

Population Genetics for Sequential Monte Carlo Algorithms

Suzie Brown

University of Warwick

with Paul Jenkins, Adam Johansen & Jere Koskela

SMC Genealogies

The selection step of the SMC algorithm induces a genealogy, since each particle has a "parent" in the previous generation. These particle lineages coalesce backwards in time.

Coalescing lineages are the source of the "ancestral degeneracy" problem. That is, because the paths coalesce backwards in time, fewer distinct samples are available at earlier times. By analysing genealogies we can compare the severity of ancestral degeneracy between different SMC algorithms.

Kingman's n-coalescent

Kingman's n-coalescent (Kingman 1982 Stoch. Proc. and Appl.) is a partition-valued stochastic process on $[n]$ such that each pair of blocks in the partition merges at unit rate backwards in time.

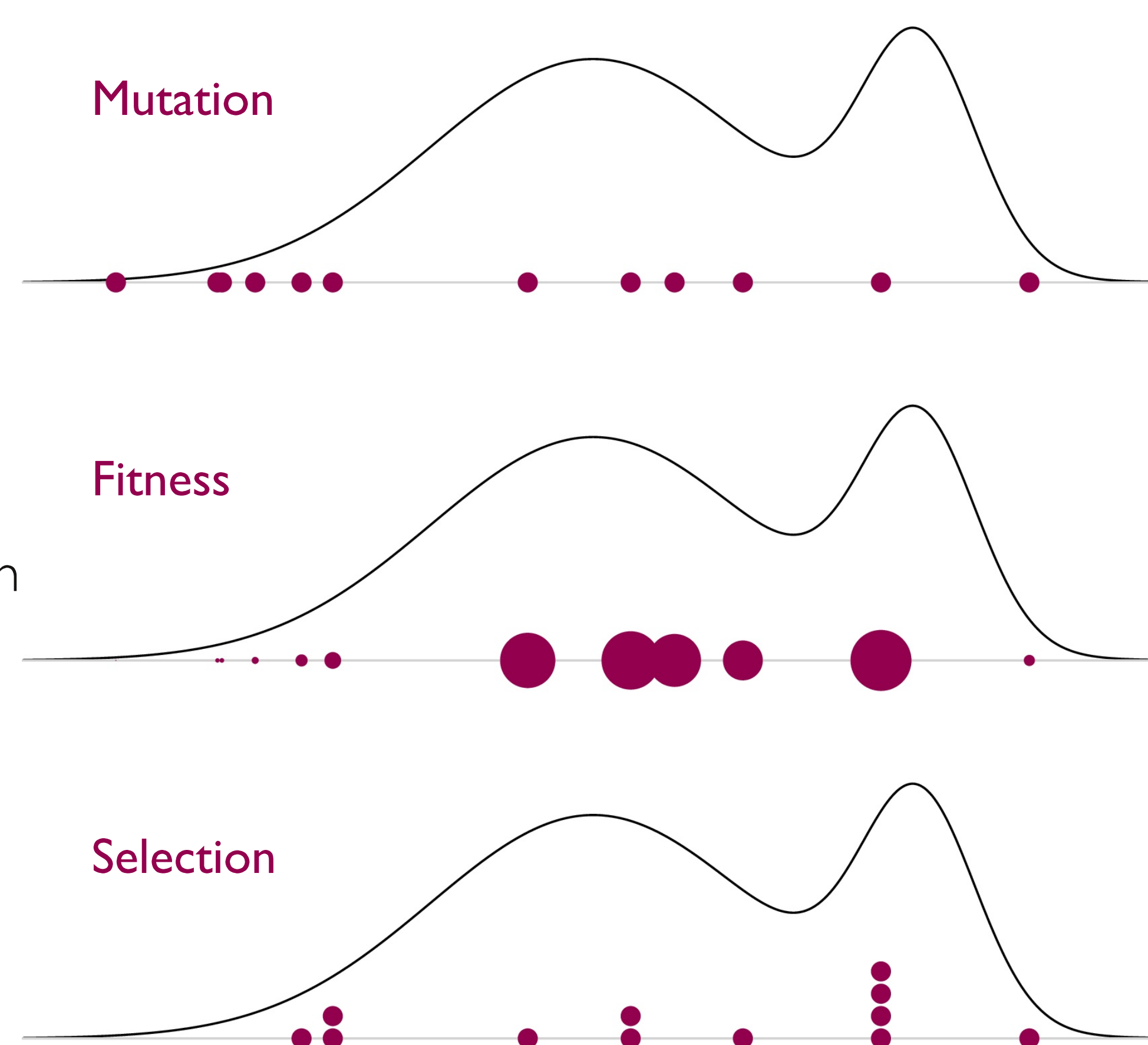
It is known to be the limiting genealogy for neutral models, but our results show it can also appear in populations with strong selection.

Conditional SMC

Conditional SMC is a variant algorithm used in particle MCMC (Andrieu et al 2010 RSS:B) in which one particular lineage is conditioned to survive all of the mutation and selection steps. This "immortal lineage" has a distinct role in the genealogical process.

For finite N , the behaviour is similar to a seed-bank model. But in the large- N limit with time rescaled appropriately, we still recover Kingman's n-coalescent for the finite-dimensional distributions. The probability of sampling the immortal lineage among n lineages vanishes sufficiently quickly.

Sequential Monte Carlo (SMC) algorithms simulate a population of N particles to approximate probability distributions of interest. The algorithm iterates through three steps, each of which has an interpretation in population genetics.



The positions of the particles are updated, according to the dynamics of the system of interest.

A fitness value is calculated for each particle, according to how important the particle is in the approximation.

Particles reproduce in such a way that particles with high fitness are replicated many times, while particles with low fitness are killed off.

Selection Mechanisms

The selection step requires a vector of fitness values to be stochastically transformed into a vector of family sizes.

Intuitively, a "good" selection mechanism should tend to replicate high-fitness particles and kill off low-fitness particles. This still leaves plenty of flexibility.

We enforce the following properties of "valid" selection mechanisms:

1. The number of particles N remains fixed.
2. After resampling each particle has equal fitness.
3. The expected number of offspring of each parent is proportional to its fitness.

Multinomial Selection

The simplest selection mechanism is for each offspring to choose its parent independently in proportion to the parents' fitness values. That is, the parents are sampled from a Categorical distribution parametrised by the parents' fitnesses. The resulting family sizes follow a Multinomial distribution.

Low-Variance Selection

Multinomial selection is known to be sub-optimal in terms of the resulting Monte Carlo variance. Proposed alternatives include residual, stratified and systematic selection (Whitley 1994 Stat. and Comp., Kitagawa 1996 Comp. and Graph. Stat.). We study a general class of selection mechanisms based on "stochastic rounding".

Stochastic Rounding

We show that all selection mechanisms based on stochastic rounding yield Kingman genealogies in the large- N limit.

Compared to the n-coalescent, the genealogies of the SMC algorithm are rescaled by a random factor. This rescaling factor has the same distribution whichever stochastic rounding is used.

