Topological Data Analysis of Financial Time Series: Landscapes of Crashes

Groupe 49
Louis-Marie Lovichi lm.lovichi@me.com
Benjamin Dahan Monsonego benjamin.dahan_monsonego@ens-paris-saclay.fr

December 18, 2023

1 Introduction and contributions

Financial crises resulting from crashes can have disastrous economic consequences, causing global economic destabilization and socio-political impacts. Detecting these sudden changes poses a significant challenge. Marian Gidea and Yuri Katz's article « Topological Data Analysis of Financial Time Series: Landscapes of Crashes » suggests using persistence homology to highlight potential early warning signals. The field of Topological Data Analysis allows the extraction of new features from multidimensional time series using Persistent Homology. Furthermore, the underlying shape of the data is often far more robust to noise than the data itself. Analyzing the topology of financial data could then be an enriching feature, all the more that this procedure can easily be applied to high-dimensional problems.

In the context of this report, we propose an implementation of the described methods to verify the authors' results, quickly explore the influence of the sliding window size on a synthetic dataset (check the github repository) and apply the method to detect early warning signals in the case of various recent stock indices and cryptocurrencies. To this end, we relied on approximately 20% of existing code, notably by using libraries in the field of topological data analysis (persim, ripser, gudhi), and, similar to the authors, accessed the available data via yfinance. We contributed equally to this mini-project; the report was co-written by both Benjamin and Louis-Marie. However, Benjamin focused more on the literature research regarding various concepts of topological analysis, while Louis-Marie concentrated more on the writing and structure of the code.

2 Method

Topological Data Analysis (TDA) is a recently introduced powerful tool in data science that uses techniques from algebraic topology and statistics to study the shape of data. It is particularly powerful to understand the structure of complex, high-dimensional finite datasets. Applications of TDA are various, and the one highlighted in the paper is using the particular case of persistence homology method.

2.1 Background

Persistent homology tracks k-dimensional holes in finite datasets that are persistent under a change of scale and resolution. In other words, persistence homology extracts from datasets information about 'birth' and 'death' of particular shapes. Some preliminary notions are needed to understand what is going on.

Definition 1 (k-simplex)**.** *A k-simplex σ of $\mathbb{R}^d$ is the convex hull of k+1 points $\{p_0, \ldots, p_k\} \in \mathbb{R}^{d\times k}$ that are affinely independent : $\sigma = conv(p_0, \ldots, p_k) = \{x \in \mathbb{R}^d | \sum_{i=0}^{k} \lambda_i p_i, \lambda_i \in [0,1], \sum_{i=0}^{k} \lambda_i = 1\}$*

As examples of simplices, we have points (0-simplex), line segment (1-simplex), triangles (2-simplex), tetrahedrons (3-simplex), and so on. A crucial underlying notion is then simplicial complexes, which are higher-order structures built from simplices. It is a set composed of simplices that are joined together in a particular way.

Definition 2 (Simplicial complex)**.** *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. The dimension of a simplex σ is $dim(\sigma) = |\Sigma_\sigma| - 1$.*

The next step is to construct a filtration of simplicial complexes, ordered with respect to some resolution (scaling) parameter.

Definition 3 (Vietoris-Rips complex)**.** *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 $d\left(x_i, x_j\right) \leq \epsilon$ for all $0 \leq i, j \leq k$.*

Given a simplicial complex K, the set of p-dimensional simplices of K forms an abelian group called p-dimensional simplicial homology, and is denoted as $H_p(K)$. For each non-zero k-dimensional homology class α, there exists a pair of values $\epsilon_1 < \epsilon_2$, such that $\alpha \in H_k(R(X, \epsilon))$ but α is not in the image of any $H_k(R(X, \epsilon_1 - \delta))$, for $\delta > 0$ and the image of α in $H_k(R(X, \epsilon')$ is non-zero for all $\epsilon_1 < \epsilon' < \epsilon_2$, but the image of α in $H_k(R(X, \epsilon_2)$ is zero. In this case, we say that α is 'born' at the parameter value $b_\alpha = \epsilon_1$, and 'dies' at the parameter value $d_\alpha = \epsilon_2$ For every k-dimensional simplicial homology group, information about this 'birth and death' process is stored in a persistence diagram P_k, which consists of representing each class by its coordinate (b_α, d_α), and representing also all points in the positive diagonal of $\mathbb{R}^2$. These points represent all trivial homology generators that are born and instantly die at every level.

