AUTOREGRESSIVE MODEL WITH PARTIAL FORGETTING WITHIN RAO-BLACKWELLIZED PARTICLE FILTER

K. Dedecius, R. Hofman[*]

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

ABSTRACT

We are concerned with Bayesian identification and prediction of a nonlinear discrete stochastic process. The fact, that a nonlinear process can be approximated by a piecewise linear function advocates the use of adaptive linear models. We propose a linear regression model within a Rao-Blackwellized particle filter. The parameters of the linear model are adaptively estimated using a finite mixture, where the weights of components are tuned with a particle filter. The mixture reflects *a priori* given hypotheses on different scenarios of (expected) parameters' evolution. The resulting hybrid filter locally optimizes the weights to achieve the best fit of a nonlinear signal with a single linear model.

Index Terms— Particle filters, Bayesian methods, Recursive estimation

1. INTRODUCTION

The theory of approximation of nonlinear signals by piecewise linear functions has attained considerable attention in the past decades [1], mainly in the field of control engineering, e.g., [2]. The method is popular, because in many technical applications, it offers a reasonable trade-off between the models' complexity and its performance [3, 4]. Our method is inspired by the fact, that if the transitions between two successive (almost) linear segments is smooth enough, it can be modelled with switching models, e.g. [5, 6, 7], or model averaging [8].

Notational conventions: $\propto$ denotes proportionality, i.e., equality up to a constant factor. A' denotes transpose of A. $p(a|b)$ is a probability density function (pdf) of a (multivariate) random variable a given b. The pdfs are distinguished by their argument. $t \in \{1, 2, \dots\}$ denotes discrete time instants. All integrations are over the maximal plausible support.

2. BAYESIAN APPROACH TO MODELLING

Assume, that we are given a time series of real observations $\boldsymbol{Y}^{t-1} = (y_1, \dots, y_{t-1})$ and our purpose is to determine its next value y_t. The statistical approach employs, among others, the parametric models describing the dependence of $\boldsymbol{y}_t$ on previous observation $\boldsymbol{Y}^{t-1}$ through conditional distributions with probability density functions (pdf)

$$p(y_t|\boldsymbol{Y}^{t-1}, \boldsymbol{\Theta}). \tag{1}$$

Under the Bayesian treatment, $\boldsymbol{\Theta}$ is a set of constant model parameters with a pdf $p(\boldsymbol{\Theta}|\boldsymbol{Y}^{t-1})$. If this distribution is properly chosen from a class conjugate to the model (1), the Bayes' theorem yields a posterior pdf of the same type, and the recursive data update reads [9, 10]

$$p(\boldsymbol{\Theta}|\boldsymbol{Y}^{t}) = \frac{p(y_t|\boldsymbol{Y}^{t-1}, \boldsymbol{\Theta})p(\boldsymbol{\Theta}|\boldsymbol{Y}^{t-1})}{p(y_t|\boldsymbol{Y}^{t-1})}. \tag{2}$$

The predictive pdf $p(y_{t+1}|\boldsymbol{Y}^{t})$ provides the Bayesian output prediction. Using Chapman-Kolmogorov equation [11], it holds

$$p(y_{t+1}|\boldsymbol{Y}^{t}) = \int p(y_{t+1}|\boldsymbol{Y}^{t}, \boldsymbol{\Theta})p(\boldsymbol{\Theta}|\boldsymbol{Y}^{t})\mathrm{d}\boldsymbol{\Theta} = \frac{\mathcal{I}_{t+1}}{\mathcal{I}_t}, \tag{3}$$

where $\mathcal{I}$ denotes the normalizing integral, see, e.g. [10].

