ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D MATH

Mohnhaupt Mona
19-946-235

The Nerve Theorem and its Applications in Topological Data Analysis

Bachelor Thesis

Bachelor of Science ETH in Mathematics
Swiss Federal Institute of Technology (ETH) Zurich

Supervision

Dr. Sara Kališnik Hintz

June 19, 2023

Abstract

All complex data analysis is driven by mathematical models. Hence, advanced mathematical modeling can enable new insights into high dimensional data. The aim of this paper is to introduce mathematical theory coming from the field of algebraic topology, in particular the Nerve Theorem. I will provide a step by step proof of this important result, which guarantees homotopy equivalence between a topological space and its nerve under certain conditions. By introducing the computational method `Mapper` (17), I will illustrate the significance of the Nerve Theorem. `Mapper` is a useful tool in the field of Topological Data Analysis (*TDA*), extracting and visualizing characteristics from high dimensional data in the form of simplicial complexes. In the last chapter of this paper, I will present two biomedical applications of *TDA* and `Mapper`. The impact of the mathematical theory and computational methods introduced earlier, become clear through astonishing findings in breast cancer and diabetes research (11; 17).

Acknowledgements

I would like to give my warmest thanks to my supervisor Dr. Sara Kališnik Hintz. Her enthusiasm about mathematics is contagious and inspiring, and in combination with her guidance and advice, she carried me through all the stages of writing my thesis. I would also like thank my parents, my best friend Hannah and my friends for their continuous support with my project.

Contents

Acronyms and Abbreviations

TDA	Topological Data Analysis
Open ball with center at x and radius r	$B_r(x)$
Point cloud	Finite metric space
$\|X\|$	Cardinality of the set X
$[n]$	The set $\{1, \ldots, n\}$
S^1	The unit circle $\{(x, y) \mid x^2 + y^2 = 1\} \subset \mathbb{R}^2$

Chapter 1

Introduction

Today, data is being collected everywhere and in enormous amounts. In a high-dimensional and complex data context it is challenging to differentiate between significant information and noise, let alone to visualize the data. Mathematical methods enable us to strategically analyze the data and look for structures, relationships and hidden features.

Topological Data Analysis (TDA) presents one approach to tackle this challenge. In order to understand the mechanisms behind *TDA*, let us see what lies behind these three words. Topology is a branch of mathematics that studies the shape of objects. More specifically, topologists analyze properties of objects that are preserved when stretching, shrinking, rotating and deforming. These operations are so called *continuous deformations* and do not include tearing, cutting, ripping or puncturing. Using these topological concepts to study the shape of data, is the key idea of *TDA*. A simple example might help to understand the topologist's mind:

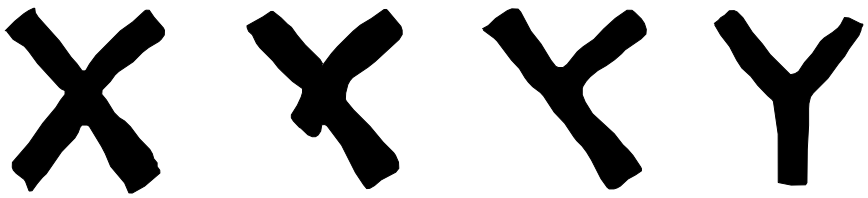


Figure 1.1: Example of a homotopy equivalence.

Figure 1.1 shows a deformation of the letter X into the letter Y. The left leg of the letter X is deformed and shrunken until only the right leg of X remains. Through bending this remaining leg we obtain the letter Y. In topology, one might not analyze those two objects separately, but rather deduce characteristics of the letter X from the letter Y, or vice versa, as they are *homotopy equivalent*. I will

discuss this in more detail in Section 2.1.

The mathematical theory presented in this paper originates from the field of algebraic topology, which uses algebraic methods to study topological problems. I will provide a step by step derivation of an important result; the Nerve Theorem. It enables us to analyze complex geometric structures in topological spaces through simpler abstract ones in form of *simplicial complexes*.

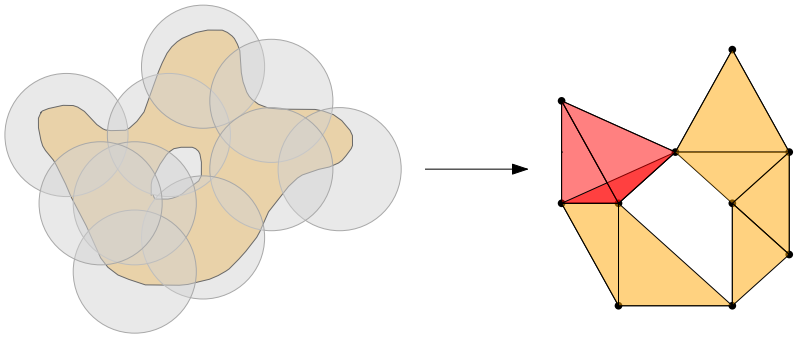


Figure 1.2: Intuition behind the Nerve Theorem.

Figure 1.2 illustrates the association of a topological space (orange shape) and its covering (gray disks) with a geometric simplicial complex (right figure) the Nerve Theorem. The nerve corresponding to the topological space depends on the chosen covering, as the intersections of the gray disks determine the simplices on the right. An intersection between two disks is represented by an edge, between three a triangle (orange) and between four disks a three dimensional simplex, a tetrahedron (red). As Figure 1.2 shows, the Nerve Theorem offers a simplified combinatorial depiction of complex spaces. Using this idea, the Nerve Theorem provides a foundation for `Mapper`, a computational method used to detect and visualize shapes in high-dimensional data sets that was introduced in 2007 by Singh et al. (17). `Mapper` is more than a visualization tool. Depending on the origin of the data set and the relationships one is interested in, the user can adapt multiple parameters in `Mapper` to try and find mathematically relevant structures and shapes.

This paper's aim is to provide an example of how pure mathematical theory and real-life application intertwine, and draw attention to the great potential of *TDA*. I will elaborate on the detection of breast cancer sub-types (11) and identification of different diabetes types (17) using this approach. Making use of advanced mathematical modeling in other sciences can yield powerful results and find patterns, structures, relationships and answers, that would have stayed hidden otherwise.

Chapter 2

Mathematical Background

The following two sections will be primarily based on the book 'Topological pattern recognition for point cloud data' written by Gunnar Carlsson (4) and the article 'A Unified View on the Functorial Nerve Theorem and its Variations' by Bauer, Kerber, Rolle and Roll (2). The main result of this chapter is the Nerve Theorem (Theorem 2.2.1), which provides a simplified representation of topological spaces.

2.1 An Introduction to Algebraic Topology

In preparation for the Nerve Theorem, some fundamental constructions from algebraic topology are needed. In this section, we will introduce the necessary mathematical theory accompanied by examples and illustrations that strengthen the reader's intuition.

Definition 2.1.1 (General position) Let $S = \{x_0, \ldots, x_k\} \subset \mathbb{R}^n$. Then S is said to be *in general position*, if it is not contained in any affine hyperplane of $\mathbb{R}^n$ of dimension d with $d < k$.

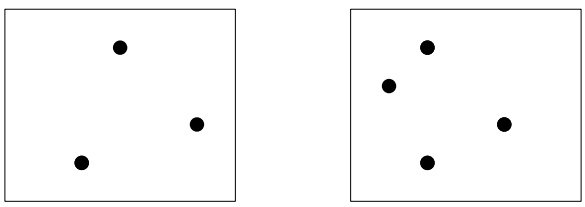


Figure 2.1: The left figure shows three points in $\mathbb{R}^2$ in general position, since the three points are contained in $\mathbb{R}^2$ but $k = 2$. The four points on the right are not in general position as they are also contained in $\mathbb{R}^2$ but $k = 3$.

Definition 2.1.2 (Simplex, face, vertex) Let $S = \{x_0, \ldots, x_k\} \subset \mathbb{R}^n$ be in general position. The *k-simplex spanned by S* is defined as the convex hull $\sigma = \sigma(S) \subset \mathbb{R}^n$. The points $x_i \in S$ are called *vertices* and the spanned simplices $\sigma(T)$ for $\emptyset \neq T \subseteq S$ are called *faces*.

Definition 2.1.3 (Geometric simplicial complex) Let χ be a finite collection of simplices, in a Euclidean space. We call χ a *geometric simplicial complex* if the following conditions hold:

1. For any simplex $\sigma \in \chi$ all faces of χ are contained in χ.
2. For any two simplices $\sigma, \tau \in \chi$ it holds $\sigma \cap \tau \in \chi$ is a simplex.

In the following we will be referring to geometric simplicial complexes as simplicial complexes.

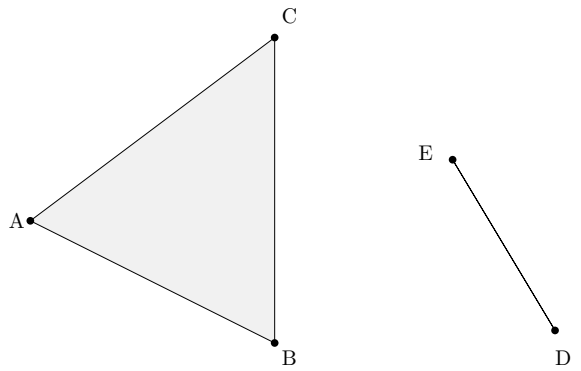


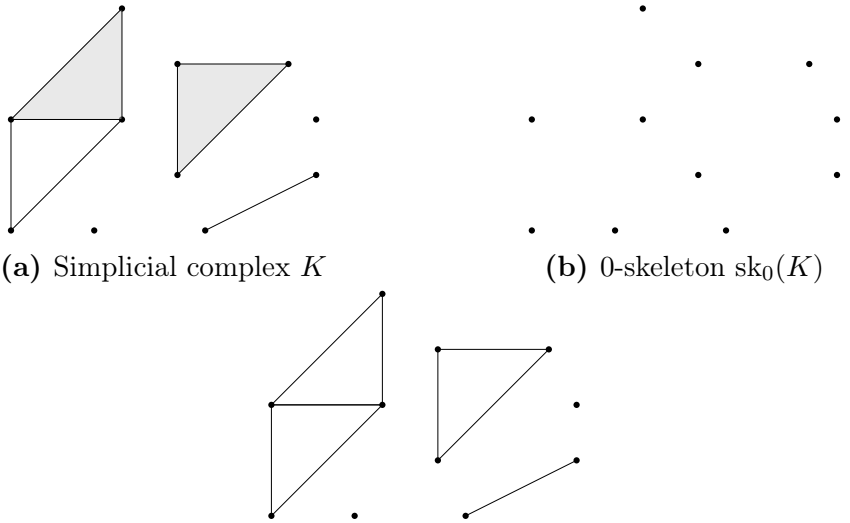
Figure 2.2: Simplicial complex with vertices A, B, C, D, E and simplices {A, B}, {A, C}, {B, C}, {A, B, C}, {D, E}.

