University of Alberta

PERSISTENT HOMOLOGY IN ANALYSIS OF POINT-CLOUD DATA

by

Violeta Kovacev–Nikolic

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science

in

Statistics

Department of Mathematical and Statistical Sciences

Fall 2012

Edmonton, Alberta

To my husband Dragan for his love and support, even at those difficult moments that challenged us both.

Also, to my parents, Aleksandar and Maria-Zsuzsanna for their love and encouragement through my entire life.

Abstract

The main goal of this thesis is to explore various applications of persistent homology in statistical analysis of point-cloud data. In the introduction, after a brief historical overview, we provide some of the underlying concepts of persistence. Starting from Chapter 2, the focus is on analysis of point-clouds sampled from a surface of a torus and a sphere; our first exploratory tool is a homology plot. In Chapter 3 we calculate the Wasserstein distances in order to visualize existing relationships among samples of data. Chapter 4 introduces a new approach in topological statistical inference, based on the notion of persistence landscapes. In Chapter 5 the method of persistence landscapes is applied to non-perturbed data; following that, data in Chapter 6 involve a component of noise which allows us to demonstrate the efficiency of the new method. To test hypotheses, we implement suitable permutation tests. Last but not least, in Chapter 7 we work with real data of samples of HIV-1 protease some of which feature drug resistance. We truly hope that with the results presented, we offer convincing evidence that testifies in favor of applications of topology in statistical data analysis.

Preface

At one occasion, when I just began to embark on the path that eventually lead to this thesis, a fellow colleague asked about my research. While trying to provide an accurate and possibly interesting answer, it seemed that saying "applications of topology to data analysis" could be a good way to start. Indeed, the apparent contradiction conveyed by my brief statement turned into an exciting discussion on the topic, all due to the fact that topology has always been considered as a highly theoretical area in mathematics and under no circumstances a part of applied mathematics or statistics. Yet, there exists a significant overlap between the two areas, and in this intersection lies *computational topology*.

To be able to study and learn about methods of computational topology has certainly been an honor and a great challenge for me. Hopefully, in this thesis I will be able to fulfill the main objective and demonstrate the unique and useful applications of computational topology, in particular persistent homology, to analysis of point-cloud data.

Acknowledgements

First and foremost, I wish to express most sincere gratitude to both research supervisors, Professor Giseon Heo and Professor Terry Gannon, without whom this thesis would not be possible. Looking back at the past two years of our work together, I understand how much I have learned. The reading courses and seminars have been a wonderful experience and the numerous meetings extremely helpful. I very much appreciate Professor Heo's valuable guidance and assistance, useful ideas conveyed via exchanged messages, and time spent discussing various approaches in our data analysis. Her work enthusiasm and persistence have kept my research focus in the right direction and up to date in terms of current developments in the area. I also most sincerely appreciate Professor Gannon's encouragement and kindness; even when half a world away, the good professor has always replied to an important message. With his expertise, Professor Gannon has never left a question unanswered and provided examples and comments that helped me grasp on ideas and better understand the underlying theoretical concepts. Let me also emphasize my appreciation for corrections obtained from both professors in proofreading the thesis.

I would also like very much to acknowledge the great scientific input received from Professor Peter Bubenik; I cannot express enough how thankful and honored I am for having full access to his recent research results on persistence landscapes; the initial codes and the detailed theory provided have added an invaluable component to this thesis and a solid base for future research.

My special thanks to Henry Adams on his kindness and great help with the javaPlex persistent homology software. I appreciate the exhaustive answers

to my inquires; the guidelines shared have been most useful. Many thanks to Andrew Tausz for resolving bugs encountered in the software.

I am also grateful to Professor Yuriy Mileyko for assistance regarding the Wasserstein distance.

Let me also acknowledge Yi Mao for providing correlations for the protease data. Many thanks to Dragan Nikolic on useful directions for exploring the concept of protease and for proofreading my work.

My deepest appreciation to the graduate studies chair, Professor Thomas Hillen, who has been a helpful and caring advisor, someone that students can always turn to for academic support. I also thank the friendly and supportive staff at the Department of Mathematical and Statistical Sciences.

This is also an opportunity to thank the committee chair, Professor Peter Hooper for supporting our research project. I will never forget my professors, present and past, who conveyed their knowledge and taught me unselfishly and with enthusiasm.

Last but not least, let me express my love and gratitude to my husband and parents, for their understanding, encouragement, and above all, endless love throughout all these years.

Edmonton, September 2012.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.A A Historical Overview

In order to better understand applications of computational topology, it would be worthwhile to make a brief historical retrospective.

From the history of mathematics we know it was in 1735 when Leonhard Euler solved the Königsberg Bridges problem; the solution published a year later had the title *"Solutio Problematis ad Geometriam Situs Pertinentis"*, or, in translation, *"The Solution to a Problem Pertaining to the Geometry of Position"*. Though Euler's explanation of the famous puzzle undoubtedly represents a moment when graph theory was introduced into mathematics [2], there may be an additional interpretation of the event [30]; as indicated in the title of Euler's paper, this is also when a new mathematical concept emerged, sprouting from the idea that it is rather the relative position i.e. arrangement of objects and not the actual coordinates that describes a set of objects. Hence, the year 1736 when Euler's solution was published might as well be regarded as the early beginning of topology.

However, the majority of authors, e.g. [20], agree that on the historical timeline of mathematics a more appropriate moment that marks the birth of topology would be the years 1894-95, when Henri Poincaré developed and

systematically established the theory of algebraic topology by publishing a series of six papers called *"Analysis Situs"*. When translated as *"Analysis of Position"*, the given phrase reveals a meaning somewhat similar to the title of the earlier mentioned Euler's paper [31]. We also note that the expression *"Analysis Situs"* actually represented the initial name for topology, referring to the fact that the main concern of this theory are properties of geometrical objects that remain unchanged under continuous elastic deformations such as bending, twisting, or stretching, whereas shape and size are omitted from consideration as irrelevant features.

Therefore topology has been present in mathematics for quite a long time; however, until recently, topology used to be exclusively perceived as a field in pure mathematics, without anticipating applications to real-world problems. Nonetheless, starting from the beginning of this century, the situation has changed. As Carlsson explains in his survey article [8], with breakthroughs of modern science and technology, an increasing number of researchers frequently encounter large datasets where each data-point is described by a long vector that may contain even thousands of coordinates. An efficient way to deal with such high-dimensional data is to use dimensionality reduction so that only a few of the most important coordinates remain for further statistical analysis. To implement this method a similarity measure i.e. a notion of a distance function is needed but, unlike physics, where coordinates describe motion of three-dimensional objects in the space-time continuum, in many other areas, e.g. biology and medicine, "coordinates" are difficult to interpret. Even more, it often happens that neither the choice of coordinates nor the metrics are clearly defined which may lead to contradicting conclusions, depending on a researcher's choice. In such situations, as suggested by G. Carlsson, "we

should not restrict ourselves to studying properties of the data which depend on any particular choice of coordinates." By focusing on properties that are independent on the choice of coordinates and the metrics used, we arrive to concepts of topology.

Applications of topology to data analysis have given rise to a new field within applied mathematics called *computational topology*. A powerful tool of computational topology follows from the concept of *persistent homology*, since it allows us to obtain information on topological and geometrical properties of an object based on a point-cloud dataset sampled from the object. Moreover, some most recent results developed on the notion of *persistence landscapes* show that a detailed statistical analysis may be further performed on data obtained from persistent homology. This has brought forth a yet another new area of research which we call *topological data analysis*.

Some of the most renown researchers in topological approach to statistical analysis of data are Herbert Edelsbrunner, David Letscher, Afra Zomorodian, Gunnar Carlsson, Robert Grist, Peter Bubenik, Vin de Silva, Robert Adler, Patrizio Frosini, Massimo Ferri, and many others. The number of contributing authors to computational topology continues to grow as the the method finds its implementation not only in mathematics, but also in other applied areas of modern science and engineering such as computational biology, medicine and biostatistics, computer graphics and image processing, complex dynamical networks in physics, etc.

1.B From a Point-Cloud to a Stream

To introduce the main ideas, we start with the notion of a point-cloud. As explained in [13], when a dataset is sampled, the goal is to obtain information about the underlying phenomenon represented by the data. When the object of study is a three-dimensional object, it is important to detect global features such as the geometric shape, number of components, loops and holes through the surface, or voids inside the object. For that purpose we sample points from a given object to obtain a point-cloud dataset.

> *Point-Cloud Dataset*: usually represents a large finite dataset sampled from a geometrical object in a three dimensional space, possibly with some noise. In general, a point-cloud can be sampled in an n-dimensional metric space.

Modern techniques for obtaining point-clouds involve laser scanning in which the distance of a ray of light from an object is measured as the ray travels on the object's surface [21]. That way up to 750,000 datapoints per second can be recorded. A simulation of a point-cloud on the Klein bottle is shown below.

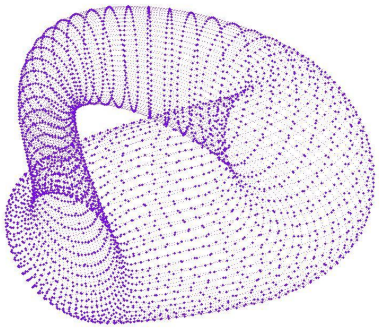

Figure 1.1: *a point-cloud obtained by sampling points from a Klein bottle.*

The sampled points (or their subset) represent vertices that mutually connect to form a structure called a simplicial complex or a stream. As Edelsbrunner explains [9], it takes less effort to construct an abstract simplicial complex so only afterwards we assign coordinates to embed the complex into a metric space. Using this guideline, we start with the following definition.

> *Affinely Independent Points*: Let $x_0, x_1, ..., x_k$ be points in an n-dimensional Euclidean space $\mathbb{E}^n$; these points are affinely independent if and only if vectors $x_i - x_0$, $1 \leq i \leq k$, are linearly independent.

An n-dimensional space can have at most n linearly independent vectors so there can be at most $n + 1$ affinely independent points. Based on this, we define a k-simplex.

> *k-simplex*: A k-simplex of $k+1$ affinely independent points $x_0, x_1, ..., x_k$ in an n-dimensional Euclidean space $\mathbb{E}^n$ is defined as the set of all linear combinations of the following form
>
> $$\sigma\{x_0, x_1, ..., x_k\} = \sum_{i=0}^{k} \lambda_i \, x_i \tag{1.1}$$
>
> where all λ_i are nonnegative and $\sum_{i=0}^{k} \lambda_i = 1$.

For $k = 0$, 1, 2, and 3, the corresponding k-simplex is just a regular vertex, edge, triangle, and tetrahedron, respectively.

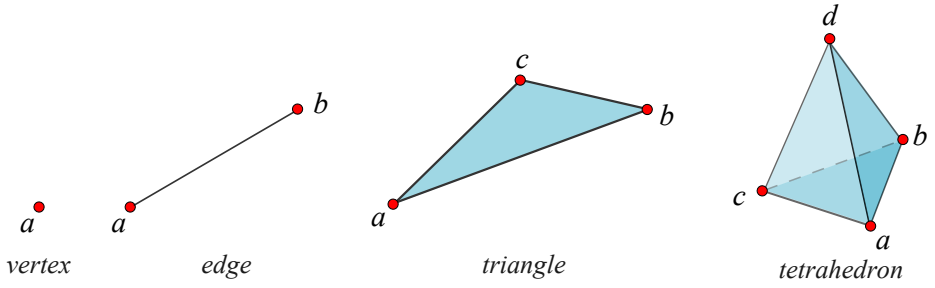

Figure 1.2: *Illustration of k-simplices for $k = 0, 1, 2, 3$.*

Now we can define an abstract simplicial complex.

Abstract Simplicial Complex: Let σ be a simplex with its nonempty subset τ that we will call a *face*. Then an abstract simplicial complex K represents a finite collection of simplices such that it is closed under taking faces and has no improper intersections. More formally,

- $\sigma \in K$ and $\tau \leq \sigma$ implies $\tau \in K$
- $\sigma_1, \sigma_2 \in K$ implies $\sigma_1 \cap \sigma_2$ is either an empty set or a face of both

To illustrate the above definition, we depict collections of simplices that do *not* represent a simplicial complex because one of the two necessary requirements is not satisfied.

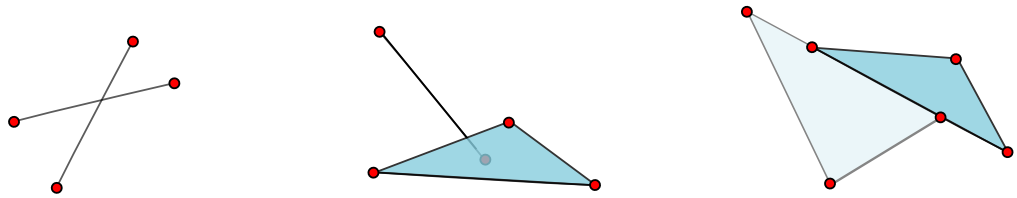

Figure 1.3: *Collections of simplices that do not form a simplicial complex. In the first case (left), the two edges intersect at a vertex that does not belong to the complex. In the second case (middle) an edge passes through a triangle at a point that is not a vertex in the complex. Last example shows two triangles that intersect along an edge that is not a face of any of the triangles.*

An example of a simplicial complex is given below.

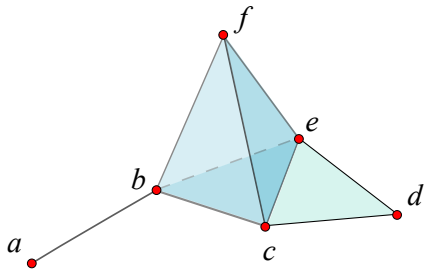

Figure 1.4: *A simplicial complex that consists of 6 vertices (denoted by letters a, b, c, d, e, f, g), 9 edges (ab, bc, be, bf, cd, ce, cf, de, ef), 5 triangles (bce, bcf, bef, cde, cef), and 1 tetrahedron (bcef). The total number of simplices in this simplicial complex is 21.*

A realization of an abstract simplicial complex is obtained assigning actual coordinates to the vertices. We also define a rule, called *filtration*, which determines when an edge, triangle, or tetrahedron (if we are in a $3d$-space, these are all the possibilities) are formed. Such a filtered simplicial complex is called a *stream*. Although a filtration represents the distance at which two vertices bond, it is common to refer to this parameter as if it were time.

Among several existing approaches, we will implement two types of streams: the Vietoris-Rips and the Lazy Witness. Since the latter is associated with the notion of the Witness stream, we provide the definition for all the three mentioned types of streams. For that purpose, we consider a point-cloud P in a metric space whose subset L assembled from points $[l_0, \ldots, l_{n_L}]$ is called a *landmark set*. Furthermore, let t represent the filtration time. Then the following streams can be defined, as shown in [43].

- *Vietoris-Rips stream*:
  - The vertices of the Vietoris-Rips stream are assembled from the entire point-cloud set P.
  - An edge between two vertices x_i and x_j appears at filtration
    $$t = d\left(x_i, x_j\right), \tag{1.2}$$
    where d is the distance between the two vertices.
  - Higher order simplices enter the stream as soon all their edges have been formed. Hence this stream represents the maximal simplicial complex that can be constructed over a set of existing edges; such a complex is called a *flag*.

- *Witness Stream*:
  - The vertices of this stream are the landmark points in L.
  - A k-simplex $[l_0, \ldots, l_k]$ appears in the stream at time t if all its faces are formed and there exists a witness point $\omega \in P$ such that:

$$t_{simplex} + m_k(w) \geq \max\{d(l_0, w), \ldots, d(l_k, w)\}, \tag{1.3}$$

    where $m_k(w)$ represents the distance of of the witness point from the $(k+1)$-th closest point in L.

- *Lazy Witness Stream*:
  - The vertices are members of the landmark set L.
  - An edge between vertices l_i and l_j appears at filtration t if there exists a witness point $w \in P$ such that:

$$t_{ij} + m(w) \geq \max\{d(l_i, w), d(l_j, w)\}. \tag{1.4}$$

    The variable $m(w)$ represents the distance of the witness point from the ν-th closest landmark point, where ν is an input parameter with values 0, 1, or 2. If $\nu = 0$, then $m(w) = 0$.
  - Higher order simplices appear when all their edges are formed.

The above defined streams can be easily constructed by implementing codes provided in a software package called *javaPlex* [44]. In order to build a stream, the user is required to input the point-cloud data, and in the case of Witness or Lazy Witness stream the number of landmark points must be specified as well. For more details, see [43].

1.C Homology Groups

Before introducing the concept of persistent homology, we shall define the underlying ideas of homology groups. To limit the extent of definitions, we present only the most important ideas. Our main references in this section are the *Elements of Algebraic Topology* by Munkres [29], the *Computational Topology* by Edelsbrunner and Harer [9], and the *Topology for Computing* by Zomorodian [49]. We start with the notion of an *oriented* simplex.

> *Oriented Simplex:* Let σ be a simplex with two different orderings of its vertices that differ from each other by an odd number of permutations; then the two orderings fall into two different equivalence classes where each class represents one *orientation* of σ. The simplex σ together with its orientation represents an oriented simplex.

An oriented simplex spanned by vertices $x_0, x_1, x_2, ..., x_p$ will be denoted by $[x_0, x_1, x_2, ..., x_p]$. Figure 1.5 shows examples of oriented simplices in a three dimensional space.

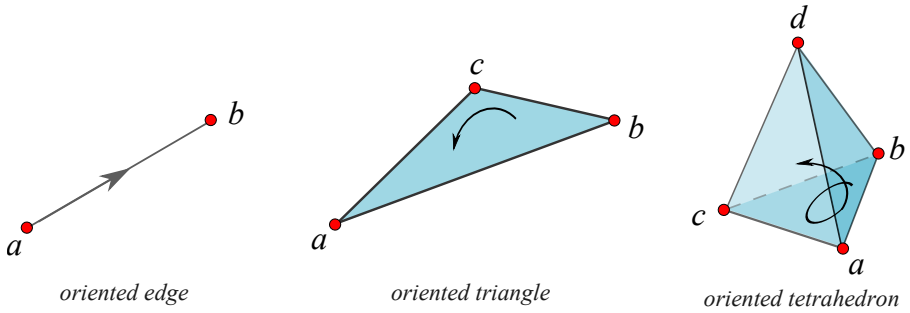

Figure 1.5: *Oriented edge* $[a, b]$*, triangle* $[a, b, c]$*, and tetrahedron* $[a, b, c, d]$*.*

The orientation of an edge is denoted by an arrow; in the case of a triangle a circular arrow is used. For a tetrahedron the "right-hand screw" was used i.e.

as explained by Munkres [29], if fingers of right hand follow the direction from a to b to c, the thumb should be pointing toward d. The opposite orientation would follow a "left-hand screw" rule. Note that a vertex as a 0-simplex has no orientation. Let us now define a p-chain.

> *p-chain* in a simplicial complex K is defined as the sum of oriented p-simplices in K, that is,
>
> $$C = \sum_i a_i \, \sigma_i \tag{1.5}$$
>
> where each σ_i is an oriented p-simplex. Coefficients a_i can be integers, rational numbers, or in general, elements of a ring or a field.

Think of a p-chain as a function C_p that assigns integers to p-simplices in K. If σ and σ' are the same simplex with opposite orientations, then we have $C_p(\sigma') = -C_p(\sigma)$; based on this, the set of p-chains is an Abelian group with respect to addition. Now we define the *boundary operator*.

> *Boundary operator:* a homomorphism $\partial_p : C_p \to C_{p-1}$ defined as:
>
> $$\partial_p \sigma = \sum_{j=0}^{p} \sigma(x_0, ..., \hat{x}_j, ..., x_p) \tag{1.6}$$
>
> where σ is an oriented p-simplex spanned by vertices $x_0, x_1, x_2, ..., x_p$, and $\hat{x}_j$ means the vertex x_j is omitted. In particular,
>
> $$\partial_p \, [x_0, x_1, ..., x_p] = \sum_{j=0}^{p} (-1)^j \, [x_0, ..., \hat{x}_j, ..., x_p] \tag{1.7}$$
>
> describes the action of ∂ on a simplex $\sigma = [x_0, x_1, x_2, ..., x_p]$.

Using the above rule, we apply the boundary operator to oriented simplices from Figure 1.5. From the results below and the accompanying illustration in Figure 1.6, we see that the outcome corresponds to the following:

$$\begin{aligned} \partial_1 [a,b] &= b - a \\ \partial_2 [a,b,c] &= [b,c] - [a,c] + [a,b] \\ \partial_3 [a,b,c,d] &= [b,c,d] - [a,c,d] + [a,b,d] - [a,b,c] \end{aligned} \tag{1.8}$$

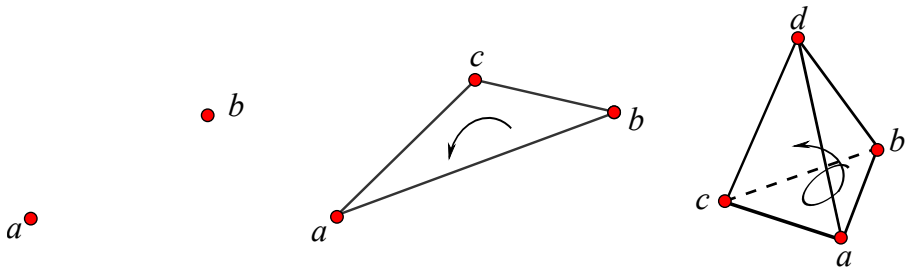

Figure 1.6: *Illustration for boundaries of oriented simplices.*

Some important results related to properties of the boundary operator are given as follows:

- The operation of taking a boundary commutes with addition, that is, for two simplices σ and σ', we have $\partial_p (\sigma + \sigma') = \partial_p \sigma + \partial_p \sigma'$.
- Taking the boundary of a p-chain $C = \sum_i a_i \, \sigma_i$ yields boundaries of its simplices, that is, $\partial_p C = \sum_i a_i \, \partial \sigma_i$
- Taking a boundary of a boundary produces a zero, that is, $\partial_{p-1} \partial_p (\sigma) = 0$, which in shorthand syntax, can be expressed as $\partial^2 = 0$.

Next, the definition of a chain complex follows.

A chain complex is a sequence of Abelian chain groups, connected with their boundary homomorphisms:

$$\ldots \overset{\partial_{P+2}}{\to} C_{P+1} \overset{\partial_{P+1}}{\to} C_P \overset{\partial_P}{\to} C_{P-1} \overset{\partial_{P-1}}{\to} \ldots \tag{1.9}$$

We also introduce p-cycles and p-boundaries, and explain their relationship.

Define a *p-cycle* as a p-chain with zero boundary: $\partial_p Z_p = 0$. The group Z_p of p-cycles is the kernel of the p-boundary homomorphism:

$$Z_p = \ker \partial_p \tag{1.10}$$

Define a *p-boundary* as the boundary of a $(p+1)$-chain. The group B_p of p-boundaries is the image of the p-boundary homomorphism:

$$B_p = \mathrm{im}\, \partial_\mathrm{p} \tag{1.11}$$

Then, as illustrated in 1.7, the p-boundary is also a p-cycle, that is, $B_p = \mathrm{im}\, \partial_{\mathrm{p}+1}$ is a subgroup of $Z_p = \ker \partial_p$:

$$\mathrm{im}\, \partial_\mathrm{p} \subset \ker\, \partial_{\mathrm{p}+1}. \tag{1.12}$$

Note that the image is contained in the kernel because of the property of the boundary operator that $\partial^2 = 0$.

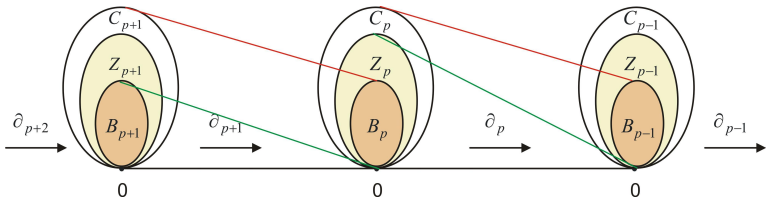

Figure 1.7: *Illustration of the boundary homomorphism, as in* [9].

Finally, we arrive to the definition of a homology group:

The p-th homology group is a quotient group denoted as the p-th cycle group modulo of the p-th boundary group:

$$H_p = Z_p/B_p = \ker \partial_p / \mathrm{im}\, \partial_{\mathrm{p+1}} \tag{1.13}$$

Furthermore, the p-th *Betti number* is the rank of this group:

$$\beta_p = rank\ H_p = rank\ Z_p - rank\ B_p \tag{1.14}$$

We will illustrate the implementation of the above equations at the end of the next section. Now, in terms of applications, it is important to point out that a homology group can be expressed in the following form:

$$H_p = Z^{\beta_p}, \tag{1.15}$$

where Z is a field, e.g. the set $\mathbb{Q}$ of rational numbers or a multiplicative group $\mathbb{Z}_n$ of integers modulo n, where n is a prime number. The rank or the group, denoted by β_p, represents the *Betti number* in the p-th homological dimension. Thus if we have $H_0 = \mathbb{Q}^1$, then $\beta_0 = 1$ (we will later understand that this means that our object of interest is a single connected component).

Furthermore, as we explain later (see end of section 1.E), a Betti number describes topological properties of an object in a particular homological dimension. Due to this important role of Betti numbers, it is very common to refer to the given homological dimension as the *Betti dimension*. This notation will be used throughout this thesis.

1.D Calculating the Torus Homology

Now we illustrate by example how a homology group can be calculated. Our main reference are lecture notes from a reading class with T. Gannon [12]. In the example we implement the fact that the mapping of the boundary operator ∂_p corresponds to the action of an incidence matrix Λ_p that characterizes the boundary of each $(p+1)$-cell in terms of lower dimensional p-cells. The p-th Betti number of a finite chain complex is then:

$$\beta_p = rank\, C_p - rank\, \Lambda_p - rank\, \Lambda_{p+1} \tag{1.16}$$

We work on a simple complex (called Δ-complex [16]) depicted in Figure 1.8.

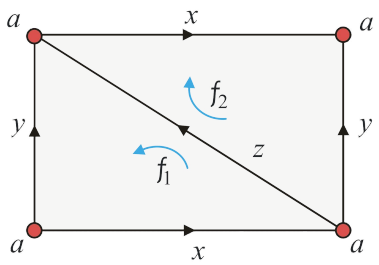

Figure 1.8: *Topological representation of a torus*

The given rectangle, in which two opposite sides are identified and all the vertices are equivalent, corresponds to the torus. After triangulation, the components of the complex are a vertex a, three edges, x, y, and z, and two triangles with faces f_1 and f_2. The orientation of the edges and triangles is denoted in the picture. Then the chain complex of the torus is:

$$0 = C_3 \xrightarrow{\partial_3} C_2 \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} C_{-1} = 0 \tag{1.17}$$

Next, we describe the nontrivial chain groups C_0, C_1, and C_3.

The chain groups C_0, C_1, C_3 for Betti dimensions 0, 1, 2, respectively, are:

$$\begin{aligned} &\text{0-dimensional chain group (vertices):} && C_0 = \{a\} \\ &\text{1-dimensional chain group (edges):} && C_1 = \{x,\ y,\ z\} \\ &\text{2-dimensional chain group (faces):} && C_2 = \{f_1,\ f_2\} \end{aligned}$$

Now consider the incidence matrices associated with our chain complex. For the 1-chain, the corresponding incidence matrix Λ_1 has one row (due to the vertex a) and there are three columns corresponding to edges x, y, and z:

$$\Lambda_1 = a \begin{pmatrix} 0 & 0 & 0 \end{pmatrix} \tag{1.18}$$

For the 2-chain, the associated incidence matrix Λ_2 consists of three rows, corresponding to edges x, y, and z, respectively, and there are two columns that correspond to faces f_1 and f_2. Then:

$$\Lambda_2 = \begin{pmatrix} 1 & 1 \\ -1 & -1 \\ 1 & 1 \end{pmatrix} \tag{1.19}$$

Note that the entries of an incidence matrix are obtained by moving along the given rectangle and observing the directions of simplices. For example, the first element of the matrix Λ_2 corresponds to the edge x which has positive orientation when we travel along the face f_1 so we assign a value of 1 to this entry. On the other hand, the edge y is traveled in the opposite direction of its orientation so the first entry in the second row takes the value -1. For given orientations of the edges x, y, z that comprise the face f_1, we obtain respective numbers 1, -1, 1, for the first column of the matrix. In the same manner, moving along the fact f_2, we obtain the second column of Λ_2.

