Stat Comput (2008) 18: 47–57
DOI 10.1007/s11222-007-9037-8

Particle methods for maximum likelihood estimation in latent variable models

Adam M. Johansen · Arnaud Doucet · Manuel Davy

Received: 15 May 2007 / Accepted: 31 August 2007 / Published online: 28 September 2007

Abstract Standard methods for maximum likelihood parameter estimation in latent variable models rely on the Expectation-Maximization algorithm and its Monte Carlo variants. Our approach is different and motivated by similar considerations to simulated annealing; that is we build a sequence of artificial distributions whose support concentrates itself on the set of maximum likelihood estimates. We sample from these distributions using a sequential Monte Carlo approach. We demonstrate state-of-the-art performance for several applications of the proposed approach.

Keywords Latent variable models · Markov chain Monte Carlo · Maximum likelihood · Sequential Monte Carlo · Simulated annealing

1 Introduction

Performing Maximum Likelihood (ML) parameter estimation in latent variable models is a complex task. First, in many cases, the likelihood for the parameters of interest does not admit a closed-form expression. Second, even when it does, it can be multimodal. When the likelihood can be evaluated, the classical approach to problems of this sort is the Expectation-Maximisation (EM) algorithm (Dempster et al. 1977), which is a numerically well-behaved algorithm. However the EM algorithm is a deterministic algorithm, which is sensitive to initialization and can become trapped in severe local maxima. To avoid getting trapped in local maxima and to deal with cases where the E-step cannot be performed in closed-form, some Monte Carlo variants of the EM algorithm have been proposed.

More recently, an algorithm has been proposed to solve, simultaneously, this joint integration/maximization problem; see Doucet et al. (2002) or Gaetan and Yao (2003), Jacquier et al. (2007) for an independent derivation. The main idea of this algorithm is related to Simulated Annealing (SA) and consists of building a sequence of artificial distributions whose support concentrates itself on the set of ML estimates. In cases where the likelihood does not admit a closed-form expression, these artificial distributions are not standard and rely on the introduction of an increasing number of artificial copies of the latent variables. To sample from this sequence of distributions, the authors of (Doucet et al. 2002) use non-homogeneous Markov chain Monte Carlo (MCMC) algorithms which they term State Augmentation for Marginal Estimation (SAME). Although these iterative stochastic algorithms typically perform better than deterministic EM and its variants (Robert and Casella 2004, Chap. 5), they can also get stuck in severe local maxima. We propose, here, original Sequential Monte Carlo (SMC) methods to address this problem. In this approach, the distributions are approximated by a large cloud of interacting random samples. The performance of these methods is much less sensitive to initialization than EM and MCMC algorithms. We demonstrate their efficiency on a variety of problems.

A.M. Johansen (✉)
Department of Mathematics, University Walk,
University of Bristol, Bristol, BS8 1TW, UK
e-mail: Adam.Johansen@bristol.ac.uk

A. Doucet
Department of Statistics & Department of Computer Science,
University of British Columbia, Vancouver, Canada
e-mail: Arnaud@cs.ubc.ca

M. Davy
LAGIS UMR 8146, BP 48, Cité scientifique,
59651 Villeneuve d'Ascq Cedex, France
e-mail: Manuel.Davy@ec-lille.fr

The remainder of the paper is organized as follows. In Sect. 2, we formally introduce the statistical model and a sequence of artificial probability distributions which concentrates itself on the set of ML estimates. In Sect. 3, we describe two generic SMC algorithms to sample from these distributions: the first algorithm assumes the likelihood is known pointwise whereas the second algorithm considers the most general case. Finally in Sect. 4, we provide a number of example applications.

2 Maximum likelihood estimation in latent variable models

Let $y \in \mathcal{Y}$ denote the observed data, $z \in \mathcal{Z}$ the latent variables and $\theta \in \Theta$ the parameter vector of interest. The marginal likelihood of θ is given by

$$p(y|\theta) = \int p(y, z|\theta) dz, \tag{1}$$

where $p(y, z|\theta)$ is the complete likelihood. The complete likelihood is known pointwise but the marginal likelihood might not be tractable. We are interested in the set of ML estimates

$$\Theta_{\mathrm{ML}} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, p(y|\theta). \tag{2}$$

Instead of an EM approach to maximize $p(y|\theta)$, we propose an alternative related to SA. Let $p(\theta)$ be an instrumental prior distribution whose support includes the maximisers of the likelihood function, then the probability distribution

$$\overline{\pi}_\gamma(\theta) \propto p(\theta) p(y|\theta)^\gamma \tag{3}$$

concentrates itself on the set of ML estimates as $\gamma \to \infty$ under weak assumptions. Indeed, asymptotically the contribution from this instrumental prior vanishes; this term is only present to ensure that the distribution $\overline{\pi}_\gamma(\theta)$ is a proper distribution—it may be omitted in those instances in which this is already the case. If we could obtain samples from a distribution $\overline{\pi}_\gamma(\theta)$ where γ is large, then the simulated samples would be concentrated around Θ_{ML}. However, Monte Carlo methods such as MCMC and SMC require that it is possible to evaluate the distributions of interest up to a normalizing constant: such methodology cannot be applied directly if $p(y|\theta)$ does not admit a closed-form expression.

2.1 Algorithms

To circumvent this problem, it has been proposed in Doucet et al. (2002), in a Maximum a Posteriori (MAP) rather than ML setting, to build an artificial distribution known up to a normalizing constant which admits as a marginal distribution the target distribution $\overline{\pi}_\gamma(\theta)$ for an integer power γ greater than one. A similar scheme was subsequently proposed by Gaetan and Yao (2003), Jacquier et al. (2007) in the ML setting. We note that closely related approaches have also appeared in the literature to perform full ML estimation (in the absence of latent variables) (Robert and Titterington 1998) and in an optimal design context (Müller et al. 2004; Amzal et al. 2006). The basic idea consists of introducing γ artificial replicates of the missing data and defining

$$\pi_\gamma(\theta, z_{1:\gamma}) \propto p(\theta) \prod_{i=1}^{\gamma} p(y, z_i|\theta), \tag{4}$$

with $z_{i:j} = (z_i, \ldots, z_j)$. Indeed it is easy to check that the marginal in θ of (4) denoted $\pi_\gamma(\theta)$ is equal to (3). Note that it is straightforward to modify the distribution $\pi_\gamma(\theta, z_{1:\gamma})$ so that it concentrates itself on the set of the Maximum A Posteriori (MAP) estimates of θ associated with the prior $p(\theta)$ and the likelihood $p(y|\theta)$ by using a different sequence of distributions

$$\pi_\gamma(\theta, z_{1:\gamma}) \propto \prod_{i=1}^{\gamma} p(\theta) p(y, z_i|\theta). \tag{5}$$

