University of Innsbruck

Master's Thesis

An Application of Topological Data Analysis to Fibrin Networks

Author
Martin Berger

Supervisors
Tobias Hell, PhD
Univ.-Prof. Dr. Tim Netzer

Faculty of Mathematics, Computer Science and Physics
University of Innsbruck

December 2017

"Wherever there is number, there is beauty."

—Proclus

Acknowledgements

First I express my gratitude to my supervisors Tobias Hell, PhD and Univ.-Prof. Dr. Tim Netzer. Both of them always had an open door for me whenever I ran into problems with my thesis or had a question. Furthermore I thank them for caring so much about my future career and PhD studies.

I also thank the university hospital for anaesthesia and intensive care for providing the images of fibrin nets which allowed me to apply TDA to real world data. In particular I thank Mirjam Bachler, PhD for the idea of using TDA to analyse fibrin networks, Dr. Martin Hermann for creating the images of the nets and Ass.-Prof. Dr. Judith Martini for providing these 3D greyscale images.

I thank all my colleagues at the department of mathematics of the university of Innsbruck for the great atmosphere. I always enjoyed studying there and look forward to conduct my PhD studies in Innsbruck. I especially want to mention Hessel Tuinhof who was the one making me aware of the field of topological data analysis.

On a final note, I want to express my gratitude to my family for their love and support.

Contents

1 Introduction **1**

2 Introduction to Algebraic Topology **3**

2.1 Simplicial Complexes 3

2.2 Simplicial Homology 6

2.3 Calculating Homology Groups 10

3 Topological Data Analysis **13**

3.1 Persistent Homology 13

3.2 Graded Rings and Modules 18

3.3 Decomposition of the Persistence Module 19

3.4 Computing the Persistence Module 22

3.5 Persistence Diagrams 26

3.6 Persistence Landscapes 29

3.7 Weighted Silhouettes 36

3.8 Topological Data Analysis for Image Data 38

4 Application to Fibrin Nets **43**

4.1 Fibrin Nets 43

4.2 Approach and Results 43

5 Conclusion **59**

References **61**

Affidavit **63**

1 Introduction

Topological data analysis is a branch of mathematics which combines topology and statistics in order to analyse datasets by their topological structure. The main idea hereby is that given a point cloud in euclidean space, one assumes that the data was drawn from a manifold and by using algebraic topology tries to measure the persistence of its homology classes, allowing to classify the point cloud geometrically. This approach has been gaining more popularity over the past decade so its ideas have been extended to more abstract situations like greyscale images. Furthermore there are already various software packages available in order to compute persistence. We will mainly use the C++ software package DIPHA[1] and the R package TDA[2] to compute persistent homology and visualise the calculations properly.

The objective of this thesis is to accomplish two goals. First of all it gives an introduction into this modern field of mathematics and secondly we apply the presented methods to real world data. We therefore start with a short discussion of algebraic topology, mainly simplicial homology, in Section 1. Subsequently Section 2 presents topological data analysis. We begin with defining persistent homology goups, followed by proving a main theorem about the decomposition of the persistence module and its calculation. Afterwards we discuss popular methods of visualising and analysing the persistence module. Persistence diagrams, persistence landscapes and weighted silhouettes will hereby be the main topics. We finish Section 3 by describing how topological data analysis is applied to greyscale images. In Section 4 we will use topological data analysis in order to analyse 3D greyscale images of fibrin nets. Fibrin nets are essential for blood clotting and therefore their structure has an immense impact on how fast an open wound stops bleeding. Typically, in a severly injured person, different types of dilutions need to be administered into the bloodstream of the patient. Those dilutions can have a distinct effect on coagulation, which can be observed in the structure of fibrin nets. We will analyse the effect of two such dilutions used on pigs by comparing the topological structure of their natural fibrin nets and their diluted ones.

[1] `https://github.com/DIPHA/dipha`
[2] `https://CRAN.R-project.org/package=TDA`

2 Introduction to Algebraic Topology

Algebraic Topology is a branch of mathematics which studies topological spaces by means of abstract algebra. Results like the Brouwer fixed point Theorem, Jordan-Brouwer separation Theorem, the Borsuk-Ulam Theorem or even the birth of category theory are some classical examples of its achievements. But also modern mathematics still carries on research in this field. For instance the Poincarè conjecture has just been proven in 2006. The main algebraic tool we will focus on is the so-called homological algebra. While there are several homology theories in algebraic topology like singular, simplicial or cell homology, we mainly make use of the simplicial homology due to its computability and its applicability on point cloud data. A nice introduction to those three homologies can be found in [1], which we will mainly follow during the next pages. But before we start we need to briefly repeat the notion of affine simplices and define some notation.

2.1 Simplicial Complexes

In the following if $(a_i)_{i\in I}$ is a family in a given group, for some arbitrary set I, we denote by $\langle (a_i)_{i\in I}\rangle$ its generated subgroup. We use the same notation for generated submodules or ideals, respectively. Given points $p_0, p_1, \ldots, p_m \in \mathbb{R}^n$ we write $[p_0, p_1, \ldots, p_m]$ for the convex set spanned by these points, i. e. the set of all convex combinations of $p_0, p_1, \ldots, p_m$. Furthermore we call the family $(p_0, p_1, \ldots, p_m)$ affine independent if $(p_1 - p_0, \ldots, p_m - p_0)$ is linearly independent in $\mathbb{R}^n$. This is equivalent to the property that each x in the affine set spanned by $\{p_0, \ldots, p_m\} \in \mathbb{R}^n$, i. e. the set of all affine combinations, has a unique expression as an affine combination $\sum_{i=0}^m t_i p_i$, where $t_i \in \mathbb{R}$ and $\sum_{i=0}^m t_i = 1$. This shows in particular that affine independence does not depend on the ordering of the points. Therefore one calls the unique $(m+1)-$tuple $(t_0, \ldots, t_m)$ the barycentric coordinates of x with respect to $(p_0, p_1, \ldots, p_m)$.

Definition 2.1 (**Simplex**)
Let $(p_0, p_1, \ldots, p_m) \in \mathbb{R}^n$ be affine independent. The convex set $s = [p_0, \ldots, p_m]$ is called the (affine) *m-simplex* with vertices $\mathrm{Vert}(s) = \{p_0, p_1, \ldots, p_m\}$. We call m

its dimension and define the *face opposite to* p_i as

$$[p_0, \dots, \hat{p_i}, \dots, p_m] = \left\{ \sum_{j=0}^{m} t_j p_j \,\middle|\, \sum t_j = 1, t_j \geq 0, t_i = 0 \right\}$$

for $i = 0, \dots, m$. More generally we call a simplex s' a *face* of s if $\mathrm{Vert}(s') \subseteq \mathrm{Vert}(s)$ and write $s' \leq s$. If $\mathrm{Vert}(s') \subsetneq \mathrm{Vert}(s)$ we call s' a *proper face* and write $s' < s$.

Remark. We call the set

$$\Delta^n = \left\{ (x_1, x_2, \dots, x_{n+1}) \in \mathbb{R}^{n+1} \,\middle|\, x_i \geq 0 \text{ and } \sum_{i=1}^{n+1} x_i = 1 \right\}$$

the standard n-simplex. ◇

Definition 2.2 (**Simplicial Complex**)
A *simplicial complex* $\mathcal{K}$ is a finite set of simplices in some euclidean space satisfying for every $s, t \in \mathcal{K}$ that

1) every face of s belongs to $\mathcal{K}$,

2) $s \cap t$ is either empty or a common face of s and t.

Furthermore we define its *underlying space* $|\mathcal{K}|$ as

$$|\mathcal{K}| = \bigcup_{s \in \mathcal{K}} s$$

and its *dimension*

$$\dim \mathcal{K} = \max_{s \in \mathcal{K}} (\dim s) \,.$$

Remark. Obviously if $\mathcal{K}$ is a simplical complex, then $|\mathcal{K}|$ is a compact subspace of the given euclidean space. ◇

Example 2.3. The set consisting of the simplices $[A, B, C], [D, E, F]$, see Figure 1, and all their faces, is not a simplicial complex since the second condition does not hold, whereas $\mathcal{K} = \{[A, B, D], [A, D, C], [D, C, F], [C, E, F]$ and all their faces$\}$ is a simplicial complex. ◇

Definition 2.4 (**Polyhedron**)
Let X be a topological space then X is called a *polyhedron* if there exists a simplicial

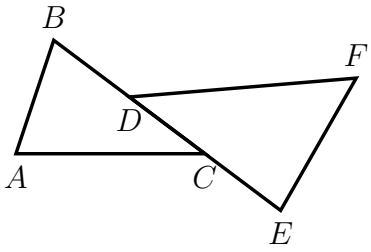


Figure 1: Illustration of Example 2.3

complex $\mathcal{K}$ and a homeomorphism $\varphi\colon |\mathcal{K}| \to X$. We say the pair $(\mathcal{K}, \varphi)$ is a triangulation of X.

Example 2.5. Obviously every simplex defines a simplicial complex $\mathcal{K}$ where $(\mathcal{K}, \mathrm{id})$ is a triangulation. Hence every simplex is a polyhedron. ◇

Example 2.6. We define an equivalence relation $\sim$ on the cartesian product $[0,1]^2$ by identifying $(t,0)$ with $(t,1)$ and identifying $(0,t)$ with $(1,t)$ for every $t \in [0,1]$ and equip $\mathbb{T}^2 = [0,1] \times [0,1]/\sim$ with the final topology. Since

$$\varphi\colon \mathbb{T}^2 \to \mathbb{S}^1 \times \mathbb{S}^1 \colon \overline{(s,t)} \mapsto \Big(\big(\sin(2\pi s), \cos(2\pi t)\big), \big(\sin(2\pi t), \cos(2\pi s)\big)\Big)\,,$$

where $\overline{(s,t)}$ denotes the class generated by (s,t) and $\mathbb{S}^n$ the n-dimensional sphere, is a well defined homeomorphism, $\mathbb{T}^2$ is a torus. Then a triangulation of the torus is given in Figure 2. ◇

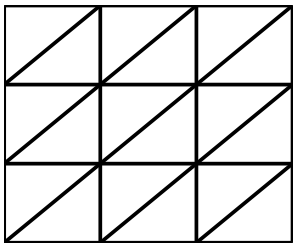


Figure 2: A triangulation of $\mathbb{T}^2$.

Since we will be dealing with more abstract spaces than the euclidean space, we will also need the definition of an abstract simplicial complex.

Definition 2.7 (**Abstract Simplicial Complex**)
Let V be a finite set. An *abstract simplicial complex* $\mathcal{K}$ is a family of non-empty

subsets of V, called *simplices*, such that

1) if $v \in V$, then $\{v\} \in \mathcal{K}$,

2) if $s \in \mathcal{K}$ and $t \subseteq s$, then $t \in \mathcal{K}$.

We call V the *vertex set* of $\mathcal{K}$ and denote it by $\mathrm{Vert}(\mathcal{K})$. And again we call a simplex with $n+1$ distinct vertices an *n-simplex*. A subset of $\mathcal{K}$ which is an abstract simplicial complex itself is called a *subcomplex*.

Remark. Obviously every simplicial complex defines an abstract simplicial complex. Additionally one can show that all simplicial complexes as well as all abstract simplicial complexes form equivalent categories. Hence we won't distinguish between simplicial complexes and abstract simplicial complexes for the rest of this thesis. ◇

2.2 Simplicial Homology

Before we start with the rigorous definition of the simplicial homology we will look at a simple example which will motivate the geometrical ideas of this theory. Consider the topological space X consisting of the vertices a, b, c and the edges $\gamma_1, \gamma_2, \gamma_3, \gamma_4$ given as the graph in Figure 3. We additionally define an orientation on X as shown by the arrows in Figure 3. Our goal is to calculate the number of

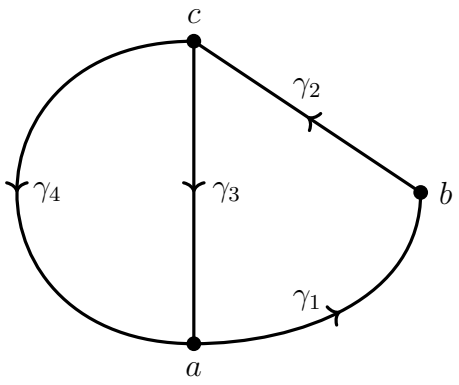


Figure 3: The topological space X.

holes in X. We therefore want to identify closed loops in X. In order to do this in an algebraically well defined way we define the set of zero-dimensional chains

$C_0(X)$ as the free abelian group generated by the vertices a, b and c, i. e. every $x \in C_0(X)$ has the form $x = \alpha a + \beta b + \gamma c$ for some $\alpha, \beta, \gamma \in \mathbb{Z}$. And analogously we define the set of one-dimensional chains $C_1(X)$ to be the free abelian group generated by $\gamma_1, \gamma_2, \gamma_3$ and γ_4. In order to detect closed loops in X we define the boundary operator

$$\partial_1 \colon C_1(X) \to C_0(X)$$

via

$$\partial_1(\gamma_1) = b - a\,, \quad \partial_1(\gamma_2) = c - b\,, \quad \partial_1(\gamma_3) = a - c\,, \quad \partial_1(\gamma_4) = a - c$$

and extend by linearity. Considering the closed loop $\gamma_4 - \gamma_3$ it follows that

$$\partial_1(\gamma_4 - \gamma_3) = a - c - (a - c) = 0\,,$$

thus $\gamma_4 - \gamma_3 \in \ker \partial_1$. We therefore refer to $\gamma_4 - \gamma_3$ as a 1-cycle and define the subgroup of simplicial 1-cycles as $Z_1(X) = \ker \partial_1$. A straightforward calculation shows

$$Z_1(X) = \langle \gamma_1 + \gamma_2 + \gamma_3, \gamma_1 + \gamma_2 + \gamma_4 \rangle\,,$$

hence $\operatorname{rank} Z_1(X) = 2$ which corresponds to the two holes in the given space X. Lets consider a slightly different situation by attaching a closed 2-cell[3] ζ into the left hole resulting in a new space Y which is illustrated in Figure 4. Repeating the

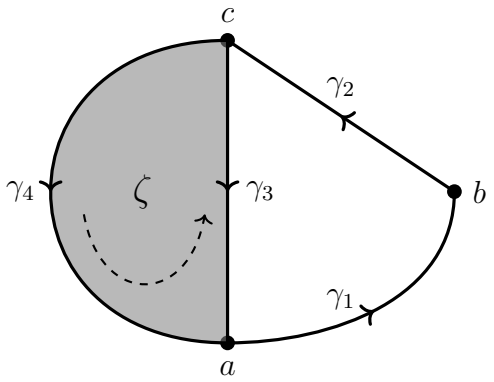


Figure 4: The topological space Y.

[3] A closed n-cell is a homeomorphic copy of the n-dimensional closed unit ball.

above computation in terms of Y now yields the same results as for X implying two holes. But obviously we do not want the cycle $\gamma_4 - \gamma_3$ to capture a hole anymore. Hence we now define $C_2(Y)$ as the free abelian group generated by ζ and another boundary operator

$$\partial_2 \colon C_2(Y) \to C_1(Y)$$

by setting $\partial_2(\zeta) = \gamma_4 - \gamma_3$ and extending by linearity. Clearly $\operatorname{im} \partial_2 = \langle \gamma_4 - \gamma_3 \rangle$ and since

$$\gamma_1 + \gamma_2 + \gamma_3 \equiv \gamma_1 + \gamma_2 + \gamma_4 \bmod \operatorname{im} \partial_2$$

it follows

$$Z_1(X)/\operatorname{im} \partial_2 = \langle \overline{\gamma_1 + \gamma_2 + \gamma_3} \rangle \,.$$

Hence we define the first homology group as $H_1(Y) = Z_1(Y)/\operatorname{im}(\partial_2)$ and showed that $\operatorname{rank} H_2(Y) = 1$ which corresponds to the one hole in Y. This observation motivates the following construction of the simplicial homology.

Definition 2.8 (**Oriented Simplicial Complex**)
A simplicial complex $\mathcal{K}$ is called *oriented* if there is a partial order on $\operatorname{Vert}(\mathcal{K})$ whose restriction to the vertices of any simplex in $\mathcal{K}$ is a linear order.

Definition 2.9 (n-**Chains**)
Let $\mathcal{K}$ be an oriented simplicial complex. For $n \in \mathbb{N}$ we define $C_n(\mathcal{K})$ to be the abelian group generated by all $(n+1)$-tuples $(p_0, \ldots, p_n)$ with $p_i \in \operatorname{Vert}(\mathcal{K})$ for $i = 0, \ldots, n$ such that $\{p_0, \ldots, p_n\}$ spans a simplex in $\mathcal{K}$ and fulfilling the following properties:

1) $(p_0, \ldots, p_n) = 0$ if $p_i = p_j$ for some $i \neq j$,

2) $(p_0, \ldots, p_n) = \operatorname{sign}(\sigma)(p_{\sigma(0)}, \ldots, p_{\sigma(n)})$ where σ is a permutation of $\{0, \ldots, n\}$.

