Topological Data Analysis: Persistent Homology of Uniformly Distributed Points

by

Ranjit Sohal

Bachelor of Science (Honours), University of Southampton, UK, 2019

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Mathematics and Statistics

We acknowledge and respect the lək̓ʷəŋən peoples on whose traditional territory the university stands and the Songhees, Esquimalt and W̱SÁNEĆ peoples whose historical relationships with the land continue to this day.

Topological Data Analysis: Persistent Homology of Uniformly Distributed Points

by

Ranjit Sohal

Bachelor of Science (Honours), University of Southampton, UK, 2019

Supervisory Committee

Dr. Ryan Budney, Supervisor
(Department of Mathematics and Statistics)

Dr. Farouk Nathoo, Committee Member
(Department of Mathematics and Statistics)

Dr. Alan Mehlenbacher, Outside Member
(Department of Economics)

ABSTRACT

Topological Data Analysis (TDA) is a branch of computational topology that provides methods to extract qualitative information from high-dimensional, noisy, and incomplete data. TDA combines techniques from various fields, such as algebraic topology, computational geometry, algorithms, statistics, and graph theory. Persistent Homology (PH), based on homology theory from algebraic topology, is the principal tool used in TDA; PH tracks the evolution of topological features of the data across multiple scales through persistent homological bars, which represent the creation (birth) and disappearance (death) of these features. These bars are graphically depicted through persistence diagrams and persistence barcodes. The challenge in using PH for the analysis of noisy real-world data is to separate the bars generated by noise from the bars that provide meaningful topological information of the underlying geometric object from which the data is sampled; this problem remains unresolved despite various proposed techniques. A limited number of papers analyzed the PH of noise by considering points in $\mathbb{R}^d$ generated using probability distributions. This thesis introduces persistent homology concentrating on the computational side, and it examines the birth and death times of persistent homology bars generated by Vietoris-Rips complexes of uniformly distributed points in three spaces: a unit interval, a unit square, and a unit cube. Through numerical simulations, it is identified that the birth and death times of the persistent homology bars adhere to distinct statistical distributions, whose precise nature varies according to the space from which the points are sampled and the homological dimension of the persistent homology bars; the research examines the behaviour of their parameters as the number of points increases, providing insights into the persistent homology of noise and laying the groundwork for further research.

Table of Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I want to express my heartfelt thanks to my supervisor Ryan Budney for taking a chance on an ambitious graduate student from Italy/England and supporting me, even as we navigated the challenges of the Covid pandemic. Ryan's belief in me, his patience and mentorship, and his encouragement throughout my time as a graduate student in Victoria have been invaluable.

I am also grateful to the Capital Planning Division in the Department of Finance at Metro Vancouver for allowing me to work as a co-op student, hiring me and becoming a stepping stone in my professional career.

I am thankful to my family in Canada for providing me with a supportive and welcoming environment during the Covid pandemic. Their presence helped me socialize and make lasting memories I will always cherish.

Lastly, I would like to thank my fiancé for coming into my life and for being a source of constant support, care, and motivation throughout this journey. I am truly grateful for her unwavering belief in me.

DEDICATION

I would like to dedicate this thesis to Salvatore Iadicicco.

Chapter 1

Introduction

The Global Datasphere[1] is increasing in size at an exponential rate. The growth of data is due to a large number of factors that need to be taken into consideration, such as human development, population growth, and technological advancements. More precisely, according to the International Data Corporation (IDC), the Datasphere is estimated to be 80 ZB (equivalent to $80 \cdot 10^{21}$ bytes) in the year 2022, and it is forecasted to reach 175 ZB ($80 \cdot 10^{21}$ bytes) by 2025 [30].

The data growth has led to various difficulties in comprehending the collected data. This is primarily due to the increased computing power required for analysis, storage demands, and the need to optimize existing algorithms and software. Additionally, the complexity of the data, including its high dimensionality, incompleteness, and the presence of noise, constitutes an additional issue [29]. The use of mathematical and statistical methodologies in data analytics has been proven to impact addressing the aforementioned challenges considerably. One of the branches of mathematics that have gained significant interest in comprehending the underlying structure of data is topology, specifically topological data analysis (TDA). Nevertheless, TDA is frequently perceived as a field on its own as it incorporates techniques from various domains such as algebraic topology, computational geometry, computer science, and statistics.

1.1 Topological Data Analysis

From a mathematical point of view, algebraic topology is interested in studying the properties of topological spaces using algebraic structures; in particular, homology describes a topological space in terms of homology groups which are properties that are invariant under

[1] The Global Datasphere is a measure of all new data that is captured, created, and replicated in any given year across the globe.

homeomorphism. Informally, the homology groups $H_k(X)$ of a topological space X describe the number of k-dimensional holes (i.e. connected components, cycles, and voids) in the space X.

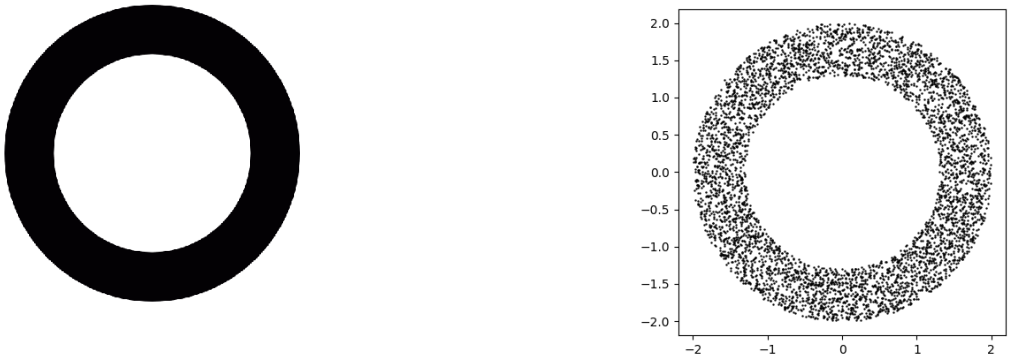
The primary tool used in TDA from algebraic topology is persistent homology; the interest in this field by researchers in non-mathematical fields is due to its large number of applications [16], the ability to understand the general idea without the necessity of solid background in mathematics and the availability of software to compute the persistent homology of real-world data.

Mathematical formalism will be introduced in Chapter 1, so we start with a simple example that can provide an intuitive understanding of the scope of this thesis.

1.2 Motivation

Suppose user A transmits Figure 1.1a to user B, who is interested in studying the topological properties of the annulus in order to detect the presence of *holes*, but a certain amount of data is lost during the process, and user B receives Figure 1.1a in the form of point cloud data (Figure 1.1b). User B can utilize persistent homology for retaining the original figure's topological information through the point cloud data.

More generally, the homology of a manifold $\mathcal{M}$ embedded in a Euclidean Space could



(a) (b)

Figure 1.1: An annulus (a) and IID random sample from an annulus (b).

be recovered from a set of independent and identically distributed random sample $\mathcal{X} = \{X_1, ..., X_n\}$ from $\mathcal{M}$. The general approach involves considering the homology of

$$U = \bigcup_{k=1}^{n} B_\epsilon (X_k) \tag{1.1}$$

where $B_\epsilon(X_k)$ is the Euclidean ball of radius ϵ centered at X_k.

The main problem is the determination of the ϵ for which U and $\mathcal{M}$ could have the same homology groups. However, the existence of such ϵ depends on n, which must be large enough.

Persistent homology solves this problem by considering a range of values for ϵ. The general idea behind persistent homology is to interpret ϵ as a time variable because the homology of the union of balls changes as ϵ increases (for instance, the number of connected components decreases over time): homological features will appear and disappear varying ϵ and their relation to the original manifold is determined by their 'lifetimes'. These lifetimes are encoded using *persistent homology bars* (or *bars*); more precisely, each bar is an interval $(\epsilon_i, \epsilon_j) \in \mathbb{R}^2$ where the coordinates represent the *birth time* and *death time* time of a homological feature. The persistent homology bars are graphically depicted through *Persistence Diagrams* (Figure 1.3) and *Persistence Barcodes*.

The assumption is that among all the homological features that are recorded over time, the ones that *persist* for a relatively long time have a higher probability of representing the homology of $\mathcal{M}$, and the remaining ones are interpreted as *topological noise*. Another issue arises when $\mathcal{X}$ contains additional *noise* or points that are not sampled from the manifold (Figure 1.2c and 1.2d).

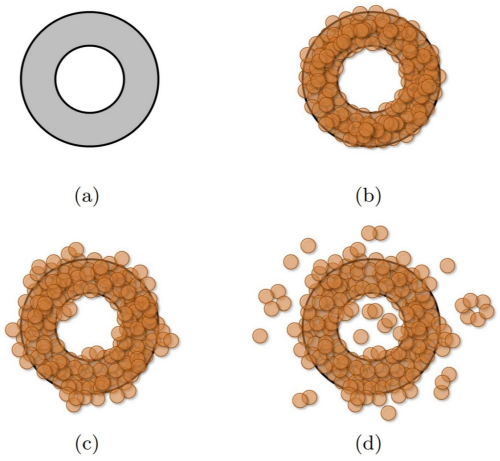


Figure 1.2: (a): an annulus; (b): a point cloud sampled from an annulus; (c), (d): point clouds sampled from an annulus with additional noise.

Adler, Bobroski and Weinberger have studied this phenomenon in [2], where it has been shown that the overall behaviour of persistent homology bars depends on the type of statistical distribution from which the additional noise is sampled. More precisely, [1, 2] investigate

the homology of simplicial complexes built from a random set of vertices sampled from Gaussian, exponential, and power-law distributions in $\mathbb{R}^d$.

In this master's thesis, we examine the persistent homology properties of uniform noise, precisely, points produced by a uniform distribution within a unit interval, unit square, and unit cube. We aim to estimate the likelihood of a persistent homology bar being born or dead at a specific time. Precisely, we explore the distribution of persistence diagrams for uniformly distributed points in subsets of a Euclidean space $\mathbb{R}^d$ and investigate the relationships between the persistence diagrams of different numbers of points in each homological dimension. Our experiments suggest that the birth and death times of the persistent homology bars of these points follow probability distributions for all $k < d \leq 3$, where k is the homological dimension, and d is the ambient dimension of the Euclidean space. It is worth noting that our focus is not on identifying the exact statistical distribution of the birth and death times but on understanding the behaviour of the parameters of the probability distributions that fit well with a specific dataset as the number of points in the dataset increases.

In this study, the choice of uniform distribution provides a foundation for understanding the distribution of the persistent homology bars of noise. In its fundamental nature, the uniform distribution is often a local approximation for more complex continuous parametric distributions. By focusing on the effect of uniformly distributed noise on persistent homology bars, this work lays the groundwork for more comprehensive research. Indeed, the ultimate goal is to develop an integral transform that describes the distribution of noise bars in the Vietoris-Rips complex from points selected from continuous parametric distributions. Thus, focusing on uniform distribution, this work is the building block for a broader research.

1.3 Thesis Overview

Chapter 2 provides an introduction to persistent homology and a brief overview of the computational approach.

Chapter 3 outlines the scope of the research, its methodology, data collection procedures, and a description of the statistical techniques employed to determine the distributions that best describe the data.

Chapter 4 presents the results of the analysis of the simulations. We present persistence diagrams, histograms, tables containing distributions tested and the quality of their fits, estimated parameters of the selected distributions, and regression models.

Chapter 5 provides a summary of the thesis and presents the main results and future work.

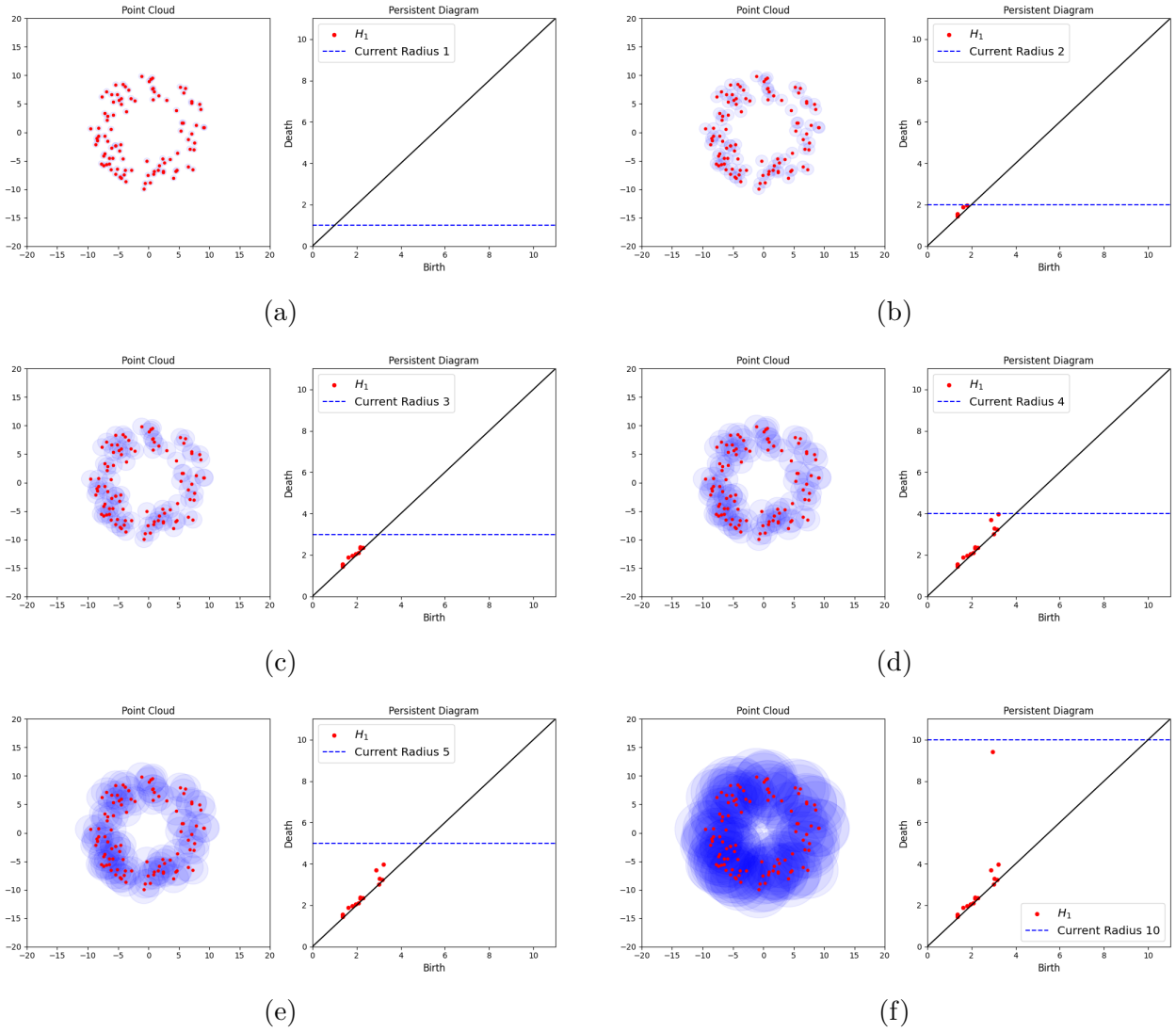


Figure 1.3: Persistence Diagrams in homological dimension 1 of a point cloud sampled from an annulus at various stages as the radii of the balls around each point increase. A union of balls with a 'very small' radius has the same homology as n distinct points (a); a union of balls with a 'very large' radius has the same homology as one point (f). The union of balls in (c), (d), and (e) have the same homology as an annulus; indeed, the presence of the one-dimensional hole (born at $radius \approx 3$, and dead at $radius \approx 9$) is represented as a point in the persistence diagram in (f) that is far away from the diagonal with respect to the other points, more precisely, this is a topological feature that persists for a relatively longer time than all the other features which are assumed not to provide any relevant information about the annulus.

Chapter 2

An Introduction to Persistent Homology

This chapter provides an overview of persistent homology, using consistent terminology from Edelsbrunner and Harer [12], Hatcher [17], and Carlsson [40]. The mathematical concepts presented in this chapter serve as the cornerstone of our research and are essential for comprehending the outcomes. However, to develop the theory of persistent homology, we must first revisit certain concepts from algebraic topology.

2.1 Algebraic Topology

Let $u_0, u_1, ..., u_k \in \mathbb{R}^{k+1}$ be affinely independent, or equivalently, let $u_1{-}u_0, ..., u_k{-}u_0$ be linearly independent. Then a k-simplex is a set of the form

$$\sigma = \Big\{ \lambda_0 u_0 + ... + \lambda_k u_k \;\Big|\; \sum\nolimits_{i=0}^{k} \lambda_i = 1 \text{ and } \lambda_i \geq 0 \text{ for } i = 0, 1, ..., k \Big\}, \tag{2.1}$$

equivalently, a k-simplex is the convex hull of $k+1$ affinely independent points, $\sigma = conv(u_{i_0}, ..., u_{i_k})$. The boundary of each simplex is composed of lower dimensional simplices: vertices form the boundary of an edge, edges and vertices form the boundary of a triangle, and triangles, edges, and vertices form the boundary of a tetrahedron (Figure 2.1). We refer to these lower dimensional simplices as *faces* of the higher dimensional simplex. Simplices are building blocks for more complicated algebraic structures (*simplicial complexes*).

Definition 2.1. A *(geometric) simplicial complex* is the union of a finite set K of simplices with the property that for any two simplices, their intersection $\sigma \cap \tau$ is either empty or a common face of both (a simplex of K).

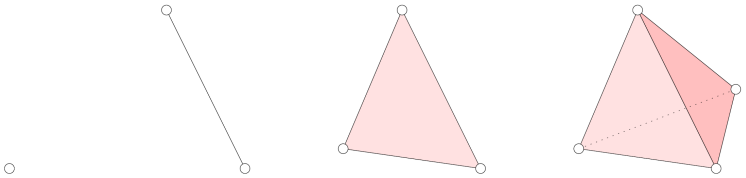


Figure 2.1: A 0-simplex or *vertex*, a 1-simplex or *edge*, a 2-simplex or *triangle*, and a 3-simplex or *tetrahedron* (from left to right) [17].

By constructing a simplicial complex from a set of points in a Euclidean space, we can conveniently establish a connection between a geometric object and these points. Nevertheless, it should be noted that simplicial complexes can also be assigned to any arbitrary collection of sets.

Definition 2.2. An *abstract simplicial complex* is a set K of nonempty finite subsets of some finite set V, with the property that for every $\sigma \in K$, and every nonempty $\tau \subseteq \sigma$, we also have $\tau \in K$.

We can construct an abstract simplicial complex from a simplicial complex. More formally, let K be a simplicial complex, then an abstract simplicial complex K^* can be obtained by considering K's set of vertices. In this scenario, K^* is defined as a *geometric realization* of K and K is a *vertex scheme* of K^*. More precisely,

$$K^* = \{\sigma \mid \sigma \text{ is the set of all vertices of some simplex in } K\}.$$

Given an abstract simplicial complex K^* on the vertex set V with $|V| = n$, we relabel V with affinely independent points $u_0, ..., u_k \in \mathbb{R}^{k+1}$ and let

$$K = \bigcup_{[u_{i_0},...,u_{i_k}] \in K^*} conv(u_{i_0}, ..., u_{i_k})$$

where $conv(u_0, ..., u_k)$ denotes the convex hull of $u_0, ..., u_k$.
Therefore, the word simplicial complex will be used to refer interchangeably to either an abstract simplicial complex or to its geometric realization.

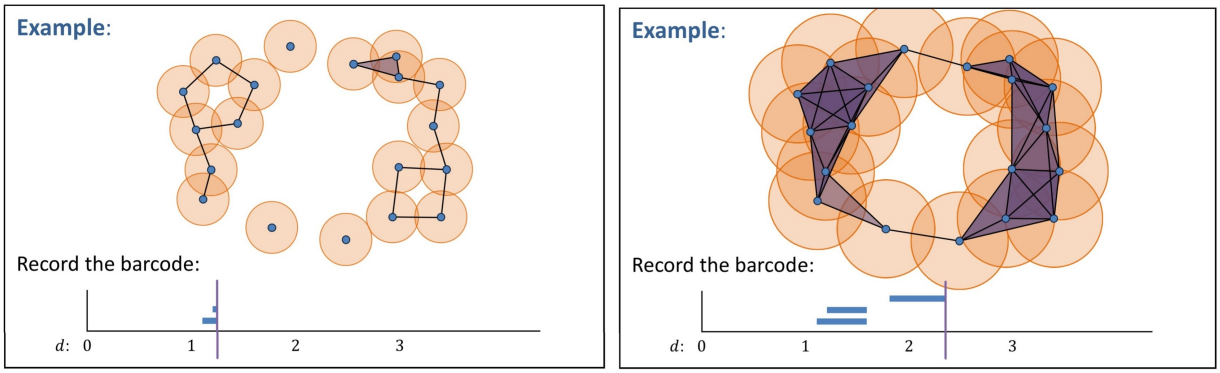
2.2 Persistent Homology Theory

Data is often represented as an unordered sequence of points in a Euclidean space $\mathbb{E}^n$. One way to convert a set of points x_α in a metric space into a global object is to use the point cloud as the vertices of a combinatorial graph whose edges are determined by proximity. The

graph is then completed to a simplicial complex.
The choice of how to fill in the higher dimensional simplices of the proximity graph allows for different global representations. According to the method used to build these objects, the final simplicial complex might be referred to as *Čech*, *Alpha*, *Delaunay* or *Vietoris-Rips Complex* [15]. The study of these complexes is useful for developing the theory of persistent homology, and a detailed description of these complexes can be found in the works of Edelsbrunner-Harer [12]. However, in computational topology, the most common complex used to derive topological features of the data is the Vietoris-Rips complex, which is closely related to the Čech complex.

Definition 2.3. Vietoris-Rips Complex. Given a collection of points $\{x_\alpha\}$ in Euclidean space $\mathbb{E}^n$, the *Rips complex* (or *Vietoris-Rips complex*) $\mathcal{R}_\epsilon$, is the abstract simplicial complex whose k-simplices correspond to unordered $(k+1)$-tuples of points $\{x_\alpha\}_0^k$ which are pairwise within distance ϵ (Figure 2.2).

Suppose we have a metric space $\mathbb{X}$ with metric d, then the Vietoris-Rips complex for $\mathbb{X}$, attached to the parameter ϵ, will be the simplicial complex whose vertex set is $\mathbb{X}$, and where $\{x_0, x_1, ..., x_k\}$ spans a k-simplex if and only if $d(x_i, x_j) \leq \epsilon$ for all $0 \leq i, j \leq k$.



(a) (b)

Figure 2.2: An illustration of simplicial complexes (Vietoris-Rips) built from point-cloud data for two different scale parameters where the barcode is drawn as the scale parameter increases [39].

2.2.1 Persistent Homology

In informal terms, persistent homology describes the changes in homology that occur to an object which evolves with respect to a parameter. More precisely, suppose we have a collection of points in a Euclidean space, and we consider all the possible associated Vietoris-Rips complexes for all distances ϵ, then all these complexes are subcomplexes of a Vietoris

complex for a sufficiently large distance ϵ. We can then generate a sequence of subcomplexes and examine its topological properties.

Definition 2.4. Filtred Simplicial Complex. A *filtred simplicial complex* K constructed from a set X is a family of subcomplexes $(K_i | i \in \mathbb{N})$ of a simplicial complex L with vertex set V such that $K_i \subseteq K_j$ for all $i \leq j$.

Example 2.5. Let $(\mathbb{X}, d_{\mathbb{X}})$ be a metric space. The Vietoris-Rips filtration is the filtered simplicial complex $\mathcal{R}_\epsilon$ defined by: for all $\epsilon \in \mathbb{R}$, we filter the complete simplicial complex on the set of vertices $\mathbb{X}$ and the filtration at parameter ϵ is the simplicial complex $\mathcal{R}_\epsilon$, where

$$[x_0, x_1, \ldots, x_k] \in \mathcal{R}_\epsilon \Leftrightarrow d_{\mathbb{X}}(x_i, x_j) \leq \epsilon \text{ for all } i, j.$$

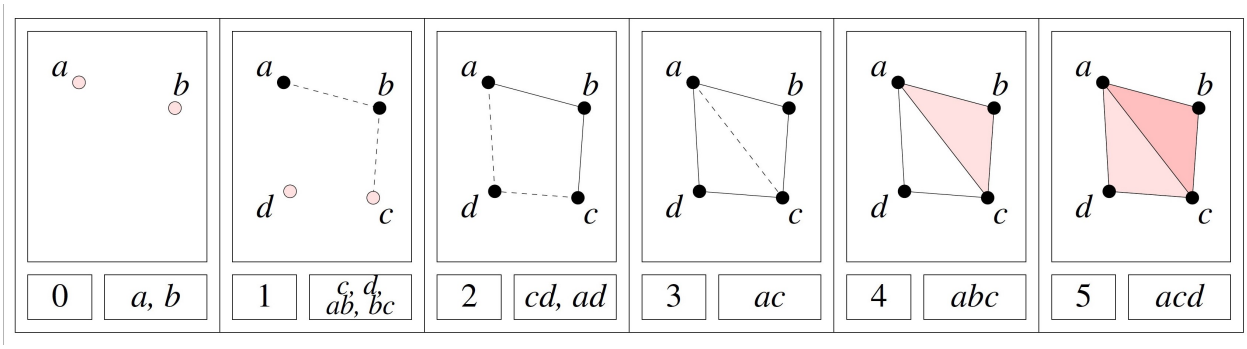


Figure 2.3: A filtred simplicial complex built by adding the vertices c and d to the vertex set $\{a, b\}$ [40].

A filtration over a simplicial complex K is an ordering of the simplices of K such that all prefixes in the ordering are subcomplexes of K (Figure 2.3).

Definition 2.6. Filtration. Let K be a simplicial complex. A *filtration* of K is a nested sequence of subcomplexes

$$\emptyset = K_0 \subseteq K_1 \subseteq ... \subseteq K_n = K.$$

Remark 2.7. More generally, a *filtration* of a topological space $\mathcal{M}$ is a nested family of subspaces $(\mathcal{M}_r)_{r \in T}$, where $T \subset \mathbb{R}$, such that for any $r, r' \in T$, if $r \leq r'$ then $\mathcal{M}_r \subset \mathcal{M}_{r'}$ and, $\mathcal{M} = \bigcup_{r \in T} \mathcal{M}_r$. The subset T might be either finite or infinite; in practice, even if the index set is infinite, all the considered filtrations are built on finite sets and are indeed finite [7]. In this thesis, we only consider filtrations indexed by a totally ordered set (e.g., $T = \mathbb{N}$ or $T = \mathbb{R}$); however, one can consider other types of filtration, for instance, indexed by a poset that is not totally ordered, or by a product T of n totally ordered sets (e.g., $T = \mathbb{R}^n$ or $T = \mathbb{N}^{op} \times \mathbb{R}$), and they are referred as *multifiltrations* [5].

Since $K_{i-1} \subseteq K_i$, the inclusion map $f(x) = x$ induces homomorphisms between the homology groups of these subcomplexes, $f_* : H_p(K_{i-1}) \to H_p(K_i)$. Therefore, a filtration corresponds to sequences (one for each dimension p) of homology groups related by homomorphisms

$$0 = H_p(K_0) \to H_p(K_1) \to ... \to H_p(K_n) = H_p(K).$$

Since homology groups are defined for more than one dimension, and there are often several subcomplexes, we are going to denote the induced homomorphism f_* as $f_p^{i,j} : H_p(K_i) \to H_p(K_j)$. As we go from K_{i-1} to K_i, we gain new homology classes and lose some when they become trivial or merge. We collect the classes born at or before a given threshold and die after another threshold in groups.

Definition 2.8. The image of the induced homomorphism $f_p^{i,j}$ is called the *p-th persistent homology group*, and it is denoted with $H_p^{i,j}$ for $0 \leq i \leq j \leq n$.

Remark 2.9. $H_p^{i,i}$ is simply the p-th homology group of the simplicial complex K_i, $H_p^{i,i} = H_p(K_i)$. Notice that we generally compute homology groups with $\mathbb{Z}_2$ coefficients for simplicity; more precisely, $H_p(K_i)$ is a free abelian group when defined over $\mathbb{Z}_2$.

The elements of $H_p^{i,j}$ are homology classes of K_i that are still alive at K_j; more formally, it is the quotient of K_i by the intersection of the boundaries and cycles of K_j and K_i respectively, $H_p^{i,j} = Z_p(K_i)/(B_p(K_j) \cap Z_p(K_i))$. Notice that $(B_p(K_j) \cap Z_p(K_i))$ is a group (subgroup of $Z_p(K_i)$) since $B_p(K_j)$ and $Z_p(K_i)$ are subgroups of the chain group C_p^j, so this definition is well-defined [40].

Definition 2.10. The rank of the p-th persistent homology group is called *p-th persistent Betti number*, and it is denoted with $\beta_p^{i,j}$.

Similarly to *Betti Numbers* (from an algebraic topology point of view), the p-th persistent Betti number counts the p-dimensional holes that exist all the way from K_i to K_j.
As previously mentioned, persistent homology provides more information about a shape than classical homology. While homology captures cycles in a shape, persistent homology allows for the retrieval of cycles that are non-boundary elements in a particular filtration step, which will become boundaries in subsequent steps. The persistence of a cycle during the filtration gives quantitative information about the relevance of the cycle itself for the shape [14].

Definition 2.11. *Birth* and *Death* of a homology class. A p-th homology class γ is said to be born at K_i if $\gamma \in H_p(K_i)$, but $\gamma \notin H_p(K_{i-1})$. It dies entering K_j if it merges with a class that is born earlier (Figure 2.4). More precisely, $f_p^{i,j-1}(\gamma) \notin H_p^{i-1,j-1}$, but $f_p^{i,j-1}(\gamma) \in H_p^{i-1,j}$.

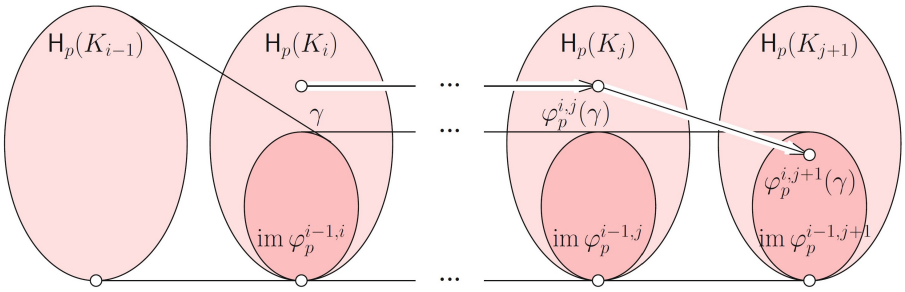


Figure 2.4: γ is born at K_i as it does not belong to the image of $H_p(K_{i-1})$. γ dies entering K_{j+1} because this is the first time its image merges with the image of $H_p(K_{i-1})$ [17].

Definition 2.12. Persistence. The *index persistence* of γ is $j-i+1$. We often have a function that describes the construction of the filtration, and we call the difference between the function values at birth and death the *persistence* of the class.

Let $\mu_p^{i,j}$ denote the number of p-dimensional classes born at K_i and dying entering K_j, then we have

$$\mu_p^{i,j} = (\beta_p^{i,j-1} - \beta_p^{i,j}) - (\beta_p^{i-1,j-1} - \beta_p^{i-1,j})$$

for all $i \leq j$ and all p. The term on the right-hand side of the equation represents the homology classes that are born at or before K_i and die entering K_j, and the second term is the number of homology classes that are born at or before K_{i-1} and die entering K_j. One of the most important results in persistent homology is the following:

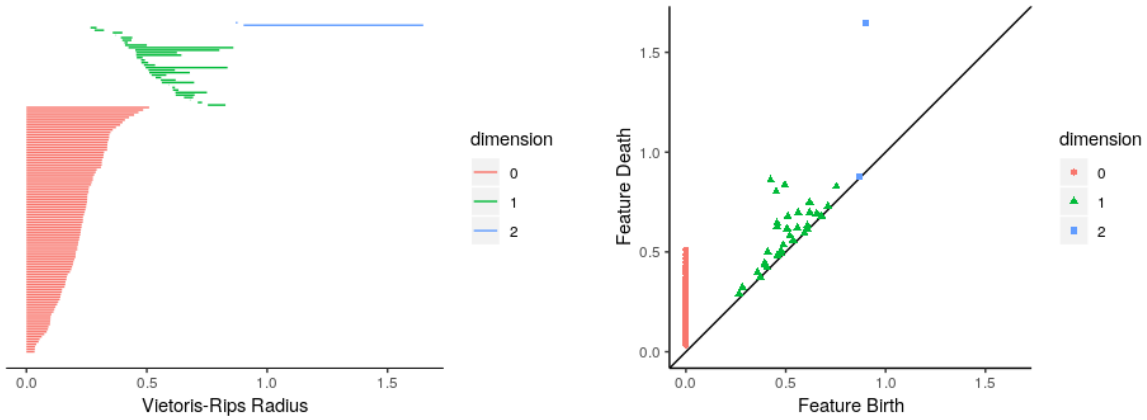
Theorem 2.13. ***Fundamental Lemma of Persistent Homology***. *Let $\emptyset = K_0 \subseteq K_1 \subseteq ... \subseteq K_n = K$ be a filtration. For every pair of indices $0 \leq k \leq l \leq n$ and every dimension p, the p-th persistent Betti number is $\beta_p^{k,l} = \sum_{i \leq k} \sum_{j > l} \mu_p^{i,j}$.*

Edelbrunner and Harer [12] provides a detailed explanation and proof of this lemma.

2.2.2 Persistence Diagrams and Barcodes

The concept of persistent homology involves tracking the creation and destruction of topological features in a given sequence of spaces. For example, when analyzing the filtration of a union of balls, there may be instances where the union contains a hole at specific radii that persists until the balls reach a larger radius and eventually fill in the hole. Specifically, for each hole in the filtration, it is possible to determine the exact radius at which the hole initially emerges and at which point it ceases to exist. These distinct radii can then serve as a set of coordinates to establish a point in the two-dimensional plane that accurately

represents the corresponding hole. Persistent homology groups can be encoded in a graph called *persistence diagram*. A point in the graph represents a homology class, where the x and the y coordinates represent the birth and the death parameter of this class, respectively. For real-world data, the distance of a point from the diagonal indicates the importance of the corresponding feature, the usual interpretation being that points close to the diagonal are likely due to noise (Figure 2.5b).
An alternative way to represent the persistence of a homology class is through *barcodes*: a graph where the length of each bar represents the *life* of a homology class (different colours represent different dimensions), more precisely, the beginning of the barcode is the *birth time*, and the end is the *death time* (Figure 2.5a).



(a) (b)

Figure 2.5: A barcode (a) and a persistence diagram (b) [36].

2.2.3 Persistent Modules

In this section, we revise some basics of abstract algebra and study persistence homology from a different perspective. We see that the persistent homology of a filtred simplicial complex is the homology of a graded module over a polynomial ring [40].
Let R be a commutative ring; if R has no divisors of zero, and all its ideals are principal (an ideal $I \triangleleft R$ is a principal ideal if $I = \langle a \rangle$ for some $a \in R$), it is a principal ideal domain (PID), e.g. $\mathbb{R}, \mathbb{Q}, \mathbb{Z}, \mathbb{Z}_p$ for p prime, and $F[t]$ for F a field. All polynomial $f(t)$ over R also form a commutative ring $R[t]$ with unity, therefore the set of the polynomial $F[t]$ over a field F forms a ring since every field is a ring, to be more precise $F[t]$ is a PID. A *graded ring* is a ring equipped with a direct sum decomposition of abelian groups $R \cong \oplus_i R_i, i \in \mathbb{Z}$ and the elements of each R_i are called *homogenous*. Similarly, a *graded module* M over a **graded** ring R is a module equipped with a direct sum decomposition $M \cong \oplus_i M_i, i \in \mathbb{Z}$.

Theorem 2.14. [40] *Every finitely generated module over a principal ideal domain D is isomorphic to a direct sum of cyclic modules over D. It decomposes uniquely into the form*

$$D^{\beta} \oplus \left(\bigoplus_{i=1}^{m} D/d_i D \right)$$

for $d_i \in D$, $\beta \in \mathbb{Z}$, such that $d_i | d_{i+1}$.
Every graded module M over a graded PID decomposes uniquely into the form

$$\left(\bigoplus_{i=1}^{n} \sum^{\alpha_i} D \right) \oplus \left(\bigoplus_{j=1}^{m} \sum^{\gamma_j} D/d_j D \right)$$

where $d_j \in D$, are homogeneous elements so that $d_j | d_{j+1}$, $\alpha_i, \gamma_j \in \mathbb{Z}$, and $\sum^{\alpha}$ denotes an α-shift upward in grading.

Now, we bring our attention back to persistent homology.

Definition 2.15. Persistent Complex. A *persistence complex* is a sequence of chain complexes $C = (C_*^i)_i$ together with chain maps $\psi : C_*^i \to C_*^{i+1}$.

Remark 2.16. A filtred simplicial complex with the inclusion homomorphisms is a persistent complex.

Suppose we have a PID over R and C is equipped with a graded $R[x]$-module structure, where x acts as a shift map (a unit monomial $x^n \in R[x]$ which sends C_*^i to C_*^{i+n} via n applications of x). Additionally, we assume that all C_*^i are finitely generated as $R[x]$-modules and that the sequence stabilizes in i. This implies that C is free as an $R[x]$-module since we are filtering C through chain maps x. Therefore $H_*(C)$ also has an $R[x]$-module structure, but it is not necessarily free.
Since the only graded ideals of $F[x]$ over a field F are of the form $x^n \cdot F[x]$, one can classify $F[x]$-modules [40]. A direct consequence of the Structure Theorem for PID (Theorem 2.14) is the following:

Theorem 2.17. [40] *For a finite persistence complex C with coefficients over a field F,*

$$H_*(C; F) \cong \bigoplus_i x^{t_i} \cdot F[x] \oplus \left(\bigoplus_j x^{r_j} \cdot (F[x]/(x^{s_j} \cdot F[x])) \right).$$

This equation's free part (left) is in bijective correspondence with the homology generators, which appear at parameter t_i and persist for all future parameter values. The torsional

(right) correspond to the homology generators, which appear at parameter r_j and disappear at parameter $r_j + s_j$. $H_*(C; F)$ is often referred as a *persistence module* over a field F

Remark 2.18. From a computational point of view, we mostly work with $F = \mathbb{Z}_p$ for p a prime; indeed, the persistent homology computations presented in the following chapter involve exclusively $\mathbb{Z}_2$.

