MAXIMUM LIKELIHOOD PARAMETER ESTIMATION FOR LATENT VARIABLE MODELS USING SEQUENTIAL MONTE CARLO

Adam Johansen[1], *Arnaud Doucet*[2] *and Manuel Davy*[3]

1 - University of Cambridge Department of Engineering, Trumpington Street, Cambridge, CB2 1PZ, UK

2 - Department of Statistics & Department of Computer Science, University of British Columbia, Vancouver, Canada

3 - LAGIS UMR 8146, BP 48, Cité scientifique, 59651 Villeneuve d'Ascq Cedex, France

ABSTRACT

We present a sequential Monte Carlo (SMC) method for maximum likelihood (ML) parameter estimation in latent variable models. Standard methods rely on gradient algorithms such as the Expectation-Maximization (EM) algorithm and its Monte Carlo variants. Our approach is different and motivated by similar considerations to simulated annealing (SA); that is we propose to sample from a sequence of artificial distributions whose support concentrates itself on the set of ML estimates. To achieve this we use SMC methods. We conclude by presenting simulation results on a toy problem and a nonlinear non-Gaussian time series model.

1. INTRODUCTION

1.1. Problem Formulation

The situation in which we are interested is that in which one has some likelihood function $p(y, z|\theta)$ in which $(y, z) \in \mathcal{Y} \times \mathcal{Z}$ for observed data y and latent variables (often called "hidden data"), z. Although this joint likelihood is known, as z is not observed, the *marginal likelihood*

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

is the quantity of interest, and as this integral is generally not tractable, it is not straightforward to maximise it with respect to the parameters to obtain the (marginal) maximum likelihood (ML) estimator:

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

1.2. Previous Approaches

When the marginal likelihood $p(y|\theta)$ can be evaluated, the classical approach to problems of this sort is the EM algorithm [1], which is a numerically well-behaved gradient-based algorithm. For complex models, $p(y|\theta)$ cannot be computed analytically and typically the expectation step of the EM algorithm cannot be performed in closed-form either. In such scenarios, Monte Carlo variants of EM have been proposed – including stochastic EM (SEM), Monte Carlo EM (MCEM) and stochastic approximation EM (SAEM). See [2] for a comparative summary of these approaches. Note that all of these (stochastic) gradient-based approaches are susceptible to trapping in local modes.

An alternative approach related to SA is to build a sequence of distributions which concentrates itself on the set of the required estimate. Let $p(\theta)$ be an instrumental prior distribution whose support includes the ML estimate then the distributions

$$p_\gamma^{ML}(\theta|y) \propto p(\theta)\, p(y|\theta)^\gamma$$

concentrate themselves on the set of ML estimates as $\gamma \to \infty$. Indeed asymptotically the contribution from this instrumental prior vanishes. The term $p(\theta)$ is here only present to ensure that the distributions $\left\{p_\gamma^{ML}(\theta|y)\right\}$ are integrable – it may be omitted in those instances in which this is already the case. To sample from these distributions, one would like to use Markov chain Monte Carlo (MCMC) methods. Unfortunately, this is impossible whenever $p(y|\theta)$ is not known pointwise up to a normalizing constant.

To circumvent this problem, it has been proposed in [3] (in a Maximum a Posteriori, rather than ML, setting) to build a sequence of artificial distributions known, up to a normalizing constant, which admit as a marginal distribution the target distribution $p_\gamma^{ML}(\theta|y)$ for an integer power γ greater than one. A similar scheme was subsequently been proposed by [4, 5] in the ML setting. This is achieved by simulating a number of replicates of the missing data where one defines

$$p_\gamma(\theta, z_{1:\gamma}|y) \propto p(\theta) \prod_{i=1}^{\gamma} p(y, z_i|\theta) \tag{3}$$

with $z_{i:j} = (z_i, ..., z_j)$. Indeed it is easy to check that

$$\int \cdots \int p_\gamma(\theta, z_{1:\gamma}|y)dz_{1:\gamma} = p_\gamma^{ML}(\theta|y).$$

