Filtering via Simulation: Auxiliary Particle Filters

Michael K. PIT\textsc{t} and Neil SHEPHARD

This article analyses the recently suggested particle approach to filtering time series. We suggest that the algorithm is not robust to outliers for two reasons: the design of the simulators and the use of the discrete support to represent the sequentially updating prior distribution. Here we tackle the first of these problems.

**KEY WORDS:** Filtering; Markov chain Monte Carlo; Particle filter; Sampling/importance resampling; Simulation; State space.

1. INTRODUCTION

In this article we model a time series $y_t, t = 1, \ldots, n$, as being conditionally independent given an unobserved sufficient state $\alpha_t$, which is itself assumed to be Markovian. The task is to use simulation to carry out on-line filtering—that is, to learn about the state given contemporaneously available information. We do this by estimating the difficult-to-compute density (or probability distribution function) $f(\alpha_t | y_1, \ldots, y_t) = f(\alpha_t | Y_t)$, $t = 1, \ldots, n$. We assume parametric forms for both the “measurement” density $f(y_t | \alpha_t)$ and the “transition” density of the state $f(\alpha_{t+1} | \alpha_t)$. The state evolution is initialized by some density $f(\alpha_0)$.

Filtering can be thought of as the repeated application of a two-stage procedure. First, the current density must be propagated into the future via the transition density $f(\alpha_{t+1} | \alpha_t)$ to produce the prediction density

$$f(\alpha_{t+1} | Y_t^t) = \int f(\alpha_{t+1} | \alpha_t) f(\alpha_t | Y_t^t) dF(\alpha_t | Y_t^t). \ (1)$$

Second, one moves to the filtering density via Bayes theorem,

$$f(\alpha_{t+1} | Y_{t+1}^t) = \frac{f(y_{t+1} | \alpha_{t+1}) f(\alpha_{t+1} | Y_t^t)}{f(y_{t+1} | Y_t^t)} ,$$

$$= \int f(y_{t+1} | \alpha_{t+1}) f(\alpha_{t+1} | Y_{t+1}^t) dF(\alpha_{t+1} | Y_{t+1}^t). \ (2)$$

This implies that the data can be processed in a single sweep, updating our knowledge about the states as we receive more information. This is straightforward if $\alpha_t, \alpha_{t-1}$ has a finite set of known discrete points of support, as the previous calculations can be computed exactly. When the support is continuous and the integrals cannot be analytically solved, then numerical methods must be used.

Numerous attempts have been made to provide algorithms that approximate the filtering densities. Important recent work includes that of Gerlach, Carter, and Kohn (1996), Kitagawa (1987), West (1992), and those papers reviewed by West and Harrison (1997, chaps. 13 and 15).

In this article we use simulation to perform filtering following an extensive recent literature. Our approach is to extend the particle filter that has recently been suggested independently by various authors. In particular, it was used by Gordon, Salmond, and Smith (1993) on non-Gaussian state-space models. The same algorithm, with extensions to the smoothing problem, has been independently proposed by Kitagawa (1996) (and generalized in Hürzeler and Künsch 1995) for use in time series problems. It reappeared and was then discarded by Berzuni, Best, Gilks, and Larizza (1997) in the context of a real-time application of the sequential analysis of medical patients. It was again proposed by Isard and Blake (1996) in the context of robustly tracking motion in visual clutter, under the term the “condensation” algorithm. Some statistical refinements of this general class of algorithm, generically called particle filters, have been given by Carpenter, Clifford, and Fearnhead (1998), Doucet (1998), and Liu and Chen (1998) (which were written independently of this article). The idea of calling this class of algorithm “particle filters” is from Carpenter et al. (1998), although Kitagawa (1996) used the term “particles.” Similar ideas (but using stronger assumptions) are used on the blind deconvolution problem of Liu and Chen (1995) and in the sequential importance sampling algorithms of Hendry and Richard (1991) and Kong, Liu, and Wong (1994).

Here we discuss the particle filtering literature and extend it in a number of directions so that it can be used in a much broader context. The article is organized as follows. In Section 2 we analyze the statistical basis of particle filters and focus on its weaknesses. In Section 3 we introduce our main contribution, an auxiliary particle filter method. We give some numerical examples in Section 4 and state some conclusions in Section 5.

2. PARTICLE FILTERS

2.1 Definition of Particle Filters

Particle filters are the class of simulation filters that recursively approximate the filtering random variable $\alpha_t | Y_t = (y_1, \ldots, y_t)$ by “particles” $\alpha_{t,1}^1, \ldots, \alpha_{t,M}^M$, with discrete probability mass of $\pi_{t,1}^1, \ldots, \pi_{t,M}^M$. Hence a continuous variable is...
approximated by a discrete one with random support. These discrete points are viewed as samples from \( f(\alpha_t | Y_t) \). In the literature all of the \( \pi_i \) are assumed to equal \( 1/M \). Throughout, \( M \) is taken to be very large. Then we require that as \( M \to \infty \), the particles can be used to increasingly well approximate the density of \( \alpha_t | Y_t \).

