Ann Inst Stat Math (2014) 66:553–575
DOI 10.1007/s10463-014-0450-4

SPECIAL ISSUE: BAYESIAN INFERENCE AND STOCHASTIC COMPUTATION

Parallel sequential Monte Carlo samplers and estimation of the number of states in a Hidden Markov Model

Christopher F. H. Nam · John A. D. Aston · Adam M. Johansen

Received: 12 March 2013 / Revised: 8 November 2013 / Published online: 16 March 2014

Abstract The majority of modelling and inference regarding Hidden Markov Models (HMMs) assumes that the number of underlying states is known a priori. However, this is often not the case and thus determining the appropriate number of underlying states for a HMM is of considerable interest. This paper proposes the use of a parallel sequential Monte Carlo samplers framework to approximate the posterior distribution of the number of states. This requires no additional computational effort if approximating parameter posteriors conditioned on the number of states is also necessary. The proposed strategy is evaluated on a comprehensive set of simulated data and shown to outperform the state of the art in this area: although the approach is simple, it provides good performance by fully exploiting the particular structure of the problem. An application to business cycle analysis is also presented.

Keywords Hidden Markov Models · Model selection · Sequential Monte Carlo

JADA was supported in part by the HEFCE/EPSRC CRiSM Grant; AMJ by EPSRC Grant EP/I017984/1.

Electronic supplementary material The online version of this article (doi:10.1007/s10463-014-0450-4) contains supplementary material, which is available to authorized users.

C. F. H. Nam
224 Pontius Avenue North, Apt 337, Seattle, WA 98109, USA
e-mail: c.f.h.nam@gmail.com

J. A. D. Aston
Statistical Laboratory, University of Cambridge, Wilberforce Road, Cambridge CB3 0WB, UK
e-mail: jada2@cam.ac.uk

A. M. Johansen (✉)
Department of Statistics, University of Warwick, Coventry CV4 7AL, UK
e-mail: a.m.johansen@warwick.ac.uk

1 Introduction

Hidden Markov Models (HMMs) provide a rich framework to model non-linear, non-stationary time series. Applications include modelling DNA sequences (Eddy 2004), speech recognition (Rabiner 1989) and modelling daily epileptic seizure counts for a patient (Albert 1991).

Much of the inference and applications for HMMs such as estimating the underlying state sequence (Viterbi 1967), parameter estimation (Baum et al. 1970) and changepoint inference (Chib 1998; Aston et al. 2011; Nam et al. 2012), assume that the number of states the underlying Markov Chain (MC) can take, H, is known a priori. However, when real time series data is analysed, H is often *not* known.

Assuming a particular number of underlying states without performing any statistical analysis can sometimes be advantageous if the states correspond directly to a particular phenomena. For example in Econometric GNP analysis (Hamilton 1989), two states are assumed a priori, "Contraction" and "Expansion", with recessions being defined as two consecutive contraction states in the underlying state sequence. Without such an assumption, this definition of a recession and the conclusions we can draw from the resulting analysis may be lost.

However, it may be necessary to assess whether such an assumption on the number of underlying states is adequate, and typically, we are presented with time series data for which we are uncertain about the appropriate number of states to assume. This paper concerns model selection for HMMs when H is unknown. Throughout this paper, we use "model" and the "number of states in a HMM" interchangeably to denote the same statistical object.

Several methods for determining the number of states of a HMM currently exist. In general model selection problems, techniques are often formulated in terms of penalised likelihood or information criteria [see for example, Konishi and Kitagawa (2008)]. However, methods such as Akaike's and Bayesian Information Criteria are not suitable for HMMs because we can always optimise these criteria via the introduction of additional states (Titterington 1984). In light of this, Mackay (2002) proposes an information theoretic approach which yields a consistent estimate of the number of states via a penalised minimum distance method. This frequentist approach appears to work well, although the uncertainty regarding the number of states is not explicit and relies on asymptotic arguments to obtain consistent estimates which may not be appropriate for short sequences of data.

Bayesian methods appear to dominate the model selection problem of interest, and quantify more explicitly the model uncertainty by approximating the model posterior distribution. A reversible jump Markov chain Monte Carlo [RJMCMC, Green (1995)] approach seems natural for such a model selection problem where the sample space varies in dimension with respect to the number of underlying states assumed and has been applied in the HMM setting by Robert et al. (2000). This is an example of variable-dimension Monte Carlo as discussed in Scott (2002). However, RJMCMC is often computationally intensive and care is required in designing moves such that the sampling MC mixes well both within model spaces (same number of states, different parameters) and amongst model spaces (different number of states).

Chopin and Pelgrin (2004) and Chopin (2007) propose the sequential HMM (SHMM) framework where the number of distinct states visited by the latent MC up to that time point is considered. It is this augmented MC that is considered, sampled via sequential Monte Carlo (SMC) and used to determine the model posterior. SMC algorithms for dealing with (general state space) HMMs are usually termed *particle methods*. In the first instance these methods focussed upon solving the optimal filtering problem Gordon et al. (1993); Kitagawa (1996); later work, dating back at least to Kitagawa (1998), has attempted to address the problem of parameter estimation—a problem rather closer in spirit to that of model selection. By reformulating the problem in terms of this new augmented underlying MC and constructing the corresponding new HMM framework, Chopin (2007) essentially turns the problem into a filtering problem and thus the use of standard particle filtering techniques becomes possible. This setup also alleviates the problem of state identifiability as states are labelled in the order in which they are identified in the data sequence. This approach is particularly suited to online applications with respect to incoming observations.

The approach of Chopin (2007) is probably the state of the art. It avoids including the latent state sequence within the target posterior, but does rely upon the simulation of these variables within the Gibbs sampling transitions. It consequently benefits partially from Rao-Blackwellisation of the state sequence *but* the correlation between parameters and the latent state sequence means that this kernel need not necessarily enjoy good mixing properties: given a particular sequence of parameters, it is possible for the conditional state distribution to be highly concentrated and vice versa. Unlike the other methods discussed here, the sequential nature of this approach allows its use in online applications.

Scott (2002) proposes a standard Markov chain Monte Carlo (MCMC) methodology to approximate the model posterior. This uses a parallel Gibbs sampling scheme where each Gibbs sampler assumes a different number of states and is used to approximate the conditional marginal likelihood, and then combining to approximate the posterior of number of states. The use of MCMC, similarly to RJMCMC, requires good algorithmic design to ensure the MC is mixing well and converges.

The Bayesian methods outlined above approximate the model posterior, by jointly sampling the parameter and the state sequence, and marginalising as necessary. However, sampling the underlying state sequence can be particularly difficult, due to its high dimension and correlation, and is wasteful if the state sequence is not of interest. Alternative sampling techniques may thus be more suitable and appropriate if they can avoid having to sample the state sequence.

We take a similar approach to the parallel MCMC sampler approach of Scott (2002), in that we approximate the model posterior via the use of parallel SMC samplers, where each SMC sampler approximates the marginal likelihood and parameter posterior conditioned on the number of states. We combine these to approximate the model posterior of interest. A major advantage of the proposed approach is that the underlying state sequence is not sampled and thus less complex sampling designs can be considered. Below, we demonstrate that the SMC sampler approach can work well even with simple, generic sampling strategies. We note that we have been interested, particularly, in settings in which simple HMMs with a small number of states and a particular associated observation structure is featured. Such problems arise naturally in Econo-

metrics (see Sect. 4.2) and Neuroscience [Højen-Sørensen et al. (2000); Nam et al. (2012)]. In such settings, the benefits of simple automation would typically outweigh those available from more sophisticated strategies with an appreciable implementation cost and, as shown below, the simple strategies considered here can outperform more sophisticated techniques, which might be more appropriate in more complex settings, when considering these simple models.

If we are already required to approximate the model parameter posteriors conditioned on several different numbers of states (as would be the case for sensitivity analysis, for example), the framework requires no additional computational effort and leads to parameter estimates with smaller standard errors than competing methods.

The structure of this paper is as follows: Sect. 2 provides background on the statistical methods used. Section 3 outlines the proposed method. Section 4 applies the methodology to both simulated data and an Econometric GNP example. Section 5 concludes the paper.

