Serik Sagitov

28 May 2009, Luminy

Quenched effective population size

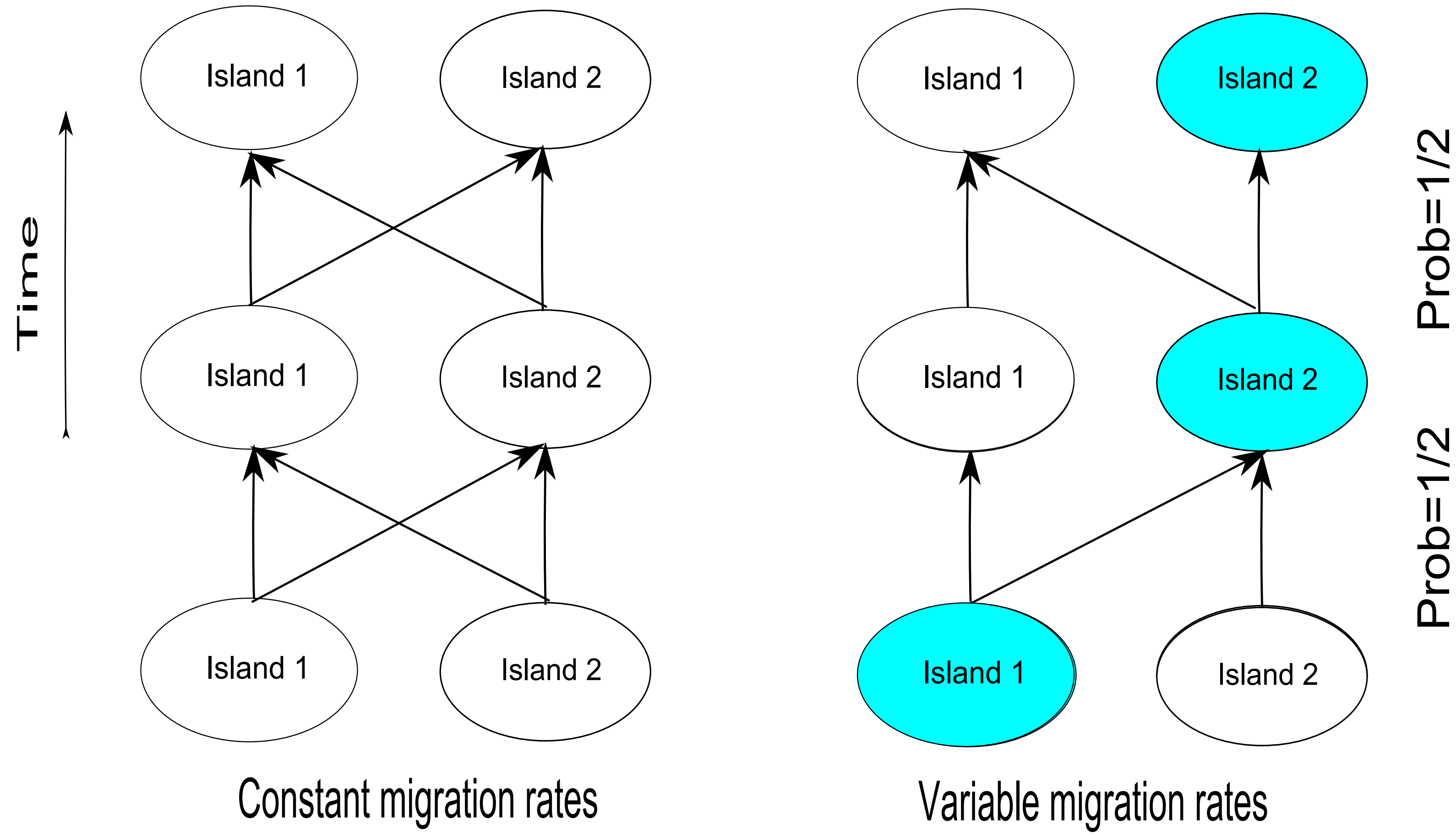
(work in progress jointly with P.Jagers and V.Vatutin)

