TDA for financial time series: persistent homology and landscapes

HEMMER Arthur, SEMENENKO Aleksandra

Abstract

In this paper we replicate the experiments of a Topological Data Analysis method presented in TDA of financial time series: Landscapes of crashes by Marian Gidea and Yuri Katz (1). Specifically, we applied TDA to analyze financial data before the crises of 10/03/2000 (dotcom crash) and 15/09/2008 (Lehman bankruptcy). After reproducing the original results we experiment with different parameters and we propose a new method for calculating the distance between two persistent landscapes.

Introduction

In this paper we analyze topological properties of US stock closing prices around dotcom crisis of year 2000 and mortgage crisis of 2008. Our work followed the pipeline below:

1. Build a persistence diagrams of the Rips filtration
2. Build corresponding persistence landscapes
3. Calculate L1 and L2 norms from persistence landscapes
4. Perform standard techniques of time-series analysis on the norms: calculate variance and ACF
5. Repeat Steps 3. and 4. replacing the L^p norms with the Bottleneck distance

The structure of this paper follows the pipeline above, along with discussions about differences and similarities with the results of the original paper (1). All the code can be found in section Attachments.

Dataset

All experiments in this paper have been conducted on the daily stock closing prices of the DJIA, NASDAQ, Russell 2000 and S&P 500 from December 1987 until December 2016. Before running the experiments, the prices are converted to log-returns $R_i = \ln(P_i/P_{i-1})$ for every index. The resulting timeseries dataset consists of 7300 rows of 4 dimensions each.

Persistent Homology

In order to analyze the homology of a point cloud we must first create a filtration of simplicial complexes, sorted by a scaling parameter ϵ. As this scaling parameter changes, one can observe the birth and death of multidimensional holes in the point cloud which can be used to describe the persistence of such a hole. More specifically, the difference between the birth and death of a topological feature describes it's longevity, meaning that the greater the difference, the greater the significance of that feature.

After acquiring the birth/death rates we can visualize the significance of various features by constructing the corresponding persistent diagram and landscape. The diagram can be found by plotting the birth on one axis and death on the other. More significant features can be found further from the diagonal of the diagram as opposed to less significant features which are closer to the diagonal.

Using this diagram we can construct a persistence landscape by computing a sequence of continuous, piece-wise functions. This function is defined as follows:

$$f_{(b_\alpha,d_\alpha)} = \begin{cases} x - b_\alpha & \text{if } x \in (b_\alpha, \frac{b_\alpha+d_\alpha}{2}]; \\ -x + d_\alpha & \text{if } x \in (\frac{b_\alpha+d_\alpha}{2}, d_\alpha); \\ 0 & \text{if } x \notin (b_\alpha, d_\alpha) \end{cases}$$

Some key properties of persistent homology is that its results are robust to (minor) changes in the underlying data and it is able to separate signal from noise.

Persistence diagrams and landscapes

For persistence homology calculations we used GUDHI library with Python as the main language. We took several sliding windows (w = 40, 80, 120) before estimating persistence diagrams and landscapes for each date of interest. As we can see from Figure 1, the longer is the sliding window the more topological noise we have, which was expected to be the result, because with more historical data included in analysis we have less chances for loops and voids to remain until the current observation.

Regarding changes in persistence landscapes, the only obvious thing to notice was that with increase of the window size we would expect average life-time of the feature to expand. For the parameter value over which the feature exists we see that with expansion of the window general peaks remain the same, although it is not mandatory that parameter value, on average, will increase.

Resulting diagrams differed from the one in the original paper (1), however, the diagrams acquired from the analysis replicated the same general trend. Taking a look at the diagram for January 2000 with w = 40 the difference in terms of depicted persistence characteristics is almost indistinguishable, however the scaling of y-axis is 10 times less than in the original paper. Possible reason for this could be that we set a window of size w exactly w days before the date of interest, while in original paper (1) authors set the window "around" the investigated day, so we are not sure how they adjusted window to an observation. As was expected, the difference between our results and original paper becomes obvious with increase of the sliding window.