As it is usually impossible to sample from $\pi_\gamma(\theta, z_{1:\gamma})$ directly, MCMC algorithms have been proposed in the literature to achieve this. However, using an MCMC kernel to sample directly from this distribution for a large integer γ can be very inefficient as, by construction, the marginal distribution $\pi_\gamma(\theta)$ is sharply peaked and the mixing properties of MCMC kernels usually deteriorate as γ increases. Such approaches can perform rather well if the likelihood is unimodal but are likely to fail if it is multimodal. A popular approach, which alleviates this problem to some degree, is adopted in the SAME algorithm. It consists of sampling from a sequence of distributions $\{\pi_{\gamma_t}(\theta, z_{1:\gamma_t})\}_{t \geq 1}$ evolving over time, t, such that γ_1 is small enough for $\pi_{\gamma_1}(\theta, z_{1:\gamma_1})$ to be easy to sample from and $\{\gamma_t\}_{t \geq 1}$ is an increasing sequence going to infinity. However, in practice, this approach suffers from two major drawbacks. First, in contrast to standard SA, we are restricted to integer inverse temperatures, $\{\gamma_t\}_{t \geq 1}$. Hence the discrepancy between successive target distributions can be high and this limits the performance of the algorithm. Second, a very slow (logarithmic) annealing schedule is necessary to ensure convergence towards Θ_{ML}. In practice, a faster (linear or geometric) annealing schedule is used, but, consequently, the MCMC chain tends to become trapped in local modes.

To solve the first problem, we introduce for any real-valued $\gamma > 0$ the target distribution

$$\pi_\gamma(\theta, z_{1:\lceil\gamma\rceil}) \propto p(\theta) p(y, z_{\lceil\gamma\rceil}|\theta)^{\gamma^\sharp} \prod_{i=1}^{\lfloor\gamma\rfloor} p(y, z_i|\theta), \tag{6}$$

where $\lfloor\gamma\rfloor \triangleq \sup\{\alpha \in \mathbb{Z} : \alpha \leq \gamma\}$, $\lceil\gamma\rceil \triangleq \inf\{\alpha \in \mathbb{Z} : \alpha \geq \gamma\}$ and $\gamma^\sharp \triangleq \gamma - \lfloor\gamma\rfloor$. Distribution (6) coincides with (4) for any integer γ; for general γ, the marginal $\pi_\gamma(\theta)$ of (6) is not equal to $\overline{\pi}_\gamma(\theta)$ but still concentrates itself on Θ_{ML} as $\gamma \to \infty$.

To solve the second problem, we propose to employ SMC methods. The sequence of distributions is approximated by a collection of random samples termed particles which evolve over time using sampling and resampling mechanisms. The population of samples employed by our method makes it much less prone to trapping in local maxima.

3 SMC sampler algorithms

SMC methods have been used primarily to solve optimal filtering problems; see, for example, Doucet et al. (2001) for a review of the literature. They are used here in a completely different framework, that proposed by Del Moral et al. (2006). This framework involves the construction of a sequence of artificial distributions which admit the distributions of interest (in our case those of the form of (4)) as particular marginals.

SMC samplers allow us to obtain, iteratively, collections of weighted samples from a sequence of distributions $(\pi_t(x_t))_{t\geq 1}$. These distributions may be defined over essentially any random variables X_t on some measurable spaces $(E_t, \mathcal{E}_t)$. Such sampling is facilitated by the construction of a sequence of auxiliary distributions $(\tilde{\pi}_t)_{t\geq 1}$ on spaces of increasing dimension, $\tilde{\pi}_t(x_{1:t}) = \pi_t(x_t)\prod_{s=1}^{t-1} L_s(x_{s+1}, x_s)$, by defining a sequence of Markov kernels $\{L_s\}_{s\geq 1}$ which operate, in some sense backwards in time as, conditional upon a point x_{s+1} in E_{s+1}, L_s provides a probability distribution over E_s. This sequence is formally arbitrary but critically influences the estimator variance. In the present application we are concerned with distributions over the collections of random variables $X_t = (\theta_t, Z_{t,1:\lceil\gamma_t\rceil})$. See Del Moral et al. (2006) for further details and guidance on the selection of these kernels. Standard SMC techniques can then be applied to the sequence of synthetic distributions $\{\tilde{\pi}_t\}_{t\geq 1}$.

We distinguish here two cases: that in which the likelihood $p(y|\theta)$ is known analytically and the general case in which it is not.

3.1 Marginal likelihood available

It is interesting to consider an analytically convenient special case, which leads to Algorithm 3.1. This algorithm is applicable when we are able to sample from particular conditional distributions, and evaluate the marginal likelihood pointwise.

We note that the details of this algorithm can be understood by viewing it as a refinement of a particular case of the general algorithm proposed below. However, we present it first as it is relatively simple to interpret and provides some insight into the approach which we would ideally like to adopt. Intuitively, one can view this algorithm as applying an importance weighting to correct for the distributional mismatch between $\pi_{\gamma_{t-1}}$ and π_{γ_t} and updating $Z^{(i)}_{t,1:\lceil\gamma_t\rceil}$ then θ at each step by applying Gibbs sampler moves which are π_{γ_t} invariant.

Although the applicability of the general algorithm to a much greater class of problems is potentially more interesting we remark that the introduction of a latent variable structure can lead to kernels which mix much more rapidly than those used in a direct approach (Robert and Casella 2004, p. 351). Here and throughout, we write $z_t = z_{1:\lceil\gamma_t\rceil}$ and $Z^{(i)}_t = Z^{(i)}_{t,1:\lceil\gamma_t\rceil}$ to denote the collection of

Algorithm 3.1 SMC MML with Marginal Likelihoods

Initialisation: $t = 1$:

Sample $\{\theta_1^{(i)}\}_{i=1}^N$ independently from some importance distribution, $\nu(\cdot)$.

Calculate importance weights $W_1^{(i)} \propto \frac{\pi_{\gamma_1}(\theta_1^{(i)})}{\nu(\theta_1^{(i)})}$, $\sum_{i=1}^N W_1^{(i)} = 1$.

for $t = 2$ to T **do**

Calculate importance weights:

$$W_t^{(i)} \propto W_{t-1}^{(i)} p(y|\theta_{t-1}^{(i)})^{\gamma_t - \gamma_{t-1}}, \quad \sum_{i=1}^N W_t^{(i)} = 1.$$

if Effective Sample Size (ESS, see Liu and Chen 1998) < Threshold, **then**

Resample from $\{W_t^{(i)}, \theta_{t-1}^{(i)}\}_{i=1}^N$.

end if

Sample $\{(\theta_t^{(i)}, Z_t^{(i)})\}_{i=1}^N$ such that:

$$Z_t^{(i)} \sim \pi_{\gamma_t}(\cdot|\theta_{t-1}^{(i)}) \text{ and } \theta_t^{(i)} \sim \pi_{\gamma_t}(\cdot|Z_t^{(i)}).$$

end for

Algorithm 3.2 A general SMC algorithm for MML estimation

Initialisation: $t = 1$:

Sample $\{(\theta_1^{(i)}, Z_1^{(i)})\}_{i=1}^N$ independently from some importance distribution, $\nu(\cdot)$.

