STAT4770 Final Report

Darren Zheng

April 2024

Abstract

Topological analysis uses new tools in persistent homology to capture geometric information of data. Since its invention in the early 2000's, it has been used to tackle primarily high-dimensional data, due to its computational efficiency and resistance to noise. TDA has found useful applications in biomedical data, sensor network analysis, etc. Only recently in the past decade have some tried applied it to financial data, particular with developing early warning signs of the stock market. Gidea and Katz's method using persistence landscapes seems promising and I want to test them out to detect other historical market crashes, specifically the general stock market crash in 2020 and cryptocurrency crash of 2018.

Contents

1 Introduction

Topological data analysis is a relatively new field that borrows tools from algebraic topology to extract geometric features of data. In a precise sense, topological data analysis is quite good at tackling high-dimensional, high noise data. Persistent homology is the key concept that captures topological data and can be used to construct various diagrams, such as classical persistence barcodes. In the past couple of years, barcodes have been eschewed in favor of other representations. Some notable applications, include the algorithm Mapper, which creates groupings of data. Mapper can be seen as an alternative to k-means clustering that is more computationally efficient on high-dimensional data; a famous example of its use is in studies of the brain neuron interactions, or "connectome", to extract patterns. Another development is that of a persistence landscape, a visual representation that captures the "birth" and "death" of features of data over time utilizing simple piecewise linear functions.

Financial markets and actors have always been concerned with predictors of market behavior. Expectations of significant growth/loss drive day-to-day decisions. Significant declines in the stock market due to crashes represent a particularly difficult problem to detect, given its sudden catastrophic nature. Personal traders and firms have a personal profit-driven interest, but even governments/policy makers want methods to predict impending financial crises. Much research goes into developing early warning signals (EWS for short) and testing their accuracy.

These aforementioned persistence landscapes play a key role in Marian Gidea and Yuri Katz's 2018 paper [5]. In it, they focused on the dotcom (technology) crash of 2000 and financial crisis of 2007-2008. Their comparison relied on computations involving the persistence landscape in the L_1 and L_2 norm, as the space of all persistence diagrams (surprisingly) admit a distance metric [1]. In the paper, Gidea and Katz computed the variance (VAR), auto-correlation function at lag time 1 (ACF1), and mean spectral density (MPS). The conclusions were drawn based on visual comparison of the L_1 time series with the Chicago Board Options Exchange Volatility Index (we will refer to it as CBOE VIX, or VIX). They also used the Mann-Kendall test to determine whether there were statistically significant upward/downward trends in the VAR, ACF1, and MPS values leading up until the date of financial collapse.

In another paper [6], a number of other authors (Ismail ,etc) applied similar techniques to the markets of the United States, Singapore, and Malaysia. This second paper introduced analysis using the variance, auto-correlation function, and mean power spectrum of these aforementioned L_1, or L_2, norms. The goal of this project is to apply these techniques to study the time-series of stock market indicators in analyzing the 2020 stock market crash and 2018 cryptocurrency crash. The aim is to verify whether these approaches actually work on other datasets.

2 Background

For more general background, refer to [3] and [2]. For a comprehensive technical treatment, the two papers [5], [6] present a good brief summary. Below, we will only give an less-than-rigorous account of the most important technical ideas for the sake of space.

