Topological Data Analysis of a Financial Time Series for insights from Landscapes for detecting Financial Crashes

Hammoutene, Sarah Lina; Mohan, Siddharth

ESSEC Business School, Cergy, France,
CentraleSupélec, Gif-sur-Yvette, France

February 2018

Abstract: *The goal of this paper was to replicate the results of the original paper by Gidea and Katz (2017) and draw special inferences on the usage of Topological Data Analysis (TDA) as a novel method to identify warning signs for impending financial crashes. This was done by a multivariate time series analysis of four major US stock market indices, namely the S&P 500, Dow Jone Industrial Average (DJIA), NASDAQ, and Russell 2000. The usage of persistent homology to identify and quantify topological patterns is illustrated here. Persistence of detected transient loops are measured and then encoded into real-valued functions referred to as 'persistence landscapes'. Their temporal changes are quantified via their L^p norms.*

1 Introduction

A growing set of techniques that reveal interesting inferences from the 'shape' of the data encompasses what is also as TDA. The use of this set of techniques has been inhibited due to issues in combining the main concept, that is 'persistence diagrams', with machine learning and statistical analysis. This paper resolves this by looking into the usage of a 'persistence landscape'. Since this is a function, it allows the usage of the vector space structure of its underlying function space. Allowing this topological summary to be viewed as random variables with values in a Banach space, it ends up obeying the strong law of large numbers and the central limit theorem. Hence, they can be defined as a series of piece-wise linear functions thereby allowing faster calculations compared to persistence diagrams [1].

TDA has already been shown to be successful in understanding the topology of the space of natural images (Carlson et al., 2008), and for the discovery of breast cancer subgroups (Nicolau et al., 2011).

These recent breakthroughs with the use of TDA set the question for its application in the field of quantitative finance. Its previously studied applications make it an interesting and novel method to identify the rising risk surrounding financial markets especially the light of the recent unprecedented financial crashes. Predicting these crashes are of vital importance to multiple stakeholder. While multiple hypothesis have been postulated regarding the volatility trend preceding financial shocks, there is no agreed upon consensus. This coupled with the lack of a predictive model that works within an acceptable accuracy range.

2 Background

The standard flow for TDA involves starting with data that one encodes as a finite set of points in R^n or more generally in some metric space. Following this, there is an application of some geometric construction to which algebraic topological tools are applied. The end result is a topological summary of the data. The two standard topological summaries of data are the barcode and the persistence diagram (Edelsbrunner et al., 2002; Zomorodian and Carlsson, 2005; Cohen-Steiner et al., 2007). A similar topological summary referred to as a persistence landscape is used in this paper. And all these summaries are obtained from something called the persistence module.

The persistence module is the main algebraic object of TDA studies. A persistence module M consists of a vector space M_a for all a $\in$ R and linear maps M(a $\leq$ b) : M_a M_b for all a $\leq$ b such that M(a $\leq$ a) is the identity map and for all a $\leq$ b $\leq$ c, M(b $\leq$ c) M(a $\leq$ b) = M(a $\leq$ c). The persistence module is constructed by looking at the creation of the singular homology groups [2]. An increase in the size of the homology groups induces the creation of maps between corresponding groups. The images of these maps are referred to as persistent homology groups. To be more specific, the calculation of simplical and peristent homology involves the creation of a simplical complex; a space built from simple pieces (simplicies) identified combinatorially along the faces [3]. The most frequently referred to complex is the Cech complex. To obtain the singular homology of a union of balls, one can calculate the simplicial homology of the corresponding Cech complex. The Cech complexes, together with the inclusions, form a filtered simplicial complex. Applying simplicial homology we obtain a persistence module. There exist efficient algorithms for calculating the persistent homology of filtered simplicial complexes (Edelsbrunner et al., 2002; Milosavljevic et al., 2011; Chen and Kerber, 2013).

The Cech complexes and other sub-complexes, which are equivalent to it, such as the alpha complex, are computationally heavy. Hence, they are often substituted by using the Rips complex, which is larger but simpler. It has x_i points as the vertices along with having k-simplices corresponding to k + 1 balls with all pairwise intersections being nonempty [5]. Simply put, both the Cech and Rips complexes are data point similarity visualizations up to to 'filtration value [4].

