, Can approximate Bayesian estimation be consistent with the ideal solution?, in: Proc. of the 12th IFAC World Congress, vol. 4, Sydney, Australia, 1993, pp. 225–228.
[24] R. Kulhavý, Implementation of Bayesian parameter estimation in adaptive control and signal processing, Statistician 42 (1993) 471–482.
[25] R. Kulhavý, M. Kárný, Tracking of slowly varying parameters by directional forgetting, in: Preprints of the 9th IFAC World Congress, vol. X, IFAC, Budapest, 1984, pp. 178–183.
[26] R. Kulhavý, M.B. Zarrop, On a general concept of forgetting, Int. J. Control 58 (4) (1993) 905–924.

[27] S. Kullback, R. Leibler, On information and sufficiency, Ann. Math. Stat. 22 (1951) 79–87.
[28] H.J. Kushner, Stochastic Stability and Control, Academic Press, 1967.
[29] F. Liu, D. Qian, C. Liu, An artificial physics optimized particle filter, Kong. yu Juece/Control Decis. 27 (8) (2012) 1145–1149.
[30] M. Loeve, Probability Theory, van Nostrand, Princeton, New Jersey, Russian translation, Moscow, 1962.
[31] P. Del Moral, A. Doucet, S.S. Singh, Forward smoothing using sequential Monte Carlo, Technical Report CUED/F-INFENG/TR 638, Cambridge University, 2010.
[32] V. Peterka, Bayesian system identification, in: P. Eykhoff (Ed.), Trends and Progress in System Identification, Pergamon Press, Oxford, 1981. pp. 239–304.
[33] M.M. Rao, Measure Theory and Integration, John Wiley, NY, 1987.
[34] I.N. Sanov, On probability of large deviations of random variables, Matemat. Sborn. 42 (1957) 11–44. also in Selected Translations Mathematical Statistics and Probability I (1961) 213–244 (in Russian)..
[35] J. Shore, R. Johnson, Axiomatic derivation of the principle of maximum entropy & the principle of minimum cross-entropy, IEEE Trans. Inform. Theory 26 (1) (1980) 26–37.
[36] E. Simpson, S. Roberts, I. Psorakis, A. Smith, Dynamic Bayesian combination of multiple imperfect classifiers, in: T.V. Guy, M. Kárný, D.H. Wolpert (Eds.), Decision Making and Imperfection, vol. 28, Springer, Berlin, 2013, pp. 1–38. Studies in Computation Intelligence.
[37] V. Šmídl, A. Quinn, The Variational Bayes Method in Signal Processing, Springer, 2005.