BAYESIAN ESTIMATION OF FORGETTING FACTOR IN ADAPTIVE FILTERING AND CHANGE DETECTION

Václav Šmídl

Department of Adaptive Systems,
Institute of Information Theory and Automation,
Czech Republic, smidl@utia.cas.cz

Fredrik Gustafsson

Department of Electrical Engineering,
Linköping University, Linköping,
Sweden, fredrik@isy.liu.se

ABSTRACT

An adaptive filter is derived in a Bayesian framework from the assumption that the difference in the parameter distribution from one time to another is bounded in terms of the Kullback-Leibler divergence. We show an explicit link to the general concepts of exponential forgetting, and outline the details for a linear Gaussian model with unknown parameter and covariance. We extend the problem to an unknown forgetting factor, where we provide a particular prior that allows for abrupt changes in forgetting, which is useful in change detection problems. The Rao-Blackwellized particle filter is used for the implementation, and its performance is assessed in a simulation of system with abrupt changes of parameters.

Index Terms— Adaptive filtering, exponential forgetting, maximum entropy, Rao-Blackwellized particle filtering

1. INTRODUCTION

Adaptive filtering concerns recursive estimation of parameters in a parametric signal model, where the underlying goal is to minimize the residuals, that is, the ability of the model to predict the observations. The most common approaches are based on optimization with algorithms such as least mean square (LMS), normalized LMS and recursive least squares (RLS) with exponential forgetting [1]. The classical technique of exponential forgetting [2] has many applications in adaptive signal processing [3]. Its theoretical justification ranges from hypothesis testing [2] to maximum entropy arguments [4].

Another approach is based on optimal filtering, where a random walk model for the parameter evolution is assumed, leading to Kalman filter (KF) based algorithms. An interesting link between these two approaches is presented in [5], where it is shown that LMS, NLMS and RLS can be interpreted as KF for a particular random walk. However, the Euclidian size of a move in the parameter space does not correspond to how much the predictive distribution of the model changes. For instance, the last parameter in an autoregressive model affects the prediction much more than the first parameter.

In this contribution, we take a conceptually completely different approach to adaptive filtering, where we assume a bound (rather than a random walk) on the change in the parameter posterior density (rather than the parameter vector) at each time step. The bound is given in terms of the Kullback-Leibler divergence in a Bayesian framework. Surprisingly, this leads to an algorithm with exponential forgetting, which bears much in common with RLS. In contrast to RLS, we also embed the noise covariance matrix and later the forgetting parameter itself into the Bayesian framework. The latter leads to a self-adjusting adaptive algorithm that with a suitable choice of prior of forgetting allows abrupt changes. The computational efficient marginalized particle filter [6] is used for the joint estimation of the forgetting factor [7].

2. FORGETTING AS MAXIMUM ENTROPY ESTIMATION

Consider observation model in the exponential family

$$y_t \sim p(y_t|\theta_t) = \rho(y_t)\exp(\eta(\theta_t)\cdot\tau(y_t) - \phi(\theta_t))), \qquad (1)$$

where y_t is the vector of observations, θ_t is the vector of unknown parameters, $\eta(\theta_t)$ and $\phi(\theta_t)$ are vector and scalar valued functions of the parameters, respectively; $\rho(y_t)$ and $\tau(y_t)$ are scalar and vector valued functions of the realization y_t; the symbol $\cdot$ denotes scalar product of two vectors. Since θ_t is time-varying, (1) may be complemented by an evolution model $p(\theta_t|\theta_{t-1})$ to form a complete state-space model. However, since it is typically unknown, we follow the maximum entropy approach suggested in [4].