More generally, let F be a field, and let **Vect** denote the category of F-vector spaces and linear maps. For P a poset, a (*P-indexed*) *persistent module*, or simply a *P-module*, is a functor $M : P \to$**Vect**. For $x \leq y$ we let $M_{x,y}$ denote the morphism $M_x \to M_y$. If $P = T_1 \times \cdots \times T_n$, where each T_i is a totally ordered set, then M, is called a *multiparameter* or *n-parameter persistence module*, and when $n = 2$, M is referred to as *bipersistence module*. A persistence module is said to be *pointwise finite-dimensional*, or *p.f.d.* if $\dim(M_x) < \infty$ for all $x \in P$ [5]. Therefore, the sequence of homology groups of a filtration over a field yields a persistence module. For an interval I in a poset P, the interval module k_I is defined by

$$(k_I)_x = \begin{cases} k & \text{if } x \in I, \\ 0 & \text{otherwise}, \end{cases} \qquad (k_I)_{x,y} = \begin{cases} \mathrm{Id}_k & \text{if } x \leq y \in I, \\ 0 & \text{otherwise.} \end{cases}$$

In particular, the sequence of homology groups associated with the Vietoris Rips filtration introduced in Example 2.5 yields a persistence module, and the Structure Theorem of Persistence Modules tells us that this module has a well-defined barcode, which is a is a collection of intervals in a totally ordered set. Indeed, persistent homology with field coefficients provides invariants of data called barcodes.

Theorem 2.19. ***Structure of Persistence Modules*** [5]. *If M is p.f.d. persistence module indexed by a totally ordered set T, then there exists a unique multiset $\mathcal{B}(M)$ of intervals in T, such that*

$$M \cong \bigoplus_{I \in \mathcal{B}(M)} k_I.$$

We call $\mathcal{B}(M)$ the barcode of M.

The Structure Theorem of Persistent Modules was proven for $\mathbb{Z}$-indexed modules, $\mathbb{R}$-indexed modules, and the case for finitely presented modules follows from Theorem 2.14 [40].

2.2.4 Stability Theorem

The Stability Theorem states that the persistence diagrams, or the persistence barcodes, are stable under perturbations in the filtration: small changes in the filtration imply only minor

changes in the persistence diagram. In other terms, if the input data is perturbed slightly, the resulting diagram or barcode will only experience small changes [8]. For example, consider a topological space being analyzed using persistent homology; suppose a small change is made to the space, such as adding or removing a single point. The resulting persistence diagram would only exhibit slight modifications rather than significantly restructuring its overall form.

The Stability Theorem does not play a direct role in the analysis presented in the following chapter, but it is a significant source of motivation for the study. Specifically, the persistence diagrams described in the next chapter are anticipated to exhibit similar shapes, as predicted by the Stability Theorem.

A detailed description and proof of the Stability Theorem can be found in the paper by Cohen-Harer-Edelsbrunener [12]. However, additional notations and mathematical information related to the previously introduced filtration are necessary to understand this result.

Let $\mathbb{X}$ be a topological space, for a function $f : \mathbb{X} \to \mathbb{R}$, let $\mathbb{X}_x = f^{-1}(-\infty, x]$ denote the sublevel set for the function value x. Now, let K be a simplicial complex and $f : K \to \mathbb{R}$ be a real-valued function on it. We require that f is *monotonic*: for every $\sigma^{'} \subseteq \sigma$, we have $f(\sigma^{'}) \leq f(\sigma)$. This property ensures that $f^{-1}(-\infty, x]$ are subcomplexes of K for every $x \in \mathbb{R}$. Denoting $K_i = f^{-1}(-\infty, x_i]$, we obtain a a filtration:

$$\emptyset = K_0 \subseteq K_1 \subseteq ... \subseteq K_n = K.$$

The *bottleneck distance* provides a method for measuring the similarities between two persistence diagrams. More precisely, it is the smallest distance d such that there exists a perfect matching between the points (completed with all the points on the diagonal) of two persistence diagrams such that any couple of matched points are at a distance at most d, where the distance between points is the *sup norm* in $\mathbb{R}^2$. In the following definition and theorem, $D_p(f)$ and $D_p(g)$ denote two persistence diagrams for two monotonic functions f and g on a complex K, and the persistent diagrams are regarded as a multiset of points in $\overline{\mathbb{R}}^2$

Definition 2.20. Bottleneck Distance. Let $\Psi = \{\psi\}$ denote the set of all bijections $\psi : D_p(f) \to D_p(g)$, possibly adding points to the diagonal in order to define bijections for diagrams with different cardinalities. Consider the distance between two points $x = (x_1, x_2)$ and $y = (y_1, y_2)$ in $\overline{\mathbb{R}}^2$ equipped with the L_∞-norm as $|x{-}y|_\infty = max\{|x_1{-}y_1|, |x_2{-}y_2|\}$). The *bottleneck distance* between the two diagrams is defined as:

$$W_\infty(D_p(f), D_p(g)) = \inf_{\psi \in \Psi} \sup_{x \in D_p(f)} |x{-}\psi(x)|_\infty$$

W_∞ is a metric on the space of persistence diagrams. Clearly, $W_\infty(X, Y) = 0$ iff $X = Y$.

Moreover, $W_\infty(X,Y) = W_\infty(Y,X)$ and $W_\infty(X,Y) \leq W_\infty(X,Z) + W_\infty(Z,Y)$.

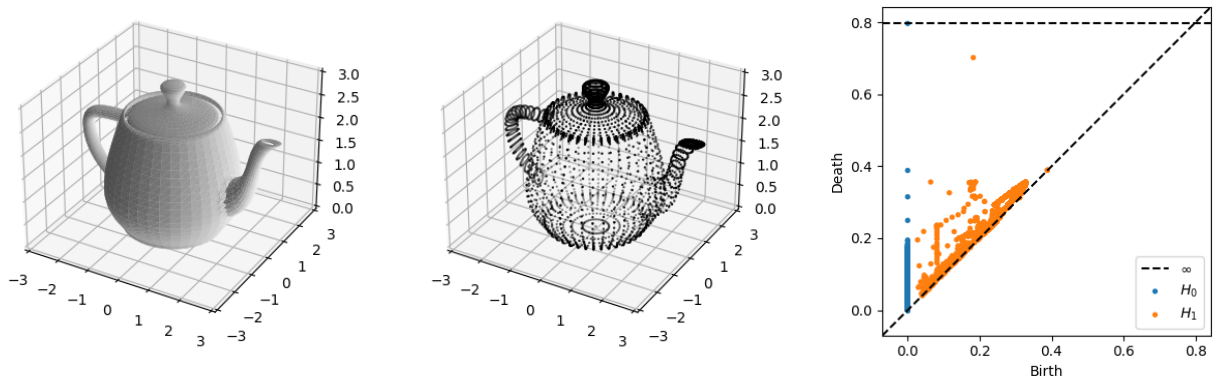
Theorem 2.21. ***Stability Theorem.*** *Let $f, g : K \to \mathbb{R}$ be two monotonic functions defined on a simplicial complex K. Then, for every $p \geq 0$,*

$$W_\infty(D_p(f), D_p(g)) \leq |f-g|_\infty.$$

2.3 Computational Approach

While the mathematical formalisms presented in the previous section provide the foundation for analyzing the topological features of datasets, the actual implementation of this framework requires the use of specialized computer software. Computing the persistent homology of a dataset is a challenging task, even for moderately sized datasets, due to the computational complexity of the algorithms involved. Consequently, we rely on dedicated software packages to perform the necessary computations. Some of the most commonly used software packages for computing persistent homology include `JavaPlex`, `Perseus`, `PHAT`, `DIPHA`, `Ripser`, `Dionysus`, `TDAstats`, and `GUDHI` [21]. In this thesis, we use `Ripser.py` [3], a Python library, to investigate the persistent homology of uniformly distributed points. In the following subsection, we introduce a classical persistent homology computation and an interesting example to demonstrate the power of applying persistent homology to real-world data.

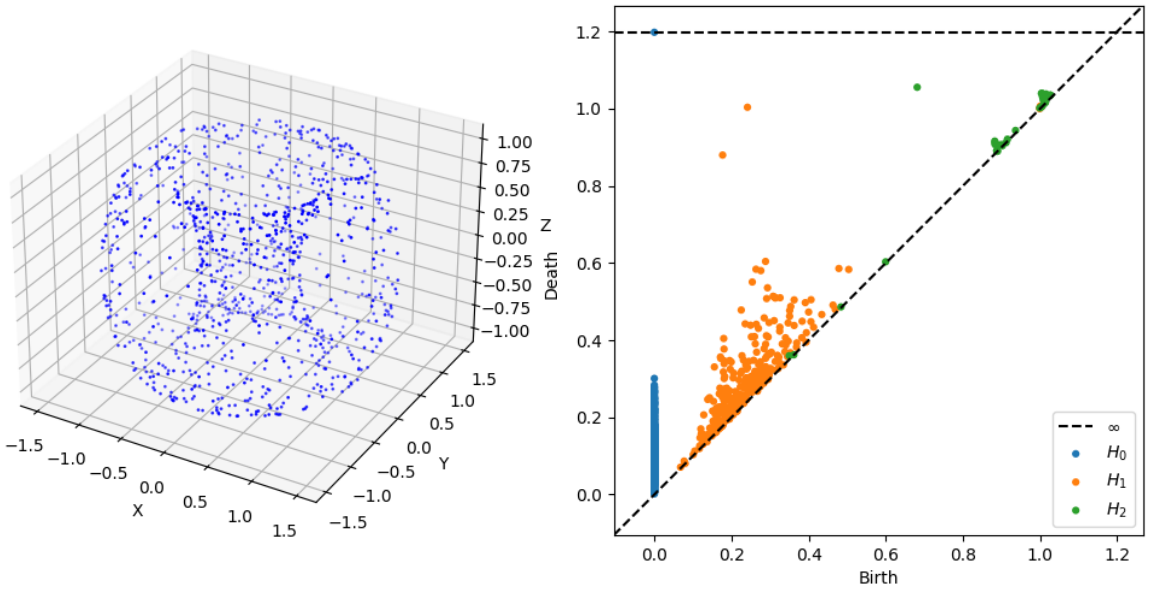
2.3.1 Ripser: PH of the Utah Teapot and a Torus



(a) (b) (c)

Figure 2.6: `ripser` demonstration: the triangular mesh (a), the 3D plot of the vertices of the triangular mesh of the Utah Teapot (b), and its persistence diagram (c).

Example 2.22. The Utah Teapot. The shape of a donut is a typical example used in topology that resembles a torus. Similarly, in computer graphics, the Utah teapot is an iconic object that is used as a standard reference for modelling and rendering [1]. To explore the teapot's topological features, we retrieved its vertices from its `OBJ` file and computed its persistent homology using `ripser`. Then, we generated a 3D rendering of the teapot and plotted its vertices in a scatter plot, visually representing its structure. We also developed a persistence diagram to identify its topological features, which revealed the teapot's handle as a 1-dimensional hole, represented by a point away from the diagonal in the persistence diagram in Figure 2.6. Unfortunately, due to computational constraints, the persistence diagram was limited to homological dimension one. However, these results provide insights into the teapot's shape, highlighting the usefulness of PH for studying topological features in computer graphics.



(a) (b)

Figure 2.7: `ripser` demonstration: a point cloud sampled from a torus (a) and its persistence diagram (b).

Example 2.23. Torus. Figure 2.7 depicts a point cloud that was sampled from the torus and its corresponding persistence diagram. In this case, the persistence diagram has two H_1 points away from the diagonal, indicating the presence of two 1-dimensional holes, and one H_2 point away from the diagonal, indicating the presence of a 2-dimensional hole. A Python script is included in Appendix A.1.

[1] The teapot, complete with its handle, spout, and lid, was created in 1975 by computer scientist Martin Newell at the University of Utah to showcase new computer graphics techniques, and it is available for download as an OBJ file from the Stanford Computer Graphics Lab [33].

Chapter 3

Persistent Homology of Uniform Noise

Understanding the distribution of persistence diagrams presents one of the most significant challenges in topological data analysis, yet it still needs to be solved despite extensive research efforts. One of the main challenges in TDA is distinguishing between significant structures (signal) and random errors (noise) when calculating topological features for a dataset. Although researchers have explored various approaches to address this issue, the distribution of topological noise can take different shapes and forms and is highly sensitive to data generation. As a result, comprehensively understanding noise is an ongoing challenge [4]. In this thesis, we delve into a detailed exploration of noise, focusing on the persistence diagrams of uniformly distributed points.

3.1 Objective

The primary objective of this thesis is to comprehend the properties of the persistence diagrams generated by uniform noise; more precisely, we are interested in relationships between the persistence diagrams of different numbers of points uniformly distributed in a Euclidean Space $\mathbb{R}^d$. Empirical observations derived from simulated data point to similarities between the persistence diagrams of different numbers of points in each homological dimension. This observation leads to the assumption that the birth and death times of the H_k persistent homology bars of $N \in \mathbb{N}$ uniformly distributed points in $\mathbb{R}^d$ both follow probability distributions for all $k < d \leq 3$; our analysis has identified potential distributions that describe the birth and death times across various homological dimension(s) in various dimensions of $\mathbb{R}^d$. It is worth mentioning that our work is focused on something other than identifying the exact statistical distribution but on understanding the behaviour of the parameters of the probability distributions that fit well a specific dataset as the number of points N increases. Moreover, we are interested in relationships between the models that describe these param-

eters, the ambient dimension of the Euclidean Space and the homological dimension. We start by defining the uniform distribution, whose probability density function is given by:

$$f(x; a, b) = \frac{1}{b-a} \quad \text{for} \quad a \leq x \leq b, \tag{3.1}$$

and the cumulative density function is given by:

$$F(x; a, b) = \begin{cases} 0 & \text{if} \quad x \leq a \\ \frac{x-a}{b-a} & \text{if} \quad a \leq x \leq b \\ 1 & \text{if} \quad b \leq x \end{cases} . \tag{3.2}$$

The results presented in the following chapter are derived from simulated data. Thus, in the upcoming sections, we describe the procedures for data collection and the methods employed for analyzing the persistence diagrams. Throughout the research, we utilized Python as the programming language. Appendix A provides the technical details and the various Python libraries used for the analysis. In this thesis, we will use the notations H_0 PH, H_1 PH, and H_2 PH to refer to the 0-dimensional persistent homology, 1-dimensional persistent homology, and 2-dimensional persistent homology, respectively, and these notations will be used interchangeably.

3.2 Data Collection

In order to investigate the PH bars of uniformly distributed points, we decided to take into consideration the following three spaces:

1. The unit interval $[0, 1]$.
2. The unit square $[0, 1]^2$.
3. The unit interval $[0, 1]^3$.

In each of these spaces, we distributed N points uniformly for all $N \in [2, 1000]$ and collected 40 samples for each N. We omitted $N = 1$ for all spaces as it does not provide any meaningful information: a 2-simplex and 3-simplex require a minimum of 2 and 3 points each, and the number of connected components never decreases unless a threshold for the radius ϵ of the 'growing ball' is set. Notice that, for all N, not setting a threshold for ϵ implies the existence of one 0-dimensional persistent homology bar with death time converging to infinity as all the persistent homology bars across all homological dimensions eventually die.

3.3 Persistent Homology Computation

According to [29], the best-performing library in terms of memory usage for the computation of persistent homology with the Vietoris-Rips complex is considered to be `Ripser`. This library is available in Python as `Ripser.py` [3, 34], which was our choice to compute the PH of the data sets described in the previous section. The data has been stored in a Python Dictionary due to its complexity and to facilitate the analysis as we have multiple numbers of simulations, spaces, and homological dimensions (Figure 3.1).

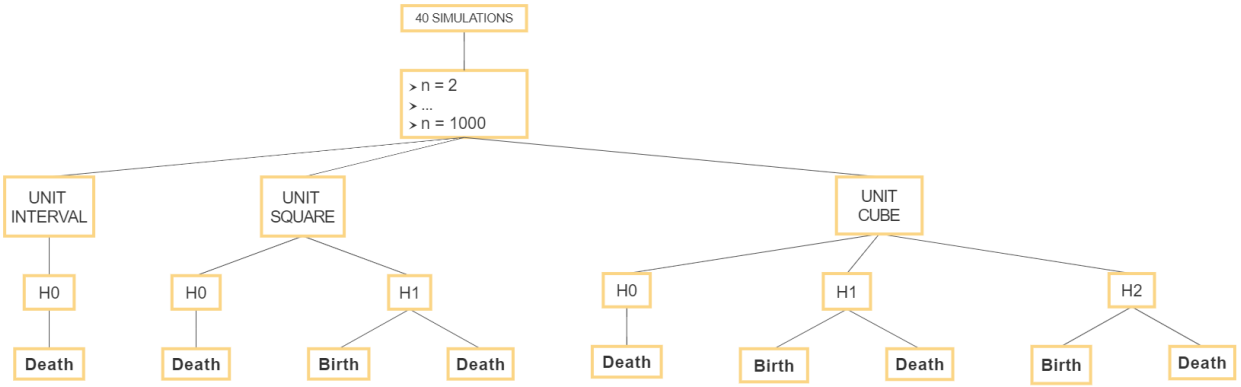


Figure 3.1: Tree Data Structure: hierarchical tree of the data collected for analysis.

3.4 Exploratory Analysis

3.4.1 Persistence Diagrams

In the study of Persistent Homology, two primary methods are used to visualize persistent homology bars: persistence diagrams and persistent barcodes, the former introduced in Section 2.2.2. Although both methods have their uses, barcodes have limited usefulness for our research purposes. In contrast, we have noticed that noise bars tend to cluster in specific regions on the persistence diagram. Thus, persistence diagrams enable us to analyze the distributions of birth and death times separately. Therefore, in the upcoming chapter, we will focus exclusively on presenting persistence diagrams to understand better the likelihood of bars being born and dying at specific times during a filtration.

3.4.2 Histograms

Our initial step involved generating histograms to examine persistent homology bars' birth and death times; the Python libraries `seaborn` and `matplotlib` were utilized to generate

the plots. Histograms are statistical tools utilized to visualize and summarize data, providing a classical nonparametric density estimator that can reliably estimate the underlying probability density function. One can make assumptions about the potential distribution that could describe the data, estimate their parameters, typically using maximum likelihood estimation, and plot the density function on the histogram. Moreover, histograms can also be used to assess the quality of a fitted density by considering additional criteria, such as mean square error, which we introduce in the following section. However, it is important to note that the quality of the fit is subject to the bin width or the number of bins used in constructing the histogram; therefore, to address this issue, a common approach is to use the Freedman-Diaconis rule, which will be introduced in Section 3.5.4.

3.5 Statistical Analysis

The following subsections present the methods used to identify the most suitable statistical distributions for the datasets presented in Section 3.2. Initially, we evaluate various distributions and estimate their parameters using Maximum Likelihood Estimation (MLE). After fitting the model to the data, we calculate each distribution's Sum Square Error (SSE). Based on the SSE values, we select the distributions that best fit the data for further analysis. However, this approach has limitations, as the number of bins used in the histograms impacts the SSE values. To mitigate the impact of the number of bins on the SSE values, we apply appropriate rules that consider the data's size and shape. Moreover, we utilize two well-established criteria, the Bayesian Information Criterion (BIC) [18] and the Akaike Information Criterion (AIC) [37], to compare the distributions with the lowest SSE values and select the most appropriate one for further analysis.

3.5.1 Statistical Distributions used for testing

Our analysis utilizes a variety of distributions available in the `scipy.stats` library [35] for testing, which are listed in Table A1 in Appendix A. It is important to note that the presented probability density functions have been reparametrized using location and scale parameters. The location parameter represents the shift in the center of the distribution, where the distribution is centred around the point indicated by the parameter. Similarly, the scale parameter represents the amount of stretch or shrink in the spread of the distribution. Let X be a random variable with pdf $g(x, \theta_1, \theta_2, \ldots, \theta_k)$ with parameters θ_i, and let μ be the location parameter. Then, the pdf of the reparametrized distribution with location parameter

μ is given by:

$$f(x;\mu,\theta_1,\theta_2,\ldots,\theta_k) = g(x-\mu;\theta_1,\theta_2,\ldots,\theta_k). \tag{3.3}$$

Similarly, let X be a random variable with pdf $g(x;\theta_1,\theta_2,\ldots,\theta_k)$, and let λ be the scale parameter. Then, the pdf of the reparametrized distribution with scale parameter λ is given by:

$$f(x;\lambda,\theta_1,\theta_2,\ldots,\theta_k) = \frac{1}{\lambda} g\left(\frac{x}{\lambda};\theta_1,\theta_2,\ldots,\theta_k\right). \tag{3.4}$$

Therefore, the reparametrization of the pdf $g(x)$ using location μ and scale λ parameters can be written as:

$$f(x;\mu,\lambda,\theta_1,\theta_2,\ldots,\theta_k) = \frac{1}{\lambda} g\left(\frac{x-\mu}{\lambda};\theta_1,\theta_2,\ldots,\theta_k\right). \tag{3.5}$$

As many distributions are being considered, their probability density functions or other characteristics are not included. Several distributions have standard forms, including one or both of a location and scale parameter. For this study, we have denoted the scale parameter with λ and the location parameter with μ. Chapter 4 will provide the probability density function for the distributions that best fit our data. These functions will be presented with their reparametrized versions, which include location and scale parameters.

3.5.2 Maximimum Likelihood Estimation

The maximum likelihood method is a statistical technique used to estimate the parameters of a given probability distribution based on observed data. Let $X_1,\ldots,X_n$ be an iid sample from a population with pdf $f(x;\boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1,\theta_2,...,\theta_k)$ is the vector of parameters, then the likelihood function is defined as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta}), \tag{3.6}$$

where $x_1, x_2, ..., x_n$ are the observed data values and n is the total number of observations. The MLEs are obtained by taking the partial derivatives of the natural logarithm of the likelihood function (3.6) with respect to θ_i, setting them equal to zero and solving each of them for θ_i:

$$\frac{\partial}{\partial\theta_i}\ln(L(\boldsymbol{\theta})) = \frac{\partial}{\partial\theta_i}\left(\sum_{i=1}^{n}\ln f(x_i;\boldsymbol{\theta})\right) = 0. \tag{3.7}$$

Solving (3.7) primarily involves identifying the likelihood function's critical points (maxima, minima, or saddle points). There are instances when the likelihood function is not unimodal, meaning it may have multiple local maxima. In such cases, the optimization process could converge to a local maximum rather than the global maximum. However, the aim is to deter-

mine the parameter values that maximize the likelihood function, corresponding to finding the global maximum. In situations where the solution cannot be obtained analytically, resorting to numerical optimization techniques may be necessary to determine the value of θ_i that maximizes $L(\boldsymbol{\theta})$. Assuming the selected model is true, the MLE has many desirable statistical properties, including consistency and asymptotical efficiency. Consistency means that the MLE converges in probability to the true parameter value as the sample size increases. Asymptotical efficiency means that the MLE has the smallest possible variance among all unbiased estimators, making it the most precise estimator in large samples. It is worth noting that the maximum likelihood assumes the existence of an identifiable probability model, and violating this assumption can lead to biased or inefficient estimates [24].

3.5.3 Sum Square Error

The sum squared error (SSE), also known as *residual sum of squares* [11], is a standard measure of the difference between a dataset's predicted and observed data values; more precisely, the SSE is defined as

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.8}$$

where n is the total number of observations, y_i and $\hat{y}_i$ represent the value and the predicted value of the i-th observation, respectively. The SSE cannot be used for model selection since it is monotone decreasing with model size; more precisely, increasing the model size might provide a better fit, but it could also capture noise and lead to overfitting. Additionally, when the observed data comes from a histogram, the number of bins used can affect the SSE, and selecting the best model can become challenging.

Various methods and rules can be used to determine the optimal number of bins, such as Scott's, Freedman-Diaconis, Sturges, Doane's, Stone's, Rice or Square Root Rules. However, we will mainly use the Freedman-Diaconis Rule, as varying the number of bins has not significantly impacted selecting the best fits based on the SSE. Moreover, we will use two other methods, AIC and BIC, to evaluate the best models that describe our data. We will be focusing on distributions with the lowest amount of parameters because we are interested in understanding how the number of points uniformly distributed in our initial spaces (Unit Interval, Unit Square, and Unit Cube) affects the estimated parameters of the selected distribution.

3.5.4 Freedman-Diaconis and Rice Rule

The Freedman-Diaconis rule, proposed in 1981 by David Freedman and Persi Diaconis, employs the dataset's interquartile range (IQR) instead of the standard deviation to determine the optimal bin size for histogram construction. The IQR, defined as the difference between the 75th and 25th percentiles, is a robust measure of data spread that is less susceptible to the influence of outliers than the standard deviation. The formula for the bin size calculation according to the Freedman-Diaconis rule [13] is given by:

$$bin\ size = \frac{2 \times IQR}{\sqrt[3]{n}} \tag{3.9}$$

This formula provides a reliable criterion for selecting an appropriate bin width for the dataset's variability and size, resulting in a visually informative histogram. However, the Freedman-Diaconis rule only applies to approximately symmetric datasets for which the underlying distribution is unknown [13]. Therefore, other methods, such as Scott's or the Sturges rule, may be more appropriate for datasets that are not symmetric or for which the underlying distribution is known. For datasets with exponential distributions, one alternative method that can be used is the Rice rule [32], which is based on the sample size and the standard deviation of the data. The formula for the Rice rule is given by:

$$bin\ size = \frac{2 \times \sigma}{n^{\frac{1}{3}}} \tag{3.10}$$

Choosing an appropriate bin size is crucial for distribution fitting accuracy. However, deciding on the best distribution can be challenging when multiple distributions fit the data well. To assist with this decision-making process, we use BIC and AIC, two techniques for evaluating the goodness of fit of different distributions.

3.5.5 BIC and AIC

Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) are selection criteria used to determine the best model among a set of candidates by balancing the complexity and the models' fit to the data. The BIC is more helpful in selecting a correct model, while the AIC is more appropriate in finding the best model for predicting future observations, and the model with the lowest BIC or AIC value is considered the best among the set of candidate models [18].

The BIC criterion, also referred to as the *Schwarz information criterion* and *Schwarz Bayesian information criterion*, is a large sample approximation to the Bayes Factor [18],

and it is defined as:

$$BIC = k\ln(n) - 2\ln(L)$$

where k represents the number of parameters in the model, n is the number of observations in the dataset, and L is the likelihood of the data under the model. The BIC criterion penalizes models with more parameters than the Akaike Information Criterion, which helps address the problem of overfitting. The BIC balances model complexity and goodness of fit by including the penalty term $k\ln(n)$.

The AIC criterion [37] is defined as:

$$AIC = 2k - 2\ln(L)$$

where k is the number of parameters in the model, and L is the likelihood of the data under the model. AIC criterion is based on minimizing the *Kullback-Leibler divergence*[1] and penalizes models with more parameters but less heavily than BIC. More precisely, the term $2k$ is considered a *penalty term* since a higher number of estimated parameters leads to an increase in the AIC value penalizing the model; therefore, among multiple models, the model with providing the best fit to the data might not necessarily be the model with the lowest AIC value due to its complexity.

It is essential to note that BIC and AIC are not always the best criteria for model selection. Evaluating the models using other criteria and visually inspecting the results is always recommended [6].

Chapter 4 aims to identify universally suitable distributions for datasets (birth and death times) with identical space and homological dimension. To achieve this, we employ the Bayesian Information Criterion (BIC) as our primary criterion for model selection. BIC balances model complexity and goodness of fit by penalizing complex models with more parameters. By favouring simpler models, we can avoid overfitting and emphasize identifying the distributions that best represent the birth and death times. Through our analysis, we expect that as the sample size (or N) increases, the distributions with the lowest BIC values will converge and become consistent across datasets with identical space and homological dimension. In contrast, the AIC values will tend to select different distributions for each dataset. Although we will report Akaike Information Criterion values for completeness, AIC tends to select complex models, which may not align with our goal of finding standard

[1]The Kullback-Leiber divergence is a measure of the difference between two probability distributions P and Q; it can be interpreted as a measure of information gained or lost when using a probability distribution Q to approximate P. It is worth mentioning that the KL divergence is not a distance metric since it does not satisfy the triangle inequality.

distributions. Thus, BIC is better suited to our purpose to ensure robust and reliable findings. In summary, the convergence of the best-fitting distributions with increasing sample size supports the effectiveness of this approach.

Chapter 4

Experiments

In this chapter, we present the data collected and results of our experiments using the methodologies presented in the previous chapter; we distributed N points uniformly in $\mathbb{R}^d$, computed the persistent homology and recorded the exact birth and death time of each persistent homology bar. We are interested in estimating the chance that a given persistent homology bar is being created (or destroyed) at a given time. We know that the number of persistent homology bars changes according to the location of the N points in the unit interval, square or cube, and we also know that the persistent homology bars that persist for a relatively long time are interpreted as important features of the data, and short bars can be considered noise. Distributing N points uniformly aims to treat all these points as noise and minimize the presence of important features. However, we have seen from experimental results that the persistent homology bars are more likely to be located in a particular range in the persistence diagram. Therefore, we considered the union of 40 samples (persistence diagrams) for each N, homological dimension, and space to identify probability density functions for each birth and death parameter, providing the probability that a PH bar is created and destroyed at certain times. Due to the low number of persistent homology bars in each computation and to ensure a sufficient number of data points for subsequent analysis, we gathered the birth and death times separately for each combination of space, homological dimension, and N from the 40 simulations to form combined datasets. However, the assumption that these times are independently and identically distributed (iid) might not hold: while the birth and death times are likely independent across different simulations, they may not be independent within a single simulation due to the spatial distribution of points. In our current analysis, we estimated the parameters of the underlying distributions of birth and death times by applying the Maximum Likelihood Estimation (MLE) method directly to the standard probability functions. This approach does not explicitly account for potential dependencies within simulations or the varying number of birth and death times across simulations. It is

essential to consider this limitation when interpreting the presented results. In future work, a mixed-effects model should be considered, which would allow for the inclusion of random effects that represent simulation-specific variations and provide a more nuanced understanding of the distributions. While our current approach provides a simplified view of the data structure, it effectively captures the overall trends in birth and death times across varying spaces and homological dimensions.

4.1 Data Visualization

In this section, we visualize the persistent homology bars of our data using persistence diagrams, which we described in Chapters 1 and 2. Notice that the 0-dimensional homological features are all born at $t = t_0$, so the use of persistence diagrams does not provide meaningful information for this specific case.

A visual inspection of the persistence diagrams presented in Figures 4.1, 4.2, and 4.3 suggest that the PH bars (represented as two-dimensional points in the persistence diagrams) are clustered in a particular area. This characteristic appears to be valid for a small and large number of points. However, it is worth noticing that this cluster scales as the number of point N in the spaces increase; more precisely, the highest death time of a sample decreases as N increases.
The phenomena raise some questions:

1. How are the number of PH bars changing as N increases?

2. Are the number of PH bars changing similarly across various homological dimensions in a certain space?

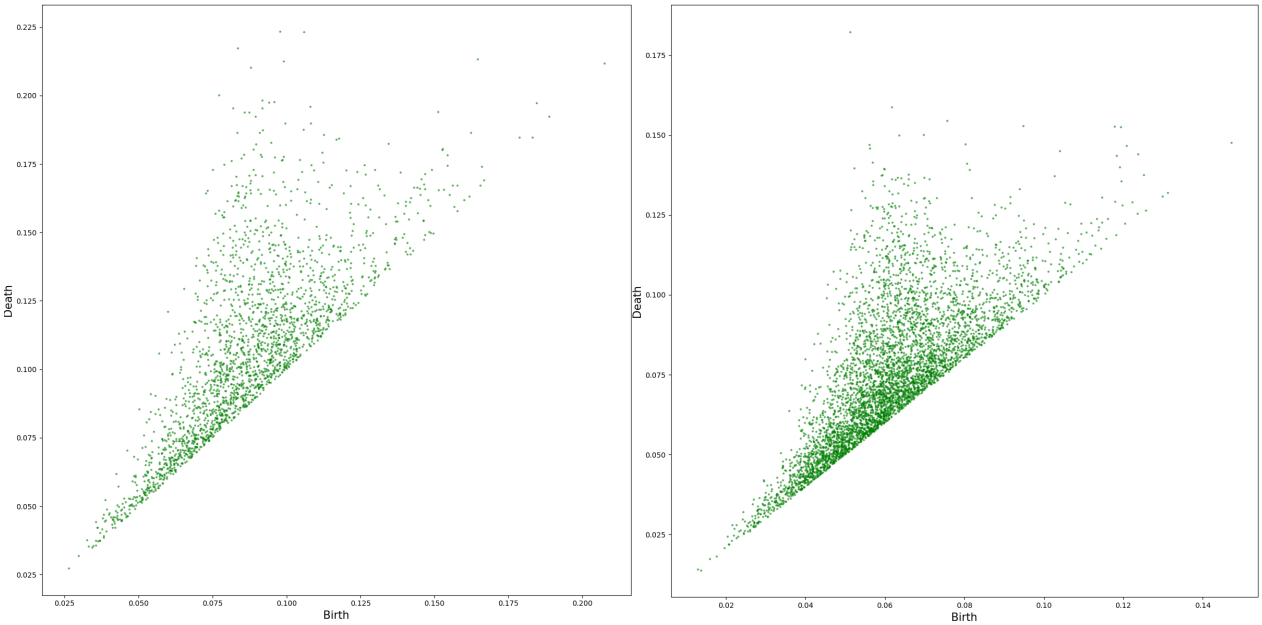
3. Are the PH bars of a specific homological dimension changing similarly across different dimensions of $\mathbb{R}^d$?

4. What is the likelihood that a certain PH feature is born or dead at a given time for a given N?

5. Suppose the birth or death of a certain PH feature in a certain space follows a statistical distribution; what is the distribution?

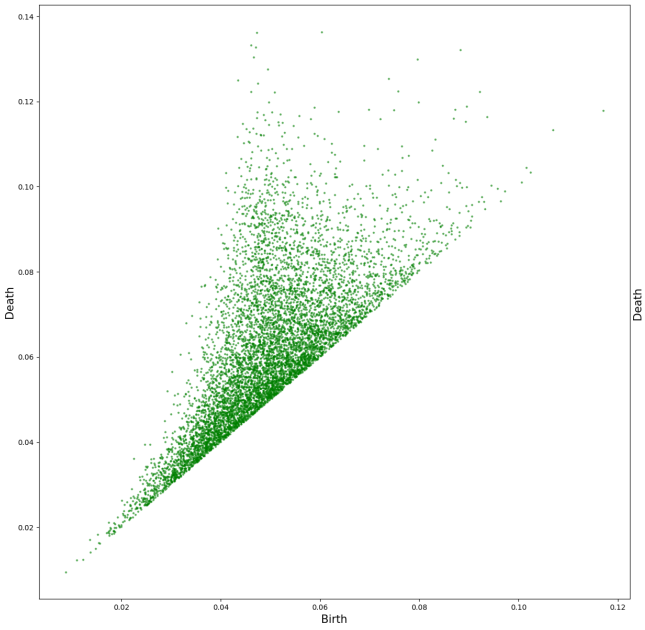
6. How does this distribution change as N increases?

7. Are there any interesting relationships among the distributions of the birth and death times for a given space and homological dimension?