In the last step we determine the ranks of matrices Λ_1 and Λ_2. It is easy to see that the $rank\,\Lambda_1 = 0$, whereas $rank\,\Lambda_2 = 1$, since all rows of the Λ_2 matrix are linearly dependent, i.e. there is only one linearly independent row. Then, by the rule $\beta_p = rank\,C_p - rank\,\Lambda_p - rank\Lambda_{p+1}$, we can finally determine the Betti numbers of the torus:

$$\beta_0 = rank\,C_0 - rank\,\Lambda_0 - rank\Lambda_1 = 1 - 0 - 0 = 1$$

$$\beta_1 = rank\,C_1 - rank\,\Lambda_1 - rank\Lambda_2 = 3 - 0 - 1 = 2$$

$$\beta_2 = rank\,C_2 - rank\,\Lambda_2 - rank\Lambda_3 = 2 - 1 - 0 = 1$$

Therefore, the Betti numbers for the torus are $(1, 2, 1)$. Now, a different way to calculate these numbers would be to express the homology group as a quotient group $H_p = Z_p/B_p$ as given in equation (1.13). For that purpose, we first calculate the two underlying quantities, using the following guidelines:

$$\begin{aligned} Z_p = &\quad \ker\,\partial_p \quad = \quad \mathrm{NullSp}\,(\Lambda_p) \\ B_p = &\quad \mathrm{im}\,\partial_{\mathrm{p+1}} = \quad \mathrm{ColSp}\,(\Lambda_{\mathrm{p+1}}) \end{aligned} \tag{1.20}$$

The first expression means that the kernel of the p-boundary operator is the null space of the incidence matrix; the second expression implies that the image of the $(p+1)$-boundary operator corresponds to the column space of the incidence matrix. For $p = 0$, the boundary homomorphism ∂_0 represents the zero map, so it takes every vector in $C_0 \subset Z$ to zero. Hence,

$$Z_0 = \quad \ker\partial_0 = Z \tag{1.21}$$

Since all the columns in Λ_1 are zeros, their span is also zero, so we have:

$$B_0 = \quad \mathrm{ColSp}\,(\Lambda_1) = \{0\} \tag{1.22}$$

Thus the homology group of the torus in Betti dimension 0 takes the form of the following quotient space:

$$H_0 = Z_0/B_0 = Z/\{0\}, \tag{1.23}$$

The final expression means that H_0 is a set that consists of a single element x in a field Z, where x and $x + 0$ are identified. Because for every member of any field the property $x + 0 = x$ holds, we conclude that the zeroth homology group of the torus is the entire field Z:

$$H_0 = Z \tag{1.24}$$

Now, according to equation (1.15), the zeroth homology group can be expressed in the form $H_0 = Z^{\beta_0}$. Here Z represents any field, for instance, it could be the set $\mathbb{R}$ of real numbers, the set $\mathbb{Q}$ of rational numbers, or a multiplicative group $\mathbb{Z}_n$ of integers modulo n where n is a prime number (note that we cannot consider the set of integers $\mathbb{Z}$ because it does not represent a field, due to lack of a multiplicative inverse). Therefore, we can finally conclude that:

$$\beta_0 = rank\,(H_0) = 1 \tag{1.25}$$

which corresponds to the results we have already found. In the similar manner we can calculate Betti numbers β_1 and β_2 which respectively correspond to the torus homology groups in homological dimensions $p = 1$ and $p = 2$.

1.E Persistent Homology & Betti numbers

The concept of persistence was developed by H. Edelsbrunner, D. Letscher, and A. Zomorodian [10]. The main idea of persistence is that important topological properties last over long filtration intervals, whereas short-lived features may be ignored as noise. A formal approach would require sophisticated definitions as shown in *Computational Topology* by Edelsbrunner and Harer [9]. We will not go into such detail, but rather just summarize the underlying ideas.

From previous section we have seen that an n-dimensional manifold M like a torus or a sphere is associated to a *single* chain complex, and thus to homology groups H_p for $0 \leq p \leq n$ (with coefficients in some field Z). We also know that these homology groups H_p are vector spaces which can be expressed in the form Z^{β_p}, where the rank of the group, β_p, represents the p-th Betti number whose index p denotes the Betti (or homological) dimension.

Let us now make a transition from a manifold to a point-cloud dataset. As shown in [12], when we deal with a point-cloud dataset P whose mathematical representation is a stream (e.g. Vietoris-Rips, Witness, or Lazy Witness), then, instead of a single chain complex, we have a *sequence* of chain complexes. In that case, the associated *persistent* homology $H_p^{i \to j}(C)$ is depicted in a suitable plot which we call a *homology plot* or a barcode.

Now, in a homology plot horizontal bars denote topological features that change through filtration time. Each bar corresponds to an interval (b_i, d_i). The beginning $t = b_i$ of the interval denotes a moment in time when the given feature is "born;" the other endpoint of the bar denotes the "death" time $t = d_i$ when the given feature ceases to exist. The longer a bar, the more important the associated feature.

Figure 4.1 provides a simple illustration of a barcode, defined on a sequence of three intervals $\{(0,16),(2,6),(4,14)\}$.

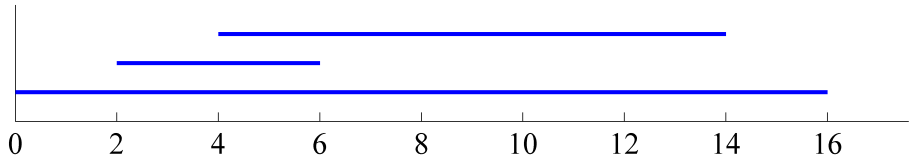

Figure 1.9: *Barcode for intervals* {(0,16), (2,6), (4,14)}

The two longer bars imply an important topological feature whereas the short bar does not carry much relevance. A point-cloud in an n-dimensional space can contain maximally n barcodes, one for each dimension. Hence, a point-cloud sampled on a torus or a sphere in the regular 3d-space can be associated with three barcodes, one for each Betti dimension 0, 1, and 2.

An equivalent form of representation is a *persistence diagram* in which the endpoints of intervals define coordinates (b_i, d_i), $i = 1, 2, ..., n$, of points scattered in the upper-half plane above the main diagonal.

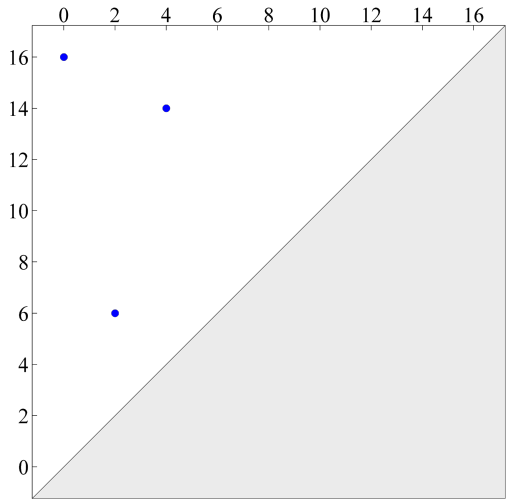

Figure 1.10: *Persistence diagram for intervals* {(0,16), (2,6), (4,14)}

Last but not least we consider a typical homology plot obtained on a point-cloud sampled from an object whose topological features we wish to describe.

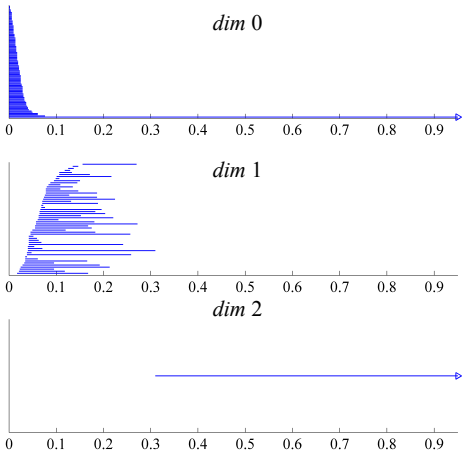

Figure 1.11: *Illustration of a homology plot across three Betti dimensions.*

Observe that in Betti dimension 0 we have one long bar, accompanied by several short bars, so the zeroth Betti number is $\beta_0 = 1$. Since in this homological dimension we count the number of components, if follows that the object from which the data are sampled consists of a single component. Considering Betti dimension 1, all bars are relatively short i.e. no long bar appears so $\beta_1 = 0$; because in this dimension we count the number of loops on a surface or the number of holes through a surface, it follows that the associated object has no loops or holes. In Betti dimension 2, a single long bar implies that $\beta_2 = 0$; since in this dimension we count the number of voids or hollow spaces inside an object, it follows that the given object has one void. Therefore, the Betti numbers corresponding to the dataset form a sequence $(1, 0, 1)$ which is the known homology of the sphere. That way the underlying object is identified as the sphere.

In the following chapters we will illustrate by example various applications of persistent homology.

Chapter 2: Two Homologies

To demonstrate the methods of topological data analysis, we will work first with data sampled from two well-known topological objects: the torus and the sphere. Our goal is to explore the homology of these manifolds, based on simplicial complexes constructed on the data.

2.A Point-Cloud Datasets

Consider two topologically different manifolds, for instance, a torus and a sphere, embedded in a three-dimensional space. Sampling points uniformly randomly from these surfaces, two point-cloud datasets are obtained.

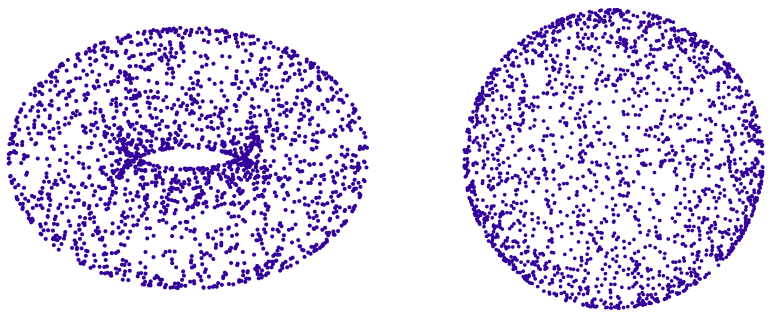

Figure 2.1: *Point-cloud datasets obtained by sampling 2000 points uniformly randomly from a torus and a sphere.*

In our case, each dataset is of size 2000. The radius of the sphere is chosen relative to the torus so that the surface densities of the point-clouds are equal.

This particular requirement yields homology plots on approximately the same scale which is important for making a statistically valid comparison between the two manifolds (see Introduction). Due to the same number of points sampled from each the torus and the sphere, the general requirement of equal surface densities is reduced to having equal surface areas. Therefore,

$$4\pi^2 ac = 4\pi r^2, \tag{2.1}$$

where a is the radius of the toroidal tube, c is the distance of the center of the tube from the center of the torus, and r is the radius of the sphere. From the equation (2.1), we obtain $r = \sqrt{\pi ac}$. Now consider a torus described by parameters $a = 1$ and $c = 2$, as illustrated below.

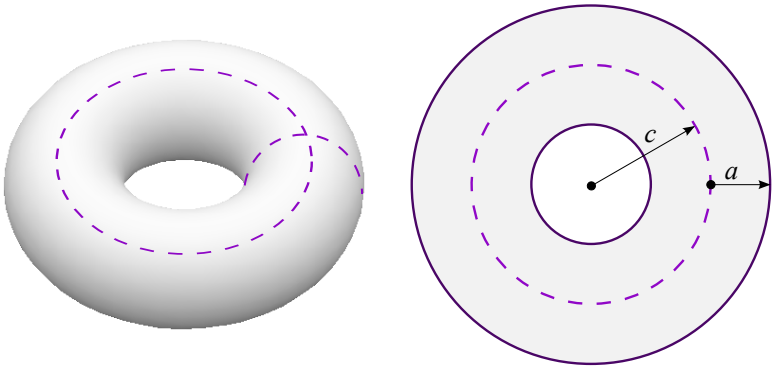

Figure 2.2: *Torus with tube of radius $a = 1$ and distance $c = 2$ from the center of the torus to the center of the tube.*

Then $r = \sqrt{2\pi} \approx 2.507$ represents the radius of the sphere that has the same area as the given torus. Note that this is only an initial estimate and a small correction should be made. We implement a trial-and-error approach which reveals the radius $r = 2.8$ as a more appropriate choice[1].

[1] The homological properties of the torus and the sphere are well-known; we use this to find that the $r = 2.8$ value provides the expected results.

2.B Building a Simplicial Complex

On each sampled point-cloud we construct a stream, i.e. a filtered simplicial complex. For this purpose we implement *javaPlex* [44], the latest in a series of similar open-source software packages for computing persistent homology. As illustrated in the next section, the number of higher-dimensional simplices rapidly increases with filtration time, i.e. the larger the number of vertices, the greater the computational challenge. Consequently, it would be impractical to assign the whole point-cloud to the vertex set of a stream; instead, a small subset of well-spaced landmark points is selected to represent the vertices of a given simplicial complex. The landmarks are obtained via an iterative optimization procedure called the *sequential maxmin* [43], described below.

> *Sequential maxmin*: Let the initial landmark l_1 be randomly selected from a point cloud P and $L_k = \{l_1, l_2, l_3, ..., l_k\}$ be the set of landmark points after k iterations of the maxmin algorithm. Then the next chosen landmark is the point $l_{k+1} \in P$ whose distance $d\left(l_{k+1}, L_k\right)$ from the set L_k is maximal.

Applying this algorithm, we select 100 well-spaced landmarks from our datasets.

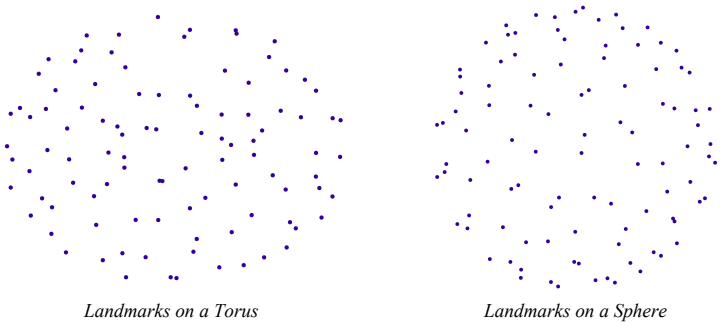

Figure 2.3: *100 well-spaced landmark points for the torus and the sphere, obtained by applying the maxmin algorithm to the initial point-clouds.*

A landmark set defines vertices of a corresponding *Lazy Witness* stream. This type of a simplicial complex is especially convenient because non-landmark points may also be involved in the construction to serve as "witnesses" to the binding process, so more information on data is captured at a relatively low computational cost. Further, there is an additional parameter ν with values in $\{0, 1, 2\}$, which allows more options in terms of building the complex. We set $\nu = 1$, thus the *first* neighbors of interlacing vertices witness the formation of edges. For more detail on the Lazy Witness complex, see [6] and [43].

2.C Evolution of Streams and Homology Plots

For each the torus and the sphere, a homology plot (barcode) is generated.

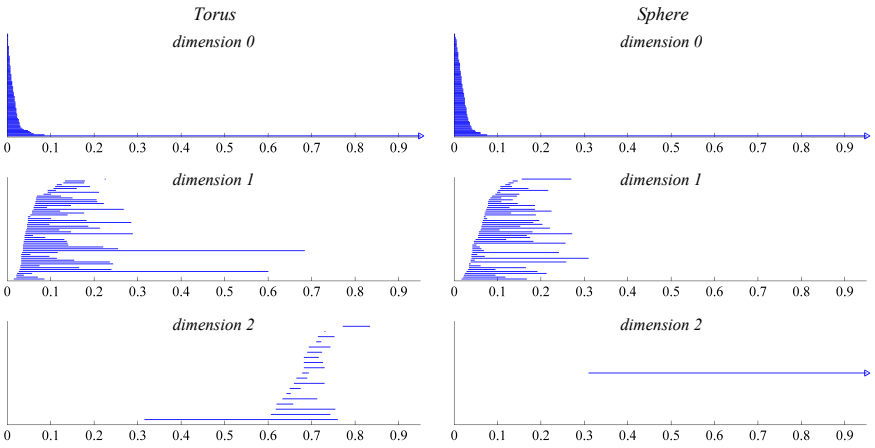

Figure 2.4: *Homology plots for Lazy Witness streams on a torus and a sphere.*

Comparing the two plots, we observe the following:

(i) At Betti dimension 0, both the torus and the sphere feature a long bar which indicates a single-component manifold.

(ii) At Betti dimension 1, the two loops on the torus yield two long bars while the sphere has none.

(iii) A long bar at dimension 2 implies a hollow space enclosed by each surface.

Thus the torus has Betti numbers $(1, 2, 1)$ and for the sphere they are $(1, 0, 1)$ which agrees with theory (e.g. [49]). Nonetheless a thorough inspection reveals a discrepancy in terms of shorter bars; though usually considered as noise, these bars may matter in statistical analysis. To understand their significance, we explore the underlying phenomena. For that purpose we think of a filtration value as a *time* parameter that describes the evolution of a given stream. The expression "time" is used for illustration only; a filtration value actually represents the maximal edge length between connecting vertices. For instance, at filtration $t = 0.02$ connections are formed among vertices that lie within a radius of 0.02 from each other. Now observe that the diameter of the sphere is almost three times greater than the diameter of the toroidal tube. Thus vertices on the torus lie closer to each other in the 3d-space so edges, triangles, and tetrahedra are sooner generated than on the sphere. Moreover, at any filtration time the total number of simplices is greater on the torus.

Table 2.1: Count of simplices through 15 filtration times.

Filtration time	**Torus**					**Sphere**				
	n_0	n_1	n_2	n_3	N	n_0	n_1	n_2	n_3	N
0.00	100	0	0	0	100	100	0	0	0	100
0.02	100	72	2	0	174	100	61	5	0	166
0.04	100	139	26	1	266	100	122	23	2	247
0.06	100	181	55	3	339	100	162	48	3	313
0.08	100	222	92	6	420	100	203	76	3	382
0.10	100	241	113	7	461	100	222	92	3	417
0.15	100	285	180	20	585	100	265	152	11	528
0.20	100	310	226	31	667	100	295	201	16	612
0.30	100	378	369	101	948	100	336	280	45	761
0.40	100	443	536	219	1298	100	390	405	123	1018
0.50	100	557	929	629	2215	100	468	635	311	1514
0.60	100	716	1629	1677	4122	100	570	1021	767	2458
0.70	100	961	3151	5081	9293	100	698	1623	1731	4152
0.80	100	1239	5670	13023	20032	100	891	2796	4315	8102
0.90	100	1526	9407	31055	42088	100	1086	4305	8732	14223

n_0 – number of vertices; n_1 – number of edges; n_2 – number of triangles;
n_3 – number of tetrahedra; N – total number of simplices.

The evolution of the streams is illustrated in the figure below.

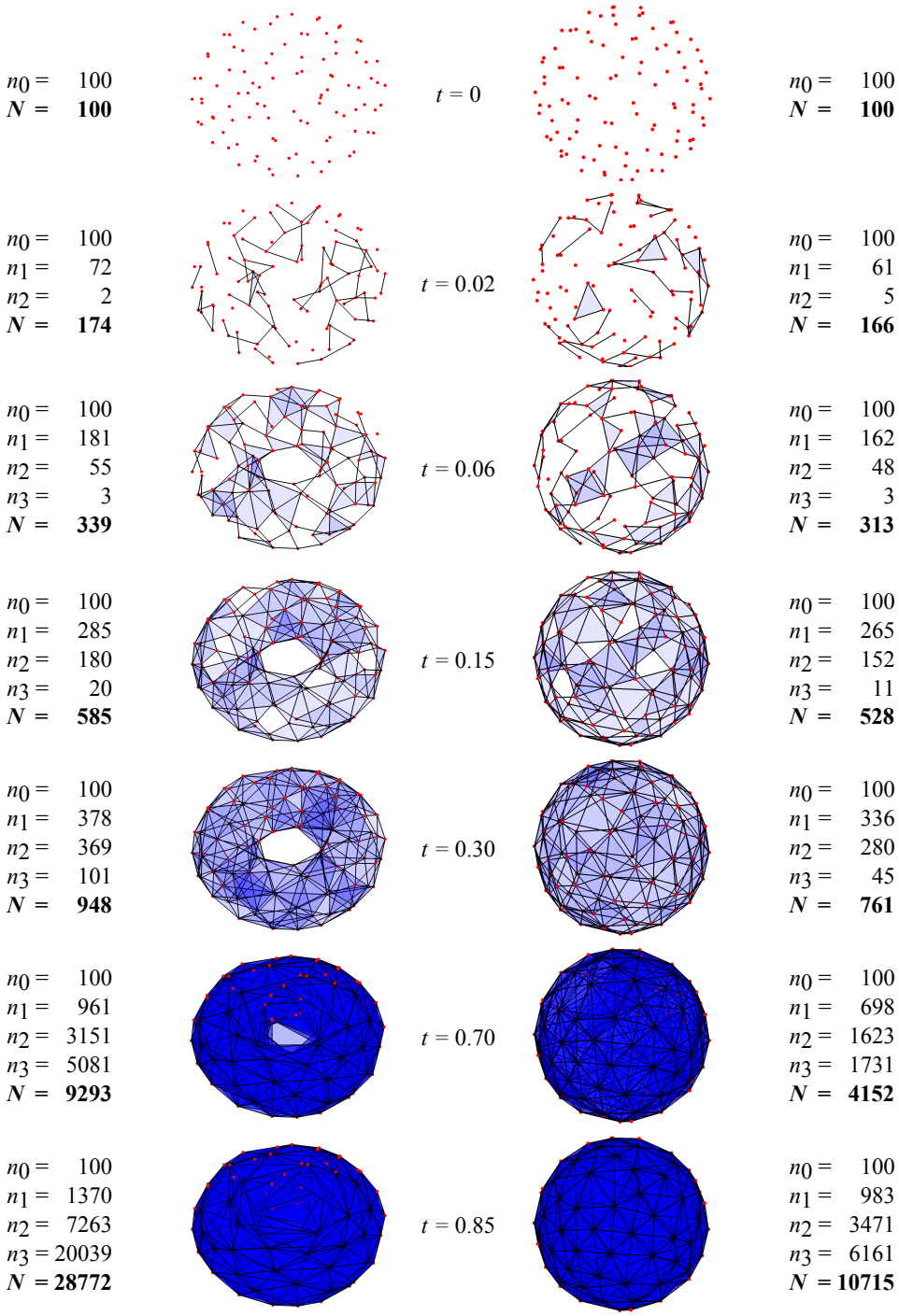

Figure 2.5: *Evolution of Lazy Witness streams on a torus and a sphere through seven filtration times. Variables n_0, n_1, n_2, and n_3, respectively, denote the number of vertices, edges, triangles, and tetrahedra at a given filtration time. The variable N represents the total number of simplices. Note the consistently greater value of N in the case of the torus.*

Thus on evolutionary timeline, the torus progresses faster than the sphere. Based on these results, we interpret the homology plots.

First, we explore Betti dimension 0, in which we count the number of components in a given stream. With maximal value at initial time $t = 0$, this number rapidly decreases until each stream becomes connected into a single component; the zoomed-in-view shows this happens around time $t = 0.085$.

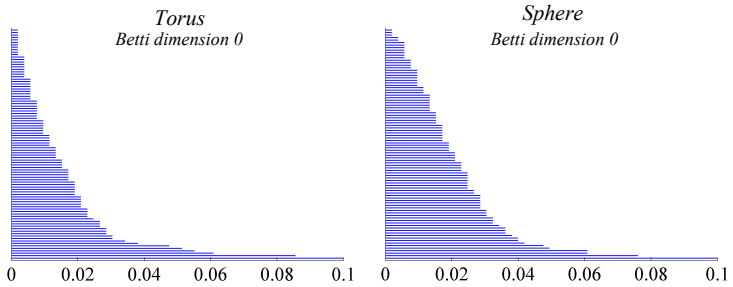

Figure 2.6: *Homology plots of the torus and the sphere for Betti dimension 0.*

Observe a slight difference in short-lived bars which appear longer for the sphere. This is not necessarily a rule; another point-cloud may yield an opposite outcome. Later statistical analysis will prove the difference insignificant.

Next, we focus on Betti dimension 1, associated with non-trivial loops on a given stream. A major difference occurs from about $t = 0.30$ through $t = 0.60$, since the torus has two independent loops and the sphere has none. The situation changes around $t = 0.6$; at $t = 0.7$, the torus loses its features.

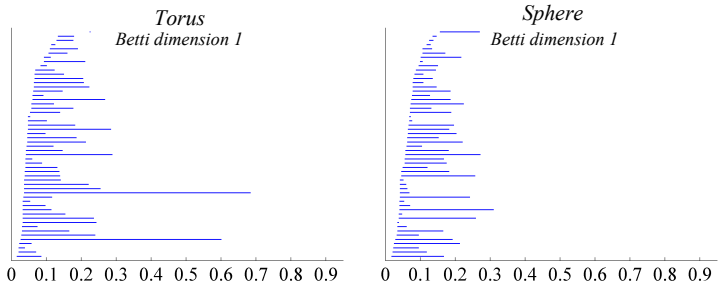

Figure 2.7: *Homology plots of the torus and the sphere for Betti dimension 1.*

An explanation may be that until time $t = 0.6$ the formation of simplices on a torus mostly occurs in the vicinity of the surface. As time progresses and more distant vertices start to connect, simplices start to build up across the hole in the center of the torus. At first only the rim of the hole is affected but as the process continues, the hole is gradually reduced and a bit before time $t = 0.7$ it ceases to exists, as illustrated in the evolution Figure 2.5.

At last we explore Betti dimension 2, associated with the number of voids enclosed by a given stream. Though initially no cavities are observed, around time $t = 0.30$ the streams are connected enough to form a closed surface that encloses a hollow space inside. However this configuration lasts only until $t = 0.6$ when several short-lived bars appear in the homology plot of the torus, while no such bars show up in the case of the sphere.

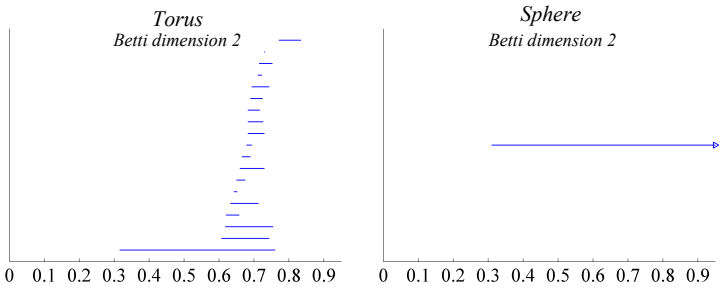

Figure 2.8: *Homology plots of the torus and the sphere for Betti dimension 2.*

An explanation may be that about time $t = 0.6$ simplices with relatively long edges start to form, passing through the toroidal tube. This process reduces the size of the initial void, though in some regions small pockets of void appear between the newly formed simplices. These remaining cavities are responsible for the appearance of short-lived bars in the barcode of the torus. When the last minuscule cavity disappears, the whole inner space of the torus is partitioned by solid tetrahedra; this happens around time $t = 0.85$, when the

last short bar in the homology plot meets its end. At that moment the stream transforms to a non-interesting structure similar to a piece of solid described by Betti numbers (1, 0, 0). Note that at the same time the void inside the sphere still exists, as indicated by the long bar in the sphere's homology plot.

Now we understand that even though in second Betti dimension the torus and the sphere feature the same homology, a geometrical difference exists due to the fact that the void inside the torus is smaller than the one in the sphere. In other words, it is rather the geometrical component that makes the difference and not the topological one. Therefore, not only that persistent homology captures information about topological properties of objects, but it also detects geometrical properties of objects. Therefore, in our statistical analysis we will not be surprised if some difference between the torus and the sphere is observed in the second Betti dimension.

2.D Chapter Summary

Retracing the work shown in this chapter, we recall that starting from the initial point-cloud data, we constructed Lazy Witness streams and then, using tools of persistent homology, the corresponding homology plots were generated. Since the Betti numbers associated with these plots match the Betti numbers of the torus and the sphere, we may conclude that persistent homology recovered the topological and geometrical features of original manifolds from which the data were sampled. This power of persistent homology to retrieve structural properties of objects is one of the most important results of the chapter.

Chapter 3: Analyzing Distances

For the purpose of statistical analysis, we generate 15 point-clouds of size 2000 for each the torus and the sphere. Using the maxmin procedure, we select 100 landmarks from every point-cloud and construct a Lazy Witness stream with parameter $\nu = 1$ and maximal filtration value 1. After generating persistence intervals at Betti dimensions 0, 1, and 2, we wish to measure the proximity of these sets; as shown in [27], a convenient measure of distance would be the Wasserstein distance.

3.A Wasserstein Distance Matrix

For the purpose of measuring the Wasserstein distance between two sets of barcodes we represent them in a plane in a form of *persistence diagrams*. Then for persistence diagrams d_1 and d_2, the p^{th} Wasserstein distance is defined as

$$W_p\left(d_1, d_2\right) = \left(\inf_{\gamma} \sum_{x \in d_1} \left\|x - \gamma\left(x\right)\right\|_{\infty}^{p}\right)^{\frac{1}{p}}, \tag{3.1}$$

where γ represents all bijections from d_1 to d_2. Considering the implementation, there are two main stages in the process of obtaining the Wasserstein distance between two sets of barcodes; in the first stage, as explained in [9],

bipartite graph matching is used. In the last step a Hungarian algorithm for optimal assignment problem is implemented. Applying the described algorithm, we obtain three Wasserstein distance matrices, one for every Betti dimension. Each matrix is symmetric and of size 30×30, with the (i, j)th entry representing the distance between persistence diagrams d_i and d_j. We use these results to visualize the relationship between the 30 samples.