2.1. Measurement Update in Exponential Family

Since the likelihood function (1) for the unknown parameter θ_t, is in the exponential family, we assume that the prior on θ_t is in the form conjugate to (1), i.e.

$$p(\theta_t|y_{1:t-1}) = \frac{\exp(\eta(\theta_t)V_{t|t-1} - \nu_{t|t-1}\phi(\theta_t))}{\gamma(V_{t|t-1},\nu_{t|t-1})}, \qquad (2)$$

Work supported by: GAČR GAP102/11/0437

where $V_{t|t-1}$ is a vector of sufficient statistics and $\nu_{t|t-1}$ is a scalar counter of the effective number of samples in the statistics. The normalization factor $\gamma(V_{t|t-1}, \nu_{t|t-1})$ is uniquely determined by the statistics $V_{t|t-1}$ and $\nu_{t|t-1}$. Then, the posterior density $p(\theta_t|y_{1:t})$ is in the form (2) with statistics

$$V_{t|t} = V_{t|t-1} + \tau(y_t). \tag{3}$$

$$\nu_{t|t} = \nu_{t|t-1} + 1, \tag{4}$$

Recursive nature of (3)–(4) is advantageous for on-line evaluation of sufficient statistics starting from a prior defined by V_0, ν_0. The predictive distribution of y_t is analytically available as:

$$p(y_t|y_{1:t-1}) = \gamma^{-1}(V_{t|t-1}, \nu_{t|t-1})\gamma(V_{t|t}, \nu_{t|t})\rho(y_t). \tag{5}$$

2.2. Time Update in Exponential Family

Bayesian estimation of non-stationary parameters θ_t requires formalization of the parameter evolution model $p(\theta_{t+1}|\theta_t)$. The predictive density of the parameter θ_{t+1} is obtained by marginalization

$$p(\theta_{t+1}|y_{1:t}) = \int p(\theta_{t+1}|\theta_t)p(\theta_t|y_{1:t})d\theta_t. \tag{6}$$

Since the transition model $p(\theta_{t+1}|\theta_t)$ is unknown, we seek an estimate of the marginal $p(\theta_{t+1}|y_{1:t})$ for many possible transition models. This has been achieved by the forgetting operator [2]. Recently, the same result has been derived using the maximum entropy arguments in [4], which we follow. Specifically, we consider $p(\theta_{t+1}|\theta_t)$ to be unknown, but implicitly limited by the constraint that

$$\text{KL}(p(\theta_{t+1}|y_{1:t})||p_\delta(\theta_{t+1}|y_{1:t})) \leq \kappa, \tag{7}$$

where KL is the Kullback-Leibler divergence defined as

$$\text{KL}(p_1||p_2) = \int_{-\infty}^{\infty} p_1(x)\log\left(\frac{p_1(x)}{p_2(x)}\right)dx. \tag{8}$$

$0 \leq \kappa < \infty$, is a known constant and

$$p_\delta(\theta_{t+1}|y_{1:t}) = \int \delta(\theta_{t+1} - \theta_t)p(\theta_t|y_{1:t})p\theta_t, \tag{9}$$

where $\delta()$ is the Dirac delta function. Equation (9) gives the predictive density for the case of time-invariant parameters. The interpretation of (7) is that we obtain an implicit definition of a class of transition models $p(\theta_{t+1}|\theta_t)$ giving predictive densities $p(\theta_{t+1}|y_{1:t})$ which are close to $p_\delta(\theta_{t+1}|y_{1:t})$, where the closeness is measured in the Kullback-Leibler sense. A deeper discussion is provided in Section 2.3.

Following the principle of maximum entropy, we choose to approximate (6) by a distribution $\hat{p}(\theta_{t+1}|y_{1:t})$ that has the maximum entropy of all distributions satisfying (7).

Theorem 1 ***(Maximum entropy under KL divergence constraint)*** *For a given $p_\delta(\theta_{t+1}|y_{1:t})$, the probability distribution*

$$\hat{p}(\theta_{t+1}|y_{1:t}, \lambda_t) = p_\delta(\theta_{t+1}|y_{1:t})^{\lambda_t} p_u(\theta_{t+1})^{1-\lambda_t}, \tag{10}$$

has maximum entropy of all densities $p(\theta_{t+1})$ defined on the same support as $p_\delta(\theta_{t+1}|y_{1:t})$ which satisfy (7) for a given value of κ. $p_u(\theta_{t+1})$ is an invariant measure of the entropy [8] and λ_t is a solution to the equation

$$\text{KL}(\hat{p}(\theta_{t+1}|y_{1:t}, \lambda_t)||p_\delta(\theta_{t+1}|y_{1:t})) = \kappa. \tag{11}$$

Proof: based on direct application of the Karush–Kuhn–Tucker conditions [4].

