THE RESTRICTED VARIATIONAL BAYES APPROXIMATION IN BAYESIAN FILTERING

Václav Šmídl

Academy of Sciences, Prague, Czech Republic

Anthony Quinn

Trinity College Dublin, Ireland

ABSTRACT

The Variational Bayes (VB) approach is used as a one-step approximation for Bayesian filtering. It requires the availability of moments of the free-form distributional optimizers. The latter may have intractable functional forms. In this contribution, we replace these by appropriate fixed-form distributions yielding the required moments. We address two scenarios of this Restricted VB (RVB) approximation. For the first scenario, an application in identification of HMMs is given. In the second, the fixed-form distribution is generated via Particle Filtering (PF). It is shown that a new approximation of Rao-Blackwellized particle filtering is obtained in this scenario of RVB. Its performance is assessed for a simple non-linear model.

1. THE VB APPROXIMATION

The VB approximation is a deterministic free-form optimization technique. It was first used in off-line inference problems [1] and extended to on-line inference of time-invariant parameters in [2]. The use of VB in Bayesian filtering was first discussed in [3]. The key theory is now reviewed.

Theorem 1 (Variational Bayes (VB)) *Let $f(\theta|D)$ be the posterior distribution of multivariate parameter, $\theta = [\theta_1', \theta_2']'$, and $\breve{f}(\theta|D)$ be an approximate distribution restricted to the set of conditionally independent distributions*

$$\breve{f}(\theta|D) = \breve{f}(\theta_1, \theta_2|D) = \breve{f}(\theta_1|D)\,\breve{f}(\theta_2|D). \quad (1)$$

The minimum of the Kullback-Leibler divergence

$$\tilde{f}(\theta|D) = \arg\min_{\breve{f}(\cdot)} KL\left(\breve{f}(\theta|D)\,||f(\theta|D)\right), \quad (2)$$

is reached for

$$\tilde{f}(\theta_i|D) \propto \exp\left(\mathsf{E}_{\tilde{f}(\theta_{/i}|D)}\left[\ln\left(f(\theta, D)\right)\right]\right), \quad i = 1, 2, \quad (3)$$

where $\theta_{/i}$ denotes the complement of θ_i in θ. We will refer to $\tilde{f}(\theta|D)$ (7) as the VB-approximation*, and $\tilde{f}(\theta_i|D)$ (3) as the* VB-marginals.

The above theorem provides a powerful tool for approximation of joint pdfs in *separable form* [3]:

$$\ln f(\theta_1, \theta_2, D) = g(\theta_1, D)'\, h(\theta_2, D). \quad (4)$$

Here, $g(\theta_1, D)$ and $h(\theta_2, D)$ are finite-dimensional vectors of compatible dimension. Using (4) in (3),

$$\tilde{f}(\theta_1|D) \quad \propto \quad \exp\left(g(\theta_1, D)'\, h\widehat{(\theta_2, D)}\right), \quad (5)$$

where $\widehat{h(\cdot)} = \mathsf{E}_{\tilde{f}(\theta_2|D)}[h(\cdot)]$ are the *VB-moments* for θ_2, and similarly for θ_1. An Iterative VB (IVB) [3] moment-swapping algorithm is implied. In many non-linear cases of g and/or h, the VB-marginals (3) are non-standard in form, and so the required VB-moments are difficult to evaluate. In this contribution, we aim to replace any such non-standard VB-marginal with a tractable alternative, as follows.

Corollary 1 (of Theorem 1: Restricted VB) *Let $\breve{f}(\theta|D)$ be a conditionally-independent approximation of $f(\theta|D)$ of the kind*

$$\breve{f}(\theta|D) = \breve{f}(\theta_1, \theta_2|D) = \breve{f}(\theta_1|D)\,\overline{f}(\theta_2|D), \quad (6)$$

where $\overline{f}(\theta_2|D)$ is a fixed-form *distribution. Then, the minimal KL divergence (2), under (6), is reached for*

$$\tilde{f}(\theta_1|D) \propto \exp\left(\mathsf{E}_{\overline{f}(\theta_2|D)}\left[\ln\left(f(\theta, D)\right)\right]\right). \quad (7)$$

$\overline{f}(\theta_2|D)$ needs to be chosen judiciously, such that its moments—required in (7)—are available. These moments are substituted just once, without the need for IVB cycles. Some standard distributional approximation methods may be interpreted as special cases of RVB; *e.g.* (i) *certainty equivalence*, where $\overline{f} \equiv \delta(\theta_2 - \hat{\theta}_2)$, in which case (7) becomes the conditional, $f\left(\theta_1|D, \hat{\theta}_2\right)$; and (ii) the *Quasi-Bayes (QB) approximation*, where $\overline{f} \equiv f(\theta_2|D)$, the exact marginal, if available [3].

