Chapter 6

Sequential Monte Carlo algorithms

So far we have been concerned with DLMs, dynamic linear models, however other models are possible and in fact most real systems are not linear and Gaussian. In some cases, a DLM is good-enough an approximation to be useful, especially given that it is closed form and easily computed. Yet, in many other cases, it is mandatory to introduce non-linear and non-Gaussian models in order to make useful forecasts. Recall that we initially introduced general evolution and observation equations of the form

$$\begin{aligned}\boldsymbol{\theta}_k &= f_k(\boldsymbol{\theta}_{k-1}, \boldsymbol{u}_k)\\ \boldsymbol{y}_k &= h_k(\boldsymbol{\theta}_k, \boldsymbol{v}_k),\end{aligned}$$

for a sequence of random variables $(\boldsymbol{\theta}_k)_{k\geq 0}$ with the Markov property.

Example 6.1. The stochastic volatility model is a very common non-linear model used in mathematical finance. If $\boldsymbol{\theta}_k$ is the log-volatility of the return of a financial asset at time k then

$$\begin{aligned}\boldsymbol{\theta}_0 &\sim \mathrm{N}\Big(\cdot\,; \mu, \frac{1}{1-\phi_1^2}\Big)\\ \boldsymbol{\theta}_k &= \mu + \phi_1(\boldsymbol{\theta}_{k-1} - \mu) + \boldsymbol{u}_k, \qquad k \geq 1\\ \boldsymbol{y}_k &= \boldsymbol{v}_k \exp\Big(\frac{1}{2}\boldsymbol{\theta}_k\Big), \qquad k \geq 0\end{aligned}$$

with $\boldsymbol{u}_k \sim \mathrm{N}(\cdot\,; 0, \beta^2)$ and $\boldsymbol{v}_k \sim \mathrm{N}(\cdot\,; 0, 1)$ and with μ, ϕ_1 and $\beta > 0$ some given scalar parameters. The state equation is simply an AR(1) process but the relation between the state and the observation is clearly non-linear. In particular, it is the variance of the observation noise that is related to the state. The reason for considering the log-volatility $\boldsymbol{\theta}_k$ instead of the volatility itself, say $\boldsymbol{\xi}_k$, is to ensure that the posterior distribution does not suggest non-positive volatilities (which do not make sense). Indeed, even if the posterior distribution of $\boldsymbol{\theta}_k$ is supported by the whole real line, the corresponding distribution of $\boldsymbol{\xi}_k = \exp(\boldsymbol{\theta}_k)$ will be supported by $(0, \infty)$.

For simplicity, we can consider that the noise sequences $(\boldsymbol{u}_k)_{k\geq 0}$ and $(\boldsymbol{v}_k)_{k\geq 0}$ are additive in the model, that is

$$\begin{aligned}\boldsymbol{\theta}_k &= \tilde{f}_k(\boldsymbol{\theta}_{k-1}) + \boldsymbol{u}_k\\ \boldsymbol{y}_k &= \tilde{h}_k(\boldsymbol{\theta}_k) + \boldsymbol{v}_k,\end{aligned}$$

for some arbitrary functions $\tilde{f}_k$ and $\tilde{h}_k$ which are not linear in general. In this situation, even if $\boldsymbol{\theta}_{k-1}$, $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$ where Gaussian, the predicted and posterior distributions would be non-Gaussian in general. However, it is straightforward to determine the transition and likelihood functions: for instance, if $\boldsymbol{u}_k \sim \mathrm{N}(\cdot\,; 0, U_k)$ and $\boldsymbol{v}_k \sim \mathrm{N}(\cdot\,; 0, V_k)$, then

$$q_k(\theta \,|\, \theta') = \mathrm{N}(\theta; \tilde{f}_k(\theta'), U_k) \qquad \text{and} \qquad \ell_k(y_k \,|\, \theta) = \mathrm{N}(y_k; \tilde{h}_k(\theta), V_k).$$

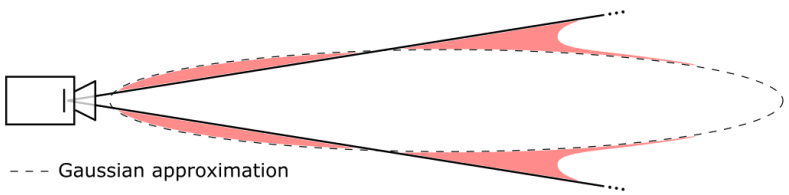

Figure 6.1: Approximation of the uncertainty associated with the observation from a camera by a Gaussian in the 2-dimensional Cartesian plane.

