

Auxiliary particle filtering within adaptive Metropolis-Hastings sampling

Michael K. Pitt

Economics Department

University of Warwick

m.pitt@warwick.ac.uk

Ralph S. Silva

School of Economics

University of New South Wales

r.silva@unsw.edu.au

Paolo Giordani

Research Department

Sveriges Riksbank

paolo.giordani@riksbank.se

Robert Kohn*

School of Economics

University of New South Wales

r.kohn@unsw.edu.au

May 15 2010

Abstract

Our article deals with Bayesian inference for a general state space model with the simulated likelihood computed by the particle filter. We show empirically that the partially or fully adapted particle filters can be much more efficient than the standard particle filter, especially when the signal to noise ratio is high. This is especially important because using the particle filter within Markov chain Monte Carlo sampling is $O(T^2)$, where T is the sample size. We also show that an adaptive independent Metropolis Hastings proposal for the unknown parameters based on a mixture of normals can be much more efficient than the usual optimal random walk methods because the simulated likelihood is not continuous in the parameters and the cost of constructing a good adaptive proposal is negligible compared to the cost of evaluating the simulated likelihood. Independent Metropolis-Hastings proposals are also attractive because they are easy to run in parallel on multiple processors. The article also shows that the proposed adaptive independent Metropolis Hastings sampler converges to the posterior distribution. We also show that the marginal likelihood of any state space model can be obtained in an efficient and unbiased manner by using the particle filter making model comparison straightforward. Obtaining the marginal likelihood is often difficult using other methods. Finally, we prove that the simulated likelihood obtained by the auxiliary particle filter is unbiased. This result is fundamental to using the particle filter for Markov chain Monte Carlo sampling and is first obtained in a more abstract

*Corresponding author.

and difficult setting by Del Moral (2004). However, our proof is direct and will make the result accessible to readers.

Keywords: Auxiliary variables; Bayesian inference; Bridge sampling; Full and partial adaptation; Marginal likelihood.

1 Introduction

Our article builds on the work of Andrieu et al. (2010) and develops general simulation methods that make Bayesian inference for time series state space models feasible and efficient. Our first contribution is to show empirically that partially or fully adapted auxiliary particle filters using the simulated likelihood defined in Pitt (2002) can be much more efficient statistically than the standard standard particle filter of Gordon et al. (1993), especially when the signal to noise ratio is high, because they reduce the noise in the simulated likelihood. It is very important to carry out the particle filter as efficiently as possible, because particle filtering within Markov chain Monte Carlo sampling is $O(T^2)$, where T is the sample size as explained in Section 2.1.

Adaptive sampling methods are simulation methods for carrying out Bayesian inference that use previous iterates of the simulation to form proposal distributions, that is, the adaptive samplers learn about some aspects of the posterior distribution from previous iterates. See for example Haario et al. (2001), Atchadé and Rosenthal (2005) and Roberts and Rosenthal (2009) who consider adaptive random walk Metropolis proposals and Giordani and Kohn (2010) who base their proposal on a mixture of normals.

The second contribution of the article is to show that when working with a simulated likelihood it is worthwhile constructing adaptive independent Metropolis Hastings proposals that provide good approximations to the posterior density. The first reason for this claim is that the simulated likelihood is not continuous in the unknown parameters. This means that standard methods for constructing proposals such as Laplace approximations based on analytic or numerical derivatives are usually infeasible. It also means that the usual optimal random walk methods do not perform as well as expected as the probability of acceptance does not tend to 1 as a proposed move becomes more local moves or even if the parameter does not change at all. The second reason is that the cost of constructing a good adaptive proposal is often negligible compared to the cost of running the particle filter to obtain the simulated likelihood. Third, an adaptive sampling scheme that consists entirely or mostly of independent Metropolis-Hastings steps is attractive because a large part of the computation can be run in parallel thus substantially reducing computing time.

Our article uses the adaptive independent Metropolis Hastings sampler of Giordani and Kohn (2010) which approximates the posterior density by a mixture of normals. We show that this proposal density can be much more efficient than the adaptive random walk Metropolis proposal of Roberts and Rosenthal (2009) for the reasons just outlined. We also show that this adaptive sampler converges to the correct posterior distribution. We note, however, that in our experience, the adaptive random walk Metropolis algorithm of Roberts and Rosenthal (2009) is important because it converges reliably for a diverse set of problems and provides a good way to initialize other more efficient adaptive sampling schemes.

The third contribution of our article is to show that the marginal likelihood of any state

space model can be estimated in an efficient and unbiased manner by combining particle filtering with bridge or importance sampling. This makes it straightforward to compare the marginal likelihoods of two or more models each of which can be expressed in state space form.

The final contribution of the article is to show that the simulated likelihood obtained the auxiliary particle filter is unbiased. This result is obtained in an abstract setting in Proposition 7.4.1 in Section 7.4.2 in Del Moral (2004). Andrieu et al. (2010) show that the unbiasedness of the simulated likelihood allows Bayesian inference using Markov chain Monte Carlo simulation. This is because the simulated likelihood can be viewed as the density of the observations conditional on the unknown parameters and a set of auxiliary latent variables. We believe that our derivation of unbiasedness is more direct and accessible than that of Del Moral (2004).

Computational algorithms for state space models such as the Kalman filter and particle filter are useful because many time series models can be expressed in state space form. Computational methods for Bayesian inference for Gaussian state space models are well developed (see Cappé et al., 2005) and there is a literature now on Bayesian computational methods for non-Gaussian state space models. Markov chain Monte Carlo computational methods based on the particle filter have the **potential to greatly increase the number and complexity of time series models amenable to Bayesian analysis**. An early use of the particle filter within a Markov chain Monte Carlo framework is by Fernández-Villaverde and Rubio-Ramírez (2007) who applied it to macroeconomic models as an approximate approach for obtaining the posterior distribution of the parameters.

Particle filtering (also known as sequential Monte Carlo) was proposed by Gordon et al. (1993) for online filtering and prediction of nonlinear or non-Gaussian state space models. The auxiliary particle filter method was introduced by Pitt and Shephard (1999) to **improve the performance of the standard particle filter when the observation equation is informative relative to the state equations, that is when the signal to noise ratio is moderate to high**. There is an extensive literature on online filtering using the particle filter, see for example Kitagawa (1996), Liu and Chen (1998), Doucet et al. (2000), Doucet et al. (2001), Andrieu and Doucet (2002), Fearnhead and Clifford (2003) and Del Moral et al. (2006). Our article considers only the standard particle filter of Gordon et al. (1993) and the fully and partially adapted particle filters proposed by Pitt and Shephard (1999).

The literature on using the particle filter to learn about model parameters is more limited. Pitt (2002) proposes the smooth particle filter to estimate the parameters of a state space using maximum likelihood. Storvik (2002) and Polson et al. (2008) consider online parameter learning when sufficient statistics are available. Andrieu et al. (2010) provide a framework for off-line parameter learning using the particle filter. Flury and Shephard (2008) give an insightful discussion of the results of Andrieu et al. (2010) and use single parameter random walk proposals for off-line Bayesian inference.

Our article is an updated version of Silva et al. (2009), which contains some extra examples.

2 State space models

Consider a state space model with observation equation $p(y_t|x_t; \theta)$ and state transition equation $p(x_t|x_{t-1}; \theta)$, where y_t and x_t are the observation and the state at time t and θ is a vector of unknown parameters. The distribution of the initial state is $p(x_0|\theta)$. See Cappé et al. (2005) for a modern treatment of general state space models. The **filtering equations** for the state space model (for $t \geq 1$) are (West and Harrison, 1997, pp. 506-507)

$$p(x_t|y_{1:t-1}; \theta) = \int p(x_t|x_{t-1}; \theta)p(x_{t-1}|y_{1:t-1}; \theta)dx_{t-1}, \quad (1a)$$

$$p(x_t|y_{1:t}; \theta) = \frac{p(y_t|x_t; \theta)p(x_t|y_{1:t-1}; \theta)}{p(y_t|y_{1:t-1}; \theta)}, \quad (1b)$$

$$p(y_t|y_{1:t-1}; \theta) = \int p(y_t|x_t; \theta)p(x_t|y_{1:t-1}; \theta)dx_t. \quad (1c)$$

where $y_{1:t} = \{y_1, \dots, y_t\}$. Equations (1a)–(1c) allow (in principle) for filtering for a given θ and for evaluating the **likelihood of the observations $y = y_{1:T}$** ,

$$p(y|\theta) = p(y_1|\theta) \prod_{t=1}^{T-1} p(y_{t+1}|y_{1:t}; \theta). \quad (2)$$

If the likelihood $p(y|\theta)$ can be computed, maximum likelihood and MCMC methods can be used to carry out inference on the parameters θ , with the states integrated out. When both the observation and state transition equations are linear and Gaussian the likelihood can be evaluated analytically using the Kalman filter (Cappé et al., 2005, pp. 141-143). More general state space models can also be estimated by MCMC methods if auxiliary latent variables are introduced, e.g. Kim et al. (1998) and Frühwirth-Schnatter and Wagner (2006) and/or the states are sampled in blocks as in Shephard and Pitt (1997). See section 6.3 of Cappé et al. (2005) for a review of Markov chain Monte Carlo methods applied to general state space models.

In general, however, **the integrals in equations (1a)–(1c) are computationally intractable** and the standard particle filter algorithm (SIR) was proposed by Gordon et al. (1993) as a method for approximating them with the approximation becoming exact as the number of particles tends to infinity. Pitt and Shephard (1999) propose the auxiliary particle filter method (ASIR) which is more efficient than standard particle filter when the observation density is informative relative to the transition density. The general auxiliary particle filter is described in Section 2.1.

2.1 General ASIR method

The general auxiliary SIR (ASIR) filter of Pitt and Shephard (1999) may be thought of as a generalisation of the SIR method of Gordon et al. (1993). We therefore focus on this, more general, approach. To simplify notation in this section, we omit to show dependence on the unknown parameter vector θ . The following algorithm describes the one time step ASIR update and is initialized with samples $x_0^k \sim p(x_0)$ with mass $1/M$ for $k = 1, \dots, M$.

Algorithm 1. Given samples $x_t^k \sim p(x_t|y_{1:t})$ with mass π_t^k for $k = 1, \dots, M$.

For $t = 0, \dots, T - 1$:

1. For $k = 1 : M$, compute $\omega_{t|t+1}^k = g(y_{t+1}|x_t^k)\pi_t^k$, $\pi_{t|t+1}^k = \frac{\omega_{t|t+1}^k}{\sum_{i=1}^M \omega_{t|t+1}^i}$.
2. For $k = 1 : M$, sample $\tilde{x}_t^k \sim \sum_{i=1}^M \pi_{t|t+1}^i \delta(x_t - x_t^i)$.
3. For $k = 1 : M$, sample $x_{t+1}^k \sim g(x_{t+1}|\tilde{x}_t^k; y_{t+1})$.
4. For $k = 1 : M$, compute

$$\omega_{t+1}^k = \frac{p(y_{t+1}|x_{t+1}^k)p(x_{t+1}^k|\tilde{x}_t^k)}{g(y_{t+1}|\tilde{x}_t^k)g(x_{t+1}^k|\tilde{x}_t^k; y_{t+1})}, \quad \pi_{t+1}^k = \frac{\omega_{t+1}^k}{\sum_{i=1}^M \omega_{t+1}^i}.$$

Note that in Step 2, $\delta(x - a)$ is the delta function with unit mass at $x = a$. In addition, in Step 2, multinomial sampling may be employed but stratified sampling is generally to be preferred and is employed throughout, see Kitagawa (1996), Carpenter et al. (1999) and Pitt and Shephard (2001).