Remark. One can show that $C_n(\mathcal{K})$ is a free abelian group with the basis consisting of all symbols $<p_0, \ldots, p_n>$, where $p_0 < p_1 < \ldots < p_n$ and $\{p_0, \ldots, p_n\}$ spans an n-simplex in $\mathcal{K}$. Hence the above definition corresponds to our construction at the beginning of this subsection. For an easy proof of this fact see [1, Lemma 7.10, p. 143]. In particular this means that $C_n(\mathcal{K})$ is a free $\mathbb{Z}$-module. $\diamond$

Definition 2.10 (**Boundary Operator**)
Let $\mathcal{K}$ be an oriented simplicial complex. For every $n \in \mathbb{N}_{\geq 1}$ we define the nth *boundary operator* $\partial_n \colon C_n(\mathcal{K}) \to C_{n-1}(\mathcal{K})$ by setting

$$\partial_n(<p_0, \ldots, p_n>) = \sum_{i=0}^{n} (-1)^i <p_0, \ldots, \hat{p_i}, \ldots, p_n>$$

and extending by linearity. Furthermore we set ∂_0 as the constant zero mapping on $C_0(\mathcal{K})$.

Theorem 2.11
Let $\mathcal{K}$ be an oriented simplicial complex of dimension n, then

$$0 \to C_n(\mathcal{K}) \xrightarrow{\partial_n} \ldots \xrightarrow{\partial_2} C_1(\mathcal{K}) \xrightarrow{\partial_1} C_0(\mathcal{K}) \to 0$$

is a (chain) complex, i. e. $\partial_k \partial_{k+1} = 0$ for all $k \in \mathbb{Z}$, which we denote by (C_*, ∂).

Proof. The proof is a tedious calculation similar to [1, Theorem 4.6, p. 65]. ■

Remark. Note that the above statement is equivalent to $\operatorname{im} \partial_{k+1} \subseteq \ker \partial_k$. ◇

Definition 2.12 (**Simplicial Homology**)
Let $\mathcal{K}$ be an oriented simplicial complex and $n \in \mathbb{N}$. Then we call

$$Z_n(\mathcal{K}) = \ker \partial_n$$

the group of *simplicial n-cycles*,

$$B_n(\mathcal{K}) = \operatorname{im} \partial_{n+1}$$

the group of *simplicial q-boundaries* and

$$H_n(\mathcal{K}) = Z_n(\mathcal{K}) / B_n(\mathcal{K})$$

the nth *simplicial homology group*. Furthermore we define the nth *Betti number* as $\beta_n = \operatorname{rank} H_n(\mathcal{K})$.

Remark. 1) As we have seen in our initial motivational example, the nth Betti

number counts the number of $(n+1)$-dimensional holes. For example β_0 counts the number of connected components, β_1 counts the number of holes and β_2 gives us the number of voids.

2) Since every abelian group is a $\mathbb{Z}$-module and vice versa we will often refer to $H_n(\mathcal{K})$ as the nth homology module.

3) While we are mainly dealing with simplicial complexes, it is possible to define the homology for more arbitrary topological spaces X. One therefore changes the definition of n-chains from formal sums of simplices to formal sums of continuous mappings

$$\sigma\colon \Delta^n \to X$$

and adapts the boundary operator accordingly, for the technical details see [1, Chapter 4]. Then the definition of the homology groups $H_n(X)$ is the same as in Definition 2.12. It is possible to show that for a simplicial complex $\mathcal{K}$ it holds that

$$H_n(\mathcal{K}) = H_n(|\mathcal{K}|)$$

for every $n \geq 0$, see [1, Theorem 7.22, p. 151]. Hence we can compute the homology of a polyhedron by calculating the simplicial homology groups of the given triangulation. This additionally shows that the simplicial homology groups do not depend on the partial order of $\mathrm{Vert}(\mathcal{K})$. ◇

2.3 Calculating Homology Groups

Let $\mathcal{K}$ be a simplicial complex. Obviously $H_n(\mathcal{K})$ is finitely generated for every n and thus is completely classified by its rank and torsion coefficients by the *Structure Theorem*.

Theorem 2.13 (**Structure Thm. for Finitely Generated Abelian Groups**) *Let G be a finitely generated commutative group. Then it holds*

1) There exists a free abelian group F of finite rank $r \in \mathbb{N}$ and a finite group T such that $G \cong F \oplus T$. We call F the free part and T the torsion part of G.

2) There exist cyclic groups $C_1, \ldots, C_k$ for a unique $k \in \mathbb{N}$ such that for $b_i = |C_i|$ for $i = 1, \ldots, k$ it holds that $b_1 \,|\, b_2 \,|\, \ldots \,|\, b_k$ and

$$T = \bigoplus_{i=1}^{k} C_i \,.$$

We call $b_1, \ldots, b_k$ the torsion coefficients of G.

3) Two finitely generated abelian groups are isomorphic if and only if they have the same rank and the same torsion coefficients.

Proof. See for instance [2, Section 9.1]. ■

Remark. Given a principal ideal domain (PID) R and a finitely generated R-module M of $\operatorname{rank} M = r$ the above theorem takes the following form. There exist non invertible elements $b_1, \ldots, b_k \in R$ with $b_1 \,|\, b_2 \,|\, \ldots \,|\, b_k$ such that

$$M \cong R^{r-k} \times R/\langle b_1 \rangle \times \ldots \times R/\langle b_k \rangle \,.$$

Again $k \in \mathbb{N}$ is unique and the torsion coefficients $b_1, \ldots, b_k$ are unique up to multiplication with units of R. ◇

In order to calculate the above decomposition we use the well known *Smith Normal Form* of matrices. We recall that if R is a euclidean ring, for every matrix A over R there exist invertible matrices P and Q over R such that $A = PSQ$ where S has the form

$$S = \left[\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array}\right]$$

and $D = \operatorname{diag}(b_1, \ldots, b_{\operatorname{rank}(A)})$ and $b_1 \,|\, b_2 \,|\, \ldots \,|\, b_{\operatorname{rank}(A)}$. We call S the Smith Normal Form of A and $b_1, \ldots, b_{\operatorname{rank}(A)}$ the elementary divisors of A which are unique up to multiplication with units of R.

Theorem 2.14 (**Computation of Homology Groups**)
For any oriented simplicial complex $\mathcal{K}$, there exists an algorithm to compute the homology groups of $\mathcal{K}$.

Proof. See [3, p. 60]. ■

Remark. The algorithm is given in the following way. Every $C_n(\mathcal{K})$ is finitely generated. Hence each boundary operator ∂_n can be identified with a matrix M_n with entries 0, 1 and -1. Let S_n be the Smith Normal Form of M_n and set r_n as the number of non-zero rows and c_n as the number of zero columns of S_n. Then the elementary divisors of M_n are the torsion coefficients of $H_n(\mathcal{K})$ and it holds

$$\beta_n = \operatorname{rank} H_n(\mathcal{K}) = c_n - r_{n+1}\,.$$

The reason for the latter equality is the following. Since $C_n(\mathcal{K})$ is a free $\mathbb{Z}$-module, submodules of $C_n(\mathcal{K})$ are also free since $\mathbb{Z}$ is a principial ideal domain. Thus Z_n and B_{n+1} are free. Hence we can apply the Rank-nullity Theorem on the projection map $Z_n \to Z_n/B_{n+1}$ implying that

$$\operatorname{rank} H_n(\mathcal{K}) = \operatorname{rank} Z_n - \operatorname{rank} B_{n+1}\,,$$

which corresponds to the above statement. ◇

3 Topological Data Analysis

The aim of topological data analysis is to identify the geometric structure within some finite statistical data points set D. Therefore the idea is to use the given data points as vertices of a simplicial complex and subdivide this complex into a family of increasing subcomplexes. This should allow to measure the persistence of certain topological features with respect to the given subcomplexes and hence gives us the possibility to classify the data topologically. Since we get finitely many points, this gives an upper bound of $|D|-1$ on the dimension of the simplices and additionally we have an upper limit of $\binom{|D|}{n+1}$ on the number of n-simplices. Therefore we will assume that in the following all simplicial complexes are finite.

3.1 Persistent Homology

We start by constructing the main tool of topological data analysis, the persistent homology, where we mainly follow the approach in [4]. Given a metric space one way of defining a simplicial complex $\mathcal{K}$ from a point cloud is by using the distances between the given points. We introduce two popular methods in this setting.

Definition 3.1 (**Čech[4] and Vietoris[5] -Rips[6] Complex**)
Let $\varepsilon > 0$, (M, d) be a metric space and $F \subset M$ finite.

i) We define the *Čech Complex* $\mathcal{C}_\varepsilon$ as the abstract simplicial complex whose n-simplices are given as unordered $(n+1)$-tuples of points of F whose closed $\varepsilon/2$-ball neighborhoods have a point of common intersection.

ii) We define the *Vietoris-Rips Complex* $\mathcal{R}_\varepsilon$ as the abstract simplicial complex whose n-simplices are given as unordered $(n+1)$-tuples of points of F which are pairwise within distance ε.

Example 3.2. Figure 5 illustrates a Čech as wells as a Vietoris-Rips complex from a point cloud for some fixed $\varepsilon > 0$. ◇

[4] *Eduard Čech*, 1893—1960, Czech mathematician
[5] *Leopold Vietoris*, 1891—2002, Austrian mathematician
[6] *Eliyahu Rips*, born 1948, Israeli mathematician

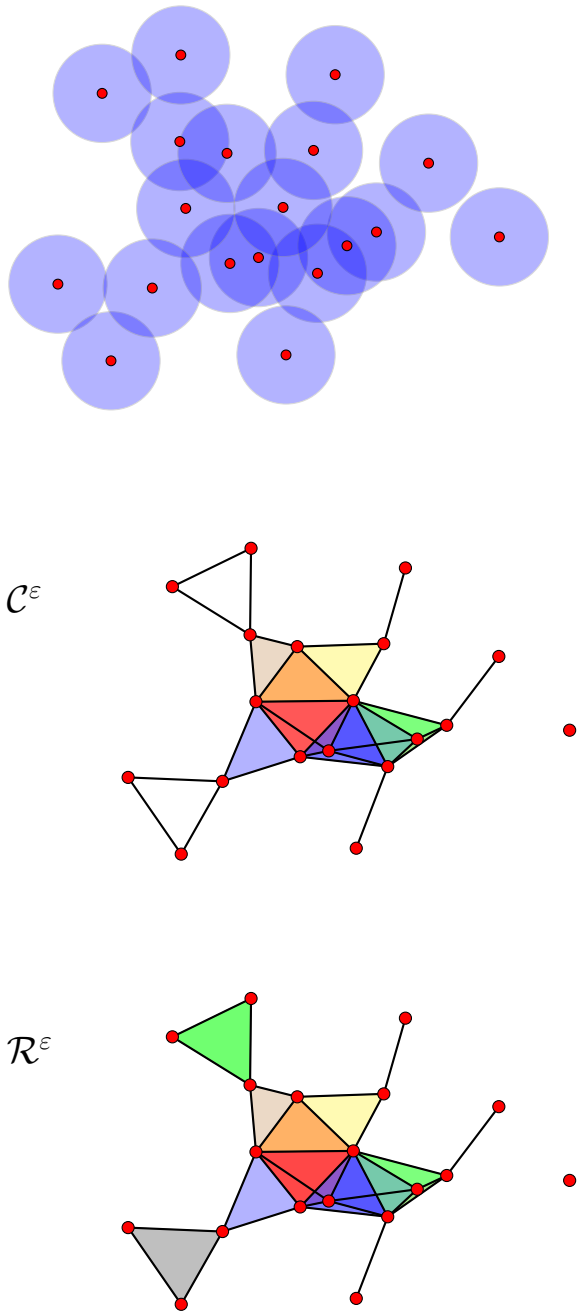


Figure 5: A point cloud in $\mathbb{R}^2$ where every point is surrounded by an ε neighbourhood (top) and its corresponding Čech Complex (middle) and Vietoris-Rips Complex (bottom) for a fixed $\varepsilon > 0$.

Definition 3.3 (**Filtration**)
Given a simplicial complex $\mathcal{K}$, a *filtration* is a totally ordered set of subcomplexes $\mathcal{K}^i$ of $\mathcal{K}$, for $i \in \mathbb{N}$, such that $i \leq j$ implies $\mathcal{K}^i \subseteq \mathcal{K}^j$.

Remark. 1) Assume there is a function $f\colon \mathcal{K} \to \mathbb{R}$ defined on a simplicial complex such that it is monotonic on faces of $\mathcal{K}$, i. e. whenever $s' \leq s$ holds it follows $f(s') \leq f(s)$. This implies that $\mathcal{K}^a = f^{-1}((-\infty, a])$ is a subcomplex $\mathcal{K}$ for every $a \in \mathbb{R}$. Since we assume that $\mathcal{K}$ is finite, f takes finitely many values. Let $a_1 < \ldots < a_N$ for $N \in \mathbb{N}$ be the function values of f, then we get an increasing sequence

$$\emptyset = \mathcal{K}^0 \subseteq \mathcal{K}^1 \subseteq \ldots \subseteq \mathcal{K}^N = \mathcal{K}$$

where $a_0 = -\infty$ and $\mathcal{K}^i = \mathcal{K}^{a_i}$. We call this sequence the filtration of f.

2) Using the Čech or Vietoris-Rips complex we can generate a filtration by choosing an increasing sequence $(\varepsilon_i)_{i=1}^N$. This gives

$$\mathcal{C}^0 \subseteq \mathcal{C}^{\varepsilon_1} \subseteq \ldots \subseteq \mathcal{C}^{\varepsilon_N}$$

and

$$\mathcal{R}^0 \subseteq \mathcal{R}^{\varepsilon_1} \subseteq \ldots \subseteq \mathcal{R}^{\varepsilon_N},$$

respectively. ◇

Definition 3.4 (**Subcomplex**)
Let (S_*, ∂) be a complex, then we say that the complex (S'_*, ∂') is a *subcomplex* of S_* if the diagram

$$\begin{array}{ccc} S'_n & \xrightarrow{\partial'_n} & S'_{n-1} \\ {\scriptstyle \mathrm{inj}^n}\downarrow & & \downarrow{\scriptstyle \mathrm{inj}^{n-1}} \\ S_n & \xrightarrow{\partial_n} & S_{n-1} \end{array}$$

where inj^n is an injection from S'_n into S_n, commutes for every $n \in \mathbb{Z}$, i. e. for every n S'_n is a submodule of S_n and $\partial_n | S'_n = \partial'_n$.

Remark. In the following, especially when dealing with subcomplexes, we will omit certain indices in order to improve readability. Furthermore we sometimes write

S_* instead of (S_*, ∂) for a given complex. ◇

From now on we assume that $\mathcal{K}$ is a simplicial complex equipped with a filtration $(\mathcal{K}^i)_{i\in\mathbb{N}}$. Then every i defines its own chain complex $(\mathcal{K}^i_*, \partial^i)$ and it's own homology module. We therefore write $\mathcal{K}^i_n, Z^i_k, B^i_n$ and H^i_n for the nth chain, cycle, boundary and homology module of $\mathcal{K}^i$ respectively as well as β^i_n for the corresponding Betti number. Since for every $p > 0$ it holds $\mathcal{K}^i \subseteq \mathcal{K}^{i+p}$ and we can embed $\mathcal{K}^i_n$ into $\mathcal{K}^{i+p}_n$. Thus Z^i_n is a submodule of $\mathcal{K}^{i+p}_n$. In particular $(\mathcal{K}^i_*, \partial^i)$ is a subcomplex of $(\mathcal{K}^{i+p}_*, \partial^{i+p})$. This allows us to make the following definition.

Definition 3.5 (**Persistent Homology Modules**)
Let $j \geq i$ then we call the modules

$$H^{i,j}_n = Z^i_n/(B^j_n \cap Z^i_n)$$

the nth *persistent homology modules*. Additionally we set the corresponding *nth persistent Betti numbers* as

$$\beta^{i,j}_n = \operatorname{rank} H^{i,j}_n \,.$$

Remark. Instead of measuring holes in $\mathcal{K}^i$ which are not generated as boundaries from an $(n+1)$-chain, $H^{i,j}_n$ characterizes the n-cycles in $\mathcal{K}^i$ which are not the boundary of an $(n+1)$-chain in the larger complex $\mathcal{K}^j$. So $H^{i,j}_n$ characterises $(n+1)$-dimensional holes in $\mathcal{K}^j$ created in $\mathcal{K}^i$. Note that these holes exist in every complex $\mathcal{K}^\ell$ where $i \leq \ell \leq j$. ◇

Definition 3.6 (**Chain Map**)
Let (S_*, ∂) and (S'_*, ∂') be complexes, then we call a sequence of morphisms $(f^n \colon S'_n \to S_n)_{n\in\mathbb{Z}}$ a *chain map* if the diagram

$$\begin{array}{ccccccccc}
\cdots & \longrightarrow & S'_{n-1} & \xrightarrow{\partial'_{n+1}} & S'_n & \xrightarrow{\partial'_n} & S'_{n-1} & \longrightarrow & \cdots \\
 & & \downarrow{\scriptstyle f^{n+1}} & & \downarrow{\scriptstyle f^n} & & \downarrow{\scriptstyle f^{n-1}} & & \\
\cdots & \longrightarrow & S_{n-1} & \xrightarrow{\partial_{n+1}} & S_n & \xrightarrow{\partial_n} & S_{n-1} & \longrightarrow & \cdots
\end{array}$$

commutes for every $n \in \mathbb{Z}$ and shortly write $f = (f_n) \colon (S'_*, \partial') \to (S_*, \partial)$.

