ISSN 1471-0498

DEPARTMENT OF ECONOMICS

DISCUSSION PAPER SERIES

LEARNING AND FILTERING VIA SIMULATION: SMOOTHLY JITTERED PARTICLE FILTERS

Thomas Flury and Neil Shephard

Number 469
December 2009

Manor Road Building, Oxford OX1 3UQ

Learning and filtering via simulation: smoothly jittered particle filters

Thomas Flury
Oxford-Man Institute, University of Oxford,
Eagle House, Walton Well Road, Oxford OX2 6ED, UK
& Department of Economics, University of Oxford
thomas.flury@economics.ox.ac.uk

Neil Shephard
Oxford-Man Institute, University of Oxford,
Eagle House, Walton Well Road, Oxford OX2 6ED, UK
& Department of Economics, University of Oxford
neil.shephard@economics.ox.ac.uk

December 11, 2009

Abstract

A key ingredient of many particle filters is the use of the sampling importance resampling algorithm (SIR), which transforms a sample of weighted draws from a prior distribution into equally weighted draws from a posterior distribution. We give a novel analysis of the SIR algorithm and analyse the jittered generalisation of SIR, showing that existing implementations of jittering lead to marked inferior behaviour over the basic SIR algorithm. We show how jittering can be designed to improve the performance of the SIR algorithm. We illustrate its performance in practice in the context of three filtering problems.

Keywords: Importance sampling, particle filter, random numbers, sampling importance resampling, state space models

1 Introduction

Particle filters are now a standard way of carrying out non-linear, non-Gaussian filtering. Reviews of the topic include Doucet, de Freitas, and Gordon (2001) and Cappe, Godsill, and Moulines (2007). Since Gordon, Salmond, and Smith (1993) various researchers have explored how adding small amounts of randomness might improve aspects of particle filters. This is particularly important for slowly moving or time invariant states where particle degeneracy is a serious issue. The initial suggestion was improved by the shrinkage method of Liu and West (2001) and the kernel density estimator approach advocated by Musso, Oudjane, and LeGland (2001) and Stravropoulos and Titterington (2001). However, all these methods seem to lead to significant bias which damages the performance of the particle filter.

Here we give a fresh analysis of jittering. It will be based on a finite sample study of the sampling importance resampling (SIR) algorithm, which was introduced by Rubin (1987) and Rubin (1988).

We provide a simple interpretation of resampling in the context of the classic approach to SIR and then study the jittered alternative. We show how to sensibly jitter and prove that the traditional approaches in the particle filter literature deliver serious bias. Our alternative, which we call the "smoothly jittered particle filter", avoids these features, is dimensionless and we think should be used in practice. We illustrate these methods.

This paper is a contribution to the literature trying to deal with on-line Bayesian inference for static model parameters which index some state space model. Alternative approaches consist in using sufficient statistics, Fearnhead (2002) and Storvik (2002), or a combination of sufficient statistics and the auxiliary particle filter from Pitt and Shephard (1999), which has been proposed in Johannes, Polson, and Stroud (2009) and Johannes and Polson (2009). There have been some recent developments in the recursive maximum likelihood estimation framework. Poyiadjis, Doucet, and Singh (2009) derive an algorithm for on-line parameter estimation using on-line estimates of the score to guide recursive maximum likelihood estimation. Del Moral, Doucet, and Singh (2009) provide a forward smoothing algorithm which permits recursive parameter estimation using an on-line expectation maximisation (EM) type algorithm.

This paper draws some insights from the literature on smoothed bootstraps, although the details and motivation differ. Leading papers there include Efron (1982) and Silverman and Young (1987). We were also influenced in our thinking by the smooth likelihood particle filter by Pitt (2002), although our objectives and scope are rather different. West (1993) is somewhat related to our paper, but our focus is on the particle filtering context.

The structure of this paper is as follows: In Section 2 we analyse the finite sample properties of standard resampling. In Section 3 we extend this to smooth jittering. In Section 4 we provide numerical examples, while Section 5 delivers some extra remarks. Conclusions are drawn in Section 6. An Appendix holds proofs of various results stated in the paper.

2 Understanding SIR

2.1 Properties

We begin this section by recalling the standard SIR algorithm. Suppose we wish to learn about the p-dimensional α given data y. Recall Bayes theorem

$$f(\alpha|y) = \frac{1}{c} f(y|\alpha) f(\alpha), \quad c = f(y).$$

Here $f(\alpha)$ is the prior, $f(y|\alpha)$ the likelihood, $f(y)$ the marginal likelihood and $f(\alpha|y)$ the posterior.

Our desire is to obtain equally weighted draws representing the posterior $f(\alpha|y)$ by using the SIR algorithm. The idea of importance sampling is to represent a target distribution via weighted

draws from an importance density, see Liu (2001, p. 31) for a review. By resampling – i.e. sampling with replacement from these weighted draws – we obtain an equally weighted sample from the target distribution.

Suppose $\alpha^{(1)}, ..., \alpha^{(n)}$ are independent and identically distributed (i.i.d.) draws from the prior $f(\alpha)$. After assigning each particle $\alpha^{(i)}$ the corresponding weight $f(y|\alpha^{(i)})/(n\widehat{c})$, where $\widehat{c} = \frac{1}{n}\sum_{i=1}^{n} f(y|\alpha^{(i)})$, we have a weighted sample from the posterior $f(\alpha|y)$. Rubin's resampling step can be thought of as sampling with replacement from the "empirical posterior distribution function"

$$F_n(\alpha|y) = \frac{\widehat{d}}{\widehat{c}}, \quad \widehat{d} = \frac{1}{n}\sum_{i=1}^{n} f(y|\alpha^{(i)}) I(\alpha^{(i)} \leq \alpha). \tag{1}$$

This works because as $n \to \infty$ the strong law of large numbers implies $\widehat{c} \xrightarrow{p} f(y)$ and $\widehat{d} \xrightarrow{p} f(y)F(\alpha|y)$, which implies that

$$F_n(\alpha|y) \xrightarrow{p} F(\alpha|y),$$

where the convergence is in fact uniform over α.

The preceding discussion should make it clear why the performance of the resampling step is intrinsically linked to the statistical properties of (1). For the univariate state case, instead of thinking about the resampling step as sampling with replacement from (1), think of applying the inverse of F_n to *i.i.d.* uniform $[0, 1]$ random numbers. By the above results we have

$$\alpha_n = F^{-1}_{\alpha|y,n}(U) \xrightarrow{L} \alpha|y, \quad U \sim U(0, 1).$$

This U-th quantile can be compared to an idealised draw $\alpha = F^{-1}_{\alpha|y}(U) \overset{L}{=} \alpha|y$. In Figure 1 we illustrate how deviations of F_n from F lead to deviations in α_n from α. The quality of the resampled particles will be determined by the quality of the estimator of F.

If we suppose α is univariate we can provide a strong theoretical justification for that link. Let the resampling error be $v_n = \sqrt{n}\,(\alpha_n - \alpha)$. It is easy to see that

$$v_n = -\frac{1}{f\,(\alpha)} u_n + O_p(n^{-1/2}), \quad u_n = \sqrt{n}\,\{F_n(\alpha|y) - F(\alpha|y)\}\,.$$

The properties of v_n are inherited from u_n. Typically v_n will be imprecise in the tails. Reducing the size of u_n should improve v_n. In the next section we show how to do this.

Having motivated the analysis in terms of the distribution function estimation error

$$u_n = \sqrt{n}\,(F_n - F)\,,$$

from the standard SIR algorithm, we provide in Proposition 1 the ingredients to compute the mean squared error (mse) of u_n.

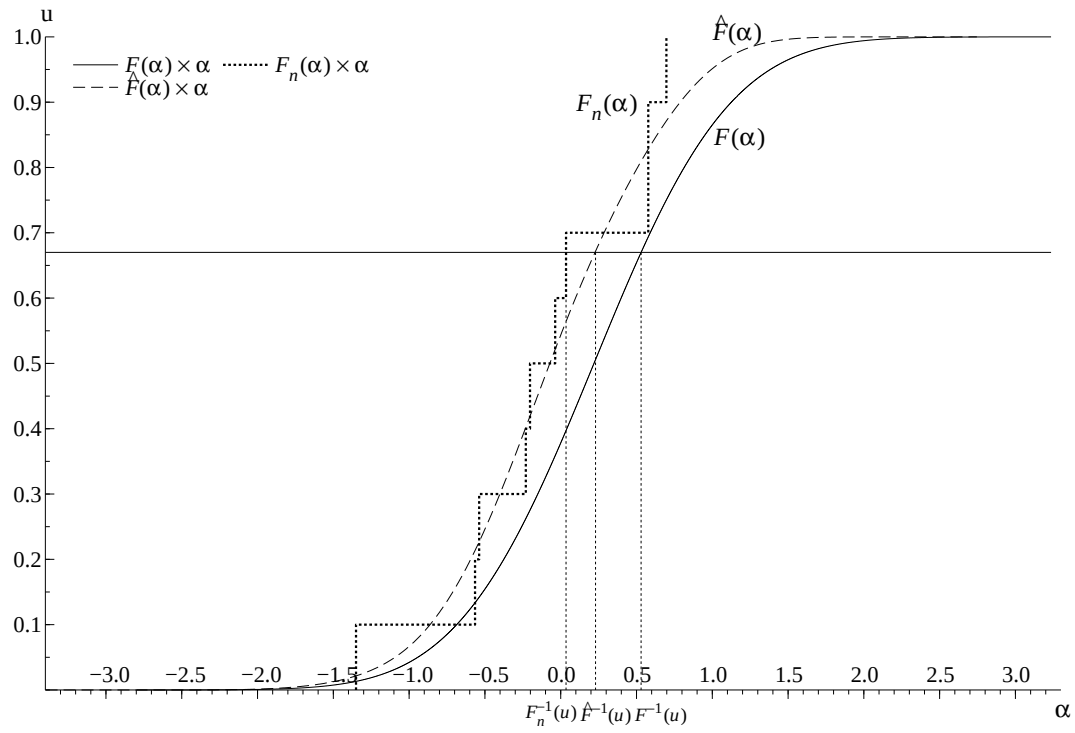

Figure 1: The solid line represents the true cumulative distribution function, dotted is the empirical distribution function $F_n(\alpha)$ based on $M = 10$ particles, and dashed the smooth distribution function estimate $\widehat{F}(\alpha)$ based on the same $M = 10$ particles. The horizontal line at $u = 0.67$ indicates which point is resampled when using SIR or the smooth jitter.

Proposition 1 *Define* $u_n = \sqrt{n}(F_n - F)$*, then*

$$\mathrm{E}[u_n] = 0, \quad \mathrm{Var}(u_n) = \{1 - 2F(\alpha|y)\} R(\alpha|y) + F(\alpha|y)^2 R(\infty|y),$$

with $R(\alpha|y) = \frac{1}{f(y)} \int_{-\infty}^{\alpha} f(y|x) \mathrm{d}F(x|y)$*, where we assume* $R(\infty|y) < \infty$*.*

Proof. Given in the Appendix.

In the case where $y \perp\!\!\!\perp \alpha$ then $f(y) = f(y|\alpha)$, so $R(\alpha|y) = F(\alpha)$, then this simplifies to the well known result on the empirical distribution function that

$$\mathrm{Var}(u_n) = \{1 - F(\alpha)\} F(\alpha).$$

We conclude this section by presenting a useful link to a performance measure often used for particle filters. This link provides an insight about the dependence of the asymptotic performance of the F_n (in terms of variance) on the quality of the sampling scheme. To do so we define $r(\alpha|y) = f(y|\alpha)f(y|\alpha)/f(y)$ and compute

$$R(\infty|y) = \int r(\alpha|y) \mathrm{d}\alpha = \frac{1}{f(y)} \int f(y|\alpha) \mathrm{d}F(\alpha|y) = \int \left(\frac{f(y|\alpha)}{f(y)} \right)^2 \mathrm{d}F(\alpha).$$

This term will be important later. If $\alpha^{(i)} \overset{i.i.d.}{\sim} F(\alpha)$ then the estimator

$$\widehat{R}(\infty|y) = n\sum_{i=1}^{n}\left(W^{(i)}\right)^2 \xrightarrow{p} R(\infty|y), \quad W^{(i)} = \frac{w^{(i)}}{\sum_{j=1}^{n} w^{(j)}}, \quad w^{(i)} = f(y|\alpha^{(i)}),$$

has the property that $\frac{1}{n} \leq \sum_{i=1}^{n}\left(W^{(i)}\right)^2 \leq 1$. In the particle filter literature one often calls $\left(\sum_{i=1}^{n}\left(W^{(i)}\right)^2\right)^{-1}$ the "effective sample size", which is taken as a measure of sample impoverishment. Linking this to the variance of u_n we see that it will be small if $\widehat{R}(\infty|y)$ is close to 1 – i.e. the weights are quite even. However, the variance can be larger for very uneven weights, when $\widehat{R}(\infty|y)$ is close to n. $\widehat{R}(\infty|y)$ will reappear later when we select a bandwidth.

