Conditional SMC genealogies

Suzie Brown

supervised by Adam Johansen, Jere Koskela, Paul Jenkins & Dario Spanò

June 1, 2018

1 Introduction

2 Background

The Wright-Fisher model (Wakeley, 2009, Chapter 3), popular in the analysis of population genetics, includes some simplifying assumptions that make it somewhat unrealistic for that application. Namely, it is assumed that the population size is constant, and that generations do not overlap. While these constraints may hamper the model's applicability to population genetics, it is pleasing to note that when re-purposed for the analysis of SMC genealogies, both of these rather restrictive assumptions apply automatically. At least in the most common SMC algorithms, the number of particles (population size) remains constant at each iteration, and the resampling (reproduction) procedure is applied to all particles at once.

However, a significant alteration must be made to the standard Wright-Fisher model before it can be applied to SMC. The offspring distributions of each particle are not exchangeable, because broadly speaking, offspring of parents with high weight will tend to have high weight themselves. This follows intuitively from the notion that propagating a particle from a high density region at time t will result in a new state that is close to the high density region at time $t+1$.

There are generalisations of the Wright-Fisher model that do not require exchangeability among individuals, since it is also an important feature in population models. For instance, it can incorporate the existence of a hereditary trait whose value affects fertility. To model SMC genealogies, it is necessary to go one step further by allowing dependence between offspring distributions at different generations. This alteration amounts to a significant qualitative change since, while the forward process is still Markovian, the reverse (coalescent) process is not.

Whilst the population genetics literature has typically focused on the proliferation of hereditary traits (Wakeley, 2009, Chapter 3), our interest lies only with the genealogical process itself. The advancement of the trait, say "being in a high density region" is taken care of in the construction of the SMC algorithm. However, it is well-known that the problem of "ancestral degeneracy" is the source of much pain for practitioners; hence our interest in quantifying this problem.

2.1 Previous work

Kingman (Kingman, 1982a,b,c) studied the asymptotic properties of the genealogies arising in a number of exchangeable population models, including the standard Wright-Fisher model. He showed that they each converge in the limit $N \to \infty$ (with appropriate time rescaling) to a Markov process termed the n-coalescent. By studying this limiting process, he was able to establish some new results about the behaviour of such populations, and produce novel derivations of some known results.

Möhle (1998) extends this result to a class of non-exchangeable models, with the assumption that offspring distributions are independent between different generations. This allows application to more complex population models; but is still too restrictive to cover SMC genealogies, which by construction have strong dependence between generations.

Allowing for dependence between generations significantly complicates the computations, because the coalescent process is no longer Markovian. Koskela et al. (2018) addresses a first case of such models, applicable to a range of standard SMC procedures, under some reasonable conditions.

We would like to extend to the case of conditional SMC, which differs substantially from the case of Koskela et al. (2018) because there is a particular ancestral line that is conditioned to survive. This extension is important to make the results applicable to the particle Gibbs algorithm (Andrieu et al., 2010), which is popular across a range of applications.

2.2 Sequential Monte Carlo

2.2.1 Conditional SMC

3 Theoretical results

We now consider extending the results of Koskela et al. (2018) to the case of conditional SMC. In particular, the SMC updates will be conditioned on a particular trajectory surviving. We concentrate on the exchangeable model, so we may take WLOG that the "immortal line" is the trajectory containing individual 1 from each generation. We first assume the simplest case, with multinomial resampling; analogous to the standard SMC case where

$$v_t^{(i)} \overset{d}{=} \mathrm{Bin}(N, w_t^{(i)}), \qquad i = 1, \dots, N$$

yielding the coalescence rate

$$c_N(t) := \frac{1}{(N)_2} \sum_{i=1}^N \mathbb{E}\left[(v_t^{(i)})_2 \right] = \sum_{i=1}^N \mathbb{E}\left[(w_t^{(i)})^2 \right]. \tag{1}$$

