THE VARIATIONAL BAYES APPROXIMATION IN BAYESIAN FILTERING

Václav Šmídl

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

Anthony Quinn

Trinity College Dublin,
Ireland
aquinn@tcd.ie

ABSTRACT

The Variational Bayes (VB) approximation is applied in the context of Bayesian filtering, yielding a tractable on-line scheme for a wide range of non-stationary parametric models. This VB-filtering scheme is used to identify a Hidden Markov Model with an unknown non-stationary transition matrix. In a simulation study involving soft-bit data, reliable inference of the underlying binary sequence is achieved *in tandem* with estimation of the transition probabilities. The performance compares favourably with a proposed particle filtering approach, and at lower computational cost.

1. INTRODUCTION

Bayesian filtering is the formal framework for on-line identification of parametric non-stationary processes. The resulting computations are tractable only for a limited class of models, such as the Gaussian model assumption of the Kalman filter. Particle filtering is currently the approximation of choice for more realistic models, but the computational overheads are often prohibitive. The Variational Bayes (VB) approximation is an attractive deterministic alternative. In Section 2, we review Bayesian filtering, and in Section 3 we develop its VB variant, *i.e.* VB-filtering. The method is applied to the inference of Hidden Markov Models (HMMs) in Section 4 and its performance is compared to particle filtering via simulations.

2. BAYESIAN FILTERING

Consider an observation process, $d_t \in \mathbb{R}^p$, $t = 1, 2, ...$, for which the parametric *observation model*, $f(d_t|\theta_t, D_{t-1})$, is an explicitly time-varying function. Hence, new parameters, $\theta_t \in \mathbb{R}^r$(of fixed dimension, r), are required to explain new data, d_t. The *parameter evolution model* is expressed via $f(\theta_t|\theta_{t-1})$. In Bayesian filtering, we are interested in *on-line* identification of θ_t, $\forall t$, via the *filtering distribution*, $f(\theta_t|D_t)$ [1, 2], where $D_t = [D_{t-1}, d_t]$ is the matrix of accumulated data, and $D_0 \equiv \{\}$. Bayes' rule is used to update knowledge of θ_t in the light of new data, d_t:

$$f(\theta_t|D_t) \propto f(d_t|\theta_t, D_{t-1})\, f(\theta_t|D_{t-1}). \quad (1)$$

The parameter predictor is obtained by marginalization**:**

$$f(\theta_t|D_{t-1}) = \int f(\theta_t|\theta_{t-1}, D_{t-1})\, f(\theta_{t-1}|D_{t-1})\, d\theta_{t-1}. \quad (2)$$

This two-step update for Bayesian on-line inference of θ_t is known as *Bayesian filtering* [2]. Kalman filtering is a special case, where the observation model and parameter evolution model are linear Gaussian [3].

Tractable recursive evaluation of (1) and (2), $\forall t$, requires functional invariance of $f(\theta_t|D_t)$ under both Bayes rule (1) and marginalization (2). This can be achieved in only a limited class of models [4]. The linear Gaussian models of Kalman filtering are one of the few such cases. Clearly, therefore, approximations are necessary. Sequential Monte Carlo methods (particle filtering) [2] are currently popular, but many other approaches have been proposed in the literature. A comprehensive review is available [1]. The use of the VB approximation [5] in Bayesian filtering was developed in [6], and is reviewed next.

3. VARIATIONAL BAYESIAN (VB) FILTERING

The VB-approximation imposes conditional independence between parameters of a joint distribution. In the context of Bayesian filtering, we impose conditional independence between θ_t and θ_{t-1}:

$$f(\theta_t, \theta_{t-1}|D_t) \approx \tilde{f}(\theta_t|D_t)\, \tilde{f}(\theta_{t-1}|D_t). \quad (3)$$

