Universidad Autónoma del Estado de Hidalgo
Instituto de Ciencias Básicas e Ingeniería
Área Académica de Matemáticas y Física

An “effective dimensionality” facilitates stochastic modeling of structured populations

Tesis que para obtener el título de

Maestra en Matemáticas

presenta

Mónica Villarroel Ramírez

bajo la dirección de

Dra. Erika Elizabeth Rodríguez Torres
Dr. Philip John Gerrish

Mineral de la Reforma, Hidalgo., julio de 2025

UAEH

Universidad Autónoma del Estado de Hidalgo
Instituto de Ciencias Básicas e Ingeniería
School of Engineering and Basic Sciences
Área Académica de Matemáticas y Física
Department of Physics and Mathematics

Mineral de la Reforma, Hgo., a 26 de junio de 2025

Número de control: ICBI-AAMyF/2777/2025
Asunto: Autorización de impresión de tesis

MTRA. OJUKY ROCÍO ISLAS MALDONADO
DIRECTORA DE ADMINISTRACIÓN ESCOLAR

El Comité Tutorial de la tesis titulada **"An "effective dimensionality" facilitates stochastic modeling of structured populations"**, realizada por la sustentante **Mónica Villarroel Ramírez**, con número de cuenta **266540**, perteneciente a la **Maestría en Matemáticas**, una vez que ha revisado, analizado y evaluado el documento recepcional de acuerdo a lo estipulado en el Artículo 110 del Reglamento de Estudios de Posgrado, tiene a bien extender la presente:

AUTORIZACIÓN DE IMPRESIÓN

Por lo que la sustentante deberá cumplir los requisitos del Reglamento de Estudios de Posgrado y con lo establecido en el proceso de grado vigente.

Atentamente
"Amor, Orden y Progreso"

El Comité Tutorial

Universidad Autónoma del Estado de Hidalgo
Instituto de Ciencias Básicas e Ingeniería
UAEH
Área Académica de Matemáticas y Física

Dra. Erika Elizabeth Rodríguez Torres
Directora

Dr. Philip John Gerrish
Co Director

Dr. Benjamín Alfonso Itzá Ortiz
Miembro del comité

Dr. Raúl Temoltzi Ávila
Miembro del comité

Resumen

La genética de poblaciones es el estudio de los cambios en las frecuencias génicas a lo largo del tiempo debido a la selección natural y los efectos de muestreo aleatorio (deriva[1]). Comparar la selección y la deriva en dos poblaciones diferentes puede ser difícil, ya que cada población puede tener su propio conjunto de factores adicionales que inducen ruido. Ejemplos de tales factores adicionales incluyen el comportamiento, las fluctuaciones en el tamaño de la población (por ejemplo, estacionalidad), la estructura de la población y la endogamia. Para facilitar la comparación entre poblaciones, se introdujo el concepto de "tamaño efectivo de la población" (denotado como $\mathcal{N}_e$), definido como el tamaño constante de una población idealizada para la cual las propiedades estocásticas de las frecuencias génicas son equivalentes a las de la población real que se modela. Esta equivalencia estocástica ha sido muy exitosa al modelar poblaciones "bien mezcladas". Sin embargo, la mayoría de las poblaciones reales no cumplen con el criterio de "bien mezcladas", ya que están distribuidas en el espacio físico. Para abordar la heterogeneidad espacial, los modelos previos han asumido simplemente que diferentes regiones del espacio tienen diferentes valores de $\mathcal{N}_e$. Pero entonces surgen preguntas, como ¿qué tan granular debe ser el valor de $\mathcal{N}_e$ en el espacio? En otras palabras, ¿cuántas poblaciones distintas existen, cada una con su propio tamaño efectivo, y cada una bien mezclada? Debido a factores intrínsecos a la biología, es realmente difícil elaborar modelos espaciales de otra manera, y como se puede observar, tales modelos no son realmente modelos espaciales. Por estas razones, muchos científicos creen que la genética de poblaciones en un espacio verdaderamente continuo sigue siendo un problema abierto. Proponemos el uso de una segunda equivalencia estocástica: el número efectivo de dimensiones ($\mathcal{D}_e$), para absorber las complejidades espaciales de manera similar a cómo $\mathcal{N}_e$ absorbe las complejidades demográficas. Hemos desarrollado un método para estimar $\mathcal{D}_e$ utilizando la teoría de la coalescencia modificada y el cálculo estocástico. Demostramos la utilidad práctica de este resultado al estimar la dimensionalidad efectiva utilizando datos de la influenza aviar altamente patógena (HPAI) subtipo A(H5N1). Finalmente, discutimos nuestros hallazgos en un contexto más amplio y preguntamos si podrían representar un paso clave hacia una teoría de la genética de poblaciones en un espacio continuo.

[1] Observamos que la palabra "drift" puede resultar confusa, ya que en física se refiere a un movimiento direccional; sin embargo, en genética de poblaciones hace referencia a fluctuaciones aleatorias en las frecuencias génicas de una población, causadas por su tamaño finito. Por lo tanto, en realidad se asemeja más a un proceso de difusión.

Abstract

Population genetics is the study of changes in gene frequencies over time due to natural selection and random sampling effects (drift[2]). Comparing selection and drift in two different populations can be difficult because each population can have its own set of additional noise-inducing factors. Examples of such additional factors include behavior, fluctuations in population size (e.g., seasonality), population structure, and inbreeding. To facilitate comparison between populations, the concept of "effective population size" (denoted $\mathcal{N}_e$) was introduced, defined as the constant size of an idealized population for which the stochastic properties of gene frequencies are equivalent to the real population being modeled. This stochastic equivalence has been very successful in modeling populations that are "well-mixed". However, most real populations do not meet the "well-mixed" criterion because they are distributed in physical space. To address spatial heterogeneity, previous models have simply assumed that different regions in space have different $\mathcal{N}_e$. But then questions arise, like how granular should we make $\mathcal{N}_e$ over space? In other words, how many distinct populations exist, each with its own effective size, and each one well-mixed? For reasons unique to biology, it is really difficult to make spatial models in any other way, and as we can see, such models are not really spatial models. For these reasons, many scientists believe that population genetics in a truly continuous space remains an open problem. Here, we propose the use of a second stochastic equivalence, the effective number of dimensions ($\mathcal{D}_e$, a real number), to absorb spatial complexities in a way similar to how $\mathcal{N}_e$ absorbs demographic complexities. We have developed a method for estimating $\mathcal{D}_e$ from spatially-sampled DNA sequences, using correlations in allele frequencies, modified coalescent theory and stochastic calculus. We demonstrate the practical utility of this result by estimating the effective dimensionality using data from highly pathogenic avian influenza (HPAI) subtype A(H5N1). We discuss our findings in the larger context and ask if they might provide a useful step towards a theory of population genetics in continuous space.

[2] We note that the word "drift" can be confusing because in physics it means directional movement; in population genetics, however, it means random fluctuations in gene frequencies in a population because of the finite size of the population, so it is actually more like diffusion.

Agradecimientos

Quiero comenzar agradeciendo al Área Académica de Matemáticas y Física por toda su ayuda en estos años de estudio, así como a cada uno de los profesores que me impartieron clase. Expreso también mi agradecimiento al Consejo Nacional de Ciencia y Tecnología (CONACYT), hoy Secretaría de Ciencias, Humanidades, Tecnología e Innovación (SECIHTI), por la beca de posgrado otorgada durante mi posgrado.

Agradezco al Dr. Raúl Temoltzi Ávila y al Dr. Benjamín Alfonso Itzá Ortiz, integrantes del comité sinodal, por cada una de sus aportaciones, así como por su tiempo y dedicación al supervisar este trabajo de investigación.

Una parte fundamental de este proyecto fue el apoyo del Dr. Gideon Bradburd, Dr. Nick Hengartner y Dr. Pablo Padilla Longoria, por sus aportaciones académicas, sus consejos, y por su apoyo brindado para la realización de esta tesis.

Quisiera reconocer de manera especial a la Dra. Erika Elizabeth Rodríguez Torres y al Dr. Philip John Gerrish, mis asesores de este trabajo de tesis, por invitarme a formar parte de este proyecto. Su dedicación, el tiempo compartido en cada sesión, y la apertura para explorar nuevas ideas fueron fundamentales en el desarrollo de este trabajo.

Mi gratitud a mi familia y amigos. Gracias por su apoyo a través de una oración, una palabra de aliento o un consejo; he recibido cada uno con profundo cariño. A mi abuelita, por enseñarme el valor de la gratitud, el compromiso y el servicio. Desde el primer momento, ha estado presente en cada uno de mis pasos, acompañándome con amor y paciencia.

A mis papás, por enseñarme el valor de la fe y la perseverancia. Su apoyo incondicional, expresado de tantas formas, ha sido esencial para completar este trabajo. Finalmente, agradezco a Dios, quien nunca me dejó sola en este camino, especialmente en los momentos más difíciles, confiando el futuro a sus manos.

Contents

Introduction

Studying biological sequence data, particularly genetic variation, is important for understanding the evolutionary processes that shape populations. By analyzing genetic variation, we can infer how populations evolve in response to mechanisms such as natural selection, genetic drift, migration, and mutation. This field, known as *population genetics*, aims to describe the dynamics of populations over time and space.

The origins of population genetics can be traced to the need to reconcile Charles Darwin's theory of natural selection with Gregor Mendel's work on inheritance. Initially, Darwin's theory was in conflict with the inheritance mechanisms that were widely accepted during his time. It was only through the mathematical work of G. Hardy and W. Weinberg that this apparent paradox was resolved, forming the basis of population genetics. From these early foundations, the field has developed to focus on the stochastic nature of evolution, particularly how genetic variation changes over time due to random and directional processes.

A central concept in population genetics is the *effective population size* ($\mathcal{N}_e$), first introduced by Sewall Wright (1931) in [78]. $\mathcal{N}_e$ refers to the *size of an idealized population that experiences the same level of genetic drift as the actual population.* This concept has since been extended to various scenarios involving overlapping generations and inbreeding. However, while $\mathcal{N}_e$ has proven to be a powerful tool in studying genetic drift, it may not fully capture the complexities of populations that are spatially structured and vary not only over time but also across physical space.

Traditional models in population genetics rely on summary statistics such as $\mathcal{N}_e$ to characterize the entire population. However, when populations have spatial structure, these models become inadequate. We hypothesize that a **second summary statistic** is necessary to account for the additional stochasticity introduced by spatial distribution. This new statistic, which we introduce as the *effective dimensionality* ($\mathcal{D}_e$), will complement $\mathcal{N}_e$ and provide a more comprehensive framework for understanding spatially structured populations.

The *effective dimensionality* $\mathcal{D}_e$ is a novel concept that extends the idea of effective population size into the spatial domain. Unlike traditional dimensions, $\mathcal{D}_e$ can take non-integer values and represents the number of spatial dimensions that effectively describe the population's distribution. This concept utilizes stochastic processes, including Bessel processes and mathematical theory, to model spatial evolution more accurately.

In summary, the hypothesis explored in this thesis is as follows:

> The single stochastic equivalence, $\mathcal{N}_e$, is not sufficient to model populations evolving in continuous space. A second stochastic equivalence, $\mathcal{D}_e$, is needed to capture the effects of spatial structure.

The applications of this work have the potential to benefit various fields. For example, consider a population in which a highly pathogenic avian influenza virus spreads across North America, affecting dairy cattle as well as wild birds. The spread of the virus is not uniform; certain regions experience rapid transmission, while others remain largely unaffected due to geographic barriers and varying population densities. Migratory wild bird populations play an important role in the virus spread. Our aim is to understand how these genetic differences evolve over time and spread across avian and mammalian populations.

In classical population genetics, we often rely on the concept of effective population size ($\mathcal{N}_e$), which summarizes the genetic drift experienced by a population. However, this measure assumes a well-mixed population and does not account for spatial structure, such as the fact that wild bird populations are spread across different geographic locations. This is an example that will be studied in the thesis.

The objectives of this thesis are to introduce the concept of *effective dimensionality* as a second summary statistic alongside $\mathcal{N}_e$ for spatially structured populations, to extend *coalescent theory* to include spatial structure, allowing us to better understand genetic variation in populations distributed across physical space, and to apply this framework to real-world population data, such as high pathogenicity avian influenza HPAI A(H5N1), by estimating $\mathcal{D}_e$ and its relevance to different fields such as conservation biology and epidemiology.

The structure of this thesis is organized as follows: Chapter 1 introduces the mathematical concepts and tools used throughout the thesis, including probability theory, stochastic processes, Brownian motion (BM), and Bessel processes. Chapter 2 explores classical population genetics, covering the historical background, genetic concepts, random genetic drift, and models such as the Wright–Fisher and Moran models, as well as the effective population size (EPS) and the coalescent process. In Chapter 3, we analyze spatial population genetics, discussing the spatial movement of genetic lineages, challenges in modeling evolution in continuous space. Chapter 4 presents the results, detailing the extraction of probabilities of neutral mutations and employing methods like Fourier analysis to derive *pgfs*. Chapter 5 applies the mathematical models to real-world data, specifically in the context of HPAI A(H5N1) and the genetic analysis of influenza A virus genes (HA, NA, and MP), and interprets the results. Finally, the thesis concludes by suggesting future research directions and potential applications, with supplementary code included in the appendices.

Chapter 1

Mathematical preliminaries

This chapter introduces the mathematical concepts and theoretical background necessary for the developments presented in this thesis. It begins with a review of fundamental concepts from probability theory. We then provide an introduction to stochastic processes and Brownian motion, which we will refer to as BM throughout the rest of the text. This includes the strong Markov property and the reflection principle, both of which will be essential to the results that follow. Finally, we introduce Bessel processes, discuss their main properties, and present methods for their simulation.

Finally, we discuss topics such as recurrence and transience, as well as the Skorokhod embedding theorem, all of which are fundamental to the probabilistic modeling of biological systems. As an interdisciplinary study, this thesis makes use of these mathematical concepts to analyze genetic variation within biological populations.

1.1 Probability theory concepts

For this section, we will be consulting several texts, such as [64], [25] and [21].

Definition 1.1.1 (σ-algebra)**.** A *σ-algebra* $\mathcal{F}$ over a sample space Ω is a collection of subsets of Ω that satisfies the following properties:

1. $\Omega \in \mathcal{F}$,

2. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$,

3. If $A_1, A_2, \dots \in \mathcal{F}$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

Definition 1.1.2. A *measurable space* is a pair $(\Omega, \mathcal{F})$, where Ω is a sample space and $\mathcal{F}$ is a σ-algebra over Ω.

Definition 1.1.3 (σ-algebra generated)**.** Let $\mathcal{A}$ be a collection of subsets of a set Ω. The *σ-algebra generated* by $\mathcal{A}$, denoted by $\sigma(\mathcal{A})$, is the smallest σ-algebra on Ω that contains every set in $\mathcal{A}$. In other words, $\sigma(\mathcal{A})$ is the intersection of all σ-algebras on Ω that contain $\mathcal{A}$, i.e.,

$$\sigma(\mathcal{A}) = \bigcap \{\mathcal{F} \mid \mathcal{F} \text{ is a } \sigma\text{-algebra on } \Omega \text{ and } \mathcal{A} \subseteq \mathcal{F}\}.$$

Definition 1.1.4 (Borel σ-algebra)**.** The *Borel σ-algebra* on the real numbers $\mathbb{R}$, denoted $\mathcal{B}(\mathbb{R})$, is the σ-algebra generated by the collection of *open intervals* in $\mathbb{R}$.

Formally, let

$$A = \{(a,b) \subset \mathbb{R} \mid a, b \in \mathbb{R}, a < b\}.$$

be the collection of all open intervals in $\mathbb{R}$.

Then, the *Borel σ-algebra* is the smallest σ-algebra containing all open intervals, i.e.,

$$\mathcal{B}(\mathbb{R}) = \sigma(A),$$

where $\sigma(A)$ denotes the σ-algebra generated by A.

Definition 1.1.5 (Probability measure)**.** Let $(\Omega, \mathcal{F})$ be a measurable space. A *probability measure* $\mathbb{P}$ on $(\Omega, \mathcal{F})$ is a function $\mathbb{P} : \mathcal{F} \to [0,1]$ that satisfies:

1. $\mathbb{P}(\Omega) = 1$,

2. For all $A \in \mathcal{F}$, we have $\mathbb{P}(A) \geq 0$,

3. If $A_1, A_2, \cdots \in \mathcal{F}$ are disjoint sets, then:

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

It is said in [29] that the probability measure is a special example of what is called a measure on the pair $(\Omega, \mathcal{F})$.

Definition 1.1.6 (Measure)**.** A *measure* is a function $\mu : \mathcal{F} \to [0, \infty]$ that satisfies the following properties:

1. $\mu(\emptyset) = 0$,

2. For all $A \in \mathcal{F}$, we have $\mu(A) \geq 0$,

3. If $A_n \in \mathcal{F}$ for each $n \in \mathbb{N}$, and $\{A_n\}_{n\in\mathbb{N}}$ is a countable collection of disjoint sets, then:

$$\mu\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} \mu(A_n).$$

We can observe that a measure μ is a probability measure if $\mu(\Omega) = 1$.

Definition 1.1.7 (Measure space)**.** A *measure space* is a triple $(\Omega, \mathcal{F}, \mu)$, where:

- Ω is a non-empty set,
- $\mathcal{F}$ is a σ-algebra of subsets of Ω,
- μ is a measure defined on $\mathcal{F}$.

Additionally, for any set $B \in \mathcal{F}$, the value $\mu(B)$ is called the measure of the set B. When $\mu(\Omega)$ is finite, meaning $\mu(\Omega) < \infty$, we refer to μ as a finite measure. Specifically, if $\mu(\Omega) = 1$, μ becomes a probability measure, often denoted by $\mathbb{P}$. In this case, $\mathbb{P}(B)$ represents the probability of the event B for any $B \in \mathcal{F}$.

Definition 1.1.8 (Probability space)**.** A *probability space* is a triple defined by $(\Omega, \mathcal{F}, \mathbb{P})$, where:

1. Ω is the sample space,
2. $\mathcal{F}$ is a sigma-algebra of events (a collection of subsets of Ω),
3. $\mathbb{P}$ is a probability measure, which assigns a probability to each event in $\mathcal{F}$.

Definition 1.1.9 (Random variable)**.** A *random variable* is a measurable function $X : \Omega \to \mathbb{R}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, that is, X is a function that satisfies:

$$X^{-1}((-\infty, x]) = \{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}.$$

for all $x \in \mathbb{R}$.

Definition 1.1.10 (Probability measure induced by a random variable)**.** Let X be a random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $\mathbb{R}$. The *probability measure induced by* X is defined for any Borel set $B \in \mathcal{B}(\mathbb{R})$ as:

$$\mathbb{P}_X(B) = \mathbb{P}\left(\{\omega \in \Omega \mid X(\omega) \in B\}\right).$$

This represents the probability that X takes values in B, with $\mathbb{P}_X(B)$ being the probability of the set $\{\omega \mid X(\omega) \in B\}$ under the original measure $\mathbb{P}$.

Definition 1.1.11 (Cumulative distribution function (*cdf*)). The *cdf* of a random variable X is defined as:

$$F_X(x) = \mathbb{P}(X \leq x).$$

This function gives the probability that X will take a value less than or equal to x.

Definition 1.1.12 (Probability density function (*pdf*)). For continuous random variables, the *pdf* $f_X(x)$ describes the likelihood of X taking a particular value. The *cdf* can be obtained by integrating the *pdf*:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt.$$

The probability that X lies in the interval $[a, b]$ is:

$$\mathbb{P}(a \leq X \leq b) = \int_{a}^{b} f_X(x)\,dx.$$

Definition 1.1.13 (Probability mass function (*pmf*)). For discrete random variables, the *pmf* $p_X(x)$ gives the probability that $X = x$:

$$p_X(x) = \mathbb{P}(X = x).$$

Definition 1.1.14 (Expected value (mean)). The *expected value* $\mathbb{E}[X]$ is the "average" or "central value" of the random variable X. It is given by:

- For discrete random variables:

$$\mathbb{E}[X] = \sum_{x} x \cdot p_X(x).$$

- For continuous random variables:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f_X(x)\,dx.$$

Notation: The expectation of X is also denoted as $\mu_X = \mathbb{E}(X)$.

Definition 1.1.15 (Variance). The *variance* $\mathrm{Var}(X)$ measures the spread of the distribution of X. It is expressed as:

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

The standard deviation $\sigma(X)$ is the square root of the variance:

$$\sigma(X) = \sqrt{\mathrm{Var}(X)}.$$

Definition 1.1.16 (Covariance)**.** The *covariance* between two random variables X and Y is a measure of how much the two variables change together. It is given by:

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

If $\mathrm{Cov}(X, Y) > 0$, then X and Y tend to increase together. If $\mathrm{Cov}(X, Y) < 0$, when one increases, the other tends to decrease.

Definition 1.1.17 (Independence)**.** Two events A and B are independent if:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B).$$

For random variables X and Y, independence means:

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B).$$

Definition 1.1.18 (Conditional probability)**.** The *conditional probability* of event A given event B is:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

The *conditional expectation* of X given Y is:

$$\mathbb{E}[X|Y] = \sum_x x \cdot p_{X|Y}(x). \quad \text{(discrete case)}$$

or

$$\mathbb{E}[X|Y] = \int_{-\infty}^{\infty} x f_{X|Y}(x)\, dx. \quad \text{(continuous case)}$$

Definition 1.1.19 (Moment-generating function *(mgf)*)**.** The *mgf* of a random variable X is:

$$M_X(t) = \mathbb{E}[e^{tX}].$$

The *mgf* is useful for deriving the moments (mean, variance, etc.) of X.

Definition 1.1.20 (Characteristic function)**.** The *characteristic function* of a random variable X is:

$$\varphi_X(t) = \mathbb{E}[e^{itX}].$$

This function is always well-defined, even when the *mgf* is not.

Definition 1.1.21 (Law of large numbers)**.** The *law of large numbers* states that the sample average of independent, identically distributed random variables converges to the expected value as the sample size grows:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mathbb{E}[X].$$

with probability 1 (strong law) or in probability (weak law). We note the implicit assumption that $\mathbb{E}[X] < \infty$.

Definition 1.1.22 (Central limit theorem)**.** The *central limit theorem* is a consequence of the *law of large numbers.* It states that, if the convergence stated in the *law of large numbers* holds true, then it necessarily follows that the sum (or average) of a large number of independent, identically distributed random variables has a normal distribution:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \mathbb{E}[X]) \xrightarrow{d} N(0,\sigma^2).$$

Now, having established the fundamental probability concepts, we move on to the definition and properties of stochastic processes, which are central to understanding BM and Bessel processes.

1.2 Introduction to stochastic processes

The framework used in this work is based on a branch of probability known as stochastic processes. These processes are important in population genetics, as they allow us to model and analyze the randomness and uncertainty in genetic variation and evolutionary processes. In this work, we will make use of a family of stochastic processes collectively known as BM and Bessel processes. The main reference for this chapter is [63]. The texts that we used as a guides are: [4], [69], [49], [19], [61].

A *stochastic process* is a collection of random variables indexed by time (or space) that describes the evolution of a system over time under uncertainty. It is a mathematical model used to represent random phenomena that evolve over time. In a stochastic process, the index set can be either discrete or continuous.

Definition 1.2.1 (Stochastic process)**.** A *stochastic process* is a collection of random variables, denoted as $\{X(t)(\omega) : t \in T, \omega \in \Omega\}$, where t takes values in the index set T and Ω is the state space.

The variable t represents time, and $X(t)$ denotes the state of the process at time t. The index set T is the set of possible times at which the process can be evaluated. A *discrete-time stochastic process* is indexed by a discrete set of time points, typically $t \in \mathbb{Z}_+$ or $t \in \mathbb{N}$, such as the random walk. A *continuous-time stochastic process* is indexed by a continuous set of time points, typically $t \in \mathbb{R}_+$, such as BM. We will discuss some types of stochastic processes.

1.2.1 Types of stochastic processes

1. A *Markov process* is a special type of stochastic process with the *Markov property.* This property is that the future state of the system depends only on the present state and not on the sequence of events that preceded it. Past events do not influence future states; they are independent of the past. For a process $\{X(t)\}$, it holds that:

$$\begin{aligned}\mathbb{P}(X(t+1) = x \mid X(t) = x_t, X(t-1) = x_{t-1}, \dots)\\ = \mathbb{P}(X(t+1) = x \mid X(t) = x_t),\end{aligned}$$

   meaning the future is independent of the past, given the present.

2. A *stationary process* is a process whose statistical properties do not change over time. For example, the joint distribution of the process values remains the same regardless of the time shift. This means that the behavior of the process is invariant to time shifts.

   Let $X(t), X(t+h), \dots, X(t+nh)$ represent a sequence of random variables. If the joint distribution of $X(t)$ is the same for all $h > 0$ and for any increasing sequence $t_1 < t_2 < \dots < t_n$, where $t_j \in T$ and $t_i + h \in T$, then the process $\{X(t) : t \in T\}$ is called a stationary process.

3. A process has *independent increments* if the increments over disjoint time intervals are independent. Formally, let $X(t)$ be a stochastic process. The process $X(t)$ has independent increments if, for any set of disjoint time intervals $[t_1, t_2], [t_3, t_4], \dots, [t_n, t_{n+1}]$, the random variables corresponding to the increments,

$$X(t_2) - X(t_1), X(t_4) - X(t_3), \dots, X(t_{n+1}) - X(t_n),$$

   are independent. That is, for any set of indices $1 \leq i, j \leq n$, we have:

$$\mathbb{P}\big((X(t_2) - X(t_1)) \in A_1, (X(t_4) - X(t_3)) \in A_2, \ldots,$$
$$(X(t_{n+1}) - X(t_n)) \in A_n\big) = \prod_{i=1}^{n} \mathbb{P}\big((X(t_{2i}) - X(t_{2i-1})) \in A_i\big). \quad (1.1)$$

where $A_1, A_2, \ldots, A_n$ are sets in the respective sample spaces.

For example, in *Poisson processes*, the number of events occurring in non-overlapping intervals are independent of each other.

4. *Poisson process*: Let $X(t)$ be a stochastic process in continuous time where $t \geq 0$. Let $T_1, T_2, \ldots$ model the number of events occurring in a fixed time interval, where the events occur independently. The number of events that occur in separate time intervals does not affect each other.

   Let $X(t)$ be a non-negative integer-valued stochastic process, where $t \in T = [0, \infty)$, satisfying the following conditions:

   (a) $X(0) = 0$ and $P(X(0) = 0) = 1$. This means that no events can occur at time $t = 0$.

   (b) $\{X(t) : t \in T\}$ has independent increments; i.e., for any time points $0 \leq t_0 < t_1 < \ldots < t_n$ in T, the random variables $X(t_1) - X(t_0), X(t_2) - X(t_1), \ldots, X(t_n) - X(t_{n-1})$ are independent. This means that the counts of events in non-overlapping intervals are independent of each other.

   (c) $\{X(t) : t \in T\}$ has stationary increments; i.e., for $t > s$, $X(t) - X(s)$ has the same distribution as $X(t+h) - X(s+h)$ for all h such that $t+h$ and $s+h$ are in T. This means that the distribution of the number of events that occur in an interval does not depend on its position in time.

   (d) For $t > s$, the probability $P[X(t) - X(s) = k] = \frac{e^{-\lambda(t-s)}(\lambda(t-s))^k}{k!}$, where $\lambda > 0$ and $k = 0, 1, 2, \ldots$.

5. *Brownian motion*: A continuous-time process where the increments over disjoint intervals are independent and normally distributed. In the next section, we will define this process in more detail and explore its properties.

These are just a few examples of stochastic processes; there are many more. For further study, the following books can be consulted: [62] and [5].

1.3 Brownian motion

We will now introduce one of the most important stochastic processes, widely used in fields such as physics, finance, and biology. This process models the random movement of particles suspended in a fluid, as well as various other types of random motion. Over the years, it has been extensively studied, leading to significant advances in both theoretical research and practical applications. We will base our discussion on the following references: [52], [6], [42], [67], and [16].

1.3.1 Definition of Brownian motion

Definition 1.3.1 (*One-dimensional BM*)**.** A *Brownian motion* (BM) is a continuous-time stochastic process $\{\mathcal{B}(t), t \geq 0\}$ taking real values. It is called a *one-dimensional BM* (or *linear process*) started at $x \in \mathbb{R}$ if it satisfies the following properties:

1. *Initial condition*:
$$\mathcal{B}(0) = x, \quad x \in \mathbb{R}.$$
This means that the process starts at an arbitrary real value x, which is not necessarily zero.

2. *Independent increments*: The increments of the process over disjoint time intervals are independent. Specifically, for any sequence of times $0 \leq t_1 < t_2 < \cdots < t_n$, the random variables
$$\mathcal{B}(t_n) - \mathcal{B}(t_{n-1}), \quad \mathcal{B}(t_{n-1}) - \mathcal{B}(t_{n-2}), \quad \ldots, \quad \mathcal{B}(t_2) - \mathcal{B}(t_1)$$
are independent.

3. *Normal increments*: For any $t > s$, the increment $\mathcal{B}(t) - \mathcal{B}(s)$ is normally distributed:
$$\mathcal{B}(t) - \mathcal{B}(s) \sim \mathcal{N}(0, t - s).$$
This means that the increment over any interval $[s, t]$ has mean zero and variance $t - s$, proportional to the length of the interval.

4. *Continuity of paths*: With probability 1, the function $t \mapsto \mathcal{B}(t)$ is continuous, meaning that BM has continuous sample paths. However, these paths are almost surely nowhere differentiable.

If $x = 0$, we refer to the process $\{\mathcal{B}(t), t \geq 0\}$ as a *standard BM.*

We now define the d-dimensional BM as a generalization of the *one*-dimensional case to higher dimensions.

Definition 1.3.2 (*d-dimensional BM*)**.** A *d-dimensional BM* is a continuous-time stochastic process

$$\mathcal{B}(t) = (\mathcal{B}_1(t), \mathcal{B}_2(t), \ldots, \mathcal{B}_d(t)), \quad t \geq 0,$$

where each component $\mathcal{B}_i(t)$ (for $i = 1, \ldots, d$) is an independent one-dimensional BM. The process is said to be *started at* $\mathbf{x} \in \mathbb{R}^d$ if

$$\mathcal{B}(0) = \mathbf{x}.$$

The process satisfies the following properties:

1. *Initial condition*:

$$\mathcal{B}(0) = \mathbf{x}, \quad \text{where} \quad \mathbf{x} = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d.$$

2. *Independent increments*: For any sequence of times $0 \leq t_1 < t_2 < \cdots < t_n$, the increments

$$\mathcal{B}(t_n) - \mathcal{B}(t_{n-1}), \quad \mathcal{B}(t_{n-1}) - \mathcal{B}(t_{n-2}), \quad \ldots, \quad \mathcal{B}(t_2) - \mathcal{B}(t_1)$$

   are independent random vectors in $\mathbb{R}^d$.

3. *Normal increments*: The increments of the process are normally distributed:

$$\mathcal{B}(t) - \mathcal{B}(s) \sim \mathcal{N}(\mathbf{0}, (t-s)I_d), \quad \text{for } t > s,$$

   meaning that each component $\mathcal{B}_i(t) - \mathcal{B}_i(s) \sim \mathcal{N}(0, t-s)$ independently, and the covariance matrix is $(t-s)I_d$, where I_d is the $d \times d$ identity matrix.

