On the use of Markov chains and Perron-Frobenius 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
- $\nu_{tg}$ nr. of offspring of $g$, time $t$.
- Freq. of black, time $t + 1$, is

\[
X_{t+1} = \frac{1}{2N} |\{g; g \text{ is black at } t + 1\}| = \frac{1}{2N} \sum_{g; g \text{ black at time } t} \nu_{tg}.
\]

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

\[
\mathcal{X} = \{0, \frac{1}{2N}, \ldots, \frac{2N-1}{2N}, 1\}
\]

\[
= \{0\} \cup \{1\} \cup \left\{\frac{1}{2N}, \ldots, \frac{2N-1}{2N}\right\}
\]

\[
= \mathcal{X}_0 \cup \mathcal{X}_1 \cup \mathcal{X}_2.
\]

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

Transient states: $\mathcal{X}_2$
Ideal population: Wright-Fisher (WF) model

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

\[
P = \left( P(x, y) \right)_{x,y \in \chi},
\]

\[
P(x, y) = P(X_{t+1} = y|X_t = x) = \binom{2N}{2Ny} x^{2Ny} (1 - x)^{2N(1-y)}.
\]

Feller (1951) showed that the eigenvalues of \( P \) 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

\[
\lambda = \lambda_3 = 1 - \frac{1}{2N}.
\]

It gives the asymptotic rate of fixation of one variant;

\[
\lim_{t \to \infty} \frac{P(X_t \in \chi_2)}{\lambda^t} = C, \quad (0 < C < \infty).
\]

\(^1\)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,

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

with coalescence probability

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

Eigenvalue effective size\(^2\) \( N_{eE} = \) size of a WF population with fixation rate \( \lambda \):

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

Gene diversities

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

\[
H_t = 2X_t(1 - X_t),
\]
\[
= P(\text{two genes picked with repl. have diff. variants}|X_t)
\]
\[
h_t = E(H_t)
\]
\[
= P(\text{two genes picked with repl. have diff. variants}).
\]

It can be shown that

\[
h_{t+1} = \lambda h_t.
\]

Hence gene diversities

\[
h_t = \lambda^t h_0 = (1 - p)^t h_0,
\]

(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

$$
2N_i = \text{number of genes in subpop. } i, \ (\sum_i N_i = N)
$$

$$
X_{ti} = \text{fraction of black variants in subpop. } i \text{ at time } t,
$$

$$
\nu_{tkig} = \text{nr. of offspring of gene } g \text{ of subpop. } k \text{ at time } t \text{ that end up in subpop. } i \text{ at time } t+1,
$$

exchangeable for $g = 1, \ldots, 2N_k$.

This gives dynamics

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

Find, under suitable conditions,

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

where $\lambda$ 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

\[ X_t = (X_{t1}, \ldots, X_{ts}). \]

If reproduction is time invariant, \{X_t\} is Markov chain with state space

\[ \mathcal{X} = \{0, \frac{1}{2N_1}, \ldots, \frac{2N_1-1}{2N_1}, 1\} \times \ldots \times \{0, \frac{1}{2N_s}, \ldots, \frac{2N_s-1}{2N_s}, 1\} \]

\[ = \mathcal{X}_0 \cup \mathcal{X}_1 \cup \mathcal{X}_2. \]

where

\[ \mathcal{X}_0 = \{(0, \ldots, 0)\}, \]
\[ \mathcal{X}_1 = \{(1, \ldots, 1)\}, \]
\[ \mathcal{X}_2 = \mathcal{X} \setminus (\mathcal{X}_0 \cup \mathcal{X}_1), \]

and

\[ P = (P(x, y))_{x, y \in \mathcal{X}}, \]
\[ P(x, y) = P(X_{t+1} = y | X_t = x), \]
\[ \lambda = \text{3rd largest eigenvalue of } P. \]
Gene diversities for structured population

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

\[
H_{tij} = X_{ti}(1 - X_{tj}) + (1 - X_{ti})X_{tj},
\]

\[
= P(\text{two genes picked with repl. from } i \text{ and } j \text{ have diff. variants}|X_t)
\]

\[
h_{tij} = E(H_{tij})
\]

\[
= P(\text{two genes picked with repl. from } i \text{ and } j \text{ have diff. variants})
\]

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

\[
h_t = \text{vec}\left((h_{tij})_{i,j=1}^s\right)
\]

with \(s^2\) predicted gene diversities satisfies

\[
h_{t+1} = Ah_t \implies h_t = A^th_0,
\]

where

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

is a square matrix of order \(s^2\), and\(^4\)

\[
\lambda = \lambda_{\text{max}}(A).
\]

---


\(^4\)By Perron-Frobenius Theorem applied to \(P\).
Backward migration and coalescence theory to find $A$

If children pick parental subpopulations independently, with

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

$$p_{ijk} = \text{coalescence probability within } k = P(\text{two genes from } i, j \text{ with parents in } k, \text{ have same parent})$$

$$= \frac{N_k}{2N_i N_j B_{ik} B_{jk}} \left( E(\nu_{tki1}(\nu_{tki1}-1)) \right)^{\{i=j\}} \cdot E(\nu_{tki1} \nu_{tkj1})^{\{i\neq j\}}.$$

Then (3) holds, with

$$A_{ij,kl} = \left(1 - \frac{1}{2N_i}\right)^{\{i=j\}} \left(1 - \frac{1}{2N_k}\right)^{\{k=l\}} B_{ik} B_{jl}. \quad \text{(5)}$$
Motivation of (3) and (5)

We have that

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

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

\[
h_{t+1} = A_t h_t.
\]
Computation of $N_{eE}$

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

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

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

\[
\text{Repr.} = \left\{ (\nu_{tgk})^{2N_k}_{g=1} \right\}_{k=1}^s
\sim \text{Mult} \left( 2N_i, \frac{B_{i1}}{2N_1}, \ldots, \frac{B_{i1}}{2N_1}, \ldots, \frac{B_{is}}{2N_s}, \ldots, \frac{B_{is}}{2N_s} \right)
\text{independently for } i = 1, \ldots, s,
\]

\[p_{ijk} = 1/(2N_k),\]

\[N_{eE} = 970,\]
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 $s^2$:

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

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

$$h_t = P(\text{the two genes have different variants})$$

$$= \sum_{1 \leq i,j \leq s} W_{ij} h_{tij}$$

$$= Wh_t$$

$$= WA_t h_0.$$ 

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

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

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

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

Constant subpop. sizes  Local bottleneck in 1  Blocked migration 1-2

Horizontal: $N_{eE}$
Solid: $t \rightarrow N_{eG}([0, t])$ 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

$$h_t = Wh_t = WA^t h_0 = C \lambda_{\text{max}}(A)^t + o \left( \lambda_{\text{max}}(A)^t \right) \quad (7)$$

as $t \to \infty$. But also

$$h_t = \sum_{ij} W_{ij} h_{tij}$$
$$= \sum_{ij} W_{ij} E \left[ X_{ti} (1 - X_{tj}) + X_{tj} (1 - X_{ti}) \right]$$
$$= E \left[ \phi(X_t) \right]$$
$$= E \left[ E(\phi(X_t) | X_0) \right]$$
$$= \sum_{x, y} \pi(x) P^{(t)}(x, y) \phi(y), \quad (8)$$

where

$$\phi(x) = \sum_{i,j} W_{ij} [x_i (1 - x_j) + x_j (1 - x_i)],$$
$$\pi(x) = P(X_0 = x),$$
$$P^t = (P^{(t)}(x, y); x, y \in \mathcal{X}).$$
Perron-Frobenius

Block decompose transition matrix as

\[
P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ P_{20} & P_{21} & P_{22} \end{pmatrix} \implies P^t = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ P_{20}^t & P_{21}^t & P_{22}^t \end{pmatrix},
\]

Since \( P_{22} \) is non-negative, irreducible and aperiodic, it has unique largest eigenvalue \( \lambda = \lambda_3 \), and therefore

\[
P^{(t)}(x, y) = \lambda^t r(x)l(y) + o(\lambda^t), \quad x, y \in \mathcal{X}_2,
\]  

(9)

where

\[
l = (l(x); x \in \mathcal{X}_2)
\]

\[
r = (r(x); x \in \mathcal{X}_2)'
\]

are left and right eigenvectors of \( P_2 \) with eigenvalue \( \lambda \) and components

\[
l(x) > 0, \\
r(x) > 0,
\]  

(10)

for all \( x \in \mathcal{X}_2 \).
Proof of (4), contd.

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

\[ \phi(x) = 0, \quad x \in X_0 \cup X_1, \]
\[ \phi(x) > 0, \quad x \in X_2 \text{ (if all } W_{ij} > 0), \]
\[ \pi(x) > 0, \quad \text{for some } x \in X_2, \]

to conclude

\[
    h_t = \lambda^t \sum_{x,y \in X_2} \pi(x) r(x) l(y) \phi(y) + o(\lambda^t) \\
    = \lambda^t \sum_{x \in X_2} \pi(x) r(x) \cdot \sum_{y \in X_2} l(y) \phi(y) + o(\lambda^t) \\
    = C\lambda^t + o(\lambda^t),
\]

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

\[ \lambda = \lambda_{\text{max}}(A). \]
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:

\[
N_{eE} = \frac{N}{C_2B} + o(B^{-1}) \text{ as } B \to 0,
\]

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

▶ Leading right eigenvector of \(A\) to assess subpopulation differentiation\(^6\)


\(^6\)\(F_{ST}\) of Wright (1943), \(G_{ST}\) of Nei (1973).


THANKS!