Note that, for the special case of stationary parameters, $\kappa = 0$, yielding $\lambda = 1$. For sudden changes of the parameter, $\kappa \to \infty$, $\lambda \to 0$ and the invariant measure $p_u(\theta_{t+1})$ has the role of the prior density. In other cases $p_u(\theta_{t+1})$ acts as regularization.

The solution (10) is particularly advantageous in the exponential family, since (10) preserves the exponential form with statistics

$$\nu_{t+1|t} = \lambda\nu_{t|t} + (1-\lambda)\nu_u, \tag{12}$$

$$V_{t+1|t} = \lambda V_{t|t} + (1-\lambda)V_u, \tag{13}$$

where we assume that the invariant measure is also in the exponential form (2) with statistics ν_u, V_u.

2.3. Interpretation of forgetting

Equation (12) is known as exponential forgetting, [9][2]. The maximum entropy interpretation [4] allows a new interpretation of the forgetting factor as a measure on the parameter evolution model. Note from (7) that a single value of κ_t determines a class of parameter evolution models of various kinds, including state-dependent models. The value of κ may be considered as a user specific parameter, or it may be adapted on-line by an approximate solution of (7).

From practical point of view, forgetting has significant algorithmic advantage in the closed form solution of the time update equation (6), a property that is very rare for explicit random walk models. However, the classical variant with constant forgetting factor can be hard to tune and even then inadequate for abrupt changes in the parameter values. In this work, we interpret λ as an unknown parameter of the transition model which would yield (10) under proper marginalization (6).

3. BAYESIAN ESTIMATION OF THE FORGETTING FACTOR

Under the assumption that the forgetting factor λ_t is a parameter of the transition model[1], λ_t can be estimated from data using Bayesian techniques.

[1] This interpretation is exact in special cases of model (1), [10].

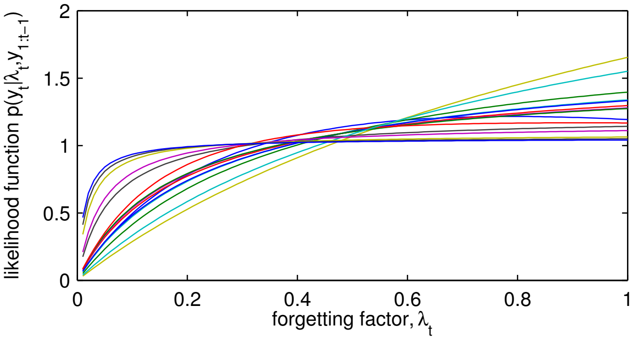

Fig. 1. Example of numerically evaluated likelihood functions for the example in Section 4.

3.1. Likelihood functions and prior distribution

Formally, joint estimation of θ_t and λ_t is

$$p(\theta_t, \lambda_t | y_{1:t}) \propto p(y_t|\theta_t)p(\theta_t|\theta_{t-1}, \lambda_t)p(\theta_{t-1}|y_{1:t-1})p(\lambda_t).$$

however, since marginalization over θ_t is analytically tractable, the marginal posterior distribution of λ_t has the form

$$p(\lambda_t|y_{1:t}) \propto p(y_t|\lambda_t, y_{1:t-1})p(\lambda_t), \tag{14}$$