But now, since the first line is immortal, in each time step the first individual must have at least one offspring. The remaining $N-1$ offspring are assigned multinomially to the N possible parents as usual, giving the offspring numbers:

$$\begin{aligned} \tilde{v}_t^{(1)} &\overset{d}{=} 1 + \mathrm{Bin}(N-1, w_t^{(1)}) \\ \tilde{v}_t^{(i)} &\overset{d}{=} \mathrm{Bin}(N-1, w_t^{(i)}), \qquad i = 2, \dots, N. \end{aligned}$$

We therefore have the following moments (using tower property):

$$\begin{aligned} &\mathbb{E}[\tilde{v}_t^{(i)}] = (N-1)\mathbb{E}[w_t^{(i)}] \\ &\mathbb{E}[(\tilde{v}_t^{(i)})^2] = (N-1)(N-2)\mathbb{E}[(w_t^{(i)})^2] + (N-1)\mathbb{E}[w_t^{(i)}] \qquad\qquad i = 2, \dots, N \\ &\mathbb{E}[\tilde{v}_t^{(1)}] = (N-1)\mathbb{E}[w_t^{(1)}] + 1 \\ &\mathbb{E}[(\tilde{v}_t^{(1)})^2] = (N-1)(N-2)\mathbb{E}[(w_t^{(1)})^2] + 3(N-1)\mathbb{E}[w_t^{(1)}] + 1 \end{aligned}$$

and we can derive the altered coalescence rate:

$$\begin{aligned} \tilde{c}_N(t) &= \frac{1}{(N)_2} \sum_{i=1}^N \mathbb{E}\left[(\tilde{v}_t^{(i)})_2 \right] \\ &= \frac{1}{(N)_2} \mathbb{E}\left[(\tilde{v}_t^{(1)})^2 - \tilde{v}_t^{(1)} \right] + \frac{1}{(N)_2} \sum_{i=2}^N \mathbb{E}\left[(\tilde{v}_t^{(i)})^2 - \tilde{v}_t^{(i)} \right] \\ &= \frac{1}{(N)_2} \left[(N-1)(N-2)\mathbb{E}[(w_t^{(1)})^2] + 2(N-1)\mathbb{E}[w_t^{(1)}] \right] + \frac{1}{(N)_2} \sum_{i=2}^N (N-1)(N-2)\mathbb{E}[(w_t^{(i)})^2] \\ &= \frac{1}{(N)_2} \sum_{i=1}^N (N-1)(N-2)\mathbb{E}[(w_t^{(i)})^2] + \frac{1}{(N)_2} 2(N-1)\mathbb{E}[w_t^{(1)}] \\ &= \frac{N-2}{N} c_N(t) + \frac{2}{N}\mathbb{E}[w_t^{(1)}] \end{aligned} \tag{2}$$

Under the conditions of Koskela et al. (2018, Corollary 2), we have that $\mathbb{E}[w_t^{(1)}] = O(N^{-1})$, and hence

$$\tilde{c}_N(t) - c_N(t) = O(N^{-2}).$$

Koskela et al. (2018) gives the following bounds on $c_N(t)$:

$$\frac{C_*}{N-1} \leq c_N(t) \leq \frac{C}{N-1}$$

Then, since $\tilde{c}_N(t)$ differs from $c_N(t)$ by $O(N^{-2})$, for sufficiently large N there exist constants $\tilde{C}, \tilde{C}_*$ such that

$$\frac{\tilde{C}_*}{N-1} \leq \tilde{c}_N(t) \leq \frac{\tilde{C}}{N-1}$$

and we can thus derive bounds analogous to Koskela et al. (2018, (5)-(6)):

$$\frac{N-1}{\tilde{C}_*} t \leq \tilde{\tau}_N(t) \leq \frac{N-1}{\tilde{C}} t \tag{3}$$

$$\frac{N-1}{\tilde{C}_*}(s-t) \leq \tilde{\tau}_N(s) - \tilde{\tau}_N(t) \leq \frac{N-1}{\tilde{C}}(s-t) \tag{4}$$