Having introduced the intuition of thinking of resampling as inverting a distribution function and having established the asymptotic properties of the F_n used in SIR, the way forward is now clear. We will introduce an estimator of F which is superior to F_n and show that this will lead to improved asymptotic properties of the resampling step. Before we do this we recall how SIR is used in traditional particle filters and illustrate the particle degeneracy problem.

2.2 Particle filter version: iterating

The SIR method can be used iteratively when analysing state space models (e.g. Durbin and Koopman (2001)), which have conditionally independent observations $y_t|\alpha_t$ from the known $f(y_t|\alpha_t)$, while the hidden states α_t are Markovian. The particle filter sequentially generates samples from $\alpha_t|\mathcal{F}_t$ for $t = 1, \ldots, T$, where $\mathcal{F}_t$ is the history of the y process up to time t, starting with a prior on $\alpha_1|\mathcal{F}_0$. The particle filter is an extension of the famous Kalman filter to non-Gaussian, non-linear models.

Suppose we can evaluate $f(y_t|\alpha_t)$ and form a simulator $\alpha_t|\alpha_{t-1}$. We start by drawing a sample of size n from $\alpha_1|\mathcal{F}_0$ and weight these, employing $f(y_1|\alpha_1)$ as the likelihood, to get an approximation for $\alpha_1|\mathcal{F}_1$. After sampling n times with replacement from this we have completed the first SIR step. For each sampled value we use the simulator $\alpha_t|\alpha_{t-1}$ as importance density, delivering n draws from $\alpha_2|\mathcal{F}_1$. These draws are weighted using the likelihood $f(y_2|\alpha_2)$, which produces weighted draws representing $\alpha_2|\mathcal{F}_2$, from which we resample to complete the second SIR step. These steps are then iterated through time. This standard particle filter has been tremendously successful in many applications and a great deal of statistical theory has been developed for it, see e.g. Del Moral (2004).

2.3 Particle degeneracy

A main weakness of particle filters is the well known particle degeneracy problem, which is important when states change very slowly or not at all through time. To get to the heart of the problem consider Example 1.

Example 1 *Let*

$$y_t = \alpha_t + \varepsilon_t, \quad \varepsilon_t \sim N\left(0, \sigma_\varepsilon^2\right), \quad \alpha_t = \alpha_{t-1}, \quad t = 1, \ldots, T, \quad \alpha_1 | \mathcal{F}_0 \sim N\left(\mu_0, \sigma_0^2\right), \tag{2}$$

so α_t is a time-invariant parameter and we wish to learn about the posterior density $f(\alpha_t|\mathcal{F}_t)$. The true posterior is given by

$$\alpha_t | \mathcal{F}_t \sim N\left(\sigma_t^2\left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{s=1}^t y_s}{\sigma_\varepsilon^2}\right), \sigma_t^2\right), \quad \sigma_t^{-2} = \frac{1}{\sigma_0^2} + \frac{t}{\sigma_\varepsilon^2}, \tag{3}$$

which the population of particles will approximate at each t. We simulate t observations of y, using $\alpha = 0.439$ and then run the particle filter on the state space model (2) with $\sigma_\varepsilon^2 = \sigma_0^2 = 1$, $\mu_0 = 0$, using $n = 20$ particles. The left hand side of Figure 2 shows the particle paths for $t = 20$. The draws from the prior indicate quite a lot of diversity, but many distinct particles die out and by the time we reach $t = 20$ only one unique particle value is left. Due to the SIR structure there is no chance any new values of α can appear as the particle filter iterates. This is the classic degeneracy problem. Some improvement can be obtained by using various types of stratified sampling of the particles, but this does not overcome the problem that the support of the particles is entirely determined at $t = 1$. Figure 3 illustrates the impact of particle degeneracy on the estimate of the posterior standard deviation σ_t. In the left graph we plot the 5%, 50% and 95% quantiles of the SIR based particle filter ($n = 100$) estimates of σ_t from $N = 1,000$ Monte Carlo replications. The true σ_t is the same in each replication, as the posterior standard deviation does not depend upon y. It only takes $t = 100$ for the median $\widehat{\sigma}_t$ to become degenerate. In the right graph of Figure 3 we plot the inefficiency of the SIR particle filter relative to i.i.d. sampling from the true posterior (dotted line). Inefficiency is measured by the ratio of the mean squared error for estimating σ_t with SIR to that of using i.i.d. draws from the true posterior. It suggests that 800 particles have roughly the same information about σ_t as a single draw from the posterior when $t = 500$. We will return to this example later.

3 Smooth SIR

We now return to the static case of having draws from a prior and wishing to sample from a posterior $F(\alpha|y)$. We propose to use a smooth distribution function estimator instead of F_n in the resampling step. This is equivalent to jittering the particles obtained from the SIR algorithm.

Since Gordon, Salmond, and Smith (1993) various authors have suggested adding small amounts of randomness to each resampled particle $\alpha^{(i)}$ — that is jittering. The introduction of all these approaches has been driven by the desire to overcome particle degeneracy for slow moving or static states in dynamic state space models. Unfortunately, these previous ways of jittering have been found not to work very well in practice and we will show that they do indeed damage the asymptotic properties of the SIR algorithm. We deduce a method which overcomes this problem.

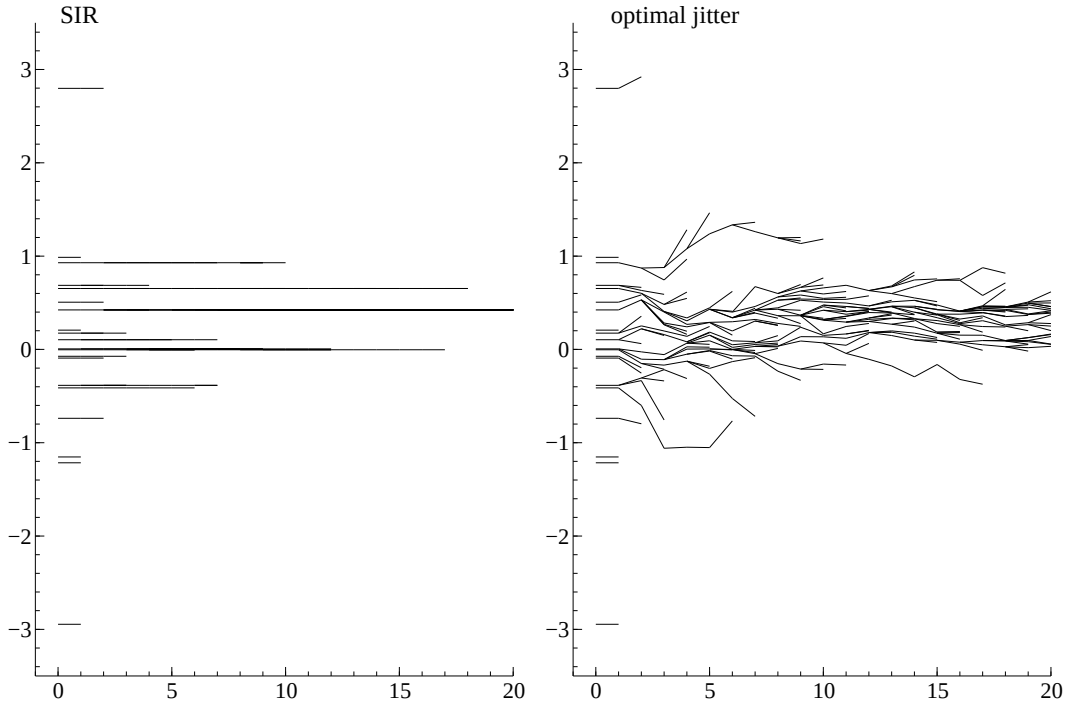

Figure 2: The particle paths from a SIR (left) and a jittered (right) particle filter for learning $\alpha_t|\mathcal{F}_t$. The underlying model is that of Example 1, with $n = 20$ particles and $t = 20$ observations. The same random numbers have been used for both graphs.

3.1 Smooth distribution function estimation

We replace $F_n(\alpha|y)$ by the alternative estimator

$$\widehat{F}(\alpha|y) = \frac{1}{\widehat{c}}\frac{1}{n}\sum_{i=1}^{n} f(y|\alpha^{(i)})G\left(\frac{\alpha_1 - \alpha_1^{(i)}}{h_1}, ..., \frac{\alpha_p - \alpha_p^{(i)}}{h_p}\right). \tag{4}$$

where h_1,..., h_p are called the bandwidths and are tuning parameters. We will assume G is absolutely continuous and non-decreasing such that

$$G(t) = \int_{-\infty}^{t} g(u)\mathrm{d}u, \quad g(u) = \prod_{j=1}^{p} g_1(u_j), \quad G(t) = \prod_{j=1}^{p} G_1(t_j),$$

where g is a p-dimensional probability density, g_1 is a univariate density which we assume symmetric about zero and scaled so that

$$\int u_j g_1(u_j)\mathrm{d}u_j = 0, \quad \int u_j u_k g(u)\mathrm{d}u = 1_{j=k}.$$

If $f(y|\alpha) = f(y)$ then $\widehat{F}(\alpha|y)$ is the smoothed distribution function estimator of $F(\alpha)$, which was introduced by Nadaraya (1964) and extensively studied by Azzalini (1981), Jones (1990), Cheng and Sun (2006) and many others.

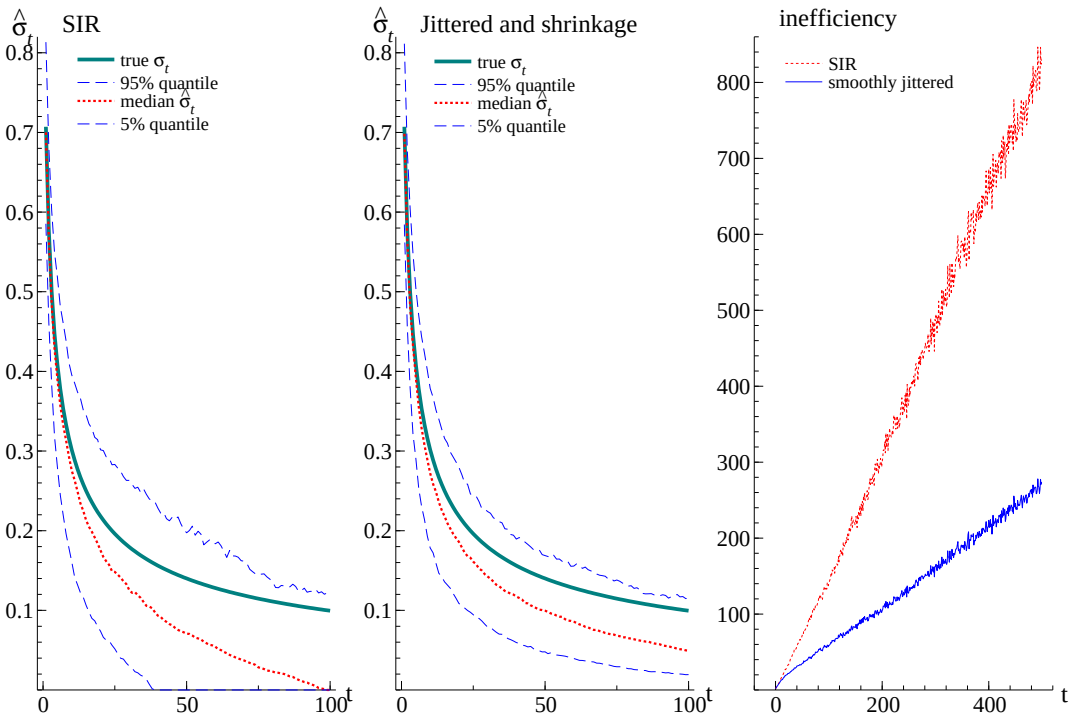

Figure 3: Left: true posterior standard deviation σ_t (solid), the 5% and 95% quantiles (dashed) and median (dotted) of $\widehat{\sigma}_t$ from SIR across $N = 1,000$ Monte Carlo experiments. Middle: The same from the smoothly jittered particle filter. The underlying model is that of Example 1, with $n = 100$ particles and $t = 100$ observations. The same random numbers have been used for both graphs. Right: Inefficiency relative to i.i.d. samples, measured as mse of the estimators of σ_t. The underlying model is that of Example 1, with $n = 5,000$ particles and $t = 500$ observations, based on $N = 4,000$ replications. The same type of results appear if n is much larger.

