Position papers of the 2014 Federated Conference on Computer Science and Information Systems pp. 61–66
DOI: 10.15439/2014F483
ACSIS, Vol. 3

Transdimensional sequential Monte Carlo for hidden Markov models using variational Bayes - SMCVB

Clare A. McGrory
Centre for Applications in Natural Resource Mathematics
University of Queensland
Brisbane
Queensland, Australia
Email: c.mcgrory@uq.edu.au

Daniel C. Ahfock
Centre for Applications in Natural Resource Mathematics
University of Queensland
Brisbane
Queensland, Australia
Email: daniel.ahfock@uqconnect.edu.au

***Abstract*—In this paper we outline a transdimensional sequential Monte Carlo algorithm - SMCVB - for fitting hidden Markov models. Sequential Monte Carlo (SMC) involves generating a weighted sample of particles from a sequence of probability distributions with the aim of converging to the target Bayesian posterior distribution. SMCVB makes use of variational Bayes (VB) in combination with SMC principles to create an algorithm which targets the posterior distribution more efficiently thereby saving on time and computational storage requirements. Another key feature of our methodology is that the variational-Bayes-generated proposals can vary in dimension. We have found in our simulation studies that we are able to obtain sensible estimates of the model dimensionality in this one-step procedure. This introduces very valuable additional flexibility in the modelling approach and opens up the potential for use of the algorithm in on-line settings where efficient and reliable estimation of dimensionality and parameters is required.**

I. Introduction

SEQUENTIAL Monte Carlo (SMC) approaches for Bayesian inference were first introduced to meet the requirement for efficient and tractable methods for analysing large amounts of data that arose sequentially over time (see [1] for an overview). In SMC, the procedure begins by initially proposing a population of samples, which are referred to as particles, from an initial target posterior distribution, reweighting these particles through importance sampling and then resampling from them to approximate the next target posterior density in the sequence. Subsequently there has also been a significant amount of research into the application of SMC to static problems, i.e. the data are treated as if they had arisen sequentially even though the whole dataset is available at the start of the analysis. For example, [2] and [3], provide examples of SMC in a static setting. Another example of a static SMC algorithm is given in [4]. This is a data-tempering SMC approach and this will form the basis for our new proposed algorithm.

Within the context of finite mixture estimation, [5] proposed a new transdimensional SMC algorithm based on the idea of using the variational Bayes (VB) approach [6], [7], [8] within an SMC framework. The resulting hybrid algorithm is called SMCVB. The SMC algorithm is initialised with particles drawn from a VB approximation to the posterior distribution rather than from a prior distribution; the aim of this is to make the algorithm more efficient. The underlying SMC algorithm takes the form of the data-tempering algorithm described in [4] as noted above. A significant advantage of the SMCVB algorithm is that it is not restricted to fixed-dimensional space. This is a highly useful feature for practical application since estimating a suitable dimension for the model is usually an important part of the analysis. In particular, in applications where new batches of data arise over time, the dimension size that achieves the most appropriate fit might change throughout the analysis as new information becomes available. Our approach has the advantage over existing schemes that it is able to adapt to such changes in an automated fashion. This feature means that there is a lot of potential for application and extension of the hybrid approach to modern applications where datasets are ever increasingly large and there is a demand for fast or even online analysis capabilities.

In this paper we describe how the algorithm proposed in [5] can be extended to the context of hidden Markov modelling, and we show that this leads to a novel scheme which is time-efficient and provides reliable results.

The article is organised as follows. In Section 2 we outline the model. In section 3, we describe the VB approach for hidden Markov modelling with Gaussian noise. In Section 4 we present the transdimensional VB-based SMC (SMCVB) algorithm. In Section 5 we show some results from the analysis of simulated data, and Section 6 concludes the paper.

II. Variational Bayesian Inference for Hidden Markov Models with Gaussian Noise

Following the approach that is described in [9], we assume a Gaussian HMM where the system can be in any one of K states at any time-point i, but the actual state sequence is hidden. Our observations correspond to a noisy realisation of the actual state sequence. We assume a discrete first-order Markovian dependence structure, therefore the current state depends only on the state occupied at the last time-point. We will follow the notation set out in [8] for specifying the HMM and we will apply the algorithm described in that article for estimation of the model. Given that the system is in state j_1 at time-point i, the transition matrix π represents the probability of moving to state j_2 at time-point $i+1$.