Calculate importance weights $W_1^{(i)} \propto \frac{\pi_{\gamma_1}(\theta_1^{(i)}, Z_1^{(i)})}{\nu(\theta_1^{(i)}, Z_1^{(i)})}$, $\sum_{i=1}^N W_1^{(i)} = 1$.

for $t = 2$ to T **do**

  if ESS $<$ Threshold, **then**

    Resample from $\{W_{t-1}^{(i)}, (\theta_{t-1}^{(i)}, Z_{t-1}^{(i)})\}_{i=1}^N$.

  end if

  Sample $\{(\theta_t^{(i)}, Z_t^{(i)})\}_{i=1}^N$ such that $(\theta_t^{(i)}, Z_t^{(i)}) \sim K_t((\theta_{t-1}^{(i)}, Z_{t-1}^{(i)}), \cdot)$.

  Set importance weights,

$$\frac{W_t^{(i)}}{W_{t-1}^{(i)}} \propto \frac{\pi_{\gamma_t}(\theta_t^{(i)}, Z_t^{(i)}) L_{t-1}((\theta_t^{(i)}, Z_t^{(i)}), (\theta_{t-1}^{(i)}, Z_{t-1}^{(i)}))}{\pi_{\gamma_{t-1}}(\theta_{t-1}^{(i)}, Z_{t-1}^{(i)}) K_t((\theta_{t-1}^{(i)}, Z_{t-1}^{(i)}), (\theta_t^{(i)}, Z_t^{(i)}))}.$$

end for

replicates of latent variables associated with the ith particle at time t.

When the marginal likelihood is known, it is unnecessary to introduce a sequence of distributions $\pi_{\gamma_t}(\theta, z_t)$. It can be seen that, as $\gamma \to \infty$, this algorithm resembles a stochastic variant of EM and, indeed, it would be computationally more efficient to switch from this algorithm to conventional EM updates after some number of iterations (by employing this approach initially, one hopes to alleviate some of the difficulties caused by the presence of local optima).

3.2 Marginal likelihood unavailable

Algorithm 3.2 introduces the general framework which we propose. We then show that Algorithm 3.1 is an important special case within this framework. Finally we present a generic form of the algorithm which can be applied to a broad class of problems, although it will often be less efficient to use this generic formulation than to construct a dedicated sampler for a particular class of problems.

Algorithm 3.1 is a particular case of this algorithm where we move the particles according to a π_{γ_t}-invariant Markov kernel given by

$$K_t((\theta_{t-1}, z_{t-1}), (\theta_t, z_t)) = \pi_{\gamma_t}(z_t|\theta_{t-1})\pi_{\gamma_t}(\theta_t|z_t)$$

and

$$L_{t-1}((\theta_t, z_t), (\theta_{t-1}, z_{t-1})) = \pi_{\gamma_t}(\theta_{t-1}|z_t)\pi_{\gamma_{t-1}}(z_{t-1}|\theta_{t-1})$$

leading to the weight expression shown in the algorithm. As the importance weight depends only upon θ_{t-1}, resampling can be carried out *before*, rather than *after* the sampling step.

In order to understand this choice of kernel, it is useful to consider an alternative interpretation of the algorithm which targets the marginal distribution (3) directly, employing Z_t as an auxiliary variable in order to sample from the proposal kernel

$$K_t(\theta_{t-1}, \theta_t) = \int \pi_{\gamma_t}(z_t|\theta_{t-1})\pi_{\gamma_t}(\theta_t|z_t)dz_t$$

which is clearly π_{γ_t}-invariant. In this case, using the time reversal kernel as its auxiliary counterpart:

$$L_{t-1}(\theta_t, \theta_{t-1}) = \frac{\pi_{\gamma_t}(\theta_{t-1})K_t(\theta_{t-1}, \theta_t)}{\pi_{\gamma_t}(\theta_t)},$$

leads to the weight expression shown in the algorithm. This is a well known approximation to the optimal auxiliary kernel (Del Moral et al. 2006).

To obtain an algorithm which may be applied to a wider range of scenarios, we can select $(\mathcal{K}_t)_{t\geq 1}$ as a collection of Markov kernels with invariant distributions corresponding to $(\pi_{\gamma_t})_{t\geq 1}$. We then employ, in Algorithm 3.2, proposal kernels of the form:

$$K_t((\theta_{t-1}, z_{t-1}), (\theta_t, z_t)) = \begin{cases} \mathcal{K}_{t-1}((\theta_{t-1}, z_{t-1}), (\theta_t, z_t)) & \text{if } \lceil \gamma_{t-1} \rceil = \lceil \gamma_t \rceil, \\ \mathcal{K}_{t-1}((\theta_{t-1}, z_{t-1}), (\theta_t, z_{t,1:\lceil \gamma_{t-1} \rceil})) & \\ \quad \times q_{\gamma_t^\sharp}(z_{t,\lceil \gamma_t \rceil}|\theta_t) \prod_{j=\lceil \gamma_{t-1} \rceil + 1}^{\lceil \gamma_t \rceil - 1} q(z_{t,j}|\theta_t) & \\ \text{otherwise,} & \end{cases}$$

where it is understood that $q_0(\cdot|\theta) = 1$, and select auxiliary kernels

$$L_{t-1}((\theta_t, Z_t), (\theta_{t-1}, Z_{t-1})) = \frac{\pi_{\gamma_{t-1}}(\theta_{t-1}, Z_{t-1})\mathcal{K}_{t-1}((\theta_{t-1}, Z_{t-1}), (\theta_t, Z_{t,1:\lceil \gamma_{t-1} \rceil}))}{\pi_{\gamma_{t-1}}((\theta_t, Z_{t,1:\lceil \gamma_{t-1} \rceil}))}.$$

As kernel selection is of critical importance to the performance of SMC algorithms, a few comments on these choices

Algorithm 3.3 A generic SMC algorithm for MML estimation

Initialisation: $t = 1$:

Sample $\{(\theta_1^{(i)}, Z_1^{(i)})\}_{i=1}^N$ independently from some importance distribution, $\nu(\cdot)$.

