Sequential Monte Carlo for state space models

Geir Storvik

October 18, 2017

This note discusses sequential Monte Carlo methods in somewhat more details than what is covered in the textbook (Givens and Hoeting, 2012). We will focus on inference with state space models and *online* applications where computation is performed/updated recursively as new observations arrive. Doucet et al. (2001) is a main reference in sequential Monte Carlo methods. More recent reviews are Doucet and Johansen (2009) and Creal (2012).

1 State space models

Consider the state space model

$$
\begin{aligned}
X_1 \sim& p(x_1;\theta) && && \text{(1a)}\\
X_t \sim& p(x_t|x_{t-1};\theta) && \text{State process} && \text{(1b)}\\
Y_t \sim& p(y_t|x_t;\theta) && \text{Observation process} && \text{(1c)}
\end{aligned}
$$

where $\{y_t\}$ are observed while $\{x_t\}$ are hidden. Note that we here deviate somewhat from the notation in the book in that we use $p(\cdot)$ generically for distributions. Further, both x_t and y_t (as well as θ) may be vectors. We will however save a boldface notation to *sequences* of variables $\mathbf{x}_{1:t} = (x_1, ..., x_t)$.

Our aim will be inference on $\{x_t\}$ and/or θ. The basic sequential relationships are then

$$p(x_t|\mathbf{y}_{1:t-1};\theta) = \int_{x_{t-1}} p(x_t|x_{t-1})p(x_{t-1}|\mathbf{y}_{1:t-1};\theta)dx_{t-1}; \tag{2}$$

$$p(x_t|\mathbf{y}_{1:t};\theta) = \frac{p(x_t|\mathbf{y}_{1:t-1};\theta)p(y_t|x_t;\theta)}{p(y_t|\mathbf{y}_{1:t-1};\theta)}. \tag{3}$$

2 Proper samples

We say that a weighted random pair (X, W) is *properly weighted with respect to* π if for any (square integrable) function h

$$E[Wh(X)] = c \cdot E_\pi[h(X)]$$

for some constant c. A weighted random sample $\{(X^i, W^i), i = 1, ..., N\}$ is properly weighted with respect to π if each (X_i, W_i) are properly weighted.

A consequence of this is that if $\{(X^i, W^i), i = 1, ..., N\}$ are iid random pairs, then

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} W^i h(X^i) \tag{4}$$

is a consistent estimator of $\mu = E_\pi[h(X)]$ (provided that $\mathsf{Var}[Wh(X)] < \infty$ where this variance is with respect to the distribution that (W, X) is generated from).

3 Sequential Monte Carlo for known θ

In this section we will, for simplicity, skip θ in the notation. Assume we want to simulate from $p(\mathbf{x}_{1:t}|\mathbf{y}_{1:t})$ where we can write

$$\begin{aligned} p(\mathbf{x}_{1:t}|\mathbf{y}_{1:t}) \propto & p(\mathbf{x}_{1:t}, \mathbf{y}_{1:t}) \\ = & p(x_1)p(y_1|x_1)\prod_{s=2}^{t} p(x_s|x_{s-1})p(y_s|x_s). \end{aligned}$$

Assume now a proposal distribution

$$q(\mathbf{x}_t) = q(x_1)\prod_{s=2}^{t} q(x_s|x_{s-1})$$

which allows for sequential simulation. Then, using the *importance sampling* idea, we obtain importance weights

$$\begin{aligned} w_t(\mathbf{x}_{1:t}) \propto & \frac{p(x_1)p(y_1|x_1)\prod_{s=2}^{t} p(x_s|x_{s-1})p(y_s|x_s)}{q(x_1)\prod_{s=2}^{t} q(x_s|x_{s-1})} \\ = & \frac{p(x_1)p(y_1|x_1)\prod_{s=2}^{t-1} p(x_s|x_{s-1})p(y_s|x_s)}{q(x_1)\prod_{s=2}^{t-1} q(x_s|x_{s-1})} \frac{p(x_t|x_{t-1})p(y_t|x_t)}{q(x_t|x_{t-1})} \\ \propto & w_{t-1}(\mathbf{x}_{1:t-1})\frac{p(x_t|x_{t-1})p(y_t|x_t)}{q(x_t|x_{t-1})} \end{aligned}$$

showing that also the importance weights can be easily updated sequentially. The missing proportionality constants makes no problem here since they will cancel out in the Monte Carlo estimates (4). In practice we usually normalize the weights to sum to one.