The transition matrix is defined as $\pi = \{\pi_{j_1 j_2}\}$ where $\pi_{j_1 j_2} = p(z_{i+1} = j_2 | z_i = j_1)$ and z_i is the latent variable representing the state at time i; all transition probabilities are non-negative and columns of the transition matrix must sum to 1. No structure is imposed on the transition matrix π, it will be estimated as part of the analysis. The observed data is denoted by $\{y_i; i = 1, \ldots, n)\}$, and the emission probabilities, i.e., the conditional probabilities of state membership at each time-point, are denoted by $p(y_i|z_i = j) = p_j(y_i|\phi_j)$. Since we are assuming Gaussian noise in the observations, the $\phi = \{\phi_j\}$ correspond to the parameters of the univariate Gaussian noise distribution corresponding to the relevant states $j = 1, \cdots, K$. Then, the model parameters are given by $\theta = (\pi, \phi)$ and we have

$$\begin{aligned} p(y, z, \theta) &= \prod_{i=1}^{n} \prod_{j=1}^{K} (p_j(y_i|\phi_j))^{z_{ij}} \\ &\times \prod_{i=1}^{n-1} \prod_{j_1=1}^{K} \prod_{j_2=1}^{K} (\pi_{j_1 j_2})^{z_{ij_1} z_{i+1 j_2}} \\ &\times \prod_{j=1}^{K} p_j(\phi_j) \prod_{j_1=1}^{K} p(\pi_{j_1}), \end{aligned}$$

where z_{ij} is a latent indicator variable such that $z_{ij} = 1$, if $z_i = j$, and $z_{ij} = 0$, if $z_i \neq j$. The terms $p_j(\phi_j)$ and $p(\pi_{j_1})$ correspond to the prior distributions over the parameters of the univariate Gaussian noise distribution, and the transition probabilities, for the relevant state j_1.

We use the same prior specifications for this model as the ones used in [8]. For more detail, we refer the reader to that paper. We follow the notation used in [8] in order to facilitate comparison with the more detailed descriptions of the corresponding derivations provided therein. The standard conjugate prior distributions are used for the model parameters. For each state j_1, there is an independent Dirichlet prior distribution for the transition probabilities $\{\pi_{j_1 j_2} : j_2 = 1, \ldots, K\}$, with hyperparameters $\{{\alpha_{j_1 j_2}}^{(0)}\}$. The noise model for the observations is univariate Gaussian with unknown means and precisions such that for each state j, we have a Gaussian prior distribution with mean μ_j and precision τ_j. Each of the means μ_j themselves have independent univariate Gaussian conjugate prior distributions, conditional on the precisions, with means and precisions given by $m_j^{(0)}$ and ${\beta_j}^{(0)}\tau_j$, respectively. The precisions τ_j have independent Gamma prior distributions with shape and scale parameters given by $\frac{1}{2}{\eta_j}^{(0)}$ and $\frac{1}{2}{\delta_j}^{(0)}$, respectively.

