ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

TOPOLOGICAL DATA ANALYSIS AND CLUSTERING ALGORITHMS IN MACHINE LEARNING

Ph.D. THESIS

İsmail GÜZEL

Department of Mathematical Engineering

Mathematical Engineering Programme

MARCH 2023

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

TOPOLOGICAL DATA ANALYSIS AND CLUSTERING ALGORITHMS IN MACHINE LEARNING

Ph.D. THESIS

İsmail GÜZEL
(509182203)

Department of Mathematical Engineering

Mathematical Engineering Programme

Thesis Advisor: Prof. Dr. Atabey KAYGUN

MARCH 2023

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

TOPOLOJİK VERİ ANALİZİ VE MAKİNE ÖĞRENİMİNDE KÜMELEME ALGORİTMALARI

DOKTORA TEZİ

İsmail GÜZEL
(509182203)

Matematik Mühendisliği Anabilim Dalı

Matematik Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Atabey KAYGUN

MART 2023

İsmail GÜZEL, a Ph.D. student of ITU Graduate School student ID 509182203 successfully defended the thesis entitled "TOPOLOGICAL DATA ANALYSIS AND CLUSTERING ALGORITHMS IN MACHINE LEARNING", which he prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Prof. Dr. Atabey KAYGUN** İstanbul Technical University	.....
Jury Members :	**Prof. Dr. Mustafa NADAR** İstanbul Technical University	.....
	Asst. Prof. Dr. Gül İNAN İstanbul Technical University	.....
	Prof. Dr. Özgür MARTİN Mimar Sinan Fine Arts University	.....
	Prof. Dr. Müge KANUNİ ER Düzce University	.....

Date of Submission : **2 January 2023**
Date of Defense : **13 March 2023**

To the memory of my father,
my spouse,
and children

FOREWORD

First of all, I want to sincerely thank my supervisor, Prof. Atabey KAYGUN, for sharing his wisdom and expertise with me during this thesis. I owe him a huge debt of gratitude for all the ways that his advice, support, constant presence, excellent leadership, encouragement, and never-ending patience have helped me. I also want to express my gratitude to Prof. Mustafa NADAR and Prof. Özgür MARTİN for their assistance with the thesis progress report. I am also grateful to Prof. Selçuk DEMİR for recommending me to pursue my Ph.D. with Atabey KAYGUN since everything about this thesis started with his advice.

I would like to dedicate a significant part of my gratitude to my colleague and dear friend Haydar Can KAYA, as our motivating coffee chats in break time had a great contribution to this thesis. I also want to express my gratitude to Dr. Alperen KARAN for our conversations on thorough topological data analysis and his really useful LaTeX thesis template. Faculty, staff, and friends at the Department of Mathematics at İTÜ are especially appreciated for giving me such a lovely workplace.

I would like to convey my thanks to Prof. Elizabeth MUNCH, who agreed to be my research supervisor during my visit to MunchLab at Michigan State University. It was a joy to work with her on our project and to attend her graduate course on Topological Data Analysis. My gratitude also goes to the MunchLab members. Our weekly meetings have advanced my knowledge of topological data analysis significantly.

This thesis was supported by Istanbul Technical University, Scientific Research Project (BAP, TDK-2020-42698). I would also like to thank the Scientific and Technological Research Council of Türkiye for supporting my research for this thesis through TÜBİTAK 2211 (1649B031701624) and TÜBİTAK 2214:A (1059B142000135), which provided research opportunities at Michigan State University.

Bu tezi en çok görmek isteyenlerden biri rahmetli babam Mehmet GÜZEL'i ve diğer zamansız ayrılanları rahmet, hüzün ve saygıyla anıyorum. Annem Ummahani GÜZEL'e ve diğer aile yakınlarıma her zaman desteklerini esirgemedikleri için teşekkürlerimi sunuyorum.

Finally, I would like to reserve my deepest gratitude to my beloved wife Berna GÜZEL who has touched my whole life and the deepest of my heart. She has been there for me through the highs and lows, and I could not have done it without her. I am so thankful for her patience and understanding as I worked late nights and weekends to complete my thesis. Aynı zamanda eşimin ailesine de teşekkürlerimi sunuyorum.

March 2023

İsmail GÜZEL
(Senior Researcher)

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

TOPOLOGICAL DATA ANALYSIS AND CLUSTERING ALGORITHMS IN MACHINE LEARNING

SUMMARY

In this dissertation, we define a new non-Archimedean metric (a.k.a. an ultra-metric) called *cophenetic metric* on persistent homology classes of all degrees using only homological information. Then, based on numerical experiments on different datasets, we statistically verify that the topological information coming from the zeroth persistent homology with our cophenetic metric is consistent with the information provided by different hierarchical clustering algorithms using different metrics. We also observe that the clusters we obtained via the cophenetic metric do yield competitive silhouette scores and the Rand indices in comparison with clusters obtained from other metrics.

The homological information about a filtered simplicial complex over the poset of positive real numbers is often presented by a barcode which depicts the evolution of the associated Betti numbers. However, there is wonderfully complex combinatorics associated with the homology classes of a filtered complex, and one can do more than just count them over the index poset. In this thesis, we show that this combinatorial information can be encoded by a filtered matroid, or even better, by rooted forests. We also show that these rooted forests can be realized as cobordisms.

The subject of this thesis is presented in Chapter 1, in which we also provide a survey of the literature and a brief introduction to topological data analysis (TDA).

In Chapter 2, we summarize the fundamental concepts from topological and metric spaces that we will need in subsequent chapters.

In Chapter 3, we will go more deeply into the mathematical foundations of clustering techniques. Clustering algorithms, especially hierarchical clustering algorithms, play a key role in this dissertation. Comparisons of various clustering strategies also play important roles in this thesis. For this reason, we also survey numerical metrics to assess such clustering strategies in the same chapter. We also work out a complete example on a collection of Turkish cities and the distances between them, and compare the resulting clusters.

The TDA calculations we make in this thesis heavily use simplicial and chain complexes, and their homologies. In all examples, we first calculate the homology of a filtered simplicial complexes in order to calculate its persistent homology. The primary computation procedure begins by distilling a point cloud into a filtered simplicial complex, followed by calculating the homology of the resulting filtered differential graded complex. Chapters 4 and 5 provide all the necessary background information on simplicial complexes, differential graded complexes, and homologies. We also present complete worked-out examples of calculating homology classes for a collection of simple simplicial complexes.

To capture the exact topological features of the point cloud, the main challenge is to find an optimal proximity parameter for the simplicial technology. *Persistent homology* is developed precisely to deal with this issue. It keeps track of how long each topological feature of the supplied data endures as the proximity parameter varies. It produces multisets of intervals represented as *barcodes*. We cover these concepts in Chapter 6.

Chapter 7 is devoted to developing a rigorous theoretical foundation for filtered simplicial and chain complexes. Along with developing such a foundation for filtered complexes, we also discovered an exciting combinatorial representation for homologies of filtered complexes in the form of *filtered matroids*, especially by employing dendrograms labeled by circuits in a matroid. In Chapter 8, we defined a new non-archimedean metric called *homological cophenetic distance* on homology classes which is the main contribution of this thesis to literature.

In Chapter 9 we are concerned with developing another presentation of the homological information coming from a filtered complex using *cobordisms* of punctured spheres, which are themselves punctured higher dimensional spheres.

In Chapter 10, we give a detailed account of how we used the cophenetic distance in our numerical experiments. In the first experiment, we compared the standard zeroth persistent homology representations (barcodes), the dendrogram we derived from the cophenetic distance, and the dendrogram coming from the hierarchical clustering algorithm with the Euclidean metric on a small synthetic dataset embedded in $\mathbb{R}^2$. In the second experiment, we studied the geographic coordinates of a small sample of Turkish cities. By following the same statistical comparison methodology as in the first experiment, we compare the dendrograms produced by cophenetic metrics on the zeroth homology and the dendrograms produced by hierarchical clustering algorithms using a variety of distance measurements in order to assess the validity of our study. In the third and last experiment, we applied the hierarchical clustering algorithms to various datasets (the dataset of Turkish cities, the Iris dataset, the Cancer Coimbra dataset, and two synthetic datasets) with different metrics including our cophenetic metric, and we statistically analyzed the clustering results.

Finally, in Chapter 11, we summarize and analyze the results we obtain, and the constraints of our approach such as need for high computational power and high memory requirements. We also discuss potential future research directions one can follow to extend this thesis such as designing suitable applications with real-world datasets, visualizing cobordisms, or developing filtered combinatorial simplicial complexes on a categorical dataset.

TOPOLOJİK VERİ ANALİZİ VE MAKİNE ÖĞRENİMİNDE KÜMELEME ALGORİTMALARI

ÖZET

Bu tez bir istatistiksel veri bilimi ve makine öğrenmesi tezidir. Tez, özel olarak, istatistiksel veri biliminin yeni bir alt disiplini olan topolojik veri bilimi (TVB) alanındadır. TVB istatistik, veri bilimi, genel ve cebirsel topoloji gibi birbirinden farklı ve uzak alanları sağlam bir teorik matematik tabanında birleştirip büyük hacimli ve yüksek boyutlu veriler konusunda yeni bakış açıları geliştiren yeni bir veri bilimi disiplini olup son yıllarda geniş kullanım alanı bulmuştur.

Bu tezde, sadece verilen sonlu örneklemler üzerinde tanımlanan komplekslerden gelen homolojik bilgi kullanılarak tüm derecelerdeki kalıcı homoloji sınıfları üzerinde *kofenetik metrik* olarak adlandırılan yeni bir Arşimedyen olmayan metrik (diğer bir deyişle ultra-metrik) tanımlıyoruz. Ardından, farklı veri kümeleri üzerinde yapılan sayısal deneylere dayanarak, kofenetik metriğimiz ile sıfırıncı kalıcı homolojiden gelen topolojik bilginin, farklı metrikler kullanan farklı hiyerarşik kümeleme algoritmaları tarafından sağlanan bilgi ile tutarlı olduğunu istatistiksel olarak doğruluyoruz. Ayrıca, kofenetik metrikle elde ettiğimiz kümelerin, diğer metriklerle elde edilen kümelere kıyasla rekabetçi silüet skorları ve Rand endeksleri verdiğini gözlemliyoruz.

Manifoldlar gibi sürekli geometrik nesnelerin simpleksel kompleksler gibi sonlu ve ayrık nesneler üzerinden incelenmesi topoloji tarihi kadar eskidir. Özellikle bir manifoldun homotopi tipi o manifold için verilmiş açık toplarla verilmiş bir örtüsünden okunabilir. Biz bu tezde, bu fikri bir manifolddan alınmış sonlu bir örneklem üzerinde tanımlanacak açık toplarla tanımlanmış reel sayılar üzerinde filtrelenmiş simpleksel komplekslere uygulayacağız.

Pozitif reel sayılar kümesi üzerinde filtrelenmiş bir simpleksel kompleks hakkındaki homolojik bilgi genellikle ilişkili Betti sayılarının gelişimini gösteren bir barkodla sunulur. Bununla birlikte, filtrelenmiş bir simpleksel kompleksin homoloji sınıflarıyla ilişkili harika bir karmaşık kombinatorik vardır ve bunları indeks poseti üzerinde saymaktan daha fazlası yapılabilir. Bu tezde, bu kombinatoryal bilginin filtrelenmiş bir matroid veya daha da iyisi köklü ağaçlar tarafından kodlanabileceğini gösteriyoruz. Ayrıca bu köklü ağaçların kobordizmalar (cobordisms) olarak gerçekleştirilebileceğini de gösteriyoruz.

Bu tezin konusu Bölüm 1'de sunulmuş olup, ve bu bölümde ayrıca topolojik veri analizine (TVA) de kısa bir giriş yapılmaktadır.

Bölüm 2'de, sonraki bölümlerde ihtiyaç duyacağımız veri analizinin temel kavramları olan topolojik ve metrik uzayları özetlenmektedir.

Bölüm 3'te, kümeleme tekniklerinin matematiksel temellerine daha derinlemesine gireceğiz. Kümeleme algoritmaları, özellikle hiyerarşik kümeleme algoritmaları bu tezde kilit bir rol oynamaktadır. Çeşitli kümeleme stratejilerinin karşılaştırmaları

da bu tezde önemli roller üstlenmektedir. Bu nedenle, kümeleme algoritmalarında kullanılan sayısal ölçütleri de araştırıyoruz. Bu tür kümeleme stratejilerini ve karşılaştırmalarını da aynı bölümde değerlendireceğiz. Ayrıca, tamamlayıcı bir örnek olarak, Türkiye'deki şehirlerin aralarındaki mesafeleri ele alarak ortaya çıkan kümeleri karşılaştırdık.

Bu tezde yaptığımız TVA hesaplamalarında ağırlıklı olarak simpleksel ve zincir kompleksleri, ve bunların homolojileri kullanılmaktadır. Tüm örneklerde, kalıcı homolojisini hesaplamak için önce filtrelenmiş bir simpleksel kompleksin homolojisini hesaplıyoruz. Birincil hesaplama prosedürü, bir nokta bulutunu filtrelenmiş bir simpleksel kompleks haline getirerek başlar ve ardından ortaya çıkan filtrelenmiş diferansiyel dereceli kompleksin homolojisini hesaplamaktır. Bölüm 4 ve 5, simpleksel kompleksler, diferansiyel dereceli kompleksler ve homolojiler hakkında gerekli tüm arka plan bilgilerini sağlamaktadır. Ayrıca, soyut simpleksel komplekslerin bir koleksiyonu için homoloji sınıflarının hesaplanmasına ilişkin eksiksiz çalışılmış örnekler sunuyoruz.

Nokta bulutunun topolojik özelliklerini tam olarak yakalamak için temel zorluk, simpleksler teknolojisi için en uygun yakınlık parametresini bulmaktır. *Kalıcı homoloji* (Persistent homology) tam olarak bu sorunla başa çıkmak için geliştirilmiştir. Yakınlık parametresi değiştikçe, sağlanan verilerin her bir topolojik özelliğinin ne kadar süre dayandığını takip eder. Bu, *barkodlar* olarak temsil edilen aralıkların çoklu kümelerini üretir. Bu kavramları Bölüm 6'da ele alıyoruz.

Bölüm 7, filtrelenmiş simpleksel ve zincir kompleksler için titiz bir teorik temel geliştirmeye adanmıştır. Filtrelenmiş kompleksler için böyle bir temel geliştirmenin yanı sıra, özellikle bir matroid'deki devreler (circuits) tarafından etiketlenen dendrogramları kullanarak, filtrelenmiş komplekslerin homolojileri için *filtrelenmiş matroidler* şeklinde heyecan verici bir kombinatoryal temsil keşfettik. Bölüm 8'de, bu tezin literatüre ana katkısı olan homoloji sınıfları üzerinde *homolojik kofenetik uzaklık* adı verilen yeni bir Arşimedyen olmayan metrik tanımladık.

Bölüm 9'de, kendileri de yüksek boyutlu küreler olan delinmiş kürelerin *kobordizmalarını* kullanarak filtrelenmiş bir kompleksten gelen homolojik bilginin başka bir sunumunu geliştirmekle ilgileniyoruz.

Bölüm 10'da, sayısal deneylerimizde kofenetik mesafeyi nasıl kullandığımıza dair ayrıntılı bir açıklama sunuyoruz. İlk deneyde, $\mathbb{R}^2$ içine gömülü küçük bir sentetik veri kümesi üzerinde standart sıfırıncı kalıcı homoloji gösterimlerini (barkodlar), kofenetik uzaklıktan türettiğimiz dendrogramı ve hiyerarşik kümeleme algoritmasından gelen dendrogramı Öklid metriği ile karşılaştırdık. İkinci deneyde, küçük bir Türkiye şehri örneğinin coğrafi koordinatlarını inceledik. İlk deneyde olduğu gibi aynı istatistiksel karşılaştırma metodolojisini izleyerek, çalışmamızın geçerliliğini değerlendirmek için sıfırıncı homoloji üzerinde kofenetik metrikler tarafından üretilen dendrogramları ve çeşitli mesafe ölçümleri kullanan hiyerarşik kümeleme algoritmaları tarafından üretilen dendrogramları karşılaştırdık. Üçüncü ve son deneyde, hiyerarşik kümeleme algoritmalarını çeşitli veri kümelerine (Türkiye şehirleri, Iris, Kanser Coimbra ve iki farklı sentetik veri kümeleri) kofenetik metriğimiz de dahil olmak üzere farklı metriklerle uyguladık ve kümeleme sonuçlarını istatistiksel olarak analiz ettik.

Son olarak, Bölüm 11'de, elde ettiğimiz sonuçları, yaklaşımımızın yüksek hesaplama gücü ve yüksek bellek gereksinimi gibi kısıtlamalarını özetliyor ve analiz ediyoruz.

Ayrıca bu bölümde, bu tezi genişletmek için, gerçek dünya veri kümeleriyle uygun uygulamalar tasarlamak, kobordizmaları görselleştirmek veya kategorik veri kümesi üzerinde filtrelenmiş kombinatoryal simpleksel kompleksler geliştirmek gibi izlenebilecek potansiyel gelecek araştırma yönlerini tartışıyoruz.

1. INTRODUCTION

1.1 Background

There is an explosion of analyzable data coming from applied sciences and engineering. However, it is not always easy to extract useful information from collected data. Particularly, examining high-dimensional or high volume data is quite challenging.

There is a plethora of statistical and machine learning approaches that aim to reveal any intrinsic geometric structure of the data at hand. Examples include regression methods, clustering algorithms, and dimensionality reduction techniques. Most of these approaches suppose that the underlying structure of the data has a very simple metric geometry. However, in most applications, the underlying intrinsic structure of the data can be highly complex, or worse, noisy. Since topology, unlike geometry, is less impervious to the underlying metric structure, data analytic tools relying on the topology instead of the metric of the ambient space might yield deeper insights into the data less sensitive to complexity or noise.

In this thesis, we concentrate on one possible avenue in analyzing high dimensional and high volume data called *topological data analysis* (TDA). TDA appeared as a new mathematically rigorous approach in statistical data analysis and machine learning that is gaining popularity. In Figure 1.1 we present the number of articles published each year using the keywords "Topological Data Analysis" and "Persistent Homology" from the SCOPUS database.

The main goal of any approach within the umbrella of TDA is to apply algebraic topological and computational geometrical tools and techniques to study *the intrinsic* or *the extrinsic shape* of data. Here, we use the term *intrinsic shape* to mean the distributional and topological structure of the data before we perform any feature engineering on the data, while we use the term *extrinsic shape* to mean the

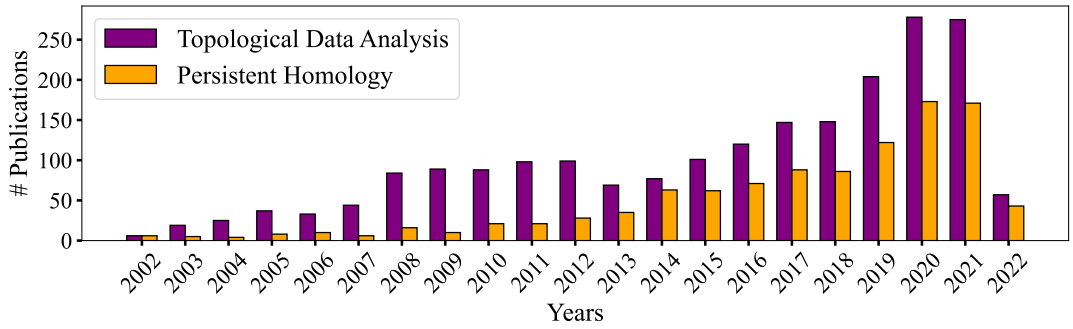


Figure 1.1 : The trends of the keywords *Topological Data Analysis* and *Persistent Homology*. The data is taken from SCOPUS.

distributional and topological structure of the data after we perform feature engineering on the data.

Suppose that we have a cloud of points X in $\mathbb{R}^n$ sampled from an embedded circle $S^1 \to \mathbb{R}^n$. TDA would aim to answer the question "How can we detect the shape of the original manifold S^1 embedded in $\mathbb{R}^n$ by only looking at the point set X we sampled?" The main difficulty lies in the fact that the finite set of sample points X tries to approximates a continuous structure, in this case, an embedded circle in $\mathbb{R}^n$.

This type of research is actually quite old. There is an old and long vein of topological research that replaces continuous structures by finite discrete ones such as *piece-wise linear manifolds* [2], *simplicial complexes* [3], *CW-complexes* [4], *Čech complexes* [5] and *Vietoris-Rips complexes* [6]. In this line of inquiry, homological invariants, such as kth Betti numbers β_k of a simplicial complex for a natural number k are import topological invariants [7]. For example, the zeroth Betti number β_0 tells us the number of connected components, and β_1 tells the number of embedded circles within a simplicial complex which is a finite approximation of a continuous manifold.

For the embedded circle $S^1 \to \mathbb{R}^n$ we considered above, we are going to construct a simplicial complex $\mathscr{R}(X)$ out of the set X we sampled within a fixed scale parameter, and then calculate its homology $H_k(\mathscr{R}(X))$ for $k = 0, 1$. If we construct $\mathscr{R}(X)$ appropriately we can detect that X was sampled from $S^1 \to \mathbb{R}^n$ if we see that $\beta_0 = 1 = \beta_1$. However, choosing *one right simplicial complex* for X is not an easy task.

To solve this problem, TDA proposes to use *persistent homology*. Instead of studying the topological features of point clouds within a fixed distance scale, one can change the scale over time and then record how long these topological features are retained.

Persistent homology was first introduced for topological simplification of alpha shapes in [8] and then was extended to arbitrary dimensional spaces over arbitrary fields by using classification methods in [7]. Since the records that persistent homology keeps are finite collections of intervals, the concept of *barcodes* are first introduced in [9] and [10] to visualize these records. Also in [9], it was presented a new approach which combined geometric and topological techniques in order to detect the sharp features of space, such as edges and corners, by using tangent complexes of clouds of data points.

TDA, specifically persistent homology, has found a place in a wide range of science and engineering applications that include time series, pattern recognition, sensor networks, medicine, time series, and more [11–27]. A number of studies [28–32] addressing surveys and various application facets of the discipline are also available for the reader.

1.2 Problem Statement

In the standard representation of persistent homology, one only keeps a record of the persistent homological classes in the form of *life-time intervals*, (a. k. a. barcodes). However, persistent homology classes carry a very rich combinatorial structure, and one can do more that just counting them. In reference to the dendrograms of hierarchical clustering algorithms, Carlsson expresses the same idea as a question in [33, Ch.8] and [33]:

> The dendrogram can be regarded as the "right" version of the invariant π_0 in the statistical world of finite metric spaces. The question now becomes if there are similar invariants that can capture the notions of higher homotopy groups or homology groups.

Since hierarchical clustering schemes, and therefore, dendrograms and non-Archimedean metrics are known to be equivalent by Carlsson [34], to answer this question it is clear that one needs to define a non-Archimedean metric for homology or homotopy classes.

1.2.1 Our contributions

The main contribution of this thesis is a new non-Archimedean metric(called as *homological cophenetic distance*) defined on persistent homology classes in all degrees using purely homological information coming from the changing scale parameter. For this metric, we analyze how persistent homology classes of a certain degree "merge" on top of recording the birth and death times of these classes as the scale parameter changes. We observe that since all data points naturally appear as zeroth degree persistent homology classes, one can compare our metric with standard distance measures with respect to their performances in machine learning algorithms that rely on distance measures on data points.

We tested the soundness of our studies by checking whether hierarchical clustering methods with different metrics do indeed yield statistically verifiable commensurate topological information. We observe that when we measure the clusters obtained from the homological cophenetic metric against the clusters coming from other metrics we obtained competitive results. Furthermore, the metric we define on the persistent homology can now be used to sketch rooted tree presentations of the persistent homology classes in all degrees that track how these classes *merge* as the scale parameter changes.

In this thesis, we also provide a brand-new combinatorial description for homological groups that fluctuate over a scale parameter, which we term *the cophenetic matroid*. This definition neatly falls in the middle between filtered vector spaces and barcodes. Henselman and Ghrist employed filtered matroids previously (See [35]), but their focus was on creating and implementing effective cosheaf homology computation techniques. Additionally, they continued to represent their persistent homology calculations using barcodes. We also demonstrate how these filtered matroids may be represented using rooted forests that originate from certain cobordisms of disjoint unions of spheres.

In forming the bridge between hierarchical clustering and persistent homology, we also found that the answer to the question raised by Carlsson [33] comes from algebraic topology: cobordisms. Dendrograms are 1-dimensional cobordism classes of disjoint

union of points. For higher homology classes, one has to resort to $n+1$-dimensional cobordisms of disjoint unions of n-spheres. For example, for the persistent homology in degree 1 such cobordisms are given by oriented genus-g Riemann surfaces with finitely many punctures, and the classification of such 2-manifolds is complete. Unfortunately, in dimensions 2- and higher such cobordisms are very difficult to classify.

1.2.2 Prior art

TDA is a new data analysis discipline whose fundamentals straddle both very abstract and concrete sub-disciplines of the mathematical research. Even though the theoretical roots TDA are firmly placed in algebraic topology, to solve its computational needs it heavily uses computational geometry and numerical linear algebra. In its simplest form, TDA aims to come up with reliable conclusions about the topological structure of a space from a random finite sample taken from this space. Since TDA relies on the topology rather than a particular metric structure of the ambient space from which data is sampled, in theory, it is more suitable for extracting information from data for which a canonical metric is not clear from the context, or worse yet, does not exist.

The fact that one can capture the homological and homotopical invariants of a space from a finite collection of points sampled from the space under certain guarantees is an old idea [36]. However, in the absence of any information whether these guarantees are satisfied, one has to construct a sample of invariants from available *local* information by playing with the notion of *proximity* via a scale parameter that we alluded above. Since we do not know which range of scale parameters truly capture the topological invariants of the underlying space, one must calculate these invariants at different scale parameters and investigate how these different calculations fit with each other.

It is perhaps a historical coincidence that the development of persistent homology mirrors that of the ordinary homology. In the beginning topologists calculated homology as Betti numbers, and it was Emmy Noether who first observed that these homological invariants had to be considered as abelian groups [37]. Similarly, in the beginning, the practitioners represented persistent homology as barcodes which are records of how Betti numbers evolve as the scale parameter varies. Then it is clear that

we must consider persistent homology as a filtered abelian group should we make the same leap.

Clustering algorithms, on the other hand, have been around for a long time and they form an important and well-understood class of machine learning algorithms [38–41]. They are known to be equivalent to non-Archimedean metrics [34]. For a given data set, these algorithms aim to deliver an optimal partition where subsets are supposed to show a high degree of heterogeneity between, and a high degree of homogeneity within each subset. However, one has to make unavoidable ad-hoc choices to determine an optimal hierarchical clustering algorithm for any data set at hand due to the impossibility result of Kleinberg [42].

Similar to clustering algorithms, the TDA methods we investigate in this paper also rely on a changing scale parameter. But instead of relying on the metric structure alone, these methods propose using *persistent homology* to compute topological invariants of a data set. The topological invariants that persistent homology identifies are *the Betti numbers* defined for every natural number n. For instance, the Betti numbers for $n = 0, 1$ and 2 indicate respectively the number of connected components, 2- and 3-dimensional holes within the data set. The information that persistent homology yields on the change in topological features as the filtration scale parameter increases is can be presented in different ways. Barcodes are the most commonly used representations of persistent homology classes in which one keeps a record of finite collections of scale parameter intervals over which individual persistent homology classes *persist* [9, 10]. Even though there is now a plethora of different representations for the output of the persistent homology, such as persistence diagrams [43], landscapes [44], images [45], terraces [46], curves [47] and persistent entropy [48], they all are derived from the barcode representation.

Hierarchical clustering algorithms that we consider in this paper extract their results based solely on the metric structure of the ambient space where the data set is embedded. One can also statistically test the stability and convergence of these methods [34, 49]. In addition, they use a convenient tree representations, called *dendrograms*, to display the information on how these clusters merge as the underlying scale parameter changes [50]. One can compare dendrograms coming from different variants of clustering algorithms using a suitable metric [40, 51, 52], or even improve

forecasts by comparing dendrogram-like features obtained from different hierarchical clustering methods on the same data set [53].

In [54], the authors develop diagrams called *mergegrams* derived out of the dendrograms of the hierarchical clustering algorithms as a replacement for the zeroth barcodes with a great application in recognizing isometry classes [55]. However, unlike the mergegrams that depend on the metric structure of the ambient space and work with the zeroth barcodes only, our non-Archimedean metric relies on purely homological information and are defined for all persistent homology degrees. More importantly, they can be used for purposes other than sketching rooted tree presentations.

1.3 Originality and Impact

Since its conception, topology had a rich toolkit to approximate continuous structures with discrete and combinatorial objects. Such tools of topology in conjunction with computational geometry have recently found practical use in statistical data analysis in the last decade, and their popularity has been showing steep rise in many areas such as time series analysis to medical image analysis. TDA utilizes persistent homology to extract information about extrinsic and intrinsic shape of high dimensional, high volume data where regular means of extracting information is difficult particularly when one does not have a canonical way of comparing similarities of data points.

By using the birth and death times of topological features, researchers constructed the vectorizations of the persistent diagrams to make sense of the topological information one extracts from a cloud of data points as we alluded above. These representations, in turn, allow researchers study statistical properties of these topological features. Moreover, one can also apply many machine learning algorithms on these representations to extract pertinent information on data [56–59].

The barcode representation of persistent homology classes tabulates homology classes individually, however, does not contain any information on relationships between these classes. As the filtration parameter $\varepsilon > 0$ grows, homology classes may appear, disappear, or merge. The information, and the resulting combinatorics of this evolution is completely absent in the barcode representation, and potentially contains relevant

information on the shape and structure of the underlying data set. More importantly, if it is done properly, one can also investigate the statistics of the newly obtained information.

In this thesis we showed, for the first time, that there is a relation between hierarchical clustering and the barcodes for the zeroth persistent homology.[1] Our methods and techniques are already suitable for studying higher Betti numbers through a combinatorics of mergers, splittings, births and deaths of persistent homology classes represented as dendrograms. The articles we wrote on the subject [60] that formed the basis of this thesis are already received well during the presentations we made from these articles.

1.4 The Problems Answered in This Thesis

Below are the problems that we answered in this thesis.

Problem 1. What are the similarities and differences between hierarchical clustering and non-archimedean metrics on zeroth persistent homology? Can one develop a visualization method for the zeroth persistent homology similar to dendrogram representations coming from hierarchical clustering models?

Problem 2. Can one extend the bridge one can build between zeroth barcodes and dendrogram representations of hierarchical clustering to higher dimensional persistent homology, and higher Betti numbers?

Problem 3. Can one construct a theoretical framework for the interactions between the persistent homology classes? And also can one find the dendrogram-like representation for persistent homology classes of higher homology degrees?

Problem 4. Are the results we are going to obtain from our theoretical findings going to be in agreement with the results we are going to obtain from our numerical experiments on the synthetic and real datasets?

[1] G. Carlsson, one of the progenitors of TDA, indicates this connection in passing in [33].

2. METRIC SPACES

In this section, we are going to summarize the basic concepts from topological spaces and metric spaces that we are going to need in later chapters.

2.1 Topological Spaces

Definition 2.1. A topology on a set X is a family $\mathscr{O}$ of subsets of X that satisfies the three following conditions:

i) the empty set $\emptyset$ and X are elements of $\mathscr{O}$

ii) any union of elements of $\mathscr{O}$ is an element of $\mathscr{O}$,

iii) any finite intersection of elements of $\mathscr{O}$ is an element of $\mathscr{O}$

The set X together with the family $\mathscr{O}$, whose elements are called open sets, is a topological space.

2.2 Metric Spaces

Definition 2.2. A metric (or distance) on a non-empty set X is a map $d : X \times X \to [0, +\infty)$ satisfying:

i) for any $x, y \in X$, $d(x,y) = d(y,x)$,

ii) for any $x, y \in X$, $d(x,y) = 0$ if and only if $x = y$,

iii) for any $x, y, z \in X$, $d(x,z) \leqslant d(x,y) + d(y,z)$

The set X together with d is called a metric space, denoted by (X,d).

2.3 Non-Archimedean Metric Spaces

The inequality in (iii) is known as the triangle inequality. There is also a strong version of that inequality which is called strong triangle inequality (is also called ultra-metric inequality) as defined by

$$d(x,y) \leqslant \max\{d(x,z), d(z,y)\}, \text{ for any } x,y,z. \tag{2.1}$$

If the last condition in the definition to be metric is change with the ultra-metric inequality, That metric is called *non-Archimedean* metric. One of the key consequences of the ultra-metric inequality is that every triangle in a non-Archimedean metric space is isosceles:

Lemma 2.3. Suppose that d is a non-Archimedean metric on a space X. For the points x, y, and z points of X,

$$d(x,y) < d(x,z) \Longrightarrow d(x,z) = d(y,z). \tag{2.2}$$

Proof. From the assumption to be non-Archimedean, we have

$$d(y,z) \leqslant \max\{d(x,y),\ d(x,z)\} = d(x,z). \tag{2.3}$$

On the other hand, we have

$$d(x,z) \leqslant \max\{d(x,y),\ d(y,z)\} = d(y,z), \tag{2.4}$$

since otherwise we would have $d(x,z) \leqslant d(x,y)$ which would be contradiction. Therefore, we have

$$d(y,z) \leqslant d(x,z) \leqslant d(y,z), \tag{2.5}$$

and thus we have $d(x,z) = d(y,z)$.

□

2.4 Open and Closed Balls

Definition 2.4. Let (X,d) be a metric space. Then open balls and closed balls with center x and a fixed radius r are defined as follows:

i) An open ball: $B_r(x) = B(x,r) = \{y \in X : d(x,y) < r\}$

ii) A closed ball: $\overline{B}_r(x) = \overline{B}(x,r) = \{y \in X : d(x,y) \leqslant r\}$

The smallest topology containing all the open balls $B(x,r)$ is called the metric topology on X induced by d. For example, the standard topology in an Euclidean space is the one induced by the metric defined by the norm: $d(x,y) = \|x-y\|$.

2.5 Homeomorphisms

Definition 2.5. Let X and Y be two topological spaces. X and Y are homeomorphic if there exists a bijection (one-to-one and onto) $f : X \to Y$ such that f and f^{-1} are continuous.

If two topological spaces are homeomorphic then they share common topological properties such as being compact, connected, or Hausdorff, or sharing the same homotopy and homology groups.

2.6 Homotopy Equivalences

Definition 2.6. Two maps $f_0, f_1 : X \to Y$ are homotopic if there exists a continuous map $H : X \times [0,1] \to Y$ such that $H(x,0) = f_0(x)$ and $H(x,1) = f_1(x)$ for all $x \in X$.

Example 2.7. If Y is a convex subset of some $\mathbb{R}^n$ and if $f_0, f_1 : X \to Y$ are any two continuous functions, we can write a homotopy equivalence

$$H(x,t) = (1-t)f_0(x) + tf_1(x), \text{ where } t \in [0,1] \tag{2.6}$$

and H would be a homotopy between f_0 and f_1.

Definition 2.8. Two topological spaces X and Y are homotopy equivalent, $X \simeq Y$ if there exists continuous maps $f : X \to Y$ and $g : Y \to X$ such that

i) $g \circ f$ is homotopic to the identity map $\mathscr{I}_X$ in X.

ii) $f \circ g$ is homotopic to the identity map $\mathscr{I}_Y$ in Y.

If the topological space Y is subset of the another topological space X and if there exists a continuous map $H : X \times [0,1] \to X$ such that

i) $H(x,0) = x.$ for all $x \in X$,

ii) $H(x,1) \in Y$ for all $x \in X$,

iii) $H(y,t) \in Y$ for all $t \in [0,1]$ and for all $y \in Y$.

then X and Y are homotopy equivalent. An important special case is when Y contains a single point. In that case we say X is contractible to a point.

Proposition 2.9. Any convex subset X of $\mathbb{R}^n$ is contractible.

Proof. Remember that a space X is said to be contractible if the identity map $1_X : X \to X$ is homotopic to a constant map. Suppose that X is a convex subset of $\mathbb{R}^n$, and let $c \in X$. Let $g : X \to X$ be the constant map such that $g(x) = c$. Then, we can define a map $H : X \times [0,1] \to X$ as follows:

$$H(x,t) = tc + (1-t)x \tag{2.7}$$

is a homotopy between the identity map $1_X = H(x,0)$ and $g = H(x,1)$. Since the identity map is homotopic to a constant map, the subset X is contractible. □

Homotopy is a very important invariant in algebraic topology. The notion of *being homotopy equivalent* can be interpreted as one space can be *continuously deformed* into the other.

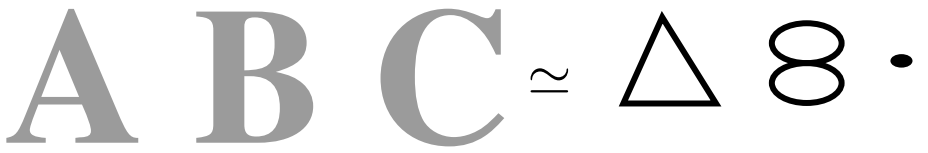


Figure 2.1 : Homotopy equivalent.

Definition 2.10. If one replaces the conditions *iii*) by $H(y,t) = y$ for all $t \in [0,1]$ and for all $y \in Y$ then H is a deformation retract of X onto Y.

In Figure 2.1, the left space consist of the first three letters A, B, and C is homotopy equivalent to the right space composed of the triangle, eight symbol, and a point, respectively.

3. CLUSTERING ALGORITHMS AND THEIR EVALUATION

Clustering algorithms, specifically hierarchical clustering algorithms and their comparisons form a crucial part of this thesis. In this chapter, we are going to dig deeper into the mathematical underpinnings of clustering algorithms. Moreover, to compare these clustering algorithms we are going to need numerical measures to evaluate the resulting clusters. To this end, in this chapter, we are also going to analyze different numerical measures of partitions of finite sets embedded in metric spaces. As an example, we are going to apply several hierarchical clustering algorithms to a set of Turkish cities using geographical distances between them, and evaluate the aforementioned cluster measures for the resulting clusters.

3.1 Hierarchical Clustering

Assume we have a connected metric space (X,d), and let $\pi_0(X)$ be the set of connected components of X. Assume we have a finite random sample of points $D \subseteq X$ taken from X whose distribution we do not know. Our aim is to deduce any information about the set of connected components of X using D. We are going to do this by finding a *finite clustering* of D which is a set function $c\colon D \to \mathbb{N}$ such that each cluster $c^{-1}(i)$ lies within a distinct connected component for each $i \in \mathbb{N}$.

Clustering algorithms are a subclass of unsupervised machine learning algorithms that aim to divide a given data set into disjoint subsets such that each group is homogeneous in itself but we observe a high degree of heterogeneity within clusters. Hierarchical clustering is a specific clustering algorithm that uses the metric structure of the ambient space in which our cloud of data points are embedded. As such, hierarchical clustering is easy to describe and implement. It also has a nice tree-based representation, called *a dendrogram*, of the clusters it produces. However, since the time complexity of the algorithm is $\mathcal{O}(n^2)$, hierarchical clustering algorithms are not particularly very efficient.

There are two types of hierarchical clustering algorithms where each variant takes a parameter p that indicates the number of clusters that our data set needs to be split. In *agglomerative hierarchical clustering*, we start from individual points as nodes and combine clusters until p groups remains. In *divisive hierarchical clustering*, we start with a single cluster and then we divide it into disjoint clusters until we arrive p groups. One can find a detailed treatment of the subject in [61].

In its simplest form, in hierarchical clustering we have a function $c_\varepsilon : D \to \mathbb{N}$ for each scale parameter $\varepsilon > 0$. This function satisfies $c_\varepsilon(x) = c_\varepsilon(y)$ for any two points $x, y \in D$ when there is a sequence of points $x_0, \ldots, x_m \in D$ such that $d(x_i, x_{i+1}) < \varepsilon$ for every $i = 0, \ldots, m-1$ where $x_0 = x$ and $x_m = y$. Notice that the clustering algorithm is monotone in the sense that if $c_\varepsilon(x) = c_\varepsilon(y)$ then $c_\eta(x) = c_\eta(y)$ for every $\eta > \varepsilon$. Moreover, since $D \subseteq X$ is finite and X is connected, there is a large enough scale parameter $\varepsilon > 0$ such that the image of c_ε is a single cluster.

Input: Observations $\{x_1, \ldots x_n\}$ and distance measure $d(G_1, G_2)$
Output: Dendrogram
Result: A hierarchy of cluster
for $i \leftarrow 1$ *to* n **do**
  $G_i \leftarrow \{x_i\}$;
end
while *# of cluster* > 1 **do**
  Keep $\arg\min_{G_i, G_j} d(G_i, G_j)$, $\forall$ cluster ;
  Extract G_i and G_j ;
  Join G_{ij};
end

Algorithm 1: Agglomerative clustering pseudo-code.

This is also called *single-linkage* clustering where we consider all pairwise of distance between the observations in G_1 and the observations in G_2. The smallest one is selected for the distance between two cluster. In other words,

$$d(G_1, G_2) := \min_{\substack{i \in G_1 \\ j \in G_2}} d_{ij} \tag{3.1}$$

One can use different metrics (linkages) to measure distances between clusters.

3.2 Linkages in Hierarchical Clustering

As we increase the scale parameter $\varepsilon > 0$ we start forming *clusters* of points. Since we replace points with clusters, we are going to need to calculate distances between clusters. See Algorithm 1. While the distance between points is given by the underlying metric of the space, the distance between two groups G_1 and G_2 can be calculated in different ways called *linkages*.

For a fixed $\varepsilon > 0$, let us use $C_i = c_\varepsilon^{-1}(i)$ to denote a cluster, and set $n_i = |C_i|$. Let us use d_{ij} for the distance between the cluster C_i and C_j. Lance and Williams [38] introduced to the following general formula for calculating distances between clusters

$$d_{(ij)k} = \alpha_{ijk} d_{ik} + \alpha_{jik} d_{jk} + \beta_{ijk} d_{ij} + \gamma |d_{ik} - d_{jk}| \tag{3.2}$$

for parameters α_{ijk}, β_{ijk} and γ to be determined. Here, $d_{(ij)k}$ denotes the distance between the clusters C_k and $C_{ij} = C_i \cup C_j$ which is merged in a single cluster. We list the parameters for commonly used methods of calculating distances between clusters in Table 3.1. See [38] for details.

Table 3.1 : Commonly used methods to determine $d_{(ij)k}$.

Linkage	α_{ijk}	β_{ijk}	γ
Single	$\frac{1}{2}$	0	$-\frac{1}{2}$
Complete	$\frac{1}{2}$	0	$\frac{1}{2}$
Average	$\frac{n_i}{n_i + n_j}$	0	0
Ward	$\frac{n_i + n_k}{n_i + n_j + n_k}$	$\frac{-n_k}{n_i + n_j + n_k}$	0

3.3 Comparisons of Clustering Schemes

As we stated above, we need to compare different clustering schemes and resulting dendrograms. For this purpose, there are several statistical tools to compare results of hierarchical clustering including Cophenetic matrix, Mantel test, and tanglegrams.

3.3.1 Cophenetic matrix

An important notion we need in studying and comparing clustering methods is *the cophenetic matrix* as defined in [39, 40, 62].

Assume we have a clustering function $c_\varepsilon : D \to \mathbb{N}$, and let $\mathscr{C} = \{C_i = c_\varepsilon^{-1}(i) \mid i \in \mathbb{N}\}$. Let ε_{ij} be the proximity level at which the clusters C_i and C_j merge to form C_{ij} for the first time. We record these numbers in the cophenetic matrix $C_\varepsilon(D) = (\varepsilon_{ij})$ for any pair of clusters C_i and C_j. The cophenetic distance is a metric under the assumption of monotonicity [63].

3.3.2 Mantel test

We are going to use the Mantel test as defined in [64]. It is commonly used in biology and ecology to compare phylogenetic trees. Mantel test is a non-parametric statistical method that computes the significance of correlation between rows and columns of a matrix through permutations of these rows and columns in one of the input distance matrices.

We consider two distance or cophenetic matrices $D_1 = (x_{ij})$ and $D_2 = (y_{ij})$ of size $n \times n$. The normalized Mantel statistic r is defined as

$$r = \frac{2}{(n-2)(n+1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \left(\frac{x_{ij} - \bar{x}}{s_x} \right) \left(\frac{y_{ij} - \bar{y}}{s_y} \right) \tag{3.3}$$

where

(i) $\bar{x}$ and $\bar{y}$ are averages of all entries of each matrix, and

(ii) s_x and s_y are the standard deviations for x and y.

The test statistic is the Pearson product-moment correlation coefficient $r \in [-1, 1]$. Having a value in the neighborhood of -1 indicates strong negative correlation whereas $+1$ indicates strong positive correlation, and 0 indicates no relation.

In order to estimate the sampling distribution of the standardized Mantel statistic under the null-hypothesis (no correlation between the distance matrices), random permutations of the rows (or equivalently columns) of the distance matrices are used to get a set of values of the statistic. Then whether the null-hypothesis is rejected depends on the value of the Mantel statistic: If the calculated statistic is unlikely to have been

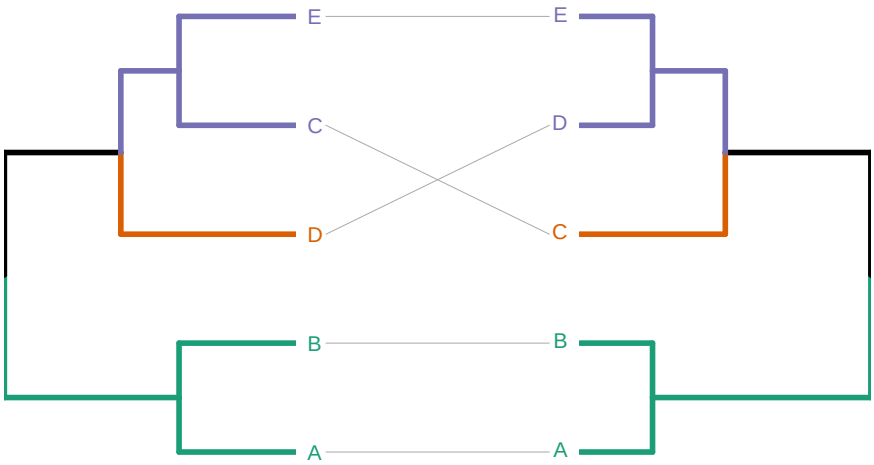


Figure 3.1 : A tanglegram example with entanglement 0.05 with L_2 norm.

obtained under the null-hypothesis then the null-hypothesis is rejected. See [41, Sect. 10.5] for details.

3.3.3 Tanglegram

To compare two different hierarchical clustering approaches, there is also a visualization tool which commonly used in biology is called tanglegram [65, 66]. In a tanglegram plot, two dendrograms with the same set of labels are presented one facing the other and with the labels connected by lines. One can see an example in Figure 3.1 that different clusters to which the observations we assinged are shown by different colors.

One can compare the tree structures using a metric derived from matches between labels placed on branches [66–68]. Entanglement is a quality measure of the alignment of two dendrograms in the tanglegram layout. It is calculated by giving the labels on the left dendrogram values ranging from 1 to the total number of labels in the dendrogram, then matching those number with the labels on the right dendrogram. So, there are two vectors of two dendrograms. Entanglement is a ratio the L_p norm between those two vectors of the tanglegrams to the worst case which is the case that the right dendrogram is completely reverse of the left dendrogram. The resulting stastic is a measure of how well the labels of two dendrograms are aligned. Entanglement values range from 0 (fully aligned labels) to 1 (fully mismatched labels). A value close to zero corresponds to a good alignment.

3.4 Metrics for Clusters

Performance evaluations in supervised learning tasks are easier and more understandable than unsupervised learning tasks such as clustering. However, there are also some useful metrics one can use to evaluate clustering tasks. In this subsection, we are going to review the metrics we are going to use to evaluate our classification and clustering algorithms. Particular comparison schemes we are going to review are the mutual information [69], Rand index [70], and homogeneity, completeness [71], and the silhouette scores [72].

For this subsection, assume X is the random variable that designates the true labels $\mathscr{X} = \{x_1, \cdots, x_n\}$ while Y is the random variable that designates the predicted labels $\mathscr{Y} = \{y_1, \cdots, y_m\}$ of a classification task. Let $\tau(d)$ be the true label and $\pi(d)$ be the predicted label of a data point d. We use $H(Z)$ to denote the Shannon entropy of a discrete random variable Z which is defined as

$$H(Z) = -\sum_{z \in Z} p(z) \log p(z). \tag{3.4}$$

3.4.1 Mutual information

The mutual information of the pair of the random variables X and Y is given as

$$MI(X,Y) = H(X) + H(Y) - H(X,Y) \tag{3.5}$$

Let $\mathrm{C}_x = \tau^{-1}(x)$ be the set of samples with true label $x \in \mathscr{X}$ of size n_x, and let $\mathrm{C}_y = \pi^{-1}(y)$ be the set of samples with predicted label $y \in \mathscr{Y}$ of size n_y. The joint distribution $p(X = x, Y = y)$ which is an observation drawn at random falls into clusters C_x and C_y turn out to be $\frac{|\mathrm{C}_x \cap \mathrm{C}_y|}{n}$ with the marginal probability $p(X = x) = \frac{n_x}{n}$ and $p(Y = y) = \frac{n_y}{n}$. So, the mutual information of the pair (X,Y) can be written in terms of these cardinalities as

$$MI(X,Y) = \sum_{x \in \mathscr{X}} \sum_{y \in \mathscr{Y}} \frac{|\mathrm{C}_x \cap \mathrm{C}_y|}{n} \log \left(\frac{n|\mathrm{C}_x \cap \mathrm{C}_y|}{n_x n_y} \right). \tag{3.6}$$

3.4.2 Homogeneity and completeness

If all clusters contain only data points that are members of a single class, then we say that our clusters are homogeneous. On the other hand, if the data points that are members of a given class are elements of the same cluster, then we say that our clusters are complete. Formally, the homogeneity and completeness scores are respectively defined by:

$$\hom(X,Y) = 1 - \frac{H(X \mid Y)}{H(X)} \quad \text{and} \quad \operatorname{comp}(X,Y) = 1 - \frac{H(Y \mid X)}{H(Y)}. \tag{3.7}$$

Notice that both homogeneity and completeness scores are not affected by permutations of labels.

3.4.3 Rand index

The Rand index counts sample of unordered distinct pairs of data points $\{u,v\}$ for which τ and π agree, and also disagree. Formally, the Rand index is defined as

$$\mathrm{RI} = \frac{a+b}{\binom{n}{2}}, \tag{3.8}$$

where

$$a = |\{\{u,v\} : \tau(u) = \tau(v) \text{ and } \pi(u) = \pi(v)\}| \tag{3.9}$$

and

$$b = |\{\{u,v\} : \tau(u) \neq \tau(v) \text{ and } \pi(u) \neq \pi(v)\}| \tag{3.10}$$

and $\binom{n}{2}$ is the total number of different possible unordered pairs in the dataset.

3.4.4 Silhouette score

We are going to use the silhouette score as defined in [72] to evaluate a clustering model on a dataset.

Assume we have a set of clusters $C_1, \ldots, C_k$, and a (dis)similarity measure $d(x,y)$ for every pair of points in our dataset. Let $U(x) = \pi^{-1}\pi(x)$ be the cluster that x belongs to and let

$$d(x,C_j) = \frac{1}{|C_j|} \sum_{y \in C_j} d(x,y). \tag{3.11}$$

The silhouette score $s(x)$ of a point x in our dataset is given by the ratio

$$s(x) = \frac{b(x) - a(x)}{\max(a(x), b(x))}, \tag{3.12}$$

where

$$a(x) = d(x, U(x)) \quad \text{and} \quad b(x) = \min_{C \neq U(x)} d(x, C). \tag{3.13}$$

The silhouette score for the whole dataset can be calculated as the mean of the silhouette score for each point in the dataset. Silhouette scores take values in the interval $[-1, 1]$ and values closer to 1 indicate that clusters are *well-formed* with low intra-class similarity while maintaining a high inter-class dissimilarity.

3.5 An Explicit Example

Let us now work out a simple example for hierarchical clustering using distances in kilometers between some Türkiye cities using the *single-linkage* method. The data was taken from the web page of Turkish General Directorate of Highways [73], and is shown in Table 3.2.

Table 3.2 : Distances between pairs of cities.

	Tekirdağ	İstanbul	Balıkesir	Manisa	İzmir	Konya	Antalya
Tekirdağ	0	132	379	511	506	794	850
İstanbul	132	0	390	529	564	662	718
Balıkesir	379	390	0	141	176	551	505
Manisa	511	529	141	0	35	534	428
İzmir	506	564	176	35	0	550	444
Konya	794	662	551	534	550	0	322
Antalya	850	718	505	428	444	322	0

The closest pair of cities is Manisa and İzmir with distance 35kms. So, they merge into a single cluster we called "Manisa-İzmir". The entry of Manisa-İzmir column (also the row) consists of the minimum distances of other cities to Manisa and İzmir since we chose single-linkage for the distance metric between two clusters. After one step, the algorithm yields the updated proximity matrix given in Table 3.3.

The second closest pair of cities is Tekirdağ and İstanbul with distance 132kms. So, they merge into a single cluster we called "Tekirdağ-İstanbul". Since we chose the

Table 3.3 : The six clusters after step one.

	Tekirdağ	İstanbul	Balıkesir	Manisa-İzmir	Konya	Antalya
Tekirdağ	0	132	379	506	794	850
İstanbul	132	0	390	529	662	718
Balıkesir	379	390	0	141	551	505
Manisa-İzmir	506	529	141	0	534	428
Konya	794	662	551	534	0	322
Antalya	850	718	505	428	322	0

single linkage, the distances between any two clusters are the minimum distance between objects of the these clusters. The resulting matrix is given in Table 3.4.

Table 3.4 : The five clusters after step two.

	Tekirdağ-İstanbul	Balıkesir	Manisa-İzmir	Konya	Antalya
Tekirdağ-İstanbul	0	379	506	662	718
Balıkesir	379	0	141	551	505
Manisa-İzmir	506	141	0	534	428
Konya	662	551	534	0	322
Antalya	718	505	428	322	0

Manisa-İzmir and Balıkesir, at a distance of 141 kilometers, are the third closest pair of clusters. Thus, they combine to form a single cluster that we named "Manisa-İzmir-Balkesir." The distances between any two clusters are the shortest distances between items in these clusters because we employed single-linkage. This process continues until last two clusters are merged into a single cluster. The resulting matrix is given in Table 3.5.

The merging of the clusters can be depicted as a dendrogram as in Figure 3.2. We obtained Figure 3.2 using the distance matrix above, and applying the function from the library **stats** [74] *dendrogram* under the R programming language. The distances when clusters are merged are shown on the $y-$axis in Figure 3.2. Using this dendrogram, we can determine the number of clusters by deciding on a *cut distance*, which in effect, determines the maximum diameter for the clusters at hand. For example, if we set

Table 3.5 : The four clusters after step three.

	Tekirdağ-İstanbul	Manisa-İzmir-Balıkesir	Konya	Antalya
Tekirdağ-İstanbul	0	379	662	718
Manisa-İzmir-Balıkesir	379	0	534	428
Konya	662	534	0	322
Antalya	718	428	322	0

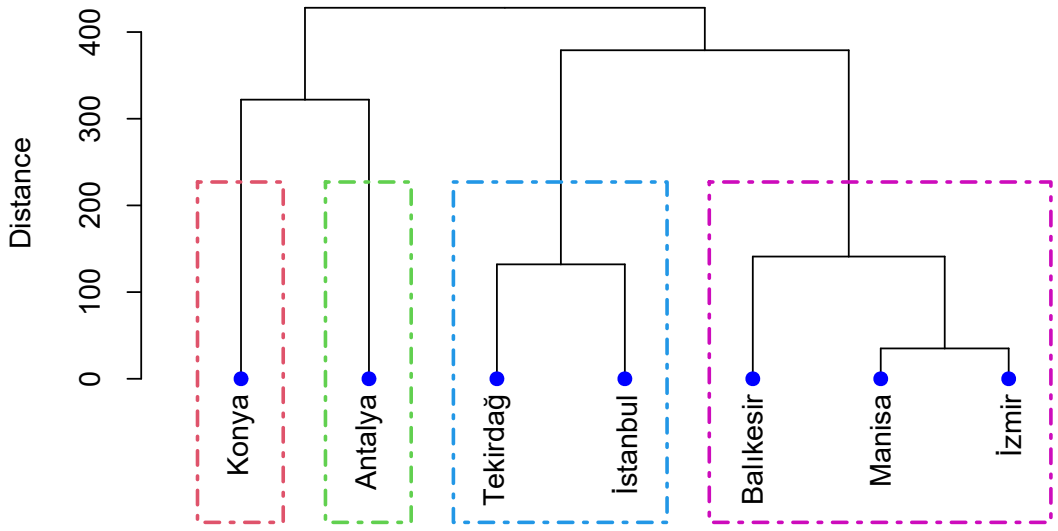


Figure 3.2 : A hierarchical clustering of Turkish cities.

the cut distance at 200kms, we get four clusters out of seven cities. The first two clusters consist of single cities (Konya and Antalya), the third cluster consist of 2 cities (Tekirdağ and İstanbul), and the last cluster consist of 3 cities (Balıkesir, Manisa and İzmir).

Let us consider the cophenetic matrix of the Turkish cities we present in Figure 3.2. We fill the matrix with the minimum merging distances that we obtain in the example above. The resulting matrix is given in Table 3.6.

The entries of the matrix comes from the following observations:

(a) We merge cluster Manisa and İzmir into cluster Manisa-İzmir at distance 35.

Table 3.6 : The cophenetic distance matrix of the dendrogram given in Figure 3.2.

	Tekirdağ	İstanbul	Balıkesir	Manisa	İzmir	Konya	Antalya
Tekirdağ	0	132	379	379	379	428	428
İstanbul	132	0	379	379	379	428	428
Balıkesir	379	379	0	141	141	428	428
Manisa	379	379	141	0	35	428	428
İzmir	379	379	141	35	0	428	428
Konya	428	428	428	428	428	0	322
Antalya	428	428	428	428	428	322	0

(b) We merge cluster Tekirdağ and İstanbul into cluster Tekirdağ-İstanbul at distance 132.

(c) We merge cluster Manisa-İzmir and Balıkesir into cluster Manisa-İzmir-Balıkesir at distance 141.

(d) We merge cluster Antalya and Konya into cluster Antalya-Konya at distance 322.

(e) We merge cluster Tekirdağ-İstanbul and Manisa-İzmir-Balıkesir into cluster Tekirdağ-İstanbul-Manisa-İzmir-Balıkesir at distance 379.

(f) We merge cluster Antalya-Konya and Tekirdağ-İstanbul-Manisa-İzmir-Balıkesir into one cluster All-Cities at distance 428.

For the dendrogram in Figure 3.2, one can also calculate the silhouette score for the clusters C_1 ={Tekirdağ, İstanbul}, C_2 ={Manisa, İzmir, Balıkesir}, C_3 = {Konya} and C_4 = {Antalya}.

There are mainly two distances to calculate the silhouette score for a point x in the dataset such as cohesion and separation. The cohesion is the average distance between each point within the cluster denoted as $a(x)$, while the separation is the minimum distance between all clusters denoted as $b(x)$.

For İstanbul, Tekirdağ is a point in the same cluster with distance 132. So, the cohesion for İstanbul is $a(\text{İstanbul}) = 132$. On the other hand, the distance between İstanbul and the second cluster C_2 is calculated as taking the average value of the distances between İstanbul and each point in cluster C_2. That distance is the average of $529, 564, 390$ which is 494.5. The distance between İstanbul and the other two clusters C_3 and C_4 are

662 and 718, respectively. Therefore, the separation $b(\text{İstanbul})$ is 494.5 which is the minimum value of $494.5, 662$, and 718. Hence, the silhouette score for a point İstanbul is approximately 0.73.

Table 3.7 : The cohesion and separation distance with the silhouette scores.

Points	Cohesion	Separation	Silhouette
Tekirdağ	132	465.3	0.72
İstanbul	132	494.5	0.73
Balıkesir	158.5	384.5	0.59
Manisa	88	428	0.79
İzmir	105.5	444	0.76
Konya	0	322	1
Antalya	0	322	1

Table 3.7 displays the silhouette scores for each point in the dataset with the cohesion and separation. Hence, the overall silhouette score for the dataset is the average of the individually silhouette scores, that is $s \approx 0.80$ when the number of clusters is 4.

4. SIMPLICIAL COMPLEXES

Persistent homology requires us to calculate the homology of filtered simplicial complexes. The main calculation scheme involves first creating a filtered simplicial complex out of a point cloud, then creating a differential graded complex, and then finally evaluating homology of the resulting differential graded complex. In this chapter we are going to cover all the required background material on simplicial complexes, differential graded complexes, and their homologies.

4.1 Simplicial Complexes

We will use simplicial complexes to construct global structures from local ones such as finite collections of data points in an affine space. The basic idea is that we choose a proximity parameter and then connect points according to this parameter to form line segments, triangles, tetrahedra etc.

Definition 4.1. Suppose that we have finite set of points $S = \{v_0, v_1, \ldots, v_k\}$ in $\mathbb{R}^n$. The set S is said to be geometrically independent if the vectors

$$v_1 - v_0, v_2 - v_0, \ldots, v_k - v_0 \tag{4.1}$$

are linearly independent.

Definition 4.2. Let $S = \{v_0, v_1, \ldots, v_k\}$ be a geometrically independent set in $\mathbb{R}^n$. A $k-$simplex σ spanned by given points S is the set of points $x \in \mathbb{R}^n$ with $|\sigma| = k+1$ such that

$$x = \sum_{i=0}^{k} c_i v_i \quad \text{and} \quad \sum_{i=0}^{k} c_i = 1 \tag{4.2}$$

for $c_i \geqslant 0$ and $i \in \{0, 1, \ldots, k\}$. In other words, a k-simplex spanned by $\sigma \subseteq S$ is the convex combination of points in σ.

The points $v_0, v_1, \ldots, v_k$ are called the vertices of σ and any simplex generated by a subset of $v_0, v_1, \ldots, v_k$ is called a face of σ.

Definition 4.3. A simplicial complex $\mathscr{K}$ in $\mathbb{R}^n$ is a collection of simplices in $\mathbb{R}^n$ such that

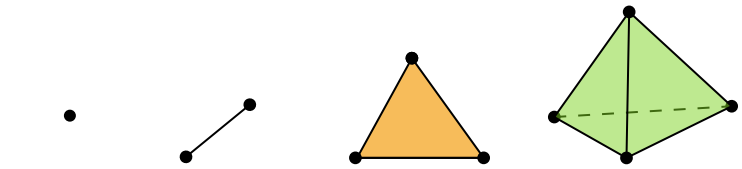


Figure 4.1 : An example k-simplex.

(i) Every face of a simplex of $\mathscr{K}$ is in $\mathscr{K}$.

(ii) The intersection of any two simplices of $\mathscr{K}$ is a face of each

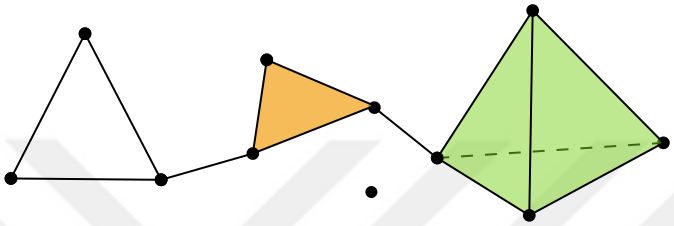


Figure 4.2 : A simplicial complex of dimension 3.

Definition 4.4. Let X be a set. An abstract simplicial complex $\mathscr{K}$ is a collection of finite non-empty subsets of X such that if $\alpha \in \mathscr{K}$ and $\beta \subseteq \alpha$ then $\beta \in \mathscr{K}$. Each set $\alpha \in \mathscr{K}$ is called its simplex, and each element of α is called a vertex of α.

4.2 Coverings of Topological Spaces

Definition 4.5. The collection of $\mathscr{U} = \{U_i \ : \ i \in I\}$ of open subsets $U_i \subseteq X, \ i \in I$, where I is a index set is called an open cover of X if

$$X = \bigcup_{i \in I} U_i \tag{4.3}$$

4.3 The Nerve of a Topological Space

Let X be a topological space with an open cover $\mathscr{U} = \{U_i \ : \ i \in I\}$. The nerve of open cover $\mathscr{U}$ is an abstract simplicial complex $N(\mathscr{U})$ whose vertex set is $\mathscr{U}$ and such that

$$\sigma = [U_{i_0}, U_{i_1}, \ldots, U_{i_k}] \in N(\mathscr{U}) \quad \text{if and only if} \quad \bigcap_{\forall j} U_{i_j} \neq \emptyset \tag{4.4}$$

Definition 4.6. Let X be a set and let $\mathscr{U}$ be a finite collection of subsets of X. The nerve $\mathscr{N}(\mathscr{U})$ of $\mathscr{U}$ is a collection of subsets of $\mathscr{U}$ that have a non-empty common

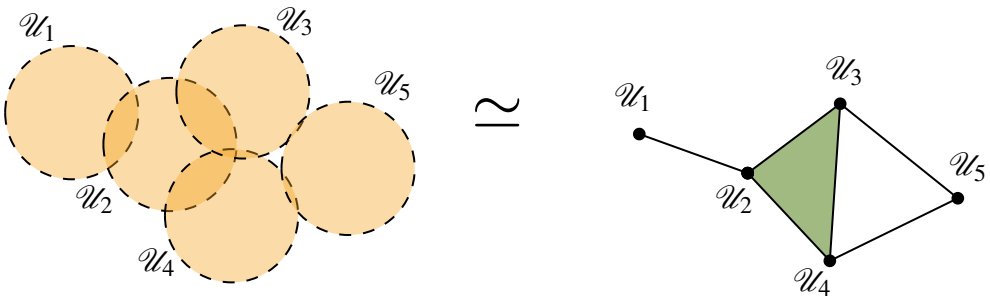


Figure 4.3 : An example of a Čech nerve.

intersection. That is,

$$\mathcal{N}(\mathcal{U}) = \left\{ B \subseteq \mathcal{U} \ : \bigcap_{V \in B} V \neq \emptyset \right\} \tag{4.5}$$

Note that, the nerve is an abstract simplicial complex since

$$B \in \mathcal{N}(\mathcal{U}) \text{ and } V \subseteq B \text{ implies } V \in \mathcal{N}(\mathcal{U}) \tag{4.6}$$

The nerve of a topological space X, rather a suitable covering $\mathcal{U}$, is a useful construction since one can recover the homotopy/homology type of X from $N(\mathcal{U})$.

Proposition 4.7 ([30]). Let X be a topological space with an open cover $\mathcal{U} = \{U_i : i \in I\}$. Assume that the intersection of elements of any subset of $\mathcal{U}$ is empty or contractible. Then, the space X and its nerve $N(\mathcal{U})$ are homotopy equivalent.

In Figure 4.3, every connected component can be contracted to a point. If two components have an intersection, we put a edge between them. If three components have a common intersection, we fill in the corresponding triangle. So, when constructing the nerve, we draw a vertex for each of the facets in the original space. Then, we fill in edges, triangles, tetrahedra etc.

4.4 A Zoo of Complexes

Now, let (X,d) be a metric space and let $D \subseteq X$ be a finite set of points (a point cloud) in X. Let us use $B_\varepsilon(x)$ for the open ball of radius ε centred at $x \in D$.

4.4.1 Clique complex

Definition 4.8. Let K be a graph. The *clique complex* of K is a simplicial complex $\mathscr{L}$ such that if any set of vertices $\{x_0,\ldots,x_k\}$ forms a *clique*, i.e. when all possible edges between these vertices are in K, then the simplex $[x_0,\ldots,x_k]$ is in $\mathscr{L}$.

4.4.2 Čech complex

Definition 4.9. Given a point cloud $D = \{x_i \in \mathbb{R}^d \,:\, i \in I,\ \text{a finite index set } I\}$, the Čech complex $\mathscr{C}(D,\varepsilon) = \mathscr{C}_\varepsilon(D)$ is the nerve of the collection of open balls $B_\varepsilon(x_i)$, formally

$$\begin{aligned}\mathscr{C}_\varepsilon(D) =& \mathscr{N}\Big(\{B_\varepsilon(x_i) \,:\, x_i \in D \text{ and given fix } \varepsilon\}\Big) \\ =& \left\{\sigma \subseteq X \mid \bigcap_{x\in\sigma} B_\varepsilon(x) \neq \varnothing\right\}.\end{aligned} \tag{4.7}$$

In other words, a collection of points $\sigma = [x_0,\ldots,x_k]$ forms an k-simplex if the set of balls of radius ε centered at these points has non-empty intersection.

Observe that if the metric space is Euclidean, the balls considered for Čech complex are necessarily convex and hence their intersections are contractible. So, from the Proposition 4.7,

$$X = \bigcup_{x\in D} B_\varepsilon(x) \simeq \mathscr{C}(D,\varepsilon). \tag{4.8}$$

In other words, the Čech complex is homotopy equivalent to the space of union of the balls.

4.4.3 Vietoris-Rips complex

The Vietoris-Rips complex $\mathscr{R}_\varepsilon(D)$ of a given point cloud D is defined to be the simplicial complex whose vertices are all points in D that are at most ε apart. In other words

$$\mathscr{R}_\varepsilon(D) = \{\sigma \subset D \mid d(x,y) \leqslant \varepsilon,\ \text{for all } x,y \in \sigma\}. \tag{4.9}$$

The clique complex of a graph K is an example of Vietoris-Rips complex if we consider a graph as a metric space via the geodesic distance and set $\varepsilon = 1$.

4.4.4 Delanuay complex

The *Voronoi region* $R(x)$ of a point $x \in D$ is defined as the points in X that are closest to x. Formally,

$$R(x) = \{y \in X \mid x \in \operatorname{argmin}_{z\in D} d(z,y)\}. \tag{4.10}$$

The *Delaunay complex* is the nerve of the covering $\{R(x) \mid x \in D\}$ of D given by Voronoi regions.

4.4.5 Alpha complex

Let $\varepsilon > 0$. The *restricted Voronoi region* of a point x is the intersection of the Voronoi region $R(x)$ and the open ball $B_\varepsilon(x)$. The *alpha complex* $\mathscr{A}_\varepsilon$ is the nerve of the covering given by the restricted Voronoi regions

$$\mathscr{A}_\varepsilon = \{B_\varepsilon(x) \cap R(x) \mid x \in D\}. \tag{4.11}$$

The alpha complex grows with ε. For instance, $\mathscr{A}_0 = \emptyset$ and if ε is big enough $\mathscr{A}_\varepsilon$ coincides with the Delaunay complex. Moreover, unlike the Vietoris-Rips and the Čech complexes, the dimension of the alpha complex is restricted to the dimension of the space the points are embedded in given that the points are in general position. For example, the dimension of the alpha complex of a set of points in $\mathbb{R}^2$ cannot exceed 2 whenever none three points are collinear.

For more kind of complexes, we refer to the readers to see the second chapter of book [5].

4.5 Vietoris-Rips vs Čech Complexes

Notice that for a given proximity $\varepsilon > 0$, the Vietoris-Rips complex of a point cloud D has more simplices compared to the Čech complex of the same cloud D as in Figure 4.4. On the other hand, computing the Čech complex requires computing all possible intersections of the balls. The Vietoris-Rips complex has reduced time complexity since we only check pairwise intersections which makes the Vietoris-Rips complex relatively less expensive to compute than the corresponding Čech complex. To be more precise, for a point cloud D in a metric space X with $|D| = m$ and for a fixed radius ε, one needs to check 2^m intersections to determine the Čech complex while to construct the Vietoris-Rips complex one would only need to check $\binom{m}{2}$ intersections.

Vietoris-Rips compelexes are generally cheaper to construct, but there is an important issue: they are not necessarily homotopy equivalent to the space they are sampled from. To deal with this issue, we have the following nice from Tamal-Yusu's book [75]:

Proposition 4.10. Given a point cloud D which is subset of the metric space (X,d) and a radius $\varepsilon > 0$, we have a inclusion relation between Vietoris-Rips and Čech complex

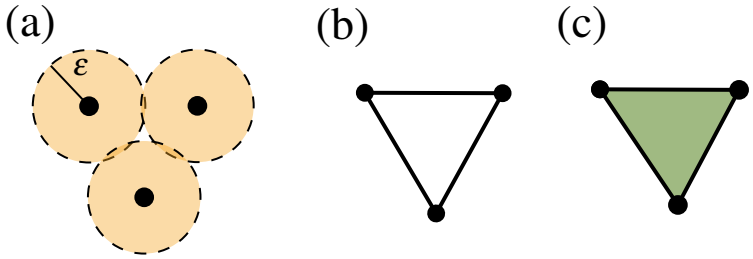


Figure 4.4 : A toy example of Čech complex (b) and Vietoris-Rips complex (c) for the point cloud (a) consist of just three points in $\mathbb{R}^2$.

as follows,

$$\mathscr{C}_\varepsilon(D) \subseteq \mathscr{R}_\varepsilon(D) \subseteq \mathscr{C}_{2\varepsilon}(D). \tag{4.12}$$

Proof. To prove the first inclusion, we take simplex $\sigma = \{v_0, v_1, \ldots, v_k\} \in \mathscr{C}_\varepsilon(D)$ and $\bigcap B_\varepsilon(v_i) \neq \emptyset$. So, all pairs $B_\varepsilon(v_i) \cap B_\varepsilon(v_j) = \emptyset$ for the distance $d(v_i, v_j) \leqslant 2\varepsilon$. That is $\sigma \in \mathscr{R}_\varepsilon$.

To prove the second inclusion, consider a simplex $\sigma = \{v_0, v_1, \ldots, v_k\} \in \mathscr{R}_\varepsilon$. That is $d(v_i, v_0) \leqslant 2\varepsilon$ for every $v_i,\ i = 0, 1, \ldots, k$. So, we have $\bigcap_{i=0}^{k} B_{2\varepsilon}(v_i) \supset v_0 \neq \emptyset$. Then, by definiton σ is also a simplex in $\mathscr{C}_{2\varepsilon}(D)$. □

Ghrist and de Silva [11] introduced a stronger version of the Proposition 4.10 which is dimension-dependent bound in $\mathbb{R}^n$ as follows:

Proposition 4.11. With the same assumptions of Proposition 4.10 we have

$$\mathscr{C}_\varepsilon(D) \subseteq \mathscr{R}_{2\varepsilon}(D) \subseteq \mathscr{C}_{2\varepsilon'}(D), \tag{4.13}$$

where $\varepsilon' = \varepsilon\sqrt{2n/(n+1)}$.

In this thesis, we primarily use Vietoris-Rips complexes since they are relatively simple and cheaper to construct compared to other available filtered complexes one can construct out of finite data clouds embedded in metric spaces, which in practice is $\mathbb{R}^n$ with the Euclidean metric.

5. CHAIN COMPLEXES AND THEIR HOMOLOGY

Homology is a functorial topological/homotopical invariant. In the context of our thesis, this means that if the two topological spaces are homeomorphic/homotopic, then they have the same homology. In this section, we explain how the homology of a simplicial complex is calculated. For more information on the homology of more general topological spaces, we refer the reader to Hatcher's book [76].

5.1 Chain Complexes

Definition 5.1. A chain complex C is a sequence of Abelian groups or vector spaces C_k connected by homomorphisms (boundary operators) $\partial_k : C_k \to C_{k-1}$ such that $\partial_{k-1} \circ \partial_k = 0$ for $k \in \mathbb{Z}$.

$$\ldots \xrightarrow{\partial_{k+2}} C_{k+1}(\mathscr{K}) \xrightarrow{\partial_{k+1}} C_k(\mathscr{K}) \xrightarrow{\partial_k} C_{k-1}(\mathscr{K}) \xrightarrow{\partial_{k-1}} \ldots \tag{5.1}$$

5.2 Homology of Chain Complexes

Definition 5.2. Assume (C_*, ∂_*) is a chain complex. We then define

$$Z_k = ker(\partial_k), \quad B_k = im(\partial_{k+1}), \quad H_k(\mathscr{K}) = Z_k/B_k \tag{5.2}$$

for every $k \geqslant 0$. The vector space Z_k is called *the space of cycles*, B_k is called *the space of boundaries*, and $H_k(\mathscr{K})$ is called *the homology* of the simplicial complex $\mathscr{K}$.

The elements of H_k are cosets of the form $[c] = c + \mathrm{B}_k = \{c + b \mid b \in \mathrm{B}_k\}$ where c is a k-cycle, which is also referred to as a generating cycle of the homology class $[c] = c + \mathrm{B}_k$. Two k-cycles c and d are homologous if $[c] = [d]$, that is, $c - d \in \mathrm{B}_k$ is the boundary of some $(k+1)$-chain. A basis of H_k is a minimal set of homology classes that generates H_k. A set of k-cycles $C = \{c_1, \ldots, c_l\}$ generates H_k, if $\{[c_i]\}$ forms a basis for H_k.

5.3 Chain Complexes From Simplicial Complexes

Definition 5.3. Let $\mathscr{K}$ be a simplicial complex and fix a dimension k. Let us denote the vector space spanned by the set of k-chains by $C_k(\mathscr{K})$. In other words, a k-chain is a formal sum of k-simplices in $\mathscr{K}$ written by

$$\alpha = \sum a_i \sigma_i \tag{5.3}$$

k-chains are added component-wise: if $\alpha = \sum a_i\sigma_i$ and $\beta = \sum b_i\sigma_i$, then $\alpha+\beta = \sum(a_i+b_i)\sigma_i$. The graded space $C_*(\mathscr{K})$ is called the associated chain complex of $\mathscr{K}$.

Let $\{v_0, v_1, \ldots, v_k\}$ span a k-simplex σ. An orientation of a simplex is given by a total ordering of the vertices $\{v_0, v_1, \ldots, v_k\}$ denoted by $[v_0, v_1, \ldots, v_k]$. Two orderings define the same orientation if and only if they differ by an even permutation.

The k-simplices of a simplicial complex span the vector space C_k, and boundary maps $\partial_k \colon C_k \to C_{k-1}$ are defined as follows: Let $k > 0$ and $\sigma = [v_0, v_1, \ldots, v_k]$ be an oriented simplex then

$$\partial_k \sigma = \sum_{i=0}^{k} (-1)^i [v_0, v_1, \ldots, \hat{v}_i, \ldots, v_k] \tag{5.4}$$

where $\hat{v}_i$ means that the vertex v_i as been dropped. We define $\partial_0 = 0$.

Theorem 5.4. Let $\mathscr{K}$ be a simplicial complex, and let $C_*(\mathscr{K})$ with differentials ∂_* be its chain complex. Then the composition $\partial_{k-1}\partial_k$ is zero for every $k \geqslant 1$.

Proof. Taken a $k-$simplex $\sigma \in C_k(\mathscr{K})$, then we have

$$\partial_k \sigma = \sum_i (-1)^i \left[v_0, \ldots \hat{v}_i \ldots v_k\right]. \tag{5.5}$$

Then,

$$\begin{aligned}
\partial_{k-1}\partial_k \sigma &= \sum_i (-1)^i \partial_{k-1} \left[v_0, \ldots \hat{v}_i, \ldots, v_k\right] \\
&= \sum_{i<j} (-1)^i (-1)^{j-1} \left[v_0, \ldots \hat{v}_i \ldots \hat{v}_j \ldots v_k\right] + \sum_{i>j} (-1)^i (-1)^j \left[v_0, \ldots \hat{v}_j \ldots \hat{v}_i \ldots v_k\right] \\
&= -\sum_{i<j} (-1)^i (-1)^j \left[v_0, \ldots \hat{v}_i \ldots \hat{v}_j \ldots v_k\right] + \sum_{i>j} (-1)^i (-1)^j \left[v_0, \ldots \hat{v}_j \ldots \hat{v}_i \ldots v_k\right] \\
&= 0
\end{aligned}$$

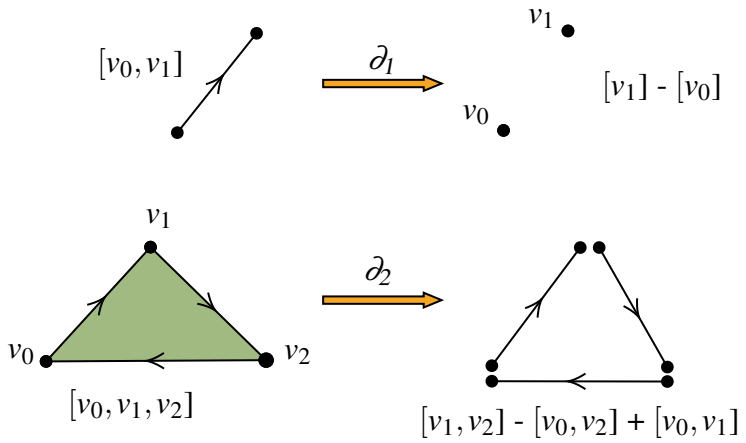


Figure 5.1 : An example of applying the boundary operator to 1- and 2-simplex.

as we wanted to show. □

A toy example of the Theorem 5.4, the simplicial identities we have $\partial_{k-1}\partial_k = 0$ in Figure 5.1

$$\begin{aligned} \partial_1[v_0, v_1] &= v_1 - v_0 \\ \partial_2[v_0, v_1, v_2] &= [v_1, v_2] - [v_0, v_2] + [v_0, v_1] \\ \partial_1\partial_2[v_0, v_1, v_2] &= (v_2 - v_1) - (v_2 - v_0) + (v_1 - v_0) = 0 \end{aligned} \tag{5.6}$$

5.4 Betti Numbers

Definition 5.5. Let $\mathscr{K}$ be a simplicial complex and $k \geqslant 0$. The p-th *Betti number* $\beta_k(\mathscr{K})$ is defined as

$$\beta_k(\mathscr{K}) = \dim H_k(\mathscr{K}) \tag{5.7}$$

the dimension of the p-th homology group of $\mathscr{K}$. In particular, β_0 counts the number of connected components β_1 counts the number of loops, and so on.

For the computer computation of the homology, there is a more efficient algorithm which utilizes the Smith Normal Form for the boundary matrix. We use it in our calculations to find homology generators. Here, we refer to the readers to [3] for more about Smith Normal Form.

5.5 A Complete Example

In the following, we provide an explicit example on how to calculate the homology of a simplicial complex using a more conventional method. Let us consider the following geometric simplicial complex given in $\mathbb{R}^2$:

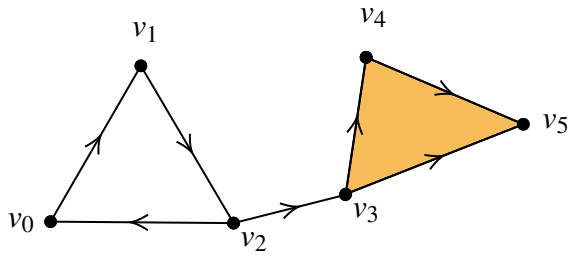


Figure 5.2 : An oriented simplicial complex K.

Our simplicial complex $\mathscr{K}$ consist the following simplicies:

$$\begin{aligned}\mathscr{K}_0 =&\{v_0, v_1, v_2, v_3, v_4, v_5\},\\ \mathscr{K}_1 =&\{[v_0, v_1], [v_1, v_2], [v_2, v_0], [v_2, v_3], [v_3, v_4], [v_4, v_5], [v_3, v_5]\},\\ \mathscr{K}_2 =&\{[v_3, v_4, v_5]\}.\end{aligned} \tag{5.8}$$

Since there are no simplices in dimensions $k < 0$ and $k > 2$, the homology $H_k(\mathscr{K})$ for those dimensions is zero. Also, since $\mathscr{K}$ is connected, $H_0(\mathscr{K})$ has to be $\mathbb{Z}$.

To find the homology generators (in other words, the basis of the vector space) and then to calculate Betti numbers β_0 and β_1, firstly we need to determine boundary maps ∂_0, ∂_1 and ∂_2 such that the chain complex is

$$0 \xrightarrow{\partial_3} C_2 \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0 \ . \tag{5.9}$$

Since $\partial_0 = 0$, $\ker(\partial_0) = C_0$. We can represent the boundary map $\partial_1 : C_1 \to C_0$ as the following matrix notion,

$$\partial_1 = \begin{array}{c} \\ v_0 \\ v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \end{array} \begin{array}{c} \begin{array}{ccccccc} [v_0,v_1] & [v_1,v_2] & [v_2,v_0] & [v_2,v_3] & [v_3,v_4] & [v_4,v_5] & [v_3,v_5] \end{array} \\ \begin{bmatrix} -1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \end{array} \tag{5.10}$$

where the columns and rows are represented for the group C_1 and C_0, respectively.

Then, one can get the row to reduce form of the boundary matrix as follows:

$$\text{Row Reduce}(\partial_1) = \begin{bmatrix} 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{5.11}$$

On the other hand, $\partial_2 : C_2 \to C_1$ then

$$\partial_2([v_3, v_4, v_5]) = [v_4, v_5] - [v_3, v_5] + [v_3, v_4], \tag{5.12}$$

in the vector form

$$\partial_2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 & -1 \end{bmatrix}^T. \tag{5.13}$$

The range of the boundary matrix ∂_1 is

$$\text{im}(\partial_1) = \text{span}\{w_0, w_1, w_3, w_4, w_5\}, \tag{5.14}$$

where w_i for $i = 0, 1, 3, 4, 5$ is the column vector corresponding to pivot elements on the row reduce form, that is,

$$\begin{aligned} w_0 &= \begin{bmatrix} -1 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T, \\ w_1 &= \begin{bmatrix} 0 & -1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}^T, \\ w_3 &= \begin{bmatrix} 0 & 0 & -1 & 1 & 0 & 0 & 0 \end{bmatrix}^T, \\ w_4 &= \begin{bmatrix} 0 & 0 & 0 & -1 & 1 & 0 & 0 \end{bmatrix}^T, \\ w_5 &= \begin{bmatrix} 0 & 0 & 0 & 0 & -1 & 1 & 0 \end{bmatrix}^T. \end{aligned} \tag{5.15}$$

Now, we will evaluate the $\ker(\partial_1)$, in the other saying we solve the unknown x in the system $\partial_1 x = 0$ via linear algebra, this yields

$$\begin{array}{rl} x_1 - x_3 &= 0 \\ x_2 - x_3 &= 0 \\ x_4 &= 0 \\ x_5 + x_7 &= 0 \\ x_6 + x_7 &= 0 \end{array} \quad \Longrightarrow \quad x = sy_1 + ty_2, \tag{5.16}$$

where

$$y_1 = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}^T \quad \text{and} \quad y_2 = \begin{bmatrix} 0 & 0 & 0 & 0 & -1 & -1 & 1 \end{bmatrix}^T. \tag{5.17}$$

Thus $\ker(\partial_1) = \text{span}\{y_1, y_2\}$. We also have $\text{im}(\partial_2) = \text{span}\{z_1\}$, where

$$z_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 & -1 \end{bmatrix}^T. \tag{5.18}$$

The homology generators for the first dimension are $H_1(\mathscr{K}) = \ker(\partial_1)/\text{im}(\partial_2) = \text{span}\{y_1\}$ since $y_2 + z_1 = 0$.

Thus,

$$H_k(\mathscr{K}) \cong \begin{cases} \mathbb{Z} & \text{for } k = 0, 1, \\ 0 & \text{for } k \neq 0, 1. \end{cases} \tag{5.19}$$

Finally, we can calculate the Betti numbers

$$\begin{aligned} \beta_0 &= \dim(\ker(\partial_0)) - \dim(\text{im}(\partial_1)) = 6 - 5 = 1, \\ \beta_1 &= \dim(\ker(\partial_1)) - \dim(\text{im}(\partial_2)) = 2 - 1 = 1. \end{aligned} \tag{5.20}$$

In fact, as we see in Figure 5.2 that there is a connected component and a loop. This means that the corresponding Betti numbers are 1 in both cases.

5.6 Another Example Coming from a Point Cloud

In this example, we present a visualization of homology generators of a simplicial complex for a point cloud. To obtain topological space from a point cloud, we use the Vietoris-Rips complex with two different scales.

In Figure 5.3, there is a noisy point cloud D sampled from a bouquet of two circles embedded in $\mathbb{R}^2$. We also display the filtered Vietoris-Rips complex with the first-degree homology generators colored with orange and red at two different scales ε and η, $\varepsilon \leqslant \eta$. While the red generator does not appear at the scale η, the orange generator persists from the scale ε to η.

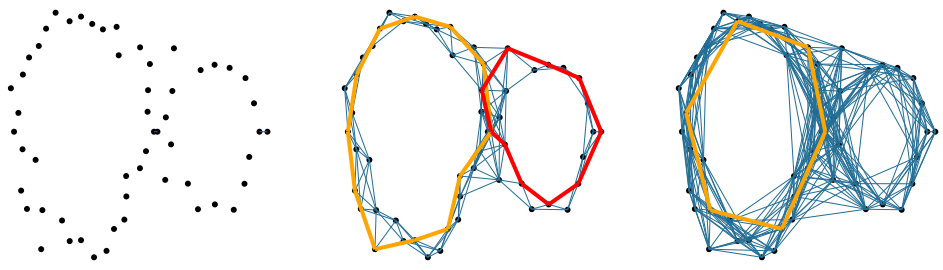


Figure 5.3 : The Vietoris-Rips complex with the first degree homology generator.

6. PERSISTENT HOMOLOGY AND BARCODES

In order to turn a point cloud into a simplicial complex by using the Vietoris-Rips or the Čech variations, we need to choose a proximity parameter ε. We then capture certain topological features of the data by changing the parameter ε. As we see in Figure 6.1, it is possible that ε may be too small or too large to capture the true topological features of the data set. But how does one choose the optimal value for ε? This is quite a challenging problem.

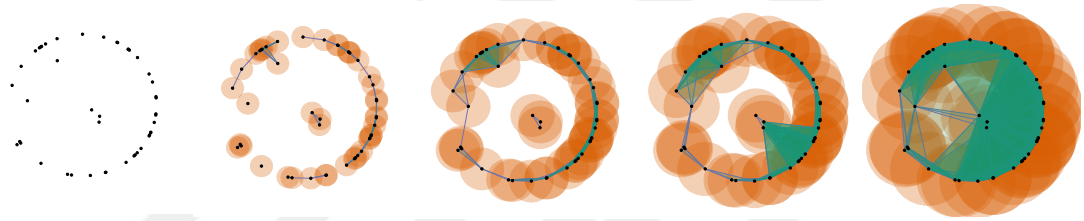


Figure 6.1 : Vietoris-Rips complexes with increasing values of the parameters $\varepsilon_1 \leqslant \varepsilon_2 \leqslant \varepsilon_3 \leqslant \varepsilon_4 \leqslant \varepsilon_5$ from left to right. Only in the case ε_4, the complex has the same topology with data sampled from S^1.

Edelsbrunner-Letscher-Zomorodian [8], and Carlsson-Zomorodian [7] proposed that persistence homology can help to solve the problem of choosing an optimal value for ε. In persistent homology, one keeps a record of how long each topological feature of a given data persists as the proximity parameter ε changes. This is computed by following the durability of topological features with respect to a filtration we associate with the observed point cloud depending on ε.

6.1 Filtered Simplicial Complexes

Definition 6.1. Given a space X, an abstract simplicial complex $\mathscr{K}$ in X is an order ideal in $(2^X, \subseteq)$. If $\mathscr{K}$ consists of finite sets we write

$$\mathscr{K}_m = \{x \in \mathscr{K} \mid |x| = m\} \tag{6.1}$$

for every $m \in \mathbb{N}$.

Definition 6.2. Let $(P,\leqslant)$ be an indexing poset. We call a collection $(\mathscr{K}_\varepsilon)_{\varepsilon\in P}$ of simplicial sets indexed by P as a *filtered simplicial complex over* P if for every comparable pair of element $\varepsilon \leqslant \eta$ in P we have a morphism of simplicial sets of the form $\imath_{\varepsilon,\eta}\colon \mathscr{K}_\varepsilon \to \mathscr{K}_\eta$ such that

$$\imath_{\eta,\nu} \circ \imath_{\varepsilon,\eta} = \imath_{\varepsilon,\nu} \tag{6.2}$$

for every $\varepsilon \leqslant \eta \leqslant \nu$ in P. One can also define a filtered complex as a functor $\mathscr{K}$ from the poset P to the category of simplicial complexes.

6.2 Persistence Modules

Definition 6.3. A persistence module over the poset P is a collection of vector spaces $\{V_\varepsilon\}$ and linear maps $\left\{f^{(\varepsilon,\eta)} : V_\varepsilon \to V_\eta\right\}$ for $\varepsilon,\eta \in P$ and $\varepsilon \leqslant \eta$ such that

- $f^{(\varepsilon,\varepsilon)} = \mathrm{id}\colon V_\varepsilon \to V_\varepsilon$,
- $f^{(\eta,\upsilon)} \circ f^{(\varepsilon,\eta)} = f^{(\varepsilon,\upsilon)}$.

Note that a persistence module is a functor from the partially ordered set P seen as a category in the obvious way, to the category of vector spaces.

6.3 Persistent Homology

All of the simplicial complexes we consider in this thesis, such as the Vietoris-Rips complex or Čech complex, are filtered over a finite subposet of $\mathbb{R}_+$ with its natural order. Then such a filtered simplicial complexes $\mathbb{K}:\ \mathscr{K}_1 \subseteq \mathscr{K}_2 \subseteq \cdots \subseteq \mathscr{K}_m$ yields a natural persistence module for every integer p,

$$H_k(\mathbb{K}) : H_k(\mathscr{K}_1) \to H_k(\mathscr{K}_2) \to \cdots \to H_k(\mathscr{K}_m)\,. \tag{6.3}$$

Now, we give the formal definition of the persistent homology groups as below:

Definition 6.4. For a filtered complex $(\mathscr{K}_\varepsilon)_{\varepsilon\in P}$, the k-th persistent homology of the filtered complex is defined as

$$\mathrm{PH}_k(\mathscr{K}) := \{H_k(\mathscr{K}_\varepsilon)\}_{\varepsilon\in P} \tag{6.4}$$

together with the collection of k-linear maps of the form $\psi^k_{\varepsilon,\eta}\colon H_k(\mathscr{K}_\varepsilon) \to H_k(\mathscr{K}_\eta)$ induced by the structure maps of the filtration $\iota_{\varepsilon,\eta}\colon \mathscr{K}_\varepsilon \to \mathscr{K}_\eta$ for all $k \in \mathbb{N}$ and $\varepsilon \leqslant \eta$ in $\mathbb{R}_+$.

Moreover, the k-th persistent Betti number is defined as $\beta^k_{\varepsilon_1,\varepsilon_2} = \operatorname{rank}(\psi^k_{\varepsilon_1,\varepsilon_2})$. That is the number of the homology class from $H_k(\mathscr{K}_{\varepsilon_1})$ survive to $H_k(\mathscr{K}_{\varepsilon_2})$.

6.4 Persistent Homology Pipeline

Throughout this thesis, our indexing poset I is going to be a finite subset of $\mathbb{R}_+$. Thus our filtrations of a simplicial complexes $\mathscr{K}$ are going to be collection of subcomplexes $\mathbb{K} = \{\mathscr{K}_\varepsilon : \varepsilon \in \mathbb{R}_+\}$ that satisfy $\mathscr{K}_{\varepsilon_1} \subseteq \mathscr{K}_{\varepsilon_2}$ whenever $\varepsilon_1 \leqslant \varepsilon_2$. We call a filtration $\mathbb{K}$ *bounded* if there is a real number N such that $\mathscr{K}_i = \mathscr{K}$ for every $\varepsilon > N$.

The persistent homology pipeline moves through the following three phases, as shown in the pipeline Figure 6.2:

Figure 6.2 : Persistent homology pipeline.

We first build a filtration from a point cloud in a way that highlights some of the data's intriguing structural features. Then to create a persistence module, we apply the homology functor to every topological space and inclusion map in the filtration to get vector spaces and linear maps.

Under certain assumptions about the persistence module, the structure theorem for the persistence modules produces a multiset of birth-death pairs, called a *barcode*, of the homology classes in the filtration. The barcode is an isomorphism invariant of the data set, and the structure theorem informs us that the persistence module decomposes into simple summands known as interval modules in an essentially unique manner. For more details about structure theorem in representation theory via quivers, please see [77].

6.5 Barcodes

Persistence modules are divided into interval modules, therefore we can effectively reveal the timing of the appearance and disappearance of features measured by homology over the module [77]. Since our persistence modules are filtered over finite subposets of $\mathbb{R}_+$, for each cycle $\gamma \in Z_k$ there is an interval that records the *life-time* of γ, i.e. the interval on which γ is non-trivial as ε ranges from 0 to ∞. We say γ is *born* at $\varepsilon = b$ when the homology class $[\gamma] \in H_k(\mathscr{K}_b)$ is not in the image of $\psi^k_{\varepsilon,b}$ for every $\varepsilon < b$. Similarly, we say γ *dies* at $\varepsilon = d$ if $\psi^k_{b,\varepsilon}([\gamma]) = 0$ for every $\varepsilon > d$. Thus we obtain a collection of pairs (b,d) that each indicates the parameter values for which a homological class occurred at time b and vanished at time d can be used to represent this data uniquely. If the data set is finite, a structure theorem [77] allows us to extract the birth-death times of the homology classes from the persistence modules. This multiset of birth-death pairs is called the *barcode*.

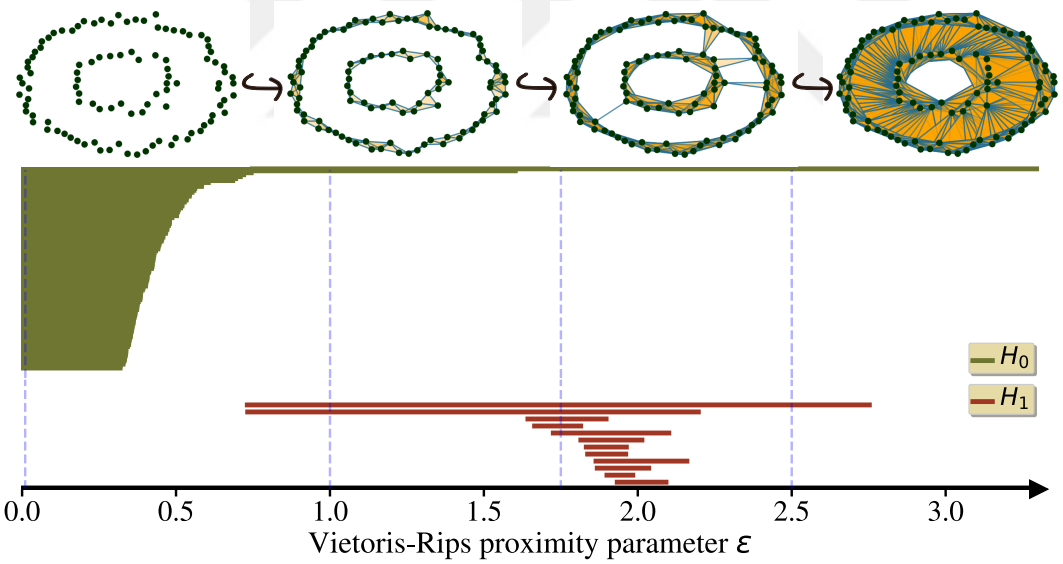


Figure 6.3 : An example barcode with the filtered Vietoris-Rips complex in the particular scale showed as blue dash vertical line for the data sampled from two intertwined-circle with a little bit noisy.

To illustrate the life-times of cycles, we use *barcodes* as introduced by Carlsson et.al. [9] and Ghrist [10]. A barcode is a representation of the kth persistent homology as a collection of horizontal line segments. In a barcode, we place the basis vectors for the homology on the vertical axis where order does not important whereas the

horizontal axis represents the life span of each basis element in terms of the scale parameter ε. When we draw the vertical line at a particular ε_i, the number of intersecting line segments in barcodes is the dimension of the corresponding homology group, i.e. the Betti number for that parameter ε_i. In Figure 6.3, one can see barcodes together with the rips complex corresponding to a particular ε. Notice that there will always be one connected component as ε grows large, the 0th Betti number β_0 is always going to be 1 eventually.

Note that, the longest living topological features in the barcode are the features that we can think of as important for the point cloud, whereas the shorter ones can be seen as noise in the point cloud.

7. FILTERED MATROIDS

In this chapter, we build a theoretical framework for filtered simplicial and chain complexes. Persistent homology and various representations of persistent homology classes are built on such complexes. While we built a solid theoretical foundation for filtered complexes, we also found an intriguing combinatorial representation we call *filtered matroids* for homologies of filtered complexes which we represent by using dendrograms labeled by circuits in the cophenetic matroid that we are going to define in the next chapter.

7.1 Posets and Order Ideals

A poset is a set P together with an anti-symmetric reflexive and transitive relation $\leqslant_P$. A function $f\colon P \to Q$ between two posets is called *order preserving* if $x \leqslant_P y$ implies $f(x) \leqslant_Q f(y)$ for every $x, y \in P$. Given two order preserving maps $f, g\colon P \to Q$ we say that g dominates f if $f(x) \leqslant_Q g(x)$ for every $x \in P$. A subset $\mathscr{I}$ of a poset P is called an *order ideal* if for every $y \in \mathscr{I}$ and $x \in P$, if $x \leqslant y$ then $x \in \mathscr{I}$.

7.2 Matroids

A *matroid* M is a pair $(E, \mathscr{I})$ where E is a non-empty set and $\mathscr{I}$ is a non-empty order ideal in the poset $(2^E, \subseteq)$ such that for every $A, B \in \mathscr{I}$ with $|A| < |B|$ there is an element $x \in B \backslash A$ such that $\{x\} \cup A \in \mathscr{I}$. Elements of $\mathscr{I}$ are called *independent sets*.

7.3 The Rank Function of a Matroid

Let M be a matroid on a finite ground set E. The rank $r(X)$ of a subset $X \subseteq E$ is the cardinality of the largest independent set contained in X. In other words

$$r(X) = \max\{|A| \in \mathbb{N} \mid A \subseteq X \text{ and } A \in \mathscr{I}\} \tag{7.1}$$

Notice that the rank function $r\colon 2^E \to \mathbb{N}$ is order preserving and is dominated by the cardinality function $|\cdot|\colon 2^E \to \mathbb{N}$.

We can convert matroids to rank functions and vice versa. To show this we need the following definition: A poset map $r\colon 2^E \to \mathbb{N}$ is called *semimodular* or *submodular* if

$$r(A \cup B) \leqslant r(A) + r(B) - r(A \cap B) \tag{7.2}$$

for every $A, B \in 2^E$. A submodular map r is called *modular* if the inequality is replaced with an equality, i.e. when r satisfies the *inclusion/exclusion* principle.

Theorem 7.1 ([78, Chap. 2.5, pg 69]). Let E be a finite set and let $r\colon 2^E \to \mathbb{N}$ be a poset map dominated by the cardinality function. Then r is the rank function of a matroid if and only if r is *submodular*.

Thus Theorem 7.1 gives us a license to replace any matroid with its rank function, and vice versa.

7.4 Morphisms of Matroids

Assume (E, r_E) and (F, r_F) are two matroids given by rank functions. A set map $f\colon E \to F$ is called *a morphism of matroids* if $r_F(f(A)) \leqslant r_E(A)$ for every finite subset A of E. One can easily see that the identity map is a morphism of matroids, and the composition of any two morphisms is again a morphism. So, we have a category of matroids.

7.5 Induced Matroids

Let (E, r_E) be a matroid and assume $\pi\colon F \to E$ is any function. Let us define

$$\pi^* r_E(A) := r_E(\pi(A)) \tag{7.3}$$

for every finite subset A of F. The following Lemma is pretty straightforward and its proof is left to the reader.

Lemma 7.2. The pair (F, r_F) is a matroid and $\pi\colon (E, r_E) \to (F, \pi^* r_E)$ is a morphism of matroids.

7.6 Circuits Sets in a Matroid

Assume (E, r_E) is a matroid. We call a subset $A \subseteq E$ as *a circuit* if $r_E(A) = |A| - 1$ and for every proper subset B of A we have $r_E(B) = |B|$.

Proposition 7.3. Given any finite subset $X \subseteq E$ with $r_E(X) < |X|$, there are circuits $A_1, \ldots, A_n$ such that $X = \bigcup_{i=1}^{n} A_i$.

Proof. We give the proof of induction on the size of X. For $|X| = 1$, X is already a circuit and the statement is obviously true. So, let us assume the statement holds for every $k \leqslant n$ and let $|X| = n+1$ with $r_E(X) \leqslant n$. Take any element $x \in X$ and consider the set $\mathscr{U}$ of all subsets $A \subseteq X$ such that $x \in A$ and A are a circuit. Since $\mathscr{U}$ is a non-empty finite poset, there are maximal elements. Let $Y \in \mathscr{U}$ be such a maximal set. If it is already $Y = X$ one can stop. Otherwise, we remove x from X and proceed by induction. □

7.7 Filtered Matroids

Let P be an indexing poset. A filtered matroid $(M_\varepsilon)_{\varepsilon \in P}$ is a set of pairs $(E_\varepsilon, r_\varepsilon \colon 2^{E_\varepsilon} \to \mathbb{N})$ indexed by P where E_ε is a set and r_ε is a rank function. We must also have functions $\psi_{\varepsilon,\eta} \colon E_\varepsilon \to E_\eta$ that satisfy the conditions

$$\psi_{\eta,\nu} \circ \psi_{\varepsilon,\eta} = \psi_{\varepsilon,\nu} \qquad r_\eta(\psi_{\varepsilon,\eta}(A)) \leqslant r_\varepsilon(A) \tag{7.4}$$

for every finite set $A \subseteq E_\varepsilon$ and for every $\varepsilon \leqslant \eta \leqslant \nu$ in P.

Here is another interpretation: Let us view P as a category such that there is a unique morphism $x \to y$ whenever $x \leqslant y$ in P. Then a filtered matroid is a functor from P into the category of matroids.

7.8 Ramification of Circuits

Let us assume P is an indexing poset and let $(E_\varepsilon, r_\varepsilon)$ be a filtered matroid over P with structure maps $\psi_{\eta,\varepsilon} \colon E_\varepsilon \to E_\eta$ for every $\varepsilon \leqslant \eta$ in P. A circuit $A \subseteq E_\varepsilon$ is said to be *ramified* at $\eta > \varepsilon$ if $r_\eta(\psi_{\eta,\varepsilon}(A)) < r_\varepsilon(A)$.

Theorem 7.4. Assume $(E_\varepsilon, r_\varepsilon, \psi_{\eta,\varepsilon})$ is a filtered matroid over the poset $\mathbb{R}_+$. For every ε and for every circuit A of E_ε, one can write the ramification information as a finite rooted tree whose edges are labeled by circuits.

Proof. Assume A is ramified at $\eta > \varepsilon$, i.e. $r_\eta(\psi_{\eta,\varepsilon}(A)) \leqslant r_\varepsilon(A) = |A| - 1$. Then by Proposition 7.3, we can write $\psi_{\eta,\varepsilon}(A)$ as a union circuits. Since A is finite, A can only ramify finitely many times. □

The rooted tree of a circuit A is going to be called the *ramification tree* or the *ramification dendrogram* of the circuit A.

Example 7.5. For every $\varepsilon \in [0,\infty)$ let us define $s_\varepsilon \colon \mathbb{R}^n \to \mathbb{R}^n$ as

$$s_\varepsilon(x_1,\ldots,x_n) = \begin{cases} (x_1,\ldots,x_n) & \text{if } 0 \leqslant \varepsilon < 1, \\ (0,\ldots,0,x_i,\ldots,x_n) & \text{if } i \leqslant \varepsilon < i+1, \\ (0,\ldots,0) & \text{if } \varepsilon \geqslant n+1. \end{cases} \tag{7.5}$$

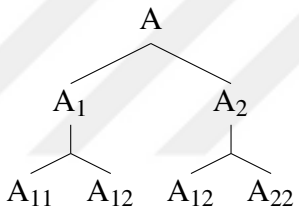


Figure 7.1 : The rooted tree representation of the matroid given in Example.

Let $\mathscr{F}_n$ be the set of all finite subsets of $\mathbb{R}^n$ and define $r_\varepsilon \colon \mathscr{F}_n \to \mathbb{N}$ by

$$r_\varepsilon(A) = \dim s_\varepsilon(A) \tag{7.6}$$

One can check that this is a filtered matroid. Consider

$$A = \{(1,1,1,1),(1,1,2,2),(1,2,3,3),(3,5,6,6)\} \tag{7.7}$$

where we have $r_0(A) = 3$ and every subset of A has rank 3 which means A is a circuit. But

$$s_1(A) = \{(0,1,1,1),(0,1,2,2),(0,2,3,3),(0,5,6,6)\} \tag{7.8}$$

has rank 2, and therefore, is not a circuit. We can write $s_1(A)$ as a union of circuits of maximal rank 2, $s_1(A) = s_1(A_1) \cup s_1(A_2)$, where

$$\begin{aligned} A_1 &= \{(1,1,1,1),(1,1,2,2),(1,2,3,3)\} \\ A_2 &= \{(1,1,2,2),(1,2,3,3),(1,5,6,6)\} \end{aligned} \tag{7.9}$$

These circuits further reduce at $\varepsilon = 2$ and we split

$$s_2(A_1) = s_2(A_{11}) \cup s_2(A_{12}) \quad \text{and} \quad s_2(A_2) = s_2(A_{12}) \cup s_2(A_{22}), \tag{7.10}$$

where

$$\begin{aligned} A_{11} &= \{(1,1,1,1),(1,1,2,2)\}, \\ A_{12} &= \{(1,1,2,2),(1,2,3,3)\}, \\ A_{22} &= \{(1,2,3,3),(3,5,6,6)\}, \end{aligned} \tag{7.11}$$

and each set has rank 1. These sets preserve their ranks until $\varepsilon = 4$ and after $\varepsilon \geqslant 4$ all subset reduce to 0. Thus we can write the tree as shown in Figure 7.1.

8. THE COPHENETIC MATROID

The previous chapter sets the necessary background on filtered matroids that we are going to need for the material we are going to cover in this chapter. The cophenetic matroid is a filtered matroid that comes out naturally as the homology of a filtered simplicial complex. Recall also that developing dendrogram-like visual representations for higher dimensional homology classes had been a formidable task. Now that we can represent how circuits evolve in a filtered matroid, we can write dendrogram-like visual representations for higher dimensional homology classes from the naturally associated cophenetic matroid of the homology of a filtered simplicial complex.

8.1 Multi-dimensional Persistence and The *no-go* Theorem of Bauer et.al.

Let $n \geqslant 1$ and let us consider the poset

$$\mathbb{R}^n_+ = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_i \geqslant 0, i = 1, \ldots, n\} \tag{8.1}$$

together with the partial ordering $(x_1, \ldots, x_n) \preceq (y_1, \ldots, y_n)$ if $x_i \leqslant y_i$ for every $1 \leqslant i \leqslant n$. Given a filtered simplicial complex $\mathscr{K}_\varepsilon$ over $\mathbb{R}^n_+$, one may try to construct barcodes similar to the ordinary barcodes of [9, 10]. Barcodes are complete invariants due to the fact that the representation theory of the poset $\mathbb{R}_+$ is rather simple. However, no such simple representations exist for filtered complexes over $\mathbb{R}^n_+$ since the representation theory of the poset $\mathbb{R}^n_+$ and its discretization $\mathbb{N}^n$ are both wild for $n \geqslant 2$ by [79].

8.2 Carlsson-Zomorodian Rank Function

The *no-go* result of [79] forces us to come up with new representations to depict evolutions of persistent homology classes over a scale parameter. One such example is by Carlsson and Zomorodian [80].

Assume M_ε is a $\mathbb{R}^n_+$-filtered vector space where we assume $\dim_{\mathbb{R}}(M_\varepsilon)$ is finite for every $\varepsilon \in \mathbb{R}^n_+$. In other words, we have finite dimensional vector spaces M_ε for each

$\varepsilon \in \mathbb{R}^n_+$ together with structure maps $\psi_{\varepsilon,\eta} \colon M_\varepsilon \to M_\eta$ for every $\varepsilon \preceq \eta$. Then the Carlsson-Zomorodian rank function of M is defined to be

$$\rho(\varepsilon,\eta) := \dim_{\mathbb{R}} \psi_{\varepsilon,\eta}(M_\varepsilon) \tag{8.2}$$

for every $\varepsilon \preceq \eta \in \mathbb{R}^n_+$ [80, Definition 6].

8.3 Carlsson-Zomorodian Matroid

There is a finer invariant than Carlsson-Zomorodian rank function given by a filtered matroid.

Proposition 8.1. Given any filtered finite dimensional vector spaces (M_ε), the function r_ε defined as

$$r_\varepsilon(A) = \dim_{\mathbb{R}} Span_{\mathbb{R}}(A) \tag{8.3}$$

for every finite subset A of M_ε yields a filtered matroid.

Proof. The function r_ε is dominated by the cardinality function, and it satisfies

$$dim_{\mathbb{R}} \psi_{\varepsilon,\eta}(Span_{\mathbb{R}}(A)) \leqslant \dim_{\mathbb{R}} Span_{\mathbb{R}}(A) \tag{8.4}$$

for every $\varepsilon \preceq \eta$, and thus, the collection $(r_\varepsilon)_{\varepsilon \in \mathbb{R}^n_+}$ is a filtered matroid. □

8.4 Cophenetic Matroid

From this point onward, we work with the poset $\mathbb{R}_+$, and a filtered simplicial complex $(\mathscr{K}_\varepsilon)_{\varepsilon \in \mathbb{R}_+}$ such that structure morphisms are inclusions. Recall that for this filtered complex we have cycles $Z^\varepsilon_k := ker(d^\varepsilon_k)$ and boundaries $B^\varepsilon_k := im(d^\varepsilon_{k+1})$ and homology groups $H_k(\mathscr{K}_\varepsilon) := Z^\varepsilon_k / B^\varepsilon_k$ for every $k \in \mathbb{N}$ and $\varepsilon \in \mathbb{R}_+$. We also have connecting linear maps $\psi^k_{\varepsilon,\eta} \colon Z^\varepsilon_k \to Z^\eta_k$ and $\psi^k_{\varepsilon,\eta} \colon B^\varepsilon_k \to B^\eta_k$ for every pair $\varepsilon \leqslant \eta$ and for every $k \in \mathbb{N}$. Note that since $\mathscr{K}_\varepsilon \subseteq \mathscr{K}_\eta$ for every $\varepsilon \leqslant \eta$, the induced maps $\psi^k_{\varepsilon,\eta}$ on cycles and boundaries are also monomorphisms. However, even if this is the case, the induced maps in homology need not be monomorphisms.

Let us write F^ε_k for the set of all finite subsets of Z^ε_k. For every $A \in F^\varepsilon_k$ define

$$\begin{aligned} c^k_\varepsilon(A) &= \dim(Span_{\mathbb{R}}(A) + B^\varepsilon_k) - \dim B^\varepsilon_k \\ &= \dim(Span_{\mathbb{R}}(A)) - \dim(Span_{\mathbb{R}}(A) \cap B^\varepsilon_k) \end{aligned} \tag{8.5}$$

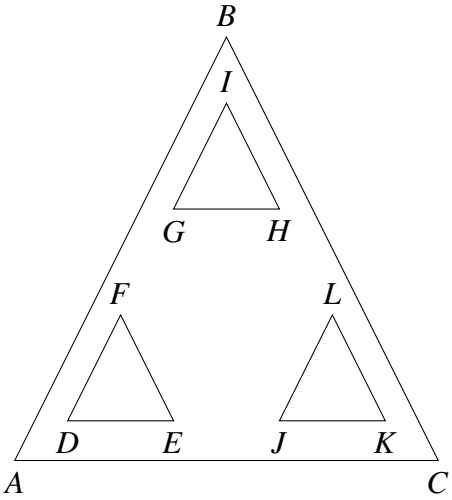


Figure 8.1 : The configuration of points for Subsection 8.3

Notice that c_ε^k is a poset map and is dominated by the cardinality function and we have

$$c_\eta^k(\psi_{\varepsilon,\eta}^k(A)) \leqslant c_\varepsilon^k(A) \tag{8.6}$$

for every $\eta \geqslant \varepsilon$ and $A \in F_k^\varepsilon$. The function c_ε^k is called the *cophenetic rank function* of the filtered complex $\mathscr{K}_\varepsilon$.

Theorem 8.2. The cophenetic rank function $c_\varepsilon^k \colon F_k^\varepsilon \to \mathbb{N}$ is submodular for every $\varepsilon \in \mathbb{R}_+$ and for every $k \in \mathbb{N}$. Thus by Theorem 7.1 for every $k \geqslant 0$ there is a filtered matroid $(M_\varepsilon^k)_{\varepsilon\in\mathbb{R}_+}$ of the filtered simplicial complex $(\mathscr{K}_\varepsilon)_{\varepsilon\in\mathbb{R}_+}$.

Proof. Given a finite set A in Z_k^ε its cophenetic rank $c_\varepsilon^k(A)$ is the dimension of $Span_\mathbb{R}(A)$ in the quotient vector space $H_k(\mathscr{K}_\varepsilon) = Z_k^\varepsilon / B_k^\varepsilon$. Now, apply Lemma 7.2. □

The matroid $(M_\varepsilon^k)_{\varepsilon\in\mathbb{R}_+}$ given in Theorem 8.2 is called the *k-th cophenetic matroid* of a filtered simplicial complex $(\mathscr{K}_\varepsilon)_{\varepsilon\in\mathbb{R}_+}$.

Example 8.3. Consider the configuration of points given in Figure 8.1. Assume we put a filtration where at $\varepsilon = 0$ we have disjoint points, and at $\varepsilon = 1$ the smaller triangles DEF, GHI, and JKL are formed. Then at $\varepsilon = 2$ the large triangle ABC is formed, and at $\varepsilon = 3$ we fill-in the region between the large triangle ABC and the three smaller triangles. Finally at $\varepsilon = 4,5,6$ we fill-in the smaller triangles DEF, GIH and JKL in order.

Consider the set of first homology classes $X = \{ABC, DEF, GIH, JKL\}$ that forms as an independent set at $\varepsilon = 2$. But at $\varepsilon = 3$ when we fill in the region between ABC and

the smaller triangles, they become linearly dependent, and we get a circuit. As we kill the smaller triangles we get

$$\begin{aligned} X &= \{ABC, GIH, JKL\} \cup \{DEF\} \\ &= \{ABC, JKL\} \cup \{GIH\} \cup \{DEF\} \\ &= \{ABC\} \cup \{JKL\} \cup \{GIH\} \cup \{DEF\} \end{aligned} \tag{8.7}$$

as unions of circuits. We represent these splittings as a tree in Figure 8.2.

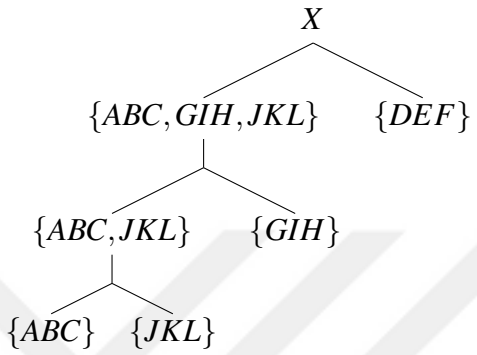


Figure 8.2 : Tree representation of the cophenetic matroid of Example 8.3.

8.5 Homological Cophenetic Distance

Now, for each pair of cycles α and β in Z_k^ε representing classes in $H_k(\mathscr{K}_\varepsilon)$, one can test the rank of the pair $\{\alpha, \beta\}$ at every $\eta > \varepsilon$. If the rank of the pair $\{\psi^k_{\varepsilon,\eta}(\alpha), \psi^k_{\varepsilon,\eta}(\beta)\}$ is less than 2, then we will say that the cycles α and β *merged* at time η. Thus we can give the following definition:

Definition 8.4. *The k-th homological cophenetic distance* is

$$d_k(\alpha, \beta) = \inf \left\{ \eta - \varepsilon \geqslant 0 \mid c^k_\eta(\{\psi^k_{\varepsilon,\eta}(\alpha), \psi^k_{\varepsilon,\eta}(\beta)\}) < 2 \right\}. \tag{8.8}$$

for every $\alpha, \beta \in H_k(\mathscr{K}_\varepsilon)$ and for every $k \geqslant 0$.

Proposition 8.5 ([60]). The homological cophenetic distance d_k on $H_k(\mathscr{K}_\varepsilon)$ is a non-Archimedean metric for every $\varepsilon \geqslant 0$ and for every $k \geqslant 0$.

Proof. Assume $\alpha, \beta, \gamma \in Z_k(\mathscr{K}_\varepsilon)$. Assume, by way of contradiction, that

$$d_k(\alpha, \beta) > \max(d_k(\alpha, \gamma), d_k(\gamma, \beta)). \tag{8.9}$$

This means there are indices $\eta > \mu$ such that the pair (α, β) becomes linearly dependent in $H_k(\mathscr{K}_\eta)$ while the pairs (α, γ) and (γ, β) become linearly dependent at

an earlier time in $H_k(\mathscr{K}_\mu)$. Then there are non-zero scalars $a, b \in k$ such that

$$\alpha = a\gamma,\ \beta = b\gamma \text{ which implies } b\alpha = a\beta \tag{8.10}$$

in $H_k(\mathscr{K}_\mu)$ which is a contradiction since α and β become linearly dependent at a later time $\eta > \mu$. □

8.6 Non-Archimedean Metrics and Hierarchical Clustering

It has been known that hierarchical clustering methods and non-Archimedean metrics are intimately related [40, 51, 52]. But it was Carlsson and Memoli who proved that hierarchical clustering methods and non-archimedean metrics are naturally equivalent in [34].

Recall that the connected components of a topological or a metric space X are encoded in the zeroth homology classes of X. This means one can naturally compare various clustering schemes on a dataset embedded in X with the homological cophenetic distance for the zeroth homology by varying the underlying metric of X if the ambient space X allows it.

The immediate corollary is that the homological cophenetic metric we defined in Definition 8.4 does correspond to a unique hierarchical clustering scheme on every homology group.

9. COBORDISMS

In the previous chapter, we related homologies of filtered complexes with filtered matroids and then represented the homological information as dendrograms (rooted trees) labeled with circuits in a special matroid built on homology of a filtered complex. In this chapter, we are going to relate the homological information one can derive from filtered complexes with cobordisms of punctured spheres, which themselves are punctured higher dimensional spheres.

There is a big challenge in examining higher dimensional homology classes by using dendrogram structures as in the zeroth-dimension. For the first dimension, there are many possible phases such as merging, splitting, creation, and annihilation; one can find an algebraic representation that correspond to multiplication, comultiplication, unit, and counit [81]. In this thesis, we propose representing higher dimensional homology classes of a filtered complex as a cobordism of punctured speheres.

9.1 Hurewicz map

Assume that our data set D is sampled from a manifold M embedded in $\mathbb{R}^n$. Assume also we created a filtered simplicial complex $(\mathscr{K}_\varepsilon)$ from D. Since we work with filtered complexes $(\mathscr{K}_\varepsilon)_{\varepsilon\in\mathbb{R}_+}$, the corresponding vector spaces of cycles $(Z_k^\varepsilon)_{\varepsilon\in\mathbb{R}_+}$ and boundaries $(B_k^\varepsilon)_{\varepsilon\in\mathbb{R}_+}$ are also filtered.

First, we recall the following version of the rational Hurewicz Theorem:

Proposition 9.1 ([82, 83]). Assume M is a simply connected topological space with $\pi_n(M) = 0$ for $1 \leqslant n < r$. Then the rational Hurewicz map $\pi_n(M)\otimes\mathbb{Q} \to H_n(M)\otimes\mathbb{Q}$ is an isomorphism for $1 \leqslant n < 2r-1$ and is a surjection for $n = 2r-1$.

We use Proposition 9.1 as follows.

Proposition 9.2. Let us assume $\pi_n(C) = 0$ for each connected component C of M for all $0 \leqslant n < r$ for some r. Then for all $0 \leqslant n \leqslant 2r-1$ all cycles in Z_n^ε, in particular, every boundary in B_n^ε comes from an embedded n-sphere in M.

Proof. If a connected component C is simply connected, i.e. when $r = 1$, then we use Proposition 9.1. If C fails to be simply-connected then the classical Hurewicz map $\pi_1(C) \to H_1(C)$ is already surjective for every path connected component C of M. □

Now, we give the definition of cobordism as follows:

Definition 9.3. [81] A cobordism between any two compact $n-1$ dimensional manifolds M_1 and M_2 is a compact n dimensional manifold M where $\partial M = M_1 \sqcup M_2$.

9.2 Dendrograms of Circuits as Cobordisms of Spheres

Assume our data D is sampled from a manifold M that satisfies the hypothesis of Proposition 9.2. Assume also that we constructed a filtered complex $(\mathscr{K}_\varepsilon)$ out of D.

Theorem 9.4. The ramification tree of every circuit A in $H_n(\mathscr{K}_\varepsilon)$ can be represented by a $n+1$ dimensional cobordism of disjoint n-spheres for every $0 \leqslant n \leqslant 2r-1$.

Proof. For every $0 \leqslant n < 2r-1$, and circuit of homology n-cycles A there is a $n+1$-sphere with k-punctures such that punctures represent classes in A and the $n+1$-sphere implements the linear dependence of elements in A. This is because every cycle $\alpha \in Z_n^\varepsilon$ and boundary $\beta \in B_{n+1}^\varepsilon$, and their every scalar multiple, is represented with a sphere via the Hurewicz map. If a collection A in $H_n(\mathscr{K}_\varepsilon)$ is a circuit, the elements in A represented by n-spheres have to be linearly dependent given by a boundary which is a $n+1$-sphere. The result follows. □

10. EXPERIMENTS

To determine if our research is sound, we performed numerical experiments on synthetic and different real datasets. In this section, we are going to summarize these experiments.

As we discussed earlier, one of the main problems we tackle in this thesis is to develop dendrogram-like visualizations for higher homology groups. One of the main objects of this thesis is the non-archimedean metric we developed for all homological degrees. We know that hierarchical clustering schemes and non-archimedean metrics are equivalent, and therefore, we can compare our results for the zeroth homology with that of dendrograms of hierarchical clustering. To make sure that our method produces results comparable with ordinary dendrograms, we need to statistically compare the dendrograms from hierarchical clustering using the homological cophenetic distance for the zeroth persistent homology with dendrograms coming from the Euclidean distance. This statistical comparison is then repeated with hierarchical clustering results from different distance metrics. Finally, we measure the quality of the clusters we obtain from the hierarchical clustering algorithm using different metrics with different clustering approaches and then compare our results.

10.1 Our Experiments in Detail

We start by constructing the Vietoris-Rips complex $\mathscr{R}_\varepsilon(D)$ for given a point cloud D and the maximum value of the proximity parameter, $\varepsilon_{\max}$. We then get equally-spaced values of the proximity parameter ε between 0 and $\varepsilon_{\max}$. Our algorithm builds the homological cophenetic distance matrix in the following steps where we repeat the process for each pair ε_1 and ε_2 with $\varepsilon_1 \leqslant \varepsilon_2$:

Step 1. For the pair ε_1 and ε_2, we have the natural map in homology,

$$\psi^k_{\varepsilon_1,\varepsilon_2} : H_k(\mathscr{R}_{\varepsilon_1}) \to H_k(\mathscr{R}_{\varepsilon_2}) \tag{10.1}$$

coming from the embedding $\mathscr{R}_{\varepsilon_1} \subset \mathscr{R}_{\varepsilon_2}$.

Step 2. Take two cycles $\alpha,\ \beta \in H_k(\mathscr{R}_{\varepsilon_1})$ and obtain α' and β' by padding α and β with suitable number of 0's to obtain $\psi^k_{\varepsilon_1,\varepsilon_2}(\alpha)$ and $\psi^k_{\varepsilon_1,\varepsilon_2}(\beta)$.

Step 3. Check whether $\psi^k_{\varepsilon_1,\varepsilon_2}(\alpha)$ and $\psi^k_{\varepsilon_1,\varepsilon_2}(\beta)$ are linearly independent by evaluating the rank of the differential matrix $\mathscr{D}$ at ε_2 with appending α', β' and α',β' together. We look at $r_\alpha := \operatorname{rank}(\mathscr{D}_\alpha) - \operatorname{rank}(\mathscr{D})$, $r_\beta := \operatorname{rank}(\mathscr{D}_\beta) - \operatorname{rank}(\mathscr{D})$ and $r_{\alpha,\beta} := \operatorname{rank}(\mathscr{D}_{\alpha,\beta}) - \operatorname{rank}(\mathscr{D})$ with the following cases:

$$\begin{cases} \alpha \text{ and } \beta \text{ both die} & \text{if } r_{\alpha,\beta} = 0, \\ \alpha \text{ and } \beta \text{ both live} & \text{if } r_{\alpha,\beta} = 2, \\ \alpha \text{ dies and } \beta \text{ lives} & \text{if } r_{\alpha,\beta} = 1 \text{ and } r_\alpha = 0, \\ \alpha \text{ lives and } \beta \text{ dies} & \text{if } r_{\alpha,\beta} = 1 \text{ and } r_\beta = 0, \\ \alpha \text{ and } \beta \text{ merge} & \text{if } r_{\alpha,\beta} = 1,\ r_\alpha = 1 \text{ and } r_\beta = 1. \end{cases} \tag{10.2}$$

We also summarize our experiment for comparing dendrograms as a pseudo-code in Algorithm 2.

Input: A point cloud D, $|D| = m$ and a list $\varepsilon = \{\varepsilon_1, \ldots, \varepsilon_{max}\}$.
Output: Dendrograms
begin
  HomDist = $[\ ]_{m\times m}$
  $\mathscr{R}_\varepsilon(P) \longleftarrow$ Vietoris-Rips filtration ;
  for *every* $\varepsilon_1, \varepsilon_2 \in \varepsilon$ **do**
    for *every* $\alpha_i, \alpha_j \in H_0(\mathscr{R}_\varepsilon)$ **do**
      $d_0(\alpha_i, \alpha_j) \longleftarrow \inf\{\varepsilon_2 - \varepsilon_1 \,|\, r_{\alpha_i,\alpha_j} = r_{\alpha_i} = r_{\alpha_j} = 1\}$;
      $\text{HomDist}_{i,j} \longleftarrow d_0(\alpha_i, \alpha_j)$;
    end
  end
  $E(D) \longleftarrow$ EuclideanDist(D);
  $Dend_1 \longleftarrow$ HierarchicalClustering(HomDist(D)) ;
  $Dend_2 \longleftarrow$ HierarchicalClustering$(E(D))$;
  Compare$(Dend_1, Dend_2)$
end

Algorithm 2: Our experiment as pseudo-code.

To compare dendrograms, we used tools coming from the **dendextend** [84] and **vegan** [85] packages of the R programming language [86]. To compute the homological cophenetic distance matrix we used SageMath [87]. In order to compute and visualize clusters, we used Python programming language [88] and its **scikit-learn** library [89]. The source code and the data of the numerical

experiments we conducted in this thesis can be found on the authors' GitHub page at `https://github.com/ismailguzel/TDA-HC`.

10.2 The First Experiment

For our experiments, we generated a synthetic data cloud D in $\mathbb{R}^2$, $|D| = 20$ with labeled the first 20 letters, using the continuous uniform distribution over $[0,1)$ using the **numpy** package [90] in Python. To get the Vietoris-Rips complexes, we used the **dionysus2** [91] package in Python. To calculate the cophenetic distance of two homology classes, we use the linear algebra and homology packages of the computer algebra system Sage [87]. Before applying the hierarchical clustering to the data D, we also calculated the distance matrix using the Euclidean distance, i.e. the L^2 norm, $E(x,y) = \|x-y\|_2$ for every $x,y \in D$.

10.2.1 Barcodes and dendrograms

Now, let us compare the zeroth ordinary barcodes and our dendrogram for the zeroth homology. The left-hand side of Figure 10.1 is the dendrogram we obtained from cophenetic homological distance matrix for the zeroth homology. The right-hand side of Figure 10.1 is the ordinary barcode obtained from the zeroth persistent homology which displays the birth and death times of each homology class, whereas the left hand side is the dendrogram that indicates which classes merge.

10.2.2 Comparison of dendrograms

Next, we apply the hierarchical clustering (with single linkage), using the Euclidean distance matrix $E(D)$, and the homological cophenetic distance matrix HomDist(D) for the zeroth persistent homology. The resulting dendrograms are given in Figure 10.2.

In a tanglegram representation as discussed in Section 3.3.3, we also align the labels from both dendrograms without changing the underlying cluster structure. In order to optimize the alignment of the labels from the two dendrograms without changing the underlying cluster structure, we use the `untangle` function from the package **dendextend**. After the alignment process, we use the function `tanglegram` to show two dendrograms in one figure side by side with their labels connected by lines.

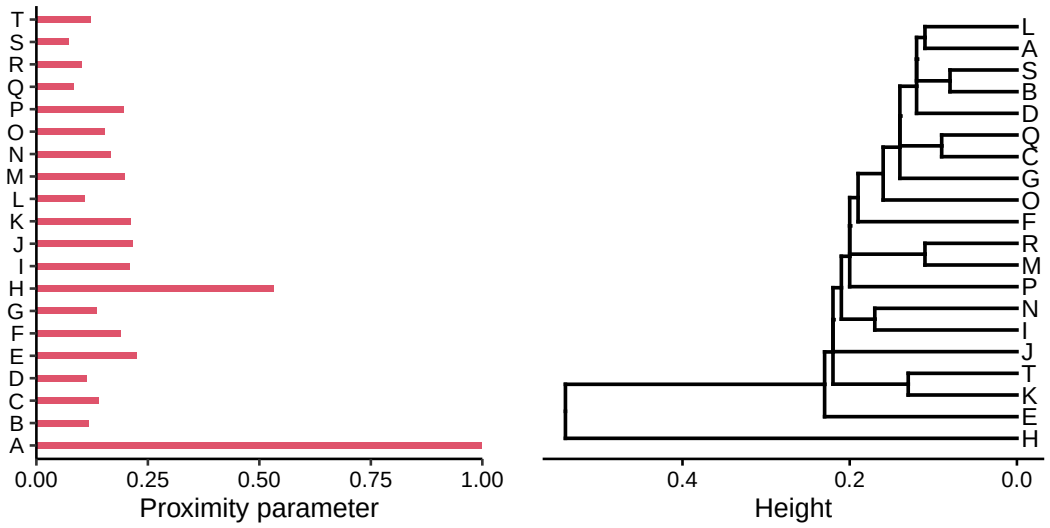


Figure 10.1 : The zeroth ordinary barcodes and hierarchical enriched barcodes in TDA.

Also, the degree of entanglement of two dendrograms is measured using the function `entanglement`. The resulting statistic is a measure of how well the labels of two dendrograms are aligned. The output of these functions is given in Figure 10.3.

10.2.3 Test results and their analysis

In the next phase, we need to statistically compare dendrograms. We use the Mantel test via the function `mantel` (See Section 3.3.2) provided in **vegan** package. But, before we use the Mantel test, we need the cophenetic distance matrix from each dendrogram by using the `cophenetic` function provided in **stats** package [74]. Since dendrograms are just representation of hierarchical clustering, we directly use the cophenetic distance matrix in the hierarchical clustering function `hclust` provided in **stats** package.

Based on the results given in Figure 10.4, the Mantel statistic value 0.9998 indicates that there is a relatively strong positive correlation between the dendrogram $Dend_1$ from the homological distance matrix HomDist(D) and the dendrogram $Dend_2$ from the Euclidean distance matrix $E(D)$. The p-value of 0.001 indicates that our results are statistically significant. Note that since this test is based on random permutations, the same code will always reach at the same observed correlation r but seldomly the same p-value. Thus, we can confidently conclude that there is a high statistically significant

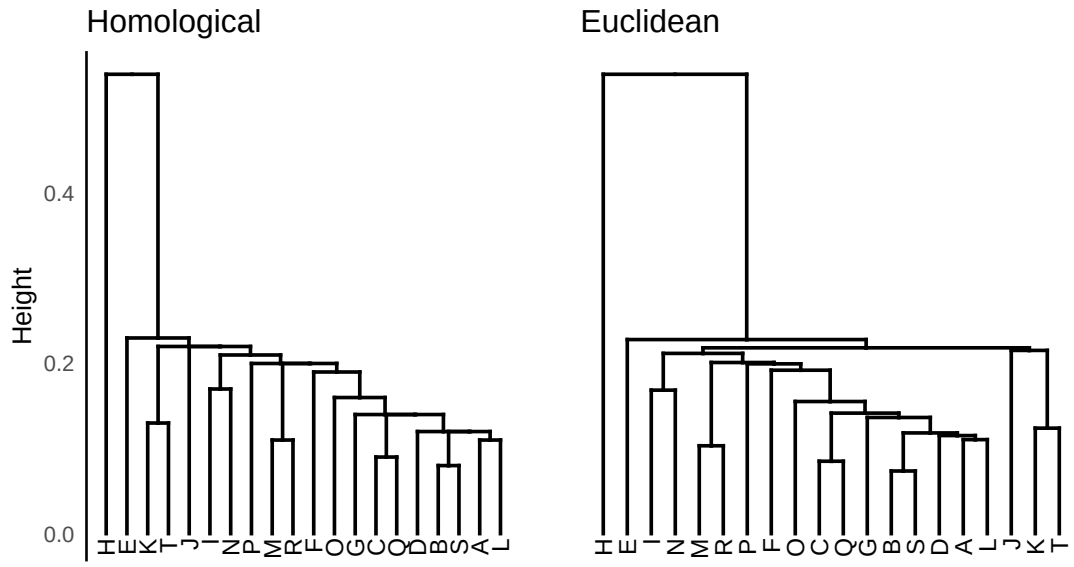


Figure 10.2 : Two dendrograms: one from homology and the other from the Euclidean distance.

relation between the dendrograms obtained from the homology and hierarchical clustering methods.

On the other hand, we also repeat our procedure 100 times to get robust statistics. For this purpose, we sampled 20 points uniformly randomly from the unit square $[0,1] \times [0,1]$, and then we compared their Euclidean distance and homological cophenetic distance matrices using the Mantel test. The median statistic was 0.94 with a p-value of 0.001. A histogram of resulting statistics is shown in Figure 10.5.

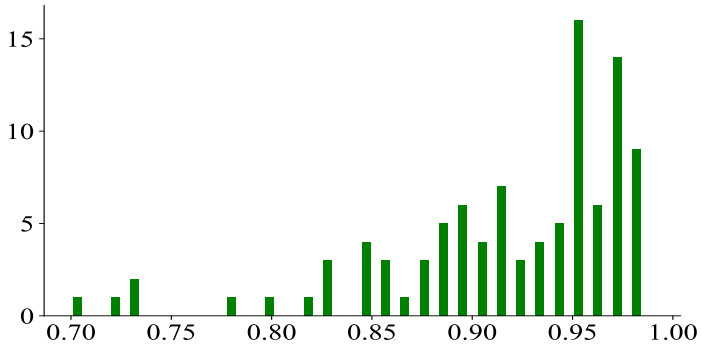


Figure 10.5 : A histogram of Mantel statistics from random point clouds.

The statistical evidence we collected supports our hypothesis that hierarchical clustering and zeroth persistent homology yield the same topological information about the connected components of the sampled manifold using completely different methods. While hierarchical clustering exclusively relies on metric structure alone,

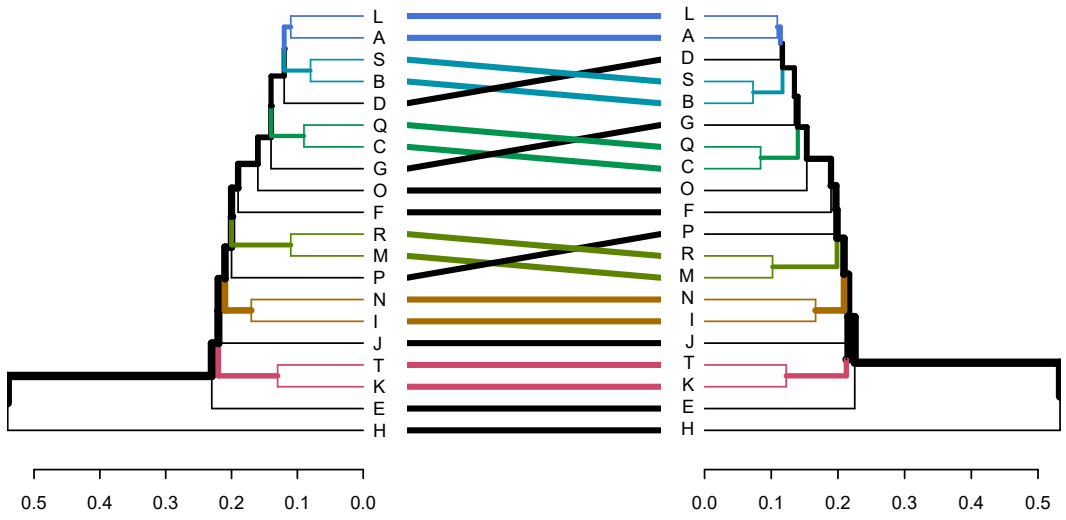


Figure 10.3 : Tanglegram of the dendrograms of the cophenetic (left) and Euclidean distances (right) with the entanglement of 0.01 using L^2 norm after applying to untangle to get optimal alignment.

```
Mantel statistic based on Pearson's product-moment correlation
Call:
mantel(xdis = cophenetic(hc1), ydis = cophenetic(hc2))

Mantel statistic r: 0,9998
      Significance: 0,001

Upper quantiles of permutations (null model):
   90%    95%  97,5%    99%
0,0754 0,8907 0,9065 0,9174
Permutation: free
Number of permutations: 999
```

Figure 10.4 : Mantel test result.

persistent homology relies on simplicial technology to derive its results. The highly correlated nature of the result comes from the fact that the Vietoris-Rips complex is derived from the same metric structure used in hierarchical clustering. However, the homological machinery opens new avenues for statistical data analysis in different directions.

10.3 The Second Experiment

For our second experiment, we first used a subset of 24 of cities in Türkiye whose coordinates are encoded as longitudes and latitudes in radians. See Figure 10.6. We, then, consider the different datasets to get a comparison of our approach vs others in the clustering algorithm with the different linkages.

Figure 10.6 : A sample of cities in Türkiye. For the map, we used Generic Mapping Tools [1].

10.3.1 A full comparison of metrics

In Section 10.2.2, we explicitly compared the cophenetic homological metric with the Euclidean metric. In this section, we extend the comparison to a variety of metrics. As we discussed in Section 3.3.2, we are going to use the Mantel test for this comparison.

For our comparisons we used the L^2-metric (also known as the *Euclidean metric*), the L^1-metric (also known as *taxicab distance*, *city-block distance*, or the *Manhattan distance*), L^p-metric (also known as the *Minkowski distance*) with $p = 1/2$, the cosine similarity converted to a dissimilarity function, and Bray-Curtis dissimilarity [92], and [41, Eq. 7.58]. Our resulting Mantel test table is given in Table 10.1.

Table 10.1 : The pairwise Mantel statistics of metrics on the cities of Türkiye dataset.

Metrics	Bray-Curtis	Cosine	Manhattan	Euclidean	Minkowski	Homological
Bray-Curtis	1.00	0.64	0.96	0.90	0.90	0.90
Cosine		1.00	0.61	0.52	0.69	0.59
Manhattan			1.00	0.96	0.87	0.97
Euclidean				1.00	0.75	0.98
Minkowski					1.00	0.78
Homological						1.00

The results again indicate that our homological cophenetic distance, produces results most similar to the Euclidean metric and the Manhattan metric, and is most dissimilar to the cosine similarity. Homology is a topological invariant, and therefore, is impervious in any perturbations of the underlying metric structure. However, the filtration structure on the Rips complexes we used relies heavily on the underlying

metric. Thus one can surmise that the main contributing factor to this similarity might be the fact that we used Euclidean metric when we formed Rips complexes.

10.3.2 Cophenetic distance on different datasets

In the previous section, we compared the dendrograms coming from different metrics on the coordinates of a small sample of cities in Türkiye. In this section, we are going to evaluate the clusters coming from hierarchical clustering algorithms by varying the metrics on 5 different datasets: all cities in Türkiye, the *Iris Dataset* [93], the *Cancer Coimbra Dataset* [94] and two synthetic datasets that we generated. One of these synthetic datasets has 4 linearly separable clusters and the other contains 4 clusters with mixing along their boundaries. For the synthetic datasets we used `make_blobs` function from the scikit-learn library of Python.

Table 10.2 : Datasets used and their properties.

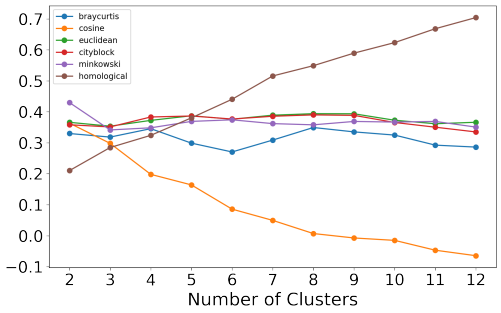
Dataset	#Instances	#Attributes	Supervised	#Classes
Turkish Cities	82	2	No	-
Iris	150	4	Yes	3
Cancer Coimbra	116	10	Yes	2
Synthetic (total separation)	100	100	Yes	4
Synthetic (with mixture)	100	2	Yes	4

10.3.3 Silhouette scores

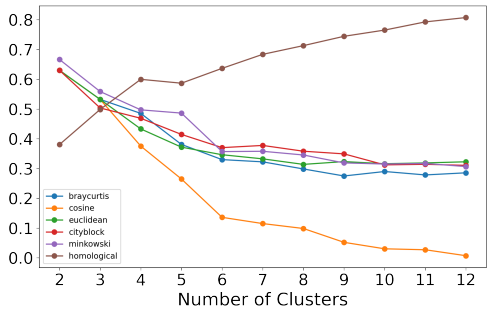
In this section, we compare the resulting clusters of the hierarchical clustering algorithm with different metrics including the homological cophenetic distance. The resulting clusters on each dataset are going to compare using the silhouette scores as discussed in Section 3.3. The results for different numbers of clusters are given in Figure 10.7.

The graphs in Figure 10.7 indicate that one must consider the silhouette score and the *marginal* silhouette scores simultaneously in order to determine the right number of clusters. For the synthetic datasets, the cophenetic metric produced the best silhouette scores. We also see that the silhouette score did indeed determine the right number of clusters for both of the synthetic datasets regardless of the choice of the metric.

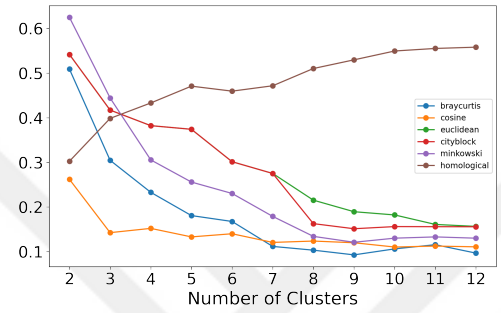
For the Iris datasets, all metrics appear to produce comparable silhouette scores. We notice that even though the original dataset has 3 preset clusters, Figure 10.7



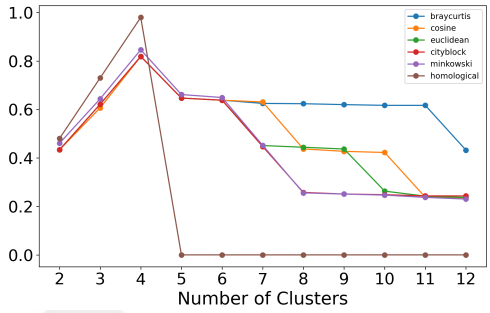
(a) Türkiye cities.



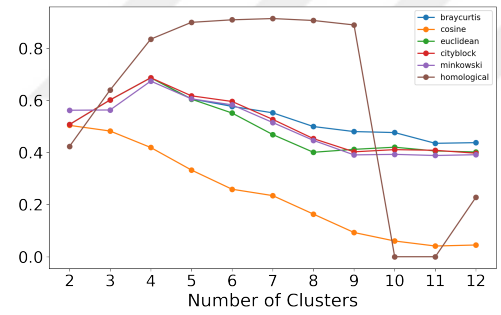
(b) Iris dataset.



(c) Cancer dataset.



(d) Separable synthetic dataset.



(e) Synthetic dataset with mixing.

Figure 10.7 : Silhouette scores for each dataset.

suggests that the optimal number of clusters for the Iris dataset is 2 for all metrics while the homological cophenetic distance suggests that it is 4. However, one must observe that the classes *iris versicolor* and *iris virginica* are intertwined [95–97]. Our computations appear to detect this phenomenon. The homological cophenetic distance result suggests splitting extra subclusters along their intersection while results from other metrics suggest merging these clusters.

A similar phenomenon appears in the Cancer Coimbra dataset. The consensus on the right number of clusters for the Cancer dataset is 2 for all metrics, while the cophenetic metric suggests 5. In [94], the authors use logistic regression, random forests, and

support vector machines to label data points as *control* or *patient* with specificity and sensitivity in the high 80%'s. Our results, on the other hand, were in the 60%'s for both specificity and sensitivity. However, we focus on using hierarchical clustering algorithms with commonly-used metrics and our cophenetic metric to split the dataset into meaningful clusters instead of labeling data points as *control* or *patient*. The results we obtained in Section 10.3.2 indicate that the dataset contains homogeneous meaningful subsets other than *control* and *patient*.

Finally, there appear to be no meaningful clusters for the Turkish cities dataset.

10.3.4 Linkages

There is one more parameter that affects the forming of clusters: the linkages we use to calculate the distances between newly merged clusters, as discussed in Section 3.2. For this set of experiments, we used hierarchical clustering on each dataset using different linkage methods.

Since hierarchical clustering is an unsupervised method, to use measures such as the F1-score and accuracy, we need to find a suitable permutation of the confusion matrix for each dataset. To deal with this problem, we use the Hungarian assignment method as defined in [98]. We then evaluate each model using F1-score (F1), accuracy (Acc.), homogeneity (Hom.), completeness (Comp.) [71], mutual information (M.Info) [69] and Rand index (Rand) [70]. We display our results in Table 10.3. In each table, we report the best-performing linkage method (single (S), complete (C), average (A), or ward (W)) with each evaluation measure. We excluded the linearly separable synthetic dataset because all of the metrics did produce the most optimal result with the appropriate linkage.

The cophenetic distance did produce the best results for the Cancer dataset across the board. For the synthetic dataset with mixing, the results for the cophenetic distance appear to be on par with the other metrics even though it produced the weakest results after the cosine distance for the Iris dataset. Our results indicate that cophenetic distance does produce competitive results on measures such as the F1-score, accuracy, homogeneity, and the Rand index while it shines on measures such as completeness and mutual information score consistently on all datasets. The results indicate that

Table 10.3 : A comparison of metrics on datasets.

(a) Iris dataset.

Metric	F1	Acc.	Hom.	Comp.	M.Info	Rand
Bray-Curtis	0.82W	0.88W	0.69W	0.95S	0.75S	0.82W
Cosine	0.68W	0.79W	0.58W	0.95S	0.64S	0.77W
Manhattan	0.88W	0.92W	0.74W	0.92S	0.82A	0.87W
Euclidean	0.88A	0.92A	0.77A	0.95S	0.84A	0.87A
Minkowski	0.85A	0.90A	0.77A	0.95S	0.80A	0.85A
Homological	0.76W	0.84W	0.58S	1.00S	0.64S	0.78S

(b) Cancer Coimbra dataset.

Metric	F1	Acc.	Hom.	Comp.	M.Info	Rand
Bray-Curtis	0.56S	0.56S	0.02A	0.14S	0.02A	0.50S
Cosine	0.55C	0.55C	0.01S	0.12S	0.01S	0.50C
Manhattan	0.53S	0.53S	0.02A	0.13A	0.02A	0.50S
Euclidean	0.54W	0.54W	0.02A	0.13A	0.02A	0.50W
Minkowski	0.53S	0.53S	0.02A	0.13A	0.02A	0.50S
Homological	0.61W	0.61W	0.03W	1.00S	0.02W	0.52W

(c) Synthetic dataset with mixing.

Metric	F1	Acc.	Hom.	Comp.	M.Info	Rand
Bray-Curtis	1.00A	1.00A	1.00A	1.00A	1.38A	1.00A
Cosine	0.83A	0.91A	0.72C	0.77S	1.38C	0.87A
Manhattan	1.00S	1.00S	1.00S	1.00S	0.99S	1.00S
Euclidean	1.00A	1.00A	1.00A	1.00A	1.38A	1.00A
Minkowski	1.00C	1.00C	1.00C	1.00C	1.38C	1.00C
Homological	0.98A	0.99A	0.95A	1.00S	1.31A	0.98A

cophenetic distance does tend to produce complete clusters that show high average inter-class dissimilarities.

11. CONCLUSIONS

We defined a non-archimedean metric, called the cophenetic metric, on persistent homology classes of all degrees. We then used this metric to sketch rooted tree presentations called dendrograms for zeroth persistent homology classes instead of sketching rooted trees on points in the data set. We note that having a non-archimedean metric persistent homology classes in all degrees allows one to visualize higher homology classes as dendrograms as well.

Since the zeroth homology classes naturally correspond to connected components of the subspace from which our data set is sampled, one can now compare the results of hierarchical clustering schemes with different metrics on data points with the results we obtain from the cophenetic distance on the zeroth homology. To test the soundness of our study, we did numerical experiments on the geographical coordinates of a small sample cities of Türkiye, all cities in Türkiye, the Iris dataset, the Cancer Coimbra dataset, and two synthetic datasets to compare the dendrograms coming the cophenetic metric on the zeroth homology and the dendrograms of hierarchical clustering algorithms by varying metrics in Section 10.3.

The results of our numerical experiments we outlined in Section 10.3 indicate that there is a statistically verifiable strong correlation between the dendrograms coming from the cophenetic distance and the dendrograms coming from other metrics. The statistical evidence we collected supports our hypothesis that hierarchical clustering and zeroth persistent homology together with the cophenetic metric yield statistically verifiable commensurate topological information about the connected components of the datasets we used in our analyses.

We also note that while hierarchical clustering algorithms exclusively rely on a metric structure on the data cloud alone, persistent homology relies on the simplicial technology to derive its results, and therefore, should be impervious to the underlying metric. On the other hand, the Mantel test results in Table 10.1 indicate that homological cophenetic distance and Euclidean distance are most similar. This may

come from the fact that Vietoris-Rips complex we used to calculate our homological invariant uses the Euclidean metric for its filtration structure.

We must add that our results must come with a word of caution: In its most basic form our thesis relies on numerical experiments on large point clouds to extract certain information about the life-time of topological features of the data. This requires us to generate large numbers of simplicial complexes and appeal heavily to computational linear algebra. Due to the high time complexity and large memory requirements of the algorithms we employ, running such numerical experiments and simulations is computationally expensive. Thus any implementation of our techniques should use state-of-the-art parallelization and serialization techniques, as we also did.

In addition to the experimental comparisons of our techniques, we also developed a completely new theoretical framework for filtered simplicial and chain complexes using a combinatorial object called the cophenetic matroid. We showed that the cophenetic metric can be derived from the information encoded in the cophenetic matroid. Moreover, we were also able to represent the same homological information as rooted trees whose vertices are labeled with the circuits of the cophenetic matroid. Our theoretical framework extended to a topological one: we were able to show that the information encoded in the cophenetic matroid can also be described as cobordisms of punctured spheres.

One can extend the results of this study in different directions. The first obvious avenue for extension is replacing the zeroth homology with higher persistent homology and finding a suitable application for the cophenetic distance, or the cophenetic matroid. As we noted above, dendrograms as cobordisms of 0-spheres are adequate in representing the relationships between zeroth persistent homology classes. As we showed in the thesis, for the higher homology classes we need to deal with higher cobordisms of punctured n-spheres [99]. For the first persistent homology, the cobordisms are given by genus-g Riemann surfaces with punctures. Fortunately, there is a complete classification of such surfaces in full [100]. Unfortunately, for higher dimensional homology, the cobordisms require higher dimensional manifolds with finitely many punctures for which there is no classification exists.

The second avenue of extension one can consider is extending our result to datasets that cannot be easily embedded in an affine space, or in general, a metric space. This is often the case when one deals with categorical data that require different techniques than numerical data [101]. Recall that abstract simplicial complexes are highly combinatorial (they require no topology or metric,) and our results strongly indicate that provided one can define a simplicial complex out of data sets whose features are purely or partially categorical, the cophenetic homological distance would yield usable information about the dataset on par with hierarchical clustering.

REFERENCES

[1] **Wessel, P.**, **Luis, J.F.**, **Uieda, L.**, **Scharroo, R.**, **Wobbe, F.**, **Smith, W.H.F. and Tian, D.** (2019). The generic mapping tools version 6, *Geochemistry, Geophysics, Geosystems*, *20*(11), 5556– 5564, https://doi.org/10.1029/2019GC008515.

[2] **Rourke, C.P. and Sanderson, B.J.** (2012). *Introduction to piecewise-linear topology*, Springer Science & Business Media, https://doi.org/10.1007/978-3-642-81735-9.

[3] **Edelsbrunner, H. and Harer, J.** (2010). *Computational topology: an introduction*, American Mathematical Society, https://doi.org/10.1090/mbk/069.

[4] **Lundell, A.T. and Weingram, S.** (2012). *The topology of CW complexes*, Springer Science & Business Media, https://doi.org/10.1007/978-1-4684-6254-8.

[5] **Ghrist, R.W.** (2014). *Elementary applied topology*, volume 1, Createspace Seattle.

[6] **Reitberger, H.** (2002). Leopold Vietoris (1891-2002), *Notices-American Mathematical Society*, *49*(10), 1232–1236.

[7] **Zomorodian, A. and Carlsson, G.** (2005). Computing persistent homology, *Discrete & Computational Geometry*, *33*(2), 249– 274, https://doi.org/10.1007/s00454-004-1146-y.

[8] **Edelsbrunner, H.**, **Letscher, D. and Zomorodian, A.** (2000). Topological persistence and simplification, *Proceedings 41st Annual Symposium on Foundations of Computer Science*, IEEE, pp.454– 463, https://doi.org/10.1109/SFCS.2000.892133.

[9] **Carlsson, G.**, **Zomorodian, A.**, **Collins, A. and Guibas, L.J.** (2005). Persistence barcodes for shapes, *International Journal of Shape Modeling*, *11*(02), 149– 187, https://doi.org/10.1142/S0218654305000761.

[10] **Ghrist, R.** (2008). Barcodes: the persistent topology of data, *American Mathematical Society, Bulletin, New Series*, *45*(1), 61– 75, http://doi.org/10.1090/S0273-0979-07-01191-3.

[11] **De Silva, V. and Ghrist, R.** (2007). Coverage in sensor networks via persistent homology, *Algebraic & Geometric Topology*, *7*(1), 339–358, https://doi.org/10.2140/agt.2007.7.339.

[12] **Carlsson, G.**, **Ishkhanov, T.**, **De Silva, V. and Zomorodian, A.** (2008). On the local behavior of spaces of natural images, *International journal of computer vision*, *76*(1), 1–12, https://doi.org/10.1007/s11263-007-0056-x.

[13] **Ferri, M. and Stanganelli, I.** (2010). Size functions for the morphological analysis of melanocytic lesions, *Journal of Biomedical Imaging*, *2010*, 5, https://doi.org/10.1155/2010/621357.

[14] **Nicolau, M.**, **Levine, A.J. and Carlsson, G.** (2011). 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*(17), 7265–7270, https://doi.org/10.1073/pnas.1102826108.

[15] **Adams, H. and Carlsson, G.** (2015). Evasion paths in mobile sensor networks, *The International Journal of Robotics Research*, *34*(1), 90–104, https://doi.org/10.1177/0278364914548051.

[16] **Emrani, S.**, **Gentimis, T. and Krim, H.** (2014). Persistent homology of delay embeddings and its application to wheeze detection, *IEEE Signal Processing Letters*, *21*(4), 459–463, http://doi.org/10.1109/LSP.2014.2305700.

[17] **Carlsson, G.** (2014). Topological pattern recognition for point cloud data, *Acta Numerica*, *23*, 289–368, https://doi.org/10.1017/S0962492914000051.

[18] **Giusti, C.**, **Pastalkova, E.**, **Curto, C. and Itskov, V.** (2015). Clique topology reveals intrinsic geometric structure in neural correlations, *Proceedings of the National Academy of Sciences*, *112*(44), 13455–13460, https://doi.org/10.1073/pnas.1506407112.

[19] **Perea, J.A.**, **Deckard, A.**, **Haase, S.B. and Harer, J.** (2015). SW1PerS: Sliding windows and 1-persistence scoring; discovering periodicity in gene expression time series data, *BMC bioinformatics*, *16*(1), 257, https://doi.org/10.1186/s12859-015-0645-6.

[20] **Perea, J.A. and Carlsson, G.** (2014). A klein-bottle-based dictionary for texture representation, *International journal of computer vision*, *107*(1), 75–97, https://doi.org/10.1007/s11263-013-0676-2.

[21] **Khasawneh, F.A.**, **Munch, E. and Perea, J.A.** (2018). Chatter classification in turning using machine learning and topological data analysis, *IFAC-PapersOnLine*, *51*(14), 195–200, https://doi.org/10.1016/j.ifacol.2018.07.222.

[22] **Amézquita, E.J.**, **Quigley, M.Y.**, **Ophelders, T.**, **Munch, E. and Chitwood, D.H.** (2020). The shape of things to come: Topological data analysis and biology, from molecules to organisms, *Developmental Dynamics*, *249*(7), 816–833, https://doi.org/10.1002/dvdy.175.

[23] **Carrière, M. and Rabadán, R.**, (2020). Topological data analysis of single-cell Hi-C contact maps, *Topological Data Analysis*, Springer, pp.147–162, https://doi.org/10.1007/978-3-030-43408-3_6.

[24] **Karan, A. and Kaygun, A.** (2021). Time series classification via topological data analysis, *Expert Systems with Applications*, *183*, 115326, https://doi.org/10.1016/j.eswa.2021.115326.

[25] **Tymochko, S.**, **Munch, E.**, **Dunion, J.**, **Corbosiero, K. and Torn, R.** (2020). Using persistent homology to quantify a diurnal cycle in hurricanes, *Pattern Recognition Letters*, *133*, 137–143, https://doi.org/10.1016/j.patrec.2020.02.022.

[26] **Yesilli, M.C.**, **Khasawneh, F.A. and Otto, A.** (2022). Topological feature vectors for chatter detection in turning processes, *The International Journal of Advanced Manufacturing Technology*, *119*(9), 5687–5713, https://doi.org/10.1007/s00170-021-08242-5.

[27] **Güzel, İ.**, **Munch, E. and Khasawneh, F.A.** (2022). Detecting bifurcations in dynamical systems with CROCKER plots, *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *32*(9), 093111, https://doi.org/10.1063/5.0102421.

[28] **Munch, E.** (2017). A user's guide to topological data analysis, *Journal of Learning Analytics*, *4*(2), 47–61, https://doi.org/10.18608/jla.2017.42.6.

[29] **Wasserman, L.** (2018). Topological data analysis, *Annual Review of Statistics and Its Application*, *5*, 501–532, https://doi.org/10.1146/annurev-statistics-031017-100045.

[30] **Chazal, F. and Michel, B.** (2021). An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists, *Frontiers in Artificial Intelligence*, *4*, 667963, https://doi.org/10.3389/frai.2021.667963.

[31] **Carlsson, G. and Vejdemo-Johansson, M.** (2021). *Topological Data Analysis with Applications*, Cambridge University Press, https://doi.org/10.1017/9781108975704.

[32] **Hensel, F.**, **Moor, M. and Rieck, B.** (2021). A survey of topological machine learning methods, *Frontiers in Artificial Intelligence*, *4*, 681108, https://doi.org/10.3389/frai.2021.681108.

[33] **Carlsson, G.**, (2020). Persistent homology and applied homotopy theory, *Handbook of Homotopy Theory*, CRC Press/Chapman and Hall Handbooks in Mathematics Series, pp.297–329, https://doi.org/10.1201/9781351251624.

[34] **Carlsson, G. and Mémoli, F.** (2010). Characterization, stability and convergence of hierarchical clustering methods, *Journal of Macine Learning Research*, *11*, 1425– 1470, http://jmlr.org/papers/v11/carlsson10a.html.

[35] **Henselman, G. and Ghrist, R.** (2016). Matroid filtrations and computational persistent homology, *arXiv preprint arXiv:1606.00199*, https://doi.org/10.48550/arXiv.1606.00199.

[36] **Borsuk, K.** (1948). On the imbedding of systems of compacta in simplicial complexes, *Fundamenta Mathematicae*, *35*(1), 217–234, http://eudml.org/doc/213158.

[37] **Hilton, P.** (1988). A brief, subjective history of homology and homotopy theory in this century, *Mathematics magazine*, *61*(5), 282–291, https://doi.org/10.2307/2689545.

[38] **Lance, G.N. and Williams, W.T.** (1967). A general theory of classificatory sorting strategies: 1. hierarchical systems, *The Computer Journal*, *9*(4), 373– 380, https://doi.org/10.1093/comjnl/9.4.373.

[39] **Sneath, P.H.**, **Sokal, R.R.** *et al.* (1973). *Numerical taxonomy. the principles and practice of numerical classification.*, W.H. Freeman and Company San Franscisco.

[40] **Jain, A.K. and Dubes, R.C.** (1988). *Algorithms for clustering data*, Prentice Hall Advanced Reference Series, Prentice Hall, Inc., Englewood Cliffs, NJ.

[41] **Legendre, P. and Legendre, L.** (2012). *Numerical ecology*, Elsevier, 3 edition.

[42] **Kleinberg, J.M.** (2002). An impossibility theorem for clustering, *Advances in neural information processing systems 32*, Neural Information Processing Systems, MIT Press, pp.446– 453.

[43] **Cohen-Steiner, D.**, **Edelsbrunner, H. and Harer, J.** (2007). Stability of persistence diagrams, *Discrete & Computational Geometry*, *37*(1), 103–120, https://doi.org/10.1007/s00454-006-1276-5.

[44] **Bubenik, P.** (2015). Statistical topological data analysis using persistence landscapes, *Journal of Machine Learning Research*, *16*(1), 77– 102, http://jmlr.org/papers/v16/bubenik15a.html.

[45] **Adams, H.**, **Emerson, T.**, **Kirby, M.**, **Neville, R.**, **Peterson, C.**, **Shipman, P.**, **... and Ziegelmeier, L.** (2017). Persistence images: A stable vector representation of persistent homology, *Journal of Machine Learning Research*, *18*(1), 218– 252, http://jmlr.org/papers/v18/16-337.html.

[46] **Moon, C.**, **Giansiracusa, N. and Lazar, N.A.** (2018). Persistence terrace for topological inference of point cloud data, *Journal of Computational and Graphical Statistics*, *27*(3), 576– 586, https://doi.org/10.1080/10618600.2017.1422432.

[47] **Chung, Y. and Lawson, A.** (2022). Persistence curves: A canonical framework for summarizing persistence diagrams, *Advances in Computational Mathematics*, *48*(1), 6, https://doi.org/10.1007/s10444-021-09893-4.

[48] **Merelli, E.**, **Rucco, M.**, **Sloot, P. and Tesei, L.** (2015). Topological characterization of complex systems: Using persistent entropy, *Entropy*, *17*(10), 6872– 6892, https://doi.org/10.3390/e17106872.

[49] **Carlsson, G. and Mémoli, F.** (2008). Persistent clustering and a theorem of J. Kleinberg, *arXiv preprint arXiv:0808.2241*, https://doi.org/10.48550/arXiv.0808.2241.

[50] **Johnson, S.C.** (1967). Hierarchical clustering schemes, *Psychometrika*, *32*(3), 241– 254, https://doi.org/10.1007/BF02289588.

[51] **Jardine, N. and Sibson, R.** (1971). *Mathematical taxonomy*, John Wiley & Sons Ltd., London-New York-Sydney, wiley Series in Probability and Mathematical Statistics.

[52] **Hartigan, J.A.** (1985). Statistical theory in clustering, *Journal of Classification*, *2*(1), 63– 76, https://doi.org/10.1007/BF01908064.

[53] **Ignacio, P.S.P.** (2020). Intrinsic hierarchical clustering behavior recovers higher dimensional shape information, *arXiv preprint arXiv:2010.03894*, https://doi.org/10.48550/arXiv.2010.03894.

[54] **Elkin, Y. and Kurlin, V.** (2020). The Mergegram of a Dendrogram and Its Stability, *45th International Symposium on Mathematical Foundations of Computer Science (MFCS 2020)*, volume170 of *Leibniz International Proceedings in Informatics (LIPIcs)*, Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Dagstuhl, Germany, pp.32:1–32:13.

[55] **Elkin, Y. and Kurlin, V.** (2021). Isometry invariant shape recognition of projectively perturbed point clouds by the mergegram extending 0D persistence, *Mathematics*, *9*(17), 2121, https://doi.org/10.3390/math9172121.

[56] **Gabrielsson, R.B.**, **Nelson, B.J.**, **Dwaraknath, A. and Skraba, P.** (2020). A Topology Layer for Machine Learning, ***S. Chiappa and R. Calandra**, editors, Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume108 of *Proceedings of Machine Learning Research*, PMLR, pp.1553–1563, https://proceedings.mlr.press/v108/gabrielsson20a.html.

[57] **Joshi Milan, Joshi Dhanajay, S.V.** (2020). Persistent Homology Techniques for Big Data and Machine Intelligence: A Survey, *Machine Intelligence and Signal Processing*, Springer Singapore, pp.97–111, https://doi.org/10.1007/978-981-15-1366-4_8.

[58] **Pun, C.S.**, **Lee, S.X. and Xia, K.** (2022). Persistent-homology-based machine learning: a survey and a comparative study, *Artificial Intelligence Review*, 1–45, https://doi.org/10.1007/s10462-022-10146-z.

[59] **Ali, D.**, **Asaad, A.**, **Jimenez, M.J.**, **Nanda, V.**, **Paluzo-Hidalgo, E. and Soriano-Trigueros, M.** (2022). A Survey of Vectorization Methods in Topological Data Analysis, *arXiv preprint arXiv:2212.09703*.

[60] **Güzel, İ. and Kaygun, A.** (2022). A new non-archimedean metric on persistent homology, *Computational Statistics*, *37*(4), 1963–1983, https://doi.org/10.1007/s00180-021-01187-z.

[61] **James, G.**, **Witten, D.**, **Hastie, T. and Tibshirani, R.** (2013). *An introduction to statistical learning*, volume112, Springer.

[62] **Sokal, R.R. and Rohlf, F.J.** (1962). The comparison of dendrograms by objective methods, *Taxon*, *11*(2), 33– 40, https://doi.org/10.2307/1217208.

[63] **Sergios, T. and Konstantinos, K.** (2009). *Pattern recognition*, Academic Press, Boston, fourth edition edition.

[64] **Mantel, N.** (1967). The detection of disease clustering and a generalized regression approach, *Cancer research*, *27*(2 Part 1), 209– 220.

[65] **Matsen, F.A.**, **Billey, S.C.**, **Kas, A. and Konvalinka, M.** (2016). Tanglegrams: a reduction tool for mathematical phylogenetics, *IEEE/ACM transactions on computational biology and bioinformatics*, *15*(1), 343–349, https://doi.org/10.1109/TCBB.2016.2613040.

[66] **Scornavacca, C.**, **Zickmann, F. and Huson, D.H.** (2011). Tanglegrams for rooted phylogenetic trees and networks, *Bioinformatics*, *27*(13), i248– i256, http://doi.org/10.1093/bioinformatics/btr210.

[67] **Fernau, H.**, **Kaufmann, M. and Poths, M.** (2010). Comparing trees via crossing minimization, *Journal of Computer and System Sciences*, *76*(7), 593– 608, https://doi.org/10.1016/j.jcss.2009.10.014.

[68] **Buchin, K.**, **Buchin, M.**, **Byrka, J.**, **Nöllenburg, M.**, **Okamoto, Y.**, **Silveira, R.I. and Wolff, A.** (2012). Drawing (complete) binary tanglegrams: hardness, approximation, fixed-parameter tractability, *Algorithmica*, *62*(1-2), 309–332, https://doi.org/10.1007/s00453-010-9456-3.

[69] **Strehl, A. and Ghosh, J.** (2002). Cluster ensembles-A knowledge reuse framework for combining multiple partitions, *Journal of Machine Learning Research*, *3*(Dec), 583–617, http://doi.org/10.1162/153244303321897735.

[70] **Hubert, L. and Arabie, P.** (1985). Comparing partitions, *Journal of Classification*, *2*(1), 193–218, https://doi.org/10.1007/BF01908075.

[71] **Rosenberg, A. and Hirschberg, J.** (2007). V-measure: A conditional entropy-based external cluster evaluation measure, *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pp.410–420, https://aclanthology.org/D07-1043.

[72] **Rousseeuw, P.J.** (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics*, *20*, 53–65, https://doi.org/10.1016/0377-0427(87)90125-7.

[73] Republic of Türkiye General Directorate of Highways, Retrieved 2019-09-20 from http://www.kgm.gov.tr/Sayfalar/KGM/SiteTr/Uzakliklar/illerArasiMesafe.aspx.

[74] **R Core Team**, (2019). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

[75] **Dey, T.K. and Wang, Y.** (2022). *Computational topology for data analysis*, Cambridge University Press.

[76] **Hatcher, A.** (2002). *Algebraic topology*, Cambridge University Press, Cambridge.

[77] **Oudot, S.Y.** (2017). *Persistence theory: from quiver representations to data analysis*, volume209, American Mathematical Society.

[78] **Gordon, G. and McNulty, J.** (2012). *Matroids: a geometric introduction*, Cambridge University Press.

[79] **Bauer, U.**, **Botnan, M.B.**, **Oppermann, S. and Steen, J.** (2020). Cotorsion torsion triples and the representation theory of filtered hierarchical clustering, *Advances in Mathematics*, *369*, 107171, 51, https://doi.org/10.1016/j.aim.2020.107171.

[80] **Carlsson, G. and Zomorodian, A.** (2009). The theory of multidimensional persistence, *Discrete & Computational Geometry. An International Journal of Mathematics and Computer Science*, *42*(1), 71–93, https://doi.org/10.1007/s00454-009-9176-0.

[81] **Kock, J.** (2004). *Frobenius algebras and 2-d topological quantum field theories*, 59, Cambridge University Press.

[82] **Dyer, M.** (1972). Rational homology and Whitehead products, *Pacific Journal of Mathematics*, *40*, 59–71.

[83] **Klaus, S. and Kreck, M.** (2004). A quick proof of the rational Hurewicz theorem and a computation of the rational homotopy groups of spheres, *Mathematical Proceedings of the Cambridge Philosophical Society*, *136*(3), 617–623, http://doi.org/10.1017/S0305004103007114.

[84] **Galili, T.** (2015). dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering, *Bioinformatics*, *31*(22), 3718–3720.

[85] **Oksanen, J.**, **Blanchet, F.G.**, **Friendly, M.**, **Kindt, R.**, **Legendre, P.**, **McGlinn, D.**, **... and Wagner, H.** (2019). vegan: Community ecology package, *R package version 2.5-6*.

[86] **R Core Team**, (2021). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria.

[87] **Developers, T.S.**, **Stein, W.**, **Joyner, D.**, **Kohel, D.**, **Cremona, J. and Eröcal, B.**, (2020), SageMath, version 9.0.

[88] **Van Rossum, G. and Drake, F.L.** (2009). *Python 3 Reference Manual*, CreateSpace, Scotts Valley, CA.

[89] **Pedregosa, F.**, **Varoquaux, G.**, **Gramfort, A.**, **Michel, V.**, **Thirion, B.**, **Grisel, O.**, **... and Duchesnay, E.** (2011). Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, *12*, 2825–2830.

[90] **Oliphant, T.E.** (2015). *Guide to NumPy*, CreateSpace Independent Publishing Platform, North Charleston, SC, USA, 2nd edition.

[91] **Dmitriy, M.**, (2018). dionysus2: Computational topology package, python package version 2.0.6.

[92] **Bray, J.R. and Curtis, J.T.** (1957). An ordination of upland forest communities of southern Wisconsin, *Ecological Monographs*, *27*, 325– 349, https://doi.org/10.2307/1942268.

[93] **Fisher, R.A.** (1936). The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, *7*(2), 179–188, https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

[94] **Patrício, M.**, **Pereira, J.**, **Crisóstomo, J.**, **Matafome, P.**, **Gomes, M.**, **Seiça, R. and Caramelo, F.**, (2018), Using Resistin, glucose, age and BMI to predict the presence of breast cancer, https://doi.org/10.1186/s12885-017-3877-1.

[95] **Ben-Hur, A.**, **Horn, D.**, **Siegelmann, H.T. and Vapnik, V.** (2002). Support vector clustering, *Journal of Machine Learning Research (JMLR)*, *2*(2), 125–137, http://doi.org/10.1162/15324430260185565.

[96] **Lumbreras, A.**, **Velcin, J.**, **Guégan, M. and Jouve, B.** (2017). Non-parametric clustering over user features and latent behavioral functions with dual-view mixture models, *Computational Statistics*, *32*(1), 145–177, https://doi.org/10.1007/s00180-016-0668-0.

[97] **Melnykov, V. and Zhu, X.** (2019). An extension of the *K*-means algorithm to clustering skewed data, *Computational Statistics*, *34*(1), 373–394, https://doi.org/10.1007/s00180-018-0821-z.

[98] **Kuhn, H.W.** (2005). The Hungarian method for the assignment problem, *Naval Research Logistics (NRL)*, *52*(1), 7–21, https://doi.org/10.1002/nav.3800020109.

[99] **Stong, R.E.** (1968). *Notes on cobordism theory*, Mathematical notes, Princeton University Press, Princeton, N.J.; University of Tokyo Press, Tokyo.

[100] **Donaldson, S.** (2011). *Riemann surfaces*, volume 22 of *Oxford Graduate Texts in Mathematics*, Oxford University Press, Oxford.

[101] **Agresti, A.** (2019). *An introduction to categorical data analysis*, Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., Hoboken, NJ, https://doi.org/10.1002/0470114754.

CURRICULUM VITAE

Name surname: İsmail Güzel

Education:

- **M.Sc.:** 2016, Dokuz Eylül University, Institute of Science, Mathematics.
- **B.Sc.:** 2015, Dokuz Eylül University, Faculty of Science, Statistics (minor).
- **B.Sc.:** 2014, Dokuz Eylül University, Faculty of Science, Mathematics (major).

Work experience:

- *Senior Researcher*, Turkish Academic Network and Information Center, TÜBİTAK-ULAKBİM, 01.2023 - Present.
- *Research Assistant*, İstanbul Technical University, 02.2018 - 01.2023.
- *Visiting Researcher*, Michigan State University, 09.2021 - 09.2022.
- *Mathematics Teacher*, İzmir Bayraklı Municipality, 10.2015 - 02.2018.

Publications and presentations on the thesis:

- **Güzel İ.**, Kaygun A. (2022). A new non-archimedean metric on persistent homology, *Computational Statistics,* 37, 1963-1983.
- **Güzel İ.**, Kaygun A. (2021). Hierarchical clustering and zeroth persistent homology. *Institute for Mathematical and Statistical Innovation - Workshop: Topological Data Analysis*, April 26-30, 2021, USA.

Other publications and presentations:

- **Güzel İ.**, Munch E., Khasawneh F. (2022). Detecting bifurcations in dynamical systems with CROCKER plots, *Chaos: An Interdisciplinary Journal of Nonlinear Science,* 32, 093111 (2022). (This paper was selected as Featured.)
- **Güzel İ.**, Munch E., Khasawneh F. (2022). A Case Study on Identifying Bifurcation and Chaos with CROCKER Plots. *SIAM International Conference on Data Mining - Applications of Topological Data Analysis to Data Science, Artificial Intelligence, and Machine Learning*, April 28, 2022, Alexandria, Virgina, USA.
- **Güzel İ.**, Kaygun A. (2022). Classification of Stochastic Processes with Topological Data Analysis. *BAŞARIM 2022 - 7th High-Performance Computing Conference*, 11-13 May, 2022, İstanbul, Türkiye.