3.B Hierarchical Clustering

Our first visualization tool is cluster analysis. To estimate the dissimilarity among the 30 samples, we use single-linkage hierarchical clustering in which each sample initially represents a cluster; via an iterative procedure the samples are grouped into clusters based on the nearest neighbor criterion. That way we obtain three dendrograms, one for each Betti dimension. The results are displayed as follows. As expected, the first dendrogram shows the 15 tori and the 15 spheres are indistinguishable from each other in Betti dimension 0.

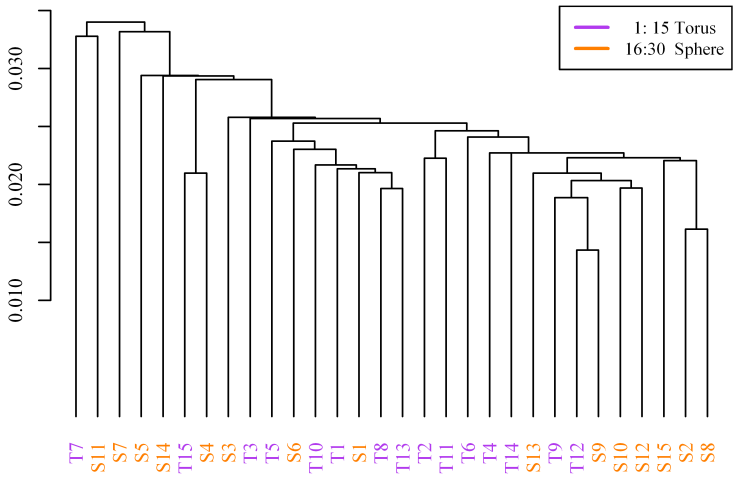

Figure 3.1: *Dendrogram for Betti dimension 0.*

At Betti dimension 1, two separate clusters verify the difference between the torus and the sphere data.

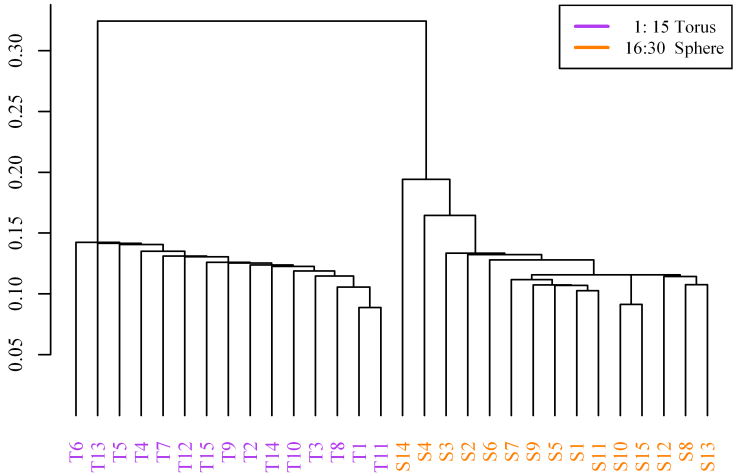

Figure 3.2: *Dendrogram for Betti dimension 1.*

The dendrogram for second Betti dimension indicates a difference between the torus and the sphere. As already mentioned, this is due to a difference in the geometry of the torus and the sphere detected from the appearance of short-lived bars in the homology plot of the torus.

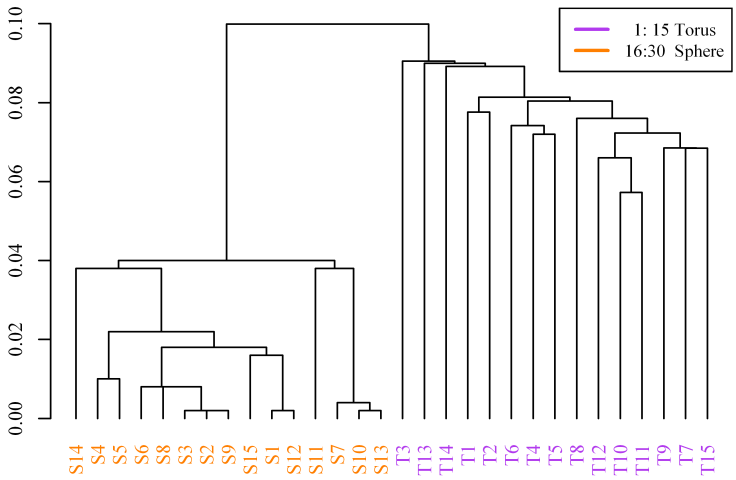

Figure 3.3: *Dendrogram for Betti dimension 2.*

Last but not least, observe that the variation among the 15 spheres is much less

than the variation among the 15 tori. This is due to the fact that at dimension 2 all the spheres in our sample feature a single persistence interval that extends until the maximal filtration time; thus the only difference within the sphere group is the difference among the 15 left endpoints. On the other hand, for all the tori in our sample in addition to a long persistence interval there are also several shorter intervals, which gives rise to more variation within the torus group.

3.C Multidimensional Scaling

Though easy to implement, hierarchical clustering has a disadvantage because it is a non-robust method where a small change in clustering criterion can yield a significantly different outcome. A more reliable statistical tool to visualize the relationship among our samples is the *multidimensional scaling* (MDS) method.

A great advantage of MDS is that *any* measure of similarity (or dissimilarity) among objects may be used, as long as there exists a monotone relationship between numerical values of the chosen measure and the actual proximity of the objects [19]. This allows us to use Wasserstein distance as a measure of dissimilarity between sequences of persistence intervals. Classical scaling methods do not have such flexibility; for instance, in principal component analysis (PCA) similarity is expressed through covariances or correlation coefficients which only measure the strength of a *linear* association between variables so there is an assumption of a linear relationship. MDS has greater range of application because no assumption on the nature of the data is needed, as shown in [18].

Using MDS, we depict the 30 samples in a 2d-plane. From the first plot, for Betti dimension 0, no major difference can be observed.

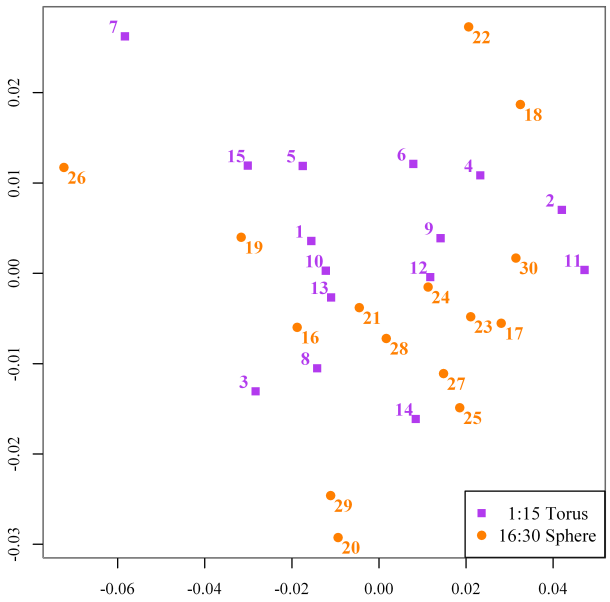

Figure 3.4: *Multidimensional scaling for Betti dimension 0.*

At Betti dimension 1, MDS shows a topological difference between the groups.

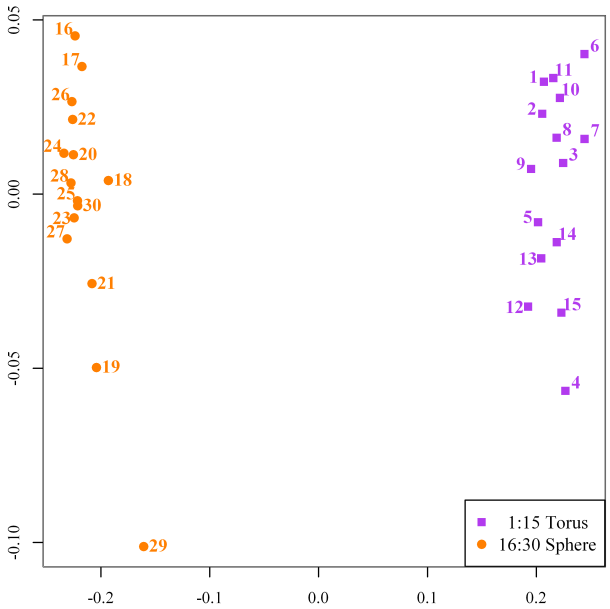

Figure 3.5: *Multidimensional scaling for Betti dimension 1.*

In the case of Betti dimension 2, MDS yields a similar result as the clustering analysis, that is, the two groups differ from each other.

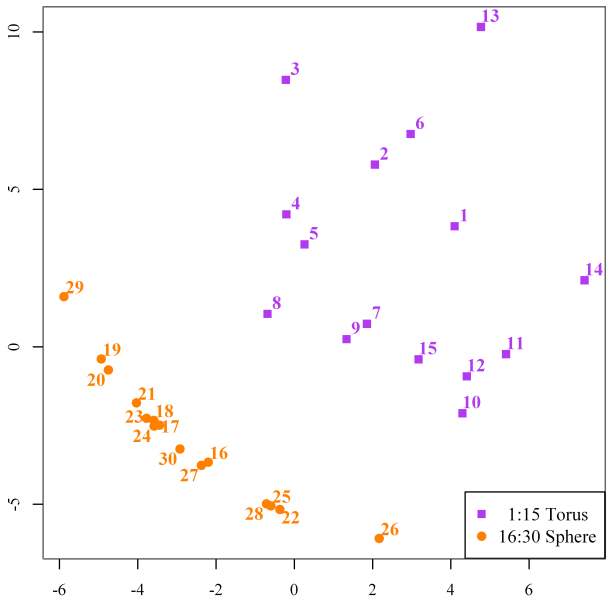

Figure 3.6: *Multidimensional scaling for Betti dimension 2.*

Again, we recall that the reason for this outcome lies in the geometrical difference between the torus and the sphere. As in the Figure 3.3 at the end of previous section, we can also observe that the variation within the group of spheres seems a bit less than that of the torus group.

Note that the visualization methods used in the current and previous sections serve only as our exploratory tools, but for a sound statistical analysis this is insufficient. We will now explore a method that involves a new data descriptor called *Persistence Landscape*.

Chapter 4: Persistence Landscapes

The theoretical framework for the method of *Persistence Landscapes* together with examples on applications in data analysis have been developed by Peter Bubenik, a mathematics professor at Cleveland State University[1]. Throughout this whole chapter, we follow the ideas from a recently published paper of Bubenik to introduce and implement persistence landscapes in our statistical data analysis.

4.A Polish Space: Complete and Separable

As pointed out in [3], when we apply persistent homology to statistical analysis of data, we would like to know if the resulting output, that is, the obtained sets of persistence intervals, allow us to calculate means, estimate variability, perform hypothesis testing, use convergence laws, etc. For implementations of the probability theory to a class of non-traditional objects such as persistence intervals, an appropriate mathematical environment is required. One such environment is a *Polish Space*, defined as a metric space that is separable and

[1]The *Persistence Landscapes* method was presented for the first time in Peter Bubenik's talk [3] during the Joint Mathematics Meetings held in Boston at the beginning of this year; another interesting presentation [4] with further developments on the topic took place during a Mathematical Biosciences Institute workshop in Columbus, Ohio at the end of last May. An article that will formally introduce Persistence Landscapes into science is expected to soon join the list of numerous publications of Bubenik.

complete. The importance of separability and completeness is explained in more detail as follows.

Separability of a metric space means that a countable dense subset exists in the given space. The crucial word here is "countable." Namely, as shown in [39], a measure μ on a sigma-field $\mathcal{A}$ (a non-empty class of sets that is closed under complements and countable unions), is a set function defined by:

1. null empty set: $\mu(\emptyset) = 0$
2. non-negativity: $\mu(A) \geq 0, \forall A \in \mathcal{A}$
3. countable additivity: $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n)$, for disjoint sets A_n in $\mathcal{A}$.

Therefore the notion of countability, and hence separability, is substantial for the measure theory; without this property we would not be able to introduce a probability measure.

Completeness of a metric space means that every Cauchy sequence in the given space converges and the limit is also in the space [41]. For example, the open unit interval $(0, 1)$ with the standard Euclidean metric is not complete because the limit of the harmonic sequence $x_n = \frac{1}{n}$, $n = 1, 2, 3, ...$, which is Cauchy, does not belong to the space. On the other hand, the closed interval $[0, 1]$ or, in general, any closed subset of $\mathbb{R}^n$ equipped with the usual Euclidean metric is complete. In any case, the property of completeness is important in probability theory because it gives rise to the notion of convergence.

Hence, in a Polish space a probability measure can be introduced. This result is crucial for applications of topology in statistical inference.

4.B Probability Space of Persistence Diagrams

As proven in recent article [27], the space of persistence diagrams, labeled as $\mathcal{D}$ together with the metric based on the Wasserstein distance represents a Polish space in which a probability measure, denoted by $\mathcal{P}$, can be defined. If $\mathcal{B}(\mathcal{D})$ is the usual Borel sigma-algebra in $\mathcal{D}$, then $(\mathcal{D}, \mathcal{B}(\mathcal{D}), \mathcal{P})$ represents a probability space in which expectations and variances are defined as follows.

Definition: The *Fréchet variance* is a quantity defined by

$$Var = \inf_{d \in \mathcal{D}} \left[\mathcal{F}(d) = \int_{e \in \mathcal{D}} W(d,e)^2 dP < \infty \right], \tag{4.1}$$

where $W(d,e)$ denotes the Wasserstein distance between persistence diagrams d and e. Then the *Fréchet mean* is defined as the quantity that minimizes the corresponding variance:

$$E = \{ d \,|\mathcal{F}(d) = Var \}. \tag{4.2}$$

The Fréchet mean defined above is a generalization of the usual mean and thus can be used in any metric space [5]. However, from the above expressions (4.2) and (4.1), it may seem that the implementation is far from trivial. Nevertheless, if a different but topologically equivalent space is used, calculations of the mean and the variance can be performed in a much easier way. Such a convenient setting for obtaining the Fréchet mean and variance is the metric space induced by a new data descriptor defined as the *Persistence Landscape*.

4.C Probability Space of Persistence Landscapes

In this section we briefly mention the results that lay out the theoretical background for statistical inference using persistence landscapes. The proof of each result is given in [5].

- The metric space induced by persistence landscapes is topologically equivalent to the Wasserstein distance. This is the main result, serving as the base for development of the theory.
- The space of persistence landscapes is separable and complete, hence it is a Polish space. Each of the results is separately stated and proved. Together, they imply that a probability measure can be introduced in the space of persistence landscapes, which is a crucial for statistical inference.
- In the probability space of persistence landscapes, the expectation of a random variable is defined as the Fréchet mean and the variance is the Fréchet variance.
- If the parameter in the definition 3.1 of Wasserstein distance is $p = 2$, then the mean and the variance can be calculated pointwise.
- In the above case when $p = 2$, the Limit Laws hold pointwise. Proofs are given for the Strong Law of Large Numbers and the Central Limit Theorem. The corollary of the Strong Law of Large Numbers is the pointwise convergence of the sample mean to the Fréchet mean.

These powerful tools give rise to a new area of research, which may be referred to as the *Statistical Topology using Persistence Landscapes* [5].

4.D Understanding Persistence Landscapes

To understand the construction of persistence landscapes [5], consider a set of finitely many persistence intervals $\{I_k\}_{k=1}^n$ where each interval has the form $I_i = (b_i, d_i)$, with finite non-decreasing endpoints so $b_i \leq d_i < \infty$. As usual, we refer to the left endpoint b_i as the "birth" time of a given bar and the right endpoint d_i is the "death" time. We depict persistence intervals using a homology plot i.e. a barcode, as shown below.

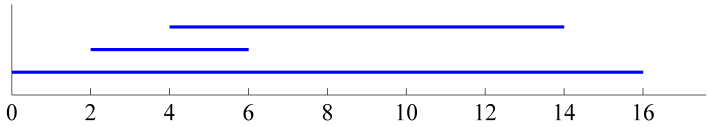

Figure 4.1: *A simple barcode with intervals* $\{(0, 16), (2, 6), (4, 14)\}$

The formation of a persistence landscape starts by constructing a triangle whose base corresponds to a generalized persistence interval (b_i, d_i) and the top vertex is in the intersection of the vertical line through the midpoint $\left(\frac{b_i+d_i}{2}, 0\right)$ and the circle passing through the endpoints, centered at the midpoint. The result is an isosceles right triangle whose catheti meet at $\left(\frac{b_i+d_i}{2}, \frac{d_i-b_i}{2}\right)$.

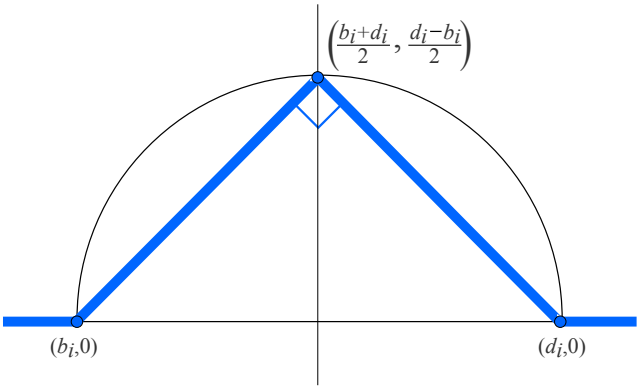

Figure 4.2: *Persistence landscape of an interval* (b_i, d_i)

Apply the same approach to all persistence intervals in a given barcode, as illustrated on our simple barcode example.

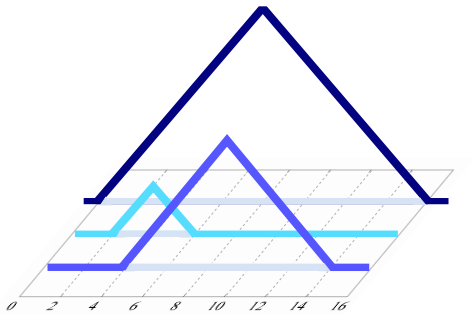

Figure 4.3: *Triangles constructed atop intervals* $\{(0,16),(2,6),(4,14)\}$.

Note that this is just the initial step in the construction of a persistence landscape; the triangles in the figure above do *not* represent a persistence landscape. The next stage of the construction can be described as the flattening all the triangles to a single vertical plane, as illustrated below.

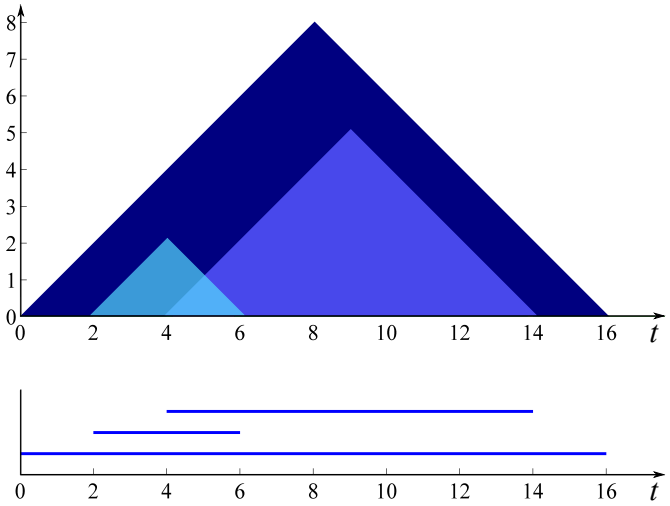

Figure 4.4: *Overlapping triangles in a vertical plane. The triangles are constructed over intervals* $\{(0,16),(2,6),(4,14)\}$*, depicted at the bottom of the image.*

With this construction the vertical plane becomes partitioned into polygon-shaped regions, each characterized by the number of overlapping triangles. As

shown in [5], let P_k, $k \in \mathbb{N}$, denote a union of regions populated by *at least* k triangles. Then the persistence landscape at a fixed value of k is a real-valued function $\lambda_k : \mathbb{R} \to \mathbb{R}$ that corresponds to the profile of P_k. Also, we take $\lambda_k(t_0) = 0$, when the vertical line positioned at $t = t_0$ does not intersect P_k. A more formal definition is given as follows.

Definition: Let (b, d) be a persistence interval, so $b \leq d$. Consider the map $f_{(b,d)} : \mathbb{R} \to \mathbb{R}$ such that

$$f_{(b,d)}(t) = \min(t - b, d - t)_+, \tag{4.3}$$

where the symbol "+" denotes the positive part, that is, $c_+ = \max(c, 0)$. Then the *persistence landscape* of a set of intervals $\{(b_i, d_i)\}_{i=1}^n$ represents a map $\lambda : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$ whose profile $\lambda_k : \mathbb{R} \to \mathbb{R}$ at a fixed $k \in \mathbb{N}$ is defined in the following way:

$$\lambda_k(t) = k^{th} \text{ largest value of } \left\{f_{(b_i,d_i)}\right\}_{i=1}^n, \tag{4.4}$$

for $t \in \mathbb{R}$. Furthermore, $\lambda_k(t) = 0$, for $k > n$.

We illustrate the definition on the simple barcode example. For intervals $\{(0, 16), (2, 6), (4, 14)\}$, the first birth time occurs at time $t = 0$ and the last death time is at $t = 16$, so times of interest lie in $(0, 16)$. Consider $t = 9$.

Table 4.1: Calculating the Persistence Landscape.

$\boldsymbol{b}$	$\boldsymbol{d}$	$\boldsymbol{t - b}$	$\boldsymbol{d - t}$	$\boldsymbol{\min(t - b, d - t)_+}$
0	16	9	7	7
2	6	7	-3	0
4	14	5	5	5

Thus $\lambda_k(9)$ takes values 7, 5, 0, for $k = 1, 2, 3$, respectively.

The persistence landscape contours for the given example are displayed below.

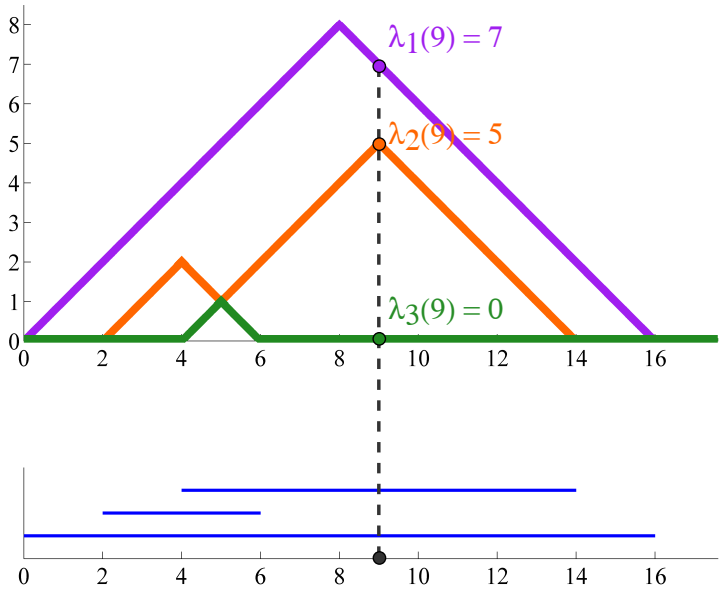

Figure 4.5: *Contours of a Persistence Landscape constructed on* $(0, 16)$*,* $(4, 14)$*, and* $(2, 6)$*.*

A 3d-plot of the corresponding persistence landscape is given below.

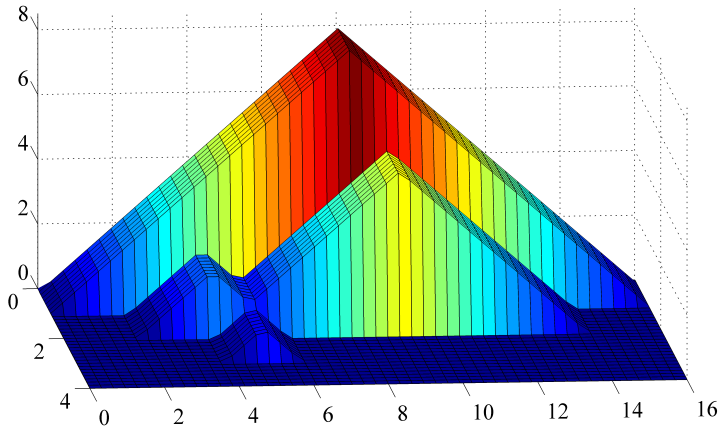

Figure 4.6: *3d-plot of a Persistence Landscape constructed on* $(0, 16)$*,* $(4, 14)$*, and* $(2, 6)$*.*

Note that the value $\lambda_k(t)$ of the persistence landscape for a particular value of $k \in \mathbb{N}$ and a fixed filtration time $t \in \mathbb{R}$, can be interpreted as the maximal

possible radius of an interval centered about t, where t belongs to k intervals in the given barcode [5].

Therefore, the persistence landscapes method allows us to capture two important pieces of information, as explained below.

- First, since longer intervals yield higher values, a persistence landscape carries information on lengths of intervals; this is important because longer bars are associated with features that persist through longer filtration times.

- Second, the information on the number of overlapping intervals at a fixed filtration time is also recorded; this is important because regions with high number of overlaps indicate short-lived bars which are usually considered as noise.

With these properties, we may conclude that the persistence landscapes method represents an excellent tool of data analysis in statistical topology. To implement the method, we use the MATLAB numerical software; though somewhat modified, our software relies on original codes generated by P. Bubenik.

Chapter 5: Application to Torus and Sphere

Upon laying out the underlying theory of persistence landscapes, we return to the statistical analysis of the fifteen point clouds sampled from each the torus and the sphere. Recall from the beginning of Chapter 3 that we had already generated persistence intervals for thirty samples; the data obtained will now serve as input for the new method. Our primary goal is to make a comparison between the torus and the sphere group. For that purpose, we first visually explore the persistence landscapes for the fifteen tori and the fifteen spheres at each of the three Betti dimensions. To avoid unnecessary repetitions, we display images of persistence landscapes for Betti dimension 1 only; note that complete sets of images across all Betti dimensions can be found in appendices A and B.

5.A Torus in Betti Dimension 1

Remember that the torus is characterized by two loops, one loop corresponding to the hole cutting through the middle and the other associated with the tunnel inside the toroidal tube. These properties of the torus were already detected

in the earlier homology plot 2.7 in Betti dimension 1. The same properties are reflected in the persistence landscapes as well; namely, for each of the fifteen tori the corresponding persistence landscape features two distinguished triangular peaks, as illustrated below.

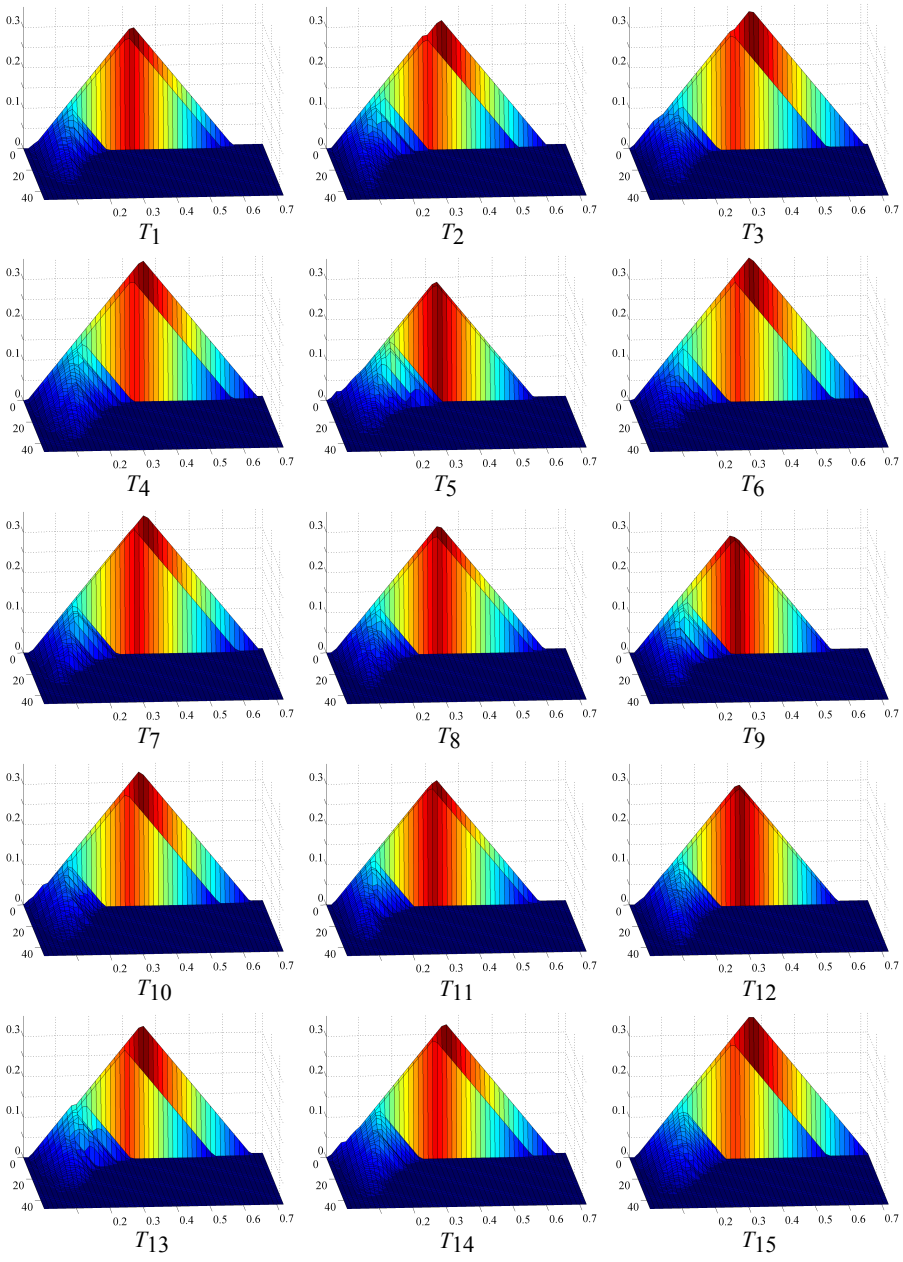

Figure 5.1: *Images of Persistence Landscape for the 15 tori in Betti dimension 1.*

As explained in section 4.C, we can obtain the Fréchet mean by pointwise averaging. The resulting 3d-plot of the average persistence landscape is displayed below.