2. BAYESIAN FILTERING

Consider the following model structure

$$d_t \sim f(d_t|\theta_t), \quad \theta_t \sim f(\theta_t|\theta_{t-1}), \quad (8)$$

where θ_t is known as the state variable. By *Bayesian Filtering* (BF), we mean the recursive evaluation of the filtering distribution $f(\theta_t|D_t)$ using Bayes' rule. $D_t = [d_1, \dots, d_t]$ denotes the history of observations. The computational flowchart of BF is displayed in Fig. 1. BF is analytically tractable if (i)

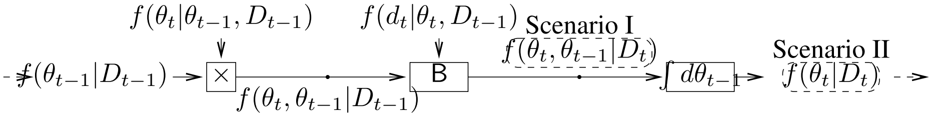
$f(\theta_t|\theta_{t-1}, D_{t-1})$
$f(d_t|\theta_t, D_{t-1})$
Scenario I
$f(\theta_t, \theta_{t-1}|D_t)$
Scenario II
$f(\theta_{t-1}|D_{t-1})$
×
B
$f(\theta_t, \theta_{t-1}|D_{t-1})$
$\int d\theta_{t-1}$
$f(\theta_t|D_t)$

Fig. 1. Two possible points ('scenarios') for VB approximation are indicated in a computational flowchart of exact Bayesian filtering. × denotes multiplication, B a Bayes' rule step, and $\int d\cdot$ marginalization over the indicated variable.

marginalization over θ_{t-1} is analytically tractable, and (ii) the resulting marginal distribution, $f(\theta_t|D_t)$, is in the same form as the previous step, $f(\theta_{t-1}|D_{t-1})$, so that the procedure can be iterated. We call this condition *conjugacy* in the BF context. (i) and (ii) are satisfied only for a very limited class of models.

2.1. Particle Filtering

Particle filtering (PF) [4] refers to a range of techniques for generating an empirical approximation of $f(\Theta_t|D_t)$ via importance sampling:

$$\tilde{f}(\Theta_t|D_t) = \textstyle\sum_{i=1}^{n} w_t^{(i)} \delta(\Theta_t - \Theta_t^{(i)}),\ w_t^{(i)} \propto \frac{f\left(\Theta_t^{(i)}|D_t\right)}{q\left(\Theta_t^{(i)}|D_t\right)} \tag{9}$$

Here, $\Theta_t = [\theta_1, \ldots, \theta_t]$ is the state trajectory, $\Theta_t^{(i)}$, $i = 1, \ldots, n$ are random samples drawn from an appropriately chosen importance function, $q(\Theta_t|D_t)$, and $w_t^{(i)}$ are the importance weights.

In the *Rao-Blackwellized* PF (RBwPF), the state variable is partitioned as $\theta_t = [\theta_{1,t}, \theta_{2,t}]$. An empirical approximation of the *marginal* is then sought:

$$\tilde{f}(\Theta_{2,t}|D_t) = \textstyle\sum_{i=1}^{n} w_t^{(i)} \delta(\Theta_{2,t} - \Theta_{2,t}^{(i)}) \tag{10}$$

$$w_t^{(i)} \propto \frac{f\left(d_t|\theta_{2,t}^{(i)}\right) f\left(\theta_{2,t}^{(i)}|\theta_{2,t-1}^{(i)}\right)}{q\left(\theta_{2,t}^{(i)}|D_t, \Theta_{2,t-1}^{(i)}\right)} w_{t-1}^{(i)}. \tag{11}$$