Example 6.2. If $\boldsymbol{\theta}_k = (\boldsymbol{x}_1, \boldsymbol{x}_2)^\intercal$ is the position of an object in the 2-dimensional Cartesian plane and if this object is observed by a camera placed at the origin, then the observation function is

$$\tilde{h}_k(x_1, x_2) = \arctan \frac{x_2}{x_1}$$

which is obviously non-linear.

There exist approximations that allow for considering non-linear evolution and observation functions while representing the predicted and posterior distributions by Gaussian distributions, but these approaches will only allow for tackling mildly non-linear functions and will not preserve the theoretical properties of the Kalman filter. The two most popular methods of this kind are the extended Kalman filter (EKF) which relies on a linearisation of the evolution and observation functions and the unscented Kalman filter (UKF) which relies on the empirical covariance computed from a deterministic set of points. For instance, in the scenario considered in Example 6.2, linearising the observation function will not be a good approximation in general, as illustrated in Figure 6.1.

Another large set of techniques is gathered under the name *Monte Carlo methods*. For a given target distribution $\pi(\cdot)$ on the space Θ, the simplest approach would be to introduce i.i.d. random variables $\boldsymbol{\theta}^{(i)} \sim \pi(\cdot)$, $i \in \{1, \dots, N\}$, in which case integrals with respect to $\pi(\cdot)$ of the form

$$I(\varphi) = \int \varphi(\theta)\pi(\theta)\,\mathrm{d}\theta$$

can be approximated, for any given integrable function $\varphi(\cdot)$, by

$$\hat{\boldsymbol{I}}(\varphi) = \frac{1}{N}\sum_{i=1}^{N} \varphi(\boldsymbol{\theta}^{(i)}).$$

For instance, if $\pi(\cdot)$ is the posterior distribution of the state $\boldsymbol{\theta}_k$ given the observations $y_0, \dots, y_k$ then $I(\varphi)$ is equal to the posterior mean $\mathbb{E}(\boldsymbol{\theta} \,|\, \boldsymbol{y}_{0:k})$ when $\varphi(\theta) = \theta$ and to the second posterior moment $\mathbb{E}(\boldsymbol{\theta}^2 \,|\, \boldsymbol{y}_{0:k})$ when $\varphi(\theta) = \theta^2$, from which the posterior variance can be recovered.

Note that $\hat{\boldsymbol{I}}(\varphi)$ is a random variable so that expectations can be taken:

$$\mathbb{E}(\hat{\boldsymbol{I}}(\varphi)) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}(\varphi(\boldsymbol{\theta}^{(i)})) = \frac{1}{N}\sum_{i=1}^{N} \int \varphi(\theta)\pi(\theta)\,\mathrm{d}\theta = I(\varphi)$$
$$\mathrm{var}(\hat{\boldsymbol{I}}(\varphi)) = \frac{1}{N}\left(\int \varphi(\theta)^2\pi(\theta)\,\mathrm{d}\theta - I(\varphi)^2 \right)$$

This is the basic Monte Carlo approach. However it is rarely the case that we can sample directly from the target distribution $\pi(\cdot)$, so that alternative methods have to be introduced.