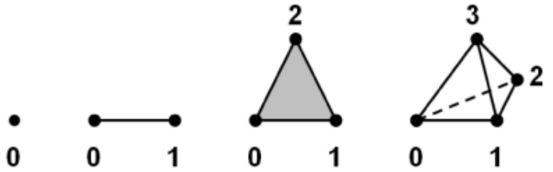

Figure 1: Examples of n-simplices for $n = 0, 1, 2, 3$ from left to right

Topology is always concerned with studying geometric features of usually continuous data. In this vein, homology groups were developed to provide an algebraic way to potentially distinguish spaces. For example, a popular saying is that a coffee cup and donut are the same in topology because both feature 1 "hole". Topology does not care about the scale of this coffee cup versus the donut. However, we cannot compute the homology of point-cloud data directly as the structure is not connected. Thus, we use a filtration of complexes. We begin with a point cloud in $\mathbb{R}^d$, where d is usually high. For a given $\epsilon > 0$, imagine drawing a d-dimensional ball denoted as $B_\epsilon(p)$ of radius ϵ in the usual Euclidean norm about each point p. If ϵ is large enough, we can imagine balls of certain pairs of points p_1 and p_2 may overlap. Whenever this occurs, draw a line segment between the two points. If it happens to be the case that 3 balls of 3 points p_1, p_2, p_3 overlap, then we have line segments between each pair of points already, so we fill in the "triangle" enclosed by those line segments. In general, whenever n balls overlap, we draw an n-dimensional simplex.

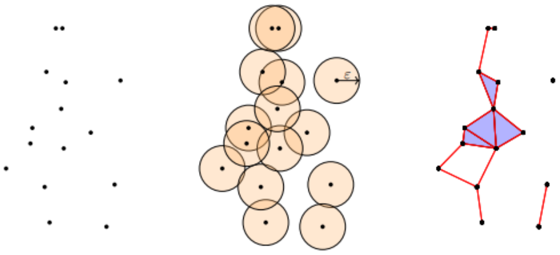

Figure 2: The Vietoris-Rips Complex from [10]

For a fixed ϵ, we end up with what is called a simplicial complex. The simplicial complex captures the data of nearness of data points. Now, given a simplicial complex, we get a chain complex from which we can compute homology groups. This is super technical, but just think of the homology groups as giving us the existence of n-dimensional holes in our space (See [] for more details). Notice however that there is an issue! How do we choose an appropriate value of ϵ? If we select ϵ to be too small, then we get little overlap; if we select ϵ to be too large, then everything in our point cloud becomes connected. The standard trick is to choose a variety of values of ϵ to create a spectrum of simplicial complexes.

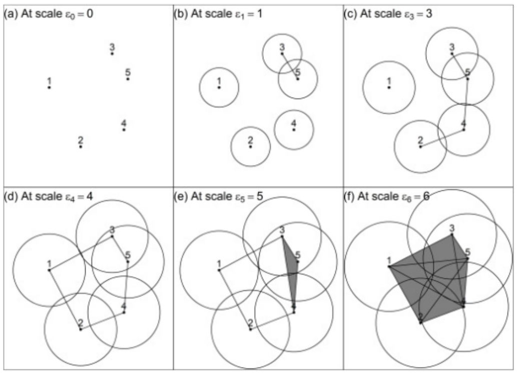
(a) At scale $\varepsilon_0 = 0$
(b) At scale $\varepsilon_1 = 1$
(c) At scale $\varepsilon_3 = 3$
(d) At scale $\varepsilon_4 = 4$
(e) At scale $\varepsilon_5 = 5$
(f) At scale $\varepsilon_6 = 6$
1
2
3
4
5

Figure 3: Simplicial Complex for different choices of ϵ from [6]

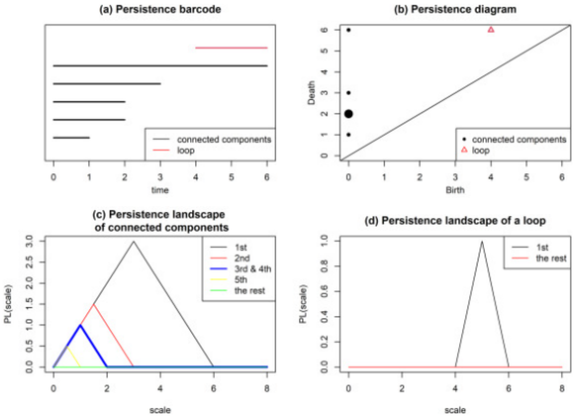

Figure 4: Source: [6]

The simplicial complex becomes what is called a filtration since as ϵ increases, we can only add more and more features to our simplicial complex. This is what gives this type of homology its name of "persistent" homology as the features persist after a certain ϵ. At each simplicial complex, we can compute the homology groups. Other data we can extract are more visual charts, such as persistence barcodes, diagrams, or landscapes. If we keep label simplices, we can keep track of the birth/death time of such features. The birth-time of a specific n-simplex is when it first appears and the death-time is when this simplex becomes the boundary of another higher dimensional simplex. See Figure 4 for visual examples.