2 Background

Let $y_1, \ldots, y_n$ denote a time series observed at equally spaced discrete points. One approach for modelling such a time series is via Hidden Markov Models (HMMs) which provide a sophisticated framework to model non-linear and non-stationary time series in particular. A HMM can be defined as in Cappé et al. (2005); a bivariate discrete time process $\{X_t, Y_t\}_{t\geq 0}$ where $\{X_t\}$ is a latent finite state Markov chain (MC), $X_t \in \Omega_X$, such that conditional on $\{X_t\}$, observation process $\{Y_t\}$ is a sequence of independent random variables where the conditional distribution of Y_t is completely determined by X_t. We consider general finite state HMMs (including Markov switching models) such that finite dependency on previous observations and states of X_t is permitted for an observation at time t. General finite state HMMs are of the form:

$$\begin{aligned} y_t|y_{1:t-1}, x_{1:t} &\sim f(y_t|x_{t-r:t}, y_{1:t-1}, \theta) \qquad \text{(Emission)} \\ p(x_t|x_{1:t-1}, y_{1:t-1}, \theta) &= p(x_t|x_{t-1}, \theta) \quad t = 1, \ldots, n \qquad \text{(Transition).} \end{aligned}$$

Without loss of generality, we assume $\Omega_X = \{1, \ldots, H\}$, $H < \infty$, with H, the number of underlying states our MC can take, often being known a priori before inference is performed. θ denotes the model parameters which are unknown and consist of the transition probabilities and the state dependent emission parameters. We use the standard notation of $U_{1:n} = (U_1, \ldots, U_n)$ for any generic sequence $U_1, U_2, \ldots$.

The Forward–Backward algorithm (Baum et al. 1970) allows us to compute the likelihood, $l(y_{1:n}|\theta, H)$, of an HMM exactly without sampling the underlying state sequence. We refer the reader to MacDonald and Zucchini (1997) and Cappé et al. (2005) for good overviews of HMMs.

In dealing with unknown θ, we take a Bayesian approach and consider the model parameter posterior conditioned on there being H states, $p(\theta|y_{1:n}, H)$. This is typically a complex distribution which cannot be sampled from directly, with numerical approximations such as Monte Carlo methods being required. We turn to SMC samplers to approximate this quantity (Del Moral et al. 2006). The SMC sampler is a

sampling algorithm used to sample from a sequence of distributions, $\{\pi_b\}_{b=1}^B$, defined over an arbitrary state sequence via importance sampling and resampling mechanisms. In addition, SMC samplers can be used to approximate the normalising constants, $\{Z_b\}_{b=1}^B$, for the sequence of distributions $\{\pi_b\}_{b=1}^B$ in a very natural way.

3 Methodology

We seek to approximate $p(H|y_{1:n})$, the posterior over the number of underlying states for a given realisation of data $y_{1:n}$ (the model posterior). Similar to the approaches of Robert et al. (2000); Scott (2002); Chopin and Pelgrin (2004); Chopin (2007), we assume a finite number of states, $H \in \{1, \ldots, H^{\max}\}$. Scott (2002) remark that this is a mild restriction; it is difficult to envisage using a model such as this without assuming $H \ll n$. Some methods, for example that of Beal et al. (2002), place no restriction on $H^{\max}$ via the use of a Dirichlet process based methodology. However, this also requires sampling the underlying state sequence via Gibbs samplers and requires approximating the likelihood via particle filters, neither of which is necessary under the proposed approach.

Via Bayes' Theorem,

$$p(H|y_{1:n}) \propto p(y_{1:n}|H)p(H)$$

where $p(y_{1:n}|H)$ denotes the marginal likelihood under model H, and $p(H)$ denotes the model prior. We are thus able to approximate the model posterior if we obtain the marginal likelihood associated with each model.

SMC samplers can be used to approximate the conditional parameter posterior, $p(\theta|y_{1:n}, H)$, and the associated marginal likelihood $p(y_{1:n}|H)$. We can define the sequence of distributions $\{\pi_b\}_{b=1}^B$ as follows:

$$\pi_b(\theta|H) = l(y_{1:n}|\theta, H)^{\gamma_b} p(\theta|H)/Z_b, \qquad b = 1, \ldots, B$$
$$Z_b = \int l(y_{1:n}|\theta, H)^{\gamma_b} p(\theta|H) d\theta$$

where conditioned on a specific model H, $p(\theta|H)$ is the prior of the model parameters and γ_b is a non-decreasing temperature schedule with $\gamma_1 = 0$ and $\gamma_B = 1$. We thus sample initially from $\pi_1(\theta|H) = p(\theta|H)$ either directly or via importance sampling, and introduce the effect of the likelihood gradually. We in turn sample and approximate the target distribution, the parameter posterior $p(\theta|y_{1:n}, H)$. As the evaluation of the likelihood does not require sampling the underlying state sequence, the distributions defined in the above equation including the parameter posterior, do not require the sampling of this quantity either. Monte Carlo error is consequently only introduced through the sampling of the parameters, leading to more accurate estimates (a Rao-Blackwellised estimate). This is one of many advantages compared to other approaches such as MCMC, where the underlying state sequence needs to be sampled.

Note that this setup is different to that proposed in Chopin and Pelgrin (2004) and Chopin (2007), where distributions are defined as $\pi_b = p(\theta|y_{1:b})$ with respect

to incoming observations. In addition to the use of a different tempering schedule, the approach of this paper has employed different proposal kernels within the SMC algorithm. The data tempering approach of Chopin (2007) facilitates online estimation; such a schedule could also be employed here but the nature of the computations involved are such that it would not lead to such substantial efficiency gains in our setting and we have preferred the geometric tempering approach which leads to a more regular sequence of distributions.

Z_B, the normalising constant for the parameter posterior $p(\theta|y_{1:n}, H) = p(\theta, y_{1:n}|H)/Z_B$, is more specifically of the following form,

$$Z_B = \int l(y_{1:n}|\theta, H)p(\theta|H)d\theta = \int p(y_{1:n}, \theta|H)d\theta = p(y_{1:n}|H).$$

That is, the normalising constant for the parameter posterior conditioned on model H, is the conditional marginal likelihood of interest. We note that here and elsewhere we have suppressed the dependence upon H from the notation as essentially every quantity in what follows is dependent upon H; of course, Z_B as described here is independent of B but the notation is consistent with that used in the algorithmic description and emphasises that it is the final value in the sequence of normalising constants computed within the algorithm. Given that we can approximate the marginal likelihood, we can thus approximate the model posterior as follows:

Algorithm outline:

1. For $h = 1, \ldots, H^{\max}$,
   (a) Approximate $p(y_{1:n}|H = h)$ and $p(\theta|y_{1:n}, H = h)$, the marginal likelihood (see Sect. 3.1) and parameter posterior [see Nam et al. (2012)] conditioned on h states, via SMC samplers.
2. Approximate $p(H = h|y_{1:n})$, the model posterior, via the approximation of $p(y_{1:n}|H = h)$ and model prior $p(H)$.

3.1 Approximating $p(y_{1:n}|H)$

SMC samplers can also be used to approximate normalising constants, Z_b, for the sequence of distributions, π_b, $b = 1, \ldots, B$. SMC samplers work on the principle of providing weighted particle approximations of distributions through importance sampling and resampling techniques. For a comprehensive exposition of SMC samplers, we refer the reader to Del Moral et al. (2006). The use of SMC to approximate normalising constants and to conduct model comparison using these approximations is well known; see Zhou et al. (2013) and references therein. The approach considered here is essentially the "SMC2" strategy described in Zhou et al. (2013) with the refinement of analytically integrating out the state sequence.

Using the SMC sampler for HMMs used within Nam et al. (2012) for parameter estimation and reproduced here as Algorithm 1, the main output of the SMC samplers algorithm is a series of weighted sample approximations of π_b, namely $\{\theta_b^i, W_b^i|H\}_{i=1}^N$, where N is the number of samples used in the SMC approximation. The approximation of the ratio between consecutive normalising constants can then be found as:

Algorithm 1 SMC algorithm for sampling from $p(\theta|y_{1:n}, H)$.