A particular simple choice of proposal distribution is

$$q(x_t|x_{t-1}) = p(x_t|x_{t-1})$$

in which case the updating equation for the importance weights reduces to

$$w_t(\mathbf{x}_{1:t}) \propto w_{t-1}(\mathbf{x}_{1:t-1})p(y_t|x_t).$$

This is called the *Bootstrap filter* (Gordon et al., 1993).

An alternative choice is

$$q(x_t|x_{t-1}) = p(x_t|x_{t-1}, y_t) = \frac{p(x_t|x_{t-1})p(y_t|x_t)}{p(y_t|x_{t-1})}.$$

In that case, the updating scheme becomes

$$w_t(\mathbf{x}_{1:t}) = w_{t-1}p(y_t|x_{t-1})$$

This choice can be seen as an optimal one (in that the variance of weights are minimized), but require both the possibility of simulating from $p(x_t|x_{t-1}, y_t)$ and evaluating $p(y_t|x_{t-1})$. How easy this is depends on the application. Pitt and Shephard (1999), through their *auxiliary particle filter* algorithm, propose the use of approximations of both these terms in order to obtain practical algorithms.

3.1 Weight degeneracy and the need for resampling

Based on the general rule:

$$\mathsf{var}[Y] = E[\mathsf{var}[Y|Z]] + \mathsf{var}[E[Y|Z]] \geq \mathsf{var}[E[Y|Z]]$$

we obtain, by choosing $Y = w_t, Z = \mathbf{X}_{1:t-1}$ (note that w_{t-1} is a deterministic function of $\mathbf{x}_{1:t-1}$) that

$$\begin{aligned} E_q[w_t|X_{1:t-1}] =& w_{t-1}E_q[\tfrac{p(X_t|X_{t-1})}{q(X_t|X_{t-1})}|X_{1:t-1}] \\ =& w_{t-1} \cdot 1 = w_{t-1} \end{aligned}$$

implying that

$$\mathsf{var}[w_t] \geq \mathsf{var}[w_{t-1}]$$

which indicates that the variance will increase at each time-step. The practical consequence of this is that only a few samples will dominate the others when time increases and thereby that the variability of the Monte Carlo estimate will increase. In order to measure the quality of Monte Carlo estimates based on weighted samples one typically use the *effective sample size.* Assume $w_i = w(\mathbf{X}_i), i = 1, ..., N$ are *normalized* weights. We then define the effective sample size by

$$\widehat{N}_{eff} = \frac{1}{\sum_{i=1}^{n} w_i^2}$$

There are some theoretical arguments behind this definition (see the textbook), but one motivation is that it gives reasonable measures in specific cases, e.g

$$\widehat{N}_{eff} = \begin{cases} N & \text{if } w_i = \frac{1}{N} \text{ for all } i; \\ N - z & \text{if } w_i = 0, i \leq z, \, w_i = \frac{1}{N-z}, i > z. \end{cases}$$

Problems occur when the effective sample size becomes too small. In that case, the variability in the associated Monte Carlo estimate will be large. A practical solution in that case is to perform *resampling.*

The simplest option of resampling is the following:

- Resample $\{\tilde{x}_t^1, ..., \tilde{x}_t^N\}$ from $\{x_t^1, ..., x_t^N\}$ with probabilities proportional to w_t^i.
- Put weights on the resampled values equal to $\tilde{w}_t^i = N^{-1}$.

(In the following we will be a bit sloppy in the notation in that we will use $\{x_t^1, ..., x_t^N\}$ also for the resampled sample.) Note that for the resampling scheme above, the number of repeats of variable x_t^i, N_t^i say, follows the Binomial(n, w_t^i) distribution and thereby that $E[N_t^i \tilde{w}_t^i] = w_t^i$.