Note that the true joint density may be written as,

$$p(y_{t+1}|x_{t+1})p(x_{t+1}|x_t) = p(y_{t+1}|x_t)p(x_{t+1}|x_t; y_{t+1}),$$

where

$$p(y_{t+1}|x_t) = \int p(y_{t+1}|x_{t+1})p(x_{t+1}|x_t)dx_{t+1},$$

$$p(x_{t+1}|x_t; y_{t+1}) = p(y_{t+1}|x_{t+1})p(x_{t+1}|x_t)/p(y_{t+1}|x_t).$$

Typically this fully adapted form is unavailable but when it is the approximating joint density may be chosen to be the true joint. That is,

$$g(y_{t+1}|x_t)g(x_{t+1}|x_t; y_{t+1}) = p(y_{t+1}|x_t)p(x_{t+1}|x_t; y_{t+1}).$$

In this case Step 4 becomes redundant as $\omega_{t+1}^k = 1$, ($\pi_{t+1}^k = 1/M$) and the method reduces to what Pitt and Shephard (2001) call the fully adapted algorithm. **The fully adapted method is the most efficient in estimating the likelihood and is generally the optimal filter a single time step ahead.**

The SIR method of Gordon et al. (1993) arises when the joint proposal is chosen as,

$$g(y_{t+1}|x_t) \times g(x_{t+1}|x_t) = 1 \times p(x_{t+1}|x_t),$$

in which case, $g(y_{t+1}|x_t)$ is constant and $g(x_{t+1}|x_t; y_{t+1}) = p(x_{t+1}|x_t)$. In this case, step (1) above leaves the weights unchanged (as $\pi_{t|t+1}^k = \pi_t^k$).

The goal of the auxiliary particle filter is to get as close to full adaption as possible, when full adaption is not analytically possible. This is achieved by making $g(y_{t+1}|x_t)$ as close to $p(y_{t+1}|x_t)$ as a function of x_t as possible (up to a constant of proportionality) and the density

$g(x_{t+1}|x_t; y_{t+1})$ as close to $p(x_{t+1}|x_t; y_{t+1})$ as possible. Various procedures are found for doing this; see for example, Pitt and Shephard (2001) and Smith and Santos (2006).

The **general ASIR estimator** of $p(y_t|y_{1:t-1})$, introduced and used by Pitt (2002), is

$$\hat{p}^A(y_t|y_{1:t-1}) = \left\{ \sum_{k=1}^M \frac{\omega_t^k}{M} \right\} \left\{ \sum_{k=1}^M \omega_{t-1|t}^k \right\}. \quad (3)$$

The two sets of weights ω_t^k and $\omega_{t-1|t}^k$ are defined above and calculated as part of the ASIR algorithm. These two quantities are again a simple by-product of the algorithm. We define the information in the swarm of particles at time t as $\mathcal{A}_t = \{x_t^k; \pi_t^k\}$. For full adaption $\omega_t^k = 1$ and $\omega_{t-1|t}^k = p(y_t|x_{t-1}^k)/M$ and the first summation in (3) disappears. For the SIR method, $\omega_t^k = p(y_t|x_t^k)$ and $\omega_{t-1|t}^k = \pi_{t-1}^k$ and the second summation in (3) disappears.

The ASIR Algorithm 1 is a flexible particle filter approach when combined with stratification. Theorem 1 establishes that this algorithm together with the estimator of (3) is unbiased. This is important as it enables very efficient likelihood estimators from the ASIR method to be used within an MCMC algorithm.

Theorem 1. *The ASIR likelihood*

$$\hat{p}^A(y_{1:t}) = \hat{p}^A(y_1) \prod_{t=2}^T \hat{p}^A(y_t|y_{1:t-1}) \quad (4)$$

is unbiased in the sense that

$$E(\hat{p}^A(y_{1:t})) = p(y_{1:t})$$

The theorem is proved in Section 7.4.2, Proposition 7.4.1 of Del Moral (2004). We give a more direct and accessible proof in Appendix A.

Our examples use the standard particle filter and the fully adapted particle filter, and the partially adapted particle filter described in Appendix B.1.

Simulated Likelihood The ASIR likelihood estimate 4 is called the simulated likelihood and Theorem 1 shows that the general ASIR particle filter provides a simulated likelihood that is an unbiased estimate of the true likelihood function. Andrieu et al. (2010) show that we can view the simulated likelihood $\hat{p}(y|\theta)$ as the density of y conditional on θ and a set of auxiliary variables u that are not a function of θ and such that $\hat{p}(y|\theta) = p_S(y|\theta, u)$, where the subscript S denotes a simulated likelihood, and

$$\int p_S(y|\theta, u)p(u)du = p(y|\theta). \quad (5)$$

The variables u represent the uniform variates used for the multinomial/statified draws and the random variates (e.g. standard Gaussian) used in simulating from $g(x_{t+1}|x_t, y_{t+1})$. It follows that the posterior $p_S(\theta|y) = p(\theta|y)$ so that a method that simulates from $p_S(\theta, u|y)$ yields iterates from the correct posterior $p(\theta|y)$. We note that the ideas of using a simulated

likelihood for Bayesian inference have also been explored, outside the area of particle filters in the work of Beaumont (2003) and Andrieu and Roberts (2009)

We note that the variance of the log of the simulated likelihood is $O(T/M)$ so it will be necessary to take the number of particles $M = O(T)$ to keep a constant standard deviation as T increases. This implies that the particle filter MCMC algorithm is of order $O(T^2)$ for T large and means that it is important to make the particle filter as efficient as possible.

3 Adaptive sampling for the simulated likelihood

The target density for posterior inference is $p_S(\theta, u|y) \propto p_S(y|\theta, u)p(\theta)p(u)$, where $p(\theta)$ is the prior for θ . It may therefore be possible to use a Metropolis-Hastings simulation method to generate samples from the target density as follows. Suppose that given some initial θ_0 , the $j - 1$ iterates $(\theta_1, u_1), \dots, (\theta_{j-1}, u_{j-1})$ have been generated. The j th iterate, (θ_j, u_j) , is generated from the proposal density $q_j(\theta; \tilde{\theta})p(u)$, which may also depend on some other value of θ which we call $\tilde{\theta}$. Let (θ_j^p, u_j^p) be the proposed value of (θ_j, u_j) generated from $q_j(\theta; \theta_{j-1})p(u)$. Then we take $(\theta_j, u_j) = (\theta_j^p, u_j^p)$ with probability

$$\alpha(\theta_{j-1}, u_{j-1}; \theta_j^p, u_j^p) = \min \left\{ 1, \frac{p_S(y|\theta_j^p, u_j^p)p(\theta_j^p)}{p_S(y|\theta_{j-1}, u_{j-1})p(\theta_{j-1})} \frac{q_j(\theta_{j-1}; \theta_j^p)}{q_j(\theta_j^p; \theta_{j-1})} \right\}, \quad (6)$$

with $p(u_j^p)$ and $p(u_{j-1})$ cancelling out, and take $(\theta_j, u_j) = (\theta_{j-1}, u_{j-1})$ otherwise. We say that the proposal is independent if $q_j(\theta; \tilde{\theta}) = q_j(\theta)$.

In adaptive sampling the parameters of $q_j(\theta; \tilde{\theta})$ are estimated from the iterates $\theta_1, \dots, \theta_{j-2}$. When the likelihood can be measured exactly, i.e., in the non-particle filter case, then it can be shown that under appropriate regularity conditions, the sequence of iterates $\theta_j, j \geq 1$, converges to draws from the target distribution. See Roberts and Rosenthal (2007), Roberts and Rosenthal (2009) and Giordani and Kohn (2010).

Our article uses the adaptive independent Metropolis Hastings scheme of Giordani and Kohn (2010) and the adaptive random walk Metropolis scheme of Roberts and Rosenthal (2009). They are discussed in Appendix C. Appendix C.3 proves that the adaptive independent Metropolis Hastings sampler converges to the target distribution under the given conditions. Both the standard particle filter and the fully adapted particle filter satisfy these conditions almost automatically. The appendix also shows that in the partially adapted case a simple mixture of a partially adapted particle filter and the standard particle filter also satisfies the conditions for convergence.

3.1 Adaptive sampling and parallel computation

Carrying out Bayesian inference using the particle filter and Markov chain Monte Carlo simulation is computationally expensive. However, parallel processing can greatly increase the speed and therefore the areas of application of our methods. For adaptive independent Metropolis Hastings proposals we can use the following three step approach. Let θ^c the current value of θ generated by the sampling scheme and $q_c(\theta)$ the current proposal density for

θ . (a) For each of J processors generate K proposed values of θ , which we write as $\theta_{j,k}^{(p)}$, $k = 1, \dots, K$, and compute the corresponding logs of the ratios $\widehat{p}(y|\theta_{j,k}^{(p)})p(\theta_{j,k}^{(p)})/q(\theta_{j,k}^{(p)})$. (b) After each K block of proposed values is generated for each processor, carry out Metropolis-Hastings selection of the JK proposed $\{\theta_{j,k}^{(p)}\}$ parameters using a single processor to obtain $\{\theta_{j,k}\}$ draws from the chain. This last step is fast because it is only necessary to draw uniform variates. (c) Use the previous iterates and the $\theta_{j,k}$ to update the proposal density $q_c(\theta)$ and θ_c . In our applications of this approach, K is chosen so that KJ is approximately the time between updates of the adaptive independent Metropolis Hastings sampling scheme.

A second approach applies to all Metropolis-Hastings sampling schemes, and in particular to the adaptive random walk Metropolis proposal. Suppose that J processors are available. The likelihood is estimated for a given θ on each of the processors using the particle filter with M particles and these estimates are then averaged to get an estimate of the likelihood based on JM particles. This approach is similar to, but faster, than using a single processor and makes it possible to estimate the likelihood using a large number of particles. However, for a given number of generated particles, the first approach can be shown to be statistically more efficient than the second.

3.2 Estimating the marginal likelihood

Marginal likelihoods are often used to compare two or more models. For a given model, let θ be the vector of model parameters, $p(y|\theta)$ the likelihood of the observations y and $p(\theta)$ the prior for θ . The marginal likelihood is defined as

$$p(y) = \int p(y|\theta)p(\theta)d\theta. \quad (7)$$

which in our case can also be written as

$$p(y) = \int p_S(y|\theta, u)p(\theta)p(u)d\theta du. \quad (8)$$

It is often difficult to evaluate or estimate $p(y)$ in non-Gaussian state space models, although auxiliary variable methods can be used in some problems. See Frühwirth-Schnatter and Wagner (2008). Appendix D briefly outlines how the marginal likelihood can be estimated using bridge or importance sampling, with the computation carried out within the adaptive sampling framework so that a separate simulation run is unnecessary.

4 Comparing the standard SIR particle filter with adapted ASIR particle filters

It is instructive to compare the performance of the standard particle filter with the fully and partially adapted particle filters for different signal to noise ratios and using different numbers of particles. We use two simulated examples. The first example compares the standard particle filter and the fully adapted particle filter for a Gaussian autoregressive signal observed with Gaussian noise. This example is also of interest because we can compute

the exact likelihood using the Kalman filter, which is equivalent to using an infinite number of particles. The second example compares the standard SIR to a partially adapted particle filter using a binomial model where we vary the signal to noise ratio by varying the number of binomial trials.