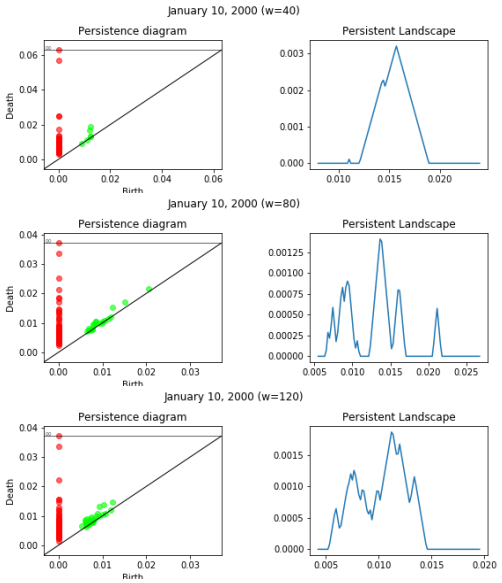

Figure 1: The different persistence diagram and landscape for the log-returns before January 10th, 2000 with window size of 40, 80 and 120 respectively.

Persistent Landscape Norms

We took windows of 40, 80 and 120 days to calculate the L-norms around both market failures (graphs are represented in Figures 3 and 4). Jump in L norms around the date of market failure is replicated by all graphs with different window sizes, which is the same result as in Section 4 of the original paper. The result is also quite intuitive: the bigger the bubble, the more topological features of historical data change, thus the distance between persistence landscapes

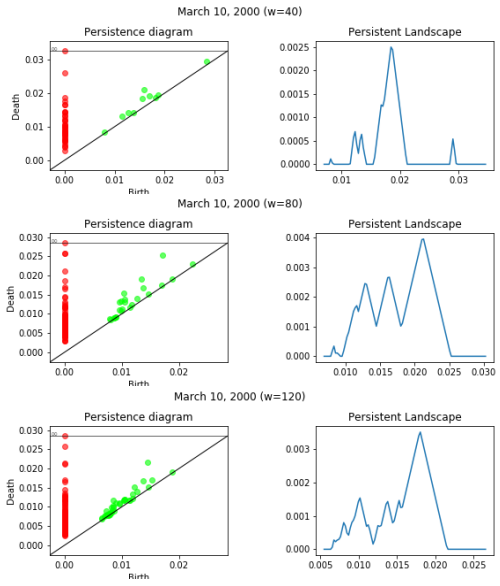

Figure 2: The different persistence diagram and landscape for the log-returns before March 10th, 2000 with window size of 40, 80 and 120 respectively.

grows, which is tracked bu corresponding jump in L^p norms. Also, we see that the L^p norms smooth when expanding the window size, but still represent the same trend. This can be seen in the growing amount of topological noise on the corresponding persistence diagrams, as with expansion of the window figures with higher persistence remain in the analysis.

One could notice the scaling differences with L1 and L2 norms estimated in (1). The original paper applied a method of normalization, but it was not specified which method was used with which parameters. Because of this we decided to use MinMax normalization to rescale the data to $[0, 1]$. Another reason for the scaling differences could be in window location, as was discussed earlier in the report.

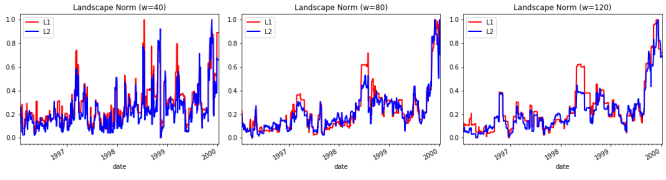

Figure 3: L1 and L2 norms before dotcom crisis

Further analysis will be performed solely with a window of 40 days for comparison with (1).

Bottleneck distance

As an alternative way to investigate financial market fluctuations with topological analysis tools, we decided to imple-

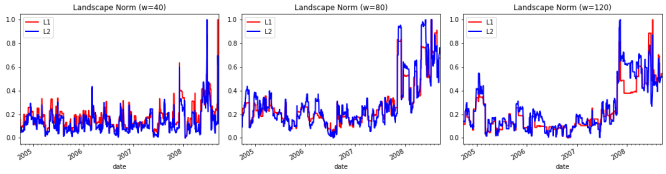

Figure 4: L1 and L2 norms before Lehman bankrupcy

ment bottleneck distance computation, which is a generalized version of the L1 and L2 norms:

$$L^q(Diag(f), Diag(g)) = \inf_{\gamma} \left(\sum_{u \in Diag(f)} \|u - \gamma(u)\|_{\infty}^{q} \right)^{\frac{1}{q}}$$

with $q \longrightarrow \infty$.

This procedure compares two persistence diagrams and returns the length of the longest edge for the best matching pairs. Bottleneck distance leads to a general stability results, but it is also sensitive to the worst edge out of best-matching and it implicitly assumes that critical points of the diagrams of interest are stable, which in practice usually is not the case. However, bottleneck distance is stable with respect to perturbations, which is why we decided to utilize it for the analysis.