where $p(y_t|\lambda_t)$ is given by (5) and $p(\lambda_t)$ is the prior on λ_t. We note that the likelihhod function $p(y_t|\lambda_t)$ is rather flat, see illustrative example in Fig. 1. Hence, the inference of time varying λ_t is highly prior dominated. This calls for detail analysis of the prior $p(\lambda_t)$. One possibility is to assume dependence of λ_t on λ_{t-1} via $p(\lambda_t|\lambda_{t-1})$, e.g. of the Dirichlet form [7]. However, this prior is inappropriate when rapid changes of the parameters are expected. In such situations, we propose to use the following mixture model

$$p(\lambda_t) = w_1 p_1(\lambda_t) + (1 - w_1)p_0(\lambda_t). \tag{15}$$

where w_1 denotes probability of stationary parameters (i.e. $p_1(\lambda_t)$ favors high values of λ_t), and $(1 - w_1)$ probability of rapid change ($p_0(\lambda_t)$ favors low values of λ_t).

3.2. Particle filtering

Estimation of the unknown forgetting factor from the mixture model (15) is then achieved by by the Rao-Blackwellized particle filter [11] (also known as marginalized particle filter [6]). Specifically, we introduce indicator variable $l_t \in \{0, 1\}$ denoting sampling from $p_0(\cdot)$ or $p_1(\cdot)$, with $p(l_t = 1) = w_1$. The full posterior on θ_t, λ_t, l_t is partitioned into

$$p(\theta_t, \lambda_t, l_t|y_{1:t}) = p(\theta_t|\lambda_t, y_{1:t}) \sum_{i=1}^{n} \omega_t^{(i)} \delta(\lambda_t - \lambda_t^{(i)})\delta(l_t - l_t^{(i)}).$$

where the distribution on parameters θ_t is the result of forgetting (10), and the weights $\omega_t^{(i)}$ are importance weights of the particles:

$$\omega_t \propto \frac{p(y_t|\lambda_t, y_{1:t-1})p(\lambda_t|l_t)p(l_t|l_{t-1})}{g(\lambda_t, l_t|\lambda_{t-1}, l_{t-1}, y_{1:t})}, \tag{16}$$

where $g(\cdot)$ is the chosen proposal function, which is important for computational efficiency of the filter.

4. EXAMPLE: LINEAR GAUSSIAN MODEL

In this section, we derive forgetting for the practically important case of normal distributed noises.

4.1. Likelihood and Conjugate Prior

Consider multivariate linear regression model

$$y_t = \theta_t' \psi_t + \Sigma_t^{\frac{1}{2}} e_t \tag{17}$$

of d_y-dimensional vector of observations y_t, with unknown matrix parameters θ_t ($d_\psi \times d_y$) and Σ_t ($d_y \times d_y$), d_ψ-dimensional regressor ψ_t containing past observations or exogeneous variables, and Gaussian disturbance e_t. Likelihood defined by (17) is conjugate with Normal-inverse-Wishart distribution prior, $[\theta_t, \Sigma_t] \sim \mathrm{NiW}(\nu_t, V_t)$. The Normal-inverse-Wishart distribution defines a hierarchical Bayesian model:

$$y_t|\theta_t, \Sigma_t \sim \mathcal{N}(\theta_t'\psi_t, \Sigma_t), \tag{18}$$

$$\theta_t|\Sigma_t, y_{1:t} \sim \mathcal{N}(\hat{\theta}_{t|t}, Z_{t|t} \otimes \Sigma_t) \tag{19}$$

$$\Sigma_t|y_{1:t} \sim \mathrm{iW}(\nu_{t|t}, \Lambda_{t|t}) \tag{20}$$

where $\mathrm{iW}(.)$ denotes the Inverse Wishart distribution with mean value $\hat{\Sigma}_t = \Lambda_{t,t}/(\nu_{t|t} - d_y - 1)$. The quantities $\hat{\theta}_{t|t}, Z_{t|t}, \Lambda_{t|t}, \nu_{t|t}$ can be recursively computed as follows [12]:

$$Z_{t|t} = Z_{t|t-1} - \frac{1}{1 + \zeta_t} Z_{t|t-1}\psi_t\psi_t' Z_{t|t-1}, \tag{21}$$