Initialisation: **Sample from prior,** $p(\theta|H)$, $b = 1$

For each $i = 1, \ldots, N$: Sample $\theta_1^i \sim p(\theta|H)$ and set $W_1^i = 1/N$.

for $b = 2, \ldots, B$ **do**

Reweighting: **For each** i compute:

$$W_b^i = \frac{W_{b-1}^i \widetilde{w}_b(\theta_{b-1}^i)}{\sum_{j=1}^N W_{b-1}^j \widetilde{w}_b(\theta_{b-1}^j)}$$

$$\text{where } \widetilde{w}_b(\theta_{b-1}^i) = \frac{\pi_b(\theta_{b-1}^i)}{\pi_{b-1}(\theta_{b-1}^i)} = \frac{l(y_{1:n}|\theta_{b-1}^i, H)^{\gamma_b}}{l(y_{1:n}|\theta_{b-1}^i, H)^{\gamma_{b-1}}}.$$

The likelihood, $l(y_{1:n}|\theta_{b-1}^i, H)$ can be computed exactly via the Forward–Backward equations (Baum et al. 1970).

Selection: **if** $ESS < T$ **then** Resample.

Mutation:

for each $i = 1, \ldots, N$: Sample $\theta_b^i \sim K_b(\theta_{b-1}^i, \cdot)$ where K_b is a π_b invariant Markov kernel.

end for

Output: Clouds of N weighted particles, $\{\theta_b^i, W_b^i|H\}_{i=1}^N$, approximating distribution $\pi_b \propto l(y_{1:n}|\theta, H)^{\gamma_b} p(\theta|H)$ for $b = 1, \ldots, B$.

$$\frac{Z_b}{Z_{b-1}} \approx \frac{\widehat{Z_b}}{Z_{b-1}} = \sum_{i=1}^N W_{b-1}^i \widetilde{w}_b(\theta_{b-1}^i) := \bar{W}_b.$$

This ratio corresponds to the normalising constant for weights at iteration b. Z_B, can thus be approximated as:

$$\widehat{Z_B} = \widehat{Z_1} \prod_{b=2}^B \bar{W}_b$$

which, remarkably, is an unbiased estimator of the true normalising constant (Del Moral 2004).

Note that the normalising constant, Z_b, corresponds to the the following quantity

$$\pi_b(\theta) = \frac{\varphi_b(\theta)}{Z_b}$$

where φ_b is the unnormalised density. We can thus approximate the marginal likelihood by simply recording the normalising constants for the weights, $\bar{W}_b$, at each iteration of Algorithm 1.

There is a great deal of flexibility with the SMC implementation and some design decisions are necessarily dependent upon the model considered. We have found that a reasonably straightforward strategy works well for the class of HMMs which we consider. An example implementation, similar to that discussed in Nam et al. (2012), is as follows: we set $\gamma_b = \frac{b-1}{B-1}$. As transition probabilities matrices are a fundamental component in HMMs, we initialise as follows: consider the transition probability vectors,

$p_h = (p_{h1}, \ldots, p_{hH}), h = 1, \ldots, H$ such that $\mathbf{P} = \{p_1, \ldots, p_H\}$, and sample from the prior $p_h \overset{\text{iid}}{\sim} \text{Dir}(\alpha_h), h = 1, \ldots, H$ where α_h is a H-long hyperparameter vector. As HMMs are associated with persistent behaviour, we choose α_h which reflects this type of behaviour. Relatively flat priors are generally implemented for the emission parameters. Random Walk Metropolis (RWM) proposal kernels are used in this paper for the mutation step of the algorithm. Details of specific implementation choices are given for representative examples in the following section.

It is appropriate to note that the choice of proposal distribution will influence, possibly very substantially, the variance of estimates obtained with the algorithm. Here we focus on dealing with simple HMMs which arise in a variety of application areas for which the inherent robustness of the SMC approach allows the use of a simple random walk proposal with scale fixed a priori (which could be set using a small pilot run). In more complicated settings it may be necessary to employ more sophisticated kernels. We have favoured this simple strategy as it requires minimal implementation effort and mixes well for problems of the type in which we are interested; in more complex settings this certainly will not be the case. First, we note that other choices, including the Gibbs sampling strategy of Chopin (2007), can be readily employed within the tempering algorithm used here. Second, it is worthwhile noting that adaptive strategies can be readily employed within SMC algorithms—see Zhou et al. (2013) for an illustration of such a technique within a Bayesian Model Selection context, including a mixture model example in which the posterior exhibits many of the same characteristics as that found here and Beskos et al. (2013) for a recent theoretical analysis demonstrating consistency (although not necessarily unbiasedness) of the evidence estimates obtained by such adaptive algorithms.

4 Results

This section applies the methodology to a variety of simulated and real data. All results have been obtained using the approach of Sect. 3 with the following settings. $N = 500$ samples and $B = 100$ iterations have been used to approximate the sequence of distributions. Additional sensitivity analysis has been performed with respect to larger values of N and B which we found reduced the Monte Carlo variability of estimates, as would be expected, but for practical purposes samples of size 500 were sufficient to obtain good results. α_h is a H-long hyperparameter vector full of ones, except in the h-th position where a 10 is present. This encourages the aforementioned persistent behaviour in the underlying MC associated with HMMs. The linear tempering schedule and proposal variances used have not been optimised to ensure optimal acceptance rates. Promising results are obtained with these simple default settings.

For the model selection results, a uniform prior has been assumed over the model space in approximating the model posterior. We consider selecting the maximum a posterior (MAP) model, that is $\arg\max_{h=1,\ldots,H^{\max}} p(H = h|y_{1:n})$, as this indicates the strongest evidence for the model favoured by the observed data.

The R language (R Core Team 2013) source code used to produce these numerical results is available as supplementary material.

4.1 Simulated data

We consider simulated data generated by two different models; Gaussian Markov Mixture (GMM) and Hamilton's Markov Switching Autoregressive model of order r [HMS-AR(r), Hamilton (1989)]. These are mathematically defined as follows:

$$Y_t|X_t \sim \mathrm{N}(\mu_{X_t}, \sigma^2_{X_t}) \qquad \text{(GMM)},$$
$$Y_t|X_{t-r:t} \sim \mathrm{N}\Big(\mu_{X_t} + \sum_{j=1}^{r} \phi_j (Y_{t-j} - \mu_{X_{t-j}}), \sigma^2\Big) \qquad \text{(HMS-AR}(r)\text{)}.$$

The first model has been chosen due to its relative simplicity and connection to mixture distributions, and the latter is a more sophisticated model which can be used to model Econometric GNP data (Hamilton 1989) and brain imaging signals (Peng et al. 2011). HMS-AR models can be seen as an extension of GMM models such that only the underlying mean switches, the variance is state invariant, and dependency on previous observations is induced in an autoregressive nature into the mean. For various scenarios under these two models, we present an example realisation of the data from the same seed (left column) and the model selection results from 50 data realisations (right column). Changes in state in the underlying state sequence occur at times 151, 301 and 451. We consider a maximum of five states, $H^{\max} = 5$, as we believe that no more than five states are required to model the business cycle data example we will consider later, and the simulations are designed to reflect this.

