Genealogies of Sequential Monte Carlo Algorithms

Suzie Brown
with Jere Koskela, Adam Johansen and Paul Jenkins
Department of Statistics, University of Warwick

Sequential Monte Carlo

Suppose we have a hidden Markov model with noisy observations $Y_{1:T}$ of unobservable states $X_{1:T}$:

$$\cdots \xrightarrow{f} X_{t-1} \xrightarrow{f} X_t \xrightarrow{f} \cdots$$

with Y_{t-1} and Y_t related to X_{t-1} and X_t through g.

We may want to infer posterior filtering densities $p(x_t|y_{1:t})$, or smoothing densities $p(x_{1:t}|y_{1:t})$. These are not available analytically, except in linear Gaussian models.

SMC [1] approximates the posterior distributions by sampling N particles from the prior distribution and iterating the following steps:

1. **Propagate:** move particles through the Markov kernel f
2. **Calculate weights:** weight particles by how likely they are to produce the observation through g
3. **Resample:** duplicate high-weight particles and kill off low-weight particles to obtain a new sample of size N.

The figure shows how the population of particles looks at each step, before resampling. The purple ribbon shows the exact posterior mode and 95% high posterior density set.
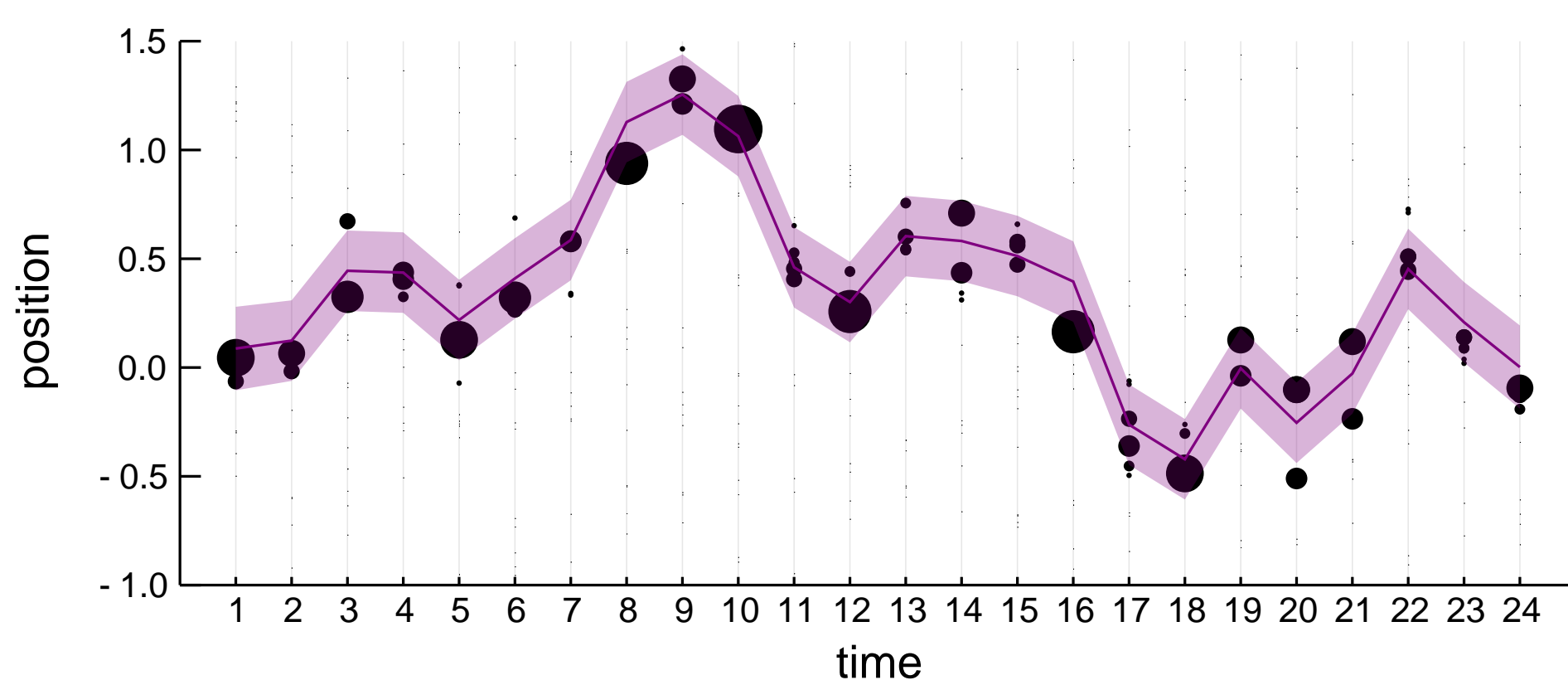

Figure: Exact posterior (purple) and weighted SMC particles before resampling (black) for a linear Gaussian model.

Ancestral Degeneracy

These particles cannot approximate the smoothing distributions: for that we need a sample of *trajectories* covering all t time steps, not just a sample of particles at each time step.

