On some properties of Markov chain Monte Carlo simulation methods based on the particle filter

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

October 6, 2011

Abstract

Andrieu et al. (2010) prove that Markov chain Monte Carlo samplers still converge to the correct posterior distribution of the model parameters when the likelihood estimated by the particle filter (with a finite number of particles) is used instead of the true likelihood. A critical issue for performance is the choice of the number of particles. We add the following contributions. First, we provide analytically derived, practical guidelines on the optimal number of particles to use. Second, we show that a fully adapted auxiliary particle filter is unbiased and can drastically decrease computing time compared to a standard particle filter. Third, we introduce a new estimator of the likelihood based on the output of a general auxiliary particle filter and use the framework of Del Moral (2004) to provide a direct proof of the unbiasedness of the estimator. Fourth, we show that the results in the article apply more generally to Markov chain Monte Carlo sampling schemes with the likelihood estimated in an unbiased manner. Fifth, we show that adaptive samplers, and in particular an adaptive independent Metropolis Hastings sampler based on a mixture of normals, can be effective in generating the parameters when working with the particle filter, and give a proof of the convergence of a Markov chain Monte Carlo sampling scheme based on this sampler.
Keywords: Auxiliary variables; Adapted filtering; Bayesian inference; Simulated likelihood.

1 Introduction

This paper is concerned with developing a methodology for Bayesian inference for general time series state space models using Markov chain Monte Carlo (MCMC) simulation with the likelihood estimated by the particle filter (PF). The particle filter gives an unbiased estimator of the likelihood, and we call this estimate the simulated likelihood. We call such methodology particle filter MCMC (PMCMC). Our paper builds on the PMCMC work of Andrieu et al. (2010).

We make a number of contributions. The first is to explore the properties of PMCMC, and in particular to give guidance for the optimal number of particles to use. More particles increase the acceptance rate of a PMCMC sampler but at an increased computational cost. To obtain analytic results on an optimal trade-off we make some simplifying assumptions which are given and discussed

in Section 4. Our analysis shows that the performance of the PMCMC is governed by the standard deviation of the log of the simulated likelihood and that the optimal value of this standard deviation is around 1, with a relatively benign region for this standard deviation being between 0.5 and 1.5. Once the standard deviation exceeds 2, performance decreases exponentially with an increase in the square of the standard deviation. Practical guidelines for choosing a reasonable value for N are given at the beginning of Section 4 and in Section 4.3.

Our second contribution is to show empirically that the log of the simulated likelihood obtained by fully adapted auxiliary particle filters can have a much smaller standard deviation than the standard deviation obtained using the standard particle filter of Gordon et al. (1993), especially when the signal to noise ratio is high. Our analytic results then suggest that it may be sufficient to use far fewer particles using adapted particle filters, especially fully adapted particle filters, to obtain the same statistical accuracy as the standard particle filter. We note that it is very important to carry out particle filtering as efficiently as possible when the sample size T is large. This issue is discussed at the end of Section 4.3.

The third contribution is to introduce a new estimator of the likelihood based on the output of a general auxiliary particle filter and use the framework of Del Moral (2004) to provide a direct proof of the unbiasedness of the resulting estimator. The fourth contribution is to show that the results in the article apply more generally to MCMC sampling schemes with the likelihood estimated in an unbiased manner, and in particular when the likelihood is estimated by importance sampling.

The simulated likelihood as constructed in our article is random and therefore not continuous in the 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 or even if the parameter does not change at all. It is therefore attractive to form proposals for the parameters using adaptive sampling methods. In our article we define adaptive sampling methods as simulation methods for carrying out Bayesian inference that use previous iterates of the parameters 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 & Rosenthal (2005) and Roberts & Rosenthal (2009) who consider adaptive random walk Metropolis proposals.

The fifth contribution of our article is to show that it is useful to generate the unknown parameters using the adaptive independent Metropolis-Hastings (AIMH) sampling scheme of Giordani & Kohn (2010), which approximates the posterior density by a mixture of normals. We also show that the resulting PMCMC converges to the correct posterior distribution. It is worthwhile constructing adaptive independent Metropolis-Hastings proposals that provide good approximations to the posterior density, because they can be appreciably more efficient than random walk proposals, while 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. Furthermore, 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.

We use simulated and real examples to illustrate a number of methodological issues addressed in the article. In particular, we show that our theory for choosing the optimal number of particles, which is developed under idealized conditions, can give a good guide to actual practice. We also illustrate the flexibility of our methods that combine particle filtering with adaptive sampling of the parameters by fitting a signal plus noise model to GDP growth data. The model for the signal is a mixture of two autoregressive experts, and it is necessary to use the particle filter in this example to make the computation tractable.

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 model are well developed (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 & 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 & 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 & Chen (1998), Doucet et al. (2000), Doucet et al. (2001), Andrieu & Doucet (2002), Fearnhead & Clifford (2003) and Del Moral et al. (2006). Our article considers only the standard particle filter of Gordon et al. (1993) and the adapted particle filters proposed by Pitt & Shephard (1999).

The literature on using the particle filter to learn about model parameters is more limited. Malik & Pitt (2011) propose the smooth particle filter to estimate the parameters of a state space model using maximum likelihood. Andrieu et al. (2010) provide a framework for off-line parameter learning using the particle filter. Flury & Shephard (2011) apply the results of Andrieu et al. (2010) to interesting econometric examples using single parameter random walk proposals for off-line Bayesian inference. Storvik (2002), Polson et al. (2008), Lopes et al. (2011) and Carvalho et al. (2010) consider online parameter learning when sufficient statistics are available. Developing online parameter learning algorithms and theory is an important research area, but one that is not covered in our article.

2 Simulated Likelihood Inference

Using an unbiased estimator of the likelihood within a Markov chain Monte Carlo (MCMC) scheme was first proposed by Beaumont (2003). The theoretical properties of such schemes are examined in Andrieu & Roberts (2009). The explanation which follows in this section and in Section 2.1 is close to that of Section 2 of Flury & Shephard (2011) who provide an elegant justification and explanation of the use of this method for both importance sampling estimators of the likelihood, discussed in Section 3.1, and estimators arising from the particle filter, discussed in Section 3.2. For now we shall assume that we have a method for generating an unbiased estimator before exploring the specific construction of such estimators. This section shows that the simulated likelihood estimator can be used to construct a valid density which admits the true posterior density as a marginal. The MCMC scheme itself becomes rather straightforward in principle and is briefly outlined in Section 2.1.

We denote all observations as $y = y_{1:T} = (y_1', ..., y_T')'$ and the parameters as θ. The true likelihood of the observations, which we consider to be analytically intractable, will be denoted as $p(y|\theta)$ whilst the prior for the parameters will be denoted as $p(\theta)$. The true posterior density for θ will be denoted as $\pi(\theta) \propto p(y|\theta)p(\theta)$. Let us denote all of the random variables generated in the construction of the likelihood estimator by the vector u. This vector consists of all the random variables used in the construction of the estimator arising from the importance sampler or from the particle filter. For the particle filter this vector includes the variables used in the resampling part of the method as well as in the transition density which should be apparent from Section 3.2. We can

consider u to be canonical, by which we mean that the distribution of u is chosen to be the same for all problems and not dependent on the parameters. For example, without any loss of generality, Flury & Shephard (2011), Section 2, consider u to consist of independent identically distributed standard uniform variates. This vector u is typically high dimensional. We include u in the notation for the likelihood estimator writing this as $\widehat{p}_N(y|\theta, u)$, with the hat indicating that this estimator is not necessarily a density in y. We will sometimes use $\widehat{p}_N(y|\theta)$ as a shorthand for $\widehat{p}_N(y|\theta, u)$. The subscript N represents the number of samples in the importance sampler estimator, or the number of particles in the particle filter, see Section 3.1 and Section 3.2 respectively. Formally, the dimension of u also depends upon N though for notational convenience we will write the density from which u arises as $p(u)$. We note that this is the density of the random variates used in the construction of the estimator, and not a prior or a proposal density.

We will now assume that the estimator is unbiased for the likelihood, i.e.

$$\int \widehat{p}_N(y|\theta, u)p(u)du = p(y|\theta). \tag{1}$$

This is a well known property of importance sampling estimators, see Section 3.1, and also holds for particle filter estimators, see Section 3.2. Given this property, it is now possible to write down a joint density in θ and u which admits the correct marginal density for θ as $\pi(\theta)$, the posterior density. If we define the joint posterior density $\pi_N(\theta, u)$ of θ and u as

$$\pi_N(\theta, u) \propto \widehat{p}_N(y|\theta, u)p(\theta)p(u), \tag{2}$$

then it may be seen that this may be written in a normalised form as

$$\pi_N(\theta, u) = \frac{\widehat{p}_N(y|\theta, u)p(\theta)p(u)}{p(y)} \tag{3}$$

$$= \frac{\widehat{p}_N(y|\theta, u)}{p(y|\theta)}\pi(\theta)p(u). \tag{4}$$

where $p(y) = \int p(y|\theta)p(\theta)d\theta$, the true marginal likelihood. We can see this joint $\pi_N(\theta, u)$ integrates to one and admits the correct marginal density for θ by integrating over u, and using (1), so that

$$\int \pi_N(\theta, u)du = \pi(\theta),$$

the posterior density of θ. This means that a Markov chain Monte Carlo scheme is relatively straightforward to implement where the target density is $\pi_N(\theta, u)$ and we use the unnormalised form in equation (2) within a Metropolis expression. This is briefly outlined in Section 2.1.

2.1 MCMC inference using the simulated likelihood

The target density for posterior inference is $\pi_N(\theta, u)$ given in unnormalised form at (2). It may therefore be possible to use a Metropolis-Hastings simulation method to generate samples from the target density as follows. We note that at present we will be generating the high dimensional random variate u from the associated density $p(u)$. This arises automatically when we generate the likelihood estimator by importance sampling or particle filter methods. As a consequence we can regard u as being proposed from $p(u)$ and this density will cancel in the Metropolis expression to be outlined.

Suppose we have a joint Markov chain in (θ_j, u_j) arising from $\pi_N(\theta, u)$. Then to move to the

next step of the chain, i.e. (θ_{j+1}, u_{j+1}) we propose u^* from $p(u)$ and θ^* from a proposal density $q(\theta|\theta_j)$. We then take $(\theta_{j+1}, u_{j+1}) = (\theta^*, u^*)$ with probability,

$$\alpha(\theta_j, u_j; \theta^*, u^*) = \min\left\{1, \frac{\widehat{p}_N(y|\theta^*, u^*)p(\theta^*)}{\widehat{p}_N(y|\theta_j, u_j)p(\theta_j)} \frac{q(\theta_j|\theta^*)}{q(\theta^*|\theta_j)}\right\}, \tag{5}$$

and take $(\theta_{j+1}, u_{j+1}) = (\theta_j, u_j)$ otherwise. It is informative to note this Metropolis expression is identical to what we would obtain using the true likelihood function, (see e.g. Chib & Greenberg, 1995), except for the appearance of the estimated likelihood rather than the true likelihood $p(y|\theta)$ in equation (5). In practical terms it should be noted that we do not normally retain or record the values of u. Instead we record the current value of the likelihood estimator $\widehat{p}_N(y|\theta_j, u_j)$ and note the new proposed value as $\widehat{p}_N(y|\theta^*, u^*)$.

The theoretical justification for using the simulated likelihood arising from importance sampling within an MCMC scheme is explored in greater detail by Andrieu & Roberts (2009). Econometric applications are considered in Section 2 of Flury & Shephard (2011) using both importance sampling, for limited dependent variable models, and particle filter methods. Section 3 outlines two commonly used methods for obtaining unbiased and consistent (in N) estimators of the likelihood. Section 3.1 considers methods based upon importance sampling and Section 3.2 considers particle filter methods for obtaining estimators. Our later applications focus exclusively on the particle filter. Our results on choosing an optimal value for the number of samples within an MCMC sampling scheme (for particle filters this is the number of particles) apply to both methods.

3 Simulated Likelihood Estimation

3.1 Importance sampling

We shall briefly outline the method of importance sampling for obtaining unbiased likelihood estimation, noting the relevant properties of such estimators. Importance sampling methods were initially discussed by Kahn & Marshall (1953) and popularised by Hammersley & Handscomb (1964) (Section 5.4). The method has been widely used in Bayesian econometrics, see Geweke (1989). For likelihood estimation, we shall assume that we have a latent variable problem so that the data $y = y_{1:T} = (y_1', ..., y_T')'$ is generated conditional upon latent variables $x = x_{1:T} = (x_1', ..., x_T')'$ and θ so that $y \sim p(y|\theta; x)$. We will assume that the latent variables are themselves generated as $x \sim p(x|\theta)$. Examples of this structure include, for instance, the Stochastic Volatility model for which the likelihood has been estimated via importance sampling by, amongst others, Danielsson (1994), Liesenfeld & Richard (2003) and Sandmann & Koopman (1998). In this case, the likelihood estimator is given as

$$\widehat{p}_N(y|\theta, u) = \frac{1}{N}\sum_{k=1}^{N} \omega(x^k; \theta), \quad \text{where} \quad \omega(x; \theta) = \frac{p(y|\theta; x)p(x|\theta)}{g(x|\theta; y)}$$

and the x^k are independent, identically distributed samples from the proposal density $g(x|\theta; y)$. Note that in this case u consists of the random variates that are used to generate $x^1, .., x^N$ from the proposal density $g(x|\theta; y)$. It is straightforward to verify that the likelihood estimator is unbiased as

$$E_{g(x|\theta;y)}[\omega(x; \theta)] = \int p(y|\theta; x)p(x|\theta)dx = p(y|\theta).$$

Furthermore, $\widehat{p}_N(y|\theta,u) \xrightarrow{p} p(y|\theta)$ as $N \longrightarrow \infty$, see (Geweke, 2005, page 114). If the proposal density is sufficiently heavy tailed for the variance of the weights to exist, so that $E_{g(x|\theta;y)}[\omega(x;\theta)^2]$ is finite, then the Lindeberg-Levy central limit theorem applies and

$$\sqrt{N}\left\{\widehat{p}_N(y|\theta,u) - p(y|\theta)\right\} \xrightarrow{d} \mathcal{N}(0;\psi^2(\theta));$$

$\mathcal{N}(a;b^2)$ is a univariate normal distribution with mean a and variance b^2. The variance $\psi^2(\theta) = E_{g(x|\theta;y)}[\omega(x;\theta)^2] - p(y|\theta)^2$ (see e.g. Cappé et al., 2005, p. 287). Applying the second order delta method (see e.g. Billingsley, 1985, p. 368), to obtain the distribution of the error of the logarithm of the estimated likelihood we have

$$\sqrt{N}\left\{\log\widehat{p}_N(y|\theta,u) - \log p(y|\theta)\right\} \xrightarrow{d} \mathcal{N}\left(-\frac{\gamma^2(\theta)}{2\sqrt{N}};\gamma^2(\theta)\right), \tag{6}$$

where $\gamma^2(\theta) = \psi^2(\theta)/p(y|\theta)^2$. This second order correction ensures that if $\sqrt{N}Z \sim \mathcal{N}\left(-\frac{\gamma^2(\theta)}{2\sqrt{N}};\gamma^2(\theta)\right)$ then $E(\exp(Z)) = 1$.

3.2 Particle filters

We briefly describe the sampling-importance-resampling (SIR) particle filter of Gordon et al. (1993). The general auxiliary sampling-importance-resampling (ASIR) filter of Pitt & Shephard (1999) may be thought of as a generalization of the SIR method. The ASIR filter is described in more detail in Section 8.2 of the appendix. In the applications we later consider we focus on the standard SIR method and what we term the fully adapted particle filter (FAPF) method, both of which are special cases of the ASIR filter. More detail on implementing the auxiliary particle filter may also be found in Pitt & Shephard (2001).

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. In the particle filter we are concerned with obtaining (possibly weighted) samples from the filtering density $p(x_t|y_{1:t};\theta)$, (e.g. West & Harrison, 1997, pages 506–507), and in obtaining estimates of the prediction density of the observations $p(y_t|y_{1:t-1};\theta)$ through time. The product of the prediction densities for the observations over time provides the likelihood $p(y|\theta)$.

To simplify the notation in this section, we omit to show dependence on the unknown parameter vector θ. The following algorithm describes the one time step SIR update and is initialized with samples $x_0^k \sim p(x_0)$ with mass $1/N$ for $k = 1, ..., N$.

Algorithm 1. *Given samples $x_t^k \sim p(x_t|y_{1:t})$ with mass π_t^k for $k = 1, ..., N$.*
For $t = 0, .., T-1$:

1. *For $k = 1:N$, sample $\widetilde{x}_t^k \sim \sum_{i=1}^N \pi_t^i \delta(x_t - x_t^i)$.*
2. *For $k = 1:N$, sample $x_{t+1}^k \sim p(x_{t+1}|\widetilde{x}_t^k)$.*
3. *For $k = 1:N$, compute $\omega_{t+1}^k = p(y_{t+1}|x_{t+1}^k)$ and $\pi_{t+1}^k = \omega_{t+1}^k / \left(\sum_{i=1}^N \omega_{t+1}^i\right)$.*

Note that in Step 1, $\delta(x-a)$ is the delta function with unit mass at $x = a$. In addition, in Step 3, multinomial sampling may be employed but stratified sampling is generally to be preferred and

is employed throughout in our applications, see Kitagawa (1996), Carpenter et al. (1999) and Pitt & Shephard (2001). The SIR method of Gordon et al. (1993) can be seen as simple to implement requiring only the ability to simulate from the transition density (in step 2) and to evaluate the observation density (in step 3).