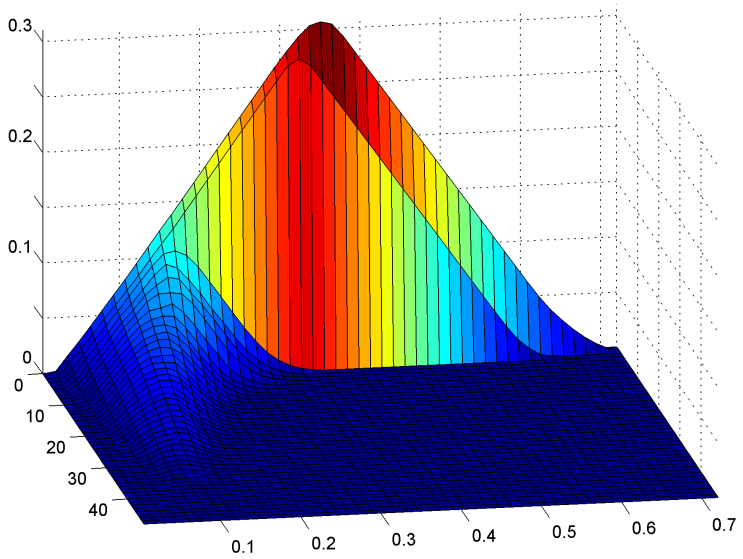

Figure 5.2: *Average Persistence Landscape for the torus group in Betti dimension 1.*

Note that the average above does *not* correspond to any particular barcode. Nonetheless, as shown in [5], it is possible to interpret the average persistence landscape.

> *Interpretation of the Average Persistence Landscape*: Consider a set $\mathfrak{B} = \{B_1, B_2, ..., B_n\}$ of n barcodes whose corresponding persistence landscapes, denoted by $\lambda^{(1)}, \lambda^{(2)}, ..., \lambda^{(n)}$, yield the average persistence landscape $\overline{\lambda}$. Then, for fixed $k \in \mathbb{N}$ and $t \in \mathbb{R}$, the value $\overline{\lambda}(k,t)$ represents the average of the maximal possible radius of an interval centered about t which belongs to k intervals of the set $\mathfrak{B}$.

This is a similar interpretation as the one given for the persistence landscape at the end of section 4.D.

Implementing the pointwise approach we calculate the Fréchet variance, that is, we rather focus on the corresponding standard deviation $s(k,t)$ which is then subtracted from and added to the empirical mean. This allows us to visualize the change in the average persistence landscape within a range of one standard deviation, as shown in the Figure 5.3 below.

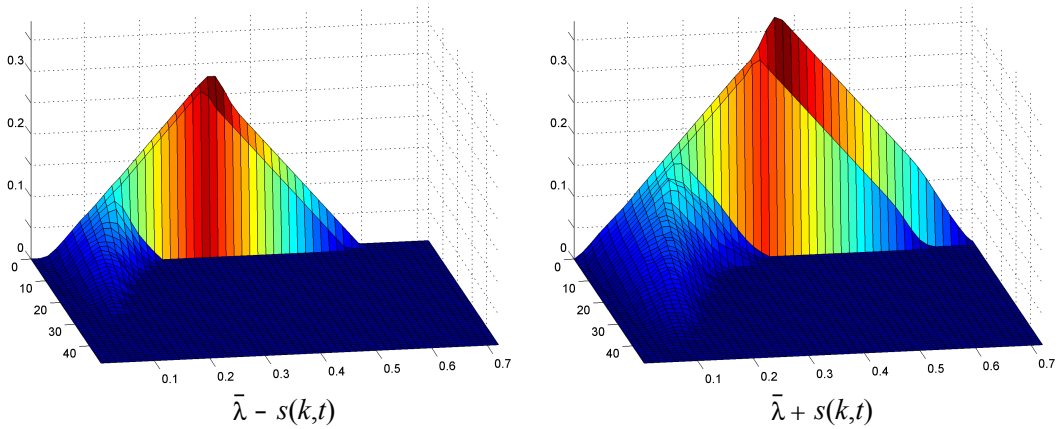

Figure 5.3: *The images to the left and right respectively show the change in the mean persistence landscape after subtracting and adding one standard deviation.*

Note that in the above calculations we ignored negative values that initially appear in the matrix corresponding to the left image; the reason for avoiding negatives is the fact that by its definition a persistence landscape is nonnegative. Nevertheless, a negative value of $\overline{\lambda} - s(k,t)$ can occur at some point $(k_0, t_0) \in \mathbb{N} \times \mathbb{R}$ if the variation in the fifteen values $\left\{\lambda^{(i)}(k_0,t_0)\right\}_{i=1}^{n}$ is such that it yields a standard deviation that exceeds the value of the mean at the given point, so $s(k_0,t_0) > \overline{\lambda}(k_0,t_0)$.

5.B Sphere in Betti Dimension 1

At Betti dimension 1, apart from smaller variations, the persistence landscapes of the fifteen spheres show no pronounced feature; this is due to the fact that the sphere has no loops, as already seen in the homology plot 2.7.

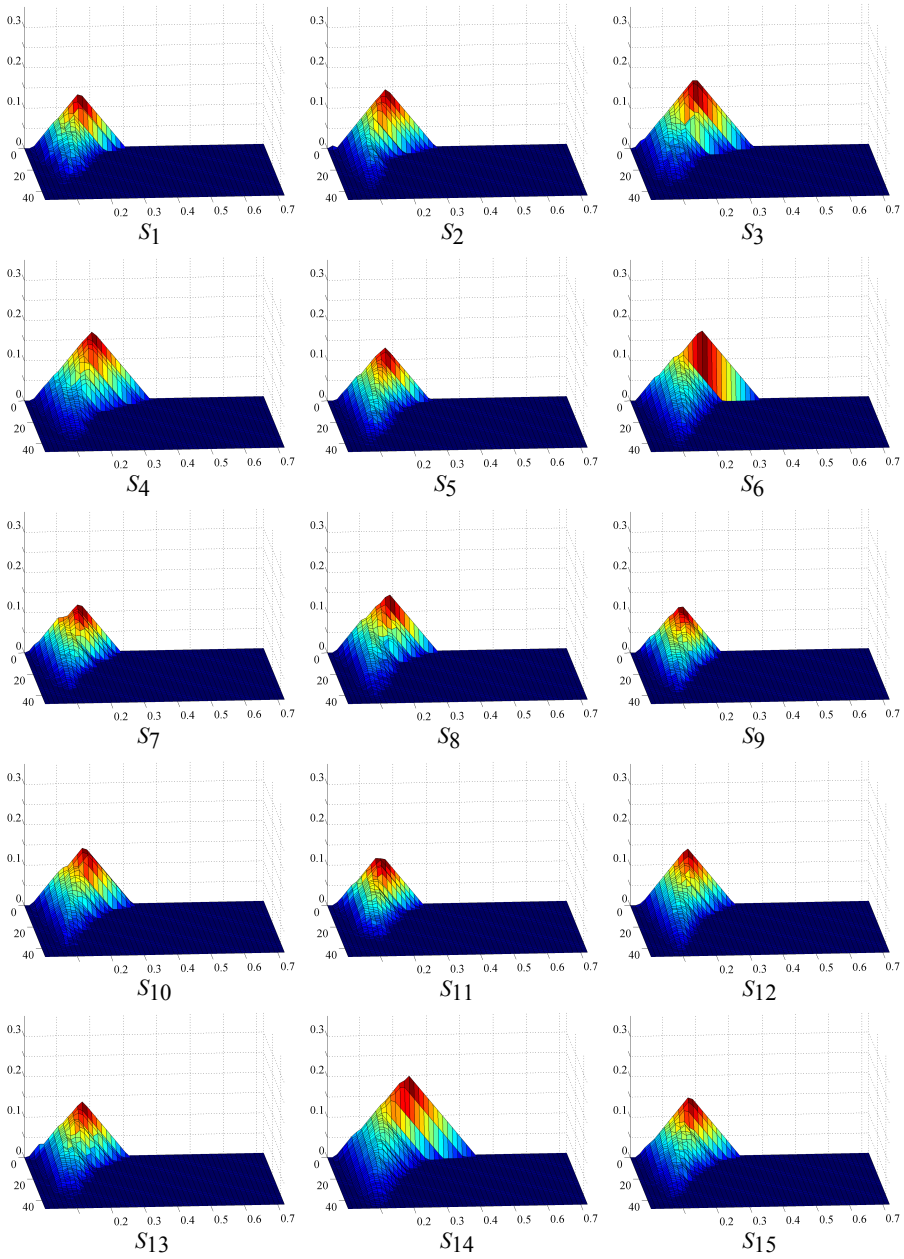

Figure 5.4: *Images of Persistence Landscape for the 15 spheres in Betti dimension 1.*

The mean persistence landscape obtained by pointwise averaging also shows a moderate structure.

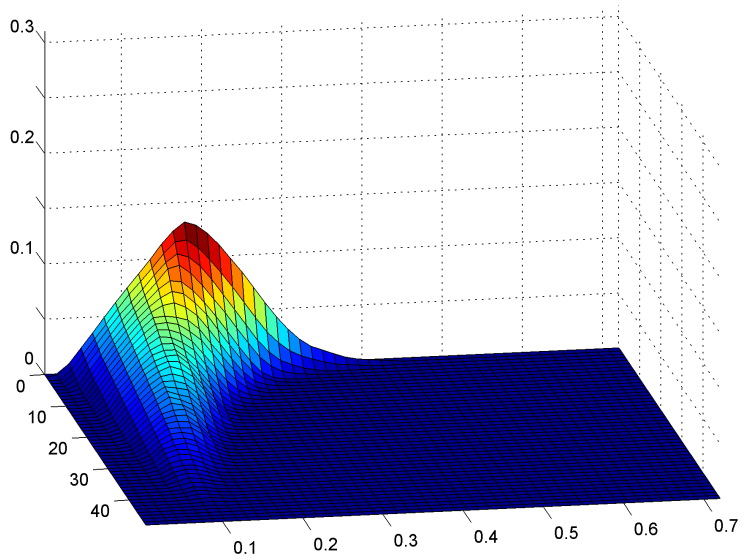

Figure 5.5: *Average Persistence Landscape for the spheres in Betti dimension 1.*

The change in the mean within a range of one standard deviation is shown below.

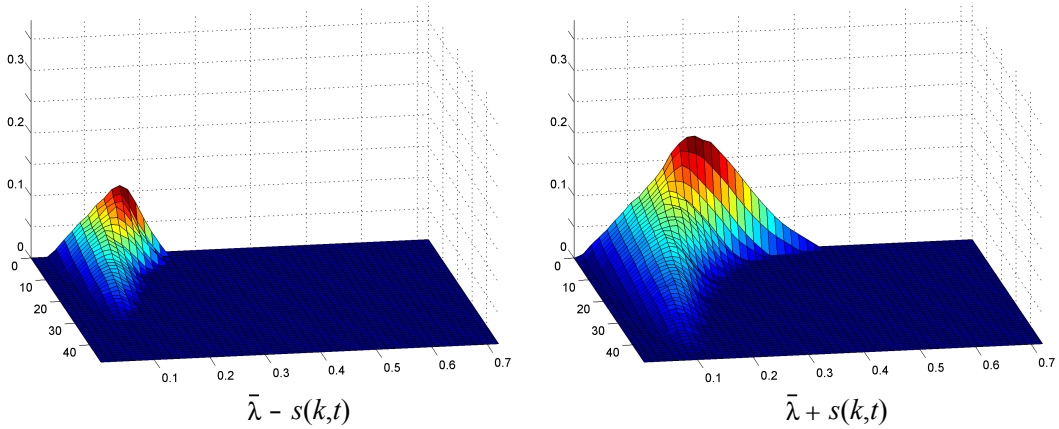

Figure 5.6: *The images to the left and right respectively show the change in the mean persistence landscape after subtracting and adding one standard deviation.*

5.C Visual Comparison of the Two Groups

Let us now visually compare the average persistence landscapes for the torus and the sphere across the three Betti dimensions.

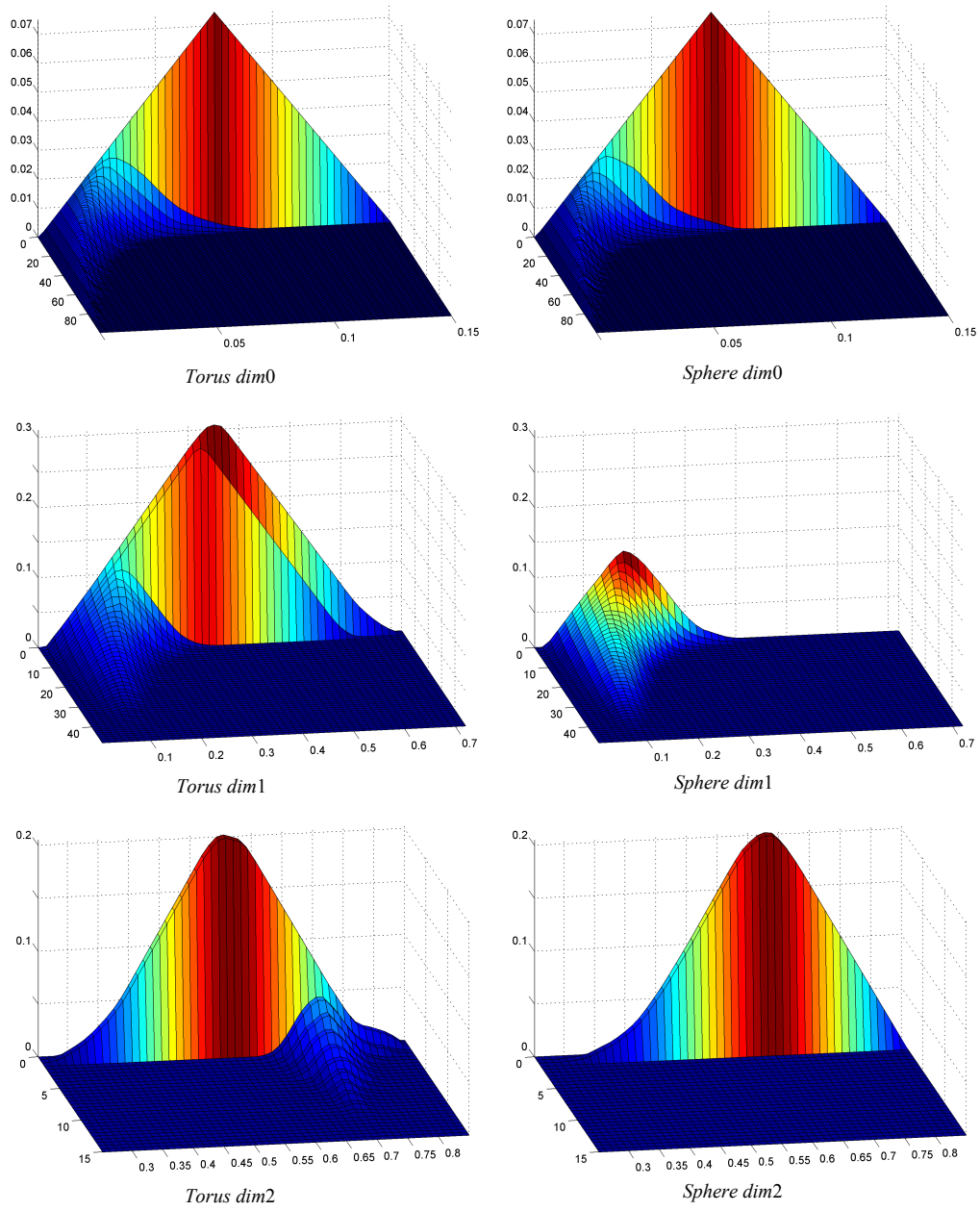

Figure 5.7: *The figure represents average persistence landscapes for the torus group (left) and the sphere group (right) for Betti dimensions 0, 1, and 2.*

Observe that there is no apparent difference between the average persistence landscapes of the torus and the sphere in Betti dimension 0. A slight difference occurs in Betti dimension 2 due to the specific geometry of each manifold. The largest difference occurs in Betti dimension 1. Nonetheless, prior to making any conclusions, methods of statistical analysis should be implemented to investigate which differences are statistically significant. Our main tool will be a permutation test.

5.D Rearranging the Data

Before the actual analysis, we appropriately rearrange the data obtained from persistence landscapes. This approach, originally proposed and implemented in data analysis by G. Heo [17], yields a vector whose entries are obtained by calculating the average number of overlapping intervals at each filtration value. The main idea is illustrated below.

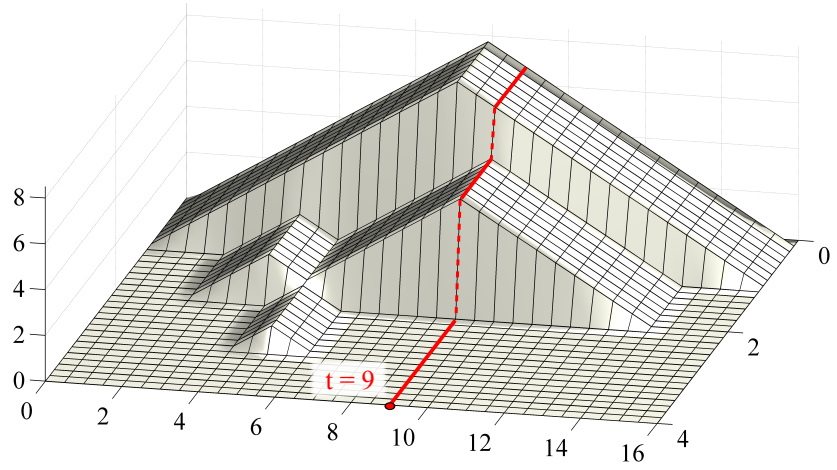

Figure 5.8: *Illustration of the idea of data rearrangement providing a more convenient input format. For instance, consider the filtration time $t_* = 9$ for which we sum up all the values of the persistence landscape function along the red line; divide the obtained sum by the maximal number of overlapping intervals $m_* = 2$. Implement this approach for all filtration values on the given grid.*

A more precise description is provided as follows.

> *Formatting of the persistence landscape data*: Consider a persistence landscape $\lambda(k,t)$ whose values are defined on a grid of size $m \times n$ with nodes (k,t). Recall that $\lambda_k(t)$ describes the contour of the persistence landscape at a particular k. Now, for a fixed filtration time $t = t_*$ at which the number of overlapping intervals k takes m_* distinct values $k = 1, 2, 3, ..., m_*$, define a new variable:
>
> $$\eta(t_*) = \frac{\sum_{k=1}^{m_*} \lambda_k(t_*)}{m_*}. \tag{5.1}$$
>
> Repeat this for every $t \in \{t_1, t_2, ..., t_n\}$ to obtain a vector of the form $\eta = [\eta(t_1), \eta(t_2), \ldots, \eta(t_n)]$. Call this vector η the *average persistence landscape curve*, since it is obtained by averaging the contours of the persistence landscape.

That way, instead of a persistence landscape in a shape of a matrix of size $m \times n$, a vector of length n is obtained which is convenient for our further analysis. Note that in this process we did not lose too much in the quality of data because every entry in the new vector stores information on the number of overlapping intervals at a given filtration value.

5.E Permutation Tests

To compare *two* groups of samples, each described by a vector of length n, some knowledge about the underlying distribution would be needed in order to construct a test statistic. When we do not posses such knowledge, we use resampling methods to gain information on the given distribution. Depending

on the way of resampling, there exist various methods, e.g. bootstrapping, jackknifing, or a permutation test. A necessary condition for implementing this methodology is that the drawn samples are representative of their populations; since our datasets are randomly generated, there is no reason to suspect the condition does not hold.

A permutation test is based on repeated sampling a large number of times, at least 1,000 or possibly 10,000. In each permutation the existing observations are randomly assigned to the two groups and an appropriate test statistic is calculated. These values represent the *null distribution* i.e. the probability distribution of the test statistic under the assumption that there is no difference among the two groups. Our goal is to test if this assumption, which we call the *null hypothesis*, may be rejected. If the observed statistic takes an extreme value in the null distribution, then it is highly unlikely that such a value occurred by chance so there is strong evidence against the null hypothesis. For a simple illustration, consider the figure below.

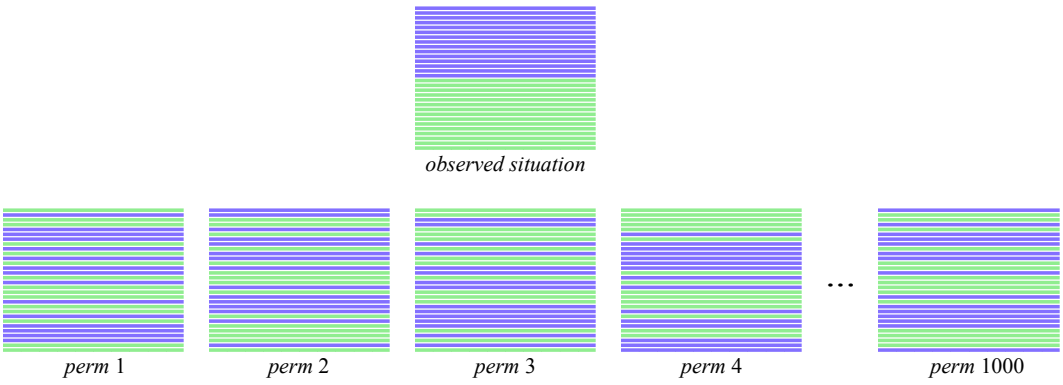

Figure 5.9: *Illustration of a permutation test in the case of a false null hypothesis. The image at the top represents the observed situation in which a group of 15 blue bars is distinguished from a group of 15 green bars (where a "bar" corresponds to a data vector). Note how extreme is the observed situation compared to configurations obtained by shuffling the bars in a random fashion; this indicates a significant difference among the groups.*

Recall that the data which we are going to submit to a permutation test are high-dimensional since each of the thirty observations is expressed in a form of a vector $\eta = [\eta(t_1), \eta(t_2), \ldots, \eta(t_n)]$ called the *average persistence landscape curve*. The length n of this vector corresponds to the number of distinct filtration times (or the number of nodes along the t-axis in the grid over which a persistence landscape is constructed). Due to such a high-dimensional input variable we need to use a multivariate statistic in our analysis; a convenient statistic can be found in Ramsay's *Functional Data Analysis* [34]. At this point, we emphasize that the implementation of methods of functional data analysis is appropriate since our data change with filtration in the same manner as some other data vary over the time continuum. That way, based on the above mentioned reference, we introduce a formula that represents the basis of our permutation test.

Permutation Test formula: Let $x_1(t)$ and $x_2(t)$ respectively denote two data vectors consisting of n_1 and n_2 observations sampled at a particular filtration time t. Define:

$$T(t) = \frac{|\overline{x}_1(t) - \overline{x}_2(t)|}{\sqrt{\frac{1}{n_1} Var(x_1(t)) + \frac{1}{n_2} Var(x_2(t))}} \tag{5.2}$$

as the test statistic of the permutation test.

The above defined statistic will serve as the main guideline in constructing our permutation test. Note that apart from the above defined test statistic some other forms of test statistics may be used, as shown in [33].

Calculating the Test Statistic

The observed test statistic and its maximal value are obtained from non-permuted data, that is, from the thirty observations given in a form of earlier defined η-vectors describing the average persistence landscape curve. Since there are $n_1 = n_2 = 15$ observations in each the torus and the sphere group across n distinct filtration times, the observations assemble a data matrix with $n_1 + n_2 = 30$ rows and n columns [1]. For this data matrix, the observed test statistic and the corresponding maximal value are calculated as follows.

- *Observed test statistic*: Applying the equation (5.2) to non-permuted data, obtain the following values:

$$T_{obs}(t), \text{ where } t \in \{t_1, t_2, \ldots, t_n\} \tag{5.3}$$

  Since $T_{obs}(t)$ corresponds to a single time point t, we may think about it as a *pointwise* statistic.

- *Maximal Observed test statistic*:

$$T_{\max\_obs} = \max_{t \in \{t_1, .., t_n\}} T_{obs}(t), \tag{5.4}$$

  Thus $T_{\max\_obs}$ represents the maximum in the array of $T_{obs}(t)$ values.

[1] The number n of filtration times is user-defined; in our codes the range of filtrations is divided into 50 equally spaced subintervals, which yields $n = 51$ distinct points.

Calculating the Null Values

As explained in [34], the null distribution is generated from the observed data by performing N different random permutations, where N is a large number. Every permutation results in a new arrangement of η-vectors but the order of components within each vector remains unchanged; in other words, only the rows in the data matrix are rearranged while the ordering of the columns is the same. In each such permutation, an array of null test statistic values and their maximum are obtained as follows.

- *Null test statistic of the i-th permutation*: Consider a random shuffling of rows in the observed data matrix; applying the equation (5.2) to the given arrangement, calculate the following values:

$$T^i_{null}(t), \text{ where } t \in \{t_1, t_2, \ldots, t_n\}. \tag{5.5}$$

  Since $T^i_{null}(t)$ corresponds to a point t in filtration time, we may think of it as a pointwise null value in the i-th permutation. Every permutation $i = 1, 2, \ldots, N$ yields an array of null values across n filtration times, hence the null values form a matrix of size $N \times n$.

- *Maximal null test statistic of the i-th permutation*:

$$T^i_{\max\_null} = \max_{t \in \{t_1, .., t_n\}} T^i_{null}(t). \tag{5.6}$$

  Thus $T^i_{\max\_null}$ is the maximal value in the array $T^i_{null}(t)$. There are N such values, one for each permutation.

p-value of the Test

If the two compared groups are statistically indistinguishable, then random permutations applied to the rows of observed data do not make a difference; in that case, the observed test statistic blends in the null distribution. On the other hand, if the two groups statistically differ, then random permutations do make a difference; in that case the observed test statistic takes an extreme value i.e. it is located in the tail of the null distribution. Therefore, the p-value of the permutation test can be obtained as the proportion of null values which exceed the observed test statistic. The smaller this value, the less likely that the observed data occurred by chance.

- *Permutation Test p-value*:

$$p\text{-value} = mean\left\{T_{\max\_obs} < T^{i}_{\max\_null}\right\}_{i=1}^{N}, \tag{5.7}$$

that is, the p-value corresponds to the average number of cases in which the maximal observed test statistic falls below a maximal null value.

The outcome of the test is obtained by comparing the p-value with the significance level $\alpha = 0.05$. When p-value < 0.05, we are able to reject the null hypothesis.

Visualization of Test Results

For the purpose of visualizing the results of the permutation test, the observed test statistic $T_{obs}(t)$ is plotted against filtration time, together with the maximal critical value $C_{\max}$ which we define as follows.

- *Maximal Critical value*:

$$C_{\max} = 0.95 \; quantile \left\{ T^i_{\max\_null} \right\}_{i=1}^{N} \tag{5.8}$$

  that is, the maximal critical value represents the 0.95-th quantile in the null distribution of maximal null values $\{T^i_{\max\_null}\}_{i=1}^{N}$.

