INTERNATIONAL JOURNAL OF ADAPTIVE CONTROL AND SIGNAL PROCESSING
Int. J. Adapt. Control Signal Process. (2008)
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/acs.1080

Use of Kullback–Leibler divergence for forgetting

Miroslav Kárný*,† and Josef Andrýsek

Adaptive Systems Department, Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, P.O. Box 18, 182 08 Prague, Czech Republic

SUMMARY

Non-symmetric Kullback–Leibler divergence (KLD) measures proximity of probability density functions (pdfs). Bernardo (*Ann. Stat.* 1979; **7**(3):686–690) had shown its unique role in approximation of pdfs. The order of the KLD arguments is also implied by his methodological result. Functional approximation of estimation and stabilized forgetting, serving for tracking of slowly varying parameters, use the reversed order. This choice has the pragmatic motivation: recursive estimator often approximates the parametric model by a member of exponential family (EF) as it maps prior pdfs from the set of conjugate pdfs (CEF) back to the CEF. Approximations based on the KLD with the reversed order of arguments preserves this property. In the paper, the approximation performed within the CEF but with the proper order of arguments of the KLD is advocated. It is applied to the parameter tracking and performance improvements are demonstrated. This practical result is of importance for adaptive systems and opens a way for improving the functional approximation. Copyright © 2008 John Wiley & Sons, Ltd.

Received 11 December 2007; Revised 6 August 2008; Accepted 12 September 2008

KEY WORDS: Bayesian estimation; Kullback–Leibler divergence; functional approximation of estimation; parameter tracking by stabilized forgetting; ARX model

1. INTRODUCTION

Recursive estimation of finitely parameterized models of random input–output relationships is the core of adaptive systems. Bayesian methodology [1, 2] solves it consistently. Essentially, the exponential family (EF) [3] of parametric models is the only class of dynamic models in which the exact Bayesian estimation can be used. This stimulates the search for estimators approximating the recursively infeasible posterior probability density function (pdf).

*Correspondence to: Miroslav Kárný, Adaptive Systems Department, Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, P.O. Box 18, 182 08 Prague, Czech Republic.
†E-mail: school@utia.cas.cz

Contract/grant sponsor: Ministry of Education of the Czech Republic; contract/grant number: 2C06001
Contract/grant sponsor: Grant Agency of the Czech Republic; contract/grant number: 102/08/0567

Within the Bayesian paradigm, the approximating pdf is to be the minimizer of a suitable expected loss. Under mild conditions, Bernardo [4] had shown that the expected loss expressing proximity of a pair of pdfs is to be the Kullback–Leibler divergence (KLD) [5]. He even recommended the order of arguments of the non-symmetric KLD. Mostly, the unrestricted optimization of the KLD provides the pdf out of the desirable class CEF of the pdfs conjugated to members of the EF [6]. This property makes developers of approximating algorithms to use the alternative ('incorrect') order of the KLD arguments. Parameter tracking via the stabilized forgetting [7] and the functional approximation of estimation [8, 9] are prominent examples in this respect. This stimulates a natural question: can we gain *practically* by minimizing the proper KLD within the CEF? The positive answer given in the paper improves parameter tracking and explains observed improvements of the functional approximation [10, 11].

Section 2 is a preparatory one. Section 3 demonstrates the conceptual preference of the KLD with the 'correct' order of arguments. Section 4 specializes this result to forgetting design elaborated for normal autoregressive model with exogenous variables (ARX). Section 5 converts this parameter tracking into a widely applicable algorithm by recommending a meaningful choice of the alternative pdf determining it. Section 6 illustrates the properties of this tracking. Section 7 provides concluding remarks.

2. PRELIMINARIES

2.1. Notation

Throughout, $\equiv$ is equality by definition; X^* denotes a set of X-values; ℓ_X means the length of vector X; $f(\cdot|\cdot)$ denotes pdf, the pdfs are distinguished by names of their arguments and by Greek subscripts α, β, ι; t labels discrete-time moments, $t \in t^* \equiv \{1, \ldots, T\}$, $T \leqslant \infty$; $d_t = (y_t, u_t)$ is the data record at time t consisting of an observed system output y_t and of an optional system input u_t, possibly void; Θ is the common symbol for unknown parameters; $d^{k:t}$ denotes the sequence $(d_k, \ldots, d_t)$; $d^t \equiv d^{1:t}$; $X_{t;i}$ is ith entry of the array X_t. No formal distinction is made between a random variable, its realization and the pdf argument. All integrals are multivariate and definite over the argument domain.

2.2. Bayesian estimation in EF

Generally, the Bayesian paradigm operates on the joint pdf of all considered uncertain variables. It composes this pdf from its conditional factors and derives its particular marginal or conditional versions. It inserts any available realization in them.

In parameter estimation, a sequence d^T of ℓ_d-dimensional *data records* d_t and an *unknown*, finite-dimensional *parameter* Θ are relevant variables. Their joint pdf is

$$\overbrace{f(d^T, \Theta)}^{\text{joint pdf}} = \overbrace{f(\Theta)}^{\text{prior pdf}} \times \overbrace{\prod_{t=1}^{T}}^{\text{time}} \overbrace{\prod_{i=1}^{\ell_d}}^{\text{entry}} \overbrace{f(d_{t;i}|\underbrace{d_{t;i+1}, \ldots, d_{t;\ell_d}}_{a_{ti}}, d^{t-1}, \Theta_i)}^{i\text{th parametric model}}, \quad \Theta \equiv \{\Theta_i\}_{i=1}^{\ell_d} \tag{1}$$

The ith parametric model in (1) is the central modelling element. It predicts ith entry $d_{t;i}$ of the data record d_t at time t, using the vector $a_{t;i}$ (containing the entries $d_{t;i+1}, \ldots, d_{t;\ell_d}$ of d_t), the

DOI: 10.1002/acs

past data records $d^{t-1}=(d_1,\ldots,d_{t-1})$ and a finite-dimensional parameter Θ_i. The factorization of the joint pdf even over entries of d_t shows that parametric models of scalar entries of d can be considered only. If, moreover, the finite-dimensional parameters Θ_i,Θ_j are *a priori* independent they stay independent even *a posteriori*. It is obvious from the above product form and from the Bayes rule [2]. We *restrict* ourselves to this usual case. Consequently, we can model and estimate respective entries of d_t individually and independently. Taking this into account, we consider a fixed i and drop this subscript hereafter.

Within the control context, the extent of the observed data records is permanently increasing. This delimits members of the EF as the predominant models.