6.1 Importance sampling

If it is not possible to sample directly from the target distribution $\pi(\cdot)$ but if we can evaluate $\pi(\cdot)$ at any point and if a proposal distribution $s(\cdot)$ is available such that it is simple to sample from $s(\cdot)$, then we can rewrite $I(\varphi)$ as

$$I(\varphi) = \int \varphi(\theta)\frac{\pi(\theta)}{s(\theta)}s(\theta)\,\mathrm{d}\theta.$$

This expression of $I(\varphi)$ suggest a different way of approximating this quantity, as described in Algorithm 1. In this method, called *importance sampling*, the weights $w(\boldsymbol{\theta}^{(i)})$, $i \in \{1, \dots, N\}$, are referred to as *importance weights*. The only assumption is that $\varphi(\theta)\pi(\theta) > 0$ implies $s(\theta) > 0$ for any $\theta \in \Theta$. This assumption has to be phrased more restrictively as "$\pi(\theta) > 0$ implies $s(\theta) > 0$" if we want the algorithm to be valid for any function φ.

Algorithm 1 Importance sampling

1: **for** $i = 1, \dots, N$ **do**
2: $\quad \boldsymbol{\theta}^{(i)} \sim s(\cdot)$
3: **end for**
4: *Output:*

$$\hat{\boldsymbol{I}}(\varphi) = \frac{1}{N} \sum_{i=1}^{N} \varphi(\boldsymbol{\theta}^{(i)}) w(\boldsymbol{\theta}^{(i)})$$

5: with $w(\theta) = \pi(\theta)/s(\theta)$

As in the basic Monte Carlo approach, we can compute the mean and variance of $\hat{\boldsymbol{I}}(\varphi)$ as

$$\mathbb{E}(\hat{\boldsymbol{I}}(\varphi)) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}(\varphi(\boldsymbol{\theta}^{(i)}) w(\boldsymbol{\theta}^{(i)})) = \frac{1}{N} \sum_{i=1}^{N} \int \varphi(\theta)\pi(\theta) \, \mathrm{d}\theta = I(\varphi)$$
$$\mathrm{var}(\hat{\boldsymbol{I}}(\varphi)) = \frac{1}{N} \left(\int \varphi(\theta)^2 \frac{\pi(\theta)^2}{s(\theta)} \, \mathrm{d}\theta - I(\varphi)^2 \right),$$

from which we conclude that the estimate $\hat{\boldsymbol{I}}(\varphi)$ is unbiased. There is however a clear limitation with importance sampling since one has to be able to evaluate the target density $\pi(\cdot)$ at any point $\theta \in \Theta$ in order to obtain the corresponding estimate $\hat{\boldsymbol{I}}(\varphi)$. In particular, expressing the target density as $\pi(\theta) = \gamma(\theta)/Z$ with γ a density function on Θ (not normalised) and Z a normalising constant, it is extremely common that the constant Z is unknown. In this case, one can instead use the *self-normalised* importance sampling described in Algorithm 2.

Algorithm 2 Self-normalised importance sampling

1: **for** $i = 1, \dots, N$ **do**
2: $\quad \boldsymbol{\theta}^{(i)} \sim s(\cdot)$
3: **end for**
4: *Output:*

$$\frac{1}{\sum_{j=1}^{N} \tilde{w}(\boldsymbol{\theta}^{(j)})} \sum_{i=1}^{N} \varphi(\boldsymbol{\theta}^{(i)}) \tilde{w}(\boldsymbol{\theta}^{(i)})$$

5: with $\tilde{w}(\theta) = \gamma(\theta)/s(\theta)$

There is indeed no need to know the value of Z in Algorithm 2 and we can even obtain an estimate of it as $\frac{1}{N} \sum_{i=1}^{N} \tilde{w}(\boldsymbol{\theta}^{(i)})$; in fact, Algorithm 2 can be seen as a version of Algorithm 1 where Z has been replaced by its estimate. However, the corresponding estimate of $I(\varphi)$ is now biased (but consistent).