Calculate importance weights $W_1^{(i)} \propto \frac{\pi_{\gamma_1}(\theta_1^{(i)}, Z_1^{(i)})}{\nu(\theta_1^{(i)}, Z_1^{(i)})}$.

for $t = 2$ to T **do**

if ESS $<$ Threshold, **then**

Resample from $\{W_{t-1}^{(i)}, (\theta_{t-1}^{(i)}, Z_{t-1}^{(i)})\}_{i=1}^N$.

end if

Sample $\{(\theta_t^{(i)}, Z_t^{(i)})\}_{i=1}^N$ such that $(\theta_t^{(i)}, Z_{t,1:\lceil\gamma_{t-1}\rceil}^{(i)}) \sim \mathcal{K}_{t-1}(\theta_{t-1}^{(i)}, Z_{t-1}^{(i)}; \cdot)$,

and if $\lceil\gamma_t\rceil > \lceil\gamma_{t-1}\rceil$, then for $j = \lceil\gamma_{t-1}\rceil + 1$ to $\lfloor\gamma_t\rfloor$, $Z_{t,j}^{(i)} \sim q(\cdot|\theta_t^{(i)})$ and, if

$\gamma_t^\sharp \neq 0$, $Z_{t,\lceil\gamma_t\rceil}^{(i)} \sim q_{\gamma_t^\sharp}(\cdot|\theta_t^{(i)})$.

Set importance weights, when $\lceil\gamma_t\rceil = \lceil\gamma_{t-1}\rceil$,

$$W_t^{(i)} / W_{t-1}^{(i)} \propto p(y, Z_{t,\lceil\gamma_t\rceil}^{(i)}|\theta_t^{(i)})^{\gamma_t^\sharp - \gamma_{t-1}^\sharp},$$

otherwise, we have that (note that the final term vanishes when $\gamma_t^\sharp = 0$):

$$\frac{W_t^{(i)}}{W_{t-1}^{(i)}} \propto \frac{p(y, Z_{t,\lceil\gamma_{t-1}\rceil}^{(i)}|\theta_t^{(i)})}{p(y, Z_{t,\lceil\gamma_{t-1}\rceil}^{(i)}|\theta_t^{(i)})^{\gamma_t^\sharp}} \left[\prod_{j=\lceil\gamma_t\rceil+1}^{\lfloor\gamma_t\rfloor} \frac{p(y, Z_{t,j}^{(i)}|\theta_t^{(i)})}{q(Z_{t,j}^{(i)}|\theta_t^{(i)})}\right] \frac{p(y, Z_{t,\lceil\gamma_t\rceil}^{(i)}|\theta_t^{(i)})^{\gamma_t^\sharp}}{q_{\gamma_t^\sharp}(Z_{t,\lceil\gamma_t\rceil}^{(i)}|\theta_t^{(i)})}.$$

end for

are justified. The proposal kernel K_t can be interpreted as the composition of two components: the parameter value and existing latent variable replicates are moved according to a Markov kernel, and any new replicates of the latent variables are obtained from some proposal distribution q. The auxiliary kernel, L_{t-1} which we propose corresponds, to the composition of the time reversal kernel associated with $\mathcal{K}_{t-1}$, and the optimal auxiliary kernel associated with the other component of the proposal.

In this case, as summarised in Algorithm 3.3, we also assume that good importance distributions, $q(\cdot|\theta)$, for the conditional probability of the variables being marginalised can be sampled from and evaluated. If the annealing schedule is to include non-integer inverse temperatures, then we further assumed that we have appropriate importance distributions for distributions proportional to $p(z|\theta, y)^\alpha, \alpha \in (0, 1)$, which we denote $q_\alpha(z|\theta)$. This is not the most general possible approach, but is one which should work acceptably for a broad class of problems.

3.3 General comments

Superficially, these algorithms appear very close to mutation-selection schemes employed in the genetic algorithms literature. However, there are two major differences: First, such methods require the function being maximized to be known pointwise, whereas the proposed algorithms do not. Second, convergence results for the SMC methods follow straightforwardly from general results on Feynman-Kac flows (Del Moral 2004).

There are a number of possible estimators associated with these algorithms. In those cases in which the marginal likelihood can be evaluated cheaply, the most obvious technique is monitoring the marginal posterior of every parameter combination which is sampled and using that set of parameters associated with the largest value seen. The only obvious advantage of this method over other approaches might be robustness in particularly complex models. We note that informal experiments revealed very little difference in the performance of this approach and the more generally applicable approach proposed below when both could be used. When the marginal likelihood cannot readily be evaluated, we recommend that the estimate is taken to be the first moment of the empirical distribution induced by the final particle ensemble; this may be justified by the asymptotic (in the inverse temperature) normality of the target distribution (see, for example, Robert and Casella 2004, p. 203) (although there may be some difficulties in the case of non-identifiable models for which more sophisticated techniques would be required).

Under weak regularity assumptions (Hwang 1980), it is possible to demonstrate that the sequence of distributions which we employ concentrates itself upon the set Θ_{ML}. Under additional regularity assumptions, the estimates obtained from the particle system converge to those which would be obtained by performing integrals under the distributions themselves—and obey a central limit theorem. The variance of this central limit theorem can be quantitatively bounded under strong regularity assumptions. All of this follows by a

rather straightforward generalisation of the results in Chopin (2004), Del Moral (2004); details are provided in Johansen (2006, Sect. 4.2.2).

4 Applications

We now show comparative results for a simple toy example and two more challenging models. We begin with a one dimensional example in Sect. 4.1, followed by a Gaussian mixture model in Sect. 4.2 and a non-linear non-Gaussian state space model which is widely used in financial modelling in Sect. 4.3.

For the purpose of comparing algorithms on an equal footing, it is necessary to employ some measure of computational complexity. We note that almost all of the computational cost associated with all of the algorithms considered here comes from either sampling the latent variables or determining their expectation. We introduce the quantity χ defined as the total number of complete replicates of the latent variable vector which needs to be simulated (in the case of the SMC and SAME algorithms) or estimated (as in the case of EM) in one complete run of an algorithm. Note that in the case of SAME and the SMC algorithm, this figure depends upon the annealing schedule in addition to the final temperature and the number of particles in the SMC case.

In those examples in which the marginal likelihood can be evaluated analytically, we present for each algorithm a collection of summary statistics obtained from fifty runs. These describe the variation of the likelihood of the estimated parameter values. We remark that, although it is common practice to employ multiple, differently replicated initialisations of many algorithms, which would suggest that the highest likelihood obtained by any run might be the important figure of merit other factors must also be considered. In many of the more complex situations in which we envisage this algorithm being useful, the likelihood cannot be evaluated and we will not have the luxury of employing this approach. The mean, variance and range of likelihood estimates in the simpler examples allow us to gauge the consistency and robustness of the various algorithms which are employed.