The numerator of (11) requires integration over $\theta_{1,t-1}$. Hence, RBwPF is feasible if there exists a partition of the state variable allowing this integration to be performed analytically. From (10), the implied approximation of the other marginal is

$$\tilde{f}(\theta_{1,t}|D_t) = \textstyle\sum_{i=1}^{n} w_t^{(i)} f\left(\theta_{1,t}|D_t, \Theta_{2,t}^{(i)}\right). \tag{12}$$

3. VARIATIONAL BAYESIAN FILTERING (VBF)

VB provides the possibility for deterministic approximation of the filtering distribution [3]. We now examine two scenarios for its application, as indicated in Fig. 1.

Scenario I of VBF: If marginalization over θ_{t-1} (Fig. 1) is intractable, it can be replaced by VB-marginalization. This is achieved by forcing conditional independence between θ_t and θ_{t-1} in $f(\theta_t, \theta_{t-1}|D_t)$ (Fig 1). The joint distribution needed in (3) is:

$$f(d_t, \theta_t, \theta_{t-1}|D_{t-1}) = \\ f(d_t|\theta_t, D_{t-1}) f(\theta_t|\theta_{t-1}) \tilde{f}(\theta_{t-1}|D_{t-1}). \tag{13}$$

Application of Theorem 1 to (13) yields VB-marginals in the form of two parallel Bayes' rule updates:

$$\tilde{f}(\theta_t|D_t) \propto f(d_t|\theta_t, D_{t-1}) \tilde{f}(\theta_t|D_{t-1}), \tag{14}$$

$$\tilde{f}(\theta_{t-1}|D_t) \propto \tilde{f}(d_t|\theta_{t-1}, D_{t-1}) \tilde{f}(\theta_{t-1}|D_{t-1}). \tag{15}$$

The following approximate distributions are involved:

$$\tilde{f}(\theta_t|D_{t-1}) \propto \exp\left\{\mathsf{E}_{\tilde{f}(\theta_{t-1}|D_t)}\left[\ln f(\theta_t|\theta_{t-1})\right]\right\}, \tag{16}$$

$$\tilde{f}(d_t|\theta_{t-1}, D_{t-1}) \propto \exp\left\{\mathsf{E}_{\tilde{f}(\theta_t|D_t)}\left[\ln f(\theta_t|\theta_{t-1})\right]\right\}. \tag{17}$$

From (16), the form of $\tilde{f}(\theta_t|D_{t-1})$ is determined by the form of the time-invariant parameter evolution model. Hence, one application of Bayes' rule (14) yields the same functional form at each time, satisfying the condition of conjugacy [3].

Scenario II of VBF: Consider the case where marginalization over θ_{t-1} is tractable (Fig 1), but yields a filtering distribution, $f(\theta_t|D_t)$, in a form different from that at $t-1$, violating conjugacy. However, if the VB approximation is then used to force conditional independence with respect to the partition $\theta_t = [\theta_{1,t}, \theta_{2,t}]$, these VB-marginals may be invariant, restoring conjugacy.

4. RESTRICTED VB IN BAYESIAN FILTERING

We now employ the RVB approximation (Corollary 1) for the two scenarios of VBF outlined in the previous section. Note, however, that there are many choices for the fixed distribution, $\overline{f}(\cdot)$, and that the final decision will depend on the particular model structure (8).

4.1. Scenario I

The moments of $\tilde{f}(\theta_{t-1}|D_t)$ (15) required in (16) may be intractable. An obvious fixed-form replacement of this distribution is the VB-filtering distribution from the previous step:

$$\overline{f}(\theta_{t-1}|D_t) \equiv \tilde{f}(\theta_{t-1}|D_{t-1}). \tag{18}$$

The VB-filtering distribution is then obtained without iterations, by substituting moments of $\tilde{f}(\theta_{t-1}|D_{t-1})$ via (16), and the result into (14).