Definition 2.1.4 (Abstract simplicial complex) A pair $X = (V(X), \Sigma(X))$ is called an *abstract simplicial complex* if $\sigma \in \Sigma(X)$ and $\emptyset \neq \tau \subseteq \sigma$ implies $\tau \in \Sigma(X)$. The set $V(X)$ is the *vertex set* and $\Sigma(X)$ denotes the *simplex set*. Simplices consisting of exactly two vertices are called *edges*.

Intuitively, an abstract simplicial complex χ_A describes a simplicial complex χ including all faces from every simplex, $\chi_A = \chi \cup \Sigma(\chi)$. Figure 2.2 visualizes an abstract simplicial complex if and only if the simplex $\{A, B, C\}$ is included. Otherwise 2.2 is a geometric, but not an abstract simplicial complex.

Definition 2.1.5 (n-skeleton) Let K be a simplicial complex. The *n-skeleton* of K, denoted as $\mathrm{sk}_n(K)$, is defined as the union of all simplices of K with dimension $m \leq n$.

Figure 2.3 illustrates the simplicial complex K with $\mathrm{sk}_0(K)$ and $\mathrm{sk}_1(K)$. Since the highest dimensional simplex in K is three dimensional, it holds $\mathrm{sk}_2(K) = K$.



(a) Simplicial complex K **(b)** 0-skeleton $\mathrm{sk}_0(K)$

(c) 1-skeleton $\mathrm{sk}_1(K)$

Figure 2.3: Visualisation of skeletons of a geometric simplicial complex.

Definition 2.1.6 (Covering and good cover (5)) Let X be a topological space. A non-empty and finite collection of sets $\mathcal{U} = \{U_i\}_{i \in I}$ with $X = \bigcup_{i \in I} U_i$ is called a *covering* of X. If all $U_i \in \mathcal{U}$ and all finite, non-empty intersections of the U_i are contractible, we call $\mathcal{U}$ a *good cover*.

A more general definition of a covering allows for a countably infinite collection of sets. In this paper, we will be using finite collections, as this is sufficient for the point cloud data application we will be considering.

Definition 2.1.7 (Map of indexed covers) Let X and Y be topological spaces, with covers $\mathcal{U} = (U_i)_{i \in I}$ and $\mathcal{V} = (V_j)_{j \in J}$, respectively. A *map of indexed covers* $\varphi \colon \mathcal{U} \to \mathcal{V}$ is defined through a map $\varphi \colon I \to J$ between the indexing sets. We say that a continuous map $f \colon X \to Y$ is *carried by* φ for all $i \in I$, if $f(U_i) \subseteq V_{\varphi(i)}$.

Definition 2.1.8 (Nerve (5)) Let X be a topological space and $\mathcal{U} = \{U_i\}_{i \in I}$ be any covering of X, so $X = \bigcup_{i \in I} U_i$. The *nerve* of $\mathcal{U}$, denoted $\mathcal{N}(\mathcal{U})$, is defined as the abstract simplicial complex whose vertices are the set I, and where any $\emptyset \neq S \subseteq X$ is a simplex in $\mathcal{N}(\mathcal{U})$ if and only if $\bigcap_{s \in S} U_s \neq \emptyset$.

Example 2.1.1 Let $X = \{(0,0),(1,0),(0,3)\} \subset \mathbb{R}^2$ and choose the covering

$$\mathcal{U} = \{B_{\frac{1}{4}}(0), B_1(3), B_{\frac{1}{8}}(1)\}$$

Since the elements of the covering do not intersect, $N(\mathcal{U}) = X$.

Example 2.1.2 Let $X = \partial([0,1] \times [0,1])$ be the boundary of the unit square. Consider the covering $\mathcal{U}$ shown in Figure 2.4 consisting of four rectangles intersecting in the corners. Each vertex in $\mathcal{N}(\mathcal{U})$ corresponds to one element of the covering and two vertices are connected by an edge if the corresponding elements of the covering intersect. For example, the blue and pink rectangle intersect, hence the blue and pink vertex are connected by an edge. The blue and orange rectangle do not intersect, hence the blue and orange vertex are not connected.

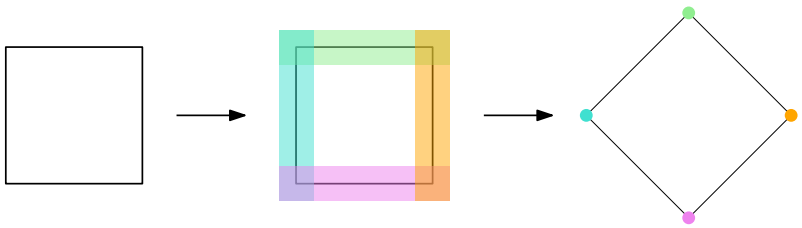


Figure 2.4: The figure on the left displays X, the unit square. The middle figure shows a covering $\mathcal{U}$ of X and the corresponding nerve $N(\mathcal{U})$ is shown on the right.

Consider a topological space X, whose elements are vectors from a data set. Nerves are the main way to represent such spaces through a combinatorial model, which is suitable for computational analyses.

Definition 2.1.9 (Geometric realization) Let K be a geometric or an abstract simplicial complex and $f : V(K) \to \mathbb{R}^d$ an injective map. Then f is called *geometric realization* of K, if $f(K)$ is a geometric simplicial complex.

Definition 2.1.10 (Barycentric subdivision (7)) Let K be a geometric simplicial complex and $\sigma = [x_1, \ldots, x_k] \in K$ a simplex spanned by the vertices $x_1, \ldots, x_k \in K$. The *barycenter* of σ is defined as $b_\Delta(\sigma) = \frac{1}{k}\sum_{i=1}^{k} x_i$. The *barycentric subdivision* $\mathrm{Sd}(K)$ is the decomposition of $[x_1, \ldots, x_k]$ into the n simplices $[b_\Delta, w_0, \ldots, w_{n-1}]$ where inductively $[w_0, \ldots, w_{n-1}]$ is a simplex in the barycentric subdivision of the face $[x_0, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n]$.

The barycenter of a filled triangle corresponds to its center of gravity. Barycentric subdivisions are a useful tool for refining simplicial complexes.

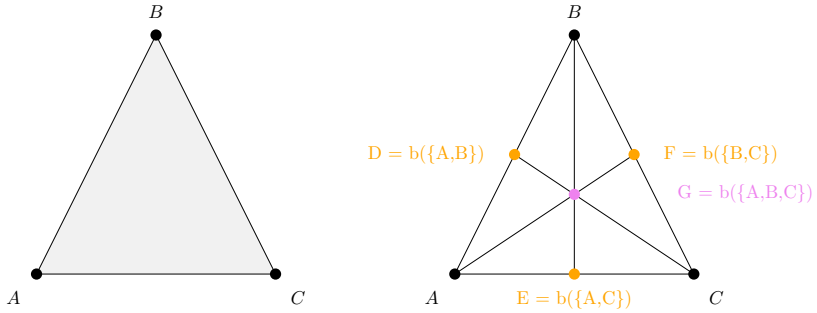


Figure 2.5: Visualization of the geometric realization of a simplicial complex (left) and the barycentric subdivision (right).

The left object in Figure 2.5 displays the geometric realization $|K|$ of a simplicial complex. The right object shows the corresponding geometric representation of the barycentric subdivision including the barycenters of the 1-simplices in orange and the barycenter of the 2-simplex in pink (12).

2.1.1 Homotopy and the Homotopy Extension Property

In the upcoming section we will introduce the definition of a homotopy and the homotopy extension property. The latter is a useful result in algebraic topology, which enables the extension of homotopies on subspaces to larger spaces. Furthermore, it provides a way of proving homotopy equivalence between spaces (2). This subsection is mainly based on the paper 'A Unified View on the Functorial Nerve Theorem and its Variations' by Bauer et al. (2) and on Hatcher's 'Algebraic Topology' (7).

Definition 2.1.11 (Homotopy and homotopic functions) Let X and Y be two topological spaces and $f, g : X \to Y$ two maps. We call f and g *homotopic*, denoted $f \cong g$ if there exists a continuous map $H\colon X \times [0,1] \to Y$ such that $H(x,0) = f(x)$ and $H(x,1) = g(x)$. The function H is called a *homotopy* between f and g.

Definition 2.1.12 (Homotopy equivalence) Let X and Y be two topological spaces. We call X and Y *homotopy equivalent*, denoted $X \simeq Y$, if there exists continuous maps $f\colon X \to Y$ and $g\colon Y \to X$ such that $f \circ g \cong id_Y$ and $g \circ f \cong id_X$.

Figure 2.6 displays the two spaces S^1 and $X = \mathbb{R}^2 \setminus \{(0,0)\}$. Intuitively, these are homotopy equivalent since we can shrink the radius of S^1 to an infinitesimal radius. Following Definition 2.1.12, we must provide two functions in order to prove homotopy equivalence. Let $f\colon S^1 \to X$ be the natural embedding and $g\colon X \to S^1$ be given by $g(x) = \frac{x}{\|x\|}$. Since $g \circ f = \mathrm{id}_{S^1}$ it follows in particular that $g \circ f \cong \mathrm{id}_{S^1}$.

Consider the function

$$H\colon X \times [0,1] \to X \qquad H(x,t) = \frac{x \cdot t}{\|x\|} + (1-t) \cdot x$$

Since $H(x,t) \neq 0$ for all $x \in X$ and all $t \in [0,1]$, it holds $H(x,t) \in X$. The function H is continuous as a composition of continuous functions and it holds $H(x,0) = x$ and $H(x,1) = (f \circ g)(x)$. Therefore H is indeed a homotopy between f and g, therefore $S^1 \simeq \mathbb{R}^2 \setminus \{(0,0)\}$.

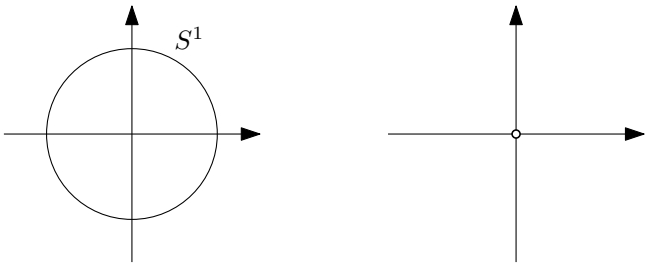


Figure 2.6: S^1 and $\mathbb{R}^2 \setminus \{(0,0)\}$ are homotopy equivalent.

Finding a specific homotopy between two functions can be quite challenging. The *straight line homotopy* $H(x,t) = t \cdot g(x) + (1-t) \cdot f(x)$ however, presents an example of a simple homotopy that we will use later on to prove the Nerve Theorem. Example 2.1.3 provides an example of such a homotopy.