The general SIR estimator of $p(y_t|y_{1:t-1})$ and the simulated likelihood are given by

$$\widehat{p}_N(y_t|y_{1:t-1}) = \sum_{k=1}^{N} \frac{\omega_t^k}{N} \quad \text{and} \quad \widehat{p}_N(y) = \prod_{t=1}^{T} \widehat{p}_N(y_t|y_{1:t-1}). \tag{7}$$

The weights ω_t^k are calculated as part of the SIR algorithm. We introduce again the parameters so that the simulated likelihood estimator $\widehat{p}_N(y)$ of (7) can be represented in our notation as $\widehat{p}_N(y|\theta, u)$ introduced in Section 2. Here u represents the standardised random variates (which do not depend upon θ) used in Steps 1 and 2 of the algorithm above.

Importantly, by Theorem 1, $\widehat{p}_N(y|\theta, u)$ is unbiased for the true likelihood $p(y|\theta)$, where we show dependence on θ for the rest of this section. Unbiasedness is discussed in the general setting of the auxiliary particle filter in the Appendix, Section 8.2. The results of Del Moral (2004) (Proposition 9.4.1, page 301) indicate that a central limit theorem applies to the likelihood estimator so that

$$\sqrt{N}\left\{\widehat{p}_N(y|\theta, u) - p(y|\theta)\right\} \xrightarrow{d} \mathcal{N}(0; \psi^2(\theta)),$$

for some $\psi(\theta)$. Applying the, second order, delta method to obtain the distribution of the error of the logarithm of the estimated likelihood we have

$$\sqrt{N}\left\{\log \widehat{p}_N(y|\theta, u) - \log p(y|\theta)\right\} \xrightarrow{d} \mathcal{N}\left(-\frac{\gamma^2(\theta)}{2\sqrt{N}}; \gamma^2(\theta)\right), \tag{8}$$

where $\gamma^2(\theta) = \psi^2/p(y|\theta)^2$.

4 Analysis of a simplified particle filter Markov chain Monte Carlo sampling scheme

This section attempts to give guidance on the number of particles, N, to choose in a particle filter Markov chain Monte Carlo (PMCMC) sampling scheme. The guidance can also be applied to the choice of the number of samples for an importance sampler within an MCMC scheme. Essentially the goal is to balance two competing costs. If N is taken to be large then we will be estimating the true likelihood quite precisely and this will result in mixing (shown through the autocorrelation in θ) of the Markov chain which will be almost as fast as if we knew the likelihood. However, the cost of doing this is that we take a large value of N and so computing each simulated likelihood $\widehat{p}_N(y|\theta, u)$ will be expensive. On the other hand, a small value of N will result in cheap evaluations of $\widehat{p}_N(y|\theta, u)$ but possibly at the cost of slow mixing (relative to knowing the true likelihood) as indicated by high autocorrelation in the draws of θ. The latter problem (N too small) can be particularly costly as the PMCMC algorithm is not geometrically ergodic, as discussed in Section 4.2, (see also Theorem 8 of Andrieu & Roberts, 2009). In practical terms this means that the resulting chain can occasionally become sticky, retaining the same value for very long periods. This is also problematic because, for values of N too small, the chain can appear to be progressing well (for the first $10,000$ draws say) and then suddenly become stuck for long periods. This feature has been noted by Flury & Shephard (2011), Section 5, who observed, for a state space form model and small N, that it was necessary

to run the PMCMC scheme for 10^6 iterations in order to observe this stickiness, whereas the first $10,000$ draws gave the misleading impression that the chain was mixing rapidly. There are clearly, therefore, gains in being able to choose N in a reasonably sensible manner prior to conducting a full and possibly expensive PMCMC analysis

The results of our analysis indicate that we should choose N so that the variance of the resulting log-likelihood is around 0.85. Of course, in practice this variance will not be constant as it is a function of the parameters as well as a decreasing function of N. A reasonable strategy, discussed in Section 4.3, is to estimate the posterior mean $E_\pi[\theta]$ from an initial short run of the PMCMC scheme with N set to a large value. The value of N could then be adjusted such that the variance of the log-likelihood $\text{Var}\{\log p_N(y|\overline{\theta}, u)\}$ evaluated at this value, say $\overline{\theta}$, is around 0.85. Alternatively, for some models there exist approximating models for which the likelihood is known. For example, the stochastic volatility model (SV) can be approximated by a state space form model, see Harvey et al. (1994). In this case, a simple strategy would be to obtain the MLE, $\widehat{\theta}$, for this approximating model and then find N such that the $Var\{\log p_N(y|\widehat{\theta}, u)\} \simeq 0.85$. The penalty for getting the variance wrong is not too severe within a certain range. Our results indicate that although a value of 0.8464 is optimal, the penalty is small provided the value is between 0.25 and 2.25. This allows for quite a large margin of error in choosing N and also suggests that the simple schemes advocated should work well.

Sections 4.1 and 4.2 we develop an approach which, with certain simplifying assumptions, allows an explicit tradeoff between the cost of running the particle filter and the mixing of the resulting Markov chain. The practical considerations are considered in Section 4.3. We examine how well the assumptions we make hold in practice in Section 5. In Section 6 a number of applications are examined in detail and a full PMCMC analysis is conducted.

4.1 Scalar representation

We will first show that the high dimensional space on which the MCMC scheme works, in Section 2.1, can be replaced by an equivalent lower dimensional space. Specifically, in Section 2.1, the Markov chain is on θ and the high dimensional object u where the joint invariant density, i.e. posterior density, from which we sample is $\pi_N(\theta, u)$ of (2). It is entirely equivalent to think of the chain, in theory, operating on a lower dimensional space consisting of θ and a scalar z.

Let $z = \log \widehat{p}_N(y|\theta, u) - \log p(y|\theta)$ be the error in the log likelihood due to the simulated likelihood estimator, where u is the vector of random variables generated in the construction of the particle filter with N particles and is defined in Section 2. We note that z is a many to one scalar function of u for any given y and θ, which we write as $z = \psi(u; \theta)$, and suppress the fixed values of y from this expression. Let $g_N(z|\theta)$ be the density of z given θ when u is generated from $p(u)$ and $z = \psi(u; \theta)$. This reduction in terms of z will be useful as we know something about the properties of this density $g_N(z|\theta)$ as N becomes large. It should be noted that whilst we can think of the MCMC scheme operating on this joint space in theory, in practice we cannot evaluate z as we do not know $\log p(y|\theta)$, the true log-likelihood at any parameter ordinate.

The following lemma expresses the proper joint posterior density of θ and z, which we denote as $\pi_N(\theta, z)$, in terms of $g_N(z|\theta)$. The density $\pi_N(\theta, z)$ is the invariant density arising from the PMCMC scheme. The conditional posterior density $\pi_N(z|\theta)$ corresponds to this joint density and is also proper.

Lemma 1. *Define $z = \log \widehat{p}_N(y|\theta, u) - \log p(y|\theta)$ for the estimator $\widehat{p}_N(y|\theta, u)$ of Sections 3.1 and 3.2. Then:*

(i) $\pi_N(z|\theta) = \exp(z) g_N(z|\theta)$.

(ii) $\pi_N(\theta, z) = \pi(\theta)\exp(z)g_N(z|\theta)$.

We can now think of proposing θ^* from a proposal density $q(\theta|\theta_j)$ and z^* from $g_N(z|\theta^*)$ (by transforming from u^*) where the current values are (θ_j, z_j), accepting the proposed pair with probability

$$\alpha(\theta_j, z_j; \theta^*, z^*) = \min\left\{1, \frac{\exp(z^*)\pi(\theta^*)}{\exp(z_j)\pi(\theta_j)}\frac{q(\theta_j|\theta^*)}{q(\theta^*|\theta_j)}\right\}. \tag{9}$$

In practice, we use the entirely equivalent expression (5) by noting that

$$\pi(\theta)\exp(z) = \widehat{p}_N(y|\theta, u)p(\theta)/p(y). \tag{10}$$

As the proposal for θ and the acceptance criteria are the same as used in the criteria (5), the reduced chain in our object of interest $\{\theta_j\}$ remains preserved.

The advantage of regarding the chain as operating on θ and the scalar z is that we can use our knowledge of the properties of z to inform us of how rapidly or slowly the reduced chain in $\{\theta_j\}$ mixes in sampling from the invariant density $\pi(\theta)$.

4.2 Asymptotic approximation of Z

Let N again be the number of samples used for the simulated likelihood estimator and $\sigma^2(\theta, N) = \text{Var}\{\psi(u;\theta)\}$, with y and θ held fixed and u arising from $p(u)$. So $\sigma^2(\theta, N) = \text{Var}_{g_N(z|\theta)}(z)$, where z and $g_N(z|\theta)$ are defined in the previous section.

Lemma 2 shows that $g_N(z|\theta)$ and $\pi_N(z|\theta)$ tend to normality as N increases and that $\sigma^2(\theta, N)$ is the only parameter that affects their distributions for large N.

Lemma 2. *Define* $\sigma^2(\theta, N) = Var(z)$ *where* $z = \log\widehat{p}_N(y|\theta, u) - \log p(y|\theta)$ *for the estimator* $\widehat{p}_N(y|\theta, u)$ *of Sections 3.1 and 3.2. Then, under the conditions for asymptotic normality of the estimators stated in Sections 3.1 and 3.2:*

(i) $N\sigma^2(\theta, N) \to \gamma^2(\theta)$ *as* N *becomes large; or, less formally, we shall write* $\sigma^2(\theta, N) = \gamma^2(\theta)/N$ *for* N *large.*

(ii) For given θ*, suppose that* $Z_N \sim g_N(z|\theta)$*. Then,*

$$\sqrt{N}\left(\frac{Z_N + \frac{\gamma^2(\theta)}{2N}}{\gamma(\theta)}\right) \xrightarrow{d} \mathcal{N}(0; 1)$$

as N *becomes large; or, less formally,*

$$Z_N \xrightarrow{d} \mathcal{N}\left(-\frac{\gamma^2(\theta)}{2N}; \frac{\gamma^2(\theta)}{N}\right),$$

as N *increases.*

(iii) For given θ*, suppose that* $Z_N \sim \pi_N(z|\theta)$*. Then,*

$$\sqrt{N}\left(\frac{Z_N - \frac{\gamma^2(\theta)}{2N}}{\gamma(\theta)}\right) \xrightarrow{d} \mathcal{N}(0; 1)$$

as N becomes large; or, less formally,

$$Z_N \xrightarrow{d} \mathcal{N}\left(\frac{\gamma^2(\theta)}{2N}; \frac{\gamma^2(\theta)}{N}\right),$$

as N increases.

As an aside, we note that via the prediction decomposition for the simulated likelihood obtained by the particle filter,

$$Z = \log \widehat{p}_N(y|\theta; u) - \log p(y|\theta) = \sum_{t=1}^{T} (\widehat{p}_N(y_t|y_{1:t-1}, \theta; u) - \log p(y_t|y_{1:t-1}, \theta)) \tag{11}$$

where $\widehat{p}_N(y_t|y_{1:t-1}, \theta; u)$ is given at (7) defined for the general auxiliary particle filter in the appendix, Section 8.2. This suggests that there is a central limit theorem for Z not only in N but also in T, the length of the time series. This observation indicates that the theoretical results should work particularly well for long time series. This is illustrated by the example considered in Section 5.1 and the results displayed in Figure 4, for which various values of N and T are considered.

From now on we shall just consider a single move in the PMCMC scheme representing the current state as (θ', z') which we consider as distributed according to the invariant joint distribution $\pi_N(\theta, z)$ given by Lemma 1(ii). We will make several assumptions at this stage.

Assumption 1. *The proposal density for θ is its posterior distribution $\pi(\theta)$, i.e., $q(\theta|\theta') = \pi(\theta)$.*

This assumption allows us to separate out the effect of the particle filter on the sampling scheme from that of the quality of the proposal density for θ. It also allows us to study the properties of the MCMC sampling schemes under ideal conditions. The joint proposal density is then $\pi(\theta)g_N(z|\theta)$. Noting equation (10), the Metropolis-Hastings expression at equation (9) reduces to

$$\alpha(\theta', z'; \theta^*, z^*) = \min\{1, \exp(z^* - z')\}. \tag{12}$$

When N is large, both the proposed value z^* and the current value z' will be close to zero and so proposals will be accepted frequently. In fact it can be seen that under Assumption 1 the resulting PMCMC scheme is a Markov chain on z where z^* is proposed from $g_N(z)$ and the current value z' arises from $\pi_N(z)$ where

$$g_N(z) = \int g_N(z|\theta)\pi(\theta)d\theta \quad \text{and} \quad \pi_N(z) = \int \pi_N(z|\theta)\pi(\theta)d\theta.$$

We may then express the Metropolis term of (12) as $\alpha(z'; z^*)$.

Assumption 2. *For a given θ and $\tau^2 > 0$, let the number of particles N be a function of θ and τ^2, which we write as $N = N(\theta, \tau^2)$, such that $N(\theta, \tau^2) = \gamma^2(\theta)/\tau^2$. The term $\gamma^2(\theta)$ is defined in* (6) *and* (8).

Constructing such an 'idealized' choice of the number of particles allows us to keep the variance of z constant across different values of θ. Thus, suppose that the target variance is τ^2. Then $\sigma^2(\theta, N(\theta, \tau^2)) = \sigma^2(\theta', N(\theta', \tau^2)) = \tau^2$ for all θ. This makes it is possible to discuss various summaries of the sampling scheme such as the probability of acceptance of the independent Metropolis Hastings proposal, the inefficiency factors, the computing time and the optimal choice of N as functions of σ only. Because of our choice of N, from now on we will use σ^2 as both the variance of z and as the function $\sigma^2(\theta, N)$. Thus, we will have that $\sigma^2(\theta, N(\theta, \tau^2 = \sigma^2)) = \sigma^2$.

Assumption 3. *Both $g_N(z|\theta)$ and $\pi_N(z|\theta)$ are normal as given by (the asymptotic in N) parts (ii) and (iii) of Lemma 2.*

Assumptions 2 and 3 mean that both $g_N(z|\theta)$ and $\pi_N(z|\theta)$ are normal and only functions of $\sigma^2(\theta, N(\theta, \sigma^2)) = \sigma^2$. For simplicity we shall write $g_N(z|\theta)$ as $g(z|\sigma)$ and $\pi_N(z|\theta)$ as $\pi(z|\sigma)$, when there is no ambiguity in such expressions. Because we deal with the difference $z^* - z'$ in the Metropolis-Hastings expression (12) we can, without loss of generality, add $\sigma^2/2$ to the mean of z in the densities $g_N(z|\theta)$ and $\pi_N(z|\theta)$ so that Assumption 3 becomes

$$g_N(z|\sigma) = \frac{1}{\sigma}\phi\left(\frac{z}{\sigma}\right), \quad \pi_N(z|\sigma) = \frac{1}{\sigma}\phi\left(\frac{z-\sigma^2}{\sigma}\right), \tag{13}$$

where $\phi(\cdot)$ is the standard normal probability density function. Note again, that we may think of $g_N(z|\sigma)$ as the density of the proposal and $\pi_N(z|\sigma)$ as the density of the accepted (current) values of z. Let

$$\Pr(A|z', \sigma) = \int \alpha(z'; z) g(z|\sigma) dz.$$

be the probability of accepting the proposal conditional on z', where $\alpha(z'; z)$ is given by the right hand side of (12) and let

$$\Pr(A|\sigma) = \int \Pr(A|z)\pi(z|\sigma) dz.$$

be the unconditional probability of accepting the proposal. The following results hold,

Lemma 3. *Under Assumptions 1 to 3, and with $z' = \psi(u'; \theta')$,*

(i)

$$\Pr(A|z', \sigma) = \Phi\left(-\frac{z'}{2}\right) + \exp\left(-z' + \frac{\sigma^2}{2}\right)\Phi\left(\frac{z'}{\sigma} - \sigma\right),$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

(ii)

$$\Pr(A|\sigma) = 2\Phi\left(-\frac{\sigma}{\sqrt{2}}\right) = 2\left\{1 - \Phi\left(\frac{\sigma}{\sqrt{2}}\right)\right\}.$$

Part (i) thus gives the probability of accepting the proposal given z' and Part (ii) gives the unconditional probability of accepting the proposal. Figure 5 plots $\Pr(A|\sigma)$ against σ and shows that the acceptance rate is virtually 0 if σ exceeds 3.

Corollary 1. $\Pr(A|Z', \sigma) \to 0$ *as* $z' \to \infty$. *This means that the Markov chain is not uniformly ergodic as the probability cannot be uniformly bounded away from zero, (see Roberts & Tweedie, 1996, Proposition 5.1, p 103).*

A more general result is established for the MCMC methods applied to importance sampling problems in Theorem 8 of Andrieu & Roberts (2009). This property means that for large σ (small N), the PMCMC scheme may reject many proposals for some values of the current state leading to stickiness in the resulting chain.

We now derive an expression for the inefficiency $\mathrm{IF}(\sigma)$ of the sampling scheme as a function of σ. The inefficiency of the sampling scheme is the factor by which it is necessary to increase the number of iterations in the Markov chain Monte Carlo in order to obtain the same accuracy as a sampling scheme that generates independent iterates. Suppose we are interested in posterior inference about some scalar functional $\zeta = h(\theta)$ of θ. Let $\{\theta_j, j = 1, \ldots, M\}$ be the iterates of the

particle filter Markov chain Monte Carlo sampling scheme after it has converged. Suppose that $\rho_\tau(\sigma)$ is the autocorrelation between ζ_j and $\zeta_{j+\tau}$. Then $\mathrm{IF}(\sigma)$ is defined as

$$\mathrm{IF}(\sigma) = 1 + 2\sum_{\tau=1}^{\infty} \rho_\tau(\sigma). \tag{14}$$

$\mathrm{IF}(\sigma)$ is also known as the integrated autocorrelation time (IACT). Let $\overline{\zeta}$ be the sample mean of the iterates $\zeta_j, j = 1, \ldots, M$, which we use as an estimate of the posterior mean of $h(\theta)$. Then,

$$M\mathrm{Var}(\overline{\zeta}) \to \mathrm{Var}(\zeta|y)\mathrm{IF}(\sigma) \text{ as } M \to \infty.$$

We note that if the ζ_j are independent then $\rho_j(\sigma) = 0$ for $j \geq 1$ and $\mathrm{IF}(\sigma) = 1$. The following lemma gives a computable expression for $\mathrm{IF}(\sigma)$.