Since $C_{\max}$ is a constant, its plot is a horizontal line. This line represents a threshold value at which the difference between the two compared groups becomes significant. Namely, for filtration times at which the observed test statistic crosses the maximal critical value, a statistically significant difference between the two groups exists. With this descriptive tool, we expect to gain more information about the homological and geometrical difference between the torus and the sphere.

Note that in all subsequent permutation test plots, the observed test statistic $T_{obs}(t)$ is represented by a blue curve, while the maximal critical value $C_{\max}$ is given by a red horizontal line.

Advantages of Permutation Tests

At the end of the section we point out some of the advantages of permutation tests. As explained in [28], one of the main advantages is that the normality condition is not required; hence, unlike the usual t-test, this method can be

applied to any data, regardless of its nature. Another advantage is that these tests are robust in the sense that accurate p-values are obtained even in cases when the two distributions compared have different standard deviations. With these properties and its easiness to implement, permutation tests are becoming increasingly popular. Nevertheless, it is interesting to note that ever since permutation methods were introduced into statistics by Fisher in 1935, there has been some scepticism on behalf of users about the reliability of the method [36]; contrary to this, the truth is that when a permutation test is properly implemented with randomization and sufficient number of permutations, the method actually represents a basis for exact inference [11]. This is explained in the following excerpt, taken from [22].

> "...tables of critical values in nonparametric statistical tests for small sample sizes are based on permutations. The authors of these tables have computed how many cases can be found, in the complete permutation distribution, that are as extreme as, or more extreme than the computed value of the statistic. Hence, probability statements obtained from small-sample nonparametric tests are exact probabilities."

Of course, for our thirty samples the idea of performing all the possible permutations is computationally unattainable since $30! \approx 2.65 \cdot 10^{32}$, but now we understand why performing more permutations may lead to better results. A usual recommendation as given in [22] is that 1,000 permutations suffice when we encounter and explore a problem but for final results it would be better to have at least 10,000 permutations.

5.F Results

To apply the described algorithm, a suitable code is written by combining the the `tperm.fd` procedure of Ramsay [34] with a code obtained from G. Heo. Using this code, we perform the permutation test on the thirty samples in order to statistically evaluate the significance of the difference among the torus and the sphere data. First, we establish the test hypotheses.

Permutation Test Hypotheses

Recall from section 5.D, that the input data for the permutation test are given as η-vectors which we call *average persistence landscape curves*. This implies that there exists a population of average persistence curves with a mean denoted by $\mu^{(\eta)}$. Given the torus and the sphere group, we differentiate among two particular populations of average persistence landscape curves:

η_T: Population of average persistence landscape curves for the torus, described by the mean $\mu_T^{(\eta)}$.

η_S: Population of average persistence landscape curves for the sphere, described by the mean $\mu_S^{(\eta)}$.

Then the fifteen η-vectors corresponding to the torus represent observations from the population η_T and similarly, the fifteen η-vectors corresponding to the sphere are observations obtained from the population η_S. Based on this, we can now provide the test hypotheses.

For the permutation test that compares the torus and the sphere data, the statistical assumptions under the null hypothesis and under the alternative hypothesis are defined as follows:

$$\begin{aligned} H_0 &: \ \mu_T^{(\eta)} = \mu_S^{(\eta)} \\ H_1 &: \ \mu_T^{(\eta)} \neq \mu_S^{(\eta)} \end{aligned} \tag{5.9}$$

The outcome of the test is determined according to the p-value. Since the test is performed at $\alpha = 0.05$ level of significance, the decision about rejecting the null hypothesis is made by comparing the p-value with the significance level:

$$\begin{aligned} p\text{-value} \ \leq \ 0.05 \ \ &\Rightarrow \ \ \text{reject} \, H_0 \\ p\text{-value} \ > \ 0.05 \ \ &\Rightarrow \ \ \text{not reject} \, H_0 \end{aligned} \tag{5.10}$$

In all permutation tests on the torus and the sphere, the same significance level will be used, i.e. our decision in all hypothesis testing will be based on the above shown criterion (5.10).

Three Approaches

To obtain more information from the permutation test, we use three approaches in terms of taking the data from persistence landscapes.

1. In the first approach, initially the data are assembled from the complete persistence landscapes of the 15 tori and 15 spheres. To implement a permutation test we need properly formatted data so we apply the method proposed by G. Heo, as shown in equation (5.1). Since all peaks

in each of the thirty persistence landscapes are taken into account, we denote this setting by "all."

2. In the second approach, instead of using all the peaks i.e. contours from an individual persistence landscape, we consider only the first (highest) peak which corresponds to the contour $\lambda_1(t)$ in a persistence landscape. Since such data already form a vector, the formatting step from section 5.D has no purpose. The gain we expect from this approach concerns Betti dimensions 0 and 2; in these dimensions, the highest peak in the persistence landscape seems to be the same for both the torus and the sphere. Since only the first peak is taken into account, we denote this approach as "peak 1."

3. In the third approach, every sample is described via all but the first two peaks in its corresponding persistence landscape; in other words, contours $\lambda_1(t)$ and $\lambda_2(t)$ are excluded from consideration. Since this includes several rows from the data matrix of a persistence landscape, we need to "flatten" the data into a vector according to the equation (5.1). We expect this approach to be helpful in our analysis of the Betti dimension 1 in which the two high peaks from persistence landscape of the torus make the difference with respect to the sphere. Our notation for this particular setting is "all but 1 and 2."

Note that the application of these different approaches is adopted from Bubenik's article [5], where similar selections of peaks from persistence landscapes was considered.

Permutation Test Results

The results obtained after 10,000 permutations for the $\alpha = 0.05$ level of significance are displayed in Table 5.1.

Table 5.1: p-values for permutation tests at $\alpha = 0.05$.

Betti dim	**Peaks considered**	**p-value**
0	all	0.1582
1	all	0.0000
2	all	0.0000
0	peak 1	1.0000
1	all but 1 and 2	0.1774
2	peak 1	0.0000

The comments on the above results are given as follows.

- At Betti dimension 0, if all peaks appearing in persistence landscapes are considered, the permutation test in Betti dimension 0 yields a p-value of 0.1582. Hence for the given $\alpha = 0.05$ (and any other) level of significance no evidence exists against the null hypothesis which means there is no statistically significant difference between the torus and the sphere group. Moreover, if only the first peak is taken into account, the p-value of 1.0000 shows a perfect match between the two groups.

- At Betti dimension 1, when all peaks are considered, a practically zero p-value shows compelling evidence *against* the null hypothesis. Thus, as expected, the torus and the sphere do significantly differ in Betti dimension 1. However, if the two highest peaks are excluded from analysis, a p-value of 0.1774 implies no significant difference among the two groups, i.e. our statistical evidence confirms that the two loops of the torus cause the difference between the two manifolds in Betti dimension 1.

- At Betti dimension 2, whether all peaks are considered or only the first peak is included into the analysis, a p-value of 0.0000 is obtained, implying a significant difference. This happens because in Betti dimension 2 the focus is on the void inside the manifolds which involves not only topological but also geometrical features. Thus, as already stated in [5], homology detects both topological and geometrical properties.

Permutation Test Plots

More information is obtained by plotting pointwise values of the observed test statistic as shown in Figure 5.10. This curve, given in blue, is compared to the maximal critical value of the given permutation test, represented by a horizontal line. Since the blue curve of the observed test statistic never crosses the red line corresponding to the maximal critical value, it follows that in the Betti dimension 0 the torus and the sphere are practically indistinguishable at *all* filtration times.

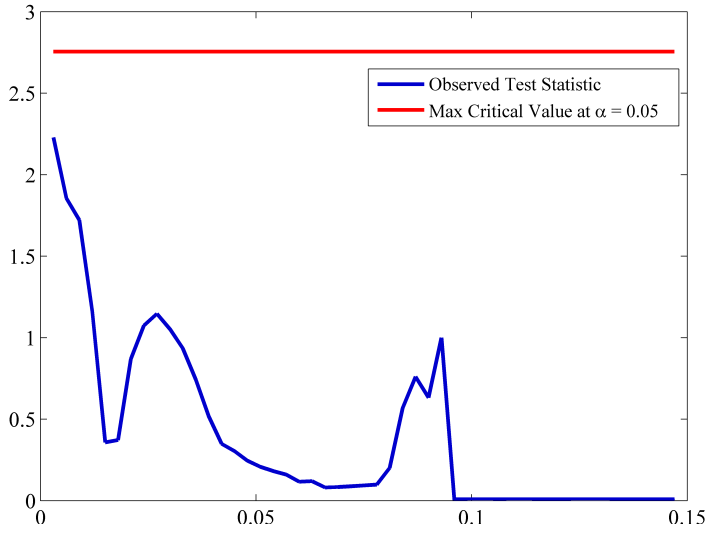

Figure 5.10: *Permutation test for the difference between the torus and the sphere in Betti dimension 0. The observed statistic never crosses the critical value. The pre-processing of data prior to the permutation test is carried out according to equation (5.1)*

Unlike this, in Betti dimension 1 a significant difference exists for filtrations $(0.23, 0.66)$, as seen in the Figure 5.11. During these times the two loops of torus are pronounced, causing a striking difference between the two manifolds.

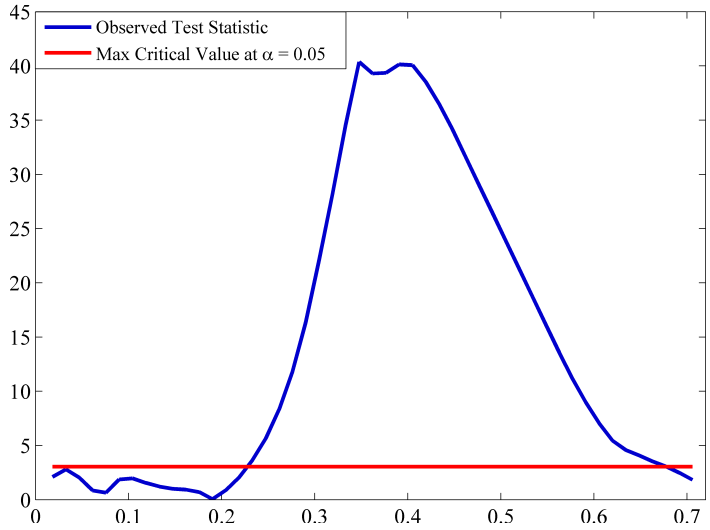

Figure 5.11: *Permutation test for the difference between the torus and the sphere in Betti dimension 1; a significant difference is present most of the time.*

However, if the two highest peaks are ignored, no difference in Betti dimension 1 can be observed, as depicted in Figure 5.12.

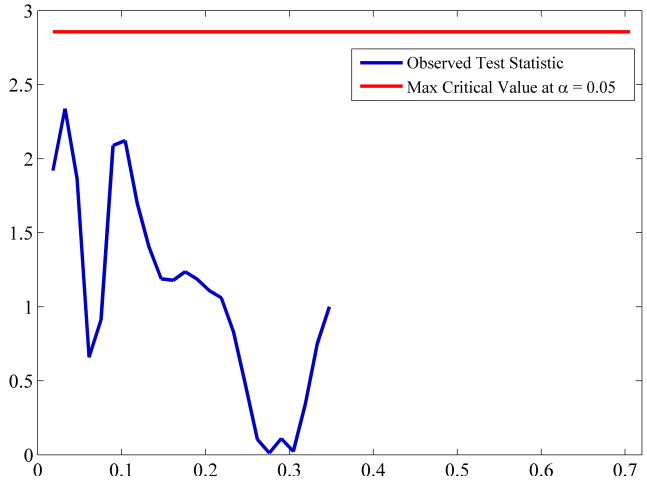

Figure 5.12: *Permutation test for the difference between the torus and the sphere in Betti dimension 1 when the two highest peaks are ignored; no significant difference is detected. At about time 0.35 small peaks disappear so no more homological activity is detected; thus there are no observed values after* $t = 0.35$.

Figure 5.13 shows that in Betti dimension 2 a significant difference repeatedly appears and disappears starting from about $t = 0.4$ until $t = 0.77$, when a permanent difference settles in, as all cavities in the torus fill in.

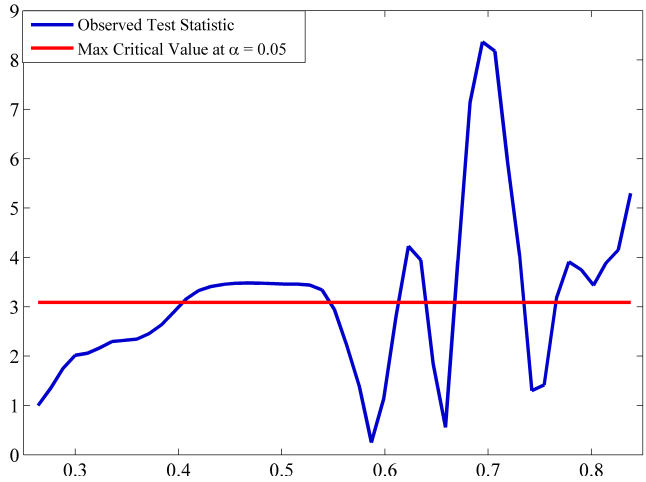

Figure 5.13: *Permutation test for the difference between the torus and the sphere in Betti dimension 2. Significant difference occurs in a repeating pattern.*

Figure 5.14 shows if only the highest peak is considered, a small difference occurs in $(0.40, 0.55)$. The major difference in a form of a large spike at about $t = 0.61$ marks the moment when the main void in the torus disappears.

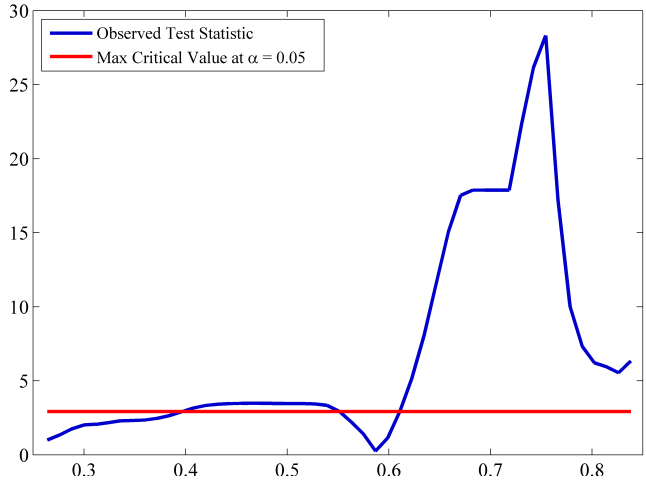

Figure 5.14: *Permutation test for the difference between the torus and the sphere in Betti dimension 2 when only the first peak is taken into account.*

At this point we emphasize that even though most of the main results have already been known, the fine details depicted in the plots obtained from the permutation tests are certainly appreciated. These plots not only prove the efficiency of a permutation test, but also demonstrate the ability of the persistence landscapes method to capture both geometrical and topological features of objects.

Note that the above results rather correspond to an ideal situation when there is no error or very little error exists among the observed data. This is not a real situation since most datasets carry errors that can go even up to 15% or higher. Therefore, our next goal is to analyze a more realistic setup which involves a noise component as well.

Chapter 6: Noisy Torus and Sphere

Upon analyzing non-perturbed point-clouds we investigate the influence of noise as an inevitable component of every real datasets. For that purpose after generating point-clouds on either a torus or a sphere, Gaussian noise is added to induce a random displacement of points. That way each point moves within a small ball whose radius is a random variable from normal distribution with zero mean and standard deviation whose value corresponds to the assigned level of noise, calculated as a fraction of the given point's distance from the origin. In addition to this, as suggested by H. Adams, each point-cloud acquires 200 sparse outliers which are generated uniformly randomly in a box with dimensions $[-3, 3] \times [-3, 3] \times [-3, 3]$. An illustration of sparse outliers scattered inside the box is provided in Figure 6.1.

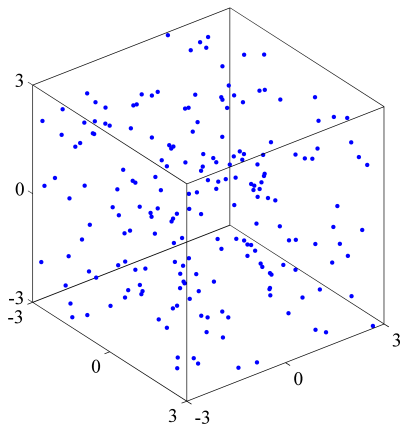

Figure 6.1: *200 sparse outliers scattered in a cube with sides of length 6.*

Typical point-clouds obtained this way are depicted below for two different levels of Gaussian noise.

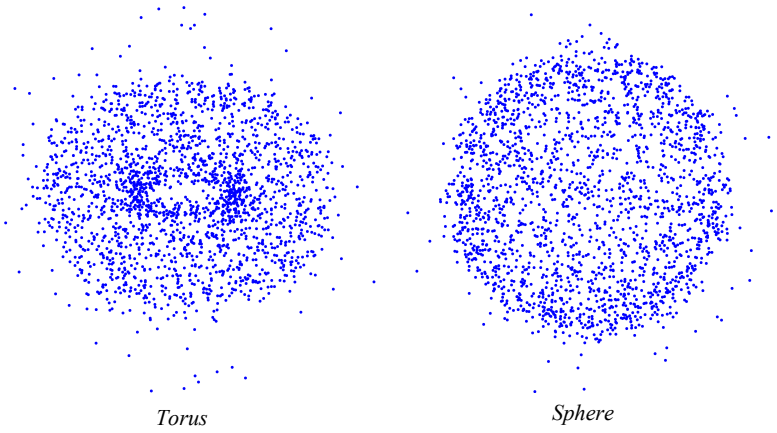

Figure 6.2: *Point-clouds after adding 7.5% of Gaussian noise and 200 sparse outliers.*

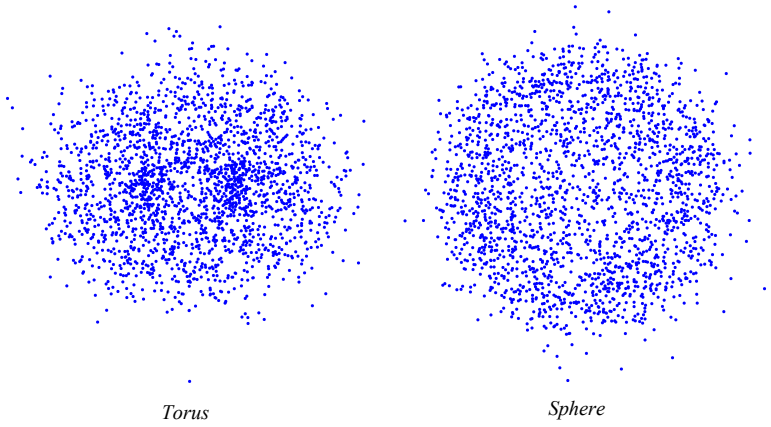

Figure 6.3: *Point-clouds after adding 15% of Gaussian noise and 200 sparse outliers.*

Observe the shrinkage of the hole through the center of the torus as the amount of the Gaussian noise increases. At the level of 7.5% of noise (Figure 6.2), the opening is still fairly visible, whereas at 15% of noise, when looking from the same viewpoint, it can be barely detected (Figure 6.3). Such noisy datasets cannot be successfully analyzed unless smoothing of the data is performed prior to the analysis; for that purpose, we implement appropriate kernel density estimation methods.

6.A Kernel Density Functions

Kernel density estimation (KDE) is a sophisticated tool of non-parametric statistics used for estimating probability density functions of variables with unknown distribution. Though plenty of literature exists for univariate data, much less is available in the multivariate case. The reasons are issues that arise in high-dimensional spaces; namely, if the domain of a density function is subdivided by a Cartesian grid, it can happen that too few observations are captured from high-density regions; in such situation, known as the *empty space phenomenon* [38], the majority of observations arrive from tails of the distribution which means low density regions gain influence on the estimate. Another issue is that the sample size needed for accuracy of the estimation, measured through the integrated mean square error, rapidly increases with the dimensionality of data [40]. Also, the choice of the smoothing parameter (bandwidth), becomes more difficult in higher dimensional spaces [48]. Thus obtaining a kernel density estimate for multivariate data is not a trivial task; to solve the problem we follow guidelines provided by Silverman [40]. First, we define a density estimator.

Multivariate Kernel Density Estimator: Let $X_1, X_2, ..., X_n$ be samples in a d-dimensional space $\mathbb{R}^d$. Then the kernel density estimator at a point $X \in R^d$ is defined as

$$f(X) = \frac{1}{nh^d} \sum_{j=1}^{n} K\left\{\frac{1}{h}(X - X_j)\right\}. \tag{6.1}$$

Here h denotes the value of the smoothing parameter and K the *kernel density function* normalized so that its integral corresponds to the total probability, i.e. $\int_{R^d} K(X)\, dX = 1$.

Considering the kernel function K, its form depends on the nature of data. A common choice is the multivariate standard normal (Gaussian) density:

$$K_G(X) = (2\pi)^{-\frac{d}{2}} e^{-\frac{1}{2}X^T X}. \tag{6.2}$$

The multivariate normal kernel may not represent the best choice for our data because each point cloud was sampled from a *uniform* distribution, and also, the small Gaussian error is insufficient for normality. Furthermore, even if the x, y, and z vectors in a point-cloud were univariate normal, their joint density could still be non-normal. To verify formally that our data are not multivariate normal we implement a suitable code [15] and perform a *Mardia test*. This test uses measures of multivariate skewness and kurtosis [1] that were introduced to statistics by K. V. Mardia [25]. Thus there are two separate tests, one for skewness and one for kurtosis; only if both do *not* reject the null hypothesis, multivariate normality of data may be assumed.

Table 6.1: Mardia test for normality of noisy point-clouds on each torus and sphere.

		Torus		**reject**	**Sphere**		**reject**
Noise	Test	***t*-stat**	***C.V.***	**H_0**	***t*-stat**	***C.V.***	**H_0**
7.5%	skewness	1.536	18.307	No	0.701	18.307	No
	kurtosis	-15.128	1.645	Yes	-23.6905	1.645	Yes
15.0%	skewness	4.699	18.307	No	2.930	18.307	No
	kurtosis	-10.700	1.645	Yes	-21.236	1.645	Yes

t-stat – test statistic; $C.V.$ – critical value at significance level $\alpha = 0.05$; H_0 – null hypothesis that data are multivariate normal.

[1] As defined by Mardia [26], suitable measures for multivariate skewness and kurtosis are $\beta_{1,d} = E\Big((X-\mu)^T \Sigma^{-1} (Y-\mu)\Big)^3$ and $\beta_{2,d} = E\Big((X-\mu)^T \Sigma^{-1} (X-\mu)\Big)^2$, respectively. Here X and Y are d-dimensional independent identically distributed (iid) random variables with mean μ and covariance matrix Σ.

From Table 6.1 we note that in all kurtosis tests the magnitude of the test statistic exceeds the critical value; therefore we may reject the null hypothesis about multivariate normality and conclude that, as expected, our data are not multivariate normal. Hence it may be more suitable to seek a non-normal kernel density estimator. One appropriate choice would be the *Epanechnikov kernel*, defined as follows:

$$K_E(X) = \begin{cases} \frac{1}{2V_d}(d+2)(1-X^TX)\ , & X^TX < 1 \\ 0 \qquad\qquad , & else \end{cases} \tag{6.3}$$

The variable V_d represents the volume of the d-dimensional unit sphere, so for dimensions $d = 1, 2, 3, ...$ the corresponding volumes are $V_d = 2, \pi, \frac{4}{3}\pi, ...,$ respectively[2].

The decision to use the radially symmetric Epanechnikov kernel is based on the symmetries of the underlying objects from which the point-clouds were sampled; since the sphere is radially symmetric and the torus possesses the property of axial symmetry, the Epanechnikov kernel seems suitable. After making the decision on the choice of the kernel, it remains to focus on the task of finding the optimal value of the smoothing parameter h which we have seen in the expression (6.1) for the general form of a density estimator.

[2] In general, the volume of the d-dimensional hypersphere is $V_d = \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)}$ where Γ denotes the gamma function defined as $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$, where values $z = 0, -1, -2, ...$ are excluded as points where the function is not analytic. Important properties of the gamma function are $\Gamma(z+1) = z\Gamma(z)$ and $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

6.B Estimating the Bandwidth

In this section we briefly show the main ideas in the process of determining the bandwidth; due to the nature of our data, we implement theory that applies to radially symmetric density functions. Our references are Silverman [40] and Rosenblatt [37].

First we start with a definition. Let $\hat{f}$ be the estimated kernel and f the true density. Then the mean square error of the estimate $\hat{f}$ is defined as:

$$\begin{aligned} MSE(\hat{f}) = & \; E\Big(\hat{f} - E(f)\Big)^2 \\ = & \; bias^2(\hat{f}) - Var\,(f) \end{aligned} \tag{6.4}$$

where $bias(\hat{f}) = E(\hat{f}) - f$, as usual. Note that for simplicity the argument is omitted from each function, but in general it would be a vector $X \in \mathbb{R}^d$. Now assume that the density f and its second derivative f'' are continuous and bounded, with integrable squares [37]; then the integrated mean square error (IMSE) is of the form:

$$IMSE(\hat{f}) = \int bias^2\left(\hat{f}(X)\right)dX + \int Var(X)dX \tag{6.5}$$

The optimal bandwidth is obtained as the value that minimizes the integrated mean squared error. Without going into more detail, we just mention that the derivation involves the multidimensional version of Taylor's theorem which is applied to the bias and the variance to yield approximations that allow us to estimate h.

As shown in [40], the optimal smoothing parameter in the case of a radially symmetric kernel is given as follows:

$$h_{opt} \;=\; \sigma h_0 = \;\sigma\, A(K) n^{-\frac{1}{d+4}} \tag{6.6}$$

Here $A(K)$ represents a constant whose value depends on the type of the kernel K and the dimensionality of data. For the Epanechnikov kernel and $d = 2$ and 3, the constant takes values $A(K_E) = 2.40$ and 2.49, respectively[3].

Considering the scaling parameter σ, there may be different choices. One possibility, as suggested by Silverman, would be to find the variance of the data averaged over the d-dimensions, that is, $\sigma^2 = \frac{1}{d}\sum_{i=1}^{d}\sigma_i^2$. Our approach is somewhat different; instead we use the standard deviation $\sigma = [\sigma_x, \sigma_y, \sigma_z]$ of the data to obtain $h = [h_x, h_y, h_z]$, where $h_i = \sigma_i\, h_0$, $i = x, y, z$. Since the torus is symmetric about the z-axis, the bandwidths corresponding to the x and y axes are the same and half of that value is obtained for the z-axis due to the specific geometry of our torus. In the case of the sphere, the bandwidth is the same for all the three coordinate axes. Our estimates are displayed in the table below.

Table 6.2: Estimated smoothing parameter values.

	Torus			**Sphere**		
Noise	$\boldsymbol{h_x}$	$\boldsymbol{h_y}$	$\boldsymbol{h_z}$	$\boldsymbol{h_x}$	$\boldsymbol{h_y}$	$\boldsymbol{h_z}$
7.5%	1.26	1.26	0.63	1.36	1.36	1.36
15.0%	1.30	1.30	0.65	1.54	1.54	1.54

[3] For higher dimensions, take $A\,(K_E) = \left\{\frac{8}{V_d}\,(d+4)\,(2\sqrt{\pi})^d\right\}^{\frac{1}{d+4}}$

6.C Applying kd-trees

Due to the advantage of fast kernel density estimations, kd-tree structures are often implemented in persistent homology, e.g. see de Silva and Carlsson [6], Carlsson et al [7], and Carlsson and Adams [1] as main references in this section. Our codes use a kd-tree library developed by Tagliasacchi [42].

To explain the main idea in applying kd-trees, consider the term $\frac{1}{h}(X-X_j)$, that appears as the argument of the kernel in equation (6.1). This means the value of the density kernel at a fixed X depends on the distance $X - X_j$ of the given point from other points $j = 1, 2, \ldots n$ in the dataset. Using the notion of distance a special structure called a kd-tree is constructed so that the data are organized on the "nearest neighbor" criterion. Moreover, a convenient parameter k is introduced so that the k-th neighbor is taken into account in the process of kernel density estimation. Sampling from the kernel density estimate the original data are replaced by a new point-cloud called the *core density subset*; depending on the choice of the parameter, different core subsets and thus different density estimates can be obtained. For low values of k the kernel density is rather locally estimated whereas high values of k yield more global estimates. This variety of perspectives can provide a valuable insight to the topology of the underlying space, so often multiple core subsets are considered.