The following notation is used to describe various probability distributions: $\mathcal{D}i(\alpha)$ the Dirichlet distribution with parameter vector α, $\mathcal{N}(\mu, \sigma^2)$ describes a normal of mean μ and variance σ^2, $\mathcal{G}a(\alpha, \beta)$ a gamma distribution of shape α and rate β, and $\mathcal{IG}(\alpha, \beta)$ the inverse gamma distribution associated with $\mathcal{G}a(\alpha, \beta)$.

4.1 Toy example

We consider first a toy example in one dimension for which we borrow Example 1 of Gaetan and Yao (2003). The model consists of a Student t-distribution of unknown location parameter θ with 0.05 degrees of freedom. Four observations are available, $y = (-20, 1, 2, 3)$. The logarithm of the marginal likelihood in this instance is given by:

$$\log p(y|\theta) = -0.525 \sum_{i=1}^{4} \log(0.05 + (y_i - \theta)^2),$$

which is not susceptible to analytic maximisation. However, the global maximum is known to be located at 1.997, and local maxima exist at $\{-19.993, 1.086, 2.906\}$ as illustrated in Fig. 1. We can complete this model by considering the Student t-distribution as a scale-mixture of Gaussians and associating a gamma-distributed latent precision parameter Z_i with each observation. The log likelihood is then:

$$\log p(y, z|\theta) = -\sum_{i=1}^{4} [0.475 \log z_i + 0.025 z_i + 0.5 z_i (y_i - \theta)^2].$$

In the interest of simplicity, we make use of a linear temperature scale, $\gamma_t = t$, which takes only integer values. We are able to evaluate the marginal likelihood function pointwise, and can sample from the conditional distributions:

$$\pi_t(z_{1:t}|\theta, y) = \prod_{i=1}^{t} \prod_{j=1}^{4} \mathcal{G}a\left(z_{i,j}|0.525, 0.025 + \frac{(y_j - \theta)^2}{2}\right), \tag{7}$$

$$\pi_t(\theta|z_{1:t}) = \mathcal{N}(\theta|\mu_t^{(\theta)}, \Sigma_t^{(\theta)}), \tag{8}$$

where the parameters,

$$\Sigma_t^{(\theta)} = \left[\sum_{i=1}^{t} \sum_{j=1}^{4} z_{i,j}\right]^{-1} = \left[1/\Sigma_{t-1}^{(\theta)} + \sum_{j=1}^{4} z_{t,j}\right]^{-1}, \tag{9}$$

$$\mu_t^{(\theta)} = \Sigma_t^{(\theta)} \sum_{i=1}^{t} y^T z_i = \Sigma_t^{(\theta)} (\mu_{t-1}^{(\theta)} / \Sigma_{t-1}^{(\theta)} + y^T z_t), \tag{10}$$

may be obtained recursively. Consequently, we can make use of Algorithm 3.1 to solve this problem. We use an instrumental uniform $[-50, 50]$ prior distribution over θ. Some simulation results are given in Table 1. The estimate is taken to be the first moment of the empirical distribution induced by the final particle ensemble.

This simple example confirms that the algorithm proposed above is able to locate the global optimum, at least in the case of extremely simple distributions. It also illustrates that it is reasonably robust to the selection of the number of particles and intermediate distributions. Generally, increasing the total amount of computation leads to very slightly more accurate localisation of the mode. Only a single simulation failed to find the global optimum—one of those with $N = 20$ and $T = 30$.

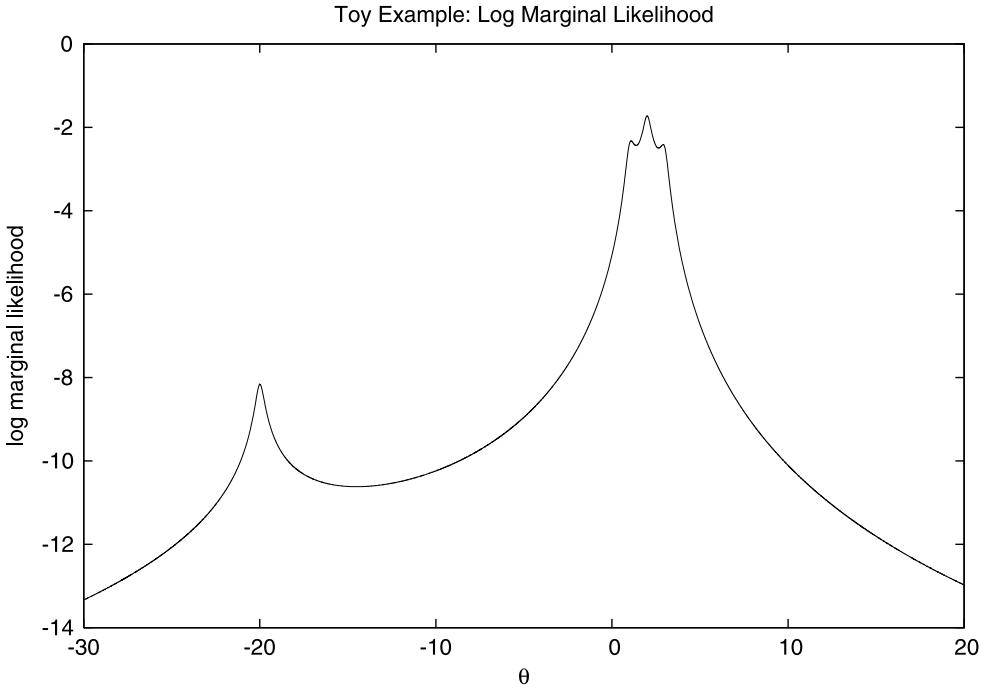

Fig. 1 The log marginal likelihood of the toy example of Sect. 4.1

Table 1 Simulation results for the toy problem. Each line summarises 50 simulations with N particles and final inverse temperature T. Only one simulation failed to find the correct mode

N	T	Mean	Std. Dev.	Min	Max
50	15	1.992	0.014	1.952	2.033
100	15	1.997	0.013	1.973	2.038
20	30	1.958	0.177	1.094	2.038
50	30	1.997	0.008	1.983	2.011
100	30	1.997	0.007	1.983	2.011
20	60	1.998	0.015	1.911	2.022
50	60	1.997	0.005	1.988	2.008