4. *Continuity of paths*: With probability 1, the function $t \mapsto \mathcal{B}(t)$ is continuous, meaning that BM has continuous sample paths in $\mathbb{R}^d$. However, these paths are almost surely nowhere differentiable.

If $\mathbf{x} = \mathbf{0}$, we call $\mathcal{B}(t)$ a *standard* d-dimensional BM.

As a note, two-dimensional BM is often referred to as *planar BM.* A d-dimensional BM can be constructed from independent one-dimensional BMs. If a process starts at $x \in \mathbb{R}^d$ instead of 0, it is given by

$$\mathcal{B}(t) + x, \quad t \geq 0.$$

This process retains all the properties of a BM but is initialized at x rather than the origin.

There are three well-known approaches to constructing a BM: those of Wiener, Kolmogorov, and Lévy. Each method provides a distinct mathematical perspective on the process. For a detailed discussion of these approaches, see [66].

1.3.2 Properties of Brownian motion

There are several important properties to consider when studying BM, such as:

Proposition 1.3.3 (*Translation*). *For every $x \in \mathbb{R}$, the process $X(t) = x+\mathcal{B}(t)$ is a BM starting at x. That is, this property means that the behavior of the process is the same regardless of the starting point.*

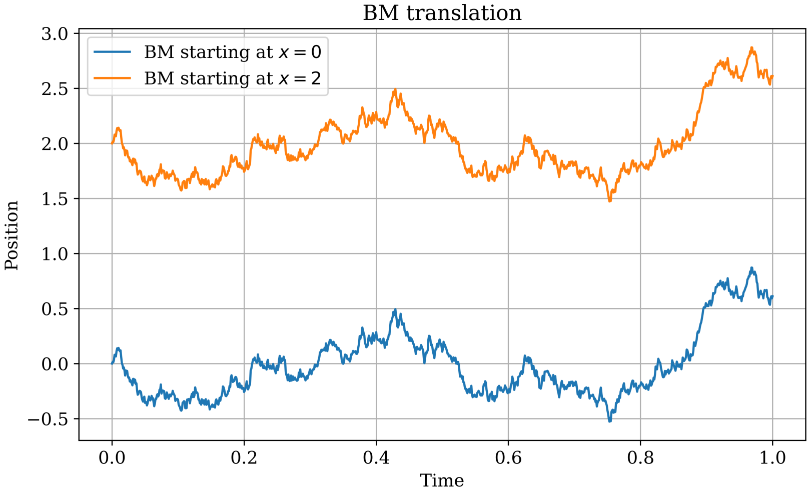

Figure 1.1: Translation. One process starts at $x = 0$, and the other starts at $x = 2$. The two trajectories were simulated with same seed for the random number generator.

Proof. We will show that the process $X(t) = x + \mathcal{B}(t)$ satisfies all four properties of a d-dimensional BM initiated at x. We need to show that $X(0) = x$ almost surely.

$$X(0) = x + \mathcal{B}(0) = x + 0 = x.$$

Then to prove that the increments of $X(t)$ are independent, let s be such that $0 \leq s < t$:

$$X(t) - X(s) = (x + \mathcal{B}(t)) - (x + \mathcal{B}(s)) = \mathcal{B}(t) - \mathcal{B}(s).$$

Since $\mathcal{B}(t)$ has independent increments, $X(t)$ also has this property.

Now we need to show that the increments $X(t) - X(s)$ are normally distributed. For $t > s$:

$$X(t) - X(s) = \mathcal{B}(t) - \mathcal{B}(s).$$

By the properties of $\mathcal{B}(t)$:

$$\mathcal{B}(t) - \mathcal{B}(s) \sim \mathcal{N}(0, (t-s)I_d).$$

Thus, $X(t) - X(s) \sim \mathcal{N}(0, (t-s)I_d)$.

And last we need to show that $X(t)$ has continuous paths. Since $\mathcal{B}(t)$ has continuous paths almost surely, the process $X(t) = x + \mathcal{B}(t)$ will also be continuous almost surely.

Then we conclude that $x + \mathcal{B}(t)$ is indeed a d-dimensional BM initiated at x. □

Proposition 1.3.4 (*Symmetry*)**.** *If $\{\mathcal{B}(t) : t \geq 0\}$ is a BM, then we have that $\{-\mathcal{B}(t) : t \geq 0\}$ is also a BM.*

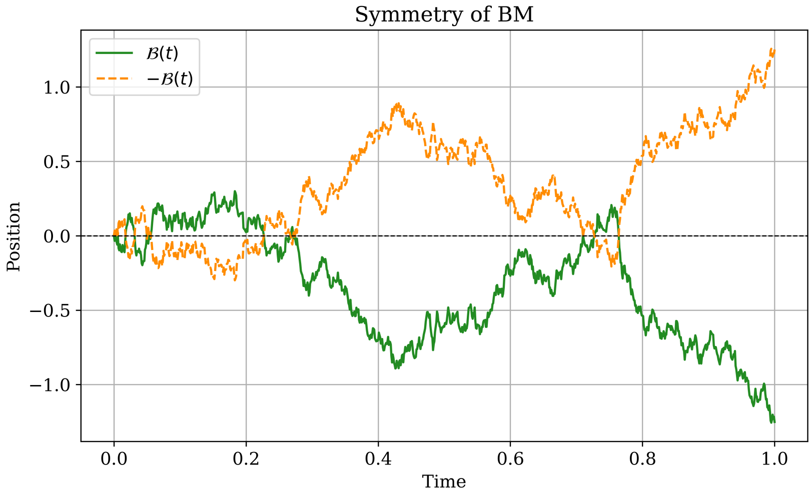

Figure 1.2: Standard BM $\mathcal{B}(t)$ and its reflection $-\mathcal{B}(t)$, illustrating the symmetry property.

Proof. The proof is similar to the previous one. In this case, we have to show that the process $X(t) = -\mathcal{B}(t)$ is also a d-dimensional BM.

We first verify the initial condition:

$$X(0) = -\mathcal{B}(0) = -0 = 0.$$

Thus, $X(0) = 0$ almost surely.

Next, we consider the increments of $X(t)$. For $0 \leq s < t$:

$$X(t) - X(s) = -\mathcal{B}(t) - (-\mathcal{B}(s)) = -[\mathcal{B}(t) - \mathcal{B}(s)].$$

Since the increments of $\mathcal{B}(t)$ are independent, it follows that $X(t)$ inherits this property and has independent increments.

We now show that the increments of $X(t)$ are normally distributed. For $t > s$:

$$X(t) - X(s) = -[\mathcal{B}(t) - \mathcal{B}(s)].$$

Given that $\mathcal{B}(t) - \mathcal{B}(s) \sim \mathcal{N}(0, (t-s)I_d)$, it follows by symmetry of the normal distribution that:

$$X(t) - X(s) \sim \mathcal{N}(0, (t-s)I_d).$$

Finally, we examine the continuity of paths. Since $\mathcal{B}(t)$ has continuous paths almost surely, it follows that $X(t) = -\mathcal{B}(t)$ also has continuous paths almost surely.

Therefore, we conclude that $X(t) = -\mathcal{B}(t)$ is indeed a d-dimensional BM.

□

The proofs of the following propositions are analogous to those presented before and follow from similar arguments. For brevity, the details are omitted and can be found in [7].

Proposition 1.3.5. *Let $\{\mathcal{B}(t) : t \geq 0\}$ be a d-dimensional BM. For a given $t > 0$, then for $\{\mathcal{B}(a) : 0 \leq a \leq t\}$,*

$$\{\mathcal{B}(t-a) - \mathcal{B}(t) : 0 \leq a \leq t\}.$$

is also a d-dimensional BM.

Proposition 1.3.6 (*Scaling*)**.** *Let $\{\mathcal{B}(t) : t \geq 0\}$ be a d-dimensional BM. For every $c > 0$, the process:*

$$\{\sqrt{c}\mathcal{B}(t/c) : t \geq 0\}.$$

is also a d−dimensional BM.

Proposition 1.3.7 (*Time Inversion*)**.** *Let $\{\mathcal{B}(t) : t \geq 0\}$ be a d-dimensional BM. The process given by:*

$$Z(t) = \begin{cases} 0, & t = 0, \\ t\mathcal{B}(\frac{1}{t}), & t > 0, \end{cases}$$

is also a d-dimensional BM.

1.3.3 The Markov property, the strong Markov property and the reflection principle

In the study of stochastic processes, the *Markov property* plays a fundamental role in understanding the evolution of processes over time. It states that the future behavior of a process depends only on its present state, independent of the past. The *strong Markov property* is an extension, particularly significant when we are dealing with random times, such as stopping times. It ensures that a process can be "restarted" at any stopping time, with the future evolution behaving as if it were starting anew, while maintaining the same statistical properties. A significant application of the strong Markov property is the *reflection principle* for BM, which is used to analyze the behavior of the process after hitting a boundary. This principle provides an important method for deriving probabilities related to the maximum of BM, making it useful in both theoretical and applied probability. We will first review the basic notions of the Markov property and the strong Markov property, and then demonstrate how the reflection principle can be derived using these properties.

The Markov property states that the process has no memory of its past once its current state is known. More formally, for BM, this means that for any fixed time s, the future evolution of the process after time s depends only on the state of the process at s, and is independent of the history before that time. This property enables us to treat the increments of the process after time s as if the process were starting anew. As illustrated in Figure 1.3, this lack of memory allows for the "restarting" of the process at any stopping time, preserving the distributional properties of BM.

Theorem 1.3.8 (Markov property)**.** *Let $\{\mathcal{B}(t) : t \geq 0\}$ be a d-dimensional BM. Then for any fixed time $s > 0$, the process*

$$\{\mathcal{B}(t+s) - \mathcal{B}(s) : t \geq 0\}.$$

is a d-dimensional BM independent of the process $\{\mathcal{B}(t) : 0 \leq t \leq s\}$.

Proof. Let $\{\mathcal{B}(t) : t \geq 0\}$ be a d-dimensional BM started at some point $\mathbf{x} \in \mathbb{R}^d$. Fix $s > 0$. Define the process

$$\tilde{\mathcal{B}}(t) := \mathcal{B}(t+s) - \mathcal{B}(s), \quad t \geq 0.$$

We aim to show that $\tilde{\mathcal{B}}(t)$ is a standard d-dimensional BM that is independent of the past process $\{\mathcal{B}(t) : 0 \leq t \leq s\}$.

1. *Initial condition:* Clearly, $\tilde{\mathcal{B}}(0) = \mathcal{B}(s) - \mathcal{B}(s) = \mathbf{0}$.

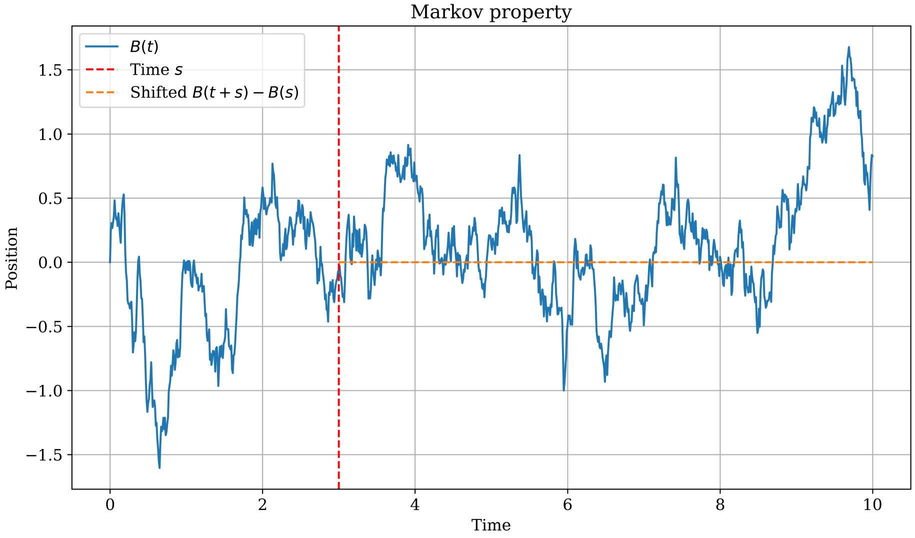

Figure 1.3: The future increments of the process from time s are independent of its past, depending only on the current state at s.

2. *Independent increments:* Let $0 \leq t_1 < t_2 < \cdots < t_n$. Since $\mathcal{B}(t)$ has independent increments, the random vectors

$$\tilde{\mathcal{B}}(t_1),\ \tilde{\mathcal{B}}(t_2) - \tilde{\mathcal{B}}(t_1),\ \ldots,\ \tilde{\mathcal{B}}(t_n) - \tilde{\mathcal{B}}(t_{n-1})$$

are just

$$\mathcal{B}(t_1+s) - \mathcal{B}(s),\ \mathcal{B}(t_2+s) - \mathcal{B}(t_1+s),\ \ldots,\ \mathcal{B}(t_n+s) - \mathcal{B}(t_{n-1}+s),$$

which are independent by the independent increments property of $\mathcal{B}(t)$.

3. *Normal increments:* For $0 \leq t_1 < t_2$, we have:

$$\tilde{\mathcal{B}}(t_2) - \tilde{\mathcal{B}}(t_1) = \mathcal{B}(t_2+s) - \mathcal{B}(t_1+s) \sim \mathcal{N}(\mathbf{0}, (t_2 - t_1)I_d),$$

since BM has stationary, normally distributed increments.

4. *Continuity:* Since $\mathcal{B}(t)$ has continuous sample paths, and we have that $\tilde{\mathcal{B}}(t) = \mathcal{B}(t+s) - \mathcal{B}(s)$, it follows that $\tilde{\mathcal{B}}(t)$ is also continuous in t.

5. *Independence from the past:* The increments $\mathcal{B}(t+s) - \mathcal{B}(s)$ are independent of $\{\mathcal{B}(u) : 0 \leq u \leq s\}$ due to the independent increments property. Therefore, $\tilde{\mathcal{B}}(t)$ is independent of $\mathcal{F}_s$, the filtration generated by $\mathcal{B}(u)$ up to time s.

Thus, $\tilde{\mathcal{B}}(t)$ is a standard d-dimensional BM, completing the proof. □

In order to define the *strong Markov property*, it is necessary to first introduce two fundamental concepts.

Definition 1.3.9 (Filtration)**.** Let $\{\mathcal{B}(t) : t \geq 0\}$ be a stochastic process. A filtration $\{\mathcal{F}_t : t \geq 0\}$ is a family of σ-algebras, where each $\mathcal{F}_t$ represents the information available up to time t. More formally, a filtration is a non-decreasing sequence of σ-algebras, meaning that for any $0 \leq s < t$, we have $\mathcal{F}_s \subseteq \mathcal{F}_t$. In the context of BM, $\mathcal{F}_t$ typically represents all the information about the path of the process up to time t.

Definition 1.3.10 (Stopping Time)**.** Let $\{\mathcal{B}(t) : t \geq 0\}$ be a BM and $\{\mathcal{F}_t : t \geq 0\}$ be a filtration associated with the process. A random variable T is called a *stopping time* if, for every $t \geq 0$, the event $\{T \leq t\}$ is in the σ-algebra $\mathcal{F}_t$, i.e., it is measurable with respect to the information available up to time t. Formally, we require that:

$$\{T \leq t\} \in \mathcal{F}_t \quad \text{for all} \quad t \geq 0.$$

This means that the decision of whether the stopping time has occurred by time t can be made using the information available up to that time.

The *strong Markov property* is a fundamental tool in the study of BM. While the Markov property asserts that the future evolution of a process depends only on its present state, the strong Markov property extends this to *random* times known as stopping times. This property allows BM to be "restarted" at any stopping time, behaving as if it were a new, independent BM starting from that point, while maintaining its fundamental distributional properties.

We do not present a full proof here, but one may be found in [40] and [22].

Theorem 1.3.11 (Strong Markov property)**.** *Let $\{\mathcal{B}(t) : t \geq 0\}$ be a BM, and let $\mathcal{F}_t$ be the filtration of the process up to time t. For any stopping time τ, the process:*

$$\{\mathcal{B}(t + \tau) - \mathcal{B}(\tau) : t \geq 0\}.$$

is a BM, independent of $\mathcal{F}_\tau$, the filtration up to time τ.

This result states that the future increments of BM after a stopping time τ depend only on the current state at that time and not on the path taken to reach it. More detailed information on this can be found in [52].

We have already stated the strong Markov property, which claims that, given a stopping time, the future evolution of a process depends only on its state at that stopping time and is independent of its past. This property plays an important role in understanding BM. In the following subsection, we will explore the *reflection principle*, which is a direct application of the strong Markov property. The reflection principle uses the fact that once a BM reaches a specific level, its future evolution can be treated as if it were starting anew from that level, independent of the path taken to reach it. Using this principle, we can compute probabilities for BM with boundaries.

The reflection principle

Imagine a particle undergoing random motion along a line, such as pollen floating on the surface of water. Consider a fixed time horizon. A natural question arises: what is the probability that this particle crosses a certain level before this specified time? The answer is not immediately obvious. However, the symmetry inherent in BM provides a tool to resolve this question: the *reflection principle.*

The reflection principle relies on the symmetry of Brownian paths. It shows that if the path crosses a given level at a stopping time, the portion of the path beyond this level can be reflected across it, resulting in a new valid Brownian path with the same probabilistic properties. This reflection allows us to compute the probability of reaching a barrier by relating it to the probability of exceeding the level at the final time. In doing so, it transforms the problem of finding the probability that the supremum of the process over time exceeding a level into a simpler problem of finding the probability of the process being above that level at the final time.

This construction is illustrated in Figure 1.4, where a Brownian path is reflected at the level a after the *first hitting time* T_a. To accurately state the reflection principle, we first introduce the concept of *first hitting time.*

Definition 1.3.12 (First hitting time)**.** Let $\{\mathcal{B}(t) : t \geq 0\}$ be a BM and $a \in \mathbb{R}$. The *first hitting time* T_a of level a is the first time that the process $\mathcal{B}(t)$ reaches or exceeds the value a. Formally, it is defined as:

$$T_a = \inf\{t \geq 0 : \mathcal{B}(t) = a\}.$$

If the process never reaches a, then $T_a = \infty$.

We now formally present the reflection principle theorem, which, in its fundamental form, applies to one-dimensional BM due to the straightforward nature of reflection across a single level in a line.

Theorem 1.3.13 (Reflection principle)**.** *Let* $(\mathcal{B}(t))_{t\geq 0}$ *be a standard one-dimensional BM, and fix* $a > 0$ *and* $T > 0$*. Then the probability that the*

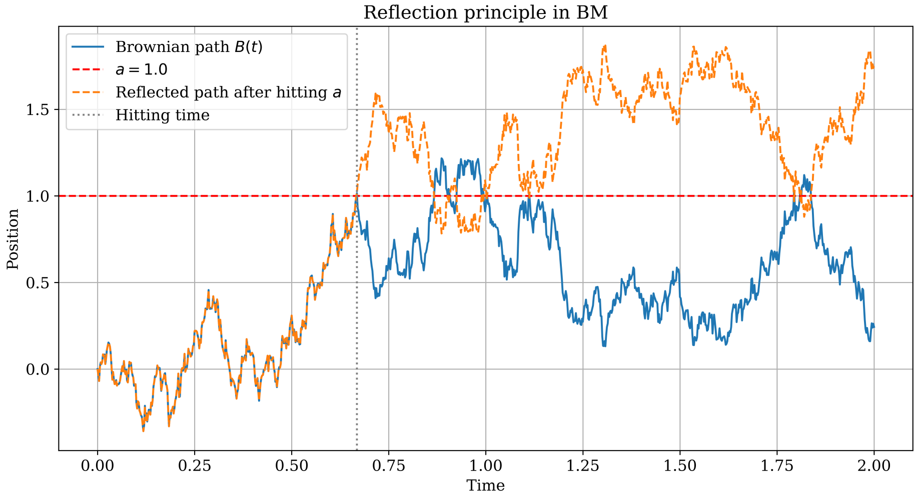

Figure 1.4: The Brownian path $\mathcal{B}(t)$ reaches level a for the first time at the hitting time T_a. The portion of the path after this time is reflected across level $a = 1$, illustrating the reflection principle.

BM reaches or exceeds level a before time T is given by:

$$\mathbb{P}\left(\sup_{0\leq t\leq T} \mathcal{B}(t) \geq a\right) = 2\,\mathbb{P}(\mathcal{B}(T) \geq a).$$

Proof. Let $T_a = \inf\{s \geq 0 : B(s) = a\}$ denote the first hitting time of the level a. Consider the event $\{\sup_{s\leq t} \mathcal{B}(s) \geq a, \mathcal{B}(t) < a\}$. For each path in this event, reflect the part of the path after T_a over the level a to construct a new path:

$$\widetilde{\mathcal{B}}(s) = \begin{cases} \mathcal{B}(s), & s \leq T_a, \\ 2a - \mathcal{B}(s), & s > T_a. \end{cases}$$

By the strong Markov property and the symmetry of BM, this new process has the same distribution as a BM starting from a. Thus, the probability of paths that reach a before time T is twice the probability that the process ends above a, yielding the result:

$$\mathbb{P}\left(\sup_{0\leq t\leq T} \mathcal{B}(t) \geq a\right) = 2\mathbb{P}(\mathcal{B}(T) \geq a).$$

□

The reflection principle is attributed to D. André, as demonstrated in André's work [2]. The original problem he addressed was as follows: if two candidates in a ballot receive a and b votes respectively, with $a > b$, what is the probability that the first candidate was always ahead throughout the counting process? More information on this problem can be found in [25]. A formulation of the reflection principle for BM was later given by Lévy in [50], though it is important to note that Lévy's formulation was established independently of the strong Markov property.

In one physical dimension, the Euclidean norm process is simply the absolute difference between positions. In this case, we can take advantage of the reflection principle to model the process efficiently. However, in higher dimensions, the situation becomes more complex. The Euclidean distance is no longer a straightforward difference between positions, and instead, it involves the root of the sum of squared positions. So, we cannot apply the reflection principle in this case. Because we are interested in this the Euclidean norm process, we see that it is equivalent to a Bessel process. This has the advantage of being not restricted to integer dimensions. This allows us to work with properties in non-integer dimensions, which we will be working on.

1.4 Bessel processes

The *Bessel process* is a type of stochastic process closely related to BM. It is often used to model random motion in multi-dimensional spaces, especially when expressed in radial coordinates. Specifically, a Bessel process represents the radial component of BM in $\mathbb{R}^d$. It describes the distance from the origin of a d-dimensional BM, describing how far the motion is from the origin over time, without keeping track of the direction. For example, consider a particle undergoing two- or three-dimensional BM. While the full process tracks its exact position in space, the associated Bessel process records only how far the particle is from its starting point. This radial component defines a stochastic process, whose dynamics depend on the dimension d of the BM.

1.4.1 Definition of Bessel processes

Definition 1.4.1 (SDE definition of Bessel process). A *δ-dimensional Bessel process* is defined as the solution $\{X_t : t \geq 0\}$ to the SDE

$$dX_t = \frac{\delta - 1}{2X_t}\, dt + d\mathcal{B}(t), \quad X_0 \geq 0,$$

where $\{\mathcal{B}(t)\}$ is a standard one-dimensional BM.

Geometric interpretation for Bessel process When $\delta = d \in \mathbb{N}$, the Bessel process coincides in distribution with the Euclidean norm of a d-dimensional BM:

$$X_t = \|\mathcal{B}(t)\| = \sqrt{\mathcal{B}_1(t)^2 + \cdots + \mathcal{B}_d(t)^2},$$

where $\mathcal{B}(t) = (\mathcal{B}_1(t), \ldots, \mathcal{B}_d(t))$ is a standard d-dimensional BM starting at the origin. Thus, in integer dimensions, the Bessel process describes the radial distance of BM from the origin.

The *SDE* formulation, can be extended to non-integer dimensions, providing a broader analytical framework. For example, the Bessel process can describe the random movement of a particle in d-dimensional space. This framework also extends naturally to modeling various phenomena such as the migration of organisms, cellular motion within a spatial domain, or the movement of animals. It effectively captures the inherent randomness present in such movement patterns, as we will see in Chapter 3.

1.4.2 Properties of Bessel processes

There are several properties of Bessel processes that are relevant both for theoretical understanding and practical modeling:

1. *Non-negativity and continuity:* Bessel processes are continuous and remain non-negative for all $t \geq 0$, making them suitable for modeling distances and other quantities that cannot take negative values.

2. *Scaling property:* Bessel processes satisfy the scaling relation, for any constant $c > 0$,
$$R(ct) = \sqrt{c}\, R(t),$$
This property is useful in both simulations and theoretical analysis.

3. *Representation from BM:* For integer $d \in \mathbb{N}$, the Bessel process can be constructed as the Euclidean norm of d independent standard BMs:
$$R(t) = \sqrt{B_1(t)^2 + \cdots + B_d(t)^2}.$$
The process $\{R(t) : t \geq 0\}$ describes the distance of a Brownian particle from the origin at time t. It represents the radial component of BM in $\mathbb{R}^d$.

4. *Behavior near the origin:* For $d < 2$, the process almost surely hits zero in finite time. For $d = 2$, the process comes arbitrarily close to the origin infinitely often but never hits it (null recurrent). For $d > 2$, the process remains strictly positive with probability one if it starts away from zero. Although the Bessel process is often introduced starting at the origin, this is not required in general. We will see this property in the following sections.

1.4.3 Simulations of Bessel processes

Simulating Bessel processes is a useful way to visualize their stochastic behavior and how it depends on the dimension parameter δ. While Bessel processes are defined for any real $\delta > 0$, they can be simulated directly when $\delta = d \in \mathbb{N}$, since they correspond to the Euclidean norm of a d-dimensional BM:

$$R(t) = \|\mathcal{B}(t)\| = \sqrt{B_1(t)^2 + \cdots + B_d(t)^2}.$$

These simulations provide a concrete sense of how the process behaves for different values of d, especially around the critical threshold $\delta = 2$, which separates recurrent from transient behavior (a topic we explore in the next section). For instance, we can observe how in low dimensions the process tends to stay close to the origin, while in higher dimensions it tends to move away. We present the Python code used to simulate Bessel processes for integer values of $\delta = d$.

```python
import numpy as np
import matplotlib.pyplot as plt

def simulate_bessel(d, T=1.0, n=1000, seed=None):
    if seed:
        np.random.seed(seed)
    dt = T / n
    B = np.cumsum(np.sqrt(dt) * np.random.randn(d, n), axis=1)
    R = np.linalg.norm(B, axis=0)
    return np.linspace(0, T, n), R

# Example for d = 1, 2, 3
dimensions = [1, 2, 3]
plt.figure(figsize=(10, 6))
for d in dimensions:
    t, R = simulate_bessel(d)
    plt.plot(t, R, label=f'd = {d}')
plt.title('Simulated Bessel Processes for Different Dimensions')
plt.xlabel('Time')
plt.ylabel('R(t)')
plt.legend()
plt.grid(True)
plt.show()
```

Listing 1.1: Simulation of a Bessel process with integer dimension d.

Figure 1.5 shows three simulated paths:

- For $d = 1$, the process behaves like reflected BM and frequently returns to the origin. It is a recurrent process.

- For $d = 2$, the process avoids the origin but still returns arbitrarily close to it with probability one. It is neighborhood-recurrent.

- For $d = 3$, the process tends to drift away from the origin and is transient, meaning it avoids the origin in the long run.

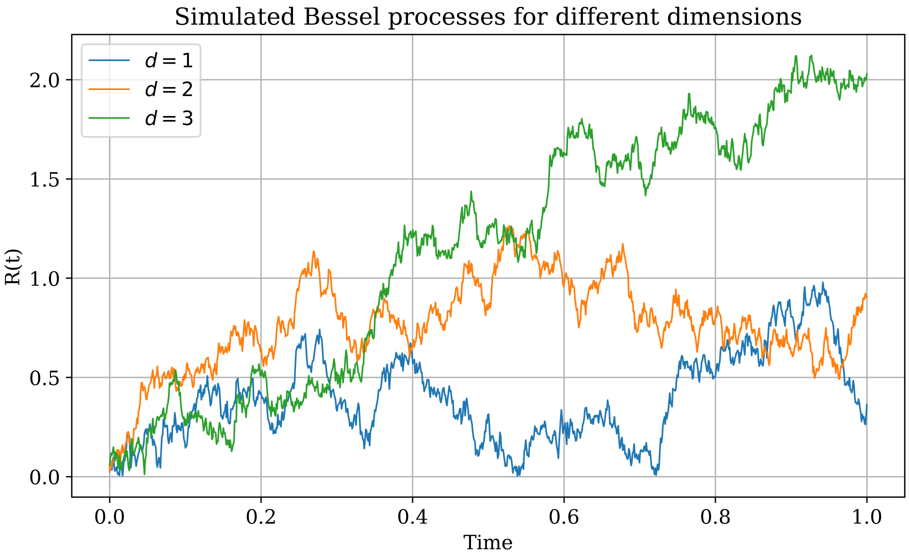

Figure 1.5: Sample paths of Bessel processes for dimensions $d = 1, 2, 3$.

1.5 Recurrence and transience of BM and Bessel processes

In the study of stochastic processes, the concepts of *recurrence* and *transience* describe whether a process tends to return to a particular state (usually the origin) or instead tends to escape to infinity. These properties depend crucially on the dimension of the space in which the process evolves.

Let $\mathcal{B}(t)$ denote BM in $\mathbb{R}^d$, starting at the origin. To determine whether the process is recurrent or transient, we first distinguish between two types of recurrence:

- *Point recurrence*: the process returns to its exact starting point with probability one:
$$\mathbb{P}(\exists t > 0 : \mathcal{B}(t) = 0) = 1.$$

- *Neighborhood recurrence*: for every $\varepsilon > 0$, the process enters the ball $B(0, \varepsilon)$ infinitely often:
$$\mathbb{P}\left(\mathcal{B}(t) \in B(0, \varepsilon) \text{ for infinitely many } t > 0\right) = 1.$$

The following properties we present demonstrate that the recurrence or transience of BM depends entirely on the spatial dimension d:

- For $d = 1$, BM is *point recurrent.* It returns to the origin infinitely often with probability one.

- For $d = 2$, BM is *neighborhood recurrent* but is not point recurrent. That is, it almost surely visits every neighborhood of the origin infinitely often, but the probability of hitting the exact origin is zero.

- For $d > 2$, BM is *transient.* With probability one, the process eventually escapes to infinity and visits any bounded region only finitely many times.

We note that the transition at $d = 2$ is critical and leads to a change in the qualitative behavior of the process.

This result highlights how spatial dimension determines the qualitative behavior of BM. In lower dimensions ($d = 1, 2$), the process exhibits a recurrent behavior, returning to (or near) the origin with probability one. In higher dimensions ($d > 2$), the process is transient and eventually escapes to infinity. This has important implications in probability theory and mathematical biology, where dimension affects whether stochastic trajectories revisit earlier states or diverge over time, as we will see in Chapter 3.

The Bessel process R_t of order ν can be defined as the Euclidean norm of BM in $\mathbb{R}^d$, where the dimension and order are related by
$$\nu = \frac{d}{2} - 1.$$

That is, $R_t = \|\mathcal{B}(t)\|$, so the Bessel process represents the *radial part* of BM, as discussed in the previous section.