In the above situation we see that for two complexes $(\mathcal{K}^i_*, \partial)$ and $(\mathcal{K}^{i+1}_*, \partial)$ the

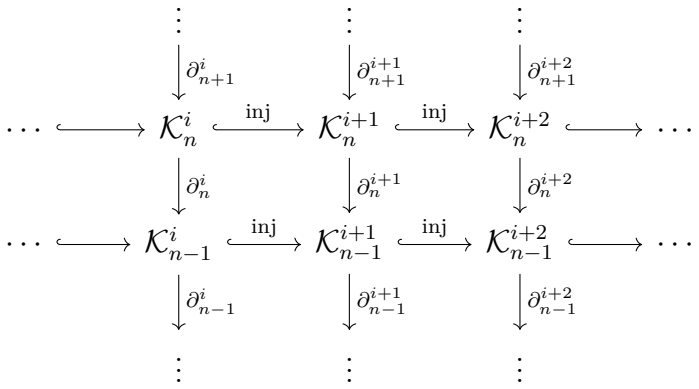


Figure 6: A section of a persistence module.

familiy of injections $\mathrm{inj}^i \colon (\mathcal{K}_*^i, \partial) \to (\mathcal{K}_*^{i+1}, \partial)$ is a chain map which induces morphisms $\eta_n^i = \mathrm{inj}_{\mathrm{ind}}^i \colon H_n^i \to H_n^{i+1}$ on the homology groups. This motivates the following definitions.

Definition 3.7 (**Persistence Complex**)
The sequence of complexes and chain maps $(\mathcal{K}_*^i, \mathrm{inj}^i)$ is called a *persistence complex*.

Remark. 1) Figure 6 shows a section of a persistent complex. Each column is a chain complex.

2) For $j \geq i$ set

$$\eta_n^{i,j} = \eta_n^{j-1,j} \circ \ldots \circ \eta_n^{i+1,i+2} \circ \eta_n^{i,i+1} \,,$$

then $H_n^{i,j} = \operatorname{im} \eta_n^{i,j}$ holds. ◇

Definition 3.8 (**Persistence Module**)
The *nth persistence module* $\mathscr{H}_n$ is the family of homology modules H_n^i and module morphisms η_n^i. We call a persistence module of *finite type* if each module is finitely generated and there exists an integer m such that for every $i \geq m$ it holds that η_n^i is an isomorphism.

Remark. 1) Since we assumed that all simplicial complexes we are dealing with are finite, all persistence modules are of finite type.

2) More generally given a function $f \colon X \to \mathbb{R}$ defined on some arbitrary topo-

logical space X one defines the associated nth persistence module $\mathscr{H}_n(f)$ by $H_n^a = H_n(f^{-1}((-\infty, a]))$ for $a \in \mathbb{R}$ and the corresponding module morphisms again induced by inclusion. One calls f tame if the associated persistence module is constant and finite-dimensional for all but finitely many $a \in \mathbb{R}$. Hence in our situation we always assume f is tame. ◇

3.2 Graded Rings and Modules

Our next goal is to classify the persistence module. We therefore need to shortly repeat the notion of graded rings and graded modules.

Definition 3.9 (**Graded Ring**)
A ring R is called $\mathbb{Z}$-*graded* if for every $k \in \mathbb{Z}$ there exists an additive subgroup R_k such that

$$R = \bigoplus_{k \in \mathbb{Z}} R_k$$

and

$$R_k \cdot R_\ell \subseteq R_{k+\ell}$$

for each $k, \ell \in \mathbb{Z}$. We call an element of $a \in R_k$ *homogenous of degree* k and write $\deg(a) = k$.

Remark. In the following we simply call a $\mathbb{Z}$-graded ring, graded. ◇

Example 3.10. Let $\mathbb{F}$ be a field. Then the polynomial ring $\mathbb{F}[t]$ is a graded ring by setting

$$\mathbb{F}[t]_k = \{ct^k \mid c \in \mathbb{F}\}$$

for $k \geq 0$ and $\mathbb{F}[t]_k = \{0\}$ for negative k. ◇

Definition 3.11 (**Homogenous Ideal**)
Let R be a graded ring, then a two-sided ideal I in R is called homogenous if one of the following equivalent conditions holds:

1) $a = \sum_{k \in \mathbb{Z}} a_k \in I$ where $a_k \in R_k$ implies $a_k \in I$ for every k,

2) $I = \bigoplus_{k \in \mathbb{Z}} I \cap R_k$,

3) I is generated by homogenous elements.

Propostion 3.12 (**Graduation of the Quotient Ring**)
If R is a graded ring and I a homogenous ideal then R/I is a graded ring by defining

$$(R/I)_k = (R_k + I)/I\,.$$

Proof. Assume that

$$\sum_{k\in\mathbb{Z}} \overline{a}_k = 0 \quad \text{in} \quad R/I$$

for some $a_k \in (R_k + I)$. Then $\sum_{k\in\mathbb{Z}} a_k \in I$ and since I is homogenous it follows $a_k \in I$ for each k. Hence

$$\sum_{k\in\mathbb{Z}} (R/I)_k = \bigoplus_{k\in\mathbb{Z}} (R/I)_k.$$

The remaining conditions are clear since R is graded. ■

Definition 3.13 (**Graded Module**)
A left *graded module* is a left module M over a graded ring R such that

$$M = \bigoplus_{k\in\mathbb{Z}} M_k$$

where each M_k is a submodule of M and

$$R_k \cdot M_\ell \subseteq M_{k+\ell}$$

for all $k, \ell \in \mathbb{Z}$. Again we call an element of M_k homogenous of degree k. Given two graded Modules M_1 and M_2 an R linear bijective map $\varphi\colon M_1 \to M_2$ is called a *graded module isomorphism* if for every $m \in M_1$ homogenous of degree k it holds $\varphi(m)$ has degree k in M_2 for every k. In this situation we call M_1 and M_2 isomorphic as graded modules.

3.3 Decomposition of the Persistence Module

Now we want to define a structure on $\mathscr{H}_n$ such that it becomes a graded module over the polynomial ring $\mathbb{F}[t]$ for some field $\mathbb{F}$. At this point we want to mention, that one can define the notion of chains, cycles, boundaries and homology classes

analogously over arbitrary rings or fields instead of $\mathbb{Z}$. We therefore identify in the following n-chains as formal sums with coefficients in $\mathbb{F}$.

By definition we can identify

$$\mathscr{H}_n = \bigoplus_{i=0}^{\infty} H_n^i \,.$$

For the variable t we now define the multiplication as

$$t \cdot \left(\sum_{i=0}^{\infty} \xi^i \right) = \sum_{i=0}^{\infty} \eta_n^i(\xi^i)$$

where $\xi^i \in H_n^i$. Then clearly $\mathscr{H}_n$ is a graded module which is also $\mathbb{F}[t]$ finitely generated. Hence our goal will be to again classify the persistence module via the Structure Theorem. But we additionally want to make use of the graduation of $\mathscr{H}_n$. Im order to do so we need to shift the graduation of $\mathbb{F}[t]$ properly. Choose a finite generating system $\gamma_1, \ldots, \gamma_r$ of homogenous elements of the persistence module with minimal cardinality and let $d_i = \deg(\gamma_i)$ for $i = 1, \ldots, r$. We define a graduation on $\mathbb{F}[t]^r$ as $\mathbb{F}[t]$-module by

$$\mathbb{F}[t]^r = \bigoplus_{\ell=0}^{\infty} \left(\bigoplus_{k=1}^{r} \mathbb{F}[t]_{\ell - d_k} \right).$$

Hence we shift the graduation in the components by the degree of the generators of $\mathscr{H}_n$. Consider the canonical surjection

$$\varphi \colon \mathbb{F}[t]^r \to \mathscr{H}_n \colon (p_1, \ldots, p_r) \mapsto \sum_{k=1}^{r} p_k \gamma_k \,.$$

Let $(p_1, \ldots, p_r)$ be homogenous, i. e. there exists an $\ell \in \mathbb{N}$ such that $p_k \in \mathbb{F}[t]_{\ell - d_k}$ for every k. Thus

$$\varphi(p_1, \ldots, p_r) = \sum_{k=1}^{r} p_k \gamma_k \in H_n^{\ell}$$

since $p_k \gamma_k \in \mathbb{F}[t]_{\ell - d_k} H_n^{d_k} \subseteq H_n^{\ell}$. Hence φ is a graded homomorphism and thus $U = \ker(\varphi)$ is a homogenous ideal. Since U is a submodule and $\mathbb{F}[t]$ is PID, U is

also finitely generated by homogenous elements, i. e.

$$U = \langle t^{c_1}, \dots, t^{c_m} \rangle$$

for some non-negative integers $c_1, \dots, c_m$ and $m \in \mathbb{N}$. Applying the first Isomorphism Theorem yields

$$\mathscr{H}_n \cong \mathbb{F}[t]^r / U$$

and since

$$\bigoplus_{\ell=0}^{\infty} \mathbb{F}[t]_{\ell-d} \longrightarrow \langle t^d \rangle \subseteq \bigoplus_{\ell=0}^{\infty} \mathbb{F}[t]_\ell \colon p \longmapsto t^d p$$

is a graded module isomorphism for every $d \in \mathbb{N}$, the above yields the following decomposition.

Theorem 3.14 (**Decomposition of the Persistence Module**)
In the above situation it holds

$$\mathscr{H}_n \cong \left(\bigoplus_{i=1}^{m} \langle t^{a_i} \rangle \right) \oplus \left(\bigoplus_{j=1}^{n} \langle t^{b_j} \rangle / \langle t^{c_j} \rangle \right)$$

for some unique non-negative integers n, m, a_i, b_j and c_j.

Proof. The existence of this graded module isomorphism was shown above hence uniqueness remains to be proven. Due to the Structure Theorem it follows that n and m are unique. Now assume there exist additional non-negative integers $\tilde{a_i}, \tilde{b_j}$ and $\tilde{c_j}$ such that

$$\mathscr{H}_n \cong \left(\bigoplus_{i=1}^{m} \langle t^{\tilde{a_i}} \rangle \right) \oplus \left(\bigoplus_{j=1}^{n} \langle t^{\tilde{b_j}} \rangle / \langle t^{\tilde{c_j}} \rangle \right).$$

We show uniqueness of the free part. Let φ denote the graded module isomorphism

$$\varphi \colon \bigoplus_{i=1}^{m} \langle t^{a_i} \rangle \longrightarrow \bigoplus_{i=1}^{m} \langle t^{\tilde{a_i}} \rangle .$$

Since φ is a graded isomorphism it holds $\varphi(t^{a_1}) = \alpha t^{a_1}$ for some $\alpha \in \mathbb{F}[t]_0 = \mathbb{F}$. Choose $\tilde{\beta_1}, \dots, \tilde{\beta_m}$ such that $\alpha t^{a_1} = \sum_{i=1}^{m} \tilde{\beta_i} t^{\tilde{a_i}}$. Since t^a is homogenous for every $a \in \mathbb{N}$ we can assume that every $\tilde{\beta_i}$ is homogenous and thus there exist $\beta_1, \dots, \beta_n \in \mathbb{F}$ such that $\tilde{\beta_i} = \beta_i t^{a_1 - \tilde{a_i}}$ for $i = 1, \dots, n$. It follows $\alpha t^{a_1} = \sum_{i=1}^{m} \beta_i t^{a_1}$

and since the above sums are direct it follows there exists a j such that $a_1 = a_j$, $\alpha = \beta_j$ and $\beta_i = 0$ for $i \neq j$. Using that φ is bijective iterating the above argument implies the claim. ■

Remark. We can interpret the above decomposition as follows. Each a_i and b_j correspond to a new $(n+1)$-dimensional hole which is born in the simplicial complexes $\mathcal{K}^{a_i}$ and $\mathcal{K}^{b_j}$, respectively, while c_j represents the index of the hole born in complex $\mathcal{K}^{b_j}$, in which it disappears. Hence the above theorem enables us to classify the topological features within our data. ◇

3.4 Computing the Persistence Module

We now quickly discuss a basic algorithm to compute the persistence modules according to Theorem 3.14. We hereby follow mainly [5, Chapter 7], [6] as well as [7, Topic 4]. Due to computational sufficiency we now slightly change the definition of n-chains. Let $\mathcal{K}$ be a finite simplicial complex as usual. In the following we define $C_n(\mathcal{K})$ as the $\mathbb{Z}/\mathbb{Z}2$ vector space generated by the n-dimensional simplices of $\mathcal{K}$, instead of being the generated $\mathbb{Z}$-module. Furthermore we define the boundary operators by setting $\partial_n(\sigma)$ as the sum of all $(n-1)$-dimensional faces where σ is an n-dimensional chain, and extend by linearity. The definitions of cycles, boundaries and homology groups stay the same.
Assume that a filtration of $\mathcal{K}$ is given and that the simplices $\sigma_1, \ldots, \sigma_m$ of $\mathcal{K}$ are totally ordered such that the faces of a simplex precede the simplex itself and that the simplices in each complex $\mathcal{K}^i$ in the filtration precede the ones in $\mathcal{K} \setminus \mathcal{K}^i$, e. g. $\mathcal{K}$ is generated by a monotonic function f. Now we encode this ordering and hence the filtration via one matrix ∂ by setting

$$\partial_{ij} = \begin{cases} 1, & \text{if } \sigma_i \text{ is a codimension one face of } \sigma_j; \\ 0, & \text{otherwise.} \end{cases}$$

Note that the rows and columns of ∂ are ordered like the simplices and that the boundary of a simplex corresponds to the entries of its column. Due to the ordering ∂ is obviously upper triangular. For a non-zero column ∂_{-j} we define $\text{low}(j) = \max\{i \in \{1, \ldots, m\} \mid \partial_{ij} = 1\}$ and call ∂ reduced if $\text{low}(j_1) \neq \text{low}(j_2)$

whenever $j_1 \neq j_2$. Algorithm 1 then computes a reduced form of ∂ by adding columns from left to right. In the worst case this algorithm needs $\mathcal{O}(m^3)$ operations, see [8]. Since R is computed by adding columns from ∂ we can write $R = \partial \cdot V$ where V encodes those elementary operations. Since Algorithm 1 adds columns only from left to right V is also upper triangular and thus R. While the matrices R and V are not unique one can show that the lowest non-zero entries do not depend on the algorithm used in order to compute a reduced form, see [5, p. 183]. R and V now contain all the information we need in order to calculate the persistent homology modules. Initially we can compute the Betti numbers of the complex $\mathcal{K}$. We write $\#\operatorname{Zero}_n(R)$ for the number of zero columns and $\#\operatorname{Low}_n(R)$ for the number of lowest ones in rows of R which correspond to an n-simplex. Additionally we denote by r_n the number of n-simplices of $\mathcal{K}$, i. e. $r_n = \operatorname{rank} C_n(\mathcal{K})$ and let D_n be the corresponding matrix to ∂_n. Since V is invertible the ranks of ∂ and R are the same. Thus it follows

Algorithm 1: Reduce Matrix [5]

Data: ∂ Boundary Matrix
Result: R reduced form of ∂

```
R = ∂
for j = 1, ..., m do
    while there exists j_0 < j with low(j_0) = low(j) do
        R_{-j} ← R_{-j} + R_{-j_0};
    end
end
return R
```

$$\operatorname{rank} B_{n-1}(\mathcal{K}) = \operatorname{rank} D_n = \#\operatorname{Low}_n(R)$$

which implies

$$\operatorname{rank} Z_n(\mathcal{K}) = r_n - \operatorname{rank} D_n = \#\operatorname{Zero}_n(R)$$

and therefore

$$\operatorname{rank} H_n(\mathcal{K}) = \#\operatorname{Zero}_n(R) - \#\operatorname{Low}_n(R)\,.$$

But the decomposition $R = \partial \cdot V$ stores more information. Foremost we observe what happens by adding the n-simplex $\sigma_j \in \mathcal{K} \setminus \mathcal{K}^{j-1}$ to $\mathcal{K}^{j-1}$ and assume for simplicity that $\mathcal{K}^j = \{\sigma_j\} \cup \mathcal{K}^{j-1}$. Then two situations are possible.