With regard to all these persistence diagrams, birth indices form the horizontal axis while the death indices from the vertical axis. A metric space structure can be used to encompass the space with the diagrams. A degree p Wasserstein distance, with p $\geq$ 1, is the standard metric. It is defined by

$$W_p(P_k^1, P_k^2) = inf_\phi \left[\Sigma \amalg x - \phi(x) \amalg_\infty^p \right]^{\frac{1}{p}} \tag{1}$$

where the summation is over all bijections $\phi : P_k^1 \to P_k^2$ and $\amalg.\amalg_\infty^p$ denotes the sup norm. When $p = \infty$, the Wasserstein distance is know as the bottleneckk' distance. The incompleteness of the metric space, formed by the space of the persistence diagrams endowed by the Wasserstein distance, makes it unsuitable for statistical testing. This is a problem for our goal of studying the diagrams with statistical analysis. This is why we resort to the usage of persistence landscapes p [1].

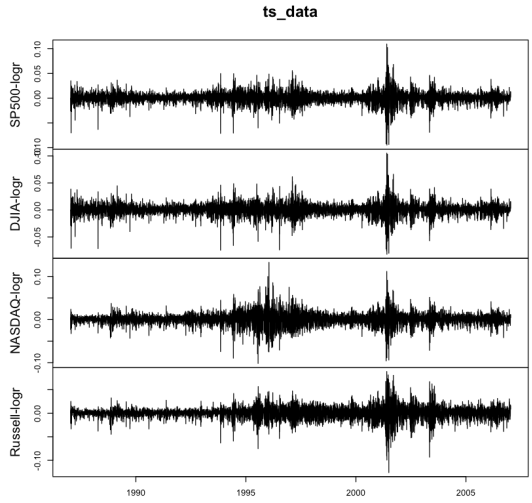

Figure 1: Line plots of the time series for the daily forward log returns of S&P 500, DJIA, Nasdaq and Russell 2000

As stated earlier, they are a piecewise linear functions which form brief overviews of the persistence diagram. The concept was introduced by Bubenik and is a useful vectorization for statistical analysis of persistence diagrams [2]. Essentially, the persistence landscape rotates the persistence diagram so that the diagonal becomes the new x-axis. The i-th order of persistence landscapes creates a piecewise linear function from the i-th largest value of the points in the persistence diagram after the rotation. For a birth-death pair p = (b, d) $\in$ D, where D is the persistence diagram, the piecewise linear functions, $\wedge_p$ (t) : R $\to$ [0, 1], are

$$\wedge_p(t) = \begin{cases} t-b & \text{if } t \in [b, \frac{b+d}{2}], \\ d-t & \text{if } t \in [\frac{b+d}{2}, d], \\ 0 & \text{otherwise} \end{cases}$$

The persistence landscape is then $\mathrm{F} : \mathrm{R} \to \mathrm{R}$. It forms a subset of the Banach space, upon being endowed with the norm. This makes it suitable for treatment by statistical techniques.

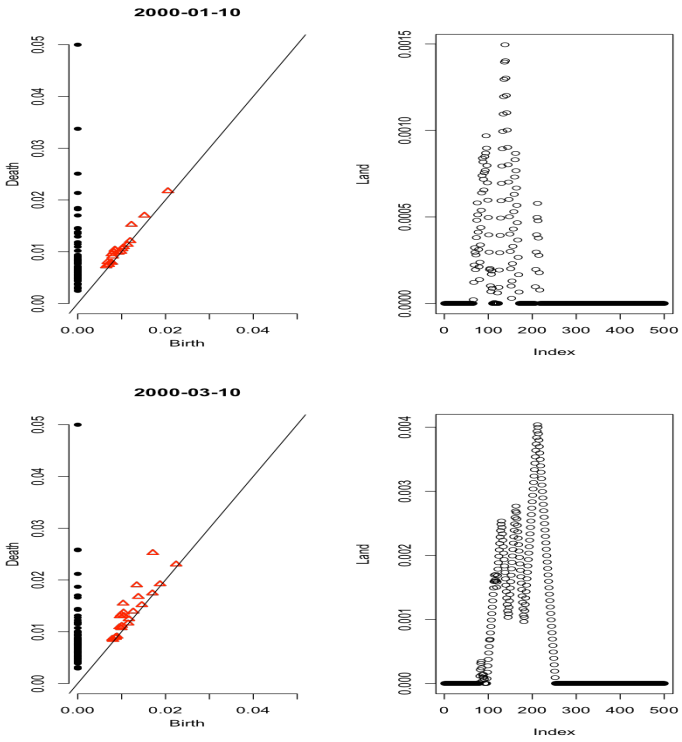

Figure 2: Persistence Diagrams with their corresponding Persistence Landscapes, obtained with a window of 80, in the vicinity of the Dotcom Crash