As with BM, the recurrence or transience of Bessel processes depends on the dimension d (or equivalently, the order ν). The difference is that Bessel processes are always non-negative.

- For $d < 2$ (i.e., $\nu < 0$), the Bessel process is recurrent at zero and hits the origin infinitely often.
- For $d = 2$ (i.e., $\nu = 0$), the Bessel process is *null recurrent*: it approaches zero arbitrarily closely infinitely often but never actually hits it.
- For $d > 2$ (i.e., $\nu > 0$), the Bessel process is transient: it drifts away from the origin over time and almost surely does not return.

We now present simulated paths of Bessel processes for three different values of the dimension parameter d, which determines the process behavior near the origin.

In the first case (Figure 1.6), where $d = 1$, the Bessel process displays recurrent behavior by frequently approaching and eventually hitting the origin. This reflects the property that when $d < 2$, the origin can be reached in finite time with probability one. Once it reaches zero, the process is absorbed, confirming its recurrence at the origin.

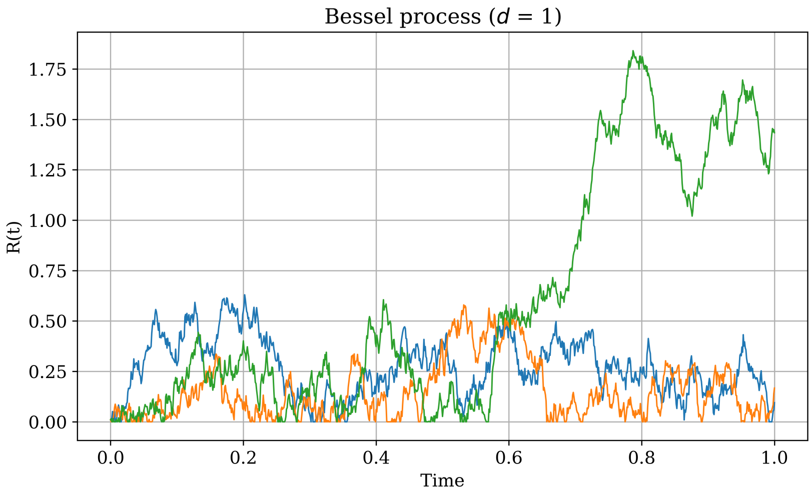

Figure 1.6: Bessel process for $d = 1$.

In the second case (Figure 1.7), where $d = 2$, we observe the critical behavior of the Bessel process. Although the process never actually hits the origin, it returns arbitrarily close to it infinitely often. This illustrates the property that for $d = 2$, the origin cannot be reached in finite time, but it remains recurrent in the sense that the process repeatedly visits neighborhoods around zero. Unlike the case $d < 2$, the origin is not absorbing, yet the process does not escape to infinity either.

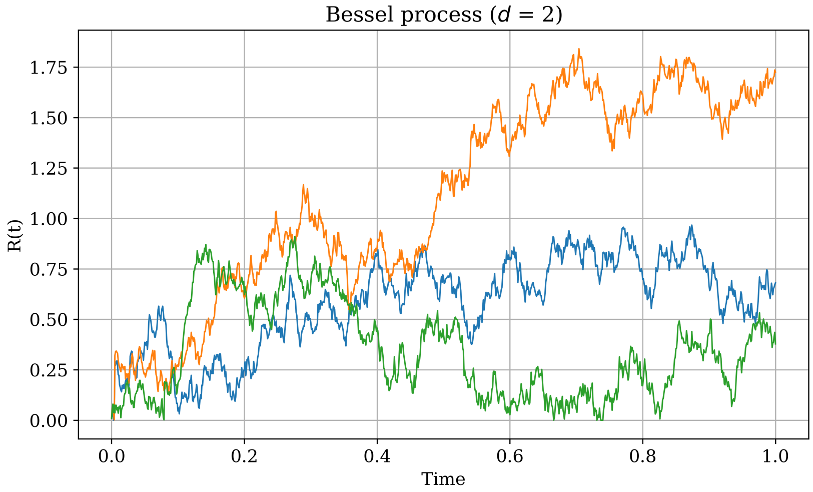

Figure 1.7: Bessel process for $d = 2$.

In the final case (Figure 1.8), where $d = 3$, the Bessel process exhibits clear transience. The path quickly drifts away from the origin and continues to grow without returning. This reflects the fact that for $d > 2$, the process tends to escape to infinity and, with probability one, does not return to any neighborhood of the origin. The process develops a persistent tendency to move away from zero, leading it to almost surely avoid the origin.

These simulations offer a clear and intuitive demonstration of how dimensionality shapes the long-term dynamics of Bessel processes. They confirm the theoretical result that recurrence and transience are completely determined by the dimension, with the critical threshold at $d = 2$ distinguishing recurrent behavior for $d \leq 2$ from transient behavior when $d > 2$.

1.6 The Skorokhod embedding theorem

Having explored the recurrence and transience properties of BM and Bessel processes, we now turn to the *Skorokhod embedding theorem*, a result in stochastic process theory.

The Skorokhod embedding theorem is an important result in probability theory. It states that if X is a real-valued random variable with mean zero

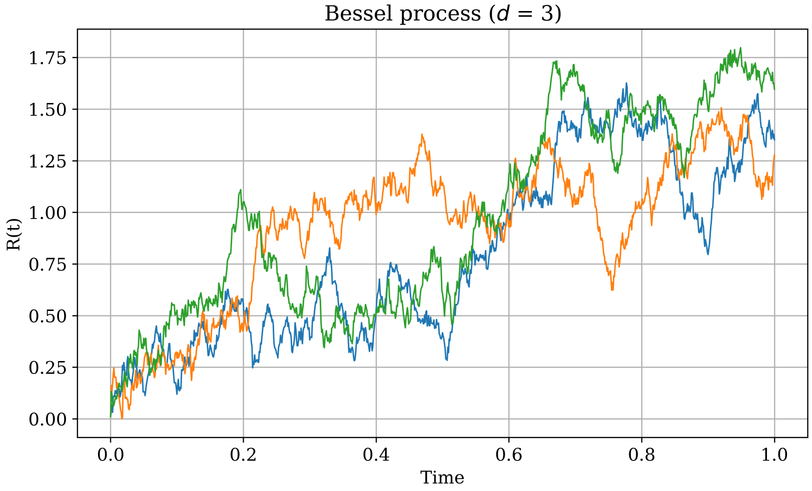

Figure 1.8: Bessel process for $d = 3$.

and finite variance, then there exists a stopping time τ such that a standard BM $\{\mathcal{B}(t)\}_{t\geq 0}$ stopped at time τ satisfies $\mathcal{B}(\tau) \sim X$.

This theorem is useful because it shows how certain random variables can be represented as the value of BM at a suitable random time. In this thesis, we mention the Skorokhod embedding theorem because it is what allows us map stochastic calculus onto the real world. It justifies the mapping from random fluctuations happening in infinitessimal time intervals (as $dt \to 0$) onto random fluctuations happening in real time intervals, so this means we can simulate stochastic processes. This theorem can be applied to any stochastic process, such as Bessel processes. The following theorem is presented in [67].

Theorem 1.6.1 (Skorokhod embedding theorem)**.** *Let X be a real-valued random variable with $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^2] < \infty$. Then there exists a stopping time τ with respect to the natural filtration of a standard BM $\{\mathcal{B}(t)\}_{t\geq 0}$ such that*

$$\mathcal{B}(\tau) \sim X \quad \textit{and} \quad \mathbb{E}[\tau] = \mathbb{E}[X^2].$$

The construction of such a stopping time provides a way to embed the distribution of X into a Brownian path, which is the idea behind many methods in stochastic simulation and optimal stopping theory. This theorem has many applications, particularly in simulating complex stochastic processes such as BM and Bessel processes.

The Skorokhod embedding theorem provides a method to transform a given random variable into a Brownian path through an appropriate stopping time. This is useful in situations where direct simulation of the process is difficult or impossible. For example, in biology and physics, the theorem helps in modeling processes such as the movement of particles (e.g., BM) or organisms (e.g., Bessel processes) under random influences.

By using the Skorokhod embedding, one can derive realistic simulations of such processes that closely mimic real-world phenomena, while maintaining mathematical rigor and ensuring computational feasibility.

1.7 SDE formulation

Having established the theory of the Skorokhod Embedding Theorem, we now present the dynamics of Bessel processes.

SDE are important for modeling systems influenced by randomness or random processes. In this section, we focus on the norm $\|\mathcal{B}(t)\|$ of n-dimensional BM $\mathcal{B}(t)$ and its Markovian properties. We explore the relationship between this norm and the squared Bessel process, deriving the corresponding SDE. Also, we provide numerical solutions to illustrate the dynamics of the Bessel process.

The norm $\|\mathcal{B}(t)\|$ of an n-dimensional BM is Markov. More precisely, letting $\mathcal{F}_t = \sigma(\mathcal{B}(s) : s \leq t)$ be its natural filtration, then $X \equiv \|\mathcal{B}\|^2$ has the following property. For times $s < t$, conditional on $\mathcal{F}_s, X(t)/(t-s)$ is distributed as $\chi^2_n (X(s)/(t-s))$. This is known as the "n-dimensional" squared Bessel process, and denoted by BES ^2_n.

Alternatively, the process X can be described by a SDE. Applying integration by parts,

$$dX = 2\sum_i \mathcal{B}^i d\mathcal{B}^i + \sum_i d\left[\mathcal{B}^i\right]. \tag{1.2}$$

As the standard BMs have quadratic variation $\left[\mathcal{B}^i\right]_t = t$, the final term on the right-hand side is equal to ndt. Also, the covariations $[\mathcal{B}^i, \mathcal{B}^j]$ are zero for $i \neq j$ from which it can be seen that

$$W_t = \sum_i \int_0^t 1^{\mathcal{B}^i}_{\{\mathcal{B}\neq 0\}} d\mathcal{B}^i. \tag{1.3}$$

is a continuous local martingale with $[W]_t = t$. By Lévy's characterization, W is a BM and, substituting this back into (2), the squared Bessel process X solves the SDE:

$$dX = 2\sqrt{X}dW + ndt.$$

The Bessel process itself (not the squared process) has SDE. If $x > 0$ then $X(t)$ satisfies:

$$dX_t = \frac{\nu}{X_t}dt + d\mathcal{B}_t, \quad X_0 = x.$$

where $\nu = (n-2)/2$ and n is the number of dimensions.

Figures 1.9 and 1.10 illustrate the ensemble dynamics of the Bessel process for two distinct values of $\mathcal{D}$, showing the numerical solution of the SDE.

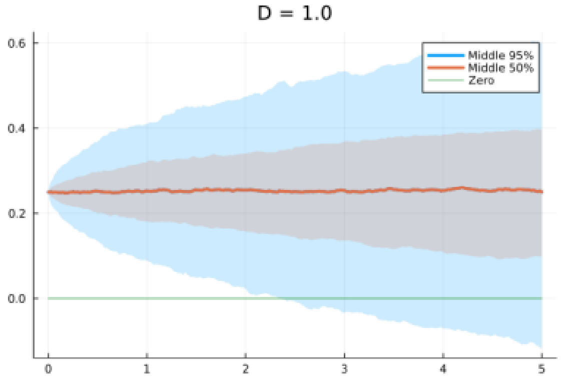

Figure 1.9: Ensemble of simulated paths for a Bessel process with dimension $\mathcal{D} = 1.0$, obtained by numerically solving the corresponding SDE given in Equation 1.2.

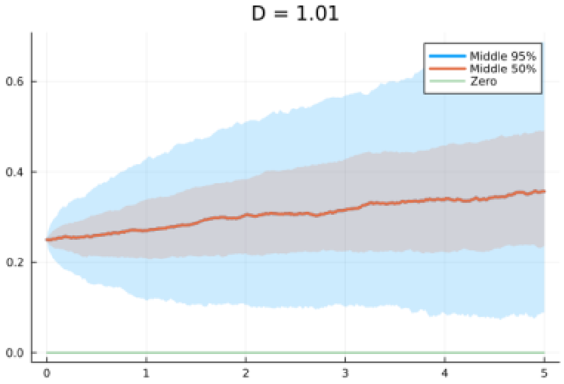

Figure 1.10: Ensemble of simulated paths for a Bessel process with dimension $\mathcal{D} = 1.01$, obtained by the numerical solution of the SDE. It is important to note the transition to non-negative process.

These Figures provide insight into how the Bessel process behaves as the dimension $\mathcal{D}$ varies. As we approach the critical dimension $\mathcal{D} = 2$, the process begins to exhibit a different qualitative behavior, transitioning from a regime that goes toward the origin to one that drifts away.

In Figure 1.9, it shows the evolution of the Bessel process in one dimension ($\mathcal{D} = 1.0$) based on a numerical solution to the corresponding SDE.

In this case, the process tends to move towards the origin, as we are in a dimension lower than 2, which is consistent with the theoretical understanding of Bessel processes in subcritical dimensions. As the process evolves, we see that the path often fluctuates, but is more likely to approach zero rather than drift away. This highlights the tendency of the process to return to the origin, which is a characteristic of Bessel processes when $\mathcal{D} < 2$.

In Figure 1.10, we simulate the Bessel process in a slightly higher dimension, $\mathcal{D} = 1.01$. At this value, the process exhibits behavior that starts to transition toward non-negative values, as the dimension is now very close to the critical value of two. While the Bessel process still tends to return to the origin in this dimension, we begin to observe a slight repulsion from the origin, which is characteristic of processes in dimensions slightly higher than two. This illustrates the shift from the recurrent behavior typical of subcritical dimensions (approaching the origin) to the more stable, outward-drifting behavior seen in dimensions greater than two.

Chapter 2

Classical population genetics

The study of *population genetics* is an important field in biology, more specifically in the study of genetic variation, DNA (Deoxyribonucleic Acid) data, coalescence processes and evolutionary biology. The concept of effective population in classical population genetics serves as a stochastic equivalence that reflects how a population behaves under certain conditions. We are going to look for this effective population size, which can fluctuate due to varying degrees of inbreeding, to represent an ideal population that operates with a constant size. In this work, we will begin by having a background on the history of population genetics and summarizing terminology that is useful to genetics, which we may refer to throughout this chapter. The following books will be particularly useful for this chapter: [27], [20], [34], [32].

2.1 Historical background

In the introduction, we have seen the historical background of population genetics. Now, we will summarize what was previously mentioned and motivate the work we have been doing. Important researchers such as Sewall Wright, Ronald A. Fisher, and J. B. S. Haldane were instrumental in shaping the field of population genetics, establishing principles of genetic variation influenced by factors such as migration, selection, and random genetic drift.

Wright's paper, see [78], introduced the concept of genetic drift and the notion of effective population size, emphasizing the role of random processes in evolution. Fisher's paper [26], provided a mathematical framework for understanding natural selection and its effects on genetic variation, merging Mendelian genetics with Darwinian theory. Haldane's article [30], expanded on Fisher's work, focusing on the mathematical models of selection processes and the dynamics of gene frequency changes over time. Together, these contributions show how important it is to integrate math with genetics, as this helps us better understand the processes of evolution.

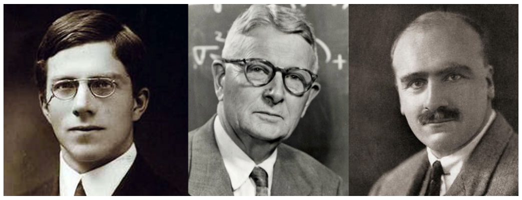

Figure 2.1: Ronald A. Fisher, Sewall Wright, and J. B. S. Haldane

2.2 Genetics basic concepts

We will introduce some fundamental terminology that will be used throughout the thesis. Although not every term may be explicitly addressed later, these basic concepts are useful to understanding genetics.

The first step might be to define *genes*, which are the fundamental units of heredity passed from parents to offspring. A *gene* is a segment of DNA (Deoxyribonucleic Acid) that contains the instructions for synthesizing a specific protein or functional RNA (Ribonucleic Acid). Each gene has a defined location, called a *locus*, on a chromosome and can influence particular characteristics, such as the color of the eyes or susceptibility to diseases. Different versions of a gene are called *alleles*, which are crucial in the following theory because we want to understand how their frequencies change over time in a population. Earlier, we mentioned *chromosomes*, the structures that contain genetic material. In humans, there are 23 pairs of chromosomes. However, the number of pairs of chromosomes varies between different organisms.

It is interesting to mention that the *genetic code* is a list of all *codons*, each of which corresponds to a specific amino acid. *Codons* are triplets of *nucleotides*, which are the basic building blocks of nucleic acids, such as DNA or RNA. Each nucleotide consists of three components: a nitrogenous base (adenine (A), guanine (G), cytosine (C), thymine (T), or uracil (U)), a sugar molecule (deoxyribose in DNA or ribose in RNA), and a phosphate group. For example, a codon could be UUU, AAU or CAU in RNA. In contrast, in DNA, thymine (T) is used instead of uracil (U), so the corresponding DNA codons would use T instead of U. The corresponding listing all the codons is provided in 2.1.

We can now explore the concept of *genotype*, which refers to the specific genetic constitution of an organism, including all alleles present for a given

Codon	**Amino Acid**	**Codon**	**Amino Acid**
TTT	Phenylalanine (Phe)	TTC	Phenylalanine (Phe)
TTA	Leucine (Leu)	TTG	Leucine (Leu)
CTT	Leucine (Leu)	CTC	Leucine (Leu)
CTA	Leucine (Leu)	CTG	Leucine (Leu)
ATT	Isoleucine (Ile)	ATC	Isoleucine (Ile)
ATA	Isoleucine (Ile)	ATG	Methionine (Met, Start)
GTT	Valine (Val)	GTC	Valine (Val)
GTA	Valine (Val)	GTG	Valine (Val)
TCT	Serine (Ser)	TCC	Serine (Ser)
TCA	Serine (Ser)	TCG	Serine (Ser)
CCT	Proline (Pro)	CCC	Proline (Pro)
CCA	Proline (Pro)	CCG	Proline (Pro)
ACT	Threonine (Thr)	ACC	Threonine (Thr)
ACA	Threonine (Thr)	ACG	Threonine (Thr)
GCT	Alanine (Ala)	GCC	Alanine (Ala)
GCA	Alanine (Ala)	GCG	Alanine (Ala)
TAT	Tyrosine (Tyr)	TAC	Tyrosine (Tyr)
TAA	Stop	TAG	Stop
CAT	Histidine (His)	CAC	Histidine (His)
CAA	Glutamine (Gln)	CAG	Glutamine (Gln)
AAT	Asparagine (Asn)	AAC	Asparagine (Asn)
AAA	Lysine (Lys)	AAG	Lysine (Lys)
GAT	Aspartic Acid (Asp)	GAC	Aspartic Acid (Asp)
GAA	Glutamic Acid (Glu)	GAG	Glutamic Acid (Glu)
TGT	Cysteine (Cys)	TGC	Cysteine (Cys)
TGA	Stop	TGG	Tryptophan (Trp)
CGT	Arginine (Arg)	CGC	Arginine (Arg)
CGA	Arginine (Arg)	CGG	Arginine (Arg)
AGT	Serine (Ser)	AGC	Serine (Ser)
AGA	Arginine (Arg)	AGG	Arginine (Arg)
GGT	Glycine (Gly)	GGC	Glycine (Gly)
GGA	Glycine (Gly)	GGG	Glycine (Gly)

Table 2.1: The complete genetic code showing all 64 DNA codons and their corresponding amino acids. The start codon (ATG) and stop codons (TAA, TAG, TGA) are indicated.

set of genes. It represents the genetic information that determines the potential traits (or *phenotypes*) of an organism. However, the genotype does

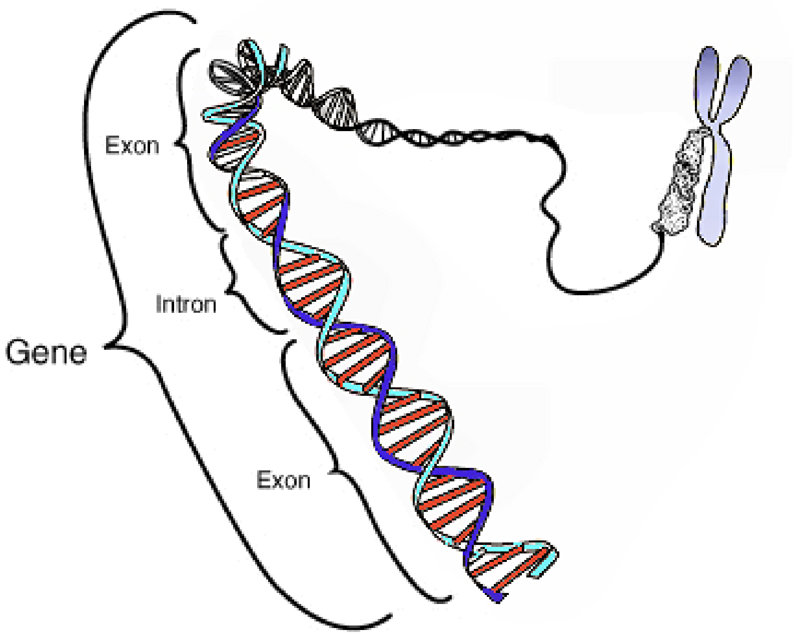

Figure 2.2: Representation of a gene illustrating its structure within DNA.

not always result in visible characteristics, as its expression can be influenced by interactions with the environment and other genetic factors. For example, an individual may have alleles for brown eyes (B) and blue eyes (b), but the expression of these traits depends on dominance relationships and environmental influences.

2.2.1 Population genetics basic concepts

This was a brief introduction to genetic concept. Now, we will introduce some of the concepts that will be using in population genetics.

One of the first concepts we will explore is the *population*. A population refers to a group of individuals of the same species living in a specific geographic area. This concept is of significant importance in population genetics because genetic variation arises within populations. The collective genetic material of all individuals in a population is known as the *gene pool*, encompassing all alleles for all genes present within the population.

We have already mentioned some concepts that are yet to be defined. *Genetic variation* refers to differences in alleles and genotypes between individuals within a population. This variation is fundamental to evolution, as it provides the raw material upon which natural selection acts. Genetic variation can arise through mutations, gene flow, and sexual reproduction, resulting in various characteristics within a population.

Next, we will look at the concept of *Hardy–Weinberg equilibrium*, a principle that describes a theoretical state in which allele and genotype frequencies in a population remain constant across generations, provided that certain conditions are met: such as no selection, mutation, migration or genetic drift. This equilibrium serves as a null hypothesis for studying evolutionary processes, allowing researchers to determine whether populations are evolving.

This next concept, *genetic drift*, will be explained in more detail in the next section. Genetic drift is a random process that causes allele frequencies to fluctuate over time due to chance events, particularly in small populations. Unlike natural selection, which is a non-random process, genetic drift can lead to the loss of genetic diversity and fixation of alleles (where only one allele is expressed in the population). As a result, genetic drift can drive significant evolutionary changes. Additionally, the term *selection*, which has not yet been defined, refers to the process by which certain characteristics become more common in a population due to their beneficial effects on an organism's survival and reproductive success. Natural selection acts on heritable traits, leading to adaptation.

Migration involves the movement of individuals between populations, leading to gene flow. This process can introduce new alleles into a population, increasing genetic diversity, and reducing differences between populations. Migration plays a crucial role in shaping genetic structure and dynamics. Next, *mutation* refers to a change in DNA sequence that introduces new genetic variations into a population. Mutations can be caused by errors in DNA replication, environmental factors, or radiation.

Gene flow refers to the transfer of alleles between populations through migration and mating. It contributed to genetic diversity within populations and can prevent populations from diverging genetically, playing a vital role in maintaining species integrity. So, the *genetic equilibrium* is a state in which allele frequencies in a population remain constant over time, typically under the Hardy-Weinberg equilibrium conditions. Finally, the *evolutionary processes* describe processes such as natural selection, genetic drift, mutation, and gene flow that drive changes in the genetic composition of populations over time.

These concepts provide a foundation for studying the dynamics of genetic variation and how it is shaped within populations over time through various evolutionary processes. For more precise definitions, see [53].

2.3 Random genetic drift

Before diving into the study of the concept of Effective Population Size, which we will refer to as EPS, it is necessary to understand the concept of *random genetic drift*. There are several mechanisms of evolution, such as natural selection, genetic drift, gene flow, and mutation. In this chapter, we will focus on random genetic drift.

In each generation, random changes can occur in the frequency of alleles in a population. These changes occur due to the random transmission of genetic material from one generation to the next. Even without selective pressures acting on the population, this randomness can cause some alleles to become more common while others decrease or even disappear. When only one allele remains in the population, this is called *fixation*. This means that all individuals in the population carry the same allele, and this characteristic becomes fixed for the entire population, meaning that genetic variation at that particular locus is lost. However, this could change if migration or mutation introduces new alleles into the population. Since these changes are not predictable or directed by any specific environmental factor, this process is known as random genetic drift.

While genetic drift is a stochastic process involving randomness, natural selection is non-random. In natural selection, certain traits provide a survival advantage in a particular environment. Organisms with those advantageous traits are more likely to survive, reproduce, and pass on their alleles. In contrast, genetic drift acts by random events rather than any specific selective pressures. For example, in an hypothetical scenario of a natural disaster, alleles that survive are chosen by chance, not because they offer any survival advantage. The alleles passed on to the next generation are not necessarily representative of the entire population, but instead are determined by random events. Some alleles might become overrepresented, while others may become less frequent or even lost entirely.

2.3.1 The Wright–Fisher and the Moran model

We have established the concept of genetic drift and its role in determining allele frequencies in populations, we can explore two foundational models that describe how this process works. Both the Wright–Fisher and Moran models illustrate the effects of genetic drift on population dynamics. While they share some similarities in their focus on randomness and stochastic processes, they differ in their assumptions about population structure and reproductive strategies. In this subsection, we will examine both models to better understand how they contribute to the random genetic drift in evolving populations.

The first model we will discuss was introduced by Sewall Wright (1931) [78] and Ronald A. Fisher (1930) [26]. Although they worked independently, they arrived at similar conclusions around the same time. This model, known as the Wright–Fisher model, provides a mathematical framework for understanding how the distribution of allele frequencies in populations changes from one generation to the next due to genetic drift.

In the Wright–Fisher model, the population we work with can be diploid or haploid, each with N individuals. A diploid population means that the individuals have two sets of chromosomes, one inherited from the mother and one from the father, having $2N$ copies of each gene, each of the individuals carrying two alleles at a specific genetic locus (for example it is usually expressed as A and a), This is the case in humans, and most of the animals. But we can also apply this model to haploid population, such as bacteria, in this case this organism has only one set of chromosomes and in this case only carrying one allele at each locus. The idea of the model is that, in each generation, individuals randomly pass on their alleles to the next generation through a process of random genetic sampling.

The second model we will briefly look at is the Moran model, introduced by Patrick Alfred Pierce Moran (1917-1988), who is known for his significant contributions to population genetics. He is best known for developing the Moran model, which was introduced in 1958 in his paper [51], a few years after the Wright-Fisher model discussed earlier. The Moran model shared similarities with the Wright-Fisher model, as both describe random genetic drift in populations. However, the Moran model also incorporates spatial structure into population genetics simulations, allowing the study of genetic drift and selection in structured populations.

The Wright-Fisher and Moran models differ in how generations are structured and how reproduction occurs. In the Wright-Fisher model, as we have seen, generations are discrete, meaning that individuals from one generation do not interact with those from the next. In contrast, the Moran model allows for overlapping generations, where individuals from different generations can coexist and interact within the same population.

Another difference between the Wright–Fisher model and the Moran models is that, in the first case, at each generation the entire population reproduces at once, meaning that all alleles randomly sampled form the individuals, and those alleles contribute to the next generation. And in the second case, individuals are chosen one at a time: one individual is randomly selected to reproduce, and the offspring may replace another individual that dies or needs to be replaced. This structure allows us to observe genetic drift on an individual level.

We will not look deeper into the details of the Wright–Fisher and Moran models, the ideas from these models will help us introduce the concept and the importance of the EPS. It is interesting to see that even though the Moran model has been applied in numerous studies. it is less popular than the Wright–Fisher model. For a more comprehensive understanding of these models and detailed examples, we recommend Hartl's book [33], which provides an interesting discussion of the topics.

2.4 The effective population size

As we discussed earlier, genetic drift plays a fundamental role in the study of population dynamics. The concept of EPS, denoted as $\mathcal{N}_e$, helps us understand how drift works in real populations by comparing them to an idealized population. An idealized population is one that experiences the same rate of genetic drift as a real population, under hypothetical conditions.

Consider a population with the following characteristics: diploid organism, has sexual reproduction, non overlapping generations, many independent subpopulations of constant size N, random mating within each subpopulation, no migration between subpopulations, no mutation and no selection, as described in [33]. This idealized situation is rare in real world populations, as most populations do not meet all of these conditons. Therefore, there must be corrections for such complications such as fluctuations in population size, unequal number of females than males, differences in age, and so on, as we could see in [20]. Therefore, to study these complexities, we need to adjust our models to reflect the real conditions under which populations evolve. One important question that arises is whether individuals who do not contribute to the next generation should be included in the population size. This brings us to the concept of EPS. This concept was introduced by Sewall Wright in 1931 in the paper title *Evolution in Mendelian populations.*

The EPS refers to the *number of individuals in a theoretically ideal population having the same magnitude of random genetic drift as the actual population.*

EPS accounts for various factors influencing genetic drift, such as fluctuations in population size, disparities in reproductive success among individuals, and the population's structure. Usually, $\mathcal{N}_e$ is smaller than the actual population size $\mathcal{N}$ because not all individuals contribute equally to the gene pool. Estimating $\mathcal{N}_e$ can sometimes be complex and often involves different ways to estimate this number, depending on how we choose to measure the magnitude. As Hartl [33] discusses, the first method involves the change in the average inbreeding coefficient, the second method examines the change in variance in allele frequencies, and the last one looks at the rate of loss in heterozygosity. Additionally, there is a method based on coalescent effective size, but we first need to introduce the concept of coalescence.

For more information on EPS , the following references can be consulted: [18], [78], [56], [76].

2.5 The coalescent process

As an introduction to this process, we can say that coalescent theory helps us to explain and analyze the history of a population and also to study the models of evolution. It is interesting to think that each gene comes from another gene and this process continues back in time, from generation to generation. But what if we randomly select two genes from the population? How can we determine whether they are related? And if they are not in the current generation, could they share a common ancestor in the past? In some cases, it may take many generations to trace back to a common ancestor, but eventually, the two lineages are expected to converge at a point in time where they coalesce. This shared ancestor is known as the *Most Recent Common Ancestor* (MRCA).

It is important to see that, as we mentioned, we want to look at the time backward in time this means that we are interested in look at the past, and not the future. Usually we are used to work with forward time and to look and predict what would happen in the future. For example, in a standard forward-time model called the Moran model, in each small interval of time, we choose an individual at random from the population and we suppose it has a number of offspring $X \sim Poisson(a)$. These offspring then form part of the next generation. This can be modeled as a branching process. A coalescent process can be viewed as a branching process run backward in time, as in Figure 2.3.

Consider two individuals sampled at random from a present-day population of fixed size N. One of these individuals must have descended from some parent in the previous generation, labeled i. The probability that the second individual also descended from the same parent is $1/N$. Therefore, the probability that the two individuals have different parents in the previous generation is $1 - 1/N$.