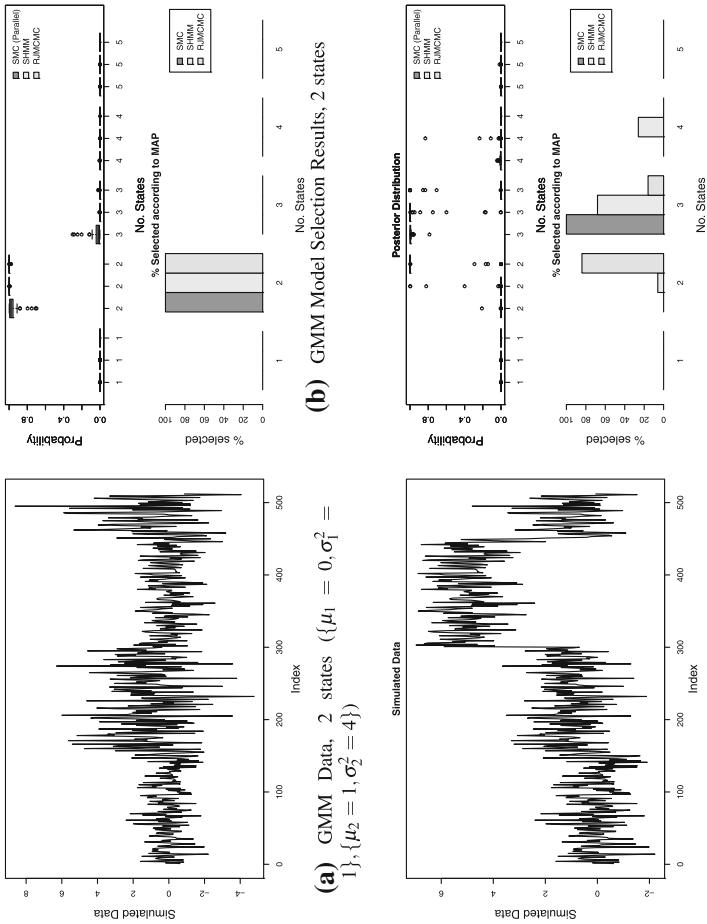
The following priors have been used for the state-dependent mean and precision (inverse of variance) parameters: $\mu_h \overset{\text{iid}}{\sim} \mathrm{N}(0, 100)$, $\frac{1}{\sigma_h^2} \overset{\text{iid}}{\sim} \mathrm{Gamma}(\text{shape} = 1, \text{scale} = 1), h = 1, \ldots, H$. For the HMS-AR model, we consider the partial autocorrelation coefficients (PAC, ψ_1) in place of AR parameter, ϕ_1, with the following prior, $\psi_1 \sim \mathrm{Unif}(-1, 1)$. The use of PAC allows us to maintain stationarity amongst the AR coefficients more efficiently, particularly in higher AR order settings.

Algorithmically, RWM transitions were used with a baseline proposal variances of 10 have been used for each parameter's mutation step which decrease linearly as a function of sampler iteration. For example, the proposal variance $\frac{\sigma^2_\mu}{b} = 10/b$ is used for μ_h mutations during iteration b.

4.1.1 Gaussian Markov mixture

Figure 1 displays results obtained for a GMM model under the proposed parallel SMC methodology. In addition, we compare our model selection results to the sequential Hidden Markov Model (SHMM) approach as proposed in Chopin (2007)[1], and a generic Reversible Jump Markov Chain Monte Carlo (RJMCMC) method via the `R` package `RJaCGH` (Rueda and Diaz-Uriarte 2011). The model posterior approx-

[1] Computer code for which was made available at: http://www.blackwellpublishing.com/rss/Volumes/Bv69p2.htm.



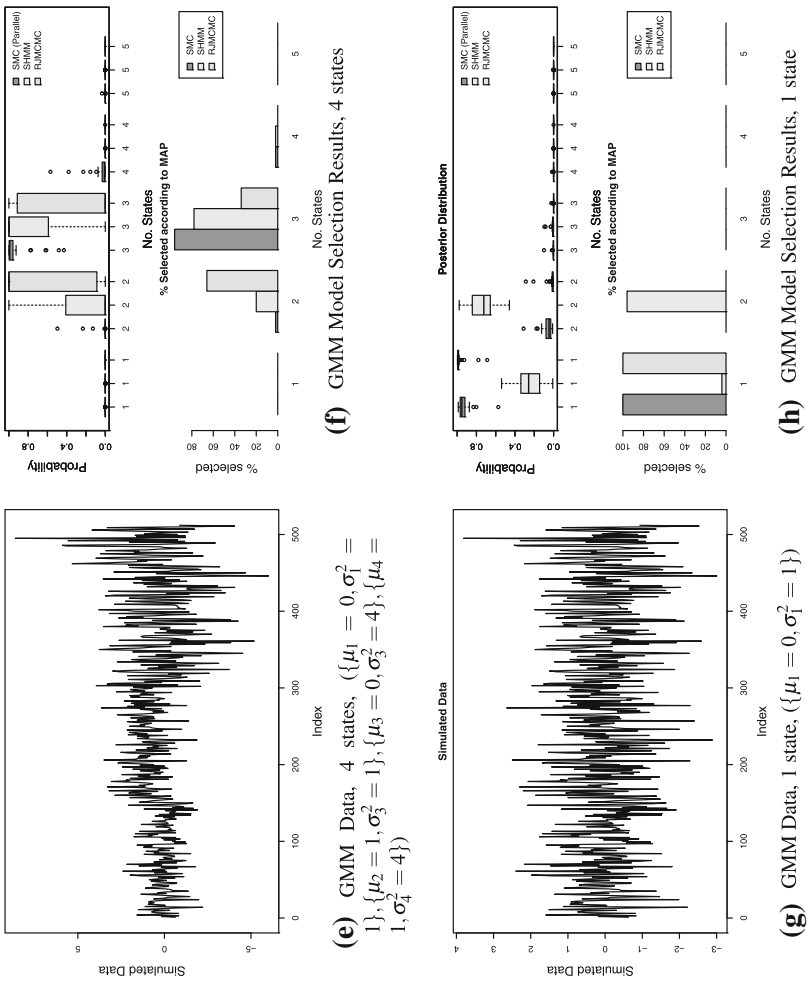
(a) GMM Data, 2 states $(\{\mu_1 = 0, \sigma_1^2 = 1\}, \{\mu_2 = 1, \sigma_2^2 = 4\})$

(b) GMM Model Selection Results, 2 states

(c) GMM Data, 3 states, $(\{\mu_1 = 0, \sigma_1^2 = 1\}, \{\mu_2 = 1, \sigma_2^2 = 1\}, \{\mu_3 = 5, \sigma_3^2 = 1\})$

(d) GMM Model Selection Results, 3 states

Fig. 1 Model selection results for a variety of Gaussian Markov Mixture Data. *Left column* shows examples of data realisations, *right column* shows the model selection results; *box plots* of the model posterior approximations under the parallel SMC, SHMM and RJMCMC approaches, and percentage selected according to maximum a posterior (*MAP*). Results are from 50 different data realisations



(e) GMM Data, 4 states, $(\{\mu_1 = 0, \sigma_1^2 = 1\}, \{\mu_2 = 1, \sigma_3^2 = 1\}, \{\mu_3 = 0, \sigma_3^2 = 4\}, \{\mu_4 = 1, \sigma_4^2 = 4\})$

(f) GMM Model Selection Results, 4 states

(g) GMM Data, 1 state, $(\{\mu_1 = 0, \sigma_1^2 = 1\})$

(h) GMM Model Selection Results, 1 state

Fig. 1 continued

imations from both approaches are displayed alongside the parallel SMC posterior approximations.

The following settings have been used for the SHMM implementation; $N = 5000$ samples have been used to approximate the sequence of distributions, $\pi'_b = p(\theta, x_{1:b}|y_{1:b})$, $H^{\max} = 5$ as the maximum number of states possible and one SMC replicate per dataset. The same prior settings under the proposed parallel SMC samplers have been implemented. Other default settings in the SHMM code such as model averaging being performed have been utilised.

The following settings have been utilised for the RJMCMC method which samples directly from the target distribution $p(\theta, x_{1:n}, H|y_{1:n})$: a 1000 burn-in with 25,000 sampler iterations thereafter. Equivalent or similar prior settings have been utilised; that is $\mu_h \sim \mathrm{N}(0, 10^2)$, $\sigma_h \sim \mathrm{Unif}(0, 5)$, for $h = 1, \ldots, H^{\max} = 5$. Other default settings in the RJMCMC package have been utilised, for example regarding jump parameter settings.

Figure 1a and b concerns a simple two-state scenario with changing mean and variance simultaneously. From the data realisation, it is evident that two or more states are appropriate in modelling such a time series. This is reflected in the model selection results with a two-state model being significantly the most probable under the model posterior from all simulations, and always correctly selected under MAP. However, uncertainty in the number of appropriate states is reflected with probability assigned to a three-state model amongst the simulations. These results indicate that the methodology works well on a simple, well-defined toy example. Results concur with the SHMM and RJMCMC framework; a two-state model is most probable for all simulations.