Example 2.1.3 Let $X = \mathbb{R}^2 \setminus \{(0,0)\}$ and consider the two functions

$$f(x) = (\cos(\pi x), \sin(\pi x)) \qquad g(x) = (\cos(\pi x), \sin(2\pi x))$$

shown in Figure 2.7. The straight line homotopy $H(x,t) = t \cdot g(x) + (1-t) \cdot f(x)$ is illustrated by the small black arrows. At time $t = 0$ the homotopy is exactly the function f and then continuously deforms until H defines the function g at time $t = 1$. Since we are considering the space $X = \mathbb{R}^2 \setminus \{(0,0)\}$, there is no straight line homotopy between f and h or g and h, as the point $(0,0)$ is not included in our space. Intuitively, one can think of f not being transformable through the hole in the space at the point $(0,0)$.

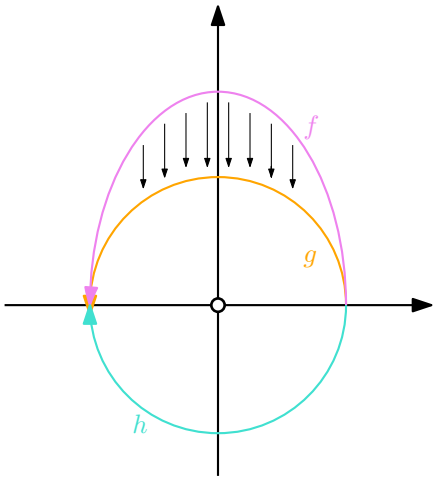


Figure 2.7: Straight line homotopy between f and g from Example 2.1.3.

Definition 2.1.13 (Contractible space) A topological space X is called *contractible*, if X is homotopy equivalent to a one point space, i.e. for $a \in X$ it holds $X \simeq \{a\}$.

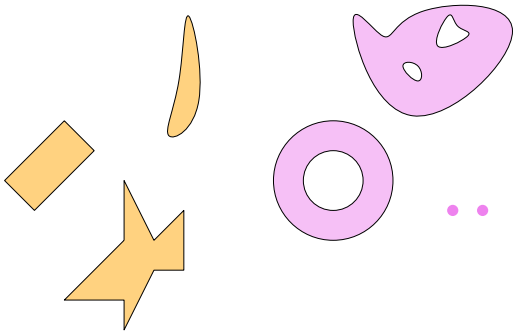


Figure 2.8: The orange spaces are contractible, the pink spaces are not.

Intuitively, two homotopy equivalent spaces can be continuously deformed into one another, without tearing. The orange spaces in Figure 2.8 can be continuously deformed into a one point space. The pink spaces on the other hand cannot be transformed into a one point space without tearing the space in order to eliminate the holes in the space.

Definition 2.1.14 (Homotopy extension property) Let X and Y be topological spaces and $A \subset X$. The pair (X, A) is said to have the *homotopy extension property*, if for every pair of maps $f\colon X \times \{0\} \to Y$ and $g\colon A \times [0,1] \to Y$ with

$$f(A \times \{0\}) = g(A \times \{0\})$$

can be extended to a map $h\colon X \times [0,1] \to Y$.

Lemma 2.1.4 Let X be a topological space and $A \subseteq X$. The pair (X, A) has the homotopy extension property if and only if $X \times \{0\} \cup A \times [0,1]$ is a retract of $X \times [0,1]$.

The proof of Lemma 2.1.4 can be found on page 533 in Hatcher's 'Algebraic Topology' (7). Next, we introduce a useful lemma which provides an alternative way of checking whether two function are homotopic to one another.

Lemma 2.1.5 Let X be a finite simplicial complex and let Y be a topological space. Let $(f, \varphi), (g, \varphi)$ be two morphisms of covered spaces

$$(f, \varphi), (g, \varphi)\colon (|X|, (|U_i|)_{i\in[n]}) \to (Y, (V_j)_{j\in J})$$

with the same map of index sets $\varphi\colon [n] \to J$, a cover of subcomplexes $\mathcal{U} = (|U_i|)_{i\in[n]}$ and a good cover $\mathcal{V} = (V_j)_{j\in J}$. Then f is homotopic to g.

Proof. Let $m = \dim(K)$. In order to construct a homotopy between f and g, we will first use induction to define homotopies between $f|_{|\mathrm{sk}_m(L_i)|}$ and $g|_{|\mathrm{sk}_m(L_i)|}$. These homotopies

$$H^m\colon |\mathrm{sk}_m(K)| \times [0,1] \to Y$$

are carried by

$$\varphi\colon (|\mathrm{sk}_m(L_i)| \times [0,1])_{i\in[n]} \to (V_j)_{j\in J} \tag{2.1}$$

which is induced by the given map $\varphi\colon [n] \to J$. Finally, we will use the homotopy extension property to deduce the homotopy between f and g.

Base Case: Let $m = 0$ and $p \in K$ be a vertex. Let $i_0, \ldots, i_k \in [n]$ be the indices i such that $|p| \in |L_i|$, where $|p|$ denotes the geometric realization of p. Since $(V_j)_{j\in J}$ is a good cover, $S \coloneqq \bigcap_{l=0}^{k} V_{\varphi(i_l)}$ is contractible. Since $f(|p|), g(|p|) \in S$, we can find a path Γ that connects $f(|p|)$ and $g(|p|)$. With $H^0 = \Gamma$, we have found the wanted homotopy. It remains to show that H^0 is carried by φ. Let $(|p|, t) \in |p| \times [0,1]$ be a point, we have to show that $H^0((|p|, t)) \in V_{\varphi(i)}$. If $(|p|, t) \in |L_i| \times [0,1]$, then the index i is one of the indices $i_0, \ldots, i_k \in [n]$ above. By construction it holds $f(|p|), g(|p|) \in S$ and therefore $H^0((|p|, t)) \in S \subseteq V_{\varphi(i_l)} = V_{\varphi(i)}$.

Induction hypothesis: Let

$$H^{m-1}\colon |\mathrm{sk}_{m-1}(K)| \times [0,1] \to Y$$

be the homotopy between $f|_{|\mathrm{sk}_{m-1}(L_i)|}$ and $g|_{|\mathrm{sk}_{m-1}(L_i)|}$, carried by

$$\varphi\colon (|\mathrm{sk}_{m-1}(L_i)| \times [0,1])_{i\in[n]} \to (V_j)_{j\in J}.$$

We now want to deduce the homotopy for m, assuming that the construction holds for $(m-1)$.

Induction step: Let $\sigma \in \mathrm{sk}_m(K)$ be an m-simplex and $i_0, \ldots, i_k \in [n]$ be those indices i such that $\sigma \in L_i$. Let $\partial\sigma$ denote the boundary of the simplex. Using the induction hypothesis and the fact that the boundary of a k-simplex is a $(k-1)$-simplex itself, we have

$$H^{m-1}(|\partial\sigma| \times [0,1]) \subseteq W \coloneqq \bigcap_{l=0}^{k} V_{\varphi(i_l)}.$$

Per assumption $\mathcal{V}$ is a good cover, thus W is contractible. Since $(\sigma\times\{0\})\cup(\partial\sigma\times[0,1])$ is a retract of $\sigma \times [0,1]$, Lemma 2.1.4 concludes that $(\sigma, \partial\sigma)$ has the homotopy extension property. Therefore, we can extend the homotopy $H^{m-1}|_{|\partial\sigma|\times[0,1]}$ to the homotopy

$$H^m|_{|\sigma|\times[0,1]}$$

between $f|_{|\sigma|}$ and $g|_{|\sigma|}$. The choice of the m-simplex σ from above was arbitrary, hence without loss of generality we can choose σ as the m-skeleton sk_m and thus extend

$$H^{m-1}\colon |\mathrm{sk}_{m-1}(K)| \times [0,1] \to Y \qquad \text{to} \qquad H^m\colon |\mathrm{sk}_m(K)| \times [0,1] \to Y.$$

It remains to confirm that the map H^m is carried by φ, defined in Equation 2.1. In order to prove this, we need to show that for every point $(x,t) \in |\mathrm{sk}_m(L_i)| \times [0,1]$ it holds $H^m(x,t) \in V_{\varphi(i)}$. Following the induction hypothesis, we already know that this holds for every $x \in |\mathrm{sk}_{m-1}(L_i)|$, points lying in the $(m-1)$-skeleton of L_i. Again, using that the boundary of an m-simplex is a $(m-1)$-simplex, it suffices to prove the claim for all points x contained in the interior of an m-simplex $\sigma \in L_i$. The index i must now be one of the indices $i_0, \ldots, i_k \in [n]$ and therefore by the

construction of H^m we can conclude

$$H^m(x,t) \in H^m(|\sigma| \times [0,1]) \subseteq W \subseteq V_{\varphi(i)}.$$

The claim follows as we have found a homotopy between the morphisms of covered spaces (f,φ) and (g,φ). □

2.2 The Nerve Theorem

We have now established the framework needed to present the Nerve Theorem. This section, and the proof of the Nerve Theorem in particular, will follow the proof published by Bauer et al in their paper 'A Unified View on the Functorial Nerve Theorem and its Variations' (2).

Theorem 2.2.1 (Nerve Theorem (2)) Let $X \subset \mathbb{R}^n$ be a topological space with a closed and convex covering $\mathcal{U}$. The map $\Gamma\colon X \to |\mathrm{Sd}(\mathcal{N}(\mathcal{U}))|$ is a homotopy equivalence. In particular X is homotopy equivalent to $N(\mathcal{U})$, denoted $X \simeq N(\mathcal{U})$.

In order to prove the Nerve Theorem, we will construct a homotopy inverse Ψ to Γ. The idea is to show that Ψ and Γ are morphisms of covered spaces and then use their properties to prove the homotopy equivalence. We will present a few lemmas in preparation for the main proof. Other constructions, such as the next definition, provide a technical features needed for the construction of the inverse.

Lemma 2.2.2 Let $\mathcal{C} = \{C_i\}_{i\in[n]}$ be a finite family of closed, convex subsets of $\mathbb{R}^d$. Then there exists a family of open sets $\mathcal{U} = \{U_i\}_{i\in[n]}$ with $C_i \subseteq U_i$ and $\mathcal{N}(\mathcal{U}) = \mathcal{N}(\mathcal{C})$.

Proof. For every closed and disjoint sets $U, V \subseteq \mathbb{R}^d$ there exist open and disjoint neighborhoods $\mathcal{O}_U$ and $\mathcal{O}_V$, i.e. $\mathbb{R}^d$ is normal. Thus for every pair of closed, convex sets C_i, C_j there exist disjoint, open sets $V_{i,j} \supseteq C_i$ and $V_{j,i} \supseteq C_j$. Let

$$U_i = \bigcap_{j:C_i\cap C_j=\emptyset} V_{i,j}$$

since finite intersections of open sets are open, this yields an open cover $\mathcal{U} = \{U_i\}_{i\in[n]}$. It remains to be shown that $\mathcal{N}(\mathcal{U}) = \mathcal{N}(\mathcal{C})$, clearly $\mathcal{N}(\mathcal{C}) \subseteq \mathcal{N}(\mathcal{U})$.

