UNIVERSIDADE DE SÃO PAULO

FFCLRP - DEPARTAMENTO DE COMPUTAÇÃO E MATEMÁTICA

PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

Topologia computacional para análise de série temporal

Computational topology for time series analysis

Vanderlei Luiz Daneluz Miranda

Dissertação apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da USP, como parte das exigências para a obtenção do título de Mestre em Ciências, Área: Computação Aplicada.

Ribeirão Preto - SP

2019

Topologia computacional para análise de série temporal

Computational topology for time series analysis

Vanderlei Luiz Daneluz Miranda

ORIENTADOR: PROF. DR. LIANG ZHAO

Versão Revisada

Ribeirão Preto - SP
2019

Miranda, Vanderlei Luiz Daneluz
Topologia computacional para análise de série temporal. Ribeirão Preto, 2019.
91 p. : il. ; 30 cm

Dissertação de Mestrado, apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto/USP. Área de concentração: Computação Aplicada.
Orientador: Zhao, Liang.

1. Análise de séries temporais. 2. Topologia computacional. 3. Descoberta de conhecimento. 4. Detecção de mudança de padrão.

Vanderlei Luiz Daneluz Miranda

Topologia computacional para análise de série temporal

Dissertação apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da USP, como parte das exigências para a obtenção do título de Mestre em Ciências, Área: Computação Aplicada.

Aprovado em:

Banca Examinadora

Prof. Dr. Liang Zhao
Orientador

Convidado 1

Convidado 2

Convidado 3

Ribeirão Preto - SP
2019

Aos Bons Espírito que sustentam nossa
caminhada na Terra.

ACKNOWLEDGEMENTS

I thank God for this incarnation, full of obstacles, helping me to grow, to Prof. Dr. Liang Zhao, my mentor for the opportunity, friendship and guidance, to my family for teaching me to mature my heart and to the Good Spirits by unceasing moral support.

"When we think a given thought, then the meaning of this thought is expressed in the shape of the corresponding neurophysiological process."

Riemann; 1876

RESUMO

Mudanças de padrão são variações nos dados da série temporal. Tais mudanças podem representar transições que ocorrem entre estados. A análise de dados topológicos (TDA) permite uma caracterização de dados de séries temporais obtidos a partir de sistemas dinâmicos complexos. Neste trabalho, apresentamos uma técnica de detecção de mudança de padrão baseada em TDA. Especificamente, a partir de uma determinada série temporal, dividimos o sinal em janelas deslizantes sem sobreposição e para cada janela calculamos a homologia persistente, ou seja, o *barcode* associado. A partir desse *barcode*, o intervalo médio e a entropia persistente são calculados e plotados em relação à duração do sinal. Resultados experimentais em conjuntos de dados reais e artificiais mostram bons resultados do método proposto: 1) Detecta mudança de padrões identificando a mudança no intervalo médio e calculando a entropia persistente para os *barcodes* gerados pelo conjunto de dados de entrada. 2) Mostra qualitativamente quão sensível é a escolha do método de filtragem para evidenciar características topológicas do espaço original sob exame. Isto é conseguido usando duas filtragens: uma filtragem métrica e uma do tipo *lower-star*. 3) Variando o tamanho da janela, o método pode caracterizar a presença de estruturas locais do conjunto de dados, como o período de convulsão nos sinais EEG. 4) O método proposto é capaz de caracterizar a complexidade pela medida de entropia persistente dos *barcodes*, uma medida de entropia baseada na definição de entropia de Shannon. Além disso, neste trabalho, mostramos a evidência de mudanças de complexidade associadas a um período de convulsão de um sinal de EEG.

Palavras-chave: Mudança de padrão, análise de série temporal, análise topológica de dados, homologia persistente, entropia persistente, complexidade, redes complexas.

ABSTRACT

Pattern changings are variations in time series data. Such changes may represent transitions that occur between states. Topological data analysis (TDA) allows characterization of time-series data obtained from complex dynamical systems. In this work, we present a pattern changing detection technique based on TDA. Specifically, starting from a given time series, we divide the signal in slicing windows with no overlapping and for each window we calculate the persistent homology, i.e., the associated barcode. From the barcode the average interval size and persistent entropy are calculated and plotted against the signal duration. Experimental results on artificial and real data sets show good results of the proposed method: 1) It detects pattern changing by identifying the change in the average interval size and calculated persistent entropy for the barcodes generated by the input data set. 2) It shows qualitatively how sensible the choice of filtration method is to evidence topological features of the original space under examination. This is accomplished by using two filtrations: a metric and a lower-star filtration. 3) By varying the slice window size, the method can characterize the presence of local structures of the data set such as the seizure period in EEG signals. 4) The proposed method can characterize complexity by the measure persistent entropy for barcodes, an entropy measure based on Shannon´s entropy definition. Moreover, in this work, we show the evidence of complexity changes associated with a seizure period of an EEG signal.

Keywords: Pattern changing detection, time series analysis, topological data analysis, persistent homology, persistent entropy, complexity, complex networks.

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

Notation	Description
$\mathcal{K}$	Simplicial complex.
$\boldsymbol{\tau}$	Collection of subsets defining a topology.
$\boldsymbol{\epsilon}$	Radius of an n-dimensional sphere around each point in the point defining an **ϵ-ball**.
$(\mathcal{K}, \tau)$	Topological space defined on set $\mathcal{K}$.
R_ϵ	Vietoris-Rips (or Rips) Complex at scale ϵ.
$\beta_i(\mathcal{K})$	Betti number *i* of the simplicial complex $\mathcal{K}$. It is also defined as the dimension of the n^{th} Homology Group $\mathcal{H}_n$, e.g., dim($\mathcal{H}_n$).
PE	Persistent Entropy.
$\boldsymbol{\mathcal{B}}(\boldsymbol{\mathcal{F}})$	Persistence barcode of filtration $\boldsymbol{\mathcal{F}}$ induced by a persistent homology $\mathcal{H}$.
$\mathcal{H}_n$	The n^{th} Homology Group $\mathcal{H}_n$.
$v \prec \sigma$	v precedes σ.
$\mathcal{H}$	Persistent homology also represented by a persistence barcode graph.
t_{max}	Maximum distance considered for rips-complex.

CONTENTS

Chapter 1

INTRODUCTION

1.1 Context

Computational topology or computational algebraic topology is a subfield of topology that includes computer science, computational geometry, and computational complexity theory. It aims to develop efficient algorithms for solving problems that relate to the understanding of the shape of real abstract spaces that can be found in computer graphics, computer-aided design (CAD), structural biology, chemistry, robotics (Zomorodian 2005). One of these algorithms, i.e., computational homology, refers to the computation of homology groups of simplicial complexes, that are used to perform analysis of the topological features of point cloud data.

As a branch of Mathematics, topology is not recent (Dieudonne 1989). It was first born under the name of **Analysis Situs,** mainly in the writings of Leibniz (Debuiche 2013). After Leibniz, Euler contributed by solving the 7 bridges problem and defining what is now termed the "Euler characteristic" χ , the main topological invariant that Leibniz was unable to find. Finally, Poincarè introduced with his **analysis situs,** most of the basic theorems and concepts in the discipline (Siersma 2012).

As a research field, however, algebraic topology started very recently mainly due to some pioneering algorithms for fast computation (Edelsbrunner et al. 2002). This was motivated by the large production of heterogeneous data, available through different means such as the Internet, or the more recent Internet of Things devices. The challenge of efficiently interpreting these data has been a key problem in science and industry. In the effort of analyzing the data, various powerful statistical methods have been used to sort through the

data and determine significant components. In this context, topological data analysis (TDA), by using algebraic topology theory, attempts to create reliable methods based on topological features of spaces in order to obtain useful information from data sets. Classical expositions can be found in Hatcher (2002) and Ghrist (2014).

Intuitively, topological features can be seen as qualitative geometric properties relating the notions of proximity and continuity. In this respect, TDA provides a powerful approach to infer robust qualitative, and sometimes quantitative, information about the structure of data that are often represented as point clouds in a metric space (Chazal and Michel 2017).

Although TDA is still under development, it provides a set of efficient tools to help analyze and interpret data that are represented as point clouds, such as persistent homology, the Mapper algorithm (Singh et al. 2007), Euler calculus (Ghrist 2014), and many more. In this work, we focus on persistent homology.

Persistent Homology, in simple terms, measures shapes of spaces and the features of functions. This may give us useful information in point clouds where the shape may be interpreted as the geometry of some underlying implicit object near which the point cloud is sampled. The simplest non-trivial example of this idea is a point cloud which has the shape of a circle (Figure 3 and Figure 4), and this shape is characterized by 1-dimensional persistence. The challenge in applying the method is that noise can reduce the persistence, and not enough points can prevent the circular shape from appearing. It is also a challenge to deal with the fact that features come on all scale-levels and can be nested or in more complicated relationships. But this is just what persistent homology deals with.

There are now many methods inspired by topological approaches, but most of them rely on the following basic pipeline (Chazal and Michel 2017) that was also used during the experiments presented in this work (see also Figure 15, Figure 16 and Figure 17):

1. We take as the input a finite set of points embedded in a metric. The metric is usually given as an input or guided by the application. It is, however, important to notice that the choice of the metric may be critical to revealing interesting topological and geometric features of the data.
2. On top of the data, we build a nested family of topological spaces, i.e., a filtration. The filtration must highlight the underlying topology or geometry of the data. In order words, the shape of this filtration reflects the shape of the data, in an incremental way. As we will verify later, this also presents a challenge: how to define such structures that are proven to reflect relevant

information about the structure of data and that can be effectively constructed and manipulated in practice (computationally viable).

3. Based on these structures, i.e., the filtration, we extract relevant topological information by using specific methods such a persistent homology. This provides us with a tool that allows the identification of interesting topological/geometric information and its visualization and interpretation. But there another important issue: it is equally important to show its relevance, in particular, its stability with respect to perturbations or presence of noise in the input data. So the statistical behavior of the inferred features should be analyzed
4. Finally, the extracted topological information can provide new families of features (descriptors) of the data. They can help to understand the data either through visualization or by combining them with other kinds of features for further analysis and machine learning tasks. At this point, the challenge is to show the benefit with respect to other features of the information provided by TDA.

Another enormous field of research is time series analysis. A time series is used for many applications such as economic forecasting, sales forecasting, stock market analysis, signal processing, electroencephalography, control engineering earthquake prediction, weather forecasting, pattern recognition and many more. Despite the variety of motivations in the use of time series, the primary goal of time series analysis is forecasting, signal detection and estimation, clustering, classification, and anomaly detection.

Briefly, a time series is a sequential set of data points indexed in time order $\{x_1 , x_2 , . . x_T \}$ or $\{ x_t \}$, t = 1, 2, . . . T, where the variable x_t is treated as a random variable (Cochrane 1977). Most commonly, it is a sequence of discrete-time data equally spaced in time intervals such as hourly, daily, weekly, monthly or yearly time separations. Therefore, one important characteristic of time series is that it is a list of observations where the ordering matters, i.e., changing the order could change the meaning of the data.

The overall aim of time series analysis is to understand the past as well as predict the future. Thus, this translates to determine a model that describes the pattern of the time series and helps:

- Describe the important features of the time series pattern.
- Explain how the past affects the future or how two time series can “interact”.

- Forecast future values of the series.

We can identify four typical characteristics in time series according to the observed data (Adhikari and Agrawal 2013):

Trends: the general tendency of some time series to increase, decrease or stagnate over a long period of time. So it is the long term effect over the mean. For example, series relating to population growth, the number of deaths in a city etc.

Seasonal variations: effects due to periodic fluctuations (monthly, yearly, etc). Example: climate and weather conditions, customs, traditional habits, etc.

Cycles or cyclical variation: medium-term changes in the series that repeat in cycles but has no automatic association with any temporal measures. For example, economic cycles.

Random variations: caused by unpredictable influences, which are not regular and do not repeat in a particular pattern. For example, variations caused by incidents such as earthquakes, floods, strikes, etc. One may notice that there is no defined statistical technique for measuring random fluctuations in a time series.

A time series is *non-deterministic* phenomenon in the sense that we cannot predict with certainty what will occur in the future. However, it is generally assumed that the time series follow certain probability model (Cochrane 1977), i.e., the sequence of observations of the series is a sample realization of the stochastic process that produced it. Moreover, it is assumed that the stochastic process is *stationary*, i.e., its statistical properties do not depend on time.

This way the mean, variance and autocorrelation are all constant over time for a stationary time series. Hence, a non-stationary series is one whose statistical properties change over time. This is important because most statistical forecasting methods are based on the assumption that the time series is approximately stationary. In this case, they are relatively easy to predict by assuming that its statistical properties will be the same in the future as they have been in the past.

There are two types of stationary process (Adhikari and Agrawal 2013). A process is *Strongly Stationary* if the joint distribution of any possible set of random variables from the process is independent of time. On the other hand, a stochastic process is *Weakly Stationary of order* k if the statistical moments of the process up to that order depend only on time differences and not on the time of occurrences of the data being used to estimate the moments.

In order to design a model useful for future forecasting, strongly stationarity is expected. However, in the real world this may not be the case. Time series with a trend or seasonal patterns are non-stationary. E.g., if the series is consistently increasing over time, the

sample mean and variance will grow with the size of the sample, and they will always underestimate the mean and variance in future periods. On the other hand, for short time span, one can reasonably model the series using a stationary stochastic process. If this is not possible, for the analysis purpose, non-stationary data should be first converted into stationary data (for example by trend removal), so that further statistical analysis can be done.

In real-world scenarios, time series may appear associated with different objects than vectors of feature-values. They may represent some complex graph structure. But the traditional statistical methods of time series analysis are focused on sequences of values representing a single numerical variable. So the representation of time series plays a key role in successful discovery of time-related patterns. E.g., the most frequently used representation of single-variable time series is the piecewise linear approximation, where the original points are reduced to a set of straight lines ("segments"). This representation has been used to support clustering, classification, indexing and association rule mining of time series (Keogh et al. 2004).

Considering the detection of pattern changing in time series, a key element is the identification of similarity in time series, i.e., we look for sets of similar data sequences that differ only slightly from each other. An intuitive notion of similarity between time series and efficient approximate algorithms that compute these similarity measures have already been provided (Vlachos et al. 2004). Another possible approach rather than segmenting a time series is to see each time series as a single object. In this scenario, classification and clustering of such complex "objects" may be particularly beneficial for e areas of process control, intrusion detection, and character recognition (Last et al. 2004).

In this thesis, we intend to explore pattern changing detection in time series in a different approach than the traditional methods mentioned above. Based on results mentioned in the literature for measuring similarities between piecewise linear functions (Rucco et al. 2017), the use of maximum persistence at the point-cloud level to quantify periodicity at the signal level (Perea and Harer 2015), and the newly-introduced quantitative concept of **persistent entropy**, which was used to derive a structural model for a complex system (Merelli et al. 2015), we propose a method that uses time series segmentation and its piecewise linear representation to compute PH features and in this way to help discovering structure in time series.

Finally, besides time series analysis, another recent field of interest is Complexity. Although a complex system has many different definitions (Edmonds 1999), it is presented invariably as a dynamical system composed of a huge number of components linked both

functionally and spatially (Piangerelli et al. 2018). These systems are also generally characterized by emerging features and behaviors that arise from the interaction of their parts and cannot be predicted from the properties of the parts. Again due to the increase of data volume associated to complex systems, actually, two alternatives approaches have been used to study complexity: multivariate time series (Nazarimehr et al. 2017) and complex networks (Albert and Barabási 2002).

In the study of complex systems, it is shown that combinatorial features and statistics of connections of networks affect their dynamical, statistical and critical behavior (Erdős and Rényi 1959) and (Watts and Strogatz 1998). Moreover, we notice that researchers in this field are also familiar with the idea of topology as a complex network where a graph is a 1-chain complex (Hatcher 2002) or (Ghrist 2014). However, topology provides a high dimensional generalization to this approach: whereas a network is an assembly of pairs of elements (of neurons for example), algebraic topology (i.e., homology) investigates assemblies with arbitrary numerous elements.

Despite the enormous amount of definitions and, consequently, the ways of measuring another possible attempt to characterize complexity involves the notion of entropy. Although this was not intended by its originators (Shannon and Wiever 1964), entropy-based measures have often been used as measures of complexity including, e.g., the regularity in noisy time series (Pincus 1995) and artificial life (Ray 1994).

In this thesis, we intend to explore topological data analysis (TDA) as a methodology for characterizing complexity by means of a new definition of entropy although based on Shannon´s approach, the already mentioned persistent entropy, essentially a measure derived from persistent barcodes (Rucco et al. 2017).

1.2 Motivation and Objectives

One of the most commonly used representations of time series is the piecewise linear approximation. This representation has been used to support clustering, classification, indexing and association rule mining of time series data (Keogh et al. 2004). This fact naturally motivated us to search for a simplicial complex filtration that fitted this type of topological space and investigate its potential use for pattern changing detection in times series, although we did not aim to evaluate performance at this time.

Also in the problem of characterizing similarity between time-series, the simplest approach is to define the distance between two sequences, i.e., to map each sequence into a vector and then use a p-norm to calculate their distance. So, most of the related work on time-series has focused on the use of some metric $L_p\ Norm$. But this simple approach cannot deal well with outliers and is very sensitive to small distortions in the time axis (Vlachos et al. 2004). In algebraic topology, there are simplicial complex filtrations that are inherently metric filtrations, such as Vietoris-Rips and that are computationally feasible. This motivated us to map the time series to a metric space and to investigate the use of a metric filtration for pattern changing detection in times series.

So these both facts motivated the need to understand the theoretical foundations of TDA, its main tools such as persistent homology and how these tools can be applied in time-series pattern changing detection. As a consequence, the choice of the filtered complex, or even the creation of a new filtered complex if necessary, may be a challenge as already mentioned by some researches as Otter et al. (2017), and one needs to be aware of the obstacles and constraints involved in this choice. This is a practical consideration that needed to be investigated more deeply.

Besides this, the choice of the right tool or implemented algorithm is by itself a challenge for a novice in this field (Otter et al. 2017). These algorithms are available through standard libraries such as Gudhi library (C++ and Python) (Maria et al. 2014) and its R software interface and many more. Each of one has its pros and cons. Otter et al. (2007) give real useful benchmarks of state-of-the-art implementations for the computation of persistence as well indications of which algorithms and implementations are best suited to different types of datasets.

Another standard library is the JavaPlex library which implements homology and related techniques from computational topology (Adams and Tausz 2018). The ease of access from Matlab as well as the functionality native to Matlab itself motivated us to use it as a prototype platform for the experiments described in this work.

In the application domain, two proposals of Robert Ghrist motivated us to better understand the algebraic topology topics. One of them is the use of topological signal processing in radar signal processing based on Euler Calculus (Curry et al. 2012), where data is too coarse or noisy to retain geometry. Another is in the sensor network technology (Silva and Ghrist 2007) and (Bhattacharya et al. 2014): the replacement of expensive sensor with swarms of small, cheap, local sensors. This may have a huge impact on the Internet of Things devices due to the amount of sensors they use and their large production of data.

Finally, the overall objective of this work is two-fold. First, it intends to provide a brief and comprehensive introduction to the mathematical foundations of TDA in order to start to use its methods and software for pattern changing detection in time series. Second, this work also aims to empirically verify the application of these techniques to characterize complexity in complex systems.

Specifically, this study aims to address the following research questions:

1. What are the basic concepts underlying TDA and persistent homology (PH)?
2. Are the proposed PH metrics, i.e., persistent entropy and average interval size, sufficient to identify pattern changing in real-world time series?
3. Can we characterize complexity in time series using these PH metrics?

1.3 Organization of this work

For the purposes outlined, the focus is put on a few selected, but fundamental topics: in Chapter 2, named “Overview of Topological Data Analysis”, presents a review on the main topics concerning algebraic topology, mainly on simplicial complexes, which serve as the background for the methods used in our study concerning pattern changing in time series.

In Chapter 3, named “Methods”, the datasets chosen for the research and the methodologies based on persistent homology that used to explore pattern changing in time series are presented. Some important considerations related to the selection of the right filtration are also remarked.

Chapter 4, named “Results”, shows some results obtained through the experiments made with both synthetic data and real-world data. The corresponding analysis of these experiments is also presented.

Finally, conclusions, and future works are described in Chapter 5, named “Conclusions”.

Chapter 2

OVERVIEW OF TOPOLOGICAL DATA ANALYSIS

This section provides an overview of topological data analysis and the way of representing persistent homology using persistence barcode. Two kinds of filtration (the ordering of the simplices in the simplicial complex) are introduced and some of the different ways to use and visualize the results of topological data analysis by considering some simple examples are presented. Finally, the concept of persistent entropy, based on the Shannon entropy, calculated for the persistence barcodes is presented. Thus, we intend to review the concepts that support the use of Topological Data Analysis according to the proposed technique.

2.1 Topology and Topological Data Analysis

Data analysis in general aims to address two fundamental tasks: inferring higher dimensional structure from lower dimensional representations and assembling discrete points into a global structure. Topological data analysis (TDA) has shown to be a powerful tool for analyzing complex data sets due to its robustness against noise in data and its ability to be a coordinate-free method of analysis.

The foundation of TDA is the construction of a simplicial complex, i.e., a kind of higher dimension triangulation. In some sense, it allows us to describe the underlying manifold of which the data are a sample (see Figure 3). Loosely speaking, the data points are the vertices, edges joining these vertices are called 1-simplices, two-simplices cover the faces, and so on (Figure 1). Abstractly, a *k*-simplex is an ordered list $\sigma = \{x_1, x_2, .., x_{k+1}\}$ of k + 1 vertices.

The mathematical challenge is to connect the data points in geometrically meaningful ways. However, a discrete set of points is only an approximate representation of a continuous shape. Therefore, this description is accurate only up to some spatial scale $\boldsymbol{\epsilon}$ (see Figure 4 and Figure 5). The choice of this scale is always part of the solution. This is both a problem and an advantage: one can gather useful information on topology from examining how the shape changes with $\boldsymbol{\varepsilon}$ (Carlsson 2009), but for large $\boldsymbol{\varepsilon}$ we can fool ourselves about the original manifold. This problem will be addressed by using **persistent homology** as we will show in the next section.

More formally, we follow (Atienza et al. 2018) and (Hatcher 2002) for further definitions.

Definition 1. A **topological space** is an ordered pair $(\mathcal{K}, \tau)$ where $\mathcal{K}$ is a set and τ is a collection of subsets of $\mathcal{K}$, satisfying the following axioms:

- $\emptyset \in \tau$ and $\mathcal{K} \in \tau$.
- Any (finite or infinite) union of members of $\boldsymbol{\tau}$ still belongs to $\boldsymbol{\tau}$
- The intersection of any finite number of members of τ still belongs to τ.

We say that $\boldsymbol{\tau}$ is a **topology on $\mathcal{K}$**. For example $\mathcal{K} = \{a, b, c\}$ and the topology $\boldsymbol{\tau} = \{\emptyset, \{a\}, \{b\}, \{a, b, c\}\}$.

An important topological space is a **simplicial complex**. Intuitively, we decompose a topological space into simple pieces that maintain the definition above, i.e., the common intersections of these pieces is lower-dimensional pieces of the same type.

2.2 Simplicial and Abstract Simplicial Complexes

TDA tools are based on **simplicial complexes**, which are complexes of a geometric structure called **simplex**. TDA uses simplicial complexes because they can approximate more complicated shapes and are much more mathematically and computationally tractable than the original shapes that they approximate.

Simply stated, a simplex is a generalization of a triangle to higher dimensions. For example (Figure 1), a vertex is a 0-simplex, an edge is a 1-simplex, a 2-simplex is an ordinary 3-sided triangle in two-dimensions (or could be embedded in higher-dimensional spaces). And a 3-simplex is a tetrahedron (with triangles as faces) in 3-dimensions, and a 4-simplex is

beyond our visualization, but it has tetrahedrons as faces and so on. Notice that the **faces** of a simplex are its boundaries. For a 1-simplex (line segment) the faces are points (0-simplices), for a 2-simplex (triangle) the faces are line segments, and for a 3-simplex (tetrahedron) the faces are triangles (2-simplices) and so on.

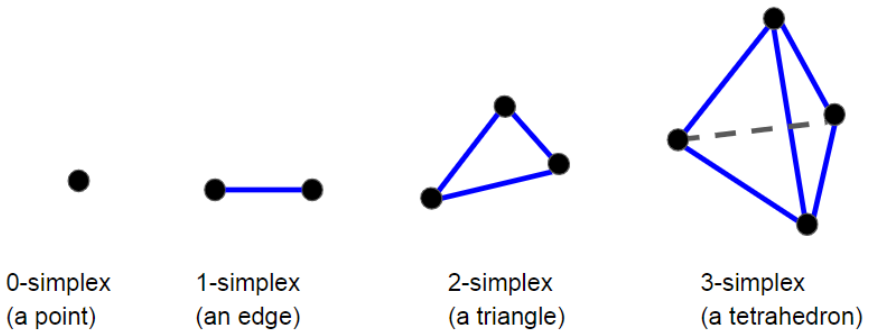


Figure 1. Visualization of simplices of several dimensions

A **simplicial complex** is formed connecting different simplices. For example, one can connect a 2-simplex (triangle) to another 2-simplex via a 1-simplex (line segment). Figure 2 shows a simplicial complex that consists of two triangles, i.e., 2 *2-simplex*, connected along one side, which are connected via a 1-simplex (line segment) to a third triangle. It is called a 2-complex because the highest-dimensional simplex in the complex is a 2-simplex (triangle).

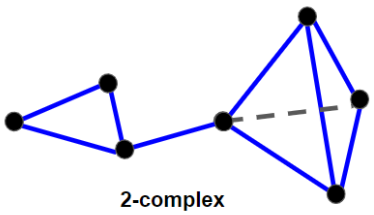


Figure 2. A simplicial complex constructed by 4 simplices.

The formal definition follows bellow based on Hatcher (2009). But before notice that an ***abstract* simplex** is any finite set of vertices. For example, the simplex $\mathcal{K}_1$= {a, b} and $\mathcal{K}_2$= {a, b, c} represent a 1-simplex (line segment) and a 2-simplex (triangle), respectively. So an abstract simplex and abstract simplicial complexes are abstract because they were not given any specific geometric realization.

Definition 2. The join of n points is a convex polyhedron of dimension n-1 is called a **simplex** (Hatcher 2002).

Definition 3. A **simplicial complex** can be described combinatorically as a set $\mathcal{K}_0$ of vertices together with sets $\mathcal{K}_n$ of n-simplices, which are (n+1)-element subsets of $\mathcal{K}_0$Erro! Indicador não definido. (Hatcher 2002).

Definition 4. Given a simplicial complex $\mathcal{K}$, a **nested sequence of simplicial complexes**

$$\emptyset = \mathcal{K}_0 \subset \mathcal{K}_1 \ldots \subset \mathcal{K}_{\text{tmax}} = \mathcal{K},$$

is called a ***filtration*** of $\mathcal{K}$. An ordering of the simplices of a simplicial complex $\mathcal{K} = \{\sigma_1, \ldots, \sigma_m\}$ is called a *filter* if it satisfies the property that s < t whenever $\sigma_s \prec \sigma_s$. Then we can create a filtration by setting: $\mathcal{K}_t = \{\sigma_1, \ldots, \sigma_t\}$, for $1 \leq t \leq t_{max}$. The filtration time (or filter value) of a simplex $\sigma \in K$ is the smallest t such that $\sigma \in \mathcal{K}(t)$.

Many applications of Computational Topology start with a cloud of points embedded in $\mathbb{R}^n$. More specifically we are interested in some properties of topological spaces that do not vary with certain types of continuous deformations) such as holes, loops, and connected components. In topology, they are called **topological invariants**. In Figure 3, for example, the two point clouds are topological identical since both have a single loop, i.e., a single hole, and only exhibit one component.

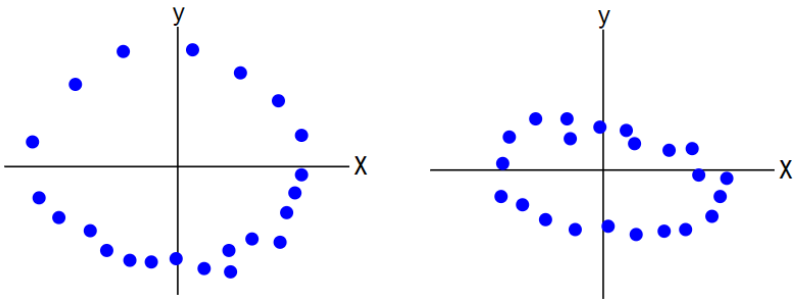
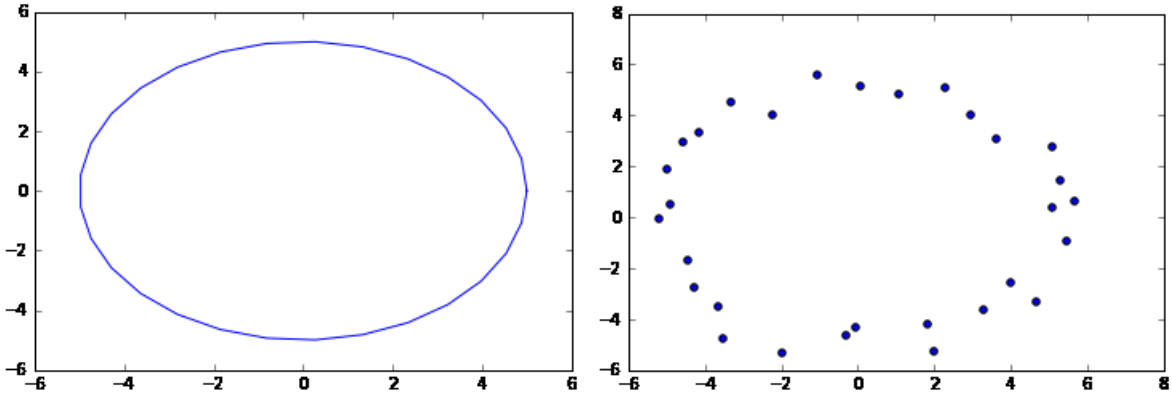


Figure 3. Two point clouds with identical topological features.

These topological characteristics may be studied by using isomorphic topological spaces of the cloud of points that are more amenable to computation. In words, we build

simplicial complexes from these point cloud and calculate their topological invariants. In fact, there are many different types of simplicial complex constructions that have differing properties. The most common simplicial complexes are **Alpha-complexes** (A_ϵ)**, Čech-complexes** (C_ϵ), and **Vietoris-Rips-complexes (or Rips-complexes,** R_ϵ).

In this work we will focus on **Vietoris-Rips** complex since it is reasonably practical from a computational standpoint.



(a) **(b)**

Figure 4. A continuous topological space (a) and an approximate representation, i.e., a data cloud with some jitter

2.3 Metric Filtrations: the Vietoris-Rips Complex

Intuitively, a **Vietoris-Rips** (VR) complex (or Rips-complex) is constructed from a point cloud by initially connecting points in the point cloud dataset with edges that are less than some arbitrarily defined distance ϵ from each other. This will construct a 1-complex, which is essentially just a graph (a set of vertices and a set of edges between those vertices as in Figure 6). Next, we need to fill in the higher-dimensional simplices, e.g. any triangles, tetrahedrons, etc. so we won't have a bunch of empty holes.

Definition 5. Given a finite set of points S = $\{x_\alpha\}$ in Euclidean space $\mathbb{E}^n$, i.e., endowed with a distance d_s, the **Rips complex** (R_ϵ)**,** at scale **ϵ** is defined as:

$$\mathrm{R}_\epsilon(\mathrm{S}) = \{\sigma \subseteq \mathrm{S} \mid d_s(x_i, x_j) \leq \epsilon \,,\, \forall x_i \neq x_j \in \sigma\}.$$

So the Rips complex at scale ϵ is the set $R_\epsilon(S)$ of all subsets σ of S such that the pair-wise distance between any non-identical points in σ is less than or equal to ϵ.

Intuitively we take the **ϵ-balls** around each point in the point cloud and build edges between that point and all other points within its ball (Figure 5). For example, the ball of a point in $\mathbb{R}^2$ is a circle, a ball around a point in $\mathbb{R}^3$ is a sphere, and so on. It is important to realize that a particular VR construction depends not only on the point cloud data but also on the parameter ϵ that is arbitrarily chosen. This raises an important question: how do we choose ϵ? Simply stating, we use various levels for ϵ and see what seems to result in a meaningful VR complex. If ϵ is too small, then the complex may just consist of the original point cloud or only a few edges between points (Figure 5-a). If ϵ is too big, then the point cloud will just become one massive ultra-dimensional simplex (Figure 5-d).

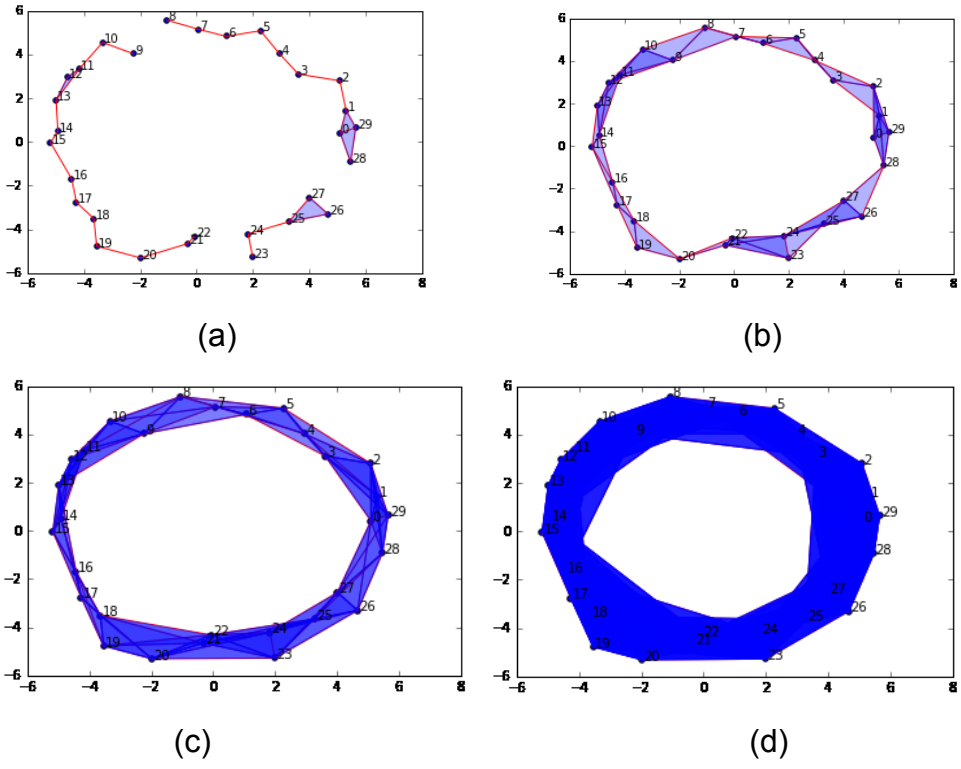


Figure 5. Vietoris-Rips Simplicial Complex. (a) $\epsilon = 1.9$. (b) $\epsilon = 3.0$. (c) $\epsilon = 4.0$. (d) $\epsilon = 7.0$.

When using rips-complex filtration, the key to actually discovering meaningful patterns is to continuously vary the ϵ parameter (and continually re-build complexes) from 0 to a maximum that results in a single massive simplex. Then a diagram that shows what topological features are born and dead as ϵ continuously increases is generated. In the analysis, we assume that features that persist for long intervals over ϵ are meaningful features

whereas features that are very short-lived are likely noise. This procedure is called **persistent homology** as it finds the homological features of a topological space (specifically a simplicial complex) that persist while ϵ varies over some specified range of interest.

Besides the selection of parameter **ϵ**, or how much it must vary during the study of a topological space, another important issue to consider is the fact that the topological space under examination must be isomorphic to the simplicial complex topology selected.

.

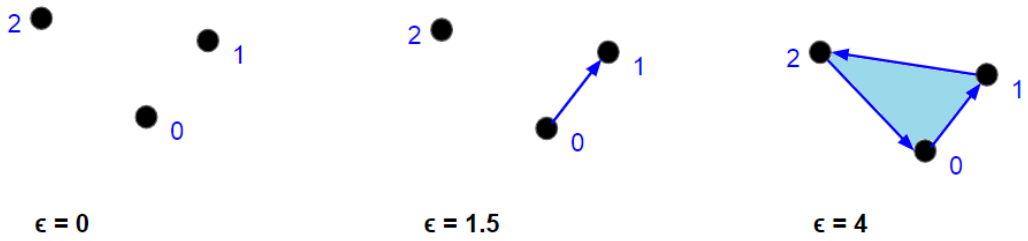


Figure 6. Rips filtration example: the sequence of simplicial complexes generated by continuously increasing the scale parameter ϵ. For ϵ = 4, the simplicial complex is $\mathcal{K}$ = {{0}, {1}, {2}, {0,1}, {2,0}, {1,2}, {0,1,2}}

2.4 Lower-star Filtrations: the Piecewise Complex

A piecewise linear function (PL) is a real-valued function defined on the real numbers (see Figure 7). In order to apply topological methods to these functions, they must be equipped with a topology. Rucco et al. (2017) reported a new methodology that allows us to associate a filtered simplicial complex to a PL.

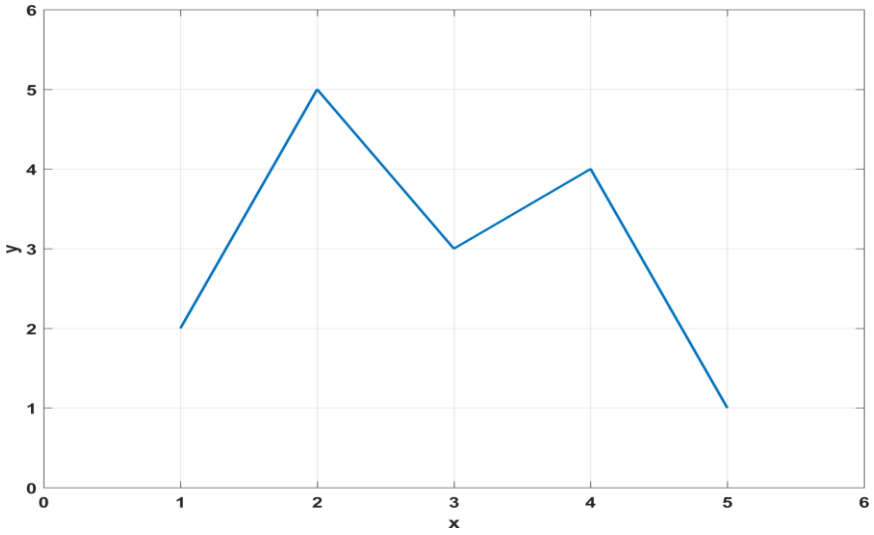


Figure 7. Graphical representation of a piecewise linear function. The graph is composed of straight-line sections

According to Edelsbrunner et al. (2002) a filtration can be created given some function on the vertices. Examples of these are the grayscale value of images or height information in geographical data.

So given a continuous function $\bar{f}$: $\mathbb{R}^n \longrightarrow \mathbb{R}$, we calculate the values of $\bar{f}$ on a finite set of points $\mathcal{K}_0 \subseteq \mathbb{R}^n$ (e.g., the vertices). Let $\mathcal{K}$ be a simplicial complex with distinct real values specified at their vertices. For example, if $\mathcal{K}_0 \subseteq \mathbb{R}$, then $\mathcal{K}$ is a line subdivided into segments with endpoints in the Domain. Then extend $\mathcal{K}$ to a continuous function by linear interpolation on the interiors of the simplices, e.g., define a piecewise linear (PL) function

$$f: \mathcal{K}_0 \longrightarrow \mathbb{R} \text{ such that } f(u) = \bar{f}(u) \text{ for } u \in \mathcal{K}_0.$$

We can then order the vertices by increasing function value as $f(u_1) < f(u_2) < \cdots < f(u_n)$, where $n = |\mathcal{K}_0|$. Each simplex σ has a unique maximum vertex v_{max}, i.e.,

$$f(v_{max}) = \max\{f(v) : v \in \mathcal{K}_0 \text{ and } v \prec \sigma\}.$$

The nested sequence of complexes $\emptyset = \mathcal{K}_0 \subset \mathcal{K}_1 \ldots \subset \mathcal{K}_p = \mathcal{K}$ is the lower-star filtration of f (Edelsbrunner and Harer 2010), such that

$$\mathcal{K}_t \setminus \mathcal{K}_{t+1} = \{\sigma \in K : \text{maximum vertex of } \sigma \text{ is } v_t\}.$$

In particular, all vertices enter at $\mathcal{K}_0$, and $\mathcal{K}_t$ and $\mathcal{K}_{t+1}$ differ by at least one simplex since each simplex has a unique maximum vertex.

The transformation of a PL function *f* into a filtered simplicial complex is shown in the Figure 8. The input consists of the five time points (Figure 8-a), with coordinates: (5, 1), (1, 2), (3, 3), (4, 4) and (2, 5). The filtered simplicial complex is formed by five 0−simplices and four 1−simplices (Figure 8-b):

- The 0-simplices: $\{v_0, v_1, v_2, v_3, v_4\}$ with filter values f(v_0) = 1, f(v_1) = 2, f(v_2) = 3, f(v_3) = 4, f(v_4) = 5.
- The 1−simplices: $\{e_1, e_2, e_3, e_4\}$, with filter values f(e_1) = f(e_2) = 4, and f(e_3) = f(e_4) = 5. The filter-value set is F = {1, 2, 3, 4, 5}.

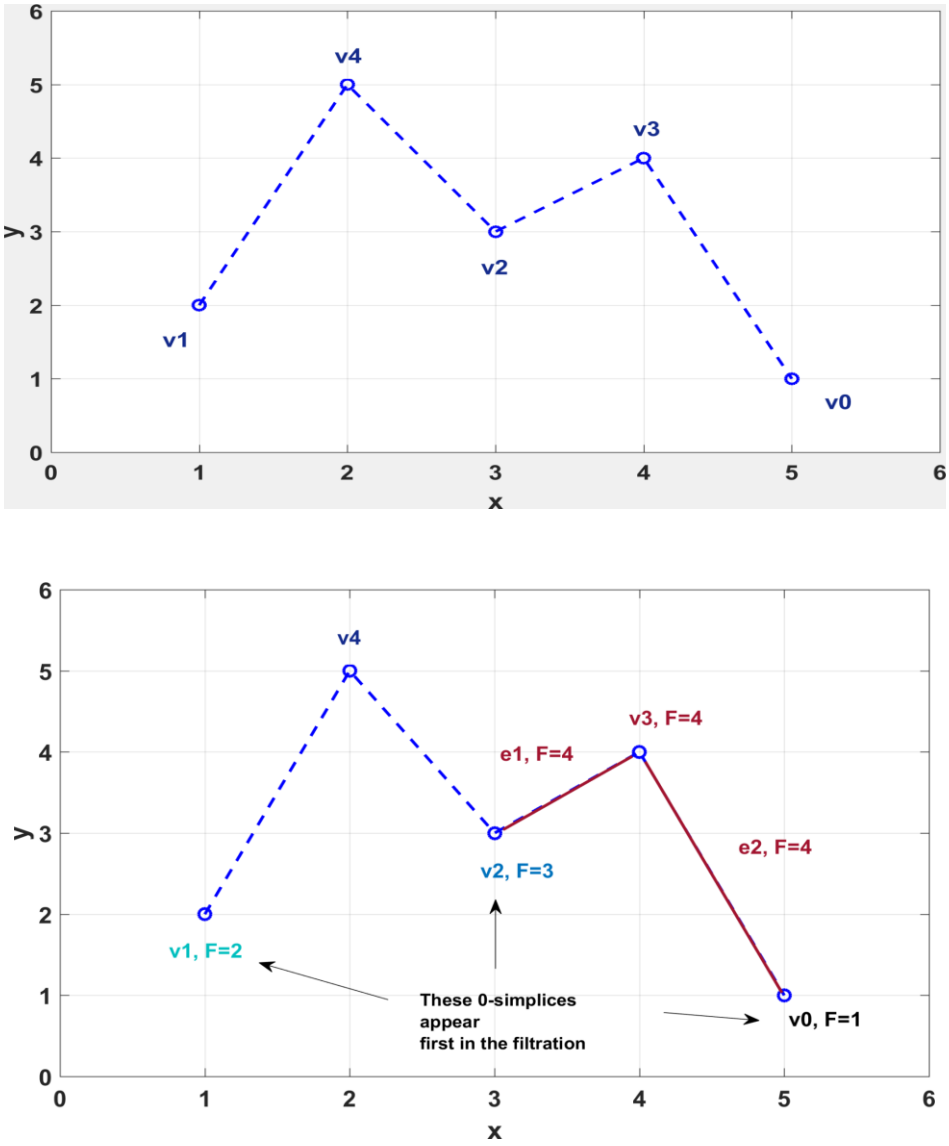


Figure 8. Graphical representation of the methodology that transforms a PL into a filtered simplicial complex. [top] The input is formed by five time points. [bottom] The filtered simplicial complex formed by five 0−simplices and two 1−simplices. Note that there will be other two 1-simplices not shown in the figure at filtration time $\mathcal{F}$ = 5.

2.5 Persistent homology: Representation and Interpretation

A topological space is described by its own topological invariants. While topology has many notions of shape, the most amenable to computation is *homology*, since it associates a sequence of algebraic objects such as abelian groups or modules to other mathematical objects such as topological spaces (Hatcher 2002). Poincaré, the founder of Algebraic Topology, in his masterwork "Analysis Situs" (Poincaré, 1895) first defined *Homology* (in part from Greek ὁμός, homos "identical"). The original motivation for defining homology groups was the observation that two shapes can be distinguished by examining their holes (Siersma 2012). Simply stated, a circle is not a disk because the circle has a hole through it while the disk is solid, and the ordinary sphere is not a circle because the sphere encloses a two-dimensional hole while the circle encloses a one-dimensional hole.

In the case of the simplicial complex $\mathcal{K}$ it might be studied by homology. This algebraic machinery gives us a way of identifying some of the invariants of a topological space through some integer parameters, the *Betti numbers*. More formally, the i-Betti number $\beta_i(\mathcal{K})$ represents the rank of the i-dimensional homology group $H_i(\mathcal{K})$ of a given simplicial complex $\mathcal{K}$. Informally,

- $\beta_0(\mathcal{K})$ is the number of connected components of $\mathcal{K}$,
- $\beta_1(\mathcal{K})$ counts the number of tunnels,
- $\beta_2(\mathcal{K})$ can be thought as the number of voids of $\mathcal{K}$ and, in general,
- $\beta_k(\mathcal{K})$ can be thought as the number of k-dimensional holes of $\mathcal{K}$.

So $\mathcal{K}$ may be described by the dimension of its homological groups. For a more formal definition, we refer to (Hatcher 2002).

In order to examine the topological invariants for a discrete set of points, we use **persistent homology (PH)**, the flagship tool of TDA, introduced in 2000 by Edelsbrunner, Letcher and Zomorodian (2002). It is the combinatorial counterpart of homology for a finite set of points, i.e., a method for computing k−dimensional holes at different spatial resolutions. Basically, it is calculated on a filtered simplicial complex $\mathcal{K}$, (it does not mind how the filtered simplicial complex has been obtained) then it describes how the homology of $\mathcal{K}$ changes along filtration.

The key idea is as follows.

- First, the space must be represented as a simplicial complex $\mathcal{K}$ and a metric must be defined on the space.
- Second, a filtration of $\mathcal{K}$, referred above as different spatial resolutions, is computed. Recall that a filtration $\boldsymbol{\mathcal{F}}$ of $\mathcal{K}$ is a collection of simplicial complexes, i.e., $\boldsymbol{\mathcal{F}} = \{ \mathcal{K}(t) \mid t \in \mathbb{R}\}$ such that $\mathcal{K}(t) \subset \mathcal{K}_s$ for $t < s$ and there exists $t_{max} \in \mathbb{R}$ such that $\mathcal{K}(t_{max}) = \mathcal{K}$. The filtration time (or filter value) of a simplex $\sigma \in \mathcal{K}$ is the smallest t such that $\sigma \in \mathcal{K}(t)$.
- Then, persistent homology $\boldsymbol{\mathcal{H}}$ describes how the homology of a given simplicial complex $\mathcal{K}$ changes along the filtration $\boldsymbol{\mathcal{F}}$. If the same topological feature (i.e., k-dimensional hole) is detected along a large number of subsets in the filtration, then it is likely to represent a true feature of the underlying space, rather than artifacts of sampling, noise, or particular choice of parameters.
- In order to represent the persistent homology, there needs to be a way to represent the data along the filtration. **Barcodes** represent the data in a way that simplices are added but never removed as $\boldsymbol{\mathcal{F}}$ increases. Graphically a barcode is a set of intervals in$\mathbb{R}$. More concretely, this situation can be represented as intervals $[t_{start}, t_{end})$, corresponding to k-dimensional holes that appears at filtration time t_{start} and remains until filtration time t_{end}. The set of intervals, or bars, $[t_{start}, t_{end})$, representing birth and death times of homology classes is called the **persistence barcode** $\boldsymbol{\mathcal{B}}(\boldsymbol{\mathcal{F}})$ of the filtration $\boldsymbol{\mathcal{F}}$ (see Figure 9-a). To sum up, a barcode shows a collection of horizontal line segments in a plane whose horizontal axis corresponds to the parameter filtration, e.g., t, and whose vertical axis represents an (arbitrary) ordering homology generators (Ghrist 2007).
- An equivalent representation is to consider the set of points $[t_{start}, t_{end}) \in \mathbb{R}^2$ plotted in a diagram with the x-axis representing birth time and y-axis representing death time (see Figure 9-b). It is then called the persistence diagram dgm($\boldsymbol{\mathcal{F}}$) of the filtration $\boldsymbol{\mathcal{F}}$. A more formal description can be read in Edelsbrunner and Harer. (2010).

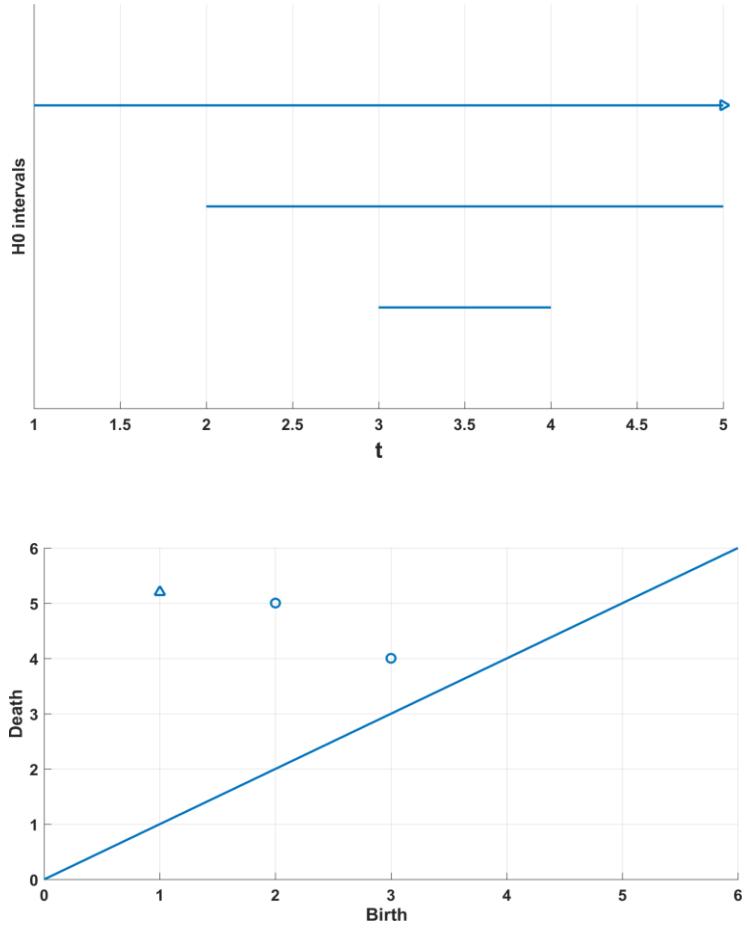


Figure 9. [top] Barcodes for $\boldsymbol{H_0}$ in the example of Figure 8. [bottom] The equivalent persistence diagram**.**

Consider Figure 9-a, which shows the barcode and persistent diagram for the persistent homology of the data points of Figure 8 using the PL filtration. Note that at $\boldsymbol{\mathcal{F}} = 1$, there is only one topological feature corresponding to v0; at $\boldsymbol{\mathcal{F}} = 2$, v0 is still in the space but v1 appears; eventually for $\boldsymbol{\mathcal{F}} = 4$, a new 0-simplex and two 1-simplices are added, {v3} and {v2, v3}, {v0, v3} respectively. Finally, for $\boldsymbol{\mathcal{F}} = 5$, a new 0-simplex and two last 1-simplices are added, {v4}, {v1, v4} and {v2, v4}. As one can see, from this filter value and successive, the space is described by only one persistent connected component, i.e., it has Betti number $\beta_0 = 1$. Visually there is only one infinite line in the barcode.

Persistent homology also allows the calculation of the símplices involved in the holes. These simplices are called *generators*. These generators may play an important role in describing the data under analysis (Merelli et al. 2016).

2.6 Persistent Entropy

How much is ordered the construction of a filtered simplicial complex? In order to answer this question, a new entropy measure has been defined (Rucco et al. 2016). It is called persistent entropy. It is based on the Shannon entropy, and has been recently successfully applied to different scenarios: characterization of the idiotypic immune network, detection of the transition between the pre-ictal and ictal states in EEG signals (Piangerelli et al. 2018), or the classification problem of real long-length noisy signals of DC electrical motors (Rucco et al. 2017), to name a few.

Definition 6. **Persistent entropy**. Given a filtered simplicial complex $\{\mathcal{K}(t): t \in \mathcal{F}\}$, and the corresponding persistence barcode $\mathcal{B}(\mathcal{F}) = \{a_i = [\, t_{start}^{(i)}, t_{end}^{(i)}): i \in I]\}$, the persistent entropy *PE* of the filtered simplicial complex is calculated as

$$PE = -\sum_{i \in I} p_i \log(p_i) \tag{1}$$

where $p_i = \frac{\ell_i}{L}$; $\ell_i = t_{end}^{(i)} - t_{start}^{(i)}$; $L = \sum_{i \in I} \ell_i$.

For example, consider the data cloud given by the five points in Figure 10 [top left]. The persistent entropy calculated using the definition (1) is 1.40 for persistent homology $\mathcal{H}_0$ represented by the barcode in Figure 10 [top right]. As more intervals of different lengths appear the value of the persistent entropy decreases, reflecting in some way more diversity in the barcode.

This way considering the definition, the maximum persistent entropy occurs when all the intervals in the barcode are of equal length (in this situation,$PE = \log(n)$, where n is the number of elements of the barcode). The minimum value is 0 and coincides with the case when there is only one interval as in Figure 10 [top right].

Figure 10 [bottom] a slightly different point cloud is shown with the corresponding persistent homologies $\mathcal{H}_0$ and $\mathcal{H}_1$. As one can see, now the different sizes of the intervals results in a lower persistent entropy (1.25).

As we will examine in the next section, the stability theorem for persistent entropy (Rucco et al. 2017) allows the comparison of entropies computed from the same simplicial complex but equipped with different filtrations.

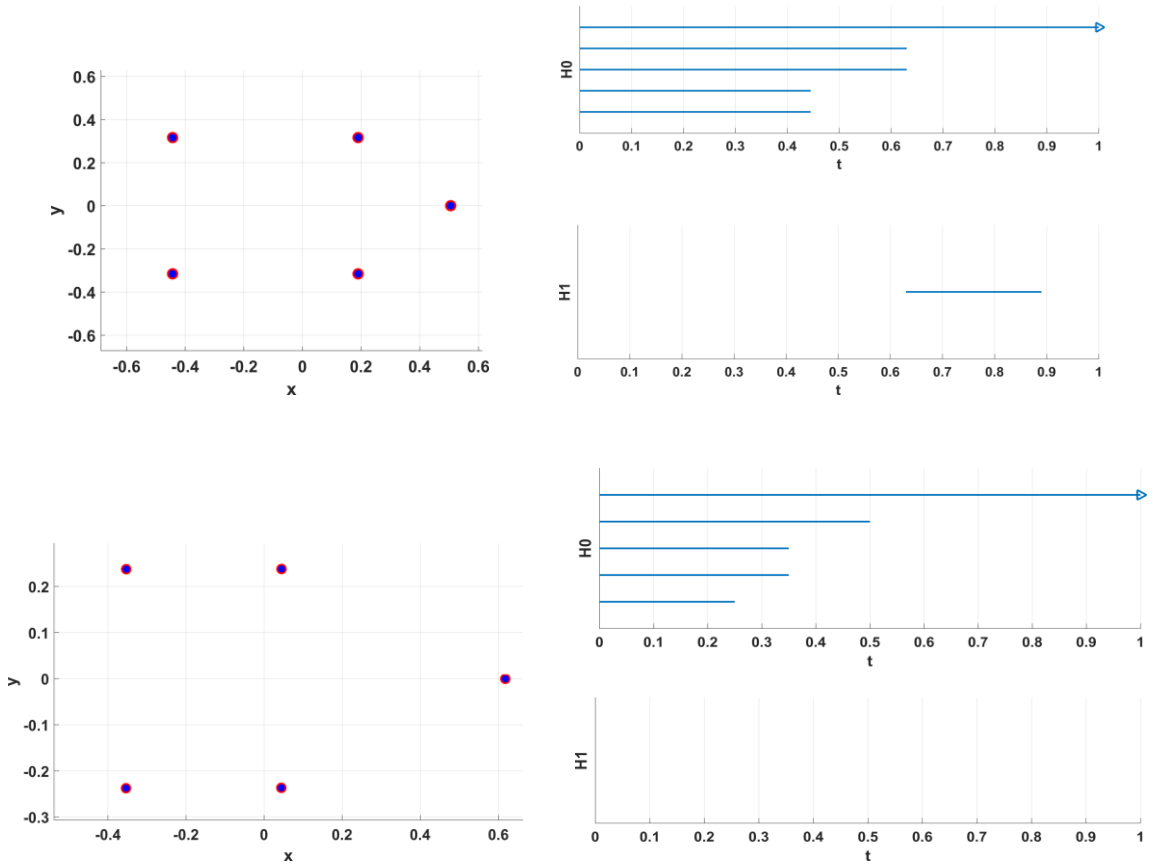


Figure 10. Point clouds and the corresponding PH. [top left] five points in $\mathbb{R}^2$. [top right] the barcodes for H0 and H1. PE for H0 is 1.40 while PE for H1 is 0 since it is a single interval. [bottom left] another five points slightly different in relative distances. [bottom right] the barcodes for H0. In this case PE for H0 is 1.25.

2.6.1 Stability of Persistent Entropy

As already known in the literature (Atienza et al. 2018) if two filter functions or two metric spaces are similar, then their corresponding persistence barcodes will be similar as well. According to Atienza et al.,2018, theorem 3.18 and 3.18 in their paper, this can be resumed by the following conclusion: Small changes in input data implies small changes in the persistent entropy.

This theorem is used in the implementation of the methodology proposes in chapter 3 where a discrete piecewise linear function, i.e., the signal under examination, is transformed into a filtered simplicial complexes. Then the persistent homology is computed and persistent

entropy is calculated. Finally, the calculated persistent entropy gives as a discriminant feature for identifying pattern changes in the signal.

The stability theorem for the persistent entropy gives us the formal support for the comparison of the persistent entropy of two topological spaces (i.e., a filtered complex), subject to the same filtration and this will be effectively used in the experiments described in Chapter 3 and 4.

2.6.2 Persistence barcodes with infinite intervals

In persistent homology theory, intervals that extend to the end of the filter are denoted by [a, ∞). This is indicated by the arrow in the persistence barcode of Figure 10-b and Figure 10-d. Although persistent entropy is defined only for persistence barcodes with intervals of finite length, the definition can be extended to infinite intervals (Atienza et al. 2018) by changing them to [a, m) where m = $\max(\mathcal{F}) + 1$. This way all intervals have finite length. According to Atienza et al., 2018, this approach sends infinite intervals to a fixed value common for all persistence barcodes we are dealing with at that moment. As those authors remark, the more importance we want to give to these intervals, the greater this value should be. Note that the persistent entropy of barcodes in Figure 8 were calculated this using this procedure

Chapter 3

METHODS

The goal of this study is to apply topological data analysis techniques to analyze complexity in time series datasets. For this purpose two situations are considered: artificial, periodic time series (i.e., sinusoids of different frequencies and real-world time series (i.e., EG signals), where the purpose is to detect pattern changing in the time series represented by change of frequency/amplitude or seizure condition in case of the EEG signals. Then the same methodology is applied to characterize complexity in a synthetic signal composed of three types of signals: random time series, chaotic time series and a periodic time series.

3.1 Time Series Analysis and Dataset used in this work

In this study, we use persistent homology as a topological method for measuring the shapes of spaces and the features of functions. Specifically, we are interested in comparing some time series, both artificial and real datasets in order to test the proposed methodologies. In this Chapter, we describe these datasets in details and outline the procedures used to calculate the features that characterize these signals.

All the experiments have been coded in MATLAB and for the topological analysis, we used the Java package Javaplex (Adams and Tausz 2018). The machine precision (minimum value between two signal values) is 2.2204e-16.

3.1.1 Time Series Analysis Overview

Time Series Analysis comprises methods for analyzing time series data in order to extract some meaningful statistics and characteristics of the data such as understand the mechanism that generates the series (explanatory) and predicts future (forecasting) (Zhang

2007). The forecast aims to capture and model underline patterns in time series such as trends, seasonality (exploratory analysis).

There are a number of approaches to modeling time series. The basics concerning time series analysis consist of designing a suitable model, i.e., a model that reflects the underlying structure of the series so that it can be used for future forecasting. In this respect, a time series model is said to be linear or non-linear depending on whether the current value of the series is a linear or non-linear function of past observations.

The most common approaches to time series modeling are (Adhikari and Agrawal 2013):

- Classical, statistical methodology, e.g., the Autoregressive Integrated Moving Average (ARIMA) (Cochrane 1977) and Exponential smoothing models, which are the two most widely used approaches to time series forecasting using linear models. By using ARIMA models we aim to describe the autocorrelations in the data, while the exponential smoothing models are based on a description of the trend and seasonality in the data (Hyndman and Athanasopoulos 2018).
- Machine learning methodologies used mainly for non-linear modeling due to their inherent capability of non-linear modeling (we make no assumptions about the statistical distribution followed by the observations), e.g., Support Vector Regression, where, as the name implies, it uses Support Vector Machine algorithm applied to the case of regression problems. In this case, since the output is a real number, it becomes very difficult to predict the information at hand, which has infinite possibilities. Therefore, a margin of tolerance (epsilon) is set in approximation to the SVM. Anyway, essentially in SVR the input vector is first mapped onto an m-dimensional feature space using some fixed (nonlinear) mapping, and then a linear model is constructed in this feature space (Smola and Schölkopf 2003). Another machine learning methodology is based on the use of Recurring Neural Networks (RNNs), a kind of artificial neural networks with an additional temporal dimension. In brief, it detects patterns in the input sequence and learns when they will probably reoccur, i.e., showing a trend (Weller 2018) and (Che et al. 2018).

In this work, we will use a different approach in order to detect pattern changes in time series. The methodology to be used, completely based on algebraic topology, transforms a

plot into a filtered simplicial complex that is further analyzed by persistent homology (PH). From the homological groups we compute the persistent entropy and the average interval size of PH. These are used as the features to characterize pattern changing in the signal.

3.1.2 Dataset used in this work

In this study, two sets of experiences are conducted to evaluate the use of TDA tools to detect pattern changing in time series. The main goal is the same: apply our pattern changing detection method using PH metrics, i.e., persistent entropy and the average interval of the barcode, to determine whether or not our method detects pattern changes that occurred in the time series:

- The first set is aimed to evaluate the use of two filtered complexes schemes inherently different, i.e., a metric filtration and a non-metric filtration. For this purpose, both filtration methods were applied to two types of time series. First, a synthetic time series, represented by two periodic sinusoids of different frequency and amplitude, gives us the exact point of pattern changing and serves as a reference for the methodology. Second, a real-world noisy signal, represented by EEG signals from a patient that suffers from epilepsy. For these time series, the exact point of the seizure period is known in advance, which gives us the opportunity to qualitatively evaluate our topological based method.

The second set aimed to complexity characterization in time series. For this purpose, a synthetic time series with chaotic behavior (logistic map) was generated. Again the exact moment and duration of this chaotic behavior are known in advance for the evaluation of the method applied.

Group I: synthetic signal and real world-signal

The synthetic signal is made of two sinusoids of different frequencies. In order to calculate some statistics, we generate 20 signals from these sinusoids. Phases are picked from a uniform distribution between 0 and 2π (see Table 1 for details).

Therefore, for the experiment, we consider that the synthetic signal has 2 different patterns (i.e., shapes), due to the frequency difference in its parts. Figure 11 shows a typical synthetic signal.

Table 1. Characteristics of the synthetic signal.

Signal	*Characteristics*	*Size N*	*Sample rate / duration*
Sinusoid 1	Frequency: 8 Hz Amplitude: 1 Phase: a random value in [0,2π]	5120	256Hz / 20s
Sinusoid 2	Frequency: 12 Hz Amplitude: 1.2 Phase: a random value in [0,2π]	5120	256Hz / 20s

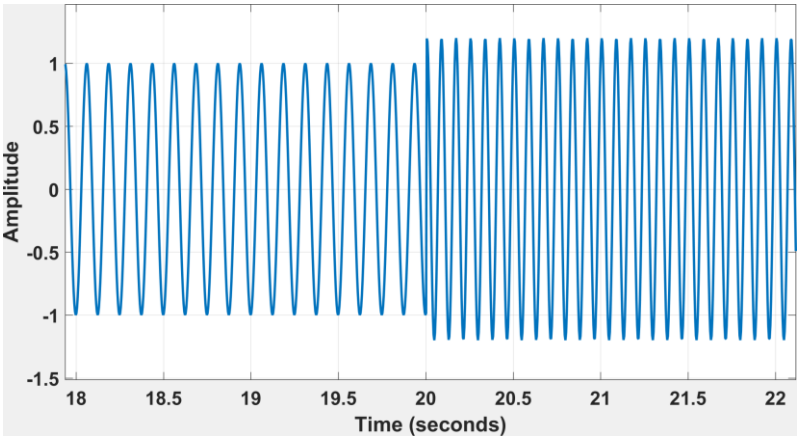


Figure 11. A synthetic signal composed of two sinusoids of 8Hz and 12 Hz. Amplitude ratio 1/1.2.

The real world signals consist of EEG recordings from pediatric subjects with intractable seizures (Shoeb 2009) made available by PhysioBank under the ODC Public Domain Dedication and License v1.0. They were collected at the Children's Hospital Boston (Goldberger Ary L. et al. 2000). Subjects were monitored for up to several days following withdrawal of anti-seizure medication in order to characterize their seizures and assess their candidacy for surgical intervention. Specifically, the EEG signals and seizure time annotations were both read from PhysioNet's Remote server.

For this study, we used the EEG signal of an individual female aged 11, identified as case chb01 in PhysioNet database. The whole EEG consists of 23 channels with 3600s of duration and sample frequency of 256Hz (921600 samples). These signals contain one seizure annotated in the database, e.g., it is possible to select the exact period of seizure occurrence. How persistent homology was calculated over individual channels 3 (T7-P7) and 5 (FP1-F3) and also over the whole set of 23 channels is explained in the following sections.

Figure 12 shows the signals used in this work. In both figures, the small red circles indicate the seizure period provided in the PhysioNet database.

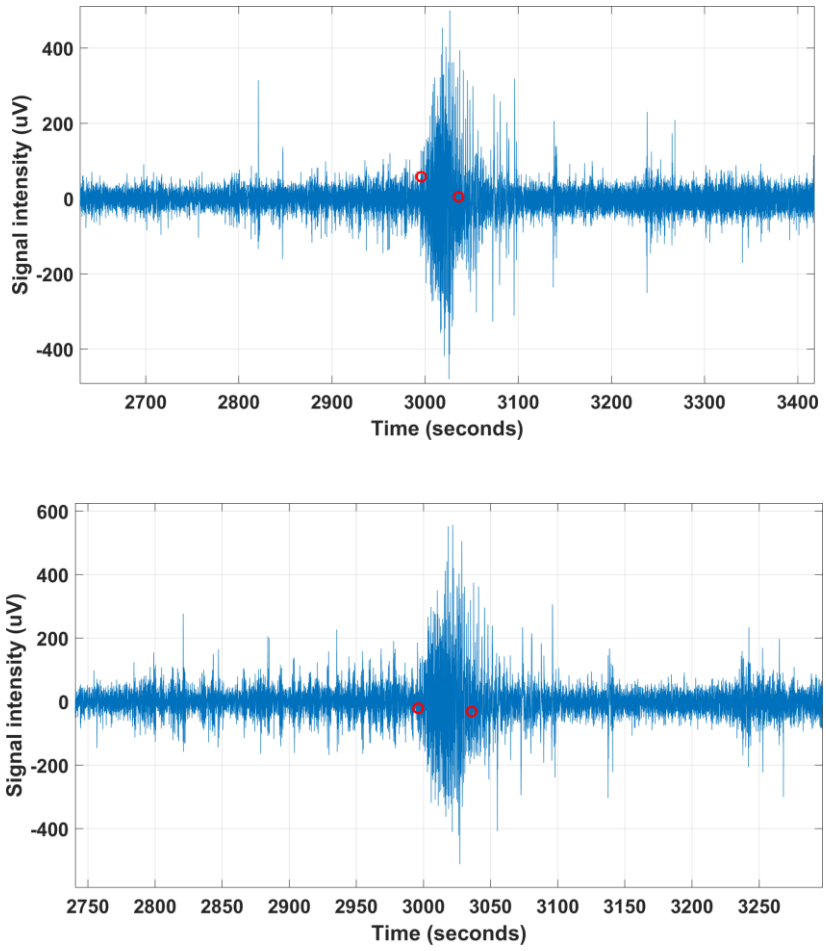


Figure 12. EEG signals. [top] Dataset for channel 1. [bottom] Dataset for channel 5. In both figures, the small circles indicate the seizure period.

Group II: synthetic signal for complexity characterization

The second group consists of a signal composed of three pieces: white noise, plus a portion generated by logistic map and a periodic signal (i.e., a sinusoid). Each piece contains N = 20480 points, with the characteristics given in Table 2. Figure 13 shows the synthetic signal used for this experiment.

Table 2. Characteristics of signal pieces used to study complexity.

Signal piece	*Characteristics*	*Size N*	*duration*
White noise	Sample rate: 256 Hz	20480	80s
Logistic map	$x_{n+1} = r * x_n * (1 - x_n)$ r = 3.9 $x_1 = 0.1$	20480	80s
Sinusoid	Sample rate: 256 Hz Frequency: 12 Hz	20480	80s

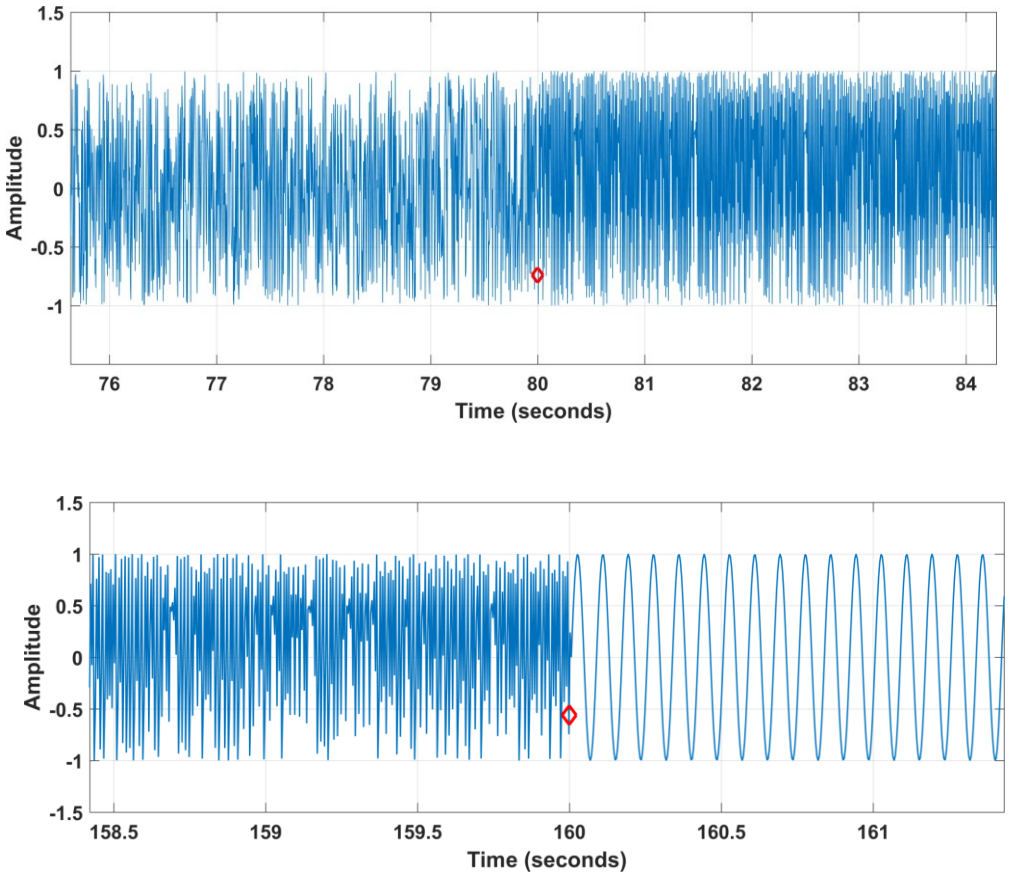


Figure 13. Synthetic signal where the small circles mark the transition of each signal piece: [top] white noise (sample rate 256Hz) and logistic map (r = 3.9, $\mathbf{x_0 = 0.1}$). [bottom] the same logistic map and the 12Hz sinusoid. All pieces have the same duration (80s).

Recall that the discrete logistic is essentially due to Verhulst (Goldstein 2015) who introduced the logistic equation for demographic modeling in his studies of population growth with limited resources

$$x_{n+1} = r * x_n * (1 - x_n), n=1, 2, 3... \quad (2)$$

Here x_n represents the population after n generations and r is the rate of growth.

The discrete logistic map is a relatively simple nonlinear map which is the basis to study chaos and it has been used as a fundamental model in several fields as cryptography, traffic control, tourism models and so on (Tricarico and Visentin 2014).

Tricarico; Visentin (2014), describe several possible behaviors for the population growth as r varies. So in the in the interval [0, 1]. Equation (2) admits two fixed points ($x_1 = 0$, $x_2 = \frac{r-1}{r}$).

For $r \in [0,1]$ population goes to extinction; for $r \in [1,3]$ the fixed point x_1 becomes unstable, while x_2 is asymptotically stable; for $r \in [3, 1+\sqrt{6}]$ the fixed point x_2 becomes unstable and the 2–periodic bifurcation begins, etc.

In this study, we are interested in studying the persistent homology of the chaotic behavior of the logistic equation. So we made r = 3.9 since in the interval [0, 1] the system shows only a chaotic (Shi and Yu 2007) as shown in Figure 14

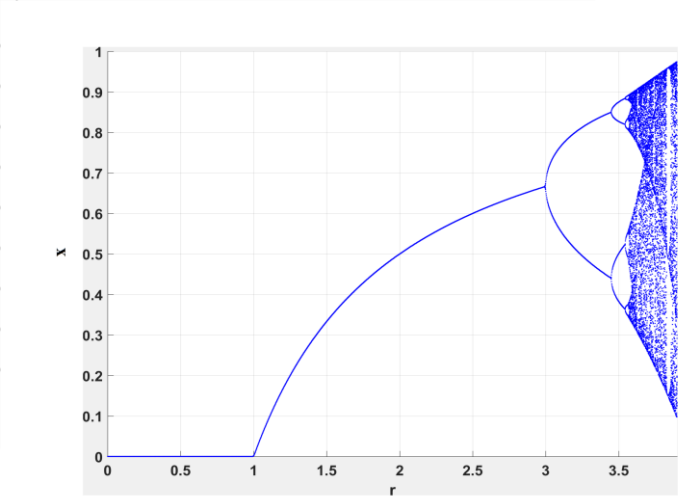


Figure 14. Logistic map 1_D bifurcation plot with initial condition $\mathbf{x_0 = 0.1}$.

After generating the logistic map, we removed its DC value and normalized to the interval [-1, 1] so that all the pieces vary in the same range.

3.2 Time Series Pattern Changing Detection Method

Time-varying dynamical systems can be studied using a sliding window (SW) embedding (Perea and Harer 2013). A sliding window embedding of a signal can be thought of as sliding a "window" of fixed size over a signal and each window can be represented as a point in a possibly high-dimensional space. Note that a SW embedding of a periodic signal has a constant shape over time. Even if the signals are noisy at some level, the shape global shape is expected to remain the same. Therefore, the proposed method is to verify if there is a pattern change along the signals using the SW embedding by extracting some relevant parameters, relevant in the sense that these parameters somehow describe invariant

geometrical (topological) local properties of the signals. So instead of treating the whole signal, that may be large, we prefer to divide it into pieces and examine the topological properties of these individual pieces as the SW "slides" along the signal.

The topological properties should reflect the topological invariance, if present, in the signal. Using Persistent Homology (PH) can accomplish this, i.e., we calculate the PH for the SW, represented by the persistent diagrams (or equivalently the associated barcodes) and calculate two features from these diagrams: the persistent entropy and the average interval size for each relevant homology. By plotting the evolution of these parameters along the SW sliding through the signal, we expect to identify when the curve levels off, providing an approach to discerning a pattern change. Critical to examine the underlying topology of the signal is the selection of a relevant filtration.

Briefly, we follow the sequence indicated in Figure 15 and Figure 16.

3.2.1 Data pre-processing

EEG signal is the only signal subject to preprocessing. We could apply a low pass filter for reducing the noise. Then decrease the sampling rate preserving its main features (decimation). During the research, a Butterworth low pass filter of order 6 (cut-off frequency of 20Hz and sampling rate fs = 256Hz) were used. Both techniques were made in Matlab using the native algorithms. In addition, the decimation factor used were both r = 10 and r = 100. Despite the preprocessing, we could not see any better result and so, the following experiments using EEG did not use any preprocessing.

3.2.2 Methodology 1: applying rips-filtration to sliding windows

Figure 15 shows the procedure used for calculating the persistent homology using rips-filtration. There are three main actions performed. Refer to the figure along the explanations in this section.

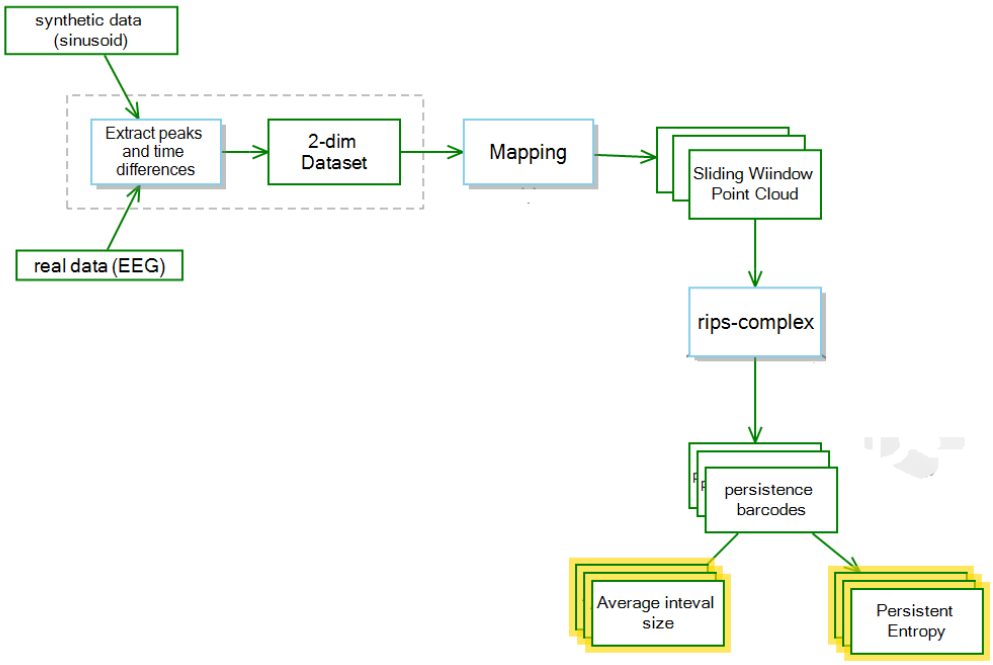


Figure 15. The procedure applied to *Group 1* dataset to generate persistent homology using rips-complex

For *Group I* dataset, the proposed method consists of mapping the original time series to the space containing only the peak values of the signal and the time differences between consecutive peaks (this corresponds to the slashed box in Figure 15). This way the space under consideration somehow reflect both the (local) amplitude changes of the original signal at the related interval time of these maximum amplitudes but in. It is then a 2-dimension vector space:

$$\begin{bmatrix} peak_1 & peak_2 & \dots & peak_m \\ \Delta t_1 & \Delta t_2 & \dots & \Delta t_m \end{bmatrix},$$

where peak values were taken *with minimum separation between peaks.*

We intend to reflect the differences in pattern between the SW in a metric space so that differences in amplitude and in frequencies are mapped to different regions in this new metric space. We do this way because in general rips-complex is used when the topology is defined through a metric.

The resulting signal is then mapped to sliding window point clouds, e.g., the signal is divided into several sliding windows and for each one a distance matrix of size *m* is generated (this corresponds to the "Mapping" box in Figure 15). Since each SW has a metric, it is possible to calculate for each one the persistent homology using the rips-filtration (it corresponds to the box named "rips-complex" in Figure 15). Notice that for each signal type,

we have to choose a value for t_{max} in the filtration. Finally, we then calculate the corresponding persistent entropy and the average interval size.

Finally, we plot the results of calculations in a graph with an x-axis given by the SW ID (slicing window identifier). This way we can follow the evolution of PE and average interval size along the signal.

3.2.3 Methodology 2: applying a lower-star filtration to sliding windows

The intuition on using PL filtration is to address the problem of the comparison between the shapes of plots. In order to satisfy this task, instead of using metric spaces an alternative is to study the shape of the plots by topology (Rucco et al. 2016). Using the technique proposed by Rucco et al (2017), representing a PL function by a topological space, i.e. a filtered simplicial complex, that is described both in qualitative and quantitative measures: using persistent homology and persistent entropy measures. This is reasonable due to the stability theorem for persistent entropy (see section 2,6.1) which is used as a unique global feature for comparing signals. Therefore, as remarked by Rucco et al (2017), given two signals with the same shape, even if they are shifted along one or both directions they have the same persistent entropy.

Based on the techniques proposed by Rucco et al (2017) and Perea and Harer (2014), e.g., the transformation of a signal into a filtered simplicial complex of dimension-1, i.e., a lower-star filtration, in order to study its topology in terms of persistent homology, we divide the signal into sliding windows (SW) of fixed size, without overlapping (refer to Figure 16). Notice that the box named “Mapping” in Figure 16 now reflects the portion of the signal composed of m points, e.g., the size of a single SW. In the end, we remain with an ordered collection of SWs over which persistent homology is calculated. More concretely, using the proposed filtration, we transform each signal portion into a filtered 1-dimensional simplicial complex (“PL Complex” box in Figure 16) according to the following steps:

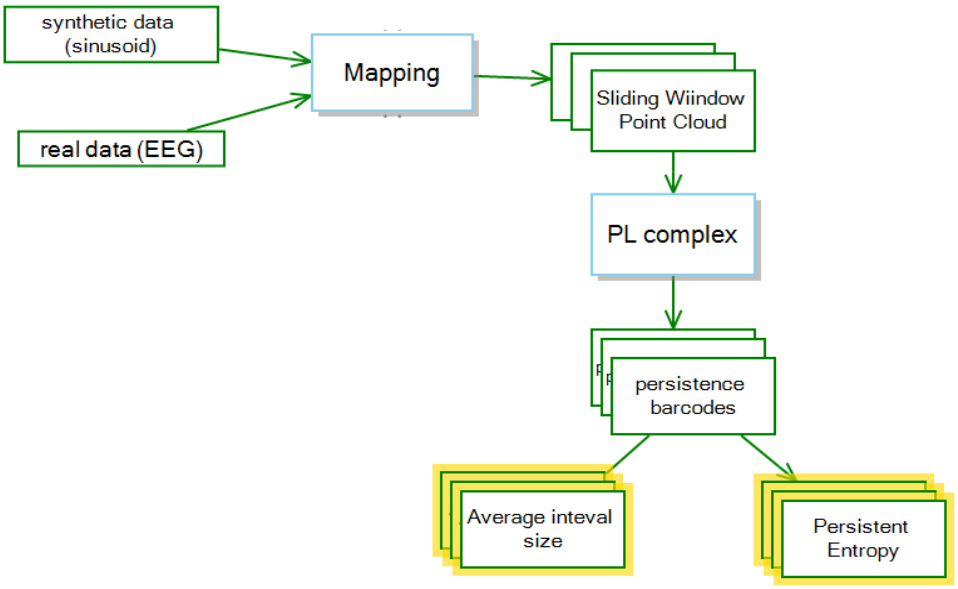


Figure 16. The procedure applied to *Group 1* dataset to generate persistent homology using the piecewise complex

Steps for calculating the piecewise filtration

Input: the value of $\bar{f}$ (the signal) on a finite set of points S $\subseteq \mathbb{R}^n$, i.e., a specific SW
Output: a lower star filtration of f

1: Order the points in S respect to their first coordinate (i.e., in time).

2: Transform S into a filtered simplicial complex:

2.1: Each point of S is a 0-simplex with filter equal to the second coordinate

2.2 A 1-simplex is formed by each pair of two consecutive points in S with filter value $f(\sigma) = \max\{y_i, y_{i+1}\}$ where (x_i, y_i) and $(x_{i+1}, y_{i+1}) \in S$ and $x_i < x_{i+1}$

3: The filtration is obtained presenting at the beginning the simplices formed with the lowest second coordinate.

Finally, the persistent homology is calculated for each SW and from the homological groups, we compute the persistent entropy and the average interval of the corresponding barcode to characterize the signal. The resulting calculations are plotted in a graph with an x-axis given by the SW id. This way we can follow the evolution of PE and average interval size along the signal.

3.3 Time Series Complexity Analysis

Complexity characterizes the behavior of a system or model with multiple interactions among themselves and, at the same time, follows local rules, meaning there is no reasonable higher instruction to define the various possible interactions.

Different formulations and measures were provided for evaluating the complexity of systems (Couture 2017). However, there is no universal definition of complexity (Edmonds 1999). Our intuition somehow places complexity as something between order and disorder, opposed to simplicity (Lopez-Ruiz et al. 2010).

At the same time, in Thermodynamics, entropy measures the level of disorder of a system. The more disordered it is, the more information is needed to describe it precisely. This way entropy-based measures have often been used as measures of complexity including the regularity in noisy time series (Pincus 1995).

In this work, we intend to explore the use of topological methods, specifically using Persistent Homology in the characterization of pattern changes in a dynamic system. For this purpose, we chose the logistic map as the representative class for our complex system.

Sliding windows (SW), or time-delay embedding has been used in the study of dynamical systems to understand the nature of their attractors (Perea and Harer 2013). Here the technique used is driven by the method proposed by Yang and Yang, (2017), but not using autocorrelation as the authors proposed. Since we base our analysis on the associated rips-complex topology, our technique is based on distances between points in each SW cloud constructed as indicated below. Refer to Figure 14 for the procedure applied in this experiment.

For analyzing *Group II* dataset, the method proposed consists of dividing the composed time series $S = \{S_1, \ldots, S_N,\}$, into non-overlapping sliding windows of length L defined by:

$$SW_m = \{S_m, S_{m+1}, \ldots, S_{m+L-1}\}, m = 1 \ldots \lfloor N/m/L \rfloor .$$

To help understand this procedure, note that we are mapping S to a network of m nodes where each one is an L-dimensional vector (in Figure 17, the box named “Mapping”). Therefore, each SW corresponds to a data cloud of m points. For each SW we compute a metric, i.e., a normalize distance matrix of size *m*, and calculate the PH for this point cloud using rips-filtration.

There are some parameters to be chosen to generate the metric for rips-filtration to be applied as shown in Table 3.

Table 3. Parameters that control the metric for rips-filtration.

m	L	r_c	*max_filt_factor*	*partition*
Number of points in the SW cloud	*Size of a point in SW*	*Filter parameter*	Percentage of max_dist to be used as t_{max}	*Number of divisions o f* t_{max}

The filter parameter r_c determines the characteristics of the resulting SW, i.e., the level of details to be examined using rips-filtration. Specifically, it impacts the distance matrix, i.e., all the distances above r_c in the distance matrix are set to *max_dist*. This directly impacts the t_{max} parameter in the persistence barcodes generated for the SW. If it is extremely small, the number of connections among the points becomes smaller and smaller, which may induce strong statistical fluctuations due to a small finite number of connections. Some of the physically meaningful connections may be filtered out as well. However, if r_c becomes extremely large the number of connections tend to generate a complete graph due to the proper nature of rips-complex and the physically meaningful connections in time series may be obfuscated by the filtration.

A second critical parameter, the *max_filt_factor*, implicitly defines t_{max}considered in the barcodes of PH according to the relation

$$t_{max} = \max\_filt\_factor \times \max\_dist$$

where $\max\_dist$ is the maximum distance in the distance matrix derived from the SW point cloud. In the end, it is used to limit the size of the constructed rips-complex for computational efficiency[1].

The parameter *partition* corresponds to the number of divisions used to compute the filtered simplicial complex, i.e., instead of computing filtered simplicial complex for all $t \geq 0$, we only compute it for

$$t \in \{0, \frac{t_{max}}{partition-1}, \frac{2*t_{max}}{partition-1}, \dots, \frac{(N-2)*t_{max}}{partition-1}, t_{max}\}.$$

[1] The complexity of the computation of the PH, in JavaPlex, is of $\Theta(s^3)$ in the worst case and $\Theta(s^2)$ in the average case, where s is the number of simplices.

.

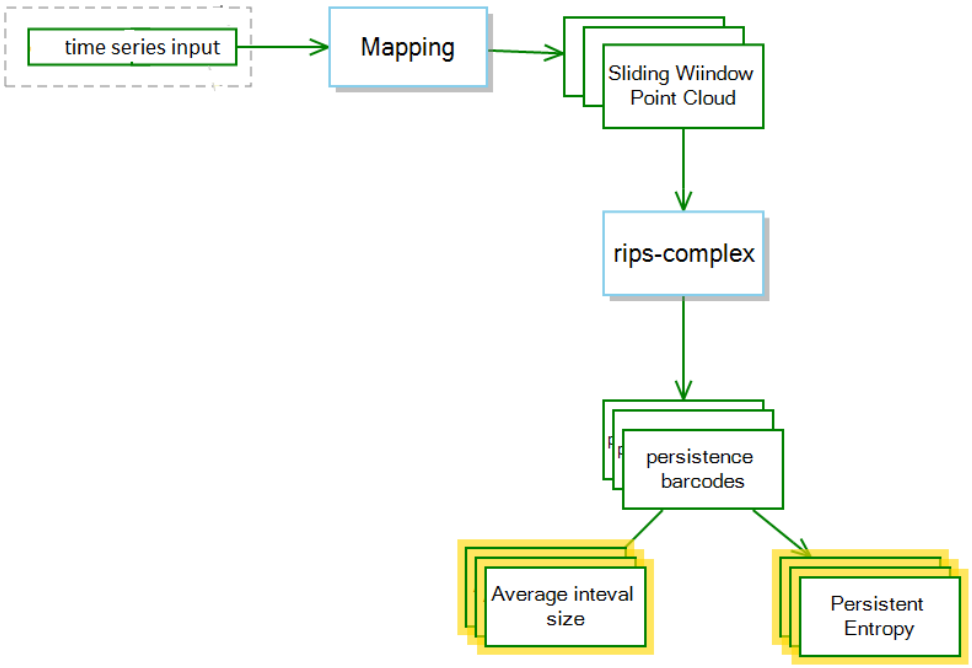


Figure 17. The procedure applied to *Group 2* dataset to generate persistent homology using rips complex

After generating the corresponding barcodes (for the homology groups) the persistent entropy and the average interval for each Sw can be calculated and plotted as a graph with the x-axis indicating the SW identification. This helps us see the changes in these measures along the signal.

Chapter 4

RESULTS

This section describes the experiment performed to evaluate the proposed pattern changing detection technique. These experiments are divided into three sections. In the first section we use synthetic data to evaluate fundamental characteristics, in the second section, we use real datasets represented by EEG signals so that we can evaluate the robustness and generality of the proposed techniques using the two filtration schemes. Finally, in the third section, we use synthetic data to characterize complexity using TDA.

4.1 Artificial Data Experiments

Initially, we present the results of applying the simplicial complex filtrations presented in Chapter 3 for the synthetic dataset case in order to evaluate how the topological analysis help detect the pattern change due to a difference in frequency and amplitude for this signal. These scenarios are used to explore the capability of the method proposed and some issues to take account when choosing the filtrations.

4.1.1 Rips-Filtration

After following the procedure indicated in Figure 15 for two sinusoids with characteristics described in Table 1, the $\mathcal{H}_0$ homology was calculated and the corresponding evolution for the average interval size and PE along the signal were plotted (Figure 18). The small circle in the figures indicates the transition between the frequency and amplitude change. The frequency of the second part of the signal is greater than the frequency of the first part. In Figure 18 (bottom), we can observe that the average interval size has increased, while

PE decreases. In both cases the transition between the sinusoids has a clear impact on PE. In Figure 18 (top right) we can see the $\mathbb{R}^2$ point cloud after the mapping using the peaks and time differences.

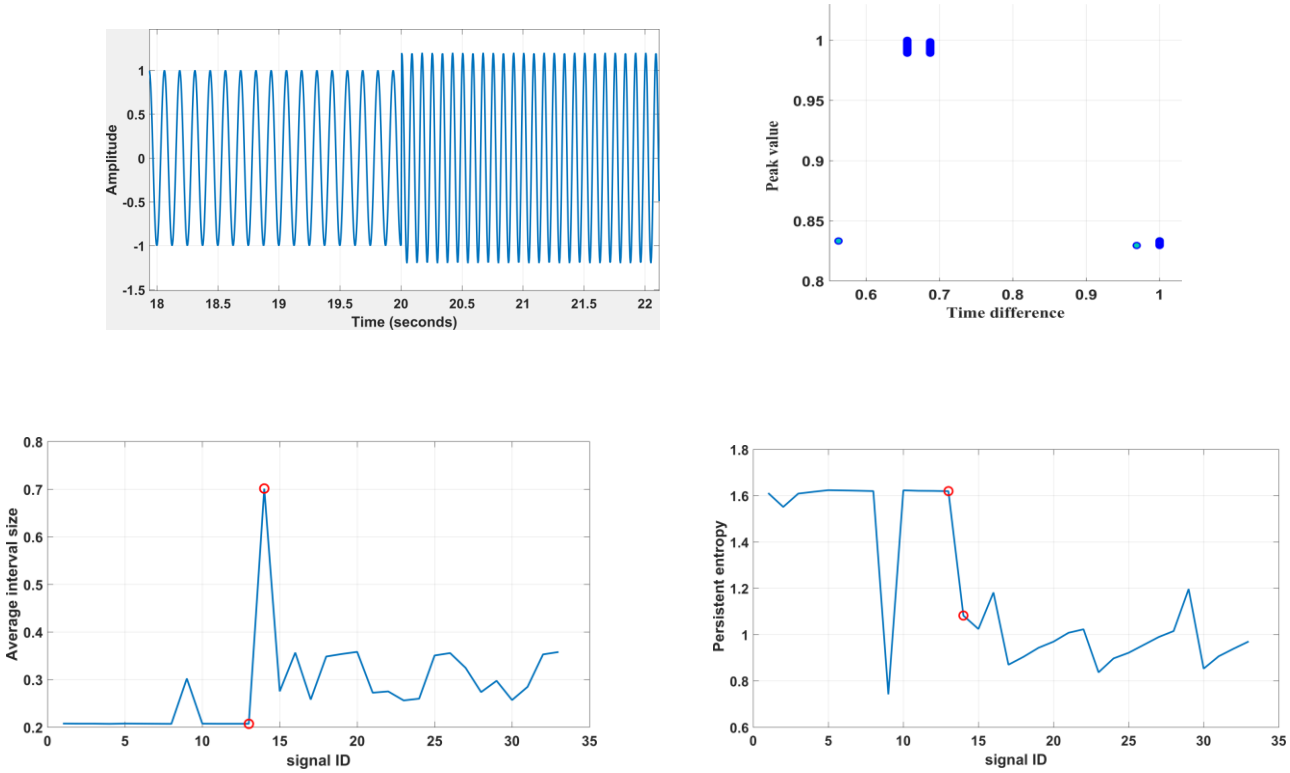


Figure 18. Variation of the average interval size and PE for H0 calculated for synthetic signal using rips-filtration (m=12, no overlapping, end interval = 0.5). [top left] Synthetic signal (2 sinusoids of 8 and 12 Hz and amplitude of 1). [top right] $\mathbb{R}^2$point cloud for the signal after mapping (N=399 peak values). [bottom left] Average interval size for H0 barcode as SW slides along the signal. [bottom right] Variation of Persistent Entropy along the signal. The small circles indicate the separation between the sinusoids. Signal ID indicates the SW identification.

In Figure 19 we can see the average value of the results of the same technique applied to 20 signals of the same size indicated in Table 1. E.g., the average value is taken for the first 20 SW, then for the next 20 SW and so on, until the end of the signals. In the end, this resulted in 33 SW's. Again the transition between the frequency changes is indicated by the small circles in Figure 19.

Now we can observe more clearly the level change for both, the average interval size and the PE along the signal.

In some way, the rips-complex can capture the topological change that occurred in the evolution of the signal. This was expected since the rips-complex is a metric filtration, and the way we mapped the original synthetic time series the frequency change produces two

different point cloud distribution as can be seen in Figure 18-c. This also confirms what was stated in Chapter 2 that the analysis of topological characteristics a point cloud by using simplicial complexes is subject to the right choice of the complex, e.g., the topological spaces must be isomorphic.

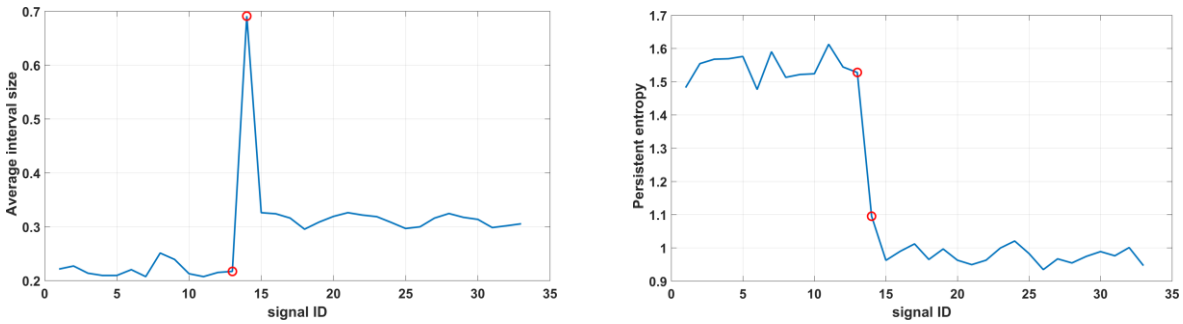


Figure 19. Variation of the average interval size and PE for H0 calculated for 20 sinusoid signals (8Hz and 12Hz, phases are uniformly distributed, N=398 peak values) using rips-filtration (m=12, no overlapping, end interval = 0.5). Signal ID indicates the SW identification. [left] Average interval size for H0 barcodes as SW slides along the signal. [right] Variation of persistent entropy along the signal. The small circles indicate the separation between the sinusoids

4.1.2 Piecewise Filtration

The second method proposed, e.g., piecewise linear filtration, for the synthetic signal following the procedure indicated in Figure 16 resulted in the plots shown in Figure 20. Again, the small circles indicate the transition between the sinusoids of different frequencies (and amplitudes). Since PL is a star filtration we only compute $\mathcal{H}_0$.

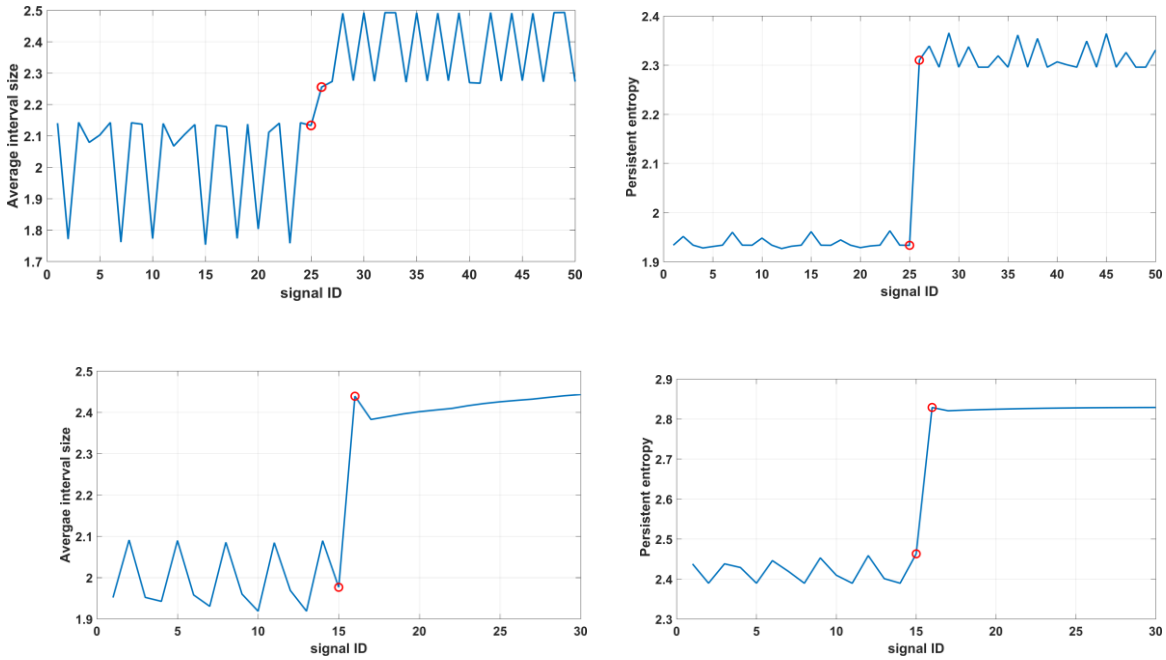


Figure 20. Variation of the average interval size and PE for H0 calculated for synthetic signal using PL-filtration (no overlapping). Signal ID indicates SW identification: [top left] Average interval size for H0 (m=204). [top right] Variation of Persistent Entropy along the signal (m=204). [bottom left] Average interval size for H0 (m=341). [bottom right] Variation of Persistent Entropy along the signal (m=341). The small circles indicate the separation between the sinusoids.

We can observe in Figure 20 a change of plateau level as the sinusoids change in frequency. Now the level change is in the opposite direction. Anyway, the level change in both charts clearly indicates the transition between the sinusoids.

Recall that the motivation for PL filtration is to address the problem of the comparison between the shapes of signals. This way, in Figure 20 (bottom), we can see the effect of increasing the number of points of the SW. We have fewer fluctuations in both plots since we now are comparing a greater portion of the signal at each time.

Figure 21 shows the mean value of the results of the same technique applied to 20 signals of the same size indicated in Table 1. E.g., the mean value is taken for the first 20 SW, then for the next 20 SW and so on, until the end of the signals. In the end, this resulted in 33 SW's. Again the transition between the frequency change is indicated by the small. And again the level change clearly appears for both plots, the average interval size, and the PE along the signal.

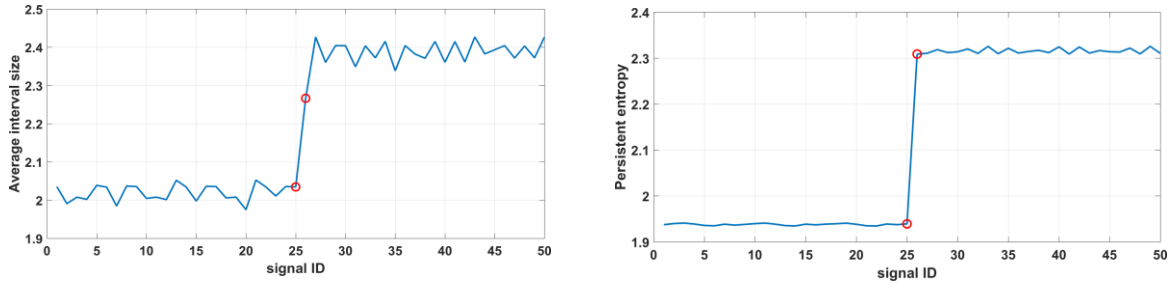


Figure 21. Variation of the average interval size and PE for H0 calculated for 20 sinusoid signals (8Hz and 12Hz, phases are uniformly distributed, N=10240 values) using PL-filtration (m=204, no overlapping). Signal ID indicates the SW identification: [left] Average interval size for H0 barcodes as SW slides along the signal. [right] Variation of Persistent Entropy along the signal. The small circles indicate the separation between the sinusoids.

4.1.3 Comparison of results for synthetic data

Although both filtrations yielded the detection of a pattern change in the artificial data, the choice of filtration is critical. One must notice that the quality of the results when using a metric filtration is constraint by how the topological characteristics of the filtration reflect the topological space under examination. Recall that the barcodes produced by the rips filtration are very sensitive to t_{max}. Besides this, rips filtration eventually produces a complete graph as t_{max} increases too much. This can mask the actual topological characteristics of the topological space under examination.

PL filtration, on the other side, shows a better result as a greater portion of the signal is taken. That is, it is typically targeted to global comparisons of shapes as already observed by Atienza et al., 2018 in their case study of DC motors comparison using PE of piecewise linear complex.

4.2 Real Data Experiments

The real data consist of EEG signals for a patient that has presented epileptic seizure, which is marked in the plots with small circles indicating the beginning and end of the seizure for reference. The calculations followed the procedure indicated in Figure 15 and Figure 16.

4.2.1 Rips-filtration for real data

This section presents the results when applying rips-filtrations to real data, e.g., the EEG dataset. The parameters used in the experiment are shown in Table 4.

Figure 22 shows the results for both channels. The dataset from other channels did not show different results and are not shown in this work.

Table 4. Parameter for calculating Persistent Homology for EEG signal.

Channel	*N*	*m*	*overlap*	t_{max}	*Method*
1	4806	50	No	0.35	rips
5	4373	50	No	0.35	rips

The SW data cloud consisted of 50 2-dimension points as presented in Chapter 3 and we set t_{max}= = 0.35 since the distance matrix of the SW dataset was normalized between 0 and 1. PE and average interval size are normalized.

Observing the results in Figure 22, there is a decrease in the average interval size during the seizure period while the PE plot does not show any detectable by examining the plot.

In Figure 23 [top left] we can observe a typical SW point cloud. The dataset does not show any topological characteristics in terms of holes but we can see the columns of points due to the same time differences. This somehow impacts the rips-complex calculation for proper t_{max} value (as shown in Figure 23 [top right] where we have set t_{max} = 0.001).

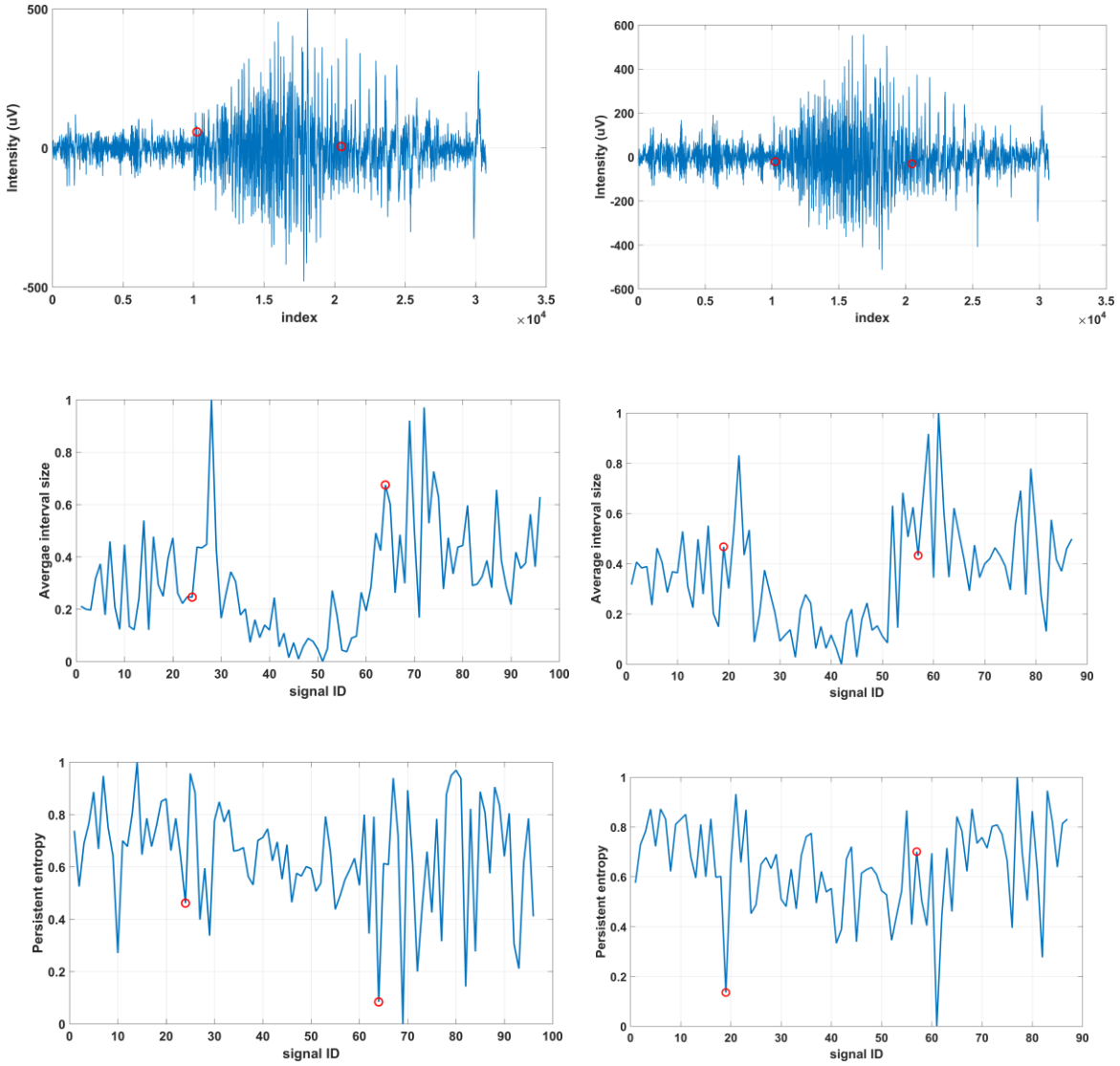


Figure 22. Variation of the average interval size and PE for H0 calculated for EEG signal with seizure using rips-filtration (m=50, no overlapping, end interval = 0.35). Signal ID indicates SW identification. [top left] EEG channel 1 signal. [middle left] variation of the average interval size of H0 along the channel 1 signal. [bottom left] persistent entropy variation for the channel 1 signal. [top right] EEG channel 5 signal. [middle right] variation of the average interval size of H0 along the channel 1 signal. [bottom right] persistent entropy variation for the channel 1 signal. The small circles indicate the seizure interval in all figures. Average interval size and persistent entropy are normalized.

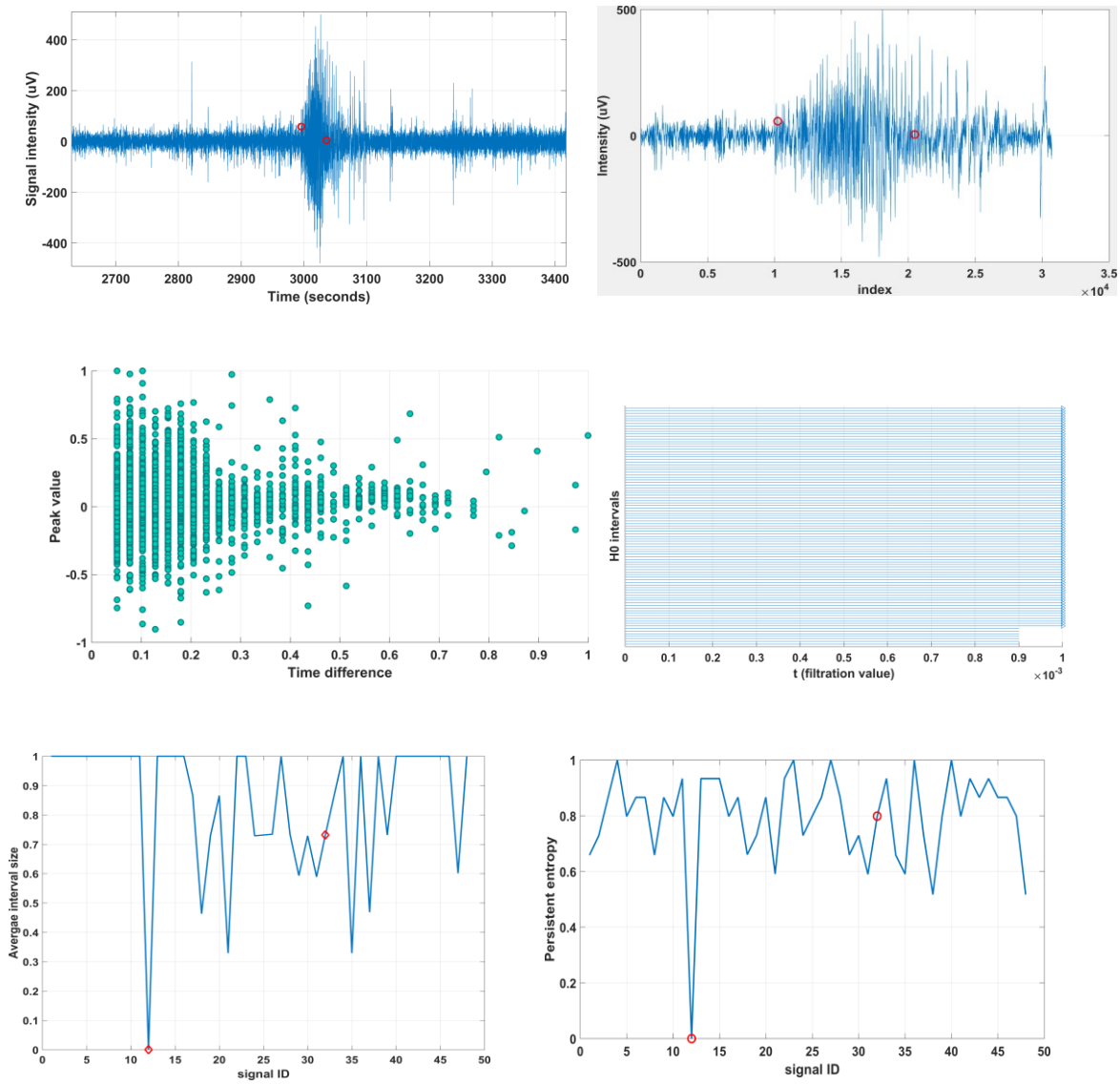


Figure 23. [top left] Portion of channel 1 EEG signal showing seizure period. [top right] The region of the signal over which the SW was applied. [middle left] The corresponding in $\mathbb{R}^2$ points clouds for all 48 SWs (each single SW contains m=100 points). Values are normalized. [midde right] Barcode for the SW-12 using rips filtration with $\boldsymbol{t_{max}}$= 0.001. [bottom left] Normalized average interval size. [bottom right] Normalized persistent entropy. Seizure period is indicated by the small circles.

4.2.2 Piecewise filtration for real data

Figure 24 shows how the average interval size and PE vary along the signal when using PL complex filtration for EEG using datasets from two channels, one and five. Each sliding windows corresponds to $^1/_{1000}$ of the signal length, which sets m = 921 points. We

have used the procedure according to Figure 17. Average interval size and persistent entropy were normalized.

In Figure 24 [left column], both average interval size and PE clearly shows the seizure period (indicated by the small circles) for channel 1 signal. This also happens for channel 5 in Figure 24 [right column]. During the seizure PE value increases and as we will see in section 4.3, it may suggest a decrease in complexity (see Figure 26).

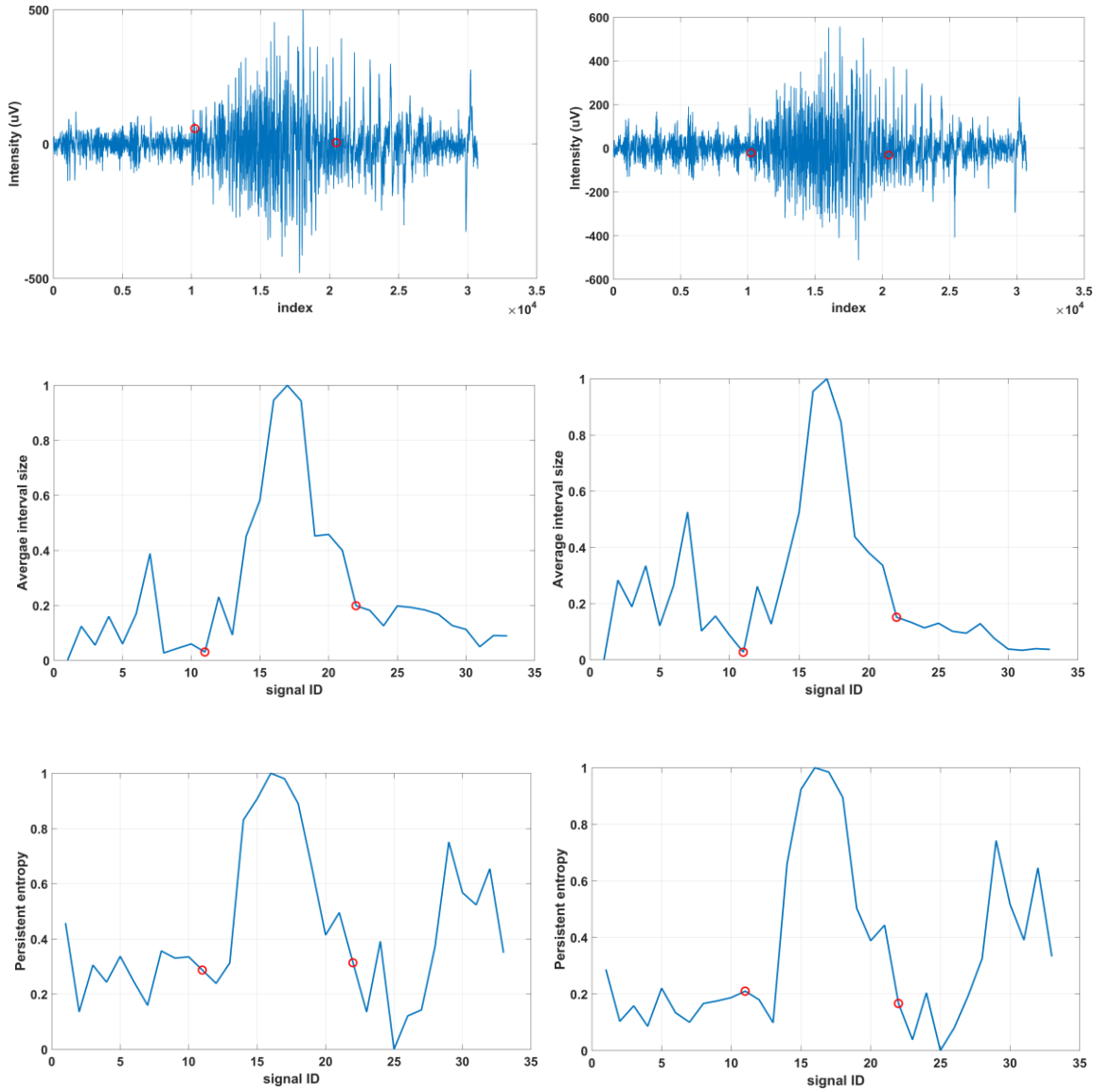


Figure 24. Variation of the average interval size and PE for H0 calculated for EEG with seizure using PL-filtration (m=921, no overlapping of SWs). Signal ID indicates SW identification. [top left] EEG channel 1 signal. [middle left] Normalized average interval and [bottom left] normalized persistent entropy for barcodes as SW slides along the channel 1 signal. [top right]] EEG channel 5 signal. [middle right] normalized average interval size and [bottom right] normalized persistent entropy for barcodes as SW slides along the channel 5 signal. The small circles indicate the seizure interval in all figures.

4.2.3 Comparison of results for real data

Figure 25 shows the values for persistent entropy calculated for EEG, channel 5. In Figure 25 (middle row), PE was calculated for rips-filtration using SW of different sizes, m=50 (left chart) and m=100 (right chart). In both cases, t_{max} =0.35. Recall that these SW point clouds are in fact the mapping of the original EEG time series to $\mathbb{R}^2$ using peak values and time differences as coordinates. Figure 25 (bottom) shows the PE calculated using PL-

filtration for m=460 (left) and m=921 (right). These four plots represent PE without normalizing the results.

Therefore, according to the results presented in Figure 25, we can observe the effect of the choice of the parameters of the sliding windows size. The charts suggest that for larger sizes of the SWs, more global changes are captured, i.e., the plots are smoother.

On the other hand, the calculated PE using PL yields to a clear level rise during the seizure, which could be interpreted as a decrease in complexity if we associate entropy as a measure for complexity in the same direction as proposed by Pincus, (1995).

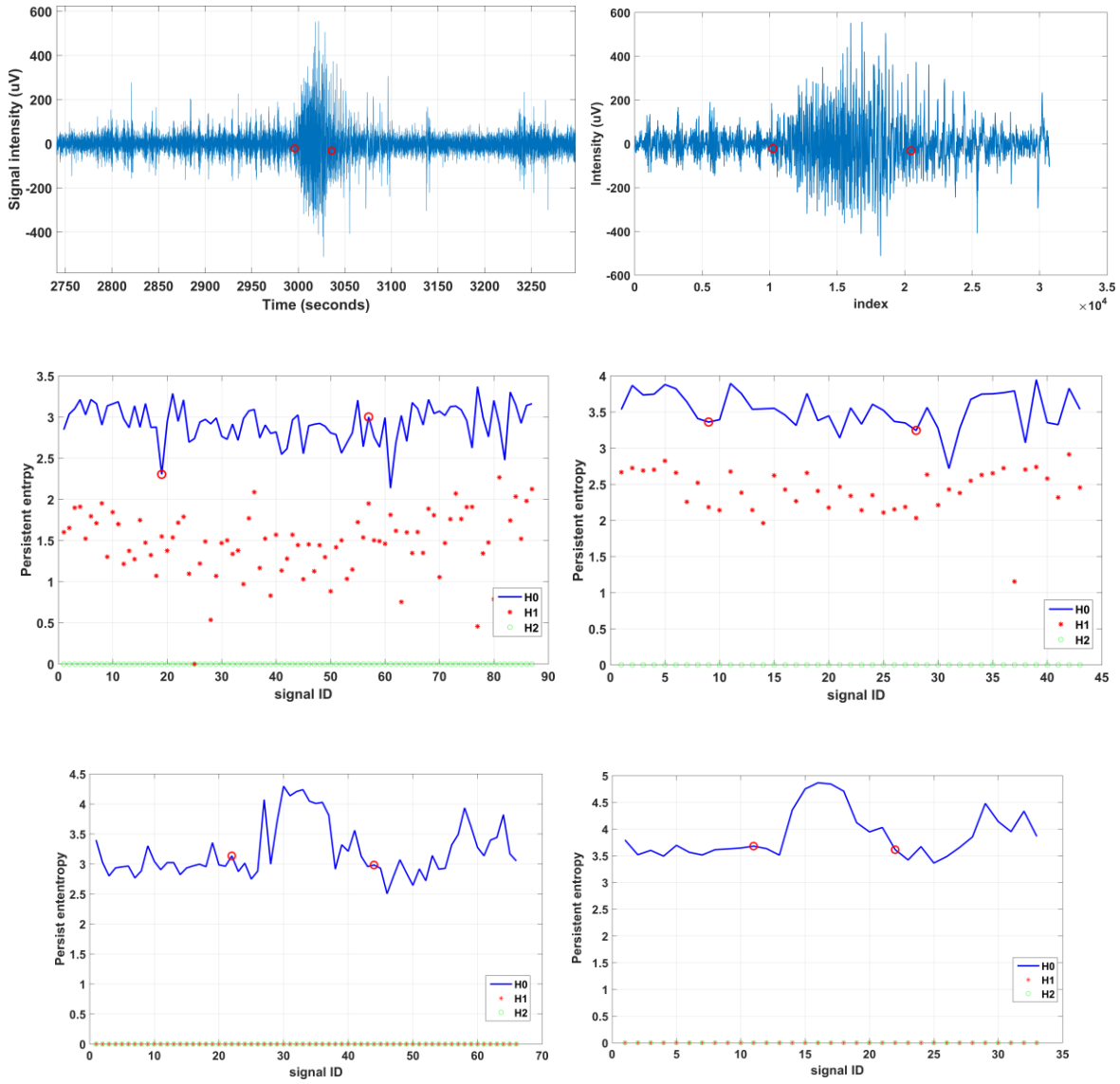


Figure 25. PE comparison. [top row, from left to right] EEG channel 5 signal and the portion over which PH was calculated. [middle row, from left to right] PE using rips filtration for two SW sizes (m=50, t_{max}= 0.35) and (m=100, t_{max}= 0.35). The red stars indicate H1 PH values as well. [bottom row, from left to right] PE using PL filtration for two SW sizes (m=460) and (m=921). Small red circles indicate seizure period.

4.3 Complexity Characterization using TDA

The procedure used to characterize complexity using persistent homology for rips-complex filtration is shown in Figure 14. The parameters used are shown in Table 5. Refer to Table 3 for their meaning.

Table 5. Parameters for calculating $\mathcal{H}_0$ Persistent Homology using rips-filtration.

m	*L*	r_c	t_{max}	*partition*
50	*4*	*0.75*	0.35	*200*

In this work, we intended to explore the use of topological methods in the characterization of pattern changes in complex systems. Figure 26 shows the average interval size and persistent entropy (PE) for a sequence of 307 SW as described in section 3.3. As can be shown in the figure, the small circles indicate the boundaries between the different signals and serves as reference points for the TDA measures. The H0 chart (middle row), shows a clear separation between the PE values for these three different signals: the sinusoid shows the lowest level of PE (the rightmost part), the white noise (leftmost part) has the highest values of PE while the logistic map (the mid part) shows clear intermediate values. In some way, PE reflects how similar in size are the intervals in the barcodes and as complexity rises in the signal, its PE raises as well,

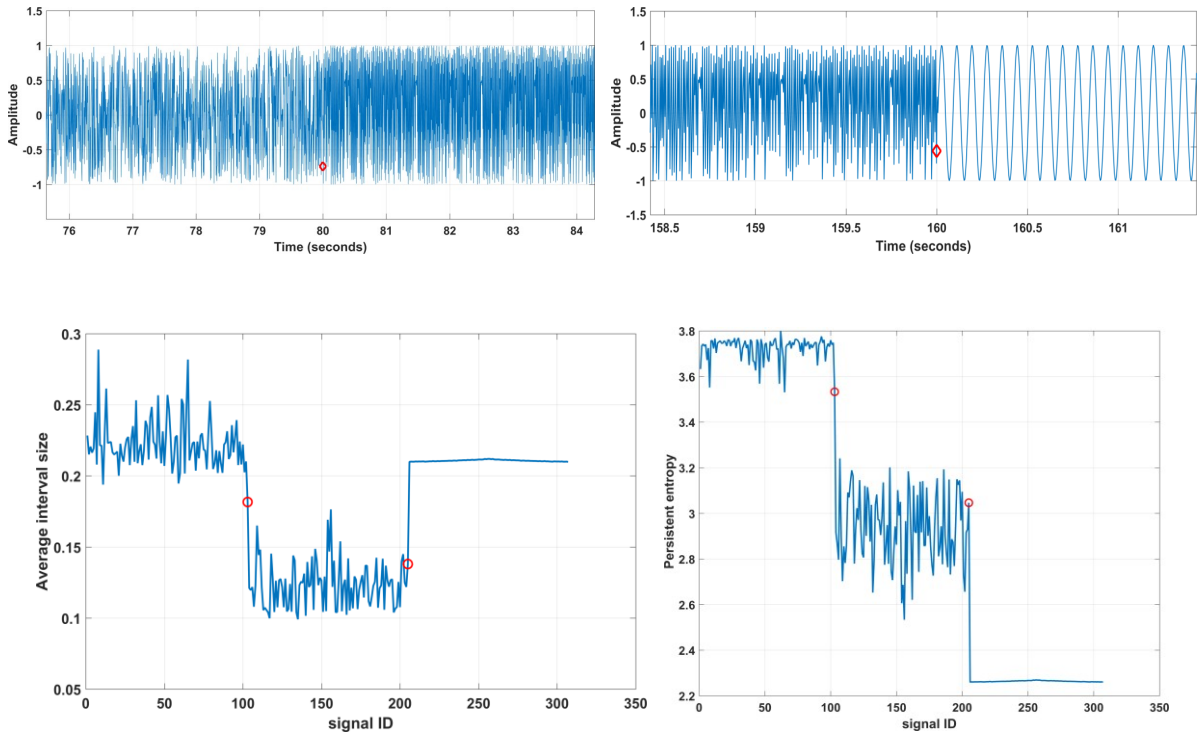


Figure 26. Variation of the average interval size and PE calculated for a synthetic signal compose white noise, logistic map and a 12Hz sinusoid. Each signal contains 20480 data points, N=61440 peak values. [top row] two portions of the synthetic signal with small red circles indicating the transition point. [bottom row] average interval size and persistent entropy variation calculated for H0. Rips-filtration: m=50, L=4, $\mathbf{r_C}$ = 0.75, $\mathbf{t_{max} = 0.35}$. Signal ID indicates the SW identification. The small circles indicate transition points in the synthetic signal**.**

Chapter 5

CONCLUSION

Here we describe the conclusions of the present study, focusing on the characteristics that were evidenced after the analysis of the computational experiments performed. Finally, future work is presented.

In this work we have reviewed the fundamentals of algebraic topology (chapter 2) showing the new topological spaces targeted to point cloud data, the simplicial complexes Moreover, these topological spaces can reveal topological invariant characteristics commonly associated to the notion of shapes such as holes, loops and connected components. In addition, we reviewed the concept of filtration, a nested sequence of simplicial complexes parameterized by a filtration value. Such a feature is important since it allows the identification of the homological features of the space that persist while the filtration value varies over some specified range of interest.

Afterward, two different filtration schemes have been discussed. One is the Vietoris-rips complex filtration that implies a metric in the filtration construction. The other is the piecewise linear filtration, basically a star filtration, where the filtration is created given some function of the vertices.

Moreover, the Persistent Homology (PH) technique has been presented along with the representation of homology using barcodes which are the encoding of PH of a data set in the form of a parameterized version of a Betti number (Ghrist 2007).

Afterward, a pipeline for PH computation was outlined: i) replace a set of data points with a family of simplicial complexes (filtration); ii) encode the PH of the data set using barcodes; iii) interpret the barcodes, i.e., using some metric.

Finally, the concept of persistent entropy has been present together with the stability theorem which gives the formal support for the comparison of the persistent entropy of two time series.

The results that have been presented through the computational experiments described in this work showed that the method proposed to detect pattern changing is functional, being able to identify pattern changing even in a real database, i.e., seizures in the EEG signals. However, some drawbacks were noticed along the comments bellow.

The charts for each simulation obtained for synthetic data could indicate interesting emerging features. In all cases, the transition between the underlying patterns has been identified (see Figure 19 and Figure 21). There is a plateau for both the average interval size and persistent entropy corresponding to the persistent shape of the underlying signal.

On the other hand, for the real-world data, i.e., the EEG time series, the two filtration schemes showed different results.

In the rips filtration case, although there is no clear plateau, the average interval size decreases during the seizure period (Figure 22), while the persistent entropy does not distinguish the seizure period. This result can be explained by examining the corresponding point clouds in Figure 23 (middle left). While many points are clearly separated by the time difference, the overlapping is evident. Since the rips filtration is a metric filtration, when applied to these point clouds, many points will be connected simply due to their proximity, without reflecting any relevant topological feature of the original time series data. This is clearly a disadvantage in using rips filtration. We must guarantee that the complex rips filtration is isomorphic to the topological space under examination. Therefore, the choice of filtration is an essential step for using PH tools.

Another important observation for rips filtration method is the effect of the choice of the maximum value for filtration, i.e., t_{max}. Again, examining the point clouds, if t_{max} is chosen too large, the rips complex will eventually produce a complete graph, misleading the real topological features of the data. On the other hand, for finer filtration values, the noise inherently present in the point clouds may mislead the eventual interpretations of PH barcodes. In these scenarios, the resulting intervals of the barcode tend to infinite lengths, resulting in an increase of persistent entropy. Since $t_{max}$is arbitrary, there must be a criteria to select an “ideal” value. What should be this criteria is something not clear for the author of this work and is another issue to be investigated in the future.

A quite interesting feature is observed in the barcode chart in Figure 23 (middle right). It represents the barcode computed for the SW identified by signal ID=12. We can observe the presence of intervals of different sizes. This causes a decrease in the persistent entropy and the average interval size, as can be seen in Figure 23 (bottom), indicated by the small circle in the, which is in fact the initial of the seizure period of the in the channel 1 EEG tie series. Future investigation will be focusing on this point in order to verify if there is any actual correlation between these results.

In the PL filtration case (applied to the real time series), both the average interval size and persistent entropy distinguished the seizure period (see Figure 24), indicated by the raised values. However, here one has to select a parameter as well, i.e., *m*, the SW size. How we should select a "good" value for *m*? PL filtration uses the shape of the entire SW, i.e., it uses piecewise approximation to the signal to generate the filtration. This suggests that it captures global information of the signal. Therefore, one should set m as large as possible. This is something we left open in this work and may lead to further research in the future.

Finally, one may notice that the results for PE values in both filtration schemes somehow indicated the seizure period but in opposite directions. While PE increased in value for PL complex, apparently, it decreases in value for rips complex filtration.

This is consistent with the fact already pointed out that in general there are many ways to construct a filter out of a collection of simplices (Chintakunta et al. 2015). Chintakunta et al. (2015) mentions it is possible to find filters that minimize persistent entropy. This is something left open in this work that deserves a thorough investigation.

Moreover, we have compared values of PE obtained from the real world data simulations using rips filtration and PL filtration, both with different values for m (the SW) (Figure 25). Two interesting features were observed. First, as mentioned before, in the PL filtration it is possible to associate the PE level rise during the seizure period, while in the rips filtration case, one cannot infer such relation. Second, the simulations used SW of different size. In both cases the larger the size of the SW, the smoother is the variation of PE. This is related to the definition of PE, which is based on how the intervals differ in size. This suggests that the effect of local variations in the data set is reduced when we choose a larger SW. For the PL case, this may also suggest that the larger SW captures larger portions of the time series, producing more trustworthy comparisons, since the PL filtration is related to the piecewise linear representation of the time series.

Finally, the simulation for complexity characterization (Figure 26) showed different plateau levels, which suggests the evidence that complexity, at least for the one-dimensional complex system case, could be distinguished using persistent entropy.

5.1 Publications

The theses presented here generated a conference paper, entitled “Topological data analysis for time series changing point detection”, which was submitted to 2019 International Joint Conference on Neural networks (IJCNN2019).

5.2 Future work

The results presented by the proposed technique were quite interesting. We intend to extend the concepts and tools identified to some other filtered complex and adapt them to the problem of the detection of pattern changing in time series, such as weighted complex filtrations (Petri et al. 2013) and witness complex for computing persistent homology (De Silva and Carlsson 2004).

Other aspects that deserve attention and can be worked in future are related to the tuning of the technique. Specifically, how to choose an optimal value for the parameter ϵ (or t_{max}) in Rips filtration and the choice of the delay and dimension parameters for the sliding windows. Another interesting approach is the use of extending windows instead of sliding windows as proposed here mainly for PL complex filtration (due to the computational complexity involved in the simplex calculation. this may not be feasible for rips complex). The extending windows would increase their size in a step-by-step way, finally covering the entire signal. This may present some interesting features not capture by sliding windows.

Finally, we would like to evaluate the use of TDA and, specifically but not restricted to, the proposed technique, in time series produced by sensor networks as proposed by Silva and Ghrist (2007).

BIBLIOGRAPHY

Adams H, Tausz A. JavaPlex [Internet]. Javaplex. 2018 [cited 2018 Dec 30]. Available from: http://appliedtopology.github.io/javaplex/

Adhikari R, Agrawal RK. An Introductory Study on Time Series Modeling and Forecasting. CoRR [Internet]. 2013;abs/1302.6613. Available from: http://dblp.uni-trier.de/db/journals/corr/corr1302.html#abs-1302-6613

Albert R, Barabási A-L. Statistical mechanics of complex networks. Rev. Mod. Phys. 2002 Jan 30;74(1):47–97.

Atienza N, Gonzalez-Diaz R, Soriano-Trigueros M. On the stability of persistent entropy and new summary functions for TDA. arXiv:1803.08304 [cs, math] [Internet]. 2018 Mar 22 [cited 2018 Dec 16]; Available from: http://arxiv.org/abs/1803.08304

Bhattacharya S, Ghrist R, Kumar V. Multi-robot coverage and exploration on Riemannian manifolds with boundaries. The International Journal of Robotics Research. 2014 Jan 1;33(1):113–37.

Carlsson G. Topology and data. Bulletin of the American Mathematical Society. 2009 Jan 29;46(2):255–308.

Chazal F, Michel B. An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists [Internet]. 2017 [cited 2019 Jan 12]. Available from: https://hal.inria.fr/hal-01614384

Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent Neural Networks for Multivariate Time Series with Missing Values. Scientific Reports [Internet]. 2018 Apr 17 [cited 2019 Jan 7];8(6085). Available from: https://www.nature.com/articles/s41598-018-24271-9

Chintakunta H, Gentimis T, Gonzalez-Diaz R, Jimenez M-J, Krim H. An entropy-based persistence barcode. Pattern Recognition. 2015 Feb;48(2):391–401.

Cochrane JH. Time Series for Macroeconomics and Finance [Internet]. Chicago; 1977 [cited 2018 Jan 11]. Available from: http://econ.lse.ac.uk/staff/wdenhaan/teach/cochrane.pdf

Couture M. Complexity and Chaos - State-of-the-Art; Formulations and Measures of Complexity [Internet]. Canada: Defence R&D Canada – Valcartier; 2017 Sep p. 78. Report No.: 2006–451. Available from: http://cradpdf.drdc-rddc.gc.ca/PDFS/unc65/p528160.pdf

Curry J, Ghrist R, Robinson M. Euler Calculus with Applications to Signals and Sensing. arXiv:1202.0275 [math] [Internet]. 2012 Jan 31 [cited 2019 Jan 7]; Available from: http://arxiv.org/abs/1202.0275

De Silva V, Carlsson G. Topological Estimation Using Witness Complexes. Proceedings of the First Eurographics Conference on Point-Based Graphics [Internet]. Aire-la-Ville,

Switzerland, Switzerland: Eurographics Association; 2004 [cited 2019 Jan 7]. p. 157–166. Available from: http://dx.doi.org/10.2312/SPBG/SPBG04/157-166

Debuiche V. Perspective in Leibniz's invention of Characteristica Geometrica: The problem of Desargues' influence. Historia Mathematica. 2013 Oct 1;40(4):359–85.

Dieudonne J. A History of Algebraic and Differential Topology, 1900 - 1960 | Jean Dieudonné | Springer [Internet]. Birkhauser; 1989 [cited 2019 Jan 6]. Available from: https://www.springer.com/la/book/9780817649067

Edelsbrunner H, Harer J. Computational Topology - an Introduction [Internet]. American Mathematical Society; 2010. Available from: http://www.ams.org/bookstore-getitem/item=MBK-69

Edelsbrunner, Letscher, Zomorodian. Topological Persistence and Simplification. Discrete Comput Geom. 2002 Nov 1;28(4):511–33.

Edmonds B. Syntactic Measures of Complexity [Internet] [Doctoral Thesis]. [Manchester]: University of Manchester; 1999 [cited 2018 Dec 23]. Available from: http://bruce.edmonds.name/thesis/

Erdős P, Rényi A. On Random Graphs I. Publicationes Mathematicae (Debrecen). 1959;6:290–7.

Ghrist R. Barcodes: The persistent topology of data. Bulletin of the American Mathematical Society. 2007 Oct 26;45(01):61–76.

Ghrist R. ELEMENTARY APPLIED TOPOLOGY [Internet]. 1.0. Createspace; 2014 [cited 2019 Jan 6]. Available from: https://www.math.upenn.edu/~ghrist/notes.html

Goldberger Ary L., Amaral Luis A. N., Glass Leon, Hausdorff Jeffrey M., Ivanov Plamen Ch., Mark Roger G., et al. PhysioBank, PhysioToolkit, and PhysioNet. Circulation. 2000 Jun 13;101(23):e215–20.

Goldstein J. The logistic map, functional iteration, and complexity – Emergence: Complexity and Organization [Internet]. 2015 [cited 2018 Dec 18]. Available from: https://journal.emergentpublications.com/article/the-logistic-map-functional-iteration-and-complexity/

Hatcher A. Algebraic Topology. Cambridge University Press; 2002.

Hyndman RJ, Athanasopoulos G. Forecasting: Principles and Practice [Internet]. 2018 [cited 2019 Jan 7]. Available from: https://Otexts.org/fpp2/

Keogh E, Chu S, Hart D, Pazzani M. Segmenting time series: a survey and novel approach. Data Mining in Time Series Databases [Internet]. WORLD SCIENTIFIC; 2004 [cited 2019 Jan 10]. p. 1–21. Available from: https://www.worldscientific.com/doi/abs/10.1142/9789812565402_0001

Last M, Kandel A, Bunke H, editors. Data Mining in Time Series Databases [Internet]. Singapore: World Scientific Publishing Co. Pte. Ltd.; 2004 [cited 2018 Dec 11]. Available from:

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.474.7183&rep=rep1&type=pdf#page=14

Lopez-Ruiz R, Mancini H, Calbet X. A Statistical Measure of Complexity. arXiv:1009.1498 [physics] [Internet]. 2010 Sep 8 [cited 2018 Dec 23]; Available from: http://arxiv.org/abs/1009.1498

Maria C, Boissonnat J-D, Glisse M, Yvinec M. The Gudhi Library: Simplicial Complexes and Persistent Homology. In: Hong H, Yap C, editors. Mathematical Software – ICMS 2014 [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2014 [cited 2019 Jan 7]. p. 167–74. Available from: http://link.springer.com/10.1007/978-3-662-44199-2_28

Merelli E, Piangerelli M, Rucco M, Toller D. A topological approach for multivariate time series characterization: the epileptic brain. Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS) [Internet]. New York City, United States: ACM; 2016 [cited 2018 Dec 16]. Available from: http://eudl.eu/doi/10.4108/eai.3-12-2015.2262525

Merelli E, Rucco M, Sloot P, Tesei L. Topological Characterization of Complex Systems: Using Persistent Entropy. Entropy. 2015 Oct;17(10):6872–92.

Nazarimehr F, Jafari S, Chen G, Kapitaniak T, Kuznetsov N, A. Leonov G, et al. A Tribute to J. C. Sprott. International Journal of Bifurcation and Chaos. 2017 Dec 1;27.

Otter N, Porter MA, Tillmann U, Grindrod P, Harrington HA. A roadmap for the computation of persistent homology. EPJ Data Science. 2017 Dec;6(1):17.

Perea J, Harer J. Sliding Windows and Persistence: An Application of Topological Methods to Signal Analysis. arXiv:1307.6188 [math, stat] [Internet]. 2013 Jul 23 [cited 2018 Dec 16]; Available from: http://arxiv.org/abs/1307.6188

Perea JA, Harer J. Sliding Windows and Persistence: An Application of Topological Methods to Signal Analysis. Found Comput Math. 2015 Jun 1;15(3):799–838.

Petri G, Scolamiero M, Donato I, Vaccarino F. Topological strata of weighted complex networks. PLoS ONE. 2013 Jun 21;8(6):e66506.

Piangerelli M, Rucco M, Tesei L, Merelli E. Topolnogical classifier for detecting the emergence of epileptic seizures. BMC Res Notes [Internet]. 2018 Jun 14 [cited 2019 Jan 12];11. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6003048/

Pincus S. Approximate entropy (ApEn) as a complexity measure. Chaos. 1995 Mar;5(1):110–7.

Ray TS. Evolution, complexity, entropy and artificial reality. Physica D: Nonlinear Phenomena. 1994 Aug 1;75(1):239–63.

Rote G, Vegter G. Computational Topology: An Introduction. In: Boissonnat J-D, Teillaud M, editors. Effective Computational Geometry for Curves and Surfaces [Internet]. Springer Berlin Heidelberg; 2006 [cited 2018 Dec 16]. p. 277–312. Available from: http://link.springer.com/10.1007/978-3-540-33259-6_7

Rucco M, Castiglione F, Merelli E, Pettini M. Characterisation of the Idiotypic Immune Network Through Persistent Entropy. In: Battiston S, De Pellegrini F, Caldarelli G, Merelli E, editors. Proceedings of ECCS 2014 [Internet]. Cham: Springer International Publishing; 2016 [cited 2018 Dec 16]. p. 117–28. Available from: http://link.springer.com/10.1007/978-3-319-29228-1_11

Rucco M, Gonzalez-Diaz R, Jimenez M-J, Atienza N, Cristalli C, Concettoni E, et al. A new topological entropy-based approach for measuring similarities among piecewise linear functions. Signal Processing. 2017 May 1;134:130–8.

Sanderson N, Shugerman E, Molnar S, Meiss JD, Bradley E. Computational Topology Techniques for Characterizing Time-Series Data. arXiv:1708.09359 [cs] [Internet]. 2017 Aug 14 [cited 2018 Dec 16]; Available from: http://arxiv.org/abs/1708.09359

Shannon C, Wiever W. The Mathematical Theory of Communication [Internet]. 10th ed. Urbana: The University of Illinois Press; 1964. Available from: http://www.magmamater.cl/MatheComm.pdf

Shi Y, Yu P. On Chaos of the Logistic Maps. DCDIS. 2007;14(2):175–17496.

Shoeb AH. Application of machine learning to epileptic seizure onset detection and treatment [Internet] [Thesis]. Massachusetts Institute of Technology; 2009 [cited 2018 Dec 16]. Available from: http://dspace.mit.edu/handle/1721.1/54669

Siersma D. Poincare´ and Analysis Situs, the beginning of algebraic topology. NAW. 2012;13(3):196–200.

Silva V de, Ghrist R. Coverage in sensor networks via persistent homology. Algebr. Geom. Topol. 2007;7(1):339–58.

Singh G, Mémoli F, Carlsson G. Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. Eurographics Symposium on Point-Based Graphics. 2007;91–100.

Smola AJ, Schölkopf B. A Tutorial on Support Vector Regression. STATISTICS AND COMPUTING; 2003.

Tricarico M, Visentin F. Logistic map: from order to chaos. Applied Mathematical Sciences. 2014;8:6819–26.

Vlachos M, Gunopulos D, Das G. Indexing time-series under conditions of noise. Data Mining in Time Series Databases [Internet]. WORLD SCIENTIFIC; 2004 [cited 2019 Jan 10]. p. 67–100. Available from: https://www.worldscientific.com/doi/abs/10.1142/9789812565402_0004

Watts DJ, Strogatz SH. Collective dynamics of ‘small-world’ networks. Nature. 1998 Jun;393(6684):440–2.

Weller M. Recurrent Neural Networks for time series forecasting [Internet]. Novatec. 2018 [cited 2019 Jan 7]. Available from: https://www.novatec-gmbh.de/en/recurrent-neural-networks-for-time-series-forecasting/

Yang Y, Yang H. Complex network-based time series analysis. Physica A: Statistical Mechanics and its Applications. 2008 Feb;387(5–6):1381–6.

Zhang GP. A Neural Network Ensemble Method with Jittered Training Data for Time Series Forecasting. Inf. Sci. 2007 Dec;177(23):5329–5346.

Zomorodian A. Topology for Computing [Internet]. New York City, United States: Cambrige University Press; 2005. Available from: http://directory.umm.ac.id/Networking%20Manual/Topology%20for%20Computing.pdf

.