These VB-marginals are found by functional optimization:

$$\tilde{f}(\theta_t|D_t), \tilde{f}(\theta_{t-1}|D_t) = \arg\min KL\left(f(\theta_t|D_t)\, f(\theta_{t-1}|D_t)\, ||f(\theta_t, \theta_{t-1}|D_t)\right), \quad (4)$$

where $KL(\cdot)$ denotes Kullback-Leibler divergence [7]. The *VB-marginals* have the following functional form [5]:

$$\tilde{f}(\theta_t|D_t) \propto \exp\left(\mathsf{E}_{\tilde{f}(\theta_{t-1}|D_t)}\left[\ln\left(f(\theta_t, \theta_{t-1}, D_t)\right)\right]\right), \quad (5)$$

$$\tilde{f}(\theta_{t-1}|D_t) \propto \exp\left(\mathsf{E}_{\tilde{f}(\theta_t|D_t)}\left[\ln\left(f(\theta_t, \theta_{t-1}, D_t)\right)\right]\right). \quad (6)$$

(5) and (6) are typically solved iteratively, via the *Iterative VB (IVB) algorithm*, where each VB-marginal is updated cyclically, by substitution of the moments of the other VB-marginal. These necessary moments are known as the *VB-moments* [6].

The joint distribution required in (5) and (6) is:

$$f\left(d_t, \theta_t, \theta_{t-1} | D_{t-1}\right) = f\left(d_t | \theta_t, D_{t-1}\right) f\left(\theta_t | \theta_{t-1}\right) \times \tilde{f}\left(\theta_{t-1} | D_{t-1}\right). \quad (7)$$

Here, the filtering distribution at time $t-1$ has already been replaced by its VB-approximation, $\tilde{f}\left(\theta_{t-1} | D_{t-1}\right)$. Substituting (7) into (5) and (6), then

$$\tilde{f}\left(\theta_t | D_t\right) \propto f\left(d_t | \theta_t, D_{t-1}\right) \exp\left\{\mathsf{E}_{\tilde{f}\left(\theta_{t-1} | D_t\right)}\left[\ln f\left(\theta_t | \theta_{t-1}\right)\right]\right\}, \quad (8)$$

$$\tilde{f}\left(\theta_{t-1} | D_t\right) \propto \exp\left\{\mathsf{E}_{\tilde{f}\left(\theta_t | D_t\right)}\left[\ln f\left(\theta_t | \theta_{t-1}\right)\right]\right\} \tilde{f}\left(\theta_{t-1} | D_{t-1}\right). \quad (9)$$

(8) and (9) are known as the *VB-filtering* and *VB-smoothing* distributions respectively. The two steps of Bayesian filtering, (1) and (2), have therefore been replaced by the following two parallel Bayes' rule updates:

$$\tilde{f}\left(\theta_t | D_t\right) \propto f\left(d_t | \theta_t, D_{t-1}\right) \tilde{f}\left(\theta_t | D_{t-1}\right), \quad (10)$$

$$\tilde{f}\left(\theta_{t-1} | D_t\right) \propto \tilde{f}\left(d_t | \theta_{t-1}, D_{t-1}\right) \tilde{f}\left(\theta_{t-1} | D_{t-1}\right). \quad (11)$$

These involve the following *VB-parameter predictor* and *VB-observation model* respectively:

$$\tilde{f}\left(\theta_t | D_{t-1}\right) \propto \exp\left\{\mathsf{E}_{\tilde{f}\left(\theta_{t-1} | D_t\right)}\left[\ln f\left(\theta_t | \theta_{t-1}\right)\right]\right\}, \quad (12)$$

$$\tilde{f}\left(d_t | \theta_{t-1}, D_{t-1}\right) \propto \exp\left\{\mathsf{E}_{\tilde{f}\left(\theta_t | D_t\right)}\left[\ln f\left(\theta_t | \theta_{t-1}\right)\right]\right\}. \quad (13)$$

Each is generated by substitution of VB-moments from (10) and (11) respectively into the parameter evolution model. Several cycles of the IVB algorithm are typically required to achieve convergence at each time, t. This communication of moments between the VB-marginals compensates for the removal of posterior correlation between θ_t and θ_{t-1} (3). Together (10) and (11) define *VB-filtering,* as illustrated in Fig. 1.

Note that the same functional form for the VB-filtering distribution, $\tilde{f}\left(\theta_t | D_t\right)$, is recovered at each time step, t, since $\tilde{f}\left(\theta_{t-1} | D_{t-1}\right)$ propagates through the scheme only via substitution of the VB-moments from $\tilde{f}\left(\theta_{t-1} | D_t\right)$ (Fig. 1). This framework therefore greatly extends the class of models for which tractable recursive inference is possible.