3 Data

The data taken into consideration were the time-series of four major US stock market indices, namely S&P 500, DJIA, NASDAQ and Russell 2000, ranging from December 23, 1987 to December 08, 2016 (7301 trading days). The datasets were downloaded from Yahoo Finance. Log-returns were calculated for each index and trading day. This is defined as the forward daily changes in the logarithm of the ratio $\mathrm{r}_{ij} = \ln P_{ij}/P_{i-1j}$, where P_{ij} represents the adjusted

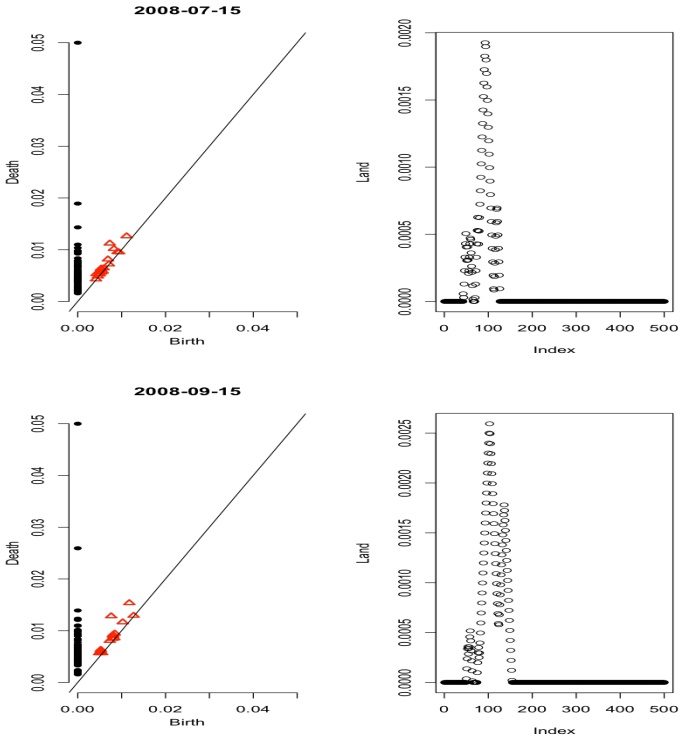

Figure 3: Persistence Diagrams with their corresponding Persistence Landscapes, obtained with a window of 80, in the vicinity of the Lehman bankruptcy

closing value of the index j at day i. An important note is that the data had missing dates, which corresponded to weekends. So applying the default time series function would shift the results. In order to resolve this, we enter values of zero for dates of weekend followed by the application of the time series function. Subsequently, the weekend entries were filtered and removed.

4 Method

The objective of applying this method is to garner some inferences from a multivariate time series, via its topological features. A point cloud is created with w points, where w represents the size of the sliding window. Each point cloud is represented by matrix of dimensions w X d, where d represents the number of columns. In our case d=4. We consider sliding windows of the following sizes: 40, 50, 80, and 120. The sliding step is set to one day, which in this case would yield (7300-w) time-ordered set of point clouds.

Following the creation of these point clouds, the Rips Diagrams were calculated. Subsequently, their persistence landscapes were obtained and their nor-