4.2 A finite Gaussian mixture model

To allow comparison with other techniques, and to illustrate the strength of the method proposed here in avoiding local maxima, we consider a finite Gaussian mixture model. A set of observations $\{y_i\}_{i=1}^P$ is assumed to consist of P i.i.d. samples from a distribution of the form:

$$Y_i \sim \sum_{s=1}^{S} \omega_s \mathcal{N}(\mu_s, \sigma_s^2), \tag{11}$$

where $0 < \omega_s < 1$; $\sum_{s=1}^S \omega_s = 1$ are the weights of each mixture component and $\{\mu_s, \sigma_s^2\}_{s=1}^S$ is the set of their means and variances. As is usual with such mixtures, it is convenient to introduce auxiliary allocation variables, Z_i which allow us to assign each observation to one of the mixture components, then we may write the distribution in the form:

$$Y_i|(\{\omega, \mu_s, \sigma_s^2\}, Z_i = z_i) \sim \mathcal{N}(\mu_{z_i}, \sigma_{z_i}^2),$$
$$p(Z_i = z_i) = \omega_{z_i}.$$

It is both well known and somewhat obvious, that the maximum likelihood estimate of all parameters of this model is not well defined as the likelihood is not bounded. However, the inclusion of prior distributions over the parameters has a bounding effect and makes MAP estimation possible (Robert and Titterington 1998). We consequently show the results of all algorithms adapted for MAP estimation by inclusion of diffuse proper priors (see, for example Robert and Casella 2004, p. 365), which are as follows:

$$\omega \sim \mathcal{D}i(\delta),$$
$$\sigma_i^2 \sim \mathcal{IG}\left(\frac{\lambda_i + 3}{2}, \frac{\beta_i}{2}\right),$$
$$\mu_i|\sigma_i^2 \sim \mathcal{N}(\alpha_i, \sigma_i^2/\lambda_i),$$

with δ, λ_i and β_i are hyperparameters, whose values are given below.

It is straightforward to adjust our Algorithm 3.1 to deal with the MAP, rather than ML case. For this application it is possible to sample from all of the necessary distributions, and to evaluate the marginal posterior pointwise and so we employ such an algorithm.

At iteration t of the algorithm, for each particle we sample the parameter estimates, conditioned upon the previous values of the latent variables according to the conditional distributions:

$$\omega \sim \mathcal{D}i(\gamma_t(\delta-1)+1+n(\lfloor\gamma_t\rfloor)+\gamma_t{}^\sharp\Delta n(\lceil\gamma_t\rceil)),$$

$$\sigma_i^2 \sim \mathcal{IG}(A_i, B_i),$$

$$\mu_i|\sigma_i^2 \sim \mathcal{N}\left(\frac{\gamma_t\lambda_i\alpha_i+\overline{y}(\lfloor\gamma_t\rfloor)_i+\gamma_t{}^\sharp\Delta\overline{y}(\lceil\gamma_t\rceil)_i}{\gamma_t\lambda_i+n(\lfloor\gamma_t\rfloor)_i+\gamma_t{}^\sharp\Delta n(\lceil\gamma_t\rceil)_i}, \frac{\sigma_i^2}{\gamma_t\lambda_i+n(\lfloor\gamma_t\rfloor)_i\gamma_t{}^\sharp\Delta n(\lceil\gamma_t\rceil)_i}\right),$$

where we have defined the following quantities for convenience:

$$n(i)_j=\sum_{l=1}^{i}\sum_{p=1}^{P}\mathbb{I}_j(Z_{l,p}),$$

$$\Delta n(i)_j=n(i)_j-n(i-1)_j,$$

$$\overline{y}(i)_j=\sum_{l=1}^{i}\sum_{p=1}^{P}\mathbb{I}_j(Z_{l,p})y_j,$$

$$\Delta\overline{y}(i)_j=\overline{y}(i)_j-\overline{y}((i-1))_j,$$

$$\overline{y^2}(i)_j=\sum_{l=1}^{i}\sum_{p=1}^{P}\mathbb{I}_j(Z_{l,p})y_j^2,$$

$$\Delta\overline{y^2}(i)_j=\overline{y^2}(i)_j-\overline{y^2}(i-1)_j,$$

and the parameters for the inverse gamma distribution from which the variances are sampled from are:

$$A_i=\frac{\gamma_t(\lambda_i+1)+n(\lfloor\gamma_t\rfloor)_i+\gamma_t{}^\sharp\Delta n(\lceil\gamma_t\rceil)_i}{2}+1,$$

$$B_i=\frac{1}{2}\left(\gamma_t(\beta_i+\lambda_i\alpha_i^2)+\overline{y^2}(\lfloor\gamma_t\rfloor)_i+\gamma_t{}^\sharp\Delta\overline{y^2}(\lceil\gamma_t\rceil)_i-\sum_{g=1}^{\lfloor\gamma_t\rfloor}\frac{(\Delta\overline{y}(g)_i+\lambda_i\alpha_i)^2}{\lambda_i+\Delta n(g)_i}-\gamma_t{}^\sharp\frac{(\Delta\overline{y}(\lceil\gamma_t\rceil)_i+\lambda_i\alpha_i)^2}{\lambda_i+\Delta n(\lceil\gamma_t\rceil)_i}\right).$$

Then we sample all of the allocation variables from the appropriate distributions, noting that this is equivalent to augmenting them with the new values and applying an MCMC move to those persisting from earlier iteration.

As a final remark, we note that it would be possible to use the proposed framework to infer the number of mixture components, as well as their parameters—by employing Dirichlet process mixtures, for example.

4.2.1 Simulated data

We present results first from data simulated according to the model. 100 data were simulated from a distribution of the form of (11), with parameters $\omega=[0.2,0.3,0.5]$, $\mu=[0,2,3]$ and $\sigma^2=[1,\frac{1}{4},\frac{1}{16}]$. The same simulated data set was used for all runs, and the log posterior density of the generating parameters was -155.87. Results for the SMC algorithm are shown in Table 2 and for the other algorithms in Table 3—two different initialisation strategies were used for these algorithms, that described as "Prior" in which a parameter set was sampled from the prior distributions, and "Hull" in which the variances were set to unity, the mixture weights to one third and the means were sampled uniformly from the convex hull of the observations.

Two annealing schedules were used for the SAME algorithm: one involved keeping the number of replicates of the augmentation data fixed to one for the first half of the iterations and then increasing linearly to a final maximum value of 6; the other keeping it fixed to one for the first 250 iterations, and then increasing linearly to 50. The annealing schedule for the SMC algorithm was of the form $\gamma_t=Ae^{bt}$ for suitable constants to make $\gamma_1=0.01$ and $\gamma_T=6$. This is motivated by the intuition that when γ is small, the effect of increasing it by some amount $\Delta\gamma$ is to change its form somewhat more than would be the case for a substantially larger value of γ. No substantial changes were found for values of γ greater than 6, presumably due to the sharply peaked nature of the distribution. Varying the forms of the annealing schedules did not appear to substantially affect the results. Hyperparameter values were shared across all simulations, with $\delta=1$, $\lambda_i=0.1$, $\beta_i=0.1$ and $\alpha_i=0$.