Sampling from (4) is easy: we perform SIR and add to each element of the p-dimensional particle some jitter $h_j\epsilon_j$, where $\epsilon_j \sim g_1$ are independently sampled over $j = 1, \ldots, p$. The variance h_j^2 will be chosen such as to optimise the asymptotic behaviour of the smooth resampling scheme and g_1 is chosen to make it easy to sample from.

For the $p = 1$ case, Figure 1 depicts that sampling from (4), or jittering the SIR samples, is equivalent to sampling from the estimated quantile function

$$\widehat{\alpha} = \widehat{F}^{-1}_{\alpha|y}(U), \quad U \sim U(0,1).$$

We now analyse the properties of such a jittered sample, which simply derive from the properties of the estimation error

$$\widehat{u} = \sqrt{n}\left(\widehat{F} - F\right).$$

To simplify notation we write ϵ_j as being drawn from g_1 and

$$\int_{-\infty}^{\infty} \epsilon_j g_1(\epsilon_j) G_1(\epsilon_j) \, \mathrm{d}\epsilon = \mathrm{E}_{g_1}\left[\epsilon_j G_1(\epsilon_j)\right].$$

Section 5.1 will give an alternative analysis of the error using a different norm.

Proposition 2 *Define* $\widehat{u} = \sqrt{n}\left(\widehat{F} - F\right)$*, then for the p-dimensional case*

$$\mathrm{E}\left[\widehat{u}\right] = \sqrt{n}\frac{1}{2}\left\{\sum_{j=1}^{p} h_j^2 \frac{\partial^2 F(\alpha|y)}{\partial a_j^2}\right\} + \sqrt{n}O(\max_j h_j^4) \tag{5}$$

and

$$\begin{aligned}\mathrm{Var}\left(\widehat{u}\right) &= \left\{1 - 2F(\alpha|y)\right\} R(\alpha|y) + F(\alpha|y)^2 R(\infty|y) \\ &\quad -2\left\{\sum_{j=1}^{p} h_j \frac{\partial R(\alpha|y)}{\partial \alpha_j}\right\} \mathrm{E}_{g_1}\left[\epsilon_1 G_1(\epsilon_1)\right] + O(\max_j h_j^2) + nO(\max_j h_j^4),\end{aligned}$$

with

$$R(\alpha|y) = \frac{1}{f(y)}\int_{-\infty}^{\alpha} f(y|x)\mathrm{d}F(x|y),$$

where we assume $R(\infty|y) < \infty$*.*

Proof. Given in the Appendix.

Comparing these properties with those of F_n we see that we can reduce the variance of the distribution function estimator at the cost of introducing some bias.

3.2 Optimal jittering

We now work towards finding a simple rule for choosing the bandwidth h. We start with a direct comparison between the SIR algorithm and our smooth jittering to state the optimal rate of convergence for h. Then we argue for independent jittering and introduce the integrated mean squared error as performance measure, which is easy to optimise over in practice. Based on this we introduce a simple bandwidth selection rule.

Combining the results from Section 2 and Section 3.1 we find the mean squared error

$$mse\left(u_n\right) - mse\left(\widehat{u}\right) = a_0 \sum_{j=1}^{p} h_j \frac{\partial R(\alpha|y)}{\partial \alpha_j} - \frac{n}{4}\left\{\sum_{j=1}^{p} h_j^2 \frac{\partial^2 F(\alpha|y)}{\partial a_j^2}\right\}^2, \quad a_0 = 2\mathrm{E}_{g_1}\left\{\epsilon_1 G_1(\epsilon_1)\right\}.$$

If $h_j = c_j n^{-1/3}$ then

$$mse\left(u_n\right) - mse\left(\widehat{u}\right) = n^{-1/3}\left[a_0 \sum_{j=1}^{p} c_j \frac{\partial R(\alpha|y)}{\partial \alpha_j} - \frac{1}{4}\left\{\sum_{j=1}^{p} c_j^2 \frac{\partial^2 F(\alpha|y)}{\partial a_j^2}\right\}^2\right],$$

The rate of the difference in the scaled mse is not influenced by p, it is always $n^{-1/3}$.

To select sensible $\{c_j\}$ we will work with the optimal bandwidth in each univariate dimension. Define

$$u_{j,n} = u_n(\infty, ..., \alpha_j, ..., \infty), \quad \widehat{u}_j = \widehat{u}\left(\infty, ..., \alpha_j, ..., \infty\right), \quad j = 1, 2, ..., p.$$

Then

$$mse(u_{j,n}) - mse(\widehat{u}_j) = a_0 h_j \frac{\partial R(\infty, ..., \alpha_j, ..., \infty|y)}{\partial \alpha_j} - \frac{n}{4}\left\{h_j^2 \frac{\partial^2 F(\alpha_j|y)}{\partial \alpha_j^2}\right\}^2.$$

The integrated mean squared error (imse), summed over j, is

$$\sum_{j=1}^{p} imse(u_{j,n}) - imse(\widehat{u}_j) = a_0 \sum_{j=1}^{p} h_j R(\infty|y) - \frac{n}{4}\sum_{j=1}^{p} h_j^4 \int f'(\alpha_j|y)^2 \mathrm{d}\alpha_j.$$

The improvement in the integrated mean squared error $imse\,(u_{j,n}) - imse\,(\widehat{u}_j)$ is maximal with bandwidth

$$\widehat{h}_j = a_0^{1/3}\left(\frac{R(\infty|y)}{\int f'(\alpha_j|y)^2 \mathrm{d}\alpha}\right)^{1/3} n^{-1/3}.$$

Example 2 *In the case where the likelihood is flat $f(y) = f(y|\alpha)$ then $R(\infty|y) = 1$ for all α. If additionally $p = 1$ then this becomes*

$$imse\,(u_n) - imse\,(\widehat{u}) = ha_0 - \frac{1}{4}nh^4 \int f'(\alpha)^2 \mathrm{d}\alpha.$$

This is the integrated version of the Azzalini (1981) result.

The improvement in accuracy $imse\,(u_{j,n}) - imse\,(\widehat{u}_j)$ is

$$\begin{aligned}
&a_0^{1/3}\left(\frac{R(\infty|y)}{\int f'(\alpha_j|y)^2 \mathrm{d}\alpha_j}\right)^{1/3} n^{-1/3} a_0 R(\infty|y) \\
&\quad - n a_0^{4/3}\left(\frac{R(\infty|y)}{\int f'(\alpha_j|y)^2 \mathrm{d}\alpha_j}\right)^{4/3} n^{-4/3}\frac{1}{4}\int f'(\alpha_j|y)^2 \mathrm{d}\alpha_j \\
&\quad = \frac{3}{4} n^{-1/3} a_0^{4/3} \frac{R(\infty|y)^{4/3}}{\left(\int f'(\alpha_j|y)^2 \mathrm{d}\alpha_j\right)^{1/3}} = \frac{3}{4^{1/3}}\mathcal{G}\frac{R(\infty|y)^{4/3}}{\left(\int f'(\alpha_j|y)^2 \mathrm{d}\alpha_j\right)^{1/3}} n^{-1/3},
\end{aligned}$$

where $\mathcal{G} = \left(\mathrm{E}_{g_1}\left\{\epsilon_1 G_1(\epsilon_1)\right\}\right)^{4/3}$. Hence the smoothing improves the accuracy of the estimated distribution function, but the improvement is not in terms of the first order theory. The most important aspect of this result is that the rate of convergence is unaffected by jittering and the precision somewhat improved.

The choice of the jittering distribution $\epsilon_1 \sim G_1$, only influences the universal constant $\mathcal{G}$. The jitter is assumed to have unit variance, so we maximise $\mathrm{E}_{g_1}\left[|\epsilon_1|\right]$ subject to this constraint. Jones (1990) showed $\mathcal{G}$ is maximal if g_1 is a uniform distribution, but that many other distributions are nearly as efficient. The same result applies here. Table 1 gives numerical approximations to $\mathcal{G}$, based on ten million random numbers from the relevant distributions. In particular the standard normal and Laplace are 3% and 11% less efficient, respectively, than the uniform. However, the normal and Laplace have the advantage for particle filters of having support on the entire real line.

	$g(x)$	$\mathrm{E}_{g_1}\left[\epsilon_1 G_1(\epsilon_1)\right]$	$\mathcal{G}$
uniform	$\frac{1}{2\sqrt{3}} I\left(x \in \left[-\sqrt{3}, \sqrt{3}\right]\right)$	0.2890	0.1911
normal	$(2\pi)^{-1/2} \exp\left(-x^2/2\right)$	0.2819	0.1849
Laplace	$\exp\left(-\sqrt{2}\,\lvert x\rvert\right)/\sqrt{2}$	0.2654	0.1706

Table 1: Performance of various jittering distributions. $\mathcal{G}$ denotes their efficiency, with higher numbers being better. Jones (1990) proved that the uniform is the most efficient.

3.2.1 Simple bandwidth rules

To compute the optimal h we need to well approximate

$$\frac{(R(\infty|y))^{1/3}}{\left(\int f'(\alpha_j|y)^2 \mathrm{d}\alpha_j\right)^{1/3}}.$$

We will use a Gaussian approximation to the posterior for $\int f'(\alpha_j|y)^2 \mathrm{d}\alpha_j$. By assuming $\alpha_j|y \sim N(\mu_{\alpha_j|y}, \sigma^2_{\alpha_j|y})$ we find (e.g. Hansen (2004, p. 2))

$$\left\{\int f'(\alpha_j|y)^2 \mathrm{d}\alpha_j\right\}^{-1} = \sigma^3_{\alpha_j|y} 4\sqrt{\pi}.$$

We can estimate $\sigma_{\alpha_j|y}$ robustly using the scaled interquartile range $\widehat{IQR}/1.349$, where

$$\widehat{IQR} = F^{-1}_{j,n}\left(0.75|y\right) - F^{-1}_{j,n}\left(0.25|y\right),$$

and $F_{j,n}\left(\alpha_j|y\right)$ is (1). Thus, if we use Gaussian jitter this suggests the rule:

$$\widehat{h}_j = 1.59 \left\{\widehat{R}(\infty|y)\right\}^{1/3} \widehat{\sigma}_{\alpha_j|y} n^{-1/3}, \quad \widehat{\sigma}_{\alpha_j|y} = \widehat{IQR}_j/1.349.$$

At the end of Section 2 we have seen how we can consistently estimate $R(\infty|y)$. Linking $\widehat{R}(\infty|y) = n\sum_{i=1}^n \left(W^{(i)}\right)^2$ to the effective sample size $\mathrm{ESS} = (\sum_{i=1}^n (W^{(i)})^2)^{-1}$, which was introduced as a rough measure of the variance of the importance weights and serves as an indicator of the quality of the weighted sample, we may write the bandwidth rule

$$\widehat{h}_j = 1.59\ \widehat{\sigma}_{\alpha_j|y}\ \mathrm{ESS}^{\ -1/3}.$$

When we are smoothing a weighted sample, we have the intuitive result that instead of using the actual number of draws n, we scale the bandwidth by the ESS, which will be smaller if the particle sample is of poor quality, i.e. the importance sampling weights are quite uneven. Hence it is adaptive to the problem.

We return to Example 1 to illustrate how jittering affects particle degeneracy in practice.