6.2 Sequential importance sampling

Considering the case where the target distribution is the smoothing distribution at time n, that is

$$\pi_n(\theta_{0:n}) = \frac{\gamma_n(\theta_{0:n})}{Z_n}$$

with

$$\gamma_n(\theta_{0:n}) = p_0(\theta_0)\ell_0(y_0 \,|\, \theta_0) \prod_{k=1}^{n} q_k(\theta_k \,|\, \theta_{k-1})\ell_k(y_k \,|\, \theta_k)$$

and with Z_n the corresponding normalising constant. The self-normalising importance sampling algorithm can be used in principle with this type of target distribution, for instance with the proposal distribution

$$s_n(\theta_{0:n}) = p_0(\theta_0) \prod_{k=1}^{n} q_k(\theta_k \,|\, \theta_{k-1})$$

which is usually easy to sample from and which yields the importance weight

$$\tilde{w}_n(\theta_{0:n}) = \prod_{k=0}^{n} \ell_k(y_k \,|\, \theta_k).$$

If the weights are computed recursively in time, it is easy to see that

$$\tilde{w}_n(\theta_{0:n}) = \tilde{w}_{n-1}(\theta_{0:n-1}) \ell_n(y_n \,|\, \theta_n).$$

Other proposal distributions are possible, but, for the sake of simplicity, we will always assume that

$$s_n(\theta_{0:n}) = s_0(\theta_0) \prod_{k=1}^{n} s_k(\theta_k \,|\, \theta_{k-1}).$$

See Algorithm 3 for the corresponding calculations. However, simply following the evolution of the process without taking into account the observations is unlikely to yield samples with a non-negligible importance weight so that a large number of samples would be required, especially when n is large. This is particularly the case if the evolution model is uninformative (i.e. large evolution noise) and if the observations are very informative (i.e. small observation noise).

Algorithm 3 Sequential importance sampling

for $i = 1, \ldots, N$ **do**
  Sample $\boldsymbol{\theta}_0^{(i)} \sim s_0(\cdot)$
  Define the importance weight

$$\boldsymbol{w}_0^{(i)} = \frac{p_0(\boldsymbol{\theta}_0^{(i)}) \ell_0(y_0 \,|\, \boldsymbol{\theta}_0^{(i)})}{s_0(\boldsymbol{\theta}_0^{(i)})}$$

end for
for $k = 1, \ldots, n$ **do**
  for $i = 1, \ldots, N$ **do**
    Sample $\boldsymbol{\theta}_k^{(i)} \,|\, \boldsymbol{\theta}_{k-1}^{(i)} \sim s_k(\cdot \,|\, \boldsymbol{\theta}_{k-1}^{(i)})$
    Define the importance weight

$$\boldsymbol{w}_k^{(i)} = \boldsymbol{w}_{k-1}^{(i)} \frac{q_k(\boldsymbol{\theta}_k^{(i)} \,|\, \boldsymbol{\theta}_{k-1}^{(i)}) \ell_k(y_k \,|\, \boldsymbol{\theta}_k^{(i)})}{s_k(\boldsymbol{\theta}_k^{(i)} \,|\, \boldsymbol{\theta}_{k-1}^{(i)})}$$

  end for
end for
Output:

$$\hat{\boldsymbol{I}}_n(\varphi) = \frac{1}{\sum_{j=1}^{N} \boldsymbol{w}_n^{(j)}} \sum_{i=1}^{N} \boldsymbol{w}_n^{(i)} \varphi(\boldsymbol{\theta}_n^{(i)})$$

and $\hat{\boldsymbol{Z}}_n = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{w}_n^{(i)}$

6.3 Resampling

In order to address the limitations of importance sampling for state-space models, one of the most popular technique is to add a step in which the samples with high weight are duplicated and the ones with low

weights are deleted. This is called the *resampling* step, also known as the interaction or selection step. To be consistent with the literature, we will henceforth refer to the samples as *particles* and we will directly approximate the target distribution $\pi_n(\cdot)$ by an *empirical distribution*

$$\hat{\boldsymbol{\pi}}_n(\theta) = \frac{1}{\sum_{j=1}^N \boldsymbol{w}_n^{(j)}} \sum_{i=1}^N \boldsymbol{w}_n^{(i)} \delta_{\boldsymbol{\theta}_n^{(i)}}(\theta)$$

with $\boldsymbol{\theta}_n^{(i)}$ the particles at time step n and $\boldsymbol{w}_n^{(i)}$ their respective weights and with $\delta_\theta(\cdot)$ the Dirac function at point $\theta \in \Theta$. This is equivalent to estimating $I(\varphi)$ as before since

$$\int \varphi(\theta) \hat{\boldsymbol{\pi}}_n(\theta) \, \mathrm{d}\theta = \frac{1}{\sum_{j=1}^N \boldsymbol{w}_n^{(j)}} \sum_{i=1}^N \boldsymbol{w}_n^{(i)} \varphi(\boldsymbol{\theta}_n^{(i)}) = \hat{\boldsymbol{I}}(\varphi).$$