4. BAYESIAN FILTERING FOR HMMS

Consider a *Hidden Markov Model (HMM)* [8] on an unobserved discrete (label) variable, l_t, with c possible states. For

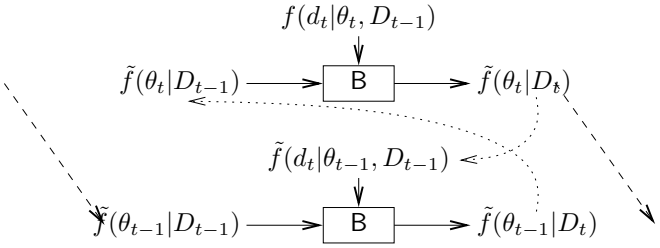

Fig. 1. VB-filtering, showing the flow of VB-moments (dotted arrows) via IVB cycles. B denotes a Bayes' rule update.

analytical convenience, we denote each state of l_t by a c-dimensional elementary basis vector, $\epsilon_c(i)$; *i.e.* $l_t \in \{\epsilon_c(1), \ldots, \epsilon_c(c)\}$. $\epsilon_c(i)$ is a length-c vector with ith element equal to 1, zero elsewhere. Let T_t denote the $c \times c$ unknown, time-variant transition matrix, with ijth element

$$t_{i,j,t} = \Pr\left[l_t = \epsilon_c(i) \,|\, l_{t-1} = \epsilon_c(j)\right].$$

T_t is modelled as a random walk process; *i.e.* the expected value of T_t is set equal to T_{t-1}. Consider the special case where this HMM is observed via a c-dimensional *observation process*, $d_t \in (0,1)^c$, $\sum_{i=1}^{c} d_{i,t} = 1$. This represents empirical probabilities of each state at time t, such as normalized counts, the output of a classifier, a soft-bit sequence, *etc*. The task is then to infer l_t from the observations, d_t.

The following Bayesian filtering models are consistent with the assumptions above:

$$f\left(l_t | l_{t-1}, T_t\right) = \mathcal{M}u_{l_t}\left(T_t l_{t-1}\right), \quad (14)$$

$$f\left(T_t | T_{t-1}\right) = \mathcal{D}i_{T_t}\left(\kappa T_{t-1} + \mathbf{1}_{c,c}\right), \quad (15)$$

$$f\left(d_t | l_t\right) = \mathcal{D}i_{d_t}\left(\rho l_t + \mathbf{1}_{c,1}\right). \quad (16)$$

Here, $\mathcal{M}u(\cdot)$ denotes the Multinomial distribution, and $\mathcal{D}i(\cdot)$ denotes the Dirichlet distribution [6]. In (15), κ controls the bandwidth of the process, T_t: *i.e.* greater values of κ favour T_t closer to T_{t-1}. In (16), ρ controls the uncertainty in inferring l_t via d_t. For large values of ρ, the observed data, d_t, have higher probability of being close to the actual labels, l_t (see Fig. 2). $\mathbf{1}$ denotes a matrix of ones, of the stated dimension.

Exact inference for (14)–(16) via Bayesian filtering is not feasible, since the required summation (from (2)) over the c states of l_t, $\forall t$, leads to an exponential growth of terms in $f(l_t, T_t | D_t)$. We overcome this problem via VB-filtering.

4.1. VB-filtering for the HMM (unknown T_t)

The log of the evolution model for $\theta_t = [l_t, T_t]$ is

$$\ln f\left(\theta_t | \theta_{t-1}\right) = \ln f\left(l_t, T_t | l_{t-1}, T_{t-1}\right) = l_t' \ln T_t l_{t-1} + \mathrm{tr}\left(\kappa \ln T_t' T_{t-1}\right), \quad (17)$$

where tr denotes the trace of the matrix, and $\ln T_t$ denotes the matrix of log-elements. VB-moments induced by this joint