Keywords

- Wright-Fisher model and Kingman's coalescent
- Coalescent effective population size N_e
- Geographically structured WFM with fast migration
- Randomly varying migration rates
- Markov chains with random transition matrices
- Random environment, $N_e^{[\text{quenched}]}$ and $N_e^{[\text{annealed}]}$

This work is supported by The Bank of Sweden Tercentenary Foundation

1 Motivating example



2 Coalescent effective population size

Wright-Fisher model: $\mathrm{Mn}(N; N^{-1}, \ldots, N^{-1})$ reproduction law.

Given $X(0) = n$, the ancestral process $X(t)$ is a Markov chain with a $n \times n$ transition matrix $\Pi = \Pi_N$.

Key decomposition: $\Pi = I + N^{-1}Q + O(N^{-2})$ with identity matrix I and

$$Q = \begin{pmatrix} 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 1 & -1 & 0 & \ldots & 0 & 0 & 0 \\ . & . & . & \ldots & . & . & . \\ 0 & 0 & 0 & \ldots & \binom{n-1}{2} & -\binom{n-1}{2} & 0 \\ 0 & 0 & 0 & \ldots & 0 & \binom{n}{2} & -\binom{n}{2} \end{pmatrix}$$

Convergence to the Kingman coalescent: $\Pi^{Nt} \to e^{tQ}$ as $N \to \infty$.

Kingman's coalescent is a robust approximation for $X(tN/c)$ in various population models. Coalescent $N_e = N/c$.

- Nordborg, M. and Krone, S. (2002) Separation of Time Scales and Convergence to the Coalescent in Structured Populations. Modern Developments in Theoretical Population Genetics, pp. 194-232, M. Slatkin and M. Veuille, editors. Oxford University Press.
- Sjödin P, Kaj I, Krone S, Lascoux M, Nordborg M (2005) On the meaning and existence of an effective population size. Genetics 169: 1061-1070.
- Jagers P. and Sagitov S. (2004) Convergence to the coalescent in populations of substantially varying size. J. Appl. Prob. 41, no. 2, 368-378.
- Sagitov S. and Jagers P. (2005) The coalescent effective size of age-structured populations. Ann. Appl. Probab. 15, 1778-1797.

Usually $c \geq 1$. Example of $c \leq 1$: offspring numbers $0, 1, 2$ with probabilities $(\alpha, 1-2\alpha, \alpha)$ imply $c = \sigma^2 = 2\alpha$.

3 Geographically structured WFM

$L \geq 2$ connected islands: migration and WF reproduction.

Subpopulations of constant sizes $N_1, \ldots, N_L$ with

$$N_1 + \ldots + N_L = N \text{ and } N_i/N \to a_i, \; N \to \infty$$

Ancestral process: lineages migrate independently over the islands until they merge according to the WFM rules of the hosting islands.

Configuration process of n lineages:

$$\mathbf{X}(t) = (X_1(t), \ldots, X_L(t))$$

$X_i(t)$ is the number of lineages located on the i-th island at t-th generation backward in time.

The total number of lineages $X(t) = X_1(t) + \ldots + X_L(t)$ is not a Markov process except for the "dummy islands" case.

$\mathbf{X}(t)$ is a Markov chain with a finite state space $S_1 \cup \ldots \cup S_n$, where S_r is the set of states $\mathbf{x}$ satisfying $x_1 + \ldots + x_L = r$.

The number of elements in S_r is $d_r = \binom{r+L-1}{r}$. The transition matrix $\mathbf{\Pi}$ of $\mathbf{X}(t)$ is of size $(d_1 + \ldots + d_n) \times (d_1 + \ldots + d_n)$.