Example 3 (Example 1 continued) *We now add jittering. At each time t we compute the "optimal" degree of jitter*

$$\widehat{h}_t = 1.59 \left\{\widehat{R}_t(\infty|\mathcal{F}_t)\right\}^{1/3} \widehat{\sigma}_{\alpha_t|\mathcal{F}_t} n^{-1/3}.$$

The resulting particle paths are given in the right hand side of Figure 2. This shows the replenishment of the support of the particles as t increases.

3.3 Damaging particles by jittering

In the particle filter literature Musso, Oudjane, and LeGland (2001) and Stravropoulos and Titterington (2001, p. 298) have suggested jittering the data by using kernel density estimation with the choice of $h = cn^{-1/5}$ in the univariate state case. Liu and West (2001) and Gordon, Salmond, and Smith (1993) have suggested methods which result in $h \propto 1$.

Lemma 1 *If we set $h_d = d_0 n^{-1/5}$, where d_0 is a constant and $p = 1$ then*

$$n^{4/5}\text{mse}\left(\widehat{F}(\alpha|y)\right) = d_0^4\left\{f'(\alpha|y)\right\}^2 + O(n^{-1/5}).$$

If $h \propto 1$ then $\widehat{F}(\alpha|y)$ is inconsistent.

Proof. Trivial from Proposition 2.

This Lemma shows that the existing suggestions in the literature lead to excessive jittering, which slows down the rate of convergence of the SIR method and the error is entirely dominated by the bias. In the literature, for higher dimensional problems, the jittering is typically increased with $h = d_0 n^{-1/(4+p)}$ where p is the dimension. The rate of convergence of the mse is not n^{-1} but $n^{-4/(4+p)}$. So if $p = 4$ then the rate at which the mse falls is $n^{-1/2}$ which is undesirable.

3.4 Further improvements from shrinkage

After the jittering step the equally weighted particle cloud no longer is an exact representation of the posterior distribution. The obvious remedy is to treat the jitter as another importance sampling step and reweight the jittered particles. These weights will be more equal than the pre-resampling weights. The post-jittering weights would then be passed forward in the particle filter to the next time period where they would be combined with the appropriate likelihood function. In sequential applications it is difficult to compute the weights since the true posterior is not easily available.

Instead of using importance sampling weights, we could weight the jittered data so that it shares some features of the unjittered data. Desired features could be the same mean and variance, for example. A principled way of weighting the data is to select the weights by maximising the empirical, Euclidian or entropy likelihoods to minimally move the weights away from $1/n$ in order to satisfy the required features. See the general discussion of these methods in Owen (2001). These types of nonparametric methods are somewhat more computationally demanding and we therefore advocate a simple yet powerful linear shrinkage rule.

The reweighting schemes look more attractive in theory than linear shrinkage, as they may be more effective at dealing with multimodality in the filtering density. Given the effectiveness of linear shrinking in this paper we intend to explore this alternative in future work.

We now introduce linear shrinkage to further improve the performance of the jittered particle filter. West (1993) has advocated this approach in the context of approximating posterior distributions by mixtures. He does not consider the particle filter approach and advocates the kernel density bandwidth choice $h = (4/(1+2p))^{1/(1+4p)} n^{-1/(1+4p)}$. Liu and West (2001) use shrinkage for the particle filter but choose $h \propto 1$.

The idea is simple and in the univariate case we can write

$$\alpha_{t+1}^{(i)} = \widehat{\mu}_t + \sqrt{\frac{\widehat{\sigma}_t^2 - \widehat{h}_t^2}{\widehat{\sigma}_t^2}} \left(\alpha_t^{(i)} - \widehat{\mu}_t \right) + \widehat{h}_t \epsilon_t^{(i)}, \quad \epsilon_t^{(i)} \overset{i.i.d.}{\sim} N(0,1),$$

where $\widehat{\mu}_t$ is an estimate of the posterior mean, $\widehat{\sigma}_t^2$ a robust estimate of the posterior variance and $\alpha_t^{(i)}$ the resampled particle. Note that

$$\frac{\widehat{\sigma}_t^2 - \widehat{h}_t^2}{\widehat{\sigma}_t^2} = 1 - 1.59^2 \left\{ \widehat{R}_t(\infty|y) \right\}^{2/3} n^{-2/3}$$

is scale free.

We can think of this being a specific example of a more general kernel distribution function estimator than the one considered in Section 3.1, given by

$$\widetilde{F}(\alpha|y) = \frac{1}{\widehat{c}} \frac{1}{n} \sum_{i=1}^{n} f(y|\alpha^{(i)}) G\left(\frac{\alpha - \mu_\alpha - \beta\left(\alpha^{(i)} - \mu_\alpha\right)}{h} \right),$$

where we now have the additional degree of freedom β. Instead of optimising over β we use it to impose the constraint that the jittered particles have the same variance as the weighted pre-resampling particles. Reverting back to the multivariate case, we advocate setting the optimal $\widehat{h}$ and $\widehat{\beta}$ to

$$\widehat{h}_{j,t} = 1.59 \left\{ \widehat{R}_t(\infty|y) \right\}^{1/3} \widehat{\sigma}_{j,t} n^{-1/3}, \quad \widehat{\beta}_{j,t} = \sqrt{\frac{\widehat{\sigma}_{j,t}^2 - \widehat{h}_{j,t}^2}{\widehat{\sigma}_{j,t}^2}}, \quad \widehat{\sigma}_{j,t} = \widehat{IQR}_{j,t}/1.349.$$

This choice avoids an artificial increase in the variance and guarantees asymptotic consistency since the $\widehat{h}_{j,t}$ converges to zero at the correct rate. For simplicity we once again jitter and shrink the states $\alpha_1, \ldots, \alpha_p$ independently. This assumption will be shown not to be too strong in Section 4, where we provide some illustrative examples.

Example 4 (Example 1 continued) *Figure 3 illustrates the impact of using jittering with shrinkage on the particle degeneracy problem. In the middle graph we plot the* 5%, 50% *and* 95% *quantiles from* $N = 1,000$ *Monte Carlo replications estimating the posterior standard deviation* σ_t *with the*

smoothly jittered particle filter with shrinkage. This particle filter does not collapse at all, despite the large number of observations $t = 100$ relative to the number of particles $n = 100$. In the right graph we see that the inefficiency relative to i.i.d. sampling is growing significantly slower (solid line) than that of SIR (dotted line).

4 Illustrations

By means of a simple Monte Carlo experiment we first analyse the univariate performance of our algorithm on the challenging static case. Then we use a linear Gaussian state space model with dynamic and static states. We finish with a stochastic volatility example.

4.1 Univariate static parameter inference

We return to the setup in Example 1. For the simulation experiment we start by constructing our observations by setting $\alpha = 0.439$ throughout and generating t observations of y. Then we draw n particles from the prior and sequentially add observations one at a time. Based on the post-resampling approximation to $F_{\alpha_t|\mathcal{F}_t}$ we obtain the estimates[1]

$$\widehat{\mu}_t = \frac{1}{n}\sum_{i=1}^{n} \alpha_t^{(i)}, \quad \widehat{\sigma}_t = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\alpha_t^{(i)} - \widehat{\mu}_t\right)^2}, \quad \widehat{q}_{t,5\%} = \overline{F}^{-1}_{\alpha_t|\mathcal{F}_t}(0.05), \quad \widehat{q}_{t,95\%} = \overline{F}^{-1}_{\alpha_t|\mathcal{F}_t}(0.95)$$

for

$$\mu_t = \mathrm{E}\left[\alpha_t|\mathcal{F}_t\right], \quad \sigma_t = \sqrt{\mathrm{Var}(\alpha_t|\mathcal{F}_t)}, \quad q_{t,5\%} = F^{-1}_{\alpha_t|\mathcal{F}_t}(0.05), \quad q_{t,95\%} = F^{-1}_{\alpha_t|\mathcal{F}_t}(0.95).$$

Here $\overline{F}_{\alpha_t|\mathcal{F}_t}(\alpha) = \frac{1}{n}\sum_{i=1}^{n} I(\alpha_t^{(i)} \leq \alpha)$. We assess the accuracy of the resampling methods by looking at the distribution of the scaled error for each statistic

$$\sqrt{n}\left(\widehat{\mu}_t - \mu_t\right), \quad \sqrt{n}\left(\widehat{\sigma}_t - \sigma_t\right), \quad \sqrt{n}\left(\widehat{q}_{t,5\%} - q_{t,5\%}\right), \quad \sqrt{n}\left(\widehat{q}_{t,95\%} - q_{t,95\%}\right), \tag{6}$$

where the true values for all our statistics are obtained through (3). The experiment is repeated $N = 1,000$ times, drawing different $y = (y_1, \ldots, y_t)$ each time. We summarise the errors by computing the square root of the average square of the scaled errors (6), e.g. for the posterior mean we would compute

$$\sqrt{n}rmse\left(\widehat{\mu}\right) = \sqrt{n}\sqrt{\frac{1}{N}\sum_{l=1}^{N}\left(\mu_t^{(l)} - \widehat{\mu}_t^{(l)}\right)^2},$$

where $l = 1, \ldots, N$ indexes the Monte Carlo iterations of our simulation experiment.

We will employ the posterior resampling methods based on

[1] We are aware that it is suboptimal to use the post-resampling approximation to obtain estimators. Here our goal is to compare the resampling algorithms and hence the interest lies in post-resampling approximations.

1. $h_t = 0$, which is the same as standard SIR;

2. $\widehat{h}_t = 1.59\ \widehat{\sigma}_t\ \text{ESS}_t^{-1/3}$, the optimal jittering approach;

3. $\widehat{h}_t$ and $\widehat{\beta}_t = \sqrt{\frac{\widehat{\sigma}^2_{\alpha_t} - \widehat{h}^2_t}{\widehat{\sigma}^2_{\alpha_t}}}$, optimal jittering with shrinkage, our preferred approach;

4. $h_{d,t}$, kernel density estimator $h_d = 1.06\widehat{\sigma}_t n^{-1/5}$ approach.

Throughout we take $\widehat{\sigma}_t = \widehat{IQR}_t/1.349$, where $\widehat{IQR}_t$ is computed from $F_n(\alpha_t|\mathcal{F}_t)$. When we rerun the above experiments for different posterior resampling methods we always use the same random numbers across methods and simply adjust h_t.

The scaled root mse ($\sqrt{n}rmse$) results are given in Table 2 for $t = 100$ and increasing n. It indicates the h_d based jittering has problems as the bias dominates. $\widehat{h}$ performs significantly better than h_d and approaches the performance of SIR from above as n grows. $\widehat{h}$ with shrinkage performs best. The error in h_d remains significantly larger, suggesting serious bias. In Figure 4 we compare

t=100	$n = 100$, $\sqrt{n}rmse$				$n = 1,000$, $\sqrt{n}rmse$				$n = 10,000$, $\sqrt{n}rmse$			
	$\widehat{\mu}_t$	$\widehat{\sigma}_t$	$\widehat{q}_{t,5\%}$	$\widehat{q}_{t,95\%}$	$\widehat{\mu}_t$	$\widehat{\sigma}_t$	$\widehat{q}_{t,5\%}$	$\widehat{q}_{t,95\%}$	$\widehat{\mu}_t$	$\widehat{\sigma}_t$	$\widehat{q}_{t,5\%}$	$\widehat{q}_{t,95\%}$
$h = 0$	1.62	0.83	2.10	2.22	1.42	0.84	2.24	2.33	1.49	0.86	2.42	2.61
$\widehat{h}$	2.23	2.25	4.49	4.27	2.20	2.04	3.97	4.10	1.95	1.52	3.16	3.35
$\widehat{h}$, $\widehat{\beta}$	1.12	0.52	1.42	1.42	1.10	0.53	1.40	1.46	1.22	0.67	1.75	1.76
h_d	2.70	2.96	5.75	5.49	4.75	5.19	9.57	9.99	7.69	7.33	13.82	14.82

Table 2: $\sqrt{n}rmse$ for different bandwidths and numbers of particles and $t = 100$ observations for the univariate static model.

the changes in the $\sqrt{n}rmse$ for $t \in \{1, 5, 25, 100, 500, 1000\}$ and $n = 100$ fixed. Overall h_d does worst and the $\sqrt{n}rmse$ increases with t more than for any other bandwidth selection. We see that $\widehat{h}$ still seems to have some problems, but much less so than h_d. We prefer $\widehat{h}$ with shrinkage as it does best of all methods. At $t = 1000$ with $n = 100$, $\widehat{h}$ with shrinkage has a $\sqrt{n}rmse$ of 68%, 99.3%, 70% and 72% relative to that of SIR, for $\widehat{\mu}_t$, $\widehat{\sigma}_t$, $\widehat{q}_{t,5\%}$ and $\widehat{q}_{t,95\%}$ respectively. The value of the true $\sigma_{1000} = 0.03$ being so small leads to a $\sqrt{n}rmse$ for SIR almost as good as $\widehat{h}$ with shrinkage, since in all of the $N = 1,000$ simulation runs the SIR particle filter has collapsed by reaching $t = 1,000$.

