Coalescent Processes With Simultaneous Multiple Collisions of Ancestral Lineages

Martin Möhle

University of Tübingen

Outline

1. An Exchangeable Population Model
2. Looking backward in time
   - Ancestral Process
   - Convergence to the Kingman Coalescent
   - Coalescent with Simultaneous and Multiple Collisions
   - Classification of Coalescent Processes
3. Looking forward in time
   - Forward Process
   - Convergence to the Wright-Fisher Diffusion
   - Other Forward Limiting Processes
   - Characterization of Forward Limiting Processes
4. Duality
5. Remarks, Discussion

An Exchangeable Population Model (Cannings)

Non-overlapping generations $r = \ldots, -1, 0, 1, \ldots$

Population size N, i.e.
N genes (or individuals) in each generation

$\nu_i^{(r)}$ = number of offspring of gene i in generation r

$$\sum_{i=1}^{N} \nu_i^{(r)} = N$$

No mutation, no selection (neutrality)

Additional Assumptions for the Offspring

Exchangeability: $(\nu_{\pi 1}^{(r)}, \ldots, \nu_{\pi N}^{(r)}) \overset{d}{=} (\nu_1^{(r)}, \ldots, \nu_N^{(r)}) \ \forall\ \pi, r$

Homogeneity: $(\nu_1^{(r)}, \ldots, \nu_N^{(r)})_r$ i.i.d.

Avoid the trivial model ($\nu_1^{(r)} \equiv 1$)

For convenience: $(\nu_1, \ldots, \nu_N) := (\nu_1^{(0)}, \ldots, \nu_N^{(0)})$

Classical Examples:
Wright-Fisher: $(\nu_1, \ldots, \nu_N) \overset{d}{=} \text{Multinomial}(N, \frac{1}{N}, \ldots, \frac{1}{N})$
Moran: $(\nu_1, \ldots, \nu_N)$ = random permutation of $(2, 0, \underbrace{1, \ldots, 1}_{N-2})$

Backward Process

Sample n ($\leq N$) genes from generation 0. For $r = 0, 1, \ldots$ define

$$\boxed{\mathcal{D}_r \ := \# \text{ of } \textbf{ancestors } r \text{ generations } \textbf{backward} \text{ in time}}$$

Backward process $\mathcal{D} := (\mathcal{D}_r)_{r=0,1,\ldots}$ is a time-homogeneous Markovian death process.

state space:	$\{0, \ldots, n\}$
initial state:	n
absorbing states:	1 (and 0)
transition matrix:	$P_N = (p_{ba})$ with entries

$$\begin{aligned} p_{ba} &:= P(\mathcal{D}_r = a \,|\, \mathcal{D}_{r-1} = b) \\ &= \frac{\binom{N}{a}}{\binom{N}{b}} \sum_{\substack{b_1,\ldots,b_a \in \mathbb{N} \\ b_1+\cdots+b_a = b}} \mathrm{E}\left(\binom{\nu_1}{b_1} \cdots \binom{\nu_a}{b_a}\right) \\ &= \frac{b!}{a!} \sum_{\substack{b_1,\ldots,b_a \in \mathbb{N} \\ b_1+\cdots+b_a = b}} \frac{\Phi_a(b_1, \ldots, b_a)}{b_1! \cdots b_a!} \end{aligned}$$

with

$$\boxed{\Phi_a(b_1, \ldots, b_a) \ := \ \frac{(N)_a}{(N)_b} \mathrm{E}\Big((\nu_1)_{b_1} \cdots (\nu_a)_{b_a}\Big)}$$

(probability for a $(b_1, \ldots, b_a)$-merger)

Notation: $(x)_k := x(x-1) \cdots (x-k+1)$

Ancestral Process

Sample n ($\leq N$) genes from generation 0. For $r = 0, 1, \ldots$ define a random relation $\mathcal{R}_r = {}_{(N)}\mathcal{R}_r^{(n)}$ on $\{1, \ldots, n\}$ via

$$\boxed{(i,j) \in \mathcal{R}_r \quad :\Longleftrightarrow \quad \begin{array}{l}\text{the chosen genes } i \text{ and } j\\ \text{have a common ancestor}\\ r \text{ generation backward in time}\end{array}}$$