Particle filters treat the discrete support generated by the particles as the true filtering density. This allows us to produce an approximation to the prediction density, \( f(\alpha_{t+1} | Y_t) \), simply by using the discrete support of the particles. We call 
\[
 f(\alpha_{t+1} | Y_t) = \sum_{j=1}^{M} f(\alpha_{t+1} | \alpha_t^j) \pi_t^j
\]
the "empirical prediction density." This is a mixture of densities and so echoes the earlier filtering work of, for example, Sorensen and Alspach (1971). This can be combined with the measurement density to produce, up to proportionality,
\[
 f(\alpha_{t+1} | Y_{t+1}) \propto f(y_{t+1} | \alpha_{t+1}) \sum_{j=1}^{M} f(\alpha_{t+1} | \alpha_t^j) \pi_t^j.
\]

2.2 Sampling the Empirical Prediction Density

One way of sampling from the empirical prediction density is to think of \( \sum_{j=1}^{M} f(\alpha_{t+1} | \alpha_t^j) \pi_t^j \) as a "prior" density \( f(y_{t+1} | Y_t) \) and so, via the prediction decomposition, the joint density of the observations; and estimating the corresponding distribution function \( F(y_{t+1} | Y_t) \), which can be a useful diagnostic measure of fit for non-Gaussian models (see, e.g., Gerlach et al. 1996, Shephard 1994; Smith 1985).

2.2.1 Sampling/Importance Resampling. Importance sampling/importance resampling (SIR) method Rubins draws \( \alpha^1, \ldots, \alpha^R \) from \( f(\alpha) \) and then associates with each of these draws the weights \( \pi_j \), where 
\[
 w_j = f(y | \alpha^j), \quad \pi_j = \frac{w_j}{\sum_{i=1}^{M} w_i}, \quad j = 1, \ldots, R.
\]

Then the weighted sample will converge, as \( R \to \infty \), to a nonrandom sample from the desired posterior \( f(\alpha | y) \) as
\[
 R \to \infty \sum_{j=1}^{R} w_i = f(y).
\]

2.2.2 Adaptation. The foregoing SIR algorithm samples from \( f(\alpha | y) \) by making blind proposals \( \alpha^1, \ldots, \alpha^R \) from the prior, ignoring the fact that we know the value of \( y \). This is the main feature of existing particle filters. We say that a particle filter is adapted if we make proposals that take into account the value of \( y \).

An adapted SIR-based particle filter has the following general structure:

1. Draw from \( \alpha^1, \ldots, \alpha^R \sim g(\alpha | y) \).
2. Evaluate \( w_j = f(y | \alpha^j) / f(\alpha^j | y) \), \( j = 1, \ldots, R \).
3. Resample among the \( \{\alpha^j\} \) using weights proportional to \( \{w_j\} \) to produce a sample of size \( M \).

Although this looks attractive, for a particle filter, \( f(\alpha) = \sum_{j=1}^{M} f(\alpha_{t+1} | \alpha_t^j) \pi_t^j \), which implies we must at least evaluate \( M \times R \) densities to generate \( M \) samples from \( f(\alpha | y) \).

2.2.3 Rejection and Markov Chain Monte Carlo Sampling. Exactly the same remarks hold for rejection sampling. A blind rejection sampling-based particle filter will simulate from \( f(\alpha) \) and accept with probability \( \pi(\alpha) = f(y | \alpha) / f(y | \alpha_{\text{max}}) \), where \( \alpha_{\text{max}} = \arg \max_{\alpha} f(y | \alpha) \). This has been proposed by Hürzeler and Künsch (1995). Again, the rejection becomes worse if the var \( \{\pi(\alpha)\} \) is high and adaption is difficult, as it will again typically involve evaluating \( f(\alpha) \) and so is computationally infeasible.
Another alternative to SIR is the use of a blind MCMC method (see Gilks, Richardson, and Spiegelhalter 1996 for a review). In this context the MCMC accepts a move from a current state \( \alpha \) to \( \alpha' \sim f(\alpha) \) with probability

\[
\min\{1, \frac{f(\alpha'|\alpha)}{f(\alpha|\alpha)} \} \cdot \frac{g(\alpha|y)}{g(\alpha'|y)}.
\]

Again, the problem with this is that evaluating \( f(\alpha) \) is very expensive.

### 2.3 Weaknesses of Particle Filter

The particle filter based on SIR has two basic weaknesses. The first is well known, that when there is an outlier, the weights \( \pi_j \) will be very unevenly distributed and so it will require an extremely large value of \( R \) for the draws to be close to samples from the empirical filtering density. This is of particular concern if the measurement density \( f(y_{t+1}|\alpha_t) \) is highly sensitive to \( \alpha_t \). Notice this is not a problem of having too small a value of \( M \). That parameter controls the accuracy of (3). Instead, the difficulty is, given that degree of accuracy, how to efficiently sample from (4).

We show how to do this in the next section.

The second weakness holds in general for particle filters for which the \( \pi_j \) are equal and where the states are updated one period at a time. As \( R \to \infty \), so the weighted samples can be used to arbitrarily well approximate (4). However, the tails of (3) usually only poorly approximate the true tails of \( \alpha_{t+1}|Y_{t+1} \), due to the use of the mixture approximation. As a result, (4) can only poorly approximate the true \( f(\alpha_{t+1}|Y_{t+1}) \) when there is an outlier. Hence the second question is how to improve the empirical prediction density's behavior in the tails. In earlier work (Pitt and Shephard 1998) we analyzed this problem using the so-called fixed lagged filter.