The results above confirm two conclusions from the above discussions: the kernel based rule of $h \propto n^{-1/5}$ behaves very poorly due to bias. Our distribution based jittering is a significant improvement over the kernel density h_d, but we require shrinkage to obtain some improvements over SIR. We will see the size of the improvements will be more substantial on more sophisticated problems.

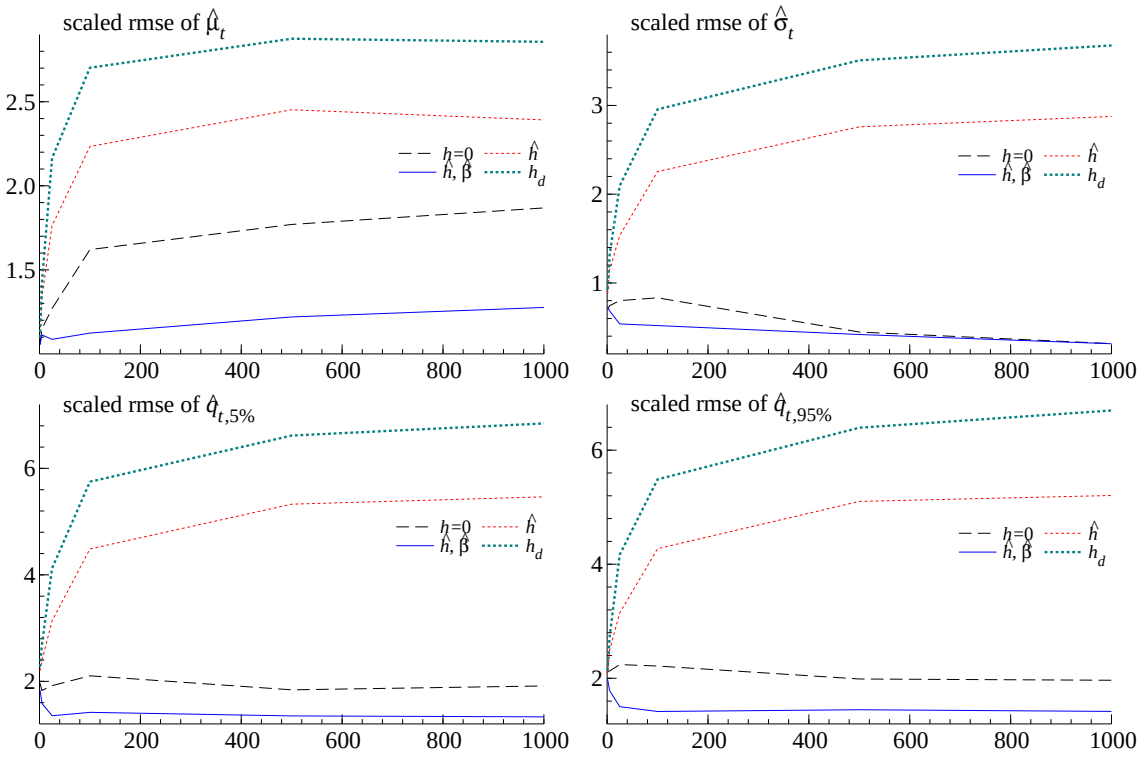

Figure 4: Univariate static model: $\sqrt{n}rmse$ for the posterior mean (top left), posterior standard deviation (top right), posterior 5% quantile (bottom left) and posterior 95% quantile. The lines are based on six points at $t = 1, 5, 25, 100, 500, 1000$. Throughout $n = 100$.

4.2 Including a dynamic state

We now consider a simple linear Gaussian state space model with unknown parameters. This enables us to benchmark the algorithm against results obtained with the Kalman filter. The model is

$$y_t = \lambda_t + \sigma_\varepsilon \varepsilon_t, \quad \varepsilon_t \overset{i.i.d.}{\sim} N(0,1), \quad t = 1, \ldots, T,$$
$$\lambda_t = \phi \lambda_{t-1} + \sigma_\eta \eta_t, \quad \eta_t \overset{i.i.d.}{\sim} N(0,1), \quad t = 1, \ldots, T, \quad \lambda_1 | \mathcal{F}_0 \sim N\left(0, \frac{\sigma_\eta^2}{1-\phi^2}\right),$$

where the unknown parameters are $\theta = \left(\sigma_\varepsilon^2, \phi, \sigma_\eta^2\right)'$. We assume priors $\sigma_\varepsilon^2 \sim$inv-$\Gamma\left(\gamma_\varepsilon, \beta_\varepsilon\right)$, $\phi \sim \mathcal{N}\left(\mu_\phi, \sigma_\phi^2\right)$ and $\sigma_\eta^2 \sim$inv-$\Gamma\left(\gamma_\eta, \beta_\eta\right)$.

The object of interest is the marginal posterior $f\left(\alpha_t | \mathcal{F}_t\right)$, where $\alpha_t = \left(\lambda_t, \theta'\right)'$. To compute the true posterior we draw from the prior $\theta^{(i)} \sim f(\theta)$, $i = 1, \ldots, M$, and evaluate

$$w_t^{(i)} = \prod_{l=1}^{t} N\left(y_l; a_{l|l-1}^{(i)}, P_{l|l-1}^{(i)} + \sigma_\varepsilon^{2(i)}\right), \quad W_t^{(i)} = \frac{w_t^{(i)}}{\sum_{j=1}^{M} w_t^{(j)}}, \quad i = 1, \ldots, M,$$

where $N(a; b, c)$ is the normal density with mean b, variance c, evaluated at a, and $a_{t|t-1}^{(i)}$ is the state prediction and $P_{t|t-1}^{(i)}$ the variance prediction, conditional on $\theta^{(i)}$, computed by the Kalman Filter.

The random measure $\left\{\theta^{(i)}, W_t^{(i)}\right\}_{i=1}^{M}$ represents $f(\theta|\mathcal{F}_t)$. For the posterior of the dynamic state we know that conditional on $\theta^{(i)}$ it is normal with mean $a_{t|t}^{(i)}$ and variance $P_{t|t}^{(i)}$. Our approximation to the true posterior will be normal with mean $\sum_{i=1}^{M} W_t^{(i)} a_{t|t}^{(i)}$ and variance $\sum_{i=1}^{M} W_t^{(i)} P_{t|t}^{(i)}$.

We perform a Monte Carlo experiment similar to the one in the previous section. We generate observations from the model with $\sigma_\varepsilon = 1$, $\phi = 0.9$, $\sigma_\eta = 0.5$ and then run each particle filter with the same random numbers on that data set. We repeat this N times, creating a new y and using different random numbers in each Monte Carlo iteration. The prior parameters are $\{\gamma_\varepsilon, \beta_\varepsilon, \mu_\phi, \sigma_\phi, \gamma_\eta, \beta_\eta\} = \{2.125, 0.9, 0.5, 0.2, 2.125, 0.9\}$. The focus lies now on comparing the best alternative to SIR, so we drop h_d and $\widehat{h}$ and maintain the bandwidth selection criteria

1. $h_{j,t} = 0$, which is the same as standard SIR;
2. $\widehat{h}_{j,t} = 1.59\, \widehat{\sigma}_{j,t}\, \mathrm{ESS}_t^{-1/3}$ and $\widehat{\beta}_{j,t} = \sqrt{\frac{\widehat{\sigma}_{j,t}^2 - \widehat{h}_{j,t}^2}{\widehat{\sigma}_{j,t}^2}}$, optimal jittering with shrinkage, our preferred approach.

Here $\widehat{\sigma}_{j,t}$ is a robust estimate of the standard deviation of $\alpha_{j,t}|\mathcal{F}_t$. For the practical implementation we jitter $\log \sigma_\varepsilon^{2(i)}$ and $\log \sigma_\eta^{2(i)}$.

Table 3 shows the $\sqrt{n}rmse$ from $N = 1,000$ Monte Carlo iterations with $t = 200$ observations and $n = 1,000$ particles for the two bandwidths considered. To compute the true Kalman-based posterior we use $M = 100,000$ draws from the prior.

$\widehat{h}$ with shrinkage does better than SIR for all statistics in all states. The best relative gain is found for $\widehat{\sigma}_t$ of ϕ where $\widehat{h}$ with shrinkage has a $\sqrt{n}rmse$ of 56% relative to that of SIR. The least relative gain is 84% for $\widehat{\mu}_t$ of σ_η^2.

t=200	α_t				σ_ϵ^2				ϕ				σ_η^2			
	$\widehat{\mu}_t$	$\widehat{\sigma}_t$	$\widehat{q}_{5\%}$	$\widehat{q}_{95\%}$	$\widehat{\mu}_t$	$\widehat{\sigma}_t$	$\widehat{q}_{5\%}$	$\widehat{q}_{95\%}$	$\widehat{\mu}_t$	$\widehat{\sigma}_t$	$\widehat{q}_{5\%}$	$\widehat{q}_{95\%}$	$\widehat{\mu}_t$	$\widehat{\sigma}_t$	$\widehat{q}_{5\%}$	$\widehat{q}_{95\%}$
$h = 0$	7.6	2.1	8.4	8.4	9.0	4.5	9.5	13.5	5.2	1.9	4.5	7.0	11.3	4.7	14.2	12.0
$\widehat{h}$, $\widehat{\beta}$	5.6	1.4	6.2	6.0	6.4	2.6	5.5	9.1	3.6	1.0	3.1	4.3	9.5	3.0	10.2	10.1

Table 3: $\sqrt{n}rmse$ for $n = 1,000$ particles and $t = 200$ for the linear Gaussian model.

In Figures 5 and 6 we compare the changes in the $\sqrt{n}rmse$ for the parameters ϕ and σ_η^2 respectively as we increase the number of observations $t \in \{1, 5, 25, 100, 200, 300, 400, 500\}$ and keep $n = 5,000$ fixed. We used $N = 400$ Monte Carlo iterations only, since the computational costs of computing the true posterior become cumbersome for $t > 200$.[2] For $t \leq 200$ we used $M = 100,000$ draws from the prior to compute the true Kalman-based posterior, for $t = 300$ we used $M = 200,000$, for $t = 400$ and $t = 500$ $M = 400,000$ draws from the prior. We find that

[2]From comparing the results for $N = 400$ with these for $N = 300$ we noticed that the ranking of the algorithms did not change and the plots do not look too differently, so we are confident that $N = 400$ is sufficient to have representative results.