In terms of applying kd-tree structures to our data, we first recall that our point-clouds contain sparse outliers. In such case a kernel density estimate of the form $\dfrac{1}{\rho_k}$, that is, the reciprocal of the distance $\rho_k(X)$ of a point X from its k-th closest neighbor, is particularly convenient. Namely, if the distance ρ_k is relatively small, then the k-th neighbor lies in the vicinity of the point

X so the density is high in that region. On the other hand, large ρ_k implies that the k-th neighbor lies far away from the point X which means the space in-between is not very populated and thus the density of points in that region is low. Thus the inverse relationship between the distance and the density function comes as a natural choice since the densest regions contribute the most in kernel density estimation.

This idea is used in a code `kDensity.m` by H. Adams [43]; implementing the code we find that $k = 75$ yields least distorted homology plots; using this value, we generate a core subset of size 2000. As shown below, the point-clouds sampled from kernel density estimate based on the concept of inverse distance contain much fewer outliers than initially in Figures 6.2 and 6.3.

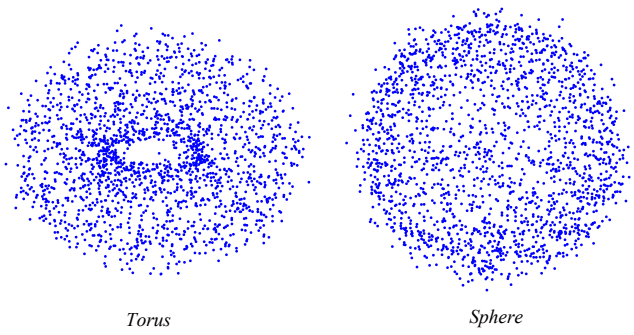

Figure 6.4: *Point-clouds with 7.5% of noise after smoothing for the outliers.*

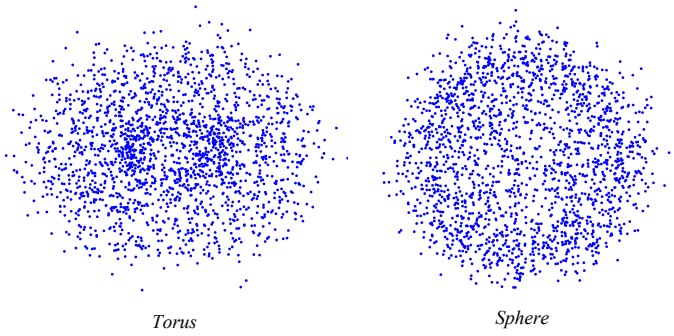

Figure 6.5: *Point-clouds with 15% of noise after smoothing for the outliers.*

The second step in processing noisy data involves smoothing based on the Epanechnikov kernel for bandwidths displayed in Table 6.2. That way point-

clouds of size 2000 are generated and 100 landmarks selected from each using the maxmin procedure in order to construct Lazy Witness streams. After computing persistent homologies of streams for the three Betti dimensions, we produce homology plots, calculate Wasserstein distance matrices, and also obtain persistence landscapes that allow us to carry out permutation tests.

Before presenting the results, let us point out that beside kd-trees there are other ways for efficient multivariate kernel density estimation. For example, in [48] a Bayesian approach is applied for bandwidth selection. Another method, mentioned in [47], involves conversion of the density estimation to a regression problem by subdividing the domain into smaller regions of equal sizes.

6.D Comparing the Results

Let us compare how properties of the torus and the sphere change as the level of Gaussian noise increases. Our first visual tool, the dendrogram in Figure 6.6 obtained from cluster analysis using the Wasserstein distance, shows that in Betti dimension 0, at noise level of 7.5%, the torus and the sphere still seem indistinguishable, as it was in Figure 3.1; for noise of 15% some clustering starts to appear indicating a difference between the torus and the sphere group.

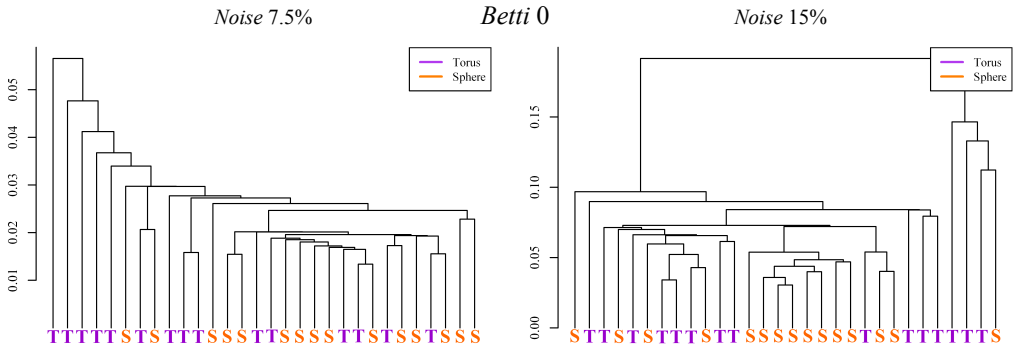

Figure 6.6: *Comparison of dendrograms for two noise levels for Betti dimension 0.*

Considering the results in Betti dimension 1, a perfect difference among the two groups still exists at noise of 7.5% just as before in Figure 3.2; however, when the amount of noise doubles and the torus starts losing its properties, the distinction among the two groups becomes less pronounced, though we can still observe some difference, as illustrated in Figure 6.7

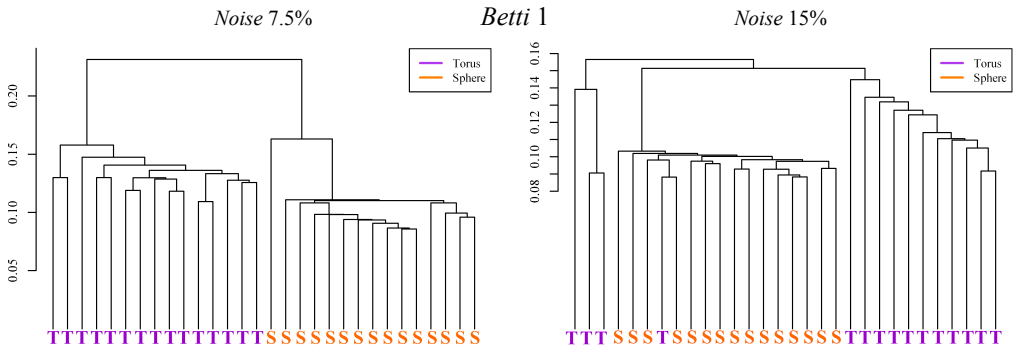

Figure 6.7: *Comparison of dendrograms for two noise levels for Betti dimension 1.*

It seems the Betti dimension 2 is most sensitive to changes as the non-noisy situation shown in Figure 3.3 does not appear anymore, not even at noise level of 7.5%, though some difference between the two groups still exists. At noise of 15% the difference gets lesser as shown in Figure 6.8.

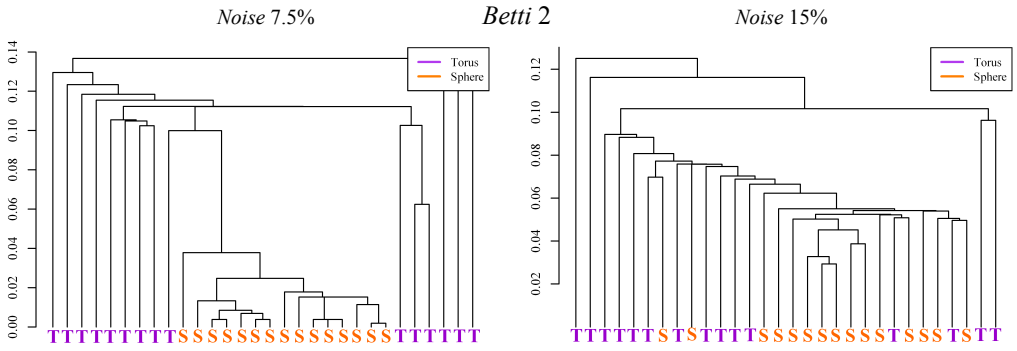

Figure 6.8: *Comparison of dendrograms for two noise levels for Betti dimension 2.*

Thus as the noise increases, differences that initially existed in first and second Betti dimensions become less pronounced.

Figures 6.9 and 6.10 display the results of multidimensional scaling across the three Betti dimensions. These results correspond to earlier dendrograms. In zeroth dimension the two groups seem similar, though less than in the non-noisy case from Figure 3.4. In Betti dimension 1, for low noise level, a difference exists like in the non-noisy case from Figure 3.5, but at higher noise this difference starts to diminish. The situation in Betti dimension 2 changes the most; in comparison to the non-noisy case from Figure 3.6, the two groups are not quite well distinguished even for low noise; also, as noise doubles the difference between the torus and the sphere practically disappears.

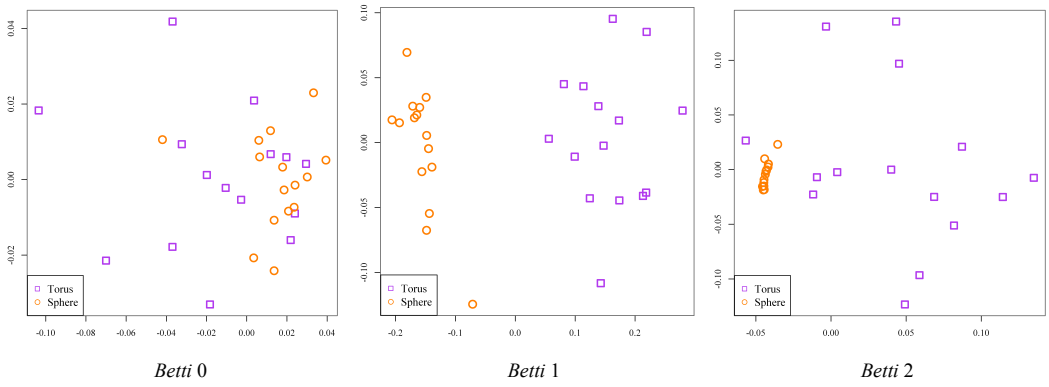

Figure 6.9: *Multidimensional scaling for noise level of 7.5%.*

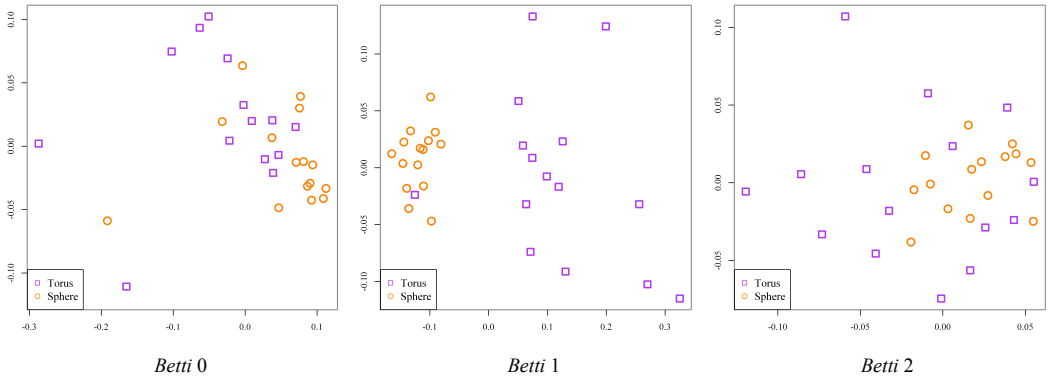

Figure 6.10: *Multidimensional scaling for noise level of 15%.*

Now we compare the homologies of noisy point-clouds. Figure 6.11 shows that at noise of 7.5% the homology plot of the torus looks different than the earlier one from Figure 2.4; in Betti dimension 1, one of the longer bars dies out sooner while the noise-related bars live longer than before. In Betti dimension 2 a long bar is not there as before. In the case of the sphere not much change happens.

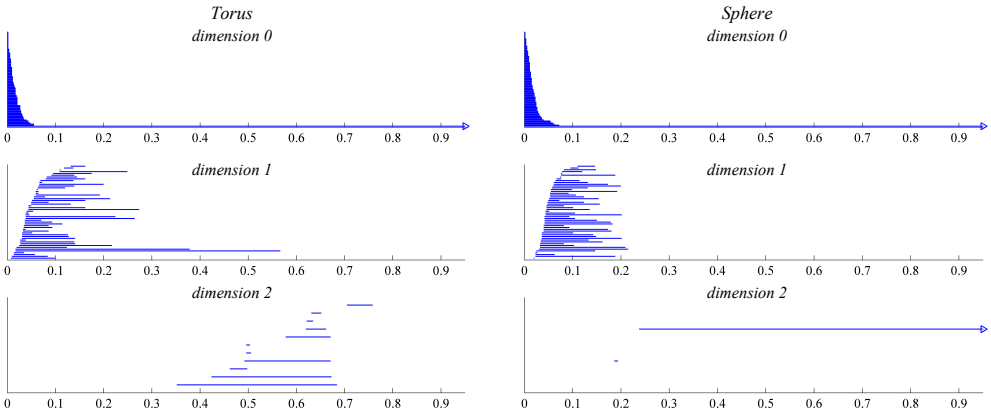

Figure 6.11: *Homology plots of a torus and a sphere for noise level of 7.5%.*

Figure 6.12 shows that for noise of 15% the homology of the torus is practically destroyed since only one bar exists in Betti dimension 1. At the same time, the sphere is much less affected.

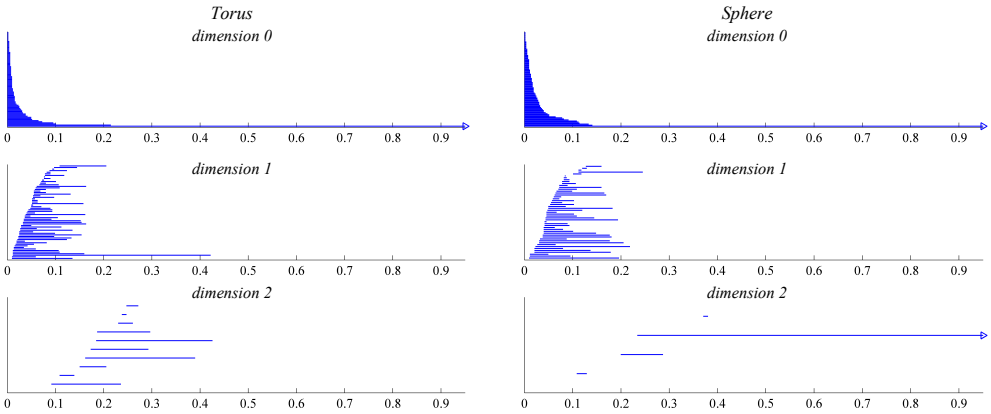

Figure 6.12: *Homology plots of a torus and a sphere for noise level of 15%.*

6.E Noisy Persistence Landscapes

Let us now consider the average persistence landscapes for the two noise levels.

Figure 6.13 shows averages for 7.5% noise.

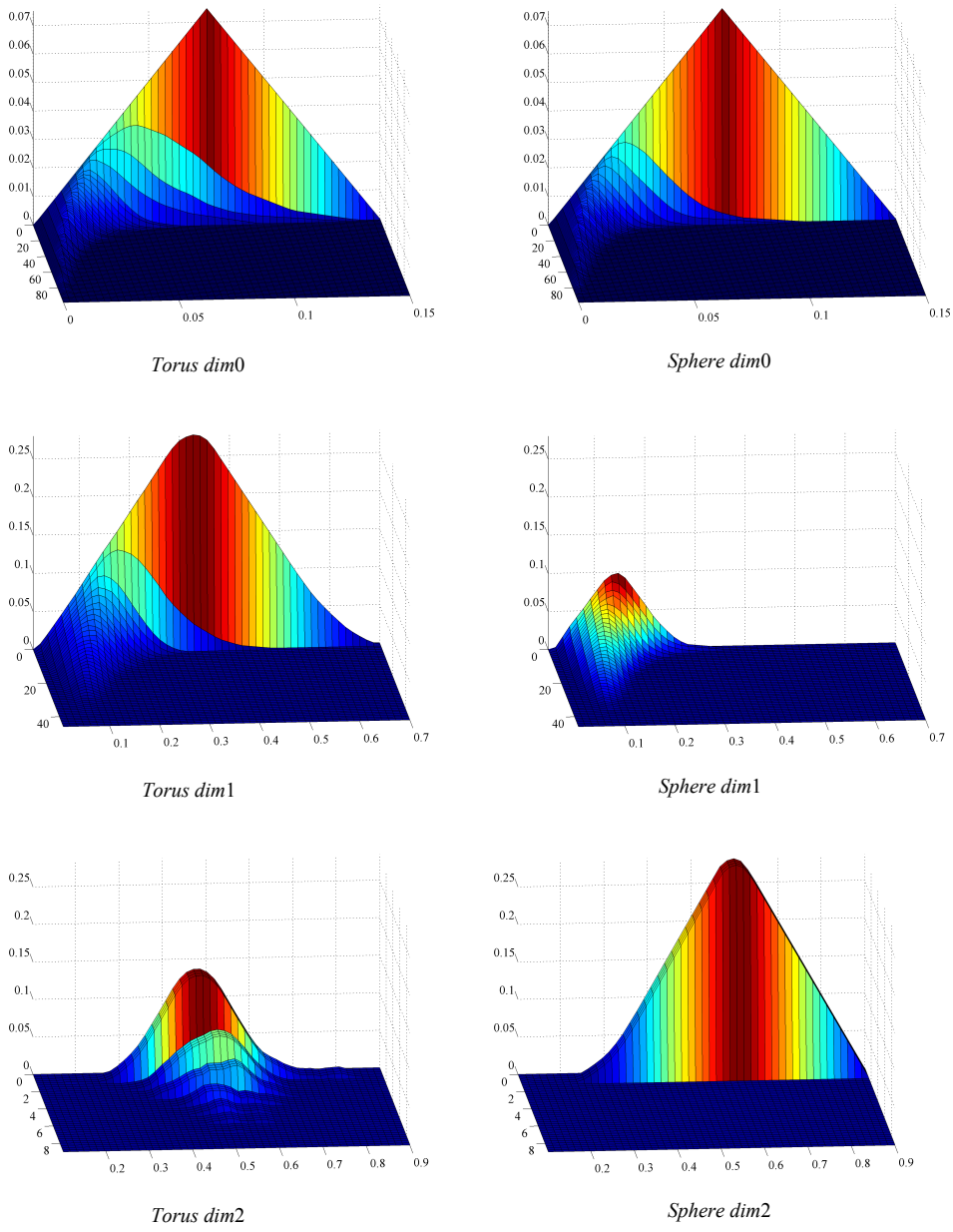

Figure 6.13: *The figure represents average persistence landscapes of the torus group (left) and the spheres group (right) across Betti dimensions for noise level of 7.5%.*

Average persistence landscapes for noise of 15% are shown in Figure 6.14

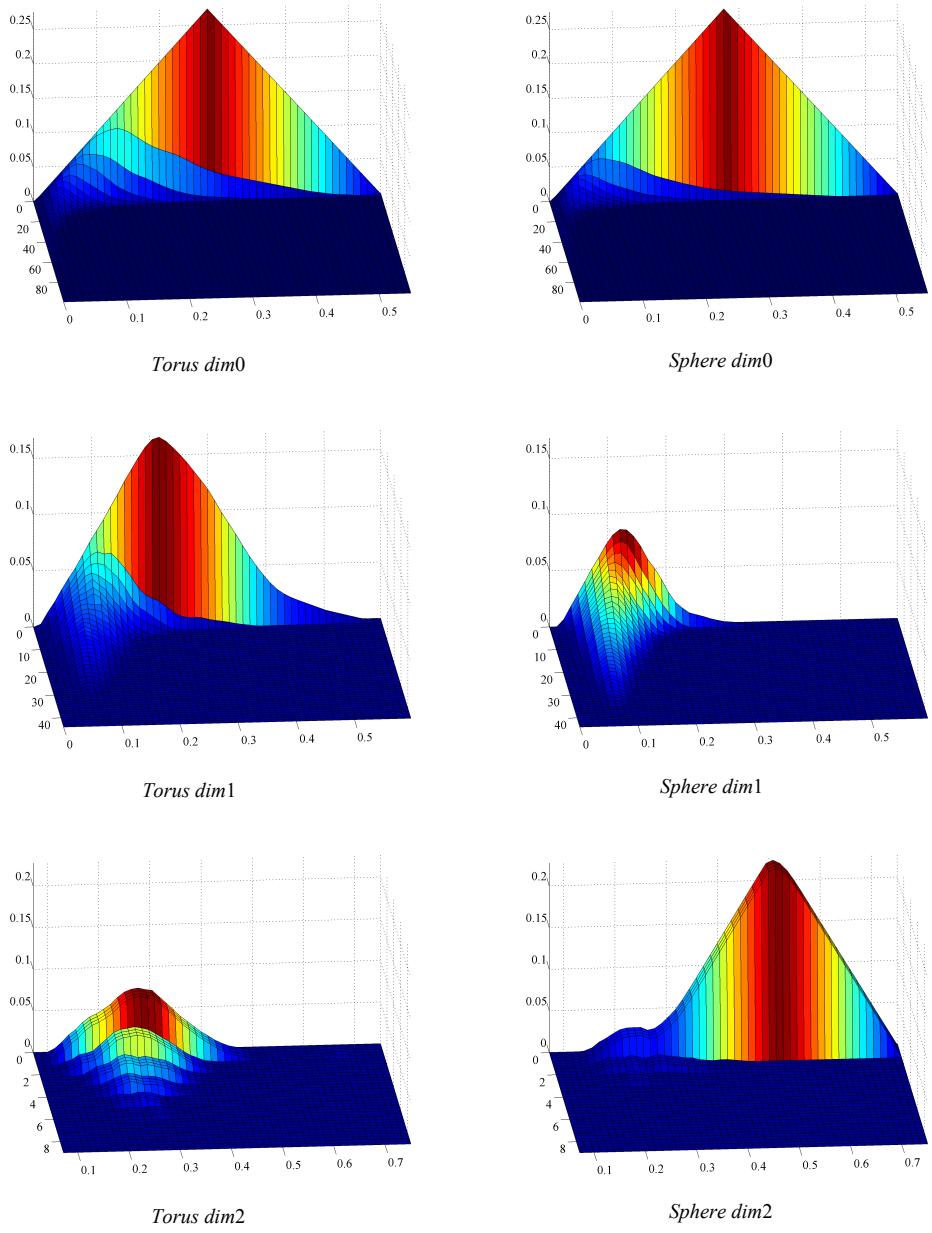

Figure 6.14: *The figure represents average persistence landscapes of the torus group (left) and the spheres group (right) across Betti dimensions for noise level of 15%.*

In comparison to the earlier situation from Figure 5.7, we can note several differences. First, in Betti dimension 0 the averages do not look as much alike as before and as noise increases, the difference between the torus and

the sphere becomes more apparent. Second, in Betti dimension 1, only a single high peak exists now in the averaged torus landscape; a much lower peak can be noticed behind the group of small peaks, but as noise doubles, this peak blends together with the small peaks. Last but not least, in Betti dimension 2 we see a drastic change - while before the main peaks were of approximately same heights, now the torus average falls to much lower heights than the sphere and as noise increases this trend continues. Considering the sphere, apart from some minor changes involving the noisy short-lived bars, not much change occurs.

Therefore, due to its specific geometry, the torus is much more sensitive to noise than the sphere. The most change that torus undergoes under the influence of noise appears in Betti dimensions 1 and 2.

6.F Tests for Noisy Data

The final step in our analysis of noisy point-clouds involves permutation tests. Again, we perform 10,000 permutations to test the null hypothesis that claims no difference between the torus and the sphere at significance level $\alpha = 0.05$. The resulting p-values for 7.5% of noise are shown in Table 6.3.

Table 6.3: p-values for permutation tests at $\alpha = 0.05$ for the difference between the torus and the sphere perturbed by sparse outliers and 7.5% of Gaussian noise.

Betti dim	**Peaks considered**	**p-value**
0	all	0.0321
1	all	0.0000
2	all	0.0000
0	peak 1	1.0000
1	all but 1 and 2	0.0007
2	peak 1	0.0000

Compared to p-values of non-perturbed data from Table 5.1, the new p-values testify that the relatively small noise of 7.5% has changed the situation.

- In Betti dimension 0, when all peaks are considered, the p-value of 0.03 indicates moderate evidence against the null hypothesis so a difference exists. If only the highest peak is considered, no difference appears.
- In Betti dimension 1, a significant difference still exists when all peaks are compared. However, when the first two peaks are ignored, the low p-value of 0.0007 indicates a difference as well. Thus, unlike before, the short-lived bars also cause a discrepancy between the two groups.
- In Betti dimension 2, both p-values of 0.0000 are compelling evidence against the null hypothesis, i.e. the torus and the sphere significantly differ, as before.

The p-values for permutation tests when the data are perturbed by sparse outliers and 15% of Gaussian noise are displayed in Table 6.4.

Table 6.4: p-values for permutation tests at $\alpha = 0.05$ for the difference between the torus and the sphere perturbed by sparse outliers and 15% of Gaussian noise.

Betti dim	**Peaks considered**	**p-value**
0	all	0.0003
1	all	0.0000
2	all	0.0000
0	peak 1	1.0000
1	all but 1 and 2	0.0002
2	peak 1	0.0000

At this level of noise, most p-values indicate a significant difference between the torus and the sphere (except for the highest peak in Betti dimension 0, which is the same for both the torus and the sphere).

To obtain more information, we compare the corresponding time-plots that allow us to determine filtrations at which a difference occurs. Figure 6.15 shows that in Betti dimension 0, when the noise is low, almost no difference appears at all times; for higher noise a difference lasts from earliest times until about $t = 0.17$ but at later times no difference exists. Thus the short-lived noisy bars cause the difference, unlike the earlier situation from Figure 5.10.

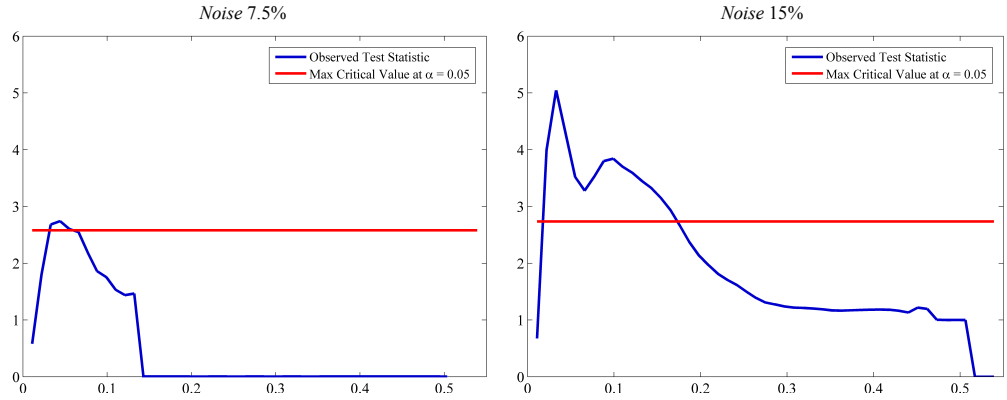

Figure 6.15: *Permutation test for the difference between the noisy torus and the noisy sphere in Betti dimension 0 at significance level* $\alpha = 0.05$.

Figure 6.16 shows that in Betti dimension 1, the low-noise situation is similar to the one from earlier Figure 5.11. At increased noise level, the difference becomes much smaller.

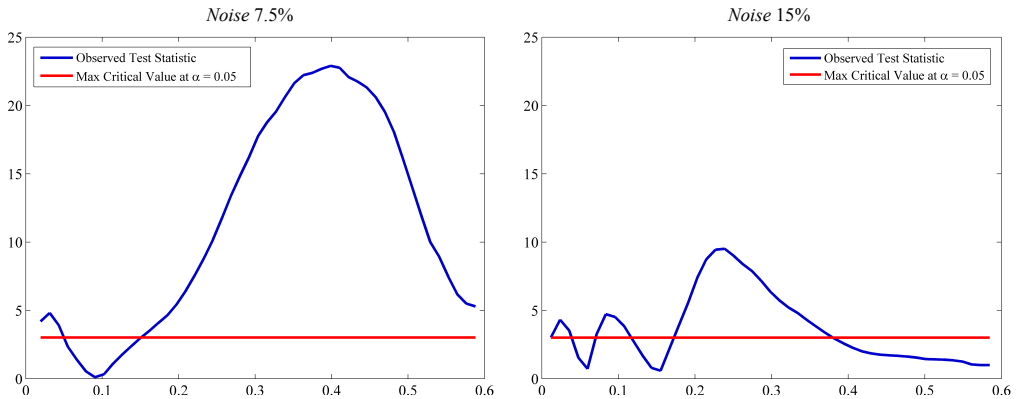

Figure 6.16: *Permutation test for the difference for noisy data in Betti dimension 1 at significance level* $\alpha = 0.05$.

Figure 6.17 shows that even if the first two peaks are ignored, a difference in Betti dimension 1 still occurs; this means the short bars cause more and more difference as noise increases, unlike before as shown in Figure 5.12.