Ancestral process $\mathcal{R} := (\mathcal{R}_r)_{r=0,1,\ldots}$ is a time-homogeneous Markov chain.

state space:	$\mathcal{E}_n :=$ set of equivalence relations on $\{1, \ldots, n\}$
initial state:	$\Delta := \{(i,i) \,\vert\, 1 \leq i \leq n\}$
absorbing state:	$\Theta := \{(i,j) \,\vert\, 1 \leq i, j \leq n\}$
typical path:	$\Delta = \mathcal{R}_0 \subseteq \mathcal{R}_1 \subseteq \cdots \subseteq \mathcal{R}_r \subseteq \Theta$
transition matrix:	$P_N := (p_{\xi\eta})_{\xi,\eta \in \mathcal{E}_n}$ with entries

$$\boxed{p_{\xi\eta} \;:=\; P(\mathcal{R}_r = \eta \,|\, \mathcal{R}_{r-1} = \xi) \;=\; \Phi_a(b_1, \ldots, b_a)} \quad \text{for } \xi \subseteq \eta$$

$a := |\eta| :=$ number of equivalence classes of η

$b := |\xi| :=$ number of equivalence classes of ξ $(= b_1 + \cdots + b_a)$

$b_1, \ldots, b_a =$ group sizes of merging classes of η

Kingman Coalescent

Latin: coalescere = to join, to unite, to combine

Definition:

The Kingman n-coalescent $(R_t)_{t\geq 0}$ is a time homogeneous Markov process with state space $\mathcal{E}_n$ and transition rates

$$q_{\xi\eta} := \begin{cases} -\frac{1}{2}k(k-1) & \text{if } \xi = \eta \\ 1 & \text{if } \xi \prec \eta \\ 0 & \text{otherwise} \end{cases}$$

where

$$k := |\xi| := \text{number of equivalence classes of } \xi$$

$$\begin{aligned} \xi \prec \eta \; &:\Longleftrightarrow \; \xi \subseteq \eta \text{ and } |\eta| = |\xi| - 1 \\ &\Longleftrightarrow \; \eta \text{ is obtained by merging two classes of } \xi \end{aligned}$$

Remark: $D_t := |R_t|$ defines a pure death process $(D_t)_{t\geq 0}$ with rates $k(k-1)/2$.

Convergence Result (Kingman 1982)

Assumptions:

- Limiting Condition: $\sigma_N^2 := \mathrm{Var}(\nu_1) \xrightarrow[N\to\infty]{} \sigma^2 \in (0,\infty)$
- Moment Condition: $\sup\limits_N \mathrm{E}(\nu_1^k) < \infty$ for all $k \in I\!N$

Then as $N \to \infty$

$$\boxed{(\mathcal{R}_{[N\sigma^{-2}t]})_{t\geq 0} \implies (R_t)_{t\geq 0}}$$

where $(R_t)_{t\geq 0}$ denotes the Kingman coalescent

Remark: Result is applicable for the Wright-Fisher model as

$$\begin{aligned} \mathrm{Var}(\nu_1) &= 1 - \tfrac{1}{N} \\ \mathrm{E}((\nu_1)_k) &= (1-\tfrac{1}{N})(1-\tfrac{2}{N})\cdots(1-\tfrac{k-1}{N}) \;\leq\; 1\,, \end{aligned}$$

but not for the Moran model ($\mathrm{Var}(\nu_1) = 2/N$).

Convergence to Kingman's n-coalescent

Coalescence Probability

$$
\begin{aligned}
c_N &:= P(\text{Two individuals have same parent}) \\
&= \Phi_1(2) \;=\; \frac{\mathrm{E}((\nu_1)_2)}{N-1} \;>\; 0
\end{aligned}
$$

Theorem: (M. and Sagitov)
The following conditions are equivalent:

(i) $\boxed{(\mathcal{R}_{[t/c_N]})_{t\geq 0} \Longrightarrow (R_t)_{t\geq 0}}$

(convergence to the Kingman coalescent in the Skorohod topology)

(ii) $\boxed{\lim_{N\to\infty} \frac{P_N - I}{c_N} = Q}$

(convergence to the generator of the Kingman coalescent)