Although the described methodology is important *per se*, its lack of adaptivity prevents successful application to non-static cases, when $\boldsymbol{\Theta}$ is not time-invariant. For a time varying $\boldsymbol{\Theta}_t$, it is necessary to perform an additional update

$$p(\boldsymbol{\Theta}_t|\boldsymbol{Y}^{t}) \to p(\boldsymbol{\Theta}_{t+1}|\boldsymbol{Y}^{t}). \tag{4}$$

Here, two significant cases can occur:

(i) The evolution model $p(\boldsymbol{\Theta}_{t+1}|\boldsymbol{\Theta}_t, \boldsymbol{Y}^{t})$ is known *a priori*.

(ii) A suitable model of parameter evolution is not known, but we can expect that they vary relatively slowly.[1]

The case (i) allows to aggregate both (1) and $p(\boldsymbol{\Theta}_{t+1}|\boldsymbol{\Theta}_t, \boldsymbol{Y}^{t})$ into a single complex model. Then, $\boldsymbol{\Theta}_t$ represents the system state and under certain conditions, the estimation task leads to the well known Kalman filtering [10].

[*]This research was supported by grants MV ČR VG20102013018 and SGS 10/099/OHK3/1T/16.

[1]However, this is a vague statement, there is no exact definition of "slowly variations", hence it is understood intuitively.

However, there exists a wide variety of cases, when the explicit model of parameter evolution is not known. If we adopt the assumption of slowly varying parameters (ii), the popular group of estimation methods using forgetting provides a solution. It is heuristically circumventing the problem of parameter model ignorance by recursive discounting the old and potentially outdated information carried by the parameter pdf. Formally, we introduce a forgetting operator $\mathfrak{F}$:

$$p(\boldsymbol{\Theta}_{t+1}|\boldsymbol{Y}^t) = \mathfrak{F}[p(\boldsymbol{\Theta}_t|\boldsymbol{Y}^t)]. \tag{5}$$

The application of the forgetting operator is equivalent to the time update in state space models [12].

3. PARTIAL FORGETTING

The use of "classical" forgetting methods, e.g., exponential forgetting [10], is limited in nonlinear cases. We present a new approach appealing to partial forgetting [13]. While it enumerates hypotheses about variability of elements of $\boldsymbol{\Theta}_t$, our modification is more general.

Let us define a finite set $\boldsymbol{\mathcal{H}}$ of hypotheses $\{H_i\}$ regarding the distribution of $\boldsymbol{\Theta}_{t+1}$ given $\boldsymbol{\Theta}_t$. The distributions induced by these hypotheses are merged together in form of a finite mixture,

$$p(\boldsymbol{\Theta}_{t+1}|\boldsymbol{Y}^t, \boldsymbol{\pi}_t) = \sum_i \pi_{i,t}\, q_i(\boldsymbol{\Theta}_{t+1}|\boldsymbol{Y}^t), \quad \sum_i \pi_{i,t} = 1, \tag{6}$$

i.e., the posterior distribution of $\boldsymbol{\Theta}_{t+1}$ is represented by a finite mixture of hypothetical pdfs $q_i(\boldsymbol{\Theta}_{t+1}|\boldsymbol{Y}^t)$.

Theoretically correct solution would express one hypothetical pdf for each (almost) linear window. In practice, this is hardly possible and a generalization of the approach is exploited. The hypotheses $H_i \in \boldsymbol{\mathcal{H}}$ enumerate several cases, which are likely to occur. Their number depends on a specific task, mainly the signal properties, and the user's ability to guess the properties of $\boldsymbol{\Theta}_t$ in each window. For instance, we may state a hypothesis about each particular element of $\boldsymbol{\Theta}_t = \{\Theta_{1,t}, \ldots, \Theta_{N,t}\}$ and about all of them in one shot:

