HOW TDA CAN ANALYZE TIME SERIES DATA

ROBIN BELTON

In this lecture, we look at how Topological Data Analysis (TDA) can measure periodicity in time series data. At the end of this lecture you should be able to:

1. Compute a Rips complex and Rips filtration for simple examples by hand.
2. Define sliding window embeddings.
3. Be able to give a high level overview of the algorithm for measuring the periodicity of time series data using TDA.

1. Preliminaries

We first describe the theory behind the method for measuring the periodicity of a time series that involves TDA and sliding window embeddings.

1.1. TDA Terminology.

An *n-simplex* is the smallest convex set of $n+1$ points $v_0, v_1, ..., v_n$ where $v_1 - v_0, v_2 - v_0, ..., v_n - v_0$ are all linearly independent (See Figure 1).

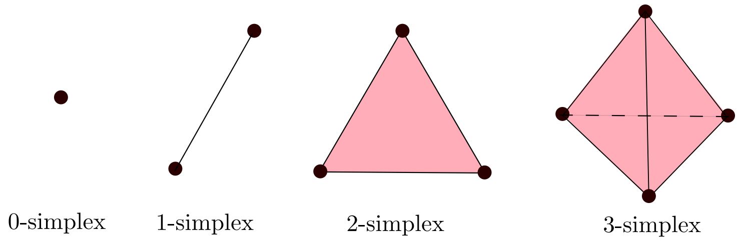

Figure 1. Simplices for dimensions 0,1,2, and 3.

An *abstract simplicial complex* is a finite collection of sets, A, that is closed under the subset relation, meaning if $a \in A$ and $b \subset a$, then $b \in A$. The elements $a \in A$ are the simplices where $\dim(a) := |a| - 1$. Here, $|a|$ is the number of points in the set, a.

Let $S \subset \mathbb{R}^n$ be a finite set of points. Let $r \geq 0$. The Rips complex S at r is the abstract simplicial complex $\mathrm{VR}(S, r)$ consisting of all subsets of diameter at most r.

$$\mathrm{V,R}(S, r) := \{\sigma \subset S \mid \mathrm{diam}(\sigma) \leq r\}$$

where the *diameter* of a set of points is the maximum distance between any two points in the set. For this lecture, we are only using the Euclidean distance. Geometrically, we compute $\mathrm{VR}(S, r)$ by centering balls of radius $r/2$ around each point. We add a $d-1$ simplex whenever we have d pairwise intersections.

A *filtration* of a simplicial complex K is a nested sequence of subcomplexes starting at the empty set and ending at the full simplicial complex. Going back to the Rips complex, we get a *Rips filtration* of nested Rips complexes if we vary the parameter r (See Figure 2).

Date: November 18, 2020.

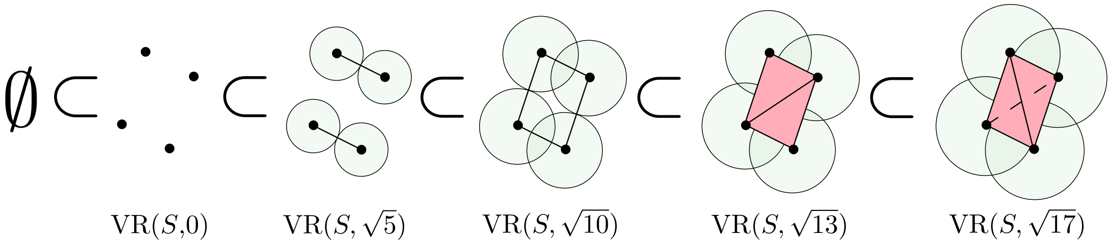

FIGURE 2. Rips filtration for $S := \{(0,0),(1,3),(2,-1),(3,2)\} \subset \mathbb{R}^2$

The *persistence diagram* for a filtered simplicial complex, K is a summary of the homology groups as the parameter ranges from $-\infty$ to ∞. At a high level, the persistence diagram is a set of birth-death pairs of the form (b,d) in $\mathbb{R}^2 \cup \{\infty, -\infty\}$, each with a corresponding dimension $k \in \mathbb{Z}_{\geq 0}$. Furthermore, all points on the diagonal, $y = x$ are also included with infinite multiplicity. Each off diagonal pair (b,d) represents an independent generator of the k-th homology group $H_k(K_t)$ for $t \in [b,d)$.

We focus on the 1-dimensional persistence diagram of a Rips filtration. We are interested in how long cycles persist in the filtration. We look at the *maximum persistence* of the 1-dimensional persistence diagram which is $\max_i d_i - b_i$ where (b_i, d_i) is an off diagonal point in the 1-dimensional persistence diagram (See Figure 3).

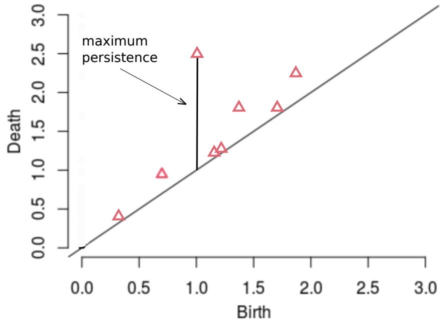

FIGURE 3. Computing the maximum persistence from a 1-dimensional persistence diagram.

1.2. **Sliding Window Embeddings.** Let $f : I \to \mathbb{R}$ where I is a closed interval in $\mathbb{R}$. Choose $M \in \mathbb{N}$ and $\tau > 0$. The *sliding window embedding* of f at $t \in \mathbb{R}$ is the point in $\mathbb{R}^{M+1}$

$$SW_{M,\tau} f(t) = [f(t), f(t+\tau), ..., f(t+M\tau)].$$

Choosing many different values for t gives a *sliding window point cloud.* The *window size* is the parameter $M\tau$. Note, we have similar definitions for a sliding window embedding in the case that $f : S \to \mathbb{R}$ where S is a finite subset of $\mathbb{R}$.

2. Measuring Periodicity in Time Series Data

Time series data is ubiquitous in our world. It is any type of data that looks at how something changes over time. Two examples include genomic time series data that tracks how gene expression changes over time, and crime activity data that records the number of crimes each day over time. We are often interested in discovering if a particular pattern repeats after a certain amount of time. For example, periodic behavior in genomic time series can help us deduce which genes are drivers in circadian rhythms. Periodicity in crime activity data may tell us which months we can expect more crime.

In [2], John Harer and Jose Perea propose a method for measuring periodicity that uses sliding window embeddings and Rips filtrations. In particular, the method is:

1. Compute a sliding window embedding for many points within a time series.
2. Compute the 1-dimensional persistence diagram of the sliding window point cloud using a Rips filtration.
3. Compute the periodicity score to be: $s = 1 - \frac{d^2 - b^2}{3}$. The score ranges between zero and one where zero means the data is perfectly periodic.

We illustrate the method in Figure 4.

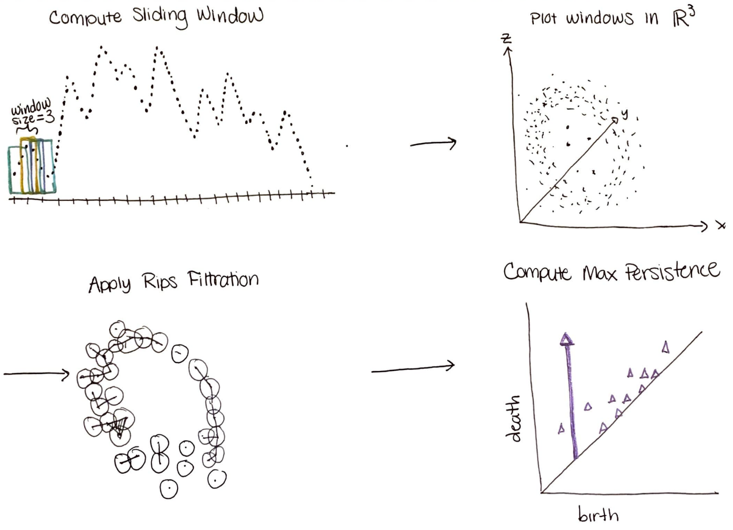

FIGURE 4. Illustration of TDA approach to measure periodicity in time series data.

The periodicity score is highly variant on the parameters. After analyzing the method on cosine and sine waves, the recommended window size for the sliding window is

$$M\tau = \frac{M}{M+1}\frac{2\pi}{L}$$

where L is the number of periods.

The TDA method above was used in [1] to see how it performed in detecting genes that drive periodicity in mice circadian rhythms. The results showed that TDA does better than other common methods for detecting genes that drive periodicity as long as the time series data has *low noise*. As more noise is added to the data, the TDA method does worse at detecting periodicity.

3. Conclusion

In summary, we can use sliding window embeddings and Rips filtrations to measure periodicity in time series that have low noise. If the time series we want to analyze has a lot of noise, then we need to apply a denoising algorithm before applying the TDA method.

References

[1] Jose Perea, Anastasia Deckard, Steven Haase, and John Harer. SW1Pers: Sliding Windows and 1-Persistence Scoring; Discovering Periodicity in Gene Expression Time Series Data. BMC Bioinformatics, 16, 2015.

[2] Jose Perea and John Harer. Sliding Windows and Persistence: An Application of Topological Methods to Signal Analysis. Foundations of Computational Mathematics, 15:799–838, 2015.