The final important notion is persistence landscape

Definition 4 (Persistance landscape)**.** *Suppose given a persistence persistence diagram $\mathcal{D} = \{(b_i, d_i)\}_{i\in I}$. 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 $\Lambda : \mathbb{N} \times \mathbb{R} \longrightarrow \mathbb{R}$ given by

$$\Lambda(k, x) = \text{k-max}\left\{f_{(b_i,d_i)}(x)\right\}_{i\in I}$$

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

Definition 5 (Persistance landscapes L^p norm)**.** *Let $\Lambda : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$ 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*

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

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

2.2 Method

The paper's authors introduced a method based on Topological Data Analysis (TDA) that uses Persistent Homology (tracking of k-dimensional features described above) to derive topological characteristics from a multivariate time series in a four-dimensional space (considering adjusted close log returns of four indices of the stock market), intending to identify and predict financial crashes. Features are obtained from data slices extracted from the original time series via a sliding time window of length w. The method involves employing Vietoris-Rips filtrations to generate a persistence diagram for each segment (or point cloud), which is then transformed into a persistence landscape. For each time window, a landscape is generated and subsequently processed (via norm L_1 or L_2) into a singular numerical value. These values are then recombined back into a comprehensive time series. The study emphasizes the capability of this final time series to reveal both current and potential future trends in the market. What is interesting about the paper's computations is that, at the end of this process, a peak appears in the obtained time series before historical financial crashes (DotCom Bubble, 2008's Crisis,..), which demonstrates a change of topology concerning the four indices studied, and by topology one has to understand the whole system of correlations and balance between studied indices.

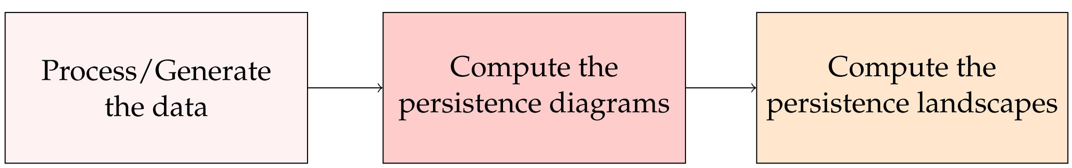

Figure 1: Method description.

3 Data

3.1 Synthetic data - noisy Hénon map

Regarding Synthetic Data, we decided to follow the paper and generate Hénon Map with increasing scale values for parameter a (from 0 to 1.4), and four different values of b: 0.27, 0.28, 0.29, 0.3. As well, we generated white noise with Gamma-distributed inverse variance : white noise is at first generated following a Gaussian with variance following Gamma distribution at some scale lambda, but at some point scale lambda is being decreased to make variance of the noise grow.

The Hénon map is notably used for its parallel with economic cycles and as a model for financial crashes. It is known that the Hénon sequence approaches an invariant set known as the Hénon attractor. Furthermore, this attractor is chaotic for certain values of the parameters a and b. For instance, it is known that for a value of $a \approx 1.06$, a chaotic attractor emerges. Then, by adding

Gaussian noise and by slowing increasing a, we obtain the following set of equations for the noisy Hénon map :

$$\begin{cases} x_{n+1} = 1 - a_n x_n^2 + b y_n + \sigma W_n \sqrt{\Delta t} \\ y_{n+1} = x_{n+1} + \sigma W_n \sqrt{\Delta t} \\ a_{n+1} = a_n + \Delta t \end{cases}$$

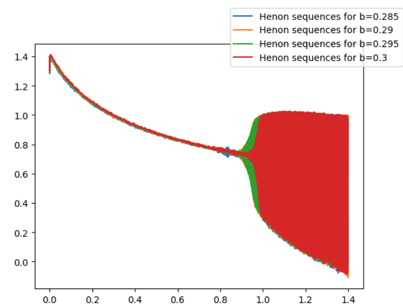

Figure 2: Noisy Hénon sequences with growing a.

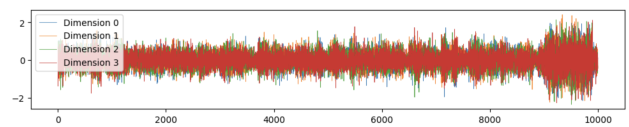

Figure 3: Gamma-inversed white noise.