(iii) $\boxed{\phi_1(3) := \lim_{N\to\infty} \frac{\Phi_1(3)}{c_N} = 0}$

(triple mergers of ancestral lines are asymptotically negligible in comparison with binary mergers)

(iv) $\boxed{\lim_{N\to\infty} c_N\,\mathrm{E}(T_{n,N}) = \mathrm{E}(T_n) = 2(1-\tfrac{1}{n}) \;\forall\; n \in I\!N}$

$T_{n,N} := \inf\{r \in I\!N_0 \,|\, \mathcal{D}_r = 1, \mathcal{D}_0 = n\}$

$T_n := \inf\{t > 0 \,|\, D_t = 1, D_0 = n\}$

(convergence of the scaled number of generations back to the most recent common ancestor)

Coalescent with Simultaneous Multiple Collisions

There exists a subsequence $(N_k)_{k\in I\!N}$, $\lim_{k\to\infty} N_k = \infty$ such that

$$c \ := \ \lim_{k\to\infty} c_{N_k} \quad \text{and} \quad Q \ := \ \lim_{k\to\infty} \frac{P_{N_k} - I}{c_{N_k}}$$

exist.

a) If $c > 0$ then

$$\boxed{({}_{(N_k)}\mathcal{R}_r)_{r\in I\!N_0} \ \Longrightarrow \ (R_r)_{r\in I\!N_0}} \quad \text{(discrete time)}$$

Initial state: $R_0 = \Delta$, Transition matrix: $I + c\,Q$

b) If $c = 0$ then

$$\boxed{({}_{(N_k)}\mathcal{R}_{[t/c_{N_k}]})_{t\geq 0} \ \Longrightarrow \ (R_t)_{t\geq 0}} \quad \text{(continuous time)}$$

Initial state: $R_0 = \Delta$, Transition matrix: e^{tQ}

Criterion: $Q = \lim_{N\to\infty} \frac{P_N - I}{c_N}$ exists iff all the limits

$$\boxed{\phi_a(b_1, \ldots, b_a) \ := \ \lim_{N\to\infty} \frac{\Phi_a^{(N)}(b_1, \ldots, b_a)}{c_N}}$$

$a \in I\!N$, $b_1 \geq \cdots \geq b_a \geq 2$, exist.

From now on assume that Q exists.

Characterization of Q

$$\boxed{q_{\xi\eta} = \begin{cases} -\sum_{j=1}^{b-1} j\phi_j(2,1,\ldots,1) & \text{if } \xi = \eta, \\ \phi_a(b_1,\ldots,b_a) & \text{if } \xi \subset \eta, \\ 0 & \text{otherwise,} \end{cases}}$$

$b := |\xi|$, $a := |\eta|$, $b_1, \ldots, b_a =$ group sizes of merging classes of η.

Lemma: There exist measures F_a, $a \in I\!N$, on the simplex

$$\Delta_a := \{(y_1,\ldots,y_a) \in [0,1]^a \,|\, y_1 + \cdots + y_a \leq 1\}$$

uniquely determined via their moments

$$\boxed{\int_{\Delta_a} x_1^{b_1-2} \cdots x_a^{b_a-2} F_a(dx_1,\ldots,dx_a) = \phi_a(b_1,\ldots,b_a),}$$

$b_1 \geq \cdots \geq b_a \geq 2$

Recursion:

$$\phi_{a+1}(b_1,\ldots,b_a,1) = \phi_a(b_1,\ldots,b_a) - \sum_{i=1}^{a} \phi_a(b_1,\ldots,b_i+1,\ldots,b_a)$$

$\Rightarrow$ Integral representation for the rates $q_{\xi\eta}$ in terms of the measures $F_1, F_2, \ldots$ (explicit formula available, but nasty).

Example: Only Multiple Collisions (F-coalescent)

- Simplest example which is more complicated than the Kingman coalescent
- Independently studied by Pitman (1999) and Sagitov (1999)

Assume that $\boxed{\phi_2(2,2) \; = \; 0}$.

