1 Sequential Monte Carlo Methods.

Sampling from a sequence of distributions that change over time is difficult task in MCMC methology. This is, however, an important problem that arises in a range of applications. For instance, the observations may be arriving sequentially in time and one could be interested in performing Bayesian inference in real time. To take full advantage of data, one should update the posterior distribution as data become available. Some real-life applications include tracking of aircrafts using radar measurements, estimating the trends and volatility of financial measurements, etc. An additional benefit of sequential methods is their computational simplicity since the data are dealt in a sequential manner.

More details can be found in the monograph Doucet et al. (2001). Also, there is a page devoted to Sequential Monte Carlo Methods/ Particle Filtering at Cambridge, `http://www-sigproc.eng.cam.ac.uk/smc/`

1.1 Definitions

For any sequence $\{a_k\}$, let $a_{i:\,j} = (a_i, a_{i+1}, \ldots, a_j)$. Consider a sequence of probability distributions $\{\Pi_n, n = 1, 2, \ldots, \}$ where The distribution Π_n is defined on the space $E_n = E^n$. We will assume that each distribution Π_n admits a probability density $\pi_n(\theta_{1:\,n})$. Each density π_n is known up to a normalizing constant; i.e. $\pi_n(\theta_{1:n}) = \frac{f_n(\theta_{1:n})}{C_n}$ where f_n is known pointwise and C_n is unknown. The index n is often referred as the time index although in applications may not have connection with real time.

SMC methodology is a set of algorithms that generate at each time instance collection of N ($N \gg 1$) weighted random samples (particles) $\{\omega_n^{(i)}, \theta_{1:n}^{(i)}; i = 1, \ldots, N\}$ where $\omega_n^{(i)} > 0$, $\sum_{i=1}^N \omega_n^{(i)} = 1$ and such that for any test function $\varphi_n : E_n \to \mathbb{R}$

$$\sum_{n=1}^{N} \omega_n^{(i)} \varphi_n(\theta_{1:n}^{(i)}) \to \int \varphi_n(\theta_{1:n}) \pi_n(\theta_{1:n}) d\theta_{1:n}$$

as $N \to \infty$. Because of in-line applications, it is desirable that the algorithms have linear (in the number of particles N) computational complexity and that the complexity is independent of n.

Two fundamental actions in SMC are *sequential importance sampling* and *resampling.*

We briefly describe the sequential importance sampling technique first. Assume that at time $n-1$ particles $\{\theta_{1:n-1}^{(i)}\}$ have been sampled from an importance density $q_{n-1}(\theta_{1:n-1})$. Since the particles are not samples from the target density they are weighted. Their weights are given by

$$\omega_{n-1}^{(i)} \propto \frac{\pi_{n-1}(\theta_{1:n-1}^{(i)})}{q_{n-1}(\theta_{1:n-1}^{(i)})}.$$

At time n, one extends each path $\theta_{1:n-1}^{(i)}$ by sampling $\theta_n^{(i)}$ according to an importance density $\theta_n^{(i)} \sim q_n(\cdot|\theta_{1:n-1}^{(i)})$; the paths till time $n-1$ are not modified in order to keep the algorithm sequential. It follows that the joint importance density of the paths $\{\theta_{1:n}^{(i)}\}$ is of the form

$$\begin{aligned} q_n(\theta_{1:n}) &= q_n(\theta_n|\theta_{1:n-1}) q_{n-1}(\theta_{1:n-1}) \\ &= q_1(\theta_1) \prod_{k=2}^{n} q_k(\theta_k|\theta_{1:k-1}). \end{aligned}$$

To correct for the discrepancy between the new target density π_n and the importance density q_n, one needs to update the weights according to

$$\begin{aligned}
\omega_n^{(i)} &\propto \frac{\pi_n(\theta_{1:n}^{(i)})}{q_n(\theta_n^{(i)}|\,\theta_{1:n-1}^{(i)})\; q_{n-1}(\theta_{1:n-1}^{(i)})} \\
&\propto \frac{\pi_n(\theta_{1:n}^{(i)})}{\pi_{n-1}(\theta_{1:n-1}^{(i)})\; q_n(\theta_n^{(i)}|\,\theta_{1:n-1}^{(i)})} \times \frac{\pi_{n-1}(\theta_{1:n-1}^{(i)})}{q_{n-1}(\theta_{1:n-1}^{(i)})} \\
&\propto \frac{\pi_n(\theta_{1:n}^{(i)})}{\pi_{n-1}(\theta_{1:n-1}^{(i)})\; q_n(\theta_n^{(i)}|\,\theta_{1:n-1}^{(i)})} \times \omega_{n-1}^{(i)}.
\end{aligned}$$

In most applications, the computational complexity required to compute $\omega_n^{(i)}$ given $\omega_{n-1}^{(i)}$ is independent of n.