In both examples, we make the comparison in terms of three criteria. The first is the acceptance rate of the adaptive independent Metropolis Hastings sampler, which we define as the percentage of accepted draws. The second is the inefficiencies of the iterates of the parameters obtained using the adaptive independent Metropolis Hastings method of Gior-dani and Kohn (2010). The third is the standard deviation of the simulated log-likelihood $p_S(y|\theta, u)$ evaluated at the true value of θ , which is a good measure of how close the particle filter likelihood is to the true likelihood $p(y|\theta)$.

We define the inefficiency of the sampling scheme for a given parameter as the variance of the parameter estimate divided by its variance if the sampling scheme generates independent iterates. We estimate the inefficiency factor for a given parameter as $IF = 1 + 2 \sum_{j=1}^{L^*} \hat{\rho}_j$, where $\hat{\rho}_j$ is the estimated autocorrelation of the parameter iterates at lag j . As a rule of thumb, the maximum number of lags L^* that we use is $L^* = \min\{1000, L\}$, where L is the lowest index j such that $|\hat{\rho}_j| < 2/\sqrt{K}$ where K is the sample size used to compute $\hat{\rho}_j$.

4.1 Example 1: Autogressive model observed with noise

Consider the following first order autoregression (AR(1)) plus noise model,

$$y_t|x_t \sim \mathcal{N}(x_t, \sigma^2)$$

$$x_{t+1}|x_t \sim \mathcal{N}(\mu + \phi(x_t - \mu), \tau^2) \tag{9}$$

$$x_0 \sim \mathcal{N}(\mu, \tau^2/(1 - \phi^2)). \tag{10}$$

The prior distributions are $\mu \sim \mathcal{N}(0, 100)$, $\phi \sim \mathcal{U}(0, 1)$, $\sigma^2 \sim \mathcal{IG}(0.1, 0.1)$ and $\tau^2 \sim \mathcal{IG}(0.1, 0.1)$. The notation $\mathcal{N}(a, b^2)$ means a normal distribution with mean a and variance b^2 , $\mathcal{U}(a, b)$ means a uniform distribution on (a, b) and $\mathcal{IG}(a, b)$ means an inverse gamma distribution with shape parameter a and scale b .

Our simulation study uses 50 replicated data sets with 500 observations each, generated by setting $\mu = 0$, $\phi = 0.6$, $\tau^2 = 1$, $x_0 \sim \mathcal{N}(\mu, \tau^2/(1 - \phi^2))$ and two values for $\sigma^2 = \{0.01, 1.0\}$, corresponding to high and low signal to noise ratios. We ran 30 000 iterations of the adaptive independent Metropolis-Hastings for the posterior distribution using the standard particle filter and the fully adapted particle filter with differing number of particles. For completeness, we also ran the adaptive sampling scheme with the Kalman filter using the exact likelihood. The update times for the adaptive independent Metropolis Hastings were at iterations 100, 200, 500, 1000, 1500, 2000, 3000, 4000, 5000, 10000, 15000 and 20000. We initialized the AIMH based on a normal proposal formed from 5000 draws of a previous run of the ARWM and the Kalman filter. For each signal to noise ratio we report the median MCMC parameter inefficiencies over the 50 replications as well as the interquartile range of the inefficiencies for differing numbers of particles. We also report results on the standard deviation of the simulated log likelihood for the standard particle filter and the fully adapted particle filter. Specifically, for each of the 50 replicated data sets we computed the log likelihood at the true parameter values 1000 times for each of the two particle filters and obtained the median and

the interquartile range of the medians and standard deviations of the log likelihoods across the 50 data replicates.

Results for the high signal to noise case Tables 1 and 2 report the results for the high signal to noise case with $\sigma^2 = 0.01$. Table 1 shows that the median variance of the simulated log likelihood $\log p_S(y|\theta, u)$ at the true parameter values for the standard particle filter with 2000 particles is over 400 times higher than the median variance of the fully adapted particle filter using 100 particles. This suggests that to get the same standard deviation for the simulated likelihood we would need approximately 8000 times as many particles for the standard particle filter as for the fully adapted particle filter as we know that the variance decreases approximately in inverse proportion to the number of particles.

Table 2 shows that the median acceptance rate of the adaptive independent Metropolis Hastings sampler using the fully adapted particle filter with 100 particles is about 1.75 times higher than the median acceptance rate of the standard particle filter using 4000 particles. The table also shows that the median parameter inefficiencies are about 3 times higher for the standard particle filter using 4000 particles than for the fully adapted particle filter using 100 particles. Finally, the table shows that the fully adapted particle filter using 500 particles performs almost as well as using the exact likelihood.

Table 1: AR(1) + noise. High signal to noise. Medians and interquantile ranges (IQR) of the estimated medians and standard deviations of the log of the simulated likelihood function at the true value for 50 different data sets.

N. Particles	Median		Standard Deviation	
	Median	IQR	Median	IQR
Standard Particle Filter				
100	-839.12	83.34	44.0381	21.0316
500	-729.43	26.09	10.5420	8.2875
1000	-719.10	20.13	5.5507	5.2818
2000	-714.95	18.16	2.8977	2.4716
Fully Adapted Particle Filter				
100	-711.69	17.74	0.1431	0.0160

Results for the low signal to noise case Tables 3 and 4 report the results for the low signal to noise case with $\sigma^2 = 1.0$. Table 3 shows that the median variance of the log of the simulated likelihood at the true parameter values for the standard particle filter using 1000 particles is about the same as the median variance of the simulated likelihood for the fully adapted particle filter using 100 particles, i.e. the variance of the simulated log likelihood of the standard particle filter is about 10 times that of the fully adapted particle filter for the same number of particles.

Table 4 shows that the median acceptance rates and parameter inefficiencies of the adaptive independent Metropolis Hastings sampler using the fully adapted particle filter with 100 particles are about the those of the standard particle filter using 4000 particles.

Table 2: AR(1) + noise. High signal to noise. Medians and interquartile range (IQR) of the acceptance rates and the inefficiencies over 50 replications of the autoregressive model using different particle filters and adaptive independent Metropolis-Hastings.

N. Particles	Ac. Rate		σ^2		τ^2		μ		ϕ	
	Median	IQR	Median	IQR	Median	IQR	Median	IQR	Median	IQR
	Kalman Filter									
	72.18	4.89	1.93	0.34	1.85	0.35	1.76	0.20	1.83	0.24
	Standard Particle Filter									
500	0.05	0.78	836.65	1727.31	1102.87	1732.95	1058.80	1733.35	979.64	1729.08
1000	9.04	5.75	70.25	35.10	63.76	30.98	59.76	58.41	64.64	42.55
2000	21.81	10.87	21.84	20.64	22.97	17.62	20.48	23.16	25.09	24.07
4000	33.27	9.19	9.33	6.33	9.11	7.20	9.66	6.58	8.95	8.13
	Fully Adapted Particle Filter									
100	58.83	3.04	3.02	0.56	2.91	0.66	2.63	0.43	2.80	0.45
500	67.64	2.30	2.23	0.31	2.08	0.31	2.02	0.20	2.10	0.24

Table 3: AR(1) + noise. Low signal to noise. Medians and interquantile ranges (IQR) of the estimated medians and standard deviations of the log of the simulated likelihood function at the true value for 50 different data sets.

N. Particles	Median		Standard Deviation	
	Median	IQR	Median	IQR
	Standard Particle Filter			
100	-904.0827	19.0013	2.4479	0.2372
500	-901.7877	18.8020	1.0793	0.1170
1000	-901.4966	18.7909	0.7629	0.0550
	Fully Adapted Particle Filter			
100	-901.4727	18.8540	0.7057	0.0398

Table 4: AR(1) + noise. Low signal to noise. Medians and interquartile range (IQR) of the acceptance rates and the inefficiencies over 50 replications of the autoregressive model using different particle filters and adaptive independent Metropolis-Hastings.

N. Particles	Ac. Rate		σ^2		τ^2		μ		ϕ	
	Median	IQR	Median	IQR	Median	IQR	Median	IQR	Median	IQR
	Kalman Filter									
	73.81	2.26	2.04	0.38	2.12	0.46	1.73	0.12	1.94	0.25
	Standard Particle Filter									
500	6.84	6.28	92.85	76.67	95.95	71.78	71.63	51.26	85.93	68.60
1000	29.86	18.15	19.67	27.30	21.10	24.91	10.87	16.07	18.18	28.90
2000	42.47	13.37	11.87	11.20	10.72	8.98	5.36	3.62	9.02	6.88
4000	52.95	11.85	6.38	6.86	6.16	4.53	3.28	2.54	5.02	4.11
	Fully Adapted Particle Filter									
100	53.94	9.24	3.48	0.88	3.53	0.84	3.27	1.32	3.62	0.92

4.2 Example 2: Binomial model with an autoregressive state equation

Consider observations generated from the following dynamic binomial model

$$y_t \sim \text{Bin}(m, \pi_t), \quad \pi_t = \exp(x_t)/(1 + \exp(x_t))$$

where m is the number of trials and π_t is the probability of success of each trial. The states x_t follow the first order AR(1) model (9) whose initial distribution is (10). The prior distributions for the parameters are $\mu \sim \mathcal{N}(0, 100)$, $\phi \sim \mathcal{U}(0, 1)$ and $\tau^2 \sim \mathcal{HN}(100)$. We use the notation $\mathcal{HN}(b^2)$ to mean a half-normal distribution with scale b .

Our simulation study is organized similarly to that in Section 4.1. The data generating process takes $\mu = 0$, $\phi = 0.97$, $\tau^2 = 0.25$ and the number of trials takes the two values $m = \{100, 500\}$. We initialized the AIMH based on a normal proposal formed from 5000 draws of a previous run of the ARWM and the partially adapted particle filter using 500 particles.

We note the following about the binomial density.

1. The standard particle filter will do worse as the number of trials m increases because the the measurement density becomes more informative and peaked so the variance of the weights increases.
2. The opposite is true for the partially adapted particle filter method. The measurement density $p(y_t|x_t)$ tends to normality as m increases by the central limit theorem, so that the partially adapted particle filter tends to a fully adapted particle filter. Hence the partially adapted particle filter method actually improves as m becomes larger and this is seen in Tables 5 and 7 below.

High signal to noise case Tables 5 and 6 report the results for the high signal to noise case, with the number of trials set at $m = 500$. Table 5 shows that the variance of the simulated log likelihood at the true parameter values for the standard particle filter with 4000 particles is about 2.5 times higher than that of the partially adapted particle filter when 100 particles are used. This means that it is necessary to have 100 times as many particles using the standard particle filter to get the same noise level for the simulated log likelihood as for the partially adapted particle filter.

Table 6 shows that the median acceptance rate of the adaptive independent Metropolis Hastings sampler using the partially adapted particle filter with 100 particles is about 1.3 times higher than the median acceptance rate of standard particle filter using 4000 particles. The table also shows that the median parameter inefficiencies are 1.5 times higher for the standard particle filter using 4000 particles than they are for the partially adapted particle filter using 100 particles.

Low signal to noise case Tables 7 and 8 report the results for the low signal to noise case, with the number of trials set at $m = 100$. Table 7 shows that the median variance of simulated log likelihood at the true parameter values for the standard particle filter with 4000 particles is about the same as that obtained by the partially adapted particle filter using

Table 5: Binomial example. High signal to noise. Medians and interquantile ranges (IQR) of the estimated medians and standard deviations of the log-likelihood function at the true value for 50 different data sets.