III. Variational Bayesian Inference for Hidden Markov Models with Gaussian Noise

In the variational Bayesian inferential approach, we do not sample from the posterior distribution, as we would in a Markov chain Monte Carlo (MCMC) based approach, instead we find a close approximation to it; this approximation to the posterior is referred to as the variational posterior distribution. The fact that the VB estimate of the posterior does not require iterative sampling makes it a very useful approach in terms of time efficiency, which is of course an important consideration when working with large datasets. We will briefly outline the key concepts of the variational approach in this section. As we have stated, our aim is to find the VB approximation to our desired posterior distribution, i.e. $p(\theta|y)$. This posterior can be obtained as the marginal distribution of $p(\theta, z|y)$; this distribution is typically a complex expression and it has to be approximated in this method. In the VB approach, we approximate $p(\theta, z|y)$ by another distribution which we call the variational approximating distribution. The variational approximating distribution is denoted by $q(\theta, z)$, and the idea is to take this distribution as being the minimiser of the the Kullback-Leibler(KL) divergence between $q(\theta, z)$ and $p(\theta, z|y)$. To make the minimisation of the KL divergence between these quantities tractable, the standard assumption made is that $q(\theta, z)$ can be factorised as $q(\theta, z) = q_\theta(\theta) q_z(z)$. Derivation of the variational function leads to a set of coupled equations for $q_\theta(\theta)$ and $q_z(z)$ for updating the estimates of the parameters and latent variables in the mode. Note that another way to view the motivation for this approach, is that the variational approximation to the posterior provides a tight lower bound on the observed-data log-likelihood. The variational Bayesian algorithm proceeds by iteratively updating these coupled expressions for the model parameters and the latent variables until they converge, at least locally, in the sense that subsequent updates no longer improve estimates. The values in the converged algorithm are then the estimates of the model parameters in the variational posterior distribution.

The Forms of the Variational Posterior Estimated Distributions over Model Parameters and Latent Variables

After applying the standard VB approximation to the Bayesian posterior of the HMM we find the following forms for the variational posterior distributions over the model parameters [8].

$$\begin{aligned} q_{j_1}(\pi_{j_1}) &= \mathrm{Dir}(\pi_{j_1}|\{\alpha_{j_1 j_2}\}), \\ q(\mu_j|\tau_j) &= \mathrm{N}(\mu_j|m_j, (\beta_j \tau_j)^{-1}), \\ q(\tau_j) &= \mathrm{Ga}\left(\tau_j|\frac{1}{2}\eta_j, \frac{1}{2}\delta_j\right). \end{aligned}$$

The parameters of these distributions can be computed by iteratively solving the set of coupled equations outlined in Algorithm 1. The well-known forward-backward algorithm [10] has to be used to make computation of the required marginal probabilities possible [11]. The forward-backward algorithm gives us estimates of the forward and backward variables, $\mathrm{fvar}_i(j)$ and $\mathrm{bvar}_i(j)$, respectively, for each i and j. In the forward-backward algorithm the $a^*_{j_1 j_2}$ are estimates of the probabilities of transition from states j_1 to state j_2, and the b^*_{ij}'s are estimates of the emission probabilities given that the system is in state j at time point i. These are then used in the update equation for $q_{ij} = q_z(z_i = j) = p(z_i = j_1|y_1, \ldots, y_n)$

Algorithm 1 VB Algorithm for Fitting a Hidden Markov Model with Gaussian Noise

Set initial values for parameters
$\alpha_{j_1 j_2}{}^{(0)}, m_j^{(0)}, \beta_j^{(0)}, \eta_j^{(0)}, \delta_j^{(0)}, K, q_z(z_i = j_1, z_{i+1} = j_2)$ and q_{ij}, for $j, j_1, j_2 \in 1, \cdots, K$, $i \in 1, \cdots, n$
while not converged **do**
Update the VB posterior parameter estimates (Ψ denotes the digamma function)

$$
\begin{aligned}
\alpha_{j_1 j_2} &= \alpha_{j_1 j_2}{}^{(0)} + \sum_{i=1}^{n-1} q_z(z_i = j_1, z_{i+1} = j_2) \\
\beta_j &= \beta_j{}^{(0)} + \sum_{i=1}^{n} q_{ij} \\
\eta_j &= \eta^{(0)} + \sum_{i=1}^{n} q_{ij} \\
\delta_j &= \delta^{(0)} + \sum_{i=1}^{n} q_{ij} {y_i}^2 + \beta_j{}^{(0)} {m_j}^{(0)\,2} \\
&\qquad -\beta_j m_j^2 \\
m_j &= \frac{\beta_j{}^{(0)} m_j{}^{(0)} + \sum_{i=1}^{n} q_{ij} y_i}{\beta_j}
\end{aligned}
$$