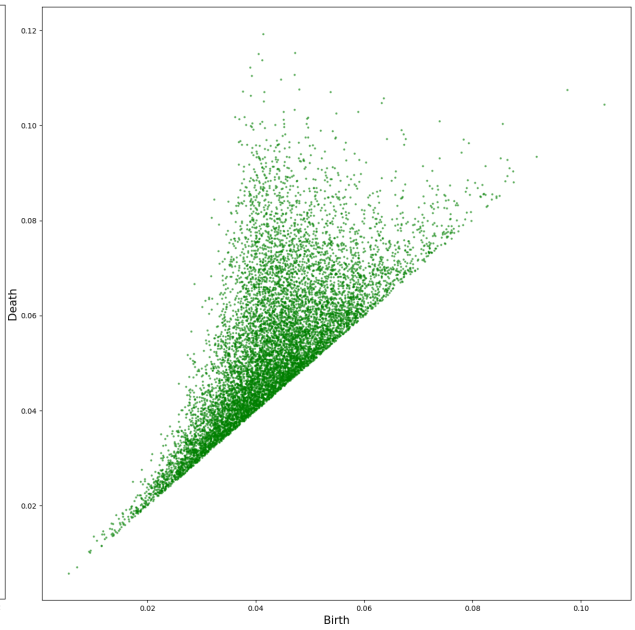


(a) $N = 250$

(b) $N = 500$

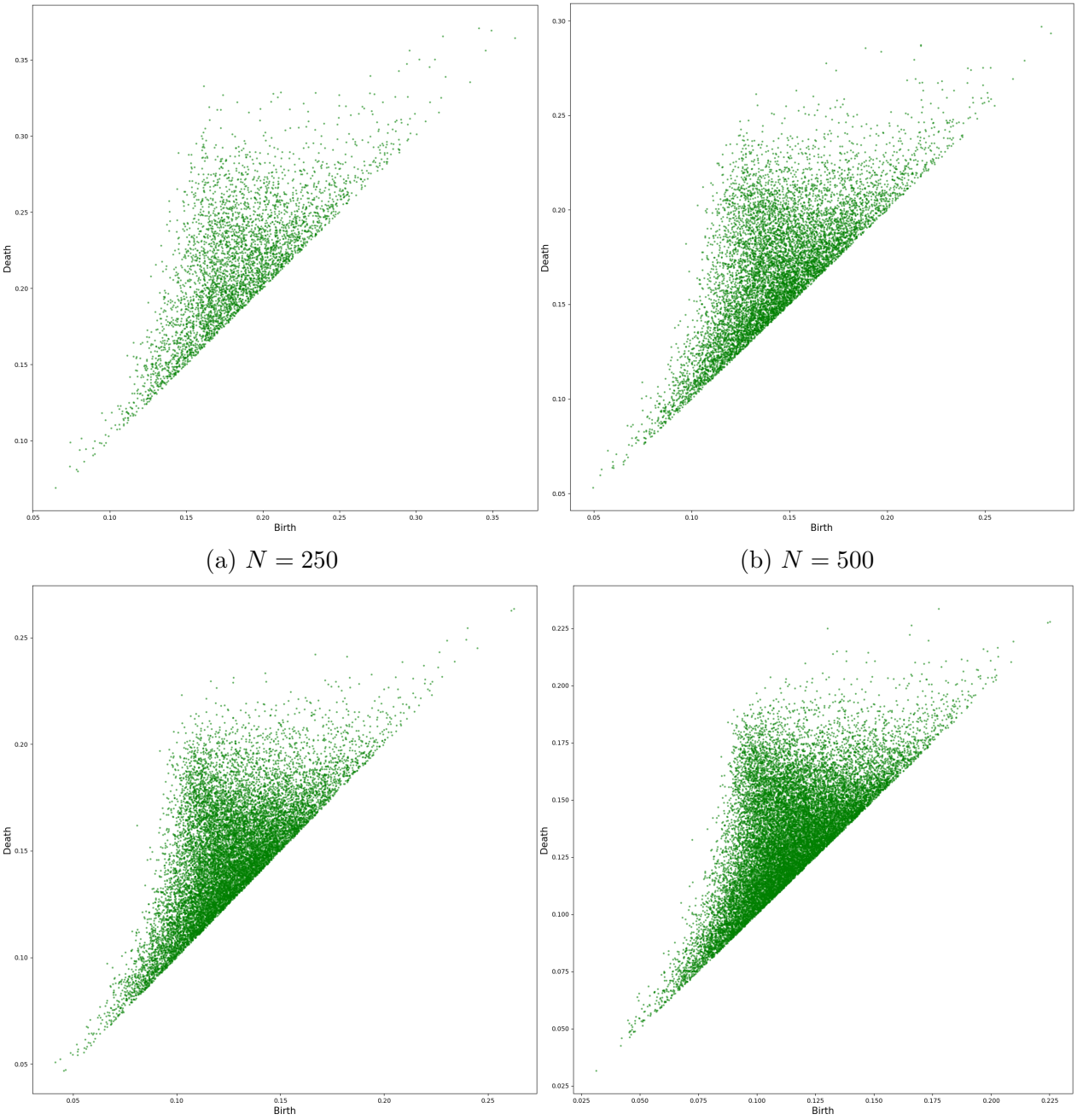


(c) $N = 750$



(d) $N = 1000$

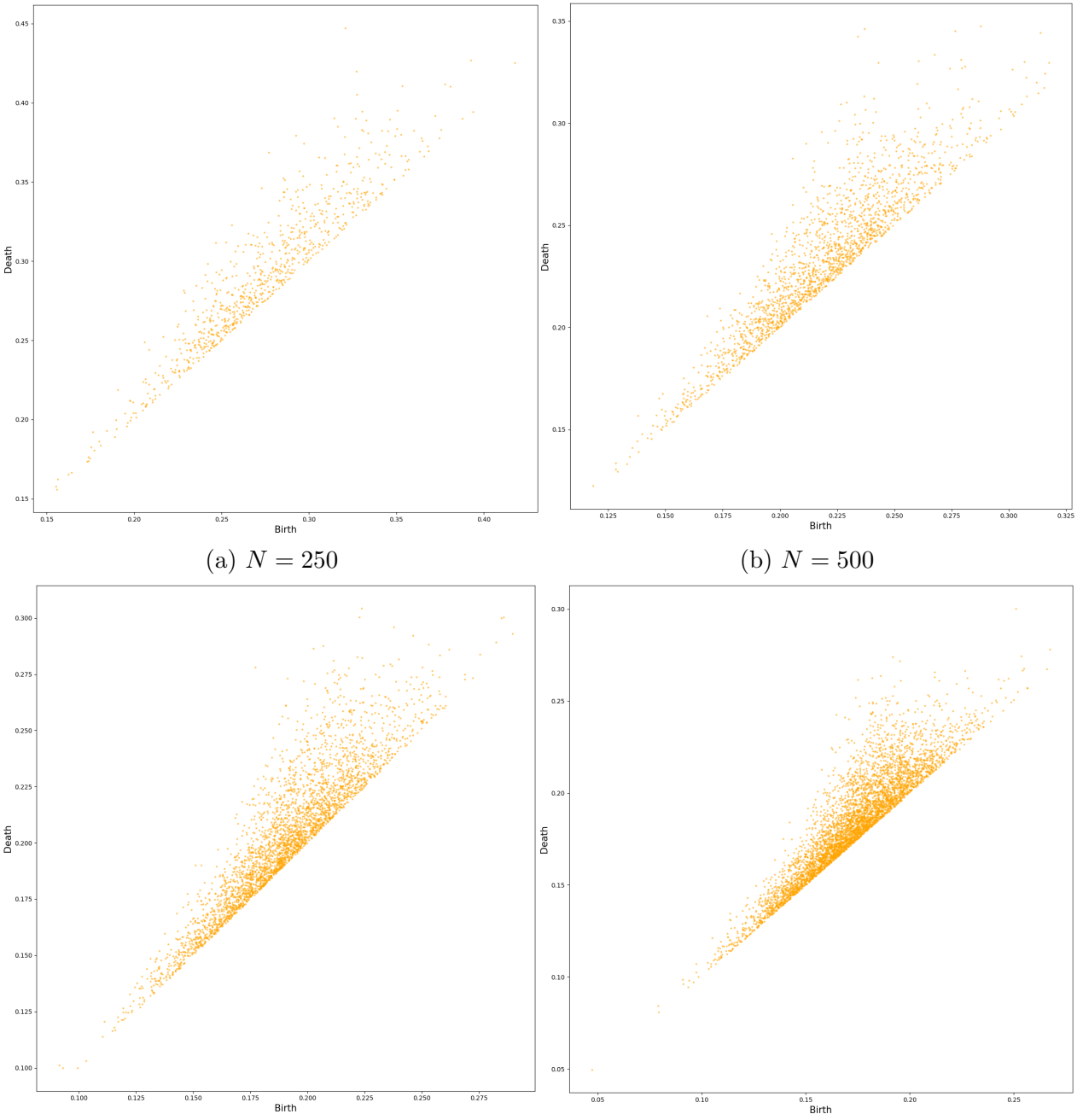
Figure 4.1: Overlapping Persistence Diagrams for homological dimension 1 of 40 samples of $N \in \{250, 500, 750, 1000\}$ uniformly distributed points in $[0, 1]^2$.



(a) $N = 250$ (b) $N = 500$

(c) $N = 750$ (d) $N = 1000$

Figure 4.2: Overlapping Persistence Diagrams for homological dimension 1 of 40 samples of $N \in \{250, 500, 750, 1000\}$ uniformly distributed points in $[0, 1]^3$.



(a) $N = 250$

(b) $N = 500$

(c) $N = 750$

(d) $N = 1000$

Figure 4.3: Overlapping Persistence Diagrams for homological dimension 2 of 40 samples of $N \in \{250, 500, 750, 1000\}$ uniformly distributed points in $[0, 1]^3$.

8. Is there any relation among the distributions of PH bars of different homological dimensions in a certain space?

9. Is there any relation among the distributions of PH bars of different spaces?

In order to address questions, we analyzed the various collections of PH bars.

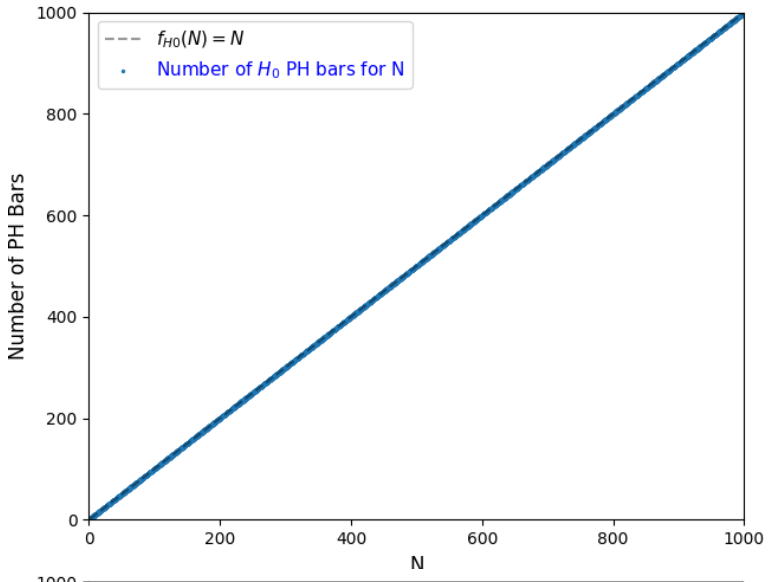
4.2 Persistent Homology Bar Count

H_i	**Space**	**Linear Model**	**Power model**
H_0	All	$f_{H0}(N) = N$	$g_{H0}(N) = N$
H_1	Square	$f_{H1}(N) = 0.25616499 \cdot N - 6.81195775$	$g_{H1}(N) = 0.14859689 \cdot N^{1.07659142}$
	Cube	$f_{H1}(N) = 0.4957729 \cdot N - 17.43039788$	$g_{H1}(N) = 0.23151776 \cdot N^{1.1074826}$
H_2	Cube	$f_{H2}(N) = 0.1210337 \cdot N - 8.404324$	$g_{H2}(N) = 0.02234173 \cdot N^{1.23980489}$

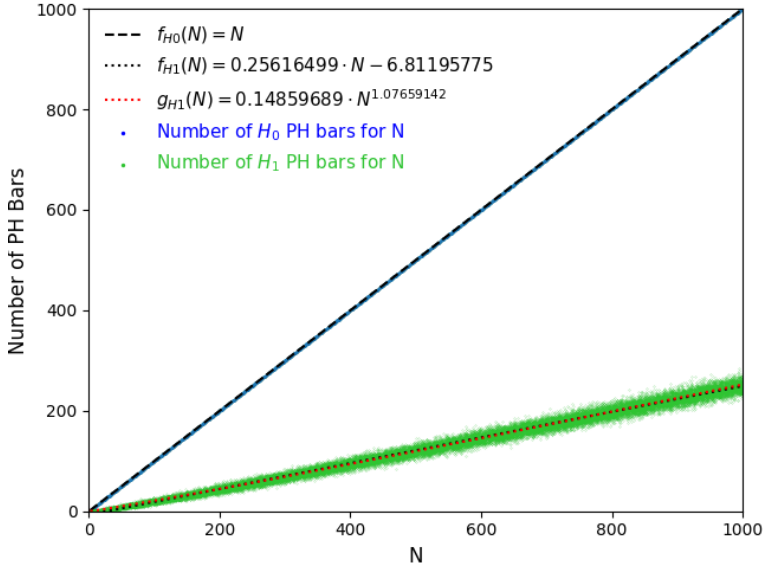
Table 4.1: Linear and power model fits describing the relationships between the number of persistent homology bars for all homological dimensions and the number of uniformly distributed points N in $[0, 1]$, $[0, 1]^2$, and $[0, 1]^3$.

In this section, we investigate the relationship between the number of persistent homology (PH) bars and the number of uniformly distributed points in the unit interval, unit square, and unit cube. As mentioned in the previous sections, we conducted 40 simulations for each value of N and analyzed the data for all homological dimensions in all spaces. We fitted first-order polynomial equations and power models aN^b to the scatter plots of PH bars against N. Specifically, we used the `curvefit` function from the `scipy` library to obtain the best-fit equations. The linear models and power models for each homological dimension and space are summarized in Table 4.1. As we can see from the table, the best-fit linear equation for H_0 is $f_{H0}(N) = N$ in all cases. For H_1, the linear model is $f_{H1}(N) \approx 0.26 \cdot N - 6.81$ in the square and $f_{H1}(N) \approx 0.49 \cdot N - 17.43$ in the cube. For H_2, the linear relantionship is $f_{H2}(N) \approx 0.12 \cdot N - 8.40$ in the cube. The power models also suggest a linear relationship between the number of PH bars and N, as the exponents are close to 1. For H_0, the power model is $g_{H0}(N) = N$ in all spaces. For H_1, the fitted model is $g_{H1}(N) \approx 0.15 \cdot N^{1.08}$ in the square and $g_{H1}(N) \approx 0.23 \cdot N^{1.12}$ in the cube. For H_2, the fitted equation is $g_{H2}(N) \approx 0.02 \cdot N^{1.24}$ in the cube. We observed a linear increase in the number of PH bars as N increases, which is consistent with the power models. However, we also observed some differences in the number of PH bars between the square and cube. For a fixed value of N, the number of PH bars decreases on average as the homological dimension increases, as visible in Figure 4.4. For H_1, the number of PH bars in the cube is higher than in the square and roughly half the number of H_0 bars for large N, whereas, in the square, it is a quarter the number of H_0 bars.

(a) Unit Interval:



(b) Unit Square:



(c) Unit Cube:

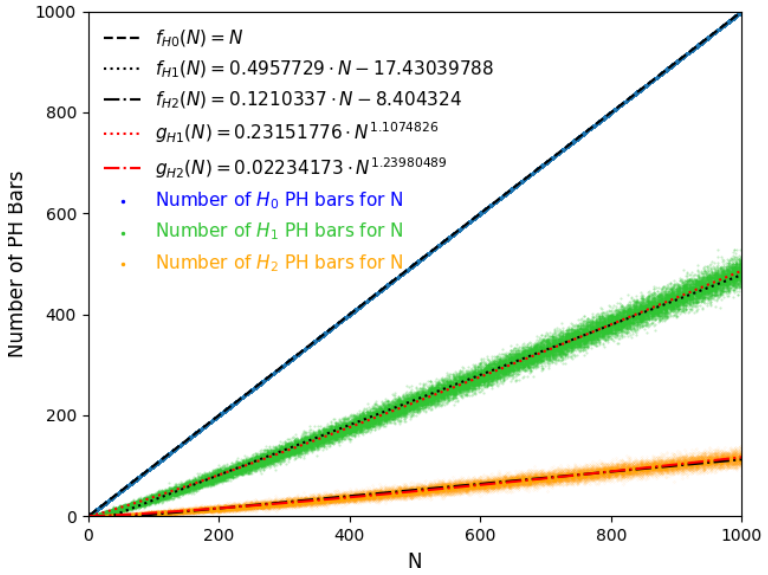


Figure 4.4: Curve fitted on the scatter plots of the number of PH bars for 40 simulations of $N \in [2, 1000]$ uniformly distributed points in $[0, 1]$, $[0, 1]^2$, and $[0, 1]^3$.

Furthermore, the number of H_2 bars in the cube is lower than the number of H_1 bars in both the square and cube. Overall, our results provide interesting insights into the behaviour of the number of PH bars in different spaces and homological dimensions for various Ns, and they can be used as a starting point for further research to develop models that describe the persistence diagrams of noisy data. We now proceed to study the distributions of birth and death times.

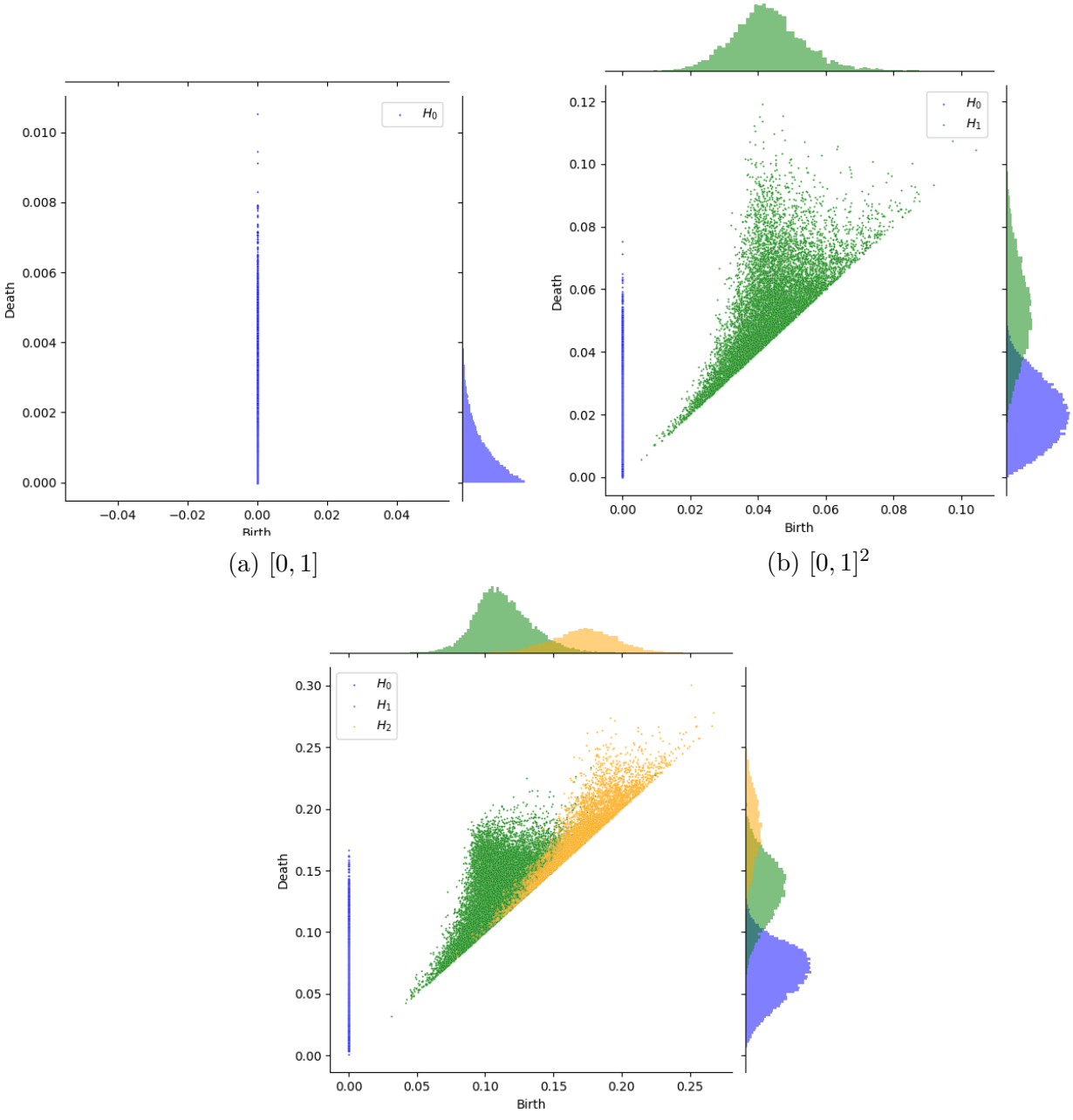
4.3 Exploratory Analysis

4.3.1 Kernel Density Estimation and Histograms

In this section, we explore the persistence diagrams of our datasets in further detail. We observed in the last section that the persistence diagrams for a specific homological dimension and space appear to be similar for different numbers of points, with the number of PH bars increasing linearly as N increases. Additionally, the points in the persistence diagrams tend to cluster in particular regions, indicating that the PH bars follow a statistical distribution. To investigate the shape of these distributions, we visually analyze the persistence diagrams. We first consider the overlapping persistence diagrams of 40 simulations for $N = 1000$ in the unit interval, square, and cube. Then, by plotting the marginal histograms of birth and death times for each homological dimension in each space, we observe that these histograms exhibit the shapes of probability density functions, with some being roughly bell-shaped and others exhibiting varying levels of kurtosis and "fat" tails (Figure 4.5).

We then plot the normalized histograms of birth and death times separately for different values of N (Figure 4.8). The resulting plots reveal distinct statistical distributions, with the histogram for death times in the unit interval suggesting an exponential distribution may be a good fit (Figure 4.8a)). In contrast, the histograms for 0-dimensional persistent homology death times in the unit square and unit cube exhibit bell-shaped curves that are positively (Figure 4.8b) and negatively skewed (Figure 4.8e), respectively. The remaining histograms are broadly symmetric but with varying levels of kurtosis, some with "fat" tails and others with less.

To better understand the relationship between the theoretical distributions of birth and death times and the persistence diagrams, we derive the kernel density estimates (KDE) of birth and death times and create 3D plots (Figures 4.6 and 4.7). This joint KDEs estimate the joint distributions of birth and death times. The resulting plots resemble the shape of the persistence diagram when viewed from the top, with different colours (*heat map*) indicating the most clustered points. Furthermore, the birth times can be visualized from the birth



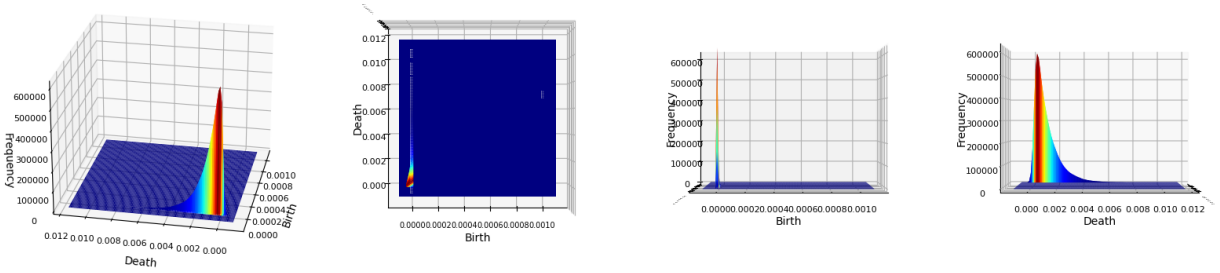
(a) $[0, 1]$ (b) $[0, 1]^2$

(c) $[0, 1]^3$

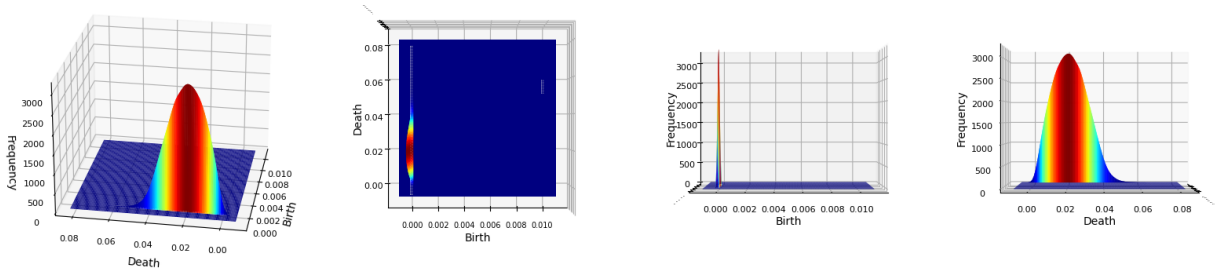
Figure 4.5: Persistence diagrams with marginal histograms for each homological dimension for 40 simulations of $N = 1000$ points uniformly distributed in $[0, 1]$, $[0, 1]^2$, and $[0, 1]^3$.

axis, while the death times can be visualized from the death axis. However, for this thesis, we focus only on the marginal distributions separately. It is worth noting that an analysis of distributions of persistence diagrams using KDE has been presented in [23] by Mike and Maroulas. Therefore, although our joint KDE plot provides insights into the behaviour of the data, it will not be used in this thesis other than visualizing the fitted distribution on

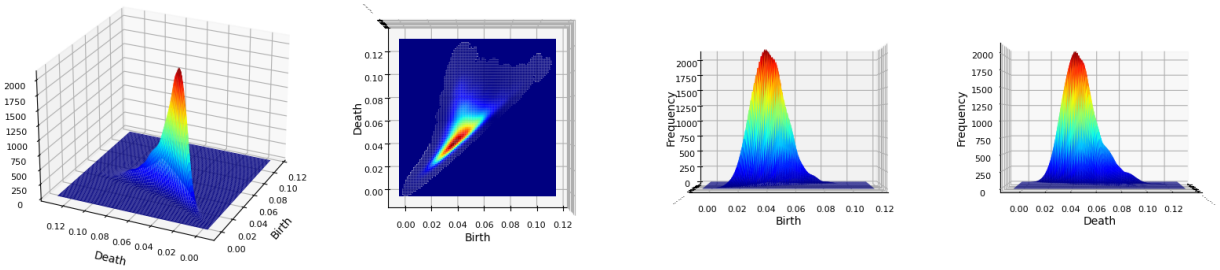
the histogram.



(a) 2D-KDE of H_0 PH bars of N points from $[0, 1]$.



(b) 2D-KDE of H_0 PH bars of N points from $[0, 1]^2$.

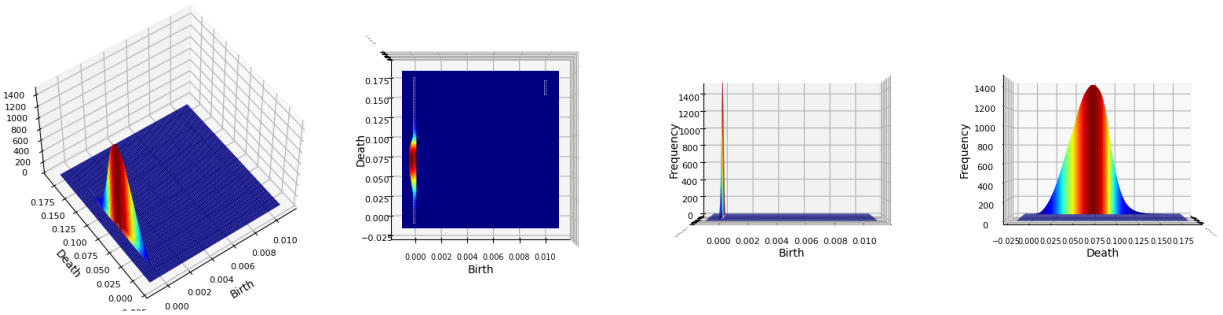


(c) 2D-KDE of H_1 PH bars of N points from $[0, 1]^2$.

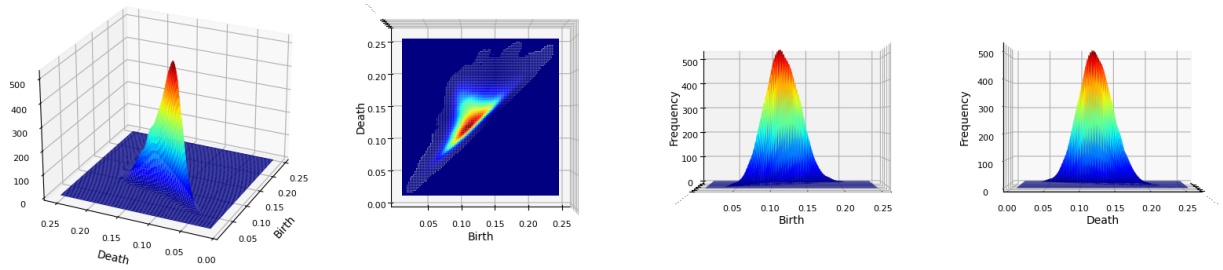
Figure 4.6: Joint KDE plots (from different angles) of the H_i Birth and Death times of the PH bars of $N = 1000$ points in $[0, 1]$, and $[0, 1]^2$.

4.3.2 Normality Test

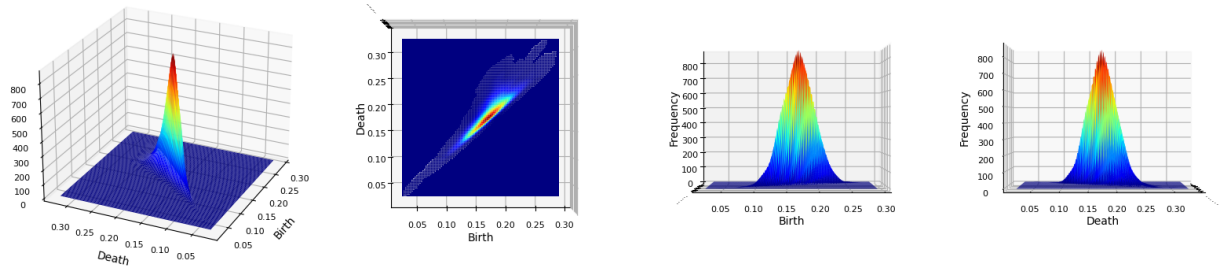
Some histograms in Figure 4.8 resemble the normal distribution shape. To test normality, we used the `stats.normaltest` function from the `scipy` library in Python; this method relies on D'Agostino's K-squared test, which is discussed in [9] and [10]. If the p-value of the normality test is less than a significance level of 0.05, we reject the null hypothesis that



(a) 2D-KDE of H_0 PH bars of N points from $[0,1]^3$.

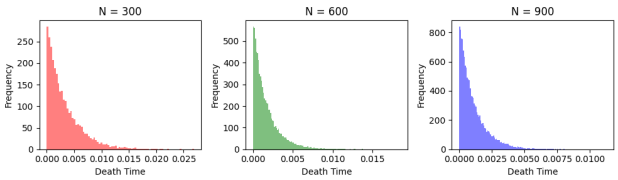


(b) 2D-KDE of H_1 PH bars of N points from $[0,1]^3$.

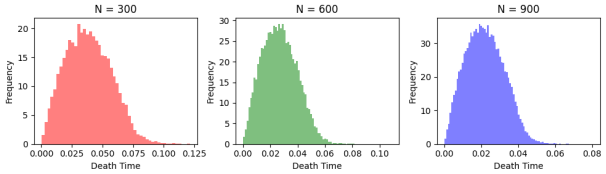


(c) 2D-KDE of H_2 PH bars of N points from $[0,1]^3$.

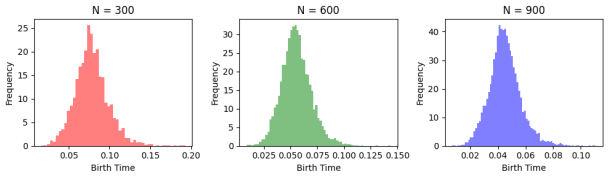
Figure 4.7: Joint KDE plots (from different angles) of the H_i Birth and Death times of the PH bars of $N = 1000$ points $[0,1]^3$.



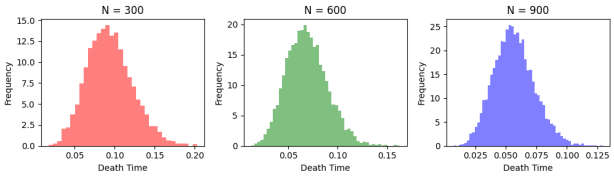
(a) Death of the H_0 PH bars of N points in $[0,1]$.



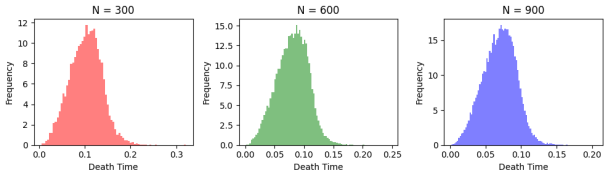
(b) Death of the H_0 PH bars of N points in $[0,1]^2$.



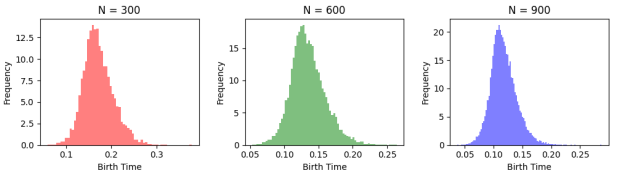
(c) Birth of the H_1 PH bars of N points in $[0,1]^2$.



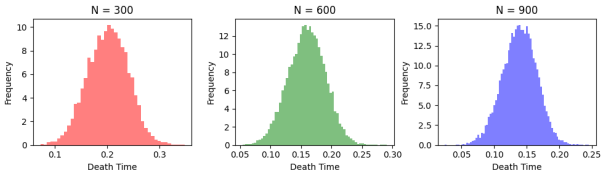
(d) Death of the H_1 PH bars of N points in $[0,1]^2$.



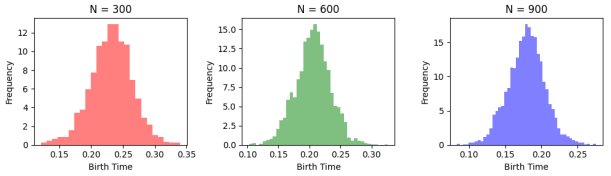
(e) Death of the H_0 PH bars of N points in $[0,1]^3$.



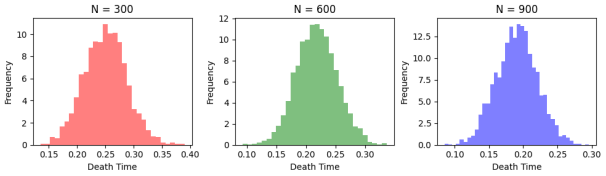
(f) Birth of the H_1 PH bars of N points in $[0,1]^3$.



(g) Death of the H_1 PH bars of N points in $[0,1]^3$.



(h) Birth of the H_2 PH bars of N points in $[0,1]^3$.



(i) Death of the H_2 PH bars of N points in $[0,1]^3$.

Figure 4.8: Normalized histograms of Birth and Death times of the PH bars for 300, 600, and 900 points in $[0,1]$, $[0,1]^2$, and $[0,1]^3$ (obtained from the union of 40 persistence diagrams).

the data follows a normal distribution. In this case, the alternative hypothesis is that the data deviates from the normal distribution. The test only failed to reject the null hypothesis for the Death times of the 2-dimensional persistent homology bars, implying that this data set might follow a Normal distribution. However, it should be outlined that the results of statistical tests are not conclusive and consider the limitations. The results are presented in Table 4.2, which includes the statistic score[1], p-value, and whether the null hypothesis is rejected; we note that the table's results represent $N = 1000$, but we have repeated these tests for different Ns and obtained similar results.

`scipy.stats.normaltest` results					
Space	Param.	H_k	Statistics	p-value	Null Hypothesis
$[0,1]^1$	`Death`	H_0	15829.392903629894	0.0	Rejected
$[0,1]^2$	`Death`	H_0	792.6801868860335	7.441910808378268e-173	Rejected
$[0,1]^2$	`Birth`	H_1	155.8778188393814	1.4176261425402775e-34	Rejected
$[0,1]^2$	`Death`	H_1	425.6761394557955	3.6779072948477658e-93	Rejected
$[0,1]^3$	`Death`	H_0	95.07096610181995	2.2677857206367554e-21	Rejected
$[0,1]^3$	`Birth`	H_1	839.7377997882116	4.5004262918175985e-183	Rejected
$[0,1]^3$	`Death`	H_1	67.03631690329354	2.774908970574887e-15	Rejected
$[0,1]^3$	`Birth`	H_2	33.17317121734176	6.25946761070106e-08	Rejected
$[0,1]^3$	`Death`	H_2	1.028448344069168	0.5979643321559095	Not Rejected

Table 4.2: Normality test performed using `scipy.stats.normaltest` for $N = 1000$.

4.4 Statistical Analysis: Distribution Fitting

In the upcoming sections, we present the analysis results for the datasets presented in the previous chapter using the techniques described earlier in this chapter. Among the many distributions utilized for testing, we only present the ones with the lowest SSE with their respective SSE, BIC and AIC. We can use BIC and AIC to identify the distributions that best fit the data; however, the central scope of this thesis is not to find the best fit but to understand how the parameters of a distribution vary as we increase the number of points N uniformly distributed in a specific space. Therefore, among the best fits, we favour the ones with the least amount of parameters.

To better understand the relationship among the number of persistent homology bars for a particular homological dimension in a specific space and N, we conducted simulations that showed that, on average, the number of bars increases linearly as N increases.

[1]Refer to the `scipy` documentation link provided in the appendix for additional information.

Next, we separately studied the distribution of birth and death times for each homological dimension in each space (unit interval, unit cube, and unit square). We used histograms to visualize the distribution and found that different distributions seem to follow. To determine the best fit, we used the maximum likelihood function (MLE) to estimate the parameters of a wide range of distributions and then reduced the sum square error (SSE) to find the best fit. The SSE was further adjusted using the Freedman-Diaconis rule (except for the 0-dimensional persistent homology for the unit interval). Our analysis showed that the distributions with the lowest SSE are almost always the same for a high N, and we calculated the estimated parameters of the best-fit distribution. We found that the scale and location parameters decrease as the number of points increase, while other parameters appear to be constant or converge asymptotically towards a constant. In this chapter, we present models describing the relationships between the parameters of fitted distributions and the number of points N uniformly distributed in the spaces. In order to achieve this goal, several models have been considered and analyzed, including linear, logarithmic, and exponential models[2]. Among the various models used, the overall best model appears to be the power model, which assumes that a relationship is given by a power law (also known as the Freundlich function):

$$y = ax^b \tag{4.1}$$

where y represents the value of a parameter, x will represents the number of points N, a is a scaling factor, and b is the power law exponent. However, in some cases, this model had to be further adjusted by adding additional constant terms. Additionally, we observed that solving (4.1) provides similar results to the case in which we optimize a while $b = -\frac{1}{d}$ where $d \in \mathbb{N}$ is the dimension of the space.