Definition 1 (*EF and CEF*)
The parametric model belongs to the dynamic EF iff

$$f(d_t|a_t,d^{t-1},\Theta)=f(d_t|\psi_t,\Theta)=\exp\langle \mathsf{B}(\Psi_t),\mathsf{C}(\Theta)\rangle \tag{2}$$

where $\Psi_t\equiv[d_t,\psi_t']'$ is the *data vector*, given by a finite-dimensional *regression vector* ψ_t, depending on a_t and d^{t-1}; $'$ denotes transposition. Data vectors, Ψ_{t-1}, can be recursively updated to Ψ_t using the pair d_t,a_t only; $\langle\cdot,\cdot\rangle$ is a functional, which is linear in its respective arguments; $\mathsf{B}(\cdot),\mathsf{C}(\cdot)$ are the functions of compatible finite dimensions. They are defined on Ψ_t^* and Θ^*, respectively. The set CEF contains pdfs on Θ^* conjugated to the pdfs in the EF [1], i.e. the pdfs having the form

$$f(\Theta|\mathscr{V})=\frac{\exp\langle\mathscr{V},\mathsf{C}(\Theta)\rangle\chi_{\Theta^*}(\Theta)}{\int\exp\langle\mathscr{V},\mathsf{C}(\Theta)\rangle\chi_{\Theta^*}(\Theta)\,\mathrm{d}\Theta}\equiv\frac{\mathscr{G}_\Theta(\mathscr{V})}{\mathsf{I}(\mathscr{V})} \tag{3}$$

For the given $\langle\cdot,\cdot\rangle$ and $\mathsf{C}(\Theta)$, a specific member of the CEF is determined by the (value of) finite-dimensional statistic $\mathscr{V}$ and by the indicator $\chi_{\Theta^*}(\Theta)$ of the set Θ^*.

Note that the *dynamic* EF differs from the standard definition EF. It models *dependence* of the data record d_t on the data compressed in the regression vector ψ_t. Moreover, recursive updating of the data vector Ψ_t is required. The first condition is inevitable in modelling of dynamic controlled systems and excludes, for instance, data having conditional exponential pdf. The second condition is needed for permanent online use and excludes, for instance, linear normal models with moving average noise.

The estimation simplicity determines practical significance of the EF and CEF [2].

Proposition 1 (*Estimation in the EF*)
Let the parametric model have the form (2). Let the prior pdf from the CEF (3), given by value of the statistic $\mathscr{V}_0$, be used. Then, the posterior pdf $f(\Theta|d^t)$ is in the CEF

$$\begin{aligned} f(\Theta|d^t)&=f(\Theta|\mathscr{V}_t)=\frac{\exp\langle\mathscr{V}_t,\mathsf{C}(\Theta)\rangle\chi_{\Theta^*}(\Theta)}{\mathsf{I}(\mathscr{V}_t)}=\frac{\mathscr{G}_\Theta(\mathscr{V}_t)}{\mathsf{I}(\mathscr{V}_t)}\\ \mathsf{I}(\mathscr{V}_t)&\equiv\int\exp\langle\mathscr{V}_t,\mathsf{C}(\Theta)\rangle\chi_{\Theta^*}(\Theta)\,\mathrm{d}\Theta,\quad t\in t^* \end{aligned} \tag{4}$$

The value of the statistic $\mathscr{V}_t$ updates recursively, $\mathscr{V}_t=\mathscr{V}_{t-1}+\mathsf{B}(\Psi_t)$, with $\mathscr{V}_0$ *a priori* chosen. The predictive pdf is $f(d_t|a_t,d^{t-1})=\mathsf{I}(\mathscr{V}_{t-1}+\mathsf{B}(\Psi_t))/\mathsf{I}(\mathscr{V}_{t-1})$.

In summary, if: (i) each parametric model in (1) belongs to the dynamic EF (2); (ii) unknown parameters of respective models are *a priori* mutually independent; (iii) the respective prior pdfs

DOI: 10.1002/acs

are chosen in the CEF (3); then, the estimation and prediction of respective entries of the data record reduce to algebraic updating of data vectors Ψ and of the finite-dimensional sufficient statistics $\mathscr{V}$.

2.3. Normal ARX model and its estimation

The normal autoregressive model with exogenous inputs in regression vector ψ (ARX) $f(d|\psi,\Theta)=\mathscr{N}_d(\theta'\psi,r)$, modelling scalar data item d, belongs to the EF with

$$\langle \mathsf{B}(\Psi),\mathsf{C}(\Theta)\rangle=-\frac{1}{2}\ln(r)+-\frac{1}{2}\mathrm{tr}\left[\Psi\Psi'\frac{[-1,\theta']'[-1,\theta']}{r}\right]$$

The unknown parameter Θ consists of the regression coefficients θ and variance r. The correspondence with (2) determines the conjugate prior pdf (3) as the Gauss-inverse-Wishart (*GiW*) pdf [12, 13], for which the function $\mathscr{G}_\Theta(\mathscr{V})$ (3) reads

$$\mathscr{G}_\Theta(\mathscr{V})=\frac{\exp\left\{-\dfrac{\mathrm{tr}[V[-1,\theta']'[-1,\theta']]}{2r}\right\}}{r^{0.5(v+\ell_\psi+2)}},\quad \mathscr{V}\equiv(V,v) \tag{5}$$

The scalar v has the meaning of the number of degrees of freedom. The (ℓ_Ψ,ℓ_Ψ)-dimensional extended information matrix V must be positive definite. Otherwise, the function $\mathscr{G}_\Theta(\mathscr{V})$ cannot be normalized to a pdf.

The matrix V can be expressed in a range of equivalent ways. Further on, it is represented in terms of the well-known least-squares (LS) quantities: LS estimate $\hat{\theta}$ of θ, LS estimate ω of the noise precision r^{-1} and LS covariance factor matrix P. The expectation $\mathsf{E}[\cdot]$, variance $\mathsf{var}[\cdot]$ and covariance $\mathsf{cov}[\cdot]$ of quantities distributed according to the *GiW* pdf, needed later on, are related to the LS representation as follows [12]:

$$\mathsf{E}[\theta|\mathscr{V}]=\hat{\theta},\quad \mathsf{E}[r|\mathscr{V}]=\frac{v}{\omega(v-2)},\quad \mathsf{E}\left[\frac{1}{r}\middle|\mathscr{V}\right]=\omega$$

$$\mathsf{cov}(\theta|\mathscr{V})=\frac{v}{\omega(v-2)}P,\quad \mathsf{var}(r^{-1}|\mathscr{V})=\frac{2}{v}\omega^2,\quad \mathsf{E}[\ln(r)|\mathscr{V}]=\ln\left(\frac{v}{2\omega}\right)-\Gamma\left[\frac{v}{2}\right]$$

where the used digamma function is defined

$$\Gamma[x]\equiv\frac{\partial\ln(\int_0^\infty z^{x-1}\exp(-z)\,\mathrm{d}z)}{\partial x}\quad \text{for } x>0$$

Proposition 1 reduces estimation to the updating $\mathscr{V}_t=\mathscr{V}_{t-1}+\mathsf{B}(\Psi_t)$, i.e. to updating of the extended information matrix V_t and the number of degrees of freedom v_t

$$V_t=V_{t-1}+\Psi_t\Psi_t',\quad v_t=v_{t-1}+1 \tag{6}$$

The prior *GiW* pdf, determined by V_0, v_0, initializes (6). The recursion (6) expressed in terms of the LS representation coincides with the recursive LS [2].