4.2. Scenario II: RVB+PF

For the first time, we propose a *stochastic* approximation as the fixed-form distribution (Corollary 1):

$$\overline{f}\left(\theta_{2,t}|D_t\right) = \textstyle\sum_{i=1}^{n} \overline{w}_t^{(i)} \delta(\theta_{2,t} - \theta_{2,t}^{(i)}). \qquad (19)$$

Following the standard RBwBF approach of Section 2.1, the particle weights are approximated as

$$\overline{w}_t^{(i)} \propto \tilde{f}\left(d_t|\theta_{2,t}^{(i)}\right) f\left(\theta_{2,t}^{(i)}|\theta_{2,t-1}^{(i)}\right) \overline{w}_{t-1}^{(i)}/q\left(\theta_{2,t}^{(i)}|D_t, \Theta_{2,t-1}^{(i)}\right). \qquad (20)$$

$$\tilde{f}\left(d_t|\theta_{2,t}\right) = \textstyle\int f\left(d_t|\theta_t\right) \tilde{f}\left(\theta_{1,t}|D_{t-1}\right) d\theta_{1,t}. \qquad (21)$$

which is obtained from (9) using (i) the assumption of independent observation models for $\theta_{1,t}$ and $\theta_{2,t}$, and (ii) VB-marginals $\tilde{f}(\theta_{t-1}|D_{t-1})$ in place of exact marginals $f(\theta_{t-1}|D_{t-1})$.

Hence, the VB-marginal, $\tilde{f}\left(\theta_{1,t}|D_t\right)$, must be tractable. From (4) and (19), it is given by

$$\tilde{f}\left(\theta_{1,t}|D_t\right) \propto \exp\left(g\left(\theta_{1,t}\right)' \textstyle\sum_{i=1}^{n} \overline{w}_t^{(i)} h\left(\theta_{2,t}^{(i)}, D_t\right)\right). \qquad (22)$$

In the sequel, we will call this the *RVB+PF approximation*.

Remark 1 *The computational flow of RBwPF and RVB+PF differ in the following respects: (i) RBwPF requires analytical marginalization of the whole trajectory $\Theta_{1,t-1}$ (*i.e. *conjugacy at each step (Section 2)), while RVB+PF is tractable at every step if the one-step marginalization in (21) is tractable (*i.e. *conjugacy is not required); and (ii) one set of sufficient statistics is stored for each particle, $\Theta_{2,t}^{(i)}$, in RBwPF (12), while one set of sufficient statistics is used for all the samples $\theta_{2,t}^{(i)}$ in RVB+PF (22).*

Remark 2 *The proposed RVB+PF scheme generalizes the Modified RBwPF method (MRBwPF) proposed in [5], where $f\left(\theta_{1,t}|D_t\right)$ was approximated by certainty equivalence;* i.e.

$$\tilde{f}\left(\theta_{1,t}|D_t\right) \propto \exp\left(g\left(\theta_{1,t}\right)' h\left(\textstyle\sum_{i=1}^{n} \overline{w}_t^{(i)} \theta_{2,t}^{(i)}, D_t\right)\right), \qquad (23)$$

and $f\left(d_t|D_{t-1}, \theta_{2,t}\right)$ in (20) was approximated using sampling. In the simulations which follow, we will use MRBwPF with (20) for easier comparison. Results using the sampling-based approximation for $w_t^{(i)}$ [5] are very similar.

5. SIMULATION STUDIES

5.1. Inference of HMM with Unknown Transition Matrix

Consider a HMM with the following two constituents: (i) a first-order Markov chain on the unobserved discrete (label) variable l_t, with c possible states and time-variant unknown transition matrix $T_t \in \Re^{c\times c}$; and (ii) a set of c known class-conditional observation models, as arises in classification. The problem of inferring l_t can be formalized as a task of Bayesian filtering as follows:

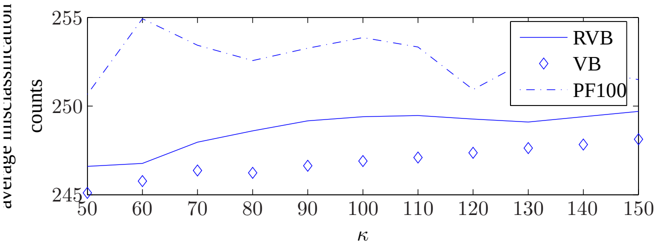

Fig. 2. Performance of competing algorithms for rapidly-varying T_t ($\kappa_{\text{true}} = 80$).