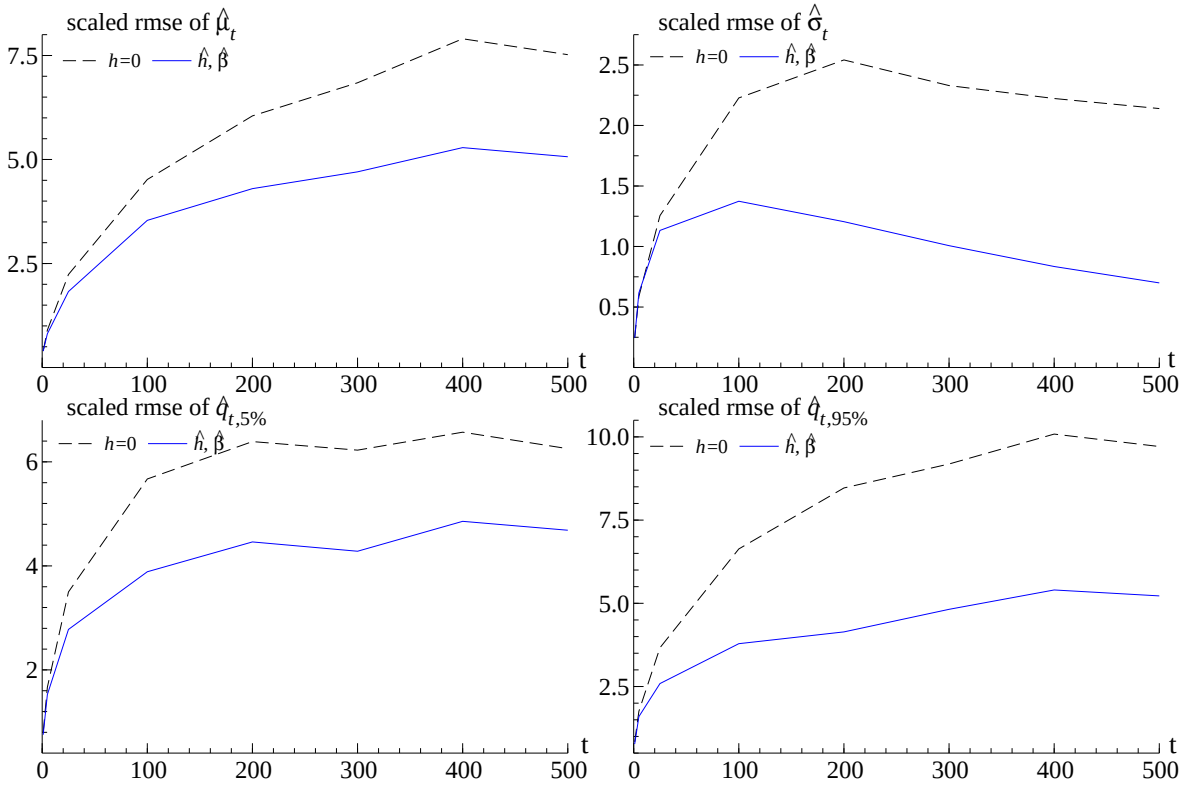

Figure 5: Linear Gaussian model, autoregressive parameter ϕ; $\sqrt{n}rmse$ for the posterior mean (top left), posterior standard deviation (top right), posterior 5% quantile (bottom left) and posterior 95% quantile. The lines are based on eight points at $t = 1, 5, 25, 100, 200, 300, 400, 500$. Throughout $n = 5,000$.

overall $\widehat{h}$ with shrinkage does best, in that it always outperforms SIR. Not depicted here for brevity are the same plots for the dynamic state α_t and σ_ε^2. It is worthwhile mentioning that for the former the performance of both bandwidths lies closer together for all t and for both $\widehat{h}$ with shrinkage consistently outperforms SIR.

In general we notice that with $n = 5,000$ particles and a small number of observations the difference in $\sqrt{n}rmse$ is negligible. As t increases, $\sqrt{n}rmse$ increases and the performance of the algorithms appears to diverge. The true power of $\widehat{h}$ with shrinkage starts to show as t increases. Comparing the $\sqrt{n}rmse$ across all states we notice that estimating σ_ε^2 and σ_η^2 are the greatest challenge and estimating the autoregressive coefficient, which happens to have the tightest prior, appears easiest.

4.3 A discrete time Gaussian stochastic volatility model

We conclude the illustrations with an empirical application to estimate the Gaussian discrete time stochastic volatility (SV) model. See for example the reviews in Ghysels, Harvey, and Renault

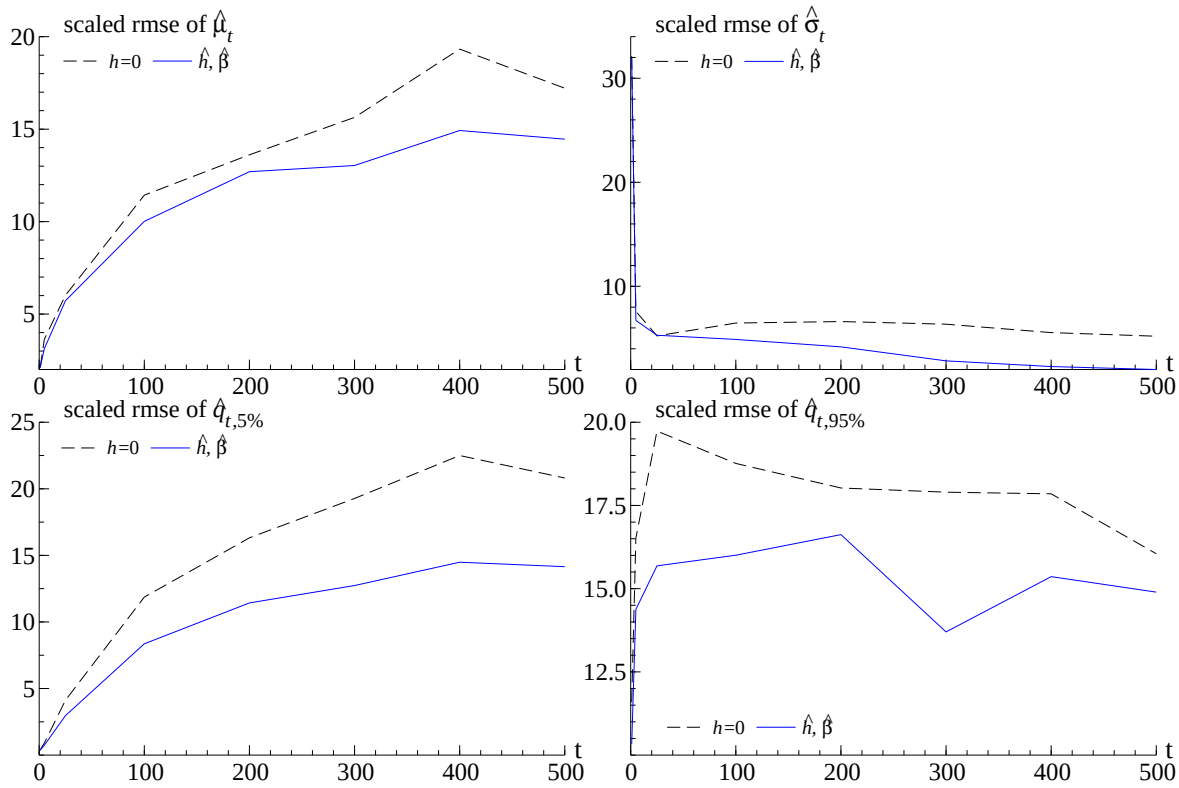

Figure 6: Linear Gaussian model, state variance parameter σ_η^2; $\sqrt{n}rmse$ for the posterior mean (top left), posterior standard deviation (top right), posterior 5% quantile (bottom left) and posterior 95% quantile. The lines are based on eight points at $t = 1, 5, 25, 100, 200, 300, 400, 500$. Throughout $n = 5,000$.

(1996), Shephard (2005) and also Kim, Shephard, and Chib (1998). The stock returns are assumed to follow the processes

$$\begin{array}{lcl} y_t & = & \mu + \exp\left\{\beta_0 + \beta_1 \alpha_t\right\} \varepsilon_t, \\ \alpha_{t+1} & = & \phi \alpha_t + \eta_t \end{array} \qquad \begin{pmatrix} \varepsilon_t \\ \eta_t \end{pmatrix} \overset{i.i.d.}{\sim} N\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right),$$

where $\alpha_0 \sim N\left(0, \left(1-\phi^2\right)^{-1}\right)$. The parameters of interest are $\theta = (\mu, \beta_0, \beta_1, \phi, \rho)'$ and the Gaussian priors are given in Table 4. We use observations of the end-of-day level of the S&P500

	μ	β_0	β_1	ϕ	ρ
prior mean	0.0	0.0	0.0	0.975	-0.6
prior stdev	0.1	0.3	0.3	0.020	0.3

Table 4: SV model: means and standard deviations for the Gaussian priors.

Composite Index (NYSE/AMEX only) from CRSP. The daily log-returns are defined as $y_t = 100\left(\log \text{S\&P500}_t - \log \text{S\&P500}_{t-1}\right)$.

To have a benchmark against which to compare the SIR and smoothly jittered particle filter we run the particle Markov chain Monte Carlo (PMCMC) algorithm from Andrieu, Doucet, and

Holenstein (2010) with $n = 2,000$ particles and $N = 100,000$ Markov chain iterations. We use the second half of the PMCMC draws from the posterior distribution to compute the "true" statistics used in the $\sqrt{n}rmse$.

In this experiment we keep the number of observations fixed and analyse the impact of changing the number of particles. We use $N = 1,000$ Monte Carlo repetitions, but this time we use different random numbers for the particle filters to avoid running out of memory. For the practical implementation we set the weight of any jittered particle to zero if $\rho^{(i)} \notin [-1, 1]$.

In Table 5 we report the usual statistics for the two algorithms when we use daily returns from 03.01.1995 until 31.12.2007, resulting in $t = 3271$. The smoothly jittered particle filter with

		α_t				μ				β_0			
		$\widehat{\mu}_t$	$\widehat{\sigma}_t$	$\widehat{q}_{5\%}$	$\widehat{q}_{95\%}$	$\widehat{\mu}_t$	$\widehat{\sigma}_t$	$\widehat{q}_{5\%}$	$\widehat{q}_{95\%}$	$\widehat{\mu}_t$	$\widehat{\sigma}_t$	$\widehat{q}_{5\%}$	$\widehat{q}_{95\%}$
n=1,000	SIR	133.0	22.3	126.9	149.7	2.0	0.4	2.5	1.7	11.4	2.3	9.5	14.1
	$\widehat{h}$, $\widehat{\beta}$	50.1	15.1	43.0	67.0	1.8	0.4	2.4	1.3	9.9	2.2	6.9	13.3
n=5,000	SIR	283.1	44.3	275.6	310.5	4.4	0.9	5.6	3.3	24.7	5.1	19.4	31.3
	$\widehat{h}$, $\widehat{\beta}$	90.9	28.5	81.4	119.2	3.3	0.5	4.1	2.6	14.7	3.7	10.7	19.6
n=10,000	SIR	366.8	57.7	374.5	385.5	5.5	1.3	7.3	4.1	31.0	7.2	24.2	40.1
	$\widehat{h}$, $\widehat{\beta}$	126.5	35.3	117.8	158.1	3.8	0.6	4.6	3.1	18.5	4.3	14.1	24.0

		β_1				ϕ				ρ			
		$\widehat{\mu}_t$	$\widehat{\sigma}_t$	$\widehat{q}_{5\%}$	$\widehat{q}_{95\%}$	$\widehat{\mu}_t$	$\widehat{\sigma}_t$	$\widehat{q}_{5\%}$	$\widehat{q}_{95\%}$	$\widehat{\mu}_t$	$\widehat{\sigma}_t$	$\widehat{q}_{5\%}$	$\widehat{q}_{95\%}$
n=1,000	SIR	5.0	0.2	5.0	4.9	1.7	0.1	1.5	1.8	12.4	1.4	14.2	10.5
	$\widehat{h}$, $\widehat{\beta}$	2.3	0.2	2.5	2.0	1.3	0.1	1.1	1.4	10.1	1.3	12.0	8.2
n=5,000	SIR	6.5	0.5	6.7	6.4	3.2	0.3	3.0	3.5	26.7	3.1	30.8	22.3
	$\widehat{h}$, $\widehat{\beta}$	3.1	0.3	3.4	2.8	1.7	0.1	1.7	1.7	11.3	1.6	12.8	9.8
n=10,000	SIR	7.0	0.7	7.5	6.7	4.0	0.4	3.7	4.3	36.1	4.4	42.0	29.8
	$\widehat{h}$, $\widehat{\beta}$	3.4	0.3	3.6	3.1	2.0	0.2	2.1	1.9	12.1	1.6	12.8	11.2

Table 5: Gaussian SV model: $\sqrt{n}rmse$ for SIR and $\widehat{h}$ with shrinkage; $n \in \{1000, 5000, 10000\}$; daily returns from the beginning of 1995 until the end of 2007, resulting in $t = 3271$.

shrinkage always outperforms SIR, with a $\sqrt{n}rmse$ relative to that of SIR of only 30% in the best case and 98% in the worst. We notice the rather large $\sqrt{n}rmse$ for the dynamic state, which can be explained by the fact that once the static states have collapsed to a single value, SIR basically runs a particle filter for the dynamic state alone, but potentially extremely poorly calibrated at whichever parameter values it collapsed onto. For $n = 1,000$ the difference in $\sqrt{n}rmse$ between the bandwidth choices is relatively small and starts to increase with n. At too small n both particle filters have a hard time, but as n starts taking reasonably large values the gains in performance are much larger for the smoothly jittered particle filter with shrinkage than for SIR. This results in the average relative $\sqrt{n}rmse$ across all statistics being 74% for $n = 1,000$, 53% for $n = 5,000$ and 49% for $n = 10,000$.