Key decomposition

$$\mathbf{\Pi} = \mathbf{B}(\mathbf{I} + N^{-1}\mathbf{C}) + o(N^{-1}).$$

Backward migration probabilities $\mathbf{B} = \mathrm{diag}(\mathbf{B}_1, \ldots, \mathbf{B}_n)$, where $\mathbf{B}_r$ is the $(d_r \times d_r)$ transition matrix for non-coalescing r lineages.
Coalescent rates

$$\mathbf{C} = \begin{pmatrix} 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ \mathbf{C}_{21} & -\mathbf{C}_2 & 0 & \ldots & 0 & 0 & 0 \\ . & . & . & \ldots & . & . & . \\ 0 & 0 & 0 & \ldots & \mathbf{C}_{n-1,n-2} & -\mathbf{C}_{n-1} & 0 \\ 0 & 0 & 0 & \ldots & 0 & \mathbf{C}_{n,n-1} & -\mathbf{C}_n \end{pmatrix}$$

$\mathbf{C}_r = \mathrm{diag}(C(\mathbf{x}), \mathbf{x} \in S_r)$, where $C(\mathbf{x}) = \sum_{k=1}^{L} \frac{1}{a_k}\binom{x_k}{2}$

$\mathbf{C}_{r,r-1}$ has $\frac{1}{a_k}\binom{x_k}{2}$ at positions $(\mathbf{x}, \mathbf{x} - \mathbf{e}_k)$ and zeros elsewhere.

In particular, if $L = 2$, then $d_r = r + 1$ and

$$\mathbf{C}_r = \begin{pmatrix} \binom{r}{2}\frac{1}{a_1} & 0 & 0 & \dots & 0 & \dots & 0 & 0 \\ . & . & . & \dots & . & \dots & . & . \\ 0 & 0 & 0 & \dots & \binom{r-k}{2}\frac{1}{a_1} + \binom{k}{2}\frac{1}{a_2} & \dots & 0 & 0 \\ . & . & . & \dots & . & \dots & . & . \\ 0 & 0 & 0 & \dots & 0 & \dots & 0 & \binom{r}{2}\frac{1}{a_2} \end{pmatrix}$$

$$\mathbf{C}_{r,r-1} = \begin{pmatrix} \binom{r}{2}\frac{1}{a_1} & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \binom{r-1}{2}\frac{1}{a_1} & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{a_2} & \binom{r-2}{2}\frac{1}{a_1} & \dots & 0 & 0 & \dots & 0 \\ . & . & . & \dots & . & . & \dots & . \\ 0 & 0 & 0 & \dots & \binom{k}{2}\frac{1}{a_2} & \binom{r-k}{2}\frac{1}{a_1} & \dots & 0 \\ . & . & . & \dots & . & . & \dots & . \\ 0 & 0 & 0 & \dots & 0 & 0 & \dots & \binom{r}{2}\frac{1}{a_2} \end{pmatrix}$$

4 Convergence to coalescent

If $(\gamma_1, \ldots, \gamma_L)$ is stationary distr. for the backward migration, then

$$\mathbf{B}_r^u \to \mathbf{P}_r, \; u \to \infty$$

where $\mathbf{P}_r$ consists of d_r equal rows $(\pi_r(\mathbf{x}), \mathbf{x} \in S_r)$ with

$$\pi_r(\mathbf{x}) = \binom{r}{x_1, \ldots, x_n} \gamma_1^{x_1} \ldots \gamma_L^{x_L}.$$

It follows

$$\mathbf{B}^u \to \mathbf{P} = \operatorname{diag}(\mathbf{P}_1, \ldots, \mathbf{P}_n), \; u \to \infty$$

and according to Möhle's lemma, with $\mathbf{G} = \mathbf{PCP}$

$$\big(\mathbf{B}(\mathbf{I} + N^{-1}\mathbf{C})\big)^{Nt} \to \mathbf{P} - \mathbf{I} + e^{t\mathbf{G}}, \; N \to \infty$$

Möhle, M. (1998) A convergence theorem for Markov chains arising in population genetics and coalescent with selfing. Adv. Appl. Prob. 30, 493–512.