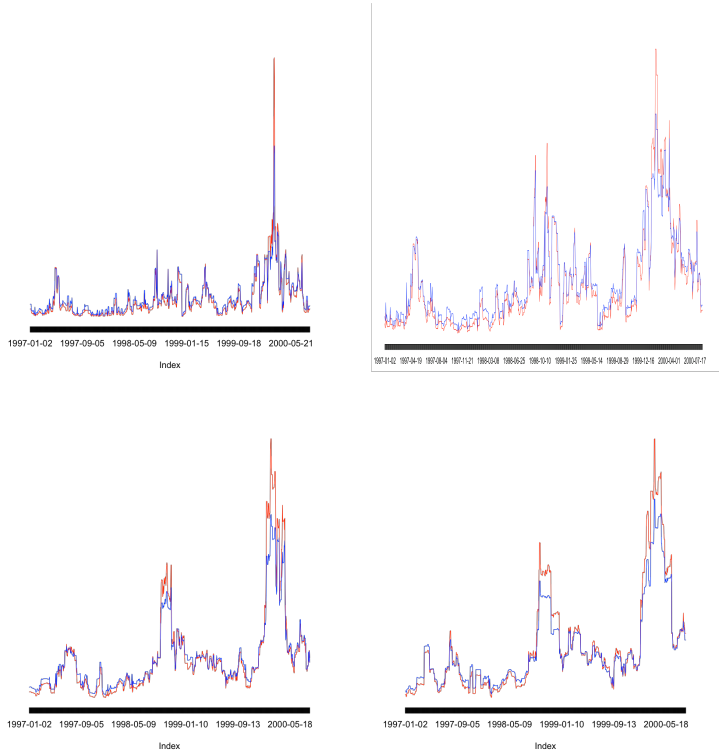

Figure 4: Normalized L1 (red line) and L2 (blue line) norms of persistence landscapes obtained with windows of size 40, 50, 80, and 120 (from top in a clockwise direction)

malized L^1, followed by, their L^2 norms were calculated. The L^p norms of the persistence landscapes allow us to check for any temporal changes in the state of the market. They allow for the quantification of the stability of topological features.

It is worth mentioning that a special focus is given to the days preceding the Dotcom crash on March 10, 2000 and the Lehman bankruptcy on September 15, 2008.

5 Results

Following the pre-processing of the data, as described previously, the Rips Complex persistence diagrams and their corresponding persistence landscapes were obtained. As is evident from Figure 3 and Figure 4, a direct distinction between the topological signals and the surrounding noise can be observed. Upon narrowing the focus to March 10, 2000 and September 15, 2008, corresponding

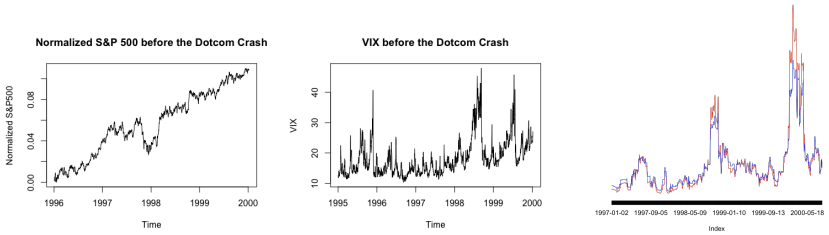

Figure 5: A look into the Normalized S&P 500, the VIX, and the L^1 (red) and L^2 (blue) norms of the persistence landscape before the Dotcom Crash

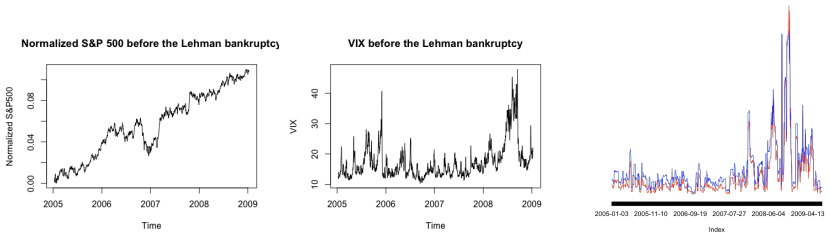

Figure 6: A look into the Normalized S&P 500, the VIX, and the L^1 (red) and L^2 (blue) norms of the persistence landscape before the Lehman bankruptcy

to the Technology crash and the Lehman crisis, we can see point clouds more clearly with an increase in volatility of the stock market.

In order to give a more robust explanation to this, the L^p norms of persistence landscapes were computed, with p=1 and p=2 being considered in this case. Figure 4 clearly illustrates the presence of an increasing trend a couple of months before the imminent crash. This is supported by both the L^1 and L^2 norms for both the financial crashes considered (only the L^p norms of the persistence landscapes in the vicinity of the Dotcom crash are shown in this paper). What is also interesting to note is that a higher resolution of the trend is noted by increasing the window size, though for a qualitative study, very window sizes are not required. As the market gets agitated with the approaching crashes, an increased persistence of loops is observed.

In order to obtain a better inference, we compare the normalized L^1 and L^2 norms of the persistence landscapes, obtained with a sliding window w of 80, with the corresponding normalized time series of the S&P 500 and the VIX. The VIX, also knows as the Chicago Board Options Exchange (CBOE) Volatility Index, is used to forecast the volatility implied by the S&P 500 index.