1) Adding σ_j to $\mathcal{K}^{j-1}$ creates a new n-cycle since there is no $(n+1)$-chain in $\mathcal{K}^j$ where σ_j is a face. This new cycle cannot be part of the boundary of an $(n+1)$-chain and hence a new homology class is born. Furthermore only one new homology class can be generated. Every newly generated cycle contains σ_j. Thus by choosing one new cycle γ and adding it as a new basis element to a basis of the previos homology module, then each new cycle can be written as a linear combination of γ and the older basis. Since therefore σ_j increases the corresponding Betti number β_n^i we shall henceforth call σ_j positive.

2) Adding σ_j does not create a new cycle. Hence the boundary $\partial_n(\sigma_j)$ was a non-trivial cycle in $\mathcal{K}^{j-1}$ and was filled by σ_j. Again only one homology class can be killed. Since σ_j now reduces the Betti number we call it negative.

Now we will see that we can decide which of the cases above occurs by simply looking at R. Firstly we realise that adding ∂_{-j_1} to ∂_{-j_2} corresponds to the sum $\sigma_{j_1} + \sigma_{j_2}$ and since Algorithm 1 only adds columns from left to right the column R_{-j} obtains its final form at the end of the jth iteration of the for loop. Again there are two situations possible at this point.

1) R_{-j} is a zero column. Hence the n-chain given by the sum of simplices indexed by the row indices of non-zero entries in V_{-j} creates a cycle. Since σ_j is a summand this cycle has to be new. Thus σ_j is positive.

2) R_{-j} is non-zero. Let γ denote the $(n-1)$-simplex accumulated in R_{-j}. Then γ is a non-trivial cycle in $\mathcal{K}^{j-1}$, otherwise R_{-j} could have been written as a linear combination of the previous columns and hence it would have become a zero column but in $\mathcal{K}^j$ it becomes a boundary. Thus adding σ_j kills the homology class $\overline{\gamma}$. Now let $k = \text{low}(j)$. Then the cycle γ was generated by adding σ_k, since it is the youngest part of γ and since we have a filtration also the homology class $\overline{\gamma}$ was born when adding σ_k. So putting it all together we get that adding σ_j kills the homology class γ which was born at $\mathcal{K}^k$. Hence σ_j is negative.

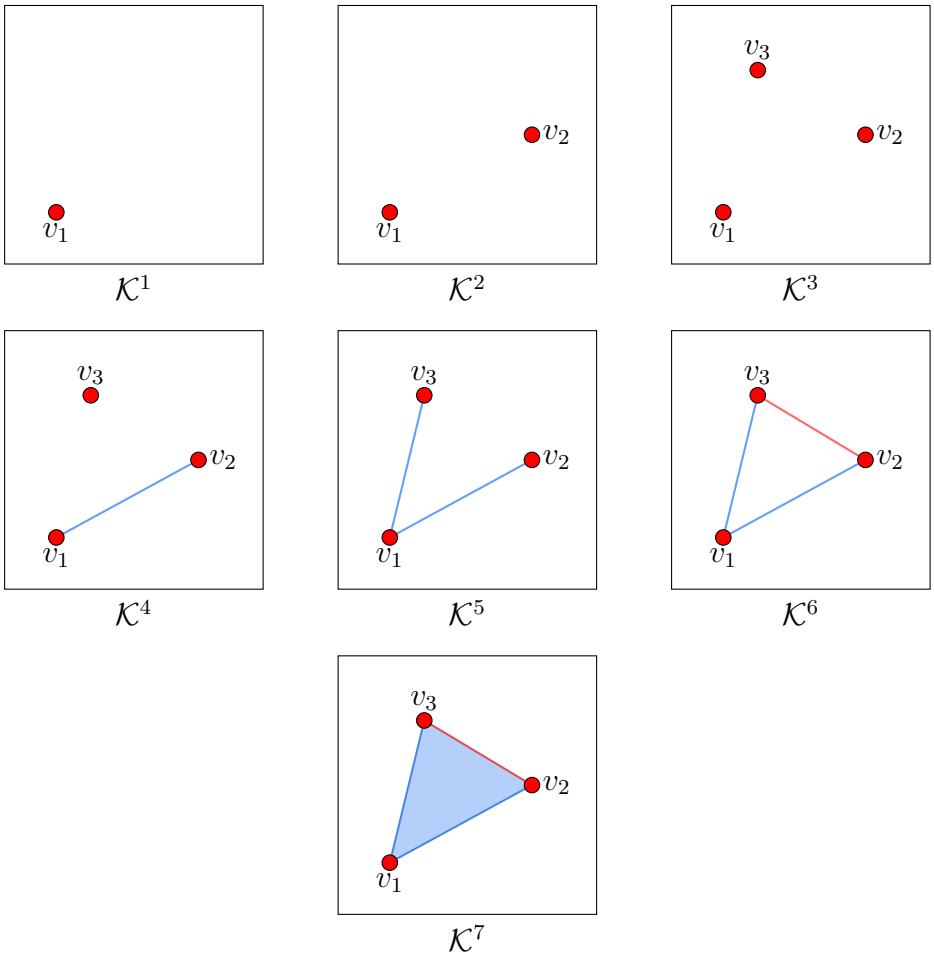


Figure 7: The filtration described in Example 3.15. Red simplices give birth to new, while blue ones kill previously existing homology classes.

Example 3.15. Consider for example $\mathcal{K}$ to be a simplicial complex consisting of a triangle and all of its faces, i. e.

$$\mathcal{K} = \Big\{[v_1], [v_2], [v_3], [v_1, v_2], [v_2, v_3], [v_1, v_3], [v_1, v_2, v_3]\Big\}$$

where v_1, v_2, v_3 denote distinctive points in some euclidean space. A filtration is given by initially adding the vertices, then the edges and finally the whole triangle and numbering them in this order from 1 to 7, see Figure 7. Application of the above algorithm on the boundary matrix yields the following decomposition

$$
\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array}
\begin{array}{c}
\begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\
\left(\begin{array}{ccccccc}
\mathbf{1} & 0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} \\
0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}\right) \\
\underbrace{\qquad\qquad\qquad}_{R}
\end{array}
=
\begin{array}{c}
\begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\
\left(\begin{array}{ccccccc}
1 & 1 & 1 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}\right) \\
\underbrace{\qquad\qquad\qquad}_{\partial}
\end{array}
\begin{array}{c}
\begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\
\left(\begin{array}{ccccccc}
1 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1
\end{array}\right) \\
\underbrace{\qquad\qquad\qquad}_{V}
\end{array}
$$

The first lowest one is in the first row and column, which corresponds to the transition from the empty set to $\mathcal{K}^1$ consisting of the first vertex. Since adding the vertices 2 and 3 creates two new 0-cycles, namely $[v_1]+[v_2]$ and $[v_1]+[v_3]$, see V_{-2} and V_{-3}, R_2 and R_3 are zero-columns. In column four we find the second lowest one. This signifies that the 0-cycle which was created by adding the vertex 2 is killed by edge $[v_1, v_2]$. Indeed it holds $\partial_1([v_1, v_2]) = [v_1]+[v_2]$ where ∂_1 denotes again the first boundary operator. Equivalently the 0-cycle born by vertex 3 vanishes due to edge $[v_1, v_3]$. Since column 6 is zero, adding edge $[v_2, v_3]$ doesn't kill anything and generates the first 1-cycle given by the sum of the edges $[v_1, v_2]+[v_2, v_3]+[v_1, v_3]$, indicated in V_{-6}. This 1-cycle born in $\mathcal{K}^6$ is then killed by adding the whole triangle in the 7th and final part of the filtration. ◇

Remark. There are already more efficient algorithms available. One variation of Algorithm 1 which relies on sparse matrix implementations can also be found in [6]. ◇

3.5 Persistence Diagrams

Since we are now able to compute the persistence module, we want to visualise it. One way of achieving this is by means of so-called persistence diagrams. We hereby follow mainly [9]. Let the filtration of the simplicial complex at hand be given by sublevel sets of a tame function f, see Subsection 3.1, then we define for an n-dimensional homology class γ which is born in $\mathcal{K}^{a_i}$ and vanishes entering $\mathcal{K}^{a_j}$, $\mathrm{b}(\gamma) = a_i$, $\mathrm{d}(\gamma) = a_j$ and $\mathrm{pers}(\gamma) = a_j - a_i$ respectively. Hence we can identify each such class by a multiset of $\mathbb{R}^2$. The multiset of all those points together with

the diagonal Δ of $\mathbb{R}^2$ is called the nth persistence diagram which we denote by $\mathrm{dgm}_n(f)$. The technical reason for adding the diagonal is clarified in the next remark. Intuitively we can interpreted it as representing trivial homology classes which arise and die in the same simplicial complex. Since $\mathrm{pers}(\gamma) \geq 0$ for every γ all points lie above or on the diagonal. Note that every persistence diagram has only finitely many points off the diagonal. Since we only consider classes which die at some point in the filtration, one often therefore chooses a filtration such that every hole vanishes at some point or one identifies a class which was born in $\mathcal{K}^{a_i}$ but never vanishes with the point (a_i, a_N) where $a_N = \sup_{\mathcal{K}} f$. A metric on the set of all persistence diagrams is given by the Wasserstein distance.

Definition 3.16 (**Wasserstein Distance**)
Let $p \geq 1$ then the pth *Wasserstein distance* of two persistence diagrams d_1 and d_2 is defined as

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

where σ ranges over all bijections from d_1 to d_2 and $\|\cdot\|_\infty$ denotes the infinity norm.

Remark. Since we added the diagonal to a persistence diagram the set of all bijections from d_1 to d_2 is non-empty. ◇

For $n \in \mathbb{N}$ define a persistence diagram $d_n = \{(0, 2^{-k}) \,|\, k = 1, \ldots, n\} \cup \Delta$. Then it holds

$$W_p(d_n, d_{n+k}) \leq \frac{1}{2^n}$$

and thus $(d_n)_{n \in \mathbb{N}}$ is a Cauchy sequence. But since the number of points above the diagonal goes to infinity as $n \to \infty$, the limit is not a persistence diagram anymore. One therefore adapts the definition of persistence diagrams.

Definition 3.17 (**Generalized Persistence Diagram**)
A *generalized persistence diagram* is a countable multiset of points in $\mathbb{R}^2$ together with the diagonal Δ where each point on the diagonal has multiplicity infinity.

Definition 3.18 (**Space of Persistence Diagrams**)
Let $d_\emptyset = \Delta$ denote the empty diagram and $p \geq 1$ then we call the subset

$$\mathcal{D}_p = \{d \,|\, W_p(d, d_\emptyset) < \infty\}$$

of all generalized persistence diagrams as the *space of persistence diagrams*.

Remark. With the above adaption it is shown in [9] that $\mathcal{D}_p$ is a Polish space, i. e. $\mathcal{D}_p$ is a completely metrisable space. This admits the definition of Fréchet means and conditional probabilities. Although the Fréchet mean may be not unique in this setting. Again see [9] for the corresponding constructions and proofs. ◇

Example 3.19. Figure 8 shows the persistence diagram of the filtration in Example 3.15. Blue points represent zero-dimensional homology classes while orange ones display one-dimensional classes. Since the component created by adding the first vertice never vanishes, we draw the according homology class as a rectangle instead of a circle with coordinates $(1, 7)$. ◇

Example 3.20. Figure 9 shows three Vietoris-Rips complexes created from a point cloud drawn from a double annulus and the corresponding persistence diagram. The upper circle has radius 1 while the circle below has 1/2. As in Example 3.19 blue points represent zero-dimensional and orange points illustrate one-dimensional homology classes and since one component never dies, we draw it as a rectangle. As we can see many components rise, but vanish with increasing radii. Furthermore the diagram captures the birth and death of two holes. For the calculation of the persistence diagram the C++ software package DIPHA was used. ◇

We finish this subsection by stating a stability theorem from [10]. Let $k \in \mathbb{R}$. We say a topological space X implies bounded degree-k persistence if for every Lipschitz continuous map f with Lipschitz constant ≤ 1 it holds

$$\sum_{\gamma \in \mathrm{Dgm}_n(f)} \mathrm{pers}(\gamma)^k \leq C_X$$

for some constant $C_X \geq 0$ only depending on X.

Theorem 3.21 (**Total Persistence Stability Theorem**)
Let X be a triangulable, compact metric space that implies bounded degree-k total persistence for $k \geq 1$, and let $f, g \colon X \to \mathbb{R}$ be two tame Lipschitz continuous

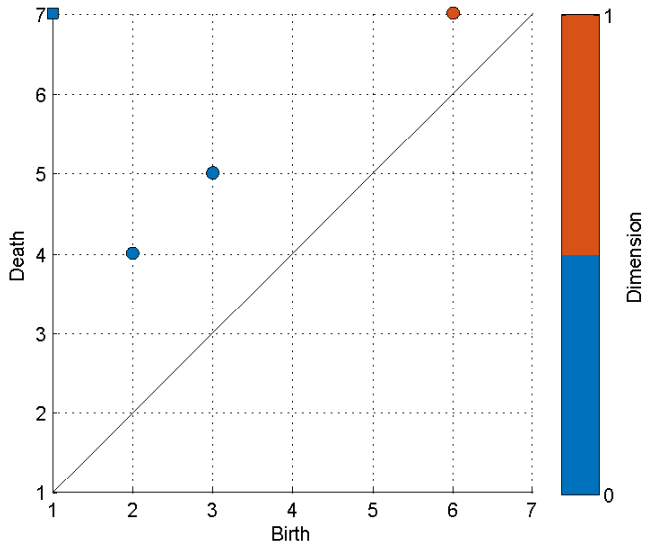


Figure 8: Persistence diagram of Example 3.19

functions. Then it holds for all dimensions $n \in \mathbb{N}$ *and* $p \geq k$ *that*

$$W_p\Big(\operatorname{Dgm}_n(f), \operatorname{Dgm}_n(g)\Big) \leq C\|f-g\|_\infty^{1-\frac{k}{p}}$$

where C *is some constant depending on* X, f, g *and* k*.*

3.6 Persistence Landscapes

Another way of analysing and visualising the persistence module is via persistence landscapes defined in [11]. In the following fix $n \in \mathbb{N}$ and again assume $\mathcal{K}$ is a simplicial complex equipped with a filtration generated by a function $f\colon \mathcal{K} \to \mathbb{R}$. Recall that for $j \geq i$ it holds $H_n^{i,j} = \operatorname{im}\eta_n^{i,j}$ and hence $\beta_n^{i,j} = \dim(\operatorname{im}\eta_n^{i,j})$. In particular we get for $i \leq k \leq \ell \leq j$ that $\beta_n^{i,j} \leq \beta_n^{k,\ell}$ since

$$H_n^{i,j} = \operatorname{im}\left(\eta_n^{\ell,j} \circ \eta_n^{k,\ell} \circ \eta_n^{i,k}\right).$$

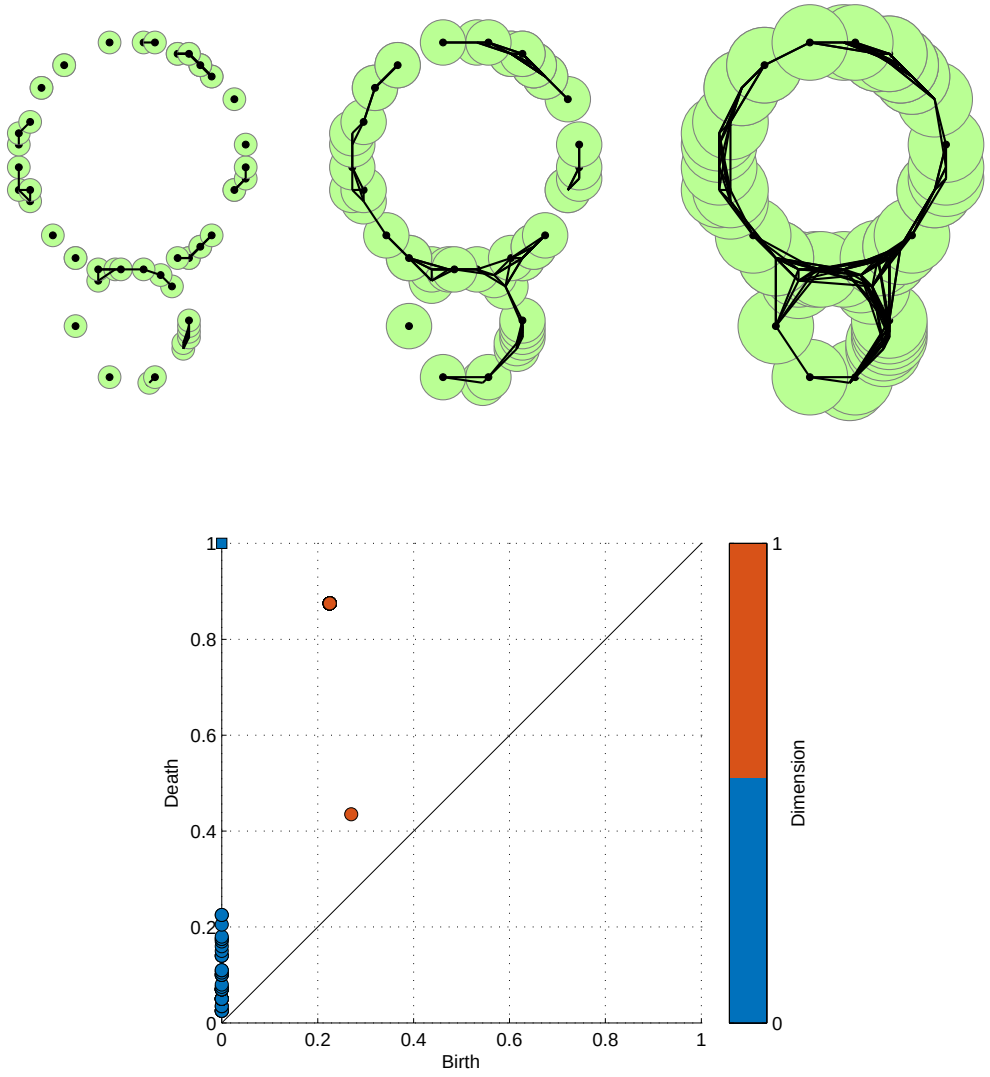


Figure 9: Three Vietoris-Rips complexes created from a point cloud uniformly drawn from two circles, only 1-simplices are drawn due to visibility (top). The corresponding persistence diagram where the radii increasingly vary from 0 to 1 (bottom).