### 3. Auxiliary Variable

#### 3.1 The Basics

A fundamental problem with existing particle filters is that their mixture structure means that it is difficult to adapt the SIR, rejection, or MCMC sampling methods without greatly slowing the running of the filter. Here we argue that many of these problems are reduced when we perform particle filtering in a higher dimension.

Our task is to sample from the joint density \( f(\alpha_{t+1}, k|Y_{t+1}) \), where \( k \) is an index on the mixture in (3). We define

\[
f(\alpha_{t+1}, k|Y_{t+1}) \propto f(y_{t+1}|\alpha_{t+1}) f(\alpha_{t+1}|k)^{\pi_k},
\]

\[
k = 1, \ldots, M.
\]

If we draw from this joint density and then discard the index, then we produce a sample from the empirical filtering density (4) as required. We call \( k \) an auxiliary variable, as it is present simply to aid the task of simulation. We call generic particle filters of this type auxiliary particle filters.

We can now sample from \( f(\alpha_{t+1}, k|Y_{t+1}) \) using SIR, rejection, or MCMC. The SIR idea is to make \( R \) proposals \( \alpha_{t+1, k}^{(i)}, k^{(i)} \sim g(\alpha_{t+1}, k|Y_{t+1}) \) and then construct resampling weights

\[
w_j = \frac{f(y_{t+1}|\alpha_{t+1, k}^{(j)}, k^{(j)}) f(\alpha_{t+1, k}^{(j)}|k^{(j)})}{g(\alpha_{t+1, k}^{(j)}|k^{(j)})}, \quad \pi_j = \frac{w_j}{\sum_{i=1}^R w_i}, \quad j = 1, \ldots, R.
\]

We have complete control over the design of \( g(\cdot) \), which can depend on \( y_{t+1} \) and \( \alpha_t^{(i)} \) to make the weights even. Thus this method is adaptable and extremely flexible. In the next section we give a convenient generic suggestion for the choice of \( g(\cdot) \).

Rejection sampling for auxiliary particle filtering could also be used in this context. An example of this appears in Section 3.3.4. We can also make proposals for an MCMC variate of the auxiliary particle filter from \( \alpha_{t+1, k}^{(i)}, k^{(i)} \sim g(\alpha_{t+1}, k|Y_{t+1}) \), where \( g(\alpha_{t+1}, k|Y_{t+1}) \) is some arbitrary density; then these moves are accepted with probability

\[
1 - \frac{f(y_{t+1}|\alpha_{t+1, k}^{(i)}|k^{(i)}) f(\alpha_{t+1, k}^{(i)}|k^{(i)})}{g(\alpha_{t+1, k}^{(i)}|k^{(i)}) Y_{t+1})} \times \frac{g(\alpha_{t+1, k}^{(i)}|k^{(i)})}{g(\alpha_{t+1, k}^{(i)}|k^{(i)}) Y_{t+1})}
\]

A special case of this argument has been given by Berzunzi et al. (1997), who put \( g(\alpha_{t+1}, k|Y_{t+1}) \propto f(\alpha_{t+1}|k^{(i)}) \), which means that their method is again blind.

#### 3.2 A Generic Sampling/Importance Resampling—Based Auxiliary Proposal

Here we give a generic \( g(\cdot) \) that can be broadly applied. We base our discussion on the SIR algorithm, although we could have used an MCMC method. We approximate (5) by

\[
g(\alpha_{t+1}, k|Y_{t+1}) \propto f(y_{t+1}|\mu_{k+1}^t) f(\alpha_{t+1}|k)^{\pi_k},
\]

\[
k = 1, \ldots, M,
\]

where \( \mu_{k+1}^t \) is the mean, the mode, a draw, or some other likely value associated with the density of \( \alpha_{t+1}|k \). The form of the approximating density is designed so that

\[
g(k|Y_{t+1}) \propto \pi_k \int f(y_{t+1} | \mu_{k+1}^t) dF(\alpha_{t+1}|k)
\]

\[
= \pi_k f(y_{t+1} | \mu_{k+1}^t).
\]

Thus we can sample from \( g(\alpha_{t+1}, k|Y_{t+1}) \) by simulating the index with probability \( \lambda_k \propto g(k|Y_{t+1}) \), and then sampling from the transition density given the mixture \( f(\alpha_{t+1}|k) \). We call the \( \lambda_k \) the first-stage weights.

The implication is that we simulate from particles associated with large predictive likelihoods. Having sampled
the joint density of \( g(\alpha_{t+1}, k|Y_{t+1}) \) \( R \) times, we perform a reweighting, putting on the draw \( (\alpha_{t+1}^j, k^j) \) the weights proportional to the so-called second-stage weights

\[
\pi_j = \frac{w_j}{\sum_{i=1}^{R} w_i}, \quad j = 1, \ldots, R.
\]

The hope is that these second-stage weights are much less variable than for the original SIR method. We might resample from this discrete distribution to produce a sample of size \( M \).

By making proposals that have high conditional likelihoods, we reduce the costs of sampling many times from particles that have very low likelihoods and so will not be resampled at the second stage of the process. This improves the statistical efficiency of the sampling procedure and means that we can reduce the value of \( R \) substantially.

