Topological Summaries for Time-Varying Data
Sintesi Topologiche per Serie Storiche

Tullia Padellini and Pierpaolo Brutti

Abstract Topology has proven to be a useful tool in the current quest for "insights on the data", since it characterises objects through their connectivity structure, in an easy and interpretable way. More specifically, the new, but growing, field of TDA (Topological Data Analysis) deals with Persistent Homology, a multiscale version of Homology Groups summarized by the Persistence Diagram and its functional representations (Persistence Landscapes, Silhouettes etc). All of these objects, however, are designed and work only for static point clouds. We define a new topological summary, the Landscape Surface, that takes into account the changes in the topology of a dynamical point cloud such as a (possibly very high dimensional) time series. We prove its continuity and its stability and, finally, we sketch a simple example.
Abstract *A causa della crescente complessità dei dati, diventa sempre più importante riuscire a sintetizzarli attraverso un numero ridotto di caratteristiche interpretabili. Lo studio delle invarianti topologiche si è dimostrato utile in questo senso, in quanto caratterizza un oggetto in termini della sua struttura di connettività. In particolare, lo studio della topologia dei dati viene condotto a partire da una versione multiscala dei gruppi omologici detti gruppi di omologia persistente, rappresentati da oggetti come il diagramma di persistenza, che rappresenta i generatori di tali gruppi, e le sue trasformazioni in spazi di funzioni. In questo lavoro introduciamo un nuovo strumento, costruito per studiare l'evoluzione delle caratteristiche topologiche di serie storiche multidimensionali, la "Landscape Surface". Dopo averne provato continuità e stabilità, accenneremo ad una sua applicazione in un semplice esempio.*

Key words: Persistent Homology, Time Series, Topological Inference

---

Tullia Padellini
Sapienza, Università di Roma, Piazzale Aldo Moro 5, e-mail: tullia.padellini@uniroma1.it

Pierpaolo Brutti
Sapienza, Università di Roma, Piazzale Aldo Moro, 5 e-mail: pierpaolo.brutti@uniroma1.it

1 Introduction to TDA

As we are dealing with increasingly complex data, our need for characterising them through a few, interpretable features has grown considerably. In recent years there has been quite some interest in the study of the "shape of data" [2]. Among the many ways a "shape" could be defined, topology is the most general one, as it describes an object in terms of its connectivity structure: connected components (topological features of dimension 0), cycles (features of dimension 1) and so on. There is a growing number of techniques (generally denoted as *Topological Data Analysis*) aimed at estimating the shape of a point-cloud through some topological invariant. In this work we extend those techniques to the case of multivariate time series, i.e. when, rather than considering only one point-cloud, we are dealing with a collection of point-clouds indexed by time, as for example in animal migration, player tracking in sports, EEG signals and most spatio-temporal data; our goal is to summarize in one object not only the shape of the data at each fixed time, but also how this shape changes with time.

Before introducing new objects, it is worth briefly reviewing what Topological Data Analysis (`TDA`) is, and how can we estimate the topology of data, or, to be more precise, the topology of the space $\mathscr{M}$ data was sampled from. As a matter of fact, data itself, when in the form of a point cloud $\mathbb{X} = \{X_1, \dots, X_n\}$, has a trivial topological structure, consisting of as many connected components as there are observations and no higher dimensional features. The basic idea in the `TDA` is thus to use data to build "shape aware" estimates of $\mathscr{M}$ and then compute topological invariants. One of the most common way of estimating $\mathscr{M}$, in `TDA`, is *Devroye-Wise support estimator* $\widehat{\mathscr{M}_\varepsilon}$ built by centering a ball of fixed radius ε in each of the observations X_i, i.e.

$$\widehat{\mathscr{M}_\varepsilon} = \bigcup_{i=1}^{n} B(X_i, \varepsilon)$$

where $B(Y, \delta)$ denotes a ball of radius δ and center Y. For each value ε we obtain a different estimate $\widehat{\mathscr{M}_\varepsilon}$, whose topology can be recovered by computing its Homology Groups. Persistent Homology, a multiscale version of Homology, then allows us to analyze how those Homology Groups change with ε.

Persistent Homology Groups can be summarized by the *Persistence Diagram*, a multiset $D = \{(b_i, d_i), i = 1, ..., m\}$ whose generic element (b_i, d_i) is the generator of the i-th Persistent Homology group. The space of persistence Diagrams $\mathscr{D}$ is a metric space, when endowed with the *Bottleneck distance*, which, given two multisets A and B, is defined as

$$d_B(A, B) = \inf_{\gamma} \sup_{x \in A} \| x - \gamma(x) \|_\infty$$

where the infimum is taken over all bijections $\gamma : A \to B$.

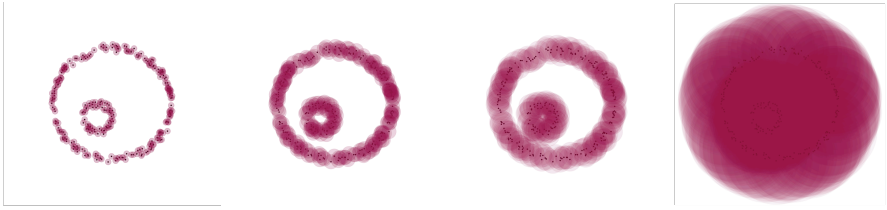

Fig. 1 $\widehat{\mathscr{M}_\varepsilon}$ for different values ε. For small values of ε (left), the topology of $\widehat{\mathscr{M}_\varepsilon}$ is close to the one of the point cloud itself. As ε grows more and more points start to be connected, until eventually (right) the corresponding $\widehat{\mathscr{M}_\varepsilon}$ is homeomorphic to a point. Values ε_b, ε_d of ε corresponding to when two components are connected for the first time (*birth-step*) and when they are connected to some other larger component (*death-step*) are the generators of a Persistent Homology Group.

The Bottleneck distance allows us to compare Persistence Diagrams and to define their most important property: *stability* [4].

Theorem 1. *Let $\mathbb{X},\mathbb{Y}$ two point clouds, and $D_\mathbb{X},D_\mathbb{Y}$ their corresponding Persistence Diagrams, then*

$$d_B\left(D_\mathbb{X},D_\mathbb{Y}\right) \leq 2d_H\left(\mathbb{X},\mathbb{Y}\right)$$

where $d_H(A,B)$ is the Hausdorff *distance between two topological spaces A and B.*

Roughly speaking, this means that if two point clouds are similar, then their Persistence Diagrams will be as well, and is therefore instrumental for using them in statistical tasks such as classification or clustering.

Since Persistence Diagrams are general metric objects, it is usually advisable to transform them in order to work with more statistics-friendly spaces. The most famous transformations of the persistence diagram are the persistence landscape [1] and the persistence silhouette[3], which are functions built by mapping each point $z=(b_i,d_i)$ of a Persistence Diagram D to a piecewise linear function called the "triangle" function T_z, defined as

$$T_z(y) = (y-b_i+d_i)\,\mathbb{1}_{[b_i-d_i,b_i]}(t) + (b_i+d_i-y)\,\mathbb{1}_{(b_i,b_i+d_i]}(y)$$

where $\mathbb{1}_A(x)=1$ if $x\in A$ and $\mathbb{1}_A(x)=0$ otherwise. Informally a triangle function links each point of the diagram to the diagonal with segments parallel to the axes, which are then rotated of 45 degrees.

The blocks T_z can be combined in many different ways. If we take their $k$max, i.e. the k-th largest value in the set $T_z(y)$, we obtain the *Persistence Landscape*

$$\lambda_D(k,y) = k\max_{z\in D} T_z(y) \qquad k\in\mathbb{Z}^+.$$

The persistence landscape is the collection of functions $\lambda_D(k,y)$. If we take the weighted average of the functions $T_z(y)$, we have the *Power Weighted Silhouette*

$$\phi_p(t) = \frac{\sum_{z \in D} w_z^p T_z(y)}{\sum_{z \in D} w_z^p}.$$

Although we are loosing some information in going from Persistence Diagrams

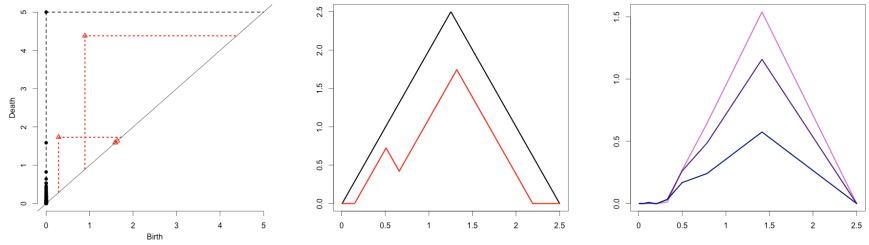

Fig. 2 Persistence Diagram (left), Persistence Landscape (center) and Persistence Silhouette for different values of p (right) of the data shown in Fig. 1

to Persistence Landscapes, the main result we had for Diagrams, i.e. stability, still holds [1].

Theorem 2. *Let $\mathbb{X}, \mathbb{Y}$ two point clouds, $D_{\mathbb{X}}, D_{\mathbb{Y}}$ their corresponding Persistence Diagrams, and $\lambda_{\mathbb{X}}, \lambda_{\mathbb{Y}}$ their corresponding Persistence Landscapes, then*

$$d_\Lambda(\lambda_{\mathbb{X}}, \lambda_{\mathbb{Y}}) \leq d_B(D_{\mathbb{X}}, D_{\mathbb{Y}}) \leq 2 d_H(\mathbb{X}, \mathbb{Y})$$

where $d_\Lambda(\lambda_{\mathbb{X}}, \lambda_{\mathbb{Y}}) = \| \lambda_{\mathbb{X}} - \lambda_{\mathbb{Y}} \|_\infty$ is the L^∞ distance in the space of Persistence Landscapes.

2 The Landscape Surface

In order to study the evolution of the topological structure of time-varying data, we think of a multidimensional Time series $\mathbb{X}(t)$ as a dynamic point cloud; for every fixed time t we can use the tools we have previously defined and build a Persistence Diagram $D(t)$, Landscape $\lambda_{\mathbb{X}(t)}(k,y)$ and Silhouette. Intuitively we can consider this Persistence Lanscape $\lambda_{\mathbb{X}(t)}$ as a function of time t as well, which means that we can work with a surface, rather than just a curve. It is important to notice that although in the following we focus on Landscapes, the same results hold for Silhouettes as well.

Definition 1. Given a dynamic point cloud $\mathbb{X}(t)$ we define the *Lanscape Surface* as the function

$$\Lambda(t,k,y) = \lambda_{\mathbb{X}(t)}(k,y) \qquad \forall t,k,y.$$

This surface is still a meaningful topological summary, as we can prove its stability.

Theorem 3. *Let $\{\mathbb{X}(t), \mathbb{Y}(t)\}$ with $t \in (0,1)$ two continuous dynamic point clouds, $\Lambda_{\mathbb{X}}$ and $\Lambda_{\mathbb{Y}}$ their corresponding Landscape Surfaces, then:*

1. *$\Lambda_{\mathbb{X}}$ and $\Lambda_{\mathbb{Y}}$ are continuous;*
2. *$I_\Lambda(\Lambda_{\mathbb{X}}, \Lambda_{\mathbb{Y}}) \leq I_H(\mathbb{X}, \mathbb{Y})$*

where $I_\Lambda = \int_0^1 d_\Lambda(\lambda_{\mathbb{X}(t)}, \lambda_{\mathbb{Y}(t)})dt$ is the Integrated L^∞ distance on the space of Persistence Landscapes and $I_H(\mathbb{X}, \mathbb{Y}) = \int_0^1 d_H(\mathbb{X}(t), \mathbb{Y}(t))dt$ is the Integrated Hausdorff distance for dynamic pointclouds.

The proof is a direct consequence of the Stability Theorem for Persistence Landscapes (2), in fact:

1. For a fixed t, consider $\mathbb{X}(t)$ and $\mathbb{X}(t+\varepsilon)$ (same applies for $\mathbb{Y}$). By 2 and the continuity of $\mathbb{X}(t)$ we have

$$0 \leq \lim_{\varepsilon \to 0} d_\Lambda \left(\lambda_{\mathbb{X}(t)}, \lambda_{\mathbb{X}(t+\varepsilon)}\right) \leq \lim_{\varepsilon \to 0} 2d_H \left(\mathbb{X}(t), \mathbb{X}(t+\varepsilon)\right) = 0.$$

2. Since for a fixed t we have, by 2 we have

$$d_\Lambda \left(\lambda_{\mathbb{X}(t)}, \lambda_{\mathbb{Y}(t)}\right) \leq 2d_H \left(\mathbb{X}(t), \mathbb{Y}(t)\right)$$

   integrating both terms is enough to prove the result.

In order to show an example of this object with real data, we consider EEG data, which are signals recorded at a very high frequency through many different electrodes (64 in our case). We build the Persistence Surface using EEG signals from an alcoholic and a control patient, both under the same stimuli. As we can clearly see from Fig. 3 and 4 these two subjects show a very different behavior. While the signal from the control patient is strongly characterized by a few persistent features, in the alcoholic patient there is less structured, as there are many features but they all have a smaller persistence, and could therefore be interpreted as noise.

References

1. Bubenik, P.: Statistical topological data analysis using persistence landscapes. The Journal of Machine Learning Research, **16(1)**, 77–102 (2015)
2. Carlsson, G.: Topology and data. Bulletin of the American Mathematical Society, **46(2)**, 255–308 (2009)
3. Chazal, F., Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L.: Stochastic convergence of persistence landscapes and silhouettes. In Proceedings of the thirtieth annual symposium on Computational geometry, ACM (2014)
4. Cohen-Steiner, D., Edelsbrunner, H., and Harer, J.: Stability of persistence diagrams. Discrete and Computational Geometry **37(1)**, 103-120 (2007)
5. Edelsbrunner, H., Letscher, D., Zomorodian, A.: Topological persistence and simplification. Discrete and Computational Geometry, **28(4)**, 511–533 (2002)
6. Munch, E.: Applications of persistent homology to time varying systems. Diss. Duke University (2013).

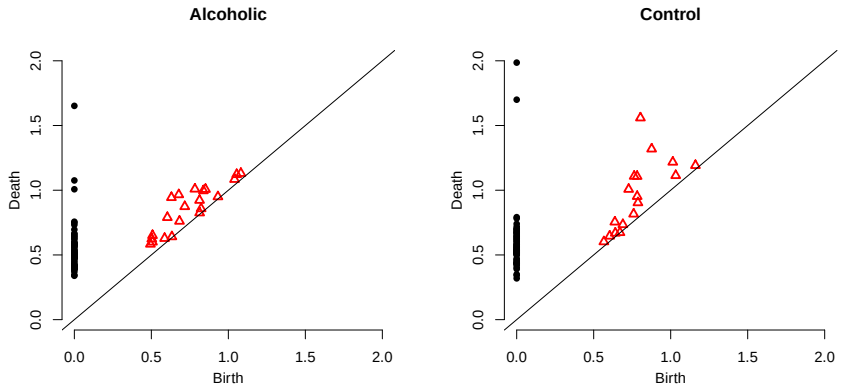

Fig. 3 Persistence Diagram of the Alcoholic and Control subjects for a fixed time t.

Fig. 4 Landscape Surface of dimension 1 for the EEG signal of a control patient (top) and an alcoholic (bottom)