Furthermore, we have that

$$\begin{aligned} \frac{\tilde{C}}{N-1} &= \frac{N-2}{N} \frac{C}{N-1} + O(N^{-2}) \\ &= \frac{C}{N-1} + O(N^{-2}) \end{aligned}$$

therefore $\tilde{C} - C = O(N^{-1})$ and similarly $\tilde{C}_* - C_* = O(N^{-1})$. Hence the bounds in (3), (4) are asymptotically equal to those of Koskela et al. (2018, (5)–(6)).

It remains to verify that the conditions (Koskela et al., 2018, (3)–(4)) can extend to this case. If so, a modified version of Koskela et al. (2018, Theorem 1) and its corollaries will hold, by the same argument, for conditional SMC.

4 Simulation study

In order to investigate how well the asymptotic results hold for finite N, we conducted a simulation study on the Ornstein-Uhlenbeck model, a "simplest case" hidden Markov model which is popular for such studies in the literature:

$$\begin{aligned} &X_0 \sim \mathcal{N}(0,1) \\ &X_{t+1} \mid X_t \sim \mathcal{N}((1-\Delta)X_t, \Delta) \\ &Y_t \mid X_t \sim \mathcal{N}(X_t, \sigma^2) \end{aligned}$$

with parameters $\Delta > 0, \sigma > 0$.

Tree height T is one of the basic properties of an ancestral sub-tree. It denotes the number of generations back one must go to find the most recent common ancestor (MRCA) of a sample of n leaves (individuals from generation N_{obs}). That is, how many time steps of the reverse process pass before the n sampled lineages all coalesce to a single lineage. If the asymptotic genealogical process is known to be an n-coalescent, moments of T are available analytically. In particular, we expect for conditional SMC to obtain the same limiting process as for standard SMC, since we have shown in Section [REF] that their coalescent rates are asymptotically equal. Therefore in the limit as $N \to \infty$ we expect the moments of T to behave as stated in Koskela et al. (2018, Corollary 1):

$$\frac{C_*}{C^2}\left(1 - \frac{1}{n}\right) + O(N^{-1}) \leq \mathbb{E}[T/N] \leq 2\frac{C}{C_*^2}\left(1 - \frac{1}{n}\right)$$

$$\left(\frac{4\pi^2}{3} - 12 + O(n^{-1})\right)\left(\frac{C_*}{C^2}\right)^2 + O(N^{-1}) \leq \mathrm{Var}(T/N) \leq \left(\frac{4\pi^2}{3} - 12 + O(n^{-1})\right)\left(\frac{C}{C_*^2}\right)^2$$

where of course T depends on n and N. In particular, the choice of immortal line does not feature in these bounds and so should have no effect as $N \to \infty$ with respect to n.

Following Koskela et al. (2018), we take $\Delta = \sigma = 0.1$ and generate one fixed sequence of observations for use in all SMC runs. We use a range of values $\{256, 512, \ldots, 4096\}$ for N, and two fixed values $n = 2, 16$ intended to show the qualitative difference in behaviour as we cross a supposed "$n << N$" threshold. The number of observations N_{obs} is taken such that, for all the choices of n and N, the N_{obs} generations of SMC particles are enough for the n sampled lineages to coalesce to one common ancestor (with high enough probability that it happens reliably on every repetition); ensuring that the tree height can always be recorded.