$$l_t \sim \mathcal{M}u_{l_t}\left(T_t l_{t-1}\right), \quad T_t \sim \mathcal{D}i_{T_t}\left(\kappa T_{t-1} + \mathbf{1}_{c,c}\right),$$
$$d_t \sim f_1\left(d_t\right)^{l_{1,t}} \times \ldots \times f_c\left(d_t\right)^{l_{c,t}}.$$

Here, scalar parameter κ governs variance of T_t. For high values of κ, the variance of the Dirichlet distribution is very tight allowing only slow evolution of T_t, while low values of κ allow faster evolution of T_t.

Under Scenario I (Section 3), the VB-approximation yields:

$$\tilde{f}\left(l_t|D_t\right) = \mathcal{M}u_{l_t}\left(\alpha_t\right), \; \tilde{f}\left(T_t|D_t\right) = \mathcal{D}i_{T_t}\left(Q_t\right), \qquad (24)$$
$$\tilde{f}\left(l_{t-1}|D_t\right) = \mathcal{M}u_{l_t}\left(\beta_t\right), \; \tilde{f}\left(T_{t-1}|D_t\right) \propto \exp\left(q\left(T_{t-1}\right)\right),$$

where shaping parameters α_t, Q_t, β_t are omitted for brevity. $q\left(T_{t-1}\right)$ is a complicated function, and so the moments of $\tilde{f}\left(T_{t-1}|D_t\right)$ are analytically intractable. Therefore, we make the RVB assignment (18),

$$\overline{f}\left(T_{t-1}|D_t\right) \equiv \tilde{f}\left(T_{t-1}|D_{t-1}\right),$$

which is Dirichlet (24).

A Monte Carlo study (100 runs) was undertaken, in order to compare (i) this RVB approach with (ii) (full) VB (*i.e.* (24), with numerical evaluation of $\tilde{f}\left(T_{t-1}|D_t\right)$ on a grid of 100×100 points), and with (iii) a particle filter with 100 particles (PF100) (Figure 2). Performance was defined in terms of the average number of misclassified labels in 1000 data samples. The RVB scheme is computationally cheaper than the competing methods, yet offers good performance even for rapid variations of T_t (Figure 2). For slowly varying T_t (high κ), the RVB performance is comparable to VB.

5.2. Performance of RVB+PF (Scenario II)

Consider the following model:

$$\begin{aligned} f\left(x_t|x_{t-1}\right) &= \mathcal{N}\left(Ax_{t-1}, Q\right), \\ f\left(C_t|C_{t-1}\right) &= \mathcal{N}\left(\arctan\left(C_{t-1}\right), P\right), \\ f\left(d_t|x_t, C_t\right) &= \mathcal{N}\left(C_t' x_t, R\right). \end{aligned}$$

Essentially, this is a standard linear-Gaussian model with unknown non-stationary C_t, for which a non-linear evolution model is defined. Here, integration over x_{t-1} is possible using standard Kalman Filtering (KF) theory, yielding the following conditional posterior of x_t:

$$f\left(x_t|D_t,C_t\right) \quad = \quad \mathcal{N}_{x_t|C_t}\left(\mu_t,\Omega_t^{-1}\right), \tag{25}$$

$$\Omega_t = \left(Q + A\Omega_{t-1}^{-1}A'\right)^{-1} + C_t'R^{-1}C_t,$$
$$\mu_t = \Omega_t^{-1}\left[\left(Q + A\Omega_{t-1}^{-1}A'\right)^{-1}A\mu_{t-1} + C_t'R^{-1}d_t\right].$$

This is written in terms of precision matrix Ω_t for analytical convenience. Exact integration over C_{t-1} is intractable. A Rao-Blackwellized PF is obtained using (10)–(11) with assignments $\theta_{1,t} = x_t$ and $\theta_{2,t} = C_t$.