DOI: 10.1002/acs

3. APPROXIMATION OF PDF WITH THE KLD AS A LOSS

We consider that parameters $\Theta \in \Theta^*$ are described by a 'true' pdf $\mathcal{T}$ and we search for its approximation $\hat{f} \in \hat{f}^*$. The choice of $\hat{f}$ is the decision to be made. Consistent Bayesian formulation specifies a loss functional $Z(\hat{f}(\cdot), \Theta)$ and minimizes the expected loss $\mathsf{E}[Z] \equiv \int \mathcal{T}(\Theta) Z(\hat{f}(\cdot), \Theta)\, \mathrm{d}\Theta$. The loss Z serves properly if its unrestricted minimum is finite and it is attained for $\hat{f} = \mathcal{T}$, $\mathcal{T}$-almost surely. We require the loss to depend only on the realized value $\hat{f}(\Theta)$, i.e.

$$Z(\hat{f}(\cdot), \Theta) = Z(\hat{f}(\Theta), \Theta) \tag{7}$$

This requirement, representing a sort of likelihood principle, was advocated by Bernardo in [4] where the following proposition is proved. The basic properties of the KLD we need [14] are attached to it.

Proposition 2 (*Recommended loss*)
Let the loss Z meet the assumption (7) and have continuous partial derivative with respect to the first argument. Then, $Z(\hat{f}(\Theta), \Theta) = D \ln(1/\hat{f}(\Theta)) + E(\Theta)$ with a constant $D > 0$ and an arbitrary function $E(\Theta)$ for which $\int \mathcal{T}(\Theta) E(\Theta)\, \mathrm{d}\Theta$ is finite. The specific choice of D and $E(\Theta)$ does not influence the minimizer found. For $D = 1$ and $E(\Theta) = \ln(\mathcal{T}(\Theta))$, the expected loss is the KLD $\mathsf{D}(\mathcal{T} || \hat{f})$ of $\mathcal{T}$ on $\hat{f}$

$$\mathsf{D}(\mathcal{T} || \hat{f}) \equiv \int \mathcal{T}(\Theta) \ln\left(\frac{\mathcal{T}(\Theta)}{\hat{f}(\Theta)}\right) \mathrm{d}\Theta, \quad \mathsf{D}(\mathcal{T} || \hat{f}) \geqslant 0 \quad \text{and} \quad \mathsf{D}(\mathcal{T} || \hat{f}) = 0 \text{ iff } \mathcal{T} = \hat{f} \tag{8}$$

The last identity is understood $\mathcal{T}$-almost surely.

The choice $D = 1$ and $E(\Theta) = 0$ makes the loss equal to the Kerridge inaccuracy $\mathcal{K}(\mathcal{T} || \hat{f}) \equiv \int T)(\Theta) \ln(1/\hat{f}(\Theta))\, \mathrm{d}\Theta$ [15] that has the same minimizer $\hat{f}$ as the KLD.

Further on, the approximated pdf $\mathcal{T}$, defined on Θ^*, is assumed to be random with a finite number of variants, labelled by $\imath \in \imath^*$

$$\mathcal{T} = f_\imath \quad \text{with probability } \lambda_\imath \in (0, 1), \quad \sum_{\imath \in \imath^*} \lambda_\imath = 1 \tag{9}$$

We inspect the best *unrestricted* approximation of the pdf $\mathcal{T}$ for both orders of the KLD arguments.

Proposition 3 (*Approximation of the random* $\mathcal{T}$)
The unrestricted minimizer $\hat{f}$ of the expected KLD (with expectation taken over the random $\mathcal{T}$)

$$\mathsf{E}[\mathsf{D}(\mathcal{T} || f)] \equiv \mathsf{E}[\mathsf{E}[\mathsf{D}(\mathcal{T} || f) | \mathcal{T} = f_\imath]] = \mathsf{E}[\mathsf{D}(f_\imath || f)] = \sum_{\imath \in \imath^*} \lambda_\imath \mathsf{E}[\mathsf{D}(f_\imath || f)]$$

over the pdfs $f \in f^*$ on Θ^* is the mixture

$$\hat{f}(\Theta) \equiv \sum_{\imath \in \imath^*} \lambda_\imath f_\imath(\Theta) \tag{10}$$

The minimizer $\hat{f}$ of $\mathsf{E}[\mathsf{D}(f || \mathcal{T})]$ is the geometric mean

$$\hat{f}(\Theta) = \frac{\prod_{\imath \in \imath^*} [f_\imath(\Theta)]^{\lambda_\imath}}{\int \prod_{\imath \in \imath^*} [f_\imath(\Theta)]^{\lambda_\imath}\, \mathrm{d}\Theta} \tag{11}$$

DOI: 10.1002/acs

If the pdfs, $f_\iota(\Theta)$, belong to the CEF with the common support Θ^* and common $\mathsf{C}(\Theta)$, i.e. $f_\iota(\Theta)=\exp\langle\psi_\iota,\mathsf{C}(\Theta)\rangle\chi_{\Theta^*}(\Theta)/\mathsf{I}(\psi_\iota)$, then the geometric mean (11) stays within the CEF with the statistic value $\hat{\psi}\equiv\sum_{\iota\in\iota^*}\lambda_\iota\psi_\iota$.

Proof
The first statement is implied by the identities

$$\mathsf{E}[\mathsf{D}(\mathscr{T}||f)|f]=\sum_{\iota\in\iota^*}\lambda_\iota\int f_\iota(\Theta)\ln\left(\frac{f_\iota}{f(\Theta)}\right)\mathrm{d}\Theta=\int\sum_{\iota\in\iota^*}\lambda_\iota f_\iota(\Theta)\ln\left(\frac{\sum_{\iota\in\iota^*}\lambda_\iota f_\iota(\Theta)}{f(\Theta)}\right)\mathrm{d}\Theta$$
$$-\int\sum_{\tilde{\iota}\in\tilde{\iota}^*}\lambda_{\tilde{\iota}}f_{\tilde{\iota}}(\Theta)\ln\left(\frac{\sum_{\iota\in\iota^*}\lambda_\iota f_\iota(\Theta)}{f_{\tilde{\iota}}(\Theta)}\right)\mathrm{d}\Theta$$

Only the first term depends on the optimized pdf $f(\Theta)$ and reaches the smallest value (8) for the claimed mixture. The second statement can be demonstrated in the same way and the last claim of the proposition is obvious. □