The probability that their lineages coalesce not in the previous generation but in the one before that is given by the product of two events, the probability that they did not coalesce in the previous generation, and the probability that they do coalesce in the generation before that. This gives the expression: $(1 - 1/N)(1/N)$.

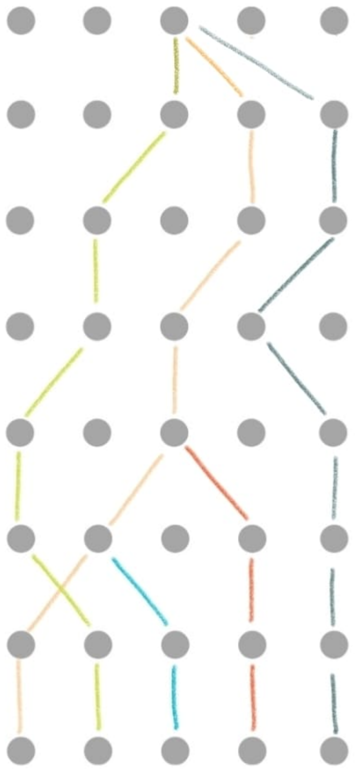

Figure 2.3: Coalescent tree showing the process of tracing the ancestry of alleles backward in time. We start with $k = 5$ alleles in generation 0 (the present), and as generations pass, the alleles coalesce into fewer lineages, ultimately leading to a single ancestral allele.

Extrapolating this logic to the probability of coalescing n generations ago, we get that this probability is:

$$p(n) = (1 - 1/N)^{n-1} \cdot (1/N).$$

In large populations, this expression can be approximated using the exponential function:

$$p(n) \approx \frac{1}{N} e^{-(n-1)/N}.$$

Thus, in a well-mixed population of constant size N, the probabilities of neutral coalescence can be computed explicitly and exhibit a simple exponential form.

We presented the easiest case, constant N and neutral coalescence. If we start to make some more realistic assumptions, the coalescence process can become complicated pretty quickly. There is a lot more interesting information to explain and review on the coalescent theory and biological applications. Here are relevant citations along with papers for further study: [44], [74], [28], [39].

Chapter 3

Spatial population genetics

In the first chapter, we developed an introduction to the mathematical concepts that we needed for developments in the second chapter and especially here in this third chapter. The goal of this chapter is to introduce and develop the main idea of this thesis and provide a general understanding of what we aim to achieve when we state that "effective dimensionality" facilitates stochastic modeling of structured populations. As we will see, the very simple coalescence process described above will become not so simple when we account for spatial distribution.

To begin, our objective is to explore how the *genetic distance* between two individuals can be an indicator of *physical distance*. The quantity that connects these two distances is *time*. The genetic distance evolves over time due to random mutations and genetic drift. Physical distance evolves over time through random walks and stochastic processes. We will show how random walks, modeled by BM, can describe the path in which the genetic evolution is embedded. Effectively the genetic process is embedded within the spatial process. Such embedding is made possible by the Skorokhod embedding theorem 1.6. The behavior of the random walk depends on the dimensional space as we shall see, and it can either be recurrent or transient. In dimensions greater than one, we will see that this is a Bessel process modeling the evolution of genetic distance between two individuals; this process represents the spatial divergence of genetic lineages over time.

In particular, we will define and explore the concept of "effective dimensionality". These observations suggest that there may be deep connections between our modeling approach and fractal structures in nature, which could have implications for understanding population structure and genetic divergence.

Finally, we will consider how stochastic models, based on the previous concepts, can be used to simulate and predict the behavior of structured populations. These models will allow us to better understand the dynamics of genetic divergence and the coalescence process over time, contributing to our overall understanding of population genetics and evolution.

3.1 Preliminaries

3.1.1 Spatial movement of genetic lineages

Most of classical population genetics (Chapter 2) makes an assumption that populations are well-mixed, meaning that any two randomly chosen individuals have the same probability of encountering each other. Here, we eliminate this "mass action" assumption and make the more realistic assumption that populations live and interact in physical space, such that relatedness and interaction probability both increase as the physical distance increases. The complexity introduced by considering spatial effects can make the study of population genetics more difficult.

The spatial movement of individuals in a population can look like a stochastic process such as BM. Over the course of many generations, the movement of a given genetic lineage in physical space is random. On a much shorter time scale, pollen grains in water move around randomly, as described by Scottish botanist Robert Brown (from which the term "Brownian motion" comes). The movement of genetic lineages in space and pollen grains in water have essentially the same statistical properties; the only difference is scale, both in time and in space. L. Bachelier and A. Einstein showed that the probability density governing the position of a pollen grain was Gaussian, with variance proportional to the time since it occupied its original position [3], [23].

One particularly important property of BM is its *time-reversibility.* This means that the process has the same statistical behavior when observed backward in time as it does forward. Formally, if $\mathcal{B}(t)$ is a BM, then the time-reversed process $\mathcal{B}(T-t)-B(T)$, for $0 \leq t \leq T$, is also a BM. This symmetry under time reversal plays a crucial role in many probabilistic constructions and is especially important for the coalescence analyses considered in this work.

3.1.2 Modeling evolution in continuous space and why it is so difficult

These considerations might give the impression that spatial population genetics are just a kind of trivial extension of physics models of things like diffusion and drift. But it is a mistake to think of spatial population genetics in this way, because there are several things that make spatial population genetics different from physics models of spatial phenomena. One example (maybe the most important example) is that organisms of interest are often obligately sexual so that an individual cannot reproduce without a mate. Because of this, potential mates have to be accessible and one way they can be accessible is by simple physical proximity. The result of this requirement is that organisms tend to live in groups, and biologists use the word "population" for such groups. So the technical meaning of "population" in biology is different from the popular use of this word. It means a group of organisms that usually mate among themselves and only rarely mate with "outsiders".

There are two important points here. The first is that evolution is a population process. It makes no sense to talk about individuals evolving on their own. Only populations of individuals evolve. So it is important to figure out how populations emerge and persist, which is a problem that is uniquely biological. The second point is about the size of populations. In really small populations there comes another uniquely biological problem which is inbreeding (endogamy) and it makes populations weak and susceptible to extinction. So populations have to be large. But not too large, because then ecological factors like food sources can be not enough for the large population. In summary, how populations form and persist in physical space is a problem that is uniquely biological. It is a spatial phenomenon that does not have an analogous counterpart in physics.

The basic problem can be summarized like this: populations are discrete entities that spontaneously emerge and live in continuous space. Most of the old models of populations living in space just assume that there are effectively a number of separate populations, each population lives in an isolated "compartment" and there is not very much migration between compartments. These are called "stepping stone" models and for a long time they were the standard model in spatial population genetics. More recently, scientists have understood that there are some important characteristics of real populations that can't be captured by stepping-stone models. Some of these characteristics are related to climate change and human migration, for example, which can cause habitat fragmentation. Sometimes this can cause habitats to have boundaries that are increasingly irregular and have a fractal (non-integer) dimension. We need to keep this point in mind because it will be useful in trying to interpret our "effective dimensionality " concept.

3.1.3 Spatial evolution in backward time

The basic framework for our modeling here is this: Over time, genetic divergence is embedded in a spatial random walk; it grows as each individual accumulates neutral mutations. However, if we look at the problem from the reverse-time perspective, we are asking a different question: When did these two individuals have a common ancestor? That is, what was their genetic coalescence time? This "spatial coalescence" process is a much more complex problem than classical (non-spatial) models of coalescence processes, because now we must consider that, for a coalescence event to happen, two lineages must be in the same neighborhood in physical space.

In reverse-time evolution, the physical distance between the two individuals forms a Bessel process, because we have to use a Euclidean distance. This process captures the random, independent evolution of their genetic compositions as they "move" closer to a common ancestor. The coalescence time—the moment their genetic histories merge—corresponds to the hitting time of the Bessel process, which tells us when the physical distance between them becomes small enough so that a coalescence event could have happened.

We start with BM (this could be described, as we have seen, as a random walk with continuous steps) and how it is important for describing random processes. Also, it is useful to study this kind of process, like random walks in multidimensional spaces (as in genetic drift in populations), because this can be modeled by BM.

Then, the Bessel process can be seen as a generalization of BM in more than one dimension, but with a specific focus on the radial distance—in this case, between two random walks. We used the Bessel process to describe the genetic divergence between two individuals over time, modeled as the distance between their genetic compositions as they evolve due to neutral mutation and random genetic drift.

3.1.4 The Bradburd "coexisting clusters" model

Some recent work by Gideon Bradburd [9] has developed a new way to model populations in continuous space. It is based on the observation that coexisting clusters of related individuals can have different patterns of movement in the same area of continuous space. Their model has a number of different coexisting "layers", and each layer represents a cluster of related individuals, and each layer can have different movement patterns.

As we shall see later, the methods we develop and the concept of "effective dimensionality" that we introduce can be combined with Bradburd's approach. We have been talking about the effective dimensionality $\mathcal{D}_e$ as a parameter with a constant numerical value. But there is no formal reason to think that $\mathcal{D}_e$ has to be a constant. If we think of $\mathcal{D}_e$ instead as a random variable, the connection to the Bradburd model becomes apparent. So, for example, we can define $\mathcal{D}_e$ to be a discrete random variable that can have, let us say, 3 possible values, $\mathcal{D}_e \in \{1.8, 2.0, 2.2\}$. We can think of each value of $\mathcal{D}_e$ as defining a "layer" like in Bradburd's model. Each layer contributes to observed patterns with probabilities $\{p_1, p_2, p_3\}$. See Figure 3.1. This definition of a layer is similar but not the same as Bradburd's definition. It is similar because, in our scheme, departure from integer dimensions can be seen as a measure of a departure from passive diffusion processes. This departure will affect spatial auto-correlations so that each layer will have a different auto-correlation structure, just like Bradburd's model. But it is different because we do not explicitly allow for some rate of admixture between the layers.

And we can go one step more. There is no reason to suppose that $\mathcal{D}_e$ has to be a discrete random variable like in the above example. It could also be a continuous random variable. This would implicitly assume that the number of layers is infinite. We are not sure if this approach would be more realistic or less realistic. But it can reduce the degrees-of-freedom of the problem. For example, if we assume $\mathcal{D}_e$ is a normally-distributed random variable, then there are only two parameters that you have to estimate: the mean and the variance. In contrast, if you use the discrete approach and assume there are 8 layers, then you have to estimate 7 parameters (the probability of each layer).

Applications using this kind of spatial modeling can address many different kinds of practical situations. Figure 3.2 illustrates an application to the inference of human migration patterns.

3.2 From Brownian motion to Bessel process

In the previous chapter, we made the assumption that populations are "well-mixed", which as we have stated earlier, means that any two individuals chosen at random have the same probability of interaction (which can include mating possibility). An example of this in a laboratory setting would be a chemostat (it is a type of bioreactor used in laboratories to maintain a continuous culture of microorganisms), where the environment is liquid and well-mixed, sometimes this is referred to as zero dimensions.

In one-dimensional space, two randomly-chosen individuals will not have a constant probability of interaction; instead, this probability will depend

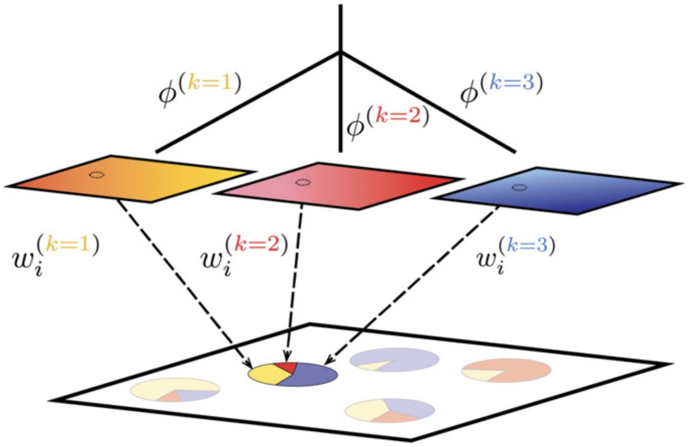

Figure 3.1: This figure is reproduced with permission from G Bradburd [10]–[12]. In this figure, there are three layers representing three clusters of related individuals coexisting in the same space. Our methods resemble this approach, but each layer will represent a different effective dimensionality $\mathcal{D}_e$ and there are an infinite number of layers because $\mathcal{D}_e$ is a continuous random variable.

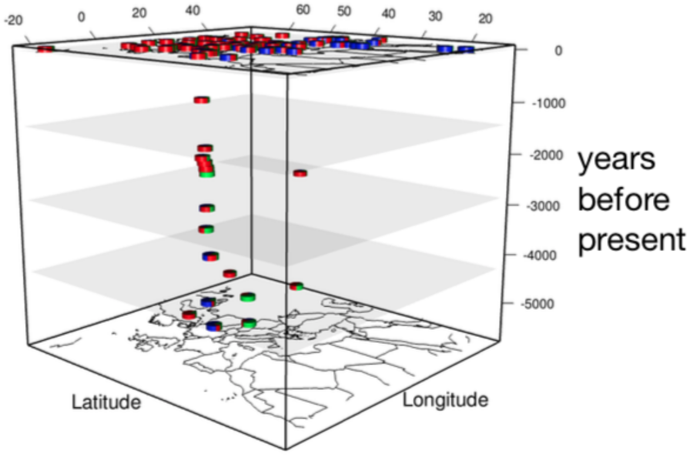

Figure 3.2: This figure is reproduced with permission from G Bradburd [10]–[12]. It shows a spatiotemporal analysis of genetic structure and ad-mixture in a set of human populations from Eurasia over the last 5,000 years.

on the physical distance between the two individuals. BM is a stochastic process that describes random movement along a number line. In one dimension, the random walk of a particle is simply described by BM, and the distance from the origin is just the absolute value of the position. There are some examples of populations that effectively live in one-dimensional space; a classic example of this is populations living along coastlines. We recall that one dimension is particularly convenient because we could use the reflection principle. However, most biological populations exist in more than one dimension. When we extend the concept to higher dimensions, the process of interest is not simply an independent collection of one-dimensional BMs. Instead, we measure the Euclidean distance between the two lineages; over time this distance forms its own stochastic process. This distance process is qualitatively different than BM because Euclidean distance is never negative. We found that this process is equivalent to a Bessel process.

A Bessel process is a type of stochastic process that appears in many different contexts. It can, for example, be used to model the distance of the path of a BM from the origin in multi-dimensional Euclidean space. More specifically, the Bessel process arises from the radial part of the BM in higher dimensions. In a multi-dimensional setting, while each individual spatial component of the BM remains independent, the radial distance follows a distinct process that has different properties than the one-dimensional case.

Thus, while the motion of the particle itself in each dimension still follows BM, the radial distance is described by a Bessel process. This distinction is what leads to the observation that, in higher dimensions, the process governing the particle's distance from the origin is referred to as the Bessel process rather than simple BM.

Let us consider two particles moving in a one-dimensional space, i.e., along a straight line, over time. This is a simple BM. Let the position of the first particle at time t be denoted as $X_1(t)$, and the position of the second particle at time t denoted as $X_2(t)$. The positions of the particles follow:

$$X_1(t) = X_1(0) + \mathcal{B}(t).$$

$$X_2(t) = X_2(0) + \mathcal{B}(t).$$

Where $\mathcal{B}(t)$ is a standard BM. The position of each particle changes by a random amount dy in the time interval $(t + dt)$. Technically speaking BM is usually defined in a stochastic calculus context, in which the usual limits of calculus apply: $dy \to 0$ and $dt \to 0$. In normal calculus, numerical solution of differential equations is straight-forward and easy to figure out: you just assume that dt is small but not zero, and then use it to deterministically compute dy for each time step. To make numerical solutions in stochastic calculus, it is not so obvious because dy in this case is like a random variable, not a deterministic value. To make numerical solutions we rely on the Skorokhod embedding theorem (1.6), which shows that, in the case of BM, dy will be a normally-distributed random variable with mean zero and variance $dt > 0$. In fact, Skorokhod proved that dt does not have to be small; it can be any value you choose. As illustrated in figure 3.3, if we track the position of the particle over time, we observe the values $X_1(t)$ and $X_2(t)$ at each time point, describing the random movement of the particle.

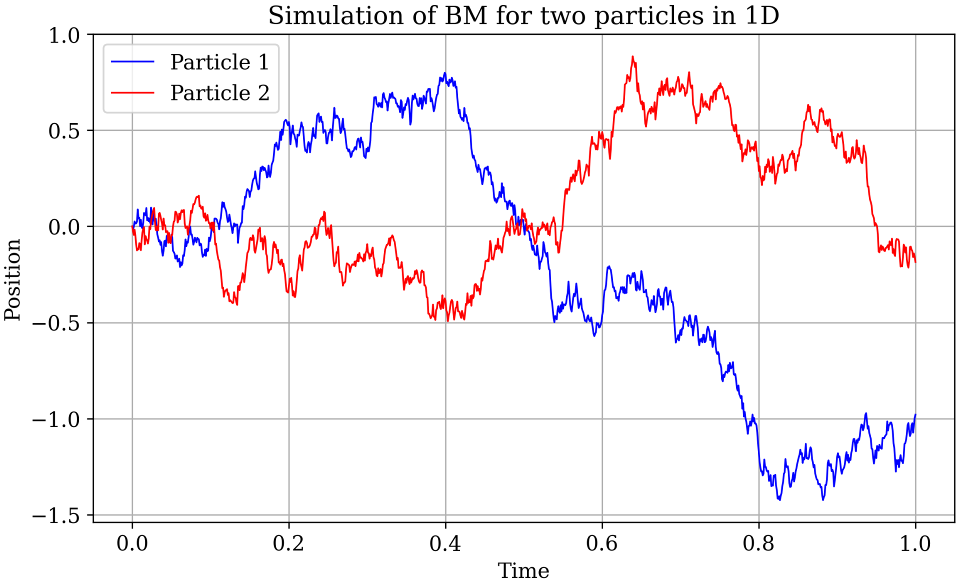

Figure 3.3: A simulation of BM of two particles in one-dimension. This figure shows the random movement of two particles along a straight line (the y-axis) over time. Both particles follow independent Brownian processes, resulting in random fluctuations in their positions. The positions $X_1(t)$ and $X_2(t)$ of the two particles are shown as they evolve over time, with particle 1 (blue) and particle 2 (red) moving randomly in one-dimensional space.

As we mentioned earlier, we are interested in the distance between the two lineages. In one dimension, this distance is simply $\Delta R(t) = |X_1(t) - X_2(t)|$, which is easy to work with because the process $X_1(t) - X_2(t)$ is itself a BM process, allowing us to take advantage of the reflection principle. The only thing that changes is the variance associated with $X_1(t) - X_2(t)$ is twice the variance associated with $X_1(t)$. So, the variance of dx is now $2dt$ and not dt as before.

Now, we imagine two particles moving in a two-dimensional space. The position of each particle can be described by a vector $(X_1(t), Y_1(t))$ for the first particle and $(X_2(t), Y_2(t))$ for the second particle, where each of these position components $X_1(t), Y_1(t), X_2(t)$, and $Y_2(t)$ are independent BMs. As illustrated in Figure 3.4, both particles follow independent BM in two dimensions, which means that the movement of each particle along the x- and y-axes is independent of the other.

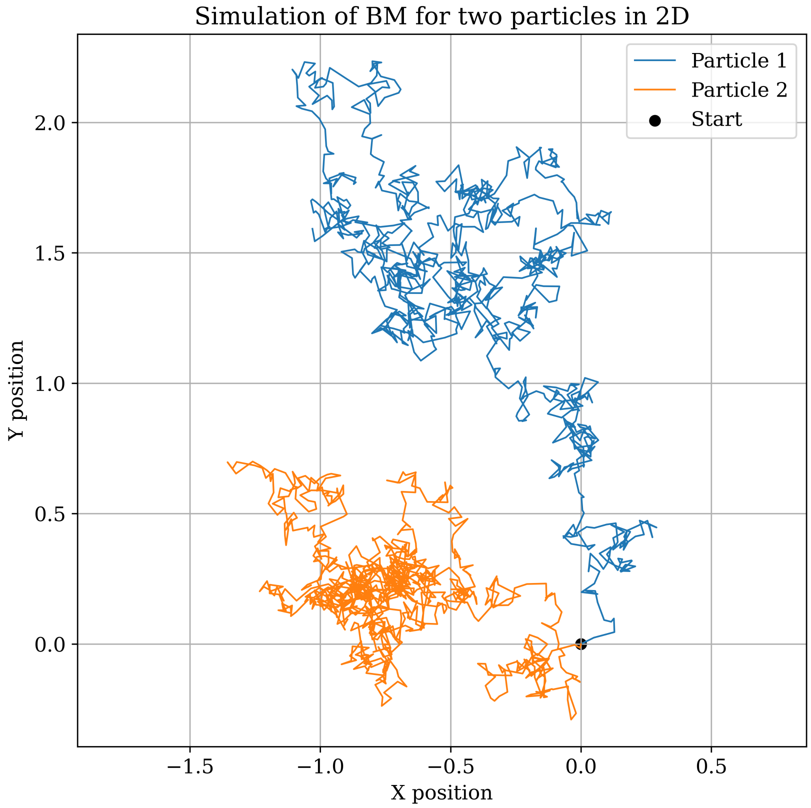

Figure 3.4: BM of two particles in two-dimension. This Figure shows the random movement of two particles along a plane. The positions $(X_1(t), Y_1(t))$ and $(X_2(t), Y_2(t))$ of the two particles are shown as they evolve over time, with particle 1 (blue) and particle 2 (orange) moving randomly in two-dimensional space.

In this case, the distance between the two particles at any given time t, denoted by $d(t)$, is given by:

$$\Delta R(t) = \sqrt{\Delta X(t)^2 + \Delta Y(t)^2}.$$

where $\Delta X(t) = X_1(t) - X_2(t)$ and $\Delta Y(t) = Y_1(t) - Y_2(t)$. We recall that $\Delta X(t)$ and $\Delta Y(t)$ are normal random variables with with variance $2dt$. This distance evolves over time as a result of the random movements of both particles. As illustrated in figure 3.5, the observation here is that the evolution of this distance between the two particles follows a Bessel process of dimension 2.

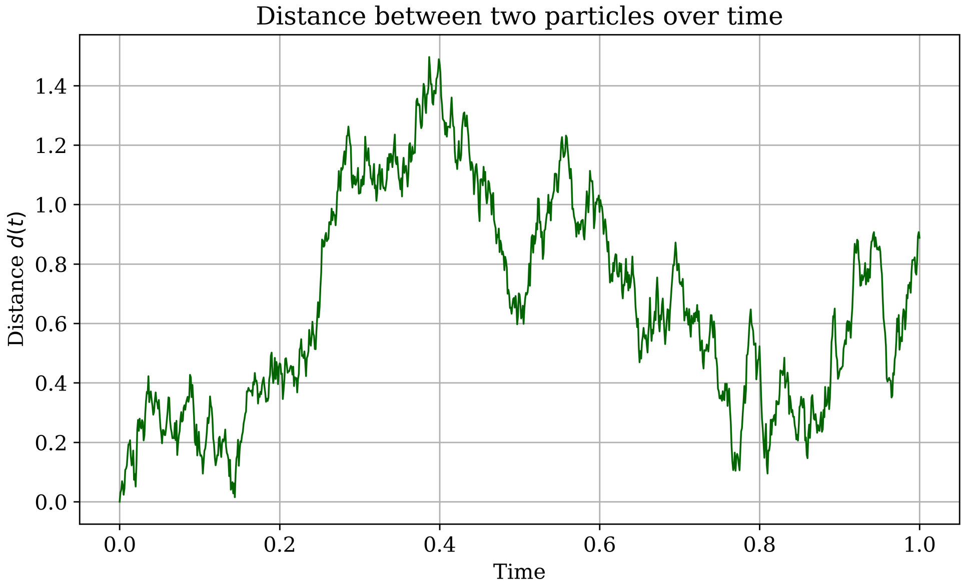

Figure 3.5: BM distances of two particles $(X_1(t), Y_1(t))$ and $(X_2(t), Y_2(t))$ in two-dimension.

In the Preliminaries section, we stated the following question from a reverse-time perspective: When did these two individuals share a common ancestor? However, before determining when, we must first evaluate whether there is a possibility for the two lineages to converge. In the following section, we will apply the theorem proved in Section 1.5.

3.3 Recurrence and transience in BM and Bessel processes

The property of being recurrent or transient in a BM or a Bessel process is determined by its dimension $\mathcal{D}$, first we are going to look at the theory with $\mathcal{D}$ an integer and then what happens like in our case $\mathcal{D}_e$ non-integer dimensionality.

We talk about recurrent and transient processes in Section 1. We saw that a random process is called *recurrent* if it eventually returns to a particular state (or neighborhood of that state) with probability 1. In other words, after leaving a state, the process is guaranteed to revisit it at some point in the future. For instance, BM is recurrent in one dimension. So if two BMs start in the same place, they will eventually meet again at some time in the future, in one dimension. In two dimensions, BM is "neighborhood recurrent", meaning that the two paths will eventually come to be less that a distance ϵ from each other, for arbitrarily small ϵ, with probability $p = 1$.

This recurrence property in one-dimension and two-dimension is tied to the geometry of the space, where the "smallness" of the space guarantees that the process will eventually revisit a state. This idea is often connected to the Poincaré recurrence theorem [59].

On the other hand, a random process is said to be *transient* if, once it leaves a state, there is a non-zero probability that it will never return to that state. Essentially, the process might drift away and never revisit certain points. For example, in three-dimensional or higher-dimensional processes, the process is transient, meaning that once the process moves away from its starting point, there is a non-zero probability that it will never return to it. This property is a result of the increased dimensionality of space, where the "escape" routes available to the process become more abundant, reducing the likelihood of returning to any specific point. More precisely, if two BMs start at the same position in three-dimensional space, the two paths may eventually become less than a distance ϵ from each, for arbitrarily small ϵ, with probability $p < 1$; in other words, there is no guarantee that they will ever meet again.

This difference arises because, in lower dimensions, the process is constrained enough that it is likely to return to previous states, whereas in higher dimensions the space becomes large enough that the process may never return to a particular point once it has moved away from it.

When we move to higher-dimensional spaces, the behavior changes significantly. In three-dimensional space or higher, the associated Bessel process becomes transient. The probability that the particles will encounter decreases, and they are more likely to "escape" indefinitely.

In biology, the study of this analysis is important. If a process is recurrent, it guarantees that two lineages will meet at some point in the past, much like how recurrent events will happen over time in a stochastic process. This concept is fundamental for studying evolutionary relationships. In fact, one of the assumptions made in phylogenetic analysis is that if two lineages meet (at some point in the past), will inevitably cross paths, ensuring a connection at some point. If this meeting occurs as a genomic "meeting", then this is called a coalescence. Then if a process is transience, lineage may drift away from a common point and never return. In evolutionary terms, this transience suggests that some lineages may diverge and never meet again.

In the context of phylogenetic trees, the idea is that when we observe relatedness between two species, it is based on the assumption that at some point in the past, their evolutionary paths intersected. If two lineages are found to be related, it is not just a coincidence – they must have shared a common ancestor at some point. For example, schematic (3.1) illustrates two ways in which two sequences can be the same.

$$A)\quad \left\{\begin{array}{l} \{1,0,1\}\rightarrow\{1,1,1\}\rightarrow\boxed{\{1,0,1\}} \\ \{1,0,1\}\rightarrow\{0,0,1\}\rightarrow\boxed{\{1,0,1\}} \end{array}\right. \tag{3.1}$$

$$B)\qquad \{0,0,0\}\rightarrow\left\{\begin{array}{r} \{1,0,0\}\rightarrow\boxed{\{1,0,1\}} \\ \{0,0,1\}\rightarrow\boxed{\{1,0,1\}} \end{array}\right.$$

where the observed pair of sequences are in a box; A represents a case of “identity by state” (IBS), and B represents a case of “identity by descent” (IBD). IBS tells us nothing about why the sequences are the same. In the illustration A shows a case where the two sequences could come different lineages and they could be the same because of chance or because natural selection is favoring that sequence (called “convergent evolution”). In case B, the two sequences are the same because they come from the same lineage. We note that IBS can reflect IBD or something else. IBD means that we have some reason to believe that the two sequences are the same because they are related by ancestry (by descent). Generally speaking, if the sequence changes are neutral (or *synonymous*) – i.e., if they do not change the encoded amino acids so they are “invisible” to natural selection – then the probability they are *not* IBD is very small. In all the data analyses we performed, we counted only synonymous differences in sequences.

We have seen that different numbers of physical dimensions can cause evolutionary differences that are quantitative and sometime even qualitative (like the transition from recurrent to transient between 2 and 3 dimensions, for example). Now we would like to have a look at these differences. We will do this by examining the relationship between diffusion processes and the Laplace operator in Euclidean space. When the number of physical dimensions is an integer, this is an easy exercise in restating some fundamental identities. But we will extrapolate to cases where the number of physical dimensions is not an integer; this extrapolation is non-trivial and informative (although probably not new). In a further extrapolation (which might be new), we will show how we can use the connection between Bessel processes and the radial Laplacian operator to derive the standard Laplacian operator in Euclidean space in non-integer dimensions. Our findings clearly partition the Laplacian into the standard integer-dimension Laplacian plus an extra term that quantifies the contribution of the non-integer part of the dimensionality.

3.4 Life in non-integer dimensions

As an introduction to non-integer dimensionality, we aim to explore both *the utility* of working with non-integer dimensions in contrast to the integer dimensions typically used in existing theory, as well as *possible interpretations* of non-integer dimensionality in this context. While our focus is primarily on the utility of non-integer dimensionality for modeling purposes, we remain curious about its interpretation. To help with both utility and interpretation, we decided to look at how the non-integer dimensionality affects Bessel processes (the radial Laplacian) in a more familiar setting, namely in Euclidean space.

As we have seen the Bessel process describes the motion of a particle (or lineage) in $\mathcal{D}_e$-dimensional space, focusing on its Euclidean distance between two particles (or lineages) that are both moving in space as Brownian processes. But here there is an obvious problem. When $\mathcal{D}_e = 2$, the distance measure is defined as $\sqrt{\Delta X^2 + \Delta Y^2}$. Likewise, when $\mathcal{D}_e = 3$, the distance measure is defined as $\sqrt{\Delta X^2 + \Delta Y^2 + \Delta Z^2}$. But what if $\mathcal{D}_e = 2.5$? In this case, how can we measure Euclidean distance?

To solve this problem, we make the following observation: Maybe a Bessel process can occur in a non-integer dimensionality, $\mathcal{D}_e$. So what does this mean? We live in integer dimensions. Most land animals, for example effectively live in two dimensions. Even most marine life effectively lives in two dimensions. But there are some exceptions, i.e., marine life that effectively lives in three dimensions. We have an intuitive idea about integer dimensions. But, how do we make sense of a dimensionality of $\mathcal{D}_e = 2.3$, for example? We do not live in 2.3 dimensions, and we cannot make measurements in 2.3 dimensions. What we can do is to consider the *projection* of dynamics in 2.3 dimensions onto dynamics in 2 dimensions. Then we are bringing our modeling back to real dynamics that make sense to us.