To measure the statistical efficiency of these procedures, we argued earlier that we could look at minimizing \( E\{\pi(\alpha)^2\} \). Here we compare a standard SIR with a SIR based on our auxiliary variable. For simplicity, we set \( \pi^k = 1/M \) in both cases. Then, for a standard SIR-based particle filter, for large \( M \),

\[
E\{\pi(\alpha)^2\} = \frac{1}{M} \sum_{k=1}^{M} \int f(y_{t+1}|\alpha_{t+1})^2 dF(\alpha_{t+1}|\alpha_{t}^k)^2
\]

where

\[
f_k = \int \left\{ \frac{f(y_{t+1}|\alpha_{t+1})}{f(y_{t+1}|\mu_{t+1})} \right\}^2 dF(\alpha_{t+1}|\alpha_{t}^k)
\]

and

\[
f_k^* = \int \left\{ \frac{f(y_{t+1}|\alpha_{t+1})}{f(y_{t+1}|\mu_{t+1})} \right\} dF(\alpha_{t+1}|\alpha_{t}^k).
\]

The same calculation for a SIR-based auxiliary variable particle filter gives

\[
E\{\pi_a(\alpha)^2\} = \frac{1}{M} \sum_{k=1}^{M} \lambda_k f_k
\]

which shows an efficiency gain if

\[
M \sum_{k=1}^{M} \lambda_k f_k < M \sum_{k=1}^{M} \lambda_k^2 f_k.
\]

If \( f_k \) does not vary over \( k \), then the auxiliary variable particle filter will be more efficient as \( \sum_{k=1}^{M} \lambda_k (1/M) = (1/M) \leq \sum_{k=1}^{M} \lambda_k^2 \). More likely is that \( f_k \) will depend on \( k \), but only mildly, as \( f(\alpha_{t+1}|\alpha_{t}^k) \) will be typically quite tightly peaked [much more tightly peaked than \( f(\alpha_{t+1}|Y_t) \)] compared to the conditional likelihood.

### 3.3 Examples of Adaptation

#### 3.3.1 Basics

Although the previous generic scheme can usually reduce the variability of the second-stage weights, other adaption schemes use the specific structure of the time series model to allow us to achieve yet more equal weights. If we can achieve exactly equal weights, then we say that we have fully adapted the procedure to the model, for now we can produce iid samples from (4). This situation is particularly interesting, as we are then close to the assumptions made by Kong et al. (1994) for their sequential importance sampler. Although full adaption is of some practical importance, we should remind ourselves that even fully adapted particle filters do not produce iid samples from \( f(\alpha_{t+1}|Y_{t+1}) \), due to their approximation of \( f(\alpha_{t+1}|Y_{t+1}) \) by a finite mixture distribution. This is inherent in the construction of this class of filter.

#### 3.3.2 Nonlinear Gaussian Measurement Model

In the Gaussian measurement case, the absorption of the measurement density into the transition equation is particularly convenient. Consider a nonlinear transition density with \( \alpha_{t+1}|\alpha_t \sim N(\mu(\alpha_t), \sigma^2(\alpha_t)) \) and \( y_{t+1}|\alpha_{t+1} \sim N(\alpha_{t+1}, 1) \). Then

\[
f(\alpha_{t+1}, k|Y_{t+1}) \propto f(y_{t+1}|\alpha_{t+1})f(\alpha_{t+1}|\alpha_{t}^k)
\]

where

\[
\frac{f(\alpha_{t+1}|\alpha_{t+1}, y_{t+1})}{f(y_{t+1}|\alpha_{t+1})} = g_k(y_{t+1})f(\alpha_{t+1}|\alpha_{t}^k, y_{t+1}),
\]

and

\[
\mu_k = \sigma_k^2 \left\{ \frac{\mu(\alpha_{t}^k)}{\sigma^2(\alpha_{t}^k)} + y_{t+1} \right\},
\]

\[
\sigma_k^{-2} = 1 + \sigma^{-2}(\alpha_{t}^k).
\]

This implies that the first-stage weights are

\[
g_k(y_{t+1}) \propto \sigma_k^2 \exp \left\{ \frac{\mu_k^2}{2\sigma_k^2} - \frac{\mu(\alpha_{t}^k)^2}{2\sigma^2(\alpha_{t}^k)} \right\}
\]

The Gaussian measurement density implies that the second-stage weights are all equal.

An example of this is a Gaussian autoregressive conditional heteroscedasticity (ARCH) model (see, e.g., Bollerslev, Engle, and Nelson 1994) observed with independent Gaussian error. So we have

\[
y_{t+1}|\alpha_t \sim N(\alpha_t, \sigma^2), \quad \alpha_{t+1}|\alpha_t \sim N(0, \beta_0 + \beta_1\alpha_t^2).
\]