For any $\mathbf{x} \in S_i$

$$\sum_{\mathbf{y} \in S_j} \mathbf{G}(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{y} \in S_j} (\mathbf{PCP})(\mathbf{x}, \mathbf{y}) = cQ_{ij}, \quad c = \sum_{k=1}^{L} \frac{1}{a_k} \gamma_k^2.$$

Writing this as $\mathbf{G}^{\downarrow} = cQ$ we conclude

$$(\mathbf{\Pi}^{Nt})^{\downarrow} \to e^{ctQ}, \; N \to \infty$$

that the total number of lineages with scaled time $X(Nt/c)$ is approximated by the number of branches in the Kingman coalescent.

Coalescent $N_e = N/c$. By Jensen's inequality $c \geq 1$

$$\sum_{k=1}^{L} \frac{1}{a_k} \gamma_k^2 = \sum_{k=1}^{L} a_k \left(\frac{\gamma_k}{a_k} \right)^2 \geq \left(\sum_{k=1}^{L} a_k \frac{\gamma_k}{a_k} \right)^2 = 1.$$

Test example: WFM with dummy islands. In this case $\gamma_i = a_i$ and $c = 1$.

5 Migration in random environment

Transition matrices of backward migration $\mathbf{B}_1^{[1]}, \mathbf{B}_1^{[2]}, \ldots$ are iid.

Takahashi, Y. (1969) Markov chains with random transition matrices.

Kodai Math. Sem. Rep. 21, 426–447.

Irreducible case: for each pair $1 \le i, j \le L$ there is a u such that

$$P(\mathbf{B}_1^{[1]} \ldots \mathbf{B}_1^{[u]}(i,j) > 0) > 0. \tag{1}$$

If furthermore, for some j and u

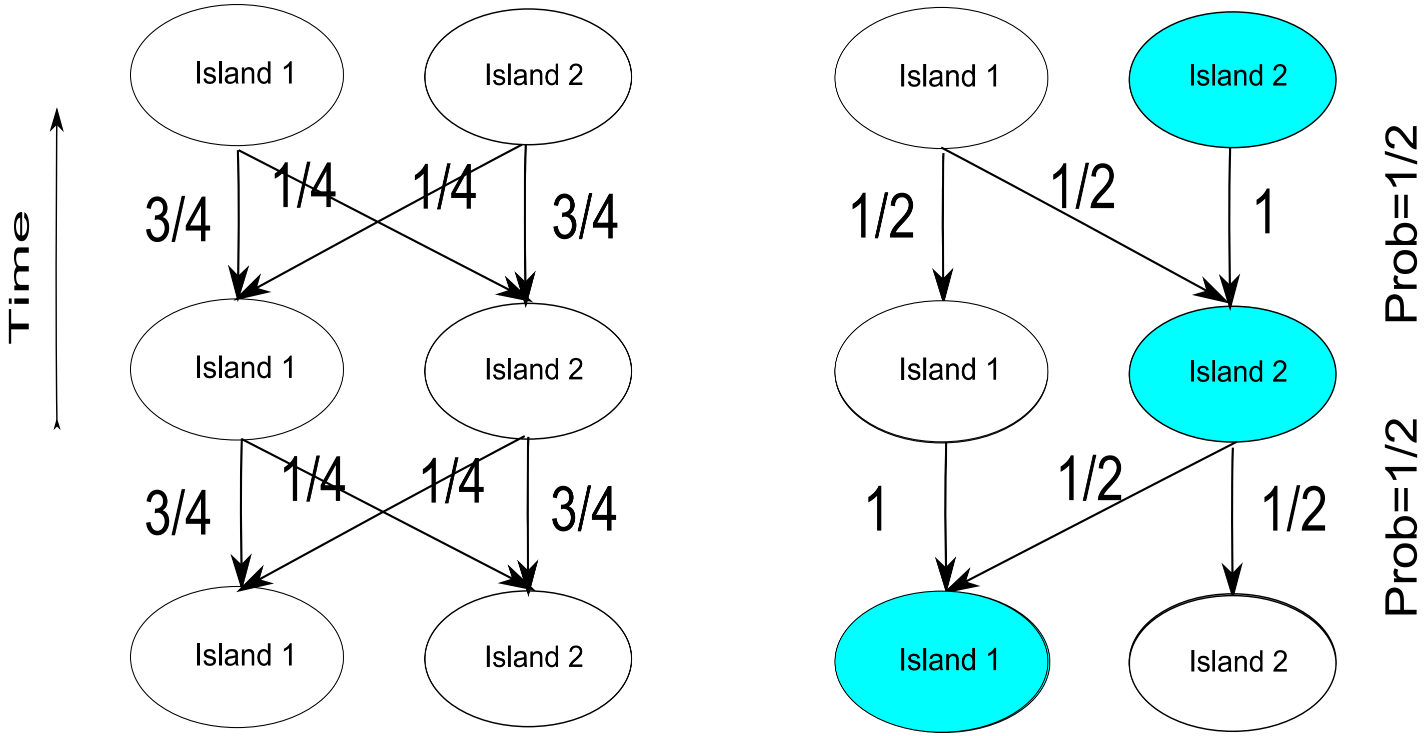
$$P(\mathbf{B}_1^{[1]} \ldots \mathbf{B}_1^{[u]}(i,j) > 0 \text{ for all } i) > 0, \tag{2}$$