Next, we need to note a few things about time-series analysis of financial data and EWSs, specifically defining certain indicators. These indicators are called CSD indicators, or critical slowing down indicators. As their name suggests, CSDs are used to represent variability and slow-down of stock market growth. Broadly, people intuitively view CSDs as a first-order predictor of near-future market behavior. There are many different types of CSDs, but the ones that we will use are autocorrelation function with lag time 1 (ACF), variance (VAR), and mean power spectrum (MPS) as mentioned in the introduction. Since our time series will be the log returns graphed day-to-day, the ACF measures how correlated a day's returns are compared to the previous day's returns. The variance is the variance of the returns. The MPS is a more complicated indicator and further details can be found at [6]. These indicators computed over "sliding windows" of time, which is detailed in the methods section.

Finally, we will be using the Mann-Kendall test. The Mann-Kendall test is a statistical test developed to determine whether time-series data is monotonically increasing, decreasing, or has no trend. It is a non-parametric hypothesis test, which works well, since financial data is not assumed to be normally distributed. To compute this, we will use an implementation from the pymannkendall package. Specific details about the test can be found in [8].

3 Data

We are work with stock market data publically available on Yahoo Finance. We use the Standard and Poor's 500 (S&P500), Dow Jones Industrial Average (DJIA), Nasdaq Composite (NASDAQ), and Russel 2000 (RUT). These are all relatively popular indices of the stock market, each meant to gauge the performance of the broader U.S. economy. More information about generally how they are calculated are available online. We also downloaded the Chicago Board Options Exchange Volatility Index (VIX). The VIX, as its name suggests, is meant to predict the market's near-future trends. It is based on the computed volatility of S&P500 and thus, forms a great basis for the comaprison of our methods. The VIX is not perfect, and there is plenty of criticism about its shortcomings, but presents a more-than-sufficient data set for our comparative purposes.

Whilst the data downloaded includes opening/closing price, high/low price, volume, etc, we mainly need to focus on closing price, more specifically the changes in closing price. Thus, we use apply the natural logarithm on the percentage change in closing price from day-to-day. This is standard across the papers I'm referencing and helps make the computational load easier due to the additive properties of log. When I refer to time-series of stock market data, what I am really referring to are these log daily returns as a function of a specific date.

The data that Yahoo finance gives us as an exported CSV excel file is well-formatted. *NaN* values are only featured in the VIX dataset. We drop these at the very beginning; they do not change the outcome of the results significantly, since they are infrequent.

In terms of pre-processing, the methods are relatively similar for both the 2020 and 2018 analysis. We first extract data for the 1300 days leading up to the respective crash. This comprises ~ 890 data points in reality as stock market data is not available on weekends. For 2018 cryptocurrency crash, we consider the start date as September 20, 2018 as that was the last day the S&P 500 peaked and similarly February 24, 2020 for the 2020 crash (These dates were taken from the wikipedia page [4]). Thus, the data of interest begins on February 28, 2015, or August 8, 2016 respectively. After computing the log daily returns for each day, we can begin to use our methods.

4 Problem and Method

Our central problem is determining whether the proposed indicators computed via topological methods work at actually predicting impending financial collapse. Below is our method for the data. More insight can be found in the commented code in Supplementary Materials.

Persistence Landscape Method Pipeline

(1) We start with 5 datasets - S&P500, DJI, NASDAQ, RUT, and VIX. Suppose we are trying to study the 2020 stock market crash. For each, we clean up the initial data to remove any missing values and extract the appropriate date range of data we need. We take data up to 1300 calendar days before the financial crash, which results in about 890 days of closing price data. We compute a log daily returns of each dataset. We then combine S&P500, DJI, NASDAQ, and RUT into a multi-dimensional time series, which we will call the 4-stock-aggregate. We leave VIX as its own time-series.