The approach of [3] is to construct an inhomogeneous Markov chain which produces samples from a sequence of such distributions for increasing values of γ. Just as in SA, this concentrates the mass on the set of global maxima of $p(y|\theta)$ as γ becomes large. Another approach proposed by [4] is to construct a homogeneous Markov chain whose invariant distribution corresponds to such a distribution for a predetermined value of γ. It can be theoretically established that these methods converge asymptotically towards the set of estimates of interest if γ grows slowly enough to ∞. However in practice, these approach suffer from several weaknesses. First, they allows only integer values for γ. Second, unless a very slow annealing schedule is used, the MCMC chain tends to become trapped in local modes.

We propose another approach to sampling from $p_\gamma(\theta, z_{1:\gamma}|y)$. We sample from this sequence of distributions using SMC. SMC methods have been used primarily to solve optimal filtering problems in signal processing and statistics. They are used here in a completely different framework which requires extensions of the methodology described further in the paper. Broadly speaking, the distributions of interest are 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, and the framework naturally allows for the introduction of bridging densities between target distributions, say, $p_\gamma(\theta, z_{1:\gamma}|y)$ and $p_{\gamma+1}(\theta, z_{1:\gamma+1}|y)$ for an integer γ. At first glance, the algorithm appears very close to mutation-selection schemes employed in the genetic algorithms literature. However, there are two major differences with these algorithms. First, they require the function being maximized to be known pointwise, whereas we do not. Second, convergence results for our method follow straightforwardly from general results on Feynman-Kac flows [6].

2. AN SMC SAMPLER APPROACH

2.1. Background

The SMC samplers framework of [7] is a very general method for obtaining a set of samples from a sequence of distributions which can exist on the same or different spaces. This is a generalisation of the standard SMC method (commonly referred to as particle filtering and summarised by [8]) in which the target distribution exists on a space of strictly increasing dimension and no mechanism exists for updating the estimates of the state at earlier times after receiving new data. It is not possible to give a thorough exposition of the SMC samplers approach here, but we will try to include sufficient detail for our purposes. We remark that convergence results, including a central limit theorem, for the particle estimates obtained by this method are available (indeed, these are applications of standard results on Feynman-Kac flows [6]) and are contained in [7].

Given a sequence of distributions $(\pi_t)_{t\geq 1}$ on a sequence of measurable spaces $(E_t, \mathcal{E})_{t\geq 1}$ from which we wish to obtain sets of weighted samples, we construct a sequence of distributions on a sequence of spaces of increasing dimension which admit the distributions of interest as marginals, by defining:

$$\widetilde{\pi}_t(x_{1:t}) = \pi_t(x_t) \prod_{s=t-1}^{1} L_s(x_{s+1}, x_s)$$

where L_s is an arbitrary Markov kernel from space E_{s+1} to E_s (these act, in some sense, backwards in time). It is clear that standard SMC methods can now be applied on this space, by propagating samples forward from one distribution to the next according to a sequence of Markov kernels, $(K_t)_{t\geq 2}$, and correcting for the discrepancy between the proposal and the target distribution by importance sampling. As always it is important to ensure that a significant fraction of the particle set have non-negligible weights. The effective sample size (ESS), introduced by [9], is an approximation obtained by Taylor expansion of a quantity which describes the effective number of iid samples to which the set corresponds. The ESS is defined as $ESS = \left[\sum_{i=1}^N W^{(i)^{-2}}\right]^{-1}$ where $\left\{W^{(i)}\right\}$ are the normalized weights. This approximation, of course, fails if the particle set does not accurately represent the support of the distribution of interest. Resampling should be carried out after any iteration which causes the ESS to fall below a reasonable threshold (typically around half of the total number of particles), to prevent the sample becoming degenerate with a small number of samples having very large weights.