Lemma 4. *Under the assumptions in this section,*

$$IF(\sigma) = \int \frac{1 + p^*(w,\sigma)}{1 - p^*(w,\sigma)} \phi(w) dw \tag{15}$$

where $p^*(w,\sigma) = \Phi(w+\sigma) - \exp(-w\sigma - \sigma^2/2)\Phi(w)$. *This result also means that* $IF(\sigma)$ *is invariant to the functional* ζ.

Thus, under our assumptions, the posterior mean of $N(\theta, \sigma^2)$ for given σ is

$$E_{\pi(\theta)}[N] = E_{\pi(\theta)}[\gamma^2(\theta)/\sigma^2] = \overline{\gamma^2}/\sigma^2.$$

Therefore, taking into account statistical inefficiency (IF), the computing time is proportional to $\mathrm{IF}(\sigma)/\sigma^2$. Thus, without loss of generality (for our purposes) we take the computing time as $\mathrm{CT}(\sigma) = \mathrm{IF}(\sigma)/\sigma^2$.

The next lemma gives the minimizing value for σ, as well as the behavior of $\mathrm{IF}(\sigma)$ and $\mathrm{CT}(\sigma)$ for large σ.

Lemma 5. *(i) The computing time* $CT(\sigma)$ *is minimized for* $\sigma = 0.92$. *For this value of* σ, $IF(\sigma) = 4.54$ *and* $\Pr(A|\sigma) = 0.5153$.

(ii) For $\sigma^2 \to \infty$

$$\frac{IF(\sigma)}{2\exp(\sigma^2) - 1} \to 1 \tag{16}$$

$$\frac{\sigma^2 CT(\sigma)}{2\exp(\sigma^2) - 1} \to 1 \tag{17}$$

which means that for σ *large*

$$IF(\sigma) \approx 2\exp(\sigma^2) - 1 \quad \textit{and} \quad CT(\sigma) \approx \frac{2\exp(\sigma^2) - 1}{\sigma^2}.$$

It is straightforward to verify that equations (16) *and* (17) *also hold for* $\sigma \to 0$ *because* $IF(\sigma) \to 1$ *as* $\sigma \to 0$.

It is necessary to interpret the results in equations (16) and (17) with some care. The asymptotics rely on $\sigma^2 \to \infty$ in which case the number of particles tends to 0, so that at extreme values of σ this

approximation and the exact analytics may not be accurate because the Gaussian approximations for $g(z|\sigma^2)$ and $\pi(z|\sigma)$ given in (13) may not hold for a small number of particles. However, over the range of σ of practical importance (practically it is problematic to have $\sigma > 3$) these approximation may be extremely good as shown in Figure 1.

Figure 1 shows that for $\sigma > 2.5$ both $\mathrm{IF}(\sigma)$ and $\mathrm{CT}(\sigma)$ are large and that small changes in σ change both $\mathrm{IF}(\sigma)$ and $\mathrm{CT}(\sigma)$ appreciably. Figure 2 plots $\mathrm{CT}(\sigma)$ and $\log \mathrm{CT}(\sigma)$ against σ for σ in the range 0.5 to 1.5 and shows that in this range, small changes in σ result in relatively small changes in $\mathrm{CT}(\sigma)$. This is important because even if we estimate σ with a small error so that we are not at the optimal σ, the effect on $\mathrm{CT}(\sigma)$ will not be serious.

4.3 Computational considerations

In the previous section we chose the number of particles $N(\theta, \sigma^2)$ as a function of θ for a given σ so that $\sigma^2(\theta, N(\theta, \sigma^2)) = \sigma^2$ and then derived the optimal choice of σ to minimize the computing time. In practice this is infeasible because obtaining estimates of the variation $\gamma^2(\theta)$ for each parameter ordinate θ is prohibitively expensive. This section discusses a practical way of choosing N and informally discusses how to relate the results for the inefficiency and computing time in Section 4.1 to the empirical measures of inefficiency and computing time we would obtain if Assumptions 1 and 3 hold.

Our solution is to hold N constant so that the standard deviation of the particle filter log-likelihood estimator evaluated at a central value of the posterior distribution of θ is 0.92. In other words, we choose $N = N_{\mathrm{opt}}$ so that

$$\sigma^2(\overline{\theta}, N_{\mathrm{opt}}) = \frac{\gamma^2(\overline{\theta})}{N_{\mathrm{opt}}} = 0.92^2,$$

where $\overline{\theta}$ might be chosen as the posterior sample mean or componentwise posterior median under $\pi(\theta)$. That is, $N_{\mathrm{opt}} = \gamma^2(\overline{\theta})/0.92^2$.

We now discuss how to obtain $\overline{\theta}$ and evaluate $\gamma^2(\overline{\theta})$. A reasonable strategy, which we have pursued, is to first run a short MCMC scheme for large N to determine an approximate value for $\overline{\theta}$. Then a second short run at this fixed ordinate for a given N allows us to estimate $\gamma^2(\overline{\theta})$. From this the optimal value N_{opt} is determined and the full MCMC scheme implemented. Alternatively, during burn-in, we can have a fairly large initial value for N and take multiple runs (say 4 or 5) of the PF for each accepted value of θ_j. By averaging the variances of the log-likelihood estimator over this period we can get an estimate of $E_{\pi(\theta)}[\gamma^2(\theta)/N]$. This again allow us to accurately deduce the correct value for N.

We now informally relate the inefficiency and computing time measures defined in the previous section to the actual measures that we use in the reported empirical work. For any θ we define

$$\begin{aligned}\sigma^2_{opt}(\theta) &= \sigma^2(\theta, N_{\mathrm{opt}}) \\ &= \frac{\gamma^2(\theta)}{\gamma^2(\overline{\theta})} \times 0.92^2 = \frac{c^2(\theta)}{c^2(\overline{\theta})} \times 0.92^2.\end{aligned}$$

Let the expression for $\mathrm{IF}\{\sigma^2_{opt}(\theta)\}$ be computed as in Lemma 4. The posterior mean of $\mathrm{IF}\{\sigma^2_{opt}(\theta)\}$ is $\overline{\mathrm{IF}} = E_{\pi(\theta)}(\mathrm{IF}\{\sigma^2_{opt}(\theta)\})$. $\overline{\mathrm{IF}}$ will be the actual inefficiency which results from $\sigma^2_{opt}(\theta) = \sigma^2(\theta, N_{opt})$ where θ arises from $\pi(\theta)$. The corresponding posterior mean of the computing time is $\overline{\mathrm{CT}} = N_{\mathrm{opt}}\overline{\mathrm{IF}}$. These relate the empirical measures of inefficiency and computing time we would obtain if Assumptions 1 and 3 hold and apply to Figures 1 and 2.

Crucially, it is important for finite T (where T is the length of the time series) that $\sigma^2_{opt}(\theta)$ does not vary outside the range $(0.25, 2.25)$ during the MCMC or equivalently $\sigma_{opt}(\theta)$ is between 0.5 and 1.5, see Figure 2. Within this region the computing time varies little. This is actually a fairly wide region, a 9:1 ratio in variance, and even for the examples we have examined for moderate T the variability of $\sigma^2_{opt}(\theta)$ meant that we were comfortably in this range.

To understand whether $\sigma^2_{opt}(\theta)$ is likely to lie in the range above we now consider the posterior distribution of $\sigma^2_{opt}(\theta)$ for finite T, and in particular when T is large. Clearly, under the usual regularity conditions for $\pi(\theta)$, the posterior $Var(\overline{\theta}|y) \propto 1/T$ and so the variance of $\sigma^2_{opt}(\theta)$ will also reduce at rate $1/T$, so that as $T \to \infty$

$$\sigma^2_{opt}(\theta)|y \xrightarrow{d} \mathcal{N}\left(0.92^2; \frac{V(\overline{\theta})}{T}\right)$$

for some function V of θ, which means that at least when T becomes large, $\sigma^2_{opt}(\theta)$ will lie in the required range.

We complete this section by giving an informal argument on why the computational load of PMCMC is $O(T^2)$. Equation (11) and Table 4 suggest that $\mathrm{Var}(Z)$ rises linearly with T. As the computational load of the particle filter with N_{opt} particles is $O(N_{\mathrm{opt}}T)$, it follows that the corresponding computational load is then approximately $O(T^2)$.

5 Performance of the theory in the 'finite' N case

This section compares the large sample (in N) results developed above with results for a small to moderate number of particles N for three models. All three are signal plus noise models for which it is possible to run the standard particle filter and the fully adapted particle filter (described in Section 8.2). The finite sample results are obtained by simulation using several different choices for the number of particles using a fixed value of θ. We take a fixed value of θ as this fixes $\sigma^2 = \gamma^2(\theta)/N$ and leads to the ideal setting corresponding to the "perfect proposal" in Assumption 1 where the Metropolis-Hastings acceptance probability is given at (12).

For the first order autoregressive (AR(1)) model plus noise example the true log likelihood is obtained by the Kalman filter. For the other two models the true log likelihood used to construct the z from the particle filter was approximated by the sample mean of $M = 50,000$ unbiased estimates, where we construct each unbiased estimate using $N = 10,000$ particles for the standard particle filter and $N = 500$ particles for the fully adapted particle filter.

5.1 AR1 plus noise model

We consider the AR(1) plus noise model as a simple example to compare the relative performance of the standard SIR method and the fully adapted particle filter (FAPF). The model is

$$y_t = x_t + \sigma_\varepsilon \varepsilon_t, \quad x_{t+1} = \phi x_t + \sigma_\eta \eta_t, \quad x_0 \sim \mathcal{N}(0, \sigma^2_\eta/(1-\phi^2)) \tag{18}$$

where ε_t and η_t are standard normal and independent. We take $\phi = 0.6$, $\sigma^2_\eta = (1-\phi^2)$, so that the marginal variance of the state x_t is $\sigma^2_x = 1$. We simulate a single series of length $T = 200$, varying the measurement noise σ^2_ε (with fixed innovations ε_t) for the experiment. We take $N = 50$ and record the bias and the variance of the logarithm of the estimator of the likelihood for the two particle filter methods, SIR and the fully adapted particle filter (FAPF) of Pitt & Shephard (1999). The algorithm for the FAPF method is specified in Section 8.2 and more details are provided

in Pitt & Shephard (2001). In this case we can evaluate $p(y_{t+1}|x_t)$ explicitly and simulate from $p(x_{t+1}|x_t, y_{t+1})$, the two requirements for using the FAPF method. As the FAPF method essentially guides the particles using the future observation and integrates over the corresponding state when estimating the predictive density of the observations, the estimation of the likelihood should be much more efficient.

The bias and the variance are computed using 400 independent runs of both the SIR and FAPF filters. The true likelihood for the data is given by the Kalman filter as the model is of simple state space form. Figure 3 plots the variance and the bias in the log-likelihood estimator against σ_ε^{-1}. It is apparent that, as expected from (8), the bias downwards is about half the variance. It is also clear that as the measurement standard deviation σ_ε becomes smaller (so σ_ε^{-1} is larger) the standard SIR method does increasingly poorly whereas the performance of the FAPF actually improves. In all cases, it is found that the FAPF has smaller variance and bias. However, in examining the relative variances, given as the last row of Figure 3, it is apparent that the FAPF is dramatically better for small σ_ε (large σ_ε^{-1}). The range of σ_ε chosen is not particularly extreme (no smaller than 1/9). We have found that this result applies to other models which may be fully adapted, see Sections 5.2 and 5.3, as the measurements becomes more informative. The variance of course, directly translates into the necessary number of particles, as we need to take N such that the variance is around 0.85. Thus, Figure 3 suggests that as σ_ε goes down towards about 1/9, the SIR method requires around $2,000$ more particles than the FAPF method.

For models where the measurements are informative but which cannot be fully adapted, the estimator resulting from applying the general auxiliary particle filter, see Section 8.2, can be used. There are several effective ways of constructing auxiliary proposals, see for example Pitt & Shephard (2001). For relatively uninformative measurement equations the gains over the standard SIR method may well be modest.

Figure 4 displays the histogram (over 10,000 replications of the filter) of the log likelihood error $z = \log \widehat{p}_N(y|\theta) - \log p(y|\theta)$ and the fitted Gaussian distribution for the standard particle filter using the estimated mean and the estimated variance σ^2 (calculated using the 10,000 replications of z) only. We use $T = 50$ and 500 and varying N for $\sigma_\varepsilon = \sqrt{2}$. The asymptotic approximating density, which is $\mathcal{N}(-\sigma^2/2; \sigma^2)$, is also displayed using the estimated variance only. Figure 4 (in the boxes) also displays the standard deviations. It is clear that even for small T, the asymptotic Gaussian approximation is good when σ is close to one for the log-likelihood estimator. As σ increases (N decreases) the approximation is a little worse. For large T, the approximation appears to be good even for small N. On the likelihood scale (left hand side), displaying $e^z = \widehat{p}_N(y|\theta)/p(y|\theta)$, N needs to be very large for the density to be close to Gaussian. Of course our approximation, in Section 4.2, relies on the asymptotics for the log of the estimator $\log \widehat{p}_N(y|\theta; u)$. In particular we want this approximation to be good around the region of optimisation for σ (close to one). We have found that the approximation holds up equally well for the other examples considered in this paper. In particular, it improves as T becomes moderately large.

5.2 A mixture of autoregressive experts observed with noise model

We consider a two-component mixture of experts model observed with noise, where each expert is modeled as a first order autoregressive process. Section 6.3 motivates the model and applies it to

GDP growth data. The model is given by

$$y_t = x_t + \sigma_\varepsilon \varepsilon_t \tag{19a}$$

$$x_t = c_{J_t} + \phi_{J_t} x_{t-1} + \tau_{J_t} \eta_t, \quad \text{for} \quad J_t = 1, 2, \tag{19b}$$

$$\Pr(J_t = 1 | x_{t-1}, x_{t-2}) = \frac{\exp(\xi_1 + \xi_2 x_{t-1} + \xi_3 (x_{t-1} - x_{t-2}))}{1 + \exp(\xi_1 + \xi_2 x_{t-1} + \xi_3 (x_{t-1} - x_{t-2}))}, \tag{19c}$$

where ε_t and η_t are standard normal and independent. The means of the first and second autoregressive experts are $\mu_1 = c_1(1 - \phi_1)$ and $\mu_2 = c_1(1 - \phi_2)$. To identify the two experts we assume that $\mu_1 < \mu_2$ so the first expert has a lower mean than the second expert. This is a state space model with a two dimensional state vector (x_t, x_{t-1}).

We generate a single series of length $T = 100$ from the model where the parameters are set as $\sigma_\varepsilon = 0.5$, $\phi_1 = 0.58$, $\phi_2 = 0.32$, $\log(\tau_1^2) = 0.54$, $\log(\tau_2^2) = 0.255$, $c_1 = -0.11$, $c_2 = 2.17$, $\xi_1 = -0.80$, $\xi_2 = -2.33$ and $\xi_3 = -1.53$. With the exception of σ_ε, they are the posterior means from the GDP growth example analyzed in Section 6.3. The parameter σ_ε is chosen to allow a moderate to high signal to noise ratio in the data.

We carry out a PMCMC scheme keeping the parameters fixed so that the Metropolis expression is given by (12). We record the proposed and accepted values of z. Table 1 summarizes the results of the simulation. From the theory of Section 4.2 we expect the mean of the proposed values of z arising from the PF to be around $-\sigma^2/2$ and the mean of the accepted values of z to be around $\sigma^2/2$, where σ^2 is the estimated variance of the proposed draws. The table shows that for the standard particle filter the results for the proposed and posterior means and variances of Z are close to the asymptotic (in N) theoretical results for $N \geq 400$ and that the standard deviation of the proposed z is close to 1 for $N = 400$. The table also shows that the empirical results for the fully adapted particle filter are close to the theoretical results for N as low as 25, and the standard particle filter needs more than 40 times the number of particles required by the fully adapted particle filter to achieve the same proposed variance as the fully adapted particle filter.

Figure 5 shows that the estimated and theoretical acceptance probabilities, given by $\Pr(A|\sigma)$ of Lemma 3 (ii), for differing values of σ are close. Figure 6 summarizes the output when the PMCMC is run for two different values of the standard deviation (σ) of z, taking $N = 200$ and $N = 1400$, with the parameters fixed at their true values and using the standard particle filter. The plots show that the empirical densities for $g(z|\sigma)$ and $\pi(z|\sigma)$ and the theoretical densities given by the asymptotic results in parts (iii) and (iv) of Lemma 2 are close, especially when σ is smaller than 1. The correlograms for the accepted draws of z are displayed and decay in the manner the theory indicates.

The results are based on a sample of $100,000$ (independent) draws from the proposal and at least $200,000$ draws from the posterior using the independent Metropolis Hastings algorithm for each number of particles.