Figure 1c and d displays results from a similar three-state model, where different means correspond to the different states with subtle changes in mean present, for example around the 151 time point. Such subtle changes are of interest due to the subtle changes in mean associated in the GNP data considered later. The correct number of states is significantly the most probable under all simulations, and always correctly identified under MAP selection. Under the SHMM approach, more variability is present amongst the simulations which can lead to differing MAP estimates. A three-state model is largely the most probable, although some approximations display a four- or two-state model also being the most probable. A two-state model is identified as the most probable in most simulations under the RJMCMC approach which evidently does not concur with the truth.

Figure 1e and f displays results from a challenging scenario of changes in both subtle mean and variance, independently of each other, with four states being present. The SMC methodology is unable to correctly identify the number of states, with three states being the most probable and most selected model from the majority of the simulations. However, given the example data realisation it seems likely that the data does not strongly support a four-state model. This underestimation of the number of states is likely to be a consequence of the final two states being adequately explained using a single state (owing in part to the shortness of the associated observation intervals). Some probability has been associated with four- and two-state models, however. In addition the variability in the approximation of the model posterior is more pronounced for this simulation scenario, a result of the challenging scenario presented. The SHMM

and RJMCMC also perform similarly, with probability being assigned to two-state and three-state models and failing to identify the correct model.

Figure 1g and h presents results from a one-state GMM model, a stationary Gaussian process. Of interest here is whether our methodology is able to avoid overfitting even though a true HMM is not present. The model selection results highlight that overfitting is successfully avoided with a one-state model being most probable under the model posterior for all simulations and always the most selected under MAP. The RJMCMC method also performs well in this setting. The SHMM method, however, attaches substantially great probability to a two-state model than to a one-state model.

We also consider comparing the samples approximating the true emission parameters under the three methods. We consider the presented data scenarios of Fig. 1a and c where the proposed SMC and SHMM both concur with respect to the number of underlying states identified via MAP and the truth. For the RJMCMC method, the same reasoning applies for the first data scenario (Fig. 1a), although for the second data scenario (Fig. 1c) we assume the true number of states and SMC and SHMM MAP estimates for a valid comparison. To perform inference regarding the emission parameters, an identifiability constraint is enforced to allow for valid comparisons. This is achieved by re-ordering the states with respect to ascending means post SMC samplers, that is: $\mu_1^{(i)} < \mu_2^{(i)} < .... < \mu_H^{(i)}$ for each SMC particle [the extent to which these means are separated relative to the posterior uncertainty will determine the degree of bias introduced by such an ordering constraint and it is appropriate to select such constraints ex post to minimise this bias; see Sect. 4.1, Celeux et al. (2000)].

Table 1 displays the averaged posterior means and standard error for each emission parameter over the 50 simulations. The SMC methodology is more accurate in estimating the true value, and the standard error is smaller compared to the estimates provided by SHMM, and on a par with the RJMCMC approach. Note that the RJMCMC's on par performance with the SMC methodology requires 25,000 sampling iterations, compared to the 2,500 samples across the five SMC samplers under the

Table 1 Averaged posterior means and standard error for each emission parameter over the 50 simulations for the two data scenarios considered

	μ_1	μ_2	μ_3	σ_1	σ_2	σ_3
Truth	0	1	–	1	2	–
SMC	0.00 (0.06)	0.99 (0.14)	–	1.00 (0.04)	2.03 (0.10)	–
SHMM	0.05 (0.18)	0.94 (0.23)	–	1.06 (0.19)	1.97 (0.21)	–
RJMCMC	0.00 (0.06)	0.99 (0.15)	–	1.01 (0.04)	2.02 (0.09)	–
Truth	0	1	5	1	1	1
SMC	0.00 (0.09)	1.00 (0.08)	5.00 (0.08)	1.00 (0.06)	1.01 (0.05)	1.01 (0.06)
SHMM	0.70 (0.19)	1.50 (0.38)	3.51 (33.86)	1.01 (0.06)	1.01 (0.06)	1.07 (0.35)
RJMCMC	−0.24 (1.14)	2.34 (0.66)	4.75 (0.44)	1.32 (0.26)	1.29 (0.24)	1.31 (0.19)

We compare the proposed parallel SMC, SHMM and RJMCMC method. Averaged standard errors are denoted in the parentheses. Results indicate that the SMC outperforms the SHMM and RJMCMC method with greater accuracy in approximating the true values and smaller standard errors

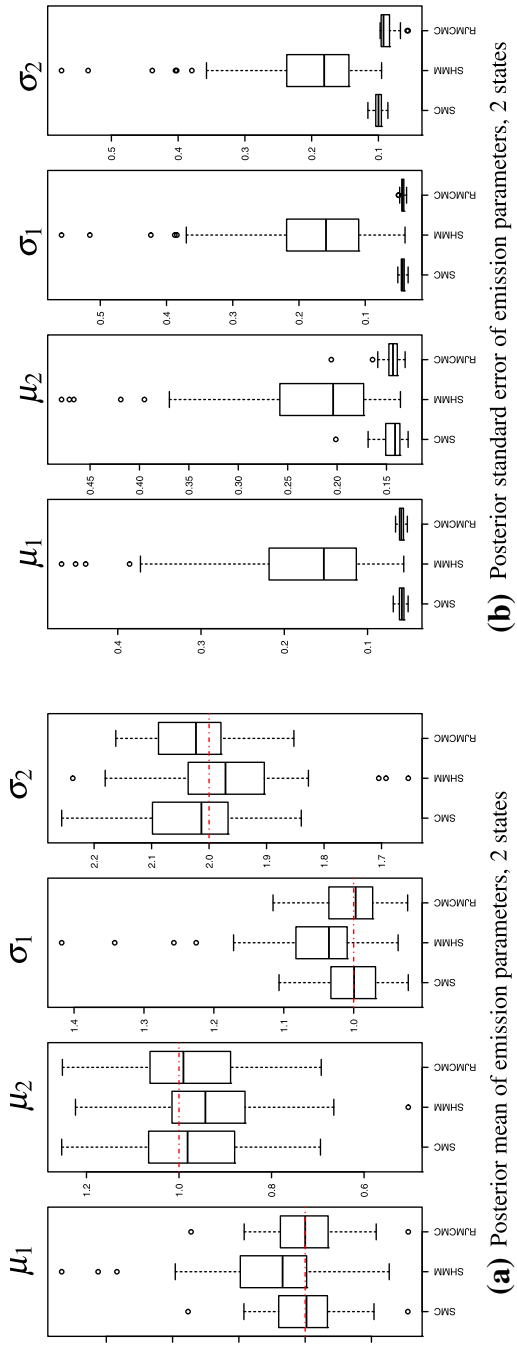
proposed approach. In addition, the poor performance of RJMCMC in the second scenario is due to its MAP estimate not coinciding with the true: it experiences relatively few visits to the three-state model and produces relatively few parameter samples from this model.

Figure 2 displays box plots of the posterior means (2a and c) and standard error (2b and d) of the emission parameter estimates for all 50 simulations. The posterior mean box plots indicate further that the proposed parallel SMC approach is generally more accurate and centred around the true emission parameter values (horizontal red dotted lines) for all simulations. The SHMM estimates are generally less precise with a greater range of values present. Similarly, the standard error box plots indicate that the standard error is less for the proposed SMC methodology compared to the SHMM method. These plots also indicate further, on par performance between the proposed SMC and RJMCMC approach in the first scenario (Fig. 2a and b), and poor performance by the RJMCMC method in the second scenario (Fig. 2c and d) due to the aforementioned reason.

The results indicates that in addition to identifying the correct model more often, more accurate estimates are obtained under the proposed SMC approach, compared to the existing SHMM and RJMCMC methods. This is presumably a result of the Rao-Blackwellised estimator provided by the SMC samplers framework, despite more samples being used under the SHMM approach. As fewer samples are required to achieve good, accurate estimates, the proposed parallel SMC method would appear to be more computationally efficient. In addition, while not directly comparable, the runtime for the SMC samplers approach for one time series was approximately 15 min to consider the five possible model orders using $N = 500$ particles [implemented in the R language (R Core Team 2013)], while it was approximately 90 min for the SHMM approach with the default N=5000 particles [implemented in MATLAB (MATLAB 2012)]. The RJMCMC methodology takes approximately 30 min to run under the 1000 burn-in and 25,000 sampling iterations setup.