The rapidly increasing dimension raises the concern that the variance of the importance weights will be extremely high. It can be shown (again, see [7]) that the optimal form for the Markov kernels L_s – in the sense that they minimise the variance of the importance weights if resampling occurs at every time step – depends upon the distributions of interest and the importance sampling proposal kernels K_t in the following way:

$$L_t^{opt}(x_{t+1}, x_t) = \frac{\pi_t(x_t) K_{t+1}(x_t, x_{t+1})}{\int \pi_t(x) K_{t+1}(x, x_{t+1}) dx} \tag{4}$$

In practice it is important to choose a sequence of kernels which are as close to the optimal case as possible to prevent the variance of the importance weights from becoming extremely large.

2.2. SMC Samplers for Parameter Estimation

We will consider the use of the sampling methodology described in the previous section for marginal ML estimation – noting that the method can be easily adapted to Bayesian marginal Maximum a Posteriori setting by considering a slightly different sequence of target distributions. The target distribution which we propose as generally admissible for this task, although any distribution from which an efficient sampler can be constructed which admit marginals of the desired form would be equally acceptable, is

$$\pi_t(\theta, z_{1:\lceil\gamma_t\rceil}|y) \propto p(\theta) \prod_{i=1}^{\lfloor\gamma_t\rfloor} p(y, z_i|\theta) p(y, z_{\lceil\gamma_t\rceil}|\theta)^{\gamma_t - \lfloor\gamma_t\rfloor}$$

where $\lfloor\gamma_t\rfloor$ denotes the largest integer not greater than γ_t and $\lceil\gamma_t\rceil$ the smallest integer not less than γ_t. This final term allows us to introduce sequence with non-integer elements, whilst having the same form as (3). Clearly, we have $\pi_t(\theta, z_{1:\lceil\gamma_t\rceil}|y) = p_{\gamma_t}(\theta, z_{1:\gamma_t}|y)$ for any integer γ_t. Again, an increasing sequence $(\gamma_t)_{t\geq 1}$ is required, corresponding in some sense to the annealing schedule of SA. To simplify notation we will denote $Z_{t,1:\lceil\gamma_t\rceil}^{(i)}$ – the values of $z_{1:\lceil\gamma_t\rceil}$ simulated at time t for the ith particle – by $Z_t^{(i)}$. Algorithm 1 describes the general framework which we propose.

Algorithm 1: The General Framework

Initialisation, t=1:

Sample, $\left\{\left(\theta_1^{(i)}, Z_1^{(i)}\right)\right\}_{i=1}^N$ independently from $\nu(\cdot)$

Calculate importance weights $W_1^{(i)} \propto \frac{\pi_1(\theta_1^{(i)}, Z_1^{(i)})}{\nu(\theta_1^{(i)}, Z_1^{(i)})}$

If ESS < Threshold, resample.

At time $t > 1$:

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

$\forall 1 \leq i \leq N : \left(\theta_t^{(i)}, Z_t^{(i)}\right) \sim K_t\left(\left(\theta_{t-1}^{(i)}, Z_{t-1}^{(i)}\right), \cdot\right)$

Calculate importance weights

$\frac{W_t^{(i)}}{W_{t-1}^{(i)}} \propto \frac{\pi_t(\theta_t^{(i)}, Z_t^{(i)}) L_{t-1}\left(\left(\theta_t^{(i)}, Z_t^{(i)}\right), \left(\theta_{t-1}^{(i)}, Z_{t-1}^{(i)}\right)\right)}{\pi_{t-1}(\theta_{t-1}^{(i)}, Z_{t-1}^{(i)}) K_t\left(\left(\theta_{t-1}^{(i)}, Z_{t-1}^{(i)}\right), \left(\theta_t^{(i)}, Z_t^{(i)}\right)\right)}$

If ESS < Threshold, resample.

It is interesting to consider an analytically convenient special case, which leads to a particularly elegant algorithm, 2, below. When we are able to sample from particular conditional distributions, and evaluate the marginal likelihood pointwise, it is possible to evaluate the optimal auxiliary kernels as given by (4). 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 [2, p. 351].