5.3 GARCH model observed with noise

This section considers the GARCH(1,1) model observed with Gaussian noise which is a more flexible version of the basic GARCH(1,1) model. This is a simplified version of the factor GARCH model, with x_t the factor; see Fiorentini et al. (2004). Section 6.2 motivates this model, applies it to UK MSCI weekly returns and explains how to run a fully adapted particle filter for it. The model is described as

$$y_t = x_t + \tau \varepsilon_t, \quad x_t | \sigma_t^2 = \sigma_t \eta_t, \quad \sigma_{t+1}^2 = \alpha + \beta x_t^2 + \gamma \sigma_t^2, \quad x_0 \sim \mathcal{N}(0, \alpha/(1 - \beta - \gamma)) \ . \tag{20}$$

We use the estimated posterior means from the analysis in Section 6.2 as the fixed parameter values. These values are $\tau^2 = 0.00027$, $\alpha = 0.0000495$, $\beta = 0.89275$ and $\gamma = 0.03779$. A single dataset is generated from this model of length $T = 526$. Table 2 summarizes the results. The table shows that for the standard particle filter the proposed and posterior means and variances are close to the asymptotic in N theoretical results only for $N \geq 2500$ and that the standard deviation of the proposed Z is 1 for $N = 2500$. For the fully adapted particle filter the empirical results are close to the theoretical results only for $N \geq 250$, and that the fully adapted particle filter needs about 1/25th the number of particles to achieve the same proposed variance as the standard particle filter.

Figure 7 compares the estimated and theoretical acceptance probability for differing values of σ. These results are based on a sample of $50,000$ (independent) draws from the proposal and $50,000$ draws from the posterior through the independent Metropolis Hastings algorithm for each number of particles (as in (12)). Figure 8 summarizes the output when the PMCMC is run for two different values of σ with the parameters fixed at their true values and using the standard particle filter. The plots show that the theoretical and empirical densities $g(z|\sigma)$ and $\pi(z|\sigma)$ are close especially when σ is smaller.

For all three models the theory appears to be remarkably close to what we observe in practice. This is particularly true for σ less than and around 1, the region where we hope our approximation is close as we optimise to achieve $\sigma = 0.92$. In this section we have kept the parameters fixed as we were concerned with the resulting behaviour in the Markov chain for Z. In Section 6, we will sample both the parameters and Z and investigate performance as we vary N and use both the standard and fully adapted particle filters.

6 Comparing the theory with empirical performance for a full PMCMC

The theory considers an idealized situation based on three assumptions. (i) the proposal is perfect in the sense that the proposal density for θ is its posterior and that the proposal is independent of previous iterates; (ii) the standard deviation of the log likelihood is kept constant by adjusting the number of particles for each θ; (iii) the log of the simulated likelihood is assumed to be Gaussian as a function of the particles.

The previous section considered the empirical properties of PMCMC when θ is kept constant. However, when θ is also generated, neither the assumption of a perfect sampling scheme for the parameters nor the adjustment of the number of particles for θ is met in practice.

This section uses simulated and real data to compare the theoretical results with the performance of a full PMCMC. We consider the performance of both the standard particle filter and the fully adapted particle filter and generate the parameters using the adaptive random walk Metropolis sampling scheme of Roberts & Rosenthal (2009) and (or) the adaptive independent Metropolis Hastings scheme of Giordani & Kohn (2010).

6.1 AR1 plus noise model

This section studies the performance of the theory as a guide to choosing the number of particles when the model parameters are also estimated. It also tries to separate out the effect on performance of the PMCMC due to using an estimated likelihood from the effect due to using an imperfect proposal.

We use data generated from the AR(1) plus noise model at (18) with $\sigma_\varepsilon = \sqrt{2}$. This model has two parameters (ϕ, σ_η^2). The prior distribution ϕ is a uniform on $(-1, 1)$; the prior distribution for

σ_η^2 is an inverse gamma density with shape and scale parameters equal to 0.1. We use a single data set with $T = 500$ observations generated from the true model having $\phi = 0.6$ and $\sigma_\eta = \sqrt{1-\phi^2}$.

This model allows us to compare the results obtained by estimating the likelihood using the standard and fully adapted particle filters with the results obtained when the likelihood is evaluated exactly using the Kalman filter. The unknown parameters are generated by two methods. The first is an independent Metropolis-Hastings scheme. The second is a random walk Metropolis scheme. Both methods are described in more detail below. The small number of parameters means that we can construct an independent Metropolis-Hastings scheme for the parameters that has a high acceptance rate when the exact likelihood is evaluated by the Kalman filter and makes it possible to separate out the effect on the acceptance rates, inefficiencies and computing times of the parameter generating part of the PMCMC from that of using the particle filters to estimate the likelihood.

We first ran 100,000 replications using the fixed true parameters of both the standard SIR and the FAPF to compute the mean, variance and standard deviation of $g_N(z|\theta)$ (the 'prior'for z) and $\pi_N(z|\theta)$ (the posterior for z) for different numbers of particles. Table 3 summarizes the results, which show that the optimal number of particles for the standard SIR filter is about 290 and for the FAPF it is about 52.

Next, we report the acceptance rates, inefficiencies (IF) and computing times (CT) when the two unknown parameters (ϕ, σ_η^2) are sampled from their posterior distributions. The acceptance rate of a sampling scheme is defined as the percentage of accepted draws; the inefficiency of the sampling scheme for a given parameter is defined as the variance of the parameter estimate divided by its variance if the sampling scheme generates independent iterates. We estimate the inefficiency factor, also known as the integrated autocorrelation time, for a given parameter as $IF = 1 + 2\sum_{j=1}^{L^\star} \hat{\rho}_j$, where $\hat{\rho}_j$ is the estimated autocorrelation of the parameter iterates at lag j, and L^* is that lag after which the estimated autocorrelations are randomly scattered about 0. This is the empirical estimator for (14). We define the computing time as CT $= N \times$ IF when the simulated likelihood is obtained by a particle filter.

We first discuss how we obtain the acceptance rates and inefficiencies when the Kalman filter is used to compute the true likelihood. We use two sampling approaches to generate the unknown parameters. In the first approach we ran the adaptive random walk Metropolis algorithm (Roberts & Rosenthal, 2009, described in Section 8.5) for 1,000 iterations followed by 100,000 iterations of the adaptive independent Metropolis-Hastings (Giordani & Kohn, 2010, described in Section 8.6) with the last update at 50,000 draws. After discarding the first 50,000 draws of the adaptive independent Metropolis-Hastings sampler, the acceptance rate and the inefficiencies were computed and are shown in Table 4.

We then fixed the proposal distribution, which is a mixture of normals, and ran 200,000 iterations of the independent Metropolis-Hastings algorithm for the standard particle filter and the fully adapted particle filter for differing number of particles. Table 4 summarizes the results, based on all draws, in terms of acceptance rates, inefficiencies and computing times. The results are as expected since they show that the optimal number of particles are close to the number that gives the standard deviation of the log likelihood as 0.92. That is, the optimal number of particles for the standard particle filter is about 290 particles and that should be about 52 particles for the fully adapted particle filter.

We also ran the random walk Metropolis algorithm to generate the parameters for all cases. The random walk proposal was also fixed based on a previous run of the adaptive random walk Metropolis with the exact likelihood computed using the Kalman filter. The results are also reported in Table 4 and suggest that the optimal computing time using the independent Metropolis-Hastings proposal is substantially smaller than that of the random walk Metropolis algorithm.

Figure 10 plots the inefficiencies for the two parameters as a function of σ (the variance of

$g_N(z|\theta)$) for the standard particle filter divided by the corresponding inefficiencies for the MCMC where the likelihood is evaluated exactly by the Kalman filter. We call this RIF. The figure also plots the $RCT = N \times RIF$, which we call the relative computing time. Figure 10 is obtained similarly for the fully adapted particle filter. The figures show that for the standard SIR the relative computing time ranges from about 1000 at the optimal σ to a maximum of about 5000 at the ends of the range for σ. For the FAPF, the corresponding figures are 300 at the optimum σ and 900 at the boundaries of the range of σ. That means that for the standard SIR, the RCT is five times as large on the boundaries as it is at the optimum and is also three to five times larger than the relative computing time of the FAPF.

6.2 GARCH model observed with noise

The GARCH(1,1) model is used extensively to model financial returns (e.g. Bollerslev et al., 1994). In this section we consider the GARCH(1,1) model observed with Gaussian noise, which is described by equation (20). Malik & Pitt (2011) show that the model can be reparameterised as a conditional Markov chain in σ_t^2, where it is easy to generate from $p(\sigma_t^2|\sigma_{t-1}^2; y_t)$. We can obtain the fully adapted version of the particle filter for this model because the the measurement density is $y_t|\sigma_t^2, \tau^2 \sim \mathcal{N}(0; \tau^2 + \sigma_t^2)$, with the factor x_t integrated out. Crucially, as the noise $\tau \to 0$, the standard particle filter becomes increasingly inefficient but the adapted form will become increasingly efficient as $\text{Var}(\sigma_t^2|\sigma_{t-1}^2; y_t)$ becomes 0. 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.

The parameters of the GARCH(1,1) model with noise in equation (20) are required to satisfy the following constraints: $\tau^2 > 0, \alpha > 0, \beta > 0, \gamma > 0$ and $\beta + \gamma < 1$. To facilitate the sampling of the parameters we transform α, β and γ as follows. Let $\phi = \beta + \gamma, \mu = \alpha/(1-\phi)$ and $\lambda = \beta/\phi$. Now put $\theta_1 = \text{logit}(\phi), \theta_2 = \log(\mu)$ and $\theta_3 = \text{logit}(\lambda)$ so that θ_1 to θ_3 are unconstrained. The parameter τ^2 is not transformed. The prior on τ^2 is a half normal with the corresponding normal having standard deviation 10, the prior for θ_1 is $\mathcal{N}(3; 1.5^2)$, the prior for θ_2 is $\mathcal{N}(-1; 4^2)$ and the prior for θ_3 is $\mathcal{N}(0; 5^2)$. These priors are mildly informative.

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. Table 5 summarizes the results of a single long run for each combination of particle filter, number of particles and adaptive Metropolis-Hastings sampling scheme for the parameters $(\tau^2, \theta_1, \theta_2, \theta_3)$. Similar results were obtained for the original GARCH parameterization. The table shows that the fully adapted particle filter performs much better than the standard particle filter both in terms of IF and CT for both adaptive sampling schemes for the parameters. For example, a fully adapted particle filter with 50 particles using the adaptive independent Metropolis Hastings sampler is more efficient than a standard particle filter using adaptive independent Metropolis Hastings and 1000 particles.

Next we consider the optimal choice of N in terms of CT for the adaptive independent Metropolis Hastings sampler. For the standard particle filter, theory combined with Table 2 suggest that the optimal N is a little above $N = 2500$. Similarly, the optimal N for the fully adapted particle filter is a little above $N = 100$. Table 5 confirms these results for the full PMCMC and it is apparent that too small or too large a value of N results in CT rising quite substantially due to many rejections in the former case and expensive computations in the latter

Table 6 summarizes the posterior distributions of the four parameters. The standard deviation of the noise is estimated to be sizable ($\sqrt{E(\tau^2)} = 0.016$ or 1.6%), which makes the model substantially different from a standard GARCH(1,1).

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 Altix XE320 with the analysis in Section 6 carried out as follows. The updating schedule of the adaptive independent Metropolis Hastings was at 100, 200, 500, 1000, 1500, 2000, 3000, 4000, 5000, 10000, 15000, 20000, and 50000 iterates.

6.3 A mixture of experts model of GDP growth observed with noise

Various nonlinear and non-Gaussian features of the business cycle have been noticed since at least Keynes (1936), who believed recessions to be more volatile than expansions, and Friedman (1964), who believed deep recessions to be followed by periods of fast growth. In more recent times several non-linear models have been estimated on real GDP growth or, less frequently, on industrial production and unemployment, in particular Markov switching models (Hamilton, 1989), and various (smooth) threshold models, (e.g. Tong, 1990; Terasvirta, 1994). To facilitate inference, this literature follows standard econometric practice in assuming that GDP growth is measured accurately. However, this seems unlikely even if revised data are used. The revisions in real GDP data from first to final release are in fact so large that one can only suspect that the final data contain sizable measurement errors (Zellner, 1992).

When adding measurement errors to a regime-switching model, sampling all the states conditional on the parameters and vice versa is a viable option as efficient MCMC samplers exist for this type of model; see Giordani et al. (2007) and Pitt et al. (2010). The same is not true of threshold models, however, which would require slow and sometimes unreliable single move samplers; (see Pitt et al., 2010).

This section elegantly solves the problem by using the particle filter to integrate out the states. It illustrates 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 assume that real GDP growth is measured with an error, and that the unobserved underlying process is a mixture of experts (Jacobs et al., 1991). Mixture of experts models are related to smooth threshold models and neural networks, but with a probabilistic rather than deterministic mixing of the experts (or components or regimes). The model is given by the equations at (19) in Section 5.2. Each of the two experts is an AR(1) process (further lags were not needed in our application) with its own intercept, persistence and innovation variance. The first expert is identified as a low growth regime and the constraint $c_1(1-\phi_1)^{-1} < c_2(1-\phi_2)^{-1}$ is imposed by rejection sampling. The probability of the low growth regime is a logistic function of x_{t-1} and $x_{t-1} - x_{t-2}$.

Like most signal plus noise models, this model is fully adapted. In our application the adaptive IMH sampler performed well, quickly reaching an acceptance rate of over 50%.

Data, priors and inference We model the seasonally adjusted US real log GDP annualized growth from 1984 quarter 2 to 2010 quarter 3 (Source: U.S. Department of Commerce: Bureau of Economic Analysis, series GDPC96, last updated 2010-12-22). Each of the ten model parameters has an independent normal prior. The gating function parameters ξ_1, ξ_2, ξ_3 have dispersed priors. The central moments of the other parameters are calibrated on an AR(1) estimated by OLS, with the OLS error variance split equally between measurement error and transition error, except that $E(c_1) < E(c_2)$ to reflect a prior of low and high growth regimes. Prior and posterior means and standard deviations are summarized in Table 7 and Figure 11. The following results stand out: (i) $\ln \sigma_\varepsilon^2$ is sizable ($\sqrt{E(\sigma_\varepsilon^2)} = 1.38$) (ii) x_t in the low growth regimes is more persistent and more volatile than in the high growth regime (iii) the probability of the low growth regime is a negative

function of both x_{t-1} and $x_{t-1} - x_{t-2}$, as expected. In the figure, the first result can be appreciated by comparing y_t and $E(x_t|y_{1:t})$, and the third by comparing y_t and $E(y_{t+1}|y_{1:t})$ for positive and negative values of y_t.

7 Conclusions

Designing an PMCMC scheme involves two considerations. First, it is important to design a good proposal $q(\theta|\theta')$ to ensure that the acceptance probability is reasonably high when the true likelihood is known, i.e. as $N \to \infty$, for the PF estimator. Second, it is crucial, as we have shown, to select the number of particles N appropriately. The choice of the optimal number of particles N in the PMCMC method depends on many factors including the time dimension T, the dimension of the state and the signal to noise ratio. Crucially, we have reduced the problem into a single scalar quantity which is the error in the log of the estimated likelihood (z), for which we know the limiting distribution, in N. This limiting distribution provides an extremely good approximation to the finite N distribution of z over the relevant range of the standard deviation σ.

In practice when the MCMC routine is running it is difficult to determine whether high rejections are a consequence of a bad proposal $q(\theta|\theta')$ or too few particles. Our approach allows N to be chosen quite separately and robustly so that for the hypothetical perfect proposal $q(\theta|\theta') = \pi(\theta)$ the acceptance rate would be around 50% (see Lemma 5). The form of the proposal $q(\theta|\theta')$ can then be determined and if the acceptance rate is very low it will be apparent that this is due to the proposal $q(\theta|\theta')$ rather than the choice of N. This separation is extremely useful in practice as it means there are two separate optimization considerations and two separate problems to resolve.

It may be possible using our framework to explore the properties of more general schemes advocated in Andrieu et al. (2010). In particular, they consider approaches for simulating the smoothed path of the states under for rather general models enabling Gibbs-type sampling methods to be employed. It is hoped that our approach may shed light on the properties of these more general schemes and provide guidelines for the choice of the number of particles, N.

In the applications, the standard particle filter and the fully adapted particle filters have been considered. The better performance (in terms of reducing the variance or equivalently reducing N) for the fully adapted particle filters suggests that for models where full adaption is not possible, the general auxiliary particle filter may prove to be successful. The performance of the general APF will of course depends on how good the state proposal scheme is. However, there are guidelines given on this in Pitt & Shephard (1999) and Pitt & Shephard (2001). Guidelines for the proposals used in the APF have also been considered by, for example, Smith & Santos (2006) in the context of volatility models and by Durham & Gallant (2002) for diffusion models.

Acknowledgement

We would like to thank the referees for improving the presentation and the rigor of the paper. Michael Pitt is grateful for helpful private conversations with Nicholas Chopin, Arnaud Doucet, Gareth Roberts and Neil Shephard as well as the participants of the conference "Hierarchical Models and Markov Chain Monte Carlo" in Crete, 2011. Robert Kohn and Ralph Silva were partially supported by an ARC Discovery grant DP0988579 .

References

Andrieu, C. & 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., & Holenstein, R. (2010). Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society, Series B* **72**, 1–33.

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

Atchadé, Y. & 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.

Billingsley, P. (1985). *Probability and Measure*. Wiley, New York, 3 edition.

Bollerslev, T., Engle, R. F., & Nelson, D. (1994). Arch models. In Engle, R. & McFadden, D., editors, *Handbook of Econometrics*, volume 4, chapter 49, pages 2959–3038. Elsevier, Amsterdam.

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

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

Carvalho, C., Johannes, M., Lopes, H., & Polson, N. (2010). Particle learning and smoothing. *Statistical Science* **25**, 88–106.

Chib, S. & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistican* **49**, 327–35.

Danielsson, J. (1994). Stochastic volatility in asset prices: estimation with simulated maximum likelihood. *Journal of Econometrics* **61**, 375–400.

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

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

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

Doucet, A., Godsill, S., & Andrieu, C. (2000). On sequential monte carlo sampling methods for bayesian filtering. *Statistics and Computing* **10**, 197–208.