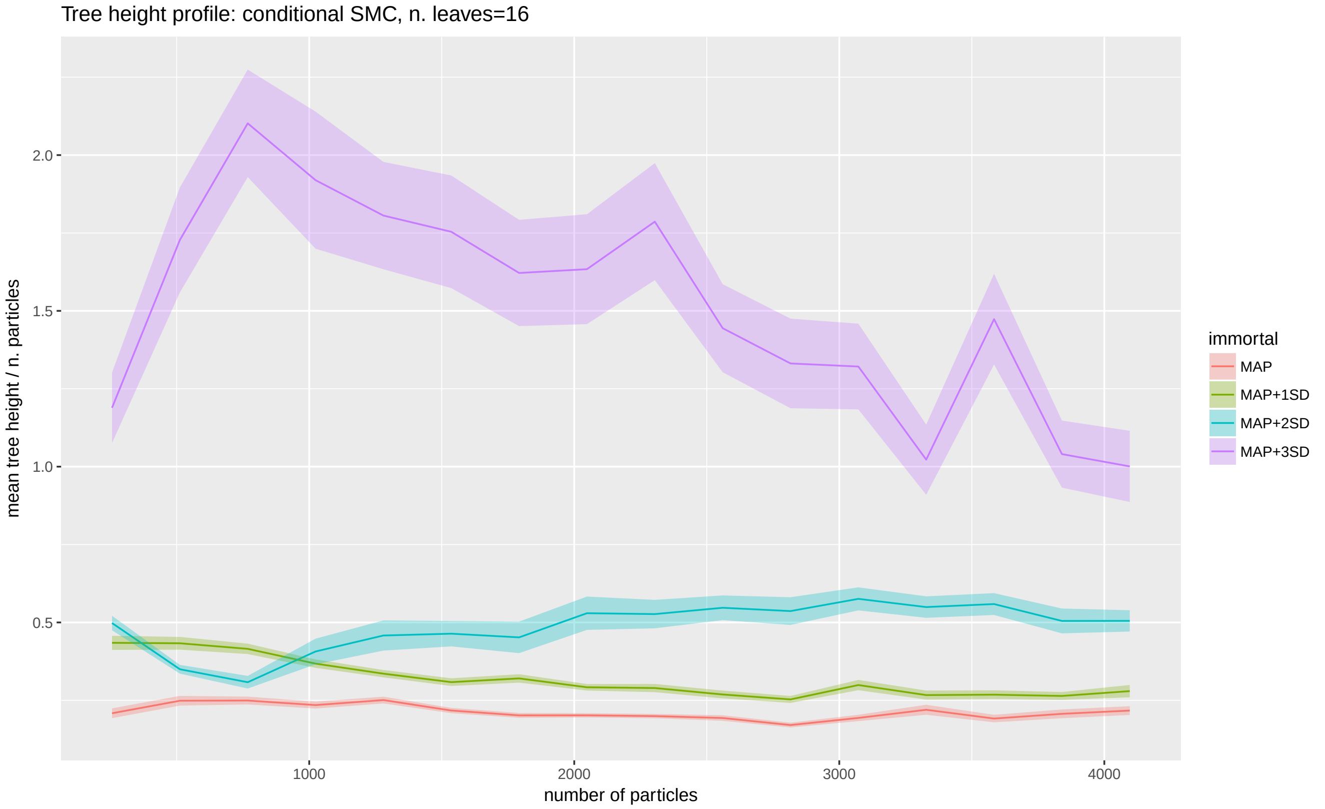

Figure 1: $\mathbb{E}(T/N)$ for samples of $n = 16$ leaves from conditional SMC where the immortal line is 0,1,2,3 standard deviations away from the MAP estimate. Each point is averaged over 100 repetitions of running the particle filter and sampling n leaves; and the same sequence of observations was used for every run. Different choices of immortal line clearly have a profound effect on tree height: in the number of generations taken for 16 lineages to coalesce, it is likely that the subtree will contain part of the immortal line.

For this toy model, the smoothing distribution is available analytically through the Rauch-Tung-Striebel (RTS) smoother (Rauch et al., 1965). We exploited this solution to choose the "immortal line" on which to condition the conditional SMC updates. Because both the MAP estimate (equal to the mean since the distributions are Gaussian) and variance are available via the RTS smoother, we were able to produce a sequence of immortal lines of decreasing likelihood, by adding multiples of the standard deviation to the mean.