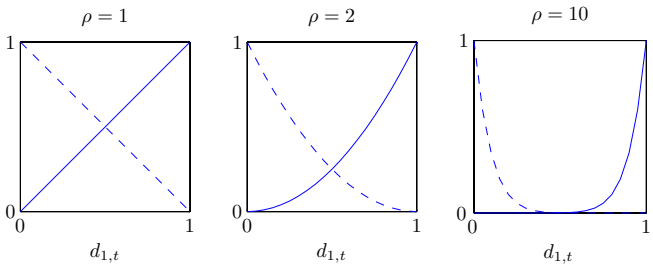

Fig. 2. The Dirichlet distribution, $f\left(d_t|l_t\right)$ (16), for $c = 2$ and $\rho = \{1, 2, 10\}$, illustrated via its scalar conditional distribution, $f\left(d_{1,t}|l_t = \epsilon_2(1)\right)$ (full) and $f\left(d_{1,t}|l_t = \epsilon_2(2)\right)$ (dashed).

evolution model are difficult to evaluate. Therefore, we extend the conditional independence assumption (3) to l_t and T_t. Under this assumption, the VB-parameter predictors and VB-observation model have the following form:

$$\tilde{f}\left(l_t|D_{t-1}\right) \propto \exp\left(l_t'\widehat{\ln T_t}\widehat{l_{t-1}}\right),$$
$$\tilde{f}\left(T_t|D_{t-1}\right) \propto \exp\left(\widehat{l_t}'\ln T_t\widehat{l_{t-1}} + \mathrm{tr}\left(\kappa \ln T_t'\widehat{T_{t-1}}\right)\right),$$
$$\tilde{f}\left(d_t|l_{t-1}, T_{t-1}\right) \propto \exp\left(l_{t-1}\widehat{\ln T_t}'\widehat{l_t} + \mathrm{tr}\left(\kappa\widehat{\ln T_t}'T_{t-1}\right)\right).$$

The resulting VB-filtering and VB-smoothing distributions have the following standard forms:

$$\tilde{f}\left(l_t|D_t\right) = \mathcal{M}u_{l_t}\left(\alpha_t\right),\ \tilde{f}\left(T_t|D_t\right) = \mathcal{D}i_{T_t}\left(Q_t\right),$$
$$\tilde{f}\left(l_{t-1}|D_t\right) = \mathcal{M}u_{l_{t-1}}\left(\beta_t\right),\ \tilde{f}\left(T_{t-1}|D_t\right) = \mathcal{D}i_{T_{t-1}}\left(R_t\right),$$

with shaping parameters

$$\alpha_t = d_t^\rho \circ \exp\left(\widehat{\ln T_t}\widehat{l_{t-1}}\right), \quad Q_t = \kappa\widehat{T_{t-1}} + \widehat{l_t}\widehat{l_{t-1}}',$$
$$\beta_t = \widehat{l_t}'\widehat{\ln T_t} + \alpha_{t-1}, \quad R_t = \kappa\widehat{T_t} + Q_{t-1},$$

where $\circ$ denotes the Hadamard product. The necessary moments of the filtering distributions are $\widehat{l_t} = \alpha_t$, $\widehat{T_t} = Q_t \circ \mathbf{1}_{c,1}\left[\mathbf{1}_{c,1}'Q_t\right]^{-1}$, $\widehat{\ln t_{i,j,t}} = \psi_\Gamma\left(q_{i,j,t}\right) - \psi_\Gamma\left(\mathbf{1}_{c,1}'Q_t\mathbf{1}_{c,1}\right)$. ψ_Γ denotes the ψ-function. Moments of the smoothing distributions are obtained in the same way.

4.2. Particle filtering (unknown T_t)

In order to evaluate the performance of this VB approximation, we now consider a simple particle filtering approach. The following properties of the HMM model, (14) and (16), allow simplification of the general framework: (i) the hidden parameter, l_t, has only c possible states. Hence, it is sufficient to generate just $n_l = c$ particles, $\{l_t\}_{n_l}$; and (ii) the space of T_t is continuous but bounded. Hence, we generate n_T particles, $\{T\}_{n_T}$, satisfying the restrictions on T_t.