Let $\sigma \in \mathcal{N}(\mathcal{U})$ be a k-simplex, per definition there exist k open sets $U_1, \ldots, U_k$ with $U_1 \cap \cdots \cap U_k \neq \emptyset$. Through renumbering the open sets we can choose the

first k sets without loss of generality. In order to prove that $\sigma \in \mathcal{N}(\mathcal{C})$, we assume the contrary. Assume that the corresponding C_j to the U_j do not intersect, i.e. $C_1 \cap \cdots \cap C_k = \emptyset$. Then there exist $\alpha, \beta \in \{1, \ldots, k\}$ such that $C_\alpha \cap C_\beta = \emptyset$ and hence $V_{\alpha,\beta} \cap V_{\beta,\alpha} = \emptyset$. Using the definition of the U_i, this contradicts the assumption that $U_1 \cap \cdots \cap U_k \neq \emptyset$. Hence $\sigma \in \mathcal{N}(\mathcal{C})$ and Lemma 2.2.2 follows.

□

Lemma 2.2.2 contributes to the construction of the homotopy inverse Ψ, as it provides an open cover containing our given closed and convex cover $\mathcal{U}$. The inverse will be a linear combination of multiple different functions, one of them being a partition of unity, introduced in the following definition. In order to maintain continuity, this function requires the covering to consist of the open sets provided by Lemma 2.2.2.

Definition 2.2.1 (Partition of unity (17)) Let X be a topological space and $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ a finite open covering. A *partition of unity* subordinate to $\mathcal{U}$ is a family of real valued functions $(f_\alpha)_{\alpha \in A}$ with the following properties for all $x \in X$:

1. $0 \leq f_\alpha(x) \leq 1 \qquad \forall \alpha \in A$;
2. $\sum_{\alpha \in A} f_\alpha(x) = 1$;
3. $\overline{\{x \in X \mid f_\alpha(x) > 0\}} \subseteq U_\alpha$.

Partitions of unity are useful for a variety of different constructions. They allow us to operate on a topological space that is divided into smaller sections, blending from one section to another whilst preserving continuity. For the purpose of proving the Nerve Theorem, the partition of unity lets us move from a covering of the topological space X to a covering of its nerve. A more detailed construction will follow in Equation 2.2.

Definition 2.2.2 (Morphism of covered spaces) Let X and Y be topological spaces with coverings $\mathcal{U} = \{U\}_{i \in I}$ and $\mathcal{V} = \{V\}_{j \in J}$, respectively. A *morphism of covered spaces*

$$(f, \varphi) : (X, \mathcal{U}) \to (Y, \mathcal{V})$$

consists of a continuous map $f : X \to Y$ which is defined through the map between index sets $\varphi : I \to J$.

Definition 2.2.3 (Closed barycentric star) Let K be a simplicial complex and $v \in K$ a vertex. The subspace

$$\mathrm{bst}(v) = |\{\sigma \in \mathrm{Sd}(K) \mid \sigma \cup \{v\} \in \mathrm{Sd}(K)\}| \subseteq |\mathrm{Sd}(K)|$$

is called the *closed barycentric star* of $v \in K$.

Figure 2.9 illustrates the closed barycentric star $\mathrm{bst}(v)$ of the orange vertex $v \in K$. The pink edges connect v with other vertices and are hence elements of $\mathrm{bst}(v)$. The gray faces contain v as a vertex and are also part of the barycentric star.

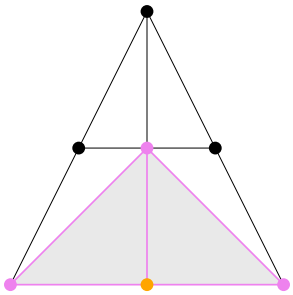


Figure 2.9: Closed barycentric star of the orange vertex.

In preparation for the next lemma, we construct a function Φ which sends an element $x \in X$ to its barycentric coordinates in $|\mathcal{N}(\mathcal{U})|$. As stated in the proof of Lemma 2.2.2, $\mathbb{R}^d$ is a normal space. It follows from the Lemma of Urysohn (8) that for each pair of disjoint sets (C_i, U_i) there exists a continuous function φ_i satisfying $\varphi_i(C_i) = 0$ and $\varphi_i(U_i) = 1$. The following is an example of such an *Urysohn function*:

$$\varphi_i : \mathbb{R}^d \to [0,1] \qquad \varphi_i(x) = \frac{d(x, \mathbb{R}^d \setminus U_i)}{d(x, C_i) + d(x, \mathbb{R}^d \setminus U_i)}. \tag{2.2}$$

Here d is defined as

$$d(x, A) = \inf_{a \in A} \tilde{d}(x, a),$$

where $\tilde{d}$ is the standard Euclidean metric on $\mathbb{R}$. Normalizing the functions φ_i yields a partition of unity on X.

$$\psi_i(x) = \frac{\varphi_i(x)}{\sum_{j=0}^{n} \varphi_j(x)}.$$

Let $v_i \in \mathcal{N}(\mathcal{U})$ be the i^{th} vertex in $\mathcal{N}(\mathcal{U})$ and $|v_i|$ the corresponding point in the geometric realization. We can now define the map $\Phi : X \to |\mathcal{N}(\mathcal{U})|$ using the constructions above:

$$\Phi(x) = \sum_{i=0}^{n} \psi_i(x) \cdot |v_i|.$$

Lemma 2.2.3 Let $\Psi\colon X \to |\mathrm{Sd}(N(\mathcal{U}))|$ be the map which sends points in the topological space to the barycentric subdivision of the $\mathcal{N}(\mathcal{U})$. The pair of maps $(\Psi, id_{[n]})$ is a morphism of covered spaces

$$(X,\mathcal{U}) \to (|\mathrm{Sd}(N(\mathcal{U}))|, \mathcal{B})$$

where $\mathcal{B} = \{\mathrm{bst}(v_i)\}_{i\in[n]}$ is the covering consisting of closed barycentric stars.

Proof. Let $\alpha\colon |N(\mathcal{U})| \to |\mathrm{Sd}(N(\mathcal{U}))|$ be the natural homeomorphism from the nerve of $\mathcal{U}$ to its barycentric subdivision. Using the construction of the map Φ from above, we can write $\Psi = \alpha \circ \Phi$. It follows from the definition of the Urysohn function φ_i that $\varphi_i(x) = 1$ for $x \in C_i$. Thus $\psi_i(x)$ is maximal among the $\psi_j(x)$. Therefore $\Psi(x) \in \mathrm{bst}(v_i)$ which concludes the claim. □

An important characteristic of the covering consisting of closed barycentric stars is its contractability, shown in Lemma 2.2.4. In order to prove this lemma, we introduce the following two definitions.

Definition 2.2.4 (Coface) Let A and B be simplices. If B is a face of A, then A is called a *coface* of B.

Definition 2.2.5 (Chain of a simplicial complex (16)) Let K be a simplicial complex. Then a nested sequence of simplices $\sigma_1 \subset \sigma_2 \subset \cdots \subset \sigma_k$ of K is called a *chain of* K. The set of all chains of K comprises an abstract simplicial complex, sometimes referred to as the barycentric subdivision $\mathrm{Sd}(K)$ of K.

Lemma 2.2.4 Let K be a simplicial complex and $\sigma \in K$ a simplex. The intersection $\bigcap_{v\in\sigma} \mathrm{bst}(v)$ is contractible.

Proof. Let $\sigma = [x_0, \ldots, x_k] \in K$ be a simplex and

$$L = \{\tau_0 \subset \cdots \subset \tau_m \mid \sigma \subseteq \tau_0\} \subseteq \mathrm{Sd}(K)$$

a subset of all chains of K. The first step is to show that $\bigcap_{i=0}^{k} \mathrm{bst}(v_i) = |L|$, where $|L|$ is the geometric realization of the subcomplex L. We prove this equality of sets by verifying both inclusions.

Let $\phi \in L$ be a simplex, per definition it holds that ϕ is contained in the chain $(\sigma \subseteq \tau_0 \subseteq \cdots \subseteq \tau_m) \in L$. Therefore, $|\phi| \subseteq \mathrm{bst}(v_i)$ for all $i \in \{0, \ldots, k\}$. Since ϕ was arbitrary, we can deduce that $|L| \subseteq \bigcap_{i=0}^{k} \mathrm{bst}(v_i)$.

Now, let $|\phi| = |(\tau_0 \subseteq \cdots \subseteq \tau_m)| \subseteq \bigcap_{i=0}^{k} \mathrm{bst}(v_i)$. Since for all $i \in \{0, \ldots, k\}$ it holds $|\phi| \subseteq \mathrm{bst}(v_i)$, it follows that $v_i \in \tau_0$ for all $i \in \{0, \ldots, k\}$ and therefore $\sigma \in \tau_0$. Using the definition of L, this implies that $\phi \in L$. Finally, this yields the other inclusion $\bigcap_{i=0}^{k} \mathrm{bst}(v_i) \subseteq |L|$, and thus proves the first claim.

It remains to show that $\mathcal{B}$ is a good cover. Using the first part of the proof it follows that every geometric simplex in $\bigcap_{v \in \sigma} \mathrm{bst}(v) \subseteq |L| \subseteq |\mathrm{Sd}(K)|$ has a coface in $\bigcap_{v \in \sigma} \mathrm{bst}(v)$ with $b_\Delta(\sigma)$, the barycenter of σ, being a vertex in the coface. Since $b_\Delta(\sigma)$ is a vertex in all these cofaces, we can deduce that $\bigcap_{v \in \sigma} \mathrm{bst}(v)$ is star-shaped[1], with center $b_\Delta(\sigma)$. Star-shaped spaces are convex and thus contractible[2] thus, the claim follows.

□

Lemma 2.2.5 Let $\Gamma \colon |\mathrm{Sd}(N(\mathcal{U}))| \to X$ be the unique and continuous map that is affine linear on each simplex $\sigma \in |\mathrm{Sd}(N(\mathcal{U}))|$. Then, the pair of maps $(\Gamma, id_{[n]})$ is a morphism of covered spaces

$$(|\mathrm{Sd}(N(X))|, \mathcal{B}) \to (X, \mathcal{U}).$$

Proof. The map Γ sends the vertices of any simplex $\sigma \in \mathrm{bst}(v_i)$ to a convex and closed element $U_i \in \mathcal{U}$. Since U_i is convex and each simplex can be written as a linear combination of its vertices, it holds $\Gamma(\sigma) \subseteq U_i$. Again using convexity and the definition of $\mathrm{bst}(v_i)$, it follows $\Gamma(\mathrm{bst}(v_i)) \subseteq U_i$, which proves the claim. □

Having set the stage using the lemmas above, we can now prove the Nerve Theorem 2.2.1.