(2) Define a sliding windows w_i of size 50. Essentially, we take subsets of data between days $(i, i+50)$ as i ranges from 1 to 840 and store this in w_i. For each w_i, using the scikit-tda package, we use the pre-built function ripser to compute the persistence diagram of this subset of the 4-stock aggregate and VIX data. Next, we use the persim subpackage to convert each collection of persistence diagrams to a collection of persistence landscapes.

(3) We compute the L_1 and L_2 norms of these persistence landscapes, using another built-in scikit-tda function. As the norms are extremely small numbers and we onyl care about relativistic data, we use max-min-normalization force them into the interval $[0, 1]$.

(4) We define ACF and MPS helper functions. Details about their formulas can be found in [6]. Again, we utilize a sliding windows approach of size 500. We compute the critical slowing down factors for each 500-consecutive-long subset of our 4-stock-aggregate and VIX persistence landscape datasets. This results in about 340 day-long data points leading up to the actual date of market failure.

5 We use SeaBorn to graph line plots of the VAR, ACF, and MPS of both the 4-stock-aggregate and VIX CSD datasets to visualize them and compare. We also use the Mann-Kendall test here to see if the trends observed in these graphs are monotonic.

6 Repeat this process 1-5 for the 2018 cryptocurrency crash.

Below is an example of what the intermediate persistence diagrams and landscapes look like on a specific day.

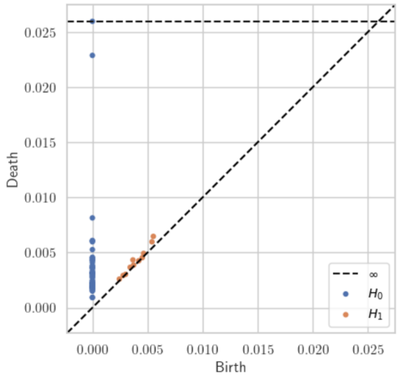

Figure 5: Persistence Diagram of the 4-aggregate-stock data before the 2020 crash from day 1 to 50

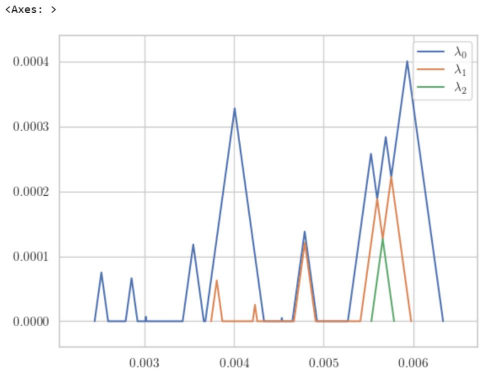

Figure 6: Persistence Landscape of the 4-aggregate-stock data before the 2020 crash from day 1 to 50

We are motivated to try topological methods on financial data due to their robustness against noise and small perturbations. The scikit-tda package is also relatively well-optimized, so it works fast despite needing to compute persistence diagrams/landscapes for many subsets of data.

This approach is largely following the model set in the original paper with slight modifications to verify whether it works. The solution to our problem will not be the most rigorous, but relies on visual analysis to see if the VIX CSD and 4-stock-aggregate are valid predictors of impending market failure. We want to observe trends that specifically occur right before the dates in question.

5 Results and Discussion

All charts generated and Mann-Kendall tests computed can be found labeled in the supplementary data's Additional Figures folder. We analyze a few specific figures here.