A sample of N trajectories is given by the lineages of the N particles alive at time t. But due to the resampling mechanism, these lineages coalesce backwards in time, leaving many fewer than N distinct samples at time 1.
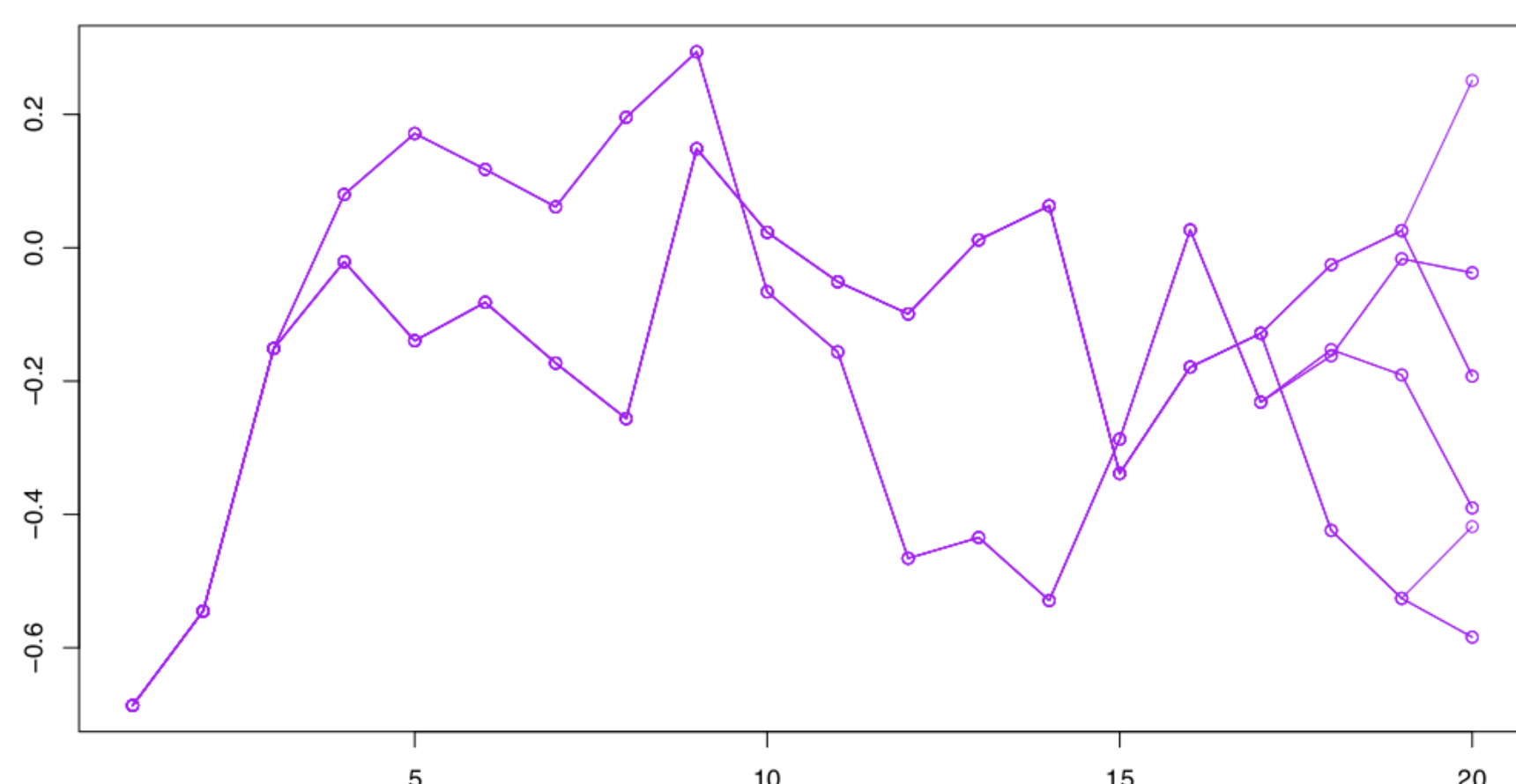
Figure: Trajectories from a sample of N=6 particles. At most times there are just one or two distinct lineages.

This phenomenon is known as *ancestral degeneracy* [2]. It can be mitigated by changing the resampling procedure, but it cannot be eradicated.

The Kingman Coalescent

The Kingman coalescent [3] is a continuous-time process which describes the asymptotic genealogies of many population models.

The process starts with a population of size N and ends when all of the lineages have merged into one. Each pair of lineages merges with rate 1, meaning that the total merger rate is higher when there are more distinct lineages.

