BAYESIAN IDENTIFICATION OF NON-STATIONARY AR MODEL WITH UNKNOWN FORGETTING FACTOR

Václav Šmídl

UTIA, Czech Academy of Sciences,
Prague, Czech Republic
smidl@utia.cas.cz

Anthony Quinn

Trinity College Dublin,
Ireland
aquinn@tcd.ie

ABSTRACT

In this paper, we study Bayesian identification of the non-stationary parameters of the AR process. It is traditionally achieved via forgetting. Numerically efficient solution is available if the forgetting factor is known a priori. In this paper, we propose a joint Bayesian identification of the AR parameters and the unknown forgetting factor. The resulting intractable posterior distribution is approximated using Variational-Bayes method. Illustration of the method on simple simulated data is presented.

1. INTRODUCTION

Identification of the AR model is analytically tractable only for stationary parameters. This assumption is rarely met in practice. In many applications, however, a complete model of parameter variations is not known. The problem is then under-determined, obviating the full analytical solution and leading to many heuristic techniques. The standard batch (off-line) algorithm uses windowing [1]. Alternatively, the concept of forgetting [2] is used in adaptive signal processing and recursive estimation. A Bayesian treatment of forgetting was developed in [3]. Real-time identification of non-stationary AR process is achieved, if the forgetting factor is known *a priori*. In this paper, we relax this constraint and develop a Bayesian on-line inference of *both* the AR parameters and the forgetting factor.

2. BAYESIAN IDENTIFICATION OF STATIONARY AR MODEL

A scalar univariate AR system is usually considered to be of the form

$$x_n = -\sum_{k=1}^{p} a_k x_{n-k} + \sigma e_n \tag{1}$$

where e_n denotes the input and x_n the output of the system, σ and $\mathbf{a} = [a_1, \ldots, a_p]'$ are parameters of the model. The classical solution to this problem is based on the Wiener MMSE criterion [4]. Parameter estimates are obtained by solution of the so-called *normal equations*. The Bayesian approach to this problem is based on the assumption that system (1) is driven by white noise with Gaussian distribution $f(e_n) = \mathcal{N}(0,1)$. The output of the model, x_n, is a random variable with conditional distribution

$$f(x_n|\mathbf{a}, \sigma, \mathbf{x}_n) = \mathcal{N}\left(-\mathbf{a}'\mathbf{x}_n, \sigma^2\right), \tag{2}$$

where $\mathbf{x}_n = [x_{n-1} \ldots x_{n-p}]$ is the vector of regressors. The task is to identify posterior distribution of system parameters $\theta = [\mathbf{a}', \sigma]'$ at each time n. The Bayes' rule is used to transform the posterior distribution of parameters at time $n-1$, into posterior distribution,

$$f(\theta|X_n) \propto f(x_n|\theta, X_{n-1})\, f(\theta|X_{n-1}), \tag{3}$$

where $X_n = [x_1, \ldots, x_n]'$ denotes history of data at time n. If a prior $f(\cdot)$ is chosen as conjugate with respect to the observation model (2) [5], then the functional form of the prior and posterior are identical.

The model (2) belongs to the exponential family, and so both a conjugate prior and sufficient statistics are available, as previously stated. The conjugate distribution of parameters for (2) is of the Normal-inverse-Gamma ($\mathcal{GW}$) type [5]:

$$\mathcal{GW}_{\mathbf{a},\sigma}(V,\nu) \equiv \frac{\left|\sigma^2\right|^{-0.5\nu}}{\zeta_{\mathcal{GW}}(V,\nu)} \times \exp\left\{-\frac{1}{2}\sigma^{-2}\left[-1, \mathbf{a}'\right] V \left[-1, \mathbf{a}'\right]'\right\}, \tag{4}$$

$$\zeta_{\mathcal{GW}}(V,\nu) = \Gamma(0.5\nu)\, \lambda^{-0.5\nu}\, |V_{\mathbf{aa}}|^{-0.5}\, 2^{0.5p}, \tag{5}$$