4.5 Unit Interval

4.5.1 Homological Dimension 0

Death Times

The analysis presented in Figure 4.10 outlines that several distributions resemble the shape of the histograms of the Death times of the H_0 PH bars of uniformly distributed points in $[0, 1]$. The Exponential distribution is often a special case of these distributions, as we will

[2]The `curvefit` function in Python was mainly used to estimate the values of the model's parameters. This function uses a non-linear least squares optimization method to fit a model to the data. The optimization method minimizes the difference between the observed data and the model predictions by adjusting the values of the model parameters.

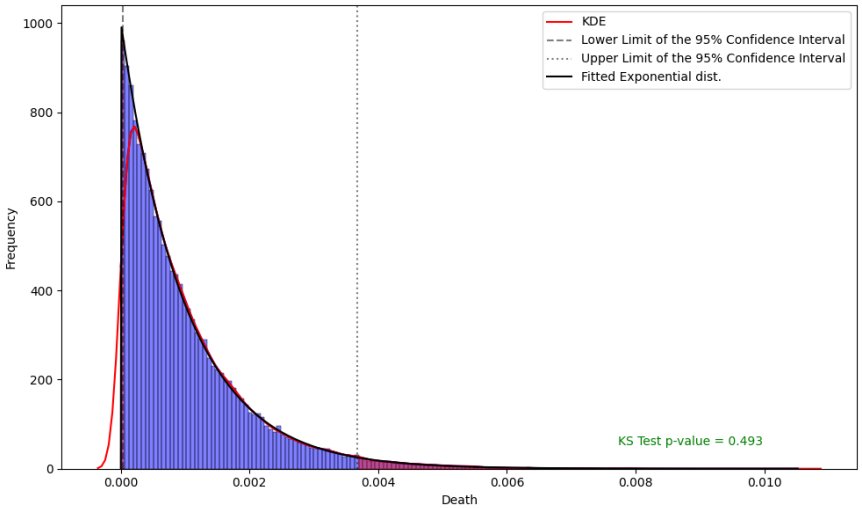


Figure 4.9: Fitted distribution on the normalized histogram of the Death times of the H_0 PH bars in $[0, 1]$ ($N = 1000$; 40 simulations).

N = 100

Distribution	SSE	AIC	BIC
Gamma	82.27464	-40.4335	-15315.9
Weibull	83.61608	-43.4295	-15251.9
Gen. Gamma	86.69681	-44.4933	-15100.3
Beta	87.0866	-35.5147	-15082.6
Gen. Normal	87.8061	-44.9772	-15058.3
Pearson Type III	88.44815	-43.4077	-15029.4
Exp. Weibulll	89.76191	-35.1014	-14962.7
Pareto	99.04239	-44.1543	-14581.4
Gen. Pareto	99.98706	-43.7486	-14543.8
Exp. Mod. Gauss.	109.9881	-36.0518	-14166.3
Gen. Exponential	111.8961	-32.7245	-14081.6
Exponential	111.9276	-38.7395	-14105.4
Gompertz	3507.582	447.1125	-455.564
Beta Prime	12002.75	-160.349	4424.347

N = 200

Distribution	SSE	AIC	BIC
Weibull	317.9849	-30.9116	-25605.7
Exponential	318.7164	-31.4849	-25596.4
Pareto	318.7164	-29.485	-25587.4
Gamma	319.8634	-32.1879	-25558.8
Beta	319.9243	-13.9918	-25548.3
Exp. Mod. Gauss.	324.416	-24.8036	-25446.3
Gen. Normal	325.1382	-23.0469	-25428.6
Gen. Exponential	325.4766	-10.9752	-25402.4
Pearson Type III	397.4201	-56.536	-23830.7
Gompertz	468.6561	203.0493	-22518.3
Beta Prime	1649.017	-199.33	-12495.1
Gen. Pareto	6719.39	-267.708	-1321.73
Gen. Gamma	75484.38	-314.616	17941.92
Exp. Weibulll	83331.71	-283.778	18729.19

N = 300

Distribution	SSE	AIC	BIC
Beta	579.1617	-259.343	-36174.2
Gen. Exponential	610.5292	-251.742	-35534
Exponential	621.3224	-272.211	-35352.6
Pareto	621.3226	-270.211	-35343.2
Gamma	623.5709	-268.947	-35300
Pearson Type III	623.6812	-268.934	-35297.9
Weibull	639.1058	-265.889	-35005.7
Exp. Mod. Gauss.	639.4526	-269.013	-34999.2
Exp. Weibulll	12005.34	-10.8133	82.8099
Gen. Pareto	17614.36	-387.283	4658.446
Gompertz	33100.26	320.0738	12203.13
Gen. Normal	115768.4	773.2538	27177.66
Gen. Gamma	213379.8	-344.726	34500.37
Beta Prime	216248.9	-338.337	34660.11

N = 400

Distribution	SSE	AIC	BIC
Gamma	1742.348	-260.811	-35320
Weibull	1921.331	-267.276	-33759.4
Beta	2008.468	-217.345	-33041.8
Gen. Exponential	2088.815	-266.826	-32406.1
Exponential	2103.381	-275.357	-32324.2
Pareto	2103.382	-273.357	-32314.5
Exp. Mod. Gauss.	2124.431	-270.569	-32155.6
Pearson Type III	3528.849	-290.986	-24056.4
Beta Prime	4349.392	-397.237	-20710.1
Gen. Pareto	39419.24	-503.083	14459.56
Gompertz	73653.66	1172.36	24436.47
Gen. Normal	248584	2278.977	43850.33
Gen. Gamma	460704.7	-474.677	53706.95
Exp. Weibulll	465540.7	-471.14	53873.61

N = 500

Distribution	SSE	AIC	BIC
Exp. Weibulll	1485.731	-64.8341	-51812.9
Gen. Gamma	1498.106	-62.9431	-51647.4
Pearson Type III	1566.433	-42.2497	-50767.1
Exp. Mod. Gauss.	1603.283	-44.732	-50303
Exponential	1605.505	-45.5362	-50285.2
Gen. Exponential	1605.562	-39.4885	-50254.8
Gompertz	1610.868	-40.9734	-50208.8
Weibull	1636.137	-45.3549	-49898.1
Gen. Normal	1694.881	-52.708	-49194
Pareto	1709.814	-74.6187	-49018.9
Beta	1798.219	14.54414	-48002.8
Gen. Pareto	63187	-601.204	23030.98
Beta Prime	769779	-656.293	72940.98
Gamma	799140	-635.424	73678.23

N = 600

Distribution	SSE	AIC	BIC
Gen. Normal	3139.09	-262.966	-48667.3
Pareto	3156.739	-263.213	-48533
Gen. Gamma	3182.662	-258.602	-48326.9
Exp. Weibulll	3207.67	-258.664	-48139.4
Weibull	3255.75	-257.367	-47793
Exponential	3264.433	-248.864	-47739.3
Gen. Exponential	3265.438	-242.926	-47701.7
Exp. Mod. Gauss.	3275.288	-247.987	-47649.7
Gamma	3326.666	-253.15	-47276.7
Beta	3454.897	-233.741	-46360.4
Pearson Type III	4010.428	-226.972	-42798
Gompertz	13919.19	-102.67	-12982.8
Gen. Pareto	85333.74	-695.438	30463.86
Beta Prime	340920	-692.307	63660.42

N = 700

Distribution	SSE	AIC	BIC
Exp. Weibulll	3814.236	-498.838	-55656.3
Gen. Gamma	3867.668	-480.994	-55267.4
Gamma	4031.964	-507.299	-54114.4
Pearson Type III	4038.123	-506.916	-54071.7
Weibull	4131.582	-508.136	-53432
Gen. Normal	4284.876	-505.945	-52413.4
Pareto	4331.652	-505.189	-52109.8
Exponential	4347.962	-504.868	-52015
Exp. Mod. Gauss.	4351.129	-499.127	-51984.4
Gen. Exponential	4347.973	-498.867	-51984.2
Gen. Pareto	121754.5	-725.706	41166.36
Beta Prime	170305.7	-742.762	50559.65
Beta	272708.1	-29.6387	63723.41
Gompertz	275181.6	748.5241	63965.62

N = 800

Distribution	SSE	AIC	BIC
Gen. Exponential	3504.023	-522.716	-70598.1
Weibull	3504.448	-528.35	-70614.9
Pearson Type III	3532.997	-530.258	-70355.6
Gamma	3537.914	-530.392	-70311.2
Beta	3712.92	-506.519	-68757.7
Exponential	3761.196	-534.986	-68365.6
Pareto	3761.202	-532.986	-68355.2
Exp. Mod. Gauss.	3903.793	-529.51	-67165.9
Beta Prime	28376.1	-701.769	-3759.76
Gen. Pareto	185196	-822.461	56182.58
Exp. Weibulll	297373.7	-249.839	71328.41
Gompertz	354855.5	1301.396	76966.05
Gen. Normal	1279277	2903.327	117949.6
Gen. Gamma	2383717	-760.185	137850.8

N = 900

Distribution	SSE	AIC	BIC
Exp. Weibulll	6240.35	-522.684	-62937.3
Gen. Gamma	6248.549	-522.435	-62890.1
Pearson Type III	6416.246	-521.033	-61948.3
Weibull	6479.76	-520.664	-61594
Exp. Mod. Gauss.	6504.704	-521.938	-61455.9
Gen. Exponential	6505.604	-517.604	-61429.9
Exponential	6505.708	-523.579	-61460.8
Gen. Normal	6522.342	-522.285	-61358.5
Pareto	6525.446	-524.961	-61341.4
Beta	8449.898	-462.778	-52037.2
Beta Prime	21207.34	-726.988	-18947.1
Gen. Pareto	238599.5	-906.508	68081.42
Gompertz	496560.7	2239.418	94437.21
Gamma	3279843	-838.053	162324.1

N = 1000

Distribution	SSE	AIC	BIC
Gen. Exponential	3511.48	-505.717	-97123.4
Beta	3513.542	-503.543	-97110.5
Weibull	3633.174	-532.939	-95783.2
Gamma	3704.983	-531.198	-95001.1
Pearson Type III	3705.496	-531.189	-94995.6
Exponential	3904.579	-538.594	-92914.9
Pareto	3904.583	-536.595	-92904.3
Exp. Mod. Gauss.	4433.501	-528.518	-87827.8
Beta Prime	40948.58	-791.163	1018.929
Gen. Gamma	241535.9	631.5257	71935.98
Gen. Pareto	305889.4	-989.402	81364.16
Gompertz	583485.1	3126.245	107170.1
Gen. Normal	2099400	6376.03	158334.4
Exp. Weibulll	4139420	-938.215	185474

Figure 4.10: BIC and AIC for the fitted distributions with the lowest SSE on the Death times of the H_0 PH bars for $N \in \{100, 200, 300, 400, 500, 600, 700, 800, 900, 1000\}$ points in $[0, 1]$.

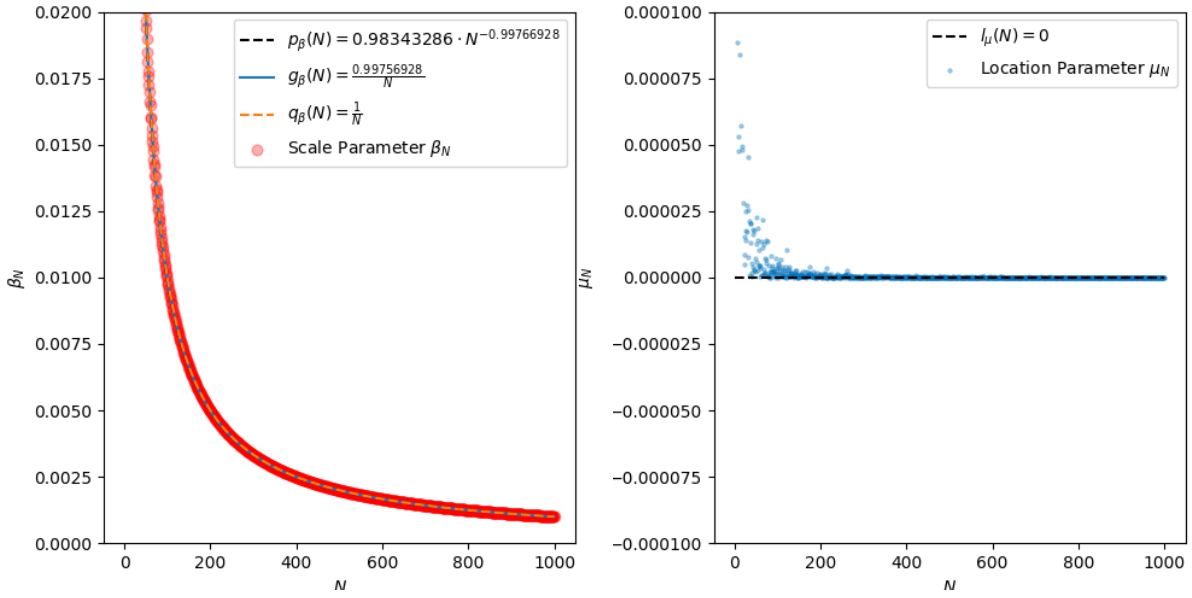
see later in the analysis. Additionally, the Exponential Distribution has the lowest number of parameters and is one of the most common statistical distributions. Moreover, it provides an excellent fit to the data while avoiding overfitting (Figure 4.9). It is among the models that provide a low BIC and AIC, which are measures of model goodness-of-fit and complexity. A lower BIC and AIC indicate a better trade-off between model fit and complexity, and the Exponential Distribution meets this criterion well. The probability distribution of the Exponential distribution [38] is defined as

$$f(x; \beta, \mu) = \frac{1}{\beta} e^{-\frac{x-\mu}{\beta}} \quad \text{for} \quad x \geq \mu; \beta > 0 \tag{4.2}$$

where β is the scale parameter and μ is the location parameter [38]. We estimated parameters β and μ for all $N \in [2, 1000]$ where $N \in \mathbb{N}$ and denoted them by β_N and μ_N respectively where the subscript N is used to denote the number of points uniformly distributed in the space (this convention will be used throughout the rest of this thesis). We created scatter plots (Figure 4.11) to understand better the relationship between an estimated parameter and N; a visual inspection of these plots indicates that β decreases as N increases. This pattern is expected since the H_0 PH bars represent the number of connected components which decrease during the filtration, and adding points to the space decreases the average distance among all the points, which lead leads to a 'quicker' death of the H_0 features that appear during the filtration. After conducting a thorough analysis of the data, we determined that the model that best describes the relationship between N and β is

$$g_\beta(N) = \frac{0.9975}{N} \approx \frac{1}{N}. \tag{4.3}$$

On the other hand, the values of the location parameter tend to be smaller than 0.000002



(a) (b)

Figure 4.11: Fitted models on the scatter plots of the estimated scale β (a) and location parameter β (b) of the exponential distribution fit for the Death times of the H_0 PH bars of $N \in [2, 1000]$ points uniformly distributed in $[0, 1]$.

for $N > 200$, as we can see from Figure 4.11b.

Let $f_N(\delta) = f(\delta; \beta_N, \mu_N)$ denote exponential distribution fit for N points where β_N and μ_N are the estimated parameters using the MLE method; additionally, let δ indicate x for the sole purpose of specifying that x represents the death parameter. Then, the following model

approximates the probability density of the Death times of the H_0 PH bars of N uniformly distributed points in $[0, 1]$:

$$f_N(\delta) \approx \tilde{f}_N(\delta; N) = \frac{N}{0.9975} e^{-\frac{N\delta}{0.9975}} \approx (N) e^{N\delta} \quad \text{for} \quad \delta \geq 0; N \in \mathbb{N}. \tag{4.4}$$

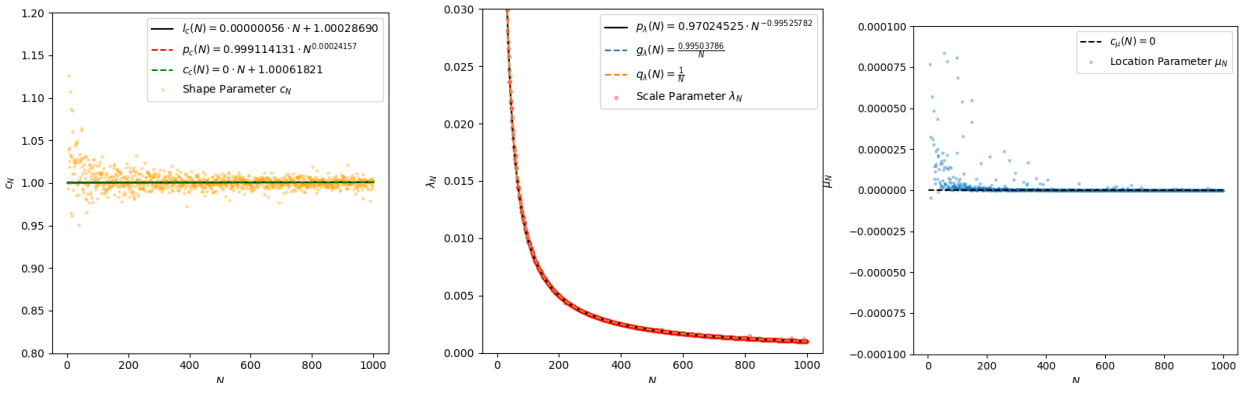
Additionally, it is worth noting that among the other distributions that fit the data well, there is the Weibull distribution, which we will present later in this thesis. The probability density function of the Weibull distribution is given by

$$f(y; c) = \frac{c}{\lambda} \left(\frac{y - \mu}{\lambda} \right)^{c-1} e^{-\left(\frac{y-\mu}{\lambda}\right)^c}, \tag{4.5}$$

where λ is the scale parameter, μ is the location parameter, and c is the shape parameter [38]. In this context, we can simplify the Weibull distribution by removing the location and scale parameters, which gives us the standard probability density function of the Weibull distribution:

$$f(x; c) = c x^{c-1} e^{-x^c}, \tag{4.6}$$

where c is the shape parameter. When $c = 1$, the Weibull distribution simplifies to the standard exponential distribution with a scale parameter equal to 1. It is worth noting that



(a) (b) (c)

Figure 4.12: Fitted models on the scatter plots of the estimated shape c (a), scale λ (b), and location μ (c) parameters of the Weibull distribution fit for the Death times of the H_0 PH bars of $N \in [2, 1000]$ points uniformly distributed in $[0, 1]$.

the estimated parameters of the shifted/scaled Weibull distribution fit on the Death times of the 0-dimensional persistent homology bars of uniformly distributed points in $[0, 1]$ (Figure 4.12) suggest that the c_N values lie close to 1, the location parameters μ_N are close to 0, and the scale parameters λ can be described by the relationship $\lambda_N = \frac{1}{N}$, in a similar manner to

the estimated scale parameters β_N of the exponential distribution; these observations further support the choice of the exponential distribution as a suitable model for the data.

4.6 Unit Square

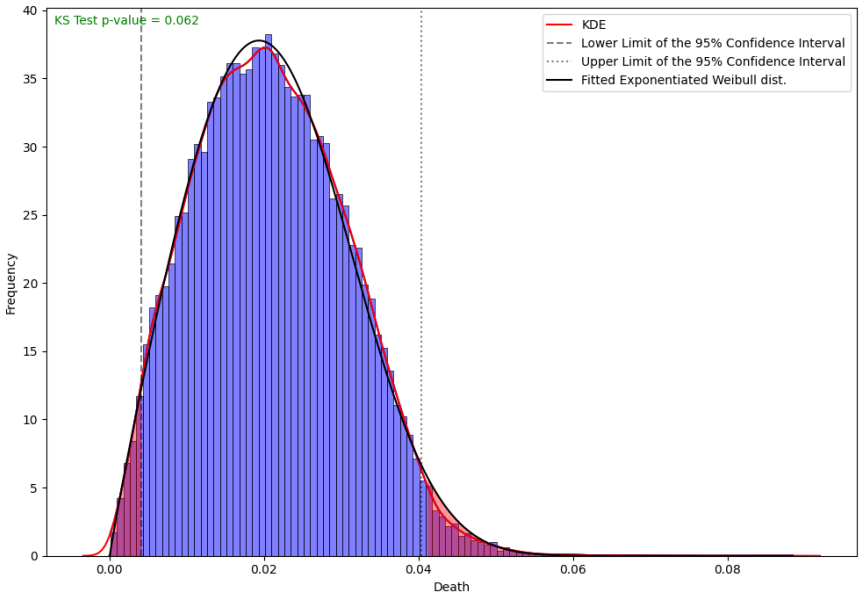


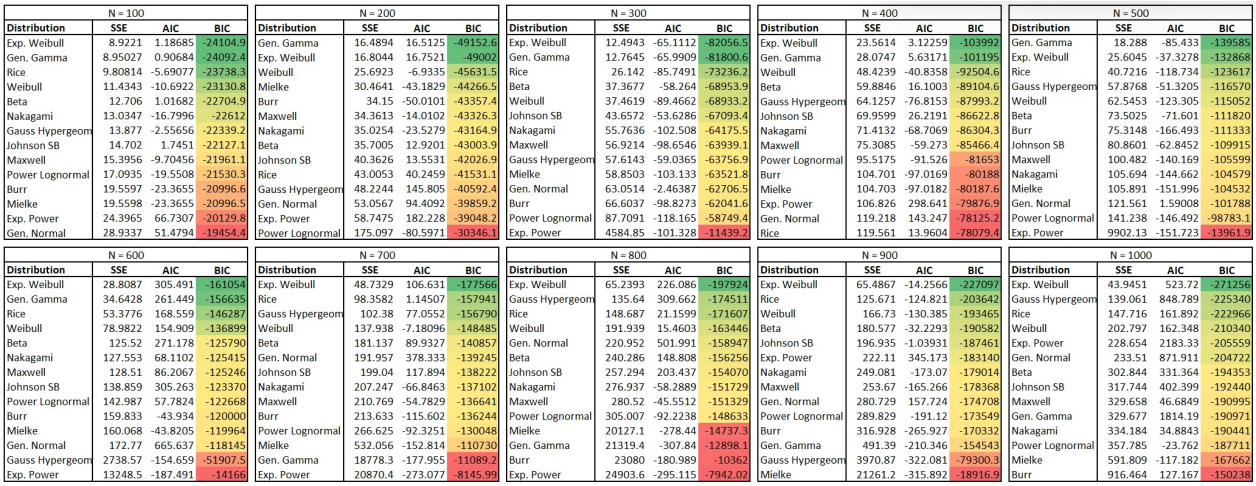
Figure 4.13: Fitted distribution on the normalized histogram of the Death times of the H_0 PH bars in $[0,1]^2$ ($N = 1000$; 40 simulations).

4.6.1 Unit Square: Homological Dimension 0

Death Times

The theoretical distribution of the Death times of the 0-dimensional persistent homology bars in $[0,1]^2$ seems to differ from that of the 0-dimensional persistent homology bars in $[0,1]$. Indeed, we can see from Figure 4.13 that the Exponentiated Weibull distribution fits well the given data. Furthermore, by comparing the sum of squared error (SSE) and the Bayesian information criterion (BIC) across multiple distributions from Figure 4.14, it is clear that the Exponentiated Weibull distribution has the lowest values for both measures, which suggests that this distribution is an appropriate model for the underlying probability distribution of the data. The probability density function for the Exponentiated Weibull distribution [28] with shape parameters α and β, and scale parameter λ is given by:

$$f(x;\alpha,\beta,\lambda) = \frac{\alpha\beta}{\lambda}\left[\frac{x}{\lambda}\right]^{\beta-1}\left[1-e^{-(x/\lambda)^{\alpha}}\right]^{\beta-1} e^{-(x/\lambda)^{\beta}} \quad \text{for} \quad x,\alpha,\beta,\lambda > 0, \tag{4.7}$$



N = 100			
Distribution	**SSE**	**AIC**	**BIC**
Exp. Weibull	8.9221	1.18685	-24104.9
Gen. Gamma	8.95027	0.90684	-24092.4
Rice	9.80814	-5.69077	-23738.3
Weibull	11.4343	-10.6922	-23130.8
Beta	12.706	1.01682	-22704.9
Nakagami	13.0347	-16.7996	-22612
Gauss Hypergeom	13.877	-2.55656	-22339.2
Johnson SB	14.702	1.7451	-22127.1
Maxwell	15.3956	-9.70456	-21961.1
Power Lognormal	17.0935	-19.5508	-21530.3
Burr	19.5597	-23.3655	-20996.6
Mielke	19.5598	-23.3655	-20996.5
Exp. Power	24.3965	66.7307	-20129.8
Gen. Normal	28.9337	51.4794	-19454.4

N = 200			
Distribution	**SSE**	**AIC**	**BIC**
Gen. Gamma	16.4894	16.5125	-49152.6
Exp. Weibull	16.8044	16.7521	-49002
Weibull	25.6923	-6.9335	-45631.5
Mielke	30.4641	-43.1829	-44266.5
Burr	34.15	-50.0101	-43357.4
Maxwell	34.3613	-14.0102	-43326.3
Nakagami	35.0254	-23.5279	-43164.9
Beta	35.7005	12.9201	-43003.9
Johnson SB	40.3626	13.5531	-42026.9
Rice	43.0053	40.2459	-41531.1
Gauss Hypergeom	48.2244	145.805	-40592.4
Gen. Normal	53.0567	94.4092	-39859.2
Exp. Power	58.7475	182.228	-39048.2
Power Lognormal	175.097	-80.5971	-30346.1

N = 300			
Distribution	**SSE**	**AIC**	**BIC**
Exp. Weibull	12.4943	-65.1112	-82056.5
Gen. Gamma	12.7645	-65.9909	-81800.6
Rice	26.142	-85.7491	-73236.2
Beta	37.3677	-58.264	-68953.9
Weibull	37.4619	-89.4662	-68933.2
Johnson SB	43.6572	-53.6286	-67093.4
Nakagami	55.7636	-102.508	-64175.5
Maxwell	56.9214	-98.6546	-63939.1
Gauss Hypergeom	57.6143	-59.0365	-63756.9
Mielke	58.8503	-103.133	-63521.8
Gen. Normal	63.0514	-2.46387	-62706.5
Burr	66.6037	-98.8273	-62041.6
Power Lognormal	87.7091	-118.165	-58749.4
Exp. Power	4584.85	-101.328	-11439.2

N = 400			
Distribution	**SSE**	**AIC**	**BIC**
Exp. Weibull	23.5614	3.12259	-103992
Gen. Gamma	28.0747	5.63171	-101195
Weibull	48.4239	-40.8358	-92504.6
Beta	59.8846	16.1003	-89104.6
Gauss Hypergeom	64.1257	-76.8153	-87993.2
Johnson SB	69.9599	26.2191	-86622.8
Nakagami	71.4132	-68.7069	-86304.3
Maxwell	75.3085	-59.273	-85466.4
Power Lognormal	95.5175	-91.526	-81653
Burr	104.701	-97.0169	-80188
Mielke	104.703	-97.0182	-80187.6
Exp. Power	106.826	298.641	-79876.9
Gen. Normal	119.218	143.247	-78125.2
Rice	119.561	13.9604	-78079.4

N = 500			
Distribution	**SSE**	**AIC**	**BIC**
Gen. Gamma	18.288	-85.433	-139585
Exp. Weibull	25.6045	-37.3278	-132868
Rice	40.7216	-118.734	-123617
Gauss Hypergeom	57.8768	-51.3205	-116570
Weibull	62.5453	-123.305	-115052
Beta	73.5025	-71.601	-111820
Burr	75.3148	-166.493	-111333
Johnson SB	80.8601	-62.8452	-109915
Maxwell	100.482	-140.169	-105599
Nakagami	105.694	-144.662	-104579
Mielke	105.891	-151.996	-104532
Gen. Normal	121.561	1.59008	-101788
Power Lognormal	141.238	-146.492	-98783.1
Exp. Power	9902.13	-151.723	-13961.9

N = 600			
Distribution	**SSE**	**AIC**	**BIC**
Exp. Weibull	28.8087	305.491	-161054
Gen. Gamma	34.6428	261.449	-156635
Rice	53.3776	168.559	-146287
Weibull	78.9822	154.909	-136899
Beta	125.52	271.178	-125790
Nakagami	127.553	68.1102	-125415
Maxwell	128.51	86.2067	-125246
Johnson SB	138.859	305.263	-123370
Power Lognormal	142.987	57.7824	-122668
Burr	159.833	-43.934	-120000
Mielke	160.068	-43.8205	-119964
Gen. Normal	172.77	665.637	-118145
Gauss Hypergeom	2738.57	-154.659	-51907.5
Exp. Power	13248.5	-187.491	-14166

N = 700			
Distribution	**SSE**	**AIC**	**BIC**
Exp. Weibull	48.7329	106.631	-177566
Rice	98.3582	1.14507	-157941
Gauss Hypergeom	102.38	77.0552	-156790
Weibull	137.938	-7.18096	-148485
Beta	181.137	89.9327	-140857
Gen. Normal	191.957	378.333	-139245
Johnson SB	199.04	117.894	-138222
Nakagami	207.247	-66.8463	-137102
Maxwell	210.769	-54.7829	-136641
Burr	213.633	-115.602	-136244
Power Lognormal	266.625	-92.3251	-130048
Mielke	532.056	-152.814	-110730
Gen. Gamma	18778.3	-177.955	-11089.2
Exp. Power	20870.4	-273.077	-8145.99

N = 800			
Distribution	**SSE**	**AIC**	**BIC**
Exp. Weibull	65.2393	226.086	-197924
Gauss Hypergeom	135.64	309.662	-174511
Rice	148.687	21.1599	-171607
Weibull	191.939	15.4603	-163446
Gen. Normal	220.952	501.991	-158947
Beta	240.286	148.808	-156256
Johnson SB	257.294	203.437	-154070
Nakagami	276.937	-58.2889	-151729
Maxwell	280.52	-45.5512	-151329
Power Lognormal	305.007	-92.2238	-148633
Mielke	20127.1	-278.44	-14737.3
Gen. Gamma	21319.4	-307.84	-12898.1
Burr	23080	-180.989	-10362
Exp. Power	24903.6	-295.115	-7942.02

N = 900			
Distribution	**SSE**	**AIC**	**BIC**
Exp. Weibull	65.4867	-14.2566	-227097
Rice	125.671	-124.821	-203642
Weibull	166.73	-130.385	-193465
Beta	180.577	-32.2293	-190582
Johnson SB	196.935	-1.03931	-187461
Exp. Power	222.11	345.173	-183140
Nakagami	249.081	-173.07	-179014
Maxwell	253.67	-165.266	-178368
Gen. Normal	280.729	157.724	-174708
Power Lognormal	289.829	-191.12	-173549
Burr	316.928	-265.927	-170332
Gen. Gamma	491.39	-210.346	-154543
Gauss Hypergeom	3970.87	-322.081	-79300.3
Mielke	21261.2	-315.892	-18916.9

N = 1000			
Distribution	**SSE**	**AIC**	**BIC**
Exp. Weibull	43.9451	523.72	-271256
Gauss Hypergeom	139.061	848.789	-225340
Rice	147.716	161.892	-222966
Weibull	202.797	162.348	-210340
Exp. Power	228.654	2183.33	-205559
Gen. Normal	233.51	871.911	-204722
Beta	302.844	331.364	-194353
Johnson SB	317.744	402.399	-192440
Maxwell	329.658	46.6849	-190995
Gen. Gamma	329.677	1814.19	-190971
Nakagami	334.184	34.8843	-190441
Power Lognormal	357.785	-23.762	-187711
Mielke	591.809	-117.182	-167662
Burr	916.464	127.167	-150238

Figure 4.14: BIC and AIC for the fitted distributions with the lowest SSE on the Death times of the H_0 PH bars of $N \in \{100, 200, 300, 400, 500, 600, 700, 800, 900, 1000\}$ points in a $[0,1]^2$.

and the probability density function for $y = x - \mu$ can be expressed in terms of x and including the location parameter μ as

$$f(x;\alpha,\beta,\mu,\lambda) = \alpha\frac{\beta}{\lambda}\left[\frac{x-\mu}{\lambda}\right]^{\beta-1}\left[1-e^{-(\frac{x-\mu}{\lambda})^{\beta}}\right]^{\alpha-1}e^{-(\frac{x-\mu}{\lambda})^{\beta}} \quad \text{for} \quad x \geq \mu. \tag{4.8}$$

The parameters α, β, λ and μ have been estimated for all $N \in [2, 1000]$ where $N \in \mathbb{N}$ and we denote them with α_N,β_N, λ_N and μ_N respectively. The values of the estimated parameters α and β decrease and increase, respectively, as N increases (Figure 4.15). The location and scale parameters (Figure 4.16) follow a trend similar to the location and scale parameters of the Exponential and Weibull fits on the Death times of the H_0 PH bars in a unit interval. Specifically, the estimation of the location parameter assigns relatively small values close to 0, and a power model also describes the scale parameter well. The fitted power models and their approximations for α, β, and λ are:

$$p_\alpha(N) = 1.2454849 \cdot N^{-0.09641111} \approx 1.24 \cdot N^{-0.097}, \tag{4.9}$$

$$p_\beta(N) = 1.68228249 \cdot N^{0.08239274} \approx 1.68 \cdot N^{0.082}, \tag{4.10}$$

$$p_\lambda(N) = 0.75371809 \cdot N^{-0.47960528} \approx 0.75 \cdot N^{-0.480}, \tag{4.11}$$

or alternatively

$$g_\lambda(N) = \frac{0.85279701}{\sqrt{N}} \approx \frac{0.853}{\sqrt{N}}. \tag{4.12}$$

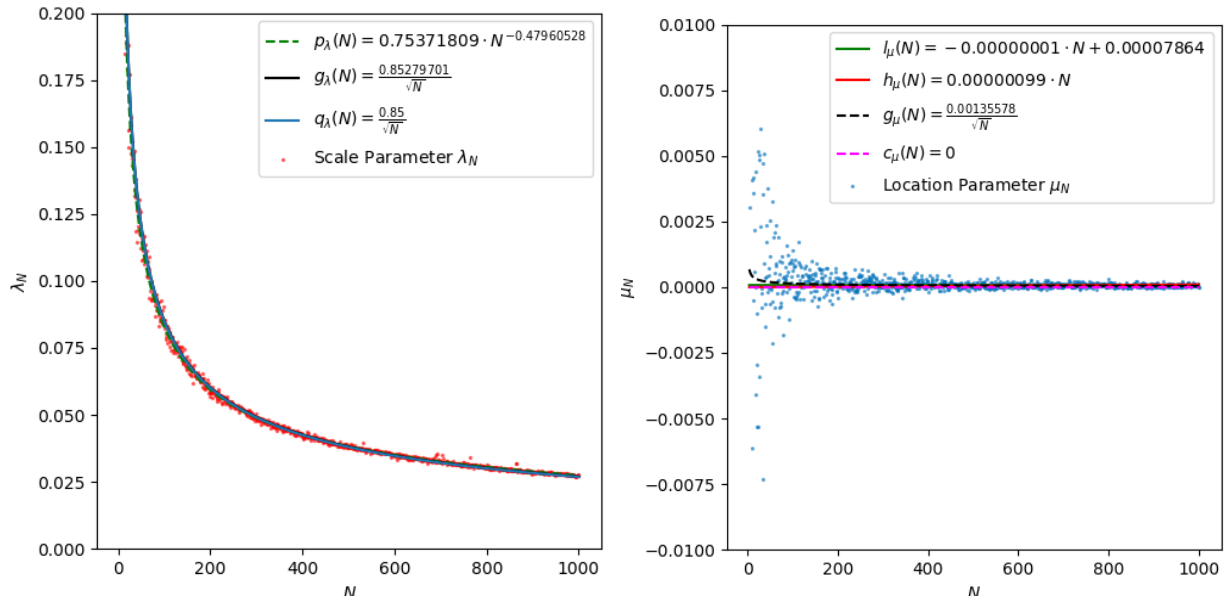
A linear model is more appropriate for the location parameter:

$$l_\mu(N) = -0.00000001 \cdot N + 0.00007864, \tag{4.13}$$

or a constant function of the form

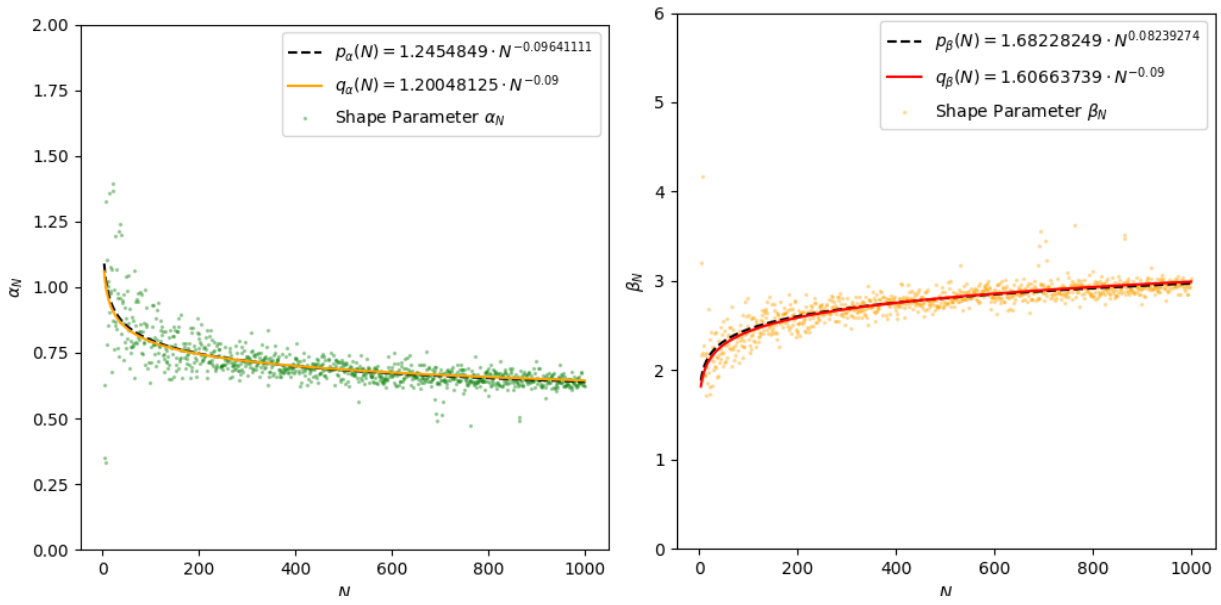
$$h_\mu(N) = 0.00000099 \cdot N. \tag{4.14}$$

However, it should be noted that while the estimated shape parameters α and β appear to converge towards asymptotic limits of 0.6 and 3, respectively, it is difficult to predict with certainty since the data collected is limited to $N = 1000$. Additionally, while power models were used to determine the trends of these parameters, other asymptotic models may provide a better fit. Therefore, the results obtained are inconclusive. Further investigation may be necessary to determine the behaviour of the parameters for larger values of N and to select the most appropriate model for the data.