$$
\begin{aligned}
a^*_{j_1 j_2} &= \exp\left(\Psi(\alpha_{j_1 j_2}) - \Psi\Big(\sum_{j=1}^{K} \alpha_{j_1 j}\Big)\right) \\
&= p(z_{i+1} = j_2 | z_i = j_1) \\
b^*_{ij} &= \exp\left(\frac{1}{2}\Psi\Big(\frac{1}{2}\eta_j\Big) - \frac{1}{2}\log\Big(\frac{\delta_j}{2}\Big) - \frac{1}{2\beta_j}\right. \\
&\qquad \left. -\frac{1}{2}\Big(\frac{\eta_j}{\delta_j}\Big)(y_i - m_j)^2\right) \\
&= p(y_{i+1} | z_{i+1} = j_2) \\
q_z(z_i = j) &= \frac{\mathrm{fvar}_i(j_1)\mathrm{bvar}_i(j_1)}{\sum_{j_2} \mathrm{fvar}_i(j_2)\mathrm{bvar}_i(j_2)} \\
q_{ij} &= q_z(z_i = j_1, z_{i+1} = j_2) \\
&= \frac{\mathrm{fvar}_i(j_1) a^*_{j_1 j_2} b^*_{i+1 j_2} \mathrm{bvar}_{i+1}(j_2)}{\sum_{j_1}\sum_{j_2} \mathrm{fvar}_i(j_1) a^*_{j_1 j_2} b^*_{i+1 j_2} \mathrm{bvar}_{i+1}(j_2)}
\end{aligned}
$$

If any state has a weighting approaching zero then eliminate this state and reduce model dimension to $K = K - 1$
end while

and $q_z(z_i = j_1, z_{i+1} = j_2)$. The $\sum_{i=1}^{n} q_{ij}$ is the VB estimate of the number of observations expected to belong to state j, and the $q_z(z_i = j_1, z_{i+1} = j_2)$ are the VB estimates of the transition probabilities.

We would like to draw the readers attention to the automatic state elimination feature that is an intrinsic part of the VB approach when estimating mixture models and HMMs. As a result of this property, given the initially chosen value for model dimension, K, the final estimated solution in the VB posterior will have dimension less than or equal to K. Provided that K is chosen sufficiently large, we expect that a suitable dimension for the model is estimated as part of the VB procedure. Note that model selection criteria could also be computed to provide an alternative way to select the most appropriate dimension size.

IV. Transdimensional Variational Bayes Sequential Monte Carlo Algorithm (SMCVB)

The SMC framework which underpins the algorithm is a modification of the SMC algorithm described in [4]. The algorithm described in [4] is initialised with a small batch of data and then proceeds to incorporate data in sequential batches of increasing size which is what is meant by data tempering. What distinguishes our SMCVB approach from other SMC algorithms is that we use a VB posterior mean estimate of the model parameters in order to generate proposal particles rather than generating them from the prior. This is an intuitively logical and sensible hybrid modification of the data-tempering SMC algorithm. The complete-data target posterior distribution that we ultimately wish to estimate is

$$\pi(\theta) = \pi(\theta | y_1, \cdots, y_n),$$

and the target posterior at each subsequent iteration t ($t = 1, \cdots, T$) is

$$\pi_t(\theta) = \pi_t(\theta | y_1, \cdots, y_{n_t}),$$

where $n_1 \leq n_2 \leq \cdots \leq n_T = n$ is an increasing set of sample sizes. This separation of the whole dataset into smaller sub-batches leads to the formation of a sequence of target posteriors which on average smoothly converge to the final complete data target posterior. Our proposed SMCVB algorithm for HMMs is outlined in Algorithm 2.

Due to the VB algorithm intrinsic state elimination property, VB solutions obtained for each batch, which are in turn used to generate new proposed sets of particles, can vary in dimension. This allows us to explore a range of models with various numbers of states at various points in the analysis. We suggest that the distribution of particles over the various dimension sizes might be used as a guide for deciding on the most appropriate number of states to include in the final model. In our analyses of simulated datasets we found that this strategy led to reliable estimates of the most appropriate number of states to include in the fitted model.