Assuming $\{x_t^1, ..., x_t^N\}$ is a properly weighted sample with respec to $p(\mathbf{x}_{1:t}|\mathbf{y}_{1:t})$, one can show that a sufficient criterion for the resampled sample to be properly weighted as well is that $E[N_t^i \tilde{w}_t^i] = w_t^i$ for all i. There are many other resampling schemes that have this property. The *optimal* one (with respect to minimizing the extra variability introduced) is the following:

- For $i = 1, ..., N$, put ($\lfloor a \rfloor$ is the largest integer smaller than a)

  $$\widetilde{N}_t^i = \lfloor N w_t^i \rfloor \quad \text{(Some will be zero)}$$

- Let $\nu_t^i = w_t^i - \widetilde{N}_t^i / N$

- Define $K = N - \sum_{i=1}^N \widetilde{N}_t^i$ (remaining particles that have not been allocated).
- Sample $(D_t^1, ..., D_t^N)$ from the multinomial distribution with probabilities proportional to $(\nu^1, ..., \nu_t^n)$.
- Put $N_t^i = \tilde{N}_t^i + D_t^i$
- Make N_t^i replicates of x_t^i, put all weights equal to $1/N$.

Remarks Resampling will introduce extra random noise at the *current* time-point, but can reduce noise at *later* time points. Introducing resampling can be beneficial when only the marginal $p(x_t|\mathbf{y}_{1:t})$ is of interest (that is only the last point in the state process). In such cases it is possible (under some regularity assumptions) to show that the Monte Carlo errors are uniformly bounded in time (that it is does not increase with t). When considering $p(\mathbf{x}_{1:t}|\mathbf{y}_{1:t})$, resampling will result in that the values of x_s for s small will degenerate to only a few (in the end a single) unique points.

4 Sequential Monte Carlo for simultaneous parameter estimation

Consider now the case where θ is unknown. There are two possible approaches in this case. We will briefly discuss maximum likelihood estimation, while go more deeply into the Bayesian approach which turns out to be somewhat simpler in this setting. For a recent review on such approaches, see Kantas et al. (2015).

4.1 Online maximum likelihood estimation

In this case, one is interested in maximizing

$$L_t(\theta) = p(\mathbf{y}_{1:t}|\theta) = \int_{\mathbf{x}_t} p(\mathbf{y}_{1:t}|\mathbf{x}_{1:t};\theta)p(\mathbf{x}_{1:t}|\theta)dx_t.$$

A main problem in this case is to calculate the likelihood function (and possibly the score function in order to do optimization). The main approach in this setting is to use that

$$p(\mathbf{y}_{1:t}|\theta) = p(y_1|\theta)\prod_{s=1}^{t} p(y_s|y_{1:s-1};\theta)$$

and then utilize that

$$p(y_s|\mathbf{y}_{1:s-1}) = \int_{x_s} p(x_s|\mathbf{y}_{1:s-1})p(y_s|x_s;\theta)dx_s \approx \sum_{i=1}^{N} w_{t-1}^i p(y_s|x_s^i;\theta)$$

where now x_s^i is drawn from $p(x_s|x_{s-1}^i)$ (the proposal in the Bootstrap filter). Alternatives are possible if x_s^i is drawn from other proposal distributions.

In Poyiadjis et al. (2011) algorithms for calculating the score function and information (matrix) recursively is proposed making it possible for online maximum likelihood estimation.