Proof of the Nerve Theorem. Let Γ and Ψ be the maps as defined above. In order to prove the Nerve Theorem, it is sufficient to show that $\Gamma \circ \Psi \cong id_{[n]}$ and

[1] A set $S \subset \mathbb{R}^n$ is called a *star-shaped*, if there exists a point $x_0 \in S$ called *center of the star*, such that for all $x \in S$ the line segment from x_0 to x is in S. In other words, if for all $t \in [0, 1]$ it holds $(x_0 \cdot t + (1 - t) \cdot x) \in S$.

[2] For more information and a proof of this statement refer to Theorem 5.19 in Armstrong's 'Basic Topology' (1).

$\Psi \circ \Gamma \cong id_{|\mathrm{Sd}(\mathcal{N}(\mathcal{U}))|}$, i.e. that Γ and Ψ are homotopy inverses of each other.

The composition of continuous maps is continuous, hence according to Lemma 2.2.3 and Lemma 2.2.5 the pair $(\Gamma \circ \Psi, id_{[n]})$ defines a morphism of covered spaces. Moreover, $\Gamma \circ \Psi \cong id_X$ by a straight line homotopy: for all $x \in C_i$ it holds $\Gamma \circ \Psi(x) \in C_i$ and because the C_i are convex, the straight line connecting x and $\Gamma \circ \Psi(x)$ is contained in C_i.

Analogously, it follows that $(\Psi \circ \Gamma, id_{[n]})$ is a morphism of covered spaces. The cover consisting of closed barycentric stars $\mathcal{B}$ is a good cover according to Lemma 2.2.4, hence it follows from Lemma 2.1.5 that the composition $\Psi \circ \Gamma$ is homotopic to $id_{|\mathrm{Sd}(\mathcal{N}(\mathcal{C}))|}$. □

Chapter 3

Introduction to `Mapper`

This upcoming section is mainly based on the introductory paper 'Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition' (17) and chapter three in 'Topology and Data' (3).

3.1 Approaches to Data Analysis

As tremendous amounts of data become more easily accessible, the use of data analysis with efficient and innovative techniques has become indispensable. Different methods allow for a visualization of data sets and their geometric structure. An example for a commonly used method is the *projection pursuit method* (9). This approach uses a high-dimensional data set and projects the data vectors onto two or three components, making it possible to visualize the data set. Whilst this can reveal important information about the data set, as we will see in Subsection 4.1.2, relevant features of the data may be overseen when projecting onto only two or three dimensions.

3.2 An Introduction to `Mapper`

`Mapper` is a computational method that was developed in 2007 by G. Singh, F. Mémoli and G. Carlsson. Its purpose is to simplify, analyze and visualize high dimensional data sets, whilst maintaining important characteristics of the data. `Mapper` helps to find and illustrate shape of data and by doing so became a very useful method for Topological Data Analysis. It's core strengths are:

- **Insensitivity to metric:** When analyzing data sets originating from a context in physics, a measure of distance can often be easily identified. In bi-

ological settings however, it is sometimes not clear how to measure distance between data points. Mapper allows for various different notions of distance, unique to the data set being analyzed.

- **Parameter flexibility:** Mapper enables the user to decide on multiple parameters constituting the analysis. Depending on the nature of the data set, the desired granularity of analysis, and the data characteristics aimed to be analyzed, the user can adapt Mapper accordingly. This allows for a very targeted analysis.

- **Multiscale representation:** As a result of the parameter flexibility, we can look at the data with various levels of resolution. Observing the effect on the data of slightly altering the parameters can reveal more about the data.

These three properties can be referred to as the topological foundation underlying Mapper and aiming for the recognition of patterns in data within a reasonable perimeter in order to extract a simplified geometric description.

3.2.1 The Topological Construction of Mapper

The theoretical construction of Mapper is carried out on a topological space. In order to use the algorithm on point cloud data, the so called statistical version is then deduced.

Let X be a topological space, Z a parameter space with a finite open covering $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ and let $f : X \to Z$ be continuous. Due to continuity, the set $\{f^{-1}(U_\alpha)\}_{\alpha \in A}$ is a covering of X. We then identify the path-connected[1] components of $f^{-1}(U_\alpha)$ for all $\alpha \in A$. Each connected component corresponds to one vertex in the simplicial complex, and if connective components from distinct pre-images $\{f^{-1}(U_\alpha)\}$ and $\{f^{-1}(U_\beta)\}$ overlap, the vertices corresponding to the path-connected components are joined by an edge. Example 3.2.1 illustrates the issue arising, if we do not separate the connected components of the pre-images.

Example 3.2.1 Consider the topological space $X = \{(x, y) \mid x^2 + y^2 = 1\} \subseteq \mathbb{R}^2$ and a covering $\mathcal{U} = \{A, B, C\}$ of X where $A = \{(x, y) \mid y > 0\}, B = \{(x, y) \mid y < 0\}$ and $C = \{(x, y) \mid y \neq \pm 1\}$. We can observe, that $\mathcal{U}$ consists of three elements A, B, C and four path-connected components $A, B, C \cap \{(x, y) \mid x < 0\}, C \cap \{(x, y) \mid x > 0\}$.

[1] A subset $A \subseteq X$ of a topological space X is called *path-connected*, if for every $x, y \in A$ there exists a continuous path $\gamma : [0, 1] \to A$ such that $\gamma(0) = x$ and $\gamma(1) = y$.

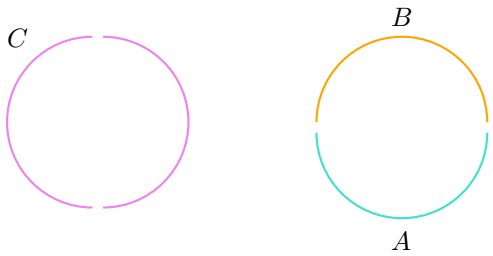


Figure 3.1: Covering from Example 3.2.1.

Not dividing the pre-images into separate path-connected components, might result in the corresponding simplicial complex not being homotopy equivalent to X. This issue happens with A, B, C from Example 3.2.1, the corresponding nerve is depicted in Figure 3.2b. This complex is not homotopy equivalent to X, as we can not transform a circle into a straight line continuously.

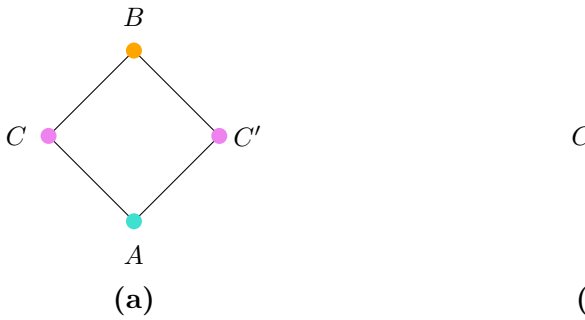


Figure 3.2: Nerves from Example 3.2.1

Figure 3.2a shows the nerve $\mathcal{N}(\mathcal{U})$, taking into account the two path-connected components of C. Mathematically, we can explain the difference between Figure 3.2a and Figure 3.2b using the Nerve Theorem 2.2.1. Without considering its two path-connected components separately, the set C is not convex. Since the Nerve Theorem requires convexity of the covering, homotopy equivalence cannot be deduced. The Nerve Theorem guarantees that the nerve is homotopy equivalent to its topological space, but only if the conditions of the theorem are met.

3.2.2 The Statistical Version of Mapper

In order to use Mapper on point cloud data, the authors introduced the statistical version of the method. In comparison to the topological method, we now implement a statistical approach to clustering the space into its connected components.

Definition 3.2.1 (Filter) Let X be a point cloud. A function $f : X \to \mathbb{R}$ is called a *filter function*.

The importance of the filter function will be highlighted later on. For now, the name filter function provides a brief intuition about what this functions does: it enables the user to analyze the data with respect to certain parameters or characteristics. Coming back to the statistical Mapper: after having chosen the filter function, the image $\mathrm{Im}(f)$ is divided into a set $\mathcal{O}$ of smaller, overlapping intervals. This cover can be thought of as the resolution under which the data is meant to be analyzed (10). The user can now use three flexible parameters to analyze the point cloud from various different angles: the filter function, the length of the intervals l_I and the percentage of overlap p_I.

Example 3.2.2 Let $I = [0, 1]$ with $l_I = \frac{1}{2}$ and $p_I = \frac{1}{2}$. Then the set of smaller intervals is defined as $\mathcal{O} = \{I_1, I_2, I_3\} = \{[0, \frac{1}{2}], [\frac{1}{4}, \frac{3}{4}], [\frac{1}{2}, 1]\}$.

The sets $U_i = \{x \in X \mid f(x) \in I_i\}$ form a covering of the point cloud X. Using any preferred clustering algorithm, the U_i are then clustered. Each cluster found in U_i corresponds to a vertex in the complex. If two distinct clusters intersect, the corresponding vertices are connected with an edge. Aligning with the topological version, the clusters here correspond to the path-connected components described above. To summarize, one can structure the approach of Mapper into four steps:

Algorithm

1. The Data is collected and pre-processed, this can include imputing missing values and normalizing or log-transforming the data. Using any efficient clustering algorithm the pre-processed data is then clustered.

2. Depending on the geometric structures to be analyzed, a filter function is selected. The range of the filter function f is divided into $n \in \mathbb{N}$ overlapping intervals $I = \{I_1, \ldots, I_n\}$.

3. For each interval $I_k \in I$ the number of clusters in the pre-image $f^{-1}(I_k)$ corresponds to the number of vertices in the resulting simplicial complex

4. An edge between the vertices is drawn, if the corresponding pre-images intersect

Figure 3.3 visualizes the method using a height function as the filter function. The image $\mathrm{Im}(f)$ is divided into four overlapping intervals $I_1, \ldots, I_4$ (yellow, blue, violet, orange). The data points $x_j \in X$ of the point cloud are colored according to the image of the filter function $f(x_j) \in I_i$. If for a point $x \in X$ the image lies in two

intervals, it is colored according to the overlapping interval (green, dark blue, red). Next, a clustering algorithm is applied to the covering $\{f^{-1}(I_1), \ldots, f^{-1}(I_4)\}$ of X. Each cluster corresponds to one vertex in the complex on the right. According to the overlap of the intervals I_j, edges are drawn between the vertices.

