Information Sciences 285 (2014) 100–111

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:
Received 19 February 2013
Received in revised form 19 January 2014
Accepted 26 January 2014
Available online 11 February 2014

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 $d\bullet$, [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/

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}

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

[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.