The bottleneck distances calculation algorithm was computationally expensive, thus we could only provide analysis for a window of 40 days.

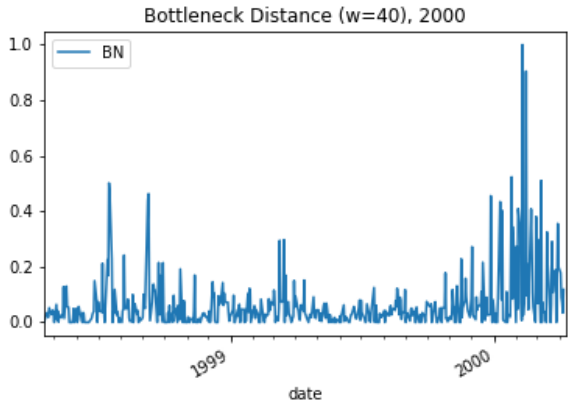

Figure 5: Bottleneck distances before dotcom crisis

We observe that L^p norms in 2000 behave more noisy than in 2008, the same tendency can be observed with bottleneck distance, with the only exception that the graph overall looks more well-behaved (we applied the same normalization technique). Bottleneck distances for thee 2000 and

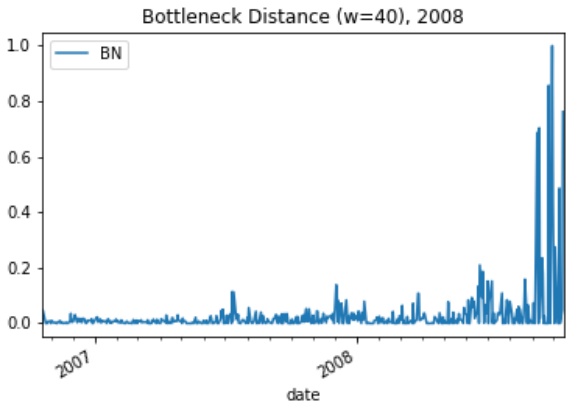

Figure 6: Bottleneck distances before Lehman bankruptcy

2008 crises are represented in Figures 5 and 6 correspondingly and clearly show a sharp increase starting from slightly before the day of market crash.

Conclusion

After seeing that the extracted topological features and their estimations are in line with (1) and show realistic trends, we proceed with examining statistical features of the time series.

In Figures 7 and 8 you can see graphs for variance and ACF(first order) estimation depending on the type of underlying time-series (L1, L2 norms and Bottleneck distance) and type of feature under estimation. We took the same sliding window as did authors in the original paper: 250 days (number of working days in a year).

For all variance graphs we see that there's a sharp rising trend around the date of a crisis, but L1 and L2 norms catch this trend earlier than Bottleneck distance - for example, for market failure of 15/09/2008 we see that L norms started obviously reacting since August and blew even larger in September, however Bottleneck distance graph shows a clearer reaction to a market bubble, but this reaction becomes distinguishable too late when it's already a middle of October.

Regarding the ACF graphs, it is hard to draw any conclusions from just a visual representation of the data. The only thing that we could notice is that ACF for Bottleneck distance shows a strong decline before the crisis of 2000 strikes. This is an interesting observation, because it shows that before the crisis stock indices become less and less correlated with their previous values, which could be a signal of an arising financial bubble that brings a different data behavior to the market.

To statistically assess the hypothesis of trends existence we run a Mann–Kendall test, which indicates whether there was a monotonic upward or downward movement of the

