

## Inference for $\Lambda$ -coalescents

Matthias Steinrücken

joint work with Matthias Birkner (Bonn)  
and Jochen Blath (Berlin)



## Introduction

- high variance in offspring distribution observed in certain marine species (HEDGECOCK '94, ÁRNASON '04)
- Hedgecock sweepstake: sometimes one individual replaces substantial fraction of whole population
- ELDON & WAKELEY '06: Kingman coalescent does not describe genealogy well, even in neutral situations

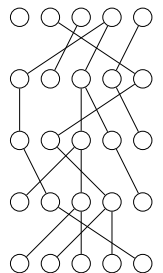
## Outline

- models for extreme reproduction events
  - ▶ lead to  $\Lambda$ -coalescents
- inference methods
- discuss suitable subclass
- apply method to Atlantic cod datasets

## Neutral Population model

### Cannings model

- non-overlapping generations of fixed size  $N$
- exchangeable offspring vector  $(\nu_1, \dots, \nu_N)$



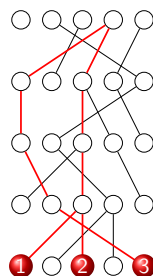
## Neutral Population model

### Cannings model

- non-overlapping generations of fixed size  $N$
- exchangeable offspring vector  $(\nu_1, \dots, \nu_N)$

### Ancestral process

- take sample of size  $n$  from present generation
- follow ancestral lines backwards in time
- sequence of partitions of  $\{1, \dots, n\}$  ( $= \mathcal{P}^{(n)}$ )
- $i$  and  $j$  in the same block if they have common ancestor



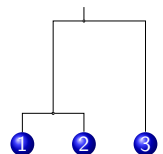
## Standard-null model: Kingman coalescent

If  $\sup_N \mathbb{E} \nu_1^k < \infty$ ,  $k \geq 1$  (e.g. Moran, Wright-Fisher), then ancestral process (rescaled by  $N$ ) converges to

### Kingman coalescent (KINGMAN '82)

$R_t^{(n)}$  is a time continuous Markov chain taking values in  $\mathcal{P}^{(n)}$

- each pair of blocks coalesce with rate 1
- only binary mergers allowed



## $\Lambda$ -coalescent

individual replaces substantial fraction of whole population

- ▶ multiple merger in limiting genealogy

### $\Lambda$ -coalescent (PITMAN '99/SAGITOV '99)

$\Pi_t^{(n)}$  is continuous time Markov chain with values in  $\mathcal{P}^{(n)}$

- $b$  blocks:  $k$  coalesce to form single block with rate  $\lambda_{b,k}$
- rates have to be consistent:  $\lambda_{b,k} = \lambda_{b+1,k} + \lambda_{b+1,k+1}$

### Theorem (PITMAN '99)

consistent rates  $\longleftrightarrow$  finite measure  $\Lambda$  on  $[0, 1]$

$$\lambda_{b,k} = \int_0^1 x^{k-2} (1-x)^{b-k} \Lambda(dx)$$

### Theorem (SCHWEINSBERG '03)

Let  $c_N := \frac{\mathbb{E}[(\nu_1)_2]}{N-1}$ . The ancestral process rescaled by  $c_N^{-1}$  converges to the  $\Lambda$ -coalescent, if

- 1  $\lim_{N \rightarrow \infty} c_N = 0$ ,
- 2  $\lim_{N \rightarrow \infty} N^{-2} c_N^{-1} \mathbb{E}[(\nu_1)_2 (\nu_2)_2] = 0$ ,
- 3  $\lim_{N \rightarrow \infty} N c_N^{-1} \mathbb{P}\{\nu_1 > Nx\} = \int_x^1 y^{-2} \Lambda(dy)$ .