$$\zeta_t = \psi_t' Z_{t|t-1}\psi_t \tag{22}$$

$$\hat{\theta}_{t|t} = \hat{\theta}_{t|t-1} + \frac{1}{1 + \zeta_t} Z_{t|t-1}\psi_t(y_t - \hat{y}_{t|t-1}), \tag{23}$$

$$\nu_{t|t} = \nu_{t|t-1} + 1, \tag{24}$$

$$\Lambda_{t|t} = \Lambda_{t|t-1} + \frac{1}{1 + \zeta_t}(\hat{y}_{t|t-1} - y_t)(\hat{y}_{t|t-1} - y_t)'. \tag{25}$$

Here, $\hat{y}_{t|t-1} = \hat{\theta}_{t|t-1}\psi_t$, and statistics of the predictive distributions are:

$$Z_{t|t-1} = \lambda^{-1} Z_{t-1|t-1}, \qquad \hat{\theta}_{t|t-1} = \hat{\theta}_{t-1|t-1}, \tag{26}$$

$$\nu_{t|t-1} = \lambda\nu_{t-1|t-1} + (1 - \lambda)\nu_u, \quad \Lambda_{t|t-1} = \lambda\Lambda_{t-1|t-1}, \tag{27}$$

The predictive distribution of y_t (5) becomes a multivariate Student-t density with $\nu_{t|t-1} - d_y + 1$ degrees of freedom

$$p(y_t|\nu_{t-1}, V_{t-1}) \propto \tag{28}$$

$$\left|1 + (y_t - \theta_{t|t-1}\psi_t)\frac{\Lambda_{t|t-1}^{-1}}{1 + \zeta_t}(y_t - \hat{\theta}_{t|t-1}\psi_t)\right|^{-\frac{1}{2}(\nu_{t|t-1}+1)},$$

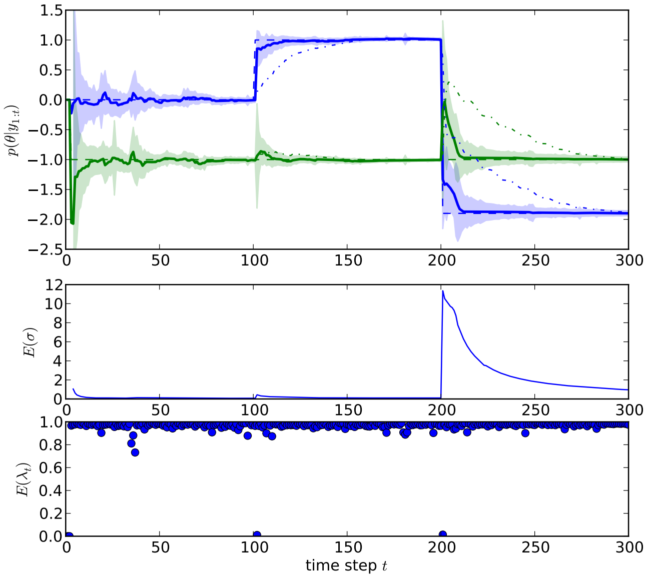

Fig. 2. Estimation of a system with rapid changes in parameters. **Top**: Simulated parameter θ_t (dashed line) and its posterior distribution via its mean value (thick full line) and 2std bounds (shaded area). Mean value of estimates with constant forgetting factor $\lambda = 0.95$ is also displayed (thin dash-dotted line). **Middle**: expected value of the parameter σ_t. **Bottom**: expected value of the forgetting factor λ_t.

with mean value $\mathrm{E}(y_t) = \hat{y}_{t|t-1}$ and variance $\mathrm{Var}(y_t) = (1+\zeta_t)\Lambda_{t,t}/(\nu_{t|t} - d_y - 1)$.

Note that with posterior factorized into Normal and Wishart parts, we can choose different forgetting factors λ_N and λ_W, for each equation (26) and (27), respectively.