Let $a_1, \dots, a_N$ denote the finitely many function values of f. We set $\beta_n^{a,b} = \dim H_n^{i,j}$ where $a \in [a_i, a_{i+1})$ and $b \in [a_j, a_{j+1})$. Then we define the rank function by

$$\mathbb{R}^2 \to \mathbb{R} \colon (a,b) \mapsto \begin{cases} \beta_n^{a,b}, & \text{if } a \leq b\,, \\ 0\,, & \text{otherwise.} \end{cases}$$

Changing the coordinates by

$$m = \frac{a+b}{2} \qquad \text{and} \qquad h = \frac{b-a}{2}$$

gives us the rescaled rank function

$$\mathbb{R}^2 \to \mathbb{R} \colon (a,b) \mapsto \begin{cases} \beta_n^{m-h,m+h}, & \text{if } h \geq 0\,, \\ 0\,, & \text{otherwise.} \end{cases}$$

Definition 3.22 (**Persistent Landscape**)
The *persistent landscape* is a sequence $(\lambda_k \colon \mathbb{R} \to [-\infty, \infty])_{k \in \mathbb{N}}$ where

$$\lambda_k(t) = \sup(\{m \geq 0 \,|\, \beta_n^{t-m,t+m} \geq k\})\,.$$

Remark. 1) Note that if $0 \leq m_1 \leq m_2$ holds, it follows $\beta_n^{t-m_2,t+m_2} \leq \beta_n^{t-m_1,t+m_1}$ for every $t \in [-\infty, \infty]$.

2) One can show that λ_k is 1-Lipschitz continuous for every $k \in \mathbb{N}$. See [11, Appendix].

3) Every persistence landscape corresponds to exactly one persistence diagram and vice versa. ◇

Example 3.23. Figure 10 shows the rank and rescaled rank function corresponding to a persistence diagram along with its persistence landscape. ◇

Since every persistence landscape is a function $\lambda \colon \mathbb{N} \times \mathbb{R} \to [-\infty, \infty]$ we can define a norm on the set of persistence landscapes by using the p-norm on $\mathbb{N} \times \mathbb{R}$ induced by the product measure of the counting measure and Lebesgue measure

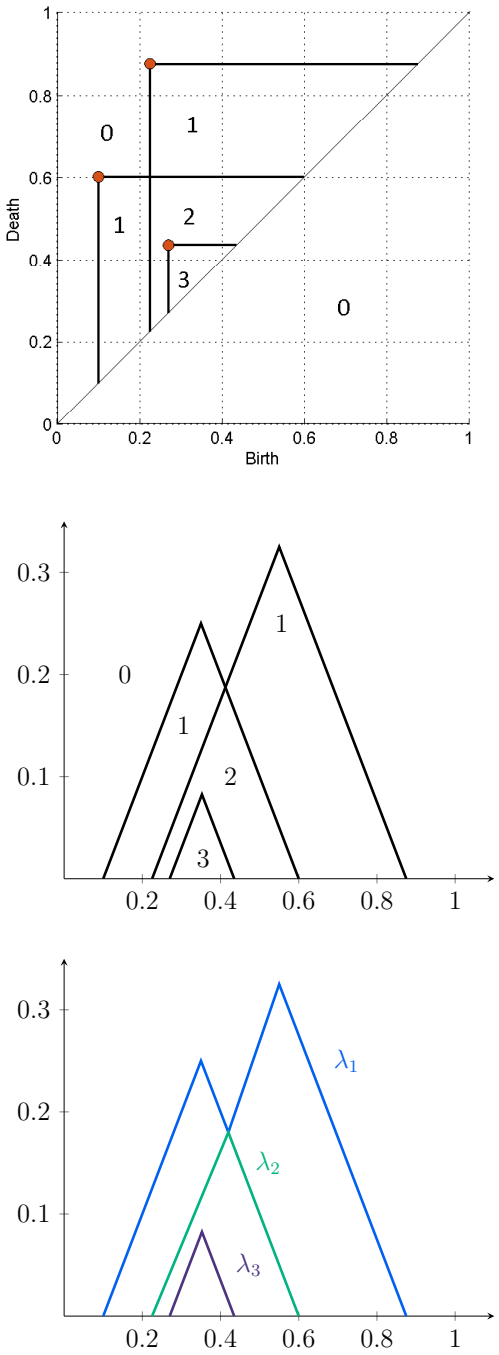


Figure 10: A persistence diagram and its corresponding rank function (top), rescaled rank function (middle) as well as the affiliated persistence landscape (bottom). The values of the rank functions are given in the specific region.

on $\mathbb{N}$ and $\mathbb{R}$ respectively for $1 \leq p < \infty$, i. e.

$$\|\lambda\|_p^p = \sum_{k=1}^{\infty} \|\lambda_k\|_p^p = \sum_{k=1}^{\infty} \int_{\mathbb{R}} |\lambda_k(t)|^p \, \mathrm{d}t$$

and since $L^p(\mathbb{N} \times \mathbb{R})$ is a Banach space, this will enable us to apply the theory of Banach space valued random variables. We therefore will quickly repeat the main notions of probability in Banach spaces.

Banach Space Valued Random Variables Let $\mathcal{B}$ be a real separable Banach space with norm $\|\cdot\|$ and $(\Omega, \mathcal{F}, \mathbb{P})$ a probality space. Let $X \colon \Omega \to \mathcal{B}$ be a Borel random variable. Then the compositions $\|X\| \colon \Omega \to \mathbb{R}$ as well as $f(X) \colon \Omega \to \mathbb{R}$, where $f \in \mathcal{B}^*$ and $\mathcal{B}^*$ denotes the topological dual, are real-valued random variables. An element $\mathbb{E}X \in \mathcal{B}$ is called Pettis integral of X if

$$\mathbb{E}f(X) = \int_{\Omega} f\Big(X(\omega)\Big) \, \mathrm{d}\mathbb{P}(\omega) = f(\mathbb{E}X)$$

holds for every $f \in \mathcal{B}^*$. One can show that if $\mathbb{E}\|X\| < \infty$ then X has a Pettis integral and $\|\mathbb{E}X\| \leq \mathbb{E}\|X\|$. We call the set of expectations

$$\mathbb{E}\Big[\Big(f(X) - \mathbb{E}f(X)\Big)\Big(g(X) - \mathbb{E}g(X)\Big)\Big]$$

for $f, g \in \mathcal{B}^*$ the covariance structure of X. Furthermore there exist versions of the Strong Law of Large Numbers and the Central Limit Theorem for Banach spaces.

Theorem 3.24 (**Strong Law of Large Numbers**)
Let $X_1, \ldots, X_n$ be i. i. d. Banach space valued random variables with $\mathbb{E}\|X_1\| < \infty$. Set $S_n = \sum_i X_i$ then $\frac{1}{n} S_n \to \mathbb{E}(X_1)$ almost surely.

Theorem 3.25 (**Central Limit Theorem**)
Let $X_1, \ldots, X_n$ be i. i. d. $L^p(\mathbb{N} \times \mathbb{R})$ valued random variables for $2 \leq p < \infty$ and $\mathbb{E}X_1 = 0$ as well as $\mathbb{E}\|X_1\|^2 < \infty$ then $\frac{1}{\sqrt{n}} S_n$ converges weakly to a Gaussian random variable G with the same covariance structure as X, i. e.

$$\lim_{n \to \infty} \mathbb{E}\, \varphi\left(\frac{1}{\sqrt{n}} S_n\right) = \varphi(G)$$

for every $\varphi \in L^p(\mathbb{N} \times \mathbb{R})^*$.

Proof. See [12]. ■

Now lets return to persistence landscapes. Let X denote a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $X(\omega)$ for $\omega \in \Omega$ is the given data and let Λ be its corresponding persistence landscape, i. e. $\Lambda\colon \Omega \to L^P(\mathbb{N} \times \mathbb{R})$ such that $\Lambda(\omega) = \lambda(X(\omega))$ for every $\omega \in \Omega$ where λ denotes the persistence landscape of the data $X(\omega)$. We shortly write λ for $\lambda(X(\omega))$ in the following. Let $X_1, \dots, X_n$ be i. i. d. random variables and $\Lambda^1, \dots, \Lambda^n$ their corresponding persistence landscapes. Then the mean landscape $\overline{\Lambda}^n = \frac{1}{n}\sum_{i=1}^n \Lambda^i$ is given pointwise, meaning

$$\overline{\Lambda}^n(\omega)(k,t) = \overline{\lambda}^n(k,t) = \frac{1}{n}\sum_{i=0}^{n} \lambda^i(k,t)$$

where $k \in \mathbb{N}$ and $t \in \mathbb{R}$. Applying the Strong Law of Large Numbers (SLLN) and the Central Limit Theorem for Banach space valued random variables now yields.

Theorem 3.26 (**SLLN for Persistence Landscapes**)
In the above situation it holds $\overline{\Lambda}^n \to \mathbb{E}\Lambda_1$ *almost surely if* $\mathbb{E}\Lambda_1 < \infty$.

Theorem 3.27 (**Central Limit Theorem for Persistence Landscapes**)
In the above situation let $p \geq 2$ *and* $\mathbb{E}\|\Lambda_1\| < \infty$. *Additionally let* $\mathbb{E}\|\Lambda_1\|^2 < \infty$ *then* $\sqrt{n}(\overline{\Lambda}^n - \mathbb{E}\Lambda_1)$ *converges weakly to a Gaussian random variable with the same covariance structure as* Λ_1.

Corollary 3.28. *In the above situation let* $p \geq 2$ *and* q *such that* $\frac{1}{p} + \frac{1}{q} = 1$. *For* $f \in L^q(\mathbb{N} \times \mathbb{R})$ *set* $Y_i = \|f\Lambda_i\|_1$ *for* $i = 1, \dots, n$. *If* $\mathbb{E}\|\Lambda_1\| < \infty$ *and* $\mathbb{E}\|\Lambda_1\|^2 < \infty$, *then* $\sqrt{n}(\overline{Y}^n - \mathbb{E}Y_1)$ *converges in distribution to a normally distributed random variable with mean zero and the same variance as* Y_1.

Proof. Since for $Z_i = \Lambda_i - \mathbb{E}\Lambda_i$ for $i = 1, \dots, n$ the Central Limit Theorem holds in $L^p(\mathbb{N} \times \mathbb{R})$, for $g \in L^p(\mathbb{N} \times \mathbb{R})^*$, $g(Z_i)$ fulfills the Central Limit Theorem in $\mathbb{R}$ and hence converges in distribution to a normally distributed random variable with mean zero and variance $\mathbb{E}(g(Z_1)^2)$. Fixing $f \in L^q(\mathbb{N}\times\mathbb{R})$ and setting $g(h) = \|hf\|_1$ then implies the statement. ■

Remark. 1) In the situation of the above corollary we set

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y}^n)^2$$

then we get for $\alpha \in (0,1)$ the $(1-\alpha)$ confidence interval for $\mathbb{E}Y_1$ by

$$I_\alpha(X_1,\dots,X_n) = \left[\overline{Y}^n \pm \Phi_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}\right]$$

where $\Phi_{1-\frac{\alpha}{2}}$ denotes the $1-\frac{\alpha}{2}$ quantile of the standard normal distribution.

2) Let $X_1,\dots,X_n$ be i. i. d. and $X'_1,\dots,X'_{n'}$ be i. i. d. random variables which satisfy the conditions of Corollary 3.28 and let Y_i and Y'_i be accordingly as above. Setting $\mu = \mathbb{E}Y_1$, $\mu' = \mathbb{E}Y'_1$ as well as $s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ and $s_{Y'}^2$ similarly one can use the two-sample t-test in order to test the null hypothesis that $\mu = \mu'$ since the random variables are asymptotically normally distributed. The test statistic is hereby given by

$$T(Y_1,\dots,Y_n,Y'_1,\dots,Y'_{n'}) = \frac{\overline{Y}^n - \overline{Y'}^n}{\sqrt{\frac{s_Y^2}{n} + \frac{s_{Y'}^2}{n'}}}$$

which is student t distributed with $n+n'-2$ degrees of freedom under the null hypothesis.

3) Typical choices of a function f in Corollary 3.28 are

$$f_1(k,t) = \begin{cases} 1, & \text{if } t \in [-B,B] \text{ and } k \le K\,, \\ 0, & \text{else} \end{cases}$$

or by setting

$$f_2(k,t) = \begin{cases} \frac{1}{k^r}, & \text{if } t \in [-B,B] \text{ and } k \le K\,, \\ 0, & \text{else} \end{cases}$$

for some $r > 1$ if we assume that every given persistence landscape has its

support in $\{1, \dots, K\} \times [-B, B]$. The first one yields $\|f_1\Lambda\|_1 = \sum_{k=1}^{K} \|\lambda_k\|_1$ while for the latter it holds $\|f_2\Lambda\|_1 = \sum_{k=1}^{K} \frac{1}{k^r} \|\lambda_k\|_1$. ◇

Again we close this subsection with a stability result similar to Theorem 3.21 given in [11, Theorem 16].

Theorem 3.29 (**Landscape Stability Theorem**)
Let X be a triangulable, compact metric space that implies bounded degree-k total persistence for $k \geq 1$ and let f, g be tame Lipschitz continous functions. Let λ_f and λ_g denote the persistence landscapes associated with the persistence diagram generated by f and g, respectively. Then for all dimensions $n \in \mathbb{N}$ and $p \geq k$ it holds

$$\|\lambda_f - \lambda_g\|_p \leq C \|f - g\|_\infty^{p-k}$$

for some constant $C \geq 0$ depending on X, f, g and k.

3.7 Weighted Silhouettes

As a last method in order to analyse the topology of the data we discuss weighted silhouettes given in [13]. Assume a persistence diagram with off diagonal points $(b_j, d_j)_{j=1}^{N}$ for $N \in \mathbb{N}$. Then we set

$$\Lambda_{p_j}(t) = \begin{cases} t - b_j, & \text{if } t \in \left[b_j, \frac{b_j + d_j}{2}\right] \\ d_j - t, & \text{if } t \in \left(\frac{b_j + d_j}{2}, d_j\right] \\ 0, & \text{else} \end{cases}$$

where $p_j = (b_j, d_j)$ for $j = 1, \dots, N$. In Figure 10 the hat functions in the middle plot correspond to $\Lambda_{p_1}, \Lambda_{p_2}$ and Λ_{p_3} where p_1, p_2, p_3 denote the three points in the persistence diagram. Using the above functions we can define the persistence landscapes of the previous subsection also as

$$\lambda(k, t) = \text{kmax}_{i=1,\dots N}\, \Lambda_{p_i}(t)$$

where kmax is the kth largest value. We want to define a function similar to the persistence landscape which additionally weighs the persistence of the points in

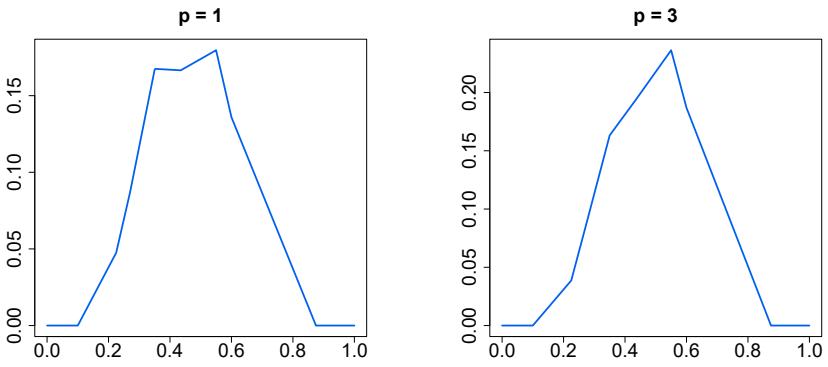


Figure 11: Two power-weighted silhouettes for $p = 1$ (left) and $p = 3$ (right) corresponding to the persistence diagram in Figure 10. The silhouettes were computed using the R package TDA.

the persistence diagram.