This model is fully adaptable. It has received a great deal of attention in the econometric literature, as it has some attractive multivariate generalizations; see the work by Diebold and Nerlove (1989), Harvey, Ruiz, and Sentana (1992) and King, Sentana, and Wadhwani (1994). As far as we know, no likelihood methods exist in the literature for the analysis of this type of model (and its various generalizations), although a number of very good approximations have been suggested.
3.3.3 Log-Concave Measurement Densities. Suppose again that \( f(\alpha_{t+1}|\alpha^{k}_t) \) is Gaussian, but the measurement density is log-concave as a function of \( \alpha_{t+1} \). Then we might extend the foregoing argument by Taylor expanding \( \log g(y_{t+1}|\alpha_{t+1}) \) to a second-order term, again around \( \mu^{k}_{t+1} \), to give the approximation

\[
\log g(y_{t+1}|\alpha_{t+1}, \mu^{k}_{t+1}) = \log f(y_{t+1}|\mu^{k}_{t+1}) + (\alpha_{t+1} - \mu^{k}_{t+1})' \nabla \log f(y_{t+1}|\mu^{k}_{t+1}) + \frac{1}{2} (\alpha_{t+1} - \mu^{k}_{t+1})' \nabla^2 \log f(y_{t+1}|\mu^{k}_{t+1}) (\alpha_{t+1} - \mu^{k}_{t+1}).
\]

Then

\[
g(\alpha_{t+1}, k|Y_{t+1}) \propto g(y_{t+1}|\alpha_{t+1}; \mu^{k}_{t+1}) f(\alpha_{t+1}|\alpha^{k}_t).
\]

Rearranging, we can express this as

\[
g(\alpha_{t+1}, k|Y_{t+1}) \propto g(y_{t+1}|\mu^{k}_{t+1}) g(\alpha_{t+1}|\alpha^{k}_t, y_{t+1}; \mu^{k}_{t+1}),
\]

which means that we could simulate the index with probability proportional to \( g(y_{t+1}|\mu^{k}_{t+1}) \) and then draw from \( g(\alpha_{t+1}|\alpha^{k}_t, y_{t+1}; \mu^{k}_{t+1}) \). The resulting reweighted sample's second-stage weights are proportional to the, hopefully fairly even, weights

\[
w_j = \frac{f(y_{t+1}|\alpha^{k}_t)f(\alpha_{t+1}|\alpha^{k}_t)}{g(y_{t+1}|\mu^{k}_{t+1}) g(\alpha^{k}_t|\alpha^{k}_t, y_{t+1}; \mu^{k}_{t+1})},
\]

\[
\pi_j = \frac{w_j}{\sum_{i=1}^{R} w_i}, \quad j = 1, \ldots, R.
\]

Thus we can exploit the special structure of the model, if available, to improve upon the auxiliary particle filter.

3.3.4 Stochastic Volatility and Rejection Sampling. The same argument carries over when we use a first-order Taylor expansion to construct \( g(y_{t+1}|\alpha_{t+1}, \mu^{k}_{t+1}) \), but in this case we know that \( g(\alpha_{t+1}|\mu_{t+1}, \mu^{k}_{t+1}) \geq f(y_{t+1}|\alpha_{t+1}) \) for any value of \( \mu^{k}_{t+1} \) due to the assumed log-concavity of the measurement density. Thus

\[
f(\alpha_{t+1}, k|Y_{t+1}) \propto f(\alpha_{t+1}|\alpha^{k}_t) f(\alpha_{t+1}|\alpha^{k}_t) \\
\leq g(y_{t+1}|\alpha_{t+1}; \mu^{k}_{t+1}) f(\alpha_{t+1}|\alpha^{k}_t) \\
= g(y_{t+1}|\mu^{k}_{t+1}) g(\alpha_{t+1}|\alpha^{k}_t, y_{t+1}; \mu^{k}_{t+1}) \\
\propto g(\alpha_{t+1}, k|Y_{t+1}).
\]

Thus we can perform rejection sampling from \( f(\alpha_{t+1}, k|Y_{t+1}) \) by simply sampling \( k \) with probability proportional to \( g(y_{t+1}|\mu^{k}_{t+1}) \) and then drawing \( \alpha_{t+1} \) from \( g(\alpha_{t+1}|\alpha^{k}_t, y_{t+1}; \mu^{k}_{t+1}) \). This pair is then accepted with probability

\[
\frac{f(y_{t+1}|\alpha_{t+1})}{g(y_{t+1}|\alpha_{t+1}; \mu^{k}_{t+1})}.\]

This argument applies to the stochastic volatility (SV) model,

\[
y_t = \varepsilon_t \beta \exp(\alpha_{t}/2), \quad \alpha_{t+1} = \phi \alpha_{t} + \eta_t,
\]

where \( \varepsilon_t \) and \( \eta_t \) are independent Gaussian processes with variances 1 and \( \sigma^2 \). Here \( \beta \) has the interpretation as the modal volatility, \( \phi \) is the persistence in the volatility shocks, and \( \sigma^2 \) is the volatility of the volatility. This model has attracted much recent attention in the econometrics literature as a way of generalizing the Black–Scholes option pricing formula to allow volatility clustering in asset returns (see, e.g., Hull and White 1987). MCMC methods have been used on this model by, for instance, Jacquier, Polson, and Rossi (1994), Kim, Shephard, and Chib (1998), and Shephard and Pitt (1997).