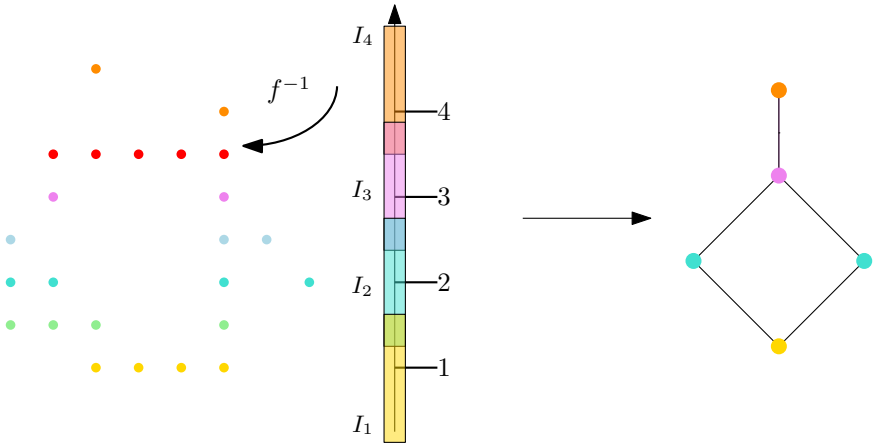


Figure 3.3: Statistical Version of `Mapper`

The Filter Function

The output of `Mapper` is highly dependent on the choice of the filter function. Depending on the geometric properties one is interested in, some filter functions are more or less likely to pick up on interesting characteristics in the data. `Mapper`'s flexibility with the filter function allows us to analyze data sets from several domains and pose various different research questions. In the following some common filter functions will be looked at in more detail.

- **The Gaussian kernel**: This is a filter function used to estimate the density of a data sample. For a given data set X, points $x, y \in X$ and a smoothing parameter $\epsilon > 0$, such a filter is given by

$$f_\epsilon(x) = C_\epsilon \cdot \sum_y \exp\left(\frac{-d(x,y)^2}{\epsilon}\right). \tag{3.1}$$

- **Eccentricity function**: When aiming for a measure describing the closeness of data points to the center of the data set, sometimes referred to as *data depth*, eccentricity functions are an appropriate choice. Given $p \in [1, \infty)$ these

functions are of the form

$$E_p(x) = \Big(\frac{\sum_{y \in X} d(x,y)^p}{|X|}\Big)^{\frac{1}{p}}. \tag{3.2}$$

- **Projection**: This filter function returns the projection onto chosen columns of the data. When wanting to extract one specific parameter from the data set, the projection filter function might be useful. Using a projection filter function with `Mapper` mimics the *projection pursuit method* mentioned in Section 3.1.
- **Entropy**: Measuring the entropy of a point cloud, yields in a description of the order of chaos of the point cloud.

In order to demonstrate the impact of the filter function on `Mapper`'s output, consider the filter functions Entropy, Eccentricity and Projection. Using the scikit-learn function *make_circles* provided by Python (13), we generate 5000 data points arranged in two concentric circles. Figure 3.4 illustrates the point cloud:

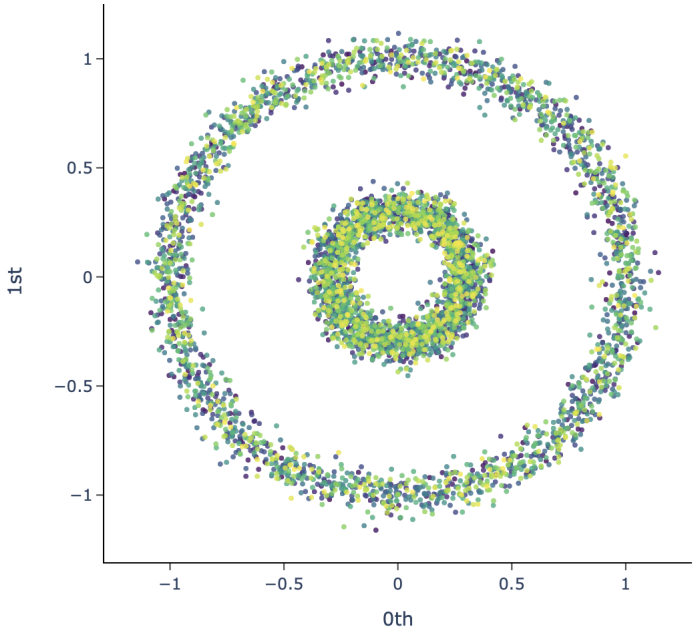
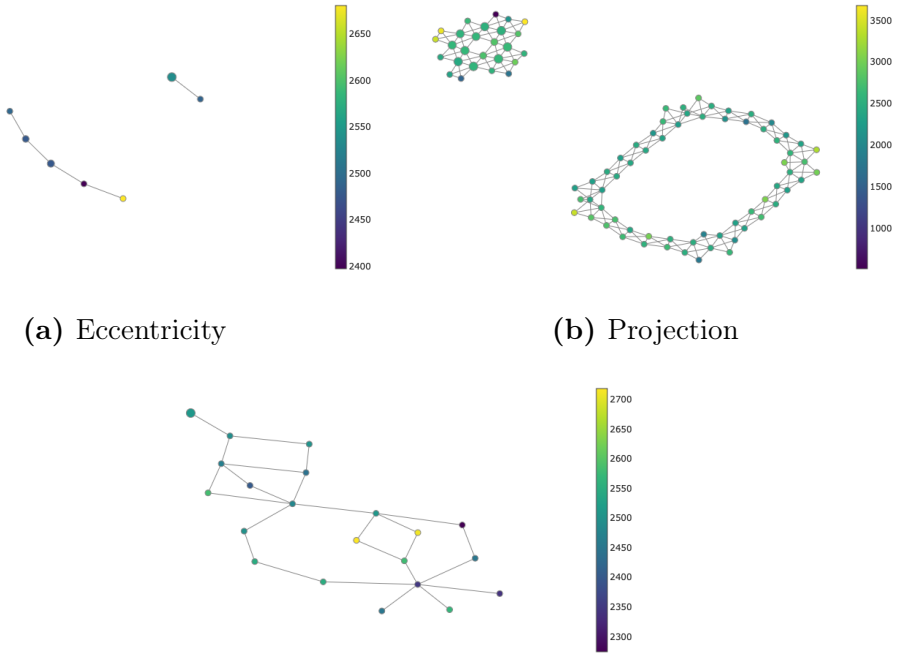


Figure 3.4: Point cloud generated from 5000 points using the *make_circles* function with a noise parameter of 0.05, scale factor of inner and outer circle of 0.3 and random state of 42.

Figure 3.5 illustrates how the different filter functions generate significant differences in the `Mapper` output. The graph resulting from the projection filter function in Figure 3.5b nicely displays an important topological feature of our point cloud; the two holes.

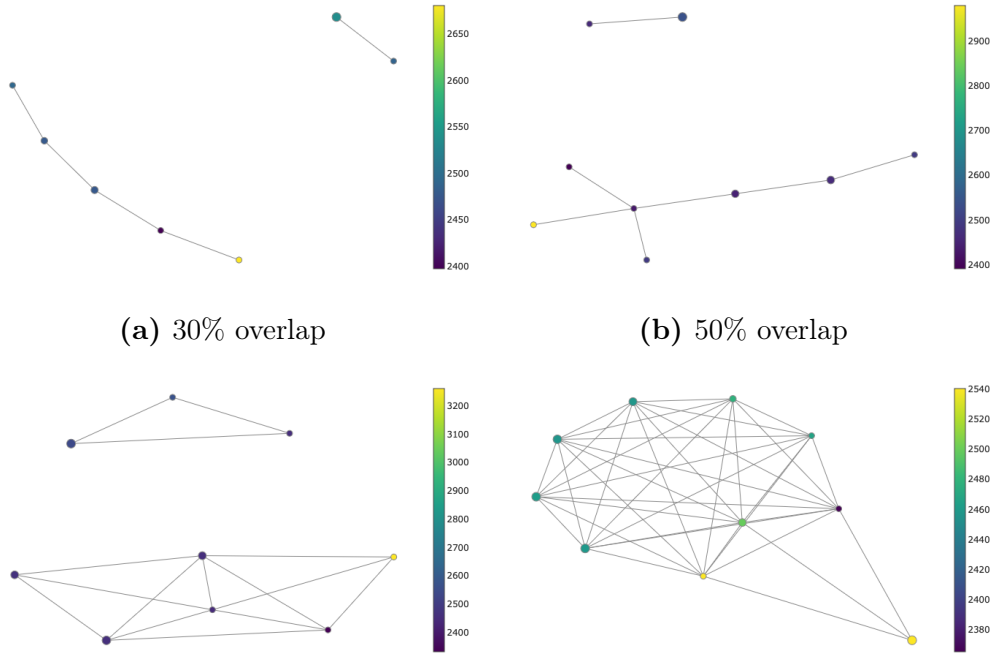


(a) Eccentricity **(b)** Projection

(c) Entropy

Figure 3.5: `Mapper` outputs using different filter functions on the same data. The cover of the image each filter function consists of 10 intervals with an overlap of 30%. By default, the vertices of the graph are colored by the mean value of the points that belong to a given vertex.

The example above illustrates that not every filter function is suitable for a specific data set and it is an essential part of the analysis to identify a useful filter function and optimize the other parameters such as interval length and overlap. Figure 3.6 shows the effect of the percentage of overlap on the `Mapper` output. The method is not able to detect any structure, if the overlap is too small. On the other hand, an overlap of close to, or even equal to 100%, will not display any interesting characteristics, as all of the vertices in the nerve will be connected by an edge.



(a) 30% overlap **(b)** 50% overlap

(c) 70% overlap **(d)** 90% overlap

Figure 3.6: `Mapper` outputs using the eccentricity filter function on the same data. The cover of the image of each filter function consists of 10 intervals with an overlap of 30%-90%.

It is important to note that mathematically relevant structures in the data are not always relevant in the context of the data. More specifically, if the `Mapper` output presents a visually distinguishable subset of a given biological point cloud, as in cancer or diabetes research, this might not be a relevant new sub-type of e.g. cancer or diabetes. A statistical analysis is required, to ensure biological relevance of the mathematically outstanding structures.

Chapter 4

Applications

4.1 Using `Mapper` to illustrate Different Diabetes Types

In 1979 G. Reaven and R. Miller conducted a study on the relationship between chemical and overt diabetes (15). This section focuses on the recreation of their results.

4.1.1 Data

Diabetes can be categorized into four intensity stages based on the human carbohydrate metabolism. The earliest stage is referred to as *prediabetes*, which describes a slight abnormal glucose tolerance. *Subclinical diabetes* denotes abnormal glucose tolerance in response to stress such as pregnancy. The third stage is referred to as *overt diabetes*, here patients show no symptoms of diabetes but present a significant abnormal glucose tolerance. Patients showing classical diabetes symptoms and elevated fasting blood glucose concentration are diagnosed with *chemical diabetes* (6).

In efforts to study the relationship between overt and chemical diabetes, 145 non-obese adults were examined. The authors measured six biological parameters during the course of the experiment.

- Age.
- Weight.
- Fasting plasma glucose, which corresponds to the blood glucose level an a fasted state.
- Area under the plasma glucose curve for the 3 hour oral glucose tolerance

test (Oral Glucose Tolerance Test - OGTT): The OGTT consists of multiple blood samples, which are being drawn first at fasted state and then in regular intervals after having consumed a glucose drink. Plotting the received glucose levels in a graph, yields the desired area under the plasma glucose curve (18).