4.2. Experimental Results

A simulated example of an autoregressive model:

$$y_t = \theta_t'[y_{t-1}, y_{t-2}]' + \sigma e_t,$$

with constant $\sigma = 0.3$ and initial parameter values $\theta_0 = [0, -1]$ was used for tests. A simulation run with rapid changes in the parameters was simulated, see Figure 2.

The estimation method was Algorithm 1, with $N = 10$ particles, partial forgetting with unknown λ_N with prior distribution of $\lambda_{N,t}$ in the form of (15) with $w_1 = 0.95$, $p_1(\lambda_t) = Be(100, 1)$ and $p_0(\lambda_t) = U(0, 1)$. Forgetting factor for the variance $\lambda_{W,t}$ is deterministically dependent on $\lambda_{N,t}$ via $\lambda_{W,t} = \frac{k\lambda_N}{1+(k-1)\lambda_N}$ with $k = 10$, which corresponds to k times longer exponential window. The proposal distribution for l_t was multinomial with $q(l_t = 1) = 0.5$.

All parameters of the invariant measure $p_u(\theta_t)$ were chosen to be as uninformative as possible, i.e. $Z_u = \mathbf{0}, L_u = 0, \nu_u = 1$. The proposal function was chosen to make sure that in each time at least one particle will have the $\lambda_t^{(i)}$ from the uniform component. This would be unlikely if the proposal density is equal to the prior. The same experiment with $\lambda_{W,t} = \lambda_{N,t}$ exhibit much longer convergence times to the true parameters after the rapid change. This was due to the fact that the jump in the observations was explained by increase of variance σ.

5. CONCLUSION

Maximum entropy interpretation of forgetting is an interesting way of looking at the classical problem that allows various extensions. Bayesian estimation of the forgetting factor was proposed and its key components—the likelihood function and the prior—were examined. It was found that the likelihood is typically very flat and the inference is often prior-dominated. Mixture-based prior on the forgetting factor was therefore proposed. It was shown in simulation that a RB particle filter with as low as 10 particles is capable of estimation of rapid changes in the parameters. However, sensitivity of the performance to the prior distribution was also observed.

6. REFERENCES

[1] A.H. Sayed, *Fundamentals of adaptive filtering*, Wiley-IEEE Press, 2003.

[2] R. Kulhavý and M. B. Zarrop, “On a general concept of forgetting,” *Int. J. of Control*, vol. 58, no. 4, pp. 905–924, 1993.

[3] F. Gustafsson, *Adaptive filtering and change detection*, vol. 5, Wiley Online Library, 2000.

[4] M. Kárný and K. Dedecius, “Approximate Bayesian recursive estimation:,” Tech. Rep. 2317, Institute of Information Theory and Automation, Prague, 2012.

[5] L. Ljung and S. Gunnarsson, “Adaptation and tracking in system identification–a survey,” *Automatica*, vol. 26, no. 1, pp. 7–21, 1990.

[6] T. Schön, F. Gustafsson, and P.-J. Nordlund, “Marginalized particle filters for mixed linear/nonlinear state-space models,” *IEEE Transactions on Signal Processing*, vol. 53, pp. 2279–2289, 2005.

[7] K. Dedecius and R. Hofman, “Autoregressive model with partial forgetting within Rao-Blackwellized particle filter.,” *Communication in Statistics – Simulation and Computation.*, vol. 41, no. 5, pp. 582–589, 2012.

[8] ET Jaynes, “Statistical physics,” *Brandeis Lectures*, vol. 316, 1963.

[9] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*, Academic Press, New York, 1979.

[10] A. E. Raftery, M. Kárný, and P. Ettler, “Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill,” *Technometrics*, pp. 52–66, 2010.

[11] A. Doucet, N. de Freitas, and N. Gordon, Eds., *Sequential Monte Carlo Methods in Practice*, Springer, 2001.

[12] V. Peterka, “Bayesian approach to system identification,” in *Trends and Progress in System identification*, P. Eykhoff, Ed., pp. 239–304. Pergamon Press, Oxford, 1981.