We choose the following importance function:

$$\pi\left(T_t, l_t|D_t\right) = \mathcal{D}i_{T_t}\left(\kappa\widehat{T}_{t-1} + \mathbf{1}_{c,c}\right)\mathcal{M}u_{l_t}\left(\hat{T}_{t-1}\hat{l}_{t-1}\right),$$

where $\hat{T}_t = \sum_{j=1}^{n_T} w_{j,t}T^{(j)}$, $\hat{l}_t = d_t^\rho \circ \hat{T}_t\hat{l}_{t-1}$. Since l_t is a discrete parameter, it can be marginalized analytically. Hence, we can use the following recursive updates for the particle weights:

$$w_{j,t} \propto w_{j,t-1}\sum_{i=1}^{c} f\left(d_t|l_t = \epsilon_c(i)\right)\pi\left(T_t, l_t = \epsilon_c(i)|D_t\right).$$

We have not performed the resampling procedure after each step; *i.e.* particles, $\{T\}_{n_T}$, were sampled at $t = 1$ from the whole support and fixed at these values, $\forall t$.

4.3. Simulation Study: Inference of Soft Bits

We now consider reconstruction of a binary Markov chain, $x_t \in \{0, 1\}$, from a soft-bit sequence $y_t \in (0, 1)$, where $y_t|x_t$ is realized as a Dirichlet observation process (16) with $\rho = 2$. The problem is a special case of (14)—(16), with $c = 2$, and

$$d_t = [y_t, 1 - y_t],\ l_t = [x_t, 1 - x_t]. \tag{18}$$

The sequence T_t is illustrated in Fig. 4 (top). A typical resulting realization of the soft-bit sequence of length $\bar{t} = 1000$ is shown in Fig. 4 (middle), and the resulting VB-moment, $\widehat{T_t}$, in Fig. 4 (bottom).

We undertook a Monte Carlo study, generating 20 soft-bit sequences, y_t, each of length $\bar{t} = 1000$. We examined the following methods, each of which infers l_t and T_t: (i) **VBT**: VB-filtering as derived in Section 3; (ii) **PF50:** particle filtering (Section 4.2), with $n_T = 50$ particles; and (iii) **PF100**: particle filtering (Section 4.2), with $n_T = 100$ particles.

For comparison, we also examined the following techniques: (iv) **Threshold**: x_t is inferred by thresholding:

$$\hat{x}_t = \mathrm{round}\left(y_t\right) = \begin{cases} 1 & \text{if } y_t > 0.5, \\ 0 & \text{if } y_t \leq 0.5; \end{cases} \tag{19}$$

this constitutes Maximum Likelihood (ML) estimation of x_t, ignoring the Markov chain model for x_t (see Fig. 2); and (v) inference with **known** T_t, for which exact Bayesian filtering is tractable [6]. We expect the latter to have the best performance. Furthermore, methods (i)–(iii) should perform better than method (iv). If not, then inaccuracy in estimation of the extended model has exceeded the benefits of using an HMM. Performance was quantified via the Total Squared Error (TSE),

$$\mathrm{TSE} = \sum_{t=1}^{\bar{t}}\left(\widehat{x_t} - x_t\right)^2,$$

where—in all but the threshold method (19)—$\widehat{x_t} = \widehat{l_{1,t}}$ (18) denotes the posterior mean of x_t. In the case of the threshold inference (19), this criterion is equal to the Hamming distance between the true and inferred bit-streams.

The results of the Monte Carlo study are displayed in Fig. 3. The TSE of each method is plotted as a function

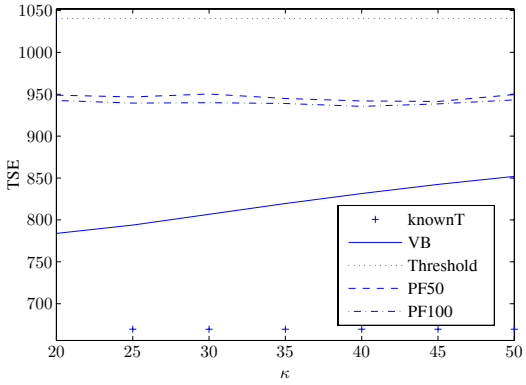

Fig. 3. Performance of HMM inference methods.

of κ. Within the illustrated range, VB-filtering performed better than either of the examined particle filters. All performed significantly better than the Threshold method. However, VB-filtering performs poorly for smaller values of κ, while the particle filters deteriorate for high κ. This sensitivity could be overcome by inferring κ from d_t on-line. Furthermore, the choice of *fixed* particles, $\{T\}_{n_T}$, limits performance of the particle filters. More sophisticated approaches to the generation of particles—such as resampling [2], kernel smoothing [2, Chapter 10], *etc.*—would probably improve performance.