3.4.1 Laplace operator

The *Laplace operator*, also known as *Laplacian*, denoted as ∇^2, is a second-order differential operator that measures the rate at which a function's value deviates from its average value in the surrounding region due to diffusion. For a scalar function f, the Laplace operator is expressed as:

$$\nabla^2 f = \nabla \cdot (\nabla f).$$

In the context of diffusion, the quantity in the parentheses is known as Fick's law. The Laplace operator appears in many different contexts; maybe the most famous context is a model describing how heat spreads throughout a material, and it is written like this:

$$\frac{\partial}{\partial t} f \ = \ \nabla^2 f \ .$$

This notation is particularly convenient because it is the same notation for any number of physical dimensions. In general, the operator is simply the sum of the second partial derivatives of the function with respect to each spatial coordinate in Cartesian space. In *Cartesian coordinates* $(x_1, x_2, \ldots, x_n)$, the Laplacian in n-dimensional space is given by the sum of the second partial derivatives with respect to each spatial variable:

$$\nabla^2 f = \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2}.$$

We note that the sum is over an integer n, which is the number of dimensions. It is not clear, at this point, how such a sum might be extrapolated to a non-integer number of dimensions.

The Laplacian can also be expressed in other coordinate systems, such as spherical coordinates and cylindrical coordinates, which are useful in problems exhibiting spherical or cylindrical symmetry.

In *spherical coordinates* (r, θ, ϕ), the Laplacian is:

$$\nabla^2 f = \frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial f}{\partial r} \right) + \frac{1}{r^2 \sin\theta} \frac{\partial}{\partial \theta} \left(\sin\theta \frac{\partial f}{\partial \theta} \right) + \frac{1}{r^2 \sin^2\theta} \frac{\partial^2 f}{\partial \phi^2}.$$

In *cylindrical coordinates*, the Laplacian ∇^2 for a scalar function $\phi(r, \theta, z)$ is given by:

$$\nabla^2 f(r, \theta, z) = \frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial f}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 f}{\partial \theta^2} + \frac{\partial^2 f}{\partial z^2}.$$

These forms are crucial in solving problems with spherical or cylindrical symmetry, such as heat conduction in a spherical object or gravitational fields. However, in problems with circular or radial symmetry, *polar coordinates* often provide a more convenient framework. In polar coordinates, the position of a point in the plane is described by the radial distance r and the angle θ, making it crucial to transform the Laplacian to accommodate these new variables. This transformation involves using the chain rule and applying the geometry of polar coordinates to express the second derivatives with respect to r and θ. By following a series of steps, we can derive the Laplacian in polar coordinates.

Deriving the Laplacian in Polar Coordinates

As we have seen, the Laplacian operator in Cartesian coordinates (x, y) is given by:

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}.$$

We aim to transform this expression into polar coordinates (r, θ), where:

$$r = \sqrt{x^2 + y^2}, \quad \theta = \tan^{-1}\left(\frac{y}{x}\right).$$

First, we need to express the Cartesian coordinates x and y in terms of polar coordinates:

$$x = r\cos\theta, \quad y = r\sin\theta.$$

Now, we calculate the first-order derivatives of r and θ with respect to x and y:

$$\frac{\partial r}{\partial x} = \frac{x}{r}, \quad \frac{\partial r}{\partial y} = \frac{y}{r}.$$

$$\frac{\partial \theta}{\partial x} = -\frac{y}{r^2}, \quad \frac{\partial \theta}{\partial y} = \frac{x}{r^2}.$$

Using the chain rule, we can express the first derivatives with respect to x and y in terms of r and θ:

$$\frac{\partial}{\partial x} = \frac{\partial}{\partial r} \cdot \frac{x}{r} - \frac{\partial}{\partial \theta} \cdot \frac{y}{r^2}.$$

$$\frac{\partial}{\partial y} = \frac{\partial}{\partial r} \cdot \frac{y}{r} + \frac{\partial}{\partial \theta} \cdot \frac{x}{r^2}.$$

Next, we compute the second derivatives by applying the product rule and chain rule:

$$\frac{\partial^2}{\partial x^2} = \frac{\partial}{\partial x}\left(\frac{\partial}{\partial r} \cdot \frac{x}{r} - \frac{\partial}{\partial \theta} \cdot \frac{y}{r^2}\right).$$

$$\frac{\partial^2}{\partial y^2} = \frac{\partial}{\partial y}\left(\frac{\partial}{\partial r} \cdot \frac{y}{r} + \frac{\partial}{\partial \theta} \cdot \frac{x}{r^2}\right).$$

After performing the necessary derivatives, the Laplacian in polar coordinates becomes:

$$\nabla^2 = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2}{\partial \theta^2}.$$

This is the Laplacian operator in polar coordinates, which consists of two terms: a radial term and an angular term. In all of our further developments, we will assume that the quantity of interest, the probability density of a diffusing particle or lineage, has radial symmetry. This means that the quantity of interest is independent of the angle, yielding:

$$\nabla^2 = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}\right).$$

3.5 Mapping our "effective dimensionality" onto Cartesian coordinates

As shown in the previous section, the radial Laplace operator looks very much like the right-hand-side of the partial-differential equation describing Bessel processes. We now rewrite the spatial component of the Bessel process as an operator. For comparison, we rewrite the two operators together. First, we have the symmetric radial Laplace operator:

$$\nabla^2 \;=\; \frac{\partial^2}{\partial r^2} + \frac{\mathcal{D}-1}{r}\frac{\partial}{\partial r}.$$

where $\mathcal{D}$ is the number of physical dimensions and is assumed to be an integer. By comparison, the Bessel operator is as follows:

$$\nabla_e^2 \;=\; \frac{\partial^2}{\partial r^2} + \frac{\mathcal{D}_e-1}{r}\frac{\partial}{\partial r}.$$

where $\mathcal{D}_e$ is the effective number of physical dimensions and can be any real number; it is not limited to integers.

The only difference is that here, in the context of a Bessel process, the number of physical dimensions is not formally limited to integers. That is why we have changed the notation: $\mathcal{D}$ in the Laplace operator must be integer, whereas $\mathcal{D}_e$ in the Bessel operator is not limited to the integers and can be any real number.

But maybe the radial Laplace operator can exist in a non-integer dimension. It makes sense because of the exact correspondence between $\mathcal{D}$ in the Laplace operator and $\mathcal{D}_e$ in the Bessel equation. In what follows, we assume that this extrapolation is valid. We note that we are not the first people to see this similarity and wonder about it. We know of a previous study that makes the same ansatz, by a physicist, Sidney Redner [46], with whom we have discussed our work. And there exists some evidence that this extrapolation can be formalized and can be shown to be more than just an ansatz [15].

The radial Laplacian operator in $\mathcal{D}$-dimensional space, where $\mathcal{D}$ is an integer is:

$$\nabla^2 \; = \; \frac{1}{r^{\mathcal{D}-1}}\frac{\partial}{\partial r}\left(r^{\mathcal{D}-1}\frac{\partial}{\partial r}\right).$$

This operator captures how a function changes with respect to the radial distance from the two particles. To better understand the connection, let us expand the radial Laplacian operator. First, we apply the product rule to the term inside the derivative:

$$\frac{\partial}{\partial r}\left(r^{\mathcal{D}-1}\frac{\partial}{\partial r}\right) = (\mathcal{D}-1)r^{\mathcal{D}-2}\frac{\partial}{\partial r} + r^{\mathcal{D}-1}\frac{\partial^2}{\partial r^2}.$$

Now, dividing this expression by $r^{\mathcal{D}-1}$, we obtain:

$$\nabla^2 = \frac{1}{r^{\mathcal{D}-1}}\left[(\mathcal{D}-1)r^{\mathcal{D}-2}\frac{\partial}{\partial r} + r^{\mathcal{D}-1}\frac{\partial^2}{\partial r^2}\right].$$

Simplifying the two terms. The first term simplifies to $(\mathcal{D}-1)\frac{1}{r}\frac{\partial}{\partial r}$ and the second term simplifies to $\frac{\partial^2}{\partial r^2}$.

Thus, we arrive at the expanded form of the radial Laplacian:

$$\nabla^2 = \frac{\partial^2}{\partial r^2} + \frac{\mathcal{D}-1}{r}\frac{\partial}{\partial r}.$$

where $\mathcal{D} \in \mathbb{N}$. We now explore the similarity, mentioned above between this operator and our proposed Bessel operator:

$$\nabla_e^2 = \frac{\partial^2}{\partial r^2} + \frac{\mathcal{D}_e - 1}{r}\frac{\partial}{\partial r}.$$

where now we have $\mathcal{D}_e \in \mathbb{R}$.

Using the Bessel operator (and its analogous radial form of the Laplacian) as our starting point, we would now like to see how we can map the radial form to Euclidean space. In particular, we are interested to see how non-integer dimensions would map onto Cartesian coordinates. When dimensions have only integer values, then we get the familiar expressions:

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}.$$

for $\mathcal{D} = 2$, and:

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}.$$

for $\mathcal{D} = 3$.

3.5.1 Generalized Cartesian Laplacian for non-integer dimensions

When the number of dimensions is not an integer, we again change notation a little and let $\mathcal{D}_e$ denote the effective number of dimensions (or effective dimensionality). Now, it is no longer straightforward to define the Laplacian operator in standard Cartesian coordinates. To address this, we derive the projection of non-integer $\mathcal{D}_e$ onto integer $\mathcal{D}$-dimensional Euclidean space, which we will denote with the operator $\nabla^2(\mathcal{D}_e \mapsto \mathcal{D})$.

We start with the radial Laplacian, which has the above mentioned similarity to our Bessel process of interest (whose dimensionality $\mathcal{D}_e$ does not have to be an integer), and we work backwards to derive the Cartesian Laplacian.

For the two-dimensional case, the generalized operator takes the form:

$$\nabla^2(\mathcal{D}_e \mapsto 2) = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{D_e - 2}{x^2 + y^2}\left(x\frac{\partial}{\partial x} + y\frac{\partial}{\partial y}\right),$$

noting that when $D_e = 2$, we recover the standard two-dimensional Cartesian Laplacian. For the following proof, this approach is valid since we are working with the radial Laplacian, which supports generalization to non-integer dimensions.

Proof. We prove how to obtain the derivation of the standard two-dimensional Cartesian Laplacian. We begin with the radial form of the Laplacian operator in effective dimension $\mathcal{D}_e$:

$$\nabla_e^2 = \frac{\partial^2}{\partial r^2} + \frac{\mathcal{D}_e - 1}{r}\frac{\partial}{\partial r},$$

where $r = \sqrt{x^2 + y^2}$, and we assume the function $u(x, y)$ is radially symmetric, i.e., $u(x, y) = f(r)$.

By the chain rule:

$$\frac{\partial u}{\partial x} = \frac{df}{dr} \cdot \frac{\partial r}{\partial x} = f'(r)\frac{x}{r},$$

$$\frac{\partial^2 u}{\partial x^2} = \frac{d}{dx}\left(f'(r)\frac{x}{r}\right) = f''(r)\left(\frac{x}{r}\right)^2 + f'(r) \cdot \frac{r^2 - x^2}{r^3}.$$

Similarly, for the y-direction:

$$\frac{\partial u}{\partial y} = f'(r)\frac{y}{r}, \quad \frac{\partial^2 u}{\partial y^2} = f''(r)\left(\frac{y}{r}\right)^2 + f'(r) \cdot \frac{r^2 - y^2}{r^3}.$$

Adding both second derivatives:

$$\begin{aligned}\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} &= f''(r)\left(\frac{x^2 + y^2}{r^2}\right) + f'(r)\left(\frac{2r^2 - (x^2 + y^2)}{r^3}\right) \\ &= f''(r) + \frac{1}{r}f'(r),\end{aligned}$$

since $x^2 + y^2 = r^2$. Thus, in standard 2D we have:

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} = \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} \quad .$$

Now we consider the radial Laplacian in non-integer dimension $\mathcal{D}_e$:

$$\nabla_e^2 = \frac{\partial^2}{\partial r^2} + \frac{\mathcal{D}_e - 1}{r}\frac{\partial}{\partial r} \; = \; \nabla^2 + \frac{\mathcal{D}_e - 2}{r}\frac{\partial}{\partial r} \quad . \tag{3.2}$$

We note that $r = \sqrt{x^2 + y^2}$, from which we have:

$$\frac{\partial r}{\partial x} \; = \; \frac{x}{\sqrt{x^2 + y^2}} \; = \; \frac{x}{r} \; , \quad \text{and} \quad \frac{\partial r}{\partial y} \; = \; \frac{y}{\sqrt{x^2 + y^2}} \; = \; \frac{y}{r} \quad ,$$

from which we use the chain rule to derive:

$$\frac{\partial u}{\partial r} = \frac{x}{r}\frac{\partial u}{\partial x} + \frac{y}{r}\frac{\partial u}{\partial y} \; .$$

We get:

$$\frac{1}{r}\frac{\partial u}{\partial r} = \frac{x}{r^2}\frac{\partial u}{\partial x} + \frac{y}{r^2}\frac{\partial u}{\partial y} = \frac{1}{x^2 + y^2}\left(x\frac{\partial u}{\partial x} + y\frac{\partial u}{\partial y}\right).$$

Inserting these expressions into Eq (3.2), we get:

$$\frac{\mathcal{D}_e - 2}{r}\frac{\partial u}{\partial r} = \frac{\mathcal{D}_e - 2}{x^2 + y^2}\left(x\frac{\partial u}{\partial x} + y\frac{\partial u}{\partial y}\right).$$

Combining terms, we obtain the expression for the projection of the radial Laplacian in $\mathcal{D}_e$ dimensions onto the two-dimensional plane:

$$\nabla^2(\mathcal{D}_e \mapsto 2) = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\mathcal{D}_e - 2}{x^2 + y^2}\left(x\frac{\partial}{\partial x} + y\frac{\partial}{\partial y}\right).$$

□

For the three-dimensional case, we find:

$$\nabla^2(\mathcal{D}_e \mapsto 3) = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} + \frac{\mathcal{D}_e - 3}{x^2 + y^2 + z^2}\left(x\frac{\partial}{\partial x} + y\frac{\partial}{\partial y} + z\frac{\partial}{\partial z}\right).$$

noting that if $\mathcal{D}_e = 3$, we recover the standard three-dimensional Cartesian Laplacian.

We can generalize this projection onto any integer dimensionality, $\mathcal{D}$. By induction, we have the general expression:

$$\nabla^2(\mathcal{D}_e \mapsto \mathcal{D}) = \sum_{i=1}^{\mathcal{D}} \frac{\partial^2}{\partial x_i^2} + \frac{\mathcal{D}_e - \mathcal{D}}{\sum_{i=1}^{\mathcal{D}} x_i^2} \sum_{i=1}^{\mathcal{D}} x_i \frac{\partial}{\partial x_i} \quad .$$

We note that this expression is clearly partitioned into: 1) the integer part (first term) and 2) the non-integer contribution (second term).

3.5.2 Some examples of numerical solutions

We begin with two points, one at location A and another at location B in the landscape, and we consider the probability of coalescence as we move backward in time. What we are effectively doing is tracking the positions of lineage one and lineage two over time. This corresponds to the product of two diffusion processes, one starting at point A and the other at point B. What interests us is the product of the probabilities associated with these two diffusion processes as time goes backward.

In the following Figure 3.6, we see the initial probability distributions, which are represented as two delta functions. This is because, based on the data, we know the exact locations of the two DNA samples. Therefore, the relevant initial probability distribution is a delta function, since each lineage starts from a known position.

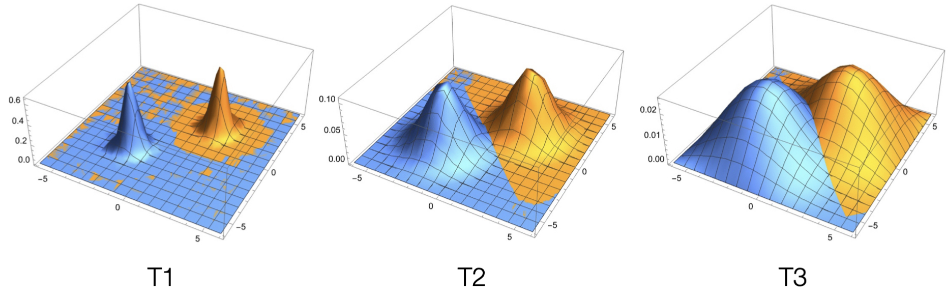

Figure 3.6: Dynamics of the probability densities two diffusion processes. This how we model spatial movement in backwards time.

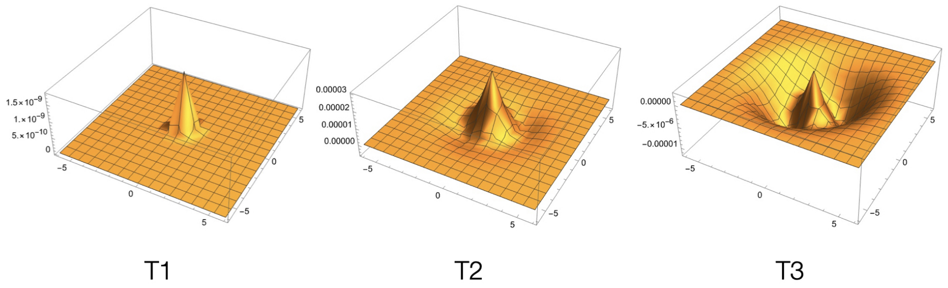

Figure 3.7: Product of the two densities as illustrated in Fig 3.6 for $\mathcal{D}_e = 2.3$ *minus* the product of two such densities for $\mathcal{D}_e = 2.0$. In essence this shows the effect of the non-integer part of $\mathcal{D}_e = 2.3$ on spatial coalescence probabilities.

At the beginning, we have two sharp peaks. As we go backward in time, the probability that a lineage is at a given position (for example, at position -1 or -5) spreads out like a diffusion process. So, each lineage follows an independent diffusion process starting from its respective location.

What we are interested in is the product of these two probabilities. These two distributions, f and g, diffuse over space. The key quantity is the probability that both lineages are at the same location at the same time. This probability is given by the product of f and g, evaluated at the same point. In essence, we are computing the probability that both lineages overlap at some location as they diffuse backward in time.

It appears that for every non-integer dimension, there exists a Cartesian mapping to all integer dimensions. While we have not been able to prove it mathematically, however we suspect that $1 > \mathcal{D}_e - n \geq 0$ must hold. Further exploration of the above expression and what we think should be the conditions, these and other things will be subjects of future study. This result is likely found in the literature, but we have not been able to locate a reference for it.

Chapter 4

Derivation of effective dimensionality from genetic distances

This chapter is divided into two sections. In the first section, we introduce a method for determining the effective number of spatial dimensions based on pairwise Hamming distances between sequences, typically derived from DNA or RNA.

In the second section, we describe how to derive probabilities from the discrete distribution of neutral differences between two sequences using a *pgf*. These probabilities will be denoted by $\mathcal{G}(\theta)$.

4.1 Spatial dimensionality and genetic distances

In this section, we present a method for determining the effective number of spatial dimensions from *pairwise Hamming distances between sequences*. We begin by establishing that if we know the coalescence time t_c for a pair of sequences and the neutral mutation rate μ, then the number of neutral mutations separating the two sequences, referred to as the *neutral Hamming distance*, follows a Poisson distribution with expectation $2\mu t_c$.

This result provides a probabilistic relationship between observable genetic differences and the unobserved coalescence time. Because coalescence time itself is influenced by the spatial structure of the population, this connection allows us to use genetic data to make inferences about spatial dimensionality.

Pairwise genetic differences between sequences can provide insight into the spatial structure of a population. In particular, we are interested in understanding how the effective number of spatial dimensions can be inferred from *pairwise Hamming distances*, a measure of genetic divergence that counts the number of nucleotide differences between sequences of equal length.

The Hamming distance is calculated by comparing two sequences of equal length and counting the number of positions where the corresponding nucleotides differ. For example, for each position in the sequences, we check if the nucleotide in the first sequence differs from the nucleotide in the second sequence. If they differ, it contributes one to the total Hamming distance. This metric provides a simple yet effective way to quantify genetic differences. For instance, when comparing the sequences `ATCG` and `ATGG`, the Hamming distance is 1 because they differ at the third position. This method allows us to quantify genetic differences by counting the positions at which the nucleotides differ. We will also discuss the neutral Hamming distance, which shares the same definition as the standard Hamming distance but accounts for only those sequence differences that do not change the encoded amino acid and is therefore invisible to selection. We will explain this concept in further detail.

As referenced in the previous section (2.1), which lists all possible nucleotide combinations, we can use this information to calculate Hamming distances more easily by comparing all pairs of sequences.

4.1.1 Coalescence time and mutation

In Chapter 2 on classical population genetics, we introduced the concept of *coalescence time*. As we have seen, the coalescence time t_c between two sequences is defined as the time in the past when they last shared a common ancestor. After this common ancestor is established, the two lineages accumulate genetic mutations independently over time. Assuming a constant *neutral mutation rate* μ.

If we know the coalescence time t_c for a pair of sequences and the neutral mutation rate μ, then the number of neutral mutations separating the two sequences, referred to as the *neutral Hamming distance*, follows a Poisson distribution with expected value $2\mu t_c$.

For example, consider two sequences with a coalescence time of $t_c = 1000$ generations and a neutral mutation rate of $\mu = 1 \times 10^{-8}$ mutations per site per generation. The expected number of neutral mutations separating the two sequences would be:

$$2\mu t_c = 2 \times 1 \times 10^{-8} \times 1000 = 2 \times 10^{-5}.$$

This means that, on average, we would expect 2×10^{-5} mutations separating the two sequences. While this number is quite small, it represents the expected number of neutral mutations that have occurred between the two sequences over the coalescence time. The actual number of mutations observed may vary, following a Poisson distribution with this expected value.

We can prove that the neutral Hamming distance follows a Poisson distribution by considering the mutation process as a Poisson process. First, we assume that mutations at each nucleotide position occur independently and at a constant rate μ per generation.

Let k represent the number of mutations that occur at a given nucleotide position over time t_c. Since the mutation rate μ is small, the probability of more than one mutation at the same position is nearly zero. Therefore, the number of mutations at each nucleotide position follows a Poisson distribution:

$$\mathbb{P}(k) = \frac{(\mu t_c)^k e^{-\mu t_c}}{k!}.$$

Now, for a pair of sequences, the total number of neutral mutations is the sum of independent Poisson random variables across all nucleotide positions. Since the sum of independent Poisson-distributed variables is also Poisson-distributed, the total number of neutral mutations (i.e., the neutral Hamming distance) follows a Poisson distribution with an expected value of:

$$\mathbb{E}[\text{Neutral Hamming Distance}] = 2\mu t_c.$$

The factor of 2 accounts for the two sequences being compared, and the result follows from the properties of the Poisson process.

When we calculate the neutral Hamming distance of two sequences, we obtain a value M. Let M denote number of neutral mutations separating a pair of sequences. The number of neutral mutations M is assumed to follow a Poisson distribution with expectation $2\mu t_c$, based on the assumption that the mutations occur independently at a constant rate over time. Then,

$$M \;\sim\; \textbf{Poisson}(2\mu t_c). \tag{4.1}$$

such that:

$$\mathbb{P}\{M = k\} \;=\; \frac{\lambda^k}{k!} e^{-\lambda}, \quad \text{where } \lambda \;=\; 2\mu t_c\ . \tag{4.2}$$

As seen in papers like [75] and [24], the number of mutations per unit length follows a Poisson distribution. This could be seen because the mutations are rare and independent events.

4.1.2 Extracting probabilities of the discrete distribution of numbers of neutral differences

Given what we previously discussed about how we obtain the value of M, it is now time to look at the way to determine the effective number of spatial dimensions.

We want to calculate the *pgf* of M, because we want to store all the probabilities associated with the values that mutations can take. The *pgf* associated with random variable M is therefore:

$$g(\theta) \;=\; e^{-2\mu t_c(1-\theta)}. \tag{4.3}$$

Proof. To see this, we could recall that the *pgf* is defined as:

$$g(\theta) = E[\theta^M] = \sum_{k=0}^{\infty} P(M = k)\theta^k. \tag{4.4}$$

For a Poisson distribution with parameter $\lambda = 2\mu t_c$, the *pmf* is:

$$\mathbb{P}(M = k) = \frac{\lambda^k}{k!}e^{-\lambda}, where \quad k = 0, 1, 2, \ldots \tag{4.5}$$

Substituting the *pmf* into the *pgf* we have that:

$$g(\theta) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}e^{-\lambda}\theta^k = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda\theta)^k}{k!}. \tag{4.6}$$

We see that this is the Taylor series expansion of $e^{\lambda\theta}$, so we could simplify this:

$$g(\theta) = e^{-\lambda}e^{\lambda\theta} = e^{-\lambda(1-\theta)}. \tag{4.7}$$

Substituting $\lambda = 2\mu t_c$, we obtain:

$$g(\theta) = e^{-2\mu t_c(1-\theta)}. \tag{4.8}$$

□

We now encounter a problem, as the value of t_c is not known beforehand. Rather than assuming t_c to be a fixed value, we assume it follows a known probability density function, denoted by $h(t_c)$. The *pgf* associated with the random variable M then becomes:

$$\mathcal{G}(\theta) = \int_0^{\infty} e^{-2\mu t_c(1-\theta)} h(t_c)\, dt_c.$$

This expression can further be rewritten in terms of the random variable T, which also represents the coalescence time, where T is the random version of t_c.

We can rewrite this expression in terms of the random variable T, which also represents also the coalescence time, were T is the random version of t_c:

$$\mathcal{G}(\theta) \;=\; E[e^{-2\mu T(1-\theta)}].$$

For $a, b \geq 0$, we denote by T the first hitting time to of the Bessel process with index ν, starting at distance a and stopping at distance $b < a$.

> At this first mention of distance parameters a and b, there are three important points to make: 1) that what really counts is the ratio of of these parameters a/b (this will become apparent later), 2) so without loss of generality in what follows we can let $b = 1$, and 3) distance a will be some function of EPS: $a = a(\mathcal{N}_e)$. Where needed for emphasis, we will write $a(\mathcal{N}_e)$; otherwise, we will just write a.

We have some approximations for the probability density of T. For the following results, we need to define the *modified Bessel function of the second kind*, denoted by K_ν. For more properties, refer to [31] and [36]. Now, we would like to evaluate the Laplace transform. We can evaluate the Laplace transform of the distribution of T by solving an eigenvalue problem. If we defined E the expectation, we have that the function:

$$x \mapsto E[e^{-\phi T}].$$

is increasing (decreasing) on $[0, b)$ (resp. (b, ∞)) and satisfies

$$G(\nu)u = \lambda u, \quad u(b) = 1.$$

The expressions for $E[e^{-\phi T}]$ is study and see in [31] [43]. If $0 < b \leq a$ and $\nu \in R$, we have the exact Laplace transform of the probability density of T:

$$f(\phi) = E\left[e^{-\phi T}\right] = \frac{a^{-\nu} K_\nu(a\sqrt{2\phi})}{b^{-\nu} K_\nu(b\sqrt{2\phi})}. \tag{4.9}$$

Where K_ν denote the modified Bessel functions of the second kinds of order ν. The modified Bessel functions K_ν are the solutions of the differential equation:

$$z^2 \frac{d^2 w}{dz^2} + z\frac{dw}{dz} - (z^2 + \nu^2)w = 0.$$

If we substitute the dummy variable ϕ with $2\mu(1-\theta)$, we get:

$$\mathcal{G}(\theta) = E\left[e^{-2\mu T(1-\theta)}\right] = \frac{a^{-\nu} K_\nu\left(2a\sqrt{\mu(1-\theta)}\right)}{b^{-\nu} K_\nu\left(2b\sqrt{\mu(1-\theta)}\right)}. \tag{4.10}$$

which is the exact *gf* associated with number of mutations, M.

We will now look at some observations about the number of neutral mutations M. Let us remember that M is a random variable that counts the number of neutral mutations between the two sequences.

4.1.3 Observations about the number of neutral mutations M

1. M is a discrete random variable, and can only take non-negative integer values in the set $\{0, 1, 2, 3, ...\}$.

2. The total probability mass of M is:

$$\sum_{k=0}^{\infty} P\{M = k\} = \lim_{\theta \to 1^-} \mathcal{G}(\theta) = \begin{cases} \left(\frac{b}{a}\right)^{2\nu}, & \nu > 0. \\ 1, & \nu \leq 0. \end{cases} \tag{4.11}$$

This indicates that M has a proper *pgf* only when $\nu \leq 0$; for $\nu > 0$, M has a *gf* but not a proper *pgf*.

Proof. We aim to compute the limit

$$\lim_{\theta \to 1^-} \mathcal{G}(\theta), \quad \text{where} \quad \mathcal{G}(\theta) = \frac{a^{-\nu} K_\nu(2a\sqrt{\mu(1-\theta)})}{b^{-\nu} K_\nu(2b\sqrt{\mu(1-\theta)})},$$

and show that

$$\lim_{\theta \to 1^-} \mathcal{G}(\theta) = \begin{cases} \left(\dfrac{b}{a}\right)^{2\nu}, & \text{if } \nu > 0, \\ 1, & \text{if } \nu \leq 0. \end{cases}$$

Let us define $x := 2a\sqrt{\mu(1-\theta)}$ and $y := 2b\sqrt{\mu(1-\theta)}$. As $\theta \to 1^-$, both $x \to 0^+$ and $y \to 0^+$. The behavior of the modified Bessel function $K_\nu(z)$ for small positive arguments depends on the value of ν, and we consider three cases.

Case 1: $\nu > 0$.

We aim to prove that:

$$\lim_{\theta \to 1^-} \mathcal{G}(\theta) = \left(\frac{b}{a}\right)^{2\nu}, \tag{4.12}$$

or in other words,

$$\lim_{\theta \to 1^-} \frac{a^{-\nu} K_\nu \left(2a\sqrt{\mu(1-\theta)}\right)}{b^{-\nu} K_\nu \left(2b\sqrt{\mu(1-\theta)}\right)} = \left(\frac{b}{a}\right)^{2\nu}.$$

Which is equivalent to:

$$\lim_{\theta \to 1^-} \left(\frac{b}{a}\right)^{\nu} \frac{K_\nu \left(2a\sqrt{\mu(1-\theta)}\right)}{K_\nu \left(2b\sqrt{\mu(1-\theta)}\right)} = \left(\frac{b}{a}\right)^{2\nu},$$

Therefore, we aim to prove this equation:

$$\lim_{\theta \to 1^-} \frac{K_\nu \left(2a\sqrt{\mu(1-\theta)}\right)}{K_\nu \left(2b\sqrt{\mu(1-\theta)}\right)} = \left(\frac{b}{a}\right)^{\nu}.$$

As recalled from [43] and [1, Eq. 9.6.9], several properties of the modified Bessel functions $I_\nu(x)$ and $K_\nu(x)$ are useful. One particularly interesting observation concerns their behavior as $z \to 0$ for $\nu > 0$. For small $z > 0$, the modified Bessel function of the second kind satisfies the following asymptotic expansion:

$$K_\nu(z) \sim \frac{\Gamma(\nu)}{2} \left(\frac{2}{z}\right)^{\nu}.$$

Applying this to both the numerator and the denominator, we obtain

$$K_\nu(x) \sim \frac{\Gamma(\nu)}{2} \left(\frac{2}{x}\right)^{\nu} = \frac{\Gamma(\nu)}{2} \left(\frac{1}{a\sqrt{\mu(1-\theta)}}\right)^{\nu},$$

$$K_\nu(y) \sim \frac{\Gamma(\nu)}{2} \left(\frac{2}{y}\right)^{\nu} = \frac{\Gamma(\nu)}{2} \left(\frac{1}{b\sqrt{\mu(1-\theta)}}\right)^{\nu}.$$

With this in mind, we can simplify our expression:

$$\lim_{\theta\to 1^-} \frac{(K_\nu(2a\sqrt{\mu(1-\theta)}))}{(K_\nu(2b\sqrt{\mu(1-\theta)}))} \sim \lim_{\theta\to 1^-} \frac{\frac{\Gamma(\nu)}{2}\left(\frac{(2a\sqrt{\mu(1-\theta)})}{2}\right)^{-\nu}}{\frac{\Gamma(\nu)}{2}\left(\frac{(2b\sqrt{\mu(1-\theta)})}{2}\right)^{-\nu}}.$$

$$\sim \lim_{\theta\to 1^-} \left(\frac{(2a\sqrt{\mu(1-\theta)})}{(2b\sqrt{\mu(1-\theta)})}\right)^{-\nu}.$$

$$\sim \lim_{\theta\to 1^-} \left(\frac{a}{b}\right)^{-\nu} = \left(\frac{b}{a}\right)^{\nu}.$$

and therefore

$$\lim_{\theta\to 1^-} \mathcal{G}(\theta) = \left(\frac{b}{a}\right)^{2\nu}.$$

Case 2: $\nu = 0$.