N. Particles	Median		Standard Deviation	
	Median	IQR	Median	IQR
Standard Particle Filter				
500	-2441.29	114.63	3.3149	1.5044
1000	-2439.64	114.39	2.0737	0.8711
2000	-2438.88	117.55	1.4106	0.3203
4000	-2438.45	118.61	0.9478	0.1593
Partially Adapted Particle Filter				
100	-2438.28	119.38	0.6182	0.1358

Table 6: Binomial example. High signal to noise. Medians and interquartile range (IQR) of the acceptance rates and the inefficiencies over 50 replications of the binomial model using different particle filters and adaptive independent Metropolis-Hastings.

N. Particles	Ac. Rate		μ		logit(ϕ)		log(τ^2)	
	Median	IQR	Median	IQR	Median	IQR	Median	IQR
Standard Particle Filter								
500	3.66	2.57	92.85	99.02	100.32	369.21	105.76	330.84
1000	13.28	4.13	39.84	47.49	38.77	32.71	40.89	23.31
2000	27.18	5.54	12.26	7.83	12.15	7.61	13.29	10.00
4000	39.73	5.85	6.01	3.86	5.93	3.00	5.82	2.14
Partially Adapted Particle Filter								
100	51.52	11.17	3.82	3.06	3.74	2.09	3.39	1.35

250 particles, i.e. the standard particle filter requires about 16 times as many particles to obtain the same standard deviation as the partially adapted particle filter.

Table 8 shows that the median acceptance rate of the adaptive independent Metropolis Hastings sampler using the partially adapted particle filter with 200 particles is higher than the median acceptance rate of standard particle filter using 2000 particles. The table also shows that the median parameter inefficiencies for the standard particle filter using 2000 particles are higher than the median inefficiencies for the partially adapted particle filter using 200 particles.

5 Performance of the adaptive sampling schemes on real examples

This section uses real data to illustrate the flexibility and wide applicability of the approach that combines particle filtering with adaptive sampling. All that is necessary for model estimation and model comparison by marginal likelihood is to code up a particle filter to evaluate the simulated likelihood and to code up the prior on the parameters. We also illustrate the difference in performance between the adaptive random walk Metropolis sampling scheme of Roberts and Rosenthal (2009) and that the adaptive independent Metropolis Hastings scheme of Giordani and Kohn (2010). This comparison is interesting for two reasons. First,

Table 7: Binomial example. Low signal to noise. Medians and interquantile ranges (IQR) of the estimated medians and standard deviations of the log-likelihood function at the true value for 50 different data sets.

N. Particles	Median		Standard Deviation	
	Median	IQR	Median	IQR
Standard Particle Filter				
500	-1729.05	109.49	1.8385	0.2960
1000	-1728.16	109.63	1.2704	0.2457
2000	-1727.69	109.72	0.8827	0.1458
4000	-1727.50	109.82	0.6300	0.0711
Partially Adapted Particle Filter				
100	-1727.81	109.84	0.9867	0.1074
200	-1727.55	109.85	0.7132	0.0810
500	-1727.41	109.90	0.4465	0.0550

Table 8: Binomial example. Low signal to noise. Medians and interquartile range (IQR) of the acceptance rates and the inefficiencies over 50 replications of the binomial model using different particle filters and adaptive independent Metropolis-Hastings.

N. Particles	Ac. Rate		μ		logit(ϕ)		log(τ^2)	
	Median	IQR	Median	IQR	Median	IQR	Median	IQR
Standard Particle Filter								
500	17.02	4.82	29.37	17.65	31.85	16.33	31.57	15.90
1000	31.01	5.46	10.85	7.54	10.51	7.14	10.33	4.73
2000	43.47	5.91	5.27	3.30	5.17	2.10	5.00	1.79
4000	54.02	4.91	3.70	3.35	3.48	1.94	3.03	0.58
Partially Adapted Particle Filter								
100	39.16	7.15	5.06	3.60	5.99	2.63	5.96	2.50
200	50.59	5.19	3.62	2.26	3.81	2.02	3.59	1.21
500	60.80	4.19	2.80	3.32	2.47	1.23	2.32	0.32

the adaptive independent Metropolis Hastings scheme tries to obtain a good approximation to the posterior density, whereas the adaptive random walk Metropolis aims for some target acceptance rate. Second, we claim that any independent Metropolis Hastings scheme (of which the adaptive independent Metropolis Hastings scheme of Giordani and Kohn (2010) is an example), is more suitable to be implemented in parallel than a Metropolis-Hastings scheme with a proposal that depends on the previous iterate (of which the adaptive random walk Metropolis scheme of Roberts and Rosenthal (2009) is an example).

The comparison between the two schemes is in terms of three criteria. The first two are the acceptance rate of the Metropolis-Hastings method and the inefficiency factors (IF) of the parameters and are defined in Section 4; they are independent of the way the algorithms are implemented. However, these two criteria do not take into account the times taken by the samplers. To obtain an overall measure of the effectiveness of a sampler, we define its equivalent computing time $ECT = 1000 \times IF \times t$, where t is the time per iteration of the sampler. We interpret ECT as the time taken by the sampler to attain the same accuracy as that attained by 1000 independent draws of the same sampler. For two samplers a and b , ECT_a/ECT_b is the ratio of times taken by them to achieve the same accuracy. We note that

the time per iteration for a given sampling algorithm depends on how it is implemented, i.e. the language used, whether operations are vectorized, etc.

The results presented are for a single processor and the two parallel methods discussed in Section 3.1. To simplify the presentation, we mainly present results for the standard particle filter.

5.1 Example 1: Stochastic volatility model with leverage and outliers

The first example considers the univariate stochastic volatility (SV) model

$$\begin{aligned} y_t &= K_t \exp(x_t/2) \varepsilon_t, & \varepsilon_t &\sim \mathcal{N}(0, 1) \\ x_{t+1} &= \mu + \phi(x_t - \mu) + \sigma_\eta \eta_t, & \eta_t &\sim \mathcal{N}(0, 1) \end{aligned} \quad (11)$$

where $\text{corr}(\varepsilon_t, \eta_t) = \rho$, $\Pr(K_t = 2.5) = \omega$ and $\Pr(K_t = 1) = 1 - \omega$, with $\omega \ll 1$. This is a state space model with a non-Gaussian observation equation and a Gaussian state transition equation for the latent volatility x_t which follows a first order autoregressive model. The SV model allows for leverage because the errors in the observation and state transition equations can be correlated. The model also allows for outliers in the observation equation because the standard deviation of y_t given x_t can be 2.5 its usual size when $K_t = 2.5$. To complete the model specification, we assume that all parameters are independent a priori with the following prior distributions: $\mu \sim \mathcal{N}(0, 10^2)$, $\phi \sim \mathcal{TN}_{(0,1)}(0.9, 0.1)$, $\sigma_\eta^2 \sim \mathcal{IG}(0.01, 0.01)$, and $\rho \sim \mathcal{TN}_{(-1,1)}(0, 10^6)$. We use the notation $\mathcal{TN}_{(c,d)}(a, b)$ to mean a truncated normal with location a and scale b restricted to the interval (c, d) and $\mathcal{IG}(a, b)$ is an inverse gamma distribution with shape parameter a , scale parameter b and mode $b/(a+1)$. We set $\omega = 0.03$ in the general model to indicate that outliers are rare apriori.

Shephard (2005) reviews SV models and a model of the form (11) is estimated by Malik and Pitt (2008) by maximum likelihood using the smooth particle filter.

S&P 500 index We apply the SV model (11) to the Standard and Poors (S&P) 500 data from 02/Jan/1970 to 14/Dec/1973 obtained from Yahoo Finance web site¹. The data consists of $T = 1\,000$ observations.

Table 9 shows the acceptance rates, the inefficiencies and the equivalent computing time over 10 replications of the stochastic volatility model using the standard particle filter and the two adaptive Metropolis-Hastings schemes. The analysis uses the SV model without leverage or outliers. In the table, SP stands for a single processor, MP₁ for multiprocessor method 1 and MP₂ for multiprocessor method 2 (where the simulated likelihood is obtained as an average) described in Section 3.1. We use eight processors for both the MP₁ and MP₂ schemes. The basic number of particles in this example is $K = 500$, which means that SP uses 4000 particles in a single processor, MP₁ uses 4000 particles in each processor and MP₂ uses 500 particles in each processor. We ran all the algorithm for 10000 iterations and took the last 5000 to compute the results. The equivalent computing time is obtained by taking the overall time divided by the number of iterations times the inefficiency times 1000. The

¹ <http://au.finance.yahoo.com/q/hp?s=GSPC>

update times for the adaptive independent Metropolis Hastings using SP or MP₂ were at 100, 200, 500, 1000, 2000, 3000, 4000, 5000, 6000 and 7500. The block sizes (also the update times) for the adaptive Metropolis-Hastings MP₁ were 15, 25, 60, 125, 250, 375, 500, 625, 750 and 940.

The table shows that the acceptance rates of the adaptive independent Metropolis Hastings sampler are about twice those of the adaptive random walk Metropolis and the inefficiencies are about 1/6 of those for the adaptive random walk Metropolis. For ECT, the best approach for adaptive independent Metropolis Hastings is MP₁ and is between 11 and 18 times better than the best available approach (which is MP₂) for adaptive random walk Metropolis. Qualitatively similar results were obtained for $K = 1000$ particles.

Table 9: SV model. Medians and interquartile range (between brackets) of the acceptance rates, the inefficiencies and the equivalent computing time ($t \times \text{inefficiency} \times 1000$) over 10 replications of the stochastic volatility model using the standard particle filter and differing adaptive Metropolis-Hastings schemes. SP = single processor, MP₁ = multiprocessor Metropolis-Hastings and MP₂ = multiprocessor averaging the likelihood function.

Algorithm	Ac. Rate	Inefficiency			Equivalent Computing Time		
		μ	$\text{logit}(\phi)$	$\log(\sigma_\eta^2)$	μ	$\text{logit}(\phi)$	$\log(\sigma_\eta^2)$
ARWM-SP	24.5 (0.6)	25.47 (16.33)	30.20 (5.12)	20.00 (2.80)	13463.7 (8589.8)	15972.9 (2655.3)	10603.0 (1514.0)
ARWM-MP ₂	22.2 (5.5)	25.04 (12.96)	31.07 (14.02)	20.51 (8.83)	3594.5 (1886.9)	4453.4 (1994.3)	2930.4 (1284.3)
AIMH-SP	51.6 (0.8)	6.45 (7.12)	3.46 (0.83)	3.08 (0.07)	3430.9 (3777.9)	1841.7 (439.7)	1638.7 (39.9)
AIMH-MP ₁	52.3 (3.3)	3.44 (2.13)	3.42 (0.76)	3.63 (0.50)	237.8 (147.4)	235.3 (50.5)	250.0 (34.6)
AIMH-MP ₂	53.6 (3.2)	4.40 (6.74)	3.52 (3.37)	3.15 (0.56)	627.1 (975.4)	504.4 (482.5)	451.1 (75.8)

Model selection We now use importance sampling and bridge sampling to compute the marginal likelihoods of the four SV models: the model with no leverage effect ($\rho = 0$) and no outlier effect ($\omega = 0$), the model that allows for leverage but not outliers, the model that allows for outliers but no leverage and the general model that allows for both outliers and leverage. Table 10 shows the logarithms of the marginal likelihoods of the four models for a single run of each algorithm. The differences between the two approaches are very small. In this example, and based on our prior distributions, the SV model with leverage effects but no outliers has the highest marginal likelihood.

Posterior estimates of model parameters The estimated posterior means and standard deviations of all four models are given in Table 11.