(a) (b)

Figure 4.15: Fitted models on the scatter plots of the estimated shape parameters α (a) and β (b) of the Exponentiated Weibull distribution fit for the Death times of the H_0 PH bars of $N \in [2, 1000]$ points uniformly distributed in $[0, 1]^2$.



(a) (b)

Figure 4.16: Fitted models on the scatter plots of the estimated scale λ (a) and location μ (b) parameters of the Exponentiated Weibull distribution fit for the Death times of the H_0 PH bars of $N \in [2, 1000]$ points uniformly distributed in $[0, 1]^2$.

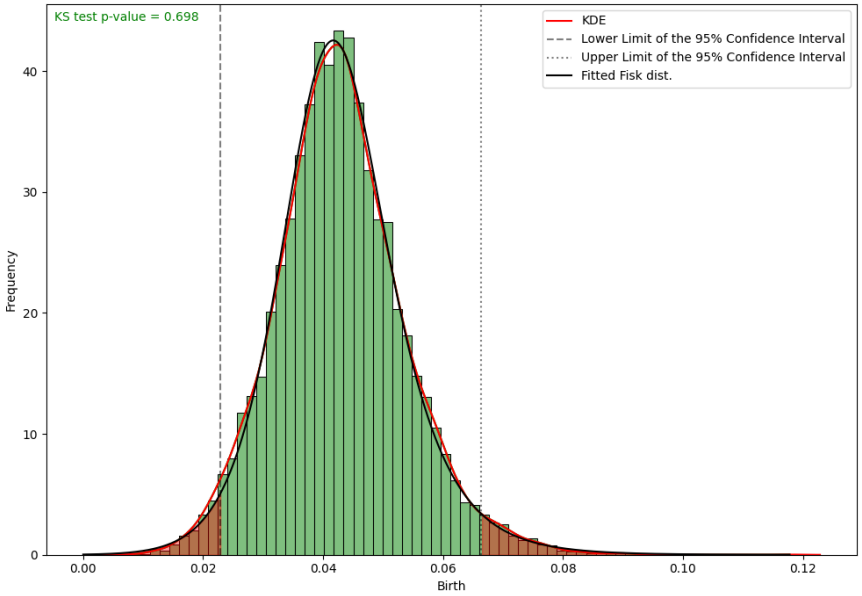


Figure 4.17: Fitted distribution on the normalized histogram of the Birth times of the H_1 PH bars in $[0, 1]^2$ ($N = 1000$; 40 simulations).

4.6.2 Unit Square: Homological Dimension 1

Birth Times

Upon analysis, we found that the Fisk and Mielke distributions have the smallest Sum of Squared Errors (SSE) and the Bayesian Information Criterion (BIC) values (Figure 4.18).

N = 100			
Distribution	**SSE**	**AIC**	**BIC**
Fisk	7.80764	-13.8181	-3588.1
Burr	7.8988	-11.6728	-3572.35
Mielke	8.32067	-10.0695	-3531.61
Gen. Logistic	8.54221	-11.6175	-3517.69
Johnson's SU	8.91155	-9.07199	-3477.89
Exp. Mod. Gauss.	10.4562	-9.13281	-3359.39
Logistic	13.7057	-10.0365	-3154.16
Tukey-Lambda	13.5931	-7.97184	-3153.95
Student's t	14.2728	-7.45331	-3115.75
Double Weibull	14.7108	-11.0595	-3092.08
Alpha	15.2212	-3.63452	-3065.38
Exp. Weibull	15.129	-1.77649	-3063.47
Inverse Gamma	15.6556	-3.08369	-3043.35
Power Log-Norm.	15.8726	-0.65965	-3025.9

N = 200			
Distribution	**SSE**	**AIC**	**BIC**
Mielke	16.6077	-48.1249	-8047.3
Fisk	16.9085	-49.9368	-8023.57
Gen. Logistic	17.3244	-48.5962	-7981.37
Burr	17.3407	-46.3094	-7972.27
Exp. Mod. Gauss.	18.9561	-43.8746	-7825.02
Johnson's SU	19.303	-40.9655	-7786.06
Logistic	19.647	-46.1717	-7770.3
Tukey-Lambda	20.309	-38.097	-7705.28
Student's t	21.0876	-38.0708	-7639.93
Exp. Weibull	24.0032	-37.2279	-7407.52
Alpha	24.5304	-39.6919	-7377.24
Power Log-Norm.	24.4615	-37.2429	-7374.67
Inverse Gamma	25.0119	-39.7153	-7343.48
Double Weibull	25.3259	-43.961	-7321.81

N = 300			
Distribution	**SSE**	**AIC**	**BIC**
Fisk	31.5416	-63.5105	-12186.9
Mielke	31.8916	-56.5627	-12148.8
Gen. Logistic	33.5122	-58.7369	-12021.1
Exp. Mod. Gauss.	35.4449	-52.55	-11867.7
Johnson's SU	38.8099	-46.8456	-11611.6
Exp. Weibull	47.4586	-37.5835	-11061.2
Alpha	47.6173	-40.5531	-11059.9
Inverse Gamma	48.6704	-39.3194	-11000.1
Power Log-Norm.	48.8197	-36.3196	-10983.8
Logistic	49.6686	-52.424	-10952.5
Tukey-Lambda	52.5559	-41.3293	-10790
Student's t	54.449	-42.4127	-10693.1
Burr	158.37	-42.4511	-7764.08
Double Weibull	245.328	-49.744	-6574.54

N = 400			
Distribution	**SSE**	**AIC**	**BIC**
Fisk	21.0259	-11.4097	-19513.4
Gen. Logistic	21.5318	12.4628	-19423.8
Burr	21.6209	-13.0647	-19400
Mielke	23.8732	-11.6457	-19026.8
Johnson's SU	25.4315	29.6872	-18788.7
Exp. Mod. Gauss.	27.2901	22.3476	-18531.3
Alpha	47.0844	67.49	-16477.3
Inverse Gamma	48.9524	74.6167	-16330.7
Logistic	52.1399	49.0224	-16101.4
Tukey-Lambda	52.504	309.599	-16067
Student's t	56.7415	57.7977	-15774.7
Exp. Weibull	56.6896	98.431	-15769.9
Double Weibull	62.4195	56.3411	-15415.5
Power Log-Norm.	1750.72	-128.145	-2851.78

N = 500			
Distribution	**SSE**	**AIC**	**BIC**
Fisk	21.0259	-11.4097	-19513.4
Gen. Logistic	21.5318	12.4628	-19423.8
Burr	21.6209	-13.0647	-19400
Mielke	23.8732	-11.6457	-19026.8
Johnson's SU	25.4315	29.6872	-18788.7
Exp. Mod. Gauss.	27.2901	22.3476	-18531.3
Alpha	47.0844	67.49	-16477.3
Inverse Gamma	48.9524	74.6167	-16330.7
Logistic	52.1399	49.0224	-16101.4
Tukey-Lambda	52.504	309.599	-16067
Student's t	56.7415	57.7977	-15774.7
Exp. Weibull	56.6896	98.431	-15769.9
Double Weibull	62.4195	56.3411	-15415.5
Power Log-Norm.	1750.72	-128.145	-2851.78

N = 600			
Distribution	**SSE**	**AIC**	**BIC**
Mielke	44.3212	-116.397	-28176.7
Fisk	45.8038	-120.367	-27994.8
Gen. Logistic	54.7911	-114.768	-26957.5
Johnson's SU	59.4666	-110.866	-26474.7
Logistic	76.7052	-109.415	-25018.2
Tukey-Lambda	77.7673	-106.349	-24929.9
Student's t	83.3882	-105.907	-24525.8
Exp. Mod. Gauss.	83.6776	-104.929	-24505.8
Double Weibull	105.318	-113.624	-23174
Exp. Weibull	136.868	-86.1282	-21648.1
Alpha	138.604	-88.5583	-21583.8
Inverse Gamma	143.232	-87.0166	-21393.6
Power Log-Norm.	163.568	-74.4437	-20616.3
Burr	531.747	-88.8954	-13790.2

N = 700			
Distribution	**SSE**	**AIC**	**BIC**
Fisk	95.6096	-154.412	-29694
Mielke	98.4265	-152.959	-29483.7
Gen. Logistic	100.613	-150.843	-29340.2
Burr	105.826	-152.204	-28980.9
Johnson's SU	115.756	-145.248	-28358.7
Exp. Mod. Gauss.	142.181	-145.509	-26941.2
Logistic	159.131	-144.869	-26168.8
Tukey-Lambda	170.904	-137.971	-25664.8
Student's t	179.476	-137.939	-25325.3
Double Weibull	188.692	-144.56	-24978
Exp. Weibull	198.174	-124.625	-24629
Alpha	201.673	-135.772	-24516.4
Power Log-Norm.	207.387	-132.3	-24313.7
Inverse Gamma	211.958	-136.946	-24171.4

N = 800			
Distribution	**SSE**	**AIC**	**BIC**
Fisk	51.1723	-119.029	-40219.6
Mielke	51.1625	-118.278	-40212.2
Burr	63.1686	-118.272	-38531.7
Gen. Logistic	64.2187	-103.309	-38409.2
Johnson's SU	82.646	-94.9798	-36389.1
Exp. Mod. Gauss.	127.329	-90.83	-32952.5
Logistic	127.558	-81.0777	-32947.2
Tukey-Lambda	137.826	-68.0492	-32321
Student's t	154.217	-74.6923	-31425.2
Double Weibull	203.966	-84.7085	-29196.3
Exp. Weibull	211.029	-31.5326	-28915.9
Alpha	214.29	-50.3906	-28802.7
Power Log-Norm.	220.963	-43.8715	-28549.2
Inverse Gamma	222.23	-49.2008	-28512.6

N = 900			
Distribution	**SSE**	**AIC**	**BIC**
Fisk	83.1111	-126.368	-41880.2
Mielke	83.6575	-123.731	-41812.4
Gen. Logistic	102.792	-110.479	-39977
Johnson's SU	109.237	-110.316	-39423.3
Logistic	151.514	-93.5244	-36511.8
Tukey-Lambda	152.508	-92.4841	-36444.2
Student's t	166.089	-96.6592	-35680.2
Exp. Mod. Gauss.	188.655	-93.1675	-34539.4
Double Weibull	212.345	-101.168	-33480.1
Alpha	311.338	-49.8507	-30053.3
Power Log-Norm.	313.317	-38.1285	-29987.5
Inverse Gamma	314.619	-44.4688	-29959.5
Exp. Weibull	367.337	-11.9883	-28563.1
Burr	1096.07	-74.0159	-18773.4

N = 1000			
Distribution	**SSE**	**AIC**	**BIC**
Fisk	101.203	-115.916	-45569.1
Mielke	107.184	-117.067	-44989.2
Burr	107.184	-117.067	-44989.2
Gen. Logistic	113.829	-98.0296	-44400.5
Johnson's SU	131.016	-93.3371	-42993.5
Logistic	181.389	-81.1597	-39778.2
Tukey-Lambda	191.525	-67.8412	-39228.5
Student's t	212.904	-76.6703	-38176.6
Exp. Mod. Gauss.	234.176	-68.5623	-37230
Double Weibull	300.03	-86.5197	-34766.8
Alpha	348.486	-10.5139	-33278.6
Power Log-Norm.	351.912	1.56032	-33172.1
Inverse Gamma	369.438	-17.1872	-32698.2
Exp. Weibull	378.075	17.2648	-32459.3

Figure 4.18: BIC and AIC for the fitted distributions with the lowest SSE on the Birth times of the H_1 PH bars of $N \in \{100, 200, 300, 400, 500, 600, 700, 800, 900, 1000\}$ points in $[0, 1]^2$.

These results indicate that both distributions are good candidates for describing the underlying distribution of our data. However, the focus is placed on the Fisk distribution due to its simplicity and flexibility (Figure 4.17). The Fisk distribution is a well-known distribution in economics and is commonly used to model income and lifetime data; in most mathematical textbooks, it is referred to as the log-logistic distribution [22]. It has several parametrizations, and we present it in the form given by (4.15), where the shape parameter is denoted by c and the scale parameter by λ:

$$f(x; c, \lambda) = \frac{c}{\lambda} \frac{\left(\frac{x}{\lambda}\right)^{c-1}}{\left(1 + \left(\frac{x}{\lambda}\right)^{c}\right)^2} \quad \text{for} \quad x \geq 0, c > 0, \lambda > 0. \tag{4.15}$$

To include a location parameter μ, we reparametrize (4.15) as:

$$f(x; \alpha, \lambda, \mu) = \frac{c\left(\frac{x-\mu}{\lambda}\right)^{c-1}}{\lambda\left(1 + \left(\frac{x-\mu}{\lambda}\right)^{c}\right)^2} \quad \text{for} \quad x > \mu. \tag{4.16}$$

The parameters of the Fisk distribution are estimated on 999 datasets, each labelled by the number of points N uniformly distributed in the unit square (Figure 4.19). The analysis reveals that the shape parameter c_N converges to a limit of approximately 15, while the scale parameter λ_N and location parameter μ_N converge to 0 as N increases. More precisely, linear and power models are fit to the estimated parameters, and it is found that the shape parameter can be described by the linear function

$$l_c(N) = 0.000000915 \cdot N + 14.9961 \approx 15. \tag{4.17}$$

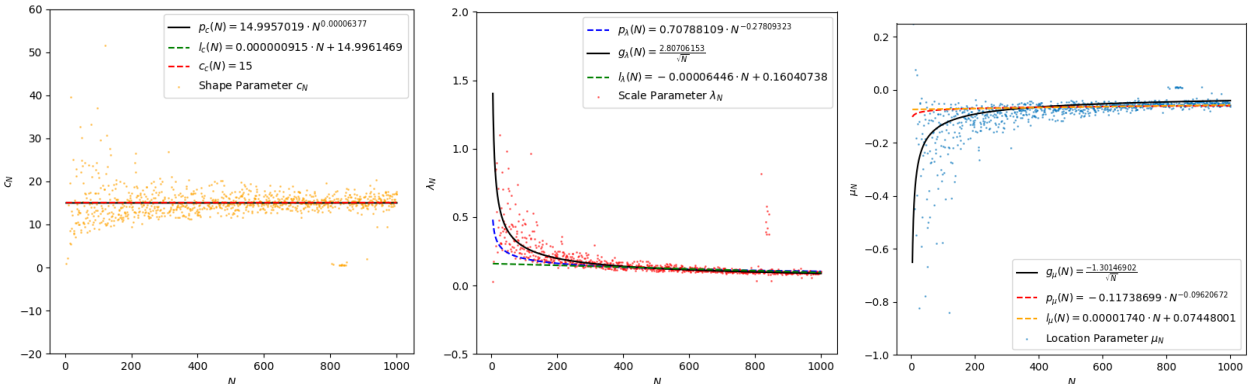
The scale parameter and location parameters are described by the power functions

$$p_\lambda(N) = 0.7079 \cdot N^{-0.2781} \quad \text{or} \quad g_\lambda(N) = \frac{2.8071}{\sqrt{N}} \approx \frac{2.81}{\sqrt{N}}, \tag{4.18}$$

and

$$g_\mu(N) = \frac{-1.30146902}{\sqrt{N}} \approx g_\mu(N) = \frac{-1.3}{\sqrt{N}}, \tag{4.19}$$

respectively. The analysis reveals that as N increases, the pdf of the Fisk distribution becomes more concentrated around its mode $\lambda(\frac{c-1}{c+1})^{1/c}$, which converges to 0. This is reflected in the decreasing values of the scale and location parameters as N increases. Additionally, the shape parameter converges to a limit of approximately 15, indicating that the theoretical distribution of the birth times of the 1-dimensional persistent homology bars maintains a constant shape as additional points are uniformly sampled from the unit square $[0, 1]^2$.



(a) (b) (c)

Figure 4.19: Fitted models on the scatter plots of the estimated shape c (a), scale λ (b), and location μ (c) parameters of the Fisk distribution fit for the Birth times of the H_1 PH bars of $N \in [2, 1000]$ points uniformly distributed in $[0, 1]^2$.

Death Times

Figure 4.21 presents a few distributions that provide similar values for the SSEs; in particular, the Generalized Gamma distribution has the lowest SSEs and BICs values and provides a good fit to the data (Figure 4.20). The Generalized Gamma distribution is a flexible distribution that is often used in the statistical literature. It has three subfamilies: the exponential, gamma, and Weibull distributions, and it can also be used to model lognormal distributions [31]. The probability density is given by

$$f(x; \alpha, \beta, \lambda) = \frac{\beta}{\lambda \Gamma(\alpha)} \left(\frac{x}{\lambda}\right)^{\alpha\beta - 1} e^{-\left(\frac{x}{\lambda}\right)^\beta} \quad \text{for} \quad x \geq 0, \alpha > 0, \beta > 0, \lambda > 0, \tag{4.20}$$

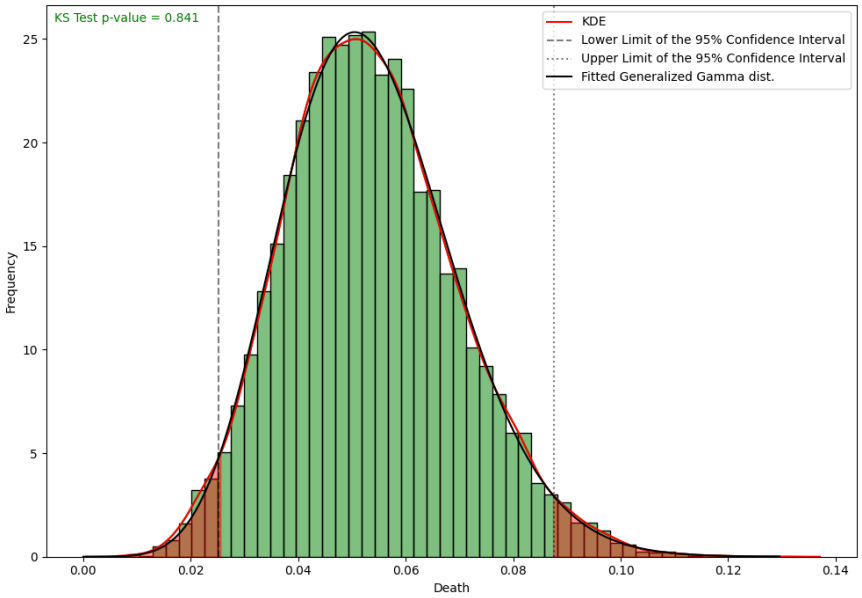


Figure 4.20: Fitted distribution on the normalized histogram of the Death times of the H_1 PH bars in $[0,1]^2$ ($N = 1000$; 40 simulations).

N = 100

Distribution	SSE	AIC	BIC
Exp. Weibull	5.158353	-16.8209	-4538.44
Gen. Gamma	5.28512	-17.1581	-4516.9
Gauss Hypergeom	5.268737	-13.1426	-4506.08
Chi	5.429849	-19.4285	-4499.73
Beta	5.525401	-16.7923	-4477.47
Johnson's SB	5.648379	-16.8927	-4457.94
Power Normal	5.758489	-19.49	-4447.61
Erlang	5.962575	-20.3648	-4416.71
Gamma	5.962581	-20.3648	-4416.71
Weibull Max	5.975005	-19.9039	-4414.87
Gen. Extreme Value	5.975412	-19.8955	-4414.81
Johnson's SU	6.007503	-18.3618	-4403.27
Exp. Normal	6.760364	-20.1557	-4305.33
Burr	7.535104	-19.5817	-4202.31
Mielke	7.535274	-19.5818	-4202.29

N = 200

Distribution	SSE	AIC	BIC
Gen. Gamma	4.178272	-27.545	-10444.3
Exp. Weibull	4.237291	-27.8288	-10419.9
Beta	4.511252	-26.98	-10311.1
Johnson's SB	4.66164	-27.0814	-10254.2
Gauss Hypergeom	4.646129	-21.9596	-10245
Chi	4.928507	-30.5789	-10164.9
Power Normal	5.087538	-30.4124	-10109.8
Erlang	5.624621	-31.6304	-9935.43
Gamma	5.624628	-31.6304	-9935.42
Johnson's SU	5.720154	-29.6498	-9898.71
Weibull Max	6.167419	-30.0494	-9775.4
Gen. Extreme Value	6.170461	-30.0559	-9774.55
Exp. Normal	6.192699	-32.0122	-9768.3
Mielke	8.288426	-32.6443	-9254.52
Burr	11.50906	-34.8373	-8684.31

N = 300

Distribution	SSE	AIC	BIC
Exp. Weibull	6.423993	-33.8301	-16723.2
Gen. Gamma	6.468193	-32.8587	-16704.3
Chi	6.565245	-32.3888	-16671.1
Beta	6.566067	-32.9656	-16662.8
Johnson's SB	6.637492	-33.1659	-16632.9
Gauss Hypergeom	6.609711	-27.8577	-16628.6
Gamma	6.716371	-38.3619	-16608.2
Erlang	6.716373	-38.3619	-16608.2
Power Normal	6.881941	-33.4833	-16540.9
Johnson's SU	6.944055	-37.6515	-16508.1
Gen. Extreme Value	7.463571	-35.1844	-16316.7
Weibull Max	7.469261	-35.1803	-16314.6
Exp. Normal	10.04513	-44.3225	-15496
Burr	16.49365	-45.3829	-14117.9
Mielke	16.49431	-45.383	-14117.8

N = 400

Distribution	SSE	AIC	BIC
Gen. Gamma	11.15378	-45.3959	-21892.7
Beta	11.2432	-44.9386	-21862.6
Exp. Weibull	11.32998	-45.6179	-21833.7
Johnson's SB	11.43261	-44.9304	-21799.7
Chi	11.62003	-49.2392	-21746.7
Power Normal	11.63474	-49.2778	-21741.9
Weibull Max	13.7883	-48.2466	-21102.4
Gen. Extreme Value	13.7942	-48.2506	-21100.8
Gauss Hypergeom	13.72087	-36.2028	-21096.1
Erlang	14.37219	-53.53	-20946.2
Gamma	14.3722	-53.53	-20946.2
Johnson's SU	15.16711	-51.9486	-20735.2
Exp. Normal	21.63294	-56.3358	-19406.2
Mielke	32.74484	-61.1774	-17836.8
Burr	41.73468	-65.7528	-16923.2

N = 500

Distribution	SSE	AIC	BIC
Gen. Gamma	12.40167	-62.9209	-29424.4
Beta	12.56968	-62.1565	-29358.2
Chi	12.65401	-65.9481	-29333.8
Johnson's SB	12.73235	-62.3023	-29294.9
Power Normal	13.02555	-66.222	-29191.3
Exp. Weibull	13.17272	-65.5831	-29127.5
Gauss Hypergeom	13.2955	-52.8145	-29064.8
Erlang	14.59503	-70.4387	-28631.2
Gamma	14.59505	-70.4387	-28631.2
Gen. Extreme Value	15.14048	-64.3142	-28450.6
Weibull Max	15.14348	-64.3326	-28449.6
Johnson's SU	15.11798	-69.1132	-28449.4
Exp. Normal	21.43802	-74.3712	-26738.4
Burr	40.32404	-78.1765	-23619.7
Mielke	43.88534	-82.6435	-23203

N = 600

Distribution	SSE	AIC	BIC
Exp. Weibull	20.2102	-39.1172	-33109.8
Gen. Gamma	20.48254	-40.8178	-33031.5
Chi	20.73825	-43.4433	-32967.7
Beta	21.25533	-39.2933	-32815
Johnson's SB	21.4952	-39.4632	-32749.3
Power Normal	21.78965	-42.8081	-32678.5
Gamma	22.25657	-52.2519	-32554.5
Erlang	22.25658	-52.2519	-32554.5
Johnson's SU	22.47275	-51.4299	-32489.3
Gauss Hypergeom	23.18061	-22.4307	-32290.5
Exp. Normal	26.94913	-61.9452	-31435.7
Weibull Max	27.45227	-34.8912	-31327.5
Gen. Extreme Value	27.45497	-34.9372	-31326.9
Burr	43.02419	-74.1118	-28691.2
Mielke	43.33378	-80.4006	-28649.3

N = 700

Distribution	SSE	AIC	BIC
Exp. Weibull	21.18584	-65.3901	-39650.3
Gen. Gamma	21.20455	-64.1304	-39644.2
Beta	21.22267	-64.0911	-39638.3
Johnson's SB	21.28952	-63.7948	-39616.7
Power Normal	21.6541	-62.5934	-39509
Erlang	21.77773	-72.6479	-39469.9
Gamma	21.77773	-72.6479	-39469.9
Chi	21.90045	-61.434	-39431.3
Gauss Hypergeom	22.39611	-54.4588	-39251.2
Johnson's SU	22.65695	-72.5752	-39189.4
Gen. Extreme Value	23.33859	-60.0593	-38994.7
Weibull Max	23.34508	-60.0617	-38992.8
Exp. Normal	33.13165	-84.6388	-36589.4
Mielke	64.25041	-99.9794	-32033.8
Burr	126.6419	-93.6788	-27375.4

N = 800

Distribution	SSE	AIC	BIC
Gen. Gamma	17.58976	-60.8231	-48702.4
Exp. Weibull	17.9458	-64.1742	-48542.8
Chi	17.96865	-57.9433	-48541.6
Power Normal	18.09481	-59.7717	-48485.8
Beta	18.19071	-65.43	-48434.7
Johnson's SB	18.41761	-64.9436	-48336
Gamma	19.15872	-72.202	-48030.6
Erlang	19.15874	-72.202	-48030.5
Johnson's SU	20.33907	-73.7017	-47545.1
Gen. Extreme Value	22.68272	-57.9326	-46685
Weibull Max	22.6866	-57.9498	-46683.7
Gauss Hypergeom	27.30233	-27.7532	-45180.9
Exp. Normal	32.13382	-88.3815	-43909.4
Mielke	56.39028	-96.8853	-39418.7
Burr	133.8356	-117.07	-32531

N = 900

Distribution	SSE	AIC	BIC
Gen. Gamma	13.24813	-76.0412	-58601.1
Beta	13.64992	-71.3553	-58332.4
Chi	13.76527	-72.4252	-58265.8
Johnson's SB	13.82278	-70.4643	-58219.2
Erlang	14.17678	-85.7357	-58000.9
Gamma	14.17679	-85.7357	-58000.9
Power Normal	14.20808	-73.6286	-57981.1
Johnson's SU	14.99522	-86.3001	-57487
Exp. Weibull	16.31487	-84.061	-56728.5
Weibull Max	16.65816	-63.246	-56550.4
Gen. Extreme Value	16.66791	-63.2636	-56545.1
Exp. Normal	27.56513	-100.207	-52021
Gauss Hypergeom	41.1353	-16.5139	-48393.7
Mielke	53.8737	-109.48	-45985.8
Burr	172.3748	-112.658	-35526.7

N = 1000

Distribution	SSE	AIC	BIC
Gen. Gamma	28.60799	-109.494	-58495.4
Johnson's SB	28.65882	-107.661	-58477.7
Beta	28.67615	-108.021	-58471.6
Gamma	28.83463	-117.4	-58425.8
Erlang	28.83464	-117.4	-58425.8
Power Normal	29.21088	-110.443	-58296.2
Exp. Weibull	29.23464	-103.427	-58278.8
Chi	29.37858	-108.907	-58239
Johnson's SU	29.44367	-117.486	-58207.6
Gen. Extreme Value	30.13319	-104.802	-57985.5
Gauss Hypergeom	33.20959	-108.479	-56986.2
Exp. Normal	43.77236	-128.039	-54253.6
Mielke	83.4394	-131.441	-47796.4
Burr	221.9439	-115.636	-38018.3
Weibull Max	8163.911	-199.832	-1994.98

Figure 4.21: BIC and AIC for the fitted distributions with the lowest SSE on the Death times of the H_1 PH bars of $N \in \{100, 200, 300, 400, 500, 600, 700, 800, 900, 1000\}$ points in $[0,1]^2$.

where Γ is the gamma function, α and τ are shape parameters, and λ is the scale parameter [25]. The pdf of $y = x - \mu$ can be defined by including the location parameter μ into (4.20):

$$f(x;\alpha,\beta,\lambda,\mu) = \frac{\beta}{\lambda\Gamma(\alpha)}\left(\frac{x-\mu}{\lambda}\right)^{\alpha\beta-1} e^{-\left(\frac{x-\mu}{\lambda}\right)^{\beta}} \quad x \geq 0, \quad \beta,\alpha,\lambda > 0. \tag{4.21}$$

We estimated power and linear models for the four parameters α, β, λ, and μ. The equations for each parameter are plotted in Figures 4.22 and 4.23. For α, we estimated both a power

and a linear model:

$$p_\alpha(N) = 2.30231411 \cdot N^{0.04190071}, \tag{4.22}$$

and

$$l_\alpha(N) = 0.00027045 \cdot N + 2.83212750. \tag{4.23}$$

For β, we also estimated both a power and a linear model:

$$p_\beta(N) = 2.07031731 \cdot N^{-0.00899737}, \tag{4.24}$$

and

$$l_\beta(N) = -0.00003429 \cdot N + 1.97767324. \tag{4.25}$$

Additionally, we tried fitting a constant model for β of the form $c_\beta(N) = 1.9574068$. For λ, we estimated a power model:

$$p_\lambda(N) = 1.94959425 \cdot N^{-0.62019282}. \tag{4.26}$$

We also tried fitting a model of the form $g_\lambda(N) = \frac{0.93589394}{\sqrt{N}}$ for p_λ. For μ, we estimated both a power and a linear model, as well as a constant fit:

$$p_\mu(N) = 0.0137425 \cdot N^{-0.46149296}, \tag{4.27}$$

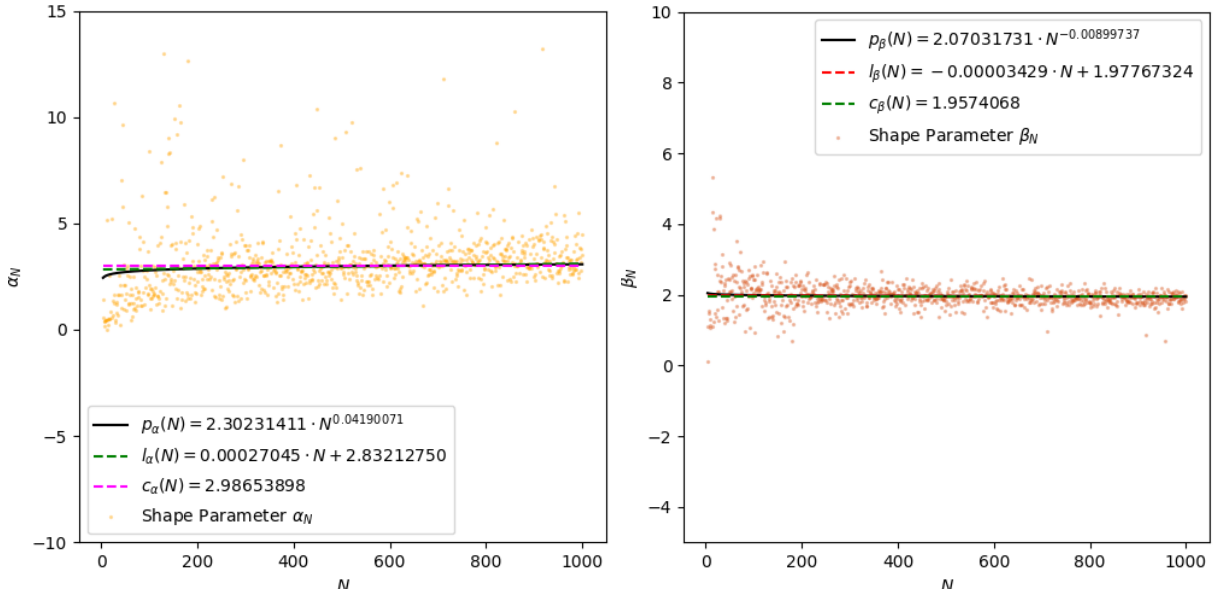
$$l_\mu(N) = -0.000001267 \cdot N + 0.00151769, \tag{4.28}$$

and

$$q_\mu(N) = \frac{0.01718616}{\sqrt{N}}. \tag{4.29}$$

Note that the power models have the form $a \cdot N^b$, while the linear models have the form $a \cdot N + b$. The constant fit approach sets the slope coefficient to zero. It estimates the intercept coefficient, representing the best-fit constant relationship. This approach is not a traditional linear or power model but can help estimate the best-fit constant relationship between the predictor and dependent variables. Particular attention was paid to constant fitting lines for α and β since we observed that the best-fit lines approximated constants: 3 for α and 2 for β, which is an interesting case since the generalized gamma function reduces to the general normal for $\beta = 2$. However, this distribution is not among the ones that provide a low SSE. In Section 4.7.3, we will introduce the generalized normal distribution, which includes other subfamilies. One is the Chi distribution, where the generalized gamma distribution with $\alpha = k/2$ for integer k and $\beta = 2$ reduces to the chi distribution with k degrees of freedom (in our case, $k = 6$). The chi distribution is indeed one of the distributions

that provide a good fit and one of the lowest SSE.



(a) (b)

Figure 4.22: Fitted models on the scatter plots of the estimated shape α (a) and β (b) parameters of the Generalized Gamma distribution fit for the Death times of the H_1 PH bars of $N \in [2, 1000]$ points uniformly distributed in $[0, 1]^2$.

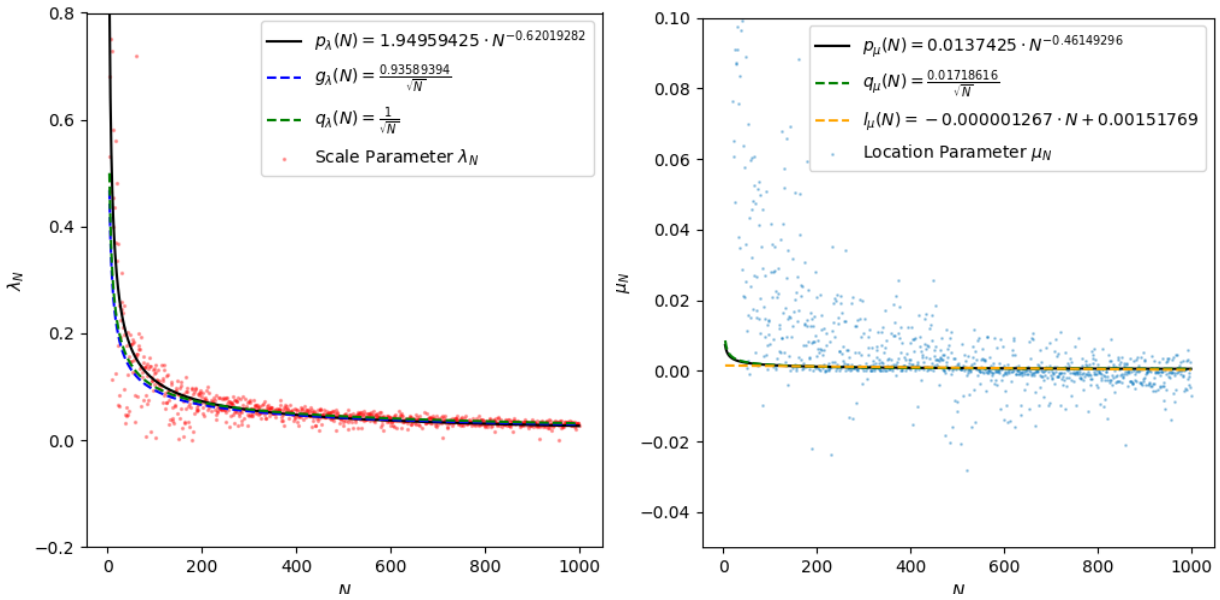
4.7 Unit Cube

4.7.1 Unit Cube: Homological Dimension 0

Death Times

Figure 4.25 presents two distributions providing the lowest SSEs scores: the Burr (Type III) and the Mielke distribution; however, the Mielke (or Dagum) distribution is simply another reparameterization of the Burr distribution (Figure 4.24). The pdf of the Burr III [19, 26] distribution is

$$f(x; \alpha, \beta) = \alpha\beta x^{-\beta-1}\left(1 + \left(\frac{x}{\lambda}\right)^{-\beta}\right)^{-\alpha-1} \quad \text{for} \quad x > 0, \alpha > 0, \beta > 0 \tag{4.30}$$



(a) (b)

Figure 4.23: Fitted models on the scatter plots of the estimated scale λ (a) and location μ (b) parameters of the Generalized Gamma distribution fit for the Death times of the H_1 PH bars of $N \in [2, 1000]$ points uniformly distributed in $[0, 1]^2$.

