1 Application: Population Genetics

Here we consider a new case study involving population genetics under the Wright-Fisher model. First we introduce a few jargons.

Jargons 1.1 (Locus). A locus is simply a location on a chromosome. It can represent a gene, a SNP (single base pair polymorphism) or simply a location.

Jargons 1.2 (Alleles). An allele is one of a number of alternative forms of the locus. A dominant allele is usually capitalized. For human and other diploid animals, there are typically two alleles of paternal and maternal origin.

Jargons 1.3 (Genotype). A genotype is the genetic makeup of an individual, in this case, the types of alleles of an individual, e.g. Aa. Having the same pair of alleles is called homozygous while having different alleles is called heterozygous.

Jargons 1.4 (Mutation). A mutation is a random genetic change. Here we refer to the change of one allele to another.

Jargons 1.5 (Fitness). Fitness is the measure of survivability and ability to reproduce of an individual possessing a certain genotype.

Jargons 1.6 (Neutral Evolution). Neutral evolution happens when all genotypes of interest have the same fitness. In this case, there is no selection from such genetic variations.

Here we introduce the idea of Genetic Drift. Consider a model with the following assumptions:

1. There is a population with finite size $N$.
2. The size of the population is constant throughout evolution.
3. There are discrete, non-overlapping generations.
4. Mating is completely random with replacements. This means, an individual from generation $i$ can give rise to, randomly, 0, 1 or more offsprings.

From this, the Wright-Fisher model consider the following evolution along the generations:

1. At generation 0, there are $2N$ alleles some A, some a.
2. At generation 1, each A or a allele from generation 0 may result in one, more or zero copies of that allele.
3. Because the population size is finite and fixed, due to uneven passing of alleles by chance, eventually there will be only one allele, A or a left. This phenomenon is called genetic drift.
2 Wright-Fisher model

We can characterize neutral evolution more mathematically. Again consider the total of \(2N\) alleles and their evolution over time. Let \(X_n\) be the number of A alleles at generation \(n\) and \(P_n = \frac{X_n}{2N}\). Set \(X_0\) to be a known quantity. Due to the fact that there are \(2N\) independent passing of alleles, all independent under assumption, the probability of generating an A allele at generation \(i\) is \(P_{i-1}\), thus:

\[ X_i \sim Binomial(2N, P_{i-1}), \]

and

\[ Pr(X_k = k, 0 \leq k \leq 2N) = \binom{2N}{k} P_{n-1}^k (1 - P_{n-1})^{2N-k} \]

Note 1. Furthermore, there are a few properties of note.

1. \(P_{\text{inf}} \to \begin{cases} 1 \ (\text{Fixed}) \\ 0 \ (\text{Lost}) \end{cases}\)

2. \(E[X_n|X_{n-1}] = 2NP_{n-1} = X_{n-1}\),

thus \(X_n\) is a Martingale.

3. \(Var[X_n] = 2NP_{n-1}(1 - P_{n-1})\)

4. Since \(X_n\) is a Martingale,

\[
\lim_{n \to \infty} E[X_n] = X_0,
\]

thus

\[
Pr(X_\infty = 2N) * (2N) + Pr(X_\infty = 0) * 0 = X_0 \Rightarrow Pr(X_\infty = 2N) = P_0.
\]

3 Continuous approximation

Since the population \(N\) can be large, any single mutation will have frequency of \(\frac{1}{2N}\), which can be very small in comparison. As such, it is reasonable to approximate the evolution by a continuous model. The SDE is in the form:

\[
dX_t = a(x)dt + \sqrt{b(x)}dW \\
X_0 = \chi
\]

Then, we want to study the dynamics of

\[
\lim_{\delta t \to 0} \frac{E[f(X_t)] - f(\chi)}{\delta t}
\]

for some \(f\). Here we apply Ito’s rule:
\[
df(X_t) = \frac{df}{dX}dX_t + \frac{1}{2} \frac{d^2f}{dX^2} (dX_t)^2 \\
= \frac{df}{dX}a(x)dt + \frac{df}{dX} \sqrt{b(x)}dW + \frac{1}{2} \frac{d^2f}{dX^2} b(x) dt
\]

Let \( f_X = \frac{df}{dX}, \frac{d}{dX} = \frac{d^2f}{dX^2} \), from the above, we can derive the dynamics of the expectation as well:

\[
E[f(X_t)] = f_X a(x) + \frac{1}{2} f_{XX} b(x)
\]