5.2 Example 2: GARCH model observed with noise

The GARCH(1,1) model is used extensively to model financial returns, see for example, Bollerslev et al. (1994). In this section we consider the GARCH(1,1) model observed with

Table 10: Logarithms of the marginal likelihoods for four different SV models for the two particle filter algorithms computed using the adaptive independent Metropolis Hastings algorithm. *BS* and *IS* mean bridge sampling and importance sampling.

Model	Standard Particle Filter		Partially Adapted Particle Filter	
	$\log(p_{BS}(y))$	$\log(p_{IS}(y))$	$\log(p_{BS}(y))$	$\log(p_{IS}(y))$
SV	-1072.9	-1072.9	-1072.9	-1072.9
SV Lev.	-1065.0	-1065.0	-1065.0	-1065.0
SV Out.	-1076.6	-1076.6	-1076.5	-1076.4
SV Lev. Out.	-1069.3	-1069.3	-1069.2	-1069.3

Table 11: S&P 500 data: Estimated posterior means and standard deviations for all four stochastic volatility models.

Parameter	SV		SV Lev.		SV Out.		SV Lev. Out	
	Mean	S. Dev.	Mean	S. Dev.	Mean	S. Dev.	Mean	S. Dev.
μ	-0.4329	1.2314	-0.5642	0.1500	-0.1786	2.2812	-0.5756	0.3497
ϕ	0.9879	0.0097	0.9811	0.0063	0.9907	0.0086	0.9830	0.0065
τ^2	0.0142	0.0068	0.0106	0.0037	0.0116	0.0053	0.0091	0.0034
ρ	—	—	-0.7608	0.0960	—	—	-0.7652	0.0960

Gaussian noise which is a more flexible version of the basic model. The model is,

$$\begin{aligned}
y_t | x_t &\sim \mathcal{N}(x_t, \tau^2) \\
x_{t+1} | \sigma_{t+1}^2 &\sim \mathcal{N}(0, \sigma_{t+1}^2) \\
\sigma_{t+1}^2 &= \alpha + \beta x_t^2 + \gamma \sigma_t^2 \\
x_0 &\sim \mathcal{N}(0, \alpha / (1 - \beta - \gamma)).
\end{aligned}$$

The priors on τ^2 and α are $\tau^2 \sim \mathcal{HN}(100)$ and $\alpha \sim \mathcal{HN}(100)$. The joint prior for β and γ is uniform in the region $\beta > 0, \gamma > 0, \beta + \gamma < 1$.

It is straightforward to show that this model is fully adapted. See Appendix B.2 Instead of using the GARCH(1,1) model with noise we can use other members of the GARCH family, e.g. an EGARCH process observed with noise. All such models are fully adapted.

MSCI UK index returns We model the weekly MSCI UK index returns from 6 January 2000 to 28 January 2010 corresponding to 526 weekly observations shown in Figure 1.

Table 12 compares different adaptive sampling algorithms and particles filters in terms of acceptance rates, inefficiencies and equivalent computing times. Medians and interquantile ranges are computed using 50 replication of each adaptive sampling, particle filter and number of particles. The adaptive sampling algorithms were run for 30 000 iterations. The first 20 000 draws were discarded and the remainder used to compute the statistics. The update times for the adaptive independent Metropolis-Hastings were at 100, 200, 500, 1000, 1500, 2000, 3000, 4000, 5000, 10000, 15000 and 20000. The initial value and proposal distribution for all algorithms were based on a single short run of the adaptive random walk and all results are for a single processor.

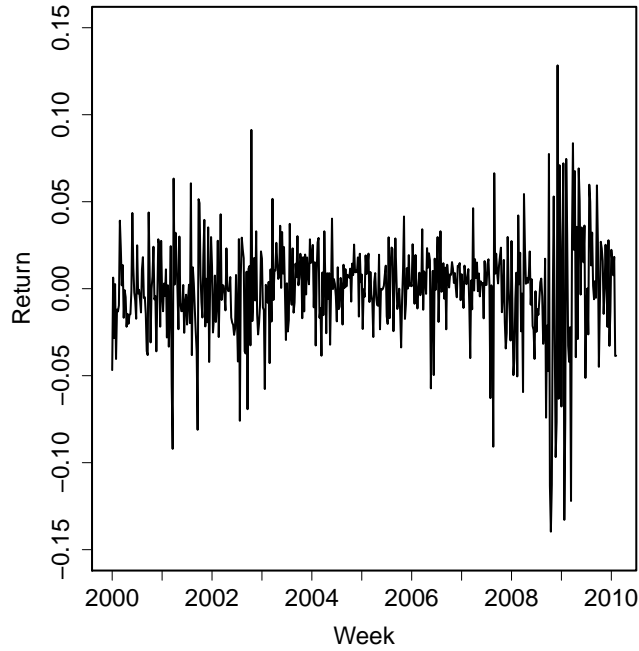


Figure 1: UK MSCI weekly returns from 6 January 2000 to 28 January 2010.

The table shows that the fully adapted particle filter is much more efficient than the standard particle filter for both the adaptive random walk Metropolis and the adaptive independent Metropolis Hastings samplers, and that the adaptive independent Metropolis Hastings sampler is much more efficient than the adaptive random walk Metropolis sampler for both the standard and the fully adapted particle filters. In particular, the table shows that adaptive independent Metropolis Hastings combined with full adaptation using 200 particles is about four times as efficient as the adaptive random walk Metropolis sampler using the standard particle filter and 10, 000 particles.

Table 13 shows the standard deviations of the simulated log-likelihood function for the particle filters using differing number of particles. The statistics are based on 1000 replications of the particle filter with the parameters fixed at their posterior means. The summary statistics of the posterior distribution is shown in Table 14. The table shows that the standard deviation of the simulated log likelihood using the fully adapted particle filter with 500 particles is smaller than the standard deviation of the simulated log likelihood using 10, 000 particles.

Table 14 is a summary of the posterior distributions of the four parameters.

Acknowledgment

The research of Robert Kohn and Ralph S. Silva was partially supported by an ARC Discovery Grant DP0667069

Table 12: Medians and interquartile range (between brackets) of the acceptance rates, the inefficiencies and the equivalent computing time for 50 replications of the Gaussian GARCH model observed with noise applied to the UK index return using differing particle filters, number of particles and adaptive Metropolis-Hastings algorithms. A single processor was used for all results.

Algorithm	# of Particles	Accept. Rate	Inefficiency			
			τ^2	α	β	γ
	Standard Particle Filter					
ARWM	1000	6.90 (0.99)	82.75 (15.40)	93.79 (19.02)	89.87 (18.98)	90.84 (24.55)
	5000	14.72 (1.07)	39.36 (5.96)	48.69 (15.40)	56.63 (11.01)	53.00 (14.48)
	10000	17.11 (1.35)	37.93 (11.42)	43.27 (6.85)	50.71 (9.73)	47.87 (16.48)
AIMH	1000	13.25 (1.94)	40.47 (24.91)	46.70 (26.91)	44.14 (14.51)	42.32 (20.83)
	5000	29.97 (2.23)	8.81 (3.49)	10.87 (2.94)	11.68 (2.94)	11.82 (3.48)
	10000	33.64 (1.51)	7.42 (1.46)	9.24 (2.25)	9.08 (3.26)	10.07 (6.39)
	Fully Adapted Particle Filter					
ARWM	200	15.35 (1.04)	43.82 (8.54)	48.42 (10.54)	56.42 (10.07)	49.56 (15.87)
	500	17.45 (1.23)	36.52 (7.25)	43.21 (12.05)	52.31 (13.67)	48.06 (15.76)
	1000	18.26 (1.22)	34.06 (5.88)	40.35 (7.70)	51.91 (7.38)	45.56 (17.74)
AIMH	200	30.28 (1.61)	10.16 (4.44)	11.72 (4.39)	11.34 (3.46)	10.44 (3.33)
	500	34.44 (1.38)	7.76 (1.97)	8.68 (3.27)	8.53 (2.02)	10.17 2 (5.40)
	1000	36.19 (1.54)	6.61 (2.05)	9.51 (3.68)	7.90 (1.66)	8.01 (5.14)

Table 13: UK MSCI index returns: Standard deviation of the simulated log-likelihood function at the posterior mean for standard particle filter and fully adapted particle filter using various numbers of particles and 1000 replications.

# Particles	SIR sd	# Particles	FAPF sd
1000	1.73	200	0.67
5000	0.71	500	0.43
10000	0.51	1000	0.31

References

- Andrieu, C. and Doucet, A. (2002), “Particle filtering for partially observed Gaussian state space models,” *Journal of the Royal Statistical Society, Series B*, 64, 827–836.
- Andrieu, C., Doucet, A., and Holenstein, R. (2010), “Particle Markov chain Monte Carlo

Table 14: Summary of statistics of the posterior distribution.

Parameter	Mean	St.Dev.
τ^2	0.0002700	0.0000462
α	0.0000495	0.0000289
β	0.8927539	0.0672126
γ	0.0377854	0.0412842

methods,” *Journal of the Royal Statistical Society, Series B*, 72, 1–33.

Andrieu, C. and Roberts, G. (2009), “The pseudo-marginal approach for efficient Monte Carlo computations,” *The Annals of Statistics*, 37, 697–725.

Atchadé, Y. and Rosenthal, J. (2005), “On adaptive Markov chain Monte Carlo algorithms.” *Bernoulli*, 11, 815–828.

Beaumont, M. (2003), “Estimation of population growth or decline in genetically monitored populations,” *Genetics*, 164, 1139.

Bollerslev, T., Engle, R. F., and Nelson, D. (1994), “ARCH models,” in *HANDBOOK OF ECONOMETRICS*, eds. Engle, R. and McFadden, D., Amsterdam: Elsevier, vol. 4, chap. 49, pp. 2959–3038.

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

Carpenter, J. R., Clifford, P., and Fearnhead, P. (1999), “An improved particle filter for non-linear problems,” *IEE Proceedings on Radar, Sonar and Navigation*, 146, 2–7.

Chen, M. H. and Shao, Q. M. (1997), “On Monte Carlo methods for estimating ratios of normalizing constants,” *The Annals of Statistics*, 25, 1563–1594.

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

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

Doucet, A., de Freitas, N., and Gordon, N. (2001), *Sequential Monte Carlo Methods in Practice*, New York: Springer.

Doucet, A., Godsill, S., and Andrieu, C. (2000), “On sequential Monte Carlo sampling methods for Bayesian filtering,” *Statistics and Computing*, 10, 197–208.

Fearnhead, P. and Clifford, P. (2003), “On-line inference for hidden Markov models via particle filters,” *Journal of the Royal Statistical Society Series B*, 65, 887–899.

Fernández-Villaverde, J. and Rubio-Ramírez, J. (2007), “Estimating macroeconomic models: a likelihood approach,” *Review of Economic Studies*, 74, 1059–1087.