For this model \( \log f(y_{t+1}|\alpha_{t+1}) \) is concave in \( \alpha_{t+1} \), so that, for \( \mu^{k}_{t+1} = \phi \alpha^{k}_t \),

\[
\log g(y_{t+1}|\alpha_{t+1}; \mu^{k}_{t+1}) = \text{const} - \frac{1}{2} \alpha_{t+1} \nabla^2 \log f(y_{t+1}|\mu^{k}_{t+1}) (1 - (\alpha_{t+1} - \mu^{k}_{t+1})).
\]

The implication is that

\[
g(\alpha_{t+1}|\alpha^{k}_t, y_{t+1}; \mu^{k}_{t+1}) \\
= \mathcal{N} \left( \mu^{k}_{t+1} + \frac{\sigma^2}{2} \left( \frac{y_t^2}{\beta^2} \exp(-\mu^{k}_{t+1}) \right) - \frac{1}{2} \left( \mu^{k}_{t+1} - \mu^{k}_{t+1} \right) \right)
\]

\[
= \mathcal{N}(\mu^{k}_{t+1}, \sigma^2).
\]

Likewise,

\[
g(\alpha_{t+1}|\alpha^{k}_t, y_{t+1}; \mu^{k}_{t+1}) = \exp \left\{ \frac{1}{2\beta^2} \left( \mu^{k}_{t+1} - \mu^{k}_{t+1} \right) \right\}
\]

\[
\times \exp \left\{ - \frac{y_t^2}{2\beta^2} \exp(-\mu^{k}_{t+1}) (1 + \mu^{k}_{t+1}) \right\}.
\]

Finally, the log-probability of acceptance is

\[
- \frac{1}{2\beta^2} \left[ \exp(-\alpha_{t+1}) - \exp(-\mu^{k}_{t+1}) (1 - (\alpha_{t+1} - \mu^{k}_{t+1})) \right].
\]

Notice that as \( \sigma^2 \) falls to 0, so the acceptance probability goes to 1.

Finally, the same argument holds when we use a SIR algorithm instead of rejection sampling. The proposals are made in exactly the same way, but now instead of computing log-probabilities of accepting, these become log-second-stage weights.

3.3.5 Limited Dependent Processes. A less trivial example of full adaption is a special case of limited dependent processes, where the observations are deterministic functions of the states. A simple example of this is a probit time series where \( y_t = I(\alpha_t > 0) \), where \( \alpha_t \) is Gaussian and univariate and \( I(\cdot) \) denotes an indicator function. Then
if \( y_{t+1} = 1 \), we have, exactly,

\[
\Pr(a_{t+1}, k|Y_{t+1}) \propto w^k \Pr(a_{t+1}| \alpha^k_{t+1}, \alpha_{t+1} > 0),
\]

\[
w^k = \Pr(a_{t+1}| \alpha^k_{t+1}, \alpha_{t+1} > 0).
\]

Hence we choose \( k \) with probability proportional to \( w^k \) and then draw from a truncated distribution conditional on \( k \). If \( y_{t+1} \) is negative, then the weights \( w^k \) would be \( \Pr(a_{t+1} < 0| \alpha^k_{t+1}, \alpha_{t+1} < 0) \), and while the truncated draw would be from \( \Pr(a_{t+1} < 0| \alpha^k_{t+1}, \alpha_{t+1} < 0) \). This style of argument carries over to ordered probit and censored models where we observe, for example, \( \min(0, \alpha_t) \).

Adaptation can be very important in these types of models, for naively implemented particle and auxiliary variable filters are generally vulnerable to tightly peaked measurement densities. In the censored model, where \( y_{t+1} = \min(0, \alpha_{t+1}) \), the measurement density is degenerate when \( y_{t+1} > 0 \), and so the particle filter will degenerate to give all of its mass on the simulation that is closest (but because they are simulated from \( \Pr(a_{t+1}| \alpha^k_{t+1}, \alpha_{t+1} > 0) \) not equal) to \( y_{t+1} \). Adaptation overcomes this problem instantly.

Adaptation is also essential for the following problem. Suppose that \( \alpha_{t+1}| \alpha_t \) is Gaussian, \( \alpha_{t+1} \) is bivariate, and we observe \( y_{t+1} = \min(\alpha_{t+1}) \). Such models are called disequilibrium models in economics. (Recent work in this area includes Larroque and Salanié 1993 and Manrique and Shephard 1998.) Then

\[
\Pr(a_{t+1}, k|Y_{t+1}) \propto \Pr(y_{t+1}| \alpha_{t+1}) \Pr(a_{t+1}| \alpha^k_{t+1}).
\]

Then we have that \( w^k \) should be proportional to the probability of \( \alpha_{t+1}| \alpha^k_{t+1} \) having its minimum exactly at \( y_{t+1} \). This can be shown to be exactly

\[
w^k = f_{a_{t+1}| \alpha_{t+1}}(y_{t+1}) \{1 - \Pr \alpha_{t+1}| \alpha_{t+1}(y_{t+1})\}
\]

\[
+ f_{a_{t+1}| \alpha_{t+1}}(y_{t+1}) \{1 - \Pr \alpha_{t+1}| \alpha_{t+1}(y_{t+1})\},
\]

while, having selected \( k \), we sample from

\[
\alpha_{t+1} = y_{t+1} \quad \text{with probability}
\]

\[
\lambda_{t+1} = \frac{f_{a_{t+1}| \alpha^k_{t+1}}(y_{t+1}) \{1 - \Pr \alpha_{t+1}| \alpha_{t+1}(y_{t+1})\}}{w^k},
\]

and then from

\[
\alpha_{t+1} = y_{t+1}, \alpha^k_{t+1}, \alpha_{t+1} > y_{t+1}.
\]

Likewise, \( \alpha_{t+1} = y_{t+1} \) with probability \( 1 - \lambda_{t+1} \).