- Area under the plasma insulin curve for the OGTT: This parameter is obtained similarly to the one above, however in this case insulin levels instead of glucose levels are being measured.
- Steady state plasma glucose response (SSPG): This parameter is measured through giving patients continuous glucose infusions for three hours until achieving constant plasma levels of glucose and insulin. In the last 90 minutes of the infusion period, blood glucose and insulin levels are measured (14).

Further information about the measurements taken can be found in the original article from Reaven and Miller (15).

4.1.2 Results

The authors used a computational projection approach to analyze the data obtained in the experiments. The *Prim-9* program at Stanford Linear Accelerator Computational Center allowed for a two-dimensional projection of three-dimensional data, which yielded the following outcome:

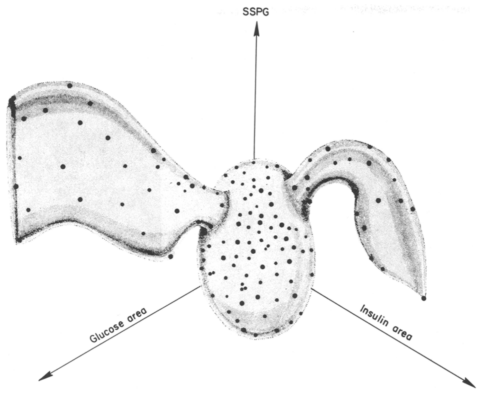
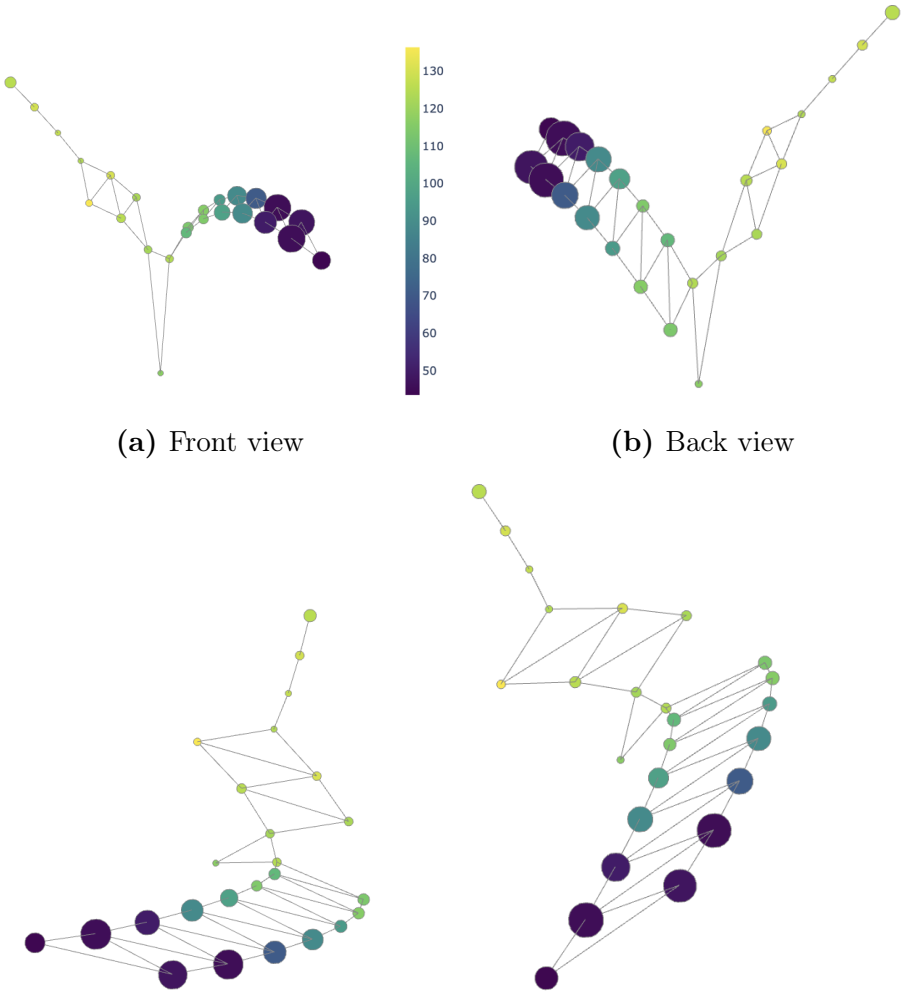


Figure 4.1: Artist's version of Prim-9 output from diabetes data collected in Reaven and Miller's experiment (15).

The authors describe the illustration in Figure 4.1 as a boomerang with two wings and a fat middle (15). The two wings represent patients with overt diabetes (left wing) and chemical diabetes (right wing), while the middle part illustrates the normal group. 28 years later, the authors Singh, Mémoli and Carlsson first used their recently developed method `Mapper` (17) on this data set.

Intending to recreate the computational analysis on the structures in the diabetes data using, we used Mapper on the data set described in Section 4.1.1. Figure 4.2 shows our findings using the eccentricity function provided by the software Giotto-tda (20). An overlap of 61% between the 25 intervals, yields a structure that has significant similarities with the original *Prim-9* illustration shown in Figure 4.1. The color of the nodes corresponds to the eccentricity value, which is calculated using the L^{∞}-norm on the distance matrix between data points. Dark blue nodes represent a distance of 48 to the core of the data, whereas yellow nodes indicate a distance from 130. The node size ranges from 2 to 48 and is proportional to the number of data points contained in the node.



(a) Front view **(b)** Back view

(c) Bottom view **(d)** Top view

Figure 4.2: Recreation of Reaven and Miller's diabetes analysis (15) using Mapper. The point of view is provided in perspective to the view in Figure 4.1.

As mentioned at the end of Section 3.2.2, after finding these mathematically interesting structures displayed in Figure 4.2, it is crucial to ensure the biological significance as well. Using statistical methods, Reaven and Miller found that the structure in Figure 4.1 does indeed illustrate diabetes subtypes (15). After recreating Reaven and Miller's results, we can thus conclude that Figure 4.2 displays the diabetes subtypes. Overt diabetes is represented by the blue wing, while the yellow wing denotes chemical diabetes. The normal subtype is represented by the pointy middle in different shades of green.

4.2 Discoveries in Breast Cancer Research using `Mapper`

In 2011, Monical Nicolau, Arnold Levine and Gunnar Carlsson published substantial findings in breast cancer data (11). Using a method called *Progression Analysis of Disease (PAD)* and `Mapper` the authors were able to identify a subgroup of breast cancers with a unique mutational profile and a 100% chance of survival, that was previously not found using standard clustering techniques. This subgroup is referred to as c-MYB$^+$ breast cancer.

4.2.1 Data

Deciding on what data to extract and use for *PAD* is very significant for obtaining meaningful results. In this breast cancer analysis the authors used data that had previously been collected at different hospitals and research institutions (21). Table 4.1 displays the parameters considered in one out of the four data batches collected. It shows an excerpt from the *Nederlands Kanker Instituut (NKI)* data consisting of 295 tumors. The tumor grade ranges from 1-3, 1 labelling slowly growing cancer cells that resemble normal cells and 3 labeling fast and aggressively growing cancer cells. Angioinvasion (A.I.) defines tumor cells infiltrating blood vessels and extensive lymphocytic infiltrate (L.I.) refers to the buildup of white blood cells. ER+ and PR+ denote that the tumor cells grow in response to estrogen and progesterone, respectively. In these rows the numbers stand for the percentage of tumor cells in the tumor with according hormone receptors (19).

Age	Diameter (mm)	Followup time (yr)	Meta-stases	Grade	A.I.	ER+	PR+	L.I.	Brca1 mutation
43	25	12,53	0	2	1	80	80	0	0
44	20	6,44	0	1	0	50	50	0	0
41	45	10,66	0	3	1	10	5	0	0
41	20	13	0	3	0	50	70	1	0
48	20	11,96	0	3	1	100	80	0	0
49	13	11,16	0	2	0	80	80	0	0
46	20	10,14	0	1	0	80	50	0	0
48	28	8,8	0	3	0	0	0	1	0
48	15	10,29	0	3	0	60	80	0	0

Table 4.1: Biological parameters used for analysis.

4.2.2 Progression Analysis of Disease (*PAD*)

In the paper (11), the authors introduce *PAD* as a method that is able to detect geometric structures of data and visualize the findings in a very accessible way. *PAD* consists of two components: working with the raw data to identify a suitable filter function, and using `Mapper` to visualize the geometric structure of the data. Nicolau et al. used *Disease-Specific Genomic Analysis (DSGA)* as a way to construct the filter function. *PAD* results in a graph which represents the geometric shape of the data set, sometimes revealing meaningful characteristics of the data.

Disease-Specific Genomic Analysis (DSGA)

DSGA is a mathematical method that displays omic data[1] $\vec{D}$ as a sum of two terms: the *normal component* and the *disease component*.

$$\vec{D} = Nc.\vec{D} + Dc.\vec{D}$$

This decomposition is defined by using the so called *Healthy State Model (HSM)*. By substituting each normal tissue vector by a linear combination of all other normal tissue vectors, the *HSM* does not only de-sparse the data, but enabled the authors to define a subspace spanned by normal tissue vectors. Figure 4.3 displays the *HSM*. The deviation from normal tissue to diseased tissue can thereby be observed through the distance of the two subspaces.

[1] Omic data refers to biological data that belongs to a biological field ending with *-omics*. For example genomics, metagenomics or metabolomics.

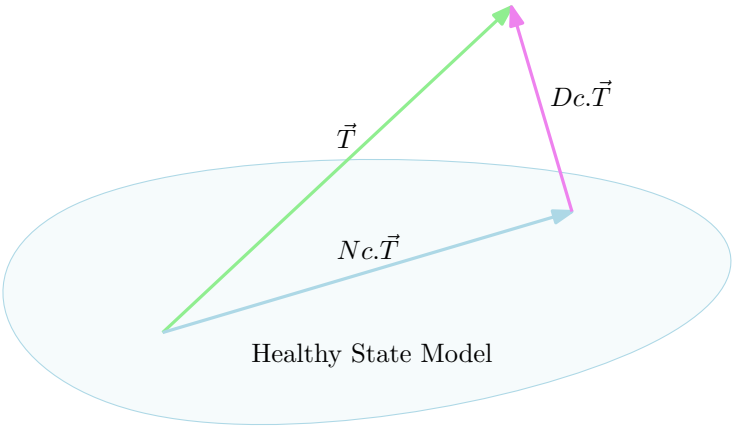


Figure 4.3: DSGA decomposition of the original tumor vector $\vec{T}$ into the normal component $Nc.\vec{T}$ as part of the Healthy State Model and the Disease component $Dc.\vec{T}$ orthogonal to the space spanned by normal components.

DSGA has the advantage of enabling us to use only the disease component for certain analyses. Working with the disease component of data outperforms methods using the original data and brings out unique biology of the tumor cells. In contrast to methods directly comparing normal and neoplastic data, DSGA shows the extent to which gene expression in the tumor data is abnormal.