Except in the cases with discrete-valued Θ, the mixture (10) of the pdfs in the CEF is out of the CEF. Thus, the unrestricted minimizer of the expected KLD with the 'correct' order of arguments lies out of the CEF. The unrestricted minimizer of the expected KLD with the reversed order may stay within the CEF. Should we use it as the approximation of the posterior pdf within the CEF? To answer this question, let us inspect the expected KLD with the 'correct' order of arguments as a function of ψ determining the pdf $f(\Theta|\psi)$ in the CEF (3). The ψ-dependent term of the expected KLD of the pdf $\mathscr{T}$ on the pdf $f(\Theta|\psi)$ is the expected Kerridge inaccuracy, cf. (4)

$$\mathsf{E}[\mathscr{K}(\mathscr{T}(\Theta)||f(\Theta|\psi))]=-\sum_{\iota\in\iota^*}\lambda_\iota\int f_\iota(\Theta)\langle\psi,\mathsf{C}(\Theta)\rangle\,\mathrm{d}\Theta+\ln[\mathsf{I}(\psi)]\tag{12}$$

The first term in (12) is linear and thus convex function of the optional value ψ. Let us inspect the second term in (12). For notational simplicity, let ψ and C be vectors and $\langle\psi,\mathsf{C}\rangle=\psi'\mathsf{C}$. Second derivative of the inspected second term with respect to ψ gives the Hessian

$$\frac{\partial^2}{(\partial\psi)^2}\ln[\mathsf{I}(\psi)]=\int\mathsf{C}(\Theta)\mathsf{C}'(\Theta)f(\Theta|\psi)\,\mathrm{d}\Theta$$
$$-\int\mathsf{C}(\Theta)f(\Theta|\psi)\,\mathrm{d}\Theta\left[\int\mathsf{C}(\Theta)f(\Theta|\psi)\,\mathrm{d}\Theta\right]'\tag{13}$$

Thus, the Hessian (13) is the covariance matrix of the vector $\mathsf{C}(\Theta)$ with respect to the pdf $f(\Theta|\psi)$. As such, it is positive semidefinite. It is positive definite in generic case. By *generic case*, we mean that sets $\Theta^*_v\equiv\{\Theta;\,v'\mathsf{C}(\Theta)=0\}$, determined by non-zero vectors v, have zero measure with respect to the inspected pdfs $f(\Theta|\psi)$.

Proposition 4 (*Proper approximation of* $\mathscr{T}$)
In the generic case, the geometric mean (11) does not minimize the 'correct' expected KLD of the random pdf $\mathscr{T}(\Theta)$ (9) on pdfs from the CEF.

Proof
The KLD is a non-negative and convex function of statistic values ψ and thus has a unique minimizer. □

DOI: 10.1002/acs

This simple result complements the paper [6], where the forgetting resulting from alternative orders of the KLD arguments was studied. We strongly recommend to focus on the version implied by the 'correct' order. More generally, it *conceptually* 'undermines' popular geometric pooling of pdfs [16] and the whole stream of algorithms based on functional approximation [9, 17].

The remainder of the paper demonstrates tracking of parameters of the normal ARX model that performance of estimation algorithms can be *practically* improved by respecting this result. Performance gains in the approximate mixture estimation are documented in [11].

4. CONSEQUENCES FOR TRACKING VIA FORGETTING

Forgetting is widely used for tracking of slowly varying parameters. We inspect the *stabilized forgetting* [6, 7] with the parametric model in the EF and the posterior pdf $f_\beta(\Theta|d^t)=f_\beta(\Theta|\mathcal{V}_{\beta t})$ in the CEF (3).

For the stabilized forgetting, an *alternative pdf* $f_\alpha(\Theta|d^t)=f_\alpha(\Theta|\mathcal{V}_{\alpha t})$ in the CEF is specified. It describes uncertainty about parameters caused by their variations within the real-time interval when no data are processed. The probability that the 'true' pdf $\mathcal{T}$ is f_β, i.e. parameters do not change, is set to $\lambda_t\in(0,1)$. The probability that the unknown parameters Θ are described by the alternative pdf f_α is $1-\lambda_t$.

We search for the pdf $\hat{f}(\Theta|d^t)=\hat{f}(\Theta|\mathcal{V}_t)$ *within the CEF* minimizing $\mathsf{E}[\mathsf{D}(\mathcal{T}||\hat{f})]$. Dropping the subscript t and the condition d^t, we search for $\mathcal{V}$ minimizing its $\mathcal{V}$-dependent part $\ln(\mathsf{I}(\mathcal{V}))-\langle\mathcal{V},\int\lambda f_\beta(\Theta)+(1-\lambda)f_\alpha(\Theta)\mathsf{C}(\Theta)\rangle$. The corresponding $\mathcal{V}$ has to solve the equation, see (3)

$$\int\mathsf{C}(\Theta)\frac{\mathcal{G}_\Theta(\mathcal{V})}{\mathsf{I}(\mathcal{V})}\,\mathrm{d}\Theta=\lambda\int\mathsf{C}(\Theta)\frac{\mathcal{G}_\Theta(\mathcal{V}_\beta)}{\mathsf{I}(\mathcal{V}_\beta)}\,\mathrm{d}\Theta+(1-\lambda)\int\mathsf{C}(\Theta)\frac{\mathcal{G}_\Theta(\mathcal{V}_\alpha)}{\mathsf{I}(\mathcal{V}_\alpha)}\,\mathrm{d}\Theta \tag{14}$$

Usefulness of the constructed $\hat{f}$ depends strongly on the possibility to evaluate the expectation of $\mathsf{C}(\Theta)$ as a function of the statistic with values $\mathcal{V}_\beta,\mathcal{V}_\alpha,\mathcal{V}$. Properties of the *GiW* pdf (6), conjugated to the inspected normal ARX model, provide the needed expectation.

Proposition 5 (*Expectation of* $\mathsf{C}(\Theta)$ *for GiW pdf*)
Let $\Theta\equiv[\theta,r]$ be described by the pdf $GiW_\Theta(\mathcal{V})=GiW_{[\theta,r]}(\hat{\theta},\omega,P,\nu)$, (5), (6). Then,

$$\begin{aligned}\mathsf{E}[\mathsf{C}(\Theta)|\mathcal{V}]&=\int\mathsf{C}(\Theta)\frac{\mathcal{G}_\Theta(\mathcal{V})}{\mathsf{I}(\mathcal{V})}\,\mathrm{d}\Theta=-\frac{1}{2}\int\begin{bmatrix}\ln(r)&0\\0&\dfrac{1}{r}\begin{bmatrix}-1\\\theta\end{bmatrix}\begin{bmatrix}-1\\\theta\end{bmatrix}'\end{bmatrix}GiW_\Theta(\mathcal{V})\,\mathrm{d}\theta\,\mathrm{d}r\\&=-\frac{1}{2}\begin{bmatrix}\ln\left(\dfrac{\nu}{2\omega}\right)-\Gamma\left[\dfrac{\nu}{2}\right]&0\\0&\omega\begin{bmatrix}-1\\\hat{\theta}\end{bmatrix}\begin{bmatrix}-1\\\hat{\theta}\end{bmatrix}'+\begin{bmatrix}0&0\\0&P\end{bmatrix}\end{bmatrix}\end{aligned} \tag{15}$$

Inserting the formula (15) into (14), we get the algorithm mapping $\mathcal{V}_\beta$ and $\mathcal{V}_\alpha$ on the optimally approximating $\mathcal{V}$.

