Topological Data Analysis of Financial Time Series

TDA Learning Seminar

Koundinya Vajjha

June 1, 2018

References

M. Gidea,Y .Katz.
Topological data analysis of financial time series: Landscapes of crashes.
Physica A: Statistical Mechanics and its Applications, 491:820 - 834, 2018.

J. Kim *et al*.
Introduction to the R package TDA.
http://arxiv.org/abs/1411.1830

V. Kovacev-Nikolic,P. Bubenik,D. Nikolec,G. Heo
Using persistent homology and dynamical distances to analyze protein binding.
arXiv:1412.1394v2 [stat.ME], 2015

Outline

Persistent Homology

Vietoris-Rips Filtration

Given point cloud data $X = \{x_1, \ldots, x_n\} \in \mathbb{R}^d$. Associate the Vietoris-Rips complex at resolution ϵ: $VR(X, \epsilon)$

- For each $k = 0, 1, \ldots$ a k-simplex of vertices $\{x_{i_1}, \ldots, x_{i_k}\}$ is in $VR(X, \epsilon)$ if and only if the mutual distance between any pair of vertices is less than ϵ.

$$d(x_{i_j}, x_{i_l}) < \epsilon \text{ for all } j, l$$

- A k-simplex is included in $VR(X, \epsilon)$ for every set of k data points that are indistinguishable from one another at resolution level ϵ.

Persistent Homology

Birth and Death

Given $X = \{x_1, \ldots, x_n\} \in \mathbb{R}^d$, if $\epsilon < \epsilon'$ then

$$VR(X, \epsilon) \subseteq VR(X, \epsilon')$$

and so

$$H_k(VR(X, \epsilon)) \hookrightarrow H_k(VR(X, \epsilon'))$$

for every k. Due to this, for every non-zero homology class α, there is a pair $b_\alpha = \epsilon_1 < \epsilon_2 = d_\alpha$ such that α is

- not in the image of any $H_k(VR(X, \epsilon'))$ for $\epsilon' < \epsilon_1$
- is non-zero in $H_k(VR(X, \epsilon'))$ for $\epsilon_1 < \epsilon' < \epsilon_2$ ("birth")
- is zero in in $H_k(VR(X, \epsilon'))$ for $\epsilon' > \epsilon_2$ ("death")

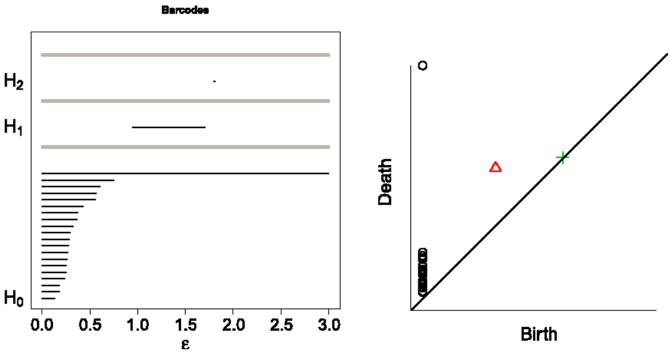
Persistence Diagrams

The information on the k-dimensional homology generators at all scales can be encoded in a "Persistence Diagram" P_k, which consists of:

- For each k-dimensional homology class α, one assigns a point $z_\alpha = (b_\alpha, d_\alpha) \in \mathbb{R}^2$, together with it's *multiplicity* $\mu(b_\alpha, d_\alpha)$ (the number of classes that are born at b_α and die at d_α).
- All points on the positive diagonal in $\mathbb{R}^2$: represents trivail homology generators that are born and instantly die at every level. Each point on the diagonal has infinite multiplicity.