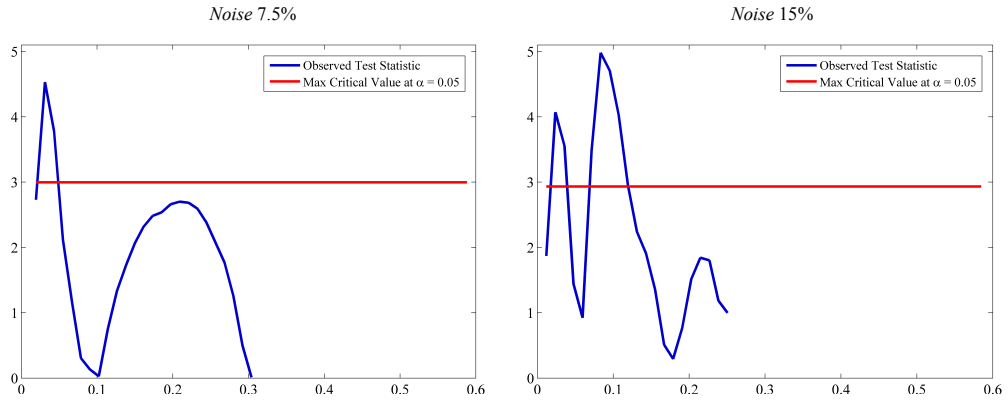

Figure 6.17: *Permutation test with* $\alpha = 0.05$ *for noisy data in Betti dimension 1 when first two peaks are ignored.*

Figure 6.18 shows that in Betti dimension 2 no difference occurs initially; a large spike at later times indicates a significant difference which increases with the level of noise. Note that in the plot for 15% noise the tip of the spike i.e. the test statistic takes a huge value of about 10^{16}.

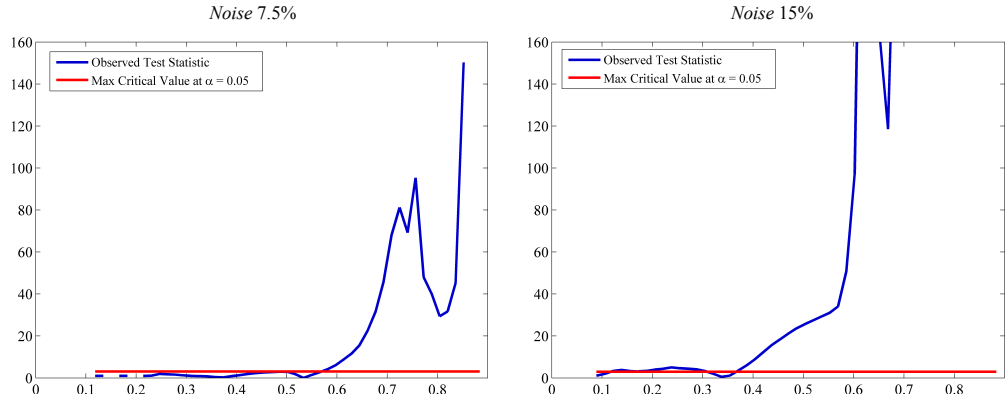

Figure 6.18: *Permutation test with* $\alpha = 0.05$ *for noisy data in Betti dimension 2.*

The above results again confirm the capability of homology to detect even the finest differences between point-cloud datasets.

Chapter 7: HIV-1 Protease Data

In this chapter we implement the aforementioned methods of topological data analysis to a real dataset of twelve samples of the HIV-1 protease, an enzyme responsible for the reproduction of the human immunodeficiency virus (HIV) of type 1. For better understanding, we first explain the role of a protease in the reproduction of HIV.

7.A Role of a Protease

As shown in literature, e.g. see [14] or [23], an essential part in the life-cycle of HIV involves a production of an immature form that consists of numerous proteins linked together into a long *polyprotein*. During an auto-catalytic process, the polyprotein releases a protease enzyme; for the two known types of HIV, the enzyme is denoted as the HIV-1 and the HIV-2 protease. Both types, as determined from X-ray crystallography, consist of two identical, mutually symmetrical protein chains, each composed of 99 amino acids. The chains act as flexible flaps enclosing a tunnel which is the "active site" i.e. the "binding pocket" of the protease; when the flaps open the protease wraps itself around the polyprotein and with a help of a water molecule cuts it into proper pieces that further assemble into a mature virus capable of infecting a new cell.

These findings gave rise to a viable approach in inhibiting the reproduction of the virus. The main idea is to apply a drug that mimics the polyprotein so the protease is prompted to tightly bond around it. The drug is a strong structure that cannot be easily cut; as long as its active site is held up by the drug, the protease cannot clip the polyprotein and the virus cannot reproduce. Unfortunately, due to viral mutations the inhibition is not permanent and after a while the protease can change its structure and become drug resilient.

7.B Our Data

Depending on the interaction between a drug and the HIV-1 protease, different complexes can arise. Our data are associated with twelve configurations of the HIV-1 protease complexed with various drugs. The Protein Data Bank [35] labels corresponding to crystal structures of these complexes are shown below, as well as an illustration of one HIV-1 structure (without showing the drug).

Table 7.1: Protein Data Bank labels for 12 complexes of HIV-1 protease

1) 1HPV	2) 1HXB	3) 1HXW	4) 1MUI
5) 3JVY	6) 1HVR	7) 2B7Z	8) 2FNS
9) 2O4K	10) 2O4P	11) 2PYN	12) 1HHP

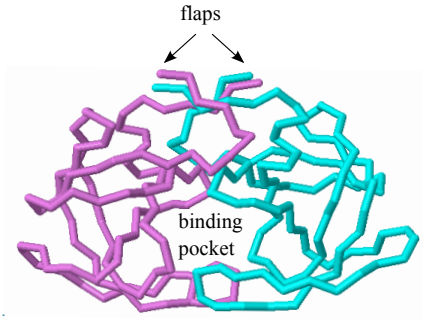

Figure 7.1: *Illustration of 3JVY structure of HIV-1 Protease Mutant.*

The image in Figure **??** is generated by downloading the appropriate 3d-coordinates from the Protein Data Bank [35] and inputting them into an online bioinformatics server at the University of Pittsburgh [32]. It is known that the 3JVY is a mutant structure, obtained in a complex with the drug *darunavir*.

Note that the data we actually work with are *not* the 3d-coordinates, but dynamical "correlations" which we obtained from Y. Mao [24] who calculated strengths of pairwise interactions between the 198 amino acids of the HIV-1 protease in various complexed environments. Thus, our analysis starts with twelve "correlation matrices" of size 198×198; in each such matrix, a correlation coefficient is a measure of proximity between two amino acids. Note that in statistics, both a distance matrix that measures the dissimilarity among data and a proximity matrix with coefficients measuring similarity can be equally used.

To implement our methods, we transform the dynamical correlations to "coordinates" using the *Isometric Mapping*, also known as *Isomap* [46]. This approach, as shown by Tenenbaum, de Silva, and Langford [45], is mostly used for efficient dimensionality reduction of non-linear datasets; an advantage is that the intrinsic geometry of data is preserved even in the case of non-Euclidean metric. Furthermore, inputting a distance or proximity matrix, the method gives back the corresponding coordinates. That way, every correlation matrix yields a set of 198 "coordinates" that (in the sense of interactions) best describe the corresponding dynamical structures of HIV-1 protease. Due to confidentiality of data that are part of an ongoing research, the Isomap transformation of the correlation matrices was performed by G. Heo.

7.C Results for Protease Data

To investigate the dynamical structures of the twelve protease samples, we start with methods of persistent homology. First, we choose an appropriate type of the stream to be constructed. Since each investigated HIV-1 protease structure consists of 198 datapoints only, a Vietoris-Rips stream which makes use of all available information from the data can be implemented. This means that every datapoint will represent a vertex in the constructed complex. Recall that in the case of earlier point clouds of size 2000 a Vietoris-Rips stream was not a feasible option, due to computational difficulties that arise when the number of simplices becomes too large. Snapshots from the evolution of the Vietoris-Rips stream built on sample 5 are shown in Figure 7.2 and the corresponding number of k-simplices is presented in Table 7.2 below.

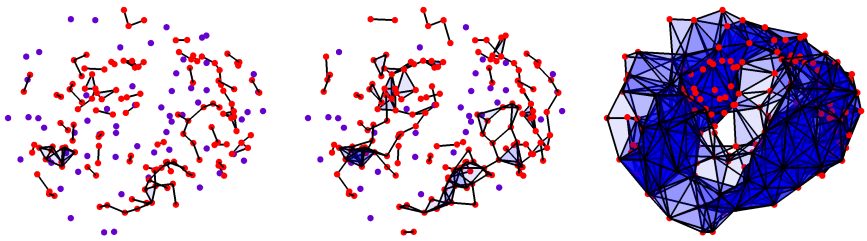

Figure 7.2: *Snapshots from the evolution of a Vietoris-Rips stream on sample 5 (3JVY structure) of HIV-1 protease at filtration times 0.4, 0.5, and 1.0.*

Table 7.2:
The number of k-simplices for $k = 0, 1, 2, 3$ at times 0.4, 0.5, and 1.0.

t	n_0	n_1	n_2	n_3
0.4	198	118	24	4
0.5	198	223	111	37
1.0	198	1413	4393	8470

In the table of k-simplices, observe the large number of tetrahedra at $t = 1.0$; at the same time, there are about half less triangles. Since each tetrahedron has four triangular faces, it follows that in most cases triangles are shared between tetrahedra, i.e. we have regions with densely packed tetrahedra.

Figure 7.3 shows homology plots of sample 5 across Betti dimensions 0, 1, and 2. In dimension 0 the initial 198 vertices mutually connect about time $t = 0.9$ at which point the stream becomes a single connected component. In dimension 1 relatively long bar appears, while all the bars in dimension 2 seem to be noise.

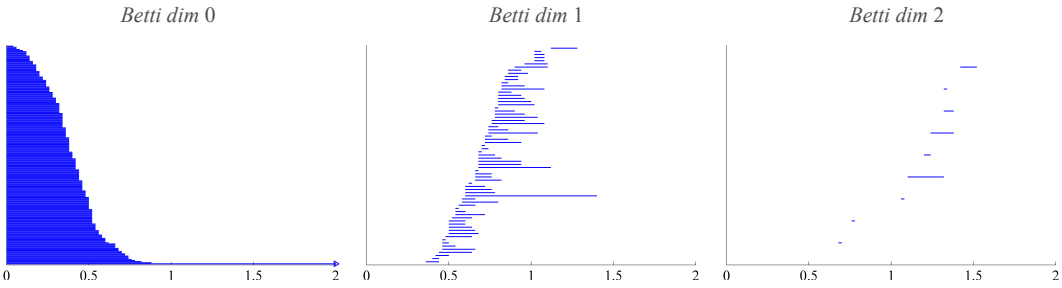

Figure 7.3: *Homology plots for sample 5 in HIV-1 protease data.*

More information follows from the persistence landscapes method. Figure 7.4 depicts mean persistence landscape across the three Betti dimensions.

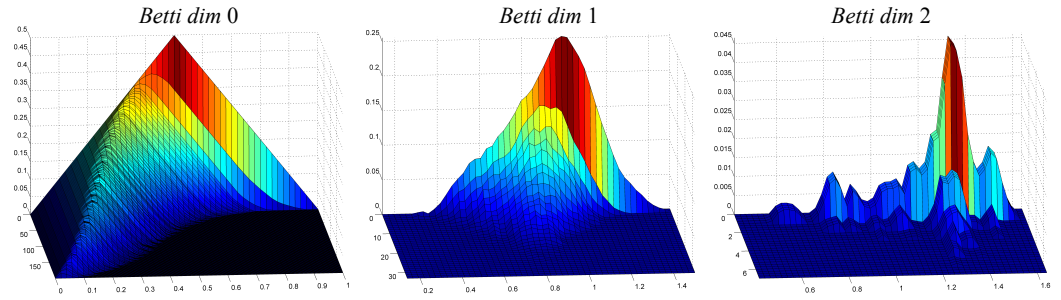

Figure 7.4: *Mean persistence landscapes for protease data.*

7.D Resistent and Non-Resistent Group

Let us point out that three out of the twelve given HIV-1 protease structures have been experimentally confirmed as drug resistant mutants. The drug-resilient structures correspond to our samples 5, 7, and 11. Using this prior information as a motivation for further analysis, we perform a permutation test. First, we define the test hypotheses.

Permutation Test Hypotheses

Due to two groups of data that we compare, we describe two populations:

η_{DR}: Population of average persistence landscape curves for the drug-resistent group of HIV-1 protease, described by the mean $\mu_{DR}^{(\eta)}$.

η_{NR}: Population of average persistence landscape curves for the non-resistent group of HIV-1 protease, described by the mean $\mu_{NR}^{(\eta)}$.

Then, for the test that compares the drug-resistent and the non-resistent group, the assumptions under the null and alternative hypothesis are:

$$\begin{aligned} H_0 &: \mu_{DR}^{(\eta)} = \mu_{NR}^{(\eta)} \\ H_1 &: \mu_{DR}^{(\eta)} \neq \mu_{NR}^{(\eta)} \end{aligned} \tag{7.1}$$

We perform the test at $\alpha = 0.05$ level of significance, so the outcome of the test is determined by comparing the p-value with the given significance level:

$$\begin{aligned} p\text{-value} \leq 0.05 \;\; &\Rightarrow \;\; \text{reject } H_0 \\ p\text{-value} > 0.05 \;\; &\Rightarrow \;\; \text{not reject } H_0 \end{aligned} \tag{7.2}$$

Results for the Permutation Test

Using data from persistence landscapes we perform a permutation test with 10,000 repetitions at significance level $\alpha = 0.05$. As stated in the previous section, our assumption under the null hypothesis is that samples 5, 7, and 11 of HIV-1 protease do not differ from the rest of the group.

Performing the test for Betti dimension 0, we obtain the p-value of 0.0430 which indicates moderate evidence against the null hypothesis. Though not very compelling, we are allowed to reject the null hypothesis and conclude that protease structures $3JVY$, $2B7Z$, $2PYN$ differ from the other nine. Plotting the test statistic we find that a small but significant difference exists during filtration times in $(0.18, 0.26)$.

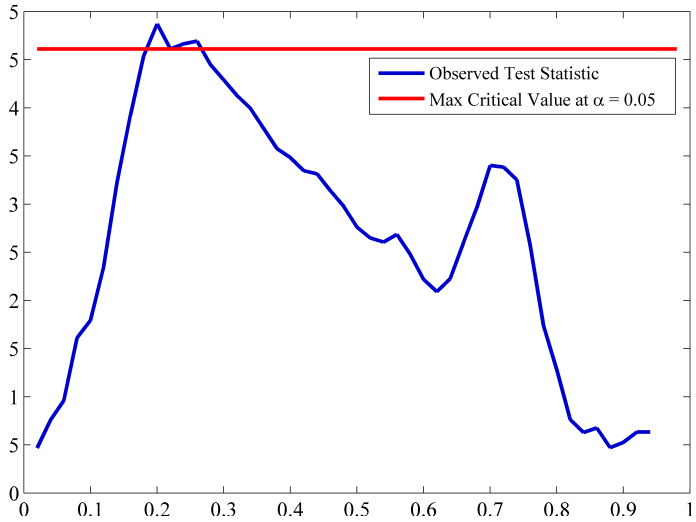

Figure 7.5: *Permutation test for the HIV-1 protease data.*

Based on Betti dimension 0 alone, the statistical analysis with sample size 12 yields a moderate evidence for difference between two groups of HIV-1 protease. These results correspond to the fact that the three protease structures are mutants which acquired drug resistance.

Chapter 8: Conclusion

In this thesis we have shown applications of persistent homology in statistical data analysis. Using homology plots and the Wasserstein distance, we obtained accurate description among our samples of data. We have also implemented a new method; our results show that apart from standard persistent homology descriptors, such as persistence diagrams and barcodes, persistence landscapes offer a qualitative topological analysis of patterns hidden in complex, high dimensional datasets. We have illustrated the strengths of this method via point cloud analysis of spheres and tori under various noise levels, as well as on an HIV-1 protease dataset.

Considering future research, there is a plenty of work to do; we have no doubts that as we will progress in our data analysis, more researches will adopt the topological approach in statistics.

Appendices

Appendix A: Landscapes for the 15 tori and 15 spheres

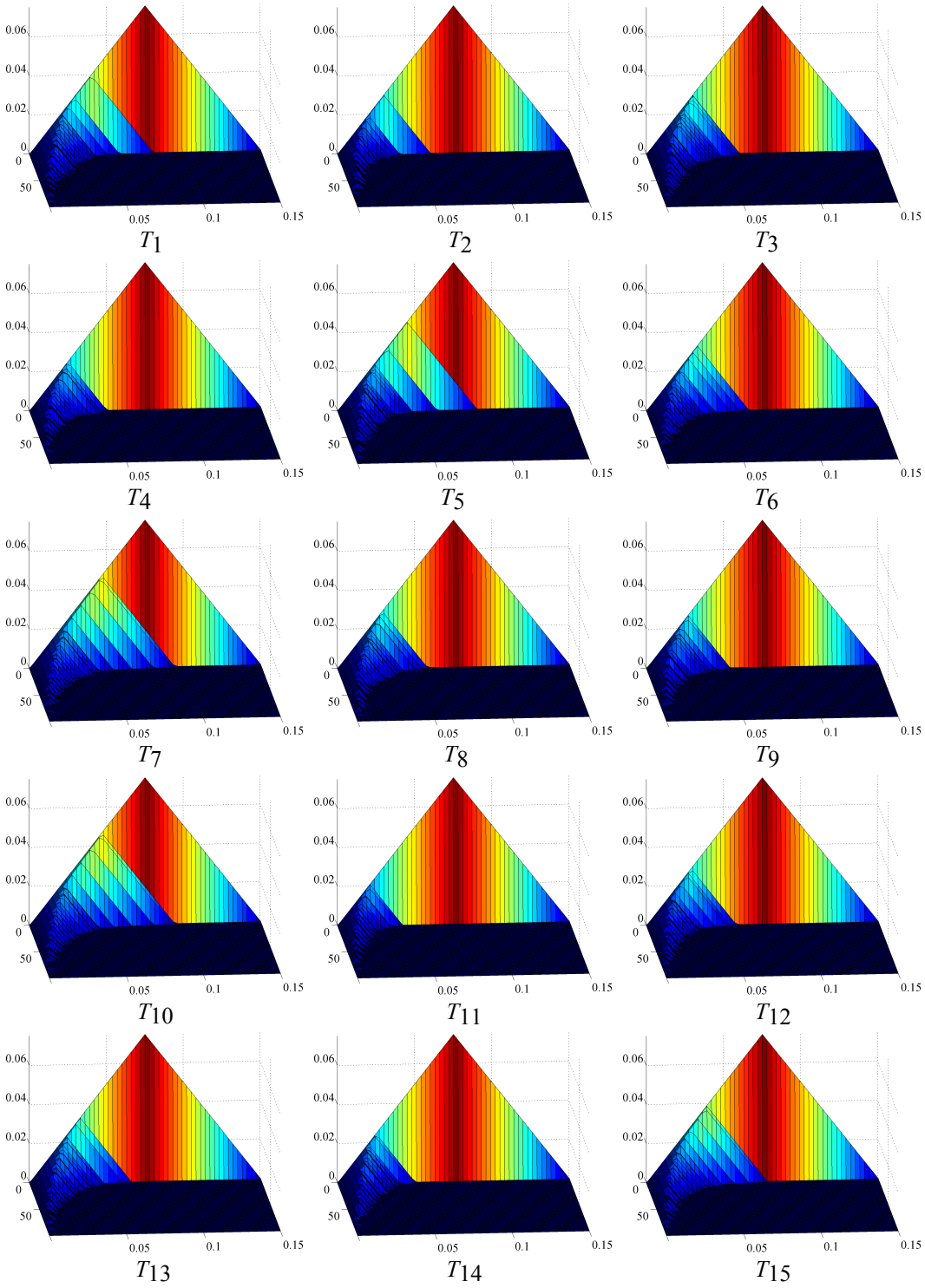

Figure A-1: *Images of Persistence Landscape for the 15 tori in Betti dimension 0.*

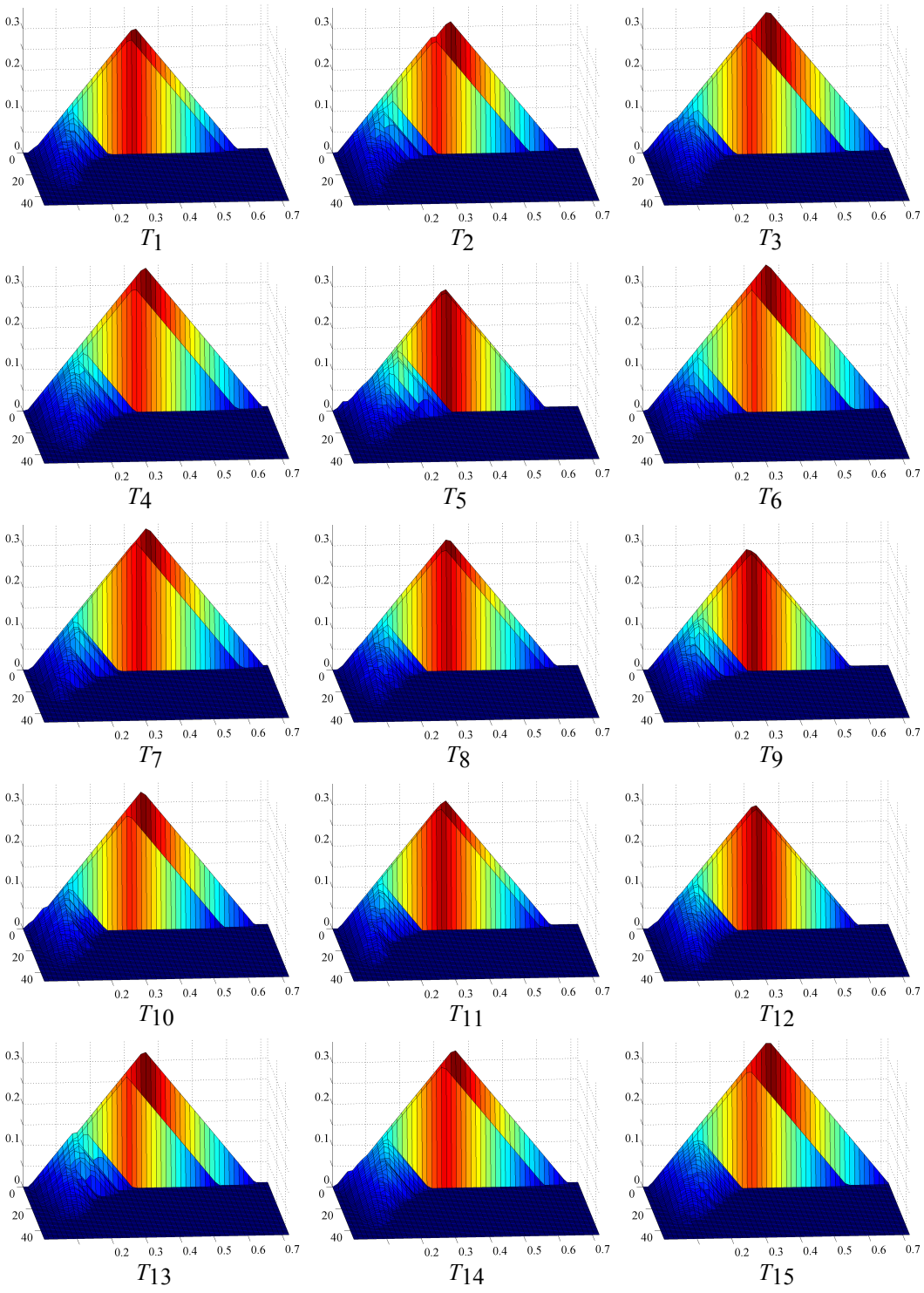

Figure A-2: *Images of Persistence Landscape for the 15 tori in Betti dimension 1.*

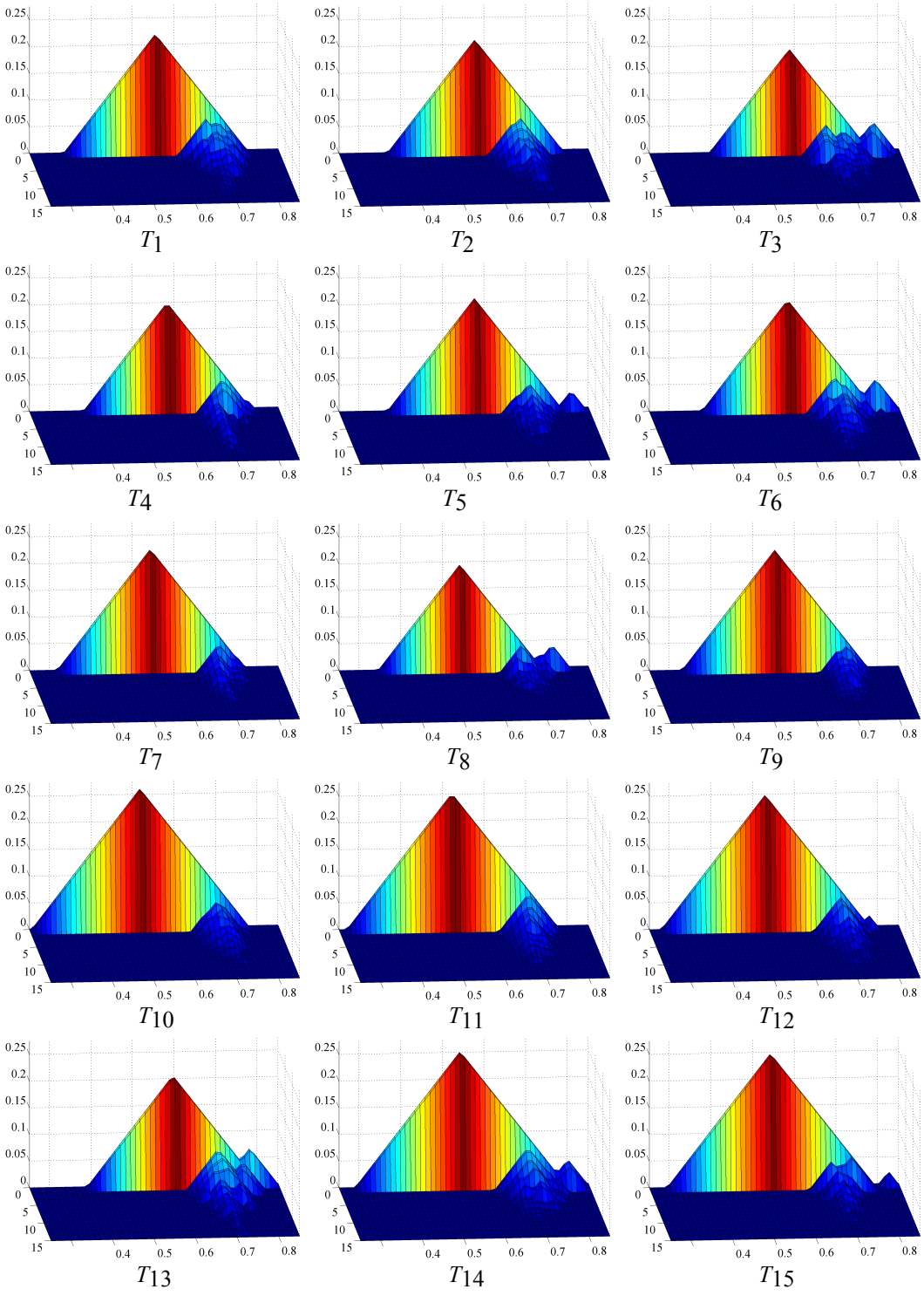

Figure A-3: *Images of Persistence Landscape for the 15 tori in Betti dimension 2.*