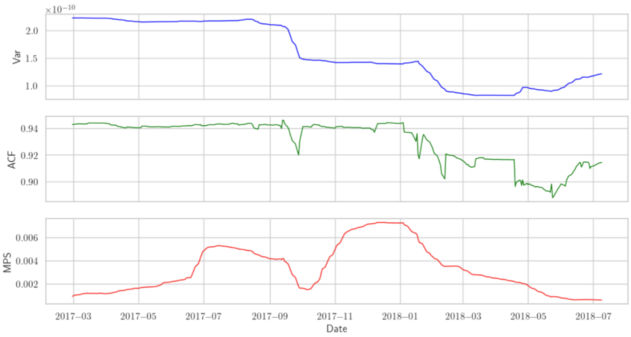

Figure 7: Plots of the critical slowing down factors leading up to the 2018 cryptocurrency crash in L_2

First, we will actually only talk about the charts computed from L_1 norms of the persistence landscapes. The reason for this is that the diagrams generated by the L_2 norms are nearly identical visually to the L_1 graphs (See Fig 7 & 8). The subtle differences can be observed in the actual magnitude in the y-axis, but the relative trend of the graphs are indistinguishable. This was also noted by the authors of both [5] and [6], but the striking similarity was more than I expected. Thus, for the rest of this section assume all charts are based on L_1.

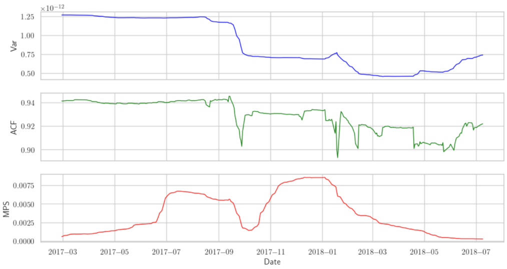

Figure 8: Plots of the CSD factors of the 4-stock-aggregate leading up to the September 20, 2018 cryptocrash

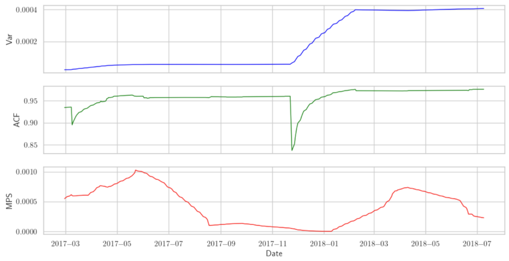

Figure 9: Plots of the CSD factors of the VIX leading up to the September 20, 2018 cryptocurrency crash

We can compare the 4-stock-aggregate versus VIX CSD plots prior to the 2018 cryptocurrency crash by looking at Figures 8 and 9. In 4-stock-aggregate, the all three CSDs are decreased ~ 8 months prior to the crash. On the other hand, we can see that before the crash, the Var and ACF of the VIX seem to have risen ~ 6 months prior. The MPS generally seems to demonstrate a small dip.

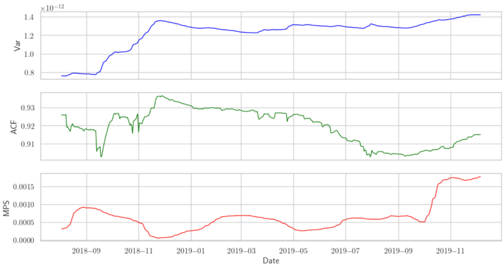

Figure 10: Plots of the CSD factors of the 4-stock aggregate leading up to the Feb 24, 2020 stock market crash

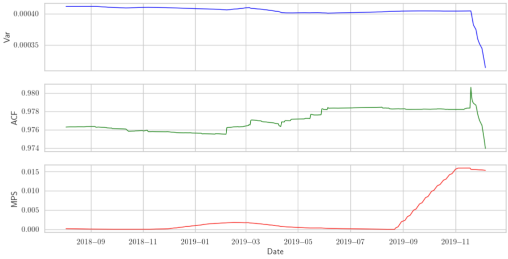

Figure 11: Plots of the CSD factors of the VIX leading up to the Feb 24, 2020 stock market crash

Based on Figure 10, the 4-stock-aggregate displays a dip in ACF prior to the crash and sharp rise in MPS prior to the crash. There seems to be no discernible trend in Var. In Figures 11, we find that VIX line plot displays a rapid decrease right before the 2020 crash in Var and ACF, while there is a sharp increase in MPS.