In this case, we use the asymptotic expansion of the modified Bessel function of the second kind of order zero. We recall from [1, Eq. 9.6.13], that the modified Bessel function of the second kind of order zero has the expansion:

$$K_0(z) \sim -\log\left(\frac{z}{2}\right) I_0(z) + R(z), \quad \text{as } z \to 0^+,$$

where $I_0(z) = 1+\mathcal{O}(z^2)$ as $z \to 0^+$, and the remainder $R(z)$ approaches zero as $z \to 0^+$. Therefore, the logarithmic term dominates, and we obtain the asymptotic behavior:

$$K_0(z) \sim -\log\left(\frac{z}{2}\right), \quad \text{as } z \to 0^+.$$

Applying this to $x = 2a\sqrt{\mu(1-\theta)}$ and $y = 2b\sqrt{\mu(1-\theta)}$, we obtain:

$$K_0(x) \sim -\log\left(a\sqrt{\mu(1-\theta)}\right), \qquad K_0(y) \sim -\log\left(b\sqrt{\mu(1-\theta)}\right).$$

Therefore,

$$\mathcal{G}(\theta) = \frac{K_0(x)}{K_0(y)} \sim \frac{-\log\left(a\sqrt{\mu(1-\theta)}\right)}{-\log\left(b\sqrt{\mu(1-\theta)}\right)}.$$

Now observe that each logarithm can be expanded as:

$$\log\left(a\sqrt{\mu(1-\theta)}\right) = \log(a) + \frac{1}{2}\log(\mu) + \frac{1}{2}\log(1-\theta),$$

and similarly for b. As $\theta \to 1^-$, the term $\log(1-\theta) \to -\infty$ dominates both the numerator and denominator, while the other terms remain constant. Thus,

$$\mathcal{G}(\theta) \to 1, \quad \text{as } \theta \to 1^-.$$

Therefore, the ratio tends to 1:

$$\lim_{\theta \to 1^-} \mathcal{G}(\theta) = 1.$$

This completes the case $\nu = 0$.

Case 3: $\nu < 0$**.**

The modified Bessel function of the second kind, $K_\nu(z)$, satisfies the symmetry

$$K_{-\nu}(z) = K_\nu(z),$$

which allows us to reduce the case $\nu < 0$ to the previously studied case $\nu' = -\nu > 0$. This symmetry, studied in [43] and [1], permits us to reuse the known asymptotic behavior for $\nu > 0$.

Let us recall the expression for the generating function:

$$\mathcal{G}(\theta) = \frac{a^{-\nu} K_\nu\left(2a\sqrt{\mu(1-\theta)}\right)}{b^{-\nu} K_\nu\left(2b\sqrt{\mu(1-\theta)}\right)}.$$

For small arguments $z \to 0^+$, the asymptotic behavior of the Bessel function for $\nu > 0$ is given by [43, pag. 762] [1, Eq. 9.6.9], where:

$$K_\nu(z) \sim \frac{\Gamma(-\nu)}{2}\left(\frac{z}{2}\right)^\nu .$$

Apply this to both numerator and denominator:

$$K_\nu\left(2a\sqrt{\mu(1-\theta)}\right) \sim \frac{\Gamma(-\nu)}{2}\left(a\sqrt{\mu(1-\theta)}\right)^\nu,$$

$$K_\nu\left(2b\sqrt{\mu(1-\theta)}\right) \sim \frac{\Gamma(-\nu)}{2}\left(b\sqrt{\mu(1-\theta)}\right)^\nu.$$

Substitute into $\mathcal{G}(\theta)$:

$$\mathcal{G}(\theta) \sim \frac{a^{-\nu} \cdot \left(a\sqrt{\mu(1-\theta)}\right)^{\nu}}{b^{-\nu} \cdot \left(b\sqrt{\mu(1-\theta)}\right)^{\nu}}.$$

We simplify the asymptotic expression:

$$\mathcal{G}(\theta) \sim \frac{a^{-\nu} \cdot \left(a\sqrt{\mu(1-\theta)}\right)^{\nu}}{b^{-\nu} \cdot \left(b\sqrt{\mu(1-\theta)}\right)^{\nu}} = \frac{a^{-\nu}a^{\nu}\left(\mu(1-\theta)\right)^{\nu/2}}{b^{-\nu}b^{\nu}\left(\mu(1-\theta)\right)^{\nu/2}}.$$

Canceling the terms we get:

$$\mathcal{G}(\theta) \sim \frac{(\mu(1-\theta))^{\nu/2}}{(\mu(1-\theta))^{\nu/2}} = 1.$$

Thus,

$$\lim_{\theta \to 1^-} \mathcal{G}(\theta) = 1.$$

Combining all this three cases, we obtain that:

$$\lim_{\theta \to 1^-} \mathcal{G}(\theta) = \begin{cases} \left(\frac{b}{a}\right)^{2\nu}, & \text{if } \nu > 0, \\ 1, & \text{if } \nu \leq 0. \end{cases}$$

□

3. The distribution of M has no defined moments. This can be seen, for example, by trying to compute its expectation

$$E[M] \;=\; \lim_{\theta \to 1^-} \mathcal{G}'(\theta) \;=\; \infty.$$

Proof. To show that $E[M] = \lim_{\theta \to 1^-} \mathcal{G}'(\theta) = \infty$, we can see that if we differentiate the *(pgf)* with respect to θ that is given by:

$$\mathcal{G}(\theta) = \frac{a^{-\nu}K_\nu(2a\sqrt{\mu(1-\theta)})}{b^{-\nu}K_\nu(2b\sqrt{\mu(1-\theta)})}.$$

we get:

$$\mathcal{G}'(\theta) = \frac{d}{d\theta}\left(\frac{a^{-\nu}K_\nu(2a\sqrt{\mu(1-\theta)})}{b^{-\nu}K_\nu(2b\sqrt{\mu(1-\theta)})}\right).$$

Using the chain rule, we have that

$$\mathcal{G}'(\theta) = R \cdot \frac{N_1 - N_2}{N_3}.$$

where:

$$R = \left(\frac{b}{a}\right)^\nu.$$

$$N_1 = K_\nu(2b\sqrt{\mu(1-\theta)}) \cdot \frac{d}{d\theta}\left(K_\nu(2a\sqrt{\mu(1-\theta)})\right).$$

$$N_2 = -K_\nu(2a\sqrt{\mu(1-\theta)}) \cdot \frac{d}{d\theta}\left(K_\nu(2b\sqrt{\mu(1-\theta)})\right).$$

$$N_3 = K_\nu(2b\sqrt{\mu(1-\theta)})^2.$$

We need to examine what happens as $\theta \to 1^-$, i.e., when $1-\theta \to 0$. Then, we analyze the behavior of $K_\nu(x)$ as $x \to 0$.

As we seen before that for $\nu > 0$ it is known that:

$$K_\nu(x) \sim \frac{\Gamma(\nu)}{2}\left(\frac{2}{x}\right)^\nu \quad \text{as } x \to 0.$$

Thus for small $1-\theta$, we substitute the value of $K_\nu(x)$:

$$K_\nu(2a\sqrt{\mu(1-\theta)}) \sim \frac{\Gamma(\nu)}{2}\left(\frac{2}{2a\sqrt{\mu(1-\theta)}}\right)^\nu.$$

and

$$K_\nu(2b\sqrt{\mu(1-\theta)}) \sim \frac{\Gamma(\nu)}{2}\left(\frac{2}{2b\sqrt{\mu(1-\theta)}}\right)^\nu.$$

We can substitute and simplify As $\theta \to 1^-$:

$$K_\nu(2a\mu(1-\theta)) \quad \text{and} \quad K_\nu(2b\mu(1-\theta)).$$

both approach infinity.

Then taking the derivative $\mathcal{G}'(\theta)$, we have $\mathcal{G}'(\theta) \sim \frac{a^{-nu}}{b^{-\nu}} \cdot (1-\theta)^{-\nu}$.

$\mathcal{G}'(\theta) \sim \dfrac{a^{-\nu}}{b^{-\nu}} \cdot$(terms involving small $(1-\theta)$) $\rightarrow$ (behavior dominated by $(1-\theta)^{-\nu}$)

And as a conclusion because $(1-\theta)^{-\nu}$ diverges to infinity as $\theta \to 1^-$ for $\nu > 0$, we conclude that:

$$\lim_{\theta \to 1^-} \mathcal{G}'(\theta) = \infty.$$

Thus, we have:

$$E[M] = \infty.$$

□

4. We can also show that the variance is undefined:

$$Var[M] \;=\; E[M^2] - (E[M])^2 \;=\; \infty.$$

Proof. We start with the definition of variance and use the *pgf* to derive the necessary limits.

The variance of the discrete random variable M is defined as:

$$Var[M] = E[M^2] - (E[M])^2.$$

Then as we know the first moment or expected value is:

$$E[M] = \lim_{\theta \to 1^-} \mathcal{G}'(\theta).$$

And the second moment $E[M^2]$ is defined as:

$$E[M^2] = \lim_{\theta \to 1^-} \mathcal{G}''(\theta) + \lim_{\theta \to 1^-} \mathcal{G}'(\theta).$$

And substituting this values into the variance formula gives:

$$Var[M] = \lim_{\theta \to 1^-} \mathcal{G}''(\theta) + \lim_{\theta \to 1^-} \mathcal{G}'(\theta) - \left(\lim_{\theta \to 1^-} \mathcal{G}'(\theta) \right)^2.$$

From the previous result, we established that:

$$\lim_{\theta \to 1^-} \mathcal{G}'(\theta) = \infty.$$

Therefore, $(\lim_{\theta\to 1^-} \mathcal{G}'(\theta))^2 = \infty$.

Now we analyze $\lim_{\theta\to 1^-} \mathcal{G}''(\theta)$. We need to show that it is also ∞ as we know that $\mathcal{G}'(\theta)$ diverges:

If $\mathcal{G}''(\theta)$ does not converge to a finite limit, then it will approach either ∞ or remain undefined.

And last, we have:

$$Var[M] = \lim_{\theta\to 1^-} \mathcal{G}''(\theta) + \infty - \infty.$$

Since both $\lim_{\theta\to 1^-} \mathcal{G}''(\theta)$ and ∞ are divergent, we can conclude that:

$$Var[M] = \infty.$$

Therefore, the variance $Var[M]$ is indeed undefined as it diverges to infinity. □

5. More generally, we can show that higher moments are also undefined:

$$E[M^n] \;=\; \lim_{\theta\to 1^-} \mathcal{G}^{(n)}(\theta) \;=\; \infty.$$

and the (*mgf*) does not exist.

Proof. To show that higher moments of M are undefined,we will consider the n-th moment $E[M^n]$ and prove it diverges.

The $\mathcal{M}(\theta)$ for a random variable M is defined as:

$$\mathcal{M}(\theta) = E[e^{\theta M}] = 1 + E[M]\theta + \frac{E[M^2]}{2}\theta^2 + \dots$$

and the n-th moment $E[M^n]$ can be derived from the *mgf*:

$$E[M^n] = \frac{d^n}{d\theta^n}\mathcal{M}(\theta)\bigg|_{\theta=0}.$$

From previous proof, we established that the first moment diverges:

$$E[M] = \lim_{\theta\to 1^-} \mathcal{G}'(\theta) = \infty.$$

Since $E[M] = \infty$, the series for the *mgf* contains an infinite term:

$$\mathcal{M}(\theta) = 1 + \infty\cdot\theta + \frac{E[M^2]}{2}\theta^2 + \dots$$

so the the *mgf* is not defined and since the *mgf* is not defined due to the presence of infinite coefficients, we conclude that the higher moments $E[M^n]$ must also diverge for $n \geq 1$:

$$E[M^n] = \infty \quad \text{for all } n \geq 1.$$

Thus, we conclude that all higher moments of M are undefined because they diverge to infinity.

□

6. While M exhibits a complex distribution with undefined moments, its characteristic function does exist:

$$E[e^{i\theta}] \;=\; \mathcal{G}(e^{i\theta}).$$

We need to proof that the characteristic function exists despite the undefined moments, we begin with the definition of the characteristic function $\varphi_M(t) = E[e^{i\theta M}]$ and see that this can be expressed in terms of the *pgf* $\mathcal{G}(\theta)$ as follows:

$$\varphi_M(\theta) = \mathcal{G}(e^{i\theta}).$$

Next, we analyze the behavior of $\mathcal{G}(e^{i\theta})$ as θ varies. We substitute $\theta = e^{i\theta}$:

$$\mathcal{G}(e^{i\theta}) = \frac{a^{-\nu} K_\nu(2a\sqrt{\mu(1-e^{i\theta})})}{b^{-\nu} K_\nu(2b\sqrt{\mu(1-e^{i\theta})})}.$$

As $\theta \to 0$, we have $e^{i\theta} \to 1$. The functions $K_\nu(2a\sqrt{\mu(1-e^{i\theta})})$ and $K_\nu(2b\sqrt{\mu(1-e^{i\theta})})$ remain bounded. Therefore, $\mathcal{G}(e^{i\theta})$ is well-defined for all θ.

In conclusion, even though the moments of M are undefined, the characteristic function $\varphi_M(\theta)$ remains valid since $\mathcal{G}(e^{i\theta})$ converges for all θ.

7. We could use a Fourier method to compute the probabilities $p_k = \mathbb{P}\{M = k\}$. In section 4.2.1, we will explain how these values are obtained.

4.2 Obtaining probabilities for neutral differences between sequences from the *pgf*, $\mathcal{G}(\theta)$

Before discussing the method we will use to compute the probabilities, we must first highlight an important issue related to $\mathcal{G}(\theta)$.

$$\lim_{\theta \to 1^-} \mathcal{G}(\theta) \;=\; \begin{cases} \left(\frac{b}{a}\right)^{2\nu}, & \nu > 0. \\ 1, & \nu \leq 0. \end{cases} \tag{4.13}$$

We recall that a *pgf* is defined as:

$$\mathcal{G}(\theta) = p_0 + p_1\theta + p_2\theta^2 + \cdots \tag{4.14}$$

where p_i is the probability that the two sequences have i neutral differences. The total probability mass is therefore given by $\mathcal{G}(1)$, which represents the sum of all probabilities.

From this we can conclude that the two lineages will meet with probability one in two dimensions or less ($\nu \leq 0$). However, in more than two dimensions ($\nu > 0$), they will encounter each other with the following probability:

$$\left(\frac{b}{a}\right)^{2\nu}, \quad a > b \tag{4.15}$$

where a and b represent the initial and final Euclidean distances, respectively. This implies that the two lineages do not meet in physical space with probability

$$1 - \left(\frac{b}{a}\right)^{2\nu} \tag{4.16}$$

which can be significant, in particular in more than two dimensions.

The associated *cf* is given by:

$$\mathcal{C}(\theta) = p_0 + p_1 e^{i\theta} + p_2 e^{2i\theta} + \cdots \tag{4.17}$$

Thus, the characteristic function is related to the *pgf* as follows:

$$\mathcal{C}(\theta) = \mathcal{G}(e^{i\theta}). \tag{4.18}$$

The above observation presents us with a practical problem: How can we determine if the observed neutral differences between two sequences are representative of the evolutionary time between two lineages that did meet in the past, or if they are erroneous differences between two lineages that did not meet in the past?

In more than two dimensions, there may be a significant probability that the lineages never met in the past. This raises the question of whether there exists a numerical "signature" that can identify lineages that have never met.

In an infinite genome, one might hypothesize that lineages that never met would have an infinite number of neutral differences, as they are separated by infinite time. However, this is not possible in reality. We can consider two potential solutions to this problem:

1. We could compare sequences with a very distantly related variant or subspecies (called an "outgroup" in phylogenetics) and declare that distances similar to those from the outgroup can be considered to be "effectively infinite".

2. We could assume that the sequences must have met some time in the past, within the subgroup or subspecies in question. In this case, we would condition on the sequences meeting, and our first guess is that the resulting *pgf* is simply:

   $$\mathcal{G}^*(\theta) \;=\; \begin{cases} \mathcal{G}(\theta)\ , & \nu \leq 0. \\ (a/b)^{2\nu}\mathcal{G}(\theta)\ , & \nu > 0. \end{cases} \tag{4.19}$$

   such that $\mathcal{G}^*(1) = 1$ is always true. Strictly speaking, this is a condition that must hold if we are to call it a *pgf*, so $\mathcal{G}(\theta)$ is, technically speaking, not a *pgf*.

   Why can we assure that $\mathcal{G}(\theta)$ is not a *pgf*? We can prove by contradiction that $\mathcal{G}(\theta)$ does not satisfy the properties to be a *pgf* by observing that it does not meet one of the three necessary conditions. Recall that the three conditions for a function to be a valid *pgf* are: The first is non-negativity, meaning it is non-negative for $\theta \in [0, 1]$. The second is normalization, that is, it satisfies $\mathcal{G}(1) = 1$. Last, power series representation, which means it is a power series in θ, which can be expressed in the form:

   $$\mathcal{G}(\theta) = \sum_{k=0}^{\infty} p_k \theta^k,$$

   where $p_k = \mathbb{P}(M = k)$ and $p_k \geq 0$. Let us check that the condition $\mathcal{G}(1) = 1$ is not satisfied.

   Recalling that the function $\mathcal{G}(\theta)$ is defined as:

$$\mathcal{G}(\theta) = \frac{a^{-\nu} K_\nu(2a\sqrt{\mu(1-\theta)})}{b^{-\nu} K_\nu(2b\sqrt{\mu(1-\theta)})}.$$

Thus, substituting $\theta = 1$ into $\mathcal{G}(\theta)$:

$$\mathcal{G}(1) = \frac{a^{-\nu} K_\nu(2a\sqrt{\mu(1-1)})}{b^{-\nu} K_\nu(2b\sqrt{\mu(1-1)})}.$$

We need to evaluate $K_\nu(0)$, and it is also known that the behavior of the Bessel function $K_\nu(x)$ as $x \to 0$ is as follows:

$$K_\nu(x) \sim \frac{\Gamma(\nu)}{2}\left(\frac{2}{x}\right)^\nu \quad \text{for } \nu > 0.$$

Therefore, $K_0(0) = \infty$, which is undefined. Hence, if $\nu > 0$, $\mathcal{G}(1)$ diverges, indicating that $\mathcal{G}(\theta)$ does not satisfy the required condition for a valid *pgf*.

4.2.1 Fourier method for deriving *pgf*

To compute the probabilities $p_k = \mathbb{P}\{M = k\}$, we employ a Fourier method for back-transforming the *pgfs*. In this section, we will explore how Fourier transforms can be used to derive the probability distributions, denoted as p_k, and present a computational analysis to demonstrate that these values indeed represent the probabilities of the random variable M. In the following subsection, we will see that, after obtaining the values of the *pgfs* using this formula, this approach allows us to utilize Bayesian methods, including maximum log-likelihood estimation, to estimate the parameter ν. From this estimation, we can derive the "effective dimensionality" $\mathcal{D}_e = 2\nu + 2$.

Fourier method for back-transforming (*pgf 's*)

Suppose X is an integer-valued random variable with support in the set $\{0, 1, \ldots, N-1\}$. The characteristic function of X is given by

$$\mathcal{C}_X(\omega) = \sum_{k=0}^{N-1} e^{i\omega k} p_k, \qquad (4.20)$$

where $p_k = P(X = k)$ is the *pmf.*

Using Euler's formula $e^{i\theta} = \cos\theta + i\sin\theta$, we can easily verify that $\mathcal{C}_X(\omega)$ has period 2π, that is, it satisfies $\mathcal{C}(\omega) = \mathcal{C}(\omega + 2\pi)$ for all ω, since

$$\begin{aligned} e^{(i(\omega+2\pi)k)} &= e^{i\omega k}e^{ik2\pi} \\ &= e^{i\omega k}(\cos(k2\pi) + i\sin(k2\pi)) \\ &= e^{i\omega k}(1+0) \\ &= e^{i\omega k}. \end{aligned}$$

Also, the characteristic function is real-valued if and only if the corresponding distribution function is symmetric around the origin. An example with a real-valued characteristic function is a random variable that has a standard normal distribution.

To see the connection with the discrete-Fourier transform, we evaluate the characteristic function at N equally spaced values in the interval $[0, 2\pi]$

$$c_m = \mathcal{C}_X\left(\frac{2\pi m}{N}\right) = \sum_{k=0}^{N-1} p_k e^{i2\pi km/N}, m = 0, 1, \ldots, N-1.$$

Here, $\mathcal{C}$ and P form a Fourier transform pair. The above equation defines the *dft* of the sequence of probabilities $p_0, \ldots, p_{N-1}$. As mentioned earlier, the c_m's are in general complex numbers. Also note that extension of the range of m outside the range $\{0, 1, \ldots, N-1\}$ will result in a periodic sequence consisting of a repetition of the sequence $c_0, \ldots, c_{N-1}$.

Our interest is in recovering the sequence of probabilities from the corresponding sequence of characteristic function values. In other words, we seek to obtain the sequence of p_k's from the sequence of c_m's. This can be accomplished by using the inverse *dft* operation which is defined by:

$$p_k = \frac{1}{N}\sum_{m=0}^{N-1} c_m e^{-i2\pi km/N}, k = 0, 1, \ldots, N-1.$$

Calculating probability distribution p_k

In this section, now we are going to describe the process of calculating the probability distribution p_k of M, the discrete random variable that represents the number of neutral differences between two sequences. Employing the Fourier method for back-transforming *pgf*, the probabilities $p_k = \mathbb{P}\{M = k\}$ quantify the occurrence of k neutral differences, where k can take integer values from 0 to $N-1$.

We can determine the p_k using the following formula:

$$p_k = \frac{1}{N}\sum_{m=0}^{N-1} c_m e^{-i2\pi km/N}, k = 0, 1, \ldots, N-1. \tag{4.21}$$

where:

$$c_m = \mathcal{C}\left(\frac{2\pi m}{N}\right) = \mathcal{G}(e^{i2\pi m/N}) = \sum_{k=0}^{N-1} p_k e^{i2\pi km/N}, m = 0, 1, \ldots, N-1. \tag{4.22}$$

Thus, we can express p_k as:

$$p_k = \frac{1}{N}\sum_{m=0}^{N-1} \mathcal{G}(e^{i2\pi m/N})\ e^{-i2\pi km/N}, k = 0, 1, \ldots, N-1. \tag{4.23}$$

where $\mathcal{G}(\theta)$ is derived above.

Computational analysis

To verify that the calculated values p_k form a valid probability distribution for the discrete random variable M, we must ensure that the following properties hold.

First, the non-negativity condition requires that each probability satisfies $p_k \geq 0$ for all $k = 0, 1, \ldots, N-1$. Second, the total probability must sum to one, that is,

$$\sum_{k=0}^{N-1} p_k = 1.$$

To demonstrate these properties, we performed a computational analysis using the code provided in the Appendix A.0.1. The results are shown for three specific values of ν, corresponding to different spatial dimensions:

1. $\nu = -\frac{1}{2}$: corresponds to one spatial dimension.
2. $\nu = 0$: corresponds to two spatial dimensions.
3. $\nu = \frac{1}{2}$: corresponds to three spatial dimensions.

Figures 4.1, 4.2, 4.3, and 4.4 show the results for different values of the parameters. These figures help us visualize how the probabilities p_k behave under various conditions. Each one corresponds to a specific spatial dimension, represented by the Bessel order ν. We can observe how changing the parameters affects the overall shape of the distribution, particularly how the values of p_k vary in response.

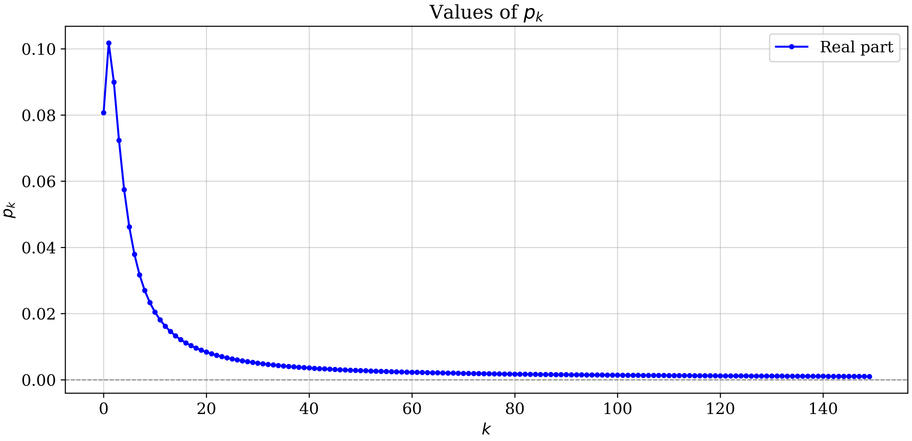

Figure 4.1: Values of p_k showing real and imaginary parts. With the following parameters: value for $a = 5.0$, value for $b = 1.0$, order of the Bessel $\nu = -0.5$, value for $m = 0.1$, number of values for $N = 150$.

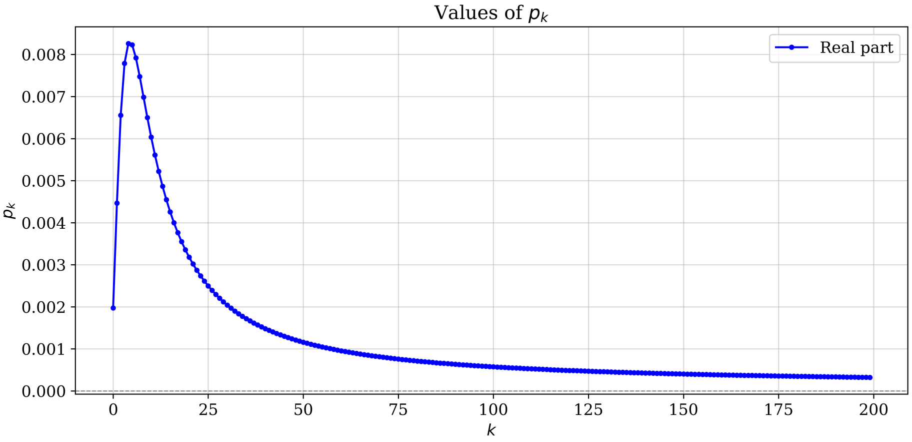

Figure 4.2: Values of p_k showing real and imaginary parts. With the following parameters: value for $a = 4.0$, value for $b = 1.0$, order of the Bessel $\nu = 0.5$, value for $m = 0.7$, number of values for $N = 200$.

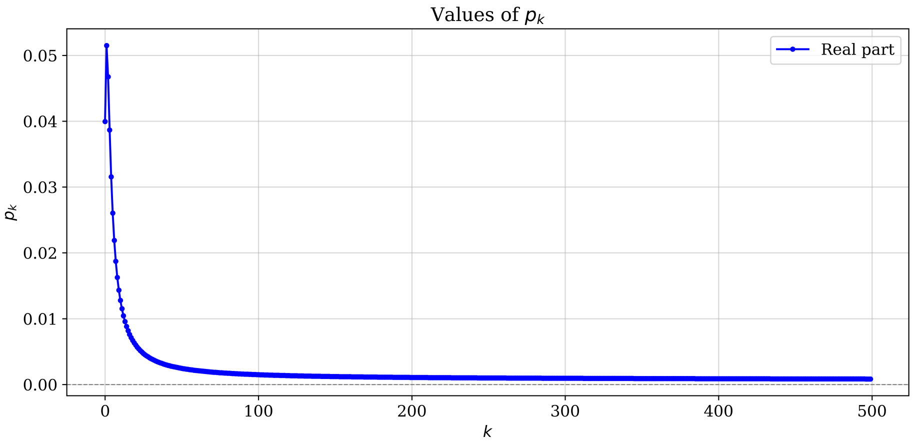

Figure 4.3: Values of p_k showing real and imaginary parts. With the following parameters: value for $a = 5.0$, value for $b = 1.0$, order of the Bessel $\nu = 0$, value for $m = 0.1$, number of values for $N = 500$.

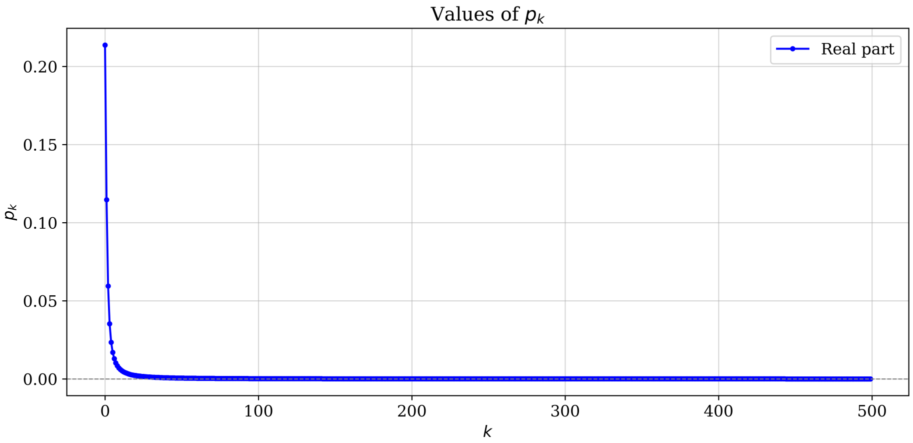

Figure 4.4: Values of p_k showing real and imaginary parts. With the following parameters: value for $a = 1.6$, value for $b = 1.0$, order of the Bessel $\nu = 0.5$, value for $m = 0.8$, number of values for $N = 500$.

Chapter 5

Estimation of physical effective dimensionality from sequence data

In this chapter, we explore the concept of effective dimensionality applied to real sequence data. Using data from the study by Nguyen et al. (2024) [55], we demonstrate the utility of this method. Specifically, we analyze sequences obtained from simulations based on datasets of avian influenza A(H5N1).