Several points are noticeable from these results:

- The SMC algorithm produce estimates whose posterior density *uniformly* exceeded that of the generating parameters (and the SAME algorithm frequently produced such estimates). Whilst this provides no guarantee that the global optimum has been located it does provide some encouragement that the parameter estimates being obtained are sensible.
- For a given computational cost, the SMC algorithm outperformed SAME in the sense that both the mean and maximum posterior is substantially increased.
- Whilst, as is well documented, the EM algorithm can perform well if favourably initialised, neither of the initialisation strategies which we employed led to a large number of good performances. Furthermore, it can be seen that taking the best result from 50 runs of the EM algorithm lead to poorer performance than a single run of the SMC algorithm with a lower cost:

Table 2 Summary of the final log posterior estimated by 50 runs of the SMC Algorithm on simulated data from a finite Gaussian mixture with varying numbers of particles, N, and intermediate distributions, T

N	T	χ	Mean	Std. Dev.	Min	Max
25	25	1325	−154.39	0.55	−155.76	−153.64
25	50	2125	−153.88	0.13	−154.18	−153.59
50	50	4250	−153.80	0.08	−153.93	−153.64
100	50	8500	−153.74	0.07	−153.91	−153.59
250	50	21250	−153.70	0.07	−153.90	−153.54
1000	50	85000	−153.64	0.04	−153.71	−153.57
100	100	20300	−153.73	0.08	−153.92	−153.61

Table 3 Performance of the EM and SAME Algorithm on simulated data from a finite Gaussian mixture. Summary of the log posterior of the final estimates of 50 runs of each algorithm

Algorithm	Init.	T	χ	Mean	Std. Dev.	Min	Max
EM	Prior	500	500	−169.79	8.50	−181.16	−160.70
EM	Hull	500	500	−158.06	3.23	−166.39	−153.85
EM	Prior	5000	5000	−168.24	8.41	−181.02	−153.83
EM	Hull	5000	5000	−157.73	3.83	−165.81	−153.83
SAME(6)	Prior	4250	8755	−155.45	0.82	−157.56	−154.06
SAME(6)	Hull	4250	8755	−155.32	0.87	−157.35	−154.03
SAME(50)	Prior	4250	112522	−154.91	0.81	−156.22	−153.94
SAME(50)	Hull	4250	112522	−155.05	0.82	−156.11	−153.98

Table 4 Summary of the final log posterior estimated by 50 runs of the SMC Algorithm on the galaxy dataset of Roeder (1990) from a finite Gaussian mixture with varying numbers of particles, N, and intermediate distributions, T

N	T	χ	Mean	Std. Dev.	Min	Max
25	25	1325	−44.21	0.13	−44.60	−43.96
50	25	2650	−44.18	0.10	−44.48	−43.95
25	50	2125	−44.14	0.10	−44.32	−43.92
50	50	4250	−44.07	0.07	−44.22	−43.96
100	50	8500	−44.05	0.06	−44.18	−43.94
250	50	21250	−44.00	0.05	−44.10	−43.91
1000	50	85000	−43.96	0.03	−44.02	−43.92
100	100	20300	−44.03	0.05	−44.15	−43.94

Table 5 Performance of the EM and SAME Algorithm on the galaxy data of Roeder (1990) from a finite Gaussian mixture. Summary of the log posterior of the final estimates of 50 runs of each algorithm

Algorithm	Init.	T	χ	Mean	Std. Dev.	Min	Max
EM	Hull	500	500	−46.54	2.92	−54.12	−44.32
EM	Hull	5000	5000	−46.91	3.00	−56.68	−44.34
SAME(6)	Hull	4250	8755	−45.18	0.54	−46.61	−44.17
SAME(50)	Hull	4250	112522	−44.93	0.21	−45.52	−44.47

– 50 runs of the EM algorithm with 500 iterations has cost slightly higher than a single run of the SMC algorithm with $N = 250$, $T = 50$ and the best result produced is significantly inferior to the poorest run seen in the SMC case;
– 50 runs of the EM algorithm with 5000 iterations has a cost more than 10 times that of the SMC algorithm with $N = 250$, $T = 50$ and the best result produced is comparable to the worst result obtained in the SMC case.

This provides us with a degree of confidence in the algorithms considered and their ability to perform well at the level of computational cost employed here, and the next step is to consider the performance of the various algorithms on a real data set.

Table 6 SMC sampler results (mean ± standard deviation) for simulated stochastic volatility data with generating parameters of $\delta = 0.95$, $\alpha = -0.363$ and $\sigma = 0.26$

N	T	γ_T	α	δ	σ
1,000	250	4	-0.45 ± 0.19	0.939 ± 0.026	0.36 ± 0.09
1,000	500	4	-0.59 ± 0.27	0.919 ± 0.037	0.43 ± 0.11
1,000	1,000	4	-0.21 ± 0.02	0.973 ± 0.003	0.25 ± 0.02
5,000	250	4	-0.33 ± 0.06	0.954 ± 0.008	0.31 ± 0.04

4.2.2 *Galaxy data*

We also applied these algorithms, with the same hyperparameters to the galaxy data of Roeder (1990). This data set consists of the velocities of 82 galaxies, and it has been suggested that it consists of a mixture of between 3 and 7 distinct components—for example, see Roeder and Wasserman (1997) and Escobar and West (1995). For our purposes we have estimated the parameters of a 3 component Gaussian mixture model from which we assume the data was drawn. Results for the SMC algorithm are shown in Table 4 and for the other algorithms in Table 5.

We are able to draw broadly the same conclusions from these results as we were from those obtained with simulated data: the SMC algorithm performs more consistently than the alternatives and provides better estimates at given computational cost. It may be possible to fine tune all of the algorithms consider to improve their performance (including the SMC algorithm) but these results illustrate that a reasonably straightforward implementation of the SMC algorithm is able to locate at least as good a solution as any of the other algorithms considered here, and that it can do so consistently.

4.3 Stochastic volatility

In order to provide an illustration of the application of the proposed algorithm to a realistic optimisation problem in which the marginal likelihood is not available, we take this more complex example from Jacquier et al. (2007). We consider the following model:

$$Z_i = \alpha + \delta Z_{i-1} + \sigma_u u_i, \qquad Z_1 \sim \mathcal{N}(\mu_0, \sigma_0^2),$$

$$Y_i = \exp\left(\frac{Z_i}{2}\right)\epsilon_i,$$

where u_i and ϵ_i are uncorrelated standard normal random variables, and $\theta = (\alpha, \delta, \sigma_u)$. The marginal likelihood of interest, $p(\theta|y)$, where $y = (y_1, \ldots, y_{500})$ is a vector of 500 observations, is available only as a high dimensional integral over the latent variables, Z, and this integral cannot be computed.