- Flury, T. and Shephard, N. (2008), “Bayesian inference based only on simulated likelihood: particle filter analysis of dynamic economic models,” <http://www.economics.ox.ac.uk/-Research/wp/pdf/paper413.pdf>.
- Frühwirth-Schnatter, S. and Wagner, H. (2006), “Auxiliary mixture sampling for parameter-driven models of time series of counts with applications to state space modelling,” *Biometrika*, 93, 827–841.
- Frühwirth-Schnatter, S. and Wagner, H. (2008), “Marginal likelihoods for non-Gaussian models using auxiliary mixture sampling,” *Computational Statistics and Data Analysis*, 52, 4608–4624.
- Geweke, J. (1989), “Bayesian inference in econometric models using Monte Carlo integration,” *Econometrica*, 57, 1317–1339.
- Giordani, P. and Kohn, R. (2010), “Adaptive Independent Metropolis-Hastings by Fast Estimation of Mixture of Normals,” *Journal of Computational and Graphical Statistics*, see <http://pubs.amstat.org/doi/abs/10.1198/jcgs.2009.07174>.
- Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993), “A novel approach to non-linear and non-Gaussian Bayesian state estimation,” *Radar and Signal Processing, IEE Proceedings F*, 140, 107–113.
- Haario, H., Saksman, E., and Tamminen, J. (2001), “An adaptive Metropolis algorithm,” *Bernoulli*, 7, 223–242.
- Kim, S., Shephard, N., and Chib, S. (1998), “Stochastic volatility: Likelihood inference and comparison with ARCH models,” *Review of Economic Studies*, 65, 361–393.
- Kitagawa, G. (1996), “Monte Carlo filter and smoother for non-Gaussian non-linear state space models,” *Journal of Computational and Graphics Statistics*, 5, 1–25.
- Liu, J. S. and Chen, R. (1998), “Sequential Monte Carlo methods for dynamic systems,” *Journal of the American Statistical Association*, 93, 1032–1044.
- Malik, S. and Pitt, M. K. (2008), “Modeling Stochastic Volatility with Leverage and Jumps: A ‘Smooth’ Particle Filtering Approach,” Available at <http://www.riksbank.com/upload/-Research/Conferences/StateSpace2008/Pitt.pdf>.
- Meng, X. L. and Wong, W. H. (1996), “Simulating ratios of normalizing constants via a simple identity: A theoretical exploration,” *Statistica Sinica*, 6, 831–860.
- Pitt, M. and Shephard, N. (1999), “Filtering via simulation: auxiliary particle filter,” 94, 590–599.
- (2001), “Auxiliary variable based particle filters,” in *Sequential Monte Carlo Methods in Practice*, eds. de Freitas, N., Doucet, A., and Gordon, N. J., New York: Springer-Verlag, pp. 273–293.
- Pitt, M. K. (2002), “Smooth particle filters for likelihood evaluation and maximization,” .

- Polson, N. G., Stroud, J. R., and Müller, P. (2008), “Practical filtering with sequential parameter learning.” *Journal of the Royal Statistical Society, Series B*, 70, 413–428.
- Roberts, G. O., Gelman, A., and Gilks, W. R. (1997), “Weak convergence and optimal scaling of random walk Metropolis algorithms.” *Annals of Applied Probability*, 7, 110–120.
- Roberts, G. O. and Rosenthal, J. S. (2007), “Coupling and ergodicity of adaptive MCMC,” *Journal of Applied Probability*, 44, 458–475.
- (2009), “Examples of adaptive MCMC,” *Journal of Computational and Graphical Statistics*, 18, 349–367.
- Shephard, N. (2005), *Stochastic Volatility: Selected Readings*, Oxford: Oxford University Press.
- Shephard, N. and Pitt, M. (1997), “Likelihood analysis of non-Gaussian measurement time series,” 84, 653–667.
- Silva, R., Giordani, P., Kohn, R., and Pitt, M. (2009), “Particle filtering within adaptive Metropolis Hastings sampling,” [Http://arxiv.org/abs/0911.0230](http://arxiv.org/abs/0911.0230).
- Smith, J. and Santos, A. (2006), “Second-Order Filter Distribution Approximations for Financial Time Series With Extreme Outliers,” *Journal of Business and Economic Statistics*, 24, 329–337.
- Storvik, G. (2002), “Particle filters for state-space models with the presence of unknown static parameters,” *IEEE Transactions on Signal Processing*, 50, 281–290.
- West, M. and Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models*, New York: Springer-Verlag, 2nd ed.

Appendices

A Proof that the AISR likelihood is unbiased

This appendix proves Theorem 1 using an iterated expectations argument on the simulated likelihood. A similar result is obtained in Proposition 7.4.1 of Section 7.4.2 by Del Moral (2004) by showing that the difference of the measure on the states induced by the particle filter and that of the limiting Feynman-Kac measure is a martingale. We believe that our proof which deals specifically with the unbiasedness of the simulated likelihood is simpler and more direct and is accessible to a much wider range of readers.

Before giving the proof we define some terms that are used in Algorithm 1,

$$\begin{aligned}
\hat{p}_M^A(x_t|y_{1:t}) &= \sum_{k=1}^M \pi_t^k \delta(x_t - x_t^k), \text{ where } \pi_t^k \text{ is given in Step (4).} \\
\hat{g}_M^A(x_t|y_{1:t+1}) &= \sum_{k=1}^M \pi_{t|t+1}^k \delta(x_t - x_t^k), \text{ where } x_t^k \sim \hat{p}_M^A(x_t|y_{1:t}), \\
\hat{g}_M^A(x_t|y_{1:t}) &= \int g(x_t|\tilde{x}_{t-1}; y_t) \hat{g}_M^A(\tilde{x}_{t-1}|y_{1:t}) d\tilde{x}_{t-1}, \\
\omega_{t|t+1}(x_t) &= g(y_{t+1}|x_t) \pi_t, \quad \omega_{t+1}(x_{t+1}; x_t) = \frac{p(y_{t+1}|x_{t+1})p(x_{t+1}|x_t)}{g(y_{t+1}|x_t)g(x_{t+1}|x_t, y_{t+1})}
\end{aligned}$$

The term $\hat{p}_M^A(x_t|y_{1:t})$ is the empirical filtering density arising from step 4 of Algorithm 1. The second term $\hat{g}_M^A(x_t|y_{1:t+1})$, is the empirical “look ahead” approximation drawn from in step 2. The expression $\hat{g}_M^A(x_t|y_{1:t})$ is the filtering approximation which we draw from in step 3 (integrating over step 2). Furthermore, we have that in Algorithm 1, $\omega_{t|t+1}^k = \omega_{t|t+1}(x_t^k) \pi_t^k$ and $\omega_{t+1}^k = \omega_{t+1}(x_{t+1}^k, x_t^k)$.

Lemma 1.

$$E[\hat{p}^A(y_t|y_{1:t-1})|\mathcal{A}_{t-1}] = \sum_{k=1}^M p(y_t|x_{t-1}^k) \pi_{t-1}^k,$$

where the swarm of particles at time t is $\mathcal{A}_t = \{x_t^k; \pi_t^k\}$.

Proof.

$$\begin{aligned}
& E[\hat{p}^A(y_t|y_{1:t-1}) | \mathcal{A}_{t-1}] \\
&= E \left[\sum_{k=1}^M \frac{\omega_t(x_t^k; \tilde{x}_{t-1}^k)}{M} \mid \mathcal{A}_{t-1} \right] \left\{ \sum_{j=1}^M \omega_{t-1|t}^j \right\} \\
&= \int \omega_t(x_t; \tilde{x}_{t-1}) g(x_t|\tilde{x}_{t-1}; y_t) \hat{g}_M^A(\tilde{x}_{t-1}|y_{1:t}) dx_t d\tilde{x}_{t-1} \left\{ \sum_{j=1}^M \omega_{t-1|t}^j \right\} \\
&= \int \sum_{k=1}^M \omega_t(x_t; x_{t-1}^k) g(x_t|x_{t-1}^k; y_t) \frac{\omega_{t-1|t}(x_{t-1}^k)}{(\sum_{j=1}^M \omega_{t-1|t}(x_{t-1}^j))} dx_t \left\{ \sum_{j=1}^M \omega_{t-1|t}^j \right\} \\
&= \int \sum_{k=1}^M \omega_t(x_t; x_{t-1}^k) g(x_t|x_{t-1}^k; y_t) \omega_{t-1|t}(x_{t-1}^k) dx_t \\
&= \sum_{k=1}^M \int \frac{p(y_t|x_t)p(x_t|x_{t-1}^k)}{g(y_t|x_t^k)g(x_t|x_{t-1}^k, y_t)} g(x_t|x_{t-1}^k, y_t) g(y_t|x_{t-1}^k) \pi_{t-1}^k dx_t,
\end{aligned}$$

so

$$\begin{aligned}
E[\widehat{p}^A(y_t|y_{1:t-1}) \mid \mathcal{A}_{t-1}] &= \sum_{k=1}^M \pi_{t-1}^k \int p(y_t|x_t)p(x_t|x_{t-1}^k)dx_t \\
&= \sum_{k=1}^M p(y_t|x_{t-1}^k)\pi_{t-1}^k.
\end{aligned}$$

□

Lemma 2.

$$E[\widehat{p}^A(y_{t-h:t}|y_{1:t-h-1})|\mathcal{A}_{t-h-1}] = \sum_{k=1}^M p(y_{t-h:t}|x_{t-h-1}^k)\pi_{t-h-1}^k$$

Proof. (by induction)

A. Show true for case $h = 1$,

$$\begin{aligned}
E[\widehat{p}^A(y_{t-1:t}|y_{1:t-2})|\mathcal{A}_{t-2}] &= E[\widehat{p}^A(y_t|y_{1:t-1})\widehat{p}^A(y_{t-1}|y_{1:t-2})|\mathcal{A}_{t-2}] \\
&= E \left[E[\widehat{p}^A(y_t|y_{1:t-1}) \mid \mathcal{A}_{t-1}] \widehat{p}^A(y_{t-1}|y_{1:t-2}) \mid \mathcal{A}_{t-2} \right].
\end{aligned}$$

The inner integral is,

$$E[\widehat{p}^A(y_t|y_{1:t-1}) \mid \mathcal{A}_{t-1}] = \sum_{k=1}^M p(y_t|x_{t-1}^k)\pi_{t-1}^k,$$

from Lemma 1. Hence,

$$\begin{aligned}
& E[\widehat{p}^A(y_{t-1:t}|y_{1:t-2})|\mathcal{A}_{t-2}] \\
&= E \left[\left\{ \sum_{k=1}^M p(y_t|x_{t-1}^k) \pi_{t-1}^k \right\} \left\{ \sum_{i=1}^M \frac{\omega_{t-1}^i}{M} \right\} \mid \mathcal{A}_{t-2} \right] \left\{ \sum_{j=1}^M \omega_{t-2|t-1}^j \right\} \\
&= E \left[\left\{ \sum_{k=1}^M p(y_t|x_{t-1}^k) \frac{\omega_{t-1}^k}{\sum_{i=1}^M \omega_{t-1}^i} \right\} \left\{ \sum_{i=1}^M \frac{\omega_{t-1}^i}{M} \right\} \mid \mathcal{A}_{t-2} \right] \left\{ \sum_{j=1}^M \omega_{t-2|t-1}^j \right\} \\
&= E \left[\left\{ \frac{1}{M} \sum_{k=1}^M p(y_t|x_{t-1}^k) \omega_{t-1}^k \right\} \mid \mathcal{A}_{t-2} \right] \left\{ \sum_{j=1}^M \omega_{t-2|t-1}^j \right\} \\
&= \left\{ \sum_{j=1}^M \omega_{t-2|t-1}^j \right\} \int p(y_t|x_{t-1}) \omega_{t-1}(x_{t-1}; \tilde{x}_{t-2}) g(x_{t-1}|\tilde{x}_{t-2}; y_{t-1}) \widehat{g}_M^A(\tilde{x}_{t-2}|y_{1:t-1}) dx_{t-1} d\tilde{x}_{t-2} \\
&= \left\{ \sum_{j=1}^M \omega_{t-2|t-1}^j \right\} \int \sum_{k=1}^M p(y_t|x_{t-1}) \omega_{t-1}(x_{t-1}; x_{t-2}^k) g(x_{t-1}|x_{t-2}^k; y_{t-1}) \frac{g(y_{t-1}|x_{t-2}^k) \pi_{t-2}^k}{\sum_{j=1}^M \omega_{t-2|t-1}^j} dx_{t-1} \\
&= \sum_{k=1}^M \pi_{t-2}^k \int p(y_t|x_{t-1}) \omega_{t-1}(x_{t-1}; x_{t-2}^k) g(y_{t-1}|x_{t-2}^k) g(x_{t-1}|x_{t-2}^k; y_{t-1}) dx_{t-1} \\
&= \sum_{k=1}^M \pi_{t-2}^k \int p(y_t|x_{t-1}) p(y_{t-1}|x_{t-1}) p(x_{t-1}|x_{t-2}^k) dx_{t-1} \\
&= \sum_{k=1}^M p(y_{t-1:t}|x_{t-2}^k) \pi_{t-2}^k \text{ as required.}
\end{aligned}$$