3.3.6 Mixtures of Normals. Suppose that \( f(a_{t+1}| \alpha_t) \) is Gaussian, but the measurement density is a discrete mixture of normals \( \sum_{j=1}^P \lambda_j f_j(y_{t+1}| \alpha_{t+1}) \). Then we can perfectly sample from \( f(a_{t+1}, k|Y_{t+1}) \) by working with

\[
f(a_{t+1}, k, j|Y_{t+1}) \propto \lambda_j f_j(y_{t+1}| \alpha_{t+1}) f(a_{t+1}| \alpha^k_{t+1})
\]

\[
= w_{t,k,j} f_1(\alpha_{t+1}| \alpha^k_{t+1}, y_{t+1}).
\]

Then we sample from \( f(a_{t+1}, k, j|Y_{t+1}) \) by selecting the index \( k, j \) with probability proportional to \( w_{t,k,j} \) and then drawing from \( f_j(\alpha_{t+1}| \alpha^k_{t+1}, y_{t+1}) \). The disadvantage of this approach is that the complete enumeration and storage of \( w_{t,k,j} \) involves \( P \times M \) calculations. This approach can be trivially extended to cover the case where \( f(a_{t+1}| \alpha_t) \) is a mixture of normals. MCMC smoothing methods

---

**Figure 1.** Plot of the Angular Measurements From Origin, the True Trajectory (Solid Line, Crosses), the Particle Filtered Mean Trajectory (Dashed Boxes), and the Auxiliary Particle Mean Trajectory (Dotted Line, Circles). Ship moving southeast: \( T = 10, M = 300, R = 500 \).
for state-space models with mixtures have been studied by, for example, Carter and Kohn (1994) and Shephard (1994).

4. NUMERICAL EXAMPLE

4.1 A Time Series of Angles

4.1.1 The Model. In this section we compare the performance of the particle and auxiliary particle filter methods for an angular time series model: the bearings-only model. We consider the simple scenario described by Gordon et al. (1993). The observer is considered stationary at the origin of the $x-z$ plane, and the ship is assumed to gradually accelerate or decelerate randomly over time.

We use the following discretisation of this system, where

$$\alpha_{t+1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \alpha_t + \begin{pmatrix} \frac{1}{2} & 0 \\ 1 & 0 \\ 0 & \frac{1}{2} \\ 0 & 1 \end{pmatrix} u_t,$$

and $u_t \sim \text{NID}(0,1)$. (7)

In obvious notation $x_t$ and $z_t$ represent the ship's horizontal and vertical positions at time $t$ and $v_x_t$ and $v_z_t$ represent the corresponding velocities. The state evolution is thus a VAR(1) of the form $\alpha_{t+1} = T\alpha_t + H u_t$. The model indicates that the state evolution error arises because the accelerations are white noise. The initial state describes the ship's starting positions and velocities $\alpha_0 \sim \text{NID}(a_0, P_i)$. This prior, together with the state evolution of (7), describe the overall prior for the states.

Our model is based on a mean direction $\mu_t = \tan^{-1}(z_t/x_t)$. The measured angle is assumed to be wrapped Cauchy with density (see, e.g., Fisher 1993, p. 46)

$$f(y_t|\mu_t) = \frac{1}{2\pi} \frac{1 - \rho^2}{1 + \rho^2 - 2\rho \cos(y_t - \mu_t)},$$

$$0 \leq y_t < 2\pi, \quad 0 \leq \rho \leq 1,$$ (8)

where $\rho$ is the mean resultant length.

4.1.2 The Simulated Scenario. To assess the relative efficiency of the particle filter and the basic auxiliary method discussed in Section 3.2, we have closely followed the setup described by Gordon et al. (1993). They considered $\sigma_x = .001$ and $\sigma_x = .005$, where $z_t|\mu_t \sim \text{NID}(\mu_t, \sigma^2_x)$. We choose $\rho = 1 - \sigma_x^2$ (yielding the same circular dispersion) for our wrapped Cauchy density. The actual initial starting vector of this is taken to be $\alpha_0 = (-.05, .001, .2, .055)'$. In contrast to the method of Gordon et al. (1993), however, we wish to have an extremely accurate and tight prior for the initial state. This is because we want the variance of quantities arising from the filtered.
posterior density to be small, allowing us to formulate reasonably conclusive evidence about the relative efficiency of the auxiliary method to the standard method. Thus we take $a_1 = a_1$ and have a diagonal initial variance $P_1$ with the elements $.01 \times (.5^2, .005^2, .3^2, .01^2)$ on the diagonal.

Figure 1 illustrates a realization of the model for the foregoing scenario with $T = 10$. The ship is moving in a south-easterly direction over time. The trajectories given by the posterior filtered means from the particle method and the auxiliary method ($M = 300$ and $R = 500$ in both cases) are both fairly close to the true path despite the small amount of simulation used.