$$\begin{aligned}
H_0: \quad p(\boldsymbol{\Theta}_{t+1}|\boldsymbol{Y}^t, H_0) &= p(\boldsymbol{\Theta}_t|\boldsymbol{Y}^t) \\
&= q_0(\boldsymbol{\Theta}_{t+1}|\boldsymbol{Y}^t) \\
H_1: \quad p(\boldsymbol{\Theta}_{t+1}|\boldsymbol{Y}^t, H_1) &= p(\Theta_{2,t}, \ldots, \Theta_{N,t}|\boldsymbol{Y}^t, \Theta_{1,t}) \\
\times \mathfrak{F}[p(\Theta_{1,t}|\boldsymbol{Y}^t)] &= q_1(\boldsymbol{\Theta}_{t+1}|\boldsymbol{Y}^t) \\
&\vdots \\
H_N: \quad p(\boldsymbol{\Theta}_{t+1}|\boldsymbol{Y}^t, H_N) &= p(\Theta_{1,t}, \ldots, \Theta_{N-1,t}|\boldsymbol{Y}^t, \Theta_{N,t}) \\
\times \mathfrak{F}[p(\Theta_{N,t})] &= q_N(\boldsymbol{\Theta}_{t+1}|\boldsymbol{Y}^t) \\
H_{N+1}: \quad p(\boldsymbol{\Theta}_{t+1}|\boldsymbol{Y}^t, H_{N+1}) &= \mathfrak{F}[p(\boldsymbol{\Theta}_t|\boldsymbol{Y}^t)] \\
&= q_{N+1}(\boldsymbol{\Theta}_{t+1}|\boldsymbol{Y}^t)
\end{aligned}$$

This particular set of $N+1$ hypotheses is an example of many possible choices. The two extreme hypotheses H_0 and H_{N+1} represent the user's belief that none or all parameter vary, respectively. The remaining hypotheses $H_1, \ldots, H_N$ concern the case of variability of $\boldsymbol{\Theta}_t$'s one element (with appropriate index). We can choose different operators $\mathfrak{F}$ or completely expert pdfs $q_i(\boldsymbol{\Theta}_{t+1}|\boldsymbol{Y}^t)$ as well.

Working with the mixture (6) would require a rather complex treatment. Instead, we prefer to find a single pdf $\tilde{p}$ of the same class as the components, minimizing the expected Kullback-Leibler divergence of p on $\tilde{p}$

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{D}\left(p||\tilde{p}\right)|\boldsymbol{\mathcal{H}}, \boldsymbol{\pi}_t, \boldsymbol{Y}^t\right] = \\
&= \underbrace{\mathbb{E}\left[\int p(\boldsymbol{\Theta}_{t+1}|\boldsymbol{Y}^t, \boldsymbol{\pi}_t) \frac{p(\boldsymbol{\Theta}_{t+1}|\boldsymbol{Y}^t, \boldsymbol{\pi}_t)}{\tilde{p}(\boldsymbol{\Theta}_{t+1}|\boldsymbol{Y}^t, \boldsymbol{\pi}_t)} \mathrm{d}\boldsymbol{\Theta} \middle| \boldsymbol{\mathcal{H}}, \boldsymbol{\pi}_t, \boldsymbol{Y}^t\right]}_{\to \min}.
\end{aligned} \tag{7}$$

It can be shown, that $\mathcal{D}\left(f||g\right) \geq 0$ with equality for $f = g$ almost everywhere [14]. A solution, for certain cases analytical [13], defines the approximate pdf $\tilde{p}(\boldsymbol{\Theta}_{t+1}|\boldsymbol{Y}^t, \boldsymbol{\pi}_t)$, which may be directly used as the next prior distribution pdf in (2).

4. RAO-BLACKWELLIZED PARTICLE FILTER

Let $\boldsymbol{\Psi}_{t+1} = (\boldsymbol{\Theta}'_{t+1}, \boldsymbol{\pi}'_t)'$ be a real column vector. Given the value of $\boldsymbol{\pi}_t$, the minimization (7) yielding the approximation $\tilde{p}(\boldsymbol{\Theta}_{t+1}|\boldsymbol{Y}^t, \boldsymbol{\pi}_t)$ of (6), can be evaluated. The approximate pdf can be used for linear recursive estimation of model (1). Since weights $\boldsymbol{\pi}_t$ are unknown, we attempt to estimate joint pdf of regression parameters and weights $p(\boldsymbol{\Psi}_{t+1}|\boldsymbol{Y}^t)$. We exploit the natural factorization of $\boldsymbol{\Psi}_{t+1}$ and decompose the pdf $p(\boldsymbol{\Psi}_{t+1}|\boldsymbol{Y}^t)$ as follows