Persistence Diagrams

Barcode and Diagram



Persistence Diagrams

Space of all Diagrams

- The space (multiset) of all such persistent diagrams $\mathcal{P}$ can be endowed with a metric W_p called the degree p Wasserstein distance ($p \geq 1$) or the Bottleneck distance ($p = \infty$).
- But these metric spaces $(\mathcal{P}, W_p)$ are not complete! Which is inconvenient for statistical purposes. (For SLLN and CLT type results.)
- A workaround is to embed the space $\mathcal{P}$ into the Banach Space $L^p(\mathbb{N} \times \mathbb{R})$ via *persistence landscapes*.

Persistence Landscapes

For each birth and death point $(b_\alpha, d_\alpha) \in \mathcal{P}_k$, first define

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

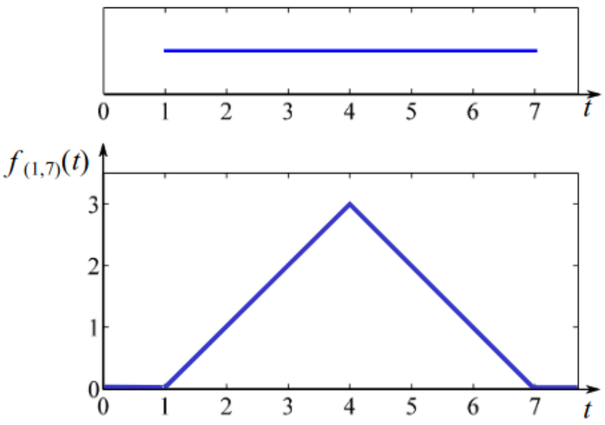
To a persistence diagram $\mathcal{P}_k$, we associate a sequence of functions $\lambda = (\lambda_n)_{n \in \mathbb{N}}$ where $\lambda_n : \mathbb{R} \to [0, 1]$ is given by

$$\lambda_j(x) = j - \max\{f_{(b_\alpha, d_\alpha)}(x) | (b_\alpha, d_\alpha) \in \mathcal{P}_k\}$$

where j-max denotes the j-th largest value of a function. $\lambda_k(x) = 0$ if the k-th largest value does not exist.

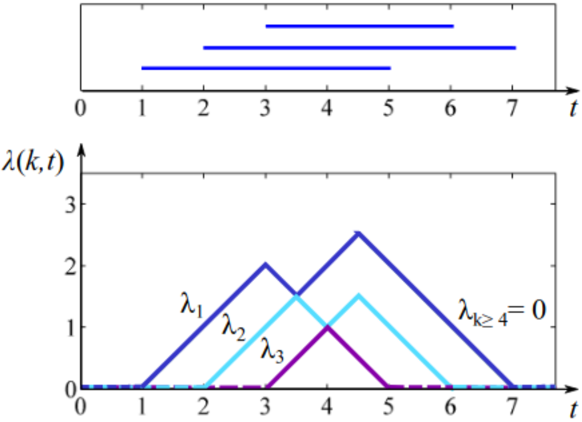
Persistence Diagrams

This is a picture of a function $f_{(1,7)}$ associated to a barcode. (Images taken from [3])



Persistence Landscapes

This is a picture of the persistence landscape associated to a barcode.(Images taken from [3])



Persistence Landscapes

- We have associated to a persistence diagram P_k a sequence of functions $\lambda = (\lambda_n)_{n \in \mathbb{N}} \in L^p(\mathbb{N} \times \mathbb{R})$ which is a Banach space.
- In general it is not possible to go back and forth between diagrams and landscapes.
- However, this whole exercise makes persistence landscapes suitable for treatment via statistical methods!

Henceforth, we shall only consider L^1, L^2 norms and only 1-dimensional homology.

TDA on Time Series

A *time series* is a series of data points indexed (or listed or graphed) in time order. Here are the general steps of the algorithm in [1].

- Consider d time series $\{x_n^k\}_n$, $k = 1, \dots, d$. So for each time instance t_n, we have a point $x(t_n) = (x_n^1, \dots, x_n^d) \in \mathbb{R}^d$.
- Pick a sliding window w. For each time-window of size w we get a point cloud data set consisting of w points in $\mathbb{R}^d$, namely $X_n = (x(t_n), x(t_{n+1}), \dots, x(t_{n+w-1}))$
- TDA is then applied on top of the time-ordered sequence of point clouds to study the time-varying topological properties of the multidimensional time series, from *window* to *window*.