4.2 Bayesian online parameter estimation

An alternative approach to maximum likelihood estimation is the Bayesian approach. In that case a prior $p(\theta)$, describing our prior knowledge about θ, is introduced into model (1). Our aim in this case can then be online inference on $p(x_t,\theta|\mathbf{y}_{1:t})$. One simple option in this case is to assume at time $t-1$ the existence of a properly weighted sample $\{(x_{t-1}^i,\theta^i,w_{t-1}^i)\}$ with respect to $p(x_{t-1},\theta|\mathbf{y}_{1:t-1})$. Then, by using that

$$\begin{aligned} p(x_t,\theta|\mathbf{y}_{1:t-1}) = &\int_{x_{t-1}} p(x_t|x_{t-1},\theta)p(x_{t-1},\theta|\mathbf{y}_{1:t-1})dx_{t-1} \\ \approx &\sum_{i=1}^{N} w_{t-1}^i p(x_t|x_{t-1}^i,\theta^i)\delta_\theta(\theta^i) \end{aligned}$$

and

$$p(x_t,\theta|\mathbf{y}_{1:t}) \approx c \cdot \sum_{i=1}^{N} w_{t-1}^i p(x_t|x_{t-1}^i,\theta^i)\delta_\theta(\theta^i)p(y_t|x_t,\theta^i)$$

we can obtain updated samples $\{(\theta^i,x_t^i,w_t^i)\}$ by simulating $x_t^i \sim p(x_t|x_{t-1}^i,\theta^i)$ and update the weights by $w_t^i \propto w_{t-1}^i p(y_t|x_t^i,\theta^i)$ (where the proportionality constant can be obtained by using normalized weights). An important feature of this approach is however that the sample $\{\theta^i\}$ is not changed over time. If resampling is introduced into the algorithm, this will lead to degeneracy of the unique values of θ, with similar problems as for samples of x_s for s small. Alternative approaches are therefore needed.

4.2.1 The Liu-West approach

One approach considered by Liu and West (2001) is to assume θ is (slowly) changing with time, introducing a model

$$\theta_t = \theta_{t-1} + \zeta_t, \quad \zeta_t \sim N(0, q)$$

and then rather focus on $p(x_t, \theta_t|\mathbf{y}_{1:t})$. In this case, assuming a weighted sample $\{(x_{t-1}^i, \theta_{t-1}^i, w_{t-1}^i)\}$ is available at time $t-1$, we can use that

$$\begin{aligned} p(x_t, \theta_t|\mathbf{y}_{1:t-1}) =& \int_{x_{t-1}} p(x_t|x_{t-1}, \theta_t) p(\theta_t|\theta_{t-1}) p(x_{t-1}, \theta_{t-1}|\mathbf{y}_{1:t-1}) dx_{t-1} d\theta_{t-1} \\ \approx& \sum_{i=1}^N w_{t-1}^i p(x_t|x_{t-1}^i, \theta_t) p(\theta_t|\theta_{t-1}^i) \end{aligned}$$

and

$$p(x_t, \theta_t|\mathbf{y}_{1:t}) \approx c \cdot \sum_{i=1}^N w_{t-1}^i p(x_t|x_{t-1}^i, \theta_t) p(\theta_t|\theta_{t-1}^i) p(y_t|x_t, \theta_t).$$

We can obtain updated samples $\{(\theta_t^i, x_t^i, w_t^i)\}$ by simulating $\theta_t^i \sim p(\theta_t|\theta_{t-1}^i)$, $x_t^i \sim p(x_t|x_{t-1}^i, \theta_t^i)$ and update the weights by $w_t^i \propto w_{t-1}^i p(y_t|x_t^i, \theta_t^i)$. Since new values $\{\theta_t^i\}$ are generated at each time point, the degeneracy problem will be reduced. A main problem is however that in this case we introduce extra variability in θ_t. The practical consequence of this is that estimation of θ_t is mainly based on the most recent observations, giving some efficiency loss.