Definition 3.30 (**Power-Weighted Silhouette**)
Let $0 < p < \infty$ then we define in the above situation the *power-weighted silhouette* as

$$\phi^{(p)}(t) = \frac{\sum_{j=1}^{N} |d_j - b_j|^p \Lambda_{p_j}(t)}{\sum_{j=1}^{N} |d_j - b_j|^p}$$

for $t \in \mathbb{R}$.

Remark. 1) The parameter p weighs the persistence of each point. When p is small the silhouette puts more emphasis on points with lower lifespan while a larger p gives more attention to points with higher persistence.

2) One can show stronger stochastical convergence results like in the previous section holding for weighted silhouettes as well as persistence landscapes. For the details see [13]. ◇

Example 3.31. In Figure 11 we see two power-weighted silhouettes corresponding to the persistence diagram in Figure 10 computed with the R package TDA version 3.4.1. As we can see the right silhouette puts more focus on the features with higher persistence since the value of p is higher. ◇

3.8 Topological Data Analysis for Image Data

We now want to apply topological data analysis on greyscale images or their higher-dimensional analogs like 3D scans. Due to the structure of such an image one adapts the definition of homology groups in order to increase performance. This yields the definition of cubical cell homology. We hereby briefly discuss [14]. A detailed construction of the cubical homology can be found in [15, Chapter 2].
For $k \in \mathbb{Z}$ we call the unit interval $[k, k+1]$ a non-degenerate elementary interval and $[k, k]$ a degenerate interval. Let $d \in \mathbb{N}$. A cube in $\mathbb{R}^d$ is a product of d elementary intervals. Its dimension is defined as the number of non-degenerate intervals in the product. One calls 0, 1, 2 and 3-dimensional cubes as vertices, edges, squares and voxels, respectively. Given two cubes C_1 and C_2 we call C_2 a face of C_1 if $C_2 \subseteq C_1$. The boundary of a d-dimensional cube is the set of all $(d-1)$-dimensional faces and again the dth boundary operator of a cube is defined as the modulo 2 sum of its boundary elements. A collection of cubes of dimension at most d which is closed under taking faces and intersections is called a d-dimensional cubical complex. One now uses cubical complexes in order to describe the given image data which consequently allows to circumvent the triangulation and therefore reduces the size of the complex drastically. A filtration is created by using sublevel sets of the function f which assigns each pixel of the given image its corresponding grey value. These grey values are interpreted as the values of vertices of a complex. Now we extended f on edges, squares and so on, by defining its value as the maximum grey value of its faces, where an edge is defined as two adjacent pixels, a square consists of four neighbouring pixels, etc. This ensures that each cube is added to the filtration after all of its faces. Hence this construction yields a filtration

$$\mathcal{K}^0 \subseteq \mathcal{K}^1 \subseteq \ldots \subseteq \mathcal{K}^{255}$$

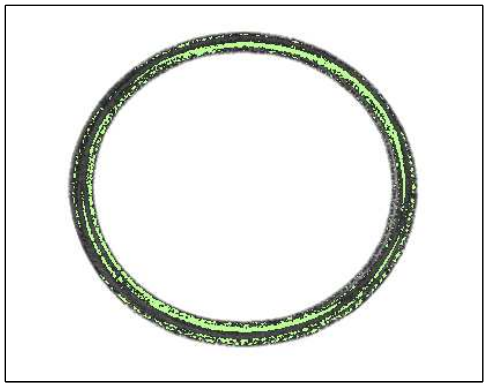
where $\mathcal{K}^n = f^{-1}((-\infty, n]) = f^{-1}([0, n])$. This filtration is called the *lower-star filtration of* f. After sorting the complex in ascending order of its function values one again creates a sorted boundary matrix and uses reduction to compute the persistence homology, see Subsection 3.4. Again we can interpret the Betti numbers as the number of components, holes, voids and so on in the data.

Example 3.32. In Figure 12 we see an example of the above construction applied on a greyscale image of an annulus, 12a. By subsequently increasing the greyscale value of the sublevel sets of f, more and more pixels are added to the complex, as we can see in 12b–12d, where the added pixels are colored green if there grey levels are smaller or equal than 20, 35 and 57 respectively. As we can see in the corresponding persistence diagram of the 1-dimensional features 12e, a lot of holes arise in this process. But due to their low persistence we can interpret most of them as noise. While clearly the homology class born at 57 and vanishing at 255 sticks out of the others. This class represents the hole of the annulus which we can see in 12d arises at a level of 57 and dies at the very end of the filtration by adding the white pixels with a value 255 to the cubical complex. Thus our geometrical interpretation of the Betti numbers is still reasonable. The diagram was computed with the C++ software package DIPHA. ◇

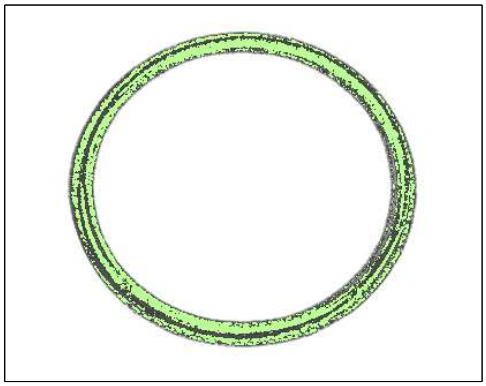
(a) Greyscale image of an annulus.

(b) The annulus for which each pixel where $f \leq 20$ holds, is colored green.

(c) The annulus for which each pixel where $f \leq 35$ holds, is colored green.



(d) The annulus for which each pixel where $f \leq 57$ holds, is colored green.



(e) Persistence diagram of the lower-star filtration of the annulus.

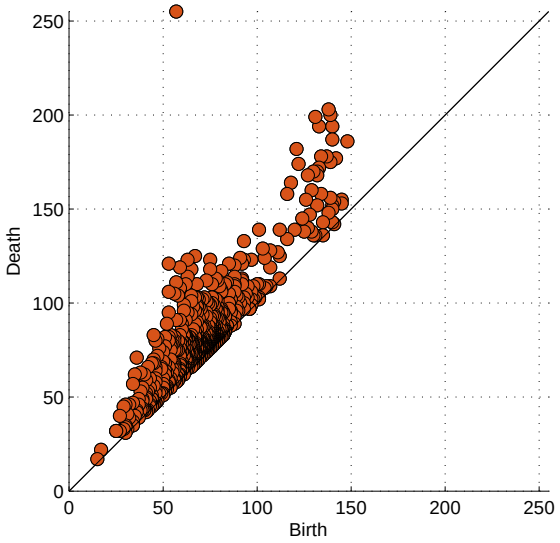


Figure 12: Visualisation of a part of the lower-star filtration and the corresponding persistence diagram of the greyscale image of an annulus 12a.

4 Application to Fibrin Nets

4.1 Fibrin Nets

Fibrin is a protein which is created from Fibrinogen. Its main task is to cause blood clotting in order to stop bleeding. Therefore, given a severly injured patient who needs volume replacement, it is important to maintain a natural fibrin net structure to enhance chances of survival. A more sophisticated and brief explanation is given in [16]: "Fibrinogen is a large protein synthesized by the liver that makes up some 5% of total plasma protein. Various stimuli, including injury, activate a sequence of clotting factors that cause soluble fibrinogen to be converted enzymically to fibrin, an insoluble polymer matrix that is the structural component of blood clots. Raised fibrinogen levels indicate an increased risk of ischemic heart disease and stroke. It is uncertain whether fibrinogen is a risk factor that causes cardiovascular disease (CVD) or a marker of developing disease. Fibrinogen is an acute-phase reactant, and the circulating level increases in response to infection, chronic inflammation, smoking, and other environmental stressors."

4.2 Approach and Results

We now want to apply topological data analysis to 3D greyscale images of fibrin nets in order to find out how two different dilutions effect the topology of the nets. To do so, we observe 3D greyscale images of fibrin nets of four different pigs indexed by 12, 14, 19 and 24. For each pig we provide an image of its natural fibrin net, called the baseline, and one of its diluted net. The blood of the pigs 12 and 19 were diluted with Hextend by Biotime[7] whereas pigs number 14 and 24 were treated with Gelofusin by Braun[8].

First of all, image processing was conducted in MATLAB Version R2017a by applying 3D box filtering for noise reduction followed by histogram equalization to improve contrast and guided filtering to perform edge-preserving smoothing, see Figure 13. Afterwards the persistence diagrams were computed as described in 3.8 using the C++ software package DIPHA. The characteristics of the topological

[7] `http://www.biotimeinc.com`
[8] `http://www.bbraun.com`

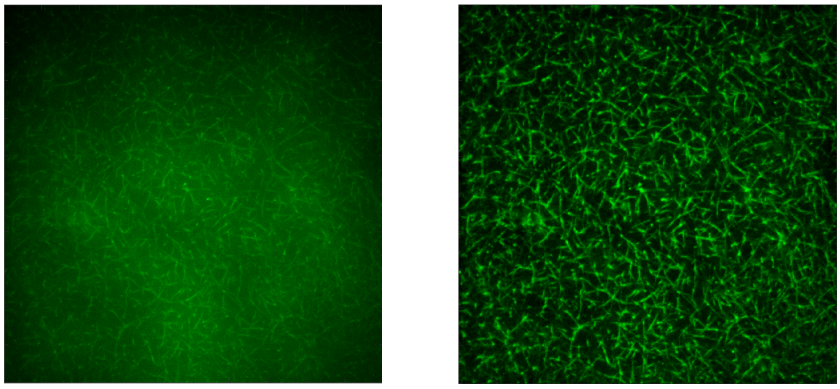


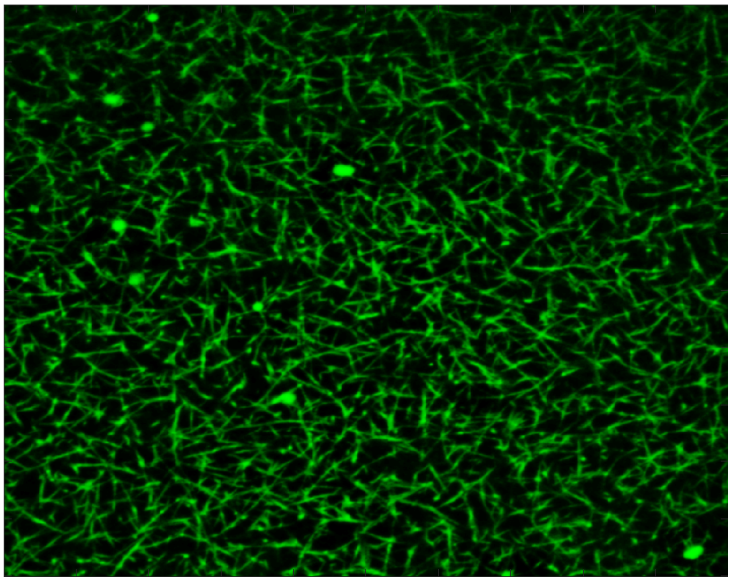
Figure 13: Left: The original image of a fibrin net. Right: The same image but denoised.

features are depicted by weighted silhouettes for $p = 20$ computed using the TDA package in R version 3.4.1. Those silhouettes give a summary of how distinct the most distinguished features are; the larger the area under the silhouette (AUC), the more distinct features are present. Therefore, we provide the relative change in AUC from baseline in the silhouette plots. In addition, given the persistence diagrams of the baseline and diluted net, we provide the percentage of the 0.05% most persistent features of both diagrams together compared to the baseline to assess the change in the topological features. This index will be called *Feature* and will be added to the plot of the silhouettes as well.

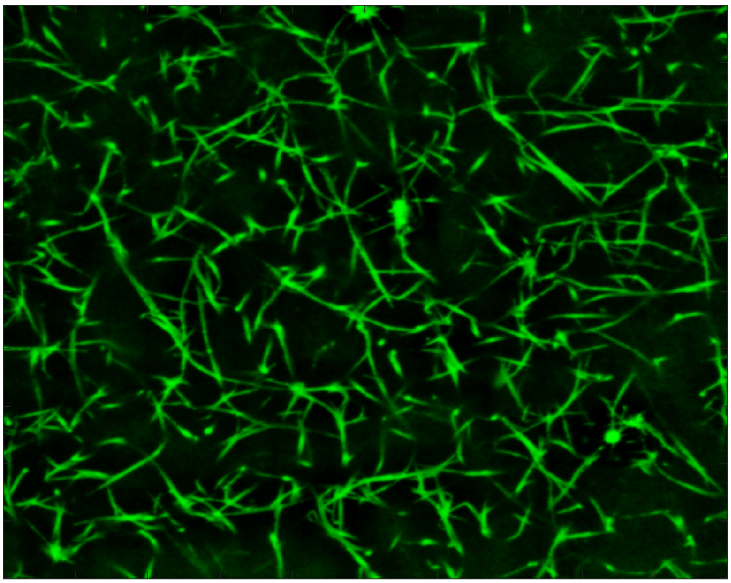
In Figure 16 we see the baseline of pig 12 and its Hextend diluted net. Figure 15 shows the corresponding persistence diagrams where again blue and orange points represent components and holes, respectively, while yellow points depict voids. Figure 16 shows the corresponding weighted silhouettes, where the black line represents the baseline and the red one corresponds to the dilution. Just by looking at the images of the nets we note quite some differences. While the baseline consists of a vast amount of thin filaments, the structure under dilution is distinctively coarser. The number of strings has drastically reduced, while their thickness increased considerably which can be interpreted as as more stiff and unflexible net than the natural one. Let us therefore observe whether the diagrams in Figure 15 and the weighted silhouettes in Figure 16 capture those differences. First we observe that looking at the diagrams both seem to have a similar structure within their components. This seems to be true when observing the most persistent

components represented by the feature rate. A feature rate of 59.56% shows that the number of high persistent components is quite balanced. However the area under the the silhouettes ascends by approximately 150%. Hence the diluted fibrin network contains more components with lower persistence. Looking at the one-dimensional features we see in the diagram that the high persistence homology classes from the dilution outnumber the classes of the baseline. This is emphasized by a higher feature rate of nearly 71%. However observing the silhouettes and the AUC change, we realise that there is only a small difference. But this again can be interpreted properly. While the holes in the diluted net have a much higher persistence, which could be due to their greater size, the holes in the natural blood sample have a shorter lifespan but due to the fine structure of the net, considerably more classes are born implying larger values of the silhouette. As final characteristic we investigate the voids encapsulated within the nets. While in the persistence diagram we don't recognise a significant difference, keep in mind that the plotted diagram does not visualise the multiplicity of the points, we note that the silhouttes vary significantly from one another. The area under curve increases by about 270% and ancillary the feature rate is nearly 97%. This high rate of long persisting features also explains why the AUC change is so enormous. The reason for the sudden change of existing and persisting voids can again be explained by the thickness of the filaments. Those thick strings encapsulate voids within while the thin threads in the baseline don't have room for such voids. Due to the drastic differences in the voids, it looks like especially those two-dimensional features correlated with the rigidy of a fibrin net.

Up next we evaluate the results concerning pig number 14, where Gelofusin was administered. Its nets are given in Figure 17, the corresponding persistence diagram in Figure 18 and the related silhouettes can be seen in Figure 19. Looking at the nets they both seem quite similar except for the big cluster in the left centre of the dilution. This cluster probably is a relict due to damaged plasma in the test specimen. The hope is therefore that this relict does not have a strong impact on our analysis, since the remaining structure of the nets looks nearly identical at least on the pictures. This similarity is visualised in the persistence diagram. Although there are more holes with higher persistence in the diluted net than in the baseline, one would say they look alike. However, observing the feature rates,

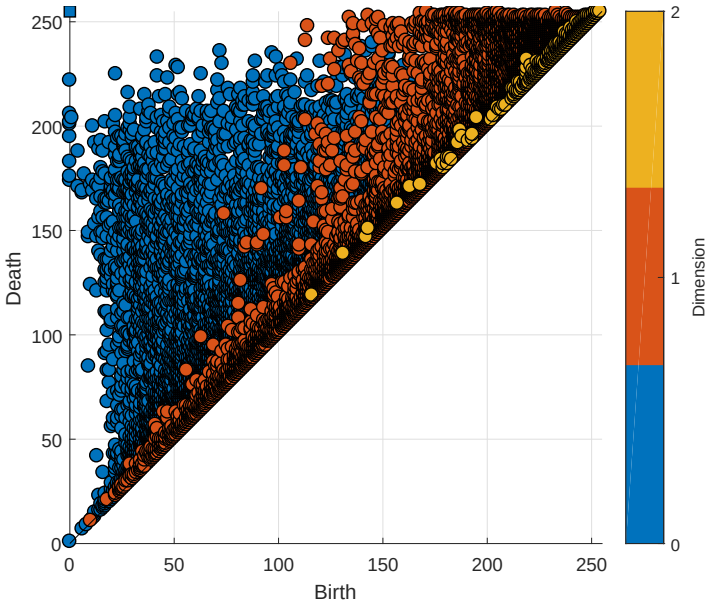


(a) Baseline of pig 12.