The efficiency of this method is highly dependent on the choice of the importance density. To minimize the conditional variance of the weights at time n, it is easy to see that the optimal importance distribution is given by

$$q_n(\theta_n|\theta_{1:n-1}) = \pi_n(\theta_n|\theta_{1:n-1}).$$

However, it might be impossible to sample easily from this density. Moreover, even if it is feasible, the incremental importance weight is given in this case by

$$\frac{\pi_n(\theta_{1:n})}{\pi_{n-1}(\theta_{1:n-1})q_n(\theta_n|\theta_{1:n-1})} = \frac{\pi_n(\theta_{1:n-1})}{\pi_{n-1}(\theta_{1:n-1})}$$

and might not admit an analytical expression as it requires to compute

$$\pi_n(\theta_{1:n-1}) = \int \pi_n(\theta_{1:n}) d\theta_n.$$

Therefore, a good alternative strategy consists of coming up with an approximation of the optimal importance sampling distribution; several approximation techniques have been presented in a nonlinear non-Gaussian state-space models context [2], [8].

Irrespective of the choice of the importance density, the main problem of sequential importance sampling is that it is just a special instance of importance sampling and degenerates when n increases. After only a few time steps, one weight approaches 1 whereas all the other weights are approach zero.

The key idea of SMC lies in the *Resampling step*. In the ideal scenario where $q_n(\theta_{1:n}) = \pi_n(\theta_{1:n})$, the weights would all be equal to N^{-1}. In practice this is obviously not the case and, roughly speaking, the approximation of π_n by $\{\omega_n^{(i)}, \theta_{1:n}^{(i)}\}$ is poor if the distribution of the weights has a high variance/small entropy. In this case, i.e. if the variance of the weights is too high/entropy of the weights is below a value specified by the user, particles with small weights are killed and particles with large weights are copied multiple times. The underlying idea is to focus the computational efforts on the promising zones of the space. Finally one assigns equal weights N^{-1} to each copy. The resampling step is what makes SMC work. Clearly it introduces locally in time additional Monte Carlo errors but it can be shown both practically and theoretically that this ensures that the algorithm does not "degenerate" over time. More formally, it consists of performing the following approximation

$$\sum_{i=1}^{N} \omega_n^{(i)} \delta_{\theta_{1:n}^{(i)}}(d\theta_{1:n}) \approx \sum_{i=1}^{N} \frac{N_n^{(i)}}{N} \delta_{\theta_{1:n}^{(i)}}(d\theta_{1:n})$$

where $N_n^{(i)} \in \mathbb{N}$ is the number of copies of the particles $\theta_{1:n}^{(i)}$ under the constraint $\sum_{i=1}^N N_n^{(i)} = N$ to keep the size of the population constant. In order to minimize the error introduced by the resampling scheme, one usually selects a stochastic mechanism to obtain $\{N_n^{(i)}\}$ such that $E[N_n^{(i)}] = N\omega_n^{(i)}$ (unbiased approximation) and with small variances $var[N_n^{(i)}]$. Several resampling schemes have been proposed in the literature including multinomial, residual and stratified resampling [3].

SMC provide an estimate of the joint distribution $\pi_n(\theta_{1:n})$ at index n. However, one can only expect to obtain "good" approximations of the most "recent" marginal distributions $\pi_n(\theta_{k:n})$ for $n-k$ say below 10. Indeed, if particles are resampled many times between time k and n, there are very few distinct paths $\{\theta_{1:k}^{(i)}\}$ at index n. Fortunately, this is the only requirement in many applications.

We have presented here a simple generic SMC method. However, like MCMC methods, SMC methods are not a black box and it is necessary to design carefully the algorithm so as to obtain good performance for a reasonable number of particles. Recently many papers have proposed various SMC methods to improve this basic scheme: construction of efficient importance sampling distributions, Rao-Blackwellised estimates, use of MCMC moves, etc. A comprehensive coverage of state-of-the-art techniques on the subject can be found in [3].