Total Proceeding of *PAD* Performed on Breast Cancer Data

1. Input: The Data matrix

$$D_{k,m} = \begin{pmatrix} | & & | & | & & | \\ N_1 & \cdots & N_k & T_1 & \cdots & T_m \\ | & & | & | & & | \end{pmatrix}$$

includes the tumor data vectors T_j and the healthy tissue vectors N_i. Table 4.1 shows the measurements taken from tumor tissue. One tumor vector T_j corresponds to one row in the table.

2. DSGA-transformation: The $Dc.mat$ matrix contains the disease components of the tumor vectors:

$$Dc.mat = \begin{pmatrix} | & & | \\ Dc.T_1 & \cdots & Dc.T_m \\ | & & | \end{pmatrix}.$$

The $L1.mat$ matrix contains the leave-one out cross validation estimates of the deviation from healthy state by normal tissue data. We denote these adapted vectors by $L1.N_j$:

$$L1.mat = \begin{pmatrix} | & & | \\ L1.N_1 & \cdots & L1.N_k \\ | & & | \end{pmatrix}.$$

$L1Dc.mat$ denotes the final matrix:

$$L1Dc.mat = \begin{pmatrix} | & & | & | & & | \\ L1.N_1 & \cdots & L1.N_k & Dc.T_1 & \cdots & Dc.T_m \\ | & & | & | & & | \end{pmatrix}.$$

3. Adapt the final matrix $L1Dc.mat$ using a suitable threshold, which eliminates all diseased components that do not show a significant deviation from the normal tissue.

4. Prepare for `Mapper`: Select a biologically relevant filter function and use the columns of $L1Dc.mat$ as data points. In this analysis the authors considered the filter functions

   $$f_{p,k}(\vec{V}) = \big(\sum_{j=1}^{s} |g_j|^p\big)^{\frac{k}{p}}$$

   which measure the overall deviation from the *HSM*. Here

   $$\vec{V} = (g_1, \ldots, g_s)^T$$

   denotes the columns of the data matrix, where g_j are the genes, proteins, or other biological parameters of the sample we wish to compare.

5. Use `Mapper` on the data obtained in Step 3, using the filter function chosen in Step 4.

In their publication, Nicolau et al. (11) used the *PAD* analysis described above to analyze subtypes of breast cancer. In the following, we will briefly summarize their findings.

The results Nicolau et al. present in their publication are impressive. In Figure 4.4 we can identify a clear structure consisting of three distinct branches. Not only is this mathematically relevant, but a statistical analysis carried out afterwards, confirmed a biological distinction between these subgroups. The vertices (small colored dots) indicate a tumor subgroup, whereas the color of each vertex corresponds to the tissue's deviation from normal tissue. As stated in the key on the top left, dark blue nodes are tissue samples closest to normal tissue, while dark red nodes denote more neoplastic tissue. Biologically, the branches visible in Figure 4.4 differ in gene activity. The high value of the filter function found in the lower right branch, recognizes these tumors as most distinct from normal tissue. The tumors show high activity in the gene groups ER+ and c-MYB$^+$, relative to normal tissue.

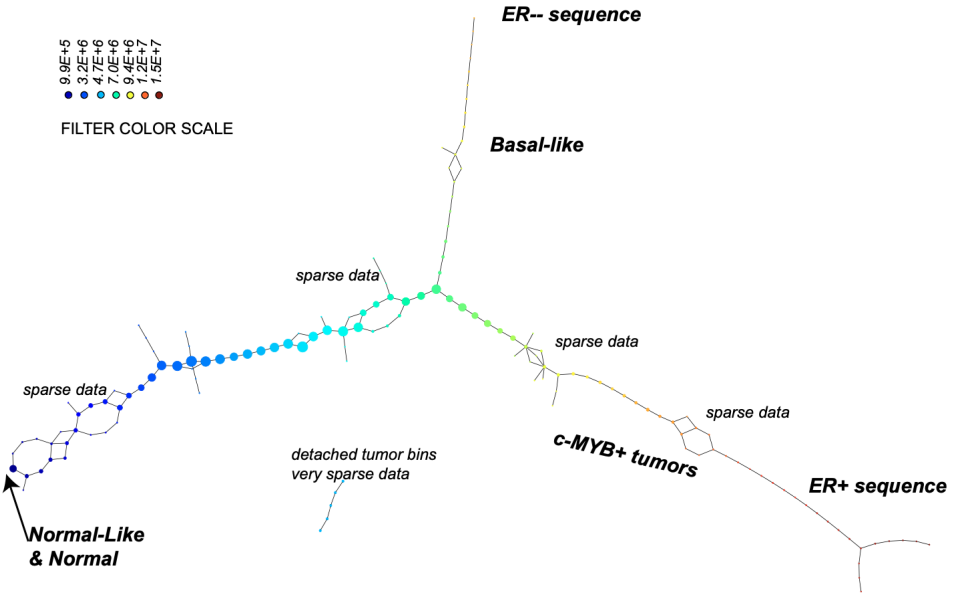


Figure 4.4: *PAD* analysis of the breast cancer data from M. Nicolau, A. Levine and G. Carlsson's paper (11).

In order to highlight the significance of this approach, Figure 4.5 shows the subgroups identified by a regular clustering approach. Clearly, the c-MYB$^+$ group (colored lines) is not distinguishable solely using clustering techniques.

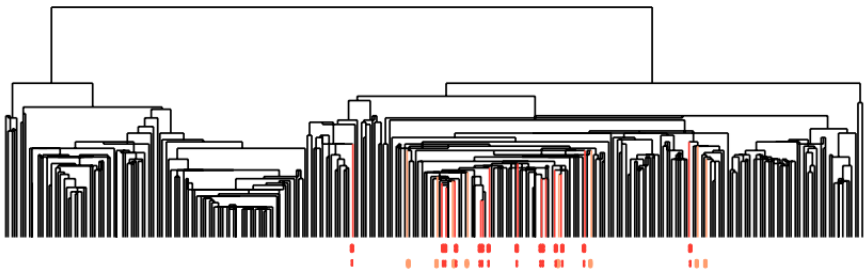


Figure 4.5: The dendrogram displays different tumor subgroups which were found using standard clustering approaches. Tumor subgroups contained in c-MYB$^+$ are colored in red (no outliers) and orange (including outliers).

Regular cluster analysis is able to find subgroups in data that are high in density, while sparse subgroups are rarely detected. The method `Mapper` presents a possible solution to this problem, as it is able to detect subtle differences in the data. Furthermore, the flexibility of the filter function allows us to analyze the data from multiple perspectives and tailor the analysis to the features we want to examine. Identifying the filter function is one of the, if not the main challenge when using `Mapper`. It requires in depth understanding of the data and the biological or medical significance of the parameters collected. Finding ways to make a suitable filter function more obvious, is a very interesting and important line of research, that could expand `Mapper`'s impact in data analysis.

Bibliography

[1] M. A. Armstrong, *Basic topology*, Springer, 1983.

[2] U. Bauer, M. Kerber, F. Roll, and A. Rolle, *A unified view on the functorial nerve theorem and its variations*, 2023.

[3] G. Carlsson, *Topology and data*, Bulletin of The American Mathematical Society, 46 (2009), pp. 255–308.

[4] ——, *Topological pattern recognition for point cloud data*, Acta Numerica, 23 (2014), pp. 289–368.

[5] G. Carlsson and M. Vejdemo-Johansson, *Topological Data Analysis with Applications*, Cambridge University Press, 2021.

[6] S. S. Fajans, *The definition of chemical diabetes*, Metabolism, 22 (1973), pp. 211–217. Chemical Diabetes Mellitus in Childhood.

[7] A. Hatcher, *Algebraic Topology*, Cambridge University Press, Cambridge, 2002.

[8] K. Jänich, *Topologie*, Springer-Verlag, 2013, p. 134.

[9] M. C. Jones and R. Sibson, *What is projection pursuit?*, Journal of the Royal Statistical Society. Series A (General), 150 (1987), pp. 1–37.

[10] C. F. Loughrey, P. Fitzpatrick, N. Orr, and A. Jurek-Loughrey, *The topology of data: opportunities for cancer research*, Bioinformatics, 37 (2021), pp. 3091–3098.

[11] M. Nicolau, A. J. Levine, and G. Carlsson, *Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival*, Proceedings of the National Academy of Sciences, 108 (2011), pp. 7265–7270.

[12] E. Paluzo-Hidalgo, R. Gonzalez-Diaz, M. A. Gutiérrez-Naranjo, and J. Heras, *Simplicial-map neural networks robust to adversarial examples*, Mathematics, 9 (2021).

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *Scikit-learn: Machine learning in Python*, Journal of Machine Learning Research, 12 (2011), pp. 2825–2830.

[14] G. M. Reaven, J. W. Farquhar, and R. H. Nakanishi, *Steady State Plasma Insulin Response to Continuous Glucose Infusion in Normal and Diabetic Subjects*, Diabetes, 18 (1969), pp. 273–279.

[15] G. M. Reaven and R. G. Miller, *An attempt to define the nature of chemical diabetes using a multidimensional analysis*, Diabetologia, 16 (1979), pp. 17–24.

[16] D. R. Sheehy, *Mesh Generation and Geometric Persistent Homology*, PhD thesis, Carnegie Mellon University, Pittsburgh, July 2011. CMU CS Tech Report CMU-CS-11-121.

[17] G. Singh, F. Mémoli, and G. Carlsson, *Topological methods for the analysis of high dimensional data sets and 3d object recognition*, 01 2007, pp. 91–100.

[18] K. Singh, *Oral glucose tolerance test (ogtt)*, the global diabetes community, (2022).

[19] T. A. C. Society, *Breast cancer hormone receptor status*, 2021.

[20] G. Tauzin, U. Lupo, L. Tunstall, J. B. Pérez, M. Caorsi, A. Medina-Mardones, A. Dassatti, and K. Hess, *giotto-tda: A topological data analysis toolkit for machine learning and data exploration*, 2020.

[21] L. J. van't Veer and et al., *Gene expression profiling predicts clinical outcome of breast cancer*, Nature, 415 (2002), pp. 530–536.

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Title of work:

The Nerve Theorem and its Applications in Topological Data Analysis

Thesis type and date:

Bachelor Thesis, June 19, 2023

Supervision:

Dr. Sara Kališnik Hintz

Student:

Name: Mohnhaupt Mona
E-mail: mmohnhaupt@student.ethz.ch
Legi-Nr.: 19-946-235

Statement regarding plagiarism:

By signing this statement, I affirm that I have read and signed the Declaration of Originality, independently produced this paper, and adhered to the general practice of source citation in this subject-area.

Declaration of Originality:

`http://www.ethz.ch/faculty/exams/plagiarism/confirmation_en.pdf`

Zurich, 11. 7. 2023: