The theorem of **?**

Suzie Brown

May 8, 2018

Framework

- Current generation is generation 0; generations enumerated backward in time.
- Generations are non-overlapping.

Notation

M_r population size at generation r

$N := M_0$ initial population size

$v_i^{(r)}$ (random) number of offspring of individual i of generation r

$n \leq N$ size of the sample of individuals in generation 0 to be considered

T_n the number of generations since the most recent common ancestor (MRCA) of the sample

$\mathcal{R}_r$ the equivalence relation that contains the pair (i, j) iff individuals i and j in the sample have a common ancestor in generation r

$\{\mathcal{R}_r\}_{r\in\mathbb{N}}$ will be referred to as the *ancestral process*

Δ the minimal relation $\{(i, i); i = 1, \ldots, n\}$

Θ the maximal relation $\{(i, j); i, j = 1, \ldots, n\}$

$p_{\xi\eta}(r)$ the transition probability $\mathbb{P}(\mathcal{R}_r = \eta \mid \mathcal{R}_{r-1} = \xi)$ of the ancestral process

c_r the probability that a random pair of distinct individuals from generation r have a common ancestor in generation $r - 1$, called the *coalescence probability*

$\sigma^2(r)$ the expected *mean crowding* of the offspring variables of generation r

X_r the number of descendants in generation r in the forward genealogical process

$\pi_{ij}(r)$ the transition probability $\mathbb{P}(X_{r-1} = j \mid X_r = i)$ of the forward genealogical process

$(x)_y$ denotes the descending factorial $x(x - 1) \cdots (x - y + 1)$

Assumptions

1. $\{v_1^{(r)}, \ldots, v_{M_r}^{(r)}\}$ is independent of $\{v_1^{(s)}, \ldots, v_{M_s}^{(s)}\}$ for $r \neq s$
2. $v_1^{(r)}, \ldots, v_{M_r}^{(r)}$ are **not** assumed to be exchangeable

Initial remarks

1. From the definitions, we have that
$$\sum_{i=1}^{M_r} v_i^{(r)} = M_{r-1} \tag{1}$$
2. Assumption 1 ensures that $\{\mathcal{R}_r\}_{r\in\mathbb{N}}$ is a Markov process (hence the applicability of transition probabilities).

Coalescence rate

A combinatorial argument allows us to derive an expression for the transition probability $p_{\xi\eta}(r)$ of the ancestral process:
$$p_{\xi\eta}(r) = \frac{1}{(M_{r-1})_b} \sum_{\substack{i_1,\dots,i_a=1 \\ \text{distinct}}}^{M_r} \mathbb{E}\left[(v_{i_1}^{(r)})_{b_1} \cdots (v_{i_a}^{(r)})_{b_a}\right] \tag{2}$$
The sum is over all the possible ordered choices of the a parents from generation r. Inside the sum is the expected number of ways to have at least the required number of offspring from each parent, given this choice of parents. Thus the whole sum represents the probability of finding a set of parents that produce the required number of offspring. Then dividing by the number of ordered ways to choose b offspring from generation $r-1$ ensures that the parents produce the *correct* ordered offspring. Overall then we have the probability that exactly the right subsets of offspring from generation $r-1$ coalesce in generation r, counting all the different parents to which they could coalesce.

We define the the *coalescence probability*, i.e. the probability that a randomly chosen pair of (distinct) individuals from generation $r-1$ have a common ancestor in generation r:
$$c_r := \frac{1}{(M_{r-1})_2} \sum_{i=1}^{M_r} \mathbb{E}\left[(v_i^{(r)})_2\right] \tag{3}$$

Algebraic tools for proof

Here are a few identities and inequalities that will be referred to when proving the theorem. (4) is obtained using a multinomial expansion, (5) using Bernoulli's inequality, and (6)–(9) by expanding factorials.
$$\sum_{i_1\dots i_m=1}^{n} \prod_{j=1}^{m} x_{i_j} = \prod_{j=1}^{m} \sum_{i=1}^{n} x_i = \left(\sum_{i=1}^{n} x_i\right)^m \tag{4}$$
$$(k-x)^n = k^n \left(1 - \frac{x}{k}\right)^n \leq k^n - nxk^{n-1} \tag{5}$$
$$n^a \geq (n)_a \tag{6}$$
$$(n)_a \leq (n)_b \, n^{a-b}, \text{ if } 0 \leq b \leq a \tag{7}$$
$$\frac{n^{a-b}}{(n)_a} = \frac{1}{(n)_b} + O(n^{-b-1}) \tag{8}$$
$$\frac{1}{(n)_b} = \frac{1}{n^b} + O(n^{-b-1}) \tag{9}$$

The theorem

Now for simplicity we assume a constant population size $M_r \equiv N$, which for the purposes of SMC will generally be satisfied.

Theorem 1. *Let $T \subset \mathbb{R}$ and suppose there is a function $\tau : T \to \mathbb{N}_0$ satisfying:*

(A) correct limiting coalescence rate
$$\forall t \in T, \lim_{N\to\infty} \sum_{r=1}^{\tau(t)} c_r = t$$

(B) variance of coalescence rate goes to zero

$$\forall t \in T,\ \lim_{N\to\infty} \sum_{r=1}^{\tau(t)} c_r^2 = 0$$

(C) no triple coalescences

$$\forall t \in T, \forall k \in \mathbb{N}_0,\ \lim_{N\to\infty} \sup_{r\le\tau(t)} \frac{1}{N^3 c_r} \sum_{i=1}^{N} \mathbb{E}\left[(v_i^{(r)})_2 (v_i^{(r)})^k\right] = 0$$

(D) only one coalescence at a time

$$\forall t \in T,\ \lim_{N\to\infty} \sup_{r\le\tau(t)} \frac{1}{N^4 c_r} \sum_{i,j=1}^{N} \mathbb{E}\left[(v_i^{(r)})_2 (v_j^{(r)})^2\right] = 0$$

Then the finite-dimensional distributions of $\{\mathcal{R}_{\tau(t)}\}_{t\in T}$ converge to those of the n-coalescent (with time restricted to T) in the limit $N \to \infty$.

Note that, since (being a probability) $c_r \ge 0$ for all r; under (A), (B) is equivalent to

$$\forall t \in T,\ \lim_{N\to\infty} \sup_{r\le\tau(t)} c_r = 0 \tag{10}$$

Proof. We first bound the transition probability $p_{\xi\eta}(r)$ as given by (2), in each of four possible cases. This will show that the only type of coalescence event to occur at any one time in the limit $N \to \infty$ is a merger of exactly one pair of lineages (**Case 1.** below). Assume the offspring numbers $v_i^{(r)}$ are known (so we drop the expectations).

Case 1. η is obtained from ξ by merging exactly one pair of lineages, i.e. $b_1 = 2, b_2 = \cdots = b_a = 1$, and $b = a + 1$. We derive an upper bound:

$$\begin{aligned}
p_{\xi\eta}(r) &= \frac{1}{(N)_b} \sum_{\substack{i_1,\ldots,i_a=1 \\ \text{distinct}}}^{N} (v_{i_1}^{(r)})_2 (v_{i_2}^{(r)})_1 \cdots (v_{i_a}^{(r)})_1 \\
&\le \frac{1}{(N)_b} \sum_{i_1,\ldots,i_a=1}^{N} (v_{i_1}^{(r)})_2 (v_{i_2}^{(r)})_1 \cdots (v_{i_a}^{(r)})_1 && \text{dropping distinctness} \\
&= \frac{1}{(N)_b} \sum_{i=1}^{N} (v_i^{(r)})_2 \sum_{i_2,\ldots,i_a=1}^{N} v_{i_2}^{(r)} \cdots v_{i_a}^{(r)} \\
&= \frac{1}{(N)_b} \sum_{i=1}^{N} (v_i^{(r)})_2 (v_1^{(r)} + \cdots + v_N^{(r)})^{a-1} && \text{using (4)} \\
&= \frac{N^{b-2}}{(N)_b} \sum_{i=1}^{N} (v_i^{(r)})_2 && \text{using (1)} \\
&= \left(\frac{1}{(N)_2} + O(N^{-3})\right) \sum_{i=1}^{N} (v_i^{(r)})_2 && \text{using (8)} \\
&= c_r + o(c_r) && \text{by (3) and (C)}
\end{aligned} \tag{11}$$

and a lower bound:

$$
\begin{aligned}
p_{\xi\eta}(r) &= \frac{1}{(N)_b} \sum_{\substack{i_1,\ldots,i_a=1 \\ \text{distinct}}}^{N} (v_{i_1}^{(r)})_2 (v_{i_2}^{(r)})_1 \cdots (v_{i_a}^{(r)})_1 \\
&= \frac{1}{(N)_b} \sum_{i=1}^{N} (v_i^{(r)})_2 \sum_{\substack{i_2,\ldots,i_a=1 \\ \text{distinct}\neq i}}^{N} (v_{i_2}^{(r)})_1 \cdots (v_{i_a}^{(r)})_1 \\
&\geq \frac{1}{(N)_b} \sum_{i=1}^{N} (v_i^{(r)})_2 \sum_{j\neq i} \sum_{\substack{i_3,\ldots,i_a \\ \neq i}} \left[v_j \cdot v_{i_3} \cdots v_{i_a} - \binom{b-2}{2} (v_j^{(r)})^2 \cdot v_{i_4} \cdots v_{i_a} \right] \qquad (12) \\
&= \frac{1}{(N)_b} \sum_{i=1}^{N} (v_i^{(r)})_2 \left[\sum_{i_2,\ldots,i_a\neq i} v_{i_2} \cdots v_{i_a} - \binom{b-2}{2} \sum_{i_3,\ldots,i_a\neq i} v_{i_3} \cdots v_{i_a} \sum_{j\neq i} (v_j^{(r)})^2 \right] \\
&= \frac{1}{(N)_b} \sum_{i=1}^{N} (v_i^{(r)})_2 \left[(N-v_i)^{b-2} - \binom{b-2}{2} \sum_{j\neq i} (v_j^{(r)})^2 N^{b-4} \right] \qquad \text{using (4)} \\
&= \frac{1}{(N)_b} \sum_{i=1}^{N} (v_i^{(r)})_2 (N-v_i)^{b-2} - \frac{1}{(N)_b} \binom{b-2}{2} \sum_{i\neq j=1}^{N} (v_i^{(r)})_2 (v_j^{(r)})^2 N^{b-4} \\
&\geq \frac{1}{(N)_b} \sum_{i=1}^{N} (v_i^{(r)})_2 (N)^{b-2} - \frac{1}{(N)_b} (b-2) \sum_{i=1}^{N} (v_i^{(r)})_2 v_i^{(r)} N^{b-3} \\
&\qquad - \frac{1}{(N)_b} \binom{b-2}{2} \sum_{i\neq j=1}^{N} (v_i^{(r)})_2 (v_j^{(r)})^2 N^{b-4} \qquad \text{using (5)} \\
&\geq \frac{1}{(N)_2} \sum_{i=1}^{N} (v_i^{(r)})_2 - \frac{(b-2)N^{b-3}}{(N)_b} \sum_{i=1}^{N} (v_i^{(r)})_2 v_i^{(r)} \\
&\qquad - \frac{N^{b-4}}{(N)_b} \binom{b-2}{2} \sum_{i\neq j=1}^{N} (v_i^{(r)})_2 (v_j^{(r)})^2 \\
&= \frac{1}{(N)_2} \sum_{i=1}^{N} (v_i^{(r)})_2 - \left[\frac{(b-2)}{N^3} + O(N^{-4}) \right] \sum_{i=1}^{N} (v_i^{(r)})_2 v_i^{(r)} \\
&\qquad - \left[\frac{1}{N^4} \binom{b-2}{2} + O(N^{-5}) \right] \sum_{i\neq j=1}^{N} (v_i^{(r)})_2 (v_j^{(r)})^2 \qquad \text{using (8), (9)} \quad (13) \\
&= c_r + o(c_r) \qquad \text{by (3),(C),(D)}
\end{aligned}
$$

Hence in this case $p_{\xi\eta}(r) = c_r + o(c_r)$. The inequality (12) is obtained by bounding the number of configurations with distinct parents by the the number of configurations with not necessarily distinct parents minus the number with at least one pair of parents chosen indistinctly. This leaves us with only the distinct-parents configurations since all indistinct choices must necessarily have a pair of parents chosen indistinctly, and the inequality arises from double-counting.

Case 2. η is obtained from ξ by merging three or more lineages into one, possibly as well as other simultaneous mergers,

i.e. $b_1 \geq 3$.

$$
\begin{aligned}
p_{\xi\eta}(r) &= \frac{1}{(N)_b} \sum_{\substack{i_1,\ldots,i_a=1 \\ \text{distinct}}}^{N} (v_{i_1}^{(r)})_{b_1} (v_{i_2}^{(r)})_{b_2} \cdots (v_{i_a}^{(r)})_{b_a} \\
&\leq \frac{1}{(N)_b} \sum_{i=1}^{N} (v_i^{(r)})_{b_1} \sum_{i_2,\ldots,i_a=1}^{N} (v_{i_2}^{(r)})_{b_2} \cdots (v_{i_a}^{(r)})_{b_a} \\
&\leq \frac{1}{(N)_b} \sum_{i=1}^{N} (v_i^{(r)})_{b_1} (v_1^{(r)} + \cdots + v_a^{(r)})^{b_2+\cdots+b_a} && \text{using (4)} \\
&\leq \frac{1}{(N)_b} \sum_{i=1}^{N} (v_i^{(r)})_{b_1} N^{b-3} && \text{since } b_2 + \cdots + b_a \leq b-3 \qquad (14) \\
&\leq \frac{N^{b-3}}{(N)_b} \sum_{i=1}^{N} (v_i^{(r)})_2 (v_i^{(r)})^{b_1-2} && \text{using (7)} \qquad (15) \\
&= o(c_r) && \text{using (C)}
\end{aligned}
$$

Case 3. η is obtained from ξ via two or more pair mergers, with no merges of more than two lineages, i.e. $2 = b_1 = b_2 \geq b_3 \geq \cdots \geq b_a \geq 1$.

$$
\begin{aligned}
p_{\xi\eta}(r) &= \frac{1}{(N)_b} \sum_{\substack{i_1,\ldots,i_a=1 \\ \text{distinct}}}^{N} (v_{i_1}^{(r)})_2 (v_{i_2}^{(r)})_2 (v_{i_3}^{(r)})_{b_3} \cdots (v_{i_a}^{(r)})_{b_a} \\
&\leq \frac{1}{(N)_b} \sum_{i=1}^{N} (v_i^{(r)})_2 \sum_{j=1}^{N} (v_j^{(r)})_2 \sum_{i_3,\ldots,i_a=1}^{N} (v_{i_3}^{(r)})_{b_3} \cdots (v_{i_a}^{(r)})_{b_a} && \text{dropping distinctness} \\
&\leq \frac{1}{(N)_b} \sum_{i=1}^{N} (v_i^{(r)})_2 \sum_{j=1}^{N} (v_j^{(r)})^2 \sum_{i_3,\ldots,i_a=1}^{N} (v_{i_3}^{(r)})_{b_3} \cdots (v_{i_a}^{(r)})_{b_a} && \text{using (6)} \\
&\leq \frac{1}{(N)_b} \sum_{i,j=1}^{N} (v_i^{(r)})_2 (v_j^{(r)})^2 (v_1^{(r)} + \cdots + v_N^{(r)})^{b_3+\cdots+b_a} \\
&= \frac{N^{b-4}}{(N)_b} \sum_{i,j=1}^{N} (v_i^{(r)})_2 (v_j^{(r)})^2 && \qquad (16) \\
&= o(c_r) && \text{using (D)}
\end{aligned}
$$

Case 4. $\eta = \xi$, i.e. $b_1 = \cdots = b_a = 1$, and $a = b$.

$$
\begin{aligned}
p_{\xi\xi}(r) &= \frac{1}{(N)_a} \sum_{\substack{i_1,\ldots,i_a=1 \\ \text{distinct}}}^{N} (v_{i_1}^{(r)})_1 \cdots (v_{i_a}^{(r)})_1 \\
&= \frac{1}{(N)_a} \sum_{\substack{i_1,\ldots,i_a=1 \\ \text{distinct}}}^{N} v_{i_1}^{(r)} \cdots v_{i_a}^{(r)} \\
&= \frac{1}{(N)_a} \left[\sum_{i_1,\ldots,i_a=1}^{N} v_{i_1}^{(r)} \cdots v_{i_a}^{(r)} - \binom{a}{2} \sum_{j=1}^{N} (v_j^{(r)})^2 \sum_{i_3,\ldots,i_a=1}^{N} v_{i_3}^{(r)} \cdots v_{i_a}^{(r)} \right] \qquad (17) \\
&= \frac{1}{(N)_a} \left[(v_1^{(r)} + \cdots + v_N^{(r)})^a - \binom{a}{2} \sum_{i=1}^{N} (v_i^{(r)})^2 (v_1^{(r)} + \cdots + v_N^{(r)})^{a-2} \right] \qquad \text{using (4)} \\
&= \frac{1}{(N)_a} \left[N^a - \binom{a}{2} N^{a-2} \sum_{i=1}^{N} (v_i^{(r)})^2 \right] \\
&\geq 1 - \binom{a}{2} \frac{N^{a-2}}{(N)_a} \sum_{i=1}^{N} (v_i^{(r)})^2 \qquad \text{using (6)} \\
&= 1 - \binom{a}{2} \left[\frac{1}{(N)_2} + O(N^{-3}) \right] \sum_{i=1}^{N} (v_i^{(r)})^2 \qquad \text{using (8)} \qquad (18) \\
&= 1 - c_r + o(c_r)
\end{aligned}
$$

The equality (17) is obtained in the same way as (12), with no double-counting. We have now shown that the only coalescence events having positive probability in the limit $N \to \infty$ are staying the same (**Case 4.**) or merging a single pair of lineages (**Case 1.**). All other possibilities have asymptotic probability $o(c_r)$.

It remains to show that the finite-dimensional distributions converge to those of the Kingman coalescent. Because the processes considered are Markov even when viewed as coalescing backwards in time, it suffices to show that the generators of the process converge to the generators of the Kingman coalescent (this will no longer be the case for the processes considered in **?**). □

For the argument of **?**, a different form is needed for the upper bound on triple mergers (**Case 2.**). Starting from (14), we obtain:

$$
\begin{aligned}
p_{\xi\eta}(r) &\leq \frac{1}{(N)_b} \sum_{i=1}^{N} (v_i^{(r)})_{b_1} N^{b-3} \\
&\leq \frac{N^{b-3}}{(N)_b} \sum_{i=1}^{N} (v_i^{(r)})^{b_1} \\
&= \left[\frac{1}{N^3} + O(N^{-4}) \right] \sum_{i=1}^{N} (v_i^{(r)})^{b_1} \qquad (19)
\end{aligned}
$$