In Figure 7 we plot the marginal posterior distribution functions estimated from one run of the SIR (dashed line) and smoothly jittered (dotted line) particle filter on a reduced data set consisting

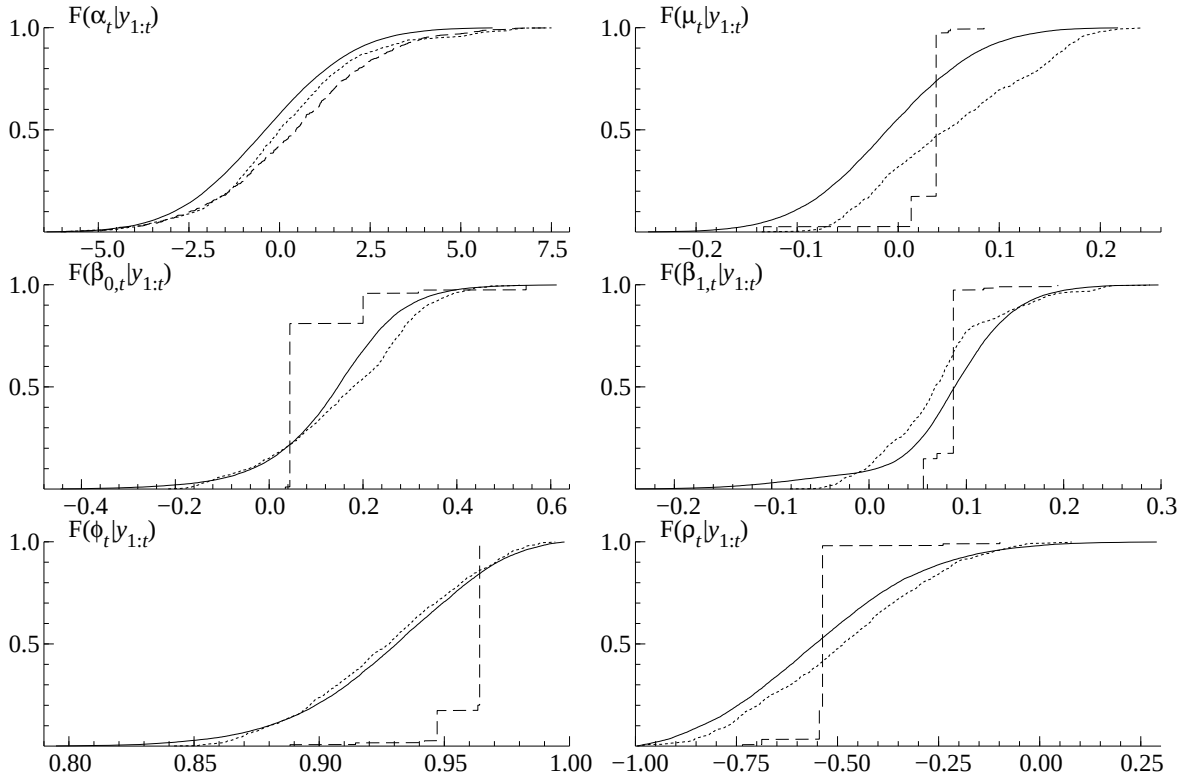

Figure 7: Gaussian SV model: estimated posterior distribution functions from one run of the SIR particle filter (dashed line) and one run of the smoothly jittered particle filter with shrinkage (dotted line) with $n = 1,000$ and $t = 100$. Solid line represents "truth" obtained from PMCMC.

of the last 100 daily observations in 2007. We also report the approximate "truth" from the PMCMC benchmark (solid line). We can see that even though we only have $t = 100$ observations and a fairly large number of particles $n = 1,000$, SIR has almost collapsed, whereas our smoothly jittered particle filter provides a smooth distribution function estimator.

This concludes the illustrations section. We have demonstrated the significantly superior performance of the smoothly jittered particle filter on a number of models. We hope to have convinced the reader that it is a good idea to jitter the resampled particles. The simplifying assumptions we made in the derivations of our preferred feasible $\widehat{h}$ with shrinkage do not appear to be too serious.

5 Further remarks

5.1 Improving accuracy by smoothing

So far our analysis is based on the quantiles. Another way of thinking about this type of problem is to use the sampling to approximate

$$B(F) = \int b(\alpha) \mathrm{d}F(\alpha|y) = \mathrm{E}_{\alpha|y}\left[b(\alpha)\right],$$

for some function b. We then replace F by $\widehat{F}$ and F_n and ask which delivers a better Monte Carlo estimate. For simplicity we will think about α as univariate. The following argument follows closely that of Silverman and Young (1987) on the smoothed bootstrap.

$$\begin{aligned} B(\widehat{F}) &= \int b(\alpha)\mathrm{d}\widehat{F}(\alpha|y) = \frac{1}{\widehat{c}nh}\sum_{i=1}^{n} f(y|\alpha^{(i)}) \int b(\alpha) g\left(\frac{\alpha-\alpha^{(i)}}{h}\right)\mathrm{d}\alpha \\ &= \frac{1}{\widehat{c}n}\sum_{i=1}^{n} f(y|\alpha^{(i)}) \int b\left(\alpha^{(i)}+h\epsilon\right) g(\epsilon)\mathrm{d}\epsilon. \end{aligned}$$

We will ignore the impact of estimating c. As

$$\int b\left(\alpha+h\epsilon\right) g(\epsilon)\mathrm{d}\epsilon = b(\alpha) + \frac{h^2}{2}b''(\alpha) + O(h^4),$$

we get

$$\begin{aligned} \mathrm{E}\left[B(\widehat{F}) - B(F_n)\right] &= \frac{h^2}{2}\int b''(\alpha) f\left(\alpha|y\right)\mathrm{d}\alpha + O(h^4), \\ \mathrm{Var}\left(B(\widehat{F})\right) - \mathrm{Var}\left(B(F_n)\right) &= \frac{h^2}{n}\int b(\alpha)b''(\alpha) r\left(\alpha|y\right)\mathrm{d}\alpha + O(h^4/n). \end{aligned}$$

This allows for reductions in mean square so long as $\int b(\alpha)b''(\alpha)r\left(\alpha|y\right)\mathrm{d}\alpha$ is negative. If this is the case then optimally $h \propto n^{-1/2}$.

If we used $h \propto n^{-1/5}$ then

$$\mathrm{mse}\left\{B(\widehat{F})\right\} \propto n^{-4/5},$$

which is worse than $\mathrm{mse}\left\{B(F_n)\right\}$ and is dominated by bias. Hence this kernel density estimation approach is problematic. Our paper has advocated using $h \propto n^{-1/3}$ in which case the first order terms of the mse are unaffected by the jittering as the squared bias is $O(n^{-4/3})$ and the change in the variance is $O(n^{-5/3})$, which are both of lower order than the usual $O(n^{-1})$ term. Hence distribution based jittering does no damage in this context.

5.2 Pitt's smoother

In the univariate state case Pitt (2002) suggested replacing $F_n(\alpha|y)$ by a smoothed version in order to remove the roughness of the resampling step, which causes jumps in the likelihood function.

His method starts by sorting the particles

$$\alpha^{[1]} < \alpha^{[2]} < ... < \alpha^{[n]}.$$

In a simple version of his methods, he then defines a "2nd nearest neighbour" type estimator

$$\widetilde{F}(\alpha|y) = \frac{1}{n\widehat{c}}\sum_{i=1}^{n-1}\left[w^{[i]} I\left(\alpha \leq \alpha^{[i-1]}\right)\right.$$

$$+ I\left(\alpha \in \left(\alpha^{[i+1]} - \alpha^{[i]}\right)\right)\left\{\frac{w^{[i]}}{2} + \frac{\alpha - \alpha^{[i]}}{\alpha^{[i+1]} - \alpha^{[i]}}\frac{w^{[i]} + w^{[i+1]}}{2}\right\}\Bigg],$$

where $w^{[i]} = f(y|\alpha^{[i]})$. This corresponds to adding uniform jitter on the interval $\left[\alpha^{[i]}, \alpha^{[i+1]}\right]$ to the modified cumulated weights. This finite nearest neighbour method would lead to an inconsistent density estimator as the neighbourhood is fixed at 2. However, we are interested in estimating the corresponding distribution function and quantile, which are consistently estimated here. Of course kernels and nearest neighbourhood methods are intimately related.

6 Conclusion

Particle filters are becoming the dominant way of carrying out on-line Bayesian inference for parametric state space models, extending the classic Kalman filter algorithm to the non-linear, non-Gaussian case. The particle degeneracy problem is extremely important and holds back the analysis of slowly varying or time invariant states.

This paper provides a solution to this problem using a particular type of jittering and shrinking. It is based on a higher order analysis of a smooth SIR algorithm, which suggests a dimensionless way of choosing the degree of jittering based upon distributional smoothing.

We explore the use of this method on a number of applied problems, showing this new procedure works in practice. It involves a trivial amount of extra computation and coding. The method is not damaging to particle filters of fast evolving states as it only affects the higher order properties of the sampling method. We show that the improvements over SIR persist as the number of particles and the number of observations increase.

A Appendix

A.1 Proof of Proposition 1

We rewrite

$$\begin{aligned}
F_n(\alpha|y) - F(\alpha|y) &= \frac{1}{\widehat{c}}\left(\frac{1}{n}\sum_{i=1}^{n} f\left(y|\alpha^{(i)}\right) I\left(\alpha^{(i)} \leq \alpha\right) - \widehat{c}F(\alpha|y)\right) \\
&= \frac{1}{\widehat{c}}\left(\frac{1}{n}\sum_{i=1}^{n} f\left(y|\alpha^{(i)}\right)\left\{I\left(\alpha^{(i)} \leq \alpha\right) - F(\alpha|y)\right\}\right) \\
&= \frac{1}{c}\left(\frac{1}{n}\sum_{i=1}^{n} f\left(y|\alpha^{(i)}\right)\left\{I\left(\alpha^{(i)} \leq \alpha\right) - F(\alpha|y)\right\}\right)\left(1 + O_p\left(n^{-1/2}\right)\right)
\end{aligned}$$

since $\widehat{c} \xrightarrow{p} f(y)$. We will ignore the $O_p\left(n^{-1/2}\right)$ term from now on.

$$\mathrm{E}[u_n] = \frac{\sqrt{n}}{c}\int f(y|x)\left\{I(x \leq \alpha) - F(\alpha|y)\right\} f(x)\,\mathrm{d}x$$

$$= \sqrt{n} \int \{I(x \leq \alpha) - F(\alpha|y)\} f(x|y) \, \mathrm{d}x = 0,$$

while

$$\begin{aligned} n\mathrm{Var}\left(F_n(\alpha|y) - F(\alpha|y)\right) &= \frac{1}{f(y)} \int f(y|x) \{I(x \leq \alpha) - F(\alpha|y)\}^2 f(x|y) \, \mathrm{d}x \\ &= \{1 - 2F(\alpha|y)\} R(\alpha|y) + F(\alpha|y)^2 R(\infty|y). \end{aligned}$$

A.2 Proof of Proposition 2

We again approximate with error $O_p(n^{-1/2})$

$$\widehat{F}(\alpha|y) - F(\alpha|y) = \frac{1}{c} \frac{1}{n} \sum_{i=1}^{n} f(y|\alpha^{(i)}) \left\{ G\left(\frac{\alpha - \alpha^{(i)}}{h} \right) - F(\alpha|y) \right\} \left(1 + O_p(n^{-1/2})\right),$$

where G is a probability distribution function. We again ignore the $O_p\left(n^{-1/2}\right)$ for the derivations that follow.