V. Some Results from Application of the SMCVB Algorithm for HMMs to Simulated Data

For illustration we present here results obtained from using our hybrid algorithm to analyse synthetic data generated from a three-state hidden Markov model. The parameter settings used to simulate the data are outlined in table I, note that this corresponds to a considerably noisy set of data. We generated 1000 datapoints from this model and the data were read in in batches of 100 points. We used vague priors in the analysis. The transition matrix was given by

Algorithm 2 SMCVB Algorithm

Initialise: estimate the VB partial posterior $\pi_{t_0}(\theta) = \pi_{VB}(\theta|y_1, \cdots, y_{n_0})$ using Algorithm 1

Particle set: generate a set of R particles $(\theta_r^{(0)}, W_r^{(0)})_{r=1,\cdots,R}$ with associated weights $\{W_r^{(0)}\}$ which target the initial posterior $\pi_{t_0}(\theta)$.

Draw: draw R particles from these estimated posteriors, which results in vectors of the form $\{\theta_R^{(0)} = (\mu_r^{(0)}, \tau_r^{(0)}, \rho_r^{(0)})\}$, with weights given by

$$W_r^{(0)} \propto \frac{p(y_1, \cdots, y_{n_0}|\theta_r^{(0)})p(\theta_r^{(0)})}{\pi_{VB}(\theta_r^{(0)}|y_1, \cdots, y_{n_0})}$$

Normalise the weights to obtain $W_r^{(0)}$.

while $n_t < n$ **do**

Reweight: update the weights at iteration t using the n_tth batch of data to give

$$W_r^{(t)} \propto W_r^{(t-1)} \times p(y_{n_{t-1}+1}, \cdots, y_{n_t}|\theta_r^{(t-1)}),$$

where $r = 1, \cdots, R$.

if Effective sample size < $\frac{n}{2}$ **then**

Resample R values from the current set of particles using multinomial sampling. i.e. we resample the $\{(\theta_r^{(t-1)}, W_r^{(t-1)})\}_{r=1,\cdots,R}$ to get $\{(\theta_r'^{(t)}, 1/R)_{r=1,\cdots,R}\}$.

end if

Move: move to a new set of particles, these become the $\{\theta_r^{(t)}\}$ to be carried forward. Propose these from distributions from the VB posterior mean of the parameters based on the current batch of data and use a standard Metropolis–Hastings update to choose the new particles.

end while

$$\pi = \begin{pmatrix} 0.15 & 0.80 & 0.05 \\ 0.50 & 0.10 & 0.40 \\ 0.30 & 0.40 & 0.30 \end{pmatrix}$$

The estimated posterior parameter values for the noise model are displayed in table II and table III. We compare the fits we obtained with our hybrid SMCVB algorithm to those obtained from a standard MCMC analysis and a standard VB analysis. Plots of the posterior distributions corresponding to the fitted parameters are shown in Figures 1 and 2.

Using our approach, we estimated the most suitable number of states by calculating the proportions of particles corresponding to the different model sizes. The majority of particles in the final set corresponded to a three-state model, which we know accurately reflects the true underlying model in this case. Further work is required to better explore the reliability and justification for the general use of the distribution of the number of states in the final particles for estimating the most appropriate number of states for the fitted model. However, our explorations of some different simulated datasets suggest good potential for this strategy. Model selection criteria could also be checked to assess the most appropriate dimension.

TABLE I
PARAMETERS OF THE GAUSSIAN NOISE DISTRIBUTIONS FOR EACH STATE IN THE MODEL USED TO SIMULATE THE EXAMPLE DATA.

State	Mean	Standard Deviation
1	1.00	0.50
2	2.00	0.15
3	2.50	0.30

TABLE II
POSTERIOR MEANS OF THE MEAN PARAMETERS OF THE GAUSSIAN NOISE DISTRIBUTIONS FOR EACH STATE ESTIMATED USING THE DIFFERENT APPROACHES

SMCVB	VB	MCMC
1.01	1.01	1.01
2.00	2.00	2.00
2.56	2.56	2.55

TABLE III
POSTERIOR MEANS OF THE STANDARD DEVIATION PARAMETERS OF THE GAUSSIAN NOISE DISTRIBUTIONS FOR EACH OF THE STATES ESTIMATED USING THE DIFFERENT APPROACHES

SMCVB	VB	MCMC
0.53	0.53	0.53
0.15	0.15	0.15
0.26	0.26	0.27

The results shown demonstrate that this approach leads to reliable estimates of model parameters. This new hybrid SMCVB scheme produces posterior estimates which are even closer to MCMC estimates for the same model than the VB approximation is. Note then that another way to view this scheme is as a way to further improve on the VB approximation to the Bayesian posterior distribution.