We hypothesised that when N is not too much bigger than n, an "unlikely" choice of the immortal line should produce qualitative differences in the tree height profile; because the$n$ lineages would often coalesce to the immortal line, which corresponds to an unlikely choice of ancestors under the unconditional algorithm. On the other hand, when N is very large with respect to n, this effect should become less significant because the sampled lineages should usually coalesce before interacting with the immortal line.

Figures 1 and 2 illustrate this distinction. Here we use a decrease in n as a proxy for increasing N: over the same range of values for N, Figure 1 shows the profile for sample size $n = 16$, and Figure 2 for $n = 2$. We see clearly that in the case of $n = 16$, the likelihood of the immortal line significantly affects the tree height profile, while for $n = 2$ it makes no appreciable difference.

In any case the mean tree height seems to be higher for conditional SMC (around 0.3) compared to standard SMC (around 0.2), although it is not yet entirely clear.

5 Conclusions

We have showed that under multinomial resampling, the coalescence probability for conditional SMC is close enough to that of standard SMC to imply the existence of a time rescaling that is asymptotically equivalent to that defined in Koskela et al. (2018, p.7). Given this result, we expect their result to be transferable to the conditional SMC setting, although conditions Koskela et al. (2018, (3)–(4)) still need to be verified.

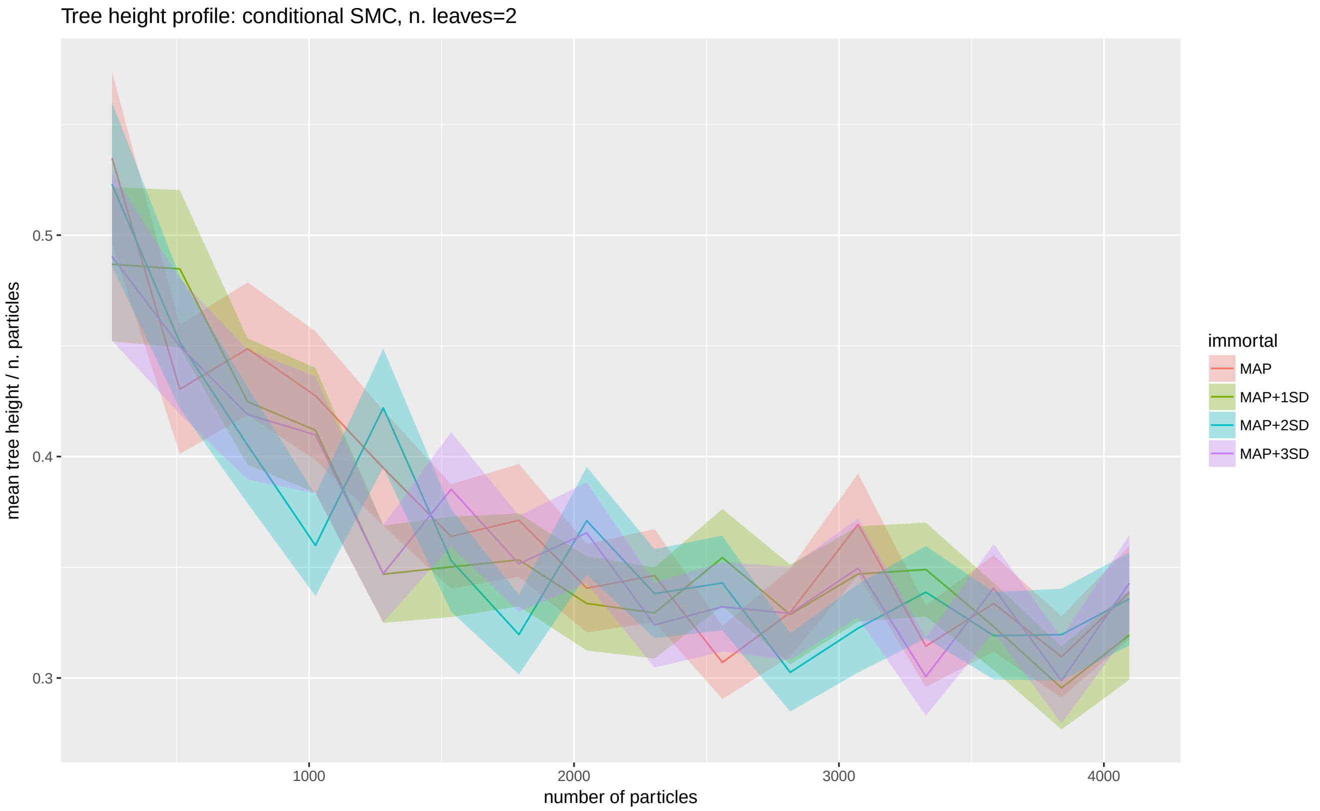

Figure 2: $\mathbb{E}(T/N)$ for samples of $n = 2$ leaves from conditional SMC where the immortal line is 0,1,2,3 standard deviations away from the MAP estimate. Each point is averaged over 1000 repetitions of running the particle filter and sampling n leaves; and the same sequence of observations was used for every run. Now the choice of immortal line does not seem to affect the tree height: n is so small compared to N that it is unlikely to sample a pair of lineages that meet the immortal line before coalescing.