$\Rightarrow$ No simultaneous mergers, only multiple mergers with rates

$$\boxed{q_{\xi\eta} \; = \; \int_{[0,1]} x^{b-a-1}(1-x)^{a-1} \, F(dx)}$$

Measure F uniquely determined via the moments

$$\int_{[0,1]} x^{k-2} \, F(dx) \; = \; \phi_1(k).$$

Examples:
$F = \delta_0$ (point measure at 0). $\Rightarrow$ Kingman coalescent
$F = U_{[0,1]}$ (uniform on $[0,1]$). $\Rightarrow$ Bolthausen-Sznitman coalescent

Forward Process

Sample n ($\leq N$) genes from generation 0. For $r = 0, 1, \ldots$ define

$$\boxed{\mathcal{X}_r \;:=\; \#\text{ of } \textbf{descendants } r \text{ generations } \textbf{forward} \text{ in time}}$$

Forward process $\mathcal{X} := (\mathcal{X}_r)_{r=0,1,\ldots}$ is a time-homogeneous Markov chain and a martingale.

state space:	$\{0, \ldots, N\}$
initial state:	n
absorbing states:	0 and N
transient states:	$1, \ldots, N-1$ (provided $P(\nu_1 = 1) < 1$)
transition matrix:	$\Pi_N = (\pi_{jk})$ with entries

$$\boxed{\pi_{jk} \;:=\; P(\mathcal{X}_r = k \,|\, \mathcal{X}_{r-1} = j) \;=\; P(\nu_1 + \cdots + \nu_j = k)}$$

(π_{jk} depends on the population size N)

Convergence to the Wright-Fisher diffusion

Notation: $(X_t)_{t\geq 0}$ Wright-Fisher diffusion with generator

$$Af(x) := \frac{x(1-x)}{2}f''(x), \qquad f \in C^2([0,1]).$$

Further define

$$A_N f(x) := \frac{T_N f(x) - f(x)}{c_N},$$

where

$$T_N f(x) \;:=\; \mathrm{E}\Big(f\Big(\frac{\mathcal{X}_{r+1}}{N}\Big)\,\Big|\,\frac{\mathcal{X}_r}{N} = x\Big)$$

Theorem:
The following conditions are equivalent:

(i) $\boxed{\left(\frac{\mathcal{X}_{[t/c_N]}}{N}\right)_{t\geq 0} \Longrightarrow (X_t)_{t\geq 0}}$

(convergence to the Wright-Fisher diffusion in the Skorohod topology)

(ii) $\lim\limits_{N\to\infty} \sup\limits_{x\in\{0,\frac{1}{N},\ldots,1\}} |A_N f(x) - Af(x)| = 0$ for all polynomials f

(uniform convergence to the generator of the Wright-Fisher diffusion)

(iii) $\boxed{\phi_1(3) := \lim\limits_{N\to\infty} \frac{\Phi_1(3)}{c_N} = 0}$

The dual generator

Criterion:
All the limits $\phi_a(b_1, \ldots, b_a) := \lim_{N\to\infty} \frac{\Phi_a(b_1, \ldots, b_a)}{c_N}$ exist
($a \in I\!N$, $b_1 \geq \cdots \geq b_a \geq 2$)

Available recursion:

$$\Phi_{a+1}(b_1, \ldots, b_a, 1) \;=\; \Phi_a(b_1, \ldots, b_a) - \sum_{i=1}^{a} \Phi_a(b_1, \ldots, b_i+1, \ldots, b_a)$$

Conclusions:

- $\phi_a(b_1, \ldots, b_a)$ exists whenever $b_1 + \cdots + b_a > a$
- $\phi_a(1, \ldots, 1) := \lim_{N\to\infty} \frac{\Phi_a(1, \ldots, 1) - 1}{c_N}$ exists

Define a linear **operator** A on the set of all polynomials via

$$\boxed{\quad Af_i \;:=\; \sum_{j=1}^{i} g_{ij} f_j \qquad i = 0, 1, 2, \ldots, \quad}$$

where $f_i(x) := x^i$ and $g_{ij} := \frac{j!}{i!} \sum_{\substack{i_1, \ldots, i_j \in I\!N \\ i_1 + \cdots + i_j = i}} \frac{\phi_j(i_1, \ldots, i_j)}{i_1! \cdots i_j!}$

Proposition: Closure $\bar{A}$ of A generates a strongly continuous contraction semigroup $\{T(t)\}$ on $C([0,1])$ (Hille-Yosida).