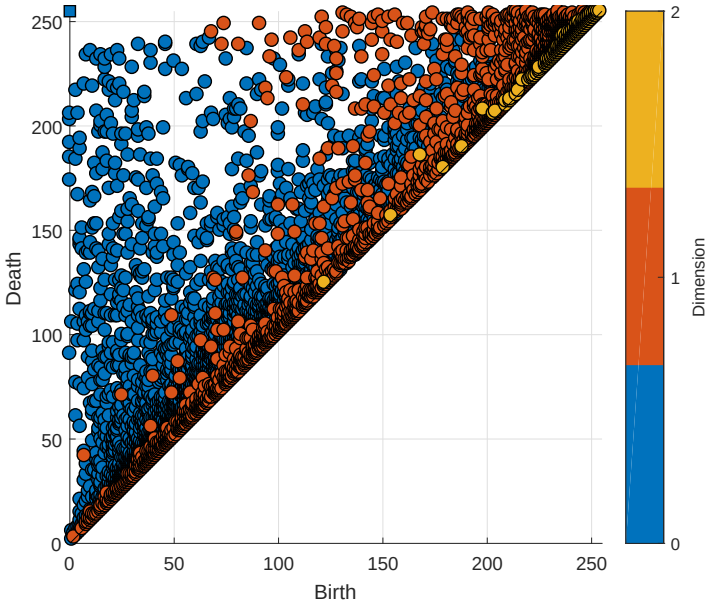


(b) Hextend diluted net of pig 12.

Figure 14: Fibrin nets of pig 12.



(a) Persistence diagram of the baseline of pig 12.



(b) Persistence diagram of the Hextend diluted net of pig 12.

Figure 15: Persistence diagrams of pig 12.

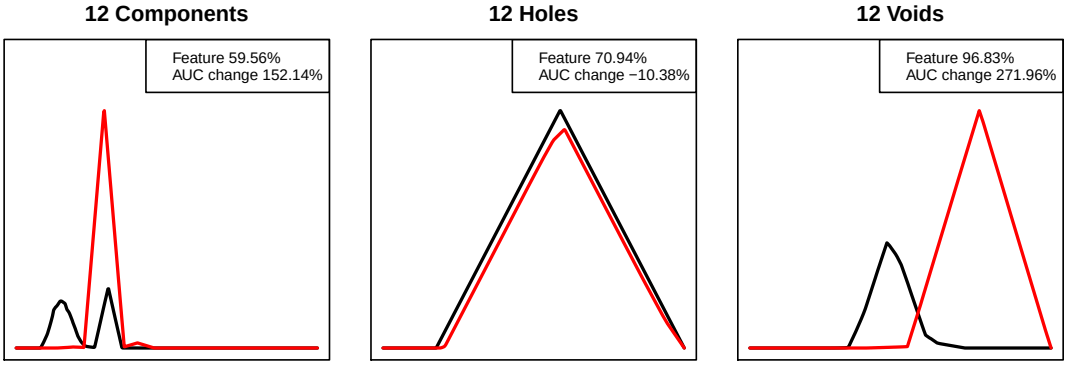
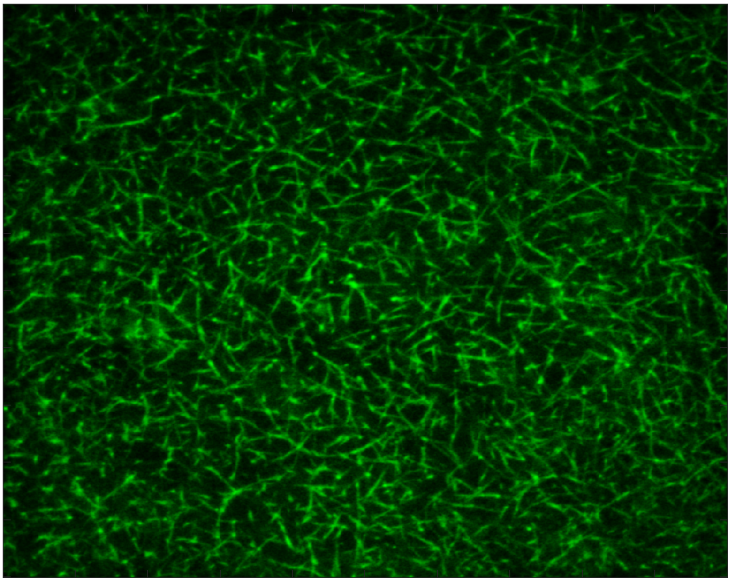


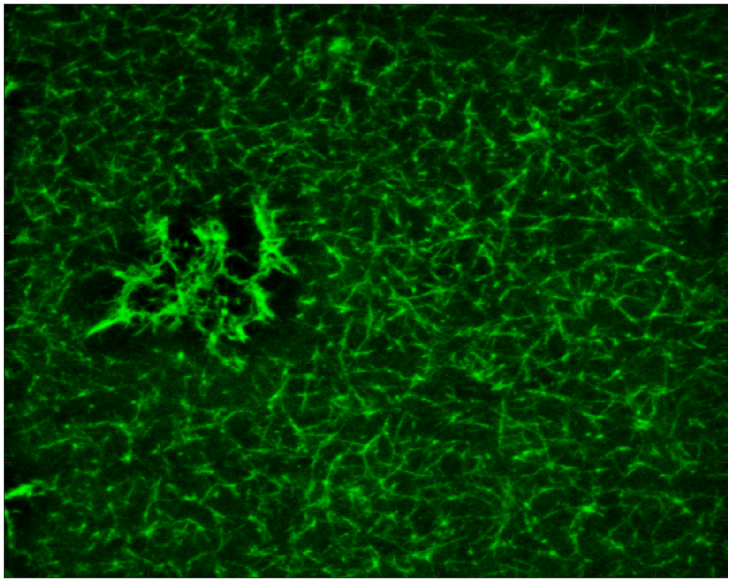
Figure 16: Weighted silhouettes corresponding to pig 12. The black line corresponds to the baseline while the red one indicates the diluted sample.

we note that tose seem to be quite little. But despite the low feature rates the AUC is relatively small and also the silhouettes themselves are comparable. We suspect that the low feature rates appear because of the cluster in the diluted net. This presumable relict obviously decreases the number of components and holes. And while the thickness of the strings within the cluster will surely encapsulate voids, wee see that the net has such a dense and fine structure outside of the cluster, even finer than the baseline of pig number 12, that also those fragile filaments probably enclose hollow spaces and therefore this cluster reduces the overall number of homology classes in each dimension. So putting those observations in contrast to the results of pig number 12, the net diluted by Gelofusin is more comparable to the baseline than the net under a Hextend dilution.

Lets move on to swine 19. As number 12 this pig has been treated with Hextend. Its nets are depicted in Figure 20, the persistence diagrams are plotted in Figure 21 and the affiliated silhouettes are shown in Figure 22. Looking at the images of the fibrin nets it again looks like Hextend changes the geometrical structure. While in the baseline the strings seem to be fragile and floating around separately, the dilution has thicker strings which appear to form circles. While the persistence diagrams show little difference, the silhouettes and feature rates encode enormous differences. For every dimension the feature rate is highly above 50%. In particular the feature rate regarding components lies at 98.18%. Hence the long persistent features appear with such a high multiplicity that they outnumber the ones in the baseline by far. Again this is not surprising. Surely also the baseline contains

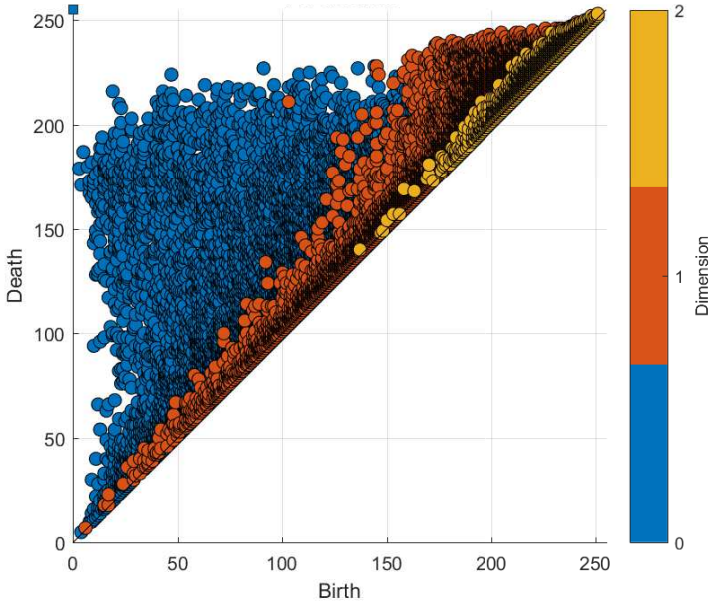


(a) Baseline of pig 14.

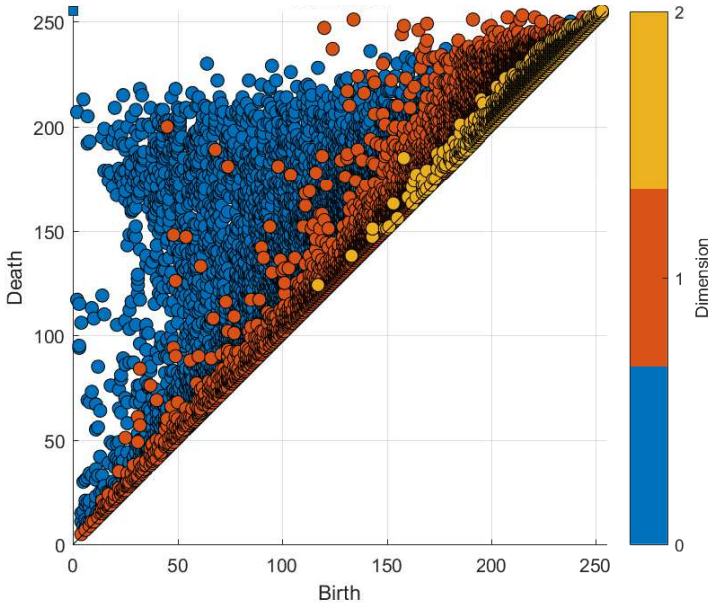


(b) Gelofusin diluted net of pig 14.

Figure 17: Fibrin nets of pig 14.



(a) Persistence diagram of the baseline of pig 14.



(b) Persistence diagram of the Gelofusin diluted net of pig 14.

Figure 18: Persistence diagrams of pig 14.

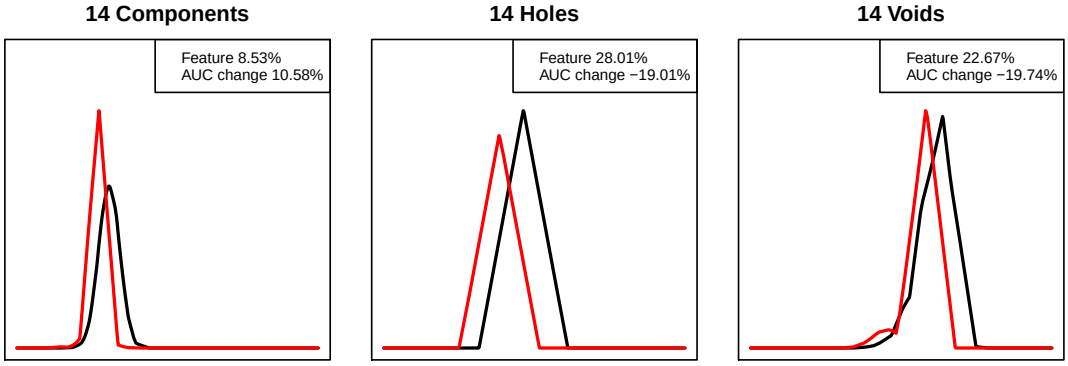
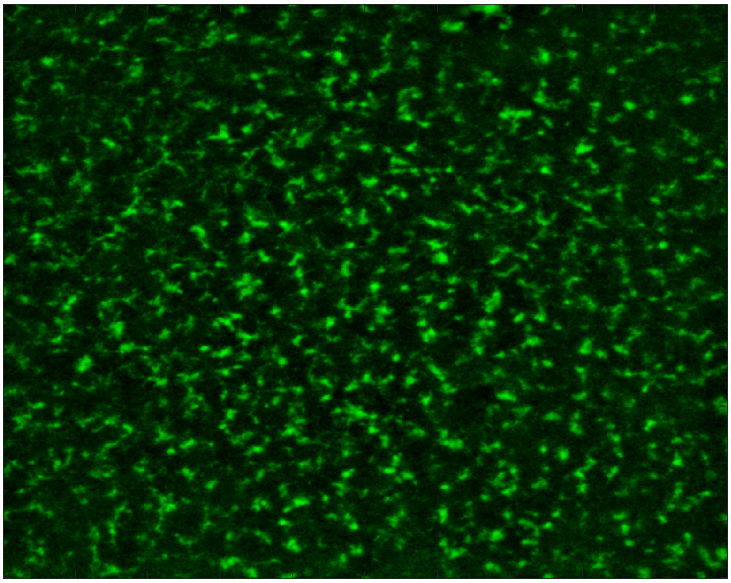


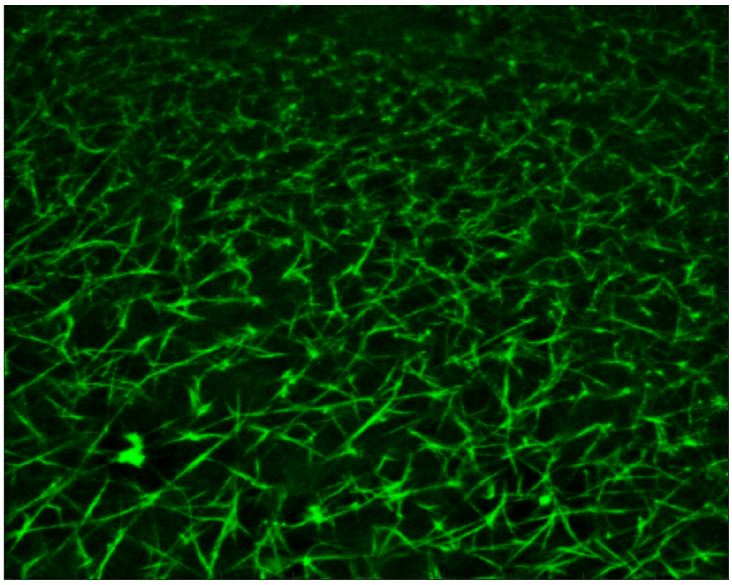
Figure 19: Weighted silhouettes corresponding to pig 14. The black line corresponds to the baseline while the red one indicates the diluted sample.

many components with high persistence due to its fine structure. But the thick pattern in the dilution generates even more long living components, also causing a high change in the area under the curve. Another tremendous discrepancy can be observed for one-dimensional classes. While the high persistent features of the Hextend diluted net exceed the ones of the baseline by far the AUC changes by nearly 1500%. So the quantity as well as the lifespans of holes in the dilution surpass the ones in the baseline dramatically, which can be again explained by the coarser structure which appears after adding Hextend. Additionally we also see the same changes, although not as enormous, for voids within the nets, i. e. a feature rate of 72.37% and a AUC change of 41.43%.

As final observation we take a look at pig number 24. This pig was treated with Gelofusin and once more we note that the corresponding fibrin nets look quite similar, see Figure 23. Also the persistence diagram in Figure 24 seem to resemble each other. Looking at the silhouettes, Figure 25, we note that the silhouettes of the components and holes also are similar. Therefore the AUC of the silhouettes corresponding to components only changes by approximately 25% despite the marginal feature rate of 2.59%. Observing the feature rate of the holes it follows that the number of high persistent one-dimensional features is nearly balanced and that the area below the curves changes by moderate 17%. Nevertheless the weighted silhouettes corresponding to voids differ substantially, although the feature rate is again relatively modest. Hence the high AUC change arises due to a higher number of voids within the diluted net. But despite the

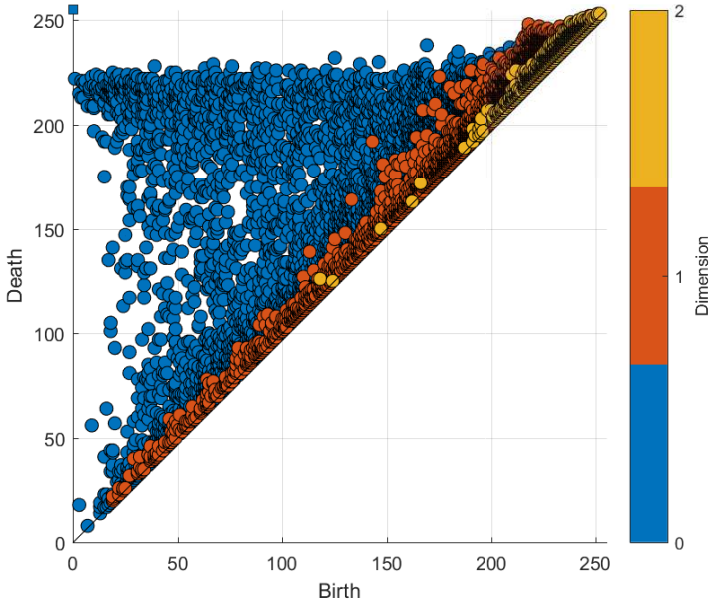


(a) Baseline of pig 19.