The resulting coalescent has many lineages for only a short time, and spends most of its time with just two or three distinct lineages.
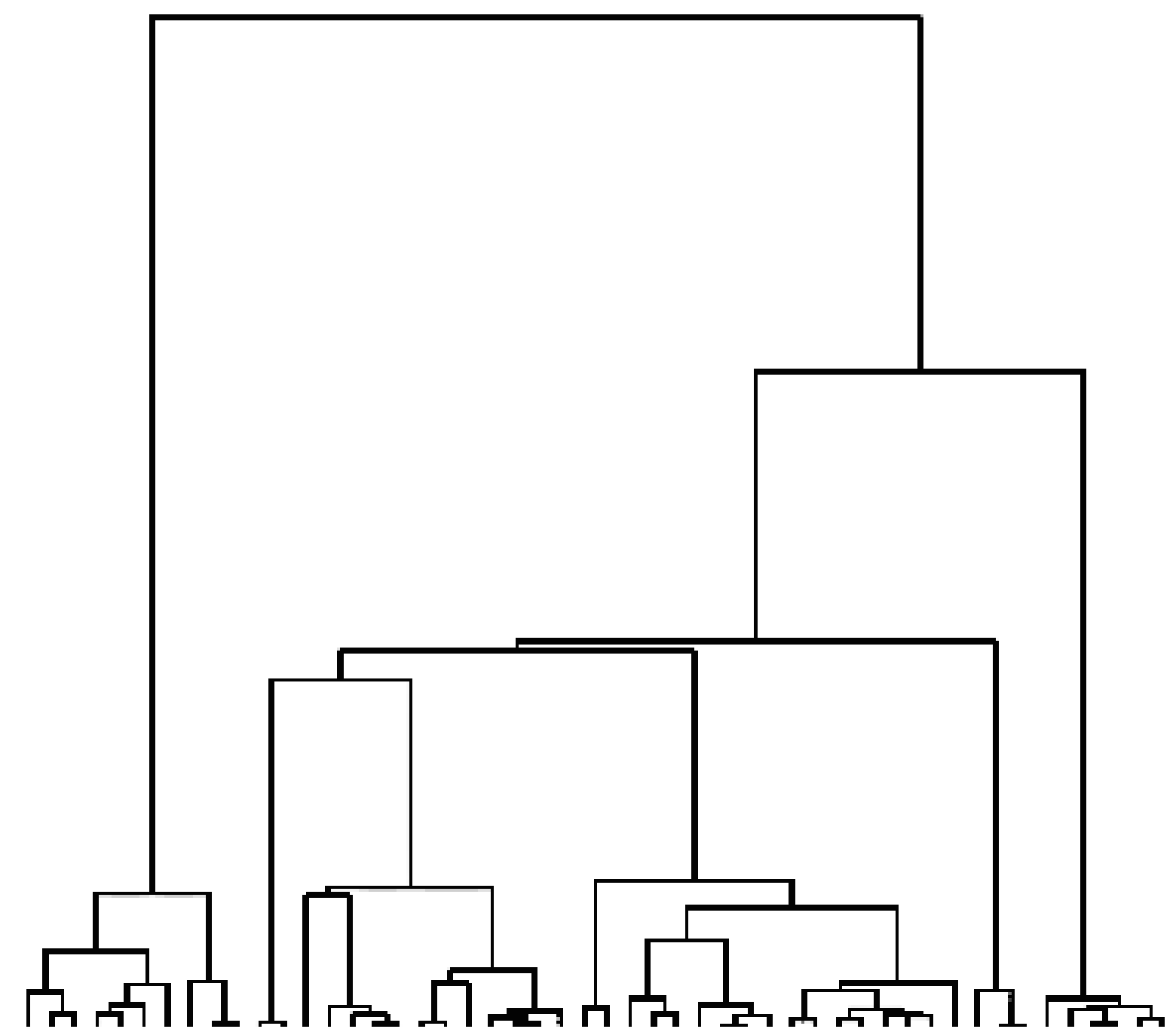
Figure: A realisation of the Kingman coalescent for a population of size N=50. (Source: Wikimedia Commons)

Asymptotic SMC Genealogies

The rate of coalescence of SMC genealogies depends on the resampling scheme used.

To obtain convergence to the Kingman coalescent, which has rate 1, we rescale time by the inverse coalescence rate, and consider the asymptotic regime as $N \to \infty$.

So far, convergence to the Kingman coalescent is only proved for the simple case of multinomial resampling [4]. We have since extended this result to cover conditional SMC [5] with multinomial resampling.

Next we would like to analyse the coalescents induced by alternative resampling schemes such as residual and stratified resampling. These schemes reduce the offspring variance, and are more popular among practitioners.

1. Gordon, N. J., Salmond, D. J., & Smith, A. F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)* 140(3), 107-113. IET Digital Library.
2. Doucet, A., & Johansen, A. M. (2011). A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, 656-704. OUP.
3. Kingman, J.F.C. (1982). The coalescent. *Stochastic processes and their applications*, 13(3), 235-248.
4. Koskela, J. et al. (2018). Asymptotic genealogies of interacting particle systems with an application to sequential Monte Carlo. *arXiv preprint* arXiv:1804.01811.
5. Andrieu, C., Doucet, A., & Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3), 269-342.