4.1.2 Hamilton's Markov switching autoregressive model of order r, HMS-AR(r)

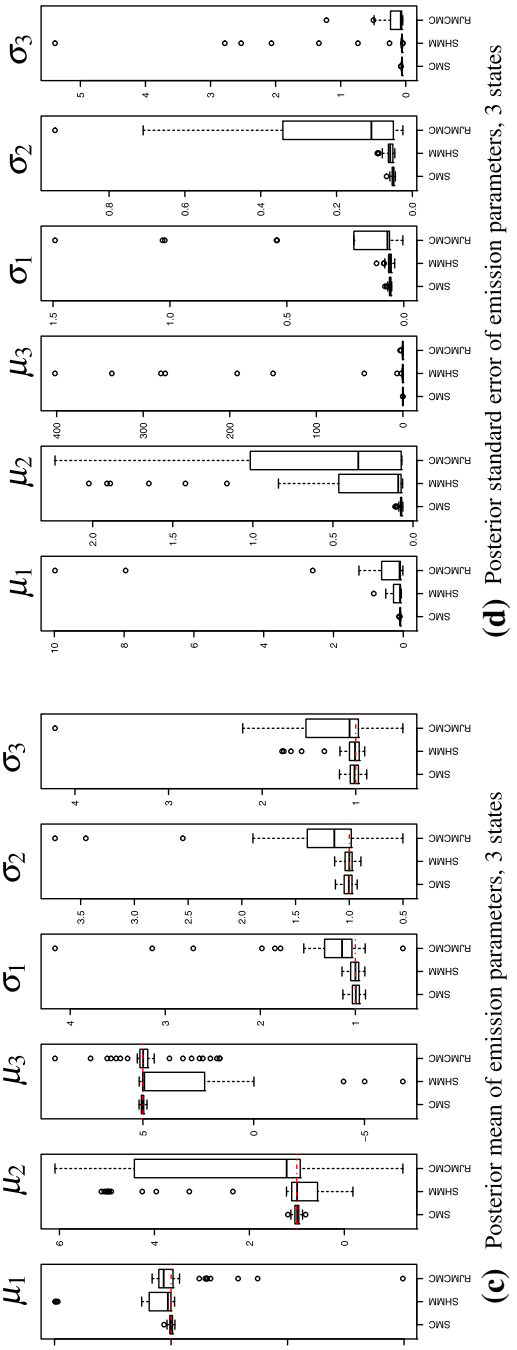
Figure 3 shows results from a HMS-AR model with autoregressive order one; we assume that this autoregressive order is known a priori although the SMC methodology could easily be extended to consider model selection with respect to higher AR orders. The following results were obtained using data generated using a two-state model, with varying autoregressive parameter, ϕ_1, and the same means and variance used for each scenario ($\mu_1 = 0, \mu_2 = 2, \sigma^2 = 1$). Interest lies in how sensitive the model selection results are with respect to ϕ_1. For small values of ϕ_1 (for example $\phi_1 = 0.1, 0.5$) indicating small dependency on previous observations, our methodology works well with the correct number of true states being highly probable and always the most selected according to MAP. Relatively little variability exists in the approximation of the model posterior. However, as ϕ_1 begins to increase and tend towards the unit root, for example $\phi_1 = 0.9$, we observe that more uncertainty is introduced into the model selection results, with greater variability in the model posterior approximations and alternative models being selected according to MAP. However, as the data realisation



(a) Posterior mean of emission parameters, 2 states

(b) Posterior standard error of emission parameters, 2 states

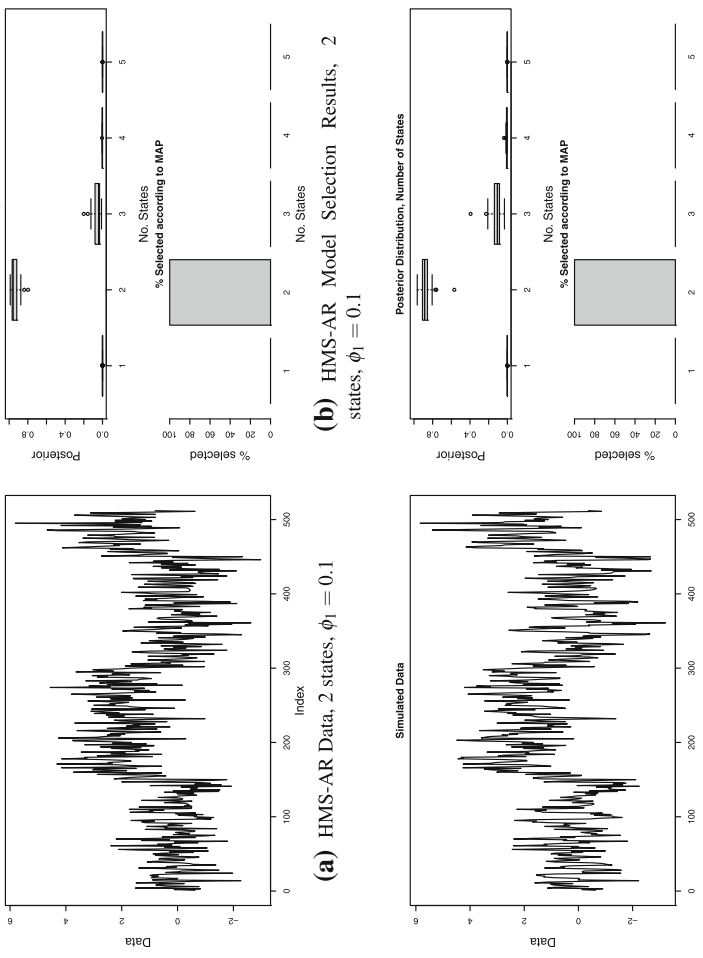
Fig. 2 *Box plots* of the posterior means and standard error for each emission parameter over 50 data realisations. *Red dotted values* denote the value of the emission parameter used to generate the simulated data in the posterior mean box plots. We compare the results under the three approaches: parallel SMC, SHMM and RJMCMC. We observe that the proposed SMC approach outperforms SHMM and RJMCMC with posterior means centred more accurately around the true values, and the standard error for the samples being smaller



(c) Posterior mean of emission parameters, 3 states

(d) Posterior standard error of emission parameters, 3 states

Fig. 2 continued



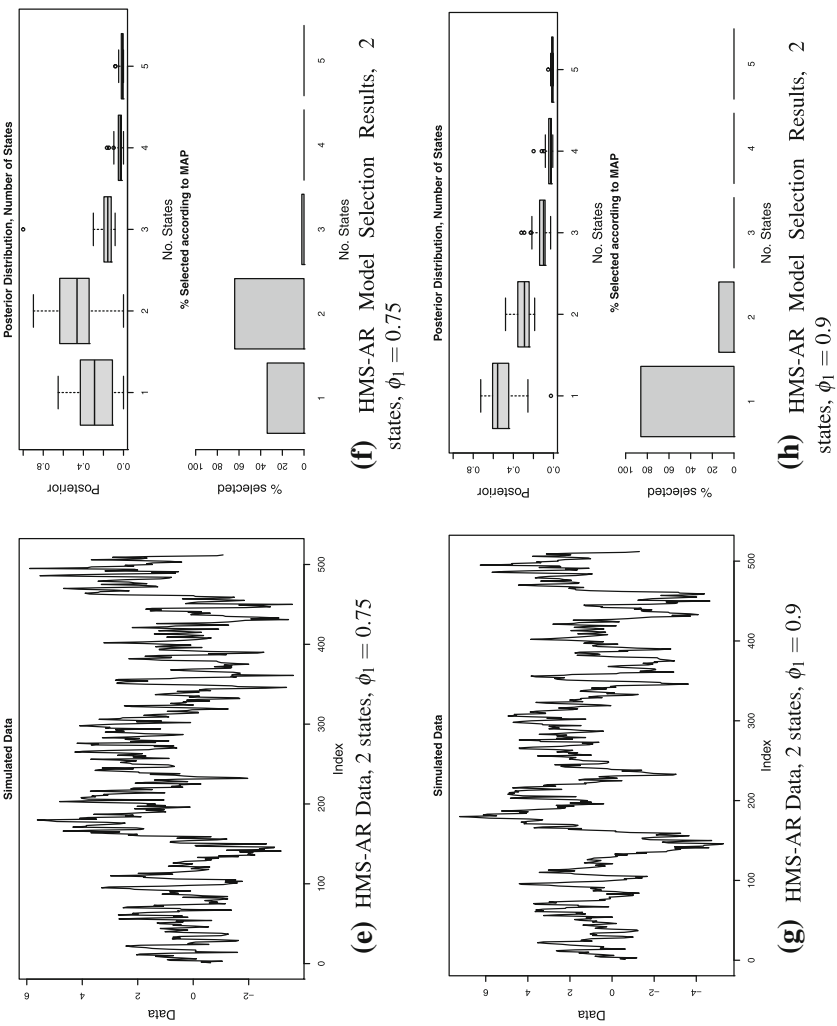
(a) HMS-AR Data, 2 states, $\phi_1 = 0.1$