DOI: 10.1002/acs

Algorithm 1 (*Alternative stabilized forgetting*)

Inputs:

- The forgetting factor $\lambda \in (0, 1)$ coinciding with the probability of the basic no-parameter-change hypothesis.
- The statistic value $\mathscr{V}_\beta \equiv (\hat{\theta}_\beta, \omega_\beta, P_\beta, \nu_\beta)$ describing the member of the CEF corresponding to the basic hypothesis that parameters do not change.
- The statistic value $\mathscr{V}_\alpha \equiv (\hat{\theta}_\alpha, \omega_\alpha, P_\alpha, \nu_\alpha)$ describing the alternative pdf in the CEF.

Outputs:

- The value $\mathscr{V} \equiv (\hat{\theta}, \omega, P, \nu)$ describing the member of the CEF minimizing the 'correct' expected KLD.

Evaluations:

1. $\omega = \lambda\omega_\beta + (1-\lambda)\omega_\alpha$.
2. $\hat{\theta} = (\lambda\omega_\beta/\omega)\hat{\theta}_\beta + (1-\lambda\omega_\beta/\omega)\hat{\theta}_\alpha$.
3. $P = \lambda P_\beta + (1-\lambda)P_\alpha + (\lambda(1-\lambda)\omega_\beta\omega_\alpha/\omega)(\hat{\theta}_\alpha - \hat{\theta}_\beta)(\hat{\theta}_\alpha - \hat{\theta}_\beta)'$.
4. $\Upsilon \equiv \lambda[\ln(\nu_\beta/\omega_\beta) - \Gamma[\nu_\beta/2]] + (1-\lambda)[\ln(\nu_\alpha/\omega_\alpha) - \Gamma[\nu_\alpha/2]] + \ln(\omega/2)$.
5. Find ν solving the equation $\ln(\nu/2) - \Gamma[\nu/2] = \Upsilon$.

Step 3 is numerically the most sensitive and demanding. The only nonlinear equation for the scalar ν, Step 5, is efficiently solvable by the standard Newton method with a good initialization resulting from a simple approximation of the digamma function $\Gamma[\cdot]$, see [18].

5. CHOICE OF THE FORGETTING CHARACTERISTICS

Algorithm 1 finds the optimal approximation of the forgetting result in the CEF. Usefulness and quality of this *alternative stabilized forgetting* (ASF) strongly depends on the quality of the approximated pdf, which should be indeed close to the 'true' pdf $\mathscr{T}$. This condition has to be achieved by a good choice of the optional inputs determining the optimal unrestricted forgetting. The discussed meaningful choice is applicable also to the *standard stabilized forgetting* (SSF) resulting as the geometric mean (11) in Proposition 3.

The pdf f_β given by $\mathscr{V}_\beta \equiv (\hat{\theta}_\beta, \omega_\beta, P_\beta, \nu_\beta)$ corresponds to the basic hypothesis that parameters are time invariant. Therefore, it has to coincide with the newest result of the parameter estimation. Thus, it remains to choose the forgetting factor $\lambda \in (0, 1)$ and the alternative pdf f_α determined by the value $\mathscr{V}_\alpha \equiv (\hat{\theta}_\alpha, \omega_\alpha, P_\alpha, \nu_\alpha)$.

5.1. Choice of the forgetting factor

For a chosen alternative pdf f_α, the forgetting factor completes the definition of the pdf

$$f(\Theta_{t+1} = \Theta | d^t) = \mathscr{T}(\Theta) \equiv \begin{cases} f_\beta(\Theta) & \text{with probability } \lambda \\ f_\alpha(\Theta) & \text{with probability } 1-\lambda \end{cases} \tag{16}$$

DOI: 10.1002/acs

We construct the alternative f_α expecting that f_α will be closer to the 'true' pdf $\mathscr{T}$ than f_β. According to Proposition 2, it means

$$\mathsf{E}[\mathsf{D}(\mathscr{T}||f_\alpha)|f_\alpha, f_\beta, \lambda] \leqslant \mathsf{E}[\mathsf{D}(\mathscr{T}||f_\beta)|f_\alpha, f_\beta, \lambda] \tag{17}$$

where expectation is taken over random 'true' pdfs (16). The left-hand side of (17) equals $\lambda\mathsf{D}(f_\beta||f_\alpha)$ and the right-hand side is $(1-\lambda)\mathsf{D}(f_\alpha||f_\beta)$. Consequently, forgetting factors meeting our expectation on the divergence $\mathsf{D}(\mathscr{T}||f_\alpha)$ are in the set

$$0 \leqslant \lambda \leqslant \frac{\mathsf{D}(f_\alpha||f_\beta)}{\mathsf{D}(f_\alpha||f_\beta)+\mathsf{D}(f_\beta||f_\alpha)} \equiv \overline{\lambda} \leqslant 1 \tag{18}$$

At the same time, λ expresses the expectation that parameters do not change. Tracking by forgetting can be successful only if this hypothesis is often acceptable, i.e. if λ is high enough. This leads us to the choice $\lambda = \overline{\lambda}$.

5.2. *Choice of the alternative* $\mathscr{V}_\alpha$

The prior pdf $f_0 \equiv f(\Theta|\mathscr{V}_0)$ should quantify all prior knowledge [19]. It delimits primarily the domain where the estimated Θ lies with a high probability. The alternative pdf will *stabilize* tracking if it respects prior knowledge, i.e. if its tails are less heavy than that of the prior pdf. For *GiW* pdfs, this condition can be met by taking $\mathscr{V}_\alpha \equiv \mathscr{V}_0$. Experience indicates that this choice can be over-conservative and that positioning of the alternative pdf on the latest point estimates of Θ with properly tuned covariance matrix [20] provides better results. It leads to the conjecture that exploitation of both prior pdf f_0 and the basic pdf f_β for constructing the alternative pdf f_α can be *universally beneficial*. The straightforward algorithm presented below rests on this conjecture. It takes the approximation of the 'true' pdf $\mathscr{T}$, generated by Algorithm 1, as an improved guess of the alternative pdf generating an improved guess of the 'true' pdf. It stops if f_α is close to f_β but still differs from it.