2 Some Applications

2.1 Kalman Filters

(Harvey, 1989; Anderson and Moore, 1979) In cases where the state space model is linear and Gaussian, the classic Kalman filter is optimal. In this case we have:

$$\begin{aligned} f(x_{t+1}|x_t) &= \mathcal{N}(x_{t+1}|Ax_t, C) \\ g(y_t|x_t) &= \mathcal{N}(y_t|Bx_t, D) \end{aligned}$$

where $\mathcal{N}(x|\mu, Q)$ is the Gaussian density function with mean vector μ and covariance matrix Q. We can write this equivalently as:

$$\begin{aligned} x_{t+1} &= Ax_t + v_t \\ y_t &= Bx_t + w_t, \end{aligned}$$

where v_t and w_t are zero mean Gaussian vectors with covariance matrices C and D, respectively. The errors v_t and w_t are independent over time and also independent of one another. We also require that the initial state be Gaussian distributed,

$$p(x_0) = \mathcal{N}(x_0|\mu_0, P_0).$$

We first require $p(x_{t+1}|y_{0:t})$, the prediction step from the above filtering recursion,

$$p(x_{t+1}|y_{0:t}) = \int p(x_t|y_{0:t})\, f(x_{t+1}|x_t)\, dx_t$$

Suppose that we have already that at time t

$$p(x_t|y_{0:t}) = \mathcal{N}(x_t|\mu_t, P_t).$$

Since $x_{t+1} = Ax_t + v_t$ the standard change of variables (linear Gaussian case) gives

$$p(x_{t+1}|y_{0:t}) = \mathcal{N}(x_{t+1}|\mu_{t+1|t}P_{t+1|t})$$

where

$$\mu_{t+1|t} = A\mu_t, \;\; P_{t+1|t} = C + AP_tA'.$$

The correction step of the above filtering recursion is

$$p(x_{t+1}|y_{0:t+1}) = \frac{g(y_{t+1}|x_{t+1})\, p(x_{t+1}|y_{0:t})}{p(y_{t+1}|y_{0:t})}.$$

Substituting the above Gaussian forms into the numerator gives

$$\begin{aligned}
p(x_{t+1}|y_{0:t+1}) &\propto \mathcal{N}(y_{t+1}|Bx_{t+1}, D)\, \mathcal{N}(x_{t+1}|\mu_{t+1|t}, P_{t+1|t}) \\
&\propto \exp\left\{-\frac{1}{2}\left((y_{t+1} - Bx_{t+1})' D^{-1}(y_{t+1} - Bx_{t+1})\right)\right\} \\
&\quad \times \exp\left\{-\frac{1}{2}\left((x_{t+1} - \mu_{t+1|t})' P_{t+1|t}^{-1}(x_{t+1} - \mu_{t+1|t})\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left((x_{t+1} - \mu_{t+1})' P_{t+1}^{-1}(x_{t+1} - \mu_{t+1})\right)\right\} \\
&\propto \mathcal{N}(x_{t+1}|\mu_{t+1}, P_{t+1})
\end{aligned}$$

where

$$\begin{aligned}
\mu_{t+1} &= P_{t+1}\,(B'\,D^{-1}\,y_{t+1} + P_{t+1|t}^{-1}\,\mu_{t+1|t}), \text{ and} \\
P_{t+1} &= (B'\,D^{-1}\,B + P_{t+1|t})^{-1}.
\end{aligned}$$

This gives, after re-expressing inverse matrix,

$$\begin{aligned}
\mu_{t+1} &= \mu_{t+1|t} + K_t\,(y_{t+1} - B\,\mu_{t+1|t}), \text{ and} \\
P_{t+1} &= (I - K_t\,B)\,P_{t+1|t},
\end{aligned}$$

where

$$K_t = P_{t+1|t}\,B'\,(B\,P_{t+1|t}\,B' + D)^{-1}.$$

Finally the complete Kalman filtering recursion can be summarized as

$$\begin{aligned}
\mu_{t+1|t} &= A\,\mu_t \\
P_{t+1|t} &= C + A\,P_t\,A' \\
K_t &= P_{t+1|t}\,B'\,(B\,P_{t+1|t}\,B' + D)^{-1} \\
\mu_{t+1} &= \mu_{t+1|t} + K_t\,(y_{t+1} - B\mu_{t+1|t}) \\
P_{t+1} &= (I - K_t\,B)\,P_{t+1|t}.
\end{aligned}$$