then there exist random stationary probabilities $(\gamma_1, \ldots, \gamma_L)$

$$\mathbf{B}_1^{[1]} \ldots \mathbf{B}_1^{[u]} \xrightarrow{d} \begin{pmatrix} \gamma_1 & \cdots & \gamma_L \\ \cdot & \cdots & \cdot \\ \gamma_1 & \cdots & \gamma_L \end{pmatrix}, \ u \to \infty.$$

6 Two examples

Example 1: $\gamma_1 = \gamma_2 = 0.5$ $\gamma_1 \overset{d}{=} \gamma_2 \sim U(0,1)$



Constant migration rates Variable migration rates

Exact distribution of $\mathbf{B}_1^{[1]} \dots \mathbf{B}_1^{[u]}$ is uniform over 2^u matrices

$$\begin{pmatrix} j2^{-u} & 1 - j2^{-u} \\ (j-1)2^{-u} & 1-(j-1)2^{-u} \end{pmatrix}, \; j = 1, \dots 2^u$$

which is verified by induction

$$\left(\frac{j}{2^u}, 1 - \frac{j}{2^u}\right) \begin{pmatrix} 1 & 0 \\ 1/2 & 1/2 \end{pmatrix} = \left(\frac{j+2^u}{2^{u+1}}, 1 - \frac{j+2^u}{2^{u+1}}\right)$$