$$V = \begin{bmatrix} V_{11} & V'_{\mathbf{a}1} \\ V_{\mathbf{a}1} & V_{\mathbf{aa}} \end{bmatrix}, \quad \lambda = V_{11} - V'_{\mathbf{a}1} V_{\mathbf{aa}}^{-1} V_{\mathbf{a}1}, \tag{6}$$

where (6) denotes partitioning of $V \in \Re^{(p+1)\times(p+1)}$ into blocks and V_{11} is the $(1,1)$ element. V, ν are the sufficient statistics of $\mathcal{GW}_{\mathbf{a},\sigma}(\cdot)$.

The statistics of the conjugate prior distribution, V_0, ν_0, are chosen to reflect our initial knowledge of parameters. If we do not have any preference, we use a very flat (non-committal) distribution. Typically $V_0 = \varepsilon I_{p+1}$, $\nu_0 = \varepsilon$, where I_{p+1} is the $(p+1) \times (p+1)$ identity matrix, and ε is a small positive scalar. Substituting (2) into (3) and invoking (4) at time $n-1$, then the posterior distribution at time $n > p$ is

$$f(\mathbf{a}, \sigma | X_n) = \mathcal{GW}_{\mathbf{a},\sigma}(V_n, \nu_n), \tag{7}$$

$$V_n = V_{n-1} + \begin{bmatrix} x_n \\ \mathbf{x}_n \end{bmatrix} [x_n, \mathbf{x}_n'] = V_{n-1} + \bar{\mathbf{x}}_n \bar{\mathbf{x}}_n' \tag{8}$$

$$\nu_n = \nu_{n-1} + 1. \tag{9}$$

Here, $\bar{\mathbf{x}}_n = [x_n, \mathbf{x}_n']'$ is the extended regression vector. $\bar{\mathbf{x}}_n \bar{\mathbf{x}}_n'$ will be called a *dyad* in this paper. Since the recursion begins at $n = p+1$, V_p and ν_p are chosen to be $V_p = V_0$ and $\nu_p = \nu_0$. It is equivalent to choosing the distribution on parameters to be stationary for $n \leq p$.

The moments of (7) are:

$$\widehat{\mathbf{a}}_n = V_{\mathbf{aa};n}^{-1} V_{\mathbf{a}1;n}, \tag{10}$$

$$\widehat{\sigma_n^2} = \frac{\lambda_n}{\nu_n - p + 2}, \tag{11}$$

$$\widehat{\ln|\sigma_n|} = -\frac{1}{4}\psi\left(\frac{1}{2}(\nu_n - p - 1)\right) + \frac{1}{4}\ln|\nu_n| \tag{12}$$

where $\widehat{\mathbf{a}_n}$ denotes the expected value of variable $\mathbf{a}$ with respect to distribution $f(\mathbf{a}|X_n)$.

3. NON-STATIONARY AR MODEL WITH KNOWN FORGETTING FACTOR

The assumption of constant parameter values is rarely met in practice. In many applications, however, a complete model of parameter variations is not known. The problem is then under-determined, obviating the full Bayesian solution and leading to many heuristic techniques. The standard batch (off-line) algorithm uses windowing [1]. Alternatively, the concept of forgetting [2] is used in adaptive signal processing [6] and recursive estimation [7].

A Bayesian treatment of forgetting was developed in [3]. There, the missing model of parameter evolution is handled via a probabilistic operator:

$$f(\theta_n | X_{n-1}, \phi_n) \propto \left[f(\theta_{n-1}|X_{n-1})_{\theta_n}\right]^{\phi_n} \times \tilde{f}(\theta_n|X_{n-1})^{1-\phi_n} \tag{13}$$

The notation $f(\cdot)_{\theta_n}$ indicates the replacement of the argument of $f(\cdot)$ by θ_n, where θ_n is the time-variant unknown parameter set at time n. $\tilde{f}(\cdot)$ is a chosen (known) alternative distribution, expressing auxiliary knowledge about θ_n at time n. This is an optimal criterion for defining $f(\theta_n|X_{n-1}, \phi_n)$ in the sense of Kullback-Leibler distance minimization from of the true posterior [3]. Coefficient ϕ_n, $0 \leq \phi_n \leq 1$ is known as a forgetting factor.