While Koskela et al. (2018, Section 3) demonstrates that in the case of standard SMC, the asymptotic results seem to hold even for $n = N$, we conclude from the simulation study that the asymptotic results for conditional SMC will require $n << N$. Although we have demonstrated empirically that $n << N$ is required to mitigate the effect of choosing an unlikely immortal line, this in itself should not raise concerns with those implementing the particle Gibbs algorithm. In particle Gibbs, the immortal line is sampled from the trajectories generated in the previous iteration, proportionally to their likelihood. Unless all the generated trajectories are unlikely (in which case there are more profound problems with the algorithm), the immortal line will therefore usually be a reasonably likely one.

We expect that less straight-forward conditional SMC algorithms, using alternative resampling schemes, should have the same limiting genealogical process (up to constants) as the multinomial scheme; as was hypothesised and demonstrated empirically in Koskela et al. (2018) for standard SMC. Proving this even for standard SMC remains an open problem. There are many other directions for generalising these results, in particular investigating their robustness under violation of the various assumptions, either theoretically or empirically. It is of particular interest to consider the classes of SMC algorithms most often used in practice. It was to this end that we began looking at conditional SMC, since the particle Gibbs algorithm is widely used in numerous applications.

References

Andrieu, C., Doucet, A. and Holenstein, R. (2010), 'Particle markov chain monte carlo methods', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**(3), 269–342.

Kingman, J. (1982*a*), 'The coalescent', *Stochastic processes and their applications* **13**(3), 235–248.

Kingman, J. (1982*b*), 'Exchangeability and the evolution of large populations'.

Kingman, J. (1982*c*), 'On the genealogy of large populations', *Journal of Applied Probability* **19**(A), 27–43.

Koskela, J., Jenkins, P. A., Johansen, A. M. and Spano, D. (2018), 'Asymptotic genealogies of interacting particle systems with an application to sequential monte carlo', *arXiv preprint arXiv:1804.01811* .

Möhle, M. (1998), 'Robustness results for the coalescent', *Journal of applied probability* **35**(2), 438–447.

Rauch, H. E., Striebel, C. and Tung, F. (1965), 'Maximum likelihood estimates of linear dynamic systems', *AIAA journal* **3**(8), 1445–1450.

Wakeley, J. (2009), *Coalescent theory: an introduction*, number 575: 519.2 WAK.