This chapter provides both the theoretical foundations and the computational framework necessary for applying the effective dimensionality method, together with the corresponding code in Appendix A.0.1. We focus on three important genes: HA, NA and MP. For each of these genes, we analyze the pairwise synonymous differences, plot the resulting graph on a log-log scale, and estimate the effective dimensionality using the maximum log-likelihood method. These estimates will be performed for each gene of the HPAI-H5N1 virus, which has been evolving in birds, cattle, and potentially humans in recent months.

5.1 Highly pathogenic avian influenza A(H5N1) in dairy cattle in 2024

The term HPAI A(H5N1) refers to a specific strain of the Highly Pathogenic Avian Influenza (HPAI) virus. HPAI is a type of virus that causes severe illness and high mortality. The letter “A” denotes the type of influenza virus. There are different types of influenza viruses (A, B, C, or D), with Type A being the most common and often causing extensive outbreaks in animals and humans. The designation (H5N1) refers to the subtype of the influenza virus. For more information on Avian influenza A (H5N1), see [58].

In the paper by Nguyen et al. (2024) [55], the authors studied the transmission of the virus from wild birds to cattle across species barriers. This type of cross-species transmission, called zoonosis, is of great concern to public health.

The emergence of HPAI A(H5N1) in dairy cattle across North America in 2024 has raised significant concerns regarding animal health, public safety, and the dairy industry.

Recent genomic analyzes have identified the H5N1 virus clade 2.3.4.4b as the primary strain that infects dairy cattle. This strain has been detected in multiple herds in North America, with reports indicating that more than 230 dairy farms have been affected since its initial identification in March 2024 [17], [80]. The virus was found to have undergone a re-assortment event, the process in which genetic material is exchanged or recombined between two or more viruses [71], resulting in the emergence of genotype B3.13. This genotype is related to both wild birds and domestic livestock [37], [38]. The Texas Panhandle, located along migratory bird routes, has been highlighted as a likely source of the outbreak, and migratory birds serve as natural reservoirs for avian influenza viruses [14], [57].

The detection of H5N1 in dairy cattle marks a significant shift in the understanding of influenza virus host range, as cattle were previously considered resistant to such infections [13], [47]. Studies have shown that the virus can bind to sialic acid receptors present in the bovine mammary gland, facilitating infection [47]. This binding capability is critical, as it suggests that H5N1 can replicate in bovine tissues, leading to potential viral shedding through milk. Indeed, high levels of the virus have been detected in raw milk, raising concerns about food safety and the risk of zoonotic transmission from animals to humans [68], [70].

The implications of H5N1 in dairy cattle extend beyond animal health, as the virus also poses a risk to human health through direct contact with infected animals and contaminated dairy products. Reports indicate that human infections have been confirmed in connection with the outbreak, highlighting the zoonotic potential of this virus. Furthermore, the persistence of H5N1 in unpasteurized milk and on milking equipment surfaces presents additional challenges for dairy workers, who may be at increased risk of exposure [65].

Environmental monitoring has also revealed the presence of H5N1 RNA in wastewater from various cities, suggesting that the virus is circulating more widely than previously understood [72]. This environmental persistence underscores the need for comprehensive surveillance and control measures to mitigate the spread of HPAI among livestock and prevent potential spillover events into human populations [55].

A signature of zoonotic potential can be found in discordant evolution across different genes, as divergence in receptors between different animal hosts can place evolutionary constraints that differ across gene segments.

5.2 Gene segments of influenza A virus

Influenza viruses are characterized by segmented genomes. Influenza Type A has eight segments, each encoding proteins essential for the virus's replication, assembly, and interaction with host cells. For more information on the names of each segment, see [8]. The location of each of these segments is shown in Figure 5.1.

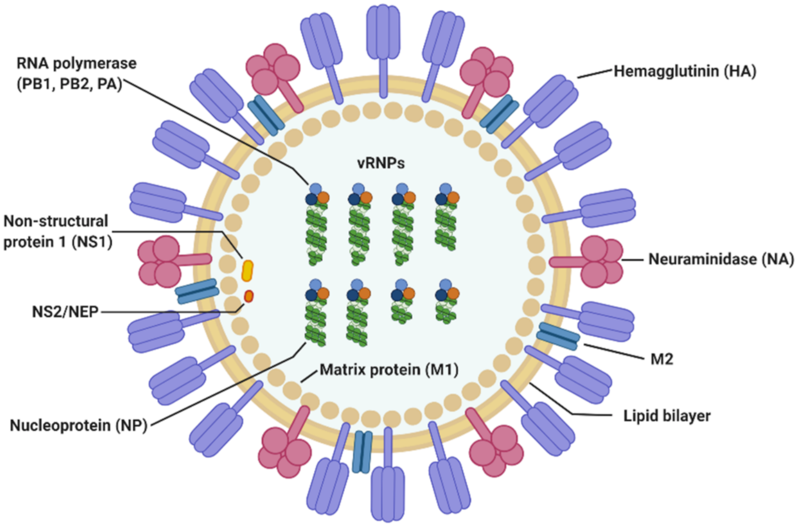

Figure 5.1: Influenza A virus genome structure. We will focus on the two segments that are expose to immune surveillance namely *Hemagglutinin* (HA), *Neuraminidase* (NA), and one segment that is not exposed to immune surveillance, *Matrix Protein* (MP). Figure source: [41]

The gene segments of the influenza A virus genome that we studied are *hemagglutinin* (HA) and *neurominidase* (NA). These two genes are crucial because they determine the ability of the virus to infect host cells and facilitate its spread. HA plays a critical role in the initial attachment of the virus to the host cell by binding to sialic acid receptors on the cell surface. This binding facilitates the fusion of the viral envelope with the host cell membrane, allowing the virus to enter the host cell and initiate infection [45]. NA, on the other hand, is involved in the release of new viral particles from infected cells. It cleaves sialic acid residues on the surface of the host cell, preventing the newly formed virions from being trapped on the host cell membrane, allowing them to spread to new cells and tissues [45].

The third gene is the *matrix protein* (MP), which encodes the M1 protein. M1 is essential for the assembly, stability, and morphogenesis of the influenza virus. It forms a structural component of the virus particle, surrounds the RNA genome, and interacts with other viral proteins to ensure the correct assembly of new virions. M1 also plays a role in the regulation of the viral replication cycle, including its transport and initiation from the host cell [8]. Given its essential role in the structural integrity of the virus, we do not expect M1 to change significantly, as alterations in this gene would likely alter the ability of the virus to replicate and spread [77].

These three genes, HA, NA and M1, are not only critical to the ability of the virus to infect and replicate, but are also targets for the development of vaccines and antiviral drugs. Vaccines often aim to generate an immune response against the surface proteins HA and NA, since these are the proteins that the immune system can recognize and attack [45]. Mutations in these genes, particularly HA, can lead to the emergence of new viral strains that can evade immunity, as seen with seasonal flu strains and pandemics [54].

The segmented nature of the influenza genome allows for a high degree of genetic diversity, particularly when different strains of the virus co-infect the same host. This can lead to re-assortment events, in which gene segments of different viral strains are exchanged, creating new strains with altered characteristics. This ability to undergo re-assortment is a major driver of influenza's rapid evolution and is a factor in its potential to cause pandemics [71]. Understanding the genetic variation within these gene segments is essential for predicting the virus's behavior and preparing for future outbreaks.

5.3 Results

In our study, we focus on the genetic diversity within the HA, NA, and MP genes of the HPAI-H5N1 virus. Changes in these genes can provide important information about the evolution of the virus, its ability to cross species barriers, and its potential for adaptation to new hosts.

In the first graphs, we develop an analysis of pairwise synonymous differences in each of HA, NA, and MP genes. A pairwise synonymous difference refers to the comparison of two genetic sequences obtained from [48] to identify synonymous mutations. This means that we want to look at changes in the RNA sequence that do not change the encoded protein. These mutations occur in the coding regions of a gene but do not alter the amino acid sequence because multiple codons can code for the same amino acid this could be seen in the table explain in 2.1. For further information, this source has an excellent explanation [60].

For example, in the sequence of H5N1, a change from GGU to GGC in the codon for glycine would be a synonymous mutations, as both codons code for the same amino acid. Then, we compare the sequences of two different strains, this means that they could be taken from samples in different geographic locations. Before the comparison, both of the sequences need to be align so they have the same length. We then count the differences between the sequences and plot a histogram of this results. This process was implemented in a program, as detailed A.0.2.

We will now explain the results we obtained for three important genes: HA, NA, and MP. These genes play crucial roles in the evolution and functionality of the HPAI-H5N1 virus, and understanding their mutations can provide insights into viral behavior, transmission, and potential impacts on different host species. In the following sections, we will present our findings for each gene, focusing on synonymous mutations, evolutionary patterns, and their inferred effective dimensionality.

5.3.1 Results for HA gene

The purpose of this analysis is to study the evolutionary dynamics of the influenza virus by examining the mutations and how the virus evolves over time in different host environments—first in birds, then in cattle (and potentially humans) during recent months, as shown in Figure 5.2. Understanding these mutations is crucial, as they can influence the virus's ability to adapt to different species, evade immune responses, or develop resistance to treatments.

The script for this figure begins by reading the sequences from the FASTA file and then using a codon-to-amino acid mapping to translate the sequences. It compares pairs of sequences by dividing them into codons and checking for synonymous mutations, where the codons differ but still code for the same amino acid.

This process is repeated 10,000 times, with random pairs of sequences selected for comparison in each iteration. The results, which represent the number of synonymous mutations detected in each comparison, are stored in a list and displayed as a histogram to visualize their distribution, as shown in Figure 5.2. The histogram helps identify patterns in mutation rates, which can reveal insights into how the virus adapts over time in different host environments.

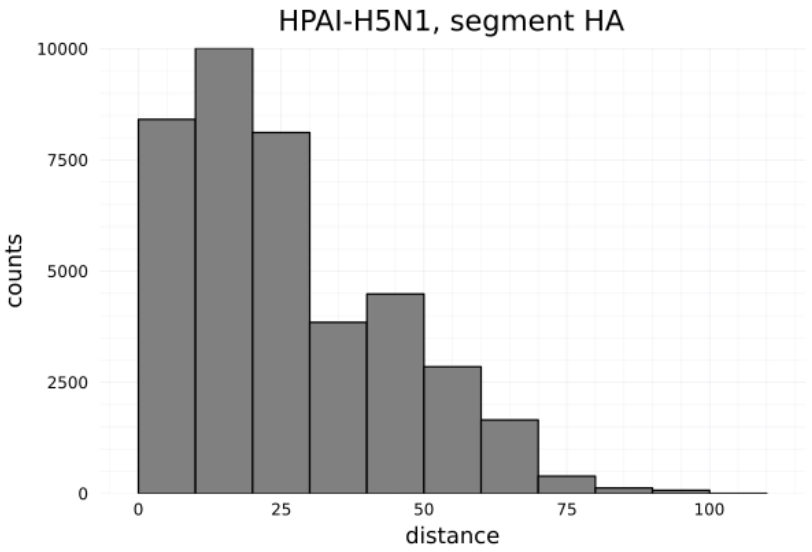

Figure 5.2: Rescaled pairwise synonymous differences (distance) in the HA gene of the Highly Pathogenic Avian Influenza (HPAI-H5N1) virus that has been evolving in birds, cattle (and potentially humans) during recent months.

Once we have the results and the histogram plot, as shown in Figure 5.2, we then plot the histogram on a log-log scale. This transformation helps us analyze the tail of the distribution more effectively, as shown in Figure 5.3.

By focusing on the tail, we can better understand the behavior of rare mutations, which are often of particular interest in evolutionary studies. This is helpful because we can observe that the tail of the distribution appears to approximate a straight line, reflecting a power-law tail, which is in agreement with theoretical predictions. A power-law distribution suggests that rare mutations occur more frequently than would be expected under a normal distribution, and this characteristic has been observed in various biological and evolutionary systems.

This observation is significant because it suggests that the influenza virus follows a scale-invariant process in terms of mutation patterns, where the frequency of mutations follows a predictable, nonlinear relationship as the virus adapts to different environments. Such insights can provide important information on how the virus evolves in response to immune pressure or environmental changes, which may have implications for vaccine development or understanding viral transmission dynamics.

For the following results, we will use the code provided in Appendix A.0.1.

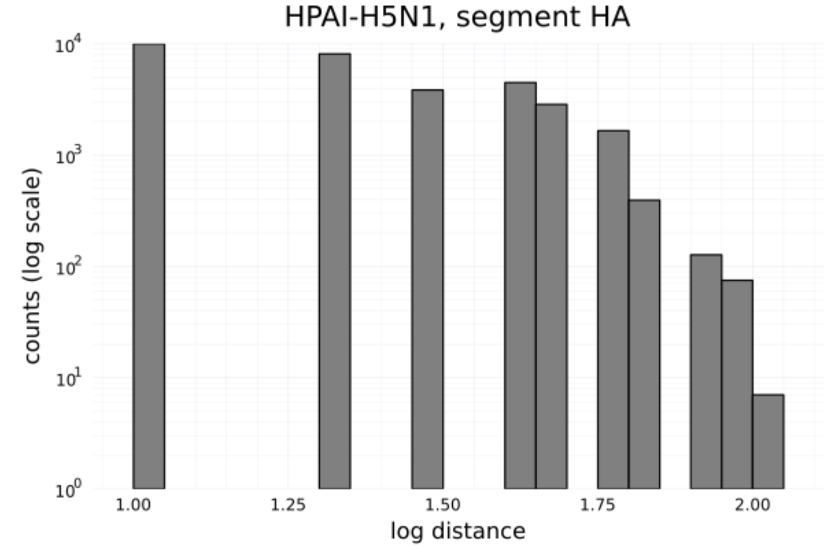

Figure 5.3: Rescaled pairwise synonymous differences in the Hemagglutinin (HA) gene. Same as Fig 5.2 but on log-log scale. We note that the tail of the distribution appears to approximate a straight line, reflecting a power-law tail which is in agreement with theoretical predictions.

This code calculates the probability distribution of the values of p_k, as discussed in Section 4.2.1. This section outlines how the values of p_k are constructed, and we have computed these values.

First, we will examine the code used to calculate the probability distribution p_k (see Appendix A.0.1). The parameters a, b, ν, and μ are defined at the beginning of the code, as they are necessary to calculate the Bessel function of the second kind, K_ν. The code initializes an array p_k to store the computed values and then iterates over the indices k and m, both of 0 to $N-1$. For each pair of k and m, it calculates θ and uses it to compute the associated terms involving K_ν. If $m=0$, a simplified formula for G_θ is applied; otherwise, the Bessel function values for both a and b are computed, ensuring their validity. The computed values are accumulated in a sum and normalized to obtain p_k. Finally, the values of p_k are plotted, providing a visual representation of the probability distribution. The resulting plot is shown in Figures 4.1, 4.2, 4.3 and 4.4.

Next, we plot the code for the maximum log-likelihood, as detailed in the appendix A.0.4. We note that maximum likelihood is inherently Bayesian (with a flat prior); this observation suggests possibilities for future work in which the prior is not flat.

The code A.0.4 calculates the maximum log-likelihood of the probability

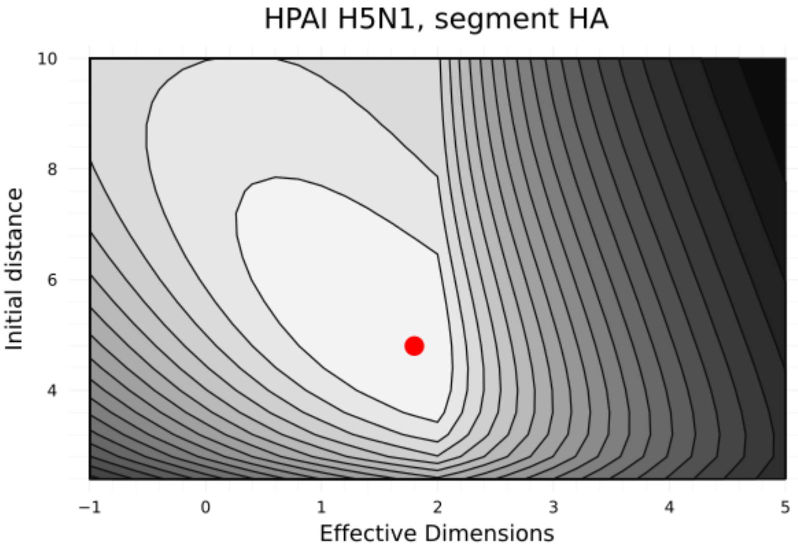

Figure 5.4: Log-likelihood surface as a function of effective physical dimensions ($\hat{\mathcal{D}}_e$) and average initial distance ($a(\mathcal{N}_e)$); the red dot signals the maximum which occurs at $\hat{\mathcal{D}}_e \approx 1.8$. Here we examine the HA gene.

distribution p_k. First, the parameters a, b, ν, and μ are defined. An array p_k is initialized to store the values. The code iterates over k and m (both from 0 to $N-1$) and computes $\theta = e^{i2\pi m/N}$. For each pair (k, m), it calculates G_θ. The values of G_θ are accumulated in sum$_G$ and normalized by dividing by N to obtain p_k. Next, the code computes the logarithms of the absolute values of the normalized p_k, ensuring that any invalid values (NaN or infinite) are set to 0. The sum of the logarithms is calculated, and the maximum log-likelihood value is found. Finally, we plot the log-likelihood for the n pairwise distances, and for each set of parameters $\{\nu, a\}$ was computed as:

$$LL = \sum_{k=1}^{n} \log p_k, \tag{5.1}$$

where the p_k are calculated by inverse Fourier transform of the *pgf*, as outlined in Section 4.2.1. In Figure 5.4, we examine the results for the HA gene, where the red dot signals the maximum, which occurs at $\hat{\mathcal{D}}_e \approx 1.8$. This is the effective dimensionality for the HA gene.

5.3.2 Results for NA gene

We then repeat the same analysis for the pairwise synonymous differences in the NA gene. As previously mentioned, a synonymous mutation occurs when a change in the RNA sequence does not alter the amino acid encoded by that sequence.

The code begins by reading the sequences from the FASTA file and then uses a codon-to-amino acid mapping to translate the sequences. It identifies synonymous mutations by comparing codons that differ but encode the same amino acid, after dividing the pairs of sequences into codons.

This process is repeated 10,000 times, with random pairs of sequences selected for comparison in each iteration. The output of the 10,000 iterations is a distribution of the number of synonymous mutations observed in pairwise comparisons. The results are stored in a list and displayed as a histogram to visualize their distribution, as shown in Figure 5.5.

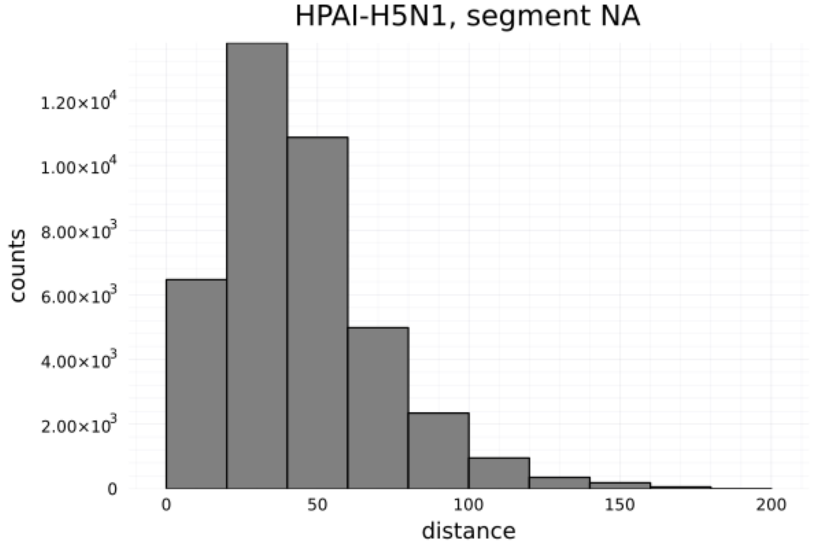

Figure 5.5: Rescaled pairwise synonymous differences in the NA gene of the HPAI-H5N1 virus that has been evolving in birds, cattle, and potentially humans during recent months.

Figure 5.6 shows the pairwise synonymous differences in the NA gene of the influenza virus, similar to Figure 5.5, but presented on a log-log scale. By transforming the data to a log-log scale, we can better visualize the tail of the distribution, which appears to approximate a straight line. This straight-line pattern in the tail suggests a power-law distribution.

The same statistical analysis used for the HA gene was repeated for the NA gene to calculate the probability distribution p_k, which represents the probability of observing k synonymous differences. As before, the parameters a, b, ν, and μ, defined at the beginning of the code, are essential for computing the modified Bessel function of the second kind, K_ν.

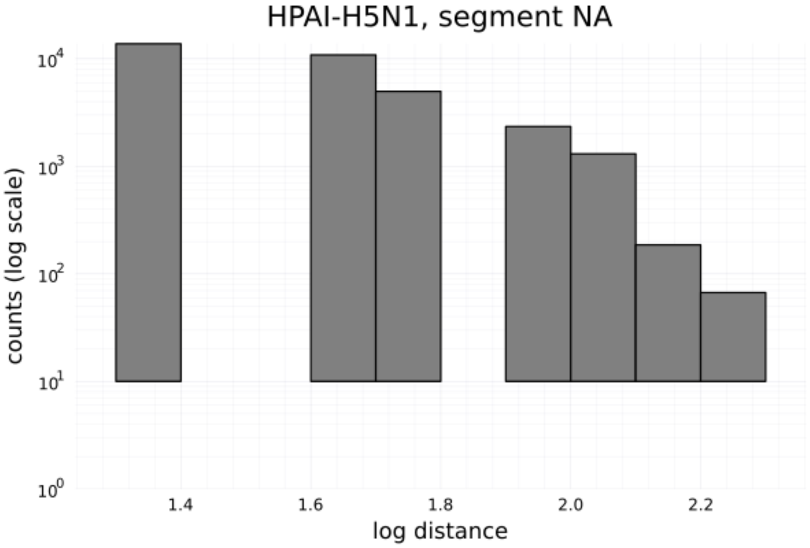

Figure 5.6: Rescaled pairwise synonymous differences NA gene. Similar to Fig 5.5, but displayed on a log-log scale. It is observed that the tail of the distribution appears to approximate a straight line, reflecting a power-law tail, which is in agreement with theoretical predictions from population genetics models.

After initializing an array to store the computed p_k values, the code iterates over indices k and m, which represent specific pairwise comparisons. For each pair, θ is calculated, and subsequently used to compute terms involving K_ν. The calculated values are then accumulated and normalized to obtain p_k, and the resulting probability distribution is plotted, as shown in Figure 5.6.

The maximum log-likelihood calculation, following the procedure described for the HA gene, was repeated for the NA gene. The code, detailed in Appendix A.0.4, computes the log-likelihood surface for the NA gene. The resulting surface and its maximum, indicated by the red dot, are shown in Figure 5.7. The effective dimensionality for the NA gene was estimated to be $\hat{\mathcal{D}}_e \approx 1.1$.

5.3.3 Results for MP gene

For the MP gene, synonymous mutations were also analyzed to understand the evolutionary mutation patterns that do not affect the encoded protein, since synonymous mutations do not alter the encoded amino acids.

The distribution of these synonymous mutations was represented in a histogram, shown in Figure 5.8. This histogram captures the frequency

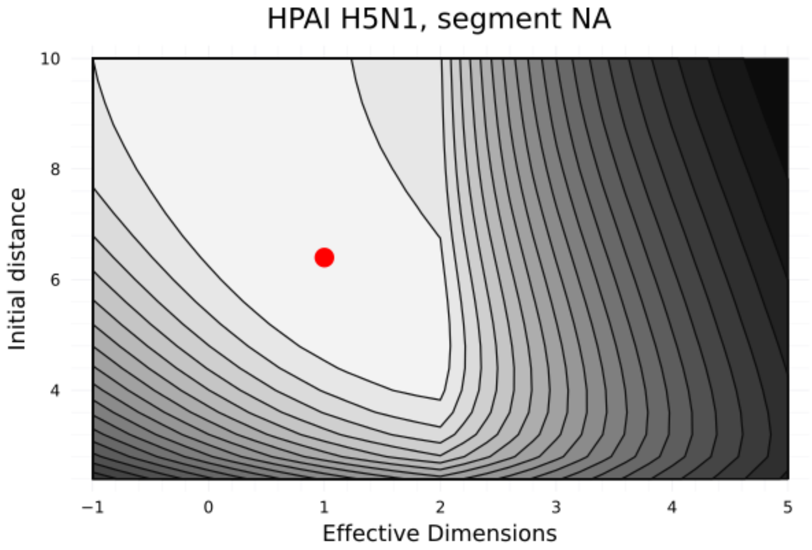

Figure 5.7: Log-likelihood surface as a function of effective physical dimensions ($\mathcal{D}_e$), representing the dimensionality of the evolutionary space, and average initial distance ($a(\mathcal{N}_e)$), reflecting the initial genetic divergence; the red dot indicates the maximum which occurs at $\hat{\mathcal{D}}_e \approx 1.1$. Here we examine the NA gene of the HPAI-H5N1 virus.

distribution of synonymous mutations between pairs of sequences from the MP gene.

The log-log scale transformation of this histogram, as shown in Figure 5.9, highlights the tail of the distribution and reveals a power-law relationship, consistent with theoretical predictions. The low estimated dimensionality, indicative of a limited evolutionary space, is consistent with the long-term evolutionary stability of the MP gene, which is expected due to its critical role in viral replication.

Lastly, the log-likelihood surface for the MP gene, presented in Figure 5.10, provides insight into how the gene evolves under different genetic conditions. The contour plot visualizes the log-likelihood values across various parameter sets, specifically effective dimensionality and average initial distance, with the red dot marking the optimal fit at $\hat{\mathcal{D}}_e \approx 1.8$. The log-likelihood values were computed by summing the logarithms of probabilities, where the p_k values were derived through the inverse Fourier transform of the *pgf*, as detailed in Section 4.2.1.

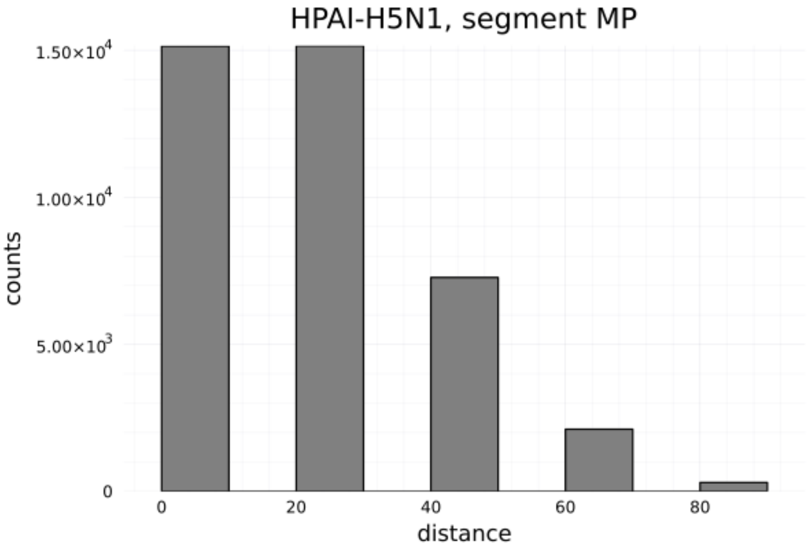

Figure 5.8: Rescaled pairwise synonymous differences in the MP gene of the HPAI-H5N1 virus that has been evolving in birds, cattle, and potentially humans during recent months.

5.4 Interpreting effective dimensions

We estimated the effective dimensionality for the HA, NA, and MP genes. The HA and NA genes, which encode surface proteins, are subject to strong immune surveillance, resulting in rapid evolutionary rates. In contrast, the MP gene, which encodes internal matrix proteins, is highly conserved due to purifying selection, leading to slower evolutionary rates. The estimated effective dimensionalities were $\hat{\mathcal{D}}_e \approx 1.8$ for the MP gene, $\hat{\mathcal{D}}_e \approx 1.1$ for the NA gene, and $\hat{\mathcal{D}}_e \approx 1.8$ for the HA gene.

Our maximum likelihood estimates revealed distinct effective dimensionalities for each gene. Notably, the NA gene's dimensionality differed significantly from both HA and MP, potentially reflecting divergent evolutionary pressures associated with zoonotic transmission. While we expected the HA gene's dimensionality to resemble that of NA due to their shared exposure to immune surveillance, we found it to be more similar to that of the MP gene. The biological mechanisms underlying this unexpected similarity remain unclear; however, a recent study reported a similar discrepancy, although not in the context of spatial structure. While this thesis focuses on estimating the effective spatial dimensionality $\hat{D}_e$ from genetic data, several intriguing questions remain regarding its broader interpretation.

An example illustrating how effective dimensionality can be interpreted to provide insight into spatial spread is as follows. In a two-dimensional

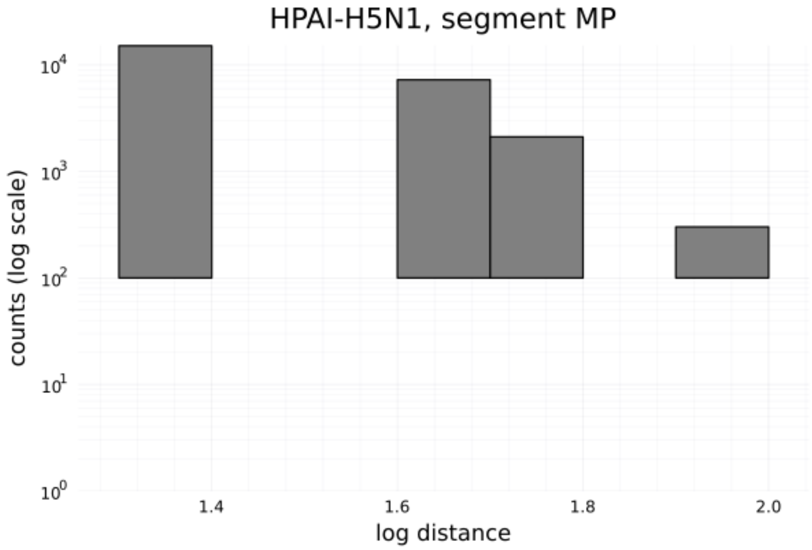

Figure 5.9: Rescaled pairwise synonymous differences in the MP gene. Similar to Fig 5.8, but displayed on a log-log scale. It is observed that the tail of the distribution appears to approximate a straight line, reflecting a power-law tail, which is in agreement with theoretical predictions from population genetics models.

context, an estimated dimensionality greater than 2—for example, 2.3—may suggest the presence of convection, meaning that gene lineages are spreading faster than expected under classical diffusion. In contrast, if the estimated dimensionality is less than 2 — for example, 1.8 or 1.1 — this indicates slower than diffusive behavior.

In the analysis of the H5N1 data presented in this study, the HA and MP gene segments both provide effective dimensionality estimates of $\hat{D}_e \approx 1.8$, consistent with near-diffusive behavior but with some spatial constraint. The NA segment, however, showed a much lower estimate of $\hat{D}_e \approx 1.1$, indicating highly restricted spatial movement. This suggests that the transmission dynamics of the NA segment may be influenced by stronger ecological or structural constraints than the other segments.