SMCVB is efficient in terms of computing time and due to the nature of the SMC structure, it is ideally suited to applications where new batches of data continually become available. That feature, combined with the improved time efficiency that is achieved through the use of the targeted VB guided proposals, has created an algorithm which has much potential to be of practical use in modern applications where large volumes of sequentially occurring data have to be processed and the traditional MCMC-based approaches may not be feasible due to computational limitations.

VI. CONCLUSION

We have extended the recently proposed transdimensional SMC algorithm, SMCVB, to the setting of estimating parameters and dimension of hidden Markov models. The algorithm allows us to explore the dimension of the posterior distribution and achieves increased computational efficiency over other SMC approaches by using a VB algorithm to generate the independent proposals at each iteration of the procedure.

Current work involves applying this algorithm to analysing large time series data with the aim of performing climate regime shift detection.

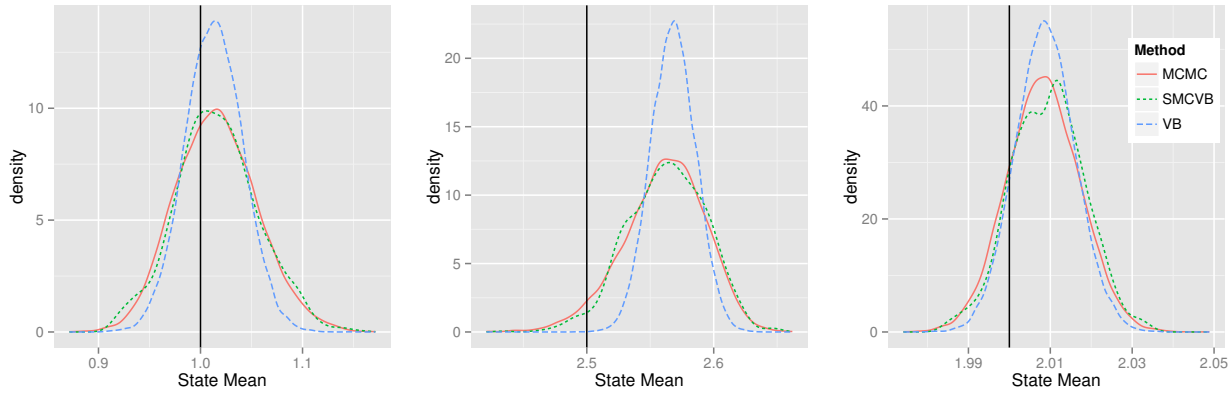

Fig. 1. Comparison of the posterior distributions for the mean parameter of the Gaussian noise distribution for the three states fitted using the different methods. The solid black line marks the true mean for the corresonding state in the model the data were simulated from.

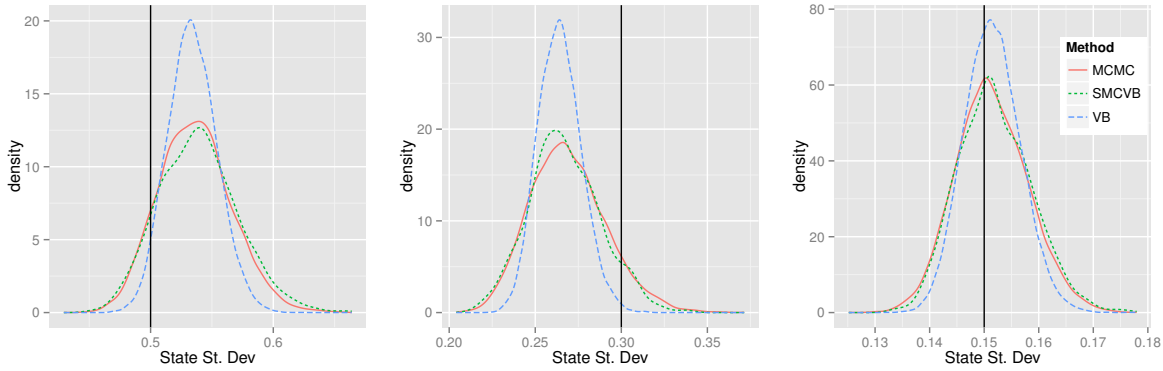

Fig. 2. Comparison of the posterior distributions for the standard deviation parameter of the Gaussian noise distribution for the three states fitted using the different methods. The solid black line marks the true standard deviation for the corresponding state in the model the data were simulated from.