Remark:

- includes Kingman coalescent:  $\Lambda = \delta_0$
- $\Lambda$ -coalescent  $\Pi_t$  taking values in  $\mathcal{P}^{(N)}$  can be constructed, dual to generalised Fleming-Viot process (BERTOIN & LE GALL '03)

## Mutation

- assumed to be neutral
- infinitely many sites (sample size  $\ll$  nr. of sites)
- every mutation occurs on a different site (completely new)

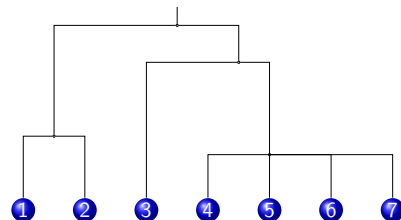
simulate sample:

## Mutation

- assumed to be neutral
- infinitely many sites (sample size  $\ll$  nr. of sites)
- every mutation occurs on a different site (completely new)

simulate sample:

- 1 take a realisation of  $\Pi_t^n$

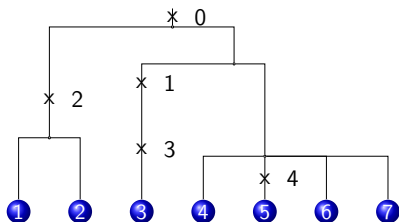


## Mutation

- assumed to be neutral
- infinitely many sites (sample size  $\ll$  nr. of sites)
- every mutation occurs on a different site (completely new)

simulate sample:

- 1 take a realisation of  $\Pi_t^n$
- 2 place mutations along branches with rate  $r$

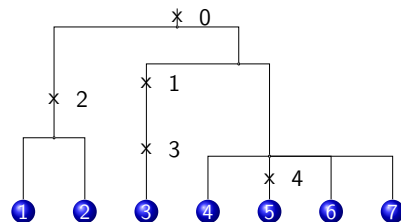


## Mutation

- assumed to be neutral
- infinitely many sites (sample size  $\ll$  nr. of sites)
- every mutation occurs on a different site (completely new)

simulate sample:

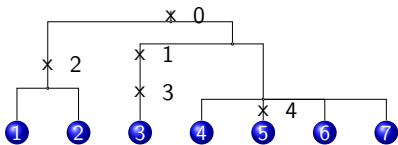
- 1 take a realisation of  $\Pi_t^n$
- 2 place mutations along branches with rate  $r$



$$(\mathbf{t}, \mathbf{n}) = \left( ((0, 2), (0, 1, 3), (0), (0, 4)), (2, 1, 3, 1) \right)$$

### Probability of a configuration

BIRKNER & BLATH '08:



$$\begin{aligned}
 p(\mathbf{t}, \mathbf{n}) = & \frac{1}{\lambda_n + nr} \sum_{i: n_i \geq 2} \sum_{k=2}^{n_i} \binom{n_i}{k} \lambda_{n,k} \frac{n_i - k + 1}{n - k + 1} p(\mathbf{t}, \mathbf{n} - (k-1)\mathbf{e}_i) \\
 & + \frac{r}{\lambda_n + nr} \sum_{\substack{i: n_i=1, x_{i0} \text{ unique,} \\ s(x_i) \neq x_j \forall j}} p(s_i(\mathbf{t}), \mathbf{n}) \\
 & + \frac{r}{\lambda_n + nr} \sum_{\substack{i: n_i=1, \\ x_{i0} \text{ unique}}} \sum_{j: s(x_k)=x_j} (n_j + 1) p(\mathbf{t}_i(\mathbf{t}), \mathbf{v}_i(\mathbf{n} + \mathbf{e}_j))
 \end{aligned}$$

### Probabilities for large datasets

- Recursion can be calculated exactly only for small datasets
- other method: Monte-Carlo based on Markov chain on space of configurations

### Beta-coalescent

number of Λ-coalescents is vast

- ▶ Which are relevant?

$$\Lambda \sim \text{Beta}(2 - \alpha, \alpha), \quad \alpha \in [1, 2)$$

SCHWEINSBERG '03:

- $\mathbb{P}(X_i \geq k) \sim ck^{-\alpha}$ , next gen.: choose  $N$  uniformly
- $\alpha = 2$  corresponds to Kingman case
- $\alpha$  shifts from binary to multiple merging

BBCEMSW '05: close relation between  $\alpha$ -stable branching and Beta-coalescent

### Beta-coalescent

number of Λ-coalescents is vast

- ▶ Which are relevant?

$$\Lambda \sim \text{Beta}(2 - \alpha, \alpha), \quad \alpha \in [1, 2)$$

SCHWEINSBERG '03:

- $\mathbb{P}(X_i \geq k) \sim ck^{-\alpha}$ , next gen.: choose  $N$  uniformly
- $\alpha = 2$  corresponds to Kingman case
- $\alpha$  shifts from binary to multiple merging

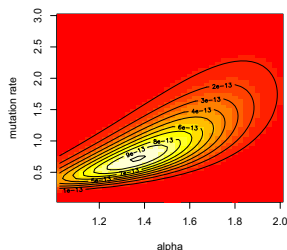
BBCEMSW '05: close relation between  $\alpha$ -stable branching and Beta-coalescent

one parameter family  $\Rightarrow$  inference possible !!!