4.2.2 The Fearnhead-Storvik approach

Consider now a more specific model where we assume the state process follows the linear Gaussian model

$$x_t = a x_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2) \tag{5}$$

in addition to that $x_1 \sim N(0, \sigma^2)$. The distribution $p(y_t|x_t)$ can be arbitrary. For simplicity, assume also σ^2 is known while $\theta = a$ needs to be estimated. We assume a prior

$$a \sim N(\mu_a, \sigma_a^2).$$

It can then be easily shown that

$$p(a|\mathbf{x}_{1:t}) = N(\mu_{a|t}, \sigma_{a|t}^2)$$

where

$$\begin{aligned}\mu_{a|t} =& \frac{\sigma_a^2\sum_{s=2}^t x_s x_{s-1}+\sigma^2\mu_a}{\sigma_a^2\sum_{s=2}^t x_{s-1}^2+\sigma^2};\\ \sigma_{a|t}^2 =& \frac{\sigma^2\sigma_a^2}{\sigma_a^2\sum_{s=2}^t x_{s-1}^2+\sigma^2}.\end{aligned}$$

A main point here is that given $\mathbf{x}_{1:t}$, the distribution of a (and simulation from this distribution) is easily obtained. Note further that $p(a|\mathbf{x}_{1:t})$ only depend on $S_{t,1} = \sum_{s=2}^t x_s x_{s-1}$ and $S_{t,2} = \sum_{s=2}^t x_{s-1}^2$ which both can be recursively updated through

$$S_{t,1} = S_{t-1,1} + x_t x_{t-1}, \quad S_{t,2} = S_{t-1,2} + x_{t-1}^2.$$

Model (5) is a special case of a class of models where the posterior distribution for the parameters involved only depends on a low-dimensional set of *sufficient* statistics that can be recursively updated. The approach by Fearnhead (2002) and Storvik (2002) utilizes this structure to construct a sequential Monte Carlo method that focus on simulating the sufficient statistics instead of the parameters. This approach, which will be described below, is sometimes (a bit unfair) called the *Storvik filter*.

Assume now all parameters are involved in the state process (so $p(y_t|x_t)$ do not involve any unknown parameters). Define S_t to be the sufficient statistics for θ given $\mathbf{x}_{1:t}$ (which might be a vector) with the property that $S_t = h(S_{t-1}, x_{t-1}, x_t)$ for some function $h(\cdot)$. The idea then is to perform simulation on $p(x_t, S_t|\mathbf{y}_{1:t})$ instead of $p(x_t, \theta|\mathbf{y}_{1:t})$. Assume a sample $\{(x_{t-1}^i, S_{t-1}^i, w_{t-1}^i), i = 1, ..., N\}$ which is properly weighted with respect to $p(x_{t-1}, S_{t-1}|\mathbf{y}_{1:t-1})$ is available at time $t-1$. Then, we can use similar recursions as before:

$$\begin{aligned}p(x_t, S_t|\mathbf{y}_{1:t-1}) =& \int_{x_{t-1}} p(x_t, S_t|x_{t-1}, S_{t-1})p(x_{t-1}, S_{t-1}|\mathbf{y}_{1:t-1})dx_{t-1}dS_{t-1}\\ \approx& \sum_{i=1}^N w_{t-1}^i p(x_t, S_t|x_{t-1}^i, S_{t-1}^i)\end{aligned}$$

and

$$p(x_t, S_t|\mathbf{y}_{1:t}) \approx c \cdot \sum_{i=1}^N w_{t-1}^i p(x_t, S_t|x_{t-1}^i, S_{t-1}^i)p(y_t|x_t).$$

Note that simulation from $p(x_t, S_t|x_{t-1}^i, S_{t-1}^i)$ can be performed through the following steps

1. Simulate $\theta^i \sim p(\theta | x_{t-1}^i, S_{t-1}^i) = p(\theta | S_{t-1}^i)$.
2. Simulate $x_t^i \sim p(x_t | x_{t-1}^i, \theta^i)$.
3. Put $S_t^i = h(S_{t-1}^i, x_{t-1}^i, x_t^i)$.