B Assume that the theorem holds for h ,

$$E[\widehat{p}^A(y_{t-h:t}|y_{1:t-h-1})|\mathcal{A}_{t-h-1}] = \sum_{k=1}^M p(y_{t-h:t}|x_{t-h-1}^k) \pi_{t-h-1}^k$$

C Show that the theorem holds for $h+1$:

$$\begin{aligned}
& E[\widehat{p}^A(y_{t-h-1:t}|y_{1:t-h-2})|\mathcal{A}_{t-h-2}] \\
&= E \left[E[\widehat{p}^A(y_{t-h:t}|y_{1:t-h-1}) \mid \mathcal{A}_{t-h-1}] \widehat{p}^A(y_{t-h-1}|y_{1:t-h-2}) \mid \mathcal{A}_{t-h-2} \right] \\
&= E \left[\sum_{k=1}^M p(y_{t-h:t}|x_{t-h-1}^k) \pi_{t-h-1}^k \sum_{i=1}^M \frac{\omega_{t-h-1}^i}{M} \sum_{j=1}^M \omega_{t-h-2|t-h-1}^j \mid \mathcal{A}_{t-h-2} \right],
\end{aligned}$$

using Lemma 1,

$$\begin{aligned}
&= E \left[\left\{ \sum_{k=1}^M p(y_{t-h:t}|x_{t-h-1}^k) \frac{\omega_{t-h-1}^k}{\sum_{i=1}^M \omega_{t-h-1}^i} \right\} \left\{ \sum_{i=1}^M \frac{\omega_{t-h-1}^i}{M} \right\} \mid \mathcal{A}_{t-h-2} \right] \\
&\quad \times \left\{ \sum_{j=1}^M \omega_{t-h-2|t-h-1}^j \right\} \tag{12}
\end{aligned}$$

$$\begin{aligned}
&= E \left[\frac{1}{M} \sum_{k=1}^M p(y_{t-h:t}|x_{t-h-1}^k) \omega_{t-h-1}^k \mid \mathcal{A}_{t-h-2} \right] \left\{ \sum_{j=1}^M \omega_{t-h-2|t-h-1}^j \right\} \\
&= \left\{ \sum_{j=1}^M \omega_{t-h-2|t-h-1}^j \right\} \int p(y_{t-h:t}|x_{t-h-1}) \omega_{t-h-1}(x_{t-h-1}; \tilde{x}_{t-h-2}) \\
&\quad g(x_{t-h-1}|\tilde{x}_{t-h-2}; y_{t-h-1}) \hat{g}_M^A(\tilde{x}_{t-h-2}|y_{1:t-h-1}) dx_{t-h-1} d\tilde{x}_{t-h-2} \tag{13}
\end{aligned}$$

$$\begin{aligned}
&= \left\{ \sum_{j=1}^M \omega_{t-h-2|t-h-1}^j \right\} \int \sum_{k=1}^M p(y_{t-h:t}|x_{t-h-1}) \omega_{t-h-1}(x_{t-h-1}; x_{t-h-2}^k) \\
&\quad g(x_{t-h-1}|x_{t-h-2}^k; y_{t-1}) \frac{g(y_{t-h-1}|x_{t-h-2}^k) \pi_{t-h-2}^k}{\sum_{j=1}^M \omega_{t-h-2|t-h-1}^j} dx_{t-h-1} \tag{14}
\end{aligned}$$

$$\begin{aligned}
&= \sum_{k=1}^M \pi_{t-h-2}^k \int p(y_{t-h:t}|x_{t-h-1}) \omega_{t-h-1}(x_{t-h-1}; x_{t-h-2}^k) \\
&\quad g(x_{t-h-1}|x_{t-h-2}^k; y_{t-1}) g(y_{t-h-1}|x_{t-h-2}^k) dx_{t-h-1}, \tag{15}
\end{aligned}$$

using the definition of ω_{t-h-1} (see step 4 of Algorithm),

$$\begin{aligned}
&= \sum_{k=1}^M \pi_{t-h-2}^k \int p(y_{t-h:t}|x_{t-h-1}) p(y_{t-h-1}|x_{t-h-1}) p(x_{t-h-1}|x_{t-h-2}^k) dx_{t-h-1} \\
&= \sum_{k=1}^M p(y_{t-h-1:t}|x_{t-h-2}^k) \pi_{t-h-2}^k \text{ as required}
\end{aligned}$$

□

Proof of 1. As a consequence we have the lemma that, with $h = t - 2$

$$E[\hat{p}^A(y_{1:t})|A_0] = \sum_{k=1}^M p(y_{1:t}|x_0^k) \pi_0^k$$

where $x_0^k \sim p(x_0)$ and $\pi_0^k = 1/M$,

$$E \left[\sum_{k=1}^M p(y_{1:t}|x_0^k) \pi_0^k \right] = \int p(y_{1:t}|x_0) p(x_0) dx_0 = p(y_{1:t}),$$

as required. □

B Fully and partially adapted particle filter

B.1 Partially adapted particle filter

The partially adapted particle filter used in our article is described in general as follows. We omit dependence on unknown parameters for clarity. Suppose that $p(x_{t+1}|x_t) \sim \mathcal{N}(\mu(x_t), \Sigma(x_t))$ and $p(y_{t+1}|x_{t+1})$ is log-concave as a function of x_{t+1} . Let $\ell(x_{t+1}) = \log p(y_{t+1}|x_{t+1})$, $\lambda(x_{t+1}, k) = \ell(x_{t+1}) + \log p(x_{t+1}|x_t^k)$ and let

$$\check{x}_{t+1}^k = \arg \max_{x_{t+1}} \lambda(x_{t+1}, k), \quad \check{\Sigma}_{t+1}^k = \left(-\frac{\partial^2 \lambda(x_{t+1}, k)}{\partial x_{t+1} \partial x_{t+1}} \right)^{-1}_{x_{t+1}=\check{x}_{t+1}^k}$$

Then, we take $g(x_{t+1}|\check{x}_t^k, y_{t+1}) = N(x_{t+1}; \check{x}_{t+1}^k, \check{\Sigma}_{t+1}^k)$ and $g(y_{t+1}|x_t^k) \propto p(y_{t+1}|\check{x}_{t+1}^k)p(\check{x}_{t+1}^k|x_t^k) \det(\check{\Sigma}_{t+1}^k)^{\frac{1}{2}}$. This is obtained from the second order approximation

$$\lambda(x_{t+1}, k) \doteq \lambda(\check{x}_{t+1}^k, k) + \log \left(\det(\check{\Sigma}_{t+1}^k)^{\frac{1}{2}} \right) - \frac{1}{2} (x_{t+1} - \check{x}_{t+1}^k)' \left(\check{\Sigma}_{t+1}^k \right)^{-1} (x_{t+1} - \check{x}_{t+1}^k) - \log \left(\det(\check{\Sigma}_{t+1}^k)^{\frac{1}{2}} \right)$$

The mode \check{x}_{t+1}^k is obtained by Newton-Raphson iteration with the starting value $\mu(x_t^k)$, or some problem specific starting value as in the binomial example.

An alternative iterative scheme to obtain \check{x}_{t+1}^k is based on solving $\partial \lambda(x)/\partial x = 0 = \partial \ell(x)/\partial x - \Sigma(x_t^k)^{-1}(x - \mu(x_t^k))$. The iteration is given by

$$x_{t+1}^k = \mu(x_t^k) + \Sigma(x_t^k) \partial \ell(x_{t+1}^k) / \partial x_{t+1}^k. \quad (16)$$

A single iteration of (16) is usually faster than a single iteration of the Newton-Raphson scheme but the actual speed of convergence depends on the problem. In practice, we can make the iterations to the mode faster by just taking a fixed small number of steps of either the Newton-Raphson or (16), or by making the convergence criterion less strict. If we take a fixed number of steps then the iterations can be vectorized over k .

Binomial example For the binomial example discussed in Section 4.2 we use the Newton-Raphson iteration with starting value $\mu(x_t^k)$ or $\text{logit}(y_t/m)$ if m is large.

B.2 Fully adaptive particle filter

Full adaptation is possible whenever $p(x_{t+1}|x_t)$ is conjugate in x_{t+1} to $p(y_{t+1}|x_{t+1})$.

Gaussian observation equation Suppose that the observation equation is Gaussian with $p(y_t|x_t) \sim N(H_t x_t, V_t)$ and the state transition equation is the same as in Section B.1. Then, from Section B.1, \check{x}_{t+1}^k and $\check{\Sigma}_{t+1}^k$ are obtained explicitly as

$$\check{\Sigma}_{t+1}^k = (H_{t+1}' V_{t+1}^{-1} H_{t+1} + \Sigma(x_t^k)^{-1})^{-1}, \quad \check{x}_{t+1}^k = \check{\Sigma}_{t+1}^k (V_{t+1}^{-1} y_{t+1} + \Sigma(x_t^k)^{-1} \mu(x_t^k)) ,$$

and $p(y_{t+1}|x_t^k)$ is obtained as in Section B.1.

Garch model We use the notation in Section 5.2. It is straightforward to show that $p(y_{t+1}|x_t, \sigma_t^2) \sim N(0, \sigma_{t+1}^2 + \tau^2)$ and that $p(x_{t+1}|y_{t+1}, x_t) \sim N(a_{t+1}, \Delta_{t+1})$, where

$$\Delta_{t+1}^{-1} = (\tau^2)^{-1} + (\sigma_{t+1}^2)^{-1}, \quad a_{t+1} = \Delta_{t+1} y_{t+1} / \tau^2.$$

C Adaptive sampling schemes

This appendix describes the two adaptive sampling schemes used in the paper.