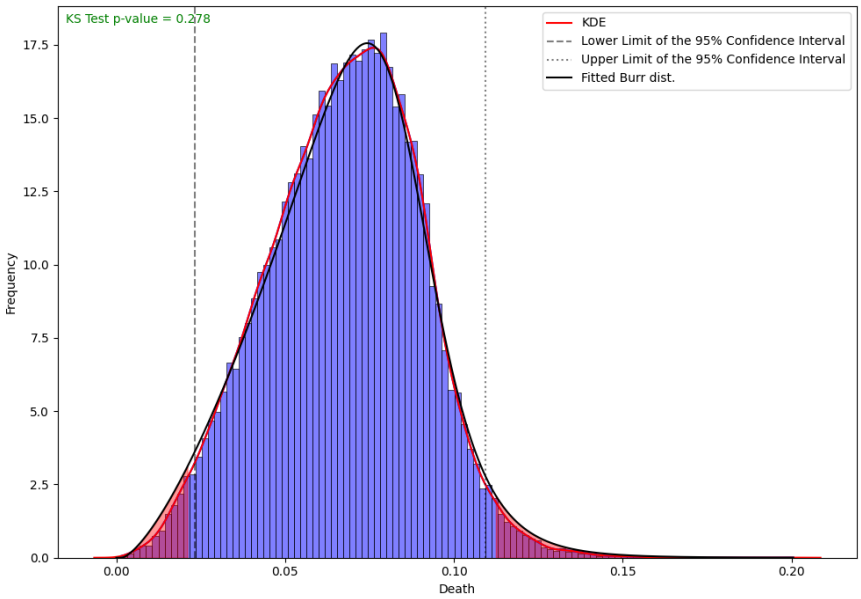


Figure 4.24: Fitted distribution on the normalized histogram of the Death times of the H_0 PH bars in a $[0, 1]^3$ ($N = 1000$; 40 simulations).

where α and β are the shapes parameters We can additionally include scale λ and location μ parameters:

$$f(x; \alpha, \beta, \lambda, \mu) = \frac{\alpha\beta\left(\frac{x-\mu}{\lambda}\right)^{-\beta-1}}{\lambda\left(1 + \left(\frac{x-\mu}{\lambda}\right)^{-\beta}\right)^{\alpha+1}} \quad \text{for} \quad x > 0, \alpha > 0, \beta > 0, \lambda > 0. \tag{4.31}$$

N = 100

Distribution	SSE	AIC	BIC
Burr	2.50169	2.82105	-29475.2
Mielke	2.50169	2.82143	-29475.1
Gen. Normal	2.69772	20.8509	-29181.7
Normal	2.85859	17.3879	-28958.3
Weibull	3.38927	22.8245	-28268.9
Exp. Weibull	3.68018	19.5468	-27931.2
Gen. Gamma	3.68161	20.2079	-27929.6
Power Normal	4.41916	11.4515	-27207.5
Johnson's SB	4.43293	13.5197	-27186.8
Pearson Type III	4.46649	9.55344	-27164.9
Beta	4.46658	11.5467	-27156.5
Power Log-Norm.	4.5472	9.80523	-27085

N = 200

Distribution	SSE	AIC	BIC
Burr	6.56192	-31.1715	-56487.2
Gen. Normal	11.62	-15.6682	-51947.5
Normal	11.8786	-18.995	-51781.3
Weibull	12.0664	-10.0171	-51647.5
Gen. Gamma	12.7746	-11.703	-51184.5
Exp. Weibull	12.8257	-12.5227	-51152.7
Power Log-Norm.	13.1714	-21.3488	-50941
Power Normal	14.51	-21.2153	-50179.5
Pearson Type III	14.7808	-22.6054	-50032.3
Johnson's SB	14.7757	-20.3414	-50026.1
Beta	15.2961	-21.4472	-49750.6
Mielke	24.6	-62.5274	-45968.4

N = 300

Distribution	SSE	AIC	BIC
Mielke	9.14307	-17.8021	-86773
Burr	9.14313	-17.8026	-86772.9
Gen. Normal	12.1433	35.8056	-83354.4
Gen. Gamma	12.2478	39.2919	-83241.4
Exp. Weibull	12.4502	38.106	-83043.4
Weibull	12.6366	34.8347	-82873.3
Normal	15.5533	14.2034	-80374
Johnson's SB	17.593	22.6322	-78866.6
Beta	17.8472	22.5249	-78693.3
Power Log-Norm.	18.0216	6.83619	-78575.9
Pearson Type III	18.2013	6.727	-78465.4
Power Normal	18.3804	7.96214	-78347.1

N = 400

Distribution	SSE	AIC	BIC
Burr	8.30813	-24.1566	-120629
Weibull	16.8892	52.0497	-109316
Exp. Weibull	17.2857	50.3708	-108936
Gen. Normal	17.794	38.5897	-108483
Power Log-Norm.	18.8581	33.481	-107546
Normal	20.4483	20.4967	-106273
Gen. Gamma	20.5722	29.4104	-106158
Power Normal	22.3187	17.5486	-104867
Johnson's SB	22.3878	26.4799	-104808
Pearson Type III	22.5256	15.6146	-104719
Beta	22.575	26.549	-104675
Mielke	42.1653	-52.3928	-94703.7

N = 500

Distribution	SSE	AIC	BIC
Burr	10.1976	7.61563	-151244
Mielke	10.204	7.55904	-151231
Gen. Gamma	21.5319	169.269	-136326
Weibull	21.9143	154.065	-135985
Exp. Weibull	22.7297	145.568	-135246
Gen. Normal	25.667	110.316	-132830
Power Normal	27.922	88.6437	-131149
Normal	27.9868	86.4594	-131112
Pearson Type III	28.6219	86.0017	-130655
Johnson's SB	28.6815	87.8498	-130603
Beta	28.7818	94.5369	-130534
Power Log-Norm.	30.5091	68.9038	-129370

N = 600

Distribution	SSE	AIC	BIC
Mielke	23.079	-7.26241	-166685
Burr	23.1345	-8.14932	-166627
Gen. Gamma	26.74	144.282	-163151
Exp. Weibull	27.2046	141.278	-162738
Weibull	27.9549	133.893	-162095
Gen. Normal	32.5506	109.509	-158442
Power Log-Norm.	37.4463	63.5081	-155069
Johnson's SB	38.5748	102.67	-154357
Beta	39.1244	100.075	-154017
Pearson Type III	39.9101	70.997	-153550
Power Normal	40.1625	68.0824	-153399
Normal	42.8246	58.8463	-151869

N = 700

Distribution	SSE	AIC	BIC
Mielke	16.3003	36.337	-208525
Burr	16.3007	36.3367	-208524
Gen. Gamma	31.1897	340.151	-190355
Weibull	31.827	284.255	-189799
Exp. Weibull	33.5185	336.337	-188339
Gen. Normal	39.9684	221.536	-183422
Power Log-Norm.	41.7316	158.234	-182203
Power Normal	43.1684	175.88	-181265
Johnson's SB	43.7597	199.173	-180874
Pearson Type III	43.922	177.262	-180781
Beta	44.0414	196.583	-180694
Normal	47.2491	160.922	-178746

N = 800

Distribution	SSE	AIC	BIC
Burr	20.0794	10.744	-236255
Mielke	20.717	6.25157	-235253
Exp. Weibull	39.1994	248.869	-214821
Weibull	39.6308	242.81	-214481
Power Log-Norm.	50.1735	126.803	-206913
Gen. Normal	51.609	178.666	-206020
Power Normal	51.9328	141.686	-205819
Johnson's SB	52.7579	169.797	-205304
Pearson Type III	52.8004	144.894	-205288
Beta	53.1423	166.518	-205071
Normal	58.7274	122.788	-201890
Gen. Gamma	326.5	-63.5996	-146904

N = 900

Distribution	SSE	AIC	BIC
Burr	24.0687	207.424	-263464
Mielke	24.0696	207.422	-263462
Weibull	31.1512	1027.69	-254178
Exp. Weibull	35.961	1041.1	-248993
Gen. Gamma	38.3659	869.279	-246660
Gen. Normal	44.0147	821.695	-241720
Power Log-Norm.	45.5246	598.133	-240494
Power Normal	48.8569	655.606	-237958
Pearson Type III	48.928	681.759	-237906
Beta	49.0411	767.955	-237812
Normal	56.5737	601.952	-232684
Johnson's SB	56.7581	607.124	-232545

N = 1000

Distribution	SSE	AIC	BIC
Burr	28.6374	-81.3139	-289305
Mielke	28.7555	-82.2694	-289140
Gen. Gamma	36.5221	49.8058	-279586
Weibull	42.5683	26.6927	-273475
Power Log-Norm.	54.8645	-26.0508	-263325
Johnson's SB	58.6235	11.7052	-260677
Beta	59.9662	7.64958	-259772
Pearson Type III	60.1827	-18.5165	-259638
Power Normal	60.8655	-24.8189	-259187
Gen. Normal	61.7553	1.90617	-258607
Normal	76.2935	-44.6723	-250170
Exp. Weibull	177.273	-29.8319	-216459

Figure 4.25: BIC and AIC for the fitted distributions with the lowest SSE on the Death times of the H_0 PH bars of $N \in \{100, 200, 300, 400, 500, 600, 700, 800, 900, 1000\}$ points in $[0, 1]^3$.

A power model provides the following fits: (4.32) for α, (4.33) for β, (4.34) for λ, and (4.35) for μ.

$$p_\alpha(N) = 6.709447176 \cdot N^{0.07087991} \approx 6.709 \cdot N^{0.071}. \tag{4.32}$$

$$p_\beta(N) = 0.25664305 \cdot N^{-0.01289085} \approx 0.257 \cdot N^{-0.013}. \tag{4.33}$$

$$p_\lambda(N) = 0.97274449 \cdot N^{-0.34760558} \approx 0.97 \cdot N^{-0.348} \quad \text{or} \quad q_\lambda(N) = \frac{0.9}{\sqrt[3]{N}}. \tag{4.34}$$

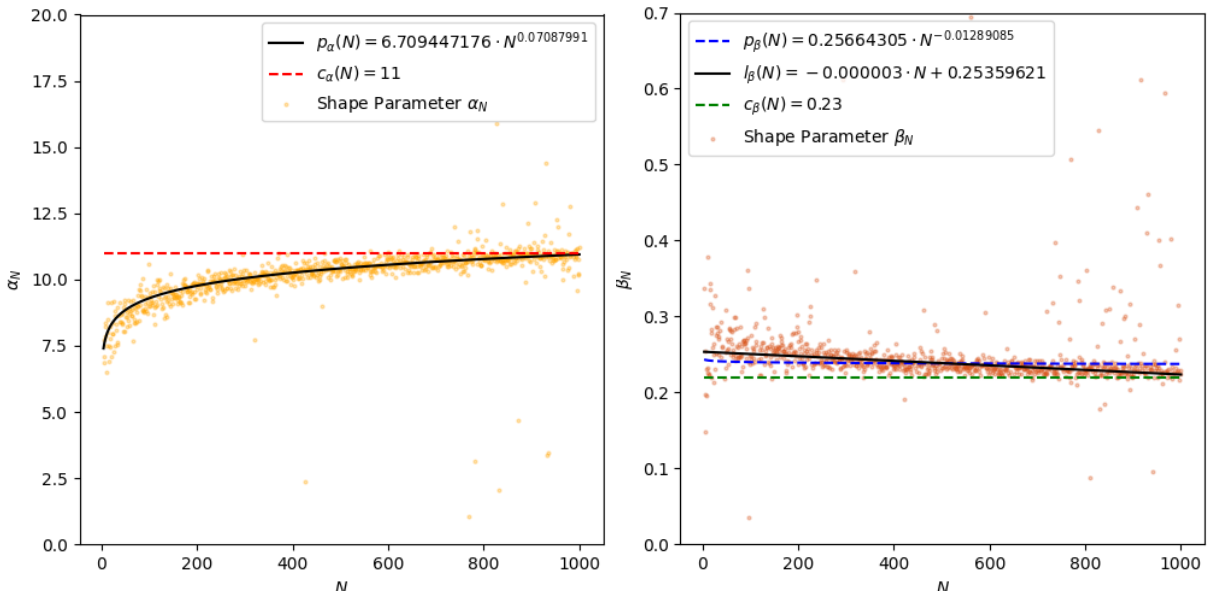
$$h_\mu(N) = 0.11660005 \cdot N^{-0.63004483} \approx 0.117 \cdot N^{-0.63}. \tag{4.35}$$

These results indicate that as N increases, the shape parameter α increases, the scale parameter λ decreases, and the shape parameter β decreases slightly, revealing that the distribution becomes less skewed as N increases. The location parameter μ also decreases, meaning that the distribution shifts towards 0 as N increases (Figures 4.26 and 4.27). Taken together, these results suggest that as N increases, the probability density function becomes more concentrated around 0, with smaller scale and location parameters, and a shape parameter that becomes more peaked and less skewed. The decreasing values of the scale and location parameters suggest that the Birth times tend to be smaller as N increases. Additionally, the increasing value of the shape parameter α suggests that the distribution becomes more peaked and concentrated around the mode.

4.7.2 Unit Cube: Homological Dimension 1

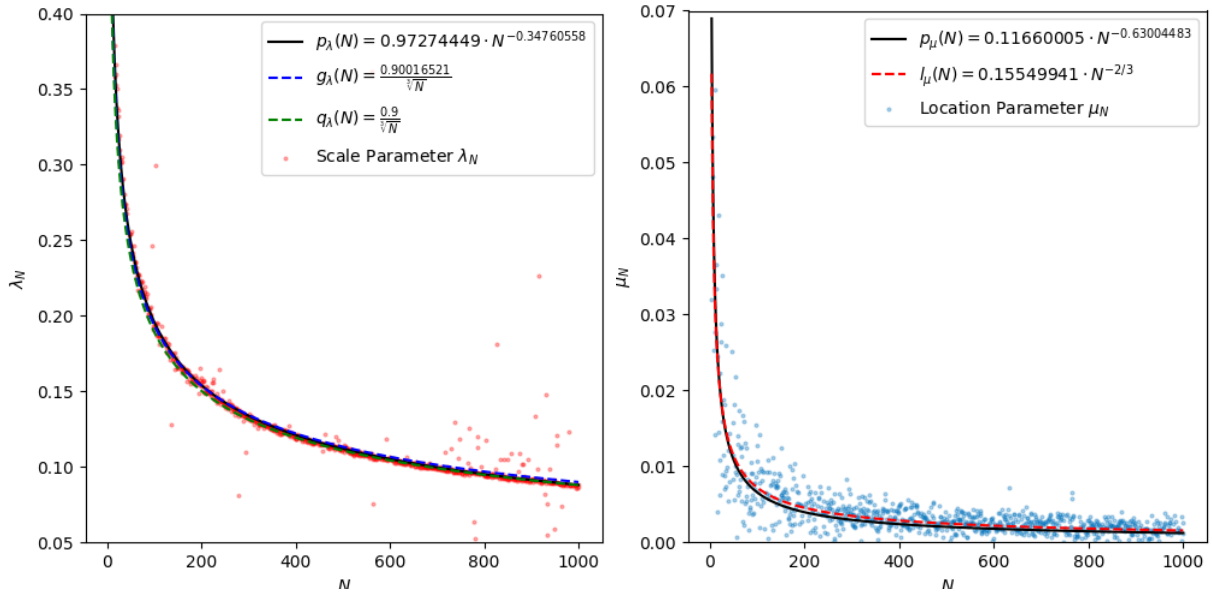
Birth Times

The Generalized Logistic distribution is one of the densities that overall provides the lowest SSE values for the Birth times of the 1-dimensional persistent homology bars (Figure 4.29).



(a) (b)

Figure 4.26: Fitted models on the scatter plots of the estimated shape α (a) and β (b) parameters of the Burr distribution fit for the Death times of the H_0 PH bars of $N \in [2, 1000]$ points uniformly distributed in $[0, 1]^3$.



(a) (b)

Figure 4.27: Fitted models on the scatter plots of the estimated scale λ (a) and location μ (b) parameters of the Burr distribution fit for the Death times of the H_0 PH bars of $N \in [2, 1000]$ points uniformly distributed in $[0, 1]^3$.

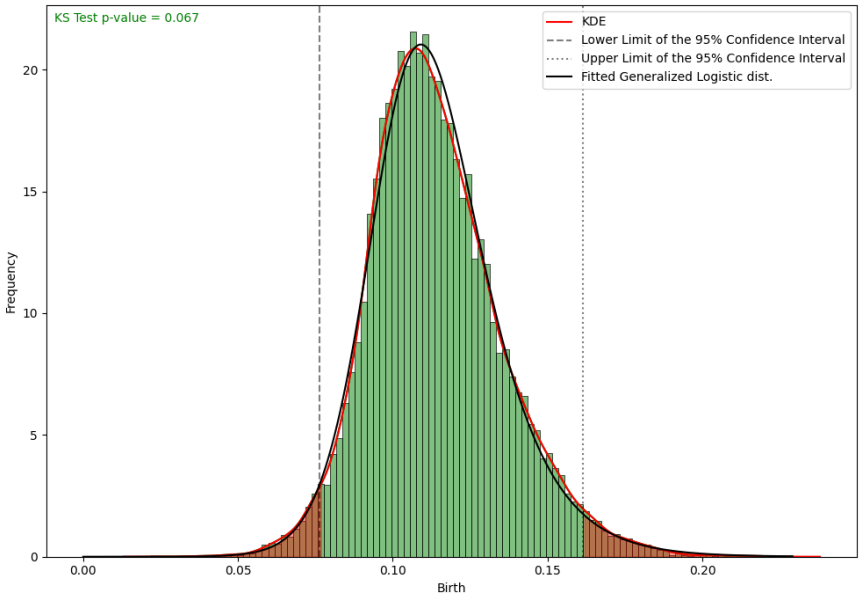


Figure 4.28: Fitted distribution on the normalized histogram of the Birth times of the H_1 PH bars in $[0,1]^3$. ($N = 1000$; 40 simulations).

The standardized pdf of the Generalized Logistic Distribution (Type I) [20] with shape parameter c is

$$f(x,c) = c\frac{e^{-x}}{(1+e^{-x})^{c+1}} \quad \text{for} \quad x \in \mathbb{R}, c > 0. \tag{4.36}$$

This expression can be reparameterized in terms of location and scale parameters as:

$$f(x;c,\lambda,\mu) = \frac{ce^{-\left(\frac{x-\mu}{\lambda}\right)}}{\lambda\left(1+e^{-\left(\frac{x-\mu}{\lambda}\right)}\right)^{c+1}} \quad x \in \mathbb{R}, c > 0, \mu \in \mathbb{R}, \lambda > 0 \tag{4.37}$$

where c, μ, and λ represent the shape, location, and scale parameters, respectively. The parameters for the Generalized Logistic distribution that are used to model the death times of H_1 PH bars (Figure 4.30) can be described as follows:

$$l_c(N) = 0.00002704 \cdot N + 2.07751305 \quad \text{or} \quad c_c(N) = 2.1, \tag{4.38}$$

$$g_\mu(N) = \frac{0.98940653}{\sqrt[3]{N}} \approx \frac{1}{\sqrt[3]{N}}, \tag{4.39}$$

and

$$g_\lambda(N) = \frac{0.14398159}{\sqrt[3]{N}} \approx \frac{0.14}{\sqrt[3]{N}}. \tag{4.40}$$

N = 100

Distribution	SSE	AIC	BIC
Gen. Logistic	7.83553	-0.47084	-7293.73
Mielke	7.82076	1.4311	-7289.14
Johnson's SU	8.05903	2.84426	-7246.85
Burr	8.32615	-0.96516	-7200.91
Fisk	8.64219	-3.10881	-7155.67
Exp. Normal	8.70397	1.23575	-7145.63
Alpha	10.2408	6.94398	-6916.52
Exp. Weibull	10.327	9.85507	-6897.47
Inv. Gamma	10.5148	7.51187	-6879.32
Beta Prime	10.6143	9.6819	-6858.8
F	10.6586	9.14223	-6852.93
Log-Normal	10.7462	7.96045	-6848.65
Power Log-Norm.	10.7494	9.93989	-6840.98
Inv. Gauss	10.9293	7.9644	-6824.84
Pearson Type III	11.443	9.12979	-6760.13

N = 200

Distribution	SSE	AIC	BIC
Gen. Logistic	7.11861	3.1089	-20048.8
Johnson's SU	7.35797	6.24756	-19932.4
Mielke	7.44279	6.60738	-19894.9
Fisk	9.19177	-3.74041	-19212
Exp. Normal	9.7106	7.20627	-19032.2
Exp. Weibull	13.9345	29.248	-17841.7
Alpha	14.0637	22.8344	-17819.6
Inv. Gamma	14.7563	24.4664	-17662.2
F	14.9563	26.2496	-17610
Inv. Gauss	15.2969	25.613	-17544.4
Log-Normal	15.339	25.7261	-17535.4
Pearson Type III	17.0614	29.0027	-17187
Power Log-Norm.	18.0984	30.7314	-16985.7
Beta Prime	63.6523	-7.6327	-12868.3
Burr	388.029	-87.6614	-6950.04

N = 300

Distribution	SSE	AIC	BIC
Gen. Logistic	17.7763	21.5658	-29269.2
Johnson's SU	18.7578	31.2492	-28983
Burr	22.5933	8.86863	-28022.1
Exp. Normal	23.1788	28.9411	-27898.5
Exp. Weibull	23.4923	66.422	-27820.6
Fisk	23.8786	7.92969	-27744.9
Alpha	25.2598	58.3323	-27454.4
Inv. Gauss	25.4838	65.7373	-27408.9
Inv. Gamma	26.0263	61.8098	-27300.1
F	26.0223	63.7907	-27292.3
Log-Normal	26.7303	64.4717	-27162.2
Power Log-Norm.	26.7294	66.4574	-27153.8
Pearson Type III	28.8965	71.7703	-26759.7
Mielke	32.6949	-4.50442	-26113.3
Beta Prime	106.806	-1.82549	-19999

N = 400

Distribution	SSE	AIC	BIC
Gen. Logistic	13.5049	44.4248	-44380.8
Johnson's SU	13.8097	48.9321	-44213.7
Mielke	13.9074	46.0624	-44163.7
Burr	14.5504	47.2218	-43843.3
Fisk	20.2733	32.7594	-41500.5
Exp. Normal	20.5464	75.6834	-41405.6
Exp. Weibull	26.0789	121.271	-39706.2
Alpha	28.6647	122.137	-39044.8
Inv. Gauss	30.0377	134.394	-38713.1
Inv. Gamma	30.0993	124.866	-38698.5
F	30.424	125.936	-38613.6
Log-Normal	31.1944	127.013	-38445.2
Power Log-Norm.	31.18	129.042	-38439.6
Pearson Type III	34.3769	131.983	-37756.4
Beta Prime	39.4883	166.776	-36764.7

N = 500

Distribution	SSE	AIC	BIC
Gen. Logistic	9.72655	1.17507	-61952.9
Burr	9.74108	3.39431	-61930.3
Johnson's SU	9.87298	1.87859	-61808.4
Mielke	9.89112	2.09592	-61791.7
Fisk	12.4313	-12.4284	-59728.8
Exp. Normal	16.9948	12.736	-56894.2
Alpha	32.1148	48.3209	-51125.2
Inv. Gamma	34.1087	52.082	-50579.2
F	34.7132	53.4993	-50410.8
Log-Normal	35.7297	54.9303	-50158.3
Power Log-Norm.	35.7457	56.8426	-50145.1
Inv. Gauss	37.397	55.8553	-49744.8
Pearson Type III	40.361	62.2224	-49053.4
Exp. Weibull	46.4395	56.5554	-47772.6
Beta Prime	207.886	-8.76977	-34185.6

N = 600

Distribution	SSE	AIC	BIC
Gen. Logistic	30.8012	-39.5632	-66077.5
Johnson's SU	31.5491	-36.5465	-65799.2
Mielke	31.6952	-38.1168	-65747.4
Fisk	39.3587	-49.243	-63329.2
Burr	40.0096	-38.9647	-63136
Exp. Normal	45.0362	-24.6882	-61818.7
Inv. Gauss	58.793	25.7174	-58830.6
Exp. Weibull	59.3873	7.45381	-58708.5
Alpha	61.5345	3.98431	-58319.7
Inv. Gamma	62.4856	6.81381	-58147.7
Beta Prime	64.0792	8.34036	-57856.1
F	64.2202	7.78759	-57831.4
Log-Normal	66.0119	7.76676	-57532.3
Power Log-Norm.	66.0691	9.71679	-57513.3
Pearson Type III	71.9564	11.7104	-56565.7

N = 700

Distribution	SSE	AIC	BIC
Gen. Logistic	25.0093	-14.0547	-83518.7
Mielke	25.3065	-12.7853	-83351.9
Burr	25.5148	-11.5373	-83242.8
Johnson's SU	26.288	-6.22233	-82845.5
Fisk	37.2548	-27.5819	-78214.2
Exp. Normal	38.5188	10.8818	-77770.1
Exp. Weibull	47.5605	57.4542	-74954.1
Inv. Gauss	49.4531	69.9548	-74444.2
Alpha	50.9934	53.5114	-74036
Inv. Gamma	53.0277	56.7314	-73515.3
Beta Prime	53.2295	59.352	-73455.3
F	53.3195	58.4443	-73432.8
Log-Normal	54.721	59.1055	-73097
Power Log-Norm.	54.8559	60.9246	-73054.7
Pearson Type III	59.6052	65.1658	-71959

N = 800

Distribution	SSE	AIC	BIC
Gen. Logistic	24.3185	19.4044	-98361.6
Burr	24.513	21.7916	-98230.2
Johnson's SU	25.0612	26.4039	-97892.5
Mielke	28.4716	-0.8363	-95944
Exp. Normal	35.484	42.5656	-92591.1
Fisk	38.8224	-2.643	-91217.9
Inv. Gauss	48.3554	141.39	-87864.5
Alpha	48.84	107.305	-87712.2
Exp. Weibull	48.967	121.456	-87662.9
F	51.6881	118.771	-86837
Inv. Gamma	51.7248	114.593	-86835.8
Log-Normal	54.216	120.138	-86117.4
Power Log-Norm.	54.2441	122.139	-86099.9
Pearson Type III	61.7548	134.555	-84129.1
Beta Prime	219.211	-11.5689	-64771.9

N = 900

Distribution	SSE	AIC	BIC
Gen. Logistic	31.0402	6.72902	-108644
Johnson's SU	32.4378	12.8331	-107877
Beta Prime	39.4782	206.903	-104498
Burr	39.7466	-16.675	-104382
Fisk	46.2008	-18.6161	-101803
Exp. Normal	48.266	25.4978	-101051
Exp. Weibull	61.1996	102.483	-96956.9
Alpha	66.5179	88.7862	-95533.2
Inv. Gauss	67.8197	106.59	-95199.8
Inv. Gamma	69.4578	95.9099	-94789.2
F	69.9038	97.8547	-94669.4
Log-Normal	72.2964	101.151	-94100.2
Power Log-Norm.	73.5922	101.95	-93784.9
Pearson Type III	80.1759	115.169	-92320.7
Mielke	286.731	-88.4467	-70390

N = 1000

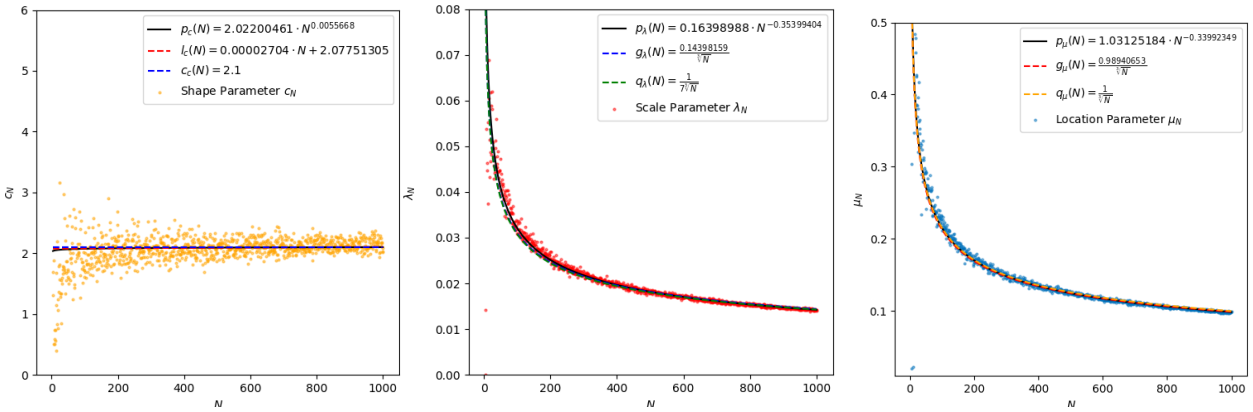
Distribution	SSE	AIC	BIC
Gen. Logistic	34.6517	10.6084	-121182
Johnson's SU	35.2942	16.853	-120820
Mielke	37.6067	10.3718	-119602
Burr	49.5542	-2.17467	-114308
Exp. Normal	53.8386	49.9418	-112727
Fisk	54.5523	-10.02	-112474
Exp. Weibull	66.1143	125.282	-108776
Inv. Gauss	71.8479	146.808	-107190
Alpha	72.9048	121.258	-106910
Inv. Gamma	75.5733	127.561	-106220
F	76.9566	129.81	-105862
Log-Normal	79.109	131.153	-105342
Beta Prime	80.2109	134.553	-105067
Pearson Type III	87.3293	141.665	-103445
Power Log-Norm.	88.3058	165.564	-103222

Figure 4.29: BIC and AIC for the fitted distributions with the lowest SSE on the Birth times of the H_1 PH bars of $N \in \{100, 200, 300, 400, 500, 600, 700, 800, 900, 1000\}$ points in $[0, 1]^3$.

The shape parameter c is a function of the number of data points N and is defined by the above equation. Alternatively, a fixed value of $c = 2.1$ can be used. This equation shows that c increases rapidly with N initially but eventually approaches a constant value of 2.1 beyond a certain value of N. This suggests that the system being modelled by the Generalized Logistic distribution reaches a steady state as the amount of data increases, with a certain level of variability that does not change significantly as more data points are added. The scale parameter λ represents the spread of the distribution, and as the number of uniformly distributed points N in the cube increases, the variance of the distribution decreases. The location parameter μ represents the center of the distribution, and as the number of uniformly distributed points N in the cube increases, the expected value of the distribution decreases. These equations show that the location parameter μ shifts to the left, and the scale parameter λ decreases as more data points are added. From a persistent homological point of view, as more uniformly distributed points are added to the cube, the Birth times of the 1-dimensional persistent homology bars decrease and become more tightly clustered around a smaller value of μ (Figure 4.28).

Death Times

Several distributions can be used to fit the histograms of the Death times of the 1-dimensional persistent homology bars of N points in a unit square (Figure 4.31). Figure 4.32 displays that the Generalized Gamma, Weibull, and Power Log-Normal distributions exhibit the lowest SSE values. However, while the estimated parameters of the Generalized Gamma and Power Log-Normal distributions for various N do not seem to follow any particular trend, the



(a) (b) (c)

Figure 4.30: Fitted models on the scatter plots of the estimated shape c (a), scale λ (b), and location μ (c) parameters of the Generalized Logistic distribution fit for the Birth times of the H_1 PH bars of $N \in [2, 1000]$ points uniformly distributed in $[0, 1]^3$.

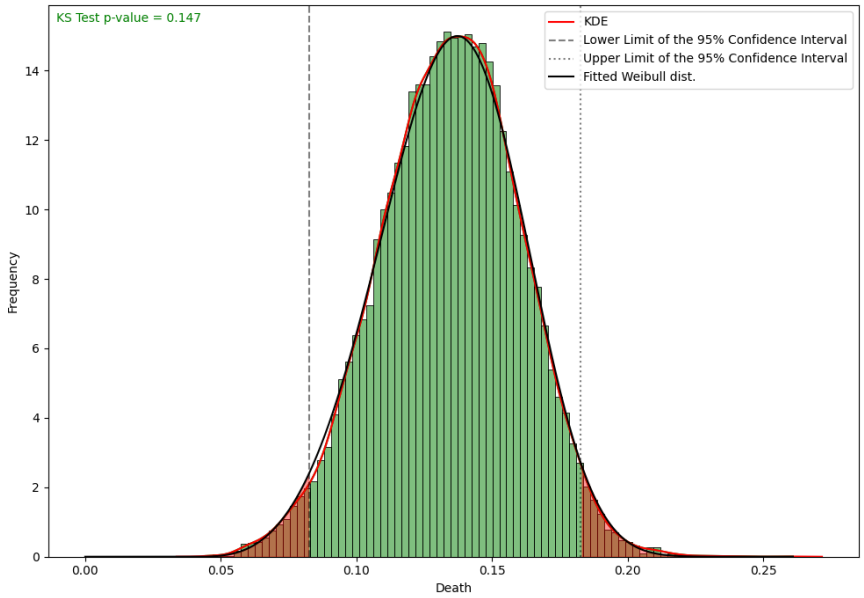


Figure 4.31: Fitted distribution on the normalized histogram of the Death times of the H_1 PH bars in $[0, 1]^3$ ($N = 1000$; 40 simulations).

Weibull distribution's parameters do exhibit a discernible trend. As a result, we will focus on this distribution since the Death times of H_0 bars in the unit interval and the Death times of the H_0 bars in the unit square can be described by the Weibull distribution and the exponentiated Weibull distribution, respectively. The Weibull distribution's probability density function (4.5) is described in Section 4.5.1. We estimated the parameters c, λ, and μ and fitted linear and power models to describe their relationship to N with power models (Figure 4.33).

N = 100

Distribution	SSE	AIC	BIC
Weibull	1.869695	-3.57033	-9587.55
Gen. Gamma	1.86192	-2.1812	-9586.3
Power Log-Norm.	1.900671	-3.86927	-9556.53
Log-Gamma	2.002705	-6.57636	-9488.25
Pearson Type III	2.029719	-6.53178	-9468.88
Johnson's SB	2.031893	-2.92031	-9460.06
Beta	2.047572	-2.95569	-9448.96
Power Normal.	2.059972	-6.68593	-9447.51
Tukey-Lambda	2.102842	-5.35955	-9417.74
Gen. Normal.	2.133373	-5.18546	-9396.91
Normal	2.171907	-8.86645	-9378.32
t	2.17211	-6.86482	-9370.91
Nakagami	2.216073	-6.95643	-9341.96
Burr	3.471097	-6.39207	-8686.26

N = 200

Distribution	SSE	AIC	BIC
Weibull	4.945372	35.36339	-21368.8
Gen. Gamma	4.939136	38.42276	-21364.9
Beta	4.942333	37.77208	-21362.8
Power Normal.	4.95552	34.93312	-21362.1
Power Log-Norm.	4.956911	35.67712	-21353.1
Pearson Type III	4.976662	35.18281	-21348.1
Normal	5.116466	31.57976	-21265
Log-Gamma	5.115913	35.54372	-21249.2
Gen. Normal.	5.146992	32.76774	-21237.3
Nakagami	5.182601	32.99679	-21214.6
t	5.216048	29.74503	-21193.5
Tukey-Lambda	5.396096	29.9878	-21081.8
Johnson's SB	5.997149	59.99004	-20734.2
Burr	10.22272	13.53811	-18970.9

N = 300

Distribution	SSE	AIC	BIC
Gen. Normal.	4.214558	59.83571	-37344.6
Weibull	4.560642	84.03189	-36930.7
Gen. Gamma	4.843667	56.47902	-36606.4
Johnson's SB	5.221753	54.86439	-36212.3
Beta	5.237559	54.57586	-36196.5
Power Log-Norm.	5.350129	45.981	-36084.9
Log-Gamma	5.502498	49.36916	-35946.2
Normal	5.706829	43.77465	-35763.6
t	5.705948	45.76713	-35755.9
Power Normal.	5.837443	43.15782	-35636.4
Pearson Type III	5.882473	42.64943	-35596.1
Nakagami	5.890588	42.51076	-35588.8
Tukey-Lambda	5.934005	56.95549	-35550.3
Burr	59.3527	22.81355	-23465.9

N = 400

Distribution	SSE	AIC	BIC
Weibull	4.71431	19.13409	-51833.8
Gen. Gamma	5.103583	17.77024	-51271.3
Power Log-Norm.	5.152531	11.26236	-51212.5
Pearson Type III	5.201872	12.45579	-51144.9
Log-Gamma	5.202432	12.28612	-51144.2
Johnson's SB	5.201958	14.97742	-51135.9
Beta	5.211492	15.29791	-51123
Gen. Normal.	5.937198	21.2334	-50207.5
Normal	6.459844	12.91933	-49618.2
t	6.460327	14.91362	-49608.8
Gen. Normal.	6.467101	15.72253	-49601.4
Tukey-Lambda	6.527258	17.66986	-49535.7
Nakagami	6.651786	15.0141	-49401.7
Burr	11.29381	2.911958	-45639.6

N = 500

Distribution	SSE	AIC	BIC
Gen. Gamma	5.339642	36.24428	-68055.4
Weibull	5.351524	34.71723	-68035.1
Log-Gamma	5.379403	29.83865	-67996.7
Beta	5.402842	31.54029	-67947.8
Pearson Type III	5.430479	29.2399	-67910.2
Power Normal.	5.436221	28.46072	-67900.6
Johnson's SB	5.434465	31.28624	-67894.4
Power Log-Norm.	5.65089	25.78314	-67555.5
Gen. Normal.	5.691413	27.23193	-67481.1
Nakagami	5.721259	27.07705	-67433.2
t	5.804186	23.15985	-67301.6
Tukey-Lambda	6.252357	25.86221	-66621.4
Normal	10.51231	76.17125	-61869.8
Burr	20.22019	-6.33301	-55878.6

N = 600

Distribution	SSE	AIC	BIC
Power Log-Norm.	6.017414	-4.72077	-82941.2
Gen. Gamma	6.034575	-1.26926	-82909.8
Weibull	6.37292	24.51342	-82316.7
Johnson's SB	6.548037	-5.31118	-82008.1
Beta	6.582085	-5.39737	-81950.8
Pearson Type III	6.733772	-9.9242	-81708.6
Log-Gamma	6.787937	-10.4555	-81620.1
Power Normal.	6.961806	-12.5496	-81340.8
Gen. Normal.	9.635988	-6.41639	-77751.4
Normal	10.33179	-13.1993	-76990.9
t	10.3354	-11.1859	-76977.7
Nakagami	10.48479	-11.6418	-76819.3
Tukey-Lambda	10.62856	-6.33676	-76668.9
Burr	20.0413	-24.9099	-69656.2

N = 700

Distribution	SSE	AIC	BIC
Weibull	10.83166	9.448956	-94054.6
Gen. Gamma	11.15566	12.50163	-93673.9
Beta	12.45845	-16.2336	-92202.4
Johnson's SB	12.46516	-17.1019	-92195.3
Pearson Type III	12.53335	-19.764	-92132.5
Log-Gamma	12.56764	-20.126	-92096.4
Power Log-Norm.	12.61594	-21.5517	-92036.1
Power Normal.	12.72318	-22.2726	-91933.5
Gen. Normal.	15.11712	-22.2088	-89651.5
Normal	16.06551	-28.3505	-88855.5
t	16.06518	-26.3456	-88846.3
Nakagami	16.51295	-27.1289	-88482.4
Tukey-Lambda	16.64382	-22.973	-88377.9
Burr	27.91041	-45.9524	-81525.4

N = 800

Distribution	SSE	AIC	BIC
Weibull	6.711973	14.95711	-117079
Gen. Gamma	6.745101	12.56217	-117004
Johnson's SB	6.984764	11.53255	-116474
Beta	7.011394	11.18291	-116417
Pearson Type III	7.291083	4.28228	-115833
Log-Gamma	7.314785	4.006481	-115784
Power Normal.	7.497562	1.915528	-115410
Gen. Normal.	8.129638	9.247874	-114182
Normal	8.629921	-0.30784	-113286
t	8.632787	1.704312	-113272
Tukey-Lambda	8.841952	10.25317	-112908
Nakagami	8.906979	1.501685	-112797
Power Log-Norm.	11.54984	59.22421	-108857
Burr	29.87157	-26.5281	-94437

N = 900

Distribution	SSE	AIC	BIC
Gen. Gamma	11.08854	396.554	-126642
Weibull	11.87668	315.6048	-125459
Power Log-Norm.	11.9346	592.0159	-125385
Beta	12.62101	340.7573	-124411
Johnson's SB	12.62105	338.816	-124411
Pearson Type III	12.64256	334.7018	-124391
Log-Gamma	12.65035	333.8481	-124381
Power Normal.	12.7559	322.6152	-124237
Gen. Normal.	13.05428	329.9156	-123839
Normal	14.70399	294.7261	-121797
t	14.70443	296.7118	-121787
Nakagami	15.13702	290.7384	-121287
Tukey-Lambda	19.41252	302.5456	-116999
Burr	37.37957	97.58641	-105695

N = 1000

Distribution	SSE	AIC	BIC
Weibull	9.362437	20.61606	-148122
Power Log-Norm.	9.393967	25.3995	-148057
Gen. Gamma	9.952543	21.11123	-146937
Beta	9.994796	20.8203	-146854
Pearson Type III	10.05081	18.16785	-146756
Log-Gamma	10.07749	17.82332	-146704
Power Normal.	10.36568	14.76847	-146157
Johnson's SB	11.13391	90.23269	-144770
Gen. Normal.	11.20411	18.72204	-144648
Normal	12.59686	9.444921	-142385
t	12.63278	11.09609	-142320
Nakagami	12.84101	11.00055	-142002
Tukey-Lambda	14.09563	18.19586	-140194
Burr	40.29662	-29.3381	-119804

Figure 4.32: BIC and AIC for the fitted distributions with the lowest SSE on the Death times of the H_1 PH bars of $N \in \{100, 200, 300, 400, 500, 600, 700, 800, 900, 1000\}$ points in $[0, 1]^3$.