This then gives the following algorithm

Algorithm 1 SMC with parameter updating

```
Simulate $\theta^i \sim p(\theta)$ for $i = 1, ..., N$.                    ▷ Initialization
Simulate $x_1^i \sim p(x_1 | \theta^i)$ for $i = 1, ..., N$.
Put weights $w_1^i = p(y_1 | x_1^i, \theta^i)$.
Put $S_1^i = 0$ for $i = 1, ..., N$.
for $t = 2, 3, ...$ do                                           ▷ Sequential Monte Carlo
    Simulate $\theta^i \sim p(\theta | S_{t-1}^i)$ for $i = 1, ..., N$.
    Simulate $x_t^i \sim p(x_t | x_{t-1}^i, \theta^i)$ for $i = 1, ..., N$.
    Put weights $w_t^i = w_{t-1}^i p(y_t | x_t^i, \theta^i)$.
    Put $S_t^i = h(S_{t-1}^i, x_{t-1}^i, x_t^i)$.
    if $\hat{N}_{eff}$ is small then                                  ▷ Resampling
        Resample $(x_t^i, S_t^i)$ with probabilities proportional to $w_t^i$.
        Put $w_t^i = 1/N$.
    end if
end for
```

Note that the simulated values of θ are not used in the following time-steps, so they are discarded. This makes it possible for the parameter values to be changed at each iteration since the sufficient statistics will have the possibility for changing. However, at time t, (θ^i, w_{t-1}^i) can be seen as a properly weighted sample with respect to $p(\theta | \mathbf{y}_{1:t-1})$, making inference about θ possible as well.

Even though this algorithm solves some of the problems that the alternative algorithm have, it has been pointed out (e.g. Andrieu et al., 2005) that also this approach will, for large t suffer with respect to degeneracy. The reason for this is that the sufficient statistic S_t depend on the whole path $\mathbf{x}_{1:t}$ where the first components will have very few unique values after several rounds of resampling. Different approaches for improving on this has been suggested (e.g. Olsson et al., 2008; Carvalho et al., 2010). Still, however, a completely satisfactory solution is lacking!

References

C. Andrieu, A. Doucet, and V. B. Tadic. On-line parameter estimation in general state-space models. In *Decision and Control, 2005 and 2005 European Control Conference. CDC-ECC'05. 44th IEEE Conference on*, pages 332–337. IEEE, 2005.

C. M. Carvalho, M. S. Johannes, H. F. Lopes, N. G. Polson, et al. Particle learning and smoothing. *Statistical Science*, 25(1):88–106, 2010.

D. Creal. A survey of sequential Monte Carlo methods for economics and finance. *Econometric reviews*, 31(3):245–296, 2012.

A. Doucet and A. M. Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, 12(656-704):3, 2009.

A. Doucet, N. De Freitas, and N. Gordon. Sequential Monte Carlo Methods in Practice, 2001. Series Statistics For Engineering and Information Science.

P. Fearnhead. Markov chain Monte Carlo, sufficient statistics, and particle filters. *Journal of Computational and Graphical Statistics*, 11(4):848–862, 2002.

G. H. Givens and J. A. Hoeting. *Computational statistics*, volume 710. John Wiley & Sons, 2012.

N. J. Gordon, D. J. Salmond, and A. F. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE Proceedings F (Radar and Signal Processing)*, volume 140, pages 107–113. IET, 1993.

N. Kantas, A. Doucet, S. S. Singh, J. Maciejowski, N. Chopin, et al. On particle methods for parameter estimation in state-space models. *Statistical science*, 30(3):328–351, 2015.

J. Liu and M. West. Combined parameter and state estimation in simulation-based filtering. In *Sequential Monte Carlo methods in practice*, pages 197–223. Springer, 2001.

J. Olsson, O. Cappé, R. Douc, E. Moulines, et al. Sequential monte carlo smoothing with application to parameter estimation in nonlinear state space models. *Bernoulli*, 14(1):155–179, 2008.

M. K. Pitt and N. Shephard. Filtering via simulation: Auxiliary particle filters. *Journal of the American statistical association*, 94(446):590–599, 1999.

G. Poyiadjis, A. Doucet, and S. S. Singh. Particle approximations of the score and observed information matrix in state space models with application to parameter estimation. *Biometrika*, 98(1):65–80, 2011. doi: 10.1093/biomet/asq062. URL +http://dx.doi.org/10.1093/biomet/asq062.

G. Storvik. Particle filters for state-space models with the presence of unknown static parameters. *IEEE Transactions on signal Processing*, 50(2): 281–289, 2002.