As mentioned before, the empirical distribution $\hat{\boldsymbol{\pi}}_n(\cdot)$ might be *degenerate*, that is the weights $\boldsymbol{w}_n^{(i)}$ might have negligible values, so that $\hat{\boldsymbol{I}}(\varphi)$ might be inaccurate. To address this issue, we consider a new collection of particles $\tilde{\boldsymbol{\theta}}^{(i)}$, $i \in \{1, \dots, N\}$, defined as duplicates of the existing particles and such that the empirical distribution

$$\tilde{\boldsymbol{\pi}}_n(\theta) = \frac{1}{N} \sum_{i=1}^N \delta_{\tilde{\boldsymbol{\theta}}_n^{(i)}}(\theta)$$

remains an estimate of $\pi_n(\cdot)$ with adequate properties. Two alternative ways to write this are

$$\hat{\boldsymbol{\pi}}_n(\theta) = \frac{1}{N} \sum_{i=1}^N \boldsymbol{o}_n^{(i)} \delta_{\boldsymbol{\theta}_n^{(i)}}(\theta) = \frac{1}{N} \sum_{i=1}^N \delta_{\boldsymbol{\theta}_n^{(\boldsymbol{a}_n^{(i)})}}(\theta)$$

where $\boldsymbol{o}_n^{(i)}$ and $\boldsymbol{a}_n^{(i)}$ are respectively the number of offsprings of particle $\boldsymbol{\theta}_n^{(i)}$ and the ancestor index for the new collection of particles. For the empirical distribution to have adequate properties, the resampling step should verify the following unbiasedness condition:

$$\mathbb{E}\big(\boldsymbol{o}^{(i)} \,|\, \boldsymbol{w}_n^{(1:N)} = w_n^{(1:N)}\big) = N \frac{w_n^{(i)}}{\sum_{j=1}^N w_n^{(j)}}.$$

A sufficient condition for this is

$$\mathbb{P}\big(\boldsymbol{a}_n^{(i)} = k \,|\, \boldsymbol{w}_n^{(1:N)} = w_n^{(1:N)}\big) = \frac{w_n^{(k)}}{\sum_{j=1}^N w_n^{(j)}}.$$

We can verify that expectations with respect to $\hat{\boldsymbol{\pi}}_n(\cdot)$ and $\tilde{\boldsymbol{\pi}}_n(\cdot)$ are identical under the unbiasedness condition. The corresponding method, referred to as the *bootstrap particle filter*, is detailed in Algorithm 4, where $\boldsymbol{a} \sim \mathrm{C}(\cdot; p_1, \dots, p_N)$ refers to the categorical distribution, that is such that $\mathbb{P}(\boldsymbol{a} = i) = p_i$.

Note that there are many other resampling procedures, which often aim at reducing the variance. It is also possible to sample from a proposal distribution instead of sampling from the transition density; this can be especially useful when the transition noise is large and/or the observation noise is small. For instance, the normal distribution given by a non-linear version of the Kalman filter can be used as a proposal distribution. An illustration of the bootstrap particle filter is given in Figure 6.2 on a linear-Gaussian model.