Durham, G. & Gallant, A. (2002). Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. *Journal of Business and Economic Statistics* **20**, 297–338.

Fearnhead, P. & 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. & Rubio-Ramírez, J. (2007). Estimating macroeconomic models: a likelihood approach. *Review of Economic Studies* **74**, 1059–1087.

Fiorentini, G., Sentana, E., & Shephard, N. (2004). Likelihood-based estimation of latent generalised ARCH structures. *Econometrica* **72**, 1481–1517.

Flury, T. & Shephard, N. (2011). Bayesian inference based only on simulated likelihood: Particle filter analysis of dynamic economic models. *Econometric Theory* **1**, 1–24.

Friedman, M. (1964). Monetary studies of the national bureau. In *The National Bureau Enters its 45th Year*, volume 44, pages 7–25. National Bureau of Economic Research, New York.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* **57**, 1317–39.

Geweke, J. (2005). *Contemporary Bayesian econometrics and statistics*, volume 537. Wiley-Interscience.

Giordani, P. & 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.

Giordani, P., Kohn, R., & van Dijk, D. (2007). A unified approach to nonlinearity, structural change, and outliers. *Journal of Econometrics* **137**, 112–133.

Gordon, N. J., Salmond, D. J., & 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., & Tamminen, J. (2001). An adaptive metropolis algorithm. *Bernoulli* **7**, 223–242.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* **57**, 357–84.

Hammersley, J. & Handscomb, D. (1964). Monte carlo methods. 1964. *Methuen, London* .

Harvey, A. C., Ruiz, E., & Shephard, N. (1994). Multivariate stochastic variance models. *Review of Economic Studies* **61**, 247–264.

Jacobs, R., Jordan, M., Nowlan, S., & Hinton, G. (1991). Adaptive mixtures of local experts. *Neural Computation* **3**, 79–87.

Kahn, H. & Marshall, A. (1953). Methods of reducing sample size in monte carlo computations. *Journal of the Operations Research Society of America* pages 263–278.

Keynes, J. (1936). *The General Theory of Employment Interest and Money*. Harcourt Brace and Company.

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.

Liesenfeld, R. & Richard, J. (2003). Univariate and multivariate stochastic volatility models: estimation and diagnostics. *Journal of Empirical Finance* **10**, 505–531.

Liu, J. S. & Chen, R. (1998). Sequential monte carlo methods for dynamic systems. *Journal of the American Statistical Association* **93**, 1032–1044.

Lopes, H., Carvalho, C., Polson, N., & Johannes, M. (2011). Particle learning for sequential bayesian computation (with discussion). *Bayesian Statistics* **9**, 317–360.

Malik, S. & Pitt, M. K. (2011). Particle filters for continuous likelihood evaluation and maximisation. *Journal of Econometrics* in press.

Pitt, M., Giordani, P., & Kohn, R. (2010). Bayesian inference for time series state space models. In Geweke, J., Koop, G., & van Dijk, H., editors, *Handbook of Bayesian Econometric*. Oxford University Pres, Oxford.

Pitt, M. & Shephard, N. (2001). Auxiliary variable based particle filters. In de Freitas, N., Doucet, A., & Gordon, N. J., editors, *Sequential Monte Carlo Methods in Practice*, pages 273–293. Springer-Verlag, New York.

Pitt, M. K. & Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association* **94**, 590–599.

Polson, N. G., Stroud, J. R., & Müller, P. (2008). Practical filtering with sequential parameter learning. *Journal of the Royal Statistical Society, Series B* **70**, 413–428.

Roberts, G. & Tweedie, R. (1996). Geometric convergence and central limit theorems for multidimensional hastings and metropolis algorithms. *Biometrika* **83**, 95.

Roberts, G. O., Gelman, A., & Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *Annals of Applied Probability* **7**, 110–120.

Roberts, G. O. & Rosenthal, J. S. (2009). Examples of adaptive mcmc. *Journal of Computational and Graphical Statistics* **18**, 349–367.

Sandmann, G. & Koopman, S. J. (1998). Estimation of stochastic volatility models via Monte Carlo maximum likelihood. *Journal of Econometrics* **87**, 271–301.

Smith, J. & 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.

Terasvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal the American Statistical Association* **89**, 208–218.

Tong, H. (1990). *Non-Linear Time Series: A Dynamical Systems Approach*. Oxford University Press, Oxford.

West, M. & Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*. Springer-Verlag, New York, 2 edition.

Zellner, A. (1992). Commentary. In Belagia, M. & Garfinkel, M., editors, *The Business Cycle: Theories and Evidence: Proceedings of the Sixteenth Annual Economic Policy Conference of the Reserve Bank of St Lois*.

8 Appendix

8.1 Proofs

This section establishes the results of Sections 4.1 and 4.2.

Proof of Lemma 1. We now write $p(u) = p_U(u)$, whenever we need to explicitly identify that we are dealing with the density of u, and recall that the scalar $z = \psi(u;\theta) = \log\{\widehat{p}_N(y|\theta,u)/p(y|\theta)\}$. To avoid ambiguity in this proof we denote the joint posterior density of u and θ as $\widetilde{\pi}_N(u,\theta)$, where this is still defined by the right side of (3). We denote the corresponding conditional density of u given θ as $\widetilde{\pi}_N(u|\theta)$. We will denote the corresponding joint density in the lower dimensional space of z and θ as $\pi_N(z,\theta)$, and the conditional density as $\pi_N(z|\theta)$. From (3), $\widetilde{\pi}_N(u|\theta) = \exp(\psi(u;\theta))p(u)$. Let z^a be an auxiliary term so that there is a one to one transformation from u to (z,z^a). Then we can write $u = u(z,z^a)$ and let $J(z,z^a) = |\det(\partial u(z,z^a)/\partial(z,z^a))|$ be the absolute value of the Jacobian of the transformation. The density of z given θ is $g_N(z|\theta)$ and is given by

$$g_N(z|\theta) = \int p_U(u(z,z^a))J(z,z^a)dz^a,$$

by integrating over the auxiliary variable z^a. Hence,

$$\begin{aligned}\pi_N(z|\theta) &= \int \widetilde{\pi}_N(u(z,z^a)|\theta)J(z,z^a)dz^a \\ &= \exp(z)\int p_U(u(z,z^a))|\theta)J(z,z^a)dz^a \\ &= \exp(z)g_N(z|\theta).\end{aligned}$$

This obtains Part (i) of the lemma. Part (ii) follows from Part (i). □

Proof of Lemma 2. Part (i) is shown in Section 3.1 for the importance sampler and is discussed in Section 3.2 for the particle filter (and also stated in Del Moral, 2004, Proposition 7.4.1, p. 236 and Proposition 9.4.1, p. 301). Part (ii) now follows from the asymptotic (in N) results stated in Section 3.1 for the importance sampler and in Section 3.2 for the particle filter which show that $g_N(z|\theta)$ tends to normality with the specified moments. Part (iii) follows directly from Part (ii) using the following informal argument based on moment generating functions. This derivation has the advantage that it gives us expressions for the mean and variance of $\pi_N(z|\theta)$, which we use as the normalizing constants in the more formal proof below. Let the moment generating function under $g_N(z|\theta)$ be $M_g(s)$. Then, as $N \to \infty$,

$$M_g(s) \longrightarrow \exp\left(-\frac{\gamma^2(\theta)}{2N}s + \frac{\gamma^2(\theta)}{2N}s^2\right).$$

To obtain Part (iii), it is straightforward to establish that the moment generating function under $\pi_N(z|\theta) = \exp(z)g_N(z|\theta)$ is given by

$$M_\pi(s) = M_g(s+1) \longrightarrow \exp\left(\frac{\gamma^2(\theta)}{2N}s + \frac{\gamma^2(\theta)}{2N}s^2\right),$$

which corresponds to a Gaussian density with mean $\gamma^{\,2}(\theta)/2N$ and variance $\gamma^2(\theta)/N$.

We now prove Part (iii) more formally. Let $a_N = \gamma(\theta)/\sqrt{N}$ and $b_N = a_N^2/2$ where we omit to show that both a_N and b_N also depend on θ. Then, $b_N/a_N = a_N/2$. For $Z_N \sim g_N(z|\theta)$, the moment

generating function of $(Z_N + b_N)/a_N$ is $M_g(s/a_N)\exp(sa_N/2)$ which tends to $\exp(s^2/2)$ as $N \to \infty$ by the central limit theorem. This means that $M_g(s/a_N) \to \exp(s^2/2)$ because $\exp(sa_N/2) \to 1$. Now suppose that $Z_N \sim \pi_N(z|\theta)$. Then the moment generating function of $(Z_N - b_N)/a_N$ is

$$\int \exp\left(\frac{s(z-b_N)}{a_N}\right)\exp(z)g_N(z|\theta)dz = M_g(s/a_N+1)(\exp(-a_N s/2) \longrightarrow \exp(s^2/2),$$

because $a_N s/2 \to 0$ and $(s/a_N+1)/(s/a_N) \to 1$ as $N \to \infty$. This completes Part (iii) of the proof. □

Proof of Lemma 3. Part (i) is obtained from

$$\begin{aligned}\Pr(A|z',\sigma) &= \int \min\{1,\exp(z-z')\}g(z|\sigma)dz \\ &= \int_{z'}^{\infty}\frac{1}{\sigma}\phi\left(\frac{z}{\sigma}\right)dz + \exp(-z')\int_{-\infty}^{z'}\exp(z)\frac{1}{\sigma}\phi\left(\frac{z}{\sigma}\right)dz \\ &= \Phi\left(-\frac{z'}{\sigma}\right) + \exp\left(-z'+\frac{\sigma^2}{2}\right)\Phi\left(\frac{z'}{\sigma}-\sigma\right).\end{aligned}$$

We use the following known result to obtain Part (ii). Suppose that $Z_1 \sim N(\mu_1,\sigma^2), Z_2 \sim N(\mu_2,\sigma^2)$ and Z_1, Z_2 are independent. Then

$$\Pr(Z_1+Z_2 \le 0) = \Phi\left(\frac{-\mu_1-\mu_2}{\sqrt{2}\sigma}\right) \quad = \int \Pr(Z_1 \le -z_2|z_2)p(z_2)dz_2 = \int \Phi\left(\frac{-z_2-\mu_1}{\sigma}\right)\frac{1}{\sigma}\phi\left(\frac{z_2-\mu_2}{\sigma}\right)dz_2.$$

From above,

$$\begin{aligned}\Pr(A|\sigma) = &\int \Phi\left(-\frac{z'}{\sigma}\right)\frac{1}{\sigma}\phi\left(\frac{z'-\sigma^2}{\sigma}\right)dz' + \\ &\exp\left(\frac{\sigma^2}{2}\right)\int \exp(-z')\Phi\left(\frac{z'}{\sigma}-\sigma\right)\frac{1}{\sigma}\phi\left(\frac{z'-\sigma^2}{\sigma}\right)dz' \\ &= P_1 + P_2,\end{aligned}$$

where

$$P_1 = \Phi\left(-\frac{\sigma}{\sqrt{2}}\right) \quad \text{and} \quad P_2 = \int \Phi\left(\frac{z'}{\sigma}-\sigma\right)\phi\left(\frac{z'}{\sigma}\right)dz' = \Phi\left(-\frac{\sigma}{\sqrt{2}}\right).$$

□

Proof Lemma 4. We need to calculate the integrated autocorrelation time for $\zeta_j = h(\theta_j)$ for a given function $h(.)$. The Markov chain is actually on the joint space $\{\theta_j, z_j\}$ where the acceptance probability is given by (12) and only depends upon the current value z_j and the proposed value z. Although the chain is on this joint space, we really only need to be concerned with the marginal chain $\{z_j\}$ as, under our assumptions, this is generated independently from θ and the acceptance probability in the Metropolis expression, at equation (12), only depends upon the current and proposed values of z.

Without any loss of generality we will consider the beginning of the chain as $\{\theta_1, z_1\}$. We shall assume that the invariant distribution, i.e. posterior distribution, has been reached so that (θ_1, z_1)

are jointly distributed according to

$$\pi(\theta, z) = \pi(\theta)\pi(z|\sigma) = \pi(\theta)g(z|\sigma)e^z,$$

where we write $\pi_N(z|\theta)$ and $g_N(z|\theta)$ as $\pi(z|\sigma)$ and $g(z|\sigma)$. Let $J = 0$ if there is no jump in the θ iterates in the period 1 to $j+1$, with $J = 1$ otherwise (at least one jump). Let $\overline{p}(z_1, \sigma) = 1 - \Pr(A|z_1, \sigma)$, where $\Pr(A|z_1, \sigma)$ is given in Lemma 3. In this context no jump means that all the Metropolis proposals are rejected and so the value of θ and z will remain the same as θ_1 and z_1. Then the probability of no jump over the interval is $\Pr(J = 0|z_1, \sigma) = \overline{p}(z_1, \sigma)^j$, the probability of j successive rejections. The probability of at least one acceptance in the Metropolis scheme is $\Pr(J = 1|z_1, \sigma) = 1 - \overline{p}(z_1, \sigma)^j$. Then

$$E(\zeta_{j+1}\zeta_1|z_1, \sigma) = E(\zeta_{j+1}\zeta_1|z_1, \sigma, J = 0)\Pr(J = 0|z_1, \sigma) + E(\zeta_{j+1}\zeta_1|z_1, \sigma, J = 1)\Pr(J = 1|z_1, \sigma).$$

We note that ζ_{j+1} and ζ_1 are independent if $J = 1$ as the proposal for θ is the true marginal posterior $\pi(\theta)$ and the proposal for Z is independent of θ. As a consequence,

$$E(\zeta_{j+1}\zeta_1|z_1, \sigma, J = 1) = E_\pi(\zeta)^2 = E_\pi[h(\theta)]^2.$$

Similarly, if there is no jump,

$$E(\zeta_{j+1}\zeta_1|z_1, \sigma, J = 0) = E_\pi(\zeta^2) = E_\pi[h(\theta)^2].$$

So

$$E(\zeta_{j+1}\zeta_1|z_1, \sigma) = E_\pi(\zeta)^2\overline{p}(z_1, \sigma)^j + E_\pi(\zeta^2)\{1 - \overline{p}(z_1, \sigma)^j\},$$

and

$$\begin{aligned}
\mathrm{Cov}(\zeta_{j+1}, \zeta_1|z_1, \sigma) &= \mathrm{E}(\zeta_{j+1}\zeta_1|z_1, \sigma) - \mathrm{E}(\zeta_{j+1}|z_1, \sigma)\mathrm{E}(\zeta_1|z_1, \sigma) \\
&= E_\pi(\zeta)^2\left(1 - \overline{p}(z_1, \sigma)\right)^j + E_\pi(\zeta^2)\overline{p}(z_1, \sigma)^j - E_\pi(\zeta)^2 \\
&= \{E_\pi(\zeta^2) - E_\pi(\zeta)^2\}\overline{p}(z_1, \sigma)^j \\
&= \mathrm{Var}_\pi(\zeta)\overline{p}(z_1, \sigma)^j,
\end{aligned}$$

where the subscripts π indicate expectations and variances under $\pi(\theta)$, the true posterior for θ is the true marginal posterior. Then,

$$\mathrm{Cov}(\zeta_{j+1}, \zeta_1|\sigma) = E_{\pi(z_1|\sigma)}[\mathrm{Cov}(\zeta_{j+1}, \zeta_1|z_1, \sigma)] = \mathrm{Var}(\zeta)E_{\pi(z|\sigma)}[\overline{p}(z, \sigma)^j].$$

Let $\rho_j(\sigma) = \mathrm{Corr}(\zeta_{j+1}\zeta_1|\sigma) = E_{\pi(z|\sigma)}[\overline{p}(z, \sigma)^j]$. Then,

$$\begin{aligned}
IF(\sigma) &= 1 + 2\sum_{j=1}^{\infty}\rho_j(\sigma) = 1 + 2\sum_{j=1}^{\infty}E_{\pi(z|\sigma)}[\overline{p}(z, \sigma)^j] \\
&= E_{\pi(z|\sigma)}\left(1 + 2\sum_{j=1}^{\infty}\overline{p}(z, \sigma)^j\right),
\end{aligned}$$

so that

$$IF(\sigma) = \int \frac{1 + \overline{p}(z, \sigma)}{1 - \overline{p}(z, \sigma)}\pi(z|\sigma)dz.$$

From Part (i) of Lemma 3

$$\begin{aligned}\overline{p}(z,\sigma) &= 1-\Phi\left(-\frac{z}{\sigma}\right)-\exp\left(-z+\frac{\sigma^2}{2}\right)\Phi\left(\frac{z}{\sigma}-\sigma\right)\\ &= \Phi(w+\sigma)-\exp\left(-\sigma w-\frac{1}{2}\sigma^2\right)\Phi(w)\\ &= p^*(w,\sigma)\end{aligned}$$

where $w=(z-\sigma^2)/\sigma$. Equation (15) follows. □

Proof of Lemma 5. Minimizing $\mathrm{CT}(\sigma)$ reduces to solving the first order condition for σ, $\frac{\partial \mathrm{IF}(\sigma)}{\partial\sigma} = \frac{2}{\sigma}\mathrm{IF}(\sigma)$. This can be solved numerically in a straightforward way to give $\sigma = 0.92$. The rest of part (i) follows. Let $p^*(w,\sigma)$ be defined as in Lemma 4 and put $G(w,\sigma) = (1+p^*(w,\sigma))/(1-p^*(w,\sigma))$. Then,

$$G(w,\sigma) = \frac{1+\Phi\left(w+\sigma\right)-\exp\left(-w\sigma-\sigma^2/2\right)\Phi\left(w\right)}{\Phi\left(-w-\sigma\right)+\exp\left(-w\sigma-\sigma^2/2\right)\Phi\left(w\right)} = \frac{\left\{1+\Phi\left(w+\sigma\right)\right\}\exp\left(w\sigma+\sigma^2/2\right)-\Phi\left(w\right)}{\exp\left(w\sigma+\sigma^2/2\right)\Phi\left(-w-\sigma\right)+\Phi\left(w\right)}.$$