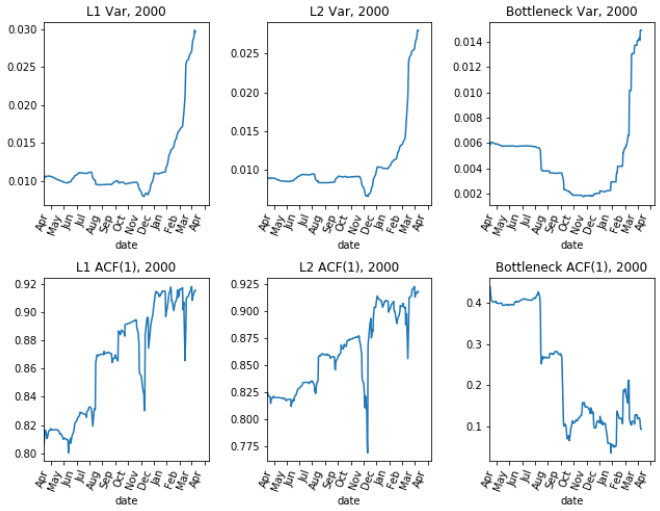

Figure 7: Bottleneck distances before Lehman bankruptcy

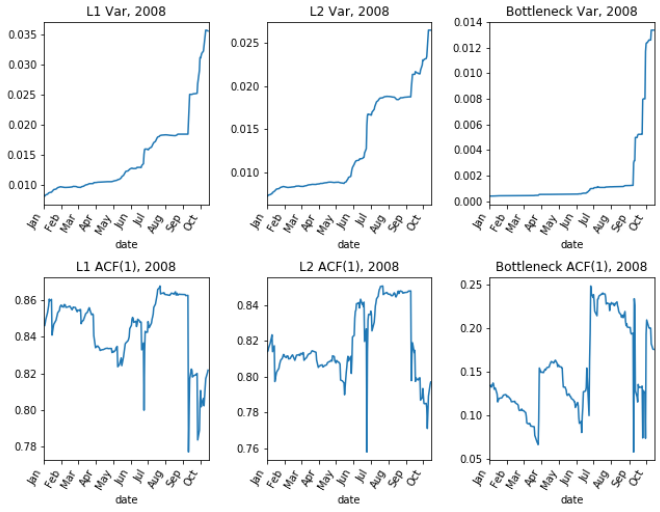

Figure 8: Bottleneck distances before Lehman bankruptcy

variance and ACF over the time. Results for variance are represented in Table 1 to Table 4:

Var:	L1	L2	Bottleneck distance
Trend	increasing	increasing	decreasing
P-value	0.0	0.0	$0.39 \cdot 10^{-5}$
Kendall tau	0.49	0.5	-0.12

Table 1: Mann–Kendall test for Var: year 2000

Trends for Var and ACF for L1 and L2 norms show statistically growing trend with high ACF average Kendall-tau correlation coefficients. Results for Bottleneck distance are more contradicting: for year 2000 that could be affected by

ACF:	L1	L2	Bottleneck distance
Trend	increasing	increasing	decreasing
P-value	0.0	0.0	0.0
Kendall tau	0.89	0.77	-0.45

Table 2: Mann–Kendall test for ACF: year 2000

Var:	L1	L2	Bottleneck distance
Trend	increasing	increasing	increasing
P-value	0.0	0.0	0.0
Kendall tau	0.67	0.76	0.74

Table 3: Mann–Kendall test for Var: year 2008

ACF:	L1	L2	Bottleneck distance
Trend	increasing	increasing	increasing
P-value	0.0	0.0	0.0
Kendall tau	0.65	0.43	0.37

Table 4: Mann–Kendall test for ACF: year 2008

variance decline prior to increase before the crisis. Also, we know that Bottleneck distance is more sensitive to highly volatile data, which is the case in year 2000.

For year 2008, although there's a statistical evidence of the presence of the trend in ACFs, we see that the correlation between data points is smaller than for year 2000, so it is still hard to draw any general conclusions from information in ACF graphs.

We study the persistence of loops in a 4D point cloud of daily log-returns of four stock-market indices. This method utilizes only one parameter: size of the sliding window to examine persistence homology. This procedure shows that there's a strong connection between state of the market and the shape of time series. When the market is approaching a bubble, persistence of loops expands to the period of crisis and the distance between consecutive persistence diagrams increases. Lp norms of persistence landscapes and the Bottleneck distance between persistence diagrams demonstrate a rise prior to dotcom crisis (10/03/2000) and Lehman bankrupcy crisis (15/09/2008). Variances of these indicators replicate the same trend. In our study the novel method (Bottleneck distance) of TDA for financial time series analysis proves to be less efficient than Lp norms due to late recognition of crisis and high bias for unstable data.

References

(1) Marian Gidea, Yuri Katz. *Topological data analysis of financial time series: Landscapes of crashes*. Physica A: Statistical Mechanics and its Applications, Elsevier, 2017.

(2) Peter Bubenik, Dlotko Pawel. *A persistence landscapes toolbox for topological statistics*. Journal of Symbolic Computation, Elsevier, 2016.

(3) Herbert Edelsbrunner, Dmitriy Morozov. *Persistent Homology: Theory and Practice*.