Proposition 3 *We will use the following result, which assumes F is twice continuously differentiable.*

$$\begin{aligned} I_1 &= \int_{-\infty}^{\infty} G\left(\frac{a-x}{h}\right) f(x) \mathrm{d}x = F(a) + \sum_{j=1}^{p} \frac{h_j^2}{2} \frac{\partial^2 F(a)}{\partial a_j^2} + O(\max_j h_j^4), \\ I_2 &= \int_{-\infty}^{\infty} G\left(\frac{a-x}{h}\right)^2 f(x) \mathrm{d}x = F(a) - 2 \sum_{j=1}^{p} h_j \frac{\partial F(a)}{\partial a_j} \int_{-\infty}^{\infty} t_j g(t) G(t) \, \mathrm{d}t + O(\max_j h_j^2). \end{aligned}$$

Proof. Let U be some absolutely continuous non-decreasing function, so we write it as

$$U(t_1, ..., t_p) = \int_{-\infty}^{t_1} ... \int_{-\infty}^{t_p} u(a_1, ..., a_p) \mathrm{d}a_1 ... \mathrm{d}a_p, \quad \frac{\partial^p U(t)}{\partial t_1 ..., \partial t_p} = u(t),$$

then Fubini's theorem implies that

$$\begin{aligned} & \int_{-\infty}^{\infty} U\left(\frac{x_1 - s_1}{h_1}, ..., \frac{x_p - s_p}{h_1} \right) \mathrm{d}F(s) \\ = & \int_{-\infty}^{\infty} u(t_1, ..., t_p) F(x_1 - t_1 h_1, ..., x_p - t_p h_1) \mathrm{d}t_1 ... \mathrm{d}t_p \end{aligned}$$

Using this and a third order Taylor approximation (as $h_1, ..., h_p \downarrow 0$) of $F(a - th)$ around $F(a)$ delivers

$$\begin{aligned} I_1 &= \int_{-\infty}^{\infty} G\left(\frac{a-x}{h}\right) f(x) \mathrm{d}x \\ &= \int_{-\infty}^{\infty} g(t_1) ... g(t_p) F(a_1 - t_1 h_1, ..., a_p - t_p h_p) \mathrm{d}t \\ &= \int_{-\infty}^{\infty} g(t) F(a) \mathrm{d}t - \sum_{j=1}^{p} h_j \frac{\partial F(a)}{\partial a_j} \int_{-\infty}^{\infty} t_j g(t) \, \mathrm{d}t \end{aligned}$$

$$+\sum_{j=1}^{p}\sum_{k=1}^{p}\frac{h_j h_k}{2}tr\left\{\frac{\partial^2 F(a)}{\partial a_j \partial a_k}\int_{-\infty}^{\infty}t_j t_k g\left(t\right)\mathrm{d}t\right\}+O(\max_j h_j^4)$$
$$=F(a)+\sum_{j=1}^{p}\frac{h_j^2}{2}\frac{\partial^2 F(a)}{\partial a_j^2}+O(\max_j h_j^4),$$

by symmetry of g about 0 and

$$\int_{-\infty}^{\infty}t_j t_k g\left(t\right)\mathrm{d}t=1_{j=k}.$$

Using the independence assumption we solve

$$\frac{\partial^p G\left(x\right)^2}{\partial x_1 ..., \partial x_p}=2^p\frac{\partial^p G\left(x\right)}{\partial x_1 ..., \partial x_p}G(x)=2^p g(x)G(x),$$

and the second integral expression therefore yields

$$I_2=\int_{-\infty}^{\infty}G\left(\frac{a-x}{h}\right)^2 f(x)\mathrm{d}x=\frac{2^p}{h^p}\int_{-\infty}^{\infty}g\left(\frac{a-x}{h}\right)G\left(\frac{a-x}{h}\right)F(x)\mathrm{d}x$$
$$=2^p\int_{-\infty}^{\infty}g\left(t\right)G\left(t\right)F(a-th)\mathrm{d}t$$
$$=2^p F(a)\int_{-\infty}^{\infty}g\left(t\right)G\left(t\right)\mathrm{d}t-2^p\sum_{j=1}^{p}h_j\frac{\partial F(a)}{\partial a_j}\int_{-\infty}^{\infty}t_j g\left(t\right)G\left(t\right)\mathrm{d}t+O(\max_j h_j^2)$$
$$=F(a)-2\sum_{j=1}^{p}h_j\frac{\partial F(a)}{\partial a_j}\int_{-\infty}^{\infty}t_j g_1\left(t_j\right)G_1\left(t_j\right)\mathrm{d}t_j+O(\max_j h_j^2),$$

where we only used a first order Taylor approximation.

□

We now tackle the properties of $\widehat{F}(\alpha|y)-F(\alpha|y)$.

For the expectation we obtain

$$\frac{1}{c}\int_{-\infty}^{\infty}f(y|x)G\left(\frac{\alpha-x}{h}\right)f(x)\mathrm{d}x=\int_{-\infty}^{\infty}G\left(\frac{\alpha-x}{h}\right)f(x|y)\mathrm{d}x$$
$$=F(\alpha|y)+\sum_{j=1}^{p}\frac{h_j^2}{2}\frac{\partial^2 F(a|y)}{\partial a_j^2}+O(\max_j h_j^4).$$

using I_1.

For the variance we recall the notation

$$\frac{f(y|x)^2 f(x)}{f(y)^2}=\frac{1}{f(y)}f(y|x)f(x|y)=r(x|y),\quad R(x|y)=\int_{-\infty}^{x}r(z|y)\mathrm{d}z$$

and solve

$$n\mathrm{Var}\big(\widehat{F}(\alpha|y)-F(\alpha|y)\big)$$
$$=\frac{1}{f(y)}\int_{-\infty}^{\infty}f(y|x)\left\{G\left(\frac{\alpha-x}{h}\right)-F(\alpha|y)\right\}^2 f(x|y)\mathrm{d}x+nO(\max_j h_j^4)$$

$$= \int_{-\infty}^{\infty} G\left(\frac{\alpha - x}{h}\right)^2 r(x|y)\mathrm{d}x - 2F(\alpha|y)\int_{-\infty}^{\infty} G\left(\frac{\alpha - x}{h}\right) r(x|y)\mathrm{d}x$$
$$+ F(\alpha|y)^2 \int_{-\infty}^{\infty} r(x|y)\mathrm{d}x + nO(\max_j h_j^4)$$

The first term yields

$$\int_{-\infty}^{\infty} G\left(\frac{\alpha - x}{h}\right)^2 r(x|y)\mathrm{d}x$$
$$= R(\alpha|y) - 2\sum h_j \frac{\partial R(\alpha|y)}{\partial a_j}\int_{-\infty}^{\infty} t_j g_1(t_j) G_1(t_j)\,\mathrm{d}t_j + O(\max_j h_j^2)$$

by I_2 and the second term

$$-2F(\alpha|y)\int_{-\infty}^{\infty} G\left(\frac{\alpha - x}{h}\right) r(x|y)\mathrm{d}x = -2F(\alpha|y)R(\alpha|y) - 2F(\alpha|y)O(\max_j h_j^2)$$

similar to I_1 (only using first order Taylor expansion). Combining the terms we get

$$\int_{-\infty}^{\infty} f(y|x)\left\{G\left(\frac{\alpha - x}{h}\right) - F(\alpha|y)\right\}^2 f(x|y)\mathrm{d}x$$
$$= \{1 - 2F(\alpha|y)\}R(\alpha|y) + F(\alpha|y)^2 R(\infty|y)$$
$$- 2\sum h_j \frac{\partial R(\alpha|y)}{\partial a_j}\int_{-\infty}^{\infty} t_j g_1(t_j) G_1(t_j)\,\mathrm{d}t_j + O(\max_j h_j^2).$$

Applying these results delivers the proof. □

References

Andrieu, C., A. Doucet, and R. Holenstein (2010). Particle Markov chain Monte Carlo methods (with discussion). *Journal of the Royal Statistical Society, Series B 72*, 1–33.

Azzalini, A. (1981). A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika 68*, 326–328.

Cappe, O., S. J. Godsill, and E. Moulines (2007). An overview of existing methods and recent advanaces in sequential Monte Carlo. *95*, 899–924. Proceedings of the IEEE.

Cheng, M.-Y. and S. Sun (2006). Bandwidth selection for kernel quantile estimation. *Journal of the Chinese Statistical Association 44*, 271–295.

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

Del Moral, P., A. Doucet, and S. S. Singh (2009). Forward smoothing using sequential Monte Carlo. Unpublished paper: Cambridge University Engineering Department.

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

Durbin, J. and S. J. Koopman (2001). *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.

Efron, B. (1982). *The Jackknife, the Bootstrap and other resampling plans*, Volume 38. SIAM. CBMS-IMF Region Conference Series in Applied Mathematics.

Fearnhead, P. (2002). MCMC, sufficient statistics and particle filter. *Journal of Computational and Graphical Statistics 11*, 848–862.

Ghysels, E., A. C. Harvey, and E. Renault (1996). Stochastic volatility. In C. R. Rao and G. S. Maddala (Eds.), *Statistical Methods in Finance*, pp. 119–191. Amsterdam: North-Holland.

Gordon, N. J., D. J. Salmond, and A. F. M. Smith (1993). A novel approach to nonlinear and non-Gaussian Bayesian state estimation. *IEE-Proceedings F 140*, 107–113.

Hansen, B. E. (2004). Nonparametric estimation of smooth conditional distributions. Unpublished paper: Department of Economics, University of Wisconsin.

Johannes, M. and N. Polson (2009). Particle filtering. In T. G. Andersen, R. A. Davis, J.-P. Kreiss, and T. Mikosch (Eds.), *Handbook of Financial Time Series*, pp. 1015–1029. Springer.

Johannes, M., N. Polson, and J. Stroud (2009). Optimal filtering of jump-diffusions: Extracting latent states from asset prices. *Review of Financial Studies 22*, 2259–2299.

Jones, M. C. (1990). The performance of kernel density functions in kernel distribution function estimation. *Statistics and Probability Letters 9*, 129–132.

Kim, S., N. Shephard, and S. Chib (1998). Stochastic volatility: likelihood inference and comparison with ARCH models. *Review of Economic Studies 65*, 361–393.

Liu, J. and M. West (2001). Combined parameter and state estimation in simulation-based filtering. In A. Doucet, N. de Freitas, and N. J. Gordon (Eds.), *Sequential Monte Carlo Methods in Practice*, pp. 197–223. Springer.

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. New York: Springer.

Musso, C., N. Oudjane, and F. LeGland (2001). Improving regularised particle filters. In A. Doucet, J. F. G. de Freitas, and N. J. Gordon (Eds.), *Sequential Monte Carlo Methods in Practice*, pp. 247–271. New York: Springer-Verlag.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and Its Applications 9*, 141–142.

Owen, A. (2001). *Empirical Likelihood*. London: Chapman and Hall.

Pitt, M. K. (2002). Smooth particle filters for likelihood maximisation. Unpublished paper: Department of Economics, Warwick University.

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

Poyiadjis, G., A. Doucet, and S. S. Singh (2009). Sequential Monte Carlo computation of the score and observed information matrix in state-space models with application to parameter estimation. *Biometrika*. forthcoming.

Rubin, D. B. (1987). A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when the fraction of missing information is modest: the SIR algorithm. Discussion of Tanner and Wong (1987). *Journal of the American Statistical Association 82*, 543–546.

Rubin, D. B. (1988). Using the SIR algorithm to simulate posterior distributions. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (Eds.), *Bayesian Statistics 3*, pp. 395–402. Oxford: Oxford University Press Press.

Shephard, N. (Ed.) (2005). *Stochastic Volatility: Selected Readings*. Oxford: Oxford University Press.

Silverman, B. W. and G. A. Young (1987). The bootstrap: to smooth or not to smooth. *Biometrika 74*, 469–479.

Storvik, G. (2002). Particle filters in state space models with the presence of unknown static parameters. *IEEE Transactions on Signal Processing, 50*, 281–289.

Stravropoulos, P. and M. Titterington (2001). Improved particle filters and smothing. In A. Doucet, J. F. G. de Freitas, and N. J. Gordon (Eds.), *Sequential Monte Carlo Methods in Practice*, pp. 465–477. Springer-Verlag: New York.

Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association 82*, 528–50.

West, M. (1993). Approximating posterior distributions by mixtures. *Journal of the Royal Statistical Society, Series B 55*, 409–42.