The power model that fits the estimated shape parameters is:

$$p_c(N) = 3.18812302 \cdot N^{0.0400306} \approx g_c(N) = 3.2 \cdot N^{0.04}. \tag{4.41}$$

For the location parameters, we have:

$$p_\mu(N) = 0.51076164 \cdot N^{-0.37219507} \approx 0.51 \cdot N^{-0.37}, \tag{4.42}$$

or

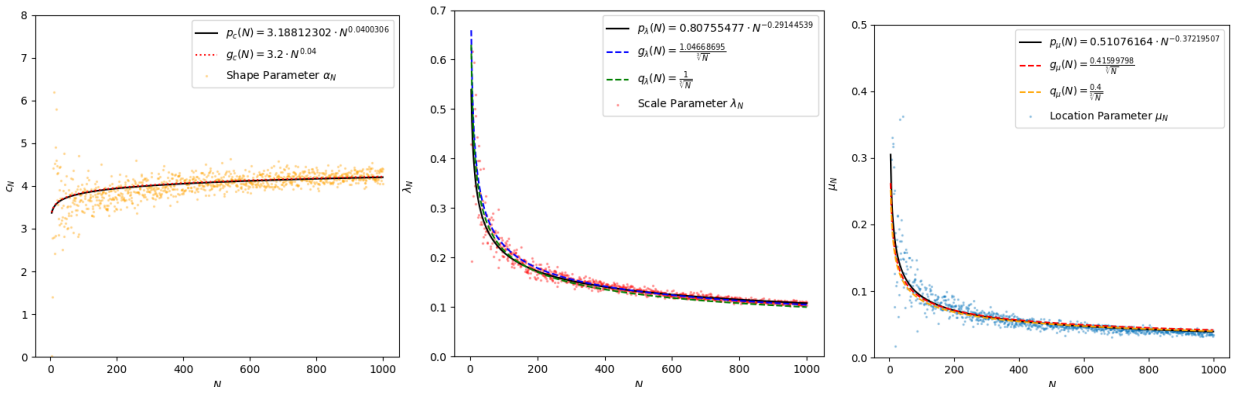
$$g_\mu(N) = \frac{0.41599798}{\sqrt[3]{N}} \approx \frac{0.42}{\sqrt[3]{N}}. \tag{4.43}$$

Similarly, the curves that fit the location parameters are:

$$p_\lambda(N) = 0.80755477 \cdot N^{-0.29144539} \approx p_\lambda(N) = 0.8 \cdot N^{-0.29}, \tag{4.44}$$

or

$$g_\lambda(N) = \frac{1.04668695}{\sqrt[3]{N}} \approx \frac{1}{\sqrt[3]{N}}. \tag{4.45}$$



(a) (b) (c)

Figure 4.33: Fitted models on the scatter plots of the estimated shape c (a), scale λ (b), and location μ (c) parameters of the Weibull distribution fit for the Death times of the H_1 PH bars of $N \in [2, 1000]$ points uniformly distributed in $[0, 1]^3$.

4.7.3 Unit Cube: Homological Dimension 2

The pdf of a Normal distribution with scale parameter λ (generally denoted with σ) and location parameter μ is given by

$$f(x; \lambda, \mu) = \frac{1}{\sqrt{2\pi}\lambda} \exp\left\{-\frac{(x-\mu)^2}{2\lambda^2}\right\} \tag{4.46}$$

for $-\infty < x < \infty, -\infty < \mu < \infty$ and $\lambda > 0$. A generalization of this pdf involves replacing 2 with an arbitrary power $c > 0$. This generalization is known as the General Normal distribution [27], and the pdf takes the form

$$f(x; c, \lambda, \mu) = K \exp\left\{-\left|\frac{x-\mu}{\lambda}\right|^c\right\} \tag{4.47}$$

where

$$K = \frac{c}{2\lambda\Gamma(1/c)}. \tag{4.48}$$

Figures 4.34 and 4.37, and the histograms in Figures 4.35 and 4.38 indicate that the Generalized Normal distribution describes the Birth and Death times of H_2 PH bars.

Birth Times

The relationships between the estimated parameters of the Generalized Normal distribution (Figure 4.36) and N can be described with the following linear and power models:

$$l_c(N) = -0.00001211 \cdot N + 1.78917136 \approx 1.8, \tag{4.49}$$

N = 100

Distribution	SSE	AIC	BIC
Gen. Normal	3.4636	-12.3049	-851.06
Tukey-Lambda	3.57401	-12.9103	-844.438
Beta	3.67001	-11.6507	-833.493
Exp. Weibull	4.10728	-11.5964	-809.742
t	4.4185	-12.8922	-799.683
Power Normal	4.45991	-13.0226	-797.714
Log-Gamma	4.4933	-13.2124	-796.14
Johnson's SU	4.41967	-10.8911	-794.275
Power Log-Norm.	4.55545	-11.0419	-787.89
Inv. Gamma	4.73643	-11.9228	-785.022
Alpha	4.88475	-11.532	-778.515
Recip. Inv. Gauss.	4.90493	-11.3307	-777.646
Gauss. Hyperg.	4.57258	-7.8235	-776.394
Inv. Gauss	5.04387	-10.9902	-771.752
Burr	6.07629	-11.9848	-727.107

N = 200

Distribution	SSE	AIC	BIC
Gen. Normal	8.78748	4.07896	-2297.3
Tukey-Lambda	8.98852	2.1501	-2284.68
t	9.09526	2.98473	-2278.09
Burr	8.99487	-1.39616	-2277.96
Johnson's SU	9.19613	4.92281	-2265.61
Log-Gamma	10.3984	10.7439	-2203.38
Power Log-Norm.	10.514	11.3172	-2190.89
Power Normal	10.7666	10.3461	-2183.96
Beta	10.9957	11.6552	-2165.89
Exp. Weibull	11.0844	10.8621	-2161.41
Inv. Gamma	12.4247	9.03188	-2104.04
Gauss. Hyperg.	12.0971	33.7584	-2099.97
Alpha	13.4858	8.78963	-2058.31
Recip. Inv. Gauss.	14.0741	8.66918	-2034.48
Inv. Gauss	17.6787	9.35324	-1907.24

N = 300

Distribution	SSE	AIC	BIC
Tukey-Lambda	12.5924	0.64088	-4488.54
Gen. Normal	12.8083	1.98134	-4471.11
t	12.8454	1.27961	-4468.15
Johnson's SU	13.1865	2.62473	-4434.35
Power Log-Norm.	14.4129	10.2945	-4343.2
Log-Gamma	14.5249	11.6866	-4342.2
Power Normal	14.809	10.2027	-4322.35
Exp. Weibull	14.7536	10.3641	-4319.25
Beta	14.871	10.9175	-4311.13
Recip. Inv. Gauss.	17.4536	9.09994	-4153.93
Inv. Gamma	18.3134	6.57043	-4104.64
Alpha	18.4887	5.38636	-4094.87
Gauss. Hyperg.	18.7538	26.654	-4059.48
Inv. Gauss	22.2507	4.96811	-3905.03
Burr	68.052	16.8174	-2752.25

N = 400

Distribution	SSE	AIC	BIC
Gen. Normal	12.4157	5.88395	-7192.63
Tukey-Lambda	12.7894	3.26787	-7148.02
t	12.8306	3.76221	-7143.18
Johnson's SU	13.2156	5.94962	-7091.41
Exp. Weibull	14.7045	20.5342	-6930.84
Burr	15.2824	-5.17353	-6872.87
Power Log-Norm.	15.2903	20.9012	-6872.09
Log-Gamma	15.5631	19.9141	-6852.81
Power Normal	15.6312	19.1429	-6846.24
Beta	15.8398	22.1359	-6818.98
Recip. Inv. Gauss.	16.0308	27.0338	-6808.27
Gauss. Hyperg.	15.8223	34.556	-6806.01
Alpha	16.0598	24.8189	-6805.55
Inv. Gamma	18.1532	22.4764	-6621.27
Inv. Gauss	22.9908	29.6277	-6265.96

N = 500

Distribution	SSE	AIC	BIC
Gen. Normal	9.32267	-23.0793	-10867.1
t	9.38722	-22.9484	-10853.1
Tukey-Lambda	9.55679	-22.2308	-10816.9
Johnson's SU	10.0326	-21.6542	-10710.9
Log-Gamma	10.9057	-15.2731	-10549.6
Power Log-Norm.	10.9801	-16.0517	-10528.2
Power Normal	11.4773	-16.328	-10446.2
Exp. Weibull	11.6752	-15.8484	-10404
Beta	13.8055	-17.7309	-10064.8
Inv. Gamma	16.8099	-20.445	-9673.86
Alpha	17.9864	-20.7299	-9536.95
Recip. Inv. Gauss.	19.1989	-19.969	-9404.91
Inv. Gauss	29.5227	-19.0599	-8533.97
Burr	100.943	-15.5626	-6038.05
Gauss. Hyperg.	162.146	-30.3187	-5063.58

N = 600

Distribution	SSE	AIC	BIC
Gen. Normal	12.462	-33.1266	-13412.8
Tukey-Lambda	13.0359	-33.4966	-13298.9
Johnson's SU	13.4368	-31.2566	-13214.5
t	13.5884	-32.9675	-13194
Burr	14.6607	-39.1429	-12994
Exp. Weibull	18.2026	-22.1787	-12446.8
Power Normal	19.024	-23.2438	-12343
Power Log-Norm.	18.9915	-20.5447	-12339.5
Beta	19.1935	-20.901	-12312.7
Log-Gamma	19.5891	-23.731	-12269
Inv. Gamma	25.7691	-14.9643	-11575.5
Alpha	26.7744	-18.2126	-11478.7
Recip. Inv. Gauss.	26.9148	-17.7736	-11465.5
Inv. Gauss	37.7316	-11.345	-10611.1
Gauss. Hyperg.	49.0766	1.35603	-9922.8

N = 700

Distribution	SSE	AIC	BIC
Gen. Normal	14.3557	-36.6092	-16123.2
Tukey-Lambda	18.6676	-35.3677	-15330.3
Johnson's SU	19.9674	-33.1048	-15119
t	20.3356	-34.8525	-15071.9
Exp. Weibull	24.9943	-24.234	-14441.1
Power Log-Norm.	25.1674	-21.9215	-14420.3
Power Normal	25.651	-25.0606	-14370.9
Beta	26.0606	-23.1942	-14315
Log-Gamma	26.224	-25.5925	-14304.2
Gauss. Hyperg.	31.8569	-13.4848	-13692.7
Inv. Gamma	33.2246	-25.7199	-13589.8
Alpha	38.8004	-26.9686	-13121.4
Recip. Inv. Gauss.	39.839	-23.8264	-13041.7
Inv. Gauss	56.5465	-22.5139	-11984.4
Burr	190.672	19.8371	-8306.81

N = 800

Distribution	SSE	AIC	BIC
Gen. Normal	8.99334	-44.4587	-22353.2
Tukey-Lambda	9.08229	-43.6424	-22316.6
Power Log-Norm.	9.14517	-40.4707	-22282.8
Johnson's SU	9.24153	-41.8294	-22243.8
Exp. Weibull	9.34004	-41.015	-22204.5
Power Normal	9.41443	-42.7588	-22183.2
Log-Gamma	9.53025	-42.7505	-22137.8
t	9.53837	-43.5577	-22134.6
Beta	9.53216	-40.4146	-22128.8
Inv. Gamma	19.7923	-34.7967	-19422.8
Recip. Inv. Gauss.	23.9371	-34.3492	-18716.4
Alpha	24.4301	-36.7679	-18640.7
Inv. Gauss	37.7794	-28.1335	-17021.1
Burr	167.13	19.7231	-11488.7
Gauss. Hyperg.	306.992	-91.5254	-9213.34

N = 900

Distribution	SSE	AIC	BIC
Burr	9.91336	-44.2426	-24549.3
Gen. Normal	15.9748	-39.3735	-22609.5
t	18.7491	-39.3833	-21955.6
Power Log-Norm.	20.619	-24.275	-21559.2
Beta	21.077	-23.892	-21469.5
Power Normal	21.16	-27.2589	-21461.7
Exp. Weibull	22.2717	-26.3456	-21244.4
Log-Gamma	22.621	-26.7506	-21189.1
Johnson's SU	24.5602	-25.0164	-20845
Gauss. Hyperg.	27.5809	-13.8769	-20354.8
Recip. Inv. Gauss.	42.0412	-22.4398	-18658.6
Alpha	46.8523	-24.6973	-18216.2
Inv. Gamma	51.3184	-24.9556	-17844.5
Inv. Gauss	63.4447	-19.0355	-16978.4
Tukey-Lambda	363.258	-10.6482	-9853.77

N = 1000

Distribution	SSE	AIC	BIC
Johnson's SU	14.7139	42.0807	-26717.3
Gen. Normal	17.2395	64.04	-25989.5
Tukey-Lambda	17.928	53.8549	-25807.5
t	18.2031	46.822	-25736.7
Power Log-Norm.	18.8339	100.007	-25569.9
Power Normal	19.4675	93.2281	-25424.5
Exp. Weibull	20.3321	89.3591	-25214.1
Beta	20.5837	98.4655	-25156.9
Log-Gamma	20.9575	98.0658	-25081.8
Burr	21.5468	17.2264	-24944.4
Recip. Inv. Gauss.	44.1636	201.747	-21617.1
Inv. Gamma	50.5235	204.288	-20991.8
Alpha	54.1463	182.459	-20669.9
Inv. Gauss	61.7168	232.47	-20061.6
Gauss. Hyperg.	784.317	-93.0616	-8219.89

Figure 4.34: BIC and AIC for the fitted distributions with the lowest SSE on the Birth times of the H_2 PH bars of $N \in \{100, 200, 300, 400, 500, 600, 700, 800, 900, 1000\}$ points in $[0,1]^3$.

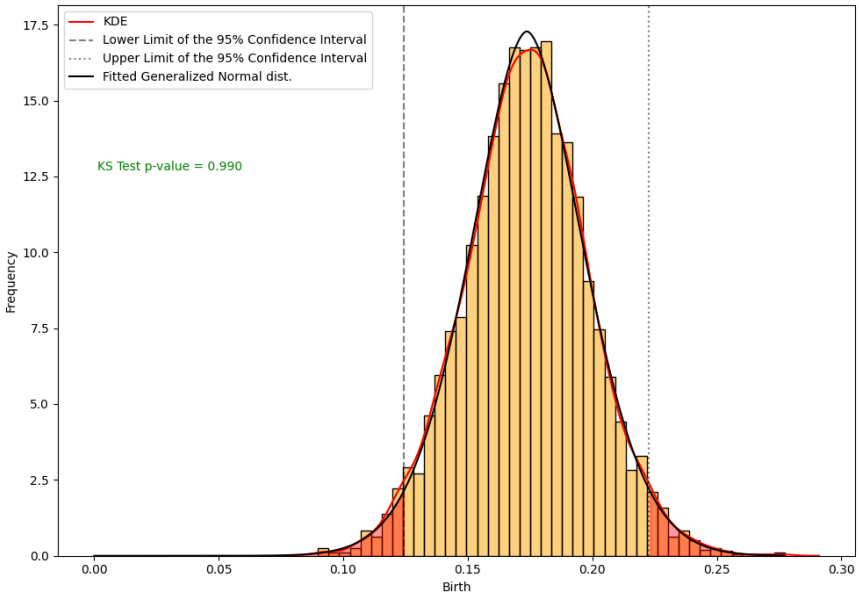
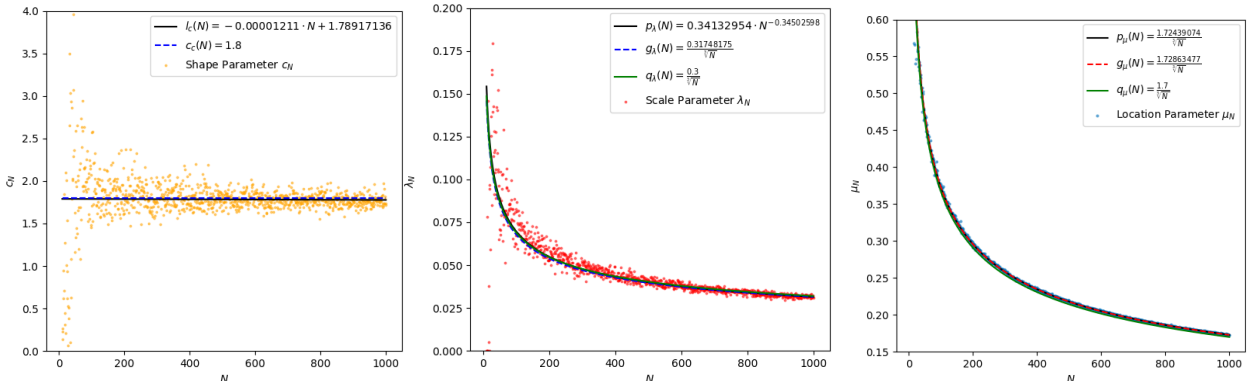


Figure 4.35: Fitted distribution on the normalized histogram of the Birth times of the H_2 PH bars in $[0,1]^3$ ($N = 1000$; 40 simulations).

$$p_\lambda(N) = 0.34132954 \cdot N^{-0.34502598} \approx \frac{0.3}{\sqrt[3]{N}}, \tag{4.50}$$

and

$$p_\mu(N) = \frac{1.72439074}{\sqrt[3]{N}} \approx \frac{1.7}{\sqrt[3]{N}}. \tag{4.51}$$



(a) (b) (c)

Figure 4.36: Fitted models on the scatter plots of the estimated shape c (a), scale λ (b), and location μ (c) parameters of the Generalized Normal distribution fit for the Birth times of the H_2 PH bars of $N \in [2, 1000]$ points uniformly distributed in $[0, 1]^3$.

Death Times

N = 100

Distribution	SSE	AIC	BIC
Gen. Normal	4.2378	-7.68462	-952.341
Power Normal	4.54978	-7.5347	-935.293
Normal	4.72346	-9.29116	-931.782
Log-Gamma	4.62107	-7.48452	-931.561
Exp. Weibull	4.52147	-5.52557	-931.31
Pearson Type III	4.62772	-7.47661	-931.216
Beta	4.6122	-5.57813	-926.541
Nakagami	4.73877	-7.23692	-925.525
Log-Normal	4.80091	-7.02502	-922.398
Fatigue Life	4.81886	-7.33741	-921.503
Johnson's SB	4.72556	-5.30173	-920.714
Beta Prime	4.8732	-5.30365	-913.33
Alpha	5.95794	-6.82316	-870.577
Recip. Inv. Gauss.	6.2311	-6.76232	-859.819
Chi-Squared	196.532	-8.06682	-31.5129

N = 200

Distribution	SSE	AIC	BIC
Gen. Normal	6.85479	-8.96825	-2740.7
Normal	7.945	-8.30544	-2656.5
Log-Gamma	7.91859	-6.69409	-2652.12
Nakagami	8.127	-5.35331	-2636.17
Pearson Type III	8.14386	-5.27396	-2634.9
Fatigue Life	8.14612	-5.26632	-2634.73
Log-Normal	8.14964	-5.23514	-2634.46
Power Normal	8.17103	-5.39475	-2632.85
Exp. Weibull	8.1011	-3.97101	-2631.71
Johnson's SB	8.14565	-3.3166	-2628.34
Beta Prime	8.14663	-3.249	-2628.27
Beta	8.22079	-3.32013	-2622.7
Alpha	9.88981	-2.44458	-2515.63
Recip. Inv. Gauss.	10.4051	-1.07862	-2484.45
Chi-Squared	517.049	-18.7847	-86.2609

N = 300

Distribution	SSE	AIC	BIC
Gen. Normal	9.03285	-24.1645	-7585.52
Exp. Weibull	9.52702	-20.9795	-7498.85
Normal	9.67374	-24.5247	-7490.69
Log-Normal	9.64325	-22.4239	-7488.09
Pearson Type III	9.64367	-22.5434	-7488.02
Nakagami	9.64724	-22.5329	-7487.47
Fatigue Life	9.65535	-22.5586	-7486.22
Power Normal	9.67853	-22.4605	-7482.65
Beta	9.64711	-20.5452	-7480.18
Johnson's SB	9.65078	-20.5368	-7479.62
Log-Gamma	9.84216	-23.0093	-7457.67
Beta Prime	10.0057	-20.4533	-7425.81
Recip. Inv. Gauss.	11.7929	-19.8452	-7188.24
Alpha	12.1864	-21.3534	-7139.34
Chi-Squared	12.2665	-22.8803	-7129.58

N = 400

Distribution	SSE	AIC	BIC
Gen. Normal	9.03285	-24.1645	-7585.52
Exp. Weibull	9.52702	-20.9795	-7498.85
Normal	9.67374	-24.5247	-7490.69
Log-Normal	9.64325	-22.4239	-7488.09
Pearson Type III	9.64367	-22.5434	-7488.02
Nakagami	9.64724	-22.5329	-7487.47
Fatigue Life	9.65535	-22.5586	-7486.22
Power Normal	9.67853	-22.4605	-7482.65
Beta	9.64711	-20.5452	-7480.18
Johnson's SB	9.65078	-20.5368	-7479.62
Log-Gamma	9.84216	-23.0093	-7457.67
Beta Prime	10.0057	-20.4533	-7425.81
Recip. Inv. Gauss.	11.7929	-19.8452	-7188.24
Alpha	12.1864	-21.3534	-7139.34
Chi-Squared	12.2665	-22.8803	-7129.58

N = 500

Distribution	SSE	AIC	BIC
Gen. Normal	10.3322	-39.9	-10489.7
Normal	10.4974	-41.6834	-10465.6
Log-Gamma	10.5777	-39.7469	-10442.8
Pearson Type III	10.6573	-40.2175	-10427.8
Fatigue Life	10.6814	-40.2692	-10423.3
Nakagami	10.6818	-40.2233	-10423.2
Power Normal	10.7036	-40.0808	-10419.2
Beta Prime	10.669	-38.2533	-10418
Log-Normal	10.7166	-40.3313	-10416.7
Johnson's SB	10.9368	-37.8447	-10368.5
Beta	10.9647	-37.8258	-10363.4
Exp. Weibull	11.3835	-37.3349	-10288.6
Recip. Inv. Gauss.	12.937	-41.5733	-10040.7
Chi-Squared	14.121	-42.2422	-9865.83
Alpha	14.7576	-41.9078	-9777.77

N = 600

Distribution	SSE	AIC	BIC
Normal	3.36425	-11.4114	-16526.7
Gen. Normal	3.35956	-9.23829	-16522.3
Power Normal	3.42666	-9.6303	-16472.9
Exp. Weibull	3.42802	-7.03051	-16464
Nakagami	3.44288	-10.0723	-16461
Log-Gamma	3.44831	-8.80684	-16457.1
Pearson Type III	3.44982	-10.2172	-16456
Log-Normal	3.45156	-10.2576	-16454.7
Johnson's SB	3.44524	-8.1477	-16451.5
Fatigue Life	3.46161	-10.1852	-16447.5
Beta	3.48138	-8.36356	-16425.4
Chi-Squared	6.16119	-12.7158	-15005
Alpha	7.27966	-11.7448	-14587.6
Recip. Inv. Gauss.	7.43663	-10.3659	-14534.2
Beta Prime	36.9285	-4.70733	-10516.8

N = 700

Distribution	SSE	AIC	BIC
Gen. Normal	5.33801	-33.9605	-18802
Normal	5.47709	-33.5637	-18733.4
Log-Gamma	5.47989	-31.6127	-18723.9
Power Normal	5.63045	-31.8202	-18643.2
Johnson's SB	5.66316	-29.9762	-18618
Nakagami	5.69428	-32.1492	-18609.7
Pearson Type III	5.70931	-32.2159	-18601.8
Exp. Weibull	5.72194	-30.492	-18587.2
Beta	5.88472	-30.5416	-18503.7
Log-Normal	5.92179	-32.4456	-18493
Fatigue Life	6.04923	-32.6837	-18429.7
Beta Prime	6.48976	-31.2508	-18212.4
Chi-Squared	10.3468	-36.1396	-16831.8
Recip. Inv. Gauss.	11.0314	-33.8078	-16641
Alpha	11.9938	-35.2579	-16392

N = 800

Distribution	SSE	AIC	BIC
Gen. Normal	7.17996	-18.7054	-22249.8
Log-Gamma	7.44826	-21.5695	-22118.3
Normal	7.57061	-23.4565	-22068.1
Beta	7.70426	-18.3943	-21989
Johnson's SB	7.70446	-18.4465	-21988.9
Power Normal	7.82822	-22.4798	-21939.9
Pearson Type III	7.86655	-23.069	-21922.4
Nakagami	7.87772	-23.0255	-21917.3
Fatigue Life	7.9886	-23.4983	-21867.2
Log-Normal	8.0269	-23.5177	-21850.1
Alpha	12.7453	-26.8876	-20192.5
Chi-Squared	12.7559	-25.8846	-20189.5
Recip. Inv. Gauss.	12.9926	-25.7426	-20123.6
Beta Prime	74.8178	-29.9807	-13839.3
Exp. Weibull	1876.61	-59.395	-2287.8

N = 900

Distribution	SSE	AIC	BIC
Gen. Normal	7.72975	-18.1496	-26447.5
Johnson's SB	8.41918	-16.7489	-26080
Beta	8.42094	-16.8043	-26079.2
Log-Normal	8.53201	-22.7927	-26032.4
Nakagami	8.54478	-22.8278	-26026.1
Normal	8.59676	-24.6348	-26009
Fatigue Life	8.62088	-22.5699	-25988.9
Pearson Type III	8.78312	-22.2223	-25910.5
Power Normal	8.7905	-22.6377	-25907
Log-Gamma	8.84681	-22.4596	-25880.1
Exp. Weibull	9.88688	-23.7709	-25404.6
Chi-Squared	13.0594	-22.5699	-24243.3
Alpha	14.3413	-22.4759	-23849.7
Recip. Inv. Gauss.	14.4608	-20.0975	-23814.9
Beta Prime	101.563	-19.6051	-15613.9

N = 1000

Distribution	SSE	AIC	BIC
Gen. Normal	11.5144	81.7996	-27865.4
Log-Gamma	11.7834	71.5443	-27758.1
Pearson Type III	11.8159	72.4329	-27745.3
Johnson's SB	11.801	75.0063	-27742.7
Beta	11.8148	74.3452	-27737.3
Normal	11.9125	71.7603	-27707.5
Power Normal	12.009	73.2478	-27678.4
Nakagami	12.1208	76.3697	-27626.9
Beta Prime	12.2517	79.8971	-27568.5
Log-Normal	12.2964	77.7271	-27560
Exp. Weibull	12.332	67.8848	-27538.1
Fatigue Life	12.5345	79.5855	-27470.9
Recip. Inv. Gauss.	19.4493	122.553	-25428.9
Alpha	21.8552	114.21	-24886.8
Chi-Squared	1788.98	-106.546	-4412.53

Figure 4.37: BIC and AIC for the fitted distributions with the lowest SSE on the Death times of the H_2 PH bars of $N \in \{100, 200, 300, 400, 500, 600, 700, 800, 900, 1000\}$ points in $[0, 1]^3$.

The parameters of the Generalized Normal distribution for the Death times (Figure 4.39) can be described in a similar way to the ones of the Birth times:

$$l_c(N) = 0.00001387 \cdot N + 2.02891051 \approx 2, \tag{4.52}$$

$$p_\lambda(N) = 0.37358781 \cdot N^{-0.32546314} \approx \frac{1.84}{\sqrt[3]{N}}, \tag{4.53}$$

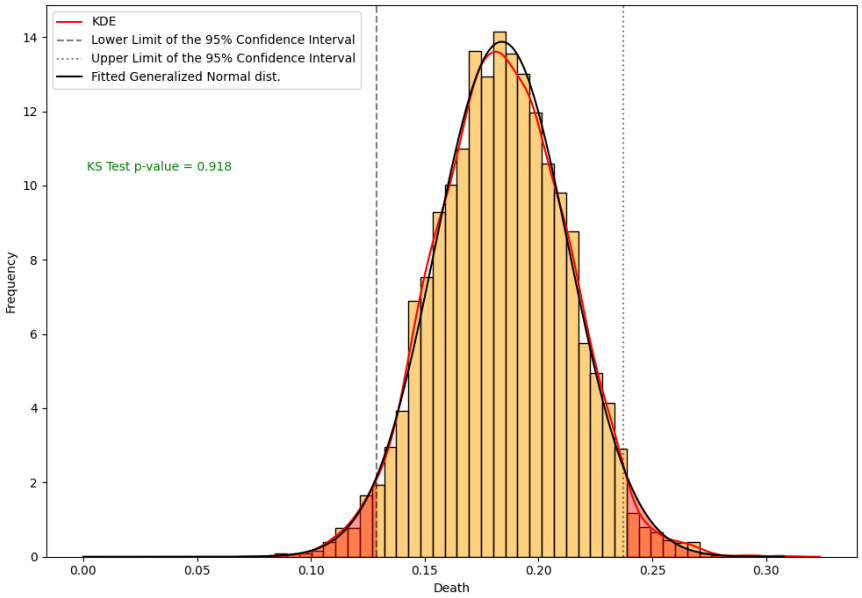
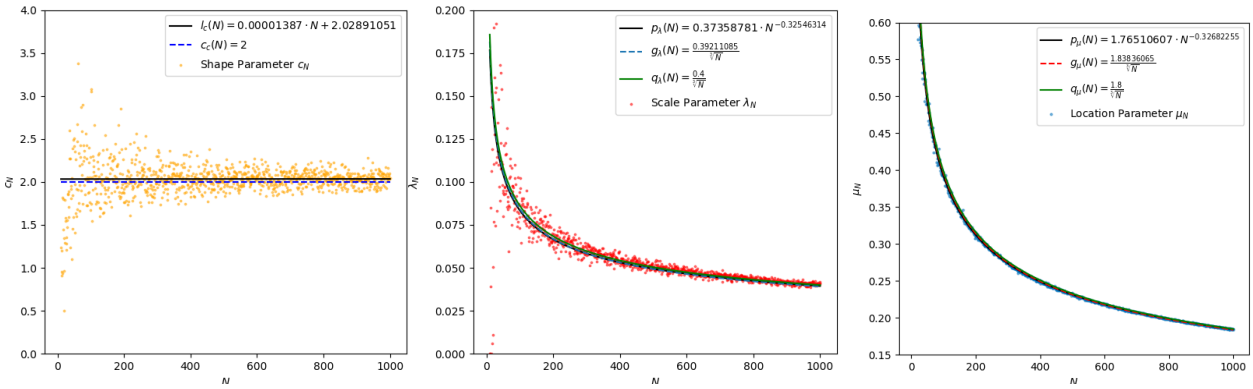


Figure 4.38: Fitted distribution on the normalized histogram of the Death times of the H_2 PH bars in $[0,1]^3$. ($N = 1000$; 40 simulations).

and

$$f_\lambda(N) = \frac{0.39211085}{\sqrt[3]{N}} \approx \frac{0.4}{\sqrt[3]{N}}. \tag{4.54}$$



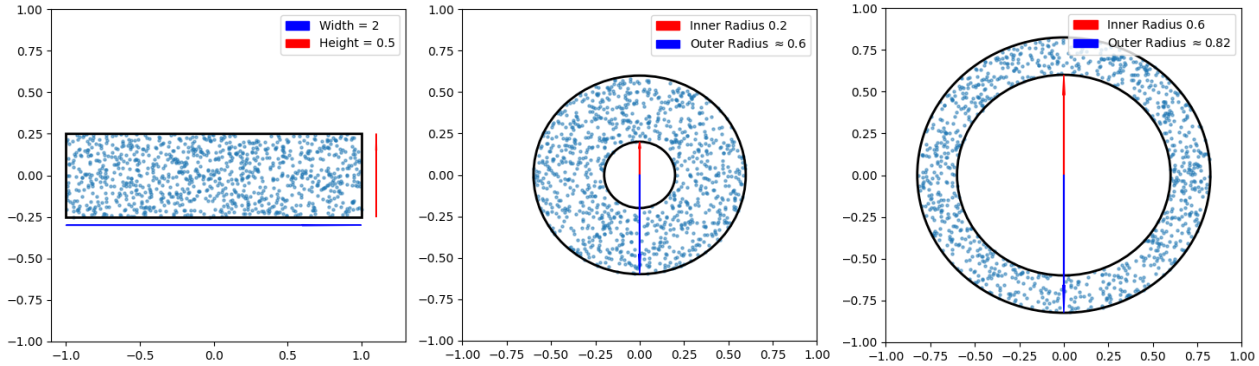
(a) (b) (c)

Figure 4.39: Fitted models on the scatter plots of the estimated shape c (a), scale λ (b), and location μ (c) parameters of the Generalized Normal distribution fit for the Death times of the H_2 PH bars of $N \in [2, 1000]$ points uniformly distributed in $[0,1]^3$.