Algorithm 2 (*ASF with automatic choice of* λ, f_α)

Inputs:

- The statistic value $\mathscr{V}_0 \equiv (\hat{\theta}_0, \omega_0, P_0, \nu_0)$ describing *a priori* the parameters Θ by the member of the CEF.
- The statistic value $\mathscr{V}_\beta \equiv (\hat{\theta}_\beta, \omega_\beta, P_\beta, \nu_\beta)$ describing parameters Θ by the member of the CEF corresponding to the basic hypothesis that parameters do not change.
- The bound $\varepsilon > 0$ determining negligibility of diminishing relative differences of forgetting factors $\lambda_{i-1}/\lambda_i - 1$.

Outputs:

- The value $\mathscr{V} \equiv (\hat{\theta}, \omega, P, \nu)$ describing the member of the CEF minimizing the 'correct' expected KLD.

Evaluations: Set iteration counter $i = 1$ and initialize the guess of the alternative pdf $f_{i\alpha} = f_0 \Leftrightarrow \mathscr{V}_{i\alpha} = \mathscr{V}_0$. Set $\Delta = 1 + \varepsilon$, $\lambda_0 = 1$, do while $\Delta > \varepsilon$.

DOI: 10.1002/acs

1. Evaluate ith guess λ_i of the forgetting factor

$$\lambda_i = \frac{\mathsf{D}(f_{i\alpha}||f_\beta)}{\mathsf{D}(f_{i\alpha}||f_\beta)+\mathsf{D}(f_\beta||f_{i\alpha})} \tag{19}$$

2. Use Algorithm 1 with inputs $\mathscr{V}_\beta$, $\mathscr{V}_\alpha = \mathscr{V}_{i\alpha}$ and $\lambda = \lambda_i$ for evaluating an improved approximation f_i of the 'true' pdf $\mathscr{T}$. The approximation f_i, found in the CEF, is given by the statistic value $\mathscr{V}_i$.
3. Take f_i as a new guess of the alternative pdf $f_{(i+1)\alpha} = f_i \Leftrightarrow \mathscr{V}_{(i+1)\alpha} \equiv \mathscr{V}_i$.
4. Set $\Delta \equiv |\lambda_{i-1}/\lambda_i - 1|$. Increment i.

Proposition 6 (*Properties of Algorithm 2*)
Let $\mathsf{D}(f_\beta||f_{0\alpha}) \equiv \mathsf{D}(f_\beta||f_0)$ be positive and finite. Then, $\mathsf{D}(f_\beta||f_{i\alpha}) \geqslant \mathsf{D}(f_\beta||f_{(i+1)\alpha})$ and $f_{i\alpha} \to f_{\infty\alpha}$.

Proof
$f_{(i+1)\alpha}$ minimizes the expected KLD of the previous guess f_i of the 'true' pdf $\mathscr{T}$ on pdfs from the CEF

$$0 \leqslant \mathsf{E}[\mathsf{D}(f_i||f_{(i+1)\alpha})|f_{i\alpha}, f_\beta, \lambda_i] \equiv \lambda_i \mathsf{D}(f_\beta||f_{(i+1)\alpha}) + (1-\lambda_i)\mathsf{D}(f_{i\alpha}||f_{(i+1)\alpha})$$

Replacement of $f_{(i+1)\alpha}$ by $f_{i\alpha}$ and non-negativity of the KLD implies that $D_{i+1} \equiv \mathsf{D}(f_\beta||f_{(i+1)\alpha}) \leqslant \mathsf{D}(f_\beta||f_{i\alpha}) \equiv D_i$. Due to non-negativity and initial boundedness, it has a finite limit D_∞ for $i \to \infty$. The same replacement implies that $\mathsf{D}(f_{i\alpha}||f_{(i+1)\alpha}) \to 0$. The properties of the KLD imply the assertion. □

Additional analysis and experiments indicate that $f_{\infty\alpha} = f_\beta$ for f_β, f_α in the CEF. Irrespective of this $\lambda_i \to 0.5$ (19). This made us to use its changes in the proposed stopping rule. It also hints that $\varepsilon \approx 0.05 = 10\%$ of the limit as a reasonable option. There are also indications that the found approximation has lighter tails than the prior pdf. Full proof is missing.

6. EXAMPLES

Theoretical analysis of Algorithm 2 is incomplete. Experimental results indicate its good properties. The examples presented here illustrate them. Prediction ($\mathscr{Q} = \mathscr{P}$) and estimation ($\mathscr{Q} = \mathscr{E}$) quality is quantified by

$$\mathscr{Q}^{t_0:t} \equiv \frac{{}^{\mathrm{SSF}}\mathrm{MSE}^{t_0:t}}{{}^{\mathrm{ASF}}\mathrm{MSE}^{t_0:t}} - 1, \quad \mathrm{MSE}^{t_0:t} \equiv \begin{cases} \sqrt{\dfrac{1}{(t-t_0+1)}\sum_{\tau=t_0}^{t}(y_\tau - \hat{\theta}'_{\tau-1}\psi_\tau)^2}, & \mathscr{Q} = \mathscr{P} \\ \sqrt{\dfrac{1}{(t-t_0+1)}\sum_{\tau=t_0}^{t}(\hat{\theta}_\tau - \theta_\tau)'(\hat{\theta}_\tau - \theta_\tau)}, & \mathscr{Q} = \mathscr{E} \end{cases} \tag{20}$$

The optional initial time t_0 allows us to judge the influence of transients. The superscripts SSF and ASF distinguish the forgetting method used. The relative norm $\mathscr{Q}$ is positive if the ASF outperforms the SSF. Whenever possible, also the ${}^{\mathrm{SSF}}\mathrm{MSE}$ and ${}^{\mathrm{ASF}}\mathrm{MSE}$ will be displayed together with the quality indicator.

DOI: 10.1002/acs

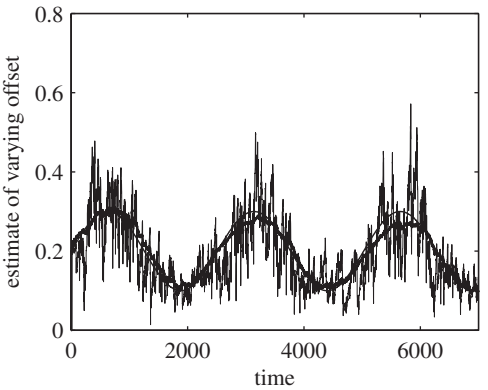

Figure 1. Estimates of time-varying offset. The sinus curve is its true value. The thin curve is its point estimate with the SSF, the thick one with the ASF.