$$p(\boldsymbol{\Psi}_{t+1}|\boldsymbol{Y}^t) = \underbrace{p(\boldsymbol{\Theta}_{t+1}|\boldsymbol{Y}^t, \boldsymbol{\pi}_t)}_{\text{linear}} \underbrace{p(\boldsymbol{\pi}_t|\boldsymbol{Y}^t)}_{\text{PF}} \tag{8}$$

where $p(\boldsymbol{\Theta}_{t+1}|\boldsymbol{Y}^t, \boldsymbol{\pi}_t)$ is analytically tractable while $p(\boldsymbol{\pi}_t|\boldsymbol{Y}^t)$ is not. The latter pdf is approximated using particle filter (PF) [15].

Particle filtering refers to a range of techniques for generating an empirical approximation of the pdf

$$p\left(\boldsymbol{\Pi}^t|\mathbf{Y}^t\right) \approx \frac{1}{M}\sum_{j=1}^{M} \delta\left(\boldsymbol{\Pi}^t - \boldsymbol{\Pi}^{t,(j)}\right), \tag{9}$$

where $\boldsymbol{\Pi}^t = (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_t)$ and $\boldsymbol{\Pi}^{t,(j)}$, $j = 1, \ldots, M$ are independent identically distributed samples from the posterior; $\delta(\cdot)$ denotes the Dirac δ-function. Therefore, this approach is feasible only if the we can sample from the exact posterior $p\left(\boldsymbol{\Pi}^t|\mathbf{Y}^t\right)$. If this is not the case, the samples can be drawn from a chosen proposal distribution (importance function), $f\left(\boldsymbol{\Pi}^t|\mathbf{Y}^t\right)$, as follows:

$$\begin{aligned}
p\left(\boldsymbol{\Pi}^t|\mathbf{Y}^t\right) &= \frac{p\left(\boldsymbol{\Pi}^t|\mathbf{Y}^t\right)}{f\left(\boldsymbol{\Pi}^t|\mathbf{Y}^t\right)} f\left(\boldsymbol{\Pi}^t|\mathbf{Y}^t\right) \\
&\approx \frac{p\left(\boldsymbol{\Pi}^t|\mathbf{Y}^t\right)}{f\left(\boldsymbol{\Pi}^t|\mathbf{Y}^t\right)} \frac{1}{M}\sum_{j=1}^{M} \delta\left(\boldsymbol{\Pi}^t - \boldsymbol{\Pi}^{t,(j)}\right)
\end{aligned} \tag{10}$$

Using the properties of the Dirac δ-function, the approximation can be written in the form of a weighted empirical distribution, as follows:

$$p\left(\mathbf{\Pi}^t|\mathbf{Y}^t\right) \approx \sum_{j=1}^{M} w_t^{(j)} \delta\left(\mathbf{\Pi}^t - \mathbf{\Pi}^{t,(j)}\right), \tag{11}$$

$$w_t^{(j)} \propto \frac{p\left(\mathbf{\Pi}^{t,(j)}|\mathbf{Y}^t\right)}{f\left(\mathbf{\Pi}^{t,(j)}|\mathbf{Y}^t\right)}. \tag{12}$$

Under this importance sampling procedure, the true posterior distribution needs only be evaluated pointwise.