Next we use the inequality that for $x>0$, $\Phi(-x)<\phi(x)/x$ to show that for $\sigma+w>0$,

$$0<\exp\left(w\sigma+\sigma^2/2\right)\Phi\left(-w-\sigma\right)<\exp\left(w\sigma+\sigma^2/2\right)\phi(w+\sigma)/(w+\sigma)=\phi(w)/(w+\sigma)=O(\sigma^{-1})$$

$$2>1+\Phi\left(w+\sigma\right)>2-\frac{\phi(-w-\sigma)}{w+\sigma}=2-O(\sigma^{-1}),$$

where $O(\sigma^{-1})$ means that $\sigma O(\sigma^{-1})$ is bounded as $\sigma\to\infty$. Hence,

$$G(w,\sigma)=\frac{(2-O(\sigma^{-1}))\exp\left(w\sigma+\sigma^2/2\right)-\Phi(w)}{O(\sigma^{-1})+\Phi(w)}=\frac{2\exp\left(w\sigma+\sigma^2/2\right)}{\Phi(w)}-1-O(\sigma^{-1}).$$

Hence,

$$\begin{aligned}\mathrm{IF}(\sigma) &= \int G(w,\sigma)\phi(w)dw\\ &= \int\frac{2\exp\left(w\sigma+\sigma^2/2\right)}{\Phi\left(w\right)}\phi\left(w\right)dw-1-O(\sigma^{-1})\\ &= 2\exp\left(\frac{\sigma^2}{2}\right)\int\exp\left(w\sigma\right)\frac{\phi\left(w\right)}{\Phi\left(w\right)}dw-1-O(\sigma^{-1})\\ &= 2\exp\left(\sigma^2\right)\int\frac{\phi\left(w-\sigma\right)}{\Phi\left(w\right)}dw-1-O(\sigma^{-1})\\ &= 2\exp\left(\sigma^2\right)-1-O(\sigma^{-1}).\end{aligned}$$

□

8.2 General ASIR particle filter

The general auxiliary SIR (ASIR) filter of Pitt & Shephard (1999) may be thought of as a generalization 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 densities $g(y_{t+1}|x_t)$ and $g(x_{t+1}|x_t; y_{t+1})$ in Algorithm 2 are approximations to $p(y_{t+1}|x_t)$ and $p(x_{t+1}|x_t; y_{t+1})$ respectively such that we can evaluate $g(y_{t+1}|x_t)$ and generate from $g(x_{t+1}|x_t; y_{t+1})$. Their choice is discussed more fully below. It should be noted that $g(y_{t+1}|x_t)$ can be specified in unnormalised form for the algorithm.

The following algorithm describes the one time step ASIR update and is initialized with samples $x_0^k \sim p(x_0)$ with mass $1/N$ for $k = 1, ..., N$.

Algorithm 2. *Given samples $x_t^k \sim p(x_t|y_{1:t})$ with mass π_t^k for $k = 1, ..., N$.*
For $t = 0, .., T-1$:

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

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

To motivate Algorithm 2, we note that the product density $p(y_{t+1}|x_{t+1})p(x_{t+1}|x_t)$ appears in its derivation and can 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

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

If we can evaluate $p(y_{t+1}|x_t)$ and generate from $p(x_{t+1}|x_t; y_{t+1})$, then we can take $g(y_{t+1}|x_t) = p(y_{t+1}|x_t)$ and $g(x_{t+1}|x_t; y_{t+1}) = p(x_{t+1}|x_t; y_{t+1})$ in 2, which Pitt & Shephard (2001) call the fully adapted form of the auxiliary particle filter. In this case Step 4 becomes redundant as $\omega_{t+1}^k = 1$, ($\pi_{t+1}^k = 1/N$) and the method reduces to what Pitt & 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.

Full adaptation is possible whenever $p(x_{t+1}|x_t)$ is conjugate in x_{t+1} to $p(y_{t+1}|x_{t+1})$. This occurs for example when 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 of the form $p(x_{t+1}|x_t) \sim \mathcal{N}(\mu(x_t), \Sigma(x_t))$.

The SIR method of Gordon et al. (1993) is a special case of Algorithm 2 when we can generate from $p(x_{t+1}|x_t)$, and we take $g(x_{t+1}|x_t) = p(x_{t+1}|x_t)$ and $g(y_{t+1}|x_t) = 1$. In this case, Step 1 above leaves the weights unchanged (as $\pi_{t|t+1}^k = \pi_t^k$).

In general, 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 available for doing this; see for example, Pitt & Shephard (2001) and Smith & Santos (2006).

The general ASIR estimator of $p(y_t|y_{1:t-1})$ that we propose is

$$\widehat{p}_N(y_t|y_{1:t-1}) = \left\{\sum_{k=1}^{N} \frac{\omega_t^k}{N}\right\}\left\{\sum_{k=1}^{N} \omega_{t-1|t}^k\right\}, \tag{21}$$

where, in this section and the next, $\widehat{p}_N(y_t|y_{1:t-1})$ and $\widehat{p}_N(y_{1:t})$ mean $\widehat{p}_N(y_t|y_{1:t-1}, u, \theta)$ and $\widehat{p}_N(y_{1:t}|u, \theta)$. 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 2. For full adaption, $\omega_t^k = 1$ and $\omega_{t-1|t}^k = p(y_t|x_{t-1}^k)/N$ and the first summation in (21) 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 (21) disappears. The derivation of this estimator is given below.

The ASIR algorithm (Algorithm 2) is a flexible particle filter approach when combined with stratification. Theorem 1 in Appendix 8.3 establishes that this algorithm together with the estimator of (21) is unbiased. This is important as it enables very efficient likelihood estimators from the ASIR method to be used within an MCMC algorithm. Our examples use the standard particle filter and the fully adapted particle filter.

We now define some terms that are used in Algorithm 2 and that will be useful for the proof of Theorem 1 and the derivation of the $\widehat{p}_N(y_t|y_{1:t-1})$.

$$\begin{aligned}
\widehat{p}_N(x_t|y_{1:t}) =& \sum_{k=1}^{N} \pi_t^k \delta(x_t - x_t^k), \text{ where } \pi_t^k \text{ is given in Step (4).} \\
\widehat{g}_N(x_t|y_{1:t+1}) =& \sum_{k=1}^{N} \pi_{t|t+1}^k \delta(x_t - x_t^k), \quad \text{where } x_t^k \sim \widehat{p}_N(x_t|y_{1:t}) \qquad (22) \\
& \text{and } \pi_{t|t+1}^k \text{ is defined in Step (1) of the algorithm.} \\
\widehat{g}_N(x_t|y_{1:t}) =& \int g(x_t|\widetilde{x}_{t-1}; y_t)\widehat{g}_N(\widetilde{x}_{t-1}|y_{1:t})d\widetilde{x}_{t-1}, \qquad (23) \\
\omega_{t|t+1}(x_t) =& g(y_{t+1}|x_t)\pi_t, \ \omega_{t+1}(x_{t+1}; x_t) = \frac{p(y_{t+1}|x_{t+1})p(x_{t+1}|x_t)}{g(y_{t+1}|x_t)g(x_{t+1}|x_t, y_{t+1})}.
\end{aligned}$$

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

We now give a derivation of $\widehat{p}_N(y_t|y_{1:t-1})$ at (21). We note that

$$\begin{aligned}
p(y_t|y_{1:t-1}) &\simeq \int\int p(y_t|x_t)p(x_t|x_{t-1})\widehat{p}_N(x_{t-1}|y_{1:t-1})dx_t dx_{t-1} \\
&= \left\{\sum_{k=1}^{N} \omega_{t-1|t}^k\right\} \int \omega_t(x_t; x_{t-1})\widehat{g}_N(x_t|y_{1:t})dx_t
\end{aligned}$$

which leads directly to $\widehat{p}_N(y_t|y_{1:t-1})$.

8.3 Proof that the ASIR likelihood is unbiased

Theorem 1. *The ASIR likelihood*

$$\widehat{p}_N(y_{1:t}) = \widehat{p}_N(y_1) \prod_{t=2}^{T} \widehat{p}_N(y_t|y_{1:t-1}),$$

is unbiased in the sense that $E(\widehat{p}_N(y_{1:t})) = p(y_{1:t}$.

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

Let us define $\mathcal{A}_t = \{x_t^k; \pi_t^k\}$ as the swarm of particles, for $k = 1, ..., N$, at time t. So the particles x_t^k with associated probability π_t^k represent the filtering density $p(x_t|y_{1:t})$ for time t.

Lemma 6.

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

Proof.

$$E[\widehat{p}_N(y_t|y_{1:t-1})|\mathcal{A}_{t-1}] = E\left[\sum_{k=1}^{N} \frac{\omega_t(x_t^k; \widetilde{x}_{t-1}^k)}{N} \mid \mathcal{A}_{t-1}\right] \left\{\sum_{j=1}^{N} \omega_{t-1|t}^j\right\},$$

as the weights $\omega_{t-1|t}^j$ are known given $\mathcal{A}_{t-1}$. So

$$\begin{aligned}
&E[\widehat{p}_N(y_t|y_{1:t-1})|\mathcal{A}_{t-1}] \\
&= \int \omega_t(x_t; \widetilde{x}_{t-1}) g(x_t|\widetilde{x}_{t-1}; y_t) \widehat{g}_N(\widetilde{x}_{t-1}|y_{1:t}) dx_t d\widetilde{x}_{t-1} \left\{\sum_{j=1}^{N} \omega_{t-1|t}^j\right\}, \\
&\text{using the terms (22) and (23),} \\
&= \int \sum_{k=1}^{N} \omega_t(x_t; x_{t-1}^k) g(x_t|x_{t-1}^k; y_t) \frac{\omega_{t-1|t}(x_{t-1}^k)}{(\sum_{j=1}^{N} \omega_{t-1|t}(x_{t-1}^j))} dx_t \left\{\sum_{j=1}^{N} \omega_{t-1|t}^j\right\} \\
&= \int \sum_{k=1}^{N} \omega_t(x_t; x_{t-1}^k) g(x_t|x_{t-1}^k; y_t) \omega_{t-1|t}(x_{t-1}^k) dx_t \\
&= \sum_{k=1}^{N} \int \frac{p(y_t|x_t) p(x_t|x_{t-1}^k)}{g(y_t|x_{t-1}^k) g(x_t|x_{t-1}^k; y_t)} g(x_t|x_{t-1}^k; y_t) g(y_t|x_{t-1}^k) \pi_{t-1}^k dx_t.
\end{aligned}$$

So

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

□

Lemma 7.

$$E[\widehat{p}_N(y_{t-h:t}|y_{1:t-h-1})|\mathcal{A}_{t-h-1}] = \sum_{k=1}^{N} p(y_{t-h:t}|x_{t-h-1}^k)\pi_{t-h-1}^k. \tag{24}$$

Proof. The proof is by induction.

Part A. This is true for $h = 0$ by Lemma 6.

Part B. We shall assume that (24) holds for h and show it then holds for $h + 1$. For the case $h + 1$ the left hand side of (24) is given by,

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

using (24) for the case h and (21). So, recalling the definition of π_t in Step (4) of the algorithm,

$$E[\widehat{p}_N(y_{t-h-1:t}|y_{1:t-h-2})|\mathcal{A}_{t-h-2}] = E\left[\left\{\sum_{k=1}^{N} p(y_{t-h:t}|x_{t-h-1}^k)\frac{\omega_{t-h-1}^k}{\sum_{i=1}^{N}\omega_{t-h-1}^i}\right\}\left\{\sum_{i=1}^{N}\frac{\omega_{t-h-1}^i}{N}\right\} \mid \mathcal{A}_{t-h-2}\right]$$
$$\times \left\{\sum_{j=1}^{N} \omega_{t-h-2|t-h-1}^j\right\},$$

due to the weights $\omega_{t-h-2|t-h-1}^j$ being known given $\mathcal{A}_{t-h-2}$,

$$= E\left[\frac{1}{N}\sum_{k=1}^{N} p(y_{t-h:t}|x_{t-h-1}^k)\omega_{t-h-1}^k \mid \mathcal{A}_{t-h-2}\right]\left\{\sum_{j=1}^{N} \omega_{t-h-2|t-h-1}^j\right\}$$
$$= \left\{\sum_{j=1}^{N} \omega_{t-h-2|t-h-1}^j\right\} \int p(y_{t-h:t}|x_{t-h-1})\omega_{t-h-1}(x_{t-h-1};\widetilde{x}_{t-h-2})$$
$$g(x_{t-h-1}|\widetilde{x}_{t-h-2};y_{t-h-1})\widehat{g}_N(\widetilde{x}_{t-h-2}|y_{1:t-h-1})dx_{t-h-1}d\widetilde{x}_{t-h-2},$$

using the terms (22) and (23),

$$= \left\{ \sum_{j=1}^{N} \omega_{t-h-2|t-h-1}^{j} \right\} \int \sum\nolimits_{k=1}^{N} p(y_{t-h:t}|x_{t-h-1}) \omega_{t-h-1}(x_{t-h-1}; x_{t-h-2}^{k})$$
$$g(x_{t-h-1}|x_{t-h-2}^{k}; y_{t-1}) \frac{g(y_{t-h-1}|x_{t-h-2}^{k}) \pi_{t-h-2}^{k}}{\sum_{j=1}^{N} \omega_{t-h-2|t-h-1}^{j}} dx_{t-h-1}$$
$$= \sum\nolimits_{k=1}^{N} \pi_{t-h-2}^{k} \int p(y_{t-h:t}|x_{t-h-1}) \omega_{t-h-1}(x_{t-h-1}; x_{t-h-2}^{k})$$
$$g(x_{t-h-1}|x_{t-h-2}^{k}; y_{t-1}) g(y_{t-h-1}|x_{t-h-2}^{k}) dx_{t-h-1},$$

using the definition of ω_{t-h-1} (see Step 4 of Algorithm 2),

$$= \sum\nolimits_{k=1}^{N} \pi_{t-h-2}^{k} \int p(y_{t-h:t}|x_{t-h-1}) p(y_{t-h-1}|x_{t-h-1}) p(x_{t-h-1}|x_{t-h-2}^{k}) dx_{t-h-1}$$
$$= \sum\nolimits_{k=1}^{N} p(y_{t-h-1:t}|x_{t-h-2}^{k}) \pi_{t-h-2}^{k} \text{ as required.}$$

□

Proof of Theorem 1. As a consequence we have from Lemma that, with $h = t - 2$,

$$E[\widehat{p}_N(y_{1:t})|\mathcal{A}_0] = \sum_{k=1}^{N} p(y_{1:t}|x_0^k) \pi_0^k,$$

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

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

□

8.4 Theory for adaptive independent Metropolis-Hastings sampling for the parameters

Our article uses the adaptive independent Metropolis Hastings scheme of Giordani & Kohn (2010) and the adaptive random walk Metropolis scheme of Roberts & Rosenthal (2009). We now show that the adaptive independent Metropolis Hastings sampling scheme described in Section 8.6 (and more fully in Giordani & Kohn, 2010), when it is combined with the ASIR particle filter converges to the posterior distribution of θ. The results generalize those of Giordani & Kohn (2010) to the simulated likelihood case. We remark that the adaptive independent Metropolis Hastings scheme of Giordani & Kohn (2010) has the potential for constructing proposals which are as close as possible to $\pi(\theta)$ as the number of iterates increases and the dimension of θ is not too large. This means that the theory of Section 4 will apply to the interpretation of the results in Section 6 when this independence proposal is used.

By equation (21) in Section 8.2, the simulated likelihood is of the form

$$\widehat{p}_N(y_{1:T}|\theta,u) = \prod_{t=1}^{T} \widehat{p}_N(y_t|y_{1:t-1};\theta,u) \tag{25a}$$

$$\widehat{p}_N(y_t|y_{1:t-1};\theta,u) = \left\{\sum_{k=1}^{N} \frac{\omega_t^k(\theta)}{N}\right\}\left\{\sum_{k=1}^{N} \omega(\theta)_{t-1|t}^k\right\} \tag{25b}$$

$$\text{where} \quad \omega_t^k(\theta) = \frac{p(y_t|x_t^k;\theta)p(x_t^k|\widetilde{x}_{t-1}^k;\theta)}{g(y_t|\widetilde{x}_{t-1}^k;\theta)g(x_t^k|\widetilde{x}_{t-1}^k;y_t;\theta)}, \quad \omega_{t-1|t}^k(\theta) = g(y_t|x_{t-1}^k;\theta)\pi_{t-1}^k \tag{25c}$$

$$\text{and} \quad \sum_{k=1}^{N} \pi_{t-1}^k = 1 \ .$$

The particles x_t^k and $\widetilde{x}_t^k$ are defined in Section 8.2. Let Θ be the parameter space of θ.

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

$$g(y_{t+1}|x_t;\theta) \le C, \tag{26}$$

$$\frac{p(y_{t+1}|x_{t+1};\theta)p(x_{t+1}|x_t;\theta)}{g(y_{t+1}|x_t;\theta)g(x_{t+1}|y_{t+1},x_t;\theta)} \le C, \tag{27}$$

$$p(\theta)/q_j(\theta) \le C \ , \tag{28}$$

where $q_j(\theta)$ is the proposal density at the jth iterate. 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*

$$\sup_{A\subset\Theta} \mid \Pr(\theta_j \in A) - \int_A p(\theta \mid y)d\theta \mid \to 0 \quad as \quad j \to \infty \tag{29}$$

   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 $g_2(\theta|\cdot)$ defined in Section 8.6. Then, almost surely,*

$$\frac{1}{n}\sum_{j=1}^{n} h(\theta_j) \to \int h(\theta)p(\theta|y)d\theta \quad as \quad n \to \infty. \tag{30}$$

Proof.

$$\widehat{p}_N(y|\theta,u) = \prod_{t=1}^{T} \widehat{p}_N(y_t|y_{1:t-1};\theta,u) \le C^{2T} \quad \text{because} \quad \widehat{p}_N(y_t|y_{1:t-1};\theta,u) \le C^2$$

from equations (25)(a)–(25)(c) and our assumptions. This shows that the simulated likelihood $\widehat{p}_N(y|\theta,u)$ is bounded and the result now follows from Giordani & Kohn (2010). □

We note that as in Giordani & 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 (28) holds. The next corollary gives a condition for equations (26) and (27) to hold for the standard particle filter and the fully adapted particle filter.

Corollary 2. *Suppose that for all $y_t, x_t, t = 1, \ldots, T$ and $\theta \in \Theta$, there exists a constant $C_1 > 0$ such that*

$$p(y_t|x_t;\theta) \leq C_1. \tag{31}$$

Then equations (26) *and* (27) *hold for the standard particle filter and the fully adapted particle filter.*

Proof. In the fully adapted case, the left side of equation (27) is identically 1, and $g(y_{t+1}|x_t;\theta) = p(y_{t+1}|x_t;\theta)$. For the standard particle filter, $g(y_{t+1}|x_t;\theta) = 1$ and $g(x_{t+1}|y_{t+1}, x_t;\theta) = p(x_{t+1}|x_t;\theta)$. The result follows from

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

□

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

We now construct a partially adapted particle filter that satisfies equations (26) and (27). 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. 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,

$$g(y_{t+1}|x_t;\theta)g(x_{t+1}|x_t, y_{t+1};\theta) = \epsilon p(x_{t+1}|x_t) + (1-\epsilon)g_0(y_{t+1}|x_t)g_0(x_{t+1}|x_t, y_{t+1};\theta) \ . \tag{32}$$

Corollary 3. *Suppose equation* (31) *holds and the partially adapted particle filter is defined by equation* (32)*. Then, equations* (26) *and* (27) *hold. The proof is straightforward.*

We would usually 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.

8.5 Adaptive random walk Metropolis sampling of the parameters

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

$$q_j(\theta;\theta_{j-1}) = \omega_{1j}\phi_d(\theta;\theta_{j-1},\kappa_1\Sigma_1) + \omega_{2j}\phi_d(\theta;\theta_{j-1},\kappa_2\Sigma_{2j}), \tag{33}$$

where d is the dimension of θ and $\phi_d(\theta;\widetilde{\theta},\Sigma)$ is a multivariate d dimensional normal density in θ with mean $\widetilde{\theta}$ and covariance matrix Σ. In (33), $\omega_{1j} = 1$ for $j \leq j_0$, with j_0 representing the initial iterations, $\omega_{1j} = 0.05$ for $j > j_0$ with $\omega_{2j} = 1 - \omega_{1j}$; $\kappa_1 = 0.1^2/d, \kappa_2 = 2.38^2/d, \Sigma_1$ is a constant covariance matrix, which is taken as the identity matrix by Roberts & 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 (5) for the adaptive random walk Metropolis simplifies to

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

8.6 Adaptive independent Metropolis-Hastings sampling of the parameters based on a mixture of normals

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

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

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.

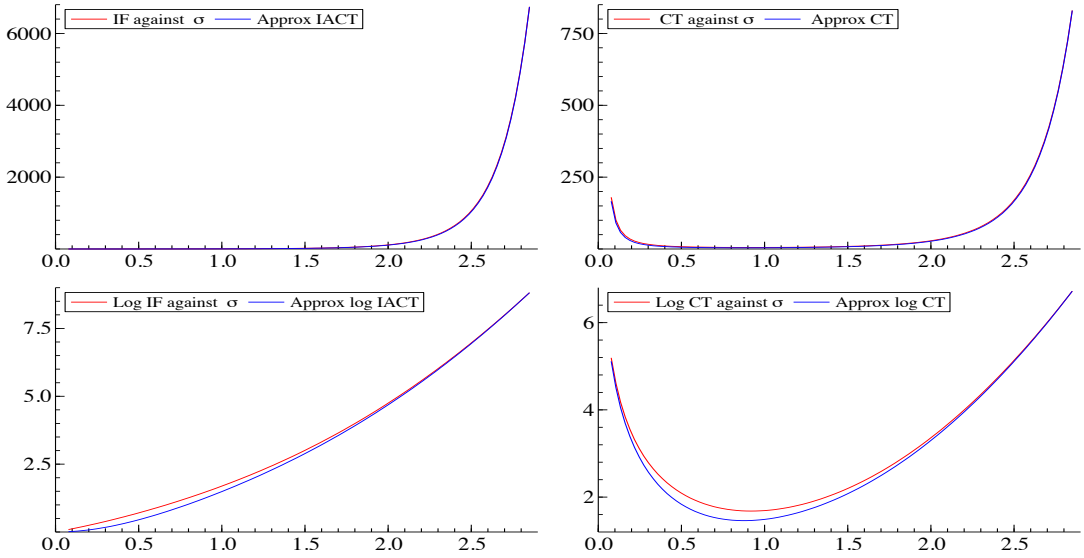


Figure 1: Plots of $\mathrm{IF}(\sigma), \mathrm{CT}(\sigma)$, $\log \mathrm{IF}(\sigma)$ and $\log \mathrm{CT}(\sigma)$ against σ. Also shown as dashed lines are the approximations for large σ given in part (ii) of Lemma 5.

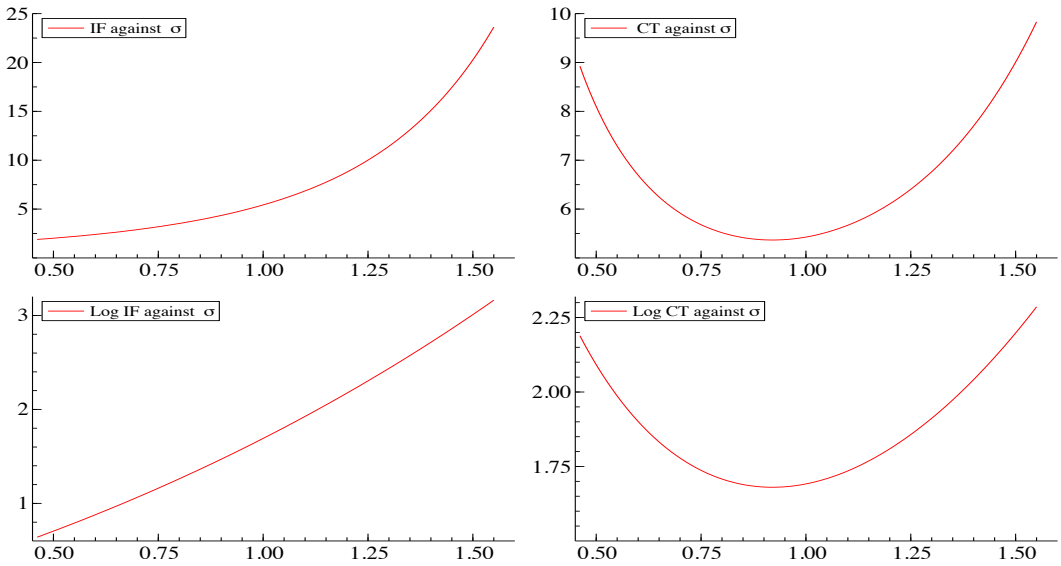


Figure 2: Plots of $\mathrm{IF}(\sigma), \mathrm{CT}(\sigma), \log \mathrm{IF}(\sigma)$ and $\log \mathrm{CT}(\sigma)$ against σ.

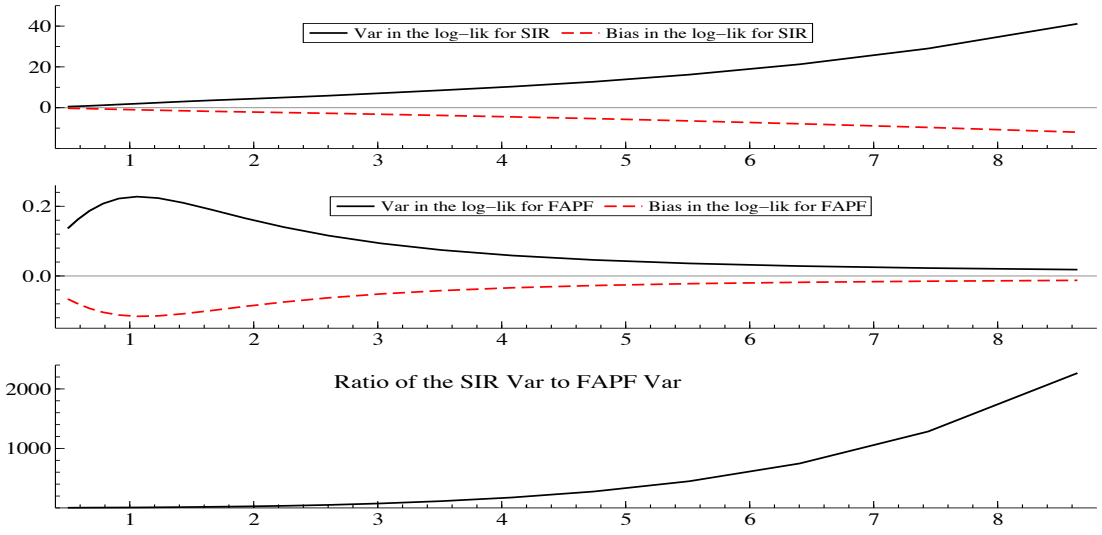


Figure 3: AR(1) plus noise model. Number of replications is 400 and $N = 50$ for both filters. Length $T = 200$, $\phi = 0.6$, $\sigma_\eta^2 = (1 - \phi^2)$. TOP: Bias and variance of SIR log-likelihood estimator against σ_ε^{-1}. MIDDLE: Bias and variance of FAPF log-likelihood estimator against σ_ε^{-1}. BOTTOM: Ratio of variance for SIR estimator to that of the FAPF estimator against σ_ε^{-1}.

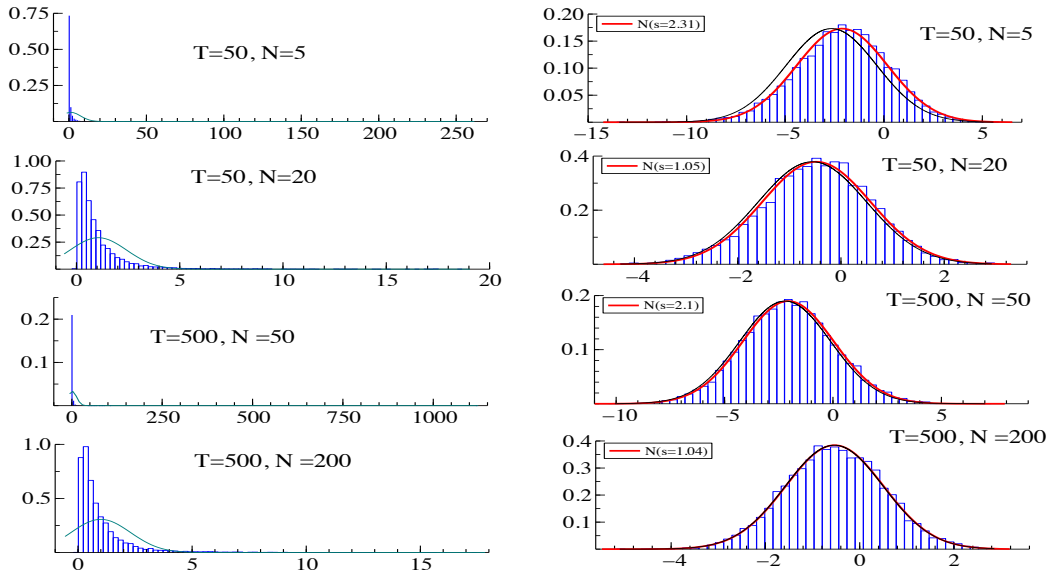


Figure 4: AR(1) plus noise model with fixed parameters. Number of replications is $10,000$. Displayed are the histograms and Gaussian approximations for the SIR likelihood estimator (divided by the true likelihood) on LEFT and for the error in the log of the SIR likelihood estimator on RIGHT. Both N and T vary as shown and $\phi = 0.6$, $\sigma_\eta^2 = (1-\phi^2)$, $\sigma_\varepsilon^2 = 2$. On RIGHT, a Gaussian is fitted to the histogram (red/solid line) using the estimated mean and variance. On RIGHT is the theoretical Gaussian density (black/dashed line) formed only from the estimated variance (mean $-\sigma^2/2$, variance σ^2).

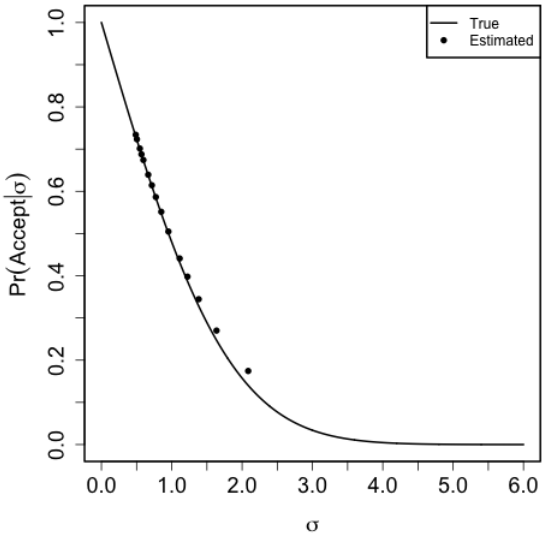


Figure 5: Probability of acceptance in Metropolis-Hastings step against the standard deviation of the log-likelihood estimator applied to the mixture of experts AR(1) model with fixed parameters. The theoretical probability of Lemma 3 is given by the solid line and the estimated probabilities are given by dots. For the estimated probabilities N is varied leading to the different values of σ.

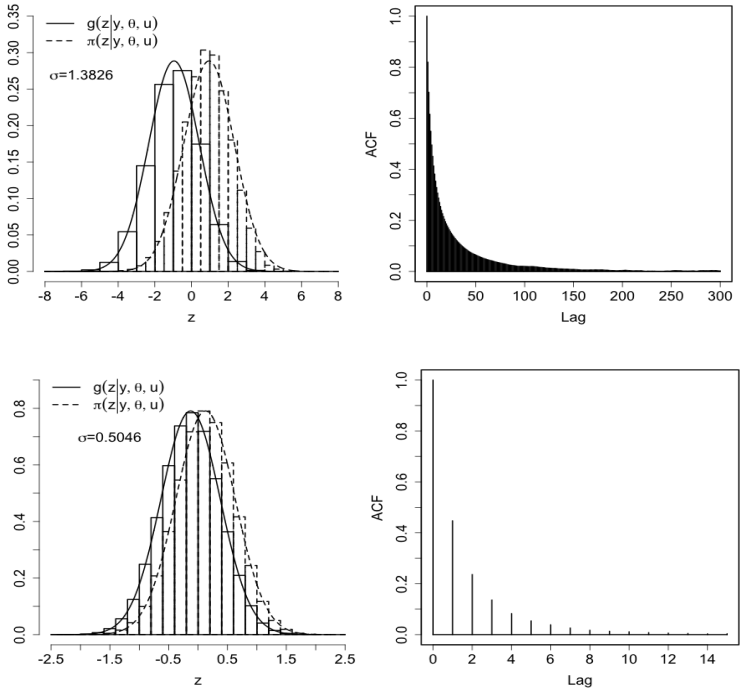


Figure 6: MCMC output using the standard particle filter for the mixture of AR(1) experts model with fixed parameters for two values of σ (and the number of particles N). The left panels are the theoretical densities of the proposal (solid line) and the posterior (dotted line) together with the histograms of the draws for Z (the error is the log-likelihood). The panels on the right are the empirical autocorrelations functions for the two values of σ. The panels in the first and second rows are for $\sigma = 1.3826$ and $\sigma = 0.5046$.

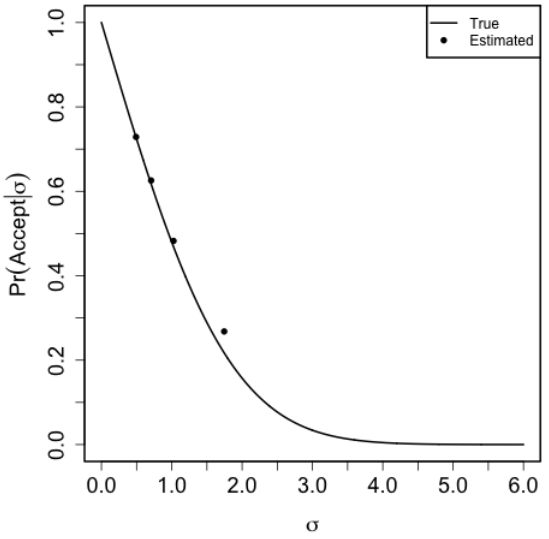


Figure 7: Probability of acceptance in Metropolis-Hastings step against the standard deviation of the log-likelihood estimator applied to the GARCH model observed with noise with fixed parameters. The theoretical probability of Lemma 3 is given by the solid line and the estimated probabilities are given by dots. For the estimated probabilities N is varied leading to the different values of σ.

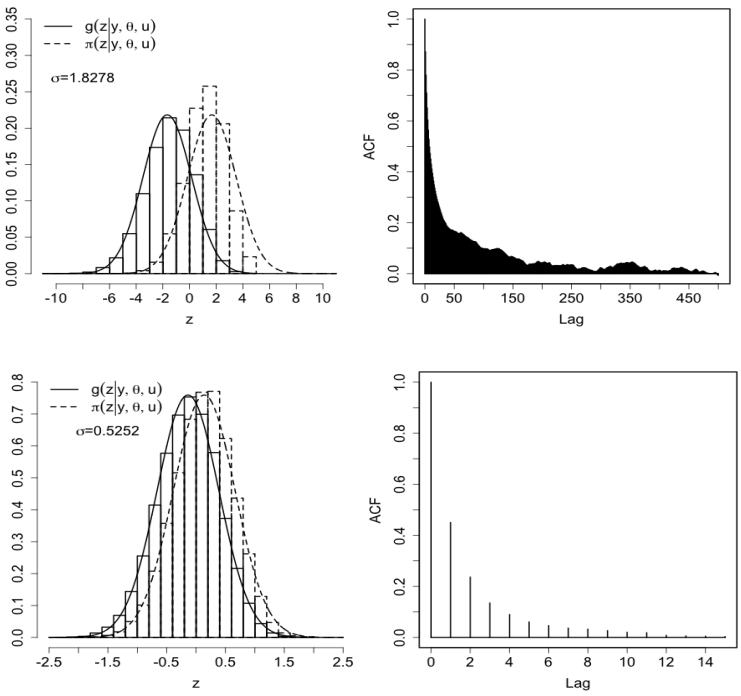
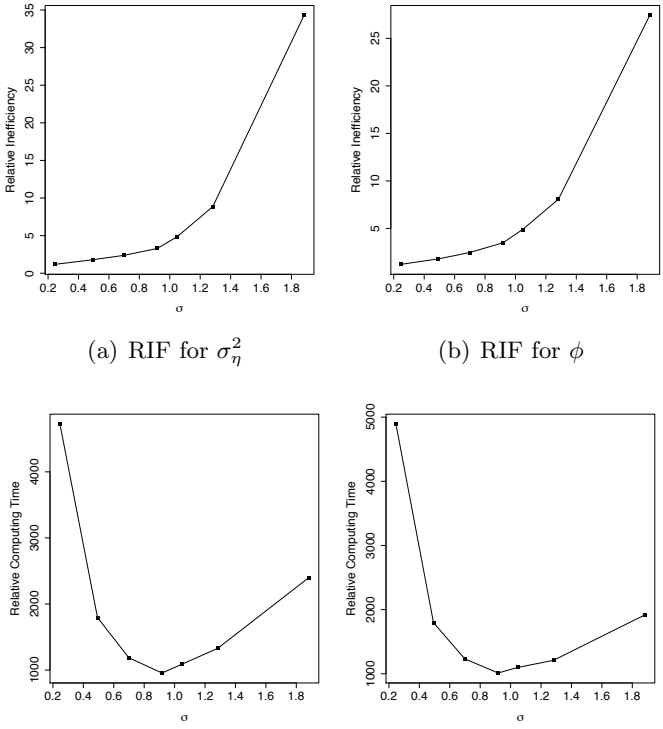


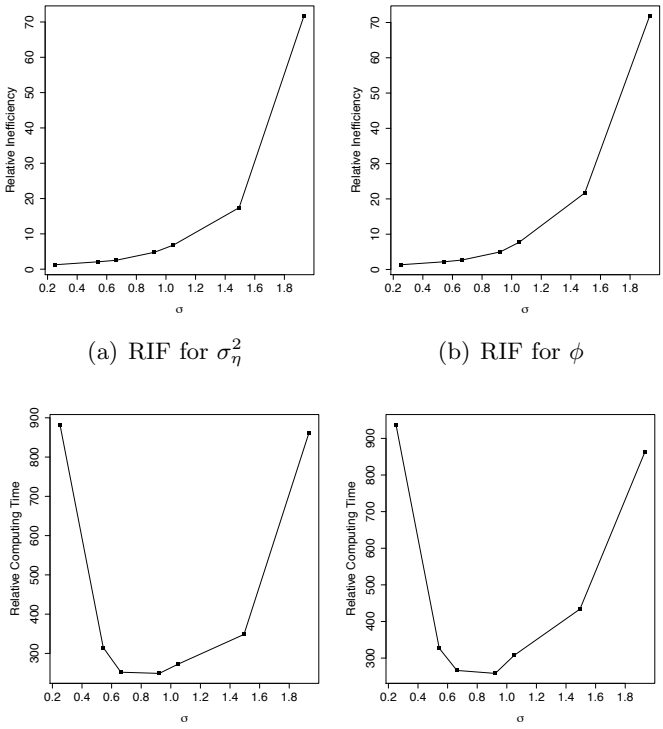
Figure 8: MCMC output using the standard particle filter for the GARCH model observed with noise with fixed parameters for two values of σ. The left panels are the theoretical densities of the proposal (solid line) and the posterior (dotted line) together with the histograms of the draws. The panels on the right are the empirical autocorrelations functions of Z for the two values of σ. The panels in the first and second rows are for $\sigma = 1.8278$ and $\sigma = 0.5252$.



(a) RIF for σ_η^2 (b) RIF for ϕ

(c) RCT for σ_η^2 (d) RCT for ϕ

Figure 9: Relative inefficiencies (RIF) and relative computing times (RCT) for the standard particle filter applied to the AR(1) plus noise model. Full MCMC applied to a single simulated time series with $T = 500$, $\phi = 0.6$ and $\sigma_\eta = 0.8$. RIF is calculated as the estimated integrated autocorrelation time from PMCMC output divided by that from the MCMC with the likelihood calculated by the Kalman filter. RCT $= N \times$ RIF. See Table 4 and Section 6.1 for details.



(a) RIF for σ_η^2 (b) RIF for ϕ

(c) RCT for σ_η^2 (d) RCT for ϕ

Figure 10: Relative inefficiencies (RIF) and relative computing times (RCT) for the fully adapted particle filter applied to the AR(1) plus noise model. Full MCMC applied to a single simulated time series with $T = 500$, $\phi = 0.6$ and $\sigma_\eta = 0.8$. RIF is calculated as the estimated integrated autocorrelation time from PMCMC output divided by that from the MCMC with the likelihood calculated by the Kalman filter. RCT $= N \times$ RIF. See Table 4 and Section 6.1 for details.

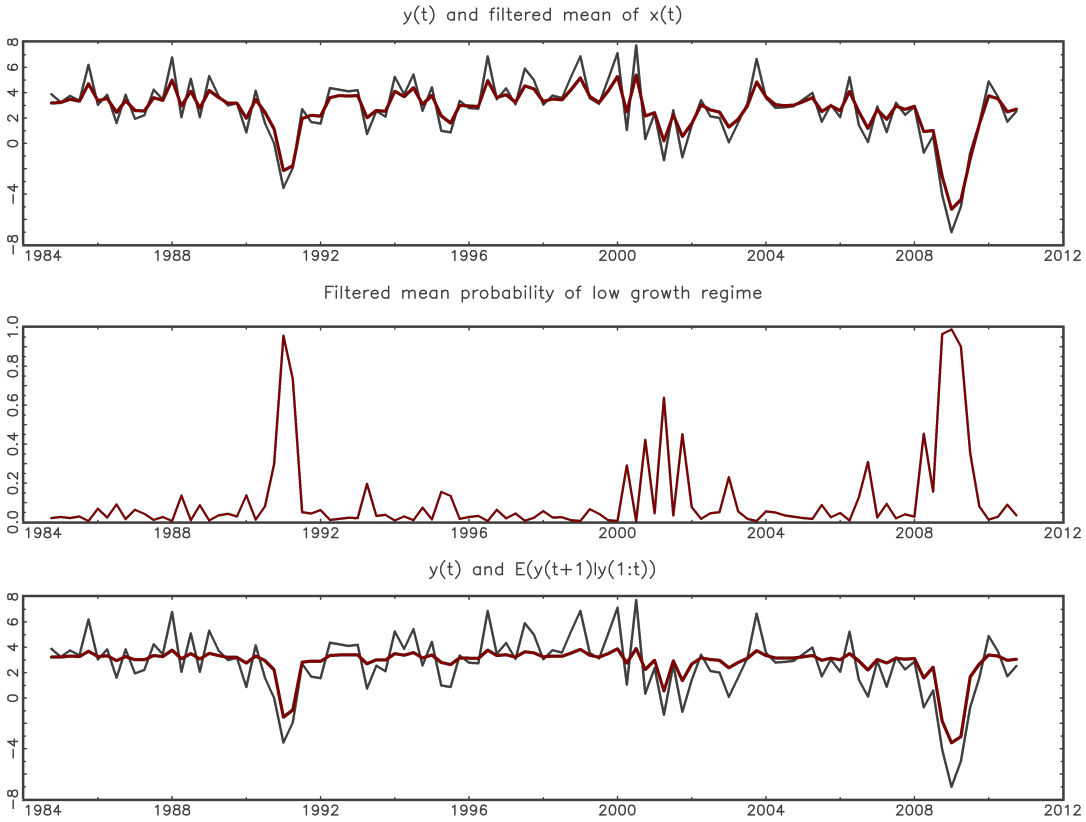


Figure 11: Mixture of experts model. Results from PMCMC analysis of seasonally adjusted US real log GDP annualized growth from 1984 Q2 to 2010 Q3 ($T = 100$). The top panel plots the dependent variable and the filtered mean of the first element x_t of the state vector against time. The middle panel plots the filtered mean probability of the low growth regime against time. The bottom panel plots y_t and $E(y_{t+1}|y_t)$ against time.

Table 1: Mixture of experts plus noise model. The table shows the mean, variance and standard deviation in the proposal and posterior of Z for various values of the number of particles for both the standard particle filter and the fully adapted particle filter. The sample size is $T = 100$ and the parameters are fixed at their true values

N	Proposal			Posterior		
	Mean	Var.	St.Dev.	Mean	Var.	St.Dev.
	Standard Particle Filter					
100	-1.9398	4.3531	2.0864	1.7775	3.2623	1.8062
200	-0.9231	1.9115	1.3826	0.8763	1.6889	1.2996
400	-0.4505	0.9064	0.9520	0.4420	0.8672	0.9312
1400	-0.1277	0.2546	0.5046	0.1251	0.2495	0.4995
	Fully Adapted Particle Filter					
4	-1.0799	3.2170	1.7936	0.8132	1.3613	1.1667
8	-0.4507	1.0537	1.0265	0.4046	0.7398	0.8601
12	-0.2860	0.6206	0.7878	0.2683	0.5061	0.7114
25	-0.1279	0.2688	0.5185	0.1273	0.2471	0.4971

Table 2: GARCH model observed with noise. The table shows the mean, variance and standard deviation in the proposal and posterior of Z for various values of the number of particles for both the standard particle filter and the fully adapted particle filter. The sample size is $T = 526$ and the parameters are fixed at their posterior mean.

N	Proposal			Posterior		
	Mean	Var.	St.Dev.	Mean	Var.	St.Dev.
	Standard Particle Filter					
1000	-1.4690	3.3408	1.8278	1.2254	2.2527	1.5009
2500	-0.5692	1.2094	1.0997	0.5266	0.9729	0.9864
5000	-0.2737	0.5729	0.7569	0.2611	0.4985	0.7061
10000	-0.1350	0.2758	0.5252	0.1334	0.2566	0.5066
	Fully Adapted Particle Filter					
50	-1.0230	2.3175	1.5223	1.1163	1.9663	1.4023
100	-0.4951	1.0973	1.0475	0.5234	0.9458	0.9725
250	-0.1926	0.4272	0.6536	0.2328	0.4197	0.6479
500	-0.1049	0.2118	0.4602	0.1024	0.2066	0.4546

Table 3: AR(1) model observed with noise. The table shows the mean, variance and standard deviation in the proposal and posterior of Z for various values of the number of particles for both the standard particle filter and the fully adapted particle filter. The artificial data set has $T = 500$ observations and the parameters are fixed at the true values.

N	Prior			Posterior		
	Mean	Var.	St.Dev.	Mean	Var.	St.Dev.
	Standard Particle Filter					
70	-1.7526	3.5433	1.8824	1.6099	3.2008	1.7891
150	-0.8101	1.6450	1.2826	0.8196	1.6214	1.2734
225	-0.5451	1.0953	1.0466	0.5284	1.0687	1.0338
290	-0.4174	0.8421	0.9176	0.4270	0.8459	0.9197
500	-0.2418	0.4881	0.6986	0.2396	0.4863	0.6973
1000	-0.1206	0.2432	0.4931	0.1216	0.2422	0.4922
4000	-0.0302	0.0607	0.2463	0.0285	0.0604	0.2458
	Fully Adapted Particle Filter					
12	-1.8514	3.7345	1.9325	1.7288	3.3914	1.8416
20	-1.1030	2.2326	1.4942	1.0784	2.0974	1.4483
40	-0.5516	1.1029	1.0502	0.5606	1.1161	1.0565
52	-0.4243	0.8501	0.9220	0.4267	0.8591	0.9269
100	-0.2213	0.4394	0.6629	0.2189	0.4387	0.6623
150	-0.1484	0.2954	0.5435	0.1437	0.2951	0.5432
700	-0.0315	0.0630	0.2510	0.0282	0.0624	0.2498

Table 4: Acceptance rates, inefficiencies (IF) and computing time (CT) for (a single run of) the Gaussian AR(1) model observed with noise applied to an artificial data set using differing particle filters and number of particles. Both random walk and independent Metropolis-Hastings proposals were fixed for all cases based on a previous run of their respective adaptive Metropolis-Hastings counterparts using the exact likelihood by the Kalman filter. We define the computing time CT = $N \times$ IF/1000 for the particle filters.

Algorithm	Number of Particles N	Accept. Rate	Inefficiency IF		Computing Time CT	
			τ^2	ϕ	τ^2	ϕ
	Kalman Filter					
RWM	-	34.32	8.50	8.52	-	-
IMH	-	94.07	1.24	1.16	-	-
		Standard Particle Filter				
RWM	70	9.81	126.93	82.90	8.885	5.803
	150	18.24	24.66	21.57	3.700	3.235
	225	22.16	16.05	15.73	3.611	3.540
	290	23.78	12.60	13.55	3.654	3.928
	500	27.38	10.27	10.45	5.137	5.227
	1000	30.58	9.38	9.17	9.375	9.168
	4000	32.95	8.70	8.57	34.788	34.289
IMH	70	19.35	42.63	31.81	2.984	2.227
	150	37.16	11.00	9.37	1.650	1.405
	225	46.64	6.01	5.67	1.351	1.275
	290	52.04	4.10	4.05	1.187	1.174
	500	61.81	2.95	2.86	1.476	1.432
	1000	71.74	2.23	2.09	2.232	2.086
	4000	84.46	1.47	1.42	5.873	5.678
	Fully Adapted Particle Filter					
RWM	12	7.40	157.03	125.49	1.884	1.506
	20	12.21	175.44	74.50	3.508	1.490
	40	19.07	21.22	22.24	0.848	0.889
	52	21.48	16.36	18.03	0.851	0.937
	100	26.58	12.06	11.87	1.206	1.187
	150	28.46	10.10	10.27	1.516	1.540
	700	32.84	8.57	8.87	6.002	6.211
IMH	12	13.26	89.22	83.44	1.070	1.001
	20	23.84	21.66	25.18	0.433	0.503
	40	38.98	8.48	8.93	0.3394	0.357
	52	45.27	5.96	5.76	0.3100	0.3000
	100	58.79	3.14	3.09	0.3136	0.3091
	150	65.40	2.61	2.53	0.3917	0.3793
	700	82.46	1.57	1.56	1.098	0.1089

Table 5: Acceptance rates, inefficiencies and computing times for (a single run of) the Gaussian GARCH model observed with noise applied to the UK index return using SIR and fully adapted particle filters, number of particles and the adaptive Metropolis-Hastings algorithms.

# of Particles	Accept. Rate	Inefficiency				Computing Time/1000			
		τ^2	θ_1	θ_2	θ_3	τ^2	θ_1	θ_2	θ_3
Standard Particle Filter									
500	8.63	110.07	169.76	231.45	255.31	55.04	84.88	115.73	127.66
1000	16.43	42.94	48.54	42.17	42.85	42.94	48.54	42.17	42.85
2500	26.10	11.39	14.40	11.89	13.51	28.48	36.01	29.74	33.77
5000	33.13	7.37	9.26	7.82	10.27	36.85	46.31	39.11	51.33
10000	37.02	5.72	8.56	5.58	7.40	57.18	85.60	55.75	74.00
Fully Adapted Particle Filter									
25	9.44	421.11	207.13	128.80	132.27	10.53	5.18	3.22	3.31
50	17.97	31.70	59.45	36.98	47.60	1.58	2.97	1.85	2.38
100	27.24	12.70	17.68	14.03	15.09	1.27	1.77	1.40	1.51
250	36.72	6.46	9.29	7.15	7.34	1.61	2.32	1.79	1.83
500	39.90	4.92	5.09	4.94	5.06	2.46	2.55	2.47	2.53

Table 6: Summary of statistics of the posterior distribution of the parameters for the GARCH(1,1) model with noise fitted to the UK MSCI index returns ($T = 526$). We used a fully adapted PF with $N = 100$ particles.

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

Table 7: Prior and posterior means and standard deviations for the mixture of AR(1) experts plus noise models. All priors are independent normals.

param	prior mean	prior std	post. mean	post. std
$\ln \sigma^2$	0.85	0.2	0.65	0.21
c_1	0	1	-0.11	0.52
ϕ_1	0.5	0.2	0.58	0.18
c_2	2	1	2.17	0.44
ϕ_2	0.5	0.2	0.32	0.14
$\ln \tau_1^2$	0.85	0.2	1.08	0.27
$\ln \tau_2^2$	0.85	0.2	0.51	0.25
ξ_1	0	10	-0.80	1.62
ξ_2	0	10	-2.33	1.31
ξ_2	0	10	-1.53	1.37