6.1. *Tracking of varying offset*

Third-order, single-output normal AR model with varying offset illustrates tracking of slow changes. The simulated pdf $f(y_t|y^{t-1}, \Theta) = \mathcal{N}_t(\mu_t, 0.001)$ had mean

$$\mu_t = \underbrace{[\overbrace{-0.5, 0.2, 0.4}^{c_{\theta'}}, 0.2 + 0.1\sin(t/400)]}_{\theta_t'} \underbrace{[y_{t-1}, y_{t-2}, y_{t-3}, 1]'}_{\psi_t}$$

7000 data records were generated and Bayesian updating, recursion (6), was applied. It was combined with the SSF and with the ASF. Both algorithms used the default prior distribution f_0 determined by

$$\mathcal{V}_0 \equiv (\hat{\theta}_0 = 0,\ P_0 = 10^6 \times \text{unit matrix},\ \omega_0 = 3 \times 10^4,\ \nu_0 = 3) \tag{21}$$

The SSF took f_0 (21) as an alternative together with the best $\lambda = 0.95$. Algorithm 2 ran with this f_0 and $\varepsilon = 0.05$.

The qualitative behavior is illustrated by Figure 1 where tracking of the time-varying offset by the compared methods is displayed. Quantitative indicators characterizing improved tracking and predictive capabilities of the ASF are in harmony with the qualitative result. The final point estimates of the time-invariant part of parameters ${}^{C}\theta = [-0.5, 0.2, 0.4]$ are ${}^{\text{ASF}}\hat{\theta} = [-0.485, 0.259, 0.379]$ and ${}^{\text{SSF}}\hat{\theta} = [-0.619, 0.184, 0.362]$, respectively. The corresponding relative prediction and estimation norms and mean square errors (20) are

$$\mathcal{P}^{1:7000} = 0.027\ ({}^{\text{SSF}}\text{MSE}^{1:7000} = 2.777,\ {}^{\text{ASF}}\text{MSE}^{1:7000} = 2.703)$$

$$\mathcal{E}^{1:7000} = 7.512\ ({}^{\text{SSF}}\text{MSE}^{1:7000} = 0.00243,\ {}^{\text{ASF}}\text{MSE}^{1:7000} = 0.000571)$$

DOI: 10.1002/acs

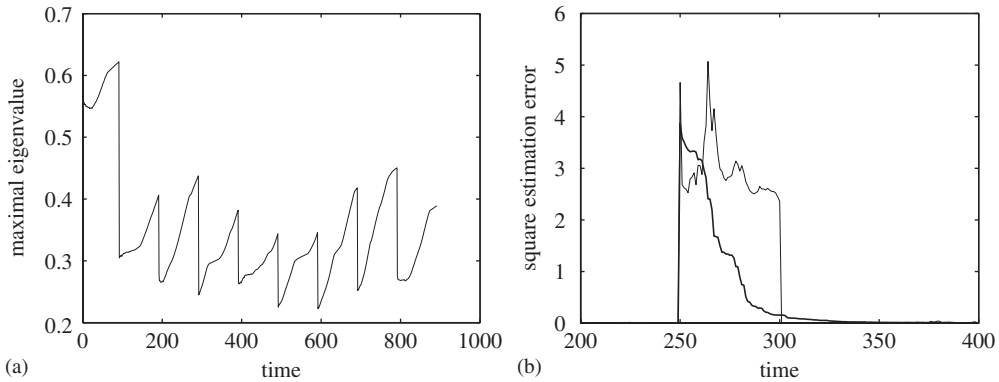

Figure 2. (a) Bounded eigenvalues of P_t demonstrate the stabilizing effect of the ASF for non-informative data and (b) convergence of the estimation square error demonstrates estimation quality of the ASF (the thick one).

6.2. *Stabilization benchmark*

This example comes from [21], where *stabilization* was addressed: bad behavior of exponential forgetting caused by non-informative data was counteracted. The simulated pdf $f(y_t|u^t, y^{t-1}, v^{t-1}, \Theta) = \mathcal{N}_{y_t}(\mu_t, 0.01)$ had mean

$$\mu_t = \theta' \times \psi_t \equiv [0.98, -0.9, 0.5, -0.25, 0.1, 0.8, 0.2][y_{t-1}, y_{t-2}, u_t, u_{t-1}, u_{t-2}, v_{t-1}, v_{t-2}]'$$

It was stimulated by white input $u_t \sim \mathcal{N}_{u_t}(0, 1)$ and by the measurable disturbance v_t alternating its values in $\{1, -1\}$ at time moments $t = 100, 200, \ldots, 1000$. At time $t = 250$, the parameter θ_1 changed to -0.98. The SSF took f_0 (21) as an alternative together with the best $\lambda = 0.9$. Algorithm 2 ran with this f_0 and the recommended $\varepsilon = 0.05$.

Figure 2(a) illustrates that the ASF, similarly to the SSF, has stabilizing effect and keeps the maximal eigenvalue of LS covariance factor matrix P_t well bounded.

The performance indices (20)

$$\mathscr{P}^{1:1000} = -0.207 \ (^{\mathrm{SSF}}\mathrm{MSE}^{1:1000} = 8.193, \ ^{\mathrm{ASF}}\mathrm{MSE}^{1:1000} = 10.331)$$

$$\mathscr{P}^{501:1000} = 0.087 \ (^{\mathrm{SSF}}\mathrm{MSE}^{501:1000} = 2.526, \ ^{\mathrm{ASF}}\mathrm{MSE}^{501:1000} = 2.325)$$

$$\mathscr{E}^{1:1000} = 0.673 \ (^{\mathrm{SSF}}\mathrm{MSE}^{1:1000} = 0.185, \ ^{\mathrm{ASF}}\mathrm{MSE}^{1:1000} = 0.110) \quad \text{and}$$

$$\mathscr{E}^{501:1000} = 3.621 \ (^{\mathrm{SSF}}\mathrm{MSE}^{501:1000} = 0.00666, \ ^{\mathrm{ASF}}\mathrm{MSE}^{501:1000} = 0.00144)$$

show that (i) the SSF predicted better in transients but the ASF outperformed it in a longer run; (ii) the estimation by the ASF was better than by the SSF. See Figure 2(b) for convergence of square error in the most important part of time sequence.

6.3. *Tests on real financial data*

This test supports our claim that the use of the ASF with judiciously chosen forgetting factor and a good alternative pdf is practically beneficial.

DOI: 10.1002/acs

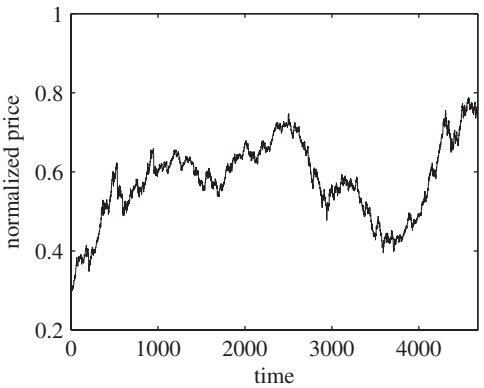

Figure 3. A typical sample of processed time series.