4.1.3 Monte Carlo Comparison. We now compare the two methods using a Monte Carlo study of the foregoing scenario with $T = 10$. The "true" filtered mean is calculated for each replication by using the auxiliary method with $M = 100,000$ and $R = 500,000$. Within each replication, the mean squared error (MSE) for the particle method for each component of the state over time is evaluated by running the method, with a different random number seed, $S$ times and recording the average of the resulting squared difference between the resulting particle filter's estimated mean and the "true" filtered mean. Hence for replication $i$, state component $j$, at time $t$, we calculate

$$\text{MSE}_{i,j,t}^P = \frac{1}{S} \sum_{s=1}^{S} (\hat{\alpha}_{i,j,s} - \tilde{\alpha}_{i,j,t})^2,$$

where $\hat{\alpha}_{i,j,s}$ is the particle mean for replication $i$, state component $j$, at time $t$, for simulation $s$ and $\tilde{\alpha}_{i,j,t}$ is the "true" filtered mean replication $i$, state component $j$, at time $t$. The log mean squared error (LMSE) for component $j$ at time $t$ is obtained as

$$\text{LMSE}_{j,t}^P = \log \frac{1}{\text{REP}} \sum_{i=1}^{\text{REP}} \text{MSE}_{i,j,t}^P.$$

The same operation is performed for the auxiliary method to deliver the corresponding quantity $\text{LMSE}_{j,t}^{AM}$. For this study, we use $\text{REP} = 40$ and $S = 20$. We allow $M = 4,000$ or $8,000$, and for each of these values we set $R = M$ or $2M$. Figure 2 shows the relative performance of the two methods for each component of the state vector over time. For each component $j$, the quantity $\text{LMSE}_{j,t}^{AM} - \text{LMSE}_{j,t}^P$ is plotted against time. Values close to 0 indicate that the two methods are broadly equivalent in performance; negative values indicate that the auxiliary method performs better than the standard particle filter.

The graphs give the expected result, with the auxiliary particle filter typically being more precise, but with the difference between the two methods falling as $R$ increases.

4.2 Stochastic Volatility

The basic SV model was defined in Section 3.3.4. Here we construct 100 times the compound daily returns on the U.S. dollar against the U.K. pound sterling from the first day of trading in 1997 and for the next 200 days of active trading. (This dataset is discussed in more detail in Pitt
Figure 4. Plot of the MSE Performance (on the Log Scale) of the Particle Filter to the Auxiliary-Based Particle Filter and an Adapted Particle Filter. The lower the number, the more efficient the method. (a) and (b) have $M = 2,000$; (c) and (d) have $M = 4,000$. (a) and (c) have $R = M$; (b) and (d) have $R = 2M$. (••••• SIR; □ — □ auxiliary SIR, ••••• adapted SIR.)

and Shephard, 1999, where we estimated the parameters of the model using Bayesian methods.) Throughout we take $\phi = .9702$, $\sigma = .178$, and $\beta = .5962$, the posterior means of the model for a long time series of returns up until the end of 1996.

Figure 3 graphs these daily returns against time. The figure displays the estimated quantiles of the filtering density, $f(\beta \exp(\alpha \beta t) | Y_t)$ computed using an auxiliary particle filter. Throughout the series, we set $M = 5,000$ and $R = 6,000$. We have also displayed the posterior mean of the filtering random variable. This is always very slightly above the posterior median, as $\alpha_t | Y_t$ is very close to being symmetric.

The figure shows that the filtered volatility jumps up more quickly than it tends to go down. This reflects the fact that the volatility is modeled on the log scale.

4.2.1 Simulation Experiment. To compare the efficiency of the simple particle filter, our basic auxiliary particle filter, and the (rejection-based) fully adapted particle filter discussed in Section 3.3.4, we again conducted a simulation experiment measuring MSE for each value of $t$ using the foregoing model and again having $n = 50$. We simulated the data using the model parameters discussed earlier. Figure 4 shows the results (using a log scale). To make the problem slightly more realistic and challenging, we set $\epsilon_{21} = 2.5$ for each series, so there is a significant outlier at that point. For this study, we set $REP = 40$ and $S = 20$. We allow $M = 2,000$ or $4,000$, and for each of these values we set $R = M$ or $2M$. For the rejection-based particle filter algorithm, it only makes sense to take $M = R$, and so when $R > M$, we repeat the calculations as if $M = R$. Finally, the rejection-based method takes approximately twice the time of the SIR-based particle filter when $M = R$.

The plot shows that the fully adapted particle filter is considerably more accurate than the other particle filters. It also has the advantage of not depending on $R$. The auxiliary particle filter is more efficient than the plain particle filter, but the difference is small, reflecting the fact that for the SV model, the conditional likelihood is not very sensitive to the state.

5. CONCLUSION

In this article we have studied the weaknesses of the very attractive particle filtering method proposed by Gordon et al. (1993). The SIR implementation of this method is not robust to outliers for two different reasons: sampling efficiency and the unreliability of the empirical prediction density in the tails of the distribution. We have introduced an auxiliary variable into the particle filter to overcome the first of these problems, providing a powerful framework that is as simple as SIR, but more flexible and reliable.
REFERENCES