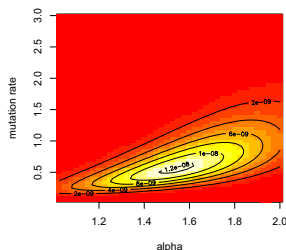
### Atlantic cod

ÁRNASON '04

- Combined Atlantic cod data from several authors
- 1278 samples, too big
- ▶ analyse small datasets



(a) Sigurgislason 2003



(b) Arnason 2000

### Atlantic cod

Dataset (size)	$(\hat{\alpha}, \hat{r})$
A96 (99)	(2.0, 0.75)
A98 (108)	(1.7, 0.85)
A00 (77)	(1.5, 0.6)
C91 (55)	(1.4, 0.75)
C95 (100)	(1.13, 0.35)
P93 (103)	(1.4, 0.65)
S03 (73)	(1.35, 0.7)

- sampled at different localities
- question of reliability for sample of size 100
- indicates extreme reproduction events

## Outlook

- improve method to treat large datasets
  - try importance sampling (STEPHENS & DONNELLY '00)
  - ...
- investigate statistical properties

Thank you for your attention!

## Monte Carlo

Markov chain  $(X_k)_{k \in \mathbb{N}}$  taking values in  $(\mathcal{T}, \mathcal{N})$

$$(\mathbf{t}, \mathbf{n}) \rightarrow \begin{cases} (s_k(\mathbf{t}), \mathbf{n}) & \text{with probability } \frac{r}{r_n f(\mathbf{t}, \mathbf{n})} \\ & \text{if } k : n_k = 1, x_{k0} \text{ unique } s_k(x_k) \neq x_j \forall j \\ (v_k(\mathbf{t}, v_k(\mathbf{n} + \mathbf{e}_j))) & \text{with probability } \frac{r(n_j + 1)}{r_n f(\mathbf{t}, \mathbf{n})} \text{ if } k : n_k = 1, x_{k0} \text{ unique} \\ (\mathbf{t}, \mathbf{n} - (k - 1)\mathbf{e}_i) & \text{with probability } \frac{1}{r_n f(\mathbf{t}, \mathbf{n})} \binom{n}{k} \lambda_{n,k} \frac{n_i - k + 1}{n - k + 1} \\ & \text{if } 2 \leq k \leq n \end{cases}$$

where

$$f(\mathbf{t}, \mathbf{n}) := \frac{1}{r_n} \left( \sum_{\substack{k: n_k = 1, x_{k0} \text{ unique,} \\ s_k(x_k) \neq x_j \forall j}} r + \sum_{\substack{k: n_k = 1, \\ x_{k0} \text{ unique}}} \sum_{j: s_k(x_k) = x_j} r(n_j + 1) \right. \\ \left. + \sum_{i: n_i \geq 2} \sum_{k=2}^{n_i} \binom{n}{k} \lambda_{n,k} \frac{n_i - k + 1}{n - k + 1} \right),$$

$$r_n := rn + \sum_{k=2}^n \binom{n}{k} \lambda_{n,k}$$

## Monte Carlo

### Lemma

If  $f$  and transition probabilities are chosen as before, then

$$p(\mathbf{t}, \mathbf{n}) = p((0), (1)) \mathbb{E}_{(\mathbf{t}, \mathbf{n})} \prod_{l=0}^{\tau-1} f(\mathbf{t}(l), \mathbf{n}(l)),$$

where  $\tau$  is hitting time of  $((0), (1))$  (the root)

- special case of lemma from GRIFFITHS & TAVARÉ '94
- simulate markov chain several times and take average  
 $\Rightarrow$  estimator of  $p(t, n)$