As far as the Mann-Kendall test results (See excel spreadsheet in Supplementary Data), they detected monotonic trends across nearly all of the graphs. The increasing/decreasing determinations also fall in line with our previous visual observations. The p-values of the test were all under 0.05 with one exception of MPS for the 4-stock-aggregate prior to 2020 in L_2. These trends stayed relatively consistent switching between across L_1 and L_2 representations, which further supports a conclusion that the CSDs changed prior to the failure event.

But, it does not seem that this change is markedly consistent on whether it is an increase of decrease in CSDs. Comparing across the 4-stock-aggregate graph before two different crashes, one featured a decrease in MPS, while the other featured a rise. This inconsistency does not seem promising and encourages the need for further testing. In fact, comparing in this manner, we only see a similarity in the dip in ACF prior to a crash when using the 4-stock-aggregate. For VIX, the situation trend prior to the two crashes seem even less promising as 2018 featured increased VAR, ACF and decreased MPS, whilst prior to 2020 featured the exact opposite changes. If this method were accurate, we'd expect similar trends.

6 Future Improvements/Research

Whilst I think the method developed is promising, there is large room for improvement. I didn't implement some of the more high-powered tests in [6]. It would be interesting to compute how the size of the sliding window effects the predictability power of this method. With a lower window comes less data, but also intuitively older data may not necessarily be relevant with predicting events farther (relatively) in the future. A weighted distribution of the importance of data based on relative time would be useful to perhaps generate more accurate predictions.

Additionally, the scope of previous studies seems particularly U.S.-centric, but we should analyze other foreign markets. We note that Singapore and Malaysia were analyzed by the authors in [6], but only for events that still corresponded to failures in the U.S (Of course, due to the nature of the global economy in this modern age, we can expect that market crashes in one will result in a decrease in another, especially for one as large as the U.S.'s). What about the analysis of financial failures that exist predominately in local markets? It is questionable whether this topological EWSs developed thus far can distinguish this.

At the same time, the classification of financial failures is not itself precise. The examples studied, such as 2008 real-estate collapse, 2000 dotcom crash, and 2018 cryptocurrency crash, have all been widely accepted as major financial events by economists and scholars alike. There is a debate of how significant a market failure needs to be to be classified as such and I'm curious whether these novel topological methods can detect the magnitude of such failure. So far, the methods seem aimed at predicting whether or not a failure will occur, rather than how significant it will be. This would be an interesting direction, but it isn't clear to me how you can expand the current methods.

In conclusion, I believe that topological data analysis provides a promising novel direction to apply to financial markets, but we should be cautious when doing so. Based on my results, it doesn't seem to be a particularly precise indicator, showing inconsistent patterns. The trends in MPS found in [5] did not appear during my testing. Unfortunately, it seems TDA currently requires more theoretical work and modification to establish its statistical rigor, before becoming a reliable tool.

References

[1] Bubenik, Peter. "Statistical topological data analysis using persistence landscapes." J. Mach. Learn. Res. 16.1 (2015): 77-102.

[2] Edelsbrunner, Herbert, and J Harer. Computational Topology : an Introduction. Providence, R.I.: American Mathematical Society, 2010.

[3] Ghrist, R. W. (2014). Elementary applied topology. Edition 1.0.

[4] "List of stock market crashes and bear markets." Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc, 22 July 2004, https://en.wikipedia.org/wiki/List-of-stock-market-crashes-and-bear-markets.

[5] Marian Gidea, Yuri Katz, *Topological data analysis of financial time series: Landscapes of crashes*, https://doi.org/10.1016/j.physa.2017.09.028.

[6] Mohd Sabri Ismail, etc., *Early warning signals of financial crises using persistent homology*, https://doi.org/10.1016/j.physa.2021.126459.

[7] Sci-kit documentation. https://docs.scikit-tda.org/en/latest/index.html

[8] Kendall, Maurice George. "Rank correlation methods." (1948).

[9] Koplik, Gary, *Persistent Homology: A Non-Mathy Introduction with Examples* https://towardsdatascience.com/persistent-homology-with-examples-1974d4b9c3d0

[10] Gowdridge, Trisan, etc., *On topological data analysis for structural dynamics: an introduction to persistent homology*, https://doi.org/10.48550/arXiv.2209.05134