On the use of Markov chains and Perron Frobenuis Theorem in Population Genetics

Ola Hössjer
Dept. of Mathematics
Stockholm University

June 2015

Motivation and Outline

- Conservation biology: Protect genetic diversity in order to
  - **Prevent inbreeding**
  - Keep species viable
  - Improve environmental adjustment of species
- Outline:
  - Populations with substructure
  - Formulas for long term rate of loss of genetic variants

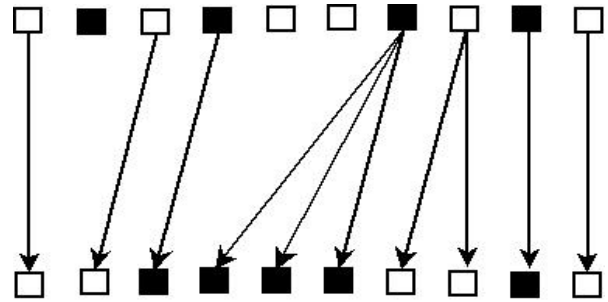
Effective population size N_e

- Population of size N
- N_e: Size of ideal population with same rate of loss of genetic variants as studied population (smaller $N_e \rightarrow$ faster rate)
- Short term protection rule: $N_e \geq 50$
- N_e/N varies between species

Genetic drift for size N population

- Discrete time $t = 0, 1, 2, \ldots$.
- Genes $g = 1, \ldots, 2N$.
- Two variants white and black
- ν_{tg} nr. of offspring of g, time t.
- Freq. of black, time $t+1$, is

$$\begin{aligned} X_{t+1} &= \tfrac{1}{2N}|\{g; g \text{ is black at } t+1\}| \\ &= \tfrac{1}{2N}\sum_{g;g \text{ black at time } t} \nu_{tg}. \end{aligned}$$



$$\begin{aligned} 2N &= 10, \\ X_t &= 0.4, \\ X_{t+1} &= 0.5, \\ \nu_{t2} &= \nu_{t5} = \nu_{t6} = 0, \\ \nu_{t1} &= \nu_{t3} = \nu_{t4} = \nu_{t9} = \nu_{t,10} = 1, \\ \nu_{t8} &= 2, \\ \nu_{t7} &= 3. \end{aligned}$$

- $\{X_t\}$ Markov chain, state space

$$\begin{aligned} \mathcal{X} &= \{0, \tfrac{1}{2N}, \ldots, \tfrac{2N-1}{2N}, 1\} \\ &= \{0\} \cup \{1\} \cup \{\tfrac{1}{2N}, \ldots, \tfrac{2N-1}{2N}\} \\ &= \mathcal{X}_0 \cup \mathcal{X}_1 \cup \mathcal{X}_2. \end{aligned}$$

Absorbing states: $\mathcal{X}_0$, $\mathcal{X}_1$
Transient states: $\mathcal{X}_2$

Ideal population: Wright-Fisher (WF) model[1]

Children pick parental genes independently;

$$\{\nu_{tg}\}_{g=1}^{2N} \sim \text{Mult}\left(2N; \frac{1}{2N}, \ldots, \frac{1}{2N}\right),$$

so that the transition kernel of $\{X_t\}$ is

$$\begin{aligned} \mathbf{P} &= (P(x,y))_{x,y\in\mathcal{X}}, \\ P(x,y) &= P(X_{t+1}=y|X_t=x) = \tbinom{2N}{2Ny} x^{2Ny}(1-x)^{2N(1-y)}. \end{aligned}$$

Feller (1951) showed that the eigenvalues of $\mathbf{P}$ are

$$\lambda_1 = \lambda_2 = 1, \lambda_j = \frac{(2N-1)(2N-2)\cdots(2N-j+2)}{(2N)^{j-2}}, \quad j = 3, \ldots, 2N+1,$$

so that the largest non-unit eigenvalue is

$$\lambda = \lambda_3 = 1 - \frac{1}{2N}. \tag{1}$$

It gives the asymtotic rate of fixation of one variant;

$$\lim_{t\to\infty} \frac{P(X_t \in \mathcal{X}_2)}{\lambda^t} = C, \quad (0 < C < \infty).$$

[1] Fisher (1921), Wright (1931).

Cannings model

Cannings (1974) showed more generally that

$$\lambda_1 = \lambda_2 = 1, \lambda_j = E\left(\prod_{g=1}^{j-1} \nu_{tg}\right), \quad j = 3, \ldots, 2N+1,$$

if $\{\nu_{tg}\}_{g=1}^{2N}$ are exchangeable, so that in particular,

$$\lambda = \lambda_3 = E(\nu_{t1}\nu_{t2}) = 1 + \mathrm{Cov}(\nu_{t1}, \nu_{t2}) = 1 - p,$$

with coalescence probability

$$\begin{aligned} p &= -\mathrm{Cov}(\nu_{t1}, \nu_{t2}) \\ &= \frac{E[\nu_{t1}(\nu_{t1}-1)]}{2N-1} \\ &= 2N \cdot E\left[\binom{\nu_{t1}}{2}\right] / \binom{2N}{2} \\ &= P\,(\text{two offspring have the same parent})\,. \end{aligned}$$

Eigenvalue effective size[2] N_{eE} = size of a WF population with fixation rate λ:

$$\lambda = 1 - \frac{1}{2N_{eE}} \Longrightarrow N_{eE} = \frac{1}{2(1-\lambda)} = \frac{N - \frac{1}{2}}{E\left[\nu_{t1}(\nu_{t1}-1)\right]}.$$

[2] Crow (1954), Ewens (1982).

Gene diversities

Introduce (predicted) gene diversities at time $t = 0, 1, 2, \ldots$

$$\begin{aligned} H_t &= 2X_t(1 - X_t), \\ &= P(\text{two genes picked with repl. have diff. variants}|X_t) \\ h_t &= E(H_t) \\ &= P(\text{two genes picked with repl. have diff. variants}). \end{aligned}$$

It can be shown that

$$h_{t+1} = \lambda h_t.$$

Hence gene diversities

$$h_t = \lambda^t h_0 = (1 - p)^t h_0, \tag{2}$$

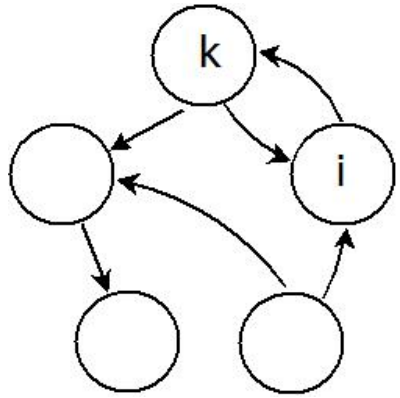
tend to zero at the same multiplicative rate as non-fixation probabilities $P(X_t \in \mathcal{X}_2)$.

But gene diversities and coalescence probabilities (p) are easier to analyze theoretically!!

Structured population

Divide into s subpopulations $i = 1, \ldots, s$. Let

$$
\begin{array}{rcl}
2N_i & = & \text{number of genes in subpop. } i, \ (\sum_i N_i = N) \\
X_{ti} & = & \text{fraction of black variants in subpop. } i \text{ at time } t, \\
\nu_{tkig} & = & \text{nr. of offspring of gene } g \text{ of subpop. } k \text{ at time } \\
 & & t \text{ that end up in subpop. } i \text{ at time } t+1, \\
 & & \text{exchangeable for } g = 1, \ldots, 2N_k.
\end{array}
$$



This gives dynamics

$$X_{t+1,i} = \frac{1}{2N_i} \sum_{k=1}^{s} \sum_{\substack{g; g \in \text{ subpop. } k \text{ at time } t, \\ g \text{ is black}}} \nu_{tkig}.$$

Find, under suitable conditions,

$$N_{eE} = \frac{1}{2(1-\lambda)},$$

where λ is rate of fixation, so that for some $0 < C < \infty$,

$$\lim_{t \to \infty} \frac{P(\text{non-fixation at time } t)}{\lambda^t} = C.$$

Structured population, contd.

Let

$$\mathbf{X}_t = (X_{t1}, \ldots, X_{ts}).$$

If reproduction is time invariant, $\{\mathbf{X}_t\}$ is Markov chain with state space

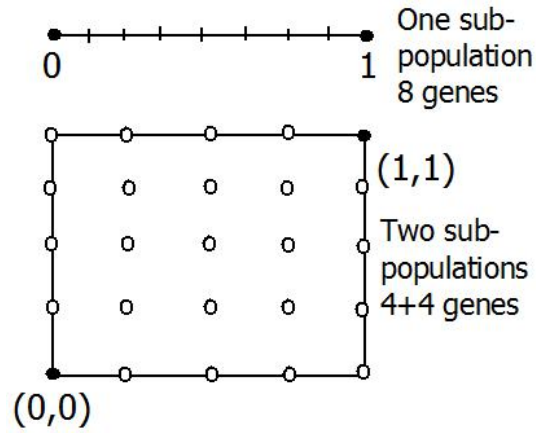
$$\begin{aligned} \mathcal{X} &= \{0, \tfrac{1}{2N_1}, \ldots, \tfrac{2N_1-1}{2N_1}, 1\} \times \ldots \times \{0, \tfrac{1}{2N_s}, \ldots, \tfrac{2N_s-1}{2N_s}, 1\} \\ &= \mathcal{X}_0 \cup \mathcal{X}_1 \cup \mathcal{X}_2. \end{aligned}$$

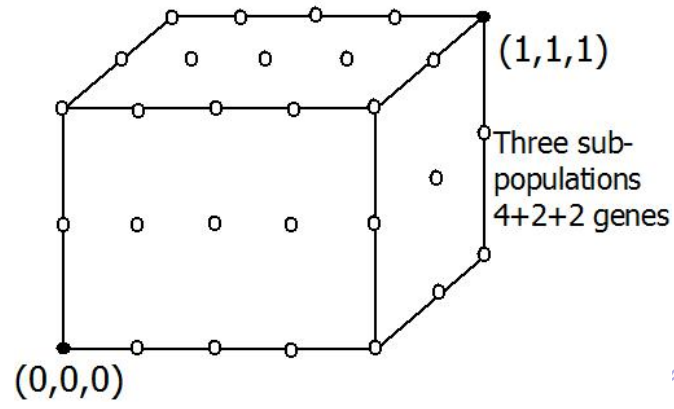
where

$$\begin{aligned} \mathcal{X}_0 &= \{(0, \ldots, 0)\}, \\ \mathcal{X}_1 &= \{(1, \ldots, 1)\}, \\ \mathcal{X}_2 &= \mathcal{X} \setminus (\mathcal{X}_0 \cup \mathcal{X}_1), \end{aligned}$$

and

$$\begin{aligned} \mathbf{P} &= (P(\mathbf{x}, \mathbf{y}))_{\mathbf{x},\mathbf{y} \in \mathcal{X}}, \\ P(\mathbf{x}, \mathbf{y}) &= P(\mathbf{X}_{t+1} = \mathbf{y} | \mathbf{X}_t = \mathbf{x}), \\ \lambda &= \text{3rd largest eigenvalue of } \mathbf{P}. \end{aligned}$$





Gene diversities for structured population

Introduce (predicted) gene diversities[3] at time $t = 0, 1, 2, \ldots$

$$\begin{aligned} H_{tij} &= X_{ti}(1 - X_{tj}) + (1 - X_{ti})X_{tj}, \\ &= P(\text{two genes picked with repl. from } i \text{ and } j \text{ have diff. variants} | \mathbf{X}_t) \\ h_{tij} &= E(H_{tij}) \\ &= P(\text{two genes picked with repl. from } i \text{ and } j \text{ have diff. variants}) \end{aligned}$$

between all pairs of subpopulations i and j. The column vector

$$\mathbf{h}_t = \text{vec}\left((h_{tij})_{i,j=1}^{s}\right)$$

with s^2 predicted gene diversities satisfies

$$\mathbf{h}_{t+1} = \mathbf{A}\mathbf{h}_t \Longrightarrow \mathbf{h}_t = \mathbf{A}^t\mathbf{h}_0, \tag{3}$$

where

$$\mathbf{A} = (A_{ij,kl})_{ij,kl \in \{1,\ldots,s\} \times \{1,\ldots,s\}}$$

is a square matrix of order s^2, and[4]

$$\boxed{\lambda = \lambda_{\max}(\mathbf{A}).} \tag{4}$$

---

[3] Maruyama (1970), Felsenstein (1972), Nei (1973).

[4] By Perron-Frobenius Theorem applied to $\mathbf{P}$.

Backward migration and coalescence theory to find **A**

If children pick parental subpopulations independently, with

$$\begin{aligned} B_{ik} &= \text{probability by which genes in } i \text{ pick} \\ &\quad \text{parental subpopulations from } k \in \{1, \ldots, s\}, \end{aligned}$$

$$\begin{aligned} p_{ijk} &= \text{coalescence probability within } k \\ &= P(\text{two genes from } i, j \text{ with parents in } k, \text{ have same parent}) \\ &= \frac{N_k}{2N_i N_j B_{ik} B_{jk}} \left(\frac{E(\nu_{tki1}(\nu_{tki1}-1))}{1-\frac{1}{2N_i}} \right)^{\{i=j\}} E\left(\nu_{tki1}\nu_{tkj1}\right)^{\{i\neq j\}}. \end{aligned}$$

Then (3) holds, with

$$A_{ij,kl} = \left(1 - \frac{1}{2N_i}\right)^{\{i=j\}} \left(\frac{1-p_{ijk}}{1-\frac{1}{2N_k}}\right)^{\{k=l\}} B_{ik} B_{jl}. \qquad (5)$$

Motivation of (3) and (5)

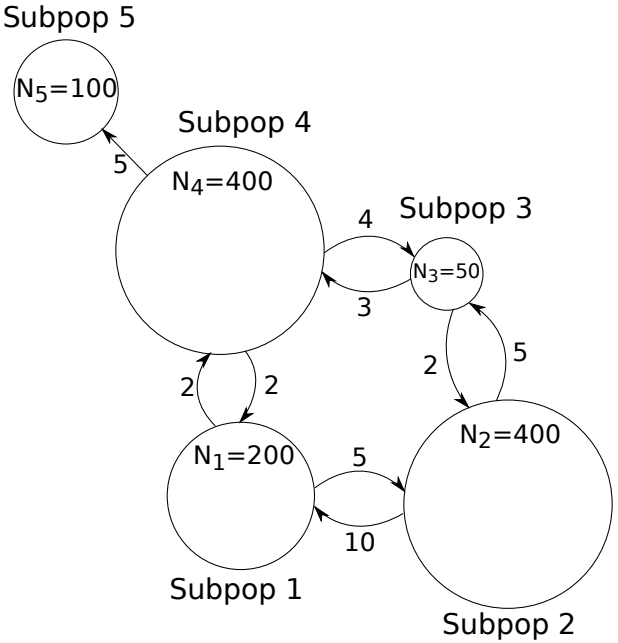
We have that

$$
\begin{aligned}
h_{t+1,ij} &= P(\text{genes from } i \text{ and } j \text{ at time } t+1 \text{ have different variants}) \\
&= P(\text{different genes picked from } i \text{ and } j \text{ at time } t+1) \\
&\quad \textstyle\sum_{k,l} \cdot P(\text{gene from } i \text{ has parent from } k \text{ at time } t) \\
&\qquad \cdot P(\text{gene from } j \text{ has parent from } l \text{ at time } t) \\
&\qquad \cdot P(\text{different parents from } k \text{ and } l) \\
&\qquad \cdot P(\text{different variants of the different parents at time } t) \\
&= (1-\tfrac{1}{2N_i})^{\{i=j\}} \textstyle\sum_{k,l} B_{ik} B_{jl} (1-p_{ijk})^{\{k=l\}} \cdot h_{t,kl} / (1-\tfrac{1}{2N_k})^{\{k=l\}} \\
&= \textstyle\sum_{k,l} A_{ij,kl} h_{t,kl}.
\end{aligned}
$$

with $A_{ij,kl}$ as in (5). In vector form this writes

$$\mathbf{h}_{t+1} = \mathbf{A}_t \mathbf{h}_t.$$

Computation of N_{eE}



$$(N_i)_{i=1}^{s} = (200, 400, 50, 400, 100),$$

$$N = \sum_{i=1}^{s} N_i = 1150,$$

$$(B_{ik})_{i,k=1}^{s} = \begin{pmatrix} 0.94 & 0.05 & 0 & 0.01 & 0 \\ 0.0125 & 0.9825 & 0.005 & 0 & 0 \\ 0 & 0.1 & 0.82 & 0.08 & 0 \\ 0.005 & 0 & 0.0075 & 0.9875 & 0 \\ 0 & 0 & 0 & 0.05 & 0.95 \end{pmatrix}$$

$$\begin{aligned} \text{Repr.} &= \{(\nu_{tkig})_{g=1}^{2N_k}\}_{k=1}^{s} \\ &\sim \text{Mult}\left(2N_i; \tfrac{B_{i1}}{2N_1}, \ldots, \tfrac{B_{i1}}{2N_1}, \ldots, \tfrac{B_{is}}{2N_s}, \ldots, \tfrac{B_{is}}{2N_s}\right) \\ &\quad \text{independently for } i = 1, \ldots, s, \end{aligned}$$

$$p_{ijk} = 1/(2N_k),$$

$$N_{eE} = 970,$$

Gene diversity effective size N_{eG}

Let

$$W_{ij} = P(\text{choose gene pair from } i \text{ and } j),$$

and collect them into row vector of length s^2:

$$\mathbf{W} = \text{vec}\left((W_{ij})_{1\le i,j\le s}\right)'.$$

Predicted gene diversity for two randomly sampled genes at time t, is

$$\begin{aligned} h_t &= P(\text{the two genes have different variants}) \\ &= \textstyle\sum_{1\le i,j\le s} W_{ij} h_{tij} \\ &= \mathbf{W}\mathbf{h}_t \\ &= \mathbf{W}\mathbf{A}^t\mathbf{h}_0. \end{aligned}$$

It follows from (1) and (2), that for Wright-Fisher model

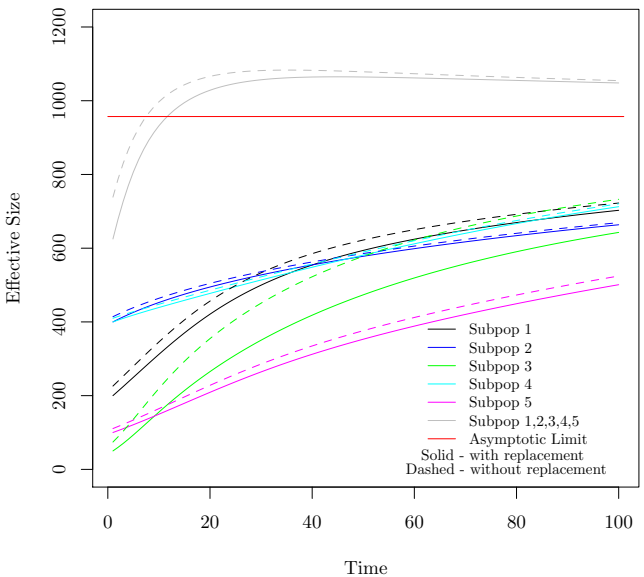
$$h_t = \left(1 - \frac{1}{2N}\right)^t \cdot h_0. \tag{6}$$

Gene diversity effective size over time interval $[0, t]$ solves (6), i.e.

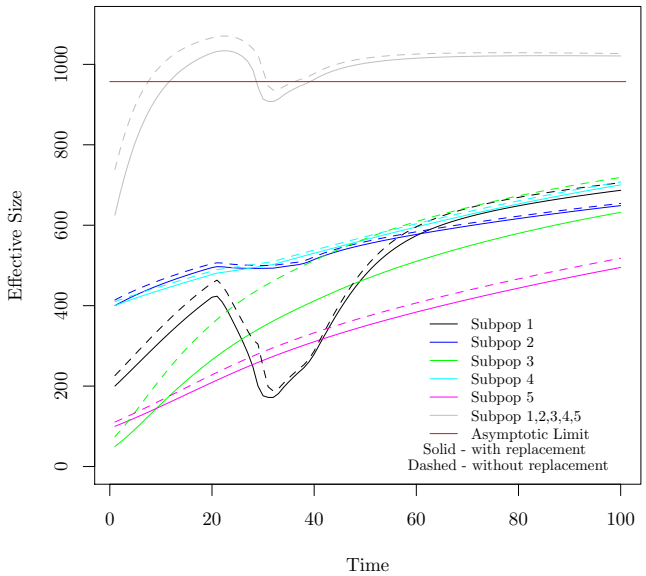
$$N_{eG}([0,t]) = \frac{1}{2\left[1 - \left(\frac{h_t}{h_0}\right)^{1/t}\right]} \overset{t\to\infty}{\to} \frac{1}{2(1-\lambda)} = N_{eE}.$$

Local/global N_{eG} and N_{eE}

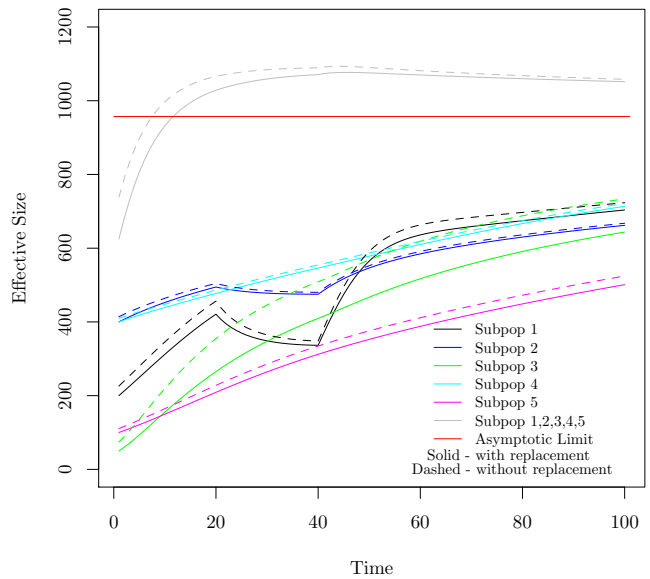
Constant subpop. sizes



Local bottleneck in 1



Blocked migration 1-2



Horizontal: N_{eE}

Solid: $t \to N_{eG}([0,t])$ for whole population and subpopulations

Global weights: $W_{ij} = 1/s^2$

Local weights, subpopulation k: $W_{ij} = 1_{\{(i,j)=(k,k)\}}$

Proof of (4)

We have that

$$h_t = \mathbf{W}\mathbf{h}_t = \mathbf{W}\mathbf{A}^t\mathbf{h}_0 = C\lambda_{\max}(\mathbf{A})^t + o\left(\lambda_{\max}(\mathbf{A})^t\right) \tag{7}$$

as $t \to \infty$. But also

$$\begin{aligned} h_t &= \sum_{ij} W_{ij} h_{tij} \\ &= \sum_{ij} W_{ij} E\left[X_{ti}(1 - X_{tj}) + X_{tj}(1 - X_{ti})\right] \\ &= E\left[\phi(\mathbf{X}_t)\right] \\ &= E\left[E(\phi(\mathbf{X}_t)|\mathbf{X}_0)\right] \\ &= \sum_{\mathbf{x},\mathbf{y}} \pi(\mathbf{x}) P^{(t)}(\mathbf{x},\mathbf{y})\phi(\mathbf{y}), \end{aligned} \tag{8}$$

where

$$\begin{aligned} \phi(\mathbf{x}) &= \sum_{i,j} W_{ij}\left[x_i(1 - x_j) + x_j(1 - x_i)\right], \\ \pi(\mathbf{x}) &= P(\mathbf{X}_0 = \mathbf{x}), \\ \mathbf{P}^t &= \left(P^{(t)}(\mathbf{x},\mathbf{y}); \mathbf{x},\mathbf{y} \in \mathcal{X}\right). \end{aligned}$$

Perron-Frobenius

Block decompose transition matrix as

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & \mathbf{0} \\ 0 & 1 & \mathbf{0} \\ \mathbf{P}_{20} & \mathbf{P}_{21} & \mathbf{P}_{22} \end{pmatrix} \Longrightarrow \mathbf{P}^t = \begin{pmatrix} 1 & 0 & \mathbf{0} \\ 0 & 1 & \mathbf{0} \\ \mathbf{P}_{20}^{(t)} & \mathbf{P}_{21}^{(t)} & \mathbf{P}_{22}^t \end{pmatrix},$$

Since $\mathbf{P}_{22}$ is non-negative, irreducible and aperiodic, it has unique largest eigenvalue $\lambda = \lambda_3$, and therefore

$$P^{(t)}(\mathbf{x}, \mathbf{y}) = \lambda^t r(\mathbf{x}) l(\mathbf{y}) + o(\lambda^t), \quad \mathbf{x}, \mathbf{y} \in \mathcal{X}_2, \tag{9}$$

where

$$\mathbf{l} = (l(\mathbf{x});\, \mathbf{x} \in \mathcal{X}_2)$$
$$\mathbf{r} = (r(\mathbf{x});\, \mathbf{x} \in \mathcal{X}_2)'$$

are left and right eigenvectors of $\mathbf{P}_2$ with eigenvalue λ and components

$$\begin{aligned} l(\mathbf{x}) &> 0, \\ r(\mathbf{x}) &> 0, \end{aligned} \tag{10}$$

for all $\mathbf{x} \in \mathcal{X}_2$.

Proof of (4), contd.

Use (8), (9), (10) and the fact that

$$\begin{aligned} \phi(\mathbf{x}) &= 0, \quad \mathbf{x} \in \mathcal{X}_0 \cup \mathcal{X}_1, \\ \phi(\mathbf{x}) &> 0, \quad \mathbf{x} \in \mathcal{X}_2 \text{ (if all } W_{ij} > 0), \\ \pi(\mathbf{x}) &> 0, \quad \text{for some } \mathbf{x} \in \mathcal{X}_2, \end{aligned}$$

to conclude

$$\begin{aligned} h_t &= \lambda^t \textstyle\sum_{\mathbf{x},\mathbf{y} \in \mathcal{X}_2} \pi(\mathbf{x}) r(\mathbf{x}) l(\mathbf{y}) \phi(\mathbf{y}) + o(\lambda^t) \\ &= \lambda^t \textstyle\sum_{\mathbf{x} \in \mathcal{X}_2} \pi(\mathbf{x}) r(\mathbf{x}) \cdot \sum_{\mathbf{y} \in \mathcal{X}_2} l(\mathbf{y}) \phi(\mathbf{y}) + o(\lambda^t) \\ &= C\lambda^t + o(\lambda^t), \end{aligned}$$

with $C > 0$. Combining this with (7), we finally deduce

$$\lambda = \lambda_{\max}(\mathbf{A}).$$

Other topics

- Various types of structure (geographic, age, sex, combinations, . . .).
- Computer program GESP (Olsson et al, 2015).
- Large population asymptotics:

$$N_{eE} = \frac{N}{C_1} + o(N) = N_{eC} + o(N) \text{ as } N \to \infty,$$

  where N_{eC} is coalescence effective size[5] and C_1 a coalescence rate.

- Small migration asymptotics:

$$N_{eE} = \frac{N}{C_2 B} + o(B^{-1}) \text{ as } B \to 0,$$

  where B = long term rate of subpopulation change in ancestral line.

- Leading right eigenvector of $\mathbf{A}$ to assess subpopulation differentiation[6]

[5] Nordborg and Krone (2002), Sjödin et al. (2005).

[6] F_{ST} of Wright (1943), G_{ST} of Nei (1973).

References

Hössjer, O. (2011). Coalescence theory for a general class of structured populations with fast migration. *Advances in Probability Theory* **43**(4), 1027-1047.

Hössjer, O., Jorde, P.E. and Ryman, N. (2013). Quasi equilibrium approximations of the fixation index under neutrality: The island model. *Theoretical Population Biology* **84**, 9-24.

Ryman, N., Allendorf, F.W., Jorde P.E., Laikre, L. and Hössjer, O. (2013). Samples from subdivided populations yield biased estimates of effective size that overestimate the rate of loss of genetic variation. *Molecular Ecology Resources* **14**, 87-99.

Olsson, F. Hössjer, O., Laikre, L. and Ryman, N. (2013). Characteristics of the variance effective population size over time using an age structured model with variable size. *Theoretical Population Biology* **90**, 91-103.

Hössjer, O. (2014). Spatial autocorrelation for subdivided populations with invariant migration schemes. *Methodology and Computing in Applied Probability* **16**(4), 777-810. DOI 10.1007/s11009-013-9321-3.

Hössjer, O. and Ryman, N. (2014). Quasi equilibrium, variance effective population size and fixation index for models with spatial structure. *Journal of Mathematical Biology* **69**(5), 1057-1128. DOI 10.1007/s00285-013-0728-9.

Hössjer, O., Olsson, F., Laikre, L. and Ryman, N. (2014). A new general analytical approach for modeling patterns of genetic differentiation and effective size of subdivided populations over time. *Mathematical Biosciences* **258**, 113-133. DOI: 10.1016/j.mbs.2014.10.001.

Hössjer, O., Olsson, F., Laikre, L. and Ryman, N. (2015). Metapopulation Inbreeding Dynamics, Effective Size and Subpopulation Differentiation - a General Analytical Approach for Diploid Organisms. *Theoretical Population Biology* **102**, 40-59. DOI:10.1016/j.tpb.2015.03.006.

Hössjer, O. (2015). On the eigenvalue effective size in structured populations. To appear in *Journal of Mathematical Biology***. Available online, DOI 10.1007/s00285-014-0832-5.**

THANKS!