(b) HMS-AR Model Selection Results, 2 states, $\phi_1 = 0.1$

(c) HMS-AR Data, 2 states, $\phi_1 = 0.5$

(d) HMS-AR Model Selection Results, 2 states, $\phi_1 = 0.5$

Fig. 3 Model selection results for variety of HMS-AR(1) data with $\mu_1 = 0$, $\mu_2 = 2$, $\sigma^2 = 1$ and varying ϕ_1. *Left column* shows examples of data realisations, *right column* shows the parallel SMC model selection results from 50 data realisations; approximations of the model posterior, and percentage selected according to maximum a posterior (MAP)



(e) HMS-AR Data, 2 states, $\phi_1 = 0.75$

(f) HMS-AR Model Selection Results, 2 states, $\phi_1 = 0.75$

(g) HMS-AR Data, 2 states, $\phi_1 = 0.9$

(h) HMS-AR Model Selection Results, 2 states, $\phi_1 = 0.9$

Fig. 3 continued

in Fig. 3g suggests, the original two-state model is hard to identify and these estimates simply reflect the associated model uncertainty. These results indicate that the proposed model selection method works for sophisticated models such as HMS-AR models, although the magnitude of the autoregressive coefficient may affect results.

4.2 Hamilton's GNP data

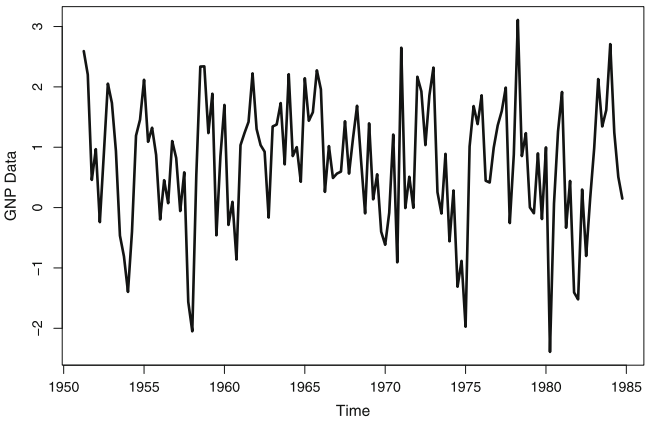
Hamilton's GNP data (Hamilton 1989) consist of differenced quarterly logarithmic US GNP between the time periods 1951:II to 1984:IV. Hamilton (1989) and Aston et al. (2011) model y_t, the aforementioned transformed data consisting of 135 observations, by a two-state HMS-AR(4) model, before performing analysis regarding identification of starts and ends of recessions. The two underlying states denote "Contraction" and "Expansion" states to correspond directly with the definition of a recession; two consecutive quarters of contraction. Whilst such a model works in practice for recession inference, we investigate whether a two-state HMS-AR(4) model is indeed appropriate. We assume the autoregressive order of four, is known a priori relating to annual dependence, and is adequate in modelling the data. We assume a maximum of five possible states in the HMM framework ($H^{\max} = 5$) as we believe that the data arises from at most five possible states for the particular time period considered.

The following priors have been used: for the means, $\mu_h \overset{\text{iid}}{\sim} \text{N}(0, 10)$, $h = 1, \ldots, H$, precision (inverse variance) $\frac{1}{\sigma^2} \sim \text{Gamma}(\text{shape} = 1, \text{scale} = 1)$, PAC coefficients $\psi_j \overset{\text{iid}}{\sim} \text{Unif}(-1, 1)$, $j = 1, \ldots, 4$. A uniform prior has been used over the number of states H.

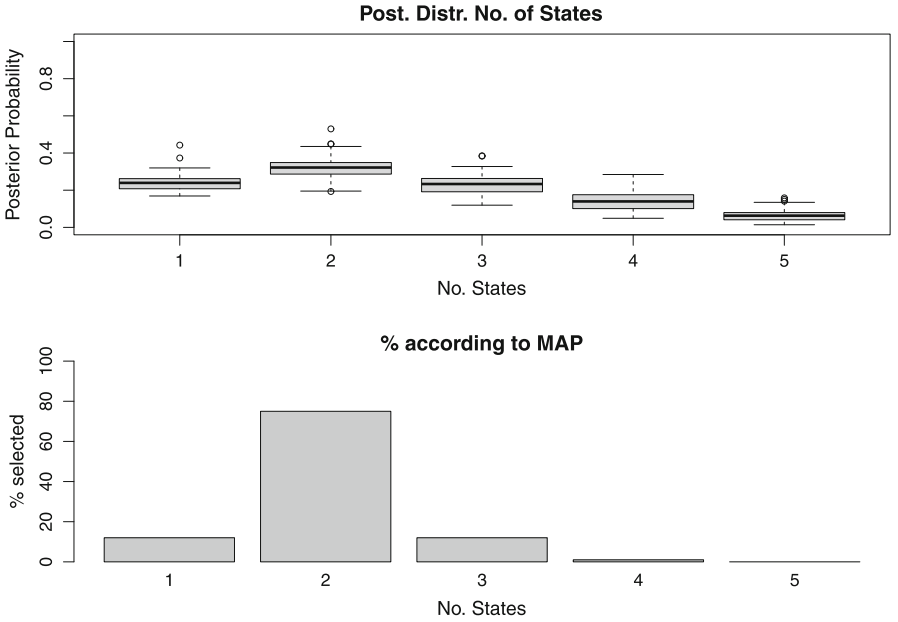
Again, Random Walk Metropolis proposals were employed. In all cases, these univariate proposals employed a baseline proposal variance of 10, and this was allowed to fall linearly with the tempering parameter.

Figure 4 displays the corresponding dataset and model selection results from 100 different SMC replicates. The model selection results, Fig. 4b, demonstrate that there is uncertainty in the appropriate number of underlying states with non-negligible probability assigned to each model considered and variability amongst the SMC replication results. Some of the alternative models seem plausible, for example a one-state model given the plot of the data and the additional underlying states modelling the subtle nuances and features in the data. However, a two-state model is the most frequently selected under a MAP criterion. In addition, the distribution appears to tail off as we consider more states, thus indicating that value of $H^{\max}$ used is appropriate. In conclusion, the two-state HMS-AR(4) model assumed by Hamilton (1989) does seem adequate in modelling the data although this is not immediately evident and uncertainty is associated with the number of underlying states.

Note that the variability in the model selected using MAP between repeated runs is due to the small but non-negligible sampling variability in the posterior model probabilities. Additional simulations (not shown) verify that moderate increases in the SMC sample size (a factor of ten suffices), N, are sufficient to eliminate this variability.



(a) Hamilton's GNP data: differenced quarterly logarithmic US GNP between 1951:II to 1984:IV.