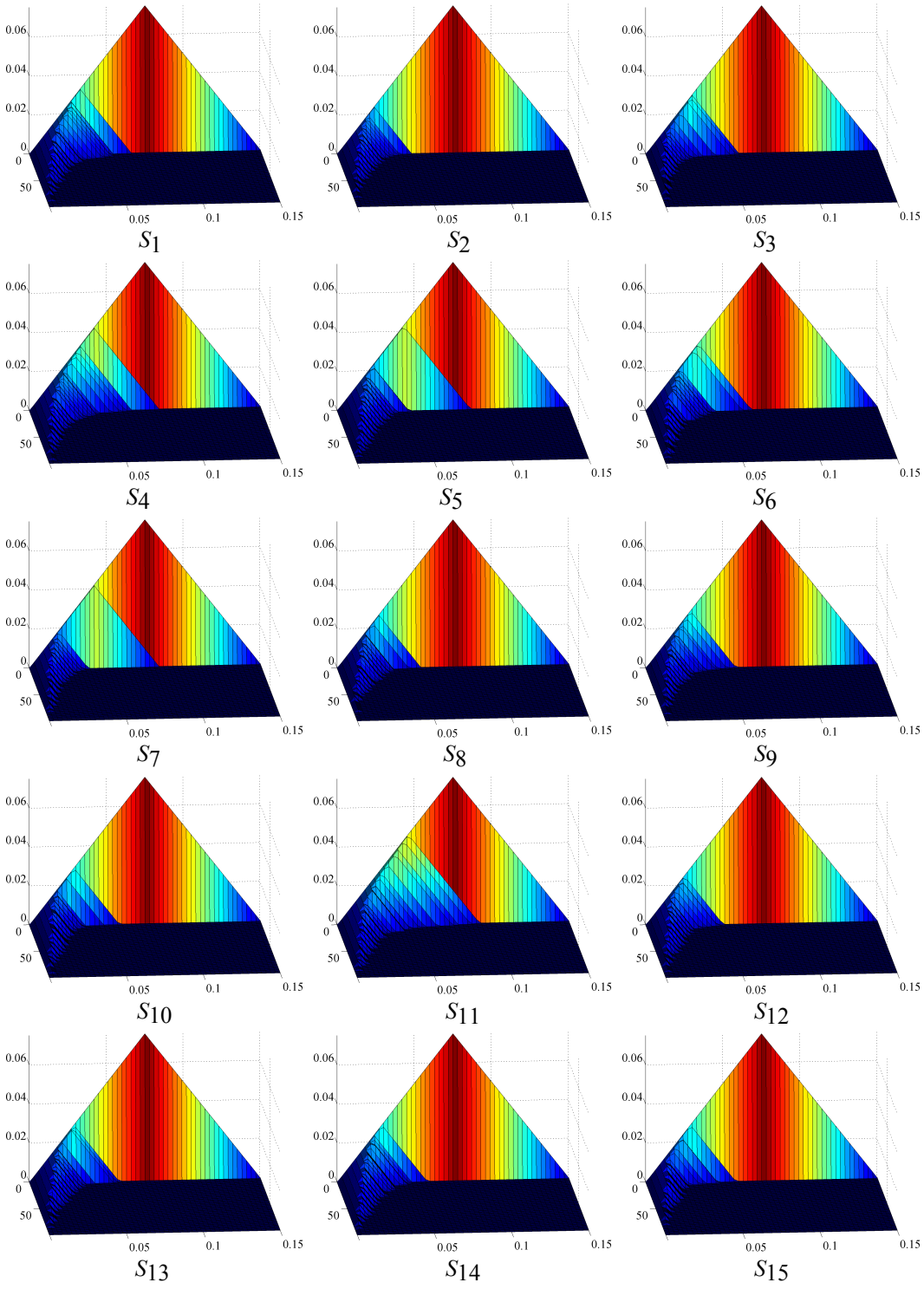

Figure A-4: *Images of Persistence Landscape for the 15 spheres in Betti dimension 0.*

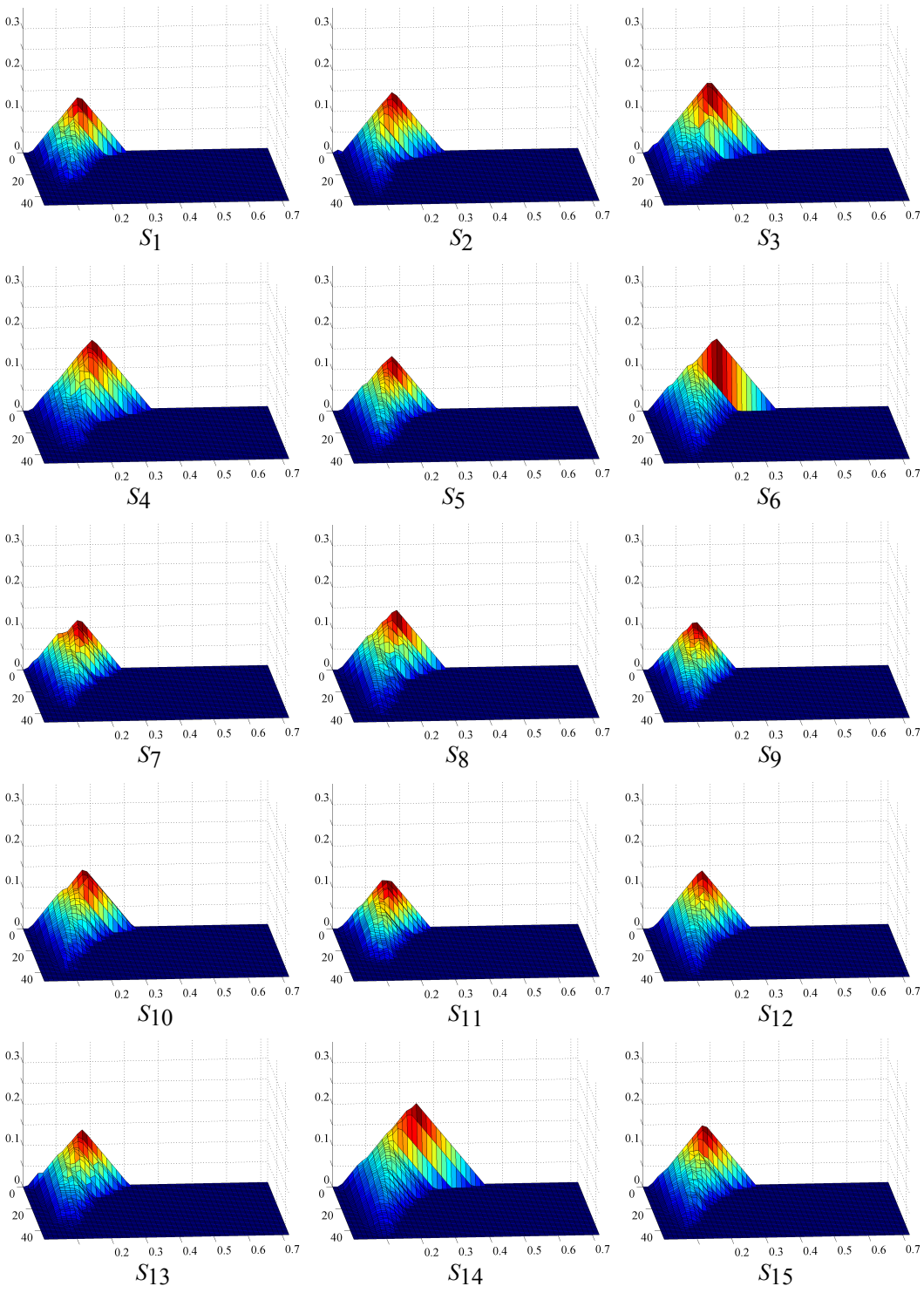

Figure A-5: *Images of Persistence Landscape for the 15 spheres in Betti dimension 1.*

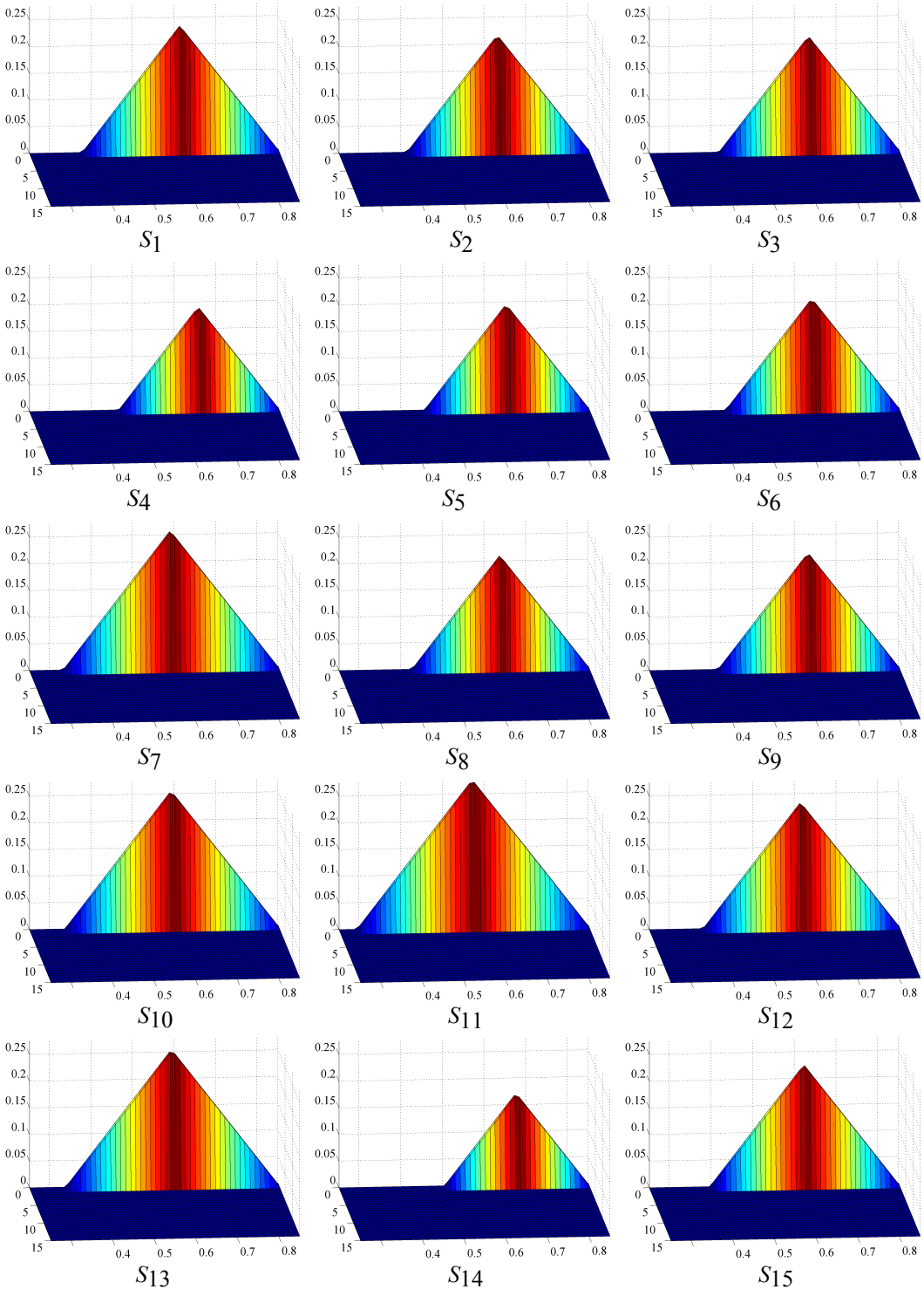

Figure A-6: *Images of Persistence Landscape for the 15 spheres in Betti dimension 2.*

Appendix B: Variation in the mean persistence landscape

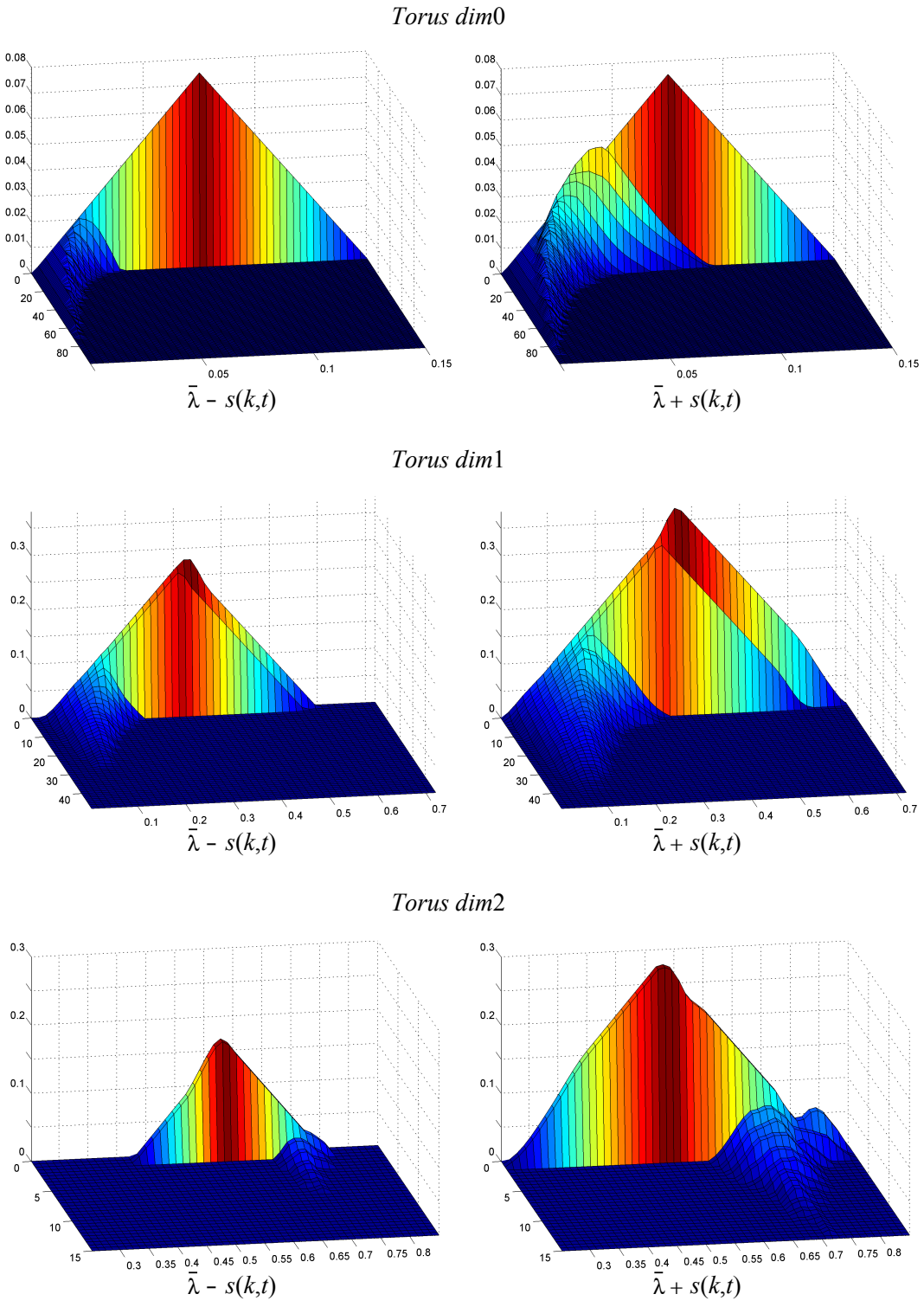

Figure B-1: *Change in the average persistence landscape within a range of one standard deviation across three Betti dimensions 0, 1, and 2 for the torus.*

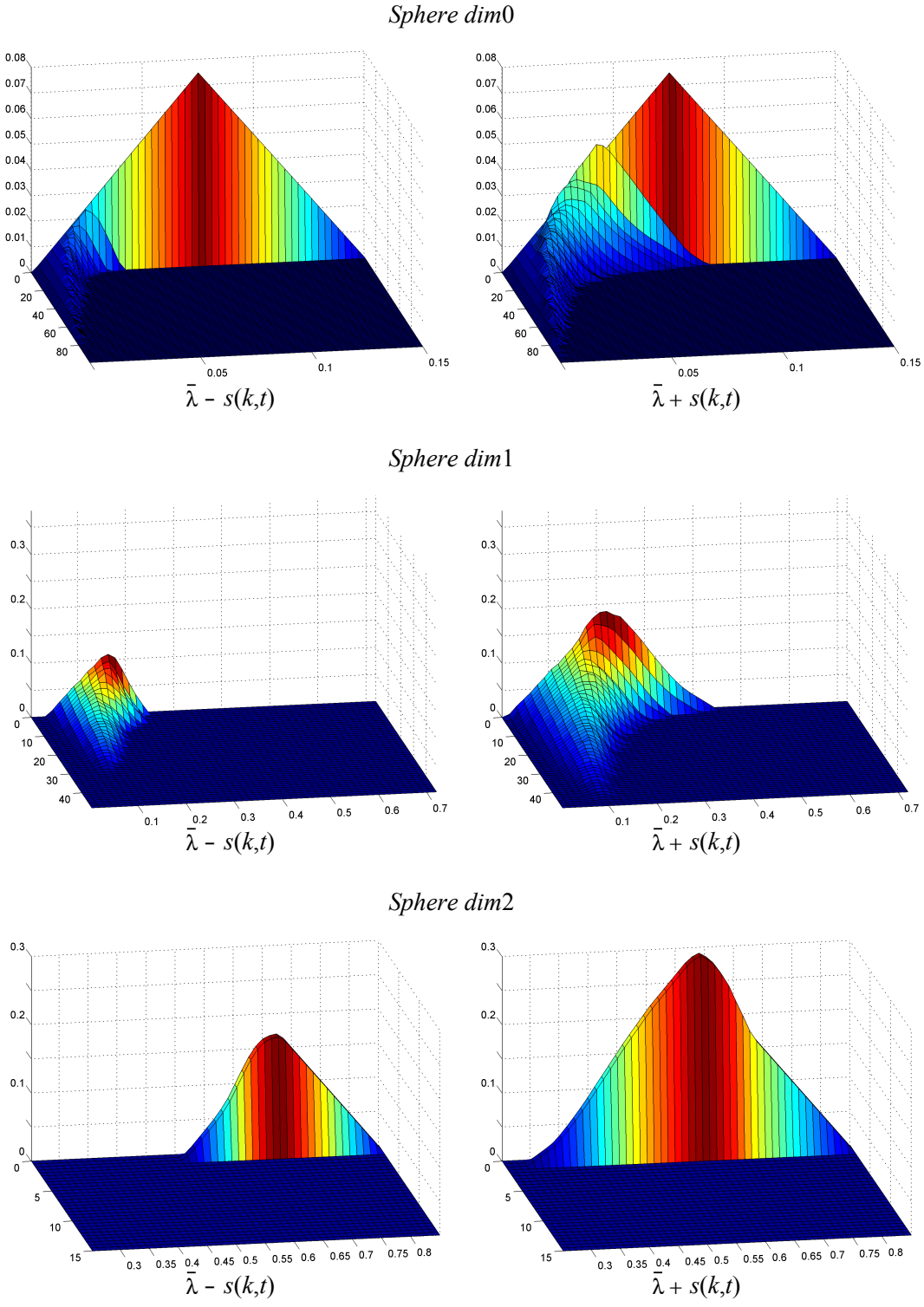

Figure B-2: *Change in the average persistence landscape within a range of one standard deviation across three Betti dimensions 0, 1, and 2 for the sphere.*

Bibliography

[1] H. Adams & G. Carlsson, *On the Nonlinear Statistics of Range Image Patches*, SIAM Journal on Imaging Sciences, **2**:110, 2009, http://dx.doi.org/10.1137/070711669. 76

[2] G. L. Alexanderson, *Euler and Königsbergs bridges: a historical view*, Bull. Amer. Math. Soc. (N.S.), **43**:567, 2006, http://dx.doi.org/10.1090/S0273-0979-06-01130-X. 1

[3] P. Bubenik, *Peristent homology and statistical inference: Persistence Landscapes*, http://academic.csuohio.edu/bubenik_p/talks/jmm2012talk.pdf, 2012. 36

[4] P. Bubenik, *Towards statistical topology: homology, persistent homology and persistence landscapes*, http://academic.csuohio.edu/bubenik_p/talks/mbi_talk.pdf, 2012. 36

[5] P. Bubenik, *Statistical Topology Using Persistence Landscapes*, SAO/-NASA Astrophysics Data System, 2012, http://arxiv.org/abs/1207.6437v1. 38, 39, 40, 42, 44, 47, 63, 65

[6] G. Carlsson & V. de Silva, *Topological estimation using witness complexes*, in *Eurographics Symposium on Point-Based Graphics*, edited by A. M. & R. S., pp. 157–166, ETH, Zürich, Switzerland, 2004, Department of Mathematics, Stanford University, California, USA, http://pages.pomona.edu/~vds04747/public/papers/witness.pdf. 24, 76

[7] G. Carlsson, T. Ishkhanov, V. de Silva, & A. Zomorodian, *On the Local Behavior of Spaces of Natural Images*, International Journal of Computer Vision, **76**:1, 2008, http://dx.doi.org/10.1007/s11263-007-0056-x. 76

[8] G. Carlsson, *Topology and Data*, Bull. Amer. Math. Soc. (N.S.), **46**:255, 2009, http://dx.doi.org/10.1090/S0273-0979-09-01249-X. 2

[9] H. Edelsbrunner & J. Harer, *Computational Topology: An Introduction*, Applied Mathematics, American Mathematical Society, Providence, RI, 2010, http://www.ams.org/bookpages/mbk-69. 5, 9, 12, 18, 30

[10] H. Edelsbrunner, D. Letscher, & A. Zomorodian, *Topological Persistence and Simplification*, Discrete & Computational Geometry, **28**:511, 2002, http://dx.doi.org/10.1007/s00454-002-2885-2. 18

[11] M. D. Ernst, *Permutation Methods: A Basis for Exact Inference*, Statistical Science, **19**:676, 2004, http://www.jstor.org/stable/4144438. 60

[12] *Lecture notes in Math 600*, Course by Terry Gannon, 2010-11. 14, 18

[13] R. Ghrist, *Barcodes: The persistent topology of data*, Bull. Amer. Math. Soc., **45**:61, 2008, http://www.ams.org/bull/2008-45-01/S0273-0979-07-01191-3. 4

[14] D. Goodsell, *HIV-1 Protease*, Protein Data Bank - Molecule of the Month, 2000, http://www.rcsb.org/pdb/101/motm.do?momID=6. 88

[15] D. Graham, *Mardiatest.m*, Loughborough University, 2006, http://www.lboro.ac.uk/research/phys-geog/confidence_regions/mardiatest.m. 72

[16] A. Hatcher, *Algebraic Topology*, Cambridge University Press, 2011. 14

[17] G. Heo, *Toplogical analysis of variance with applications in landmark data set*, in *5th ATMCS conference in Applied and Computational Topology*, ICMS, Edinburgh, UK, 2012, http://www.icms.org.uk/downloads/ATMCS5/Heo.pdf. 52

[18] S. M. Holland, *Non-Metric Multidimensional Scaling (MDS)*, CRAN R-Project, 2008, http://strata.uga.edu/software/pdf/mdsTutorial.pdf. 33

[19] A. Izenman, *Multidimensional Scaling and Distance Geometry*, in *Modern Multivariate Statistical Techniques*, edited by G. Casella, S. Fienberg, & I. Olkin, Springer Texts in Statistics, pp. 463–504, Springer New York, 2008, http://dx.doi.org/10.1007/978-0-387-78189-1_13. 33

[20] I. M. James, *Preface*, in *History of Topology*, pp. v–vi, North-Holland, Amsterdam, first edition, 1999, http://dx.doi.org/10.1016/B978-044482375-5/50000-6. 1

[21] Laser Design Inc., *3D Laser Scanning - Hard Work that Looks like Magic*, http://laserdesign.com/learn_more.aspx. 4

[22] P. Legendre & L. Legendre, *Ch.1 - Complex ecological data sets*, in *Numerical Ecology*, volume 24 of *Developments in Environmental Modelling*, pp. 1–57, Elsevier, 2012, http://dx.doi.org/10.1016/B978-0-444-53868-0.50001-0. 60

[23] P.-T. T. Ltd., *HIV-1 Protease*, online, 2012, http://www.prospecbio.com/HIV-1_Protease_9_63. 88

[24] Y. Mao, *Dynamical Basis for Drug Resistance of HIV-1 Protease*, BMC Structural Biology, **11**:31, 2011, http://www.biomedcentral.com/1472-6807/11/31. 90

[25] K. V. Mardia, *Measures of multivariate skewness and kurtosis with applications*, Biometrika, **57**:519, 1970, http://dx.doi.org/10.1093/biomet/57.3.519. 72

[26] K. V. Mardia, *Tests of unvariate and multivariate normality*, in *Analysis of Variance*, edited by P. R. Krishnaiah, volume 1 of *Handbook of Statistics*, pp. 279–320, Elsevier, 1980, http://dx.doi.org/10.1016/S0169-7161(80)01011-5. 72

[27] Y. Mileyko, S. Mukherjee, & J. Harer, *Probability measures on the space of persistence diagrams*, Inverse Problems, **27**:124007, 2011, http://stacks.iop.org/0266-5611/27/i=12/a=124007. 30, 38

[28] D. S. Moore, G. P. McCabe, W. M. Duckworth, & S. L. Sclove, *Ch.14 - Bootstrap Methods and Permutation Tests*, in *The Practice of Business Statistics: Using Data for Decisons*, Freeman, W. H., first edition, 2003, http://bcs.whfreeman.com/pbs. 59

[29] J. Munkres, *Elements Of Algebraic Topology*, The Benjamin/Cummings Publishing Company, first edition, 1986. 9, 10

[30] J. J. O'Connor & E. F. Robertson, *History of Topology*, MacTutor History of Mathematics archive, 1996, http://www-gap.dcs.st-and.ac.uk/~history/HistTopics/Topology_in_mathematics.html. 1

[31] J. J. O'Connor & E. F. Robertson, *Biography of Henri Poincaré*, MacTutor History of Mathematics archive, 2003, http://www-history.mcs.st-and.ac.uk/Biographies/Poincare.html. 2

[32] U. of Pittsburgh, *Anisotropic Network Model Web Server*, online, 2000, http://www.hsls.pitt.edu/obrc/index.php?page=URL1192629865. 90

[33] F. Pesarin & L. Salmaso, *Permutation Tests for Complex Data: Theory, Applications and Software*, Wiley Series in Probability and Statistics, John Wiley & Sons, 2010. 55

[34] J. O. Ramsay, G. Hooker, & S. Graves, *Functional Data Analysis with R and MATLAB*, Use R!, Springer, 2009. 55, 57, 61

[35] Research Collaboratory for Structural Bioinformatics (RCSB), *Protein Data Bank*, 2012, http://www.rcsb.org/pdb/home/home.do. 89, 90

[36] R. H. Riffenburgh, *Ch.26 - Methods You Might Meet, but Not Every Day*, in *Statistics in Medicine*, Elsevier Academic Press, Burlington, second edition, 2006, http://www.sciencedirect.com/science/article/pii/B9780120887705500368. 60

[37] M. Rosenblatt, *Remarks on some nonparametric estimates of a density function*, Annals of Mathematical Statistics, **27**:832, 1956. 74

[38] D. W. Scott & J. R. Thompson, *Probability density estimation in higher dimensions*, in *Proceedings of the Fifteenth Symposium on the Interface*, edited by J. E. Gentle, Computer Science and Statistics, pp. 173–179, Amsterdam, 1983, North-Holland, Rice University, Houston, Texas. 71

[39] G. Shorack, *Measures*, in *Probability for Statisticians*, edited by G. Casella, S. Fienberg, & I. Olkin, Springer Texts in Statistics, pp. 1–20, Springer London, 2000, http://dx.doi.org/10.1007/0-387-22760-1_1. 37

[40] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, CRC Monographs on Statistics and Applied Probability, Chapman & Hall, CRC Press Taylor & Francis Group, Boca Raton, FL, 1998, http://www.crcpress.com/product/isbn/9780412246203. 71, 74, 75

[41] W. A. Sutherland, *Introduction to Metric and Topological Spaces*, Oxford Mathematics, Oxford University Press, 2009, http://www.oup.com/uk/companion/metric. 37

[42] A. Tagliasacchi, *Matlab KD-Tree library*, 2010, http://www.mathworks.com/matlabcentral/fileexchange/authors/30596. 76

[43] A. Tausz & H. Adams, *JavaPlex Tutorial*, Open source software package for computing persistent homology, 2012, http://javaplex.googlecode.com/svn/trunk/reports/javaplex_tutorial/javaplex_tutorial.pdf. 7, 8, 23, 24, 77

[44] A. Tausz, M. Vejdemo-Johansson, & H. Adams, *JavaPlex: A research software package for persistent (co)homology*, 2011, http://code.google.com/javaplex. 8, 23

[45] J. B. Tenenbaum, V. d. Silva, & J. C. Langford, *A Global Geometric Framework for Nonlinear Dimensionality Reduction*, Science, **290**:2319, 2000, http://www.sciencemag.org/content/290/5500/2319.abstract. 90

[46] J. Tenenbaum, *Isomap*, 2001, http://isomap.stanford.edu. 90

[47] L. Wasserman, *All of Nonparametric Statistics*, Springer Texts in Statistics, Springer New York, 2006, http://dx.doi.org/10.1007/0-387-30623-4_6. 78

[48] X. Zhang, M. L. King, & R. J. Hyndman, *A Bayesian approach to bandwidth selection for multivariate kernel density estimation*, Computational Statistics and Data Analysis, **50**:3009, 2006, http://dx.doi.org/10.1016/j.csda.2005.06.019. 71, 78

[49] A. J. Zomorodian, *Topology for Computing*, Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, 2009, http://www.cs.dartmouth.edu/~afra/book.html. 9, 25