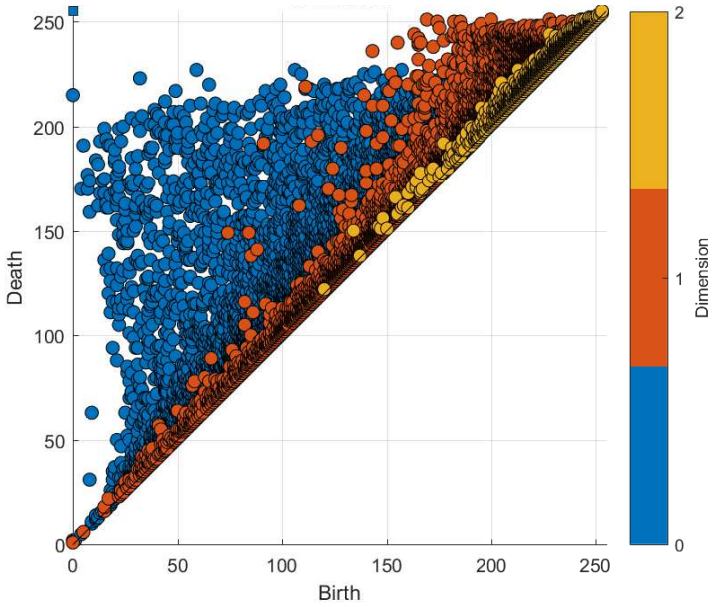


(b) Hextend diluted net of pig 19.

Figure 20: Fibrin nets of pig 19.



(a) Persistence diagram of the baseline of pig 19.



(b) Persistence diagram of the Hextend diluted net of pig 19.

Figure 21: Persistence diagrams of pig 19.

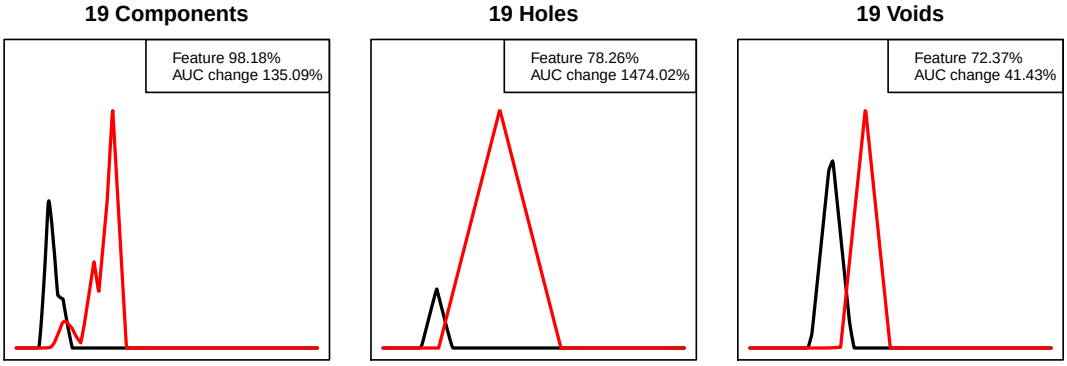
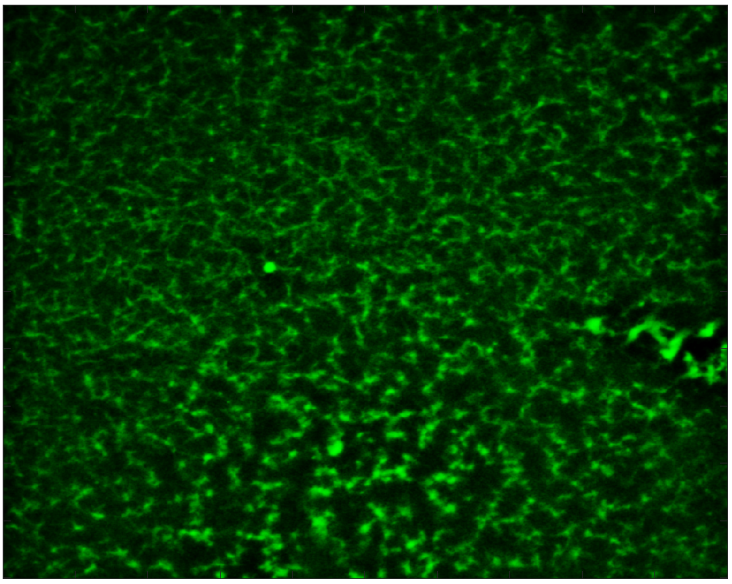


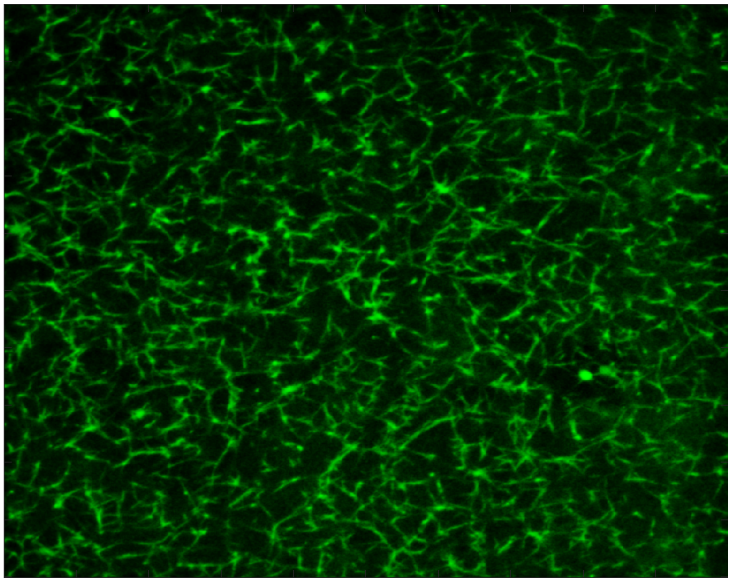
Figure 22: Weighted silhouettes corresponding to pig 19. The black line corresponds to the baseline while the red one indicates the diluted sample.

difference in this silhouette, comparing the overall results of pig 24 with the ones of pig 12 and 19, this still shows that the baseline and the dilution are more similar to each other than in the cases where Hextend was used.

Let us summarize our results. Applying topological data analysis to the fibrin networks of the four pigs yields a trend towards Gelofusin inducing a more natural fibrin net than Hextend. While Gelofusin delutions provide similar silhouettes, Hextend yields vast differences between the silhouettes, causing higher AUC change rates. Also the balance of high persistent one- and two-dimensional features and therefore most characteristic holes and voids seem to be shifted towards the diluted nets when using Hextend in comparison to pig number 24, for which the rate has been more ore less balanced. Surely pig number 14 which was treated with Gelofusin also provided unbalanced feature rates which clearly were tending towards the baseline, it is likely that this imbalance is caused by the relict of the blood plasma captured in the 3D image of the natural net.

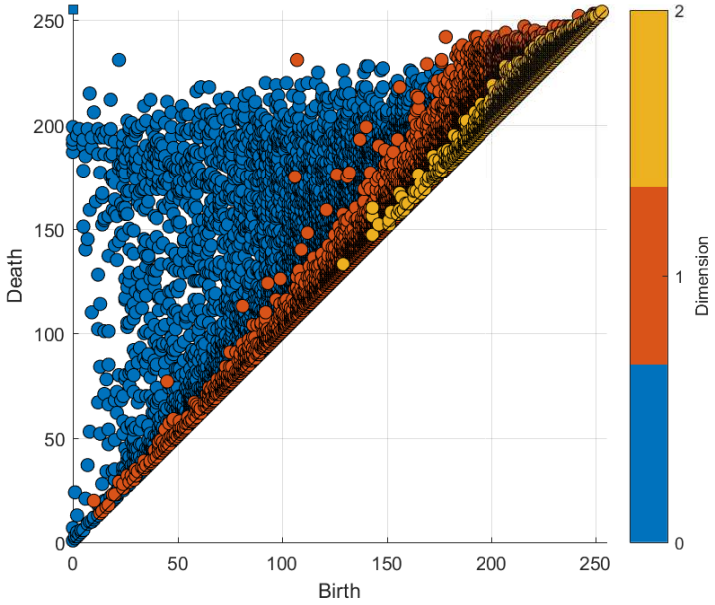


(a) Baseline of pig 24.

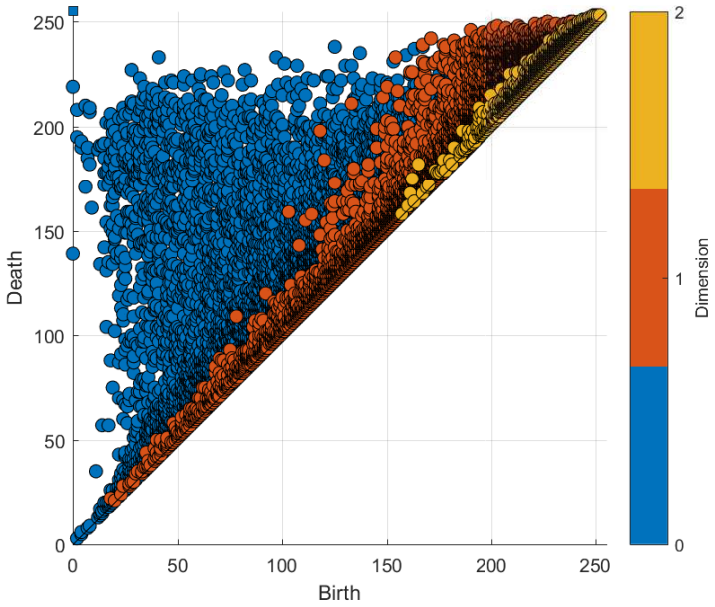


(b) Gelofusin diluted net of pig 24.

Figure 23: Fibrin nets of pig 24.



(a) Persistence diagram of the baseline of pig 24.



(b) Persistence diagram of the Gelofusin diluted net of pig 24.

Figure 24: Persistence diagrams of pig 24.

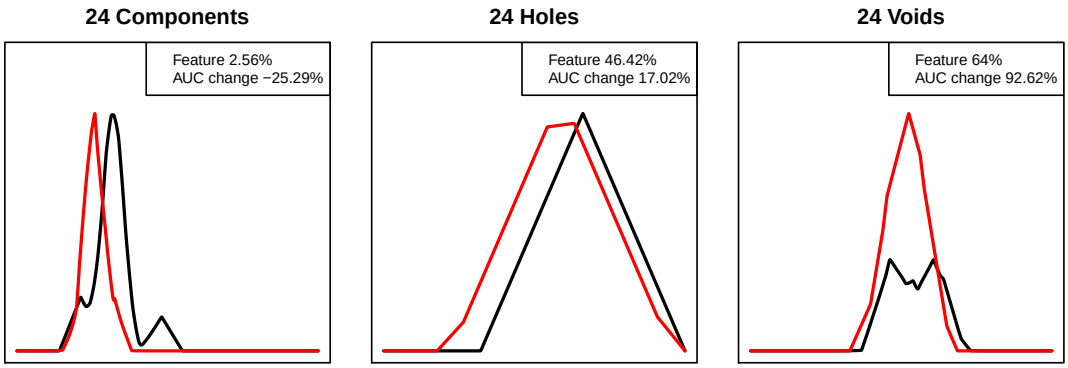


Figure 25: Weighted silhouettes corresponding to pig 24. The black line corresponds to the baseline while the red one indicates the diluted sample.

5 Conclusion

The aim of this thesis was to provide an introduction to the modern field of topological data analysis as well as applying this theory to real world data, namely 3D greyscale images of fibrin nets, which play an important role in blood clotting and henceforth are, for instance, important when it comes to severly wounded patients. We therefore gave a short introduction into algebraic topology and subsequently discussed the theoretical background of topological data analysis, that is persistent homology. Also an algorithm which computes the persistent homology groups was presented. Additionally we discussed established methods which are used to analyse the calculated results. This subsequently also lead us to stochastical convergence results which allow the computation of confidence intervals and hypothesis testing. We then showed how this theory is adapted in order to analyse high-dimensional greyscale images, which is achieved by applying the so-called lower-star filtration on the sublevel sets of the greyscale values of each pixel. Afterwards we started to use the previous techniques in order to analyse the effects of two dilutions on the geometrical structure of fibrin nets. We therefore analyzed fibrin nets caputred from blood samples of four pigs. Each providing a 3D greyscale image of its natural fibrin net and one of its diluted net. Using weighted silhouettes and the two cofficients Feature and AUC Change we stated the hypothesis that diluting with Gelofusin causes a more natural fibrin net while the second dilution Hextend causes a more rigid fibrin net. However since our sample size is so small, we cannot test our hypothesis. Hopefully this can be addressed in the future.

Overall this thesis shows that topological data analysis seems to be a promising method in modern data analysis. Despite the abstract algebraic methods it is built on, its geometrically motivated ideas allow to classify data in terms of their topological shape. Thus maybe enabling new perspectives when it comes to observing given data statistically. As Professor Gunnar Carlsson, one of the pioneers of topological data analysis, famously stated,

"Data has shape, and shape has meaning."

References

[1] Rotman, J. Joseph: *An Introduction to Algebraic Topology*. Springer-Verlag New York Inc., 1988.

[2] Rotman, J. Joseph: *Advanced Modern Algebra*. Prentice Hall, 2nd printing, 2003.

[3] Munkres, James: *Elements of Algebraic Topology*. Addison-Wesley, Reading, MA, 1984.

[4] Wang, G. Kairui: *The Basic Theory of Persistent Homology*. Participant Paper of the Research Experience for Undergraduates 2012 at the University of Chicago.

[5] Edelsbrunner H. & Harer J.: *Computational Topology: An Introduction*. American Mathematical Soc., 2010.

[6] Edelsbrunner H. & Harer J.: *Persistent Homology – a Survey*. Surveys on Discrete an Computational Geometry. Twenty Years Later, eds. J. E. Goodman, J. Pach and R Pollack, Conetmporary Mathematics 453, 257-282, Amer. Math. Soc., Providence, Rhode Island, 2008.

[7] Wang Y.: *Computational topology: Theory, algorithms, and applications to data analysis*. Lecture notes, Spring 2016, The Ohio State University.

[8] Morozov D.: *Persistence algorithm takes cubic time in worst case*. BioGeometry News, Dept. Comput. Sci., Duke Univ., 2005.

[9] Mileyko Y. & Mukherjee S. & Harer J.: *Probability measures on the space of persistence diagrams*. Inverse Problems, volume 27, December 2011.

[10] Cohen-Steiner D. & Edelsbrunner H. & Harer J. & Mileyko Y.: *Lipschitz Functions Have L_p-stable Persistence*. Foundations of Computational Mathematics,Volume 10, 127-139, 2010.

[11] Bubenik P.: *Statistical Topological Data Analysis using Persistence Landscapes*. Journal of Machine Leraning Research 16, 77-102, 2015.

[12] Hoffmann-Jørgensen J. & Pisier G.: *The Law of Large Numbers and the Central Limit Theorem in Banach Spaces*. The Annals of Probability, Vol. 4, No. 4, 587-599, 1976.

[13] Chazal F. & Fasy B. T. & Lecci F. & Rinaldo A. & Wasserman L.: *Stochastic Convergence of Persistence Landscapes and Silhouettes*. arXiv:1312.0308.

[14] Wagner H. & Chen C. & Vuçini E.: *Efficient Computation of Persistent Homology for Cubical Data*. Springer Berlin Heidelberg, 91-106, 2012.

[15] Kaczynski T. & Mischaikow K. & Mrozek M.: *Computational Homology*. Springer-Verlag New York, Inc, 2004.

[16] Brunner E.: *Encyclopedia of Stress (Second Edition)*. Elsevier Inc. 2007.

[17] Ghrist R.: *Barcodes: The Persistent Topology of Data*. Bull. Amer. Math. Soc. 45 (2008), 61-75.

[18] Netzer T.: *Algebraische Geometrie*. Lecture notes 2017, University of Innsbruck.

Affidavit

I, Martin Berger, hereby declare that this master's thesis has been written only by the undersigned and without any assistance from third parties. I confirm that no sources have been used in the preparation of this thesis other than those indicated in the thesis itself.
This master's thesis has heretofore not been submitted or published elsewhere, neither in its present form, nor in a similar version.

Signed: _____

Date: _____