5. DISCUSSION AND CONCLUSIONS

We have shown how to apply the VB approximation in order to achieve tractable Bayesian filtering. In particular, we have examined the problem of identifying HMMs with an unknown time-variant transition matrix, T_t. This extends previous work where the unknown transition matrix, T, was time-invariant [6]. This allows higher-order effects to be captured in the data, notably the seasonality evident in Fig. 4 (middle).

VB-filtering is a deterministic approximation scheme based on an assumption of conditional independence between θ_t and θ_{t-1}. This restriction is compensated by iterative exchange of moments (IVB cycles, see Fig. 1). Typically, only a few cycles are needed at each time t, providing a computationally efficient scheme for many applications where particle filtering may prove prohibitive. Further improvements in accuracy of the Variational approximation can be explored using mean field theory [9].

6. REFERENCES

[1] Z. Chen, "Bayesian filtering: From Kalman filters to particle filters, and beyond," Tech. Rep., Adaptive Syst. Lab., McMaster University, Hamilton, ON, Canada, 2003.

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

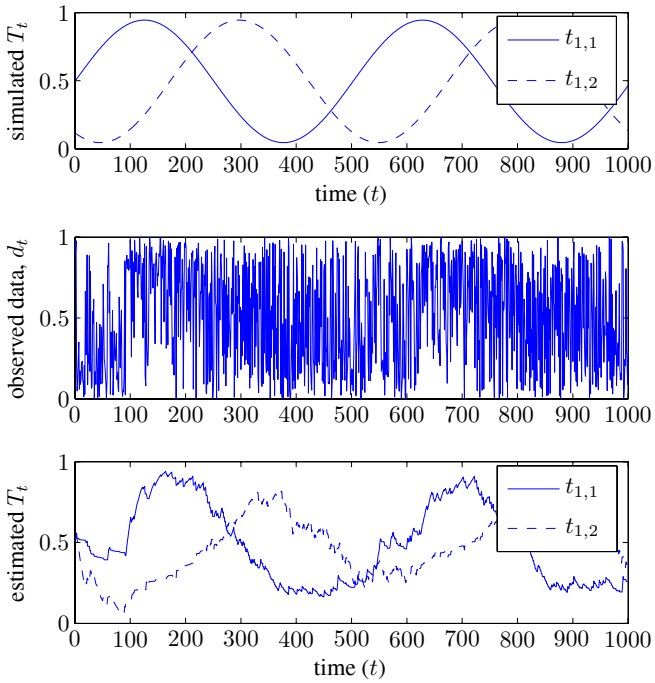

Fig. 4. Inference of T_t from a soft-bit sequence via the VB-moment, $\widehat{T}_t$.

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

[4] E.F. Daum, "New exact nonlinear filters," in *Bayesian Analysis of Time Series and Dynamic Models*, J.C. Spall, Ed. Marcel Dekker, New York, 1988.

[5] H. Attias, "A Variational Bayesian framework for graphical models.," in *Advances in Neural Information Processing Systems*, T. Leen, Ed., vol. 12. MIT Press, 2000.

[6] V. Šmídl and A. Quinn, *The Variational Bayes Method in Signal Processing*, Springer, 2005.

[7] S. Kullback and R. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 79–87, 1951.

[8] R.H. Elliot, L. Assoun, and J.B. Moore, *Hidden Markov Models*, Springer-Verlag, New York, 1995.

[9] M. Opper and D. Saad, *Advanced Mean Field Methods: Theory and Practice*, The MIT Press, Cambridge, Massachusetts, 2001.