2.2 Optimal Filtering

The application of SMC to optimal filtering was first presented in [5]. The problem of interest is estimating the state of a Markov process $\{X_k\}_{k\geq 1}$ given some observations $\{Y_k\}_{k\geq 1}$. The unobserved (hidden) Markov process is defined by

$$X_1 \sim \mu, \quad X_k|X_{k-1} \sim f(\cdot|X_{k-1})$$

whereas the observations are assumed to be independent conditional upon $\{X_k\}_{k\geq 1}$ with marginal distribution

$$Y_k|X_k \sim g(\cdot|X_k).$$

Estimating the posterior distribution of X_k given $Y_{1:k}$ is a very important problem known as optimal filtering. If the model is linear and Gaussian, the posterior distribution is Gaussian and its statistics can be computed using the Kalman filter. However in many real-world applications, these linearity and Gaussianity assumptions are not valid and one needs to use numerical methods. SMC methods can be applied directly to this problem by setting π_n as the posterior density of the collection of states $X_{1:n}$ given a realization of the observations $Y_{1:n} = y_{1:n}$. Indeed this posterior distribution satisfies

$$\pi_n(x_{1:n}) \propto \mu(x_1) \prod_{k=2}^{n} f(x_k|x_{k-1}) \prod_{k=1}^{n} g(y_k|x_k)$$

and is typically known up to a normalizing constant.

2.3 Population Monte Carlo and Static Parameter Inference

The filtering problem is characterized by the *dynamic* nature of the statistical model. However it is important to realize that SMC methods can also be used to perform inference about a *static* parameter. More generally, one is often interested in using SMC methods to sample from a sequence of distributions $\{\pi_n\}_{n\in\mathcal{N}}$ defined on a common measurable space E; each π_n being known up to a normalizing constant. For example, $\pi_n(x)$ could be the posterior distribution of a random parameter X given the observations available at time n. In a global optimization context, one could also define $\pi_n(x) \propto [\pi(x)]^{\gamma_n}$ where $\{\gamma_n\}$ is an increasing sequence such that $\gamma_n \to \infty$ so as to maximize $\pi(x)$; a similar idea is the basis of simulated annealing.

SMC methods described previously do not apply directly in this context as they address the case where π_n is defined on $E_n = E^n$ instead of E. However, it is still possible to use SMC methods by constructing an artificial sequence of distributions $\{\widetilde{\pi}_n\}_{n\in\mathcal{N}}$ where $\widetilde{\pi}_n$ is defined on E_n and satisfies

$$\int \widetilde{\pi}_n(x_{1:n})dx_{1:n} = \pi_n(x_n).$$

An obvious choice for $\widetilde{\pi}_n$ is given by

$$\widetilde{\pi}_n(x_{1:n}) = \pi_n(x_n) \prod_{k=2}^{n} L_k(x_{k-1}|x_k)$$

where $\{L_n\}_{n\in\mathcal{N}}$ is an arbitrary sequence of Markov transition kernels. The resulting SMC algorithm can be interpreted as an adaptive importance sampling resampling algorithm.

References

[1] Doucet, A. (2004) Sequential Monte Carlo Methods, Entry in Encyclopedia of Statistical Sciences, Second Edition, Wiley. In Print.

[2] Doucet, A., Godsill, S.J. and Andrieu, C. (2000) On sequential Monte Carlo sampling methods for Bayesian filtering, *Statist. Comput.*, **10**, 197-208, 2000.

[3] Doucet, A., De Freitas, J.F.G. and Gordon, N.J. (editors) (2001) *Sequential Monte Carlo Methods in Practice*. Springer Series in Statistics for Engineering and Information Science. New York: Springer-Verlag.

[4] Djurić, P. M. and J.-H. Chun, J. -H. (2002). An MCMC sampling approach to estimation of nonstationary hidden Markov models, *IEEE Transactions on Signal Processing*, **50**, 1113–1124.

[5] Gordon, N.J., Salmond, D.J. and Smith, A.F.M. (1993) Novel approach to nonlinear/non-Gaussian Bayesian state estimation, *IEE-Proc. F*, **140**, 107-113.

[6] Iba, Y. (2000) Population Monte Carlo algorithms, *Trans. Japan. Soc. Arti. Intel.*, 16, 279-286.

[7] Liu, J.S. and Chen, R. (1998) Sequential Monte Carlo methods for dynamic systems, *J. Am. Statist. Ass.*, **93**, 1032-1044.

[8] Pitt, M.K. and Shephard, N. (1999) Filtering via simulation: auxiliary particle filters, *J. Am. Statist. Ass.*, **94**, 590-599.