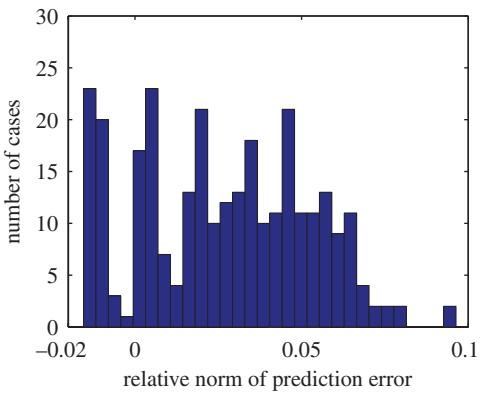

Figure 4. Relative norm of prediction errors $\mathscr{P}^{6:T}$ (20), $k \in k^*$.

We processed 49 time series describing the daily prices of various commodities. Figure 3 shows a sample of such time series normalized to zero mean and unit variance. Series contain between 3500 and 5500 records. Estimated normal AR models serve for k-days predictions, $k \in k^* \equiv \{1, 2, \ldots, 6\}$. Models have the expectations

$$\mathsf{E}[y_t | y^{t-k}, \theta] = \theta'[y_{t-k}, y_{t-1-k}]'$$

where the best 'order' 2 was chosen by trial-and-error. The SSF used the default f_0 (21) as an alternative together with the best $\lambda = 0.96$. The ASF ran Algorithm 2 with the same f_0 and the recommended $\varepsilon = 0.05$.

For one-step-ahead prediction, $k = 1$, no forgetting was necessary. In this case, both the ASF and the SSF slightly deteriorated predictions. For $k \geqslant 2$, use of forgetting was vital and the ASF systematically outperformed the SSF. Figure 4 demonstrates it showing the histogram of $\mathscr{P}^{1:T}$, see (20). The results for all delays are shown.

DOI: 10.1002/acs

Improvement of the predictive performance due to the use of the ASF is practically significant as just the better processing of data helped in crossing the boundary between financial losses and gains. Methodologically, it is important that the *relative prediction norms* $\mathscr{P}$ (20) *are predominantly positive for* $k>1$: the ASF *outperformed* the SSF *with a few exceptions*.

7. CONCLUDING REMARKS

Introduction of stabilized-type forgetting made in the 80s of the last century, e.g. [22, 23], was a small but an important step in converting academic adaptive systems into reliable practical algorithms. Since that time it has seemed that the possibilities for further progress are more or less exhausted. However, an additional insight into the underlying optimization has led to a shift in the view on 'dual' versions of stabilized forgetting [6]. Unlike there, the stabilized forgetting, corresponding to the 'correct' order of arguments in the KLD, is predicted to provide better tracking properties. The experimental results confirm it.

Methodologically, the preferred order of arguments in the KLD opens a way for improving functional approximation of Bayesian estimation. It approximates unfeasible pdfs by pdfs with a restricted dependence structure. It uses, however, the 'incorrect' version of the KLD as the proximity measure. It is possible to use the 'correct' version and optimize it over pdfs of a restricted structure. The chances for practical gains are high as was demonstrated on the non-trivial projection-based mixture estimation [10, 11].

REFERENCES

1. Bernardo JM, Smith AFM. *Bayesian Theory* (2nd edn). Wiley: Chichester, New York, Brisbane, Toronto, Singapore, 1997.
2. Peterka V. Bayesian system identification. In *Trends and Progress in System Identification*, Eykhoff P (ed.). Pergamon Press: Oxford, 1981; 239–304.
3. Barndorff-Nielsen O. *Information and Exponential Families in Statistical Theory*. Wiley: New York, 1978.
4. Bernardo JM. Expected information as expected utility. *The Annals of Statistics* 1979; **7**(3):686–690.
5. Kullback S, Leibler R. On information and sufficiency. *Annals of Mathematical Statistics* 1951; **22**:79–87.
6. Kulhavý R, Kraus FJ. On duality of regularized exponential and linear forgetting. *Automatica* 1996; **32**:1403–1415.
7. Kulhavý R, Zarrop MB. On a general concept of forgetting. *International Journal of Control* 1993; **58**(4):905–924.
8. Tipping ME, Bishop CM. Probabilistic principal component analysis. *Journal of the Royal Statistical Society*, *Series B—Statistical Methodology* 1999; **61**:611–622.
9. Šmídl V, Quinn A. *The Variational Bayes Method in Signal Processing*. Springer: Berlin, 2005.
10. Andrýsek J. Approximate recursive Bayesian estimation of dynamic probabilistic mixtures. In *Multiple Participant Decision Making*, Andrýsek J, Kárný M, Kracík J (eds). Advanced Knowledge International: Adelaide, May 2004; 39–54.
11. Andrýsek J. Estimation of dynamic probabilistic mixtures. *Ph.D. Thesis*, ÚTIA AV ČR, POB 18, 18208 Prague 8, Czech Republic, 2005.
12. Kárný M, Böhm J, Guy TV, Jirsa L, Nagy I, Nedoma P, Tesař L. *Optimized Bayesian Dynamic Advising*: *Theory and Algorithms*. Springer: London, 2005.
13. Zellner A. *An Introduction to Bayesian Inference in Econometrics*. Wiley: New York, 1976.
14. Vajda I. *Theory of Statistical Inference and Information*. Kluwer Academic Publishers: Dordrecht, 1989.
15. Kerridge DF. Inaccuracy and inference. *Journal of the Royal Statistical Society*, *Series B* 1961; **23**:284–294.
16. Berger JO. *Statistical Decision Theory and Bayesian Analysis*. Springer: New York, 1985.
17. Attias H. A variational Bayesian framework for graphical models. In *Advances in Neural Information Processing Systems*, Leen T (ed.), vol. 12. MIT Press: Cambridge, MA, 2000.
18. Nenutil P. Approximate recursive estimation of dynamic mixture models. *Technical Report*, ÚTIA AV ČR, Praha, 2004.

DOI: 10.1002/acs

19. Kárný M, Nedoma P, Khailova N, Pavelková L. Prior information in structure estimation. *IEE Proceedings—Control Theory and Applications* 2003; **150**(6):643–653.
20. Kulhavý R. Directional tracking of regression-type model parameters. Preprints of the *2nd IFAC Workshop on Adaptive Systems in Control and Signal Processing*, Lund, Sweden, 1986; 97–102.
21. Milek J. Stabilized adaptive forgetting in the recursive parameter estimation. *Vol. Diss. ETH No. 10893*, Zürich, Switzerland, 1995.
22. Kraus FJ. Stabilized least squares estimators for time-variant processes. *Proceedings of the 28th IEEE Conference on Decision and Control*, Tampa, FL, 1984.
23. Kulhavý R, Kárný M. Tracking of slowly varying parameters by directional forgetting. Preprints of the *9th IFAC World Congress*, *IFAC*, Budapest, vol. X, 1984; 178–183.

DOI: 10.1002/acs