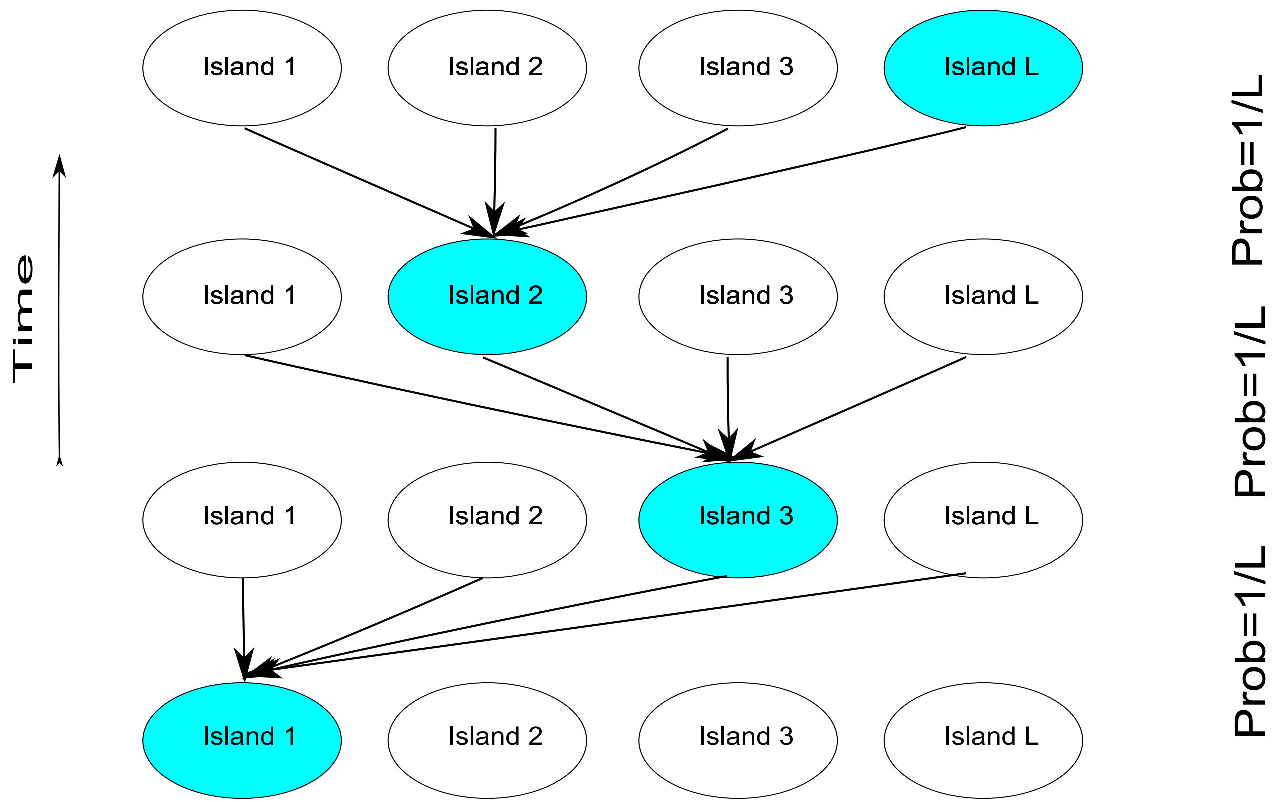
$$\left(\frac{j}{2^u}, 1 - \frac{j}{2^u}\right) \begin{pmatrix} 1/2 & 1/2 \\ 0 & 1 \end{pmatrix} = \left(\frac{j}{2^{u+1}}, 1 - \frac{j}{2^{u+1}}\right)$$

Weak convergence against almost sure convergence

$$\mathbf{B}_1^{[1]} \dots \mathbf{B}_1^{[u]} = \begin{pmatrix} Z_u & 1 - Z_u \\ Z_u - 2^{-u} & 1 - Z_u + 2^{-u} \end{pmatrix}, \; Z_{u+1} = Z_u/2 + 1/4 \pm 1/4$$

$$\mathbf{B}_1^{[u]} \dots \mathbf{B}_1^{[1]} = \begin{pmatrix} Z_u^* & 1 - Z_u^* \\ Z_u^* - 2^{-u} & 1 - Z_u^* + 2^{-u} \end{pmatrix}, \; Z_{u+1}^* = Z_u^* + 2^{-u}(1/4 \pm 1/4)$$

Example 2: $(\gamma_1, \ldots, \gamma_L) \sim \text{Mn}(1, 1/L, \ldots, 1/L)$



Conditions (1) and (2) follow from

$$P(\mathbf{B}_1(i,j) > 0 \text{ for all } i) > 0, \quad \text{for all } j.$$

7 New formula for N_e

Our main assertion: if (1) and (2) hold, then

$$(\mathbf{\Pi}^{[1]} \cdots \mathbf{\Pi}^{[Nt]})^{\downarrow} \overset{a.s.}{\to} e^{ctQ}, \qquad N \to \infty$$

so that $N_e = N/c$ with

$$c = c^{[\text{quenched}]} = \sum_{k=1}^{L} \frac{1}{a_k} E\left(\gamma_k^2\right).$$

Notice that

$$c^{[\text{quenched}]} - c^{[\text{annealed}]} = \sum_{k=1}^{L} \frac{1}{a_k} \text{Var}(\gamma_k)$$

and therefore

$$N_e^{[\text{quenched}]} \leq N_e^{[\text{annealed}]} \leq N.$$

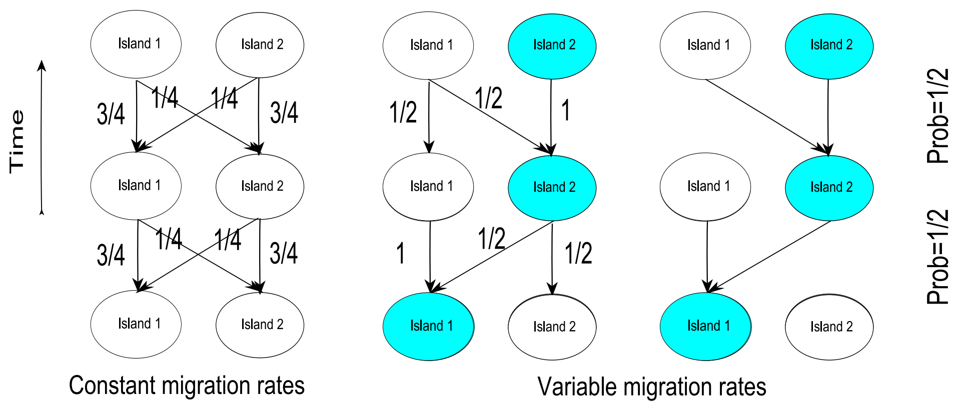
Example 1: $N_e^{[\text{quenched}]} = \frac{3}{4} N_e^{[\text{annealed}]}$ since

$$c^{[\text{annealed}]} = \frac{1}{4}\left(\frac{1}{a_1} + \frac{1}{a_2}\right),$$

$$c^{[\text{quenched}]} = \frac{1}{3}\left(\frac{1}{a_1} + \frac{1}{a_2}\right).$$

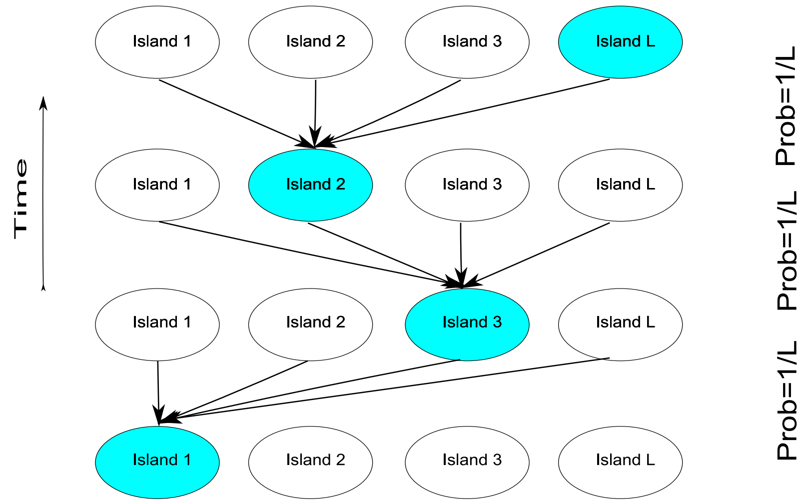
Example 2: $N_e^{[\text{quenched}]} = \frac{1}{2} N_e^{[\text{annealed}]}$ for $L = 2$

$$c^{[\text{quenched}]} = \frac{1}{2}\left(\frac{1}{a_1} + \frac{1}{a_2}\right)$$



Example 2 with general L gives the harmonic mean formula

$$\frac{1}{N_e^{[\text{quenched}]}} = \frac{1}{L}\left(\frac{1}{Na_1} + \ldots + \frac{1}{Na_L}\right).$$



Viewed backward in time the population undergoes iid fluctuations of generation sizes.

8 About the proof

Key decomposition

$$\mathbf{\Pi}^{[j]} = \mathbf{B}^{[j]}(\mathbf{I} + N^{-1}\mathbf{C}) + o(N^{-1})$$

where again $\mathbf{B}^{[j]} = \mathrm{diag}(\mathbf{B}_1^{[j]}, \ldots, \mathbf{B}_n^{[j]})$.

We have weak convergence of random matrices

$$\mathbf{B}^{[1]} \ldots \mathbf{B}^{[u]} \stackrel{d}{\to} \mathbf{P},\ u \to \infty.$$

Switching the product order

$$\mathbf{\Pi}^{[1]} \ldots \mathbf{\Pi}^{[Nt]} \stackrel{d}{=} \mathbf{\Pi}^{[Nt]} \ldots \mathbf{\Pi}^{[1]}$$

allows using a.s. convergence

$$\mathbf{B}^{[u]} \ldots \mathbf{B}^{[j]} \stackrel{a.s.}{\to} \mathbf{P}^{[j]},\ u \to \infty,\ j \geq 1.$$

Here $\mathbf{P}^{[j]} \stackrel{d}{=} \mathbf{P}$ are defined by $(\gamma_1^{[j]}, \dots, \gamma_L^{[j]}) \stackrel{d}{=} (\gamma_1, \dots, \gamma_L)$ satisfying for $i < j$

$$(\gamma_1^{[j]}, \dots, \gamma_L^{[j]})\mathbf{B}_1^{[j-1]} \cdots \mathbf{B}_1^{[i]} = (\gamma_1^{[i]}, \dots, \gamma_L^{[i]}). \quad (3)$$

An extension of Möhle's lemma implies

$$(\mathbf{\Pi}^{[Nt]} \cdots \mathbf{\Pi}^{[1]})^{\downarrow} = e^{Q \frac{1}{N} \sum_{j=1}^{[Nt]} c^{[j]}} + o_p(1)$$

where

$$c^{[j]} = \sum_{k=1}^{L} \frac{1}{a_k} (\gamma_k^{[j]})^2$$

form a strongly stationary sequence since the defining matrices $\mathbf{B}_1^{[1]}, \mathbf{B}_1^{[2]}, \dots$ are iid.

The sequence $c^{[1]}, c^{[2]}, \dots$ is mixing, because in view of (3), the vectors $(\gamma_1^{[j]}, \dots, \gamma_L^{[j]})$ and $(\gamma_1^{[i]}, \dots, \gamma_L^{[i]})$ are asymptotically independent as $j \to \infty$.

By the ergodic theorem

$$\frac{1}{Nt}\sum_{j=1}^{[Nt]} c^{[j]} \overset{a.s.}{\to} E\left(\sum_{k=1}^{L}\frac{1}{a_k}\gamma_k^2\right) =: c$$

Durrett,R. (1996) Probability: Theory and Examples. 2nd edition

we obtain convergence in probability

$$(\mathbf{\Pi}^{[1]}\cdots\mathbf{\Pi}^{[Nt]})^{\downarrow} \to e^{ctQ},\ N\to\infty. \tag{4}$$

Finally, to show that convergence in (4) is a.s. we use a monotonocity property:

for the products of transition matrices P_k the discrepancy among rows is monotone $\Delta_{u+1}\le\Delta_u$, where

$$\Delta_u = \sum_j \left(\max_i P_1\cdots P_u(i,j) - \min_i P_1\cdots P_u(i,j)\right).$$

THANK YOU!

WOW, what an audience...