It is worth paying attention to (4.52): the linear model l_c could be approximated to $c_c(N) = 2$ since the slope of this function is relatively low. However, a Generalized Normal distribution with shape parameter $c = 2$ is identical to the Normal distribution where the mean is the location parameter μ and the variance is $\frac{\lambda^2}{2}$ where λ is the location parameter of

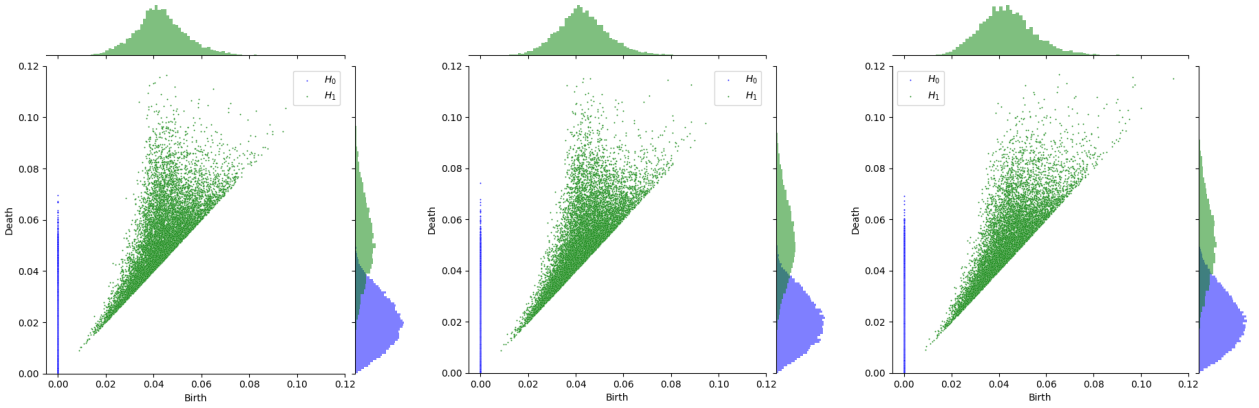
the Generalized Normal Distribution. This assumption is additionally supported by the normality test presented in Section 4.3.2, and the Normal distribution is among the distributions providing the lowest SSE values.

4.8 Discussion



(a) (b) (c)

Figure 4.40: 1000 points uniformly distributed in a rectangle (width w=2, height h=0.5) (a), a *small* annulus (inner radius r=0.2, outer radius R=0.6) (b), and a *large* annulus (inner radius r=0.6, outer radius R≈0.82)(c).

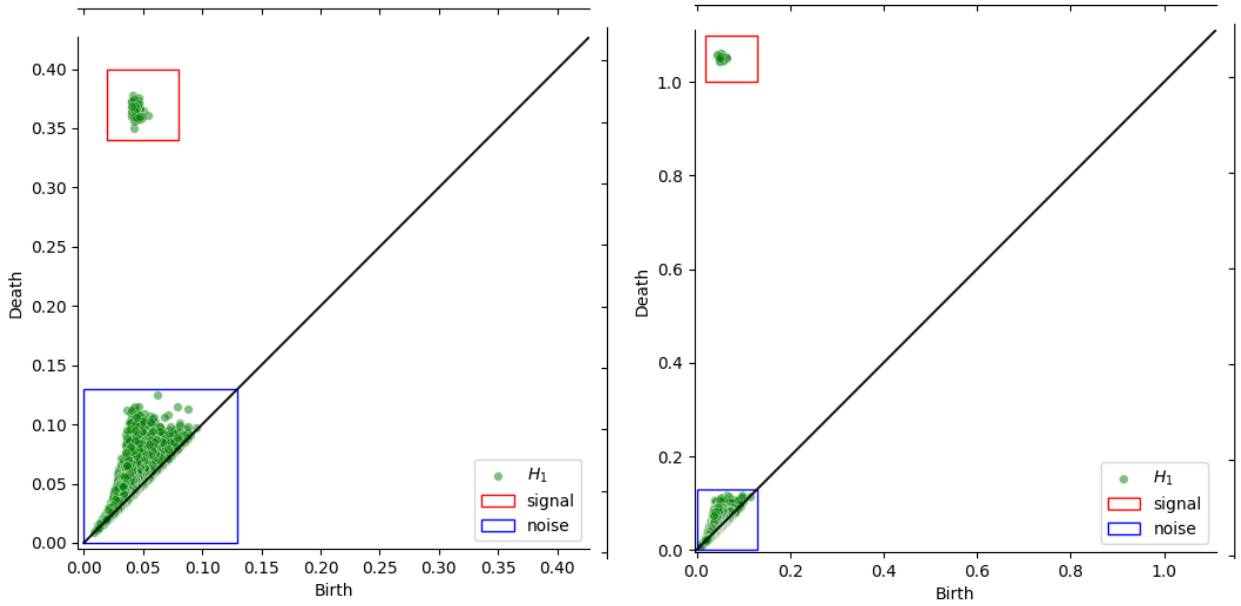


(a) (b) (c)

Figure 4.41: Persistence diagrams of noise bars with marginal histograms for each homological dimension for 40 simulations of $N = 1000$ points uniformly distributed in a rectangle (a), a *small* annulus (b), and a *large* annulus (c).

In this master's thesis, we investigated the distributions of birth and death times of the PH bars of uniformly distributed points within three different regions: a unit interval, a unit square, and a unit cube. A key question is to understand whether the identified distributions

change when we sample points uniformly from regions different from the ones mentioned above while keeping the regions' areas fixed.



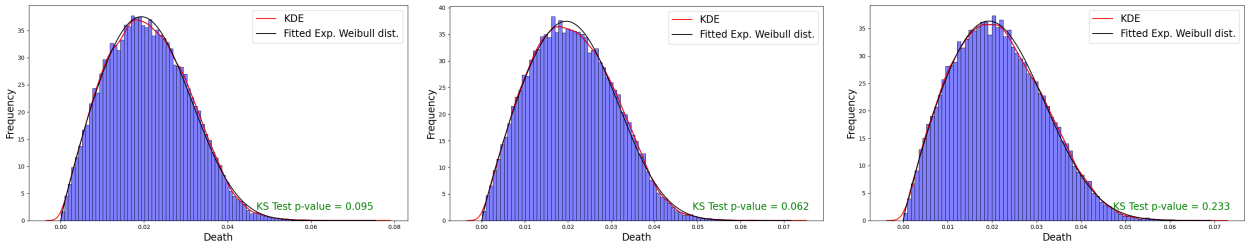
(a) (b)

Figure 4.42: Persistence diagrams of the H_1 PH bars for 40 simulations of $N = 1000$ points uniformly distributed in a *small* annulus (a), and a *large* annulus (b).

To explore this idea, we consider regions in a two-dimensional space. Intuitively, the distributions of the PH bars of points sampled from a rectangle (i.e., obtained by shrinking and stretching opposite sides of a unit square while maintaining its area) should not differ substantially from those of the PH bars of points sampled from a unit square. However, suppose we deform the square into a narrow rectangle with an infinitesimally small height; in that case, we suspect the distributions to be more related to that of points uniformly sampled from an interval. Similarly, when we select a region, such as an annulus with a relatively small radius and area 1, and consider only the noise bars, we predict that its persistent homology will be similar to that of points uniformly distributed in the unit square. Conversely, if we choose an annulus with an infinitesimally large interior radius and area fixed, we anticipate that the persistent homology will exhibit notable changes. The fundamental notion is that minor alterations to the shape of a region should yield relatively minor discrepancies in the results obtained from persistent homology.

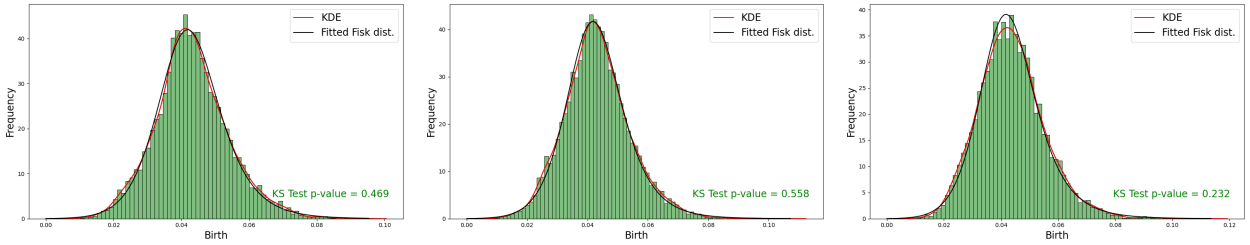
In this section, we aim to compare birth and death times distributions by simulating $N = 1000$ points uniformly sampled from two-dimensional regions with a surface area of 1. Specifically, we deform the unit square by stretching and shrinking its opposite sides into a rectangle (width $w = 2$, height $h = 0.5$), and we investigate two different annuli, one with

an interior radius of $r = 0.2$ and the other with an interior radius of $r = 0.6$ (Figures 4.40a, 4.40b, and 4.40c).



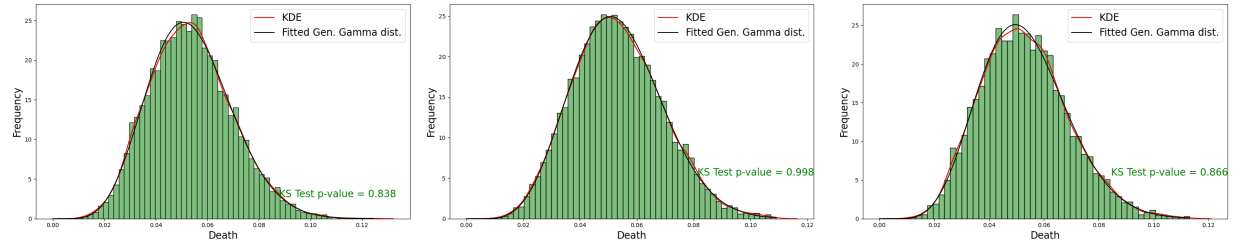
(a) (b) (c)

Figure 4.43: Fitted Exponentiated Weibull distribution on the normalized histograms (40 simulations) of the Death times of the H_0 PH bars of 1000 points uniformly distributed in a rectangle (a), a *small* annulus (b), and a *large* annulus (c).



(a) (b) (c)

Figure 4.44: Fitted Fisk distribution on the normalized histograms (40 simulations) of the Birth times of the H_1 PH bars of 1000 points uniformly distributed in a rectangle (a), a *small* annulus (b), and a *large* annulus (c).



(a) (b) (c)

Figure 4.45: Fitted Generalized Gamma distribution on the normalized histograms (40 simulations) of the Death times of the H_1 PH bars of 1000 points uniformly distributed in a rectangle (a), a *small* annulus (b), and a *large* annulus (c).

Following the methodologies described in Chapter 3, we collected 40 samples for each region, computed their persistent homology, and created persistence diagrams. The persistence

diagrams are presented in Figure 4.41 alongside the marginal distributions (H_0 Death, H_1 Birth, and H_1 Death). However, it is essential to note that the graphs of the two annuli only depict the noise bars and not the signal since the bars associated with the one-dimensional hole are far away from the diagonal, leading to difficulties in visual inspection (Figure 4.42). In persistent homology, separating noise from the signal is a significant challenge. The heuristic method identifies the furthest points from the diagonal as the signal. For the two annuli, this separation was relatively more straightforward due to specific characteristics: each annulus contains a single one-dimensional hole, and all persistence diagrams showed one H_0 bar away from the diagonal. Additionally, overlapping 40 persistence diagrams revealed that the signal clustered consistently in the same region for each simulation. Furthermore, the PH bars associated with the one-dimensional holes of the 40 samples of small and large annuli displayed a notable difference: Figure 4.42b's signal persists longer than 4.42a's signal; this difference is attributed to the size of the inner radii of the two annuli.

A first visual analysis suggests that the H_1 PH bars are clustered similarly to the H_1 PH bars for the unit square, as depicted in Figure 4.1. More precisely, the marginal histograms of all three regions show similarities to the marginal histograms of the unit square (Figure 4.5b). We aim not to identify the exact statistical distribution for each case as we did for the unit square. However, we want to verify whether the distributions identified in Section 4.6 are suitable for the birth and death times of the PH bars of points sampled from the rectangle and the annuli.

More precisely, we fitted the Exponentiated Weibull, Fisk, and Generalized Gamma distributions using the MLE method to the normalized histograms of the H_0's Death, H_1's Birth, and H_1's Death times PH bars, respectively (Figures 4.43, 4.44, and 4.45). These distributions provide a good fit for these datasets; more precisely, using p-values of the Kolmogorov-Smirnov test performed for each dataset, we fail to reject the null hypothesis, indicating that the distributions used to describe the birth and death times of the PH bars of a unit square can also be used to describe the PH bars of these spaces. To conclude, our additional findings suggest that the distributions of the persistent homology bars of a unit square remain relatively consistent for certain deformations of the region while keeping the area fixed.

4.9 Summary

This chapter built upon the experimental methodologies outlined in Chapter 3 to present results and provide an analysis. We distributed a variable number of points N (ranging from 2 to 1000) uniformly across a unit interval, a unit square, and a unit cube. These simulations

were repeated 40 times, and the persistent homology was then computed in Python using the `ripser` library. This process resulted in 40 samples of persistent homology bars in various homological dimensions depending on the space. The birth and death times were accounted for separately, except for the H_0 homological bars in all spaces, where birth times are always zero.

For our statistical analysis, we considered the union of the 40 samples for each value of N, homological dimension, and space for birth and death parameters. However, our research recognized its limitations concerning the independent and identically distributed (iid) nature of the PH bars; although the independence of the birth and death times across different simulations is apparent, dependencies could arise within a single simulation due to the spatial distribution of the points.

A preliminary data visualization of the birth and death times was performed using persistence diagrams (Figures 4.1, 4.2, and 4.3). This analysis indicated differing trends across homological dimensions and spaces. The PH bars appeared to cluster along the persistence diagrams' diagonals, a pattern amplified as N increased. This phenomenon led us to question the effect of N on the number of persistent homology bars and their statistical distribution.

Firstly, we investigated the persistent homology bars, focusing on their relationship with the number N of uniformly distributed points. A linear increase was noted in the number of bars with N, and we identified specific models for different homological dimensions and spaces (Table 4.4). Notable variations were observed between square and cube spaces. Interestingly, the number of H_0 PH bars exhibited a similar linear increase across all spaces. For example, the number of PH bars in the unit square and cube generally decreased as the homological dimension of the bars increased for a fixed N. However, the number of PH bars in the cube was higher than in the square and approximately half the number of H_0 bars in the cube for large N. We documented these relationships in Table 4.1.

We further performed a visual analysis of the underlying distributions that described each death and birth time. The birth and death time histograms (Figure 4.8) displayed unique statistical distributions across the three spaces and homological dimensions. For example, the histogram of the H_0 death bars in the unit interval suggested an Exponential distribution, whereas others demonstrated bell-shaped curves with varying skewness and kurtosis.

To gain a more comprehensive understanding of the theoretical distributions of the birth and death times and their relationship with the persistence diagrams, we derived the joint kernel density estimates plots (Figures 4.6 and 4.7), facilitating a deeper grasp of the joint distributions of the birth and death times. Although the joint KDE plots were primarily used for visual inspection, they offered valuable insights.

We also conducted a normality test using the D'Agostino's K-squared test (Table 4.2)

since specific histograms resembled the shape of a Normal distribution. More precisely, a p-value less than 0.05 indicated a deviation from the Normal distribution. The null hypothesis was not rejected only for the death times of the H_2 PH bars, suggesting a potential normal distribution. However, we acknowledged the inconclusiveness of such statistical tests. These tests were repeated for different N values, yielding similar results.

In the subsequent sections, we presented a statistical analysis. First, using the Maximum Likelihood Estimation method, we fitted many distributions (Table A.2) to each dataset. Then, only the distributions with the lowest Sum Square Error were reported, complete with their respective values for the Bayesian Information Criterion and Akaike Information Criterion. Consequently, for each *best* distribution (primarily focusing on the AIC values), we provided its probability density function, general properties, relations to other distributions, and a fit on the normalized histogram of the respective dataset alongside the KDE.

Specifically, the death times of the H_0 PH bars in the unit interval were accurately described by both Exponential and Weibull distributions. The Weibull distribution also best fitted the death times of the H_1 PH bars in the unit cube. An Exponentiated Weibull distribution best suited the death times of H_0 PH bars in the unit square. The death times of the H_1 PH bars followed a Generalized Gamma distribution (which encompasses Weibull and Exponential distributions as subfamilies). The birth times of the H_1 PH bars in the unit square conformed to either a Fisk or a Mielke distribution. For the H_1 PH bars in the unit cube, a Burr (Type III) distribution proved to be the best fit, effectively a reparametrization of the Mielke distribution. For the H_2 PH bars, we found that a Generalized Normal distribution described them, but we recommended caution due to the relatively small number of PH bars collected in this homological dimension. We provided a summary of these distributions in Table 4.3.

Distribution Fitting Results			
Space	Parameter	H_k	Distribution
$[0,1]^1$	`Death`	H_0	Exponential
$[0,1]^2$	`Death`	H_0	Exponentiated Weibull
$[0,1]^2$	`Birth`	H_1	Fisk
$[0,1]^2$	`Death`	H_1	Generalized Gamma
$[0,1]^3$	`Death`	H_0	Burr (Type III)
$[0,1]^3$	`Birth`	H_1	Generalized Logistic
$[0,1]^3$	`Death`	H_1	Weibull
$[0,1]^3$	`Birth`	H_2	Generalized Normal
$[0,1]^3$	`Death`	H_2	Generalized Normal

Table 4.3: A summary of the distributions that best fit each dataset.

While the primary goal of the thesis was not necessarily to find the best fits, we were

interested in how the parameters of a distribution changed as the number of uniformly distributed points N increased in a given space. We found that the scale and location parameters decreased as N increased, while other parameters appeared constant or asymptotically converged towards a constant. We modelled these relationships using linear, logarithmic, and exponential models, finding the power model (or Freundlich function) to be the most consistent. Specifically, the distributions' estimated location and scale parameters were described by a power model of the form $\frac{c}{\sqrt[d]{N}}$, where $c \in \mathbb{R}$, N is the number of uniformly distributed points and d is the dimension of the space.

We conclude this chapter with a proposal for future research using mixed-effects models to account for simulation-specific variations, allowing for a more nuanced understanding of the distributions. Our approach effectively captured trends across different spaces and homological dimensions, serving as a foundation for future, more complex investigations.

Chapter 5

Conclusion

This thesis has introduced topological data analysis focusing on persistent homology. Our discussion covered the mathematical foundations of persistent homology, including simplicial complexes, abstract simplicial complexes, and the Vietoris-Rips complex, commonly used in computational topology. We also examined the mathematical formalisms underlying persistent homology, including the structure theorem for PID, the definition of persistence modules, and the stability theorem. Furthermore, we provided an overview of the computational approaches to perform persistent homology computations and introduced a Python library, `Ripser.py`. Our primary objective was to investigate the persistent homology of noise. We conducted various simulations to explore the behaviour of persistent homology bars in different spaces and homological dimensions for various numbers of uniformly distributed points. Our analysis revealed valuable insights into the behaviour of persistent homology bars in different spaces and homological dimensions. Specifically, we observed a linear increase in the number of persistent homology bars as the number of data points increased, with differences observed between the square and cube. We also found that, on average, the number of persistent homology bars decreased as the homological dimension increased for a fixed value of N. The histograms of birth and death times for each homological dimension and space exhibited the shapes of probability density functions, with some being roughly bell-shaped. We derived the kernel density estimates (KDE) of birth and death times and created 3D plots to visualize the joint distributions of birth and death times. We found a few distributions that fit the data by evaluating the goodness of fit for each model using various statistical metrics, such as the SSE. This comparison allowed us to determine the best-fitting models for different scenarios and understand the limitations of our approximations. Then, we estimated their parameters using linear and power models that describe their relationship with the number of data points in the space and homological dimensions.

Despite the progress made in this thesis, there are still several areas for future research.

The computation of persistent homology in this thesis was limited to 1000 points. Increasing the number of points would likely yield more accurate results; however, doing so is computationally expensive, necessitating algorithmic optimizations to maintain efficiency. Additionally, it would be intriguing to investigate how the persistent homology bars change when points are distributed in different spaces. In this study, we employed a uniform distribution for the point cloud; exploring how the bars behave with alternative distributions would further enhance our understanding of the impact of various distributions on persistent homology.

In conclusion, this thesis has provided a solid foundation for understanding the persistent homology of noise. Through the analysis of the simulated data, we have gained valuable insights into the behaviour of persistent homology bars in different spaces and homological dimensions. These findings can serve as a starting point for further research and contribute to developing more advanced tools and techniques in topological data analysis.

Appendix A

Python Information

We conducted experiments using Python version 3.9.0 on a Windows 11 machine with 16GB of memory and 64-bit architecture. The necessary packages were installed using the `pipenv` package manager, and the codes were developed using Visual Studio Code. These details are provided for the sake of reproducibility and to facilitate the understanding of our experimental results. We present two tables with information on our experiment's Python packages. The first table lists the packages employed, their version numbers, and a brief description of their purpose. The second table lists the statistical distributions and their corresponding names in the `scipy` library. Moreover, we provide an example script that uses the `ripser` package to compute the persistent homology of a point cloud. This script is a useful starting point for those seeking to utilize `ripser` for their projects. For more detailed information on `ripser`, we refer the reader to the article *Ripser: efficient computation of Vietoris-Rips persistence barcodes* by Ulrich Bauer [3] and to the GitHub page `https://ripser.scikit-tda.org/en/latest`.

Distribution name	`scipy.stats` **function name**
Alpha	`alpha`
Anglit	`anglit`
Arcsine	`arcsine`
Beta	`beta`
Gibrat	`gilbrat`
Laplace (Double Exponential, Bilateral Exponential)	`laplace`
Log-Normal (Cobb-Douglass)	`lognorm`
	Continued on next page

Distribution name	`scipy.stats` **function name**
Pearson Type III	`pearson3`
Rice	`rice`
Tukey-Lambda	`tukeylambda`
Wrapped Cauchy	`wrapcauchy`
Generalized Exponential	`genexpon`
Gamma	`gamma`
Log-Gamma	`loggamma`
Reciprocal	`reciprocal`
Beta Prime	`betaprime`
Bradford	`bradford`
Burr	`burr`
Cauchy	`cauchy`
Chi	`chi`
Gompertz (Truncated Gumbel)	`gompertz`
Hyperbolic Secant	`hypsecant`
Lévy	`levy`
Pareto Second Kind (Lomax)	`lomax`
Power-function	`powerlaw`
Reciprocal Inverse Gaussian	`recipinvgauss`
Uniform	`uniform`
Generalized Gamma	`gengamma`
Generalized Extreme Value	`genextreme`
Generalized Half-Logistic	`genhalflogistic`
Log-Laplace	`loglaplace`
Truncated Exponential	`truncexpon`
Chi-Squared	`chi2`
Cosine	`cosine`
Double Gamma	`dgamma`
Double Weibull	`dweibull`
Erlang	`erlang`
Gumbel	`gumbel r`
Inverted Gamma	`invgamma`
Maxwell	`maxwell`
Power Log-Normal	`powerlognorm`

Continued on next page

Distribution name	`scipy.stats` **function name**
Semicircular	`semicircular`
Von Mises	`vonmises`
Generalized Logistic	`genlogistic`
Generalized Normal	`gennorm`
Half-Logistic	`halflogistic`
Normal	`norm`
Truncated Normal	`truncnorm`
Exponential	`expon`
Exponentially Modified Normal	`exponnorm`
Exponentiated Weibull	`exponweib`
Exponential Power	`exponpow`
Fratio (or F)	`f`
Gumbel Left-Skewed	`gumbel l`
Inverse Normal (Inverse Gaussian)	`invgauss`
Mielke's Beta-Kappa	`mielke`
Power Normal	`powernorm`
Student's t	`t`
Gauss Hypergeometric	`gausshyper`
Johnson SB	`johnsonsb`
Pareto	`pareto`
Weibull Minimum Extreme Value	`weibull min`
Fatigue Life (Birnbaum-Saunders)	`fatiguelife`
Fisk (Log Logistic)	`fisk`
Folded Cauchy	`foldcauchy`
Folded Normal	`foldnorm`
Half-Cauchy	`halfcauchy`
Inverted Weibull	`invweibull`
Logistic (Sech-squared)	`logistic`
Nakagami	`nakagami`
Wald	`wald`
Half-Normal	`halfnorm`
Generalized Pareto	`genpareto`
Johnson SU	`johnsonsu`
Rayleigh	`rayleigh`

Continued on next page

Distribution name	`scipy.stats` **function name**
Weibull Maximum Extreme Value	`weibull max`

Table A.1: List of distribution names used for testing and their function name in `scipy.stats`

`Package Name`	**Version**	**Purpose**
`NumPy`	1.24.1	Numerical operations and array manipulation
`Pandas`	1.5.3	Data manipulation and analysis
`SciPy`	1.10.0	Scientific computing and statistics
`Matplotlib`	3.7.0	Creating visualizations and plots
`mpl.toolkits`	1.3.0	Additional tools and utilities for Matplotlib
`Seaborn`	0.12.2	Advanced data visualization and plotting
`Scikit-learn`	1.2.1	Machine learning and predictive modelling
`Statsmodels`	0.10.2	Statistical modelling and analysis
`ripser`	0.6.4	Computing persistent homology on data
`persim`	0.3.1	Plotting persistence diagrams
`pickle`	0.0.1	Serializing and de-serializing Python objects
`distfit`	1.6.4	Fitting probability distributions to data
`math`	N/A	Mathematical operations and functions
`enum`	N/A	Creating and working with enumerated types
`operator`	N/A	Performing various operations on Python objects
`cprofile`	N/A	Profiling and analyzing the performance of Python code

Table A.2: Name, version and usage of the python packages used for experiments.

A.1 Risper demonstration

The following code generates a point cloud sampled from a torus, computes its persistent homology, and plots the point cloud and the persistence diagram. The point cloud is generated by uniformly sampling $n = 1000$ points from the surface of the torus, using the `meshgrid` method from `numpy` to create a grid of *theta* and *phi* values. Each point's x, y, and z coordinates are computed using the standard equations for a torus in three dimensions. After generating the point cloud, the code uses the `ripser` package to compute the persistent homology of the point cloud. Finally, the code creates a plot of the point cloud and the persistence diagram using the `persim` package. The point cloud is displayed as a 3D scatter plot, with the x, y, and z coordinates of each point determining its position in space. The persistence diagram is displayed in a separate subplot, with the x-axis representing the birth time of topological features and the y-axis representing the death time of those features.

```python
import numpy as np
from ripser import ripser
from persim import plot_diagrams
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D

# Set the parameters of the torus
R = 1  # Major radius
r = 0.5  # Minor radius
n = 1000  # Number of samples

# Generate the point cloud sampled from the torus
theta = 2 * np.pi * np.random.rand(n)
phi = 2 * np.pi * np.random.rand(n)
x = (R + r * np.cos(phi)) * np.cos(theta)
y = (R + r * np.cos(phi)) * np.sin(theta)
z = 1 * np.sin(phi)
data = np.stack([x, y, z], axis=1)

# Compute the persistence diagram using Ripser
diagrams = ripser(data, maxdim=2)['dgms']

# Plot the point cloud and persistence diagram
fig = plt.figure(figsize=(10, 5))
ax1 = fig.add_subplot(121, projection='3d')
ax1.scatter(x, y, z, c='blue', s=1)
ax1.set_xlabel('X')
ax1.set_ylabel('Y')
ax1.set_zlabel('Z')
ax2 = fig.add_subplot(122)
plot_diagrams(diagrams, ax=ax2)
plt.tight_layout()
plt.show()
```

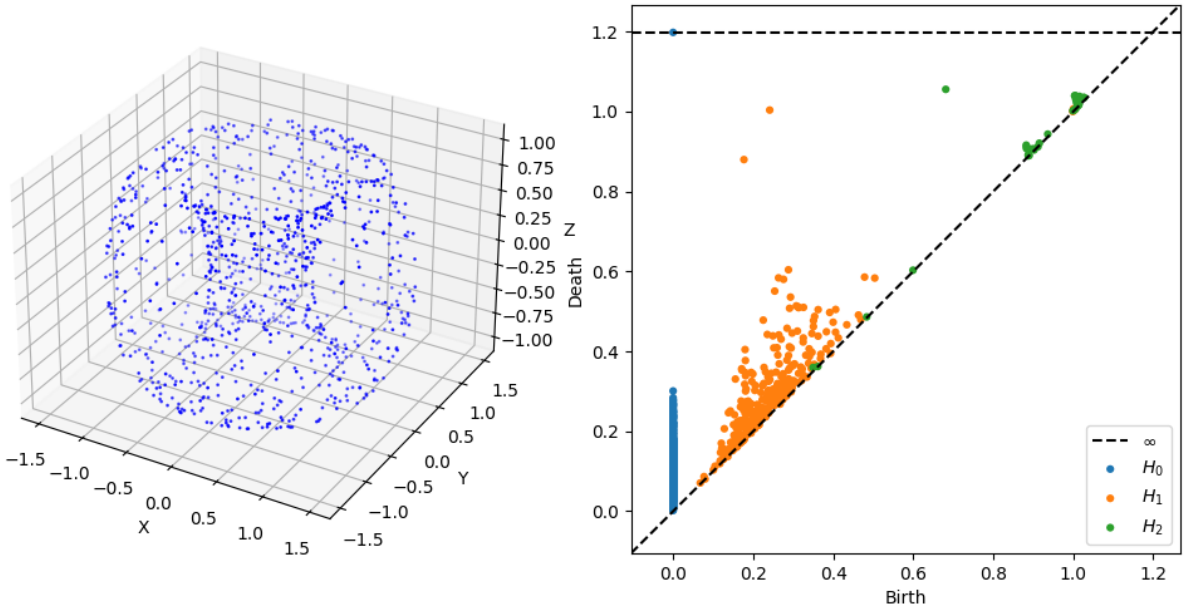


Figure A.1: Ripser demonstration: a point cloud sampled from a torus and its persistence diagram.

Bibliography

[1] R. J. Adler, *Persistent Homology for Random Fields and Complexes* (2010), 124–143 pp., available at https://doi.org/10.48550/arXiv.1003.1001.

[2] R. J. Adler, O. Bobrowski, and S. Weinberger, *Crackle: The Persistent Homology of Noise*, arXiv: Probability (1996), available at https://doi.org/10.48550/arXiv.1003.1001.

[3] U. Bauer, *Ripser: efficient computation of Vietoris-Rips persistence barcodes*, J. Appl. Comput. Topol. **5** (2021), no. 3, 391–423, available at https://doi.org/10.1007/s41468-021-00071-5. MR4298669

[4] O. Bobrowski and P. Skraba, *On the Universality of Random Persistence Diagrams* (202207), available at https://arxiv.org/abs/2207.03926.

[5] M. B. Botnan and M. Lesnick, *An Introduction to Multiparameter Persistence* (2022), available at https://arxiv.org/abs/2203.14289.

[6] A. Chakrabarti and J. K. Ghosh, *AIC, BIC and Recent Advances in Model Selection* **7** (2011), 583–605 pp., available at https://www.sciencedirect.com/science/article/pii/B9780444518620500186.

[7] F. Chazal and B. Michel, *An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists*, Frontiers in Artificial Intelligence 4:667963 (September, 29 2021), available at https://www.frontiersin.org/articles/10.3389/frai.2021.667963/full.

[8] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer, *Stability of Persistence Diagrams*, Discrete Comput. Geom. **37** (2007January), no. 1, 103–120, available at https://doi.org/10.1007/s00454-006-1276-5.

[9] R. B. D'Agostino, *An Omnibus Test of Normality for Moderate and Large Size Samples*, Biometrika **58** (1971), no. 2, 341–348, available at https://www.jstor.org/stable/2334522.

[10] R. B. D'Agostino and E. S. Pearson, *Tests for departure from normality. Empirical results for the distributions of b^2 and $\sqrt{b^1}$*, Biometrika **60** (197312), no. 3, 613–622, available at https://academic.oup.com/biomet/article-pdf/60/3/613/576953/60-3-613.pdf.

[11] N. R. Draper and H. Smith, *Applied Regression Analysis* (1998), available at https://books.google.ca/books?id=d6NsDwAAQBAJ.

[12] H. Edelsbrunner and J. L. Harer, *Computational Topology: An Introduction* (2010), available at https://www.researchgate.net/publication/220692408_Computational_Topology_An_Introduction.

[13] D. A. Freedman and P. Diaconis, *On the histogram as a density estimator: L2 theory*, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **57** (1981), 453–476, available at https://link.springer.com/article/10.1007/BF01025868citeas.

[14] U. Fugacci, S. Scaramuccia, F. Iuricich, and L. Floriani, *Persistent Homology: a Step-by-step Introduction for Newcomers* (2016), available at https://diglib.eg.org/handle/10.2312/stag20161358.

[15] R. Ghrist, *Barcodes: The persistent topology of data*, BULLETIN (New Series) OF THE AMERICAN MATHEMATICAL SOCIETY **45** (200802), available at https://www.researchgate.net/publication/228366252_Barcodes_The_persistent_topology_of_data.

[16] B. Giunti, *TDA-Applications* (2020), available at https://www.zotero.org/groups/2425412/tda-applications.

[17] A. Hatcher, *Algebraic Topology* (2001), available at https://pi.math.cornell.edu/~hatcher/AT/AT.pdf.

[18] M. Hossain, *AIC and BIC – The two competitive information criteria for model selection in economics and statistics*, Journal of Social Science: Part II **19** (199801), 133–140, available at https://www.jstor.org/stable/2347338.

[19] O. Kehinde, *A New Class of Generalized Burr III Distribution for Lifetime Data*, International Journal of Statistical Distributions and Applications **4** (2018), no. 1, 6, available at https://arxiv.org/abs/1701.00403.

[20] B. Lagos-Álvarez, N. Jerez-Lillo, J. P. Navarrete, J. Figueroa-Zúñiga, and V. Leiva, *A Type I Generalized Logistic Distribution: Solving Its Estimation Problems with a Bayesian Approach and Numerical Applications Based on Simulated and Engineering Data*, Symmetry **14** (2022), no. 4, available at https://www.mdpi.com/2073-8994/14/4/655.

[21] C. Maria, J. J. Boissonnat, M. Glisse, and M. Yvinec, *The Gudhi Library: Simplicial Complexes and Persistent Homology* (2014), available at https://www.researchgate.net/publication/267108101_The_Gudhi_Library_Simplicial_Complexes_and_Persistent_Homology.

[22] M. P. McLaughlin, *A Compendium of Common Probability Distributions* (2001), A–37 pp., available at http://www.uvm.edu/~mamclaughlin/StatPages/More_StatPages/Compendium.pdf. Retrieved 2022-02-15.

[23] J. L. Mike and V. Maroulas, *Nonparametric Estimation of Probability Density Functions of Random Persistence Diagrams* (2018), available at https://arxiv.org/abs/1803.02739.

[24] R. B. Millar and B. Russell, *Maximum likelihood estimation and inference: with examples in R, SAS, and ADMB* (2011) (eng).

[25] K. Morteza and A. Ahmadabadi, *Some properties of generalized gamma distribution*, Mathematical Sciences Quarterly Journal **4** (201003), available at https://www.researchgate.net/publication/50198287_Some_properties_of_generalized_gamma_distribution.

[26] N. N. Saralees, K. P. Tibor, and K. S. Ram, *On the characteristic function for Burr distributions*, Statistics **46** (2012), no. 3, 419–428, available at https://doi.org/10.1080/02331888.2010.513442.

[27] S. Nadarajah, *A generalized normal distribution*, Journal of Applied Statistics **32** (2005), no. 7, 685–694, available at https://doi.org/10.1080/02664760500079464.

[28] M. Nassar and F. H. Eissa, *On the Exponentiated Weibull Distribution*, Communications in Statistics - Theory and Methods **32** (2003), no. 7, 1317–1336, available at https://doi.org/10.1081/STA-120021561.

[29] N. Otter, M. A. Porter, U. Tillmann, P. Grindrod, and H. A. Harrington, *A roadmap for the computation of persistent homology*, EPJ Data Science **6** (2013), no. 17, available at `https://doi.org/10.1140/epjds/s13688-017-0109-5`.

[30] D. Reinsel, J. Gantz, and J. Rydning, *The Digitalization of the world: From Edge to Core (IDC White Paper)*, International Data Corporation (Hrsg.), Framingham MA (2021), available at `https://www.platinasystems.com/report-the-digitization-of-the-world-from-edge-to-core`.

[31] E. W. Stacy, *A Generalization of the Gamma Distribution*, The Annals of Mathematical Statistics **33** (1962), no. 3, 1187 –1192, available at `https://doi.org/10.1214/aoms/1177704481`.

[32] G. R. Terrell and D. W. Scott, *Oversmoothed Nonparametric Density Estimates*, Journal of the American Statistical Association **80** (1985), no. 389, 209–214, available at `http://www.jstor.org/stable/2288074`.

[33] teapot.obj, *Stanford Computer Graphics Lab*, available at `https://graphics.stanford.edu/courses/cs148-10-summer/as3/code/as3/teapot.obj`.

[34] C. Tralie, N. Saul, and R. Bar-On, *Ripser.py: A Lean Persistent Homology Library for Python*, The Journal of Open Source Software **3** (2018Sep), no. 29, 925, available at `https://doi.org/10.21105/joss.00925`.

[35] P. et al. Virtanen, *SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python*, Nature Methods **17** (2020), 261–272, available at `https://rdcu.be/b08Wh`.

[36] R. Wadhwa, D. Williamson, A. Dhawan, and J. Scott, *TDAstats: R pipeline for computing persistent homology in topological data analysis*, Journal of Open Source Software **3** (201808), 860, available at `https://www.researchgate.net/publication/326911777_TDAstats_R_pipeline_for_computing_persistent_homology_in_topological_data_analysis`.

[37] E. Wagenmakers and S. Farrell, *AIC model selection using Akaike weights*, Psychonomic Bulletin & Review **11** (2004), no. 1, 192–196, available at `https://www.researchgate.net/publication/8588301_AIC_model_selection_using_Akaike_weights`.

[38] C. Walck, *Hand-book on Statistical Distributions for Experimentalists* (1996), available at `https://books.google.ca/books?id=3OFYzwEACAAJ`.

[39] M. L. Wright, *Introduction to Persistent Homology* **51** (2016), 72:1–72:3 pp., available at `http://drops.dagstuhl.de/opus/volltexte/2016/5964`.

[40] A. Zomorodian and G. Carlsson, *Computing Persistent Homology*, Discrete and Computational Geometry **33** (200502), 249–274, available at `https://www.researchgate.net/publication/225181947_Computing_Persistent_Homology`.