(b) Model Selection Results for GNP data

Fig. 4 Model selection results for Hamilton's GNP data under the proposed model selection methodology. **a** displays the analysed transformed GNP data. **b** displays the model posterior approximations from 100 SMC replications, and percentage selected under maximum a posterior (MAP)

5 Conclusion and discussion

This paper has proposed a methodology in which the number of underlying states, H, in a HMM framework can be determined by the use of parallel sequential Monte Carlo samplers. Through a combination of well-known individual steps, the method outper-

forms state-of-the-art competing methods in the restricted setting of HMM model selection. Conditioned on the number of states, the conditional marginal likelihood can be approximated in addition to the parameter posterior via SMC samplers. By conditioning on a different number of states, we can obtain marginal likelihoods under each model. These marginal likelihoods can then be combined with an appropriate prior to approximate the model posterior, $p(H|y_{1:n})$, of interest. The use of SMC samplers within a HMM framework results in an computationally efficient and flexible framework such that the underlying state sequence does not need to be sampled unnecessarily compared to other methods which reduces Monte Carlo error of parameter estimates, and complex design algorithms are not required.

The SMC methodology has been demonstrated on a variety of simulated data and GNP data and shows good results, even in challenging scenarios where subtle changes in emission parameters are present. The results on the GNP data have further confirmed that a two-state HMS-AR model assumed in previous studies and analysis is appropriate, although the uncertainty associated with the number of underlying states has now been captured.

In the settings considered, the method performs at least as well as, and often rather better than, other state-of-the-art approaches in the literature such as the SHMM approach proposed in Chopin (2007) and a generic RJMCMC method (Rueda and Diaz-Uriarte 2011). It is interesting to note that, often when dealing with estimation in a time series setting an online approach leads to better performance even when the analysis is being performed in an offline setting. The marginalisation possible with the proposed method allows for better results to be obtained using a simple direct approach.

While our methodology can be applied in general situations, we have concentrated in the simulations and real data analysis on the situation here where H is small. This is often the case in many applied analyses, and we have shown that the approach performs well with considerable savings in computational efficiency and Monte Carlo error.

From a modelling perspective, the model selection results presented in this paper have assumed a uniform prior over the collection models considered but there would be no difficulty associated with the use of more complex priors. Perhaps more important in the context of model selection is the specification of appropriate priors over model parameters, which can have a significant influence on model selection results: some sensitivity analysis should always be conducted to assess the impact of parameter priors on model selection results obtained by any method. From the perspective of computational efficiency it is desirable to identify a value of $H^{\max}$ which is sufficiently large to allow for good modelling of the data but not so large that the computational cost of evaluating all possible models becomes unmanageable (noting that the cost of dealing with any given model is, with the proposed method as well as most others, an increasing function of the complexity of that model).

Finally, we note that, although this paper has focused predominantly on a retrospective, offline context, in an online context it would be possible to consider the sequence of distributions defined by $\pi'_b(\theta|H) \propto l(y_{1:b}|\theta, H)p(\theta|H)$, rather than $\pi_b(\theta|H) \propto l(y_{1:n}|\theta, H)^{\gamma_b} p(\theta|H)$ in a similar spirit to the SHMM approach but without the simulation of the latent state sequence even as auxiliary variables.

Acknowledgments The authors would like to thank Nicolas Chopin for making available the computer code for the SHMM algorithm. Two anonymous referees provided comments on an earlier version of the manuscript which led to a significantly improved final version.

References

Albert, P.S. (1991). A two-state Markov mixture model for a time series of epileptic seizure counts. *Biometrics, 47*(4), 1371–1381, http://www.jstor.org/stable/2532392.

Aston, J. A. D., Peng, J. Y., Martin, D. E. K. (2011). Implied distributions in multiple change point problems. *Statistics and Computing*, *22*, 981–993.

Baum, L.E., Petrie, T., Soules, G., Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics, 41*(1), 164–171, http://www.jstor.org/stable/2239727.

Beal, M., Ghahramani, Z., Rasmussen, C. (2002). The infinite Hidden Markov Model. *Advances in Neural Information Processing Systems*, *14*, 577–584.

Beskos, A., Jasra, A., Thiéry, A.H. (2013). On the convergence of adaptive sequential Monte Carlo methods. *Mathematics e-print* 1306.6462, ArXiv.

Cappé, O., Moulines, E., Rydén, T. (2005). *Inference in Hidden Markov Models*. New York, USA: Springer.

Celeux, G., Hurn, M., Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, *95*(451), 957–970.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics*, *86*, 221–241.

Chopin, N., Pelgrin, F. (2004). Bayesian inference and state number determination for Hidden Markov Models: an application to the information content of the yield curve about inflation. *Journal of Econometrics, 123*(2), 327–344. doi:10.1016/j.jeconom.2003.12.010, http://www.sciencedirect.com/science/article/B6VC0-4BJX37R-1/2/f34ca1f662b663107cd87cf76218159a.

Chopin, N. (2007). Inference and model choice for sequentially ordered Hidden Markov Models. *Journal of the Royal Statistical Society Series B*, *69*(2), 269–284.

Del Moral, P. (2004). *Feynman-Kac formulae: genealogical and interacting particle systems with applications. Probability and Its Applications*. New York, USA: Springer.

Del Moral, P., Doucet, A., Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society Series B*, *68*(3), 411–436.

Eddy, S. R. (2004). What is a Hidden Markov Model? *Nature Biotechnology*, *22*, 1315–1316. doi:10.1038/nbt1004-1315.

Gordon, N. J., Salmond, S. J., Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings-F*, *140*(2), 107–113.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*(4), 711–732. doi:10.1093/biomet/82.4.711.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, *57*(2), 357–384.

Højen-Sørensen, P., Hansen, L. K., Rasmussen, C. E. (2000). Bayesian modelling of fMRI time series. *Bayesian Modelling of fMRI Time Series*, *12*, 754–760.

Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics, 5*(1), 1–25, http://www.jstor.org/stable/1390750.

Kitagawa, G. (1998). A self-organizing state-space-model. *Journal of the American Statistical Association*, *93*(443), 1203–1215.

Konishi, S., Kitagawa, G. (2008). *Information criteria and statistical modeling*. New York, USA: Springer.

MacDonald, I. L., Zucchini, W. (1997). *Monographs on statistics and applied probability 70: Hidden Markov and other models for discrete-valued time series*. Boca Raton, Florida, USA: Chapman & Hall / CRC.

Mackay, R. (2002). Estimating the order of a Hidden Markov Model. *Canadian Journal of Statistics*, *30*(4), 573–589.

MATLAB (2012) version 7.14.0 (R2012a). The MathWorks Inc., Massachusetts, USA: Natick.

Nam, C. F. H., Aston, J. A. D., Johansen, A. M. (2012). Quantifying the uncertainty in change points. *Journal of Time Series Analysis*, *33*(5), 807–823. doi:10.1111/j.1467-9892.2011.00777.x.

Peng, J. Y., Aston, J. A. D., Liou, C. Y. (2011). Modeling time series and sequences using Markov chain embedded finite automata. *International Journal of Innovative Computing Information and Control*, *7*, 407–431.

R Core Team (2013) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/.

Rabiner, L. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286. doi:10.1109/5.18626.

Robert, C.P., Rydén, T., Titterington, D.M. (2000). Bayesian inference in Hidden Markov Models through the reversible jump Markov chain Monte Carlo method. *Journal of the Royal Statistical Society: Series B* (Statistical Methodology), *62*(1), 57–75. doi:10.1111/1467-9868.00219.

Rueda, O., Diaz-Uriarte, R. (2011). RJaCGH: Reversible Jump MCMC for the analysis of CGH arrays. http://CRAN.R-project.org/package=RJaCGH, R package version 2.0.2.

Scott, S. (2002). Bayesian methods for Hidden Markov Models: Recursive computing in the 21st century. *Journal of the American Statistical Association, 97*(457), 337–351. doi:10.1198/016214502753479464.

Titterington, D.M. (1984). Comments on "Application of the conditional population-mixture model to image segmentation". Pattern Analysis and Machine Intelligence, *IEEE Transactions on PAMI, 6*(5), 656–658. doi:10.1109/TPAMI.1984.4767581.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory IEEE Transactions on*, *13*(2), 260–269. doi:10.1109/TIT.1967.1054010.

Zhou, Y., Johansen, A.M., Aston, J.A.D. (2013). Towards automatic model comparison: an adaptive sequential Monte Carlo approach. *Mathematics e-print* 1303.3123, ArXiv.