In this case we are unable to use Algorithm 3.1, and employ a variant of Algorithm 3.3. The serial nature of the observation sequence suggests introducing blocks of the latent variable at each time, rather than replicating the entire set at each iteration. This is motivated by the same considerations as the previously discussed sequence of distributions, but makes use of the structure of this particular model. Thus, at time t, given a set of M observations, we have a sample of $\lfloor M\gamma_t \rfloor$ volatilities, $\lfloor \gamma_t \rfloor$ complete sets and $\lfloor M(\gamma_t - \lfloor \gamma_t \rfloor)\rfloor$ which comprise a partial estimate of another replicate. That is, we use target distributions of this form:

$$p_t(\alpha, \delta, \sigma, z_t) \propto p(\alpha, \delta, \sigma)\left[\prod_{i=1}^{\lfloor \gamma_t \rfloor} p(y, z_{t,i}|\alpha, \delta, \sigma)\right] \times p\left(y_{1:M(\gamma_t - \lfloor \gamma_t \rfloor)}, z_{t,i}^{1:M(\gamma_t - \lfloor \gamma_t \rfloor)}|\alpha, \delta, \sigma\right),$$

where $z_{t,i}^{1:M(\gamma_t - \lfloor \gamma_t \rfloor)}$ denotes the first $\lfloor M(\gamma_t - \lfloor \gamma_t \rfloor)\rfloor$ volatilities of the ith replicate at iteration t.

Making use of diffuse conjugate prior distributions (uniform over the $(-1, 1)$ stability domain for δ, standard normal for α and inverse gamma with parameters $\alpha = 1$, $\beta = 0.1$ for σ^2) for θ ensures that the prior distributions are rapidly "forgotten", leading to a maximum likelihood estimate. Our sampling strategy at each time is to sample (α, δ) from their joint conditional distribution, then to sample σ from its conditional distribution. These distributions are multivariate normal and inverse gamma, respectively. Their parameters are given by Jacquier et al. (2007). New volatilities were then sampled using a Kalman smoother obtained by a local linearisation of the model as the proposal distribution—an approach described in some detail in Doucet et al. (2006).

4.3.1 *Simulated data*

We consider a sequence of 500 observations generated from a stochastic volatility model with parameter values of $\delta = 0.95$, $\alpha = -0.363$ and $\sigma^2 = 0.26$ (suggested by Jacquier et al. 2007 as being *consistent with empirical estimates for financial equity return time series*). The parameters $\mu_0 = -7$, $\sigma_0 = 1$ were assumed known. Results are shown in Table 6.

Note that a greater number of particles and intermediate distributions are required in this case than were needed in the previous examples for a number of reasons. Unavailability of the likelihood makes the problem a little more difficult,

but the principle complication is that it is now necessary to integrate out 500 continuous-valued latent variables.

The intention of this example is to show how the algorithm can be applied in more complex settings. The results shown here do not provide rigorous evidence that the algorithm is performing well, but heuristically that does appear to be the case. Estimated parameter values are close to their true values[1] and the degree of dispersion is comparable to that observed by Jacquier et al. (2007) at small values of γ using data simulated with the same parameters. It can be seen that the results obtained are reasonably robust to variation in the number of particles and intermediate distributions which are utilised.

References

Amzal, B., Bois, F.Y., Parent, E., Robert, C.P.: Bayesian optimal design via interacting particle systems. J. Am. Stat. Assoc. **101**(474), 773–785 (2006)

Chopin, N.: Central limit theorem for sequential Monte Carlo methods and its applications to Bayesian inference. Ann. Stat. **32**(6), 2385–2411 (2004)

Del Moral, P.: Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications. Probability and its Applications. Springer, New York (2004)

Del Moral, P., Doucet, A., Jasra, A.: Sequential Monte Carlo samplers. J. R. Stat. Soc. B **63**(3), 411–436 (2006)

Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM Algorithm. J. R. Stat. Soc. B **39**, 2–38 (1977)

Doucet, A., de Freitas, N., Gordon, N. (eds.): Sequential Monte Carlo Methods in Practice. Statistics for Engineering and Information Science. Springer, New York (2001)

Doucet, A., Godsill, S.J., Robert, C.P.: Marginal maximum a posteriori estimation using Markov chain Monte Carlo. Stat. Comput. **12**, 77–84 (2002)

Doucet, A., Briers, M., Sénécal, S.: Efficient block sampling strategies for sequential Monte Carlo methods. J. Comput. Graph. Stat. **15**(3), 693–711 (2006)

Escobar, M.D., West, M.: Bayesian density estimation and inference using mixtures. J. Am. Stat. Assoc. **90**(430), 577–588 (1995)

Gaetan, C., Yao, J.-F.: A multiple-imputation Metropolis version of the EM algorithm. Biometrika **90**(3), 643–654 (2003)

Hwang, C.-R.: Laplace's method revisited: weak convergence of probability measures. Ann. Probab. **8**(6), 1177–1182 (1980)

Jacquier, E., Johannes, M., Polson, N.: MCMC maximum likelihood for latent state models. J. Econom. **137**(2), 615–640 (2007)

Johansen, A.M.: Some non-standard sequential Monte Carlo methods with applications, Ph.D. thesis. University of Cambridge, Department of Engineering (2006)

Liu, J.S., Chen, R.: Sequential Monte Carlo methods for dynamic systems. J. Am. Stat. Assoc. **93**(443), 1032–1044 (1998)

Müller, P., Sansó, B., de Iorio, M.: Optimum Bayesian design by inhomogeneous Markov chain simulation. J. Am. Stat. Assoc. **99**(467), 788–798 (2004)

Robert, C.P., Casella, G.: Monte Carlo Statistical Methods, 2nd edn. Springer, New York (2004)

Robert, C.P., Titterington, D.M.: Reparameterization strategies for hidden Markov models and Bayesian approaches to maximum likelihood estimation. Stat. Comput. **8**, 145–158 (1998)

Roeder, K.: Density estimation with cofidence sets exemplified by superclusters and voids in galaxies. J. Am. Stat. Assoc. **85**(411), 617–624 (1990)

Roeder, K., Wasserman, L.: Practical Bayesian density estimation using mixtures of normals. J. Am. Stat. Assoc. **92**(439), 894–902 (1997)

[1] Of course, there is no guarantee that the maximum likelihood estimate should correspond to the *true* parameter value in the case of a finite sample, but with a reasonably large number of observations one might expect a degree of similarity.