The Mann-Kendall statistical test for trend analysis was applied. As Guttal et al. (2016) applied it on the VIX and obtained values of 1 and 0.841 for the Kendall-tau coefficient [6], we tried the same on the normalized L^p norms of the

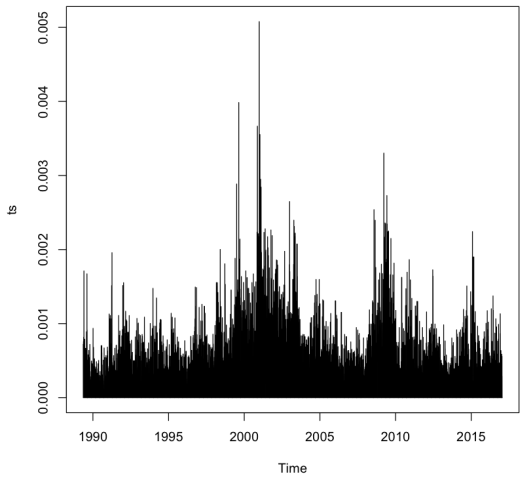

Figure 7: A line graph of the bottleneck distance of two consecutive persistence diagrams.

persistence landscapes. We obtained a coefficient value of 0.0299 associated with the Dotcom crash, in coherence with the upward trend as shown. The Kendall-τ values implied an ordinal association or high level of rank correlation. This is a classic indication of peaking volatility prior to financial crashes.

In addition to this, the bottleneck distances between two consecutive diagrams. As stated earlier, this distance is the Wasserstein distance at p being equal to infinity. Figure 7, which shows a full representation of the distances, shows the distinct peaks associated with the crash in 2000 and the crash in 2008. However, the accompanying point cloud is extremely noisy and unsuitable for the application of statistical methods.

6 Conclusion

This innovative method for assessing the environment surrounding financial crashes provides a new set of Early Warning Signs (EWS) that is worth considering.

Strong persistence of loops and sharp rise in peaks preceding an impending financial crash could be now considered as factors for future forecasting and analysis of financial time series and tracking market volatility.

Aside from this, this method has just a single parameter, which is the size of the window w. Another method, that is as the time-delay embedding by F.Takens (1981), depends on the size of the segment and the size of the window along each of those segments [7].

The application of Mann-Kendall tests shows the ability for the application of statistical methods on persistence landscapes; one of the major advantages over simple noisy persistence diagrams. This sets a precedent for the application of other statistical methods to derive an even deeper insight into market volatility.

7 Acknowledgements

We wish to thank Professor Frederic Chazal of the *Institut National de Recherche en Informatique et en Automatique* (INRIA) for guiding us with this project. We are also grateful to the original authors, Professor Gidea and Dr. Katz, for having a developed a novel method, that has numerous applications in quantitative finance, for the detecting the systemic risks leading up to a financial crash.

8 References

[1] Gidea, M., and Katz, Y., Topological data analysis of financial time series: Landscapes of crashes, *Physica A: Statistical Mechanics and its Applications*, vol 491, 2017, pp 820-834, https://www.sciencedirect.com/science/article/pii/S0378437117309202, (accessed on 1 February 2018).

[2] Bubenik, P., Statistical Topological Data Analysis using Persistence Landscapes, *Journal of Machine Learning Research*, vol 16, 2015, pp 77-102.

[3] Ghrist, R., Barcodes: The Persistent Topology of data, *Bulletin of the American Mathematical Society (New Series)*, vol 45, issue 1, 2008, pages 61–75.

[4] Karimi, HA,. and B. Karimi, *Geospatial Data Science Techniques and Applications*, Boca Raton, CRC Press: Taylor and Francis Group, 2018. Available from Google Books (accessed on 23 February 2018).

[5] Truong, P., 'An exploration of topological properties of high-frequency one-dimensional time series data using TDA', PhD Thesis, KTH Royal Institute of Technology, 2017.

[6] Guttal, V. et al., 'Lack of critical slowing down suggests that financial meltdowns are not critical transitions, yet rising variability could signal systemic risk', *PLOS ONE*, vol 11, issue 1: e0144198, https://doi.org/10.1371/journal.pone.0144198.

[7] F. Takens, Detecting strange attractors in turbulence. In Lecture Notes in Mathematics 898 Springer-Verlag (1981).