Algorithm 2: A Special Case

Initialisation, t=1:

Sample, $\left\{\left(\theta_1^{(i)}, Z_1^{(i)}\right)\right\}_{i=1}^N$ independently from $\nu(\cdot)$

Calculate importance weights $W_1^{(i)} \propto \frac{\pi_1(\theta_1^{(i)}, Z_1^{(i)})}{\nu(\theta_1^{(i)}, Z_1^{(i)})}$

If ESS $<$ Threshold, resample.

At time $t > 1$:

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

$\theta_t^{(i)} \sim \pi_t(\cdot|z_{t-1}^{(i)})$, $Z_{t,1:\lfloor\gamma_{t-1}\rfloor}^{(i)} = Z_{t-1}^{(i)}$

for $j = \lfloor\gamma_{t-1}\rfloor + 1$ to $\lfloor\gamma_t\rfloor$, $Z_{t,j}^{(i)} \sim p(\cdot|\theta_t^{(i)})$

and $Z_{t,\lceil\gamma_t\rceil}^{(i)} \sim p(\cdot|\theta_t^{(i)})^{\gamma_t - \lfloor\gamma_t\rfloor}$

Calculate importance weights

$W_t^{(i)} \propto W_{t-1}^{(i)} p(y|\theta)^{\gamma_t - \gamma_{t-1}}$

If ESS $<$ Threshold, resample.

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. We assume that a collection of Markov kernels $(\mathcal{K}_t)_{t\geq 1}$ with invariant distributions corresponding to $(\pi_t)_{t\geq 1}$ is available, and using these as a component of the proposal kernels allows the evaluation of the optimal auxiliary kernel. We assume that good importance distributions for the conditional probability of the variables being marginalised can be sampled from and evaluated, $q(\cdot|\theta)$, and that if the annealing schedule is to include non-integer inverse temperatures, then we have appropriate importance distributions for distributions proportional to $p(z|\theta)^\alpha, \alpha \in (0, 1)$, which we denote $q_\alpha(z|\theta)$. We remark that this is not the most general possible approach, but is one which should work acceptably for a broad class of problems.

Algorithm 3: A Generic Case

Initialisation, t=1:

Sample, $\left\{\left(\theta_1^{(i)}, Z_1^{(i)}\right)\right\}_{i=1}^N$ independently from $\nu(\cdot)$

Calculate importance weights $W_t^{(i)} \propto \frac{\pi_1(\theta_1^{(i)}, Z_1^{(i)})}{\nu(\theta_1^{(i)}, Z_1^{(i)})}$

If ESS $<$ Threshold, resample.

At time $t > 1$:

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

$\left(\theta_t^{(i)}, Z_{t,1:\lfloor\gamma_{t-1}\rfloor}^{(i)}\right) \sim \mathcal{K}_{t-1}\left(\theta_{t-1}^{(i)}, Z_{t-1}^{(i)}; \cdot\right)$

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

and $Z_{t,\lceil\gamma_t\rceil}^{(i)} \sim q_{\gamma_t - \lfloor\gamma_t\rfloor}(\cdot|\theta_t^{(i)})$

Calculate importance weights

$\frac{W_t^{(i)}}{W_{t-1}^{(i)}} \propto \prod_{j=\lfloor\gamma_{t-1}\rfloor+1}^{\lfloor\gamma_t\rfloor} \frac{p(y, Z_j^{(i)}|\theta^{(i)}) p(y, Z_{\lceil\gamma_t\rceil}^{(i)}|\theta^{(i)})^{\gamma_t - \lfloor\gamma_t\rfloor}}{q(Z_j^{(i)}|\theta^{(i)}) q_{\gamma_t - \lfloor\gamma_t\rfloor}(Z_{\gamma_t}^{(i)}|\theta^{(i)}}$

If ESS $<$ Threshold, resample.