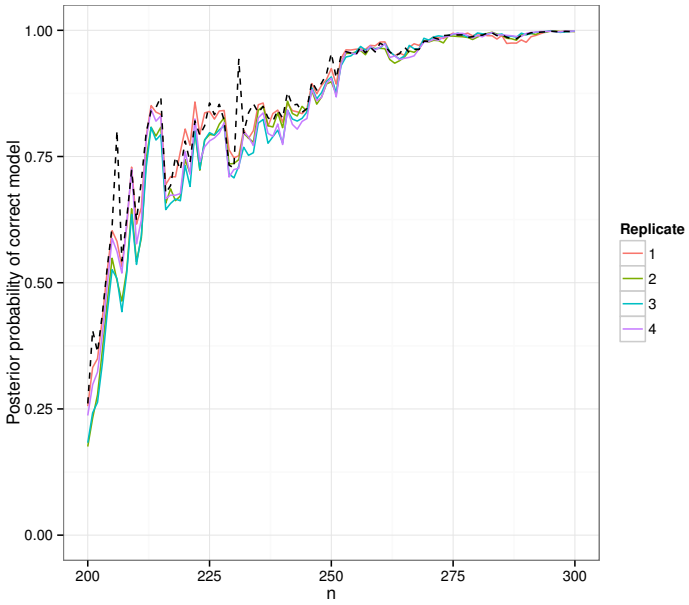

Fig. 3. Posterior probability associated with the correct number of hidden states for the model as estimated based on the proportion of particles having that dimension in the final set of particles. The plot shows results from four replicate runs of the SMCVB algorithm and for different numbers of particles.

REFERENCES

[1] A. Doucet, J.F.G. De Freitas, and N.J. Gordon, *Sequential Monte Carlo Methods in Practice.* NewYork, NY: Springer, 2001.

[2] P. Del Moral, A. Doucet, and A. Jasra, "Sequential Monte Carlo samplers," *J. Roy. Statist. Ser. B,* vol. 68, 2006, pp. 411–436.

[3] A. Jasra, D.A. Stephens, and C.C. Holmes, "On population based simulation for static inference," *Statist. Comput.* vol. 17, 2007, pp. 263–279.

[4] N. Chopin, "A sequential particle filter method for static models," *Biometrika* vol. 89, 2002, pp. 539–551.

[5] C. A. McGrory, A. N. Pettitt, D.M. Titterington, C.L. Alston, and M. Kelly, "Transdimensional Sequential Monte Carlo using Variational Bayes - SMCVB," *Unpublished,* 2014.

[6] C. A. McGrory, and D. M. Titterington, "Variational approximations in Bayesian model selection for finite mixture distributions," *Comput. Stat. Data An.* vol. 51, 2007, pp. 5352–5367.

[7] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes," *in Proceedings of the 15th International Conference on Uncertainty in Artificial Intelligence,* Morgan Kaufmann, San Francisco, 1999, pp. 21–30.

[8] M. Wand, J. Ormerod, S. Padoan, and R. Fruhwirth, "Mean Field Variational Bayes for Elaborate Distributions," *Bayesian Analysis* vol. 6, 2011, pp. 847–900.

[9] C. A. McGrory, and D. M. Titterington, "Variational Bayesian analysis for hidden Markov models," *Aust. and New Zealand J. Statist.* vol. 51, 2009, pp. 227–244.

[10] L.E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. of Math. Statist.* vol. 41, 1970, pp. 164–171.

[11] T. Rydén, T. Teräsvirta, T., and S. Åsbrink, "Stylized facts of daily return series and the hidden Markov model," *J. Appl. Econometr.* vol. 13, 1998, pp. 217–244.