Since the log of (25) can be written as $g\left(x_t\right)h\left(C_t,D_t\right)$ (4), it is amenable to RVB+PF approximation (Section 4.2). The resulting VB-marginal, $\tilde{f}\left(x_t|D_t\right)$ is Gaussian, *i.e.* in the form of (25), with assignments

$$\Omega_t = \left(Q + A\Omega_{t-1}^{-1}A'\right)^{-1} + \mathsf{E}\left[C_t'R^{-1}C_t\right], \tag{26}$$
$$\mu_t = \Omega_t^{-1}\left[\left(Q + A\Omega_{t-1}^{-1}A'\right)^{-1}A\mu_{t-1} + \widehat{C_t}'R^{-1}d_t\right].$$

where $\widehat{C_t} = \sum_{i=1}^{n} w_t^{(i)}C_t^{(i)}$. (26) is again in standard KF form, except for the term

$$\mathsf{E}\left[C_t'R^{-1}C_t\right] = \textstyle\sum_{i=1}^{n} w_t^{(i)}C_t^{(i)\,\prime}R^{-1}C_t^{(i)}. \tag{27}$$

Thus the matrix inversion lemma cannot be easily applied and the implementation is less efficient than for standard KF.

MRBwPF (Remark 2) also yields a result in the form of (26), with (27) replaced by $\widehat{C_t}'R^{-1}\widehat{C_t}$. This has the following consequences: (i) MRBwPF can use KF without any modification, while implementation of RVB+PF typically requires modification of KF; (ii) MRBwPF uses a certainty-equivalent estimate of C_t, while RVB+PF uses the second non-central moment (27), encoding information about its uncertainty.

A scalar system with one-dimensional state and output was simulated with parameters $A = 1$, $Q = 0.5$, $P = 0.1$, $R = 0.5$. The aim is to illustrate the effect of propagation of higher moments (27) within the KF scheme. Hence, we use exactly the same particle filter on C_t—*i.e.* the same importance function and re-sampling scheme—for all tested methods. The performance was assessed via the value of the posterior distribution of X_t at the simulated value. Variants of $\tilde{f}(x_t|D_t)$—obtained from (25) via, respectively, RBwPF (12), RVB+PF (22), and MRBwPF (23)—are displayed in Table 1. These numbers are obtained by averaging over 100 simulation runs for each tested setting. Similar behaviour was observed for other noise levels and numbers of particles.

Both RVB+PF and MRBwPF perform significantly worse than RBwPF but at much lower computational cost, since they require only one step of KF per time-step, in contrast to n Kalman updates for RBwPF. RVB+PF slightly outperforms MRBwPF but at the price of evaluation of the second-order moment (27) and the numerical overhead of incorporating this into the KF (26).

number of data	RBwPF	RVB+PF	MRBPF	scale
100	-0.37	-2.58	-2.68	$\times 10^3$
1000	-0.58	-3.43	-3.44	$\times 10^5$

Table 1. Results of MC studies displayed via log of the posterior distribution of X_t, evaluated at the simulated value (the greater the mantissa, the better the result).

6. CONCLUSION

The VB approximation in Bayesian filtering is constrained by the need to evaluate moments of the resulting VB-marginals, which is difficult to ensure. If at least one of the VB-marginals is tractable, the method can be combined with other approximations via the Restricted VB (RVB) approach. In this paper, we have presented two possible Scenarios for the use of VB in Bayesian filtering, and for each Scenario one possible RVB scheme to overcome the intractability of VB-marginals. The first Scenario enforced conditional independence between the current and one-step-delayed state variable, and it is expected to work well for extensions of discrete hidden Markov models. The second Scenario enforced conditional independence in the posterior distribution at each time, t. In this Scenario, particle filtering was used to generate the fixed-form distribution. This yielded a new variant of Rao-Blackwellized particle filters, generalizing previous heuristically-motivated approximations. The new RVB-based particle filter is expected to offer a computationally far cheaper scheme than standard Rao-Blackwellized particle filters.

7. REFERENCES

[1] C. M. Bishop, “Variational principal components,” in *Proceedings of the Ninth International Conference on Artificial Neural Networks*, (ICANN), 1999.

[2] M. Sato, “Online model selection based on the variational Bayes,” *Neural Computation*, vol. 13, pp. 1649–1681, 2001.

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

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

[5] F. Mustière, M. Bolić, and M. Bouchard, “A modified Rao-Blackwellised particle filter,” in *Proceedings of the IEEE conference on Acoustics, Speech, and Signal Processing*, 2006.