UNIVERSITAT DE BARCELONA

GRAU DE MATEMÀTIQUES

Treball final de grau

Departament de Matemàtiques i Informàtica

Analysis of Financial Time Series using TDA: Theoretical and Empirical Results

Autor: Lloyd Aromi Leaverton

Directors: Carles Casacuberta, Josep Vives

Barcelona, 19 de gener de 2020

Abstract

Topological Data Analysis (TDA) is a recently developed tool designed to study the geometry of finite data sets. In these notes, we describe the theory of persistent homology, which is the background underlying the application of TDA.

Our work is both practical and theoretical. We describe in detail persistence landscape functions, which are a means of visualizing persistent homology, and study some of their properties while deriving a few novel results. From a statistical approach, our theoretical work corroborates the use of TDA to measure changes in the underlying distribution of a data set.

We employ TDA to analyze the log returns of four main financial European indices throughout 2005–2015, comparing our results with the ones in the paper *Topological Data Analysis of Financial Time Series: Landscapes of Crashes* [19]. As in this article, we observe that the norms of persistence landscapes show strong growth prior to substantial financial instability.

Resum

L'Anàlisi Topològica de Dades (ATD) és una eina recent dissenyada per estudiar la geometria associada a un núvol de punts finit. En aquestes notes, descrivim la teoria d'homologia persistent, que és el marc en el qual es basa l'aplicació de l'ATD.

El nostre treball és tant pràctic com teòric. Descrivim en detall les funcions de paisatge de persistència, usades com a eines per visualitzar l'homologia persistent; estudiem algunes de les seves propietats i formulem nous resultats. Prenent un punt de vista estadístic, la nostra feina corrobora l'ús de l'ATD per mesurar fluctuacions d'una distribució subjacent en un conjunt de dades.

Emprem l'ATD per analitzar els retorns dels logaritmes de quatre índexs financers europeus entre els anys 2005 i 2015, i comparem els nostres resultats amb els de l'article *Topological Data Analysis of Financial Time Series: Landscapes of Crashes* [19]. Tal com passava en aquest article, observem que les normes dels paisatges de persistència mostren un fort creixement durant períodes d'alta inestabilitat en el mercat financer.

Acknowledgments

Vull donar les gràcies als meus tutors, per el temps que m'han dedicat i la motivació que m'han donat per seguir endevant amb aquest projecte. Gràcies en especial al Carles Casacuberta per el seguiment i l'esforç per fer que aquest treball quedi el millor possible. Gràcies en especial al Josep Vives per ajudar a formalitzar i demostrar el resultats que volíem.

Gràcies al Bruno Mazorra, per motivarme al llarg d'aquests 4 anys i una mica més, i per una última empenya per aquest treball.

I would like to specially thank Dr. Yuri Katz for motivating and helping me prove some of his results. Without our discussions and the tools you facilitated, I do not think I would have reached my goal in time.

Agraeixo a la familia i als amics més propers, les estones lliures, les converses i els bons moments.

Contents

Introduction

Back in 2017, The Economist published a story titled "The world's most valuable resource is no longer oil, but data". Refraining from any sensationalism, we state this title as a nice premise to motivate our work, since currently data is considered a paramount resource and thus, highly inclined to be the subject of gripping and innovative research.

Topological Data Analysis (TDA) is a relatively novel approach designed to study and measure certain features of discrete multidimensional data sets, commonly treated as point clouds embedded in $\mathbb{R}^n$, using a combination of statistical, computational and topological tools to find shape-like structures in data [1, 2, 3, 6, 13]. For TDA to be applied, data is usually encoded as a discrete geometrical sample naturally embedded in a predefined topological space. Intuitively, TDA explores results from studying the persistence of certain homological features, informally k-dimensional holes, that may arise when constructing simplicial complexes born from our original data set. Accordingly, we use persistent homology [12, 14, 24] as an essential tool to our application.

A standard procedure to compute the persistent homology associated to a point cloud data set relies on the construction of a filtration of simplicial complexes; though there exists various work considering different types of assemblies of complexes (see [3], among others), a valid approach, both theoretical and computationally, is given by the Vietoris-Rips scheme, which contrives complexes by setting a minimum distance parameter ϵ for an edge of a simplex to form, e.g., $\sigma = [p_0, \ldots, p_k]$ forms a k-simplex iff $d(p_i, p_j) < \epsilon$ for all i, j. The basic principle underlying this procedure relies on the fact that *altering* this distance parameter ϵ results in modifying the construction and thus, homological attributes characterizing the simplicial complex are intrinsically dependent on it. We say a feature is more significant if it *persists* for a longer range of parameters, thus considering it relevant qualitatively towards interpreting an underlying geometry; on the other hand, as features tend to persist less they are considered to be of minor importance to determine any objective shape and hence, usually referred to as *noise*. As features appear and disappear, associated parameters encode a *birth-death* pair for every k-dimensional hole. This information is captured in a concise form using the means of a *persistence diagram*. Assertively, every point in the diagram records as coordinates the birth and death of every k-dimensional feature from the corresponding simplicial complex. The ge-

ometry of the natural embedding space of persistence diagrams can be sometimes hard to work with. This is reflected mostly when we wish to compare persistence diagrams between diverging data sets.

An alternative tool to summarize the information contained in a persistence diagram is a *persistence landscape* [1]. The latter consists of a sequence of piecewise linear functions defined in the re-scaled birth-death coordinates. As opposed to the complications arisen from working in the space of persistence diagrams, persistence landscapes are naturally embedded in a Banach space, which makes their treatment much more straightforward. Indeed, one can apply classic and well known tools derived from functional analysis and statistics, e.g., compute expectations, variances and norms [1, 2].

Properly formalizing this scheme requires knowledge both from Algebraic Topology and Geometry, which gives rise to *persistent homology*. A key quality derived from this procedure is the robustness under perturbations of the data [10, 16], which allows an indiscriminate analysis of the data, no matter how noisy it may be, although taking into consideration precision issues in results are bound to occur, naturally.

Lately, much stimulating research focusing on studying the persistence of homological occurrences has sprung remarkable results in a variety of fields, such as the discovery of a subgroup of breast cancers [5], powerful means for image processing [4, 16], novel non-destructive testing methods for material evaluation [11], among others [7]. Important work dedicated to time series analysis has also been devised, studying critical transitions in financial networks [17] and time series of cryptocurrencies [18], as well as financial market crashes [19].

Motivated by these studies, we focus our notes on attempting to essentially replicate, and try to complement, article [19]. This paper, written by M. Guidea and Y. Katz, revolves around a newly conceived method which is used to employ a direct application of TDA to financial time series, while also diving into experimenting with *synthetic* time series, following an underlying random nature. We interpret that the main goal of Guidea and Katz's paper is to determine whether persistent homology can devise a tool to detect early warning signals (EWS) of imminent financial meltdowns, parallel to studying the behavior of L^1 and L^2 norms of persistence landscapes corresponding to preconceived data sets.

As expected, our work shares common goals as to those in [19]; the main contrast lies in the experiments themselves, these being determined by sampling from familiar financial markets. Thus, we analyze the time series of log returns from four relevant European stock indices, namely IBEX 35, CAC 40, FTSE 100 and GDAXI 30, belonging to Spain, France, United Kingdom and Germany, respectively. Collectively, these discrete 1-dimensional signals can be seen as embedded in a 4-dimensional space, namely $\mathbb{R}^4$, where every point has coordinates recording each log return for every index. Then, we implement a sliding window technique, enclosing continuous "pockets" of information of a certain window size $w = 50$. The sliding step is set to one day, i.e., for every shift, resulting 50-point data sets differ from one another by

two points (the "first" and the "last" one). Subsequently, we compute the persistence diagram associated to the persistence of loops (1-dimensional persistent homology) for each induced point cloud in $\mathbb{R}^4$, thus being able to draw out the corresponding persistence landscape and determine its L^1 norm. Later, we group the norms by means of a time series capturing the time dependent fluctuation derived from the considered financial market processes.

During our research process, I came in contact with Dr.Yuri Katz, namely one of the authors of the aforementioned article. His acquaintance allowed me to discuss in real-time some of the results I was gathering on the study of synthetic time series. Noticeably, our discussions revolved around studying the *growth* of norms of persistence landscapes derived from random data sets with a predefined distribution. His proposed experiments helped in expediting our research and in fact were useful to corroborate the theoretical results we were able to prove. In fact, our main result is based on an informal proposition in their notes, which states that *the growth of L^p norms of persistence landscapes derived from normal distributed data sets is proportional to the variance of the distribution.* Our discussions contemplated the veracity of this proposition against different types of distributions, which lead to formulating our (novel) result, which establishes some conditions for this behavior to take place in general.

To summarize, our theoretical study of persistent homology of random data sets is parallel to that of financial time series. Indeed, we explore the application of TDA on synthetic time series resembling active financial markets; by discerning the behavior of norms deriving from random data sets with explicit distribution, we aim to better understand the results obtained from financial data with unknown volatility. Therefore, introductory notions on modeling financial markets are given; we focus in particular on the work of Osborne [23] and Praez [25]. In addition, we were able to *prove* theoretically some empirical results gathered from the study of data born from simulated time series. This is done by working with persistence landscapes derived from randomized data sets, hence treating the persistence landscape as a proper random variable. Outcomes revolve around notions such as boundaries for expected values of L^1 norms of persistence landscapes, and behavior of these under changes of the underlying distribution of the data set. Remarkably, our work exemplifies an unexpected bond between the fields of topology and statistics, and we believe there is still a vast margin for research, both practical and theoretical.

Structure of the work

Our notes are arranged in three main parts. The first part is comprised of Chapters 1 and 2, where we state and expose the necessary mathematical background on which we base our implementations. Chapter 1 includes notions of basic Algebraic Topology, such as simplicial complexes and simplicial homology; Chapter 2 dives into the main matters describing persistent homology, where we define concepts such as persistence modules, persistence diagrams and persistence landscapes. The second part is encompassed by Chapters 3 and 4. Chapter 3 is a necessary fair description of the method employed throughout our experiments. Chapter 4 includes testing on fabricated time series and results on some properties of persistence landscapes. The final part consists of Chapter 5, which includes our case-study of the stated financial indices, as well as graphical portrayals and analysis of the fluctuation of L^1 norms of persistence landscapes subjected to time. Our conclusions are laid out in Chapter 6.

1. Simplicial homology

We assume that the reader is familiar with basic notions of topology. This chapter focuses on the concept of homology. Our aim is to motivate the theory and applications of *persistent homology*, which is the main tool of our work and is discussed in Chapter 2.

1.1 Simplicial complexes

Simplicial homology is defined for simplicial complexes. These are abstract models for polyhedra, which we assume finite. The main purpose of simplicial homology is to provide information on k-dimensional holes that a given topological space may have, for any $k \geq 1$.

A *simplicial complex* is a pair (V, Σ), where V is a finite set that we denote by $V = \{v_0, \ldots, v_n\}$, and Σ is a family of non-empty subsets of V such that if $\sigma \in \Sigma$ and $\tau \subseteq \sigma$ then $\tau \in \Sigma$. Elements of V are called *vertices* of the simplicial complex and elements of Σ are called *simplices*.

A simplicial complex (V, Σ) is *ordered* if the set V is totally ordered. In this case, simplices of (V, Σ) are denoted by $\sigma = [v_{i_0}, \ldots, v_{i_k}]$ and we implicitly assume that $v_{i_0} < \cdots < v_{i_k}$. From this point forward, we assume every given simplicial complex to be ordered.

Although the notion of a simplicial complex is purely combinatorial, we can associate a topological space to every simplicial complex (V, Σ) in such a way that the elements of Σ become simplices in Euclidean geometry, i.e., points, lines, triangles, etc., as we will see consecutively. For a set of points $\{p_0, \ldots, p_k\}$ in $\mathbb{R}^n$ such that the vectors $p_1 - p_0, \ldots, p_k - p_0$ are linearly independent, we denote

$$\Delta(p_0, \ldots, p_k) = \Big\{ \sum_{i=0}^{k} \lambda_i p_i \in \mathbb{R}^n \mid \lambda_i \geq 0, \sum_{i=0}^{k} \lambda_i = 1 \Big\}.$$

Thus $\Delta(p_0, \ldots, p_k)$ is the convex hull of $\{p_0, \ldots, p_k\}$ in $\mathbb{R}^n$, and it is referred to as a *geometrical k-dimensional simplex*. In this way (V, Σ) determines a topological space, namely the subspace of $\mathbb{R}^{n+1}$ (with the Euclidean topology) defined as

$$|(V, \Sigma)| = \bigcup_{\sigma \in \Sigma} c(\sigma),$$

where, for each simplex $\sigma = [v_{i_0}, \ldots, v_{i_k}]$ of V, we denote $c(\sigma) = \Delta(e_{i_0}, \ldots, e_{i_k})$ and $\{e_i\}_{i=0}^n$ are the unit coordinate points $e_i = (0, \ldots, 1^{(i+1)}, \ldots, 0)$ in $\mathbb{R}^{n+1}$. We call $|(V, \Sigma)|$ the *geometrical realization* of the (ordered) simplicial complex (V, Σ).

If X is a topological space such that $X \cong |(V, \Sigma)|$ for a simplicial complex (V, Σ), we say that X is *triangulable* and (V, Σ) is said to be a *triangulation* of X.

To ease notation, we denote simplicial complexes as K, L, etc., meaning that $K = (V_K, \Sigma_K)$, and for every $\sigma \in \Sigma_K$ with $\sigma = [v_{i_0}, \ldots, v_{i_k}]$ we say that σ is a *k-dimensional simplex* or *k-simplex* of K.

Given simplicial complexes $K = (V_K, \Sigma_K)$ and $L = (V_L, \Sigma_L)$, a *simplicial map* from K to L is a function $\varphi\colon V_K \longrightarrow V_L$ such that, for every k-simplex $[v_{i_0}, \ldots, v_{i_k}]$ of K, the elements $\varphi(v_{i_0}), \ldots, \varphi(v_{i_k})$ form a simplex in L (of dimension less than or equal to k).

1.2 Simplicial homology groups

For a simplicial complex K, we denote by $C_p(K)$ the free abelian group generated by the p-dimensional simplices of K,

$$C_p(K) = \Big\{ \sum \lambda_i \sigma_i \mid \lambda_i \in \mathbb{Z},\, \sigma_i = [v_{i_0}, \ldots, v_{i_p}] \in K \Big\}.$$

More generally, if R is a commutative unitary ring, we denote by $C_p(K; R)$ the free R-module generated by the p-simplices of K. Thus, $C_p(K) = C_p(K; \mathbb{Z})$.

The *p-boundary operator* is the group homomorphism $\partial_p\colon C_p(K) \longrightarrow C_{p-1}(K)$ given by

$$\partial_p[v_{i_0}, \ldots, v_{i_p}] = \sum_{0 \le k \le p} (-1)^k [v_{i_0}, \ldots, \widehat{v_{i_k}}, \ldots, v_{i_p}],$$

where $\widehat{v_{i_k}}$ denotes the fact that v_{i_k} is omitted. We also assume that $C_{-1}(K) = 0$, so $\partial_0 = 0$. For $p \ge 0$, the boundary operators have the property that

$$\partial_p \circ \partial_{p+1} = 0. \tag{1.1}$$

A sequence of abelian groups and group homomorphisms

$$\cdots C_{n+1} \xrightarrow{\partial_{n+1}} C_n \xrightarrow{\partial_n} C_{n-1} \xrightarrow{\partial_{n-1}} C_{n-2} \cdots$$

such that $\partial_n \circ \partial_{n+1} = 0$ for all n is called a *chain complex*.

Hence $(C_*(K), \partial_*)$ is a chain complex, called the *simplicial chain complex* of K. For each $p \ge 0$, the kernel $\mathrm{Ker}(\partial_p)$ is referred to as the subgroup of *p-cycles* of $C_*(K)$, and the image $\mathrm{Im}(\partial_{p+1})$ is called the subgroup of *p-boundaries* of $C_*(K)$. From (1.1) we deduce that $\mathrm{Im}(\partial_{p+1}) \subseteq \mathrm{Ker}(\partial_p)$ for all p, so we can define:

Definition 1.1. Given a simplicial complex K, we define its *p-dimensional simplicial homology* to be the abelian group

$$H_p(K) = H_p(C_*(K), \partial_*) = \mathrm{Ker}(\partial_p)/\mathrm{Im}(\partial_{p+1}).$$

Similarly, if R is a unitary commutative ring, we define the p-dimensional simplicial homology of K with coefficients in R to be the R-module

$$H_p(K; R) = H_p(C_*(K; R), \partial_*).$$

For a pair of chain complexes (C_*, ∂_*), (D_*, ∂'_*), a morphism $f_*\colon C_* \to D_*$ is a sequence of group homomorphisms $f_p\colon C_p \to D_p$ such that

$$f_{p-1} \circ \partial_p = \partial'_p \circ f_p \text{ for all } p. \tag{1.2}$$

From (1.2) it follows that a morphism $f_*\colon C_* \to D_*$ of chain complexes sends cycles to cycles and boundaries to boundaries, and thus induces a group homomorphism in homology:

$$H_*(f_*)\colon H_*(C_*, \partial_*) \longrightarrow H_*(D_*, \partial'_*).$$

Specifically, if a p-cycle $z \in C_*$ is a representative of $[z] \in H_p(C_*)$, then

$$H_*(f_*)([z]) = [f_*(z)].$$

A classical result in homology theory states that for a given simplicial complex K and its geometrical realization $|K|$, the simplicial homology groups $H_p(K)$ are isomorphic to the singular homology groups of the topological space $|K|$. For a detailed exposition on singular homology we refer the reader to [21].

Throughout our notes, we will mostly be working with the homology groups H_0 and H_1. Thus, we deem necessary to discuss an example illustrating a calculation of H_0 and H_1 for a given simplicial complex.

Example 1.2. Let $K = (V_K, \Sigma_K)$ be a simplicial complex with vertex set

$$V_K = \{v_0, v_1, v_2\},$$

and suppose that Σ_K is composed of the 0-simplices $[v_0]$, $[v_1]$, $[v_2]$ and the 1-simplices $[v_0, v_1]$, $[v_1, v_2]$, $[v_0, v_2]$. The geometrical realization of this complex is the perimeter of a triangle in $\mathbb{R}^2$.

We have the following free abelian groups:

$$\begin{aligned} C_2 &= 0, \\ C_1 &= \mathbb{Z}[v_0, v_1] \oplus \mathbb{Z}[v_0, v_2] \oplus \mathbb{Z}[v_1, v_2] \cong \mathbb{Z}^3, \\ C_0 &= \mathbb{Z}[v_0] \oplus \mathbb{Z}[v_1] \oplus \mathbb{Z}[v_2] \cong \mathbb{Z}^3, \end{aligned}$$

and the boundary operators

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

We obtain that $\partial_2 = 0$, $\partial_1[v_i, v_j] = [v_j] - [v_i]$, and $\partial_0 = 0$. Now, we can compute the matrix for the boundary operator ∂_1 by mapping the generators of C_1 to C_0:

$$\begin{aligned} [v_0, v_1] &\mapsto [v_1] - [v_0] \\ [v_0, v_2] &\mapsto [v_2] - [v_0] \\ [v_1, v_2] &\mapsto [v_2] - [v_1], \end{aligned}$$

and we have the corresponding matrix

$$M = \begin{pmatrix} -1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & 1 \end{pmatrix}.$$

Therefore, $\dim \mathrm{Im}(\partial_1) = 2$ and $\dim \mathrm{Ker}(\partial_1) = 1$, while $\dim \mathrm{Im}(\partial_1) = 2$. Additionally, $\mathrm{Ker}(\partial_0) = C_0$ and $\mathrm{Im}(\partial_2) = 0$. Hence the homology groups of K are

$$H_0 = \mathrm{Ker}(\partial_0)/\mathrm{Im}(\partial_1) \cong \mathbb{Z},$$
$$H_1 = \mathrm{Ker}(\partial_1)/\mathrm{Im}(\partial_2) \cong \mathbb{Z}.$$

Looking back to the definition of p-homology groups, we can view these as being quotients of "cycles modulo boundaries" for every dimension p.

Indeed, in our case H_1 is isomorphic to $\mathbb{Z}$. A generator for H_1 is the homology class of the 1-cycle

$$[v_0, v_1] - [v_0, v_2] + [v_1, v_2],$$

which yields a loop in the geometrical realization of K.

The group H_0 describes or measures connectedness. This is due to the fact that two vertices of K are in the same 0-homology class if and only if they are connected by an edge path. In our case, H_0 is isomorphic to $\mathbb{Z}$. This means that the geometric realization $|K|$ has precisely one connected component.

2. Persistence

In this chapter we introduce the notion of *persistence* and its role in homology theory. The idea of studying the persistence of certain homological features, such as k-dimensional holes, as a method of studying the topological structure of an underlying space was first introduced in [14]. We also refer to [3] as one of the main texts regarding this subject.

Throughout, we study the topological properties of simplicial complexes derived from a discrete set $\mathbb{X} \subset \mathbb{R}^n$, also referred to as a finite *point cloud* or *data set*. In addition, we assume $d(x, y)$ to denote the Euclidean distance between points x and y of $\mathbb{R}^n$.

2.1 Filtrations

In this section we consider *filtrations* of simplicial complexes. We start giving some necessary definitions and we proceed to introduce the Čech complex and, closely related, the Vietoris-Rips complex for a given point cloud. These are the main focus of this section and the latter will be basic from a computational perspective, since these complexes will be the ones we use for topological data analysis (TDA).

Let $K = (V_K, \Sigma_K)$ be a simplicial complex. A *simplicial subcomplex* of K is a simplicial complex $L = (V_L, \Sigma_L)$ such that $V_L \subseteq V_K$ and $\Sigma_L \subseteq \Sigma_K$. A *filtration* of K is an ascending sequence of simplicial subcomplexes

$$\emptyset \subseteq K_0 \subseteq K_1 \subseteq \ldots \subseteq K_r = K.$$

2.1.1 Čech complexes

For a finite point cloud $\mathbb{X}$, a *covering* of $\mathbb{X}$ is a collection of subsets $\{\mathcal{U}_\alpha\}_{\alpha \in I}$ of $\mathbb{R}^n$ such that $\mathbb{X} \subset \bigcup_{\alpha \in I} U_\alpha$. The following notion is classical in algebraic topology:

Definition 2.1. Let $\mathcal{U} = \{U_\alpha\}_{\alpha \in I}$ be any collection of sets in $\mathbb{R}^n$. The *nerve* of $\mathcal{U}$, denoted $N(\mathcal{U})$, is the simplicial complex $N(\mathcal{U}) = (I, \Sigma)$ with vertex set I such that:

i) $\emptyset \in \Sigma$;

ii) $\{\alpha_0, \ldots, \alpha_k\}$ spans a k-simplex if and only if $U_{\alpha_0} \cap \ldots \cap U_{\alpha_k} \neq \emptyset$.

Definition 2.2. For a point cloud $\mathbb{X} \subset \mathbb{R}^n$ and a real number $\epsilon > 0$, consider the open balls $B_\epsilon(x) = \{y \in \mathbb{R}^n \mid d(x, y) < \epsilon\}$ for $x \in \mathbb{X}$. The nerve of the covering $\{B_\epsilon(x)\}_{x \in \mathbb{X}}$ of $\mathbb{X}$ is called the *Čech complex* attached to $\mathbb{X}$ and ϵ, and will be denoted by $C_\epsilon(\mathbb{X})$.

This construction is useful towards understanding the "shape" of the point cloud $\mathbb{X}$. Indeed we have the following result:

Theorem 2.3 (Nerve Theorem)**.** *Let $\mathcal{U} = \{U_\alpha\}_{\alpha \in I}$ be a covering of a point cloud $\mathbb{X}$ in $\mathbb{R}^n$ such that the intersection of any subcollection of $\mathcal{U}$ is either empty or contractible. Then $\bigcup_{\alpha \in I} U_\alpha$ and the geometric realization of the nerve $N(\mathcal{U})$ are homotopy equivalent.*

We refer to [21, Corollary 4.G.3] for a proof of this result. Since every nonempty intersection of open balls in $\mathbb{R}^n$ is convex and hence contractible, we infer from the Nerve Theorem that the geometric realization of the Čech complex $C_\epsilon(\mathbb{X})$ of a point cloud $\mathbb{X}$ is homotopy equivalent to the union of the collection of open balls $\{B_\epsilon(x)\}_{x \in \mathbb{X}}$ for every ϵ.

2.1.2 Vietoris-Rips complexes

If we wish to compute a Čech complex, i.e., the list of simplices of a Čech complex for a given point cloud $\mathbb{X}$, we will find that it is computationally expensive, in the sense that it requires the storage of various multidimensional simplices. An idea for dealing with this problem is to construct a complex which can be recovered solely from the collection of 1-simplices. This suggests the following variant of the Čech construction, referred to as the *Vietoris-Rips* complex.

Definition 2.4. For a finite point cloud $\mathbb{X}$ and a number $\epsilon \geq 0$, the *Vietoris-Rips* complex attached to $\mathbb{X}$ and ϵ, referred to as $R_\epsilon(\mathbb{X})$, is the simplicial complex with vertex set $\mathbb{X}$ and where a set $\{x_0, \ldots, x_k\} \subseteq \mathbb{X}$ spans a k-simplex if and only if $d(x_i, x_j) \leq \epsilon$ for all $0 \leq i, j \leq k$.

We note that both complexes are closely related. Indeed, we have the following inclusions:

$$C_\epsilon(\mathbb{X}) \subseteq R_{2\epsilon}(\mathbb{X}) \subseteq C_{2\epsilon}(\mathbb{X}).$$

The first inclusion follows from the definitions, and the second one is proved in [3].

We also emphasize the following facts:

- Both the Čech complex and the Vietoris-Rips complex of a point cloud $\mathbb{X}$ in $\mathbb{R}^n$ remain invariant under translations of $\mathbb{X}$ within $\mathbb{R}^n$. More generally, they are preserved by any rigid Euclidean motion, including rotations and symmetries.

- While the Čech complex $C_\epsilon(\mathbb{X})$ involves information about the ambient Euclidean space, the Vietoris-Rips complex $R_\epsilon(\mathbb{X})$ only depends on the table of pairwise distances $d(x_i, x_j)$ between the points of $\mathbb{X}$. This is another reason why Vietoris-Rips complexes are much more often used in practical applications than Čech complexes.

For a data set $\mathbb{X}$, let us illustrate an example of the construction of both complexes. We will write ϵ to be the radius of the balls of the covering of $\mathbb{X}$ derived from the Čech construction, and $\delta = 2\epsilon$ will be the distance parameter for the construction of the Vietoris-Rips complex attached to $\mathbb{X}$.

Example 2.5. Consider $\mathbb{S}^1 = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}$, the unit circle in $\mathbb{R}^2$. Let $\mathbb{X} = \{v_0, v_1, v_2\} \subset \mathbb{S}^1$ be a sample of three points. For each ϵ, we have the attached Čech complex $C_\epsilon(\mathbb{X}) = (\mathbb{X}, \Sigma)$, and Σ is composed of the 1-dimensional simplices $[v_0, v_1]$, $[v_1, v_2]$, $[v_2, v_0]$, and 0-dimensional simplices $[v_0]$, $[v_1]$, $[v_2]$. Example 1.2 in the previous chapter yields the homology groups H_0 and H_1, namely $H_0(C_\epsilon(\mathbb{X})) \cong \mathbb{Z}$ and $H_1(C_\epsilon(\mathbb{X})) \cong \mathbb{Z}$. This captures the fact that $C_\epsilon(\mathbb{X})$ is connected and has one 1-dimensional cycle, which in relation to the underlying space $\mathbb{S}^1$ is a fair topological description.

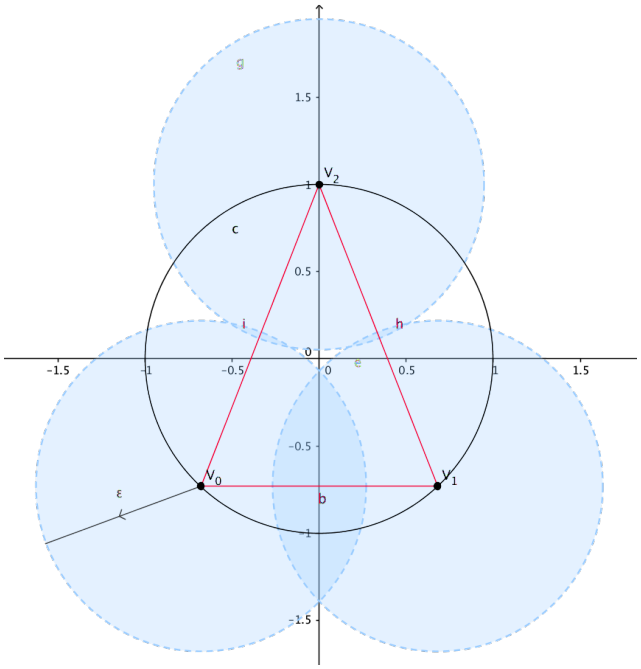


Figure 2.1: In red, Čech complex attached to $\{v_0, v_1, v_2\}$ and ϵ.

On the other hand, we have that the homology group H_1 deriving from the Vietoris-Rips complex associated to the set $\mathbb{X}$ and δ is zero. The Vietoris-Rips complex $R_\delta(\mathbb{X}) = (\mathbb{X}, \Sigma')$ attached to $\mathbb{X}$ and δ has a similar picture to the one in Fig. 2.1, the only difference being that in addition to the simplices in Σ, the set Σ' has an additional 2-dimensional simplex, namely the face $[v_0, v_1, v_2]$. Recall that $[v_0, v_1, v_2]$ forms a 2-simplex if and only if $d(v_i, v_j) \leq \delta$ for all i, j. An appropriate depiction of the Vietoris-Rips complex attached to $\mathbb{X}$ would result in filling the red triangle in Fig. 2.1 (not pictured here). Thus, its corresponding homology groups are $H_0(R_\delta(\mathbb{X})) \cong \mathbb{Z}$ and $H_1(R_\delta(\mathbb{X})) = 0$. In conclusion, in this example the structure of the Vietoris-Rips complex is not able to capture the topology of $\mathbb{S}^1$.

2.2 Persistent homology

In this section we introduce the notion of persistence of homological features, i.e., k-dimensional holes, and we give some introductory definitions that help structure the theory of persistent homology, such as *persistence modules*. We also give standard graphical interpretations of persistence, together with a classification theorem. For a more complete exposition, we refer the interested reader to [24].

Definition 2.6. A *persistence module* is a pair (V, π) where V is a collection $\{V_t\}$ $(t \in \mathbb{R})$ of finite dimensional vector spaces over a field $\mathbb{F}$, and π is a collection $\{\pi_{s,t}\}$ of $\mathbb{F}$-linear maps $\pi_{s,t} \colon V_s \to V_t$ for all $s \leq t$ in $\mathbb{R}$, such that, for $s \leq t \leq r$,

$$\pi_{s,r} = \pi_{t,r} \circ \pi_{s,t}.$$

We say that (V, π) is of *finite type* is there is a finite set $A = \{\alpha_0, \ldots, \alpha_k\} \subset \mathbb{R}$ with $\alpha_0 < \cdots < \alpha_k$ such that $V_t = 0$ for $t < \alpha_0$, and the following holds:

1. For every $\alpha \in A$ there is an $\epsilon_\alpha > 0$ so that if $\alpha \leq s < \alpha + \epsilon_\alpha$ then $\pi_{\alpha,s}$ is an isomorphism and if $\alpha - \epsilon_\alpha < s < \alpha$ then the $\pi_{s,\alpha}$ is not a isomorphism.

2. If x is any real number with $x \notin A$, then there is a $\delta > 0$ such that $\pi_{s,t}$ is an isomorphism if $x - \delta < s \leq t < x + \delta$.

The set $A = \{\alpha_0, \ldots, \alpha_k\}$ is called the *spectrum* of (V, π), and its elements are *spectral points*.

In our notes, we consider collections of Vietoris-Rips complexes $\{R_{\epsilon_j}(\mathbb{X})\}_{j \in I}$ for every point cloud $\mathbb{X}$, as described in Definition 2.4, where the parameters ϵ_j are those for which the homology groups of $R_{\epsilon_j}(\mathbb{X})$ with coefficients in $\mathbb{F}$ suffer changes. Hence we have a filtration $R_{\epsilon_j}(\mathbb{X}) \subset R_{\epsilon_{j+1}}(\mathbb{X})$ as long as $\epsilon_j < \epsilon_{j+1}$. Moreover, we have injective simplicial maps $i_{\epsilon_j, \epsilon_{j+1}} \colon R_{\epsilon_j}(\mathbb{X}) \hookrightarrow R_{\epsilon_{j+1}}(\mathbb{X})$. Taking $V_\epsilon = H_*(R_\epsilon(\mathbb{X}); \mathbb{F})$ and $\pi_{\epsilon_j, \epsilon_{j+1}} = (i_{\epsilon_j, \epsilon_{j+1}})_*$ to be the induced $\mathbb{F}$-linear maps in homology, we obtain a persistence module, which will be referred to as the *Vietoris-Rips module* of $\mathbb{X}$.

We assume that $R_\epsilon(\mathbb{X}) = \emptyset$ if $\epsilon < 0$, from which it follows that $V_\epsilon = 0$ for $\epsilon < 0$. For $\epsilon \geq 0$ we have, by definition,

$$V_\epsilon = \bigoplus_{i=0}^{\infty} H_i(R_\epsilon(\mathbb{X}); \mathbb{F}).$$

Usually, we will only consider Vietoris-Rips modules up to a certain homological dimension n. Thus, to clarify notation, we write

$$V_\epsilon^n = \bigoplus_{i=0}^{n} H_i(R_\epsilon(\mathbb{X}); \mathbb{F}).$$

The coefficient field $\mathbb{F}$ will be omitted from the notation as it is not relevant towards our applications.

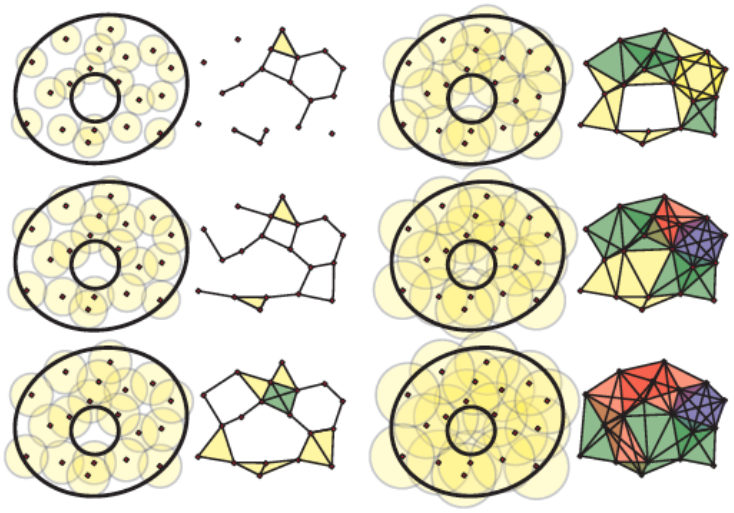


Figure 2.2: Sequence of Vietoris-Rips complexes for a point cloud surrounding an annulus. As ϵ increases holes appear and dissappear. Persistent homology studies the life-span of these holes. Figure adapted from [16].

A *morphism* $f\colon (V,\pi) \to (V^{'},\pi^{'})$ between persistence modules over a field $\mathbb{F}$ is a collection of $\mathbb{F}$-linear maps $f_t\colon V_t \to V_t^{'}$ such that

$$f_t \circ \pi_{s,t} = \pi_{s,t}^{'} \circ f_s \text{ whenever } s \leq t.$$

For a point cloud $\mathbb{X}$, the generators of the k-th homology groups $H_k(R_t(\mathbb{X}))$ for each value of the parameter t can be used as a means of representing or, better said, visualizing persistence in $\mathbb{X}$. In order to make this explicit, we introduce the following notion.

For every interval $I = [a,b) \subset \mathbb{R}$ with $a < b$ or $I = [a,\infty)$, we define a persistence module $\mathbb{F}(I)$ as follows:

$$\mathbb{F}(I)_t = \begin{cases} \mathbb{F} & \text{if } t \in I \\ 0 & \text{otherwise,} \end{cases}$$

with $\pi_{s,t} = \mathrm{id}$ if $s,t \in I$ and $\pi_{s,t} = 0$ otherwise. Such persistence modules are called *interval modules*. Their spectrum is $\{a,b\}$ if $I = [a,b)$, or $\{a\}$ if $I = [a,\infty)$. We can also consider the *direct sum* of persistence modules (V,π), $(V^{'},\pi^{'})$, which results in a persistence module (W,θ), where $W_t = V_t \oplus V_t^{'}$ for all t, and $\theta_{s,t} = \pi_{s,t} \oplus \pi_{s,t}^{'}$ for all s,t. Also, for every positive integer m we denote

$$\mathbb{F}(I)^m = \mathbb{F}(I) \oplus \overset{m}{\cdots} \oplus \mathbb{F}(I),$$

so $\mathbb{F}(I)^m$ also becomes a persistence module.

The following result states that we can decompose a persistence module (V,π) into a direct sum of interval modules.

Theorem 2.7 (Normal Form Theorem)**.** *For every persistence module (V, π) of finite type over a field $\mathbb{F}$, there is a finite collection of intervals $\{I_i\}_{i=1}^N$ with $I_i = [a_i, b_i)$ or $I_i = [a_i, \infty)$ for each i, such that $I_i \neq I_j$ if $i \neq j$, and there is an isomorphism of persistence modules*

$$V \cong \bigoplus_{i=1}^{N} \mathbb{F}(I_i)^{m_i}, \tag{2.1}$$

where $m_1, \ldots, m_N$ are positive integers.

Proof. We are going to give a proof of this theorem, based on the structure theorem for principal ideal domains.

Let R be a principal ideal domain and let M be a finitely generated R-module. Then the following holds:

1. M can be written as a direct sum of R-modules

$$M \cong R/(d_1) \oplus \cdots \oplus R/(d_s) \oplus R^k,$$

   where all d_i's are non-zero and non-units of R, and $d_i | d_j$ for all $i \leq j$.

2. The elements k and $d_1, \ldots, d_s$ are uniquely determined by M.

The summand R^k is referred to as the *free part* of M and the other summands are *torsion* submodules. Now, we want to adapt this result to finitely generated *graded* $\mathbb{F}[t]$-modules, where $\mathbb{F}[t]$ is the polynomial ring on a variable t with coefficients in $\mathbb{F}$.

We say that a module M over a graded commutative ring $R = \oplus_i R_i$ for an index set I is a *graded* module if it is a direct sum of modules $\bigoplus_{i \in I} M_i$ satisfying $R_i M_j \subseteq M_{i+j}$.

If M is a finitely generated $\mathbb{N}$-graded module over the polynomial ring $\mathbb{F}[t]$, for any field $\mathbb{F}$, then

$$M \cong \bigoplus_{i=1}^{n} T^{p_i}\mathbb{F}[t] \oplus \Big(\bigoplus_{j=1}^{m} T^{q_j}\mathbb{F}[t]/(t^{r_j}) \Big) \tag{2.2}$$

for some collection of integers $p_i \geq 0$, $q_j \geq 0$ and $r_j \geq 1$. Here $T^k\mathbb{F}[t]$ denotes an upward shifted graded module, i.e., the graded module $\mathbb{F}[t]$ with a translation of k units in its degree. Moreover, this decomposition is unique up to a permutation of the summands.

The proof of (2.2) is a variation of the standard proof in the non-graded case. Details can be found in [26]. Another useful source is [6, Theorem 2.1].

Let us focus on understanding this result in the context of persistence modules. Let (V, π) be a persistence module with spectrum $A = \{\alpha_0, \ldots, \alpha_k\}$ for $a_0 < \cdots < a_k$. Consider the vector space $V_* = V_{\alpha_0} \oplus \cdots \oplus V_{\alpha_k}$. We define an action of $\mathbb{F}[t]$ on V_* by

$$t \cdot v = \pi_{\alpha_i, \alpha_{i+1}}(v) \text{ if } v \in V_{\alpha_i}, \text{ and } t \cdot v = v \text{ if } v \in V_{\alpha_k}. \tag{2.3}$$

In this way, V_* becomes an $\mathbb{N}$-graded $\mathbb{F}[t]$-module, with V_{α_i} in degree i and V_{α_k} in degrees bigger than or equal to k. Thus, we can apply (2.3) to infer (2.1).

Indeed, applying (2.3) to the persistence module (V, π), we obtain

$$V \cong \bigoplus_{i}^{n} t^{b_i} \cdot \mathbb{F}[t] \oplus \left(\bigoplus_{j}^{m} t^{r_j} \cdot \Big(\mathbb{F}[t]/(t^{s_j}\mathbb{F}[t]) \Big) \right). \tag{2.4}$$

The free portions of (2.4) are in bijective correspondence with those basis vectors which come into existence at parameter b_i and which persist for all future parameter values. The torsional elements correspond to those basis vectors which appear at parameter r_j and disappear at parameter $r_j + s_j$. In other words, there is a bijection between the summands and the corresponding interval modules. Hence, we obtain an equation as in (2.1). □

Another proof for this theorem can be found in [9].

Let us interpret this result in the case of Vietoris-Rips modules. The spectrum of the persistence modules $\mathbb{F}(I_i)$ determines certain intervals $[a_i, b_i)$ where $0 \leq a_i \leq b_i \leq \infty$. So, each interval $[a_i, b_i)$ represents the "life-span" of a homology generator and the integers $m_0, \ldots, m_n$ represent the multiplicity of every interval with respect to the generators. In other words, if s generators are "born" at the same parameter t and "die" at the same parameter $t + \epsilon$, then we say that the interval $I = [t, t + \epsilon)$ has multiplicity $m(I) = s$. We want to remark these last two concepts, namely the *birth* and *death* of homology generators, so we will properly define them.

Definition 2.8. Let (V, π) be a persistence module. We say that a basis vector $v \in V_t$ is *born* at parameter t if, for every $\epsilon > 0$, the map $V_{t-\epsilon} \to V_t$ does not contain v in its image. Similarly, we say that $v \in V_t$ *dies* at parameter $t + s$ if for every $0 < \epsilon < s$ the map $V_t \to V_{t+\epsilon}$ sends v to a nonzero element but v is in the kernel of the map $V_t \to V_{t+s}$.

The notion of birth and death remains as one that can be understood intuitively. Figures 2.3 and 2.4 illustrate the birth and death of k-dimensional homology generators.

One of the main advantages of persistent homology is that we can *visualize* the birth and death of the homology features considered. In the following sections we will describe tools such as *barcodes*, *persistence diagrams* and *persistence landscapes*. There is a clear relationship between them, in the sense that they all store similar information extracted from a data set.

2.3 Persistence barcodes and diagrams

The parameter intervals arising from Theorem 2.7 inspire a visual snapshot of $H_*(R_{\epsilon_i}(\mathbb{X}))$ for a collection $\{\epsilon_i\}_{i\in I}$ in the form of a *barcode*. A barcode is a graphical representation of $H_*(R_{\epsilon_i}(\mathbb{X}))$ as a collection of horizontal line segments in a plane whose horizontal axis corresponds to the values of the parameter and whose vertical axis represents an (arbitrary) ordering of homology generators. Figure 2.4 gives a barcode representation of how the different k-dimensional homology generators appear and disappear as the parameters ϵ_i increase.

Now, for a fixed ϵ_i, we can pair every generator for $H_*(R_{\epsilon_i}(\mathbb{X}))$ with birth-death coordinates (b_i, d_i), and draw what we name a *persistence diagram*. Formally, from (2.1), if we write $I_i = [b_i, d_i)$ we can describe the persistence diagram as the set

$$\mathcal{D} = \{(b_i, d_i) \mid i \in I\} \subset \mathbb{R}^2.$$

It is convenient to add to each persistence diagram all the points of the diagonal $\Delta = \{(b, d) \mid b = d\}$ with infinite multiplicity. If we do so, the persistence diagram $\mathcal{D}$ is referred to as the *decorated persistence diagram*. Moreover, we denote by $\mathcal{D}' = \mathcal{D}\backslash\Delta$ the persistence diagram without the diagonal Δ, also referred to as the *undecorated persistence diagram* [9].

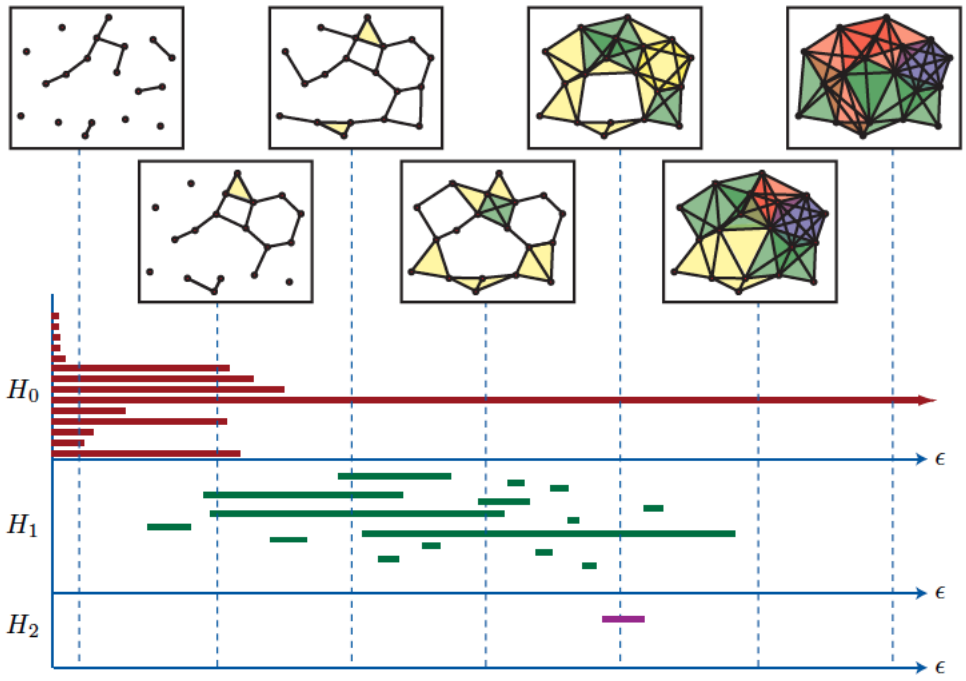


Figure 2.3: In Fig. 2.2 we can visualize a barcode for generators of $H_k(R_\epsilon(\mathbb{X}))$, for $k = 0, 1, 2$ and $\epsilon \geq 0$. By mapping each interval to its endpoints we obtain the corresponding persistence diagram. Figure adapted from [16].

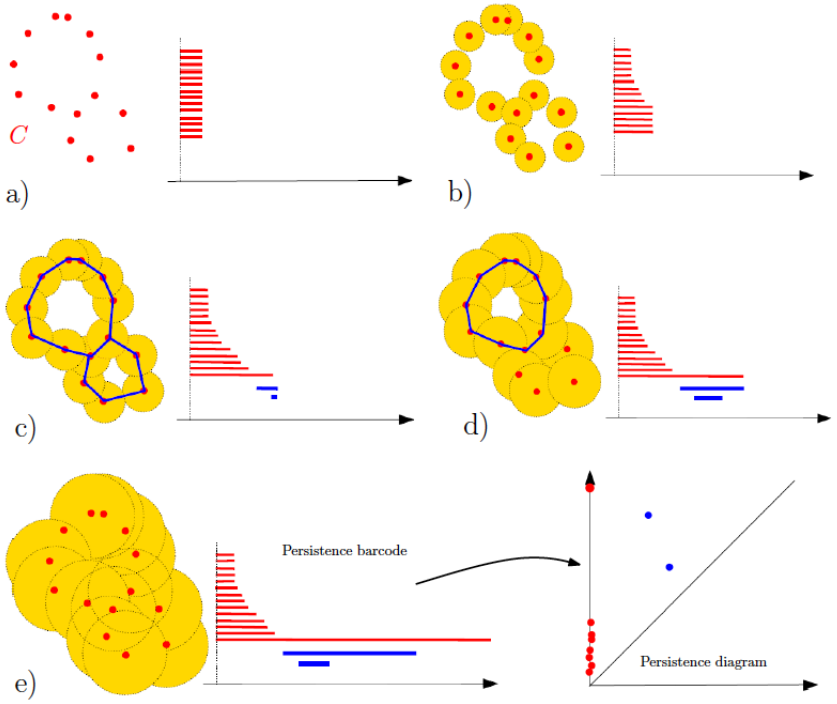


Figure 2.4: Illustration of the Vietoris-Rips scheme with attached barcode representation (details are given in the text). Figure adapted from [8].

An interpretation of Figure 2.4 can be that, for each parameter ϵ, we illustrate the corresponding "inflated" set and its *persistence barcode* on the right. The homology groups in (e) are stable; hence, for the associated parameter ϵ, we illustrate a *final* barcode representation for $H_0(R_\epsilon(\mathbb{X}))$ (red) and $H_1(R_\epsilon(\mathbb{X}))$ (blue). To its right we depict the associated *persistence diagram*, matching the ends of parameter intervals for H_0 and H_1. We note that the parameter intervals for H_0 have the property that all generators are born at $\epsilon = 0$; this is because H_0 measures connectedness and every point can be seen as a generator; thus, when we match intervals of H_0, the associated points on the persistence diagram will have the same x-coordinate, which represents the birth of a generator. Any other barcode associated to H_k (for k greater than 0) yields points $p = (b, d)$ on the persistence diagram with different coordinates, although always $d > b$, and so all points will be above Δ.

2.4 Persistence landscapes

Although the persistence landscape is a particular tool for illustrating persistent homology, it will be our main focus in these notes and we must dedicate a whole section to define some of its properties and the role of statistics in the persistent landscape paradigm. We reference the reader to [1] for a more detailed introduction to persistence landscapes, followed by [2].

Definition 2.9. Suppose given a persistence module (V, π) and let $\mathcal{D} = \{(b_i, d_i)\}_{i \in I}$ be its associated persistence diagram. For each birth-death point (b_i, d_i) in $\mathcal{D}$, we define a piecewise linear continuous function

$$f_{(b_i,d_i)}(x) = \begin{cases} x - b_i & \text{if } b_i < x \leq \frac{b_i+d_i}{2} \\ -x + d_i & \text{if } \frac{b_i+d_i}{2} < x < d_i \\ 0 & \text{otherwise.} \end{cases}$$

Then, the function $\Lambda \colon \mathbb{N} \times \mathbb{R} \longrightarrow \mathbb{R}$ given by

$$\Lambda(k, x) = \text{kmax}\{f_{(b_i,d_i)}(x)\}_{i \in I}$$

is called the *persistence landscape* function associated to the persistence diagram $\mathcal{D}$, where kmax denotes the k-th largest value of a set.

Remark 2.10. By definition, if $k > |I|$, then the value of kmax is zero.

In this definition we followed [8]. For other equivalent definitions of persistence landscapes, we refer to [1, 2]. Formally, a persistence landscape may also be viewed as a sequence of functions $\lambda_1, \lambda_2, \ldots : \mathbb{R} \to \mathbb{R}$, where $\lambda_k(x) = \Lambda(k, x)$ is called the *k-th persistence landscape function* of $\mathcal{D}$. Each function λ_k is piecewise linear with slope either 0, 1, or -1. The critical points of λ_k are those values of x at which the slope changes. The set of critical points of the persistence landscape Λ is the union of the sets of critical points of the functions λ_k.

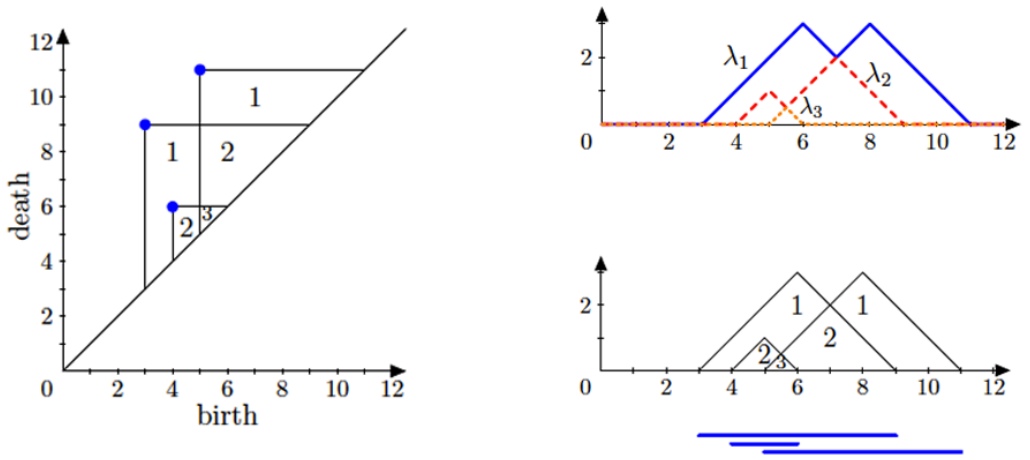


Figure 2.5: On the left, persistence diagram associated to Fig. 1 in [1]. On the right, its corresponding persistence landscape.

An overview of the persistence landscape associated to a given data set is illustrated in Fig. 2.5, top right. We can interpret the k-th persistence landscape $\lambda_k(x)$ as

the value of the largest radius interval centered at x contained in the associated barcode of Λ. This is illustrated in Fig. 2.5, bottom right. Furthermore, for a persistence landscape Λ derived from a persistence diagram $\mathcal{D}$, the k-th persistence landscape function λ_k has the following properties:

1. $\lambda_k(x) \geq 0$ for all x;
2. $\lambda_k(x) \geq \lambda_{k+1}(x)$ for all x.

These follow directly from the definition. Indeed, for every birth-death point p in $\mathcal{D}$, we have the associated piecewise linear function f_p which has image $\mathrm{Im}(f_p) \subseteq \mathbb{R}_{\geq 0}$. Thus, for every k we have that $\mathrm{kmax}\{f_{(b_i,d_i)}(x)\}_{i\in I} \geq 0$. The second property follows directly from the fact that $\lambda_k = \mathrm{kmax}\{f_{(b_i,d_i)}\}_{i\in I} \geq (k+1)\mathrm{max}\{f_{(b_i,d_i)}\}_{i\in I} = \lambda_{k+1}$. For a more detailed exposition of properties of persistence landscapes, we reference the reader to [1, 2].

Choosing to work with persistence landscapes instead of persistence diagrams is not arbitrary. Certainly, the geometry of the space of persistence diagrams makes it hard to work with. In contrast, the space of persistence landscapes can be very nice. For instance, persistence landscapes are elements of a Banach space. Hence working with norms of persistence landscapes is an advantage, as we will see below. But first, let us recall some basic concepts of real functional analysis.

Recall that, for a given measure space $(\mathcal{S}, \mathcal{A}, \mu)$ and a function $f\colon \mathcal{S} \to \mathbb{R}$ defined μ-almost everywhere, one defines, for $1 \leq p < \infty$, the *L^p-norms*

$$\|f\|_p = \Big(\int |f|^p \, d\mu \Big)^{1/p} \text{ and}$$
$$\|f\|_\infty = \sup_{x\in\mathcal{S}} |f(x)| = \inf \{a \mid \mu\{s \in \mathcal{S} : f(s) > a\} = 0\}.$$

Moreover, for $1 \leq p \leq \infty$, we have a Banach space

$$\mathcal{L}^p(\mathcal{S}) = \{f\colon \mathcal{S} \to \mathbb{R} \mid \|f\|_p < \infty\},$$

and define $L^p(\mathcal{S}) = \mathcal{L}^p(\mathcal{S})/\sim$, where $f \sim g$ if $\|f - g\|_p = 0$. Using these concepts, we can define the *persistence landscape norm*, as follows.

Definition 2.11. Let $\Lambda\colon \mathbb{N} \times \mathbb{R} \to \mathbb{R}$ be a persistence landscape function. Suppose that on $\mathbb{N} \times \mathbb{R}$ we use the product of the counting measure on $\mathbb{N}$ and the Lebesgue measure on $\mathbb{R}$. Then, for $1 \leq p < \infty$, we define

$$\|\Lambda\|_p = \sum_{k=1}^{\infty} \|\lambda_k\|_p \tag{2.5}$$

where $\lambda_k(t) = \Lambda(k, t)$, and $\|\lambda_k\|_p$ denotes the standard L^p-norm of λ_k.

Thus, we can endow the space of persistent landscapes with the norm (2.5) and the set of persistence landscapes becomes a subset of the Banach space $L^p(\mathbb{N} \times \mathbb{R})$.

2.4.1 Statistics with landscapes

Here, we discuss briefly some facts about probability in Banach spaces. These will be used in Chapter 4 to prove results on persistence landscapes derived from random data sets with a certain distribution.

Let $\mathcal{B}$ be a real separable Banach space with norm $\|\cdot\|$ and let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $V\colon (\Omega, \mathcal{F}, P) \to \mathcal{B}$ be a Borel random variable with values in $\mathcal{B}$. The composite

$$\|V\|\colon \Omega \xrightarrow{V} \mathcal{B} \xrightarrow{\|\cdot\|} \mathbb{R}$$

is a real-valued random variable. Let $\mathcal{B}^*$ denote the topological dual space of continuous linear real-valued functions on $\mathcal{B}$. For $f \in \mathcal{B}^*$, the composite

$$f(V)\colon \Omega \xrightarrow{V} \mathcal{B} \xrightarrow{f} \mathbb{R}$$

is also a real-valued random variable. If we denote $Y = f(V)$, the *mean* or *expected value* of Y is given by

$$E(Y) = \int Y\, dP = \int_\Omega Y(\omega)\, dP(\omega).$$

We call an element $E(V) \in \mathcal{B}$ the *Pettis integral* of V if $E(f(V)) = f(E(V))$ for all $f \in \mathcal{B}^*$. Now, for a sequence $(Y_n)_{n\in\mathbb{N}}$ of $\mathcal{B}$-valued random variables, we say that $(Y_n)_{n\in\mathbb{N}}$ *converges almost surely* to a $\mathcal{B}$-random variable Y if there exists a null probability set $N \in \mathcal{F}$ such that

$$\lim_{n\to\infty} Y_n(\omega) = Y(\omega) \text{ if } \omega \notin N.$$

Proposition 2.12. *Let V be a $\mathcal{B}$-valued random variable. If $E(\|V\|) < \infty$, then V has a Pettis integral and $\|E(V)\| \leq E(\|V\|)$.*

Theorem 2.13 (Strong law of large numbers)**.** *Let $V_1, \ldots, V_n$ be independent and identically distributed $\mathcal{B}$-valued random variables, and $S_n = V_1 + \cdots + V_n$. We have that $\frac{1}{n}S_n \to E(Y)$ almost surely if and only if $E(\|V\|) < \infty$.*

Details for these results are given in [22].

For a given probability space $(\Omega, \mathcal{F}, P)$, let X be a multivariate random variable with corresponding persistence landscape Λ. By this we mean that, for $\omega \in \Omega$, $X(\omega)$ is a data set and $\Lambda(\omega)$ is the corresponding persistence landscape derived from the homology generators of the associated Vietoris-Rips module of $X(\omega)$. In this sense, we interpret the persistence landscape as a Banach space valued random variable $\Lambda(\omega)\colon \Omega \to L^p(\mathbb{N}\times\mathbb{R})$.

Let $X_1, \ldots, X_n$ be independent and identically distributed copies of X, and let $\Lambda_1, \ldots, \Lambda_n$ be the corresponding persistence landscapes. Using the vector structure of $L^p(\mathbb{N}\times\mathbb{R})$, Bubenik defines in [1] the *mean landscape* $\overline{\Lambda}$, as given by the pointwise mean

$$\overline{\Lambda}(k,t) = \frac{1}{n}\sum_{i=1}^{n} \Lambda_i(k,t). \tag{2.6}$$

If $B_1, \ldots, B_n$ are the barcodes associated to the persistence landscapes $\Lambda_1, \ldots, \Lambda_n$, then, for $k \in \mathbb{N}$ and $t \in \mathbb{R}$, the value of the mean landscape $\overline{\Lambda}(k,t)$ can be interpreted as the average value of the largest radius interval centered at t that is contained in k intervals in the barcodes $B_1, \ldots, B_n$.

Now, if we consider Λ to be a $L^p(\mathbb{N} \times \mathbb{R})$-valued random variable, the composite $\|\Lambda(\omega)\|$ is a real valued random variable and we can rewrite Theorem 2.13 to be understood in terms of persistence landscapes.

Theorem 2.14. *Let $\Lambda_1, \ldots, \Lambda_n$ be independent and identically distributed copies of Λ. Then, $\overline{\Lambda} \to E(\Lambda)$ almost surely if and only if $E(\|\Lambda\|) < \infty$.*

Proof. Direct from Theorem 2.13. □

3. Topological Data Analysis of finite data sets

In this chapter we describe our method, analogous to the method used in [19], whose purpose is to analyze some given mixture of data sets using TDA; although very promising, this method still remains to be thoroughly tested.

3.1 Method

Here, we state briefly the structure of our method. This procedure is done using R as our main programming language. The main R-packages we use are "TDA" and "fda.usc", for computing persistent homology and calculating L^p norms of persistence landscapes, respectively. Regarding the former, we refer the reader to [15] for a full description of this package. Here, we do not distinguish the method for any particular p, moreover we assume the choice of p to be dependent on the application itself.

We will describe this method using a step-by-step structure, as follows:

1. *Embedding of a collection of data sets.*

2. *Studying the persistent homology of point clouds.*

3. *L^p norm process.*

3.2 Embedding of a data set

Our method requires an *embedding* of a finite collection of data sets $\{\mathcal{S}_i\}_{i\in I}$ into a space X. We note in advance that further in these notes we will be referring to the collection $\{S_i\}_{i\in I}$ as a family of discrete time series. With this in mind we describe this procedure to be applicable to any type of data sets.

Given a finite collection $\{S_i\}_{i\in I}$, we define an embedding, namely *mixture embedding*, where we wish to embed a finite number $N = |I|$ of data sets into $\mathbb{R}^N$. Here, we define a data set to be a function $S_i \colon A \subset \mathbb{N} \to \mathbb{R}$ for every $i \in I$. This mixture embedding is done using a sliding window technique, which generates point clouds of a certain window-size w, i.e., we generate point clouds of w points. Furthermore,

for a fixed $t \in A$ and a window-size w, we have a point cloud $\mathbb{X}$ in $\mathbb{R}^N$ given by

$$SW_{w,N}f(t) = \begin{bmatrix} f(t) \\ f(t+1) \\ \vdots \\ f(t+w-1) \end{bmatrix}$$

where $f\colon A \to \mathbb{R}^N$, $f(t) = (S_1(t), \ldots, S_N(t))$. Thus, every point cloud is given by a $w \times N$ matrix, where the number of columns N is the number of data sets N involved in our analysis, and the number of rows w is the size of the sliding window. A more detailed description of this method can be found in [20].

3.3 Studying the persistent homology of point clouds

For every point cloud $\mathbb{X}$ generated, we use the R-package "TDA" to compute its persistent homology. To do this, we use the function "ripsDiag", which returns, in particular, the corresponding persistence diagram up to a chosen homological dimension. Recall that we can consider the persistence module V_ϵ^n to be the Vietoris-Rips module corresponding to a certain homological dimension n. Formally, the function "ripsDiag" computes the persistent homology of the Vietoris-Rips module

$$V_\epsilon^n = \bigoplus_{i=0}^{n} H_i(R_\epsilon(\mathbb{X})),$$

ranging the scale parameter ϵ from 0 to a chosen maximum. The algorithm implemented by this package constructs the simplices using the Vietoris-Rips complex. This results in faster computations in comparisson to computing the simplices using the Čech construction. One of the outputs of this function is the encapsulation of persistent homology into its corresponding persistence diagram. Subsequently, for every stored persistence diagram we use the function "landscape" from the R-package "TDA" to compute its corresponding k-th persistence landscape.

3.4 L^p norm process

The function "landscape" receives as inputs a persistence diagram, a number k which determines which k-th persistence landscape we choose to evaluate, the dimension d of the topological features we want to consider (in our case we usually choose $d = 1$), and in addition we also choose the length l of a sequence which will determine the number of values of λ_k that the function returns. The output is a numeric $l \times 1$ matrix containing the values of the evaluated k-th landscape function.

Our method consists of computing the L^p norms for the persistence landscapes Λ derived from the point cloud $\mathbb{X}$. To do this, we need to determine the sum of L^p norms for every k-th persistence landscape function of Λ. In practice, for every resulting

table of values corresponding to a λ_k, we will use the package "fda.usc", which has the tools to approximate the L^p norm of our discretized function λ_k. In addition, if our window size is small enough, we can reduce the infinite sum $\sum_{k=1}^{\infty} \|\lambda_k\|_p$ to a very manageable size, since for a finite number of points in the persistence diagram the sum will always be finite, i.e., we can choose a fixed number of k-persistence landscape functions to compute the norm $\|\Lambda\|_p$.

The last step in our procedure consists of illustrating a time series or process to depict the L^p norm fluctuation. Once we have built our process, we can study its statistical properties, i.e., variance, spectral density, lags, etc. We note that, to ease computation, in our examples we used $p = 1$ for computing the norms to quantify the combined data sets behavior, though this can be done using any p-norm, as defined in (2.5).

```
###########INITIALIZING
###########
Norm1<-matrix(0,sdit,MC)
Norm2<-matrix(0,sdit,MC)
mc <- txtProgressBar(min = 0, max = MC, style = 3)
###########
###########MAIN LOOP
###########
for (i in 1:sdit){
  for(s in 1:MC){
    x1<-rnorm(n,0,sd=i)
    x2<-rnorm(n,0,sd=i)
    x3<-rnorm(n,0,sd=i)
    x4<-rnorm(n,0,sd=i)
    CoordMatrix<-cbind(x1,x2,x3,x4)
    max_dist<-max(dist(CoordMatrix))
    #Persistence diagram & landscape
    RipsList<-ripsDiag(CoordMatrix,maxdimension = 1,maxscale = max_dist,
                       location = TRUE,library = "DIONYSUS")
    RipsList<-RipsList[["diagram"]]
    tseq = seq(0, max_dist, length = 1000)
    for (KK_ in 1:KK_max ) {
      LandList<-landscape(RipsList,dimension = 1,KK = KK_,tseq)
      Norm1[i,s] = norm.fdata(fdata(LandList,tseq),lp=1) + Norm1[i,s]
      Norm2[i,s] = norm.fdata(fdata(LandList,tseq),lp=2) + Norm2[i,s]
    }
    setTxtProgressBar(mc,s)
  }
  print("number of simulations done:")
  print(i)
}
Normlist1=rowMeans(Norm1)
Normlist2=rowMeans(Norm2)
```

Figure 3.1: R-script drawn from a test on independent Normal distributed data sets. Details are given in Chapter 4.

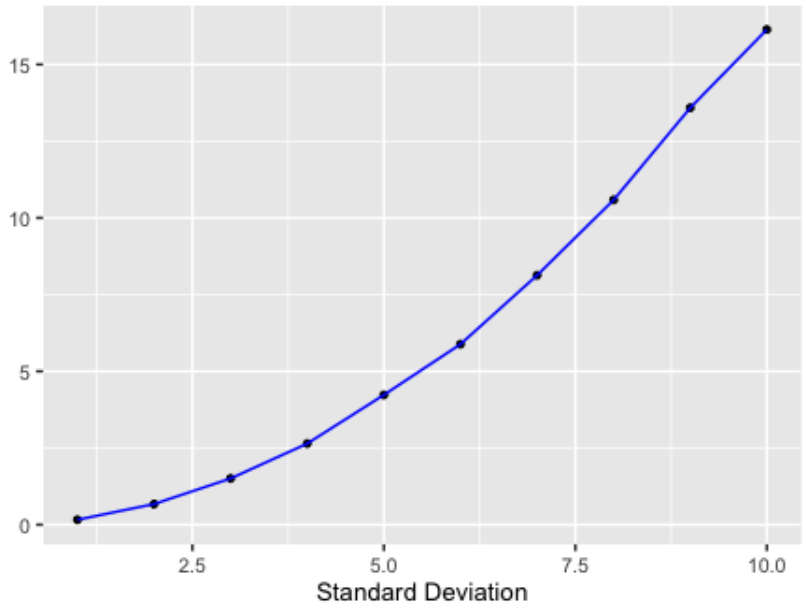
4. Testing on synthetic time series

From the practical perspective, we will test whether a growing variance of a data set born from trajectories of random distributed variables leads to an increase in values of L^1 persistence landscape norms. This is done as a means of verifying the accountability of the method proposed by Guidea and Katz [19], as they did also. Given we want to analyze financial time series, these experiments are done naturally, in the sense that as we increase the variance, we are trying to mimic a more "heated" state of the market. In addition, we prove some of the smaller results obtained experimentally in that paper and, thus, adding to the credit that persistent homology can be used as a means of studying the shape of abstract data.

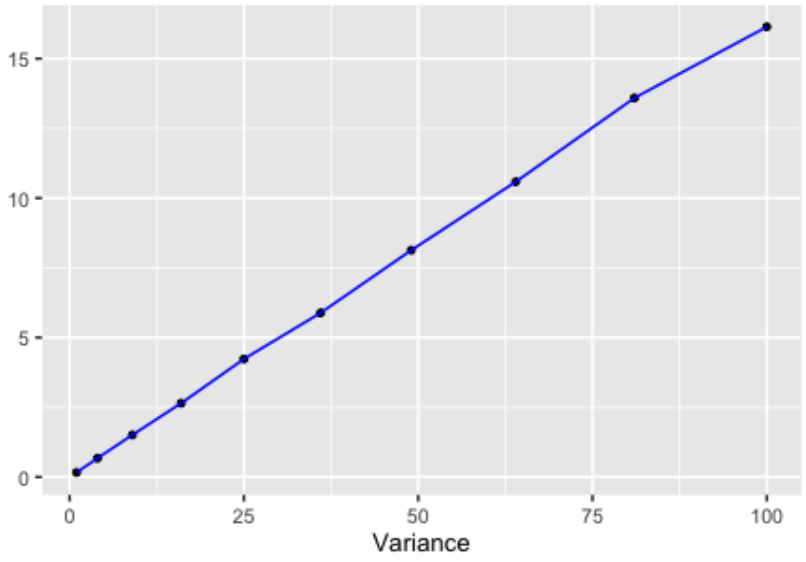
The empirical tests are done using *Monte Carlo methods*, where we perform multiple random samplings from a random variable to obtain some type of deterministic result.

4.1 Normal distributed variables with a growing variance

Closely following [19], we test the method to check the effect of a growing variance on a data set born from normal distributed variables. We start by generating four independent series born from the drawing of 50 trajectories from normal distributed variables $\mathcal{N}(0, \sigma^2)$. The idea is that the values of σ are considered in an increasing order from 1 to 10, and for every σ, we embed our series in $\mathbb{R}^4$, obtaining a 50-point data set. Now, for every point cloud born from a fixed σ we apply the method discussed and obtain the corresponding L^1 norm. Thus, these simulations are repeated various times and at the end of every realization for a given σ, we compute the mean of values of the obtained norms. We run a total of 10 realizations as described and analyze the results graphically, as seen in Figure 4.1.



(a)



(b)

Figure 4.1: Plots of Monte Carlo simulations, 400 iterations, derived from independent Normal distributed data sets. Figure (b) shows of the dependency of the L^1 norm on the growing standard deviation and (a) shows the dependency on the corresponding growing variance.

Noticeably, as σ increases, there is quadratic dependency of the L^1 norm with respect to the standard deviation σ. This implies that, as we can observe, there is associated linear growth of L^1 with respect to σ^2.

In other words, for a fixed number of simulations N, if we define $\Lambda^i_{\sigma^2}$ to be the persistence landscape derived from the i-th simulation for a given σ^2, we have that

$$\frac{1}{N}\sum_{i=1}^{N}\|\Lambda^i_{\sigma^2}\|_1 \approx \sigma^2 \cdot \frac{1}{N}\sum_{i=1}^{N}\|\Lambda^i_1\|_1,$$

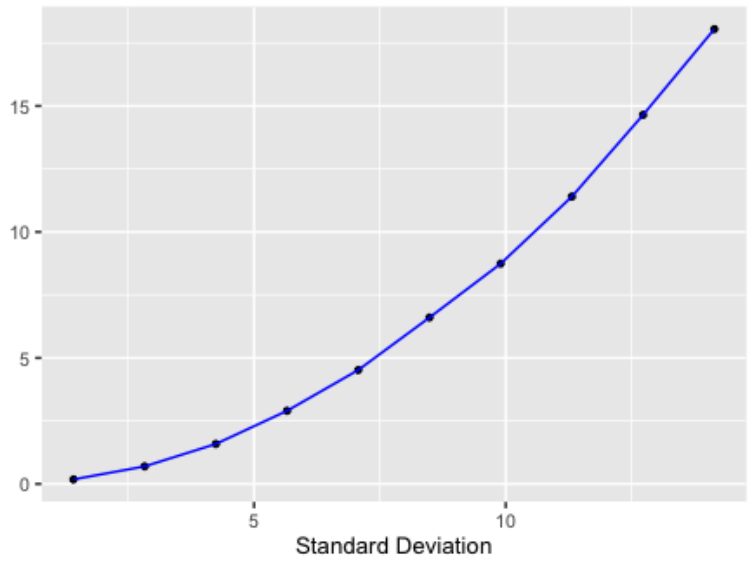
where $\|\Lambda^i_{\sigma^2}\|_1$ denotes the L^1 norm of the persistence landscape. So, as was observed in [19], for a growing number of iterations, the mean of norms of persistence landscapes converges towards a linear increasing function of the average variance of the distributions.

4.2 Scaled Student t-distributed variables

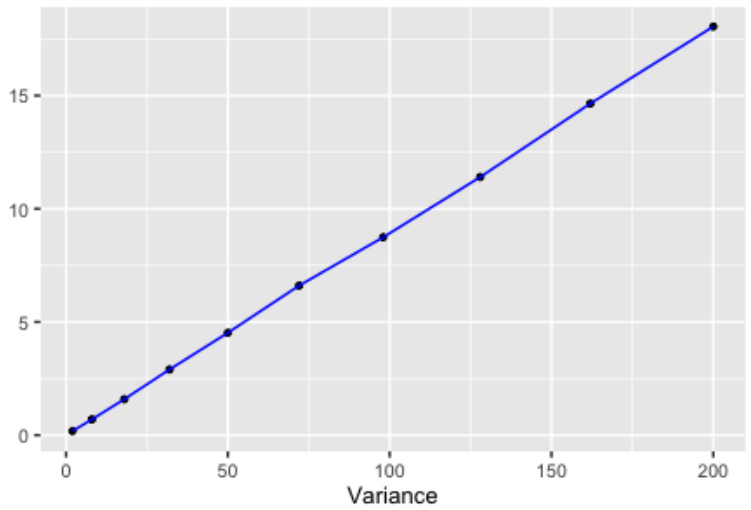
As we will see further in these notes, the Scaled t-distribution is considered to be a more accurate distribution to model the returns of financial markets. Thus, it makes sense to test whether persistence homology is accurately able to measure a growing volatility in a Scaled t-distributed variable.

First, let us start by describing a variable following a Student t-distribution. We say that T follows a Student t-distribution with ν degrees of freedom if it arises when estimating the mean of a normally distributed population when the sample size is $n = \nu + 1$ and the population's standard deviation is unknown. Its shape is bell curved, like that of the normal distribution, but it is characterized for having "heavy tails", informally this means that trajectories drawn from T are more prone to fall far from the mean. Now, if we define the random variable $X = \hat{\mu} + \hat{\sigma}T$, we have that X follows a *scaled Student* t-distribution. Moreover, this family is multiparametric, meaning the distribution of X depends on multiple parameters, namely its degrees of freedom ν, the location parameter $\hat{\mu}$ and the scaling factor $\hat{\sigma}$.

In regards to our experiment, we test whether a growing *scaling factor* for a fixed ν and location parameter $\hat{\mu} = 0$, describes a deterministic behavior for values of L^1 norms of the corresponding persistence landscapes. Aside from changing the distributions, our test is done analogously as in the previous section. In this case, we grow the scaling factor $\hat{\sigma}$ from 1 to 20 for a fixed number of degrees of freedom $\nu = 4$. We note that the choice of ν is somewhat arbitrary, as our goal is to test the behavior of L^1 norms in regards to a scaling factor. We also observe that, for $\nu < 2$ the variance is not defined, since $\mathbb{V}(X) = \hat{\sigma}\nu/(\nu - 2)$; hence we carried out tests with $\nu > 2$. Computations done with $\nu < 2$ have not resulted deterministic in value, this fact seems intuitively plausible, since for undefined variance the random nature of the variable is too high to generate predictable data sets. On the other hand and interestingly enough, *for undefined variance there still exists a real value for L^1 persistence landscape norms.*



(a)



(b)

Figure 4.2: Plots of a Monte Carlo simulation, 400 iterations, derived from independent scaled Student-t distributed data sets. For a fixed number of degrees of freedom, Figure (b) shows of the dependency of the L^1 norm on the growing standard deviation and (a) shows the dependency on the corresponding growing variance.

4.3 Some properties of persistence landscapes

Our goal is to formalize the results obtained in the previous tests, and see if we can characterize this behavior in general. We derive a theorem that, under some conditions, describes well the behavior of the expected L^1 norm of a persistence landscape constructed as in the previous tests. We note that, for this result to hold, the persistence module must be derived from the Vietoris-Rips construction.

Theorem 4.1. *Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\mathbb{X}$ denote a multivariate random variable such that $\mathbb{X}(\omega) = \{X^1(\omega), \dots, X^N(\omega)\}$ describes an N-point data set in $\mathbb{R}^d$, where $X^i \colon \Omega \to \mathbb{R}^d$. So, for $1 \leq i \leq N$, X^i is a multivariate random variable with distribution $D(\mu, \mathrm{Id} \cdot \sigma^2)$. Assume that D has the following properties:*

i) *D is symmetric;*

ii) *$\sigma^2 \in [0, \infty)$;*

iii) *for a certain scaling factor h, $hD(\mu, \sigma^2) \sim D(\mu, h^2\sigma^2)$.*

Then,

$$\mathbb{E}(\|\Lambda_{h^2\sigma^2}(\omega)\|_1) = h^2 \cdot \mathbb{E}(\|\Lambda_{\sigma^2}(\omega)\|_1).$$

Here, we consider the persistence landscape Λ_{σ^2} to be a Banach space valued random variable as described in Chapter 2, born from the data set $\mathbb{X}(\omega)$.

In order to prove this result, we will have to work on some useful propositions that will greatly simplify our explanation. From this point on, we assume $\mathbb{X}$ to be a finite point cloud in $\mathbb{R}^d$ as described in the theorem. Moreover, we note that every point X^i of $\mathbb{X}$ can be written as $X^{(i)} = (X^i_1, \dots, X^i_d)$, where for every $1 \leq j \leq d$, $X^i_j \sim D(\mu, \sigma^2)$. In other words, the coordinates of every point are also independent and identically distributed random variables, and when we write $\mathbb{V}(X) = \sigma^2$, we are denoting the variance of X^i_j.

For a given landscape Λ, we know that its L^p norm is defined by $\|\Lambda\|_p = \sum_{k=1}^{\infty} \|\lambda_k\|_p$, where λ_k is the k-th persistence landscape function of Λ. It would seem obvious that, for a finite data set, this sum is finite. Indeed, to prove this we must show that for t large enough, the terms $\|\lambda_k\|_p$ are zero for all $k \geq t$. First, let us consider the associated persistence diagram $\mathcal{D}^* = \{(b_i, d_i) \mid i \in I\} \setminus \Delta$. If $|I| < \infty$, then we have that for a certain value t such that $k > t \geq |I|$, $\Lambda(k, x) = \mathrm{kmax}\{f_{(b_i,d_i)}(x)\}_{i \in I} = 0$. In other words, if the number of points of the persistence diagram $\mathcal{D}^*$ is finite, then there exists t such that $\lambda_k = 0$ for all $k \geq t$ and hence, $\|\lambda_k\|_p = 0$. Now, from Theorem 2.7, we have that the number of points of $\mathcal{D}^*$ is directly related to the number of k-th homology generators of the persistence module, and we deduce that for a finite number of points in a data set then the number of points in the corresponding persistence diagram is also finite.

Now it would be nice to prove that, with the assumptions made on $\mathbb{X}$, the expected value $\mathbb{E}(\|\Lambda\|_1)$ is finite. To do this, we need the following result.

Proposition 4.2. *Let Λ denote the persistence landscape derived from $\mathbb{X}$. Then,*

$$\mathbb{E}(\|\lambda_1\|_1) \leq N \cdot d \cdot \sigma^2,$$

where N denotes the number of points of $\mathbb{X}$, d is the dimension of the embedding space $\mathbb{R}^d$, and σ is the standard deviation of the distribution associated to $\mathbb{X}$.

Proof. Let $\mathcal{D}^*$ denote the undecorated persistence diagram associated to Λ. From previous arguments we have that $|\mathcal{D}^*| < \infty$. Thus, there are a finite number of points $p_i = (b_i, d_i) \in \mathcal{D}^*$ so that we can define $\epsilon_b = \min_s\{b_s\}$ and $\epsilon_d = \max_l\{d_l\}$ to be the minimum parameter at which a generator is born and the maximum parameter at which a generator dies, respectively. Now, we consider the function

$$\tilde{f}(x) := f_{(\epsilon_b,\epsilon_d)}(x) = \begin{cases} x - \epsilon_b & \text{if } x \in (\epsilon_b, \frac{\epsilon_b+\epsilon_d}{2}] \\ -x + \epsilon_d & \text{if } x \in (\frac{\epsilon_b+\epsilon_d}{2}, \epsilon_d). \end{cases}$$

Recall that in Chapter 2 we defined the function f_{p_i} for every p_i so that $\lambda_k = \text{kmax}\{f_{p_i}\}_{i\in I}$. Hence, it is easily deduced that $\text{dom}(\lambda_1) = \text{dom}(\tilde{f}) =: S \subset \mathbb{R}$, since $\lambda_1 = \max\{f_{p_i}\}_{i\in I}$. Moreover, $\tilde{f}(x) \geq \lambda_1(x)$ for all $x \in S$. Taking integrals on both sides,

$$\int_S \tilde{f}(x)\, dx \geq \int_S \lambda_1(x)\, dx,$$

since $\tilde{f}(x) \geq \lambda_1(x) \geq 0$. Subsequently, we deduce the following

$$\|\tilde{f}\|_1 \overset{(1)}{=} \frac{(\epsilon_d - \epsilon_b)^2}{4} \geq \|\lambda_1\|_1. \tag{4.1}$$

Equality (1) results from the fact that we are computing the area of the triangle of basis $(\epsilon_b - \epsilon_d)$ and height $(\frac{\epsilon_d-\epsilon_d}{2})$. Now, since Vietoris-Rips complexes are invariant under translations, we may assume that the point cloud $\mathbb{X}$ is centered at 0. We have the following chain of inequalities:

$$\|\lambda_1\|_1 \leq \frac{(\epsilon_d - \epsilon_b)^2}{4} \overset{(2)}{\leq} \max_{1\leq i\leq N}(d^2(X^i, 0)) \overset{(3)}{=} \max_{1\leq i\leq N}\Big(\sum_{j=1}^{d} X_j^{i^2}\Big).$$

Inequality (2) comes directly from the fact that $(\epsilon_d - \epsilon_b) \leq 2 \cdot \max_{1\leq i\leq N}(d(X^i, 0)) = r$. In other words, the distance between two points of $\mathbb{X}$ inside the disc $D(0, r) = \{x \in \mathbb{R}^d \mid d(x, 0) \leq r\}$ will never surpass $2r$. For (3), we compute the vector norm $\|\max_{1\leq i\leq N}(d^2(X^i, 0))\|_2^2$. Now, taking expected values on $\|\lambda_1\|_1$ and $\max_{1\leq i\leq N}\big(\sum_{j=1}^{d} X_j^{i^2}\big)$,

$$\mathbb{E}(\|\lambda_1\|_1) \leq \mathbb{E}\Big(\max_{1\leq i\leq N}\Big(\sum_{j=1}^{d} X_j^{i^2}\Big)\Big) \leq \mathbb{E}\Big(\sum_{i=1}^{N}\sum_{j=1}^{d} X_j^{i^2}\Big) \leq N \cdot \sum_{j=1}^{d} \mathbb{E}(X^2) \overset{(4)}{=} N \cdot d \cdot \sigma^2,$$

where (4) comes from the fact that $\mathbb{V}(X) = \mathbb{E}(X^2) + \mathbb{E}(X)^2$. Since $\mathbb{X}$ is centered at 0, we have that $\mathbb{E}(X) = 0$ and $\mathbb{V}(X) = \sigma^2$. □

Given the propositions above, the next step is to prove that the expected value of the L^1 norm of a persistence landscape derived from $\mathbb{X}$ is finite. We came up with the following result:

Theorem 4.3. *Let $\mathbb{X}$ denote a multivariate random variable with distribution D such that $\mathbb{X}(\omega)$ describes an N-point data set in $\mathbb{R}^d$. Assume D has the following properties:*

i) *D is symmetric;*

ii) *$\sigma^2 \in [0, \infty)$;*

iii) *for a certain scaling factor h, $hD(\mu, \sigma^2) \sim D(\mu, h^2\sigma^2)$.*

Let Λ be the persistence landscape derived from $\mathbb{X}$. Then,

$$\mathbb{E}(\|\Lambda\|_1) \leq N^2 \cdot d \cdot \sigma^2.$$

Proof. From previous deductions, we have that $\sum_{i=1}^{N} \|\lambda_i\|$ has a finite number of terms N, and from Proposition 4.2, $\mathbb{E}(\|\lambda_1\|_1) < N \cdot d \cdot \sigma^2$. So, we can write the following chain of inequalities:

$$\mathbb{E}(\|\Lambda\|_1) = \mathbb{E}\big(\sum_{i=1}^{N} \|\lambda_i\|_1\big) = \sum_{i=1}^{N} \mathbb{E}(\|\lambda_i\|_1) \overset{(1)}{\leq} N \cdot \mathbb{E}(\|\lambda_1\|_1) \leq N^2 \cdot d \cdot \mathbb{V}(X).$$

Inequality (1) is due to the property $\lambda_k \geq \lambda_{k+1}$ mentioned in Chapter 2. □

Our last step before proving Theorem 4.1 involves studying what results from scaling a point cloud, in regards to its associated persistence landscape norm $\|\Lambda\|_1$.

Proposition 4.4. *Let X denote a finite point cloud in $\mathbb{R}^d$ with associated persistence landscape Λ_X, and $H\colon \mathbb{R}^d \to \mathbb{R}^d$ be a homotecy such that $H(X) = h \cdot X =: X^{'}$. Then,*

$$\|\Lambda_{X'}\|_1 = h^2 \cdot \|\Lambda_X\|_1. \tag{4.2}$$

Proof. Consider $\mathcal{D}$, $\mathcal{D}^{'}$ to be the corresponding persistence diagrams for $X, X^{'}$, respectively. Then, for every $i \in I$, $P_i \in \mathcal{D}$ becomes $hP_i \in \mathcal{D}^{'}$. Indeed, for any k-simplex $\sigma = [p_0, \ldots, p_k]$ in the Vietoris-Rips complex attached to X at parameter α, it has to be so that $d(p_i, p_j) < \alpha$ for all i, j. So, when we apply H we have that $\sigma = [p_0, \ldots, p_k]$ becomes $\sigma^{'} = [hp_0, \ldots, hp_k]$ and if σ is formed at α, then $\sigma^{'}$ is formed at $h \cdot \alpha$. Thus, $|\mathcal{D}| = |\mathcal{D}^{'}|$ and every point in $\mathcal{D}$ is scaled by h in $\mathcal{D}^{'}$. Furthermore, for every f_{P_i} associated to the persistence landscape of X, we have that $f_{P_i} \mapsto f_{hP_i}$ and so, Λ_X becomes $h\Lambda_X$. Now, for a fixed k, recall that we denote $\Lambda_X(k, x) = \lambda_k(x)$ to be the k-th persistence landscape function of Λ_X, so we can write $\Lambda_{X'}(k, x) := h\Lambda_X(k, x) = \lambda_k^{'}(x)$. Moreover, we have the associated domains in $\mathbb{R}^2$, $\Omega_k = \{(x, y) \mid x \in \mathrm{dom}(\lambda_k),\ \ y \leq \lambda_k(x)\}$ and for $X^{'}$, the new scaled domain is

$h \cdot \Omega_k =: \Omega_k' = \{(x, y) \mid x \in \mathrm{dom}(\lambda_k'),\ \ y \leq \lambda_k'(x)\}$. Since the L^1 norm measures the area, we can write

$$\|\Lambda_{X'}\|_1 = \int_{\Omega_k'} 1 \cdot d(x, y) = h^2 \cdot \int_{\Omega_k} 1 \cdot d(s, t) = h^2 \cdot \|\Lambda_X\|_1.$$

We note that this relation results from a simple change of variables,

$$x = h \cdot s \to dx = h \cdot ds$$
$$y = h \cdot t \to dy = h \cdot dt.$$

This completes the proof. □

Given these results, we can begin the proof of Theorem 4.1:

Proof. Theorem 4.1. Let H denote a homotecy such that $H \colon \mathbb{R}^d \to \mathbb{R}^d$, so $\mathbb{X} \mapsto h \cdot \mathbb{X}$. From the assumptions made, we can safely deduce that for every point $X^i \in \mathbb{X}$ we have $H(X^i) = h \cdot X^i \sim D(\mu, h^2\sigma^2)$. From this point forward we will be using the following notation terms indistinguishably: $\Lambda_{\mathbb{X}}(\omega) = \Lambda_{\sigma^2}(\omega)$ and $\Lambda_{h\cdot\mathbb{X}}(\omega) = \Lambda_{h^2\sigma^2}(\omega)$.

Let $\mathbb{X}_1, \ldots, \mathbb{X}_n$ denote independent identically distributed copies of $\mathbb{X}$, and let $\Lambda_{\mathbb{X}}^1, \ldots, \Lambda_{\mathbb{X}}^n$ be the corresponding persistence landscapes. Now, without losing generality, we can assume $\mu = 0$. Proposition 4.4 implies that

$$\frac{1}{n}\sum_{i=1}^{n} \|\Lambda_{h^2\sigma^2}^i(\omega)\|_1 = h^2 \frac{1}{n}\sum_{i=1}^{n} \|\Lambda_{\sigma^2}^i(\omega)\|_1, \tag{4.3}$$

since for every landscape $\Lambda_{\sigma^2}^i$ we have the corresponding *scaled* persistence landscape $\Lambda_{h^2\sigma^2}^i$. From Theorem 4.3 we have that the expected values $\mathbb{E}(\|\Lambda_{\sigma^2}^i\|_1)$, $\mathbb{E}(\|\Lambda_{h^2\sigma^2}^i\|_1)$ are finite. Hence, we can apply Theorem 2.13 (Strong Law of Large Numbers), and we obtain

$$\frac{1}{n}\sum_{i=1}^{n} \|\Lambda_{h^2\sigma^2}^i(\omega)\|_1 \longrightarrow \mathbb{E}(\|\Lambda_{h^2\sigma^2}\|_1) \text{ a.s.}$$
$$h^2 \frac{1}{n}\sum_{i=1}^{n} \|\Lambda_{\sigma^2}^i(\omega)\|_1 \longrightarrow h^2 \cdot \mathbb{E}(\|\Lambda_{\sigma^2}\|_1) \text{ a.s.}$$

So, taking limits in (4.3), we obtain

$$\mathbb{E}(\|\Lambda_{h^2\sigma^2}\|_1) = h^2 \cdot \mathbb{E}(\|\Lambda_{\sigma^2}\|_1),$$

as we claimed. □

5. Modeling financial markets and empirical analysis on European stock indices

Many theories have been written over the years about modeling the underlying distribution (if there is any) of stock price returns. Osborne's research in [23], together with others, led to believe that equity prices follow a geometric Brownian motion. If we consider small windows in time, i.e., days, weeks, months, this model implies that the distribution of share price changes should follow a normal distribution. Although, it is widely accepted that share price changes are not normally distributed; a modern approach is considering the Student t-distribution to model the behavior of price changes, as seen in [25]. In this chapter, we will introduce some basic concepts on Brownian motion and it's derived financial models. From a practical perspective, we will describe a new type of econometric analysis on various stock price indices. This procedure closely follows [19] and is, in fact, an attempt to replicate the results obtained in that paper.

5.1 Some financial models

In this section, we study some basic applications of *financial modeling*. Financial markets are best understood as an abstract (simplified) representation of any type of investment, where one can understand results derived from a synthetic deterministic nature and try to extrapolate results to the real world.

Modeling financial markets usually makes sense in terms of stochastic processes. In our case, we are working with models that describe the logarithmic returns of a certain price evolution, so we feel it is necessary to properly define this notion.

Definition 5.1. A *stochastic process* in a probability space $(\Omega, \mathcal{F}, P)$ is a function

$$\begin{aligned} X\colon \Omega \times T &\longrightarrow \mathbb{R} \\ (\omega, t) &\longmapsto X_t(\omega) \end{aligned}$$

such that X is measurable, that is to say $X^{-1}(B) \in \mathcal{F} \times \mathcal{B}(T)$, for all $B \in \mathcal{B}(\mathbb{R})$ where $\mathcal{B}(\mathbb{R})$ denotes the Borelians of $\mathbb{R}$.

In this sense, a stochastic process is a function that does not only depend on the time t, but also on a trajectory ω of the probability space. From this point forward,

we assume all financial processes considered to be stochastic processes, in the sense that when we describe a process of prices $S(t)$ we are assuming it to be defined as above.

5.1.1 Osborne-Samuelson model

Let's assume we have a model of a series of prices in the form of a continuous time stochastic process $S = \{S_t\}_{t>0}$ in such a way that every S_t be a random variable that represents the price at time t and so, $S_{t_i} = s_i$, for any $i = 0, \ldots, n$. If we take into account daily variations of a price, this quantity is relative to the order of magnitude of said price. Therefore, it is natural we consider the relative variation of the price change, the so called *returns*

$$r_i = \frac{s_i - s_{i-1}}{s_{i-1}}. \tag{5.1}$$

Moreover, we define the stochastic process $X = \{X_t\}_{t>0}$, where $X_t = \log S_t$. This process represents the so called *log price process*. In our approach, we are interested in considering the following series $Y = \{Y_n\}_{n\in\mathbb{N}}$, called the *log return process*, where $Y_n = X_n - X_{n-1}$; in other words, Y describes the increments of X at given times $t_0, \ldots, t_n$. Considering the process $\{Y_n\}_{n\in\mathbb{N}}$ is justified in the sense that from (5.1) we have

$$\frac{s_i - s_{i-1}}{s_{i-1}} = \frac{s_i}{s_{i-1}} - 1 \approx \log(s_i) - \log(s_{i-1}) = X_t - X_{t-1}.$$

Before we dive into the Osborne-Samuelson model, it is important we define the concept of Wiener process, also referred to as Brownian motion.

Definition 5.2. A *Wiener process* W_t is a stochastic process characterized by the following properties:

1. $W_0 = 0$.
2. For every $t > 0$, the future increments $W_{t+\upsilon} - W_t$, $u \geq 0$, are independent of the past values W_s $s < t$.
3. W has Gaussian increments: $W_{t+\upsilon} - W_t$ is normally distributed with mean $\mu = 0$ and variance υ, $W_{t+\upsilon} - W_t \sim \mathcal{N}(0, \upsilon)$.
4. W has continuous paths: W_t is continuous in t.

Osborne-Samuelson's model assumes that the price process S evolves as a geometric Brownian motion, that is,

$$S_t = s_0 e^{\alpha t} e^{\sigma W_t}\ ,\ t \geq 0,$$

where s_0 is the current price, α is a real parameter, σ is positive and W is a standard Brownian motion. Note that this is equivalent to assume

$$X_t = x_0 + \alpha t + \sigma W_t,$$

where $x_0 = \log s_0$. Now, taking equally spaced observations we have the increments of the process X,

$$Y_n = \alpha + \sigma(W_n - W_{n-1}),\ n \geq 1. \tag{5.2}$$

Thus, according to the Osborne-Samuelson model, a series of daily log returns $y_i = x_i - x_{i-1}$ must be sampled from a normal random variable with mean α and standard deviation σ. From (5.2), it is equivalent to say that Y follows a distribution with density

$$f(y) = \frac{\exp(-y^2/2\sigma^2\tau)}{\sqrt{2\pi\sigma^2\tau}}, \tag{5.3}$$

where σ^2 is the variance of y over unit time intervals and y is the difference of log returns over a time interval τ. In physics, regarding Brownian motion, σ^2 is proportional to the temperature of the gas under discussion. Thus, by analogy, we can think of the "temperature" of the share market as being a variable which represents the degree of activity or energy of the markets, namely σ^2.

Based on our results in Chapter 4, we conclude that persistent homology assertively characterizes changes in volatility (or "temperature") in the stock market, as long as we consider the return of share prices as modeled in (5.3). This experiment should be considered as a toy example, since in reality the return of share prices does not follow a normal distribution.

5.1.2 Scaled Student t-distribution

An alternative to the Osborne-Samuelson model can be derived from considering a doubly stochastic model, which takes σ^2 to be itself a random variable following a distribution f, namely a re-scaled Student t-distribution. This is considered to be a more appropriate model to describe logarithmic returns, as with other models that consider fat-tailed distributions. Lightly, the idea of a doubly stochastic model is that we consider a random variable X to be modeled in two stages. In one stage, the distribution $F(X; h_0, \dots)$ of X is represented in a standard manner, for which we use one or more parameters. The other stage consists of describing some of these parameters $(h_0, h_1, \dots)$, to be themselves random variables with a certain distribution F'.

Praez writes in [25] an extension to the work done by Osborne and Samuelson, and derives a new model that uses a scaled Student-t distribution to model logarithmic returns of financial markets. His insight is based on observing that Osborne's model assumes the volatility σ^2 to be a constant, although now it is widely considered that the underlying distribution of the variation of share price changes is not fixed. Noticeably, Praez rewrites Osborne's model (5.3) as a conditional distribution

$$f(y \mid \sigma^2) = \frac{\exp(-y^2/2\sigma^2\tau)}{\sqrt{2\pi\sigma^2\tau}}. \tag{5.4}$$

Without loss of generality, we can rewrite (5.4) by considering $\tau = 1$ (a unit time interval) and y as having a non-zero mean μ. Thus, it becomes

$$f(y \mid \sigma^2) = \frac{\exp(-(y-\mu)^2/2\sigma^2)}{\sqrt{2\pi\sigma^2}}.$$

If we now denote by $h(y)$ the distribution of y which takes into account the random nature of σ^2, and $g(\sigma^2)$ to be the underlying distribution of the variance, we obtain $h(y)$ by integrating

$$h(y) = \int_0^\infty f(y \mid \sigma^2) g(\sigma^2)\, d\sigma^2. \tag{5.5}$$

A solution is given by

$$g(\sigma^2) = \frac{\sigma_0^{2m}(m-1)^m \sigma^{-2(m+1)}}{\exp[\frac{(m-1)\sigma_0^2}{\sigma^2}] \cdot \Gamma(m)}.$$

Here, $\sigma_0^2 = \mathbb{E}(\sigma^2)$ and the variance of σ^2 is $\sigma_0^4/(m-2)$. Now, when $g(\sigma^2)$ is substituted in (5.5), we obtain $h(y)$ by integration as

$$h(y) = \frac{\Gamma(m) \cdot [(2m-2)\pi]^{1/2} \cdot \sigma_0}{[1 + (y-\mu)^2/\sigma_0^2(2m-2)]^{1/2}}.$$

This is a Student t-distribution with $\nu = 2m-n$ degrees of freedom, except for a scale factor $[n/(n-2)]^{1/2}$. Thus, the distribution of $(y-\mu)/\sigma_0$ would be that of a scaled Student t-distribution. We note that, for n small, this distribution resembles that of a normal distribution. The distribution function $g(\sigma^2)$ of the variance has mean σ_0^2 and variance $\sigma_0^4/(m-2)$. Intuitively, it represents the distribution of the variance of the share under discussion and it is discussed in [25] to be an inverted gamma distribution. Similar tests as the ones in Chapter 4 have been done with variables following an inverted gamma distribution [19]. Although the results exposed above are founded, we will not give the particulars; for a more detailed exposition we refer the reader to [25], and references therein.

Thus, we assume this distribution (scaled t-Student), to be a more rigorous approach in modeling log returns of financial markets. Moreover, resulting tests done in Chapter 4 correctly measure a change in volatility (scale) of the distribution, again proving the validity of the method considered to measure changes in variance of financial data.

5.2 Analysis on European stock indices

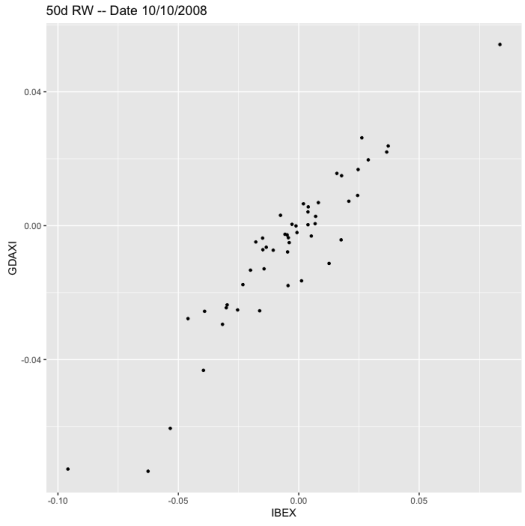
In this section, we follow [19], except that we choose financial indices from the European stock market. In our case-study, we consider IBEX 35, FTSE 100, DAX 30 and CAC 40 as notable stock indices in Spain, United Kingdom, Germany and France, respectively. Our study is designed to measure changes in the persistence of 1-dimensional loops born from this data, paying special attention during the time frames of the global economic crisis and the European debt crisis.

5.2.1 Procurement and treatment of financial data

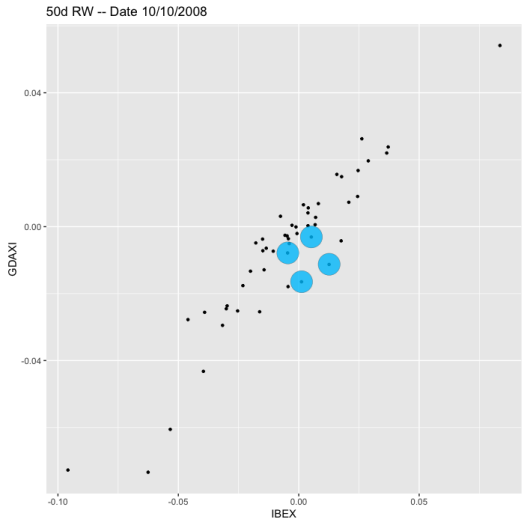
We obtain financial data using the R-package "quantmod", which extracts it from Yahoo finance. Our time frame ranges from the dates 01-01-2005 to 01-01-2015, we consider a window of 10 years to contain sufficient data to appropriately test this method. The data consists of the prices of various European stock indices and we will focus on gathering the adjusted closing value of every index. Since these values are not normalized, we use the previously defined indicator, namely the *logarithmic return*. Let us remember it is defined as the process $\log(P_{i,j}/P_{i-1,j})$, where $P_{i,j}$ is the adjusted closing value of index j at the day i. The idea now consists of considering the 4 time series independently, so we can embed them in $\mathbb{R}^4$ using the *mixture embedding* described previously. In our approach, the window size is $w = 50$ days, such that every day represents a point whose coordinates are the logarithmic returns of every index considered. Thus, for every sliding step, we have a point cloud embedded in $\mathbb{R}^4$. Figure 5.1 (a) is an illustration of a resulting point cloud on the k-th sliding step.

5.2.2 Studying the persistent homology of financial data sets

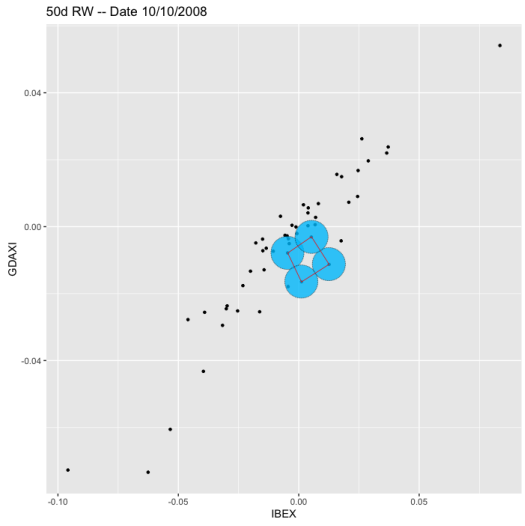
The sliding step is set to one day, which in our case yields an approximate $(2530 - w)$ time ordered set of point clouds, though variations on this number may result from missing trading dates or mismatching trading dates between indices. For every resulting point cloud, we use the R-package "TDA" to compute its persistent homology and, as mentioned previously, we construct the simplices using the Vietoris-Rips scheme. Figure 5.1 is an example illustrating this process. We recall that, in our experiment, we are computing one-dimensional persistent homology, so that for every Vietoris-Rips complex we are computing its H_1 homology group. In other words, we are only considering the persistence of 1-dimensional loops throughout. Then, for every resulting persistence diagram we use the function "landscape" from the R-package "TDA" to calculate its corresponding persistent landscape.



(a)



(b)



(c)

Figure 5.1: The Vietoris-Rips scheme; for every (fixed) radius we have a certain Vietoris-Rips complex.

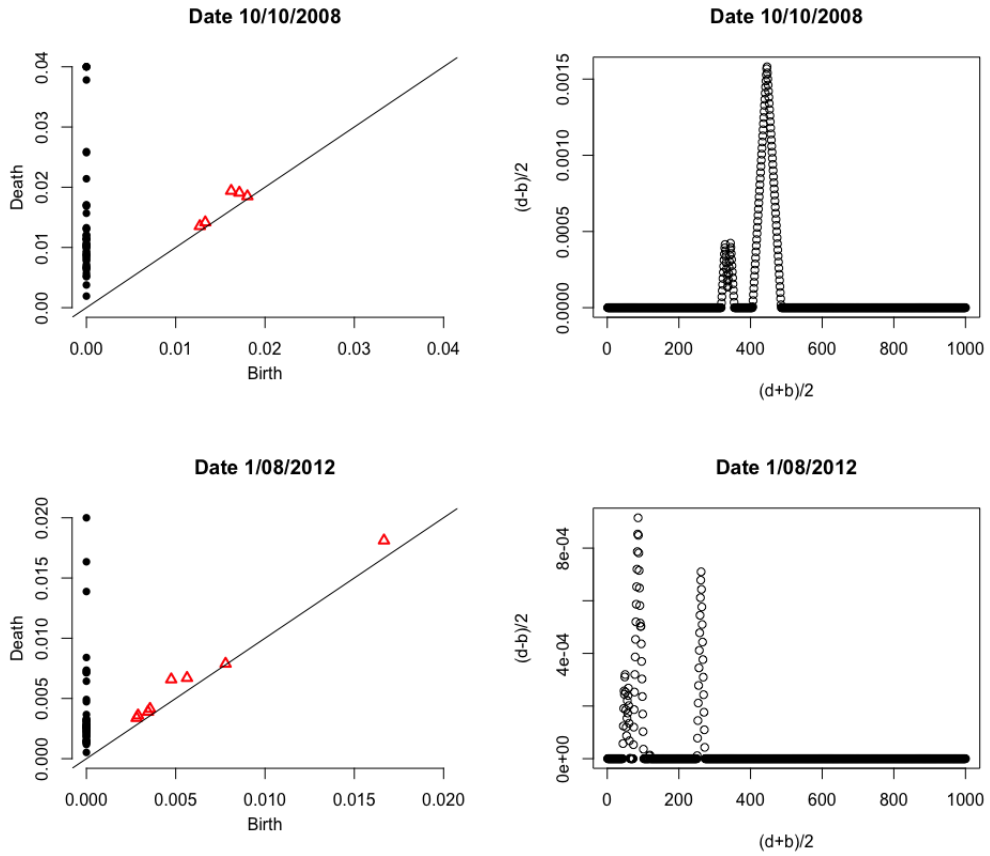
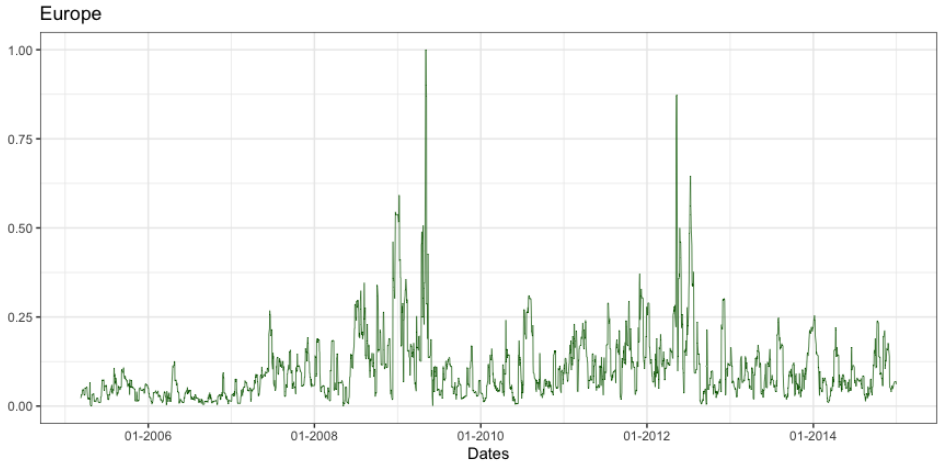


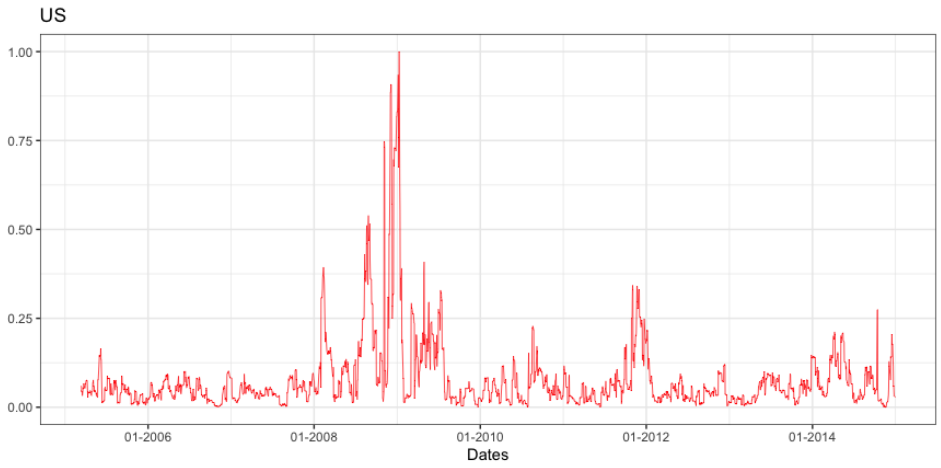
Figure 5.2: Top-down, Persistence diagram from Figure 5.1 and its corresponding 1-persistence landscape, followed by the persistence diagram and 1-persistence landscape from the same indices on different selected dates.

5.2.3 L^1 norm time series

As we know, the persistence landscape can be naturally embedded in the Banach space $L^1(\mathbb{N}\times\mathbb{R})$ and we can easily calculate its L^1 norm. The last step in our method is to construct a time series to illustrate the L^1 norm fluctuation. Our main goal is to capture variations that may have happened during the financial crisis of 2007-2008 and the European debt crisis of 2009-2013.



(a)



(b)

Figure 5.3: Time series of the normalized L^1 norms of persistence landscapes between 2005 and 2015.

Here, we interpret the L^1 norm time series in a qualitative sense. Noticeably, in Figure 5.3 we can appreciate strong peaks emerging during the period of the global economic crisis of 2007-2008. The main difference with our picture and the one derived in Guidea and Katz's paper is that of a small time shift and much higher fluctuations around the year 2012. Indeed, the financial crisis in the US is considered to have started with the Lehman Bankruptcy (15-09-2008), and regarded to have spread through the European markets a little later, around mid-October

2008. The higher peaks around 2012 we suspect are directly related to the European debt crisis, which emerged during late 2009 and had its highest impact in mid 2012. This fact can be appreciated in our picture, on the last quarter of the same year (see Fig. 5.3 (a)). We also note that, taking into consideration IBEX 35 as one of our indicators may have caused a *stronger fluctuation* during the European debt crisis period. In fact, Spain's main financial index reflects the fact that it was one of the European countries together with Greece, Portugal, Ireland and Cyprus, who were unable to repay or refinance their government debt without the assistance of third parties.

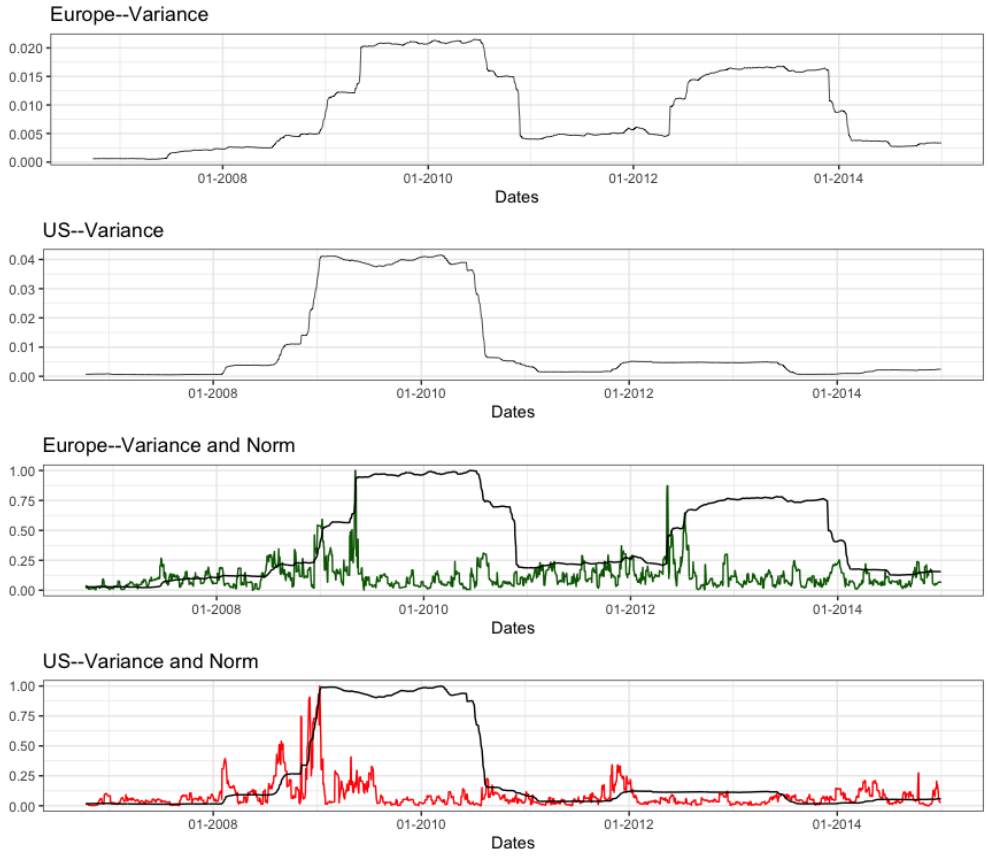


Figure 5.4: Top-down, the first two pictures show the associated variance time series (see plot title) with a rolling window of $w = 400$ days and sliding step set to one day. Las two pictures show the superposition of the normalized L^1 norm and corresponding variance time series (see plot title). Dates range from 2005 to 2015.

Still, we can perform a quantitative analysis of the time series illustrated, and furthermore we would like to be able to compare our results with the ones obtained in [19]. Though various statistical measures can be used to interpret the time series illustrated, we will restrict our study in analyzing the variance of each time series. We employ a rolling window of $w = 400$ days with the sliding step set to one day,

to the time series obtained from the European stock indices (Fig. 5.3 (a)) and US stock indices (Fig. 5.3 (b)). We also consider the variance of the L^1 norm time series derived from considering only the English, French, and German markets. We name this plot reduced Europe (see Fig. 5.5). Here, the method employed to study the *rolling variance* of a time series is analogous to that of the L^1 norm time series, namely the sliding window technique described in Chapter 3.

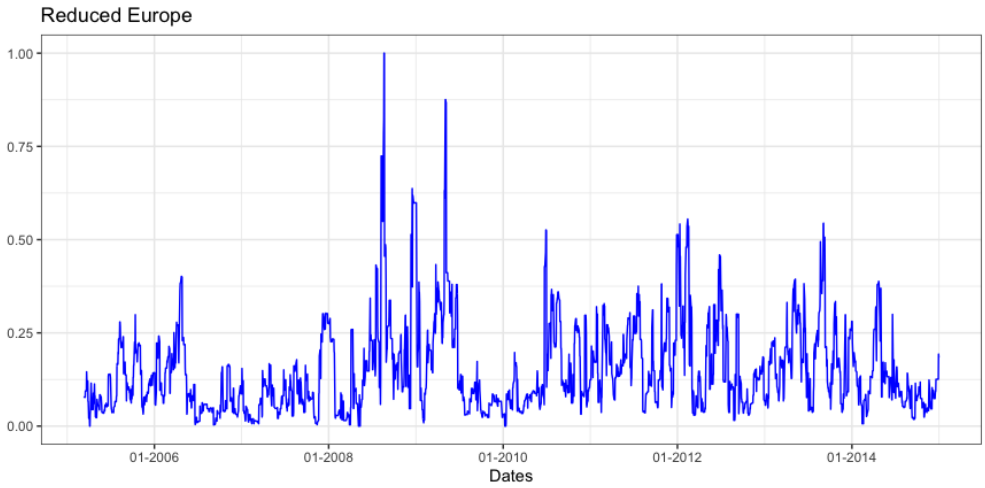


Figure 5.5: Normalized time series of the associated L^1 norm of persistence landscapes derived from three European indices.

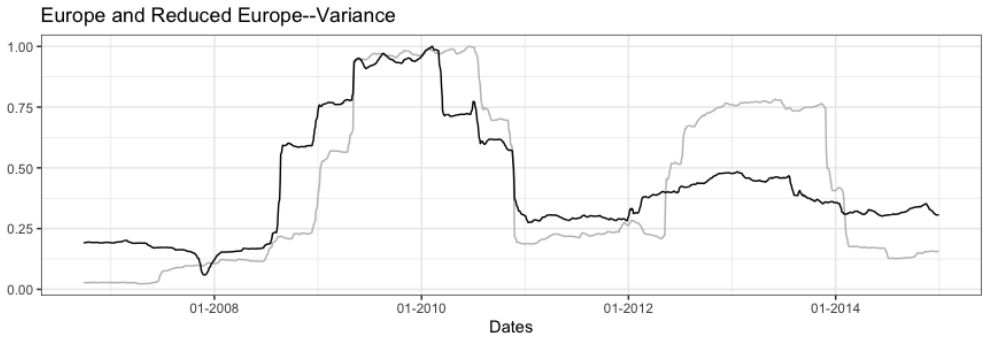


Figure 5.6: Superposition of the associated normalized variance of four European markets (in grey), in comparison to three European markets (in black). Details in text.

As was expected, variance derived from both European and US markets show strong growth prior to the financial crash of 2008. We also note that the variance associated to the time series of the four European markets starts to grow much earlier (mid 2007), in comparison to that of the US markets. It is also clearly seen that, in regards to the European financial debt crisis, variance and norms derived from

the persistence of loops of financial data during that period, show strong fluctuation and much higher peaks (Fig. 5.4 (Europe–Variance and Norms)) than the variance and norms derived from the US market. This result is expected, in the sense that different markets (US and European) have naturally different behaviors. Still, we can appreciate a slight growth in both variance and norms derived from the US market around the European financial debt crisis, an interpretation of this matter results in assuming the intrinsic dependency of global economic markets.

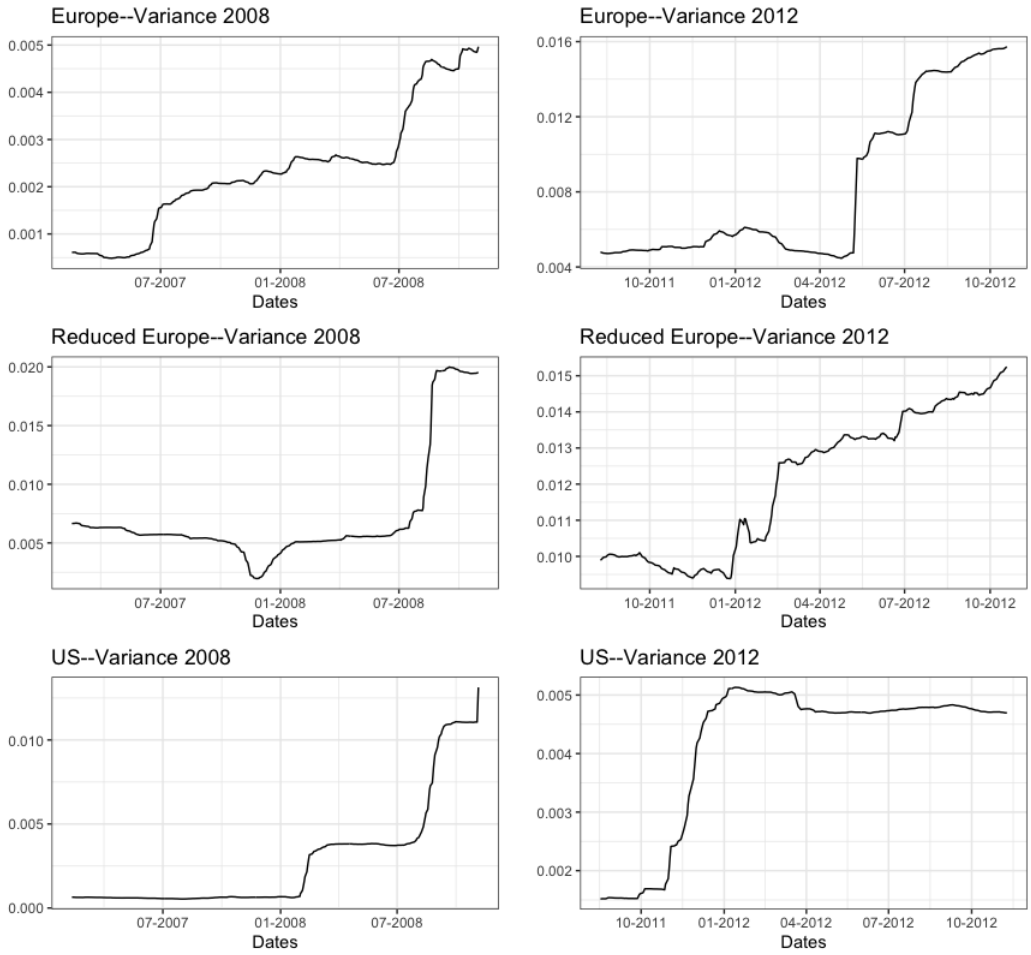


Figure 5.7: Localized associated variance for US and European markets. On the left, time series a year prior to November 2008. On the right, a year prior to the last quarter of 2012.

If we focus on the reduced Europe L^1 norm plot (see Fig. 5.5), in comparison to the illustration derived from considering the full European depiction (Fig. 5.3 (a)), we notice qualitatively that the apparent *volatility frequency* of the time series is much higher, especially from 2009 onward. On the other hand, frequencies from Fig. 5.3 (a) seem to be more stable with a distinguished higher peak around mid 2012. A

comparison of both variances derived from the corresponding L^1 norm time series is shown in Fig. 5.6. Although both variance time series seem to grow during that period, it is clearly seen that the mean variance derived from the reduced European norm time series is much lower than that of the non-reduced European plot. We suspect this matter is directly related to the fact that, adding the Spanish index IBEX 35 contributes to much higher volatility in the data set time series around mid 2012 and hence, resulting in a singular high peak in the non-reduced European plot. To synthesize this analysis, one could say that the volatility of the distribution during crisis periods from the non-reduced European data set is higher than that of reduced European data.

To conclude, Fig. 5.7 depicts the *local* behavior of the variance corresponding to each of the three norm plots considered. We notice, as in [19], the variance of each plot shows strong growth prior to both financial crashes. This is especially the case when considering the variance of the non-reduced European plot (Fig. 5.7, top left) near the financial crisis of 2008.

6. Conclusions

Regarding applications of TDA, yet again studying the persistence of 1-dimensional loops born from well-known financial data has proven successful in measuring irregularities across time in the markets considered. We want to stress that results gathered from our experiments should be seen as a continuation of the recent work done by Guidea and Katz in [19].

Hence, our approach was manifold. First, we replicated the financial time series analysis done with TDA [19] using *distinct* financial data (Fig. 5.3 (a)). Subsequently, we managed to gather the same picture derived from US financial data (Fig. 5.3 (b)), and thus we were able to put them in direct comparison. In both illustrations it can be seen that L^1 norms of persistence landscapes grow similarly when approaching periods of high financial instability. The main difference between our picture and the one derived in Guidea and Katz's paper is that of a small time shift and much higher fluctuations around the year 2012; we suspect this is due to the European financial debt crisis, which caused high volatility in European financial markets during that period. In addition, we considered a separate depiction, using only three European financial indices, namely a subset of the first four, taking out IBEX 35, and contrasted both L^1 norm time series, making use of both the qualitative norm picture and also comparing the *variance* of each time series. In regards to this last experiment, our conclusion is that the volatility of the distribution during crisis periods from the non-reduced European data set is higher than that of reduced European data. We also depict side-by-side the variance of local time-frames from each of the three data sets considered (Fig. 5.7), which shows strong growth *prior* to the financial market irregularities. This is especially the case when observing the local variance during the global financial crisis of 2008 (Fig. 5.7 left).

This approach demonstrates consistency in results and proves to be simple enough to be replicated as a means of research. As mentioned in their paper, this method behaves well enough to be applied to any type of mixture of time series. This can be seen in Chapter 3, where we break down our procedure without the need to give proper sense to the data sets themselves. Thus, this method can be seen as a means of studying the time-dependent fluctuation of the shape of *abstract* data. The novelty of this approach lies in the fact that we do not consider the data to be a discretization of an underlying shape, thus parting from the classical intent of the application of TDA, which relies on having an underlying geometry *a priori*.

We were also able to prove some empirical results gathered from the study of data born from *synthetic* time series. We note that most of the resulting propositions are based on Bubenik's work [1]. Tackling problems derived from working with

persistence landscapes as random variables is a field still considered to be in its infancy, allowing us to explore matters like the finiteness of expected values of L^1 norms derived from persistence landscapes under some conditions. Our main result, namely Theorem 4.1, should be seen as a complement to our empirical analysis of time series, since it describes the behavior of significant topological features extracted from altering data under a predefined distribution, and we believe that it could be used to study any type of data conducting in the same matter. Noticeably, our work exemplifies an unexpected bond between the fields of topology and statistics, and we believe there is still a vast margin for research, both practical and theoretical.

Bibliography

[1] Bubenik, P.: Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.* **16** (2015), 77–102.

[2] Bubenik, P.: The persistence landscape and some of its properties. arXiv: 1810.04963, 2019.

[3] Carlsson, G.: Topology and data. *Bull. Amer. Math. Soc.* **46** (2009), 255–308.

[4] Carlsson, G., et al.: Topological analysis of population activity in visual cortex. *J. Vis.* **8** (2008), 1–18.

[5] Carlsson, G., Levine A., Nicolau, M.: Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proc. Natl. Acad. Sci. U.S.A.* **108** (2011), 7265–7270.

[6] Carlsson, G., Zomorodian, A.: Computing persistent homology. *Discrete Comput. Geom.* (2005) **33**, 249–274.

[7] Ceren, N. A., et al.: Learning on blockchain graphs with topological features. arXiv:1908.06971v1, 2019.

[8] Chazal, F., Michel, B.: An Introduction to topological data analysis: fundamental and practical aspects for data scientists. arXiv:1710.04019, 2017.

[9] Chazal, F., et al.: The Structure and Stability of Persistence Modules. *SpringerBriefs in Mathematics*, Springer International Publishing, 2016.

[10] Cohen-Steiner, D., Edelsbrunner, H., Harer, J.: Stability of persistence diagrams. *Discrete Comput. Geom.* **37** (2007), 103–120.

[11] Cuenca, J., Iske, A.: Persistent homology for defect detection in non-destructive evaluation of materials. *Malaysia International NDT Conference and Exhibition 2015, Nov 22–24, Kuala Lumpur, Malaysia* (2015), NDT.net Issue: 2016-01.

[12] Edelsbrunner, H., Harer, J.: Persistent homology—a survey. Contemporary Mathematics, vol. 453, Amer. Math. Soc., 2008, 257–282.

[13] Edelsbrunner, H., Harer, J.: Computational Topology: An Introduction. Amer. Math. Soc., Providence, 2009.

[14] Edelsbrunner, H., Letscher, D., Zomorodian, A.: Topological persistence and simplification. *Discrete Comput. Geom.* **28** (2002), 511–533.

[15] Fasy, B. T., et al.: Introduction to the R package TDA. arXiv:1411.1830, 2015.

[16] Ghrist, R.: Barcodes: the persistent topology of data. *Bull. Amer. Math. Soc.* **45** (2008) 61–75.

[17] Guidea, M.: Topology data analysis of critical transitions in financial networks. *3rd International Winter School and Conference on Network Science*, Springer Proceedings in Complexity (2017), 47–59.

[18] Guidea, M., et al.: Topological recognition of critical transitions in time series of cryptocurrencies. arXiv:1809.00695v1, 2018.

[19] Guidea, M., Katz, Y.: Topological data analysis of financial time series: landscapes of crashes. *Physica A: Stat. Mech. App.* **491** (2018), 820–834.

[20] Harer, J., A. Perea, J.: Sliding windows and persistence: an application of topological methods to signal analysis. *J. Found. Comput. Math.* **15** (2015), 799–838.

[21] Hatcher, A.: Algebraic Topology. Cambridge University Press, Cambridge, 2002.

[22] Ledoux, M., Talagrand, M.: Probability in Banach Spaces. Ergebnisse der Mathematik, Springer (1991). Classics in Mathematics, Springer (2011).

[23] Osborne, M.F.M.: Brownian motion in the stock market. *Operations Research* **7** (1959), 145–173.

[24] Polterovich, L., et al.: Topological persistence in geometry and analysis. arXiv: 1904.04044, 2019.

[25] Praetz, P.: The distribution of share price changes. *J. Bus.* **45** (1972), 49–55.

[26] Skraba, P., Vejdemo-Johansson, M.: Persistence modules: algebra and algorithms. arXiv:1302.2015, 2013.