3.2 Real data from major stock indices

Regarding real-time Data, we decided at first to apply TDA on historical financial crashes to test our code. After that, we applied it on four european indices (CAC40, IBEX35, DAX, and FTSE100) during Covid19, to verify if a "bubble" happened during this period, and if there was any notable change in the market state. For those financial data, we calculate the log returns. Domain expertise asserts that integrating the signal is then approximately stationary, thus facilitating the analysis. On the other hand, one topic that attracts more and more attention is cryptocurrencies. We applied TDA on indices BTC, ETH, ADA, SOL.

Figure 4: Adjusted close prices over the whole timeline.

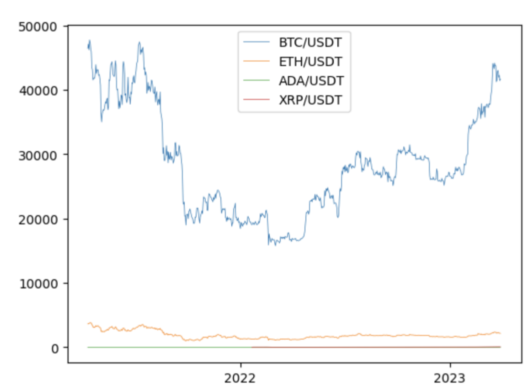

Figure 5: Adjusted close prices over the whole timeline.

4 Results

Regarding the Hénon Map, we illustrated two results. The first one is the robustness of the method regarding noise and outliers, shown in Figure 6. Indeed, we computed Wasserstein distance for 3 Diagrams with close values of parameter b (0.27, 0.28, 0.29), and the curves are very look-alike. The second one is about behavior identification. Figure 7 shows a sharp increase around the parameter value 'a' that marks the onset of chaotic behavior. It was expected as the same behavior would happen until this value of $'a'$ is reached, then the variance would increase suddenly and rapidly. Hénon Map is the perfect example to perform TDA on, as behavior suddenly and notably changes at some particular point. Regarding the example of data generated by noise with Gamma-distributed inverse variance, at some point, we modified the value of the shape parameter, which increased the variance as shown in Figure 8. Applying TDA, we can notice a peak appearing on the landscape's norms, around this period.

Regarding major European stock indices, observing Figure 9, we can see that a peak appears when we approach February 2020 and that during the same year, the norm seems to be perturbed. After that, the norm tends to go to its normal behavior, i.e before Covid19. TDA shows that Covid19 implied an upheaval of the market state. Regarding cryptocurrencies, we can notice on Figure 10 that the norms are quite unstable. It can especially be due to the limited amount of data we managed to import (1 year), but reflects the unstable aspect during past year, which perfectly reflects the instability and constantly changing behavior of this market.

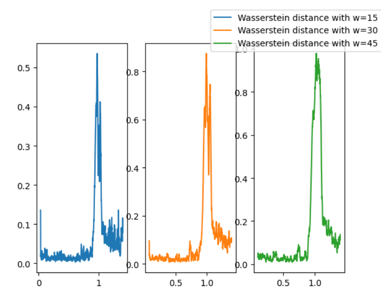

Figure 6: Sensitivity of the Wasserstein distance to the window length.

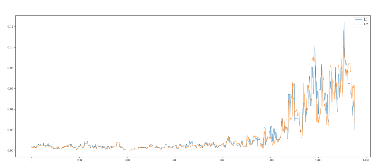

Figure 7: Persistence landscapes for Hénon map with previous parameters.

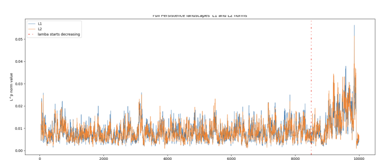

Figure 8: Persistence landscapes for Hénon map with previous parameters.

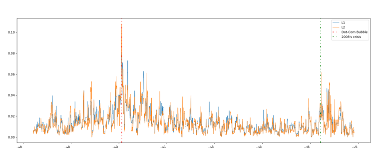

Figure 9: Persistence landscapes for Dot-Com Bubble and 2008's crisis.

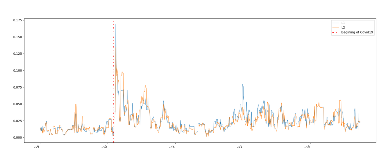

Figure 10: Persistence landscapes for Covid crisis.