Examples:

1. $\phi_1(3) = 0$. Then for $f \in C^2([0,1])$
$$Af(x) \;=\; \frac{x(1-x)}{2} f''(x)$$
(generator of the Wright-Fisher diffusion)

2. $\phi_1(3) > 0$ and $\phi_2(2,2) = 0$. Then for $f \in C^2([0,1])$
$$\boxed{Af(x) \;=\; \int_{[0,1]} \frac{xf(x+u-ux) + (1-x)f(x-ux) - f(x)}{u^2} \, F(du)}$$
(generator of the dual process of the F-coalescent)

Measure F uniquely determined via its moments
$$\int_{[0,1]} u^{k-2} \, F(du) \;=\; \phi_1(k), \qquad k = 2, 3, \ldots$$

Remark: For $F = \mathrm{Beta}(2-\alpha, \alpha)$, $0 < \alpha < 2$, the generator A arises also as that of the ratio process (properly time-changed)
$$\left(\frac{X_t}{X_t + Y_t} \right)_t .$$
$(X_t)_t, (Y_t)_t :=$ two independent copies of a continuous mass stable branching process $dX_t = X_{t-}^{1/\alpha} dZ_t$.
$Z =$ stable process with Lévy-measure $\nu(dh) = Ch^{-1-\alpha} dh$.

Convergence of the forward process

Assumptions:

(i) $c := \lim_{N\to\infty} c_N$ exists

(ii) $\mathcal{X}_0/N \Longrightarrow \nu$

(iii) All the limits $\phi_a(b_1,\ldots,b_a) := \lim_{N\to\infty} \dfrac{\Phi_a(b_1,\ldots,b_a)}{c_N}$ exist

$(a \in I\!N,\ b_1 \geq \cdots \geq b_a \geq 2)$

a) If $c > 0$ then

$$\boxed{\left(\frac{\mathcal{X}_r}{N}\right)_{r=0,1,\ldots} \Longrightarrow X := (X_r)_{r=0,1,\ldots}} \quad \text{(discrete time)}$$

Initial distribution: $P_{X_0} = \nu$

Generator: A

$(Tf(x) := \mathrm{E}(f(X_{r+1})\,|\,X_r = x)$ given via $T := I + cA)$

b) If $c = 0$ then

$$\boxed{\left(\frac{\mathcal{X}_{[t/c_N]}}{N}\right)_{t\geq 0} \Longrightarrow X := (X_t)_{t\geq 0}} \quad \text{(continuous time)}$$

Initial distribution: $P_{X_0} = \nu$

Generator: A

Limiting process X is a martingale, converges a.s., absorbing states 0 and 1

Duality

$X = (X_t)_{t \in T}$ Markov process with state space E_1

$Y = (Y_t)_{t \in T}$ Markov process with state space E_2

Definition: (Liggett 1985)
X is **dual** to Y with respect to some function $H : E_1 \times E_2 \to I\!R$ if

$$\boxed{E^x H(X_t, y) \;=\; E^y H(x, Y_t) \quad \forall\, x \in E_1, y \in E_2, t \in T}$$

Convergence and Duality

Commutative diagram

	forward in time		backward in time
finite population size	$\boxed{\text{forward process } \mathcal{X}}$	$\overset{(1)}{\longleftrightarrow}$	$\boxed{\text{backward process } \mathcal{D}}$
	(2) $\downarrow$		$\downarrow$ (3)
infinite population size	$\boxed{\text{limiting process } X}$	$\underset{(4)}{\longleftrightarrow}$	$\boxed{\text{death process } D}$

(1) = **duality** of $\mathcal{X}$ and $\mathcal{D}$ with respect to $H_N(i,k) = \binom{i}{k}/\binom{N}{k}$

(2) = **convergence** in $D_{[0,1]}([0,\infty))$

(3) = **convergence** in $D_{\{0,1,\ldots\}}([0,\infty))$

(4) = **duality** of X and D with respect to $H(x,k) = x^k$