3. EXAMPLES AND RESULTS

3.1. Toy Example

We consider first a toy example in one dimension. We borrow example 1 of [5] for this purpose. The model consists of a student t-distribution of unknown location parameter θ with 0.025 degrees of freedom. Four observations are available, $y = (-20, 1, 2, 3)$. The logarithm of the marginal likelihood in this instance is given by:

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

N	T	Mean	Std. Dev.	Min	Max
50	15	1.992	0.014	1.95	2.03
100	15	1.997	0.013	1.97	2.04
20	30	1.958	0.177	1.09	2.04
50	30	1.997	0.008	1.98	2.01
100	30	1.997	0.007	1.98	2.01
20	60	1.998	0.015	1.91	2.02
50	60	1.997	0.005	1.99	2.01

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

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

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

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

$$\pi_t(z_{1:\gamma_t}|\theta, y) = \prod_{i=1}^{\gamma_t} \mathcal{G}a\left(z_i \middle| 0.525, 0.025 + \frac{(y_i - \theta)^2}{2}\right) \quad (5)$$

$$\pi_t(\theta|z_{1:\gamma_t}) \propto \mathcal{N}\left(\theta \middle| \mu_t^{(\theta)}, \Sigma_t^{(\theta)}\right) \quad (6)$$

where the parameters

$$\Sigma_t^{(\theta)} = \left[\sum_{i=1}^{t}\sum_{j=1}^{4} z_{i,j}\right]^{-1} = \left[1/\Sigma_{t-1}^{(\theta)} + \sum_{j=1}^{4} z_{t,j}\right]^{-1} \quad (7)$$

$$\mu_t^{(\theta)} = \Sigma_t^{(\theta)} \sum_{i=1}^{t} y^T z_i = \Sigma_t^{(\mu)}\left(\mu_{t-1}^{(\theta)}/\Sigma_{t-1}^{(\theta)} + y^T z_t\right) \quad (8)$$

may be obtained recursively. Consequently, we can make use of algorithm 2 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 may be justified by the asymptotic (in the inverse temperature) normality of the target distribution (see, for example, [2, p. 203]).

3.2. Stochastic Volatility

We take this more complex example from [4]. We consider the following model:

$$Z_i = \alpha + \delta Z_{i-1} + \sigma_u u_i \qquad Z_1 \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right) \quad (9)$$

$$Y_i = \exp\left(\frac{Z_i}{2}\right)\epsilon_i \quad (10)$$

where u_i and ϵ_i are uncorrelated standard normal random variables, and $\theta = (\alpha, \delta, \sigma_u)$. The marginal likelihood of interest $p(\theta|y)$ is

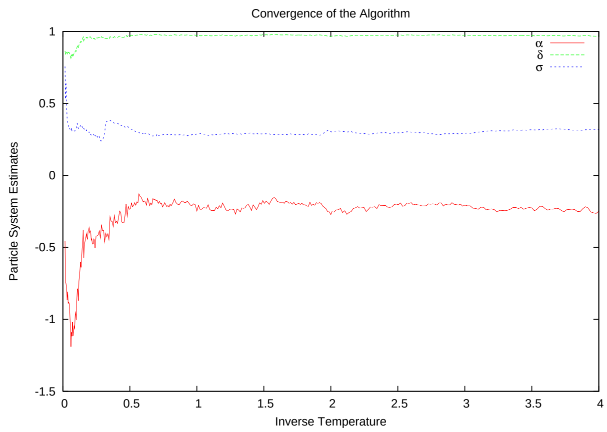

Fig. 1. Parameter estimates as a function of the inverse temperature.

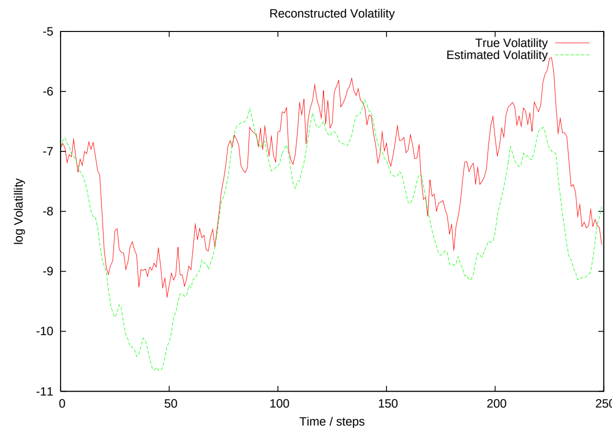

Fig. 2. Estimated volatility (dashed line) obtained using 250 particles and the true volatility (solid line).