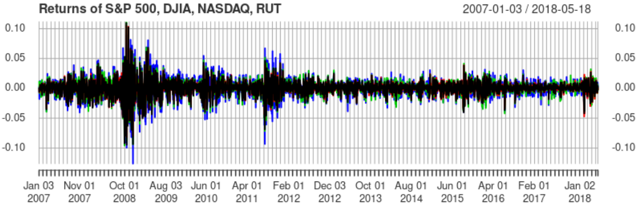
TDA on Time Series

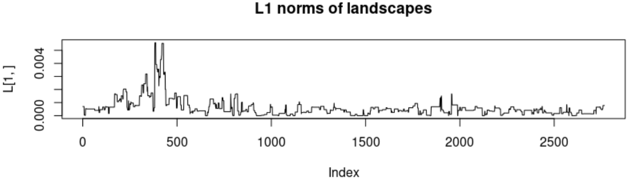
- For each point cloud, we compute the Vietoris-Rips Filtration, the corresponding persistence landscape, and it's L^p-norms for $p = 1, 2$.
- We plot the L^p-norms and observe how they behave around market crashes. General observation is the norms are sensitive to to transitions in the state of a system from regular to 'heated'.
- Using the R package "TDA", all this can be done in few lines of code!

Empirical Analysis of Stock Market Indices

I set out to replicate the results in the paper.

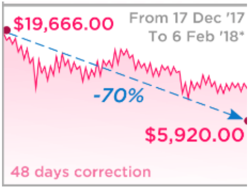
- Downloaded adjusted closing prices for four time series: S&P 500, NASDAQ, DJIA, Russel 2000. Calculated the log-returns.
- Sliding window length w=100 days.
- Applied TDA and plotted the L^1 and L^2 norms.


Returns of S&P 500, DJIA, NASDAQ, RUT
2007-01-03 / 2018-05-18
0.10
0.05
0.00
-0.05
-0.10
Jan 03 2007
Nov 01 2007
Oct 01 2008
Aug 03 2009
Jun 01 2010
Apr 01 2011
Feb 01 2012
Dec 03 2012
Oct 01 2013
Aug 01 2014
Jun 01 2015
Apr 01 2016
Feb 01 2017
Jan 02 2018


L1 norms of landscapes
L[1,]
0.000
0.004
0
500
1000
1500
2000
2500
Index

TDA on Cryptocurrencies

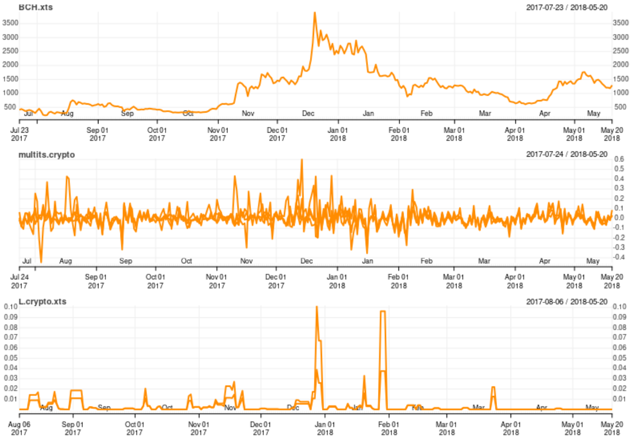
- The cryptocurrency market is extremely volatile - frequent crashes. Most cryptocurrencies seem to be highly correlated. Perfect candidate for TDA!



- Bitcoin lost nearly 70% between December 2017 and February 2018!
- What do the L^p norms show during this period?

TDA on Cryptocurrencies

Point cloud now consists of four cryptocurrencies: Bitcoin, Ethereum, Ripple and Bitcoin Cash.