Algorithm 4 Bootstrap particle filter

for $i = 1, \dots, N$ **do**
  Sample $\boldsymbol{\theta}_0^{(i)} \sim p_0(\cdot)$
  Define the importance weight $\boldsymbol{w}_0^{(i)} = \ell_0(y_0 \,|\, \boldsymbol{\theta}_0^{(i)})$
end for
for $k = 1, \dots, n$ **do**
  for $i = 1, \dots, N$ **do**
    Sample the ancestor index

$$\boldsymbol{a}_{k-1}^{(i)} \,|\, \boldsymbol{w}_{k-1}^{(1:N)} \sim \mathrm{C}\left(\cdot; \frac{\boldsymbol{w}_n^{(1)}}{\sum_{j=1}^N \boldsymbol{w}_n^{(j)}}, \dots, \frac{\boldsymbol{w}_n^{(N)}}{\sum_{j=1}^N \boldsymbol{w}_n^{(j)}}\right)$$

    Sample $\boldsymbol{\theta}_k^{(i)} \,|\, \boldsymbol{\theta}_{k-1}^{(\boldsymbol{a}_{k-1}^{(i)})} \sim q_k\big(\,\cdot \,|\, \boldsymbol{\theta}_{k-1}^{(\boldsymbol{a}_{k-1}^{(i)})}\big)$
    Define the importance weight $\boldsymbol{w}_k^{(i)} = \ell_k(y_k \,|\, \boldsymbol{\theta}_k^{(i)})$
  end for
end for
Output:

$$\hat{\boldsymbol{\pi}}_n(\theta) = \frac{1}{\sum_{j=1}^N \boldsymbol{w}_n^{(j)}} \sum_{i=1}^N \boldsymbol{w}_n^{(i)} \varphi(\boldsymbol{\theta}_n^{(i)})$$

and $\hat{\boldsymbol{Z}}_n = \prod_{k=0}^N \left(\frac{1}{N} \sum_{i=1}^N \boldsymbol{w}_n^{(i)}\right)$

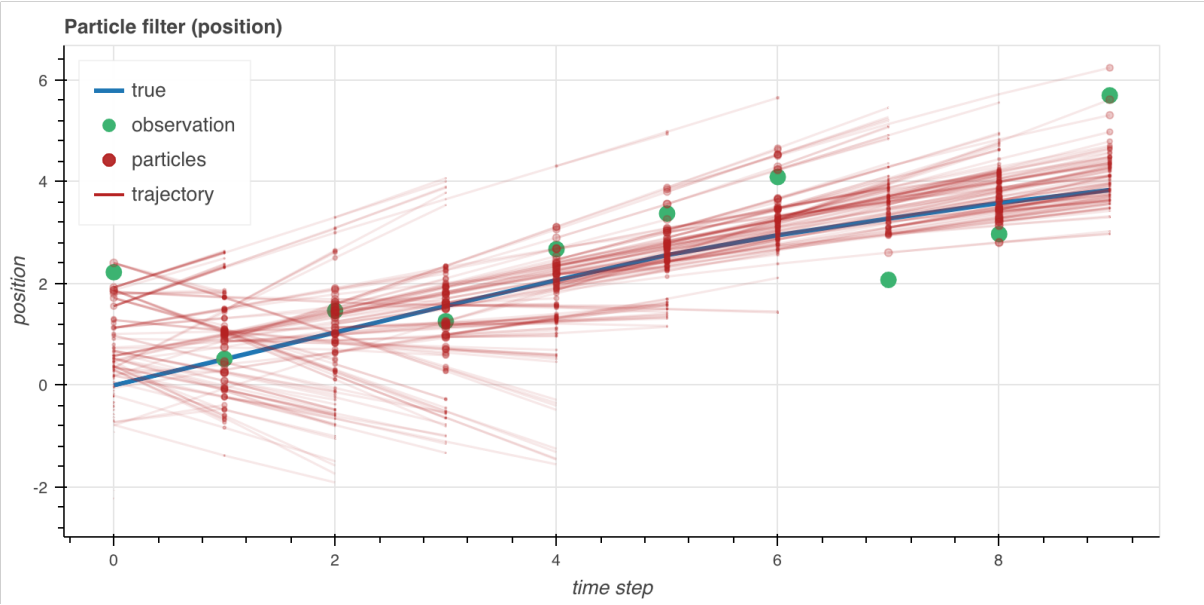

Figure 6.2: Particle filter on a linear-Gaussian model. The size of the red circles representing the particles is dependent on their weight.