This interpretation becomes especially relevant in the context of epidemiological modeling. Slower than diffusive spread implies that the virus or pathogen takes longer to reach new hosts, reducing the likelihood of transmission and potentially slowing the overall progression of the epidemic. Therefore, estimating effective dimensionality offers not only insights into spatial population structure, but also practical information for anticipating or controlling the dynamics of infectious disease.

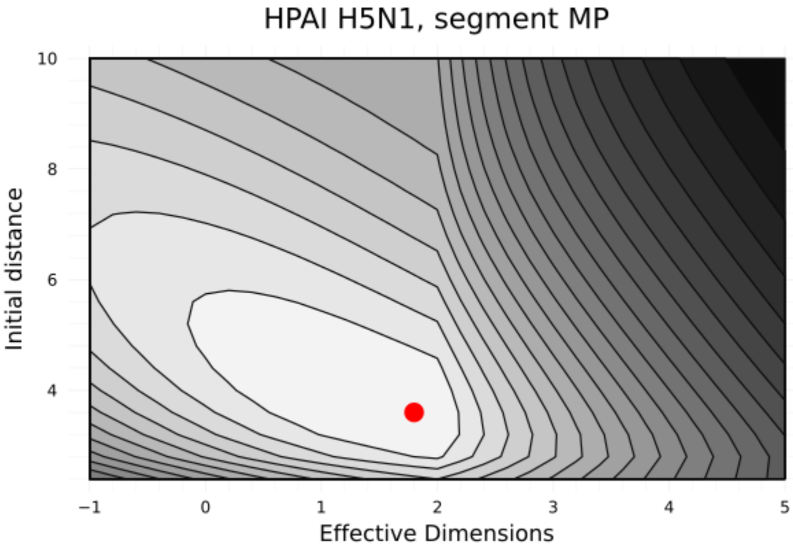

Figure 5.10: Log-likelihood surface as a function of effective physical dimensions ($\mathcal{D}_e$), representing the dimensionality of the evolutionary space, and average initial distance ($a(\mathcal{N}_e)$), reflecting the initial genetic divergence; the red dot indicates the maximum which occurs at $\hat{\mathcal{D}}_e \approx 1.8$.

Our estimate of $\hat{D}_e \approx 1.8$ lies close to the fractal dimension of well-known structures such as the Sierpinski triangle ($D_f \approx 1.6$) and critical percolation clusters in two dimensions ($D_f \approx 1.89$). While these parallels are purely heuristic, they raise interesting possibilities. In particular, spatial structures near criticality often exhibit constrained connectivity and limited dispersal paths, features that may be mirrored in our estimates of effective dimensionality.

Moreover, in the context of diffusion theory [35], the quantity D_e governs how the "mass" (or number of individuals) within a ball of radius r scales for small r. A value of $\hat{D}_e < 2$ implies that the rate at which individuals encounter one another is lower than would be expected under classical two-dimensional diffusion. This may provide insight into the epidemiological vulnerability of structured populations.

Interpreting effective dimensionality in non-diffusion terms remains a challenging task, and future research is needed to clarify how this concept relates to non-diffusive processes. Such work could provide a deeper understanding of population structure and how evolutionary forces shape genetic diversity across different spatial scales.

Conclusion

In this thesis, we studied an extension of coalescent theory to incorporate population structure, specifically spatial structure, and how this structure affects the evolution of genetic distance between two lineages as we trace their ancestral paths backward in time. By doing so, we could examine the shared history of these lineages. We can also see that the spatial coalescent process is equivalent to a Bessel process.

In particular, the times when lineages merge, known as coalescence times, can be understood as "hitting times" of a Bessel process. This approach is useful because it allows us to work with non-integer dimensions, which is essential for the analysis and simulations presented throughout this thesis. Since the distribution of spatial coalescence times does not have finite moments, moment-based fitting methods cannot be applied. Although the distribution itself does not have a simple expression, we can represent it using its Laplace transform, given by:

$$\mathcal{G}(\theta) \;=\; \mathbb{E}[e^{-2\mu T(1-\theta)}] \;=\; \frac{a^{-\nu}K_\nu(2a\sqrt{\mu(1-\theta)})}{b^{-\nu}K_\nu(2b\sqrt{\mu(1-\theta)})}$$

Using Fourier methods, we constructed an algorithm that computes the probability density of coalescence times from the Laplace transform. In addition, we developed a formula to describe the spatial distribution of coalescence events occurring at a specific time and location. This is denoted by $p(k; x, t)$, where k represents the number of coalescent events among the lineages at position x and time t.

After testing our methods with simulations, we then applied them to real data from the recent outbreak of Highly Pathogenic Avian Influenza (HPAI) in dairy cattle. We estimated the effective spatial dimensionality using the theoretical framework. We were able to determine that the genes exposed to immunity, Hemagglutinin (HA), Neuraminidase (NA), and Matrix Protein (MP), appear to have evolved in different effective spatial dimensions. The interpretation of this observation is intriguing and may have implications for public health decisions. However, these types of extrapolation are beyond the scope of this thesis.

In this work, we have taken a practical view of "effective dimensionality," treating it analogously to EPS, which can *absorb* modeling complexities such as fluctuations in population size, variations in the sex ratio, inbreeding effects, and more. Similarly, effective dimensionality can absorb the complexities introduced by the spatial structure of a population. While these two effective parameters can both be viewed as stochastic equivalences, effective dimensionality can do more than just absorb complexities; it can also introduce new properties of the model itself. We list some of these.

Non-integer dimensionality automatically introduces *memory* (both temporal and spatial). A process that has the Markov memoryless property in integer dimension becomes a non-Markovian process with memory in a non-integer dimension. This means that non-integer dimensionality could be more applicable to migrating populations and inferring past migration patterns. This point requires some mention of a potential area of future work that could build on ideas and results presented in this thesis; namely, exploring mappings between our inferred non-integer effective dimensionality and non-integer order in fractional calculus.

A useful way to understand non-integer dimensions is through *scaling.* In one dimension, the only measurable quantity is length. Suppose that an object has length r; if the length is doubled, it becomes $2r$. This corresponds to a scaling factor of 2 in the one-dimensional Lebesgue measure.

In two dimensions, consider a square with side length r, which has area r^2. Doubling the side length results in an area of $(2r)^2 = 4r^2$, leading to a scaling factor of $2^2 = 4$. Similarly, in three dimensions, a cube with side length r has volume r^3, and doubling the side we get a volume of $(2r)^3 = 8r^3$, corresponding to a scaling factor of $2^3 = 8$.

This naturally raises the question: must the exponents always be integers? What happens if they are not? For example, if the Lebesgue measure is given by $r^{1.8}$, it can be interpreted as a measure that lies between length and area, suggesting a shape that cannot easily be visualized or drawn on paper.

While visualizing a measure in non-integer dimensions may be difficult, the utility of doing so is easy to understand. One can imagine a square pasture of side length r containing 50 cows. If another square pasture has side length $2r$ and contains 350 cows, then although the area increased by a factor of $4 = 2^2$, the number of cows increased by a factor of 7, which is approximately $2^{2.8}$. This suggests a scaling dimensionality of 2.8.

Another possible interpretation of non-integer effective dimensionality is that it indicates *self-similarity across scales*, as happens with fractal dimensions. A recent study [79] presents a coherent and well-supported argument that self-similarity in phylogenies can occur as a consequence of the inevitable interaction between ecological and evolutionary dynamics. Although their analysis does not explicitly incorporate spatial structure, it is possible that spatial structure also contributes to self-similarity across scales in phylogenetic trees.

In particular, considering spatial branching processes (such as branching BM) in forward time, it is reasonable to expect that certain regions of space may become "hot spots" for communities of organisms. It is well established, for instance, that the size distribution of human communities decays according to a power law. How such spatial patterns may extend to organisms or pathogens such as influenza virus, as analyzed in this thesis, remains an open question.

It is well established that the organization of certain phenomena exhibits self-similarity when they are close to a phase transition, or criticality. Under near-critical conditions, power-law behavior often emerges, for instance, in physics, where the quantity of interest (the "order parameter") decreases according to a power law with respect to a measurable parameter as the system approaches a critical point. In the context of this thesis, inferring a non-integer effective dimensionality may suggest that the population under study is near criticality. Many researchers now believe that nature, especially biological systems, tends to operate near criticality, including evolution itself, as proposed in [73].

The proximity to criticality can have implications for conservation biology. Some people say that proximity to criticality can indicate that a population is vulnerable to collapse, while other people say that proximity to criticality is a sign that a population is healthy and not in danger. Being close to criticality can be important for fields such as ecology and conservation biology.

There remains considerable work to be done. This thesis, together with the peer-reviewed articles that have emerged from it, represents a first step in what we believe is a promising direction. This thesis offers a promising starting point for interpreting effective dimensionality in evolutionary data, with several questions left open for future exploration. It is well known that non-integer dimensions carry deep and meaningful interpretations across a wide range of scientific disciplines, including fractal theory, quantum mechanics, and computer science.

Appendix A

Coding programs

The following codes we are going to provide are in pseudo-code.

A.0.1 Code 1: Calculating probability distribution p_k

```

a = 4.0         # Value for a
b = 1.0         # Value for b
nu = -0.5       # Order of Bessel function
mu = 0.7        # Value for mu
N = 200

# Initialize p_k array for storing complex values
p_k = array of size N with zeros

# Loop over k values from 0 to N-1
for k from 0 to N-1:
    sum_G = 0  # Initialize the sum for each k

    # Loop over m values from 0 to N-1
    for m from 0 to N-1:
        # Calculate theta as e^(i * 2 * pi * m / N)
        theta = exp(1j * 2 * pi * m / N)

        if m == 0:
            if nu > 0:
                G_theta = (b/a)^(2*nu)
            else:
                G_theta = 1
        else:
            # Calculate square root term
            sqrt_term = 2 * sqrt(mu * (1 - theta))

            # Calculate K_nu for both a and b
            K_a = BesselK(nu, a * sqrt_term)
            K_b = BesselK(nu, b * sqrt_term)

            if K_a and K_b are valid and K_b != 0:
                G_theta = (a^(-nu) * K_a) / (b^(-nu) * K_b)
            else:
                print warning: Invalid value for m and k

        sum_G = sum_G + G_theta * exp(-1j * 2 * pi * k * m / N)


    p_k[k] = sum_G / N

plot(real(p_k), label="Real Part")
plot(imag(p_k), label="Imaginary Part")


add title and labels to plot
show plot
```

A.0.2 Code 2: Synonymous mutations differences

```

# Define the function to read sequences from a FASTA file
def read_fasta(file_path):
    # Initialize an empty dictionary to store sequences
    sequences = {}

    # Read the FASTA file and store sequences in the dictionary
    for record in SeqIO.parse(file_path, "fasta"):
        sequences[record.id] = str(record.seq)

    return sequences

# Define the function to provide codon to amino acid mapping
def codontoaminoacid():
    # Return the codon-to-amino-acid dictionary
    return {
        ’ATA’: ’I’, ’ATC’: ’I’, ’ATT’: ’I’, ’ATG’: ’M’, ...
        # (Full codon to amino acid mapping here)
    }

# Define the function to find synonymous mutations
def find_synonymous_mutations(original_dna, mutated_dna):
    # Divide original and mutated DNA sequences into codons
    original_codons = [original_dna[i:i+3] for i in range(0, len(original_dna) - 2, 3)]
    mutated_codons = [mutated_dna[i:i+3] for i in range(0, len(mutated_dna) - 2, 3)]

    # Get the codon-to-amino-acid mapping
    codon_to_amino_acid = codontoaminoacid()  # Call the renamed function

    # Function to convert codons to their respective proteins (amino acids)
    def codon_to_protein(codons):
        return [codon_to_amino_acid.get(codon, ’?’) for codon in codons]

    # Convert both sequences to proteins
    original_protein = codon_to_protein(original_codons)
    mutated_protein = codon_to_protein(mutated_codons)

    # Initialize an empty list to store synonymous mutations
    synonymous_mutations = []

    # Loop over the codons and compare them
    for i, (orig_codon, mut_codon) in enumerate(zip(original_codons, mutated_codons)):
        # Check if the codons are different but correspond to the same amino acid
        if orig_codon != mut_codon and codon_to_amino_acid.get(orig_codon) == codon_to_amino_acid.get(mut_codon):
            synonymous_mutations.append((i + 1, orig_codon, mut_codon))

    # Return the count of synonymous mutations
    return len(synonymous_mutations)

# Define the main function to process the FASTA file
def main(fasta_file):
    # Read sequences from the FASTA file
    sequences = read_fasta(fasta_file)

    # Check if there are at least two sequences in the file
    if len(sequences) < 2:
        print("The FASTA file must contain at least two sequences.")
        return

    # Initialize number of repeats for simulations
    num_repeats = 10000
    results = []

    # Repeat the process of comparing sequences
    for _ in range(num_repeats):
        # Randomly select two sequences from the file
        ids = list(sequences.keys())
        original_id, mutated_id = random.sample(ids, 2)
        original_dna = sequences[original_id]
        mutated_dna = sequences[mutated_id]

        # Find the number of synonymous mutations
        num_synonymous_mutations = find_synonymous_mutations(original_dna, mutated_dna)
        results.append(num_synonymous_mutations)

    # (Optional) Print results to the console for each repeat
    # for i, result in enumerate(results, 1):
    #     print(f"Run {i}: Number of synonymous mutations: {result}")

    # Plot a histogram of the results
    plt.hist(results, bins=range(min(results), max(results) + 2), edgecolor=’black’)
    plt.xlabel(’Number of Synonymous Mutations’)
    plt.ylabel(’Frequency’)
    plt.title(’Histogram of Synonymous Mutations’)
    plt.show()

    # Write the results to a text file
    with open(’synonymous_mutations_results3.txt’, ’w’) as f:
        f.write(’, ’.join(map(str, results)))

# Entry point for the program
if __name__ == "__main__":
    fasta_file = ’MP.aln.fasta’  # Specify the FASTA file
    main(fasta_file)
```

A.0.3 Code 3: Code for Hamming distances

```
# Function to read sequences from a FASTA file
Function read_fasta(file_path):
    Initialize sequences as an empty dictionary
    For each record in FASTA file:
        sequences[record.id] = record.seq
    Return sequences

# Function to calculate Hamming distance
Function Hamming_distance(seq1, seq2):
    If length of seq1 != length of seq2:
        Raise ValueError("Sequences must be of the same length")
    Initialize Hamming_distance = 0
    For each base pair (a, b) in zip(seq1, seq2):
        If a != b:
            Increment Hamming_distance by 1
    Return Hamming_distance

# Main function
Function main(fasta_file):
    sequences = read_fasta(fasta_file)

    If length of sequences < 2:
        Print "The FASTA file must contain at least two sequences."
        Return

    Initialize num_repeats = 10000
    Initialize results as an empty list

    For i from 1 to num_repeats:
        Randomly select two sequence IDs (original_id, mutated_id) from sequences
        original_dna = sequences[original_id]
        mutated_dna = sequences[mutated_id]

        Try:
            distance = Hamming_distance(original_dna, mutated_dna)
            Append distance to results
        Catch ValueError:
            Print "Skipping pair due to length mismatch."

    Plot histogram of results
    Save results to 'Hamming_distances_results.txt'

# Main entry point
If __name__ == "__main__":
    Call main with 'MP.aln.fasta' file path
```

A.0.4 Code 4: Calculating the maximum log-likelihood

```
# Import necessary libraries
Import numpy as np
Import scipy.special as ss
Import matplotlib.pyplot as plt

# Define parameters
```

```
Set a = 1.6    # Parameter a
Set b = 1.0    # Parameter b
Set nu = -0.5  # Bessel function order
Set mu = 0.8   # Parameter mu
Set N = 100    # Number of values for m and k

# Initialize an array to store p_k values
Create p_k array with N complex zeros

# Calculate p_k values
For k from 0 to N-1:  # Loop over k
    Set sum_G = 0.0  # Initialize sum_G

    For m from 0 to N-1:  # Loop over m
        Set theta = exp(1j * 2 * pi * m / N)  # Calculate the complex angle

        # Check for m == 0
        If m == 0:
            If nu > 0:
                Set G_theta = (b / a)^(2 * nu)  # Special case for m == 0
            Else:
                Set G_theta = 1
        Else:
            Set sqrt_term = 2 * sqrt(mu * (1 - theta))  # Compute square root term
            Set K_a = besselk(nu, a * sqrt_term)  # Bessel function of the first kind
            Set K_b = besselk(nu, b * sqrt_term)  # Bessel function of the first kind

            If K_a is finite and K_b is finite and K_b != 0:
                Set G_theta = (a^(-nu) * K_a) / (b^(-nu) * K_b)  # Calculate G_theta
            Else:
                Set G_theta = 0  # If invalid, set G_theta to 0

        Add G_theta * exp(-1j * 2 * pi * k * m / N) to sum_G

    # Store the result of sum_G divided by N in p_k[k]
    Set p_k[k] = sum_G / N

# Normalize p_k values
Normalize p_k by dividing by sum(p_k)

# Calculate logarithms of normalized p_k values
For each value in p_k:
    Set log_p_k = log(abs(value))

# Handle NaN and infinite values in log_p_k
For each value in log_p_k:
    If value is NaN or infinite:
        Set value to 0

# Sum all the logarithms of probabilities
Set log_product_sum = sum(log_p_k)

# Find the maximum of the log_p_k values
Set max_log_prob = max(log_p_k)

# Plot the logarithms of p_k
Create a plot with x-axis labeled ’k’ and y-axis labeled ’log(p_k) (Normalized)’
Plot log_p_k values against k
Add a horizontal line at y=0 for reference
Show gridlines on the plot

# Display the plot
Show plot
```

Bibliography

[1] Abramowitz, M. and Stegun, I. A., *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables* (Applied Mathematics Series). Dover Publications, 1965.

[2] André, D., "De l'égalité des lois des marches aléatoires," *Journal de Mathématiques Pures et Appliquées*, vol. 10, pp. 1–17, 1885.

[3] Bachelier, L., "Théorie de la spéculation," Doctoral thesis, Ph.D. dissertation, Université de Paris, 1900.

[4] Bailey, N. T., *Elements of Stochastic Processes.* New York, NY: Wiley, 1990.

[5] Basu, A. K., *Introduction to Stochastic Process.* Alpha Science, 2003.

[6] Borodin, A. and Salminen, P., *Handbook of Brownian Motion - Facts and Formulae* (Probability and Its Applications). Birkhäuser Basel, 2015.

[7] Borodin, A. N. and Salminen, P., *Handbook of Brownian Motion: Facts and Formulae.* Birkhäuser Basel, 2002.

[8] Bouvier, N. M. and Palese, P., "The biology of influenza viruses," *Vaccine*, vol. 26, no. 29-30, pp. 3657–3664, 2008.

[9] Bradburd, G. and Ralph, P., "Spatial population genetics: It's about time," *arXiv preprint arXiv:1904.09847*, Apr. 2019.

[10] Bradburd, G. S., Coop, G. M., and Ralph, P. L., "Inferring Continuous and Discrete Population Genetic Structure Across Space," en, *Genetics*, vol. 210, no. 1, pp. 33–52, Sep. 2018.

[11] Bradburd, G. S. and Ralph, P. L., "Spatial population genetics: It's about time," en, *Annu. Rev. Ecol. Evol. Syst.*, vol. 50, no. 1, pp. 427–449, Nov. 2019.

[12] Bradburd, G. S., Ralph, P. L., and Coop, G. M., "A Spatial Framework for Understanding Population Structure and Admixture," en, *PLoS Genet.*, vol. 12, no. 1, e1005703, Jan. 2016.

[13] Burki, T., “Avian influenza in cattle in the USA,” *The Lancet Infectious Diseases*, vol. 24, no. 7, e424–e425, Jul. 2024.

[14] Burrough, E. R., Magstadt, D. R., Petersen, B., *et al.*, “Highly Pathogenic Avian Influenza A(H5N1) clade 2.3.4.4b virus infection in domestic dairy cattle and cats, United States, 2024,” *Emerging Infectious Diseases*, vol. 30, no. 7, pp. 1335–1343, Jul. 2024.

[15] Caffarelli, L. and Silvestre, L., “An extension problem related to the fractional Laplacian,” en, *Comm. Partial Differential Equations*, vol. 32, no. 8, pp. 1245–1260, Aug. 2007.

[16] Calin, O., *An Informal Introduction to Stochastic Calculus with Applications.* World Scientific Publishing Company, 2015.

[17] Caserta, L. C., Frye, E. A., Butt, S. L., and al., et, “Spillover of Highly Pathogenic Avian Influenza H5N1 virus to dairy cattle,” *Nature*, vol. 634, pp. 669–676, 2024.

[18] Charlesworth, B., “Effective population size and patterns of molecular evolution and variation,” *Nature Reviews Genetics*, vol. 10, no. 3, pp. 195–205, 2009.

[19] Cinlar, E., *Introduction to Stochastic Processes.* Englewood Cliffs, NJ: Prentice Hall, 1975.

[20] Crow, J. F. and Kimura, M., *An Introduction to Population Genetics Theory.* New York: Harper and Row, 1970.

[21] Durrett, D., *Probability: Theory and Examples*, 5th. Cambridge: Cambridge University Press, 2019.

[22] Dyskin, E. B., “Inhomogeneous strong markov processes,” *Dokl. Akad. Nauk. SSSR*, vol. 113, pp. 261–263, 1957.

[23] Einstein, A., “Über die von der molekularkinetischen theorie der wärme geforderte bewegung von in ruhenden flüssigkeiten suspendierten teilchen,” *Annalen der Physik*, vol. 17, no. 8, pp. 549–560, 1905, English translation: ”On the movement of small particles suspended in stationary liquids required by the molecular-kinetic theory of heat”.

[24] Elez, M., Murray, A. W., Bi, L. J., Zhang, X. E., Matic, I., and Radman, M., “Seeing mutations in living cells,” *Current Biology*, vol. 20, no. 16, pp. 1432–1437, Aug. 2010, Epub 2010 Jul 30.

[25] Feller, W., *An Introduction to Probability Theory and Its Applications*, 3rd. New York: Wiley, 1968, vol. 1.

[26] Fisher, R. A., *The Genetical Theory of Natural Selection.* Clarendon Press, 1930.

[27] Gillespie, J. H., *Population Genetics: A Concise Guide*. Oxford: Oxford University Press, 2010.

[28] Griffiths, R. C. and Tavaré, S., "Ancestral inference in population genetics," *Statistical Science*, vol. 9, no. 3, pp. 307–319, 1994.

[29] Grimmett, G., *Probability and Random Processes*, 3rd. Oxford University Press, 2014.

[30] Haldane, J. B. S., "A mathematical theory of natural and artificial selection," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 26, no. 2, pp. 220–230, 1930.

[31] Hamana, Y. and Matsumoto, H., "The probability distributions of the first hitting times of Bessel processes," *Transactions of the American Mathematical Society*, vol. 365, no. 10, pp. 5237–5257, Oct. 2013.

[32] Hartl, D. L., *A Primer of Population Genomics and Genomics*. Hoboken, NJ: Wiley-Blackwell, 2020.

[33] Hartl, D. L. and Clark, A. G., *Principles of Population Genetics*. Sinauer Associates, an imprint of Oxford University Press, 1997.

[34] Hartl, D. L. and Jones, E. W., *Principles of Population Genetics*. Sunderland, MA: Sinauer Associates, 2016.

[35] P. Heitjans and J. Kärger, Eds., *Diffusion in Condensed Matter: Methods, Materials, Models*, 2, illustrated, revised. Springer Science Business Media, 2006, p. 965.

[36] Hernandez-Del-Valle, G. and Pacheco, C. G., "Hitting times for Bessel processes," *Communications on Stochastic Analysis*, vol. 9, no. 1, pp. 79–92, 2015.

[37] Hu, X., Saxena, A., Magstadt, D. R., *et al.*, "Genomic characterization of Highly Pathogenic Avian Influenza A H5N1 virus newly emerged in dairy cattle," *Emerging Microbes and Infections*, vol. 13, no. 1, p. 2 380 421, Dec. 2024.

[38] Hu, X., Saxena, A., Magstadt, D. R., *et al.*, "Highly Pathogenic Avian Influenza A (H5N1) clade 2.3.4.4b virus detected in dairy cattle," *bioRxiv*, 2024, Preprint.

[39] Hudson, R. R., "Gene genealogies and the coalescent process," *Oxford Surveys in Evolutionary Biology*, vol. 7, pp. 1–44, 1990.

[40] Hunt, G. A., "Some theorems concerning Brownian motion," *Transactions of the American Mathematical Society*, vol. 81, pp. 294–319, 1956.

[41] Jung, H. E. and Lee, H. K., *Image from "virus-like particle-based vaccine development: A comprehensive overview"*, CC BY 4.0, Image available at: `https://commons.wikimedia.org/w/index.php?curid=92944856`, 2020.

[42] Karatzas, I. and Shreve, S. E., *Brownian Motion and Stochastic Calculus*. Springer, 1991.

[43] Kent, J. T., "Some probabilistic properties of Bessel functions," *The Annals of Probability*, vol. 6, no. 5, pp. 762–773, 1978.

[44] Kingman, J. F., "The coalescent," *Stochastic Processes and their Applications*, vol. 13, no. 3, pp. 235–248, 1982.

[45] Krammer, F., "Emerging influenza viruses and the prospect of a universal influenza virus vaccine," *Biotechnology Journal*, vol. 10, no. 5, pp. 690–701, 2015.

[46] Krapivsky, P. and Redner, S., "Probing non-integer dimensions," *Journal of Physics: Condensed Matter*, vol. 19, no. 6, p. 065 119, 2007.

[47] Kristensen, C., Jensen, H. E., Trebbien, R., Webby, R. J., and Larsen, L. E., "The avian and human influenza a virus receptors sialic acid (sa)-2,3 and sa-2,6 are widely expressed in the bovine mammary gland," *bioRxiv*, May 2024, Preprint.

[48] Lab, A., *Avian influenza alignments*, Accessed: 2024-11-27, 2024.

[49] Lawler, G., *Introduction to Stochastic Processes*. Boca Raton, FL: CRC Press, 2016.

[50] Lévy, P., *Processus Stochastiques et Mouvement Brownien*. Gauthier-Villars, 1973.

[51] Moran, P. A., "Random processes in genetics," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 54, no. 1, pp. 60–71, 1958.

[52] Mörters, P. and Peres, Y., *Brownian Motion*. Cambridge University Press, 2010.

[53] National Human Genome Research Institute, *Genetics glossary*, n.d.

[54] Neumann, G. and Kawaoka, Y., "Host range restriction and pathogenicity in the context of influenza pandemic," *Emerging Infectious Diseases*, vol. 12, no. 6, pp. 881–886, 2006.

[55] Nguyen, T.-Q., Hutter, C., Markin, A., *et al.*, "Emergence and interstate spread of Highly Pathogenic Avian Influenza A(H5N1) in dairy cattle," *bioRxiv*, 2024, Preprint available at: `https://doi.org/10.1101/2024.05.01.591751`.

[56] Nunney, L., "The effective population size and the rate of molecular evolution," *Journal of Evolutionary Biology*, vol. 12, no. 6, pp. 1019–1026, 1999.

[57] Oguzie, J. U., Marushchak, L. V., Shittu, I., *et al.*, "Avian influenza a(h5n1) virus among dairy cattle, texas, usa," *Emerging Infectious Diseases*, vol. 30, no. 7, pp. 1425–1429, Jul. 2024.

[58] Organization, W. H., *Avian influenza a(h5n1) – United States of America*, Accessed: 2024-11-26, Apr. 2024.

[59] Poincaré, H., *Les méthodes nouvelles de la mécanique céleste*. Paris: Gauthier-Villars, 1890.

[60] ReadIAB, *Pairwise alignment*, `https://readiab.org/pairwise-alignment.html`, Accessed: 2024-11-27, 2024.

[61] Rensick, S., *Adventures in Stochastic Processes*. Mineola, NY: Dover Publications, 2011.

[62] Rincón, L., *Procesos Estocásticos*. Madrid: Ediciones Universitarias, 2013.

[63] Ross, S. M., *Stochastic Processes*. New York, NY: John Wiley Sons, 1996.

[64] Ross, S., *A First Course in Probability*, 9th. Boston: Pearson, 2010.

[65] Schafers, J., Warren, C. J., Yang, J., *et al.*, "Pasteurisation temperatures effectively inactivate influenza a viruses in milk," *Nature Communications*, vol. 16, no. 1, p. 1173, Jan. 2025.

[66] Schilling, R. L. and Partzsch, L., *Brownian motion*. De Gruyter, Berlin, 2012, pp. xiv+380, An introduction to stochastic processes, With a chapter on simulation by Björn Böttcher.

[67] Schilling, R. L. and Walet, J. C., *Brownian Motion: An Introduction to the Mathematical Theory*. Birkhäuser Basel, 2012.

[68] Singh, G., Trujillo, J. D., McDowell, C. D., *et al.*, "Detection and characterization of H5N1 HPAIV in environmental samples from a dairy farm," *Virus Genes*, vol. 60, no. 5, pp. 517–527, Oct. 2024.

[69] Smith, P., *Stochastic Processes: An Introduction*. New York, NY: Springer, 2013.

[70] Spackman, E., Jones, D. R., McCoig, A. M., Colonius, T. J., Goraichuk, I. V., and Suarez, D. L., "Characterization of highly pathogenic avian influenza virus in retail dairy products in the US," *Journal of Virology*, vol. 98, no. 7, e00881–24, Jul. 2024.

[71] Steel, J. and Lowen, A. C., "Influenza a virus reassortment," *Current Topics in Microbiology and Immunology*, vol. 382, pp. 25–46, 2014.

[72] Tisza, M. J., Hanson, B. M., Clark, J. R., *et al.*, "Virome sequencing identifies H5N1 Avian Influenza in wastewater from nine cities," *medRxiv*, 2024, Preprint.

[73] Torres, J. L., "Biological power laws and Darwin's principle," en, *J. Theor. Biol.*, vol. 209, no. 2, pp. 223–232, Mar. 2001.

[74] Wakeley, J., *Coalescent theory: An introduction.* Roberts Company Publishers, 2008.

[75] Wakeley, J., Fan, W. L., Koch, E., and Sunyaev, S., "Recurrent mutation in the ancestry of a rare variant," *Genetics*, vol. 224, no. 3, iyad049, Jul. 2023, Erratum in: *Genetics*. 2023 Jul;224(3):iyad082. doi:10.1093/genetics/iyad082.

[76] Wang, J., "Estimation of effective population sizes from data on genetic markers," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 360, no. 1459, pp. 1395–1409, 2005.

[77] Webster, R. G. and Govorkova, E. A., "Continuing challenges in influenza," *Annals of the New York Academy of Sciences*, vol. 1323, no. 1, pp. 115–139, 2014.

[78] Wright, S., "Evolution in Mendelian Populations," *Genetics*, vol. 16, no. 2, pp. 97–159, 1931.

[79] Xue, C., Liu, Z., and Goldenfeld, N., "Scale-invariant topology and bursty branching of evolutionary trees emerge from niche construction," en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 117, no. 14, pp. 7879–7887, Apr. 2020.

[80] Yang, J., Qureshi, M., Kolli, R., *et al.*, "The Haemagglutinin gene of bovine-origin H5N1 influenza viruses currently retains receptor-binding and ph-fusion characteristics of avian host phenotype," *Emerging Microbes and Infections*, vol. 14, no. 1, p. 2 451 052, 2025, Epub 2025 Jan 27.