The challenge for on-line algorithms is to achieve recursive generation of samples and evaluation of the importance weights. Using standard Bayesian calculus, (12) may be written in the following recursive form:

$$w_t^{(j)} \propto \frac{p\left(y_{t+1}|\mathbf{Y}^t\right) p\left(\boldsymbol{\pi}_t^{(j)}|\boldsymbol{\pi}_{t-1}^{(j)}\right)}{f\left(\boldsymbol{\pi}_t^{(j)}|\mathbf{\Pi}^{t-1,(j)}, \mathbf{Y}^t\right)} w_{t-1}^{(j)} \tag{13}$$

Furthermore, if $f\left(\boldsymbol{\pi}_t^{(j)}|\mathbf{\Pi}^{t-1,(j)}, \mathbf{Y}^t\right) = p\left(\boldsymbol{\pi}_t^{(j)}|\boldsymbol{\pi}_{t-1}^{(j)}\right)$, then the importance density becomes only dependent on the $\boldsymbol{\pi}_{t-1}$ and $\boldsymbol{y}_t$. This is particularly useful in the common case when only a filtered estimate of the posterior $p(\boldsymbol{\pi}_t|\boldsymbol{Y}^t)$ is required at each time step. It means, that only $\boldsymbol{\pi}_t^{(i)}$ need to be stored [16]. Then, the marginal posterior density $p(\boldsymbol{\pi}_t|\boldsymbol{Y}^t)$ can be approximated as

$$p\left(\boldsymbol{\pi}_t|\mathbf{Y}^t\right) \approx \sum_{j=1}^{M} w_t^{(j)}\ \delta\left(\boldsymbol{\pi}_t - \boldsymbol{\pi}_t^{(j)}\right). \tag{14}$$

Substituting (14) into (8) yields

$$p\left(\mathbf{\Psi}_t|\boldsymbol{Y}^t\right) = \sum_{j=1}^{M} w_t^{(j)}\ p\left(\mathbf{\Theta}_t|\boldsymbol{\pi}_t^{(j)}, \boldsymbol{Y}^t\right) \delta\left(\boldsymbol{\pi}_t - \boldsymbol{\pi}_t^{(j)}\right) \tag{15}$$

Now, we have to sample from the space of $\boldsymbol{\pi}_t$. The weights can be evaluated recursively:

$$w_t^{(j)} \propto \frac{p\left(\boldsymbol{y}_t|\boldsymbol{\pi}_t^{(j)}\right)\ p\left(\boldsymbol{\pi}_t^{(j)}|\boldsymbol{\pi}_{t-1}^{(j)}\right)}{f\left(\boldsymbol{\pi}_t^{(j)}|\boldsymbol{\pi}_{t-1}^{(j)}, y_t\right)} w_{t-1}^{(j)}. \tag{16}$$

For exact marginalization, all proofs of global convergence hold [17].

5. IMPLEMENTATION

Let the model (1) be a linear N-th order autoregressive model with Gaussian disturbances

$$p(y_t|\boldsymbol{Y}^{t-1}, \mathbf{\Theta}_t) = p(y_t|\boldsymbol{\varphi}_t, \boldsymbol{\theta}_t, \boldsymbol{Y}^{t-1}) \sim \mathcal{N}(\boldsymbol{\varphi}_t'\boldsymbol{\theta}_t, \sigma^2), \tag{17}$$

where $\boldsymbol{\varphi}_t = (y_{t-1}, \ldots, y_{t-N}, 1)'$ is a regression vector and $\boldsymbol{\theta}_t \in \mathbb{R}^{N+1}$ is a vector of regression coefficients; $\sigma^2 \in \mathbb{R}^+$ is the noise variance. The Bayesian paradigm exploits the Gauss-inverse-Wishart distribution as a suitable conjugate prior distribution [12]

$$p(\mathbf{\Theta}_t|\boldsymbol{Y}^{t-1}) \sim \mathcal{G}i\mathcal{W}(\boldsymbol{V}_t, \nu_t), \tag{18}$$

where $\boldsymbol{V}_t \in \mathbb{R}^{(N+1)\times(N+1)}$ denotes an extended information matrix, i.e., a positive definite symmetric matrix. The term $\nu_t \in \mathbb{R}^+$ stands for the degrees of freedom [10]. The data update rule (2) reads

$$\begin{aligned}\boldsymbol{V}_t &= \boldsymbol{V}_{t-1} + (y_t, \boldsymbol{\varphi}_t')'(y_t, \boldsymbol{\varphi}_t')\\ \nu_t &= \nu_{t-1} + 1\end{aligned}$$