C.1 Adaptive random walk Metropolis

The adaptive random walk Metropolis proposal of Roberts and Rosenthal (2009) is

$$q_j(\theta; \theta_{j-1}) = \omega_{1j} \phi_d(\theta; \theta_{j-1}, \kappa_1 \Sigma_1) + \omega_{2j} \phi_d(\theta; \theta_{j-1}, \kappa_2 \Sigma_{2j}) \quad (17)$$

where d is the dimension of θ and $\phi_d(\theta; \tilde{\theta}, \Sigma)$ is a multivariate d dimensional normal density in θ with mean $\tilde{\theta}$ and covariance matrix Σ . In (17), $\omega_{1j} = 1$ for $j \leq j_0$, with j_0 representing the initial iterations, $\omega_{1j} = 0.05$ for $j > j_0$ with $\omega_{2j} = 1 - \omega_{1j}$; $\kappa_1 = 0.1^2/d$, $\kappa_2 = 2.38^2/d$, Σ_1 is a constant covariance matrix, which is taken as the identity matrix by Roberts and Rosenthal (2009) but can be based on the Laplace approximation or some other estimate. The matrix Σ_{2j} is the sample covariance matrix of the first $j-1$ iterates. The scalar κ_1 is meant to achieve a high acceptance rate by moving the sampler locally, while the scalar κ_2 is considered to be optimal (Roberts et al., 1997) for a random walk proposal when the target is a multivariate normal. We note that the acceptance probability (6) for the adaptive random walk Metropolis simplifies to

$$\alpha(\theta_{j-1}, u_{j-1}; \theta_j^p, u^p) = \min \left\{ 1, \frac{p(y|\theta_j^p, u_j^p)p(\theta^p)}{p(y|\theta_{j-1}, u_{j-1})p(\theta_{j-1})} \right\}. \quad (18)$$

C.2 A proposal density based on a mixture of normals

The proposal density of the adaptive independent Metropolis-Hastings approach of Giordani and Kohn (2010) is a mixture with four terms of the form

$$q_j(\theta) = \sum_{k=1}^4 \omega_{kj} g_k(\theta|\lambda_{kj}) \quad \omega_{kj} \geq 0, \quad \text{for } k = 1, \dots, 4 \quad \text{and} \quad \sum_{k=1}^4 \omega_{kj} = 1, \quad (19)$$

with λ_{kj} the parameter vector for the density $g_{kj}(\theta; \lambda_{kj})$. The sampling scheme is run in two stages, which are described below. Throughout each stage, the parameters in the first two terms are kept fixed. The first term $g_1(\theta|\lambda_{1j})$ is an estimate of the target density and the second term $g_2(\theta|\lambda_{2j})$ is a heavy tailed version of $g_1(\theta|\lambda_{1j})$. The third term $g_3(\theta|\lambda_{3j})$ is an estimate of the target that is updated or adapted as the simulation progresses and the fourth term $g_4(\theta|\lambda_{4j})$ is a heavy tailed version of the third term. In the first stage $g_{1j}(\theta; \lambda_{1j})$ is a Gaussian density constructed from a preliminary run, of the three component adaptive random walk. Throughout, $g_2(\theta|\lambda_{2j})$ has the same component means and probabilities as $g_1(\theta|\lambda_{1j})$, but its component covariance matrices are ten times those of $g_1(\theta|\lambda_{1j})$. The term

$g_3(\theta|\lambda_{3j})$ is a mixture of normals and $g_4(\theta|\lambda_{4j})$ is also a mixture of normals obtained by taking its component probabilities and means equal to those of $g_3(\theta|\lambda_{3j})$, and its component covariance matrices equal to 20 times those of $g_3(\theta|\lambda_{3j})$. The first stage begins by using $g_1(\theta|\lambda_{1j})$ and $g_2(\theta|\lambda_{2j})$ only with, for example, $\omega_{1j} = 0.8$ and $\omega_{2j} = 0.2$, until there is a sufficiently large number of iterates to form $g_3(\theta|\lambda_{3j})$. After that we set $\omega_{1j} = 0.15, \omega_{2j} = 0.05, \omega_{3j} = 0.7$ and $\omega_{4j} = 0.1$. We begin with a single normal density for $g_3(\theta|\lambda_{3j})$ and as the simulation progresses we add more components up to a maximum of six according to a schedule that depends on the ratio of the number of accepted draws to the dimension of θ .

In the second stage, $g_1(\theta|\lambda_{1j})$ is set to the value of $g_3(\theta|\lambda_{3j})$ at the end of the first stage and $g_2(\theta|\lambda_{2j})$ and $g_4(\theta|\lambda_{4j})$ are constructed as described above. The heavy-tailed densities $g_2(\theta|\lambda_{2j})$ and $g_4(\theta|\lambda_{4j})$ are included as a defensive strategy to get out of local modes and to explore the sample space of the target distribution more effectively.

It is computationally too expensive to update $g_3(\theta|\lambda_{3j})$ (and hence $g_4(\theta|\lambda_{4j})$) at every iteration so we update them according to a schedule that depends on the problem and the size of the parameter vector.

C.3 Proof of the convergence of the adaptive independent Metropolis-Hastings sampling scheme

The following convergence results hold for the adaptive independent Metropolis Hastings sampling scheme described in Appendix C.2 (and more fully in Giordani and Kohn (2010)) when it is combined with the ASIR particle filter. They follow from Theorems 1 and 2 of Giordani and Kohn (2010). Let Θ be the parameter space of θ .

Theorem 2. *Suppose that there exists a constant $0 < C < \infty$ that does not depend on $t = 1, \dots, T, \theta \in \Theta$ and the number of iterates j such that*

$$g(y_{t+1}|x_t; \theta) \leq C, \quad (20)$$

$$\frac{p(y_{t+1}|x_{t+1}; \theta)p(x_{t+1}|x_t; \theta)}{g(y_{t+1}|x_t; \theta)g(x_{t+1}|y_{t+1}, x_t; \theta)} \leq C, \quad (21)$$

$$p(\theta)/q_j(\theta) \leq C. \quad (22)$$

Then,

1. *The simulated likelihood is bounded uniformly in $\theta \in \Theta$.*
2. *The iterates θ_j of the adaptive independent Metropolis Hastings sampling scheme converge to a sample from $p(\theta|y)$ in the sense that*

$$\sup_{A \subset \Theta} |\Pr(\theta_j \in A) - \int_A p(\theta | y) d\theta| \rightarrow 0 \quad \text{as } j \rightarrow \infty. \quad (23)$$

for all measurable sets A of Θ .

3. *Suppose that $h(\theta)$ is a measurable function of θ that is square integrable with respect to*

the density g_2 . Then, almost surely,

$$\frac{1}{n} \sum_{j=1}^n h(\theta_j) \rightarrow \int h(\theta) p(\theta|y) d\theta \quad \text{as } n \rightarrow \infty. \quad (24)$$

Proof.

$$p_S(y|\theta, u) = \prod_{t=0}^{T-1} p_S(y_{t+1}|y_{1:t}; \theta, u) \leq C^{2T} \quad \text{because } p_S(y_t|y_{1:t-1}; \theta) \leq C^2$$

from (3) and our assumptions. This shows that the simulated likelihood $p_S(y|\theta, u)$ is bounded and the result now follows from Giordani and Kohn (2010). \square

We note that as in Giordani and Kohn (2010) it is straightforward to choose the proposal density $q_j(\theta)$ as a mixture with one component that is at least as heavy tailed as $p(\theta)$ to ensure that (22) holds.

The next corollary gives a condition for equations (20) and 21 to hold for the standard particle filter and the fully adapted particle filter.

Corollary 1. *Suppose that for all $y_t, x_t, t = 1, \dots, T$ and $\theta \in \Theta$, there exists a constant $C_1 > 0$ such that*

$$p(y_t|x_t; \theta) \leq C_1. \quad (25)$$

Then equations (20) and (21) hold for the standard particle filter and the fully adapted particle filter.

Proof. We have

$$p(y_{t+1}|x_t; \theta) = \int p(y_{t+1}|x_{t+1}; \theta) p(x_{t+1}|x_t; \theta) dx_{t+1} \leq C_1$$

and the result follows for the standard particle filter and the fully adapted particle filter. \square

We note that usually $p(y_t|x_t; \theta)$ is uniformly bounded in y_t, x_t and θ for $t = 1, \dots, T$. This is true for the models in Sections 4 and 5.

We now construct a partially adapted particle filter that satisfies equations (20) and (21). Suppose that $g_0(y_{t+1}|x_t; \theta)$ and $g_0(x_{t+1}|y_{t+1}, x_t; \theta)$ correspond to a partially adapted particle filter which we refer to as g_0 , e.g. the partially adapted particle filter described in Section B.1. Let $0 < \epsilon < 1$. Now construct the partially adapted particle filter g as a mixture taking the value g_0 with probability $1 - \epsilon$ and being the standard particle filter with probability ϵ . That is,

$$g(y_{t+1}|x_t; \theta) g(x_{t+1}|x_t, y_{t+1}; \theta) = \epsilon p(x_{t+1}|x_t) + (1 - \epsilon) g_0(y_{t+1}|x_t) g_0(x_{t+1}|x_t, y_{t+1}; \theta). \quad (26)$$

Corollary 2. *Suppose equation (25) holds and the partially adapted particle filter is defined by equation (26). Then, equations (20) and (21) hold.*

The proof is straightforward.

Usually, we would take ϵ quite small so that most of the time the partially adapted particle filter g_0 is used. Using the mixture partially adapted particle filter ensures that the simulated likelihood is bounded which is important to successfully use the adaptive independent Metropolis Hastings to sample the parameters.

D Marginal likelihood evaluation using bridge and importance sampling

Suppose that $q(\theta)$ is an approximation to $p(\theta|y)$ which can be evaluated explicitly. Bridge sampling (Meng and Wong, 1996) estimates the marginal likelihood as follows. Let

$$t(\theta) = \left(\frac{p(y|\theta)p(\theta)}{U} + q(\theta) \right)^{-1},$$

where U is a positive constant. Let

$$\begin{aligned} A &= \int t(\theta)q(\theta)p(\theta | y)d\theta. \quad \text{Then,} \\ A &= \frac{A_1}{p(y)} \quad \text{where} \quad A_1 = \int t(\theta)q(\theta)p(y | \theta)p(\theta)d\theta. \end{aligned} \tag{27}$$

Suppose the sequence of iterates $\{\theta^{(j)}, j = 1, \dots, M\}$ is generated from the posterior density $p(\theta|y)$ and a second sequence of iterates $\{\theta^{(k)}, k = 1, \dots, M\}$ is generated from $q(\theta)$. Then

$$\hat{A} = \frac{1}{M} \sum_{j=1}^M t(\theta^{(j)})q(\theta^{(j)}), \quad \hat{A}_1 = \frac{1}{M} \sum_{k=1}^M t(\theta^{(k)})p(y|\theta^{(k)})p(\theta^{(k)}) \quad \text{and} \quad \hat{p}_{BS}(y) = \frac{\hat{A}_1}{\hat{A}}$$

are estimates of A and A_1 and $\hat{p}_{BS}(y)$ is the bridge sampling estimator of the marginal likelihood $p(y)$.

In adaptive sampling, $q(\theta)$ is the mixture of normals proposal. Although U can be any positive constant, it is more efficient if U is a reasonable estimate of $p(y)$. One way to do so is to take $\hat{U} = p(y|\theta^*)p(\theta^*)/q(\theta^*)$, where θ^* is the posterior mean of θ obtained from the posterior simulation.

An alternative method to estimate of the marginal likelihood $p(y)$ is to use importance sampling based on the proposal distribution $q(\theta)$ (Geweke, 1989; Chen and Shao, 1997). That is,

$$\hat{p}_{IS}(y) = \frac{1}{K} \sum_{k=1}^K \frac{p(y|\theta^{(k)})p(\theta^{(k)})}{q(\theta^{(k)})}.$$

Since our proposal distributions have at least one heavy tailed component, the importance sampling ratios are likely to be bounded and well-behaved, as in the examples in this paper.

E Implementation details

We coded most of the algorithms in MATLAB, with a small proportion of the code written using C/Mex files. We carried out the estimation on an SGI cluster with 42 compute nodes. Each of them is an SGI Altix XE320 with two Intel Xeon X5472 (quad core 3.0GHz) CPUs with at least 16GB memory. We ran parallel jobs using up to eight processors and MATLAB 2009.