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 2, and employ a variant of algorithm 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 $M\gamma_t$ volatilities, $\lfloor\gamma_t\rfloor$ complete sets and $M(\gamma_t - \lfloor\gamma_t\rfloor)$ which comprise a partial estimate of another replicate. That is, we use target distributions of this form:

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

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

Making use of diffuse conjugate prior distributions[1] 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 a Gibbs sampling kernel and to propose block-based Metropolis-Hastings moves for the existing latent variables (using a Kalman-filter and forward filtering / backward sampling proposal) before proposing new volatilities using the same proposal strategy.

Considering a sequence of 250 observations and a temperature scale which increases in a piecewise linear manner to a temperature of 4 in 340 steps, using 250 particles we obtained estimates of the three parameters, by the same mechanism as that used in the toy example, of $\delta = 0.96 \pm 0.02$, $\alpha = -0.25 \pm 0.01$ and $\sigma = 0.31 \pm 0.02$ where the true values were $\delta = 0.95$, $\alpha = -0.363$ and $\sigma = 0.26$ (suggested by [4] as being *consistent with empirical estimates for financial equity return time series*). Figure 1 shows the estimated parameter values as a function of the number of observations incorporated. Figure 2 shows the volatility estimated by its posterior mean under the empirical distribution of the final particle set – as the distribution over θ converges to a point mass at the maximum likelihood value, this amounts to the conditional expectation of the volatility.

It is clear that a longer sequence and an annealing schedule which reaches a lower temperature will have a likelihood of very similar functional form with the only difference being that the latent variables associated with each observation are replicated, rather than there simply being a greater number of latent variables. Of course longer series provide more information and hence allow better estimates, indeed, with a sequence of length 1,000 and a final temperature of 1, with just 100 particles, we obtained estimates of $\alpha = -0.40 \pm 0.14$, $\sigma = 0.28 \pm 0.03$ and $\delta = 0.95 \pm 0.02$.

[1] i.e. uniform over the $(-1, 1)$ stability domain for δ, standard normal for α and square-root inverse gamma with parameters $\alpha = 1$, $\beta = 0.1$ for σ.

4. REFERENCES

[1] A.P. Dempster, N.M. Laird, and D.B.Rubin, "Maximum likelihood from incomplete data vie the EM Algorithm" , *Journal of the Royal Statistical Society, Series B*, vol. 39, pp. 2-38, 1977.

[2] C.P. Robert, and G. Casella, *Monte Carlo Statistical Methods*. New York: Springer-Verlag, second edition,2004.

[3] A. Doucet, S.J. Godsill and C.P. Robert , "Marginal maximum a posteriori estimation using Markov chain Monte Carlo" , *Statistics and Computing*, vol. 12, pp. 77-84, 2002.

[4] E. Jacquier, M. Johannes, and N. Polson , "MCMC maximum likelihood for latent state models" , *Journal of Econometrics*, 2005, To Appear.

[5] C. Gaetan, and J.F. Yao , "A multiple-imputation Metropolis version of the EM algorithm" , *Biometrika*, vol. 90, no. 3, pp. 643-654, 2003

[6] P. Del Moral, *Feynman-Kac Formulae. Genealogical and Interacting Particle Approximations*, New York: Springer-Verlag, 2004.

[7] P. Del Moral, A. Doucet and A. Jasra, "Sequential Monte Carlo methods for Bayesian computation" (with discussion), in *Bayesian Statistics 8*, Oxford University Press, to appear 2006.

[8] A. Doucet, J.F.G. de Freitas and N.J. Gordon (eds.), *Sequential Monte Carlo Methods in Practice*. Statistics for Engineering and Information Science, New York: Springer-Verlag, 2001.

[9] A. Kong, J.S. Liu, and W.H. Wong, "Sequential imputations and bayesian missing data problems" . *Journal of the American Statistical Association*, vol. 89, no. 425, pp. 278-288, March 1994.