There are various methods accomplishing the time update based on forgetting (5), e.g. [10, 18, 19], and many others. The approximation of mixture (6) of $\mathcal{G}i\mathcal{W}$ pdfs, in the sense of minimization of the Kullback-Leibler divergence (7), is thoroughly described in [13]. The weights $\boldsymbol{\pi}_t$ are sampled from the Dirichlet distribution $\mathcal{D}ir(\boldsymbol{\pi}_t)$ by the particle filter. The evolution model $\boldsymbol{\pi}_t|\boldsymbol{\pi}_{t-1}$ is given by the following transition pdf

$$\boldsymbol{\pi}_t|\boldsymbol{\pi}_{t-1} \sim \mathcal{D}ir(\boldsymbol{\pi}_{t-1}/\Delta + s), \tag{19}$$

where Δ is the width of the random walk and s is the stabilization term. Both of them, plus the proposal distribution, are *a priori* given by the user. A popular choice of the proposal distribution $f\left(\boldsymbol{\pi}_t^{(j)}|\boldsymbol{\pi}_{t-1}^{(j)}, y_t\right) = p\left(\boldsymbol{\pi}_t^{(j)}|\boldsymbol{\pi}_{t-1}^{(j)}\right)$ simplifies (16),

$$w_t^{(j)} \propto p\left(\boldsymbol{y}_t|\boldsymbol{\pi}_t^{(j)}\right)\ w_{t-1}^{(j)}.$$

There exist more optimal choices of proposal density as well, see, e.g. [15].

6. SIMULATION

In this simulation, we analyze a time series $y_t = \mathrm{x}(t) + e_t$, where $x(t)$ is given by the x-component of the Lorenz system [20]

$$\begin{aligned}\dot{\mathrm{x}} &= \sigma(\mathrm{y} - \mathrm{x})\\ \dot{\mathrm{y}} &= \rho\mathrm{x} - \mathrm{y} - \mathrm{xz}\\ \dot{\mathrm{z}} &= -\beta\mathrm{z} + \mathrm{xy}\end{aligned}$$

and $e_t \sim \mathcal{N}(0, 1)$. We integrate the system numerically by the Runge-Kutta algorithm of the fourth order with time step 0.05 and parameters $\sigma = 10$, $\rho = 28$ and $\beta = 8/3$. The integration was initialized with $x_0 = 0$, $y_0 = 1$ and $z_0 = 1.05$. The sampling period coincides with the integration step. The system was modelled using a second-order autoregressive model (17) with $\varphi_t = (y_{t-1}, y_{t-2}, 1)$. Its parameters were estimated using partial forgetting with hypotheses formulated in Section 3, where $\mathfrak{F}$ is the exponential forgetting with factor 0.95. The result of modelling of the first 500 samples is depicted in Fig. 1.

We can see, that after the learning period (approx. first 100 samples), the estimator stabilizes and the model achieves a good performance.

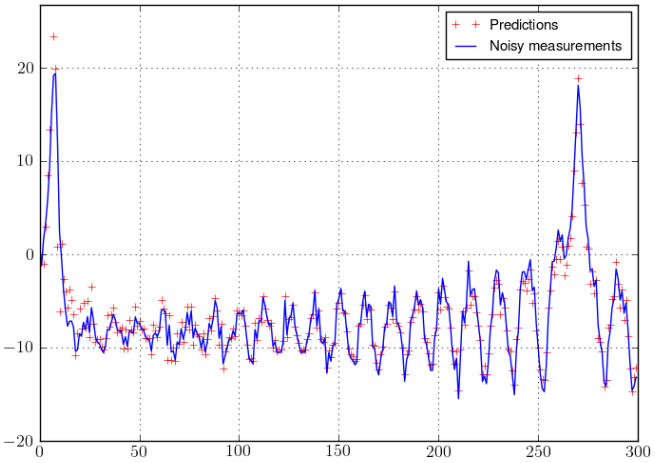

Fig. 1. Results of the numerical experiment.

7. CONCLUSION

The autoregressive model with partial forgetting within the Rao-Blackwellized particle filter was discussed. We presented a hybrid filtering method, where a subset of parameters is estimated using a particle filter. The rest of parameters is estimated conditionally linearly. The presented algorithm in its basic form performed well, however, there is a lot of space for further improvements.

8. REFERENCES

[1] H. Tong, *Non-linear time series: a dynamical system approach*, Oxford University Press, 1993.

[2] R.E. Kalman, "Physical and mathematical mechanisms of instability in nonlinear automatic control systems," *Trans. ASME*, vol. 79, no. 3, pp. 553–566, 1957.

[3] E. Sontag, "Nonlinear regulation: The piecewise linear approach," *Automatic Control, IEEE Transactions on*, vol. 26, no. 2, pp. 346–358, 2002.

[4] C. Savona, "Approximate nonlinear filtering for piecewise linear systems," *Systems & Control Letters*, vol. 11, no. 4, pp. 327–332, 1988.

[5] K. Judd and A. Mees, "On selecting models for nonlinear time series," *Physica D: Nonlinear Phenomena*, vol. 82, no. 4, pp. 426–444, 1995.

[6] J. Ragot, G. Mourot, and D. Maquin, "Parameter estimation of switching piecewise linear system," in *Decision and Control, 2003. Proceedings. 42nd IEEE Conference on*. IEEE, 2004, vol. 6, pp. 5783–5788.

[7] F. Rosenqvist and A. Karlström, "Realisation and estimation of piecewise-linear output-error models," *Automatica*, vol. 41, no. 3, pp. 545–551, 2005.

[8] A.E. Raftery, M. Kárný, J. Andrýsek, and P. Ettler, "Online Prediction under Model Uncertainty Via Dynamic Model Averaging: Application to a Cold Rolling Mill," 2007.

[9] J.M. Bernardo and A.F.M. Smith, "Bayesian theory," *Measurement Science and Technology*, vol. 12, pp. 221, 2001.

[10] V. Peterka, "Bayesian system identification," *Automatica*, vol. 17, no. 1, pp. 41–53, 1981.

[11] J. Karush, "On the Chapman-Kolmogorov equation," *The Annals of Mathematical Statistics*, vol. 32, no. 4, pp. 1333–1337, 1961.

[12] M. Kárný, *Optimized Bayesian Dynamic Advising: Theory and Algorithms*, Springer-Verlag New York Inc, 2006.

[13] K. Dedecius et al., "Parameter Estimation with Partial Forgetting Method," in *15th IFAC Symposium on System Identification*, 2009.

[14] S. Kullback and R.A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, pp. 79–86, 1951.

[15] A. Doucet, N. De Freitas, and N. Gordon, *Sequential Monte Carlo methods in practice*, Springer Verlag, 2001.

[16] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman filter: Particle filters for tracking applications*, Artech House Publishers, 2004.

[17] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statistics and computing*, vol. 10, no. 3, pp. 197–208, 2000.

[18] A.H. Jazwinski, *Stochastic processes and filtering theory*, Academic Pr, 1970.

[19] R. Kulhavý and M. Kárný, "Tracking of slowly varying parameters by directional forgetting," in *9th TFAC World Congress, Budapest, Hungary*, 1984.

[20] E.N. Lorenz, "Deterministic nonperiodic flow," *Journal of Atmospheric Sciences*, vol. 20, pp. 130–141, 1963.