In this paper, the $\mathcal{GW}$ distribution with parameters $\tilde{V}, \tilde{\nu}$ is used as the alternative. It is typically chosen as a flat distribution, e.g. with the same parameter values as the prior: $\tilde{V} = V_0, \tilde{\nu} = \nu_0$. The $\mathcal{GW}$ conjugate family (4) is closed under the convex combining (i.e. geometric mean) in (13) yielding another member of the same family. Substituting (13) and (2) into (**??**) yields the following recursive update of the $\mathcal{GW}$ statistics:

$$V_n(\phi_n) = \phi_n V_{n-1} + \bar{\mathbf{x}}_n \bar{\mathbf{x}}_n' + (1 - \phi_n)\tilde{V} \tag{14}$$

$$\nu_n(\phi_n) = \phi_n \nu_{n-1} + 1 + (1 - \phi_n)\tilde{\nu}. \tag{15}$$

When $\phi_n = 1$, the update is identical to the stationary equations (8,9).

4. NON-STATIONARY AR MODEL WITH UNKNOWN FORGETTING FACTOR

The concept of forgetting described in the previous section allows for non-stationary forgetting factor, however, it does not provide any guidance for its choise. In this Section, we introduce a Bayesian inference of the unknown forgetting factor.

Following the Bayesian methodology, the unknown forgetting factor, ϕ_n, can be treated as random variable. The task is then to estimate posterior distribution $f(\theta_n, \phi_n | X_n)$. This is formally achieved using the Bayes' rule

$$f(\theta_n, \phi_n | X_n) \propto f(x_n|\theta_n, \phi_n) f(\theta_n | X_{n-1}, \phi_n) f(\phi_n). \tag{16}$$

(16) follows from (3) using (2) and (13). Note that the process elicitate a new random variable ϕ_n at each time. Therefore, in each time we need to impose a prior on it. We have chosen the prior to be as uninformative as possible, i.e. uniform prior on the interval $[0, 1]$, $f(\phi_n) = \mathcal{U}([0, 1])$. Posterior distribution (16) is not analytically tractable. Therefore, we seek a suitable approximation. In this paper, we seek the best approximation in the class of independent distributions

$$f(\theta_n, \phi_n | X_n) \approx \breve{f}(\theta_n | X_n) \breve{f}(\phi_n | X_n), \tag{17}$$

where $\breve{f}(\cdot)$ denotes a functional (i.e. arbitrary parameterized) distribution. This assumption of conditional independence is the core of approximation known as Variational Bayes. The method is summarized by the following theorem:

4.1. Variational Bayes approximation

Theorem 1 (Variational Bayes) *Let $f(\theta_1, \theta_2 | X)$ be a posterior pdf of parameters θ_1, θ_2, given data X. Let $\overline{f}(\theta_1, \theta_2 | X)$*

be an approximate pdf restricted to the set of conditionally independent distributions on θ_1, θ_2:

$$\breve{f}(\theta_1, \theta_2|X) = \breve{f}_1(\theta_1|D)\,\breve{f}_2(\theta_2|D). \quad (18)$$

Then, the minimum of the KL distance,

$$\overline{f}_1(\theta_1|D), \overline{f}_2(\theta_2|D) = \arg\min_{\breve{f}_1,\breve{f}_2} KL\left(\breve{f}(\theta_1,\theta_2|D)\,||f(\theta_1,\theta_2|D)\right), \quad (19)$$

is reached for

$$\overline{f}_1(\theta_1|D) \propto \exp\left(\mathsf{E}_{\theta_2|D}(\ln(f(\theta_1,\theta_2,D)))\right), \quad (20)$$
$$\overline{f}_2(\theta_2|D) \propto \exp\left(\mathsf{E}_{\theta_1|D}(\ln(f(\theta_1,\theta_2,D)))\right). \quad (21)$$

We will refer to (20) as the VB-optimal *posterior. Parameters of the posterior distributions (20) will be called* VB-statistics*.*

The main computational problem of the VB approximation is that the Variational Extreme (20) is not given in closed-form. For example, the moments of $\overline{f}_1(\cdot)$, are needed for evaluation of $\overline{f}_2(\cdot)$, and vice-versa. Solution of (20) is therefore usually found via iterative algorithm that is suggestive of the EM algorithm [8].

Algorithm 4.1 (Variational EM) *Statistics of the posterior distributions (20) can be found iteratively. Cyclic iteration of the following steps converge to a solution of (20).*

***E-step**: compute approximate distribution of parameter θ_2 at iteration i:*

$$\overline{f}^{(i)}(\theta_2|X) \propto \exp\int_{\theta_1} \overline{f}^{(i-1)}(\theta_1|X)\ln f(\theta_1,\theta_2,X)\,d\theta_1. \quad (22)$$

***M-step**: using approximate distribution from the ith E-step compute approximate distribution of parameter θ_1 at iteration i:*

$$\overline{f}^{(i)}(\theta_1|D) \propto \exp\int_{\theta_2} \overline{f}^{(i)}(\theta_2|X)\ln f(\theta_1,\theta_2,X)\,d\theta_2. \quad (23)$$

Convergence of the algorithm was proven in [9] via natural gradient technique.

4.2. VB-optimal Posteriors

We intend to use the Variational Bayes method to approximate each step of the on-line update (16). This implies that the distributions of parameteres at time $n-1$ are already known and fixed. Variational Bayes is, however, a functional optimization, therefore, it is not guaranteed, that the posterior distributions at time n will be of the same form as those in time $n-1$. Then, the principle of conjugacy would be violated. Hence, it is essential to choose the approximating distributions from such a family that is closed under the VB approximation. {VB: This can be done via VB-conjugacy, but we have no space for thuis}

For the time being, let us choose

$$\overline{f}(\theta_n|X_{n-1}) = \mathcal{GW}(V_{n-1},\nu_{n-1}), \quad (24)$$
$$\tilde{f}(\theta_n|X_{n-1}) = \mathcal{GW}\left(\tilde{V},\tilde{\nu}\right). \quad (25)$$

The distribution of the forgetting factor ϕ_n is chosen as independent of the previously observed data (16). Therefore, conjugacy is not important in this case.

Substituting (24) and (25) into (13), and using the result in Theorem 1, the VB-optimal form of $\overline{f}(\theta_n|X_n)$ is found via the following assignments:

$$\overline{f}(\theta_n|X_n) = \mathcal{GW}(V_n,\nu_n), \quad (26)$$
$$V_n = \widehat{\phi}_n V_{n-1} + \bar{\mathbf{x}}_n\bar{\mathbf{x}}_n' + \left(1-\widehat{\phi}_n\right)\tilde{V} \quad (27)$$
$$\nu_n = \widehat{\phi}_n\nu_{n-1} + 1 + \left(1-\widehat{\phi}_n\right)\tilde{\nu}, \quad (28)$$

where $\widehat{\phi}_n$ denotes the expected value of $\overline{f}(\phi_n|X_n)$. However, the form of the VB-optimal distribution $\overline{f}(\phi_n|X_n)$ is very complicated and analytically intractable. The complications arise from evaluation of expectations of the term $\zeta_{\mathcal{GW}}(V(\phi_t),\nu(\phi_t))$. Therefore, we seek its suitable approximation:

$$\zeta(\phi_n) \approx \zeta_{\mathcal{GW}}(V(\phi_t),\nu(\phi_t)). \quad (29)$$

Note that we can evaluate any value of this term given ϕ_n. Specifically, we have already evaluated this term for $\phi_n = 0$, and $\phi_n = 1$:

$$\zeta_{\mathcal{GW}}(V(0),\nu(0)) = \zeta_{\mathcal{GW}}\left(\tilde{V},\tilde{\nu}\right), \quad (30)$$
$$\zeta_{\mathcal{GW}}(V(1),\nu(1)) = \zeta_{\mathcal{GW}}(V_{t-1},\nu_{t-1}), \quad (31)$$

Therefore we propose to interpolate the function between these two values as follows.

Proposition 1 (Approximation of Normalizing Constant) *Let us choose the approximation of (29) in the following form*

$$\zeta(\phi_n) = \exp(h_1 + h_2\phi_n), \quad (32)$$

where h_1 and h_2 are unknown constants. Matching (32) at points (30) and (31) we obtain:

$$h_1 = \ln\zeta_{\mathcal{GW}}\left(\tilde{V},\tilde{\nu}\right), \quad (33)$$
$$h_2 = \ln\zeta_{\mathcal{GW}}(V_{n-1},\nu_{n-1}) - \ln\zeta_{\mathcal{GW}}\left(\tilde{V},\tilde{\nu}\right). \quad (34)$$

Under this approximation, the first VB-approximation (26) is not affected, however, the second of the VB-approximations, i.e. $\overline{f}\left(\phi_n|X_n\right)$ is greatly simplified:

$$\overline{f}\left(\phi_n|X_n\right) \approx \mathcal{E}xp\left(b,[0,1]\right), \tag{35}$$

$$\begin{aligned} b = (\nu_{n-1}-\tilde{\nu})\,\widehat{\ln|\sigma_n|} - \frac{1}{2}\mathrm{tr}\left(\left(V_{\mathbf{aa},n-1}-\tilde{V}_{\mathbf{aa}}\right)\tilde{V}^{-1}_{\mathbf{aa},n}\right) \\ - \ln\zeta_{\mathcal{GW}}\left(V_{n-1},\nu_{n-1}\right) + \ln\zeta_{\mathcal{GW}}\left(\tilde{V},\tilde{\nu}\right) \\ - \frac{1}{2}\mathrm{tr}\left(\left(V_{n-1}-\tilde{V}\right)[-1,\widehat{\mathbf{a}}'_n]'\,\widehat{\sigma_n^{-2}}\,[-1,\widehat{\mathbf{a}}'_n]\right). \end{aligned} \tag{36}$$

Here, $\mathcal{E}xp\left(b,[0,1]\right)$ denotes the exponential distribution with parameter b with support restricted to interval $[0,1]$. Distribution (35), completes the analytical formulation of the VB-optimal approximation of the distribution (16).

Moments of the involved distributions—needed for evaluation of the posterior VB-statistics via the VEM algorithm—are given as follows. Moments of the $\mathcal{GW}$ distribution, namely $\widehat{\mathbf{a}_n}$, $\widehat{\sigma_n^2}$, $\widehat{\ln|\sigma_n|}$ are given by (10)–(12) respectively. Only the first moment, i.e. the mean value of the posterior distribution of the forgetting factor is required

$$\widehat{\phi_n} = \frac{\exp\left(b\right)(1-b)-1}{b\left(1-\exp\left(b\right)\right)}.$$

The main drawback of the VEM algorithm is that convergence to a solution is not assured in given number of steps. This is especially disturbing for on-line processing. This problem was addresed in [9] for stationary systems. It was proven that using only one iteration of the VEM algorithm in each step, in assymptotics, the algorithm also converge to the true value. However, no such result is available for non-stationary systems. In this paper, we experimentally compare the evaluation of the full VEM algorithm with the *on-line VEM* (i.e. the VEM algorithm with number of iterations fixed to two, at each time).

5. SIMULATED EXPERIMENT

A univariate second-order stable AR model (i.e. $\bar{\mathbf{x}}_n = [x_{n-1}, x_{n-2}]$) with parameters $\sigma = 1$, and

$$\mathbf{a} = \begin{cases} [-1.8, 0.98] & \text{if } \mathrm{mod}\left(n,30\right) = \mathrm{mod}\left(n,60\right) \\ [0.29, 0.98] & \text{if } \mathrm{mod}\left(n,30\right) \neq \mathrm{mod}\left(n,60\right) \end{cases},$$

where $\mathrm{mod}\left(\cdot\right)$ is modulo function, i.e. remainder after division. System was estimated using the VEM algorithm (Algorithm 4.1) on posterior distributions (26) and (35). We have chosen alternative in the form of

$$\tilde{V} = \mathrm{diag}\left([1, 0.001, 0.001]\right), \quad \tilde{\nu} = 10, \tag{37}$$

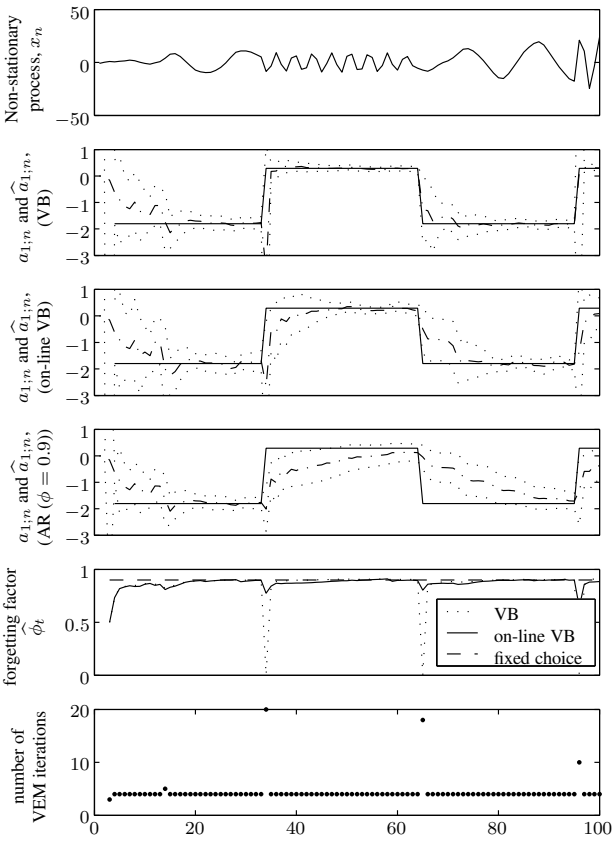

Fig. 1. Results of estimation of non-stationary process with time-variant forgetting. In sub-figures (i)–(iii), full lines denote simulated values of parameters, dashed lines denote posterior expected values, and dotted lines denote uncertainty bounds.

corresponding to initial guess $\mathbf{a}_0 = [0,0]$, $\mathrm{var}\left(a_1\right) = \mathrm{var}\left(a_2\right) = 1000$, $\sigma_0 = 10$. Prior distribution is equal to the alternative distribution. Initial value of the forgetting factor in the VEM algorithm was chosen as $\phi_n^{(1)} = 0.7$. The full VEM algorithm was stopped when $\left|\phi_n^{(m)} - \phi_n^{(m-1)}\right| < 0.001$, and number of iteration m to reach convergence was recorded. The restricted VEM algorithm was stopped after two steps. Results of identification are displayed in Figure 1.

Note that the method—within one observation—correctly detects the change of parameters and estimates forgetting factor as low as $\widehat{\phi}_{n=33} = 0.05$, which causes almost instant replacement of statistics V_n, ν_n by alternative (prior) values $\tilde{V}, \tilde{\nu}$. Thus, identification process is restarted. Note that number of iterations of the VEM algorithm is significantly higher at the point of change. Therefore, at these points, the expected value of forgetting factor, $\widehat{\phi}_n$, obtained using the restricted VEM, is still too high compared to the converged value of the full VEM (Figure 1). For comparison, results

of identification with stationary forgetting $\phi_n = 0.9, \forall n$ are displayed in Figure 1. The best parameter tracking is achieved using the VB posterior distributions evaluated via VEM with full iterations. Identification of the process using VB posteriors evaluated via VEM with restricted iterations is acceptable if the parameter variations are not too rapid.

6. DISCUSSION

The technique of forgetting is used in many estimation methods for non-stationary processes [7], where the forgetting factor is considered time-invariant and known. Attempts to relax the assumption of *a priori* known forgetting factor were made, especially for the Recursive Least Square (RLS) algorithms [7]. The method presented in [10] is the closest to our approach. It is a gradient based estimation of the forgetting factor for the RLS algorithm. We note the following differences: (i) The RLS algorithm is based on the assumption of Normal distribution of parameters. The Bayesian interpretation of forgetting can be applied to any class of posterior distributions that is closed under the geometric mean (13), such as the Markov model, (ii) we minimize the KL distance of posterior distribution at each time n. This allows for rapid changes (i.e. switching) of the model parameters. The criterion of asymptotic mean square error minimized in [10] is addressing slower variations of parameters, and (iii) the posterior estimate of the forgetting factor is sensitive to the chosen alternative pdf of the parameters (13). It plays similar role as the tuning parameters (α and β) in [10]. The alternative distribution can be chosen using the available expert knowledge of the problem. The tuning parameters of [10] must be adjusted experimentally.

We have noted that the optimal posterior distribution of the forgetting factor is not tractable and it must be approximated to achieve a numerically efficient identification algorithm (Proposition 1). The choice of approximation, of course, influences the quality of results of the inference algorithm. The proposed approximation is rather simple and it may be inappropriate for certain tasks. Other approximations might be investigated in such cases to improve performance.

For on-line processing, the maximum number of iteration of the VEM algorithm in each step must be restricted according to available computational power. However, if the number of iteration is chosen too low, performance of the algorithm may be degraded (see Figure 1).

7. CONCLUSION

In this paper, we have studied the Bayesian identification of non-stationary parameters of the AR process with unknown forgetting factor. We proposed to approximate the intractable posterior distribution by conditionally independent approximation, and developed the associated identification algorithm. Results of identification of simulated studies demonstrate improved tracking abilities of the new method, compared to stationary forgetting. This was achieved at the price of higher computational load of the new algorithm. Further work is needed to achieve faster convergence of the iterative evaluation of parameters of the posterior distributions.

8. REFERENCES

[1] R. H. Middleton, G. C. Goodwin, D. J. Hill, and D. Q. Mayne, "Design issues in adaptive control," *IEEE Transactions on Automatic Control*, vol. 33, no. 1, pp. 50–58, 1988.

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

[3] R. Kulhavý and M.B.Zarrop, "On general concept of forgetting," *International Journal of Control*, vol. 58, no. 4, pp. 905–924, 1993.

[4] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[5] J.M. Bernardo and A.F.M. Smith, *Bayesian Theory*, John Wiley & Sons, Chichester, New York, Brisbane, Toronto, Singapore, 1997, 2nd edition.

[6] G. V. Moustakides, "Locally optimum adaptive signal processing algorithms," *IEEE Transactions on Signal Processing*, vol. 46, no. 12, pp. 3315–3325, 1998.

[7] L. Ljung and T. Söderström, *Theory and practice of recursive identification*, MIT Press, Cambridge; London, 1983.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistical Society, Series B*, vol. 39, pp. 1–38, 1977.

[9] M. Sato, "Online model selection based on the variational bayes," *Neural Computation*, vol. 13, pp. 1649–1681, 2001.

[10] C.-F. So, S. C. Ng, and S. H. Leung, "Gradient based variable forgetting factor rls algorithm," *Signal Processing*, vol. 83, pp. 1163–1175, 2003.