TDA for financial time series: Persistent homology and landscapes

Foundations of Geometric Methods in Data Analysis

CentraleSupélec

Nathan Brites – Armand Margerin

May 2020

Introduction

The goal of this project is to use TDA (Topological Data Analysis) to study financial time series. This study will be done using data from SandP500, DowJones, NASDAQ, and Russell2000 between 1988 and 2016. We will reproduce the experiments done in [1], try different parameter values and propose a new method to detect approaching market crashes thanks to the topological features of the time-series.

Question 1

No specific results are expected in this question but we can summarize the approach described in the paper. The idea is to use TDA on financial time-series to detect early signs of financial crashes. The authors describe a new method and then test it on historical data : SandP500, DowJones, NASDAQ, and Russell2000 between 1988 and 2016.

Here are the main steps of this method:

- Point clouds: using a sliding window of size w, transform the d time-series of length N into a time-ordered set of $(N - w)$ point clouds, where each point cloud is a $w * d$ matrix.
- Rips complex: create the Rips complex associated to each point-cloud.
- Persistence diagram: compute the persistence diagram of the Rips complex.
- Persistence landscape: compute the persistence landscape from the diagram. This is useful because the space of persistence diagrams endowed with the Wasserstein distance is not complete, hence not appropriate for statistical treatment.
- Norms: compute the L^p norm of the landscapes. This norm is expected to be high when approaching financial meltdowns.
- Statistical treatment: compute the main statistical indicators of the L^p norms time-serie.

Below is an example of persistence diagram, and its associated persistence landscape:

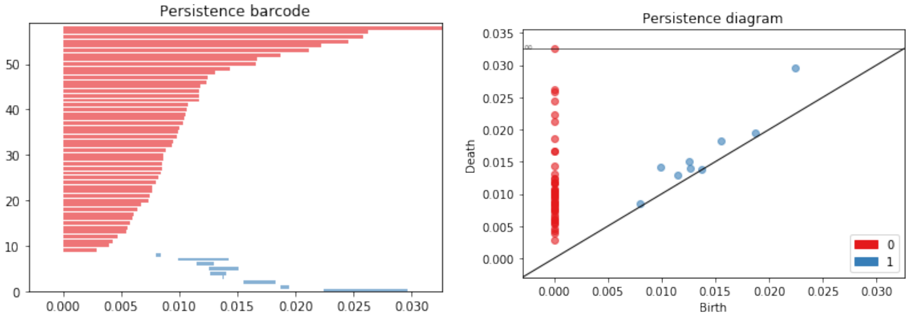

Figure 1: Barcode and persistence diagram of the point cloud of log returns during 50 trading days ending on march 3 2000

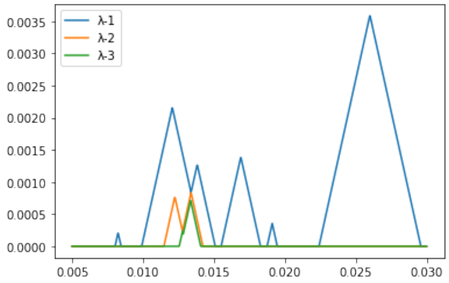

Figure 2: Persistence landscape related to the previous persistence diagram

Question 2

The definition of a persistent landscape is described in the second part of the article. For each birth-death point (b_α, d_α) in the diagram, we compute the values of its associated piecewise linear function $f_{(b_\alpha, d_\alpha)}$ on the *nbodes* points of the grid. Then, on each point of the grid, we sort the list of these values. Finally, to obtain the landscape n° k, we take the k-est element of each list.

This implementation is not optimal in terms of complexity, but we use it because it is straightforward and the datasets are small.

Question 3

Using the landscape function created before, we reproduce the experiments of the paper, and then compare the results when taking windows of different sizes.

Below are the plots of L^1 and L^2 norms of the landscapes before the "dotcom crash" for different values of w, the window size:

$w = 40$:

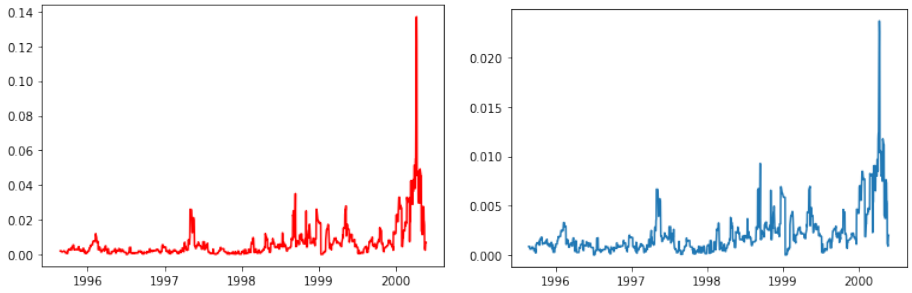

Figure 3: L^1 (red line) and L^2 (blue line) of the landscapes with w=40

$w = 80$:

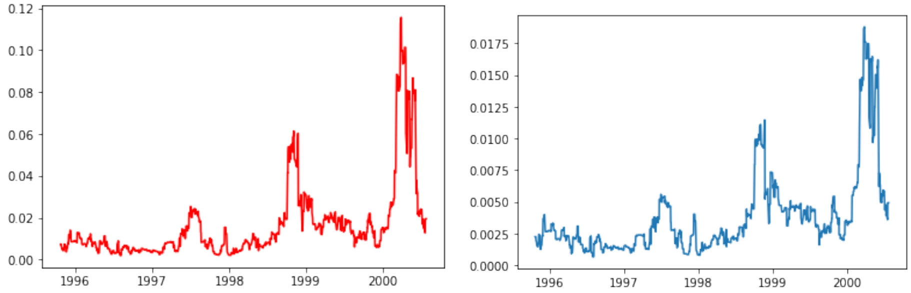

Figure 4: L^1 (red line) and L^2 (blue line) of the landscapes with w=80

$w = 120$:

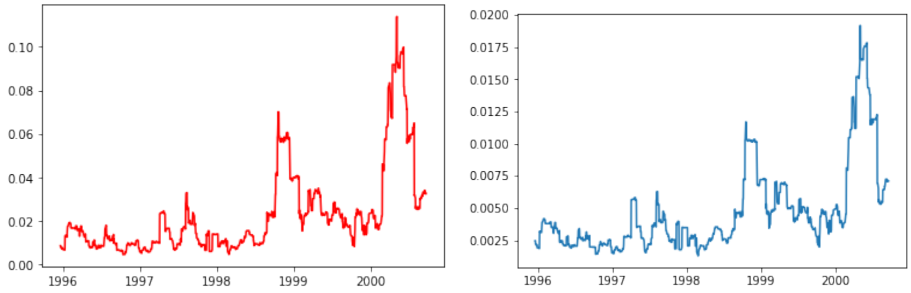

Figure 5: L^1 (red line) and L^2 (blue line) of the landscapes with w=120

Those result are very similar to the one in the paper, both L^1 and L^2 norms increase strongly before the "dotcom crash". The size of the window only has a small impact on the final results: when we increase it, the main spike is slightly flattened and a medium-size spike appears at the end of 1998.

We also have to keep in mind that the window size has a big impact on the computational cost because with bigger point clouds the computation of topological features would be more expensive.

Question 4

To achieve the previous results we calculated the L^p -norm of the landscapes. We can also try to calculate other metrics on persistent diagrams.

The first metric we tried is the bottleneck distance between two consecutive persistent diagrams:

Bottleneck distance measures the similarity between two persistence diagrams. It is the shortest distance d for which there exists a perfect matching between the points of the two diagrams such that any couple of matched points are at distance at most d. It corresponds to the degree p Wasserstein distance with $p = \infty$

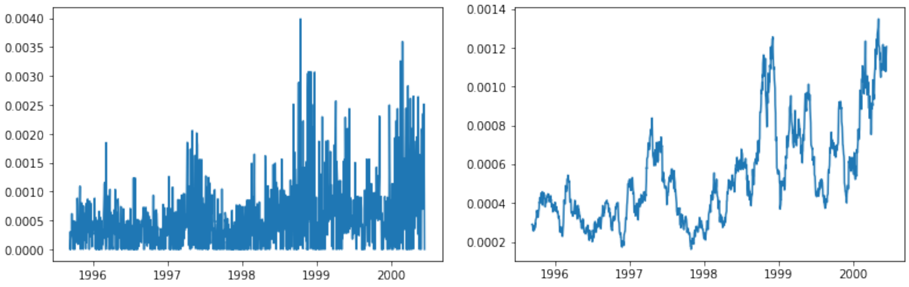

Figure 6: Bottleneck distance between consecutive persistence diagrams (real values on the left and smoothed curve on the right)

The bottleneck distance is not robust at all because this distance is actually a distance between a single pair of points. That is why the line oscillates a lot.

But, we can see the same trend than the previous analysis, the bottleneck distance between consecutive diagrams increase before the 'dotcom crash'.

Another metric which can be used is the Wasserstein distance between two persistence diagrams. This metric is the distance needed to get an exact match. For instance, on the following graph, red points represent a diagram and the blue points another diagram, the Wasserstein distance is the sum of all the edges.

Those two distances are similar, in one case it is the maximum length of the edges, in the other case it is the sum of the length of the edges.

We obtain similar results, but still, the L^p-norms calculated from the persistence landscapes seems to be more convenient to predict financial crises.

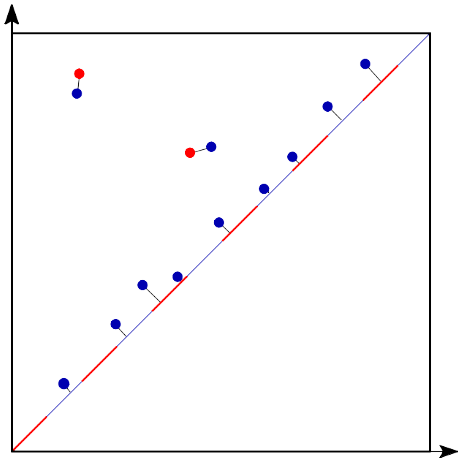

Figure 7: Graphical representation of Wasserstein distance

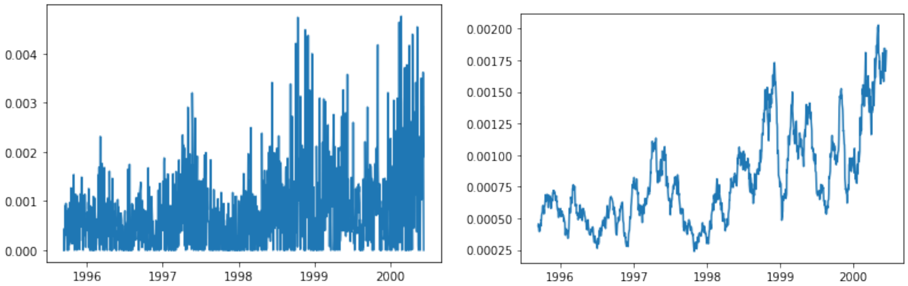

Figure 8: Wasserstein distance between consecutive persistence diagrams (real values on the left and smoothed curve on the right)

References

[1] Yuri KATZ Marian GIDEA. Topological data analysis of financial time series: Landscapes of crashes `http://geometrica.saclay.inria.fr/team/Fred.Chazal/Centrale2017.html`. 2017.