We can then define an operator which correspond to the above:

**Definition 1** (Diffusion generator). Let \( L \) be a linear operator such that

\[
Lf = \frac{E[f(X_t)]}{dt} = f_X a(x) + \frac{1}{2} f_{XX} b(x)
\]

**Example 3.1.**

\( f(x) = x, \ f_x = 1, \ f_{xx} = 0 \),

apply \( L \):

\[
Lf = a(x) * 1 + b(x) * 0
\]

\[
\frac{E[f(X_t)]}{dt} = a(x)
\]

**Remark 1.** It can be seen that \( a(x) \) represents the infinitesimal mean.

**Example 3.2.**

\( f(x) = (x - \chi)^2 \),

apply \( L \):

\[
b(x) = \frac{d}{dt} E[(X_t - \chi)^2]
\]

**Remark 2.** It can be seen that \( b(x) \) represents the infinitesimal variance.

### 4 SDE model

In the original problem, we have \( Var[P_n] = \frac{1}{2N} P_{n-1}(1 - P_{n-1}) \), we can change the scale of time to \( t \) such that each unit represent \( 2N \) generation, thus \( Var[P(t)] = P(t-1)(1-P(t-1)) \).

Let \( X(t) \) be the frequency of A allele at time \( t \), as \( N \rightarrow \infty \), there is the following SDE:

\[
dX(t) = \sqrt{X(t)(1-X(t))}dW,
\]

with the corresponding diffusion generator:

\[
LX = \frac{1}{2} X(1-X) \frac{d^2X}{dX^2}
\]
Example 4.1 (Adding selection to the model). Assume the relative fitness of A is 1 and a 
$1 - S$ (meaning during the random sampling there is $1 - S$ chance of a being selected), let 
$\gamma = S \times 2N, N \to \infty$, we can derive the following generator instead:

$$L = \gamma X (1-X) \frac{d}{dX} + \frac{1}{2} X (1-X) \frac{d^2}{dX^2}$$

If we look at each evolution of $X$, as an extention of the discrete case, there is:

$$X_{next} = \frac{X}{X + (1-X)(1-S)} = \frac{X}{1 - (1-X)S} \approx X(1 + (1-X)S + O(S^2))$$

thus:

$$X_{next} - X \approx SX(1-X)$$

from here the form of $L$ can be derived.

Example 4.2 (Adding mutation). Similarly we can add mutation to the equation, let $\mu_1 = \frac{\beta_1}{2N}$ be the chance of a allele randomly changing to A allele, and $\mu_2 = \frac{\beta_2}{2N}$ be the chance of the reverse mutation. Similarly we can derive the further extended generator with mutation included:

$$L = [\gamma X (1-X) + \beta_1 (1-X) + \beta_2 X] \frac{d}{dX} + \frac{1}{2} X (1-X) \frac{d^2}{dX^2}$$

Example 4.3 (Hitting probability). We know as $t \to \infty$, $X_t$ tends to either 1 or 0, the chance of either happening can be of interest given certain initial distributions.

Define $h(t) = Pr(T_a < T_b)$ where $T_a = \inf \{ X_t = a \}$. For example $a = 1$ would indicate the chance of fixation by certain time.

If there is no selection in the model, then $h(y)$ is Martingale and we have:

$$h(x) = \int_0^1 Pr(X(t) = y|X_0 = \chi)h(y)dy = \int_0^1 [Pr(X(t) = y|X_0 = \chi) - \delta(x - y)]h(y)dy$$

Here we can apply the Kolmogorov forward equation for $Pr$:

$$\frac{\partial Pr}{\partial t} = -\frac{\partial}{\partial y} a(y) Pr + \frac{1}{2} \frac{\partial^2}{\partial y^2} b(y) Pr$$

To address the hitting probability issue, we simply solve for $h$ at steady state given some boundary conditions, for example $h(a) = 1, h(b) = 0$ with:

$$0 = -a(y) \frac{dh}{dx} + b(x) \frac{1}{2} \frac{d^2h}{x^2}$$

This can be solved by a change of variable, let $\psi = \frac{dh}{dx}$, we can derive the analytical solution of $h$. In particular:

$$\psi = e^{\int a - \frac{b}{2} dy}$$

Similarly, the hitting probability with mutation but without any selection can be derived as well and follows a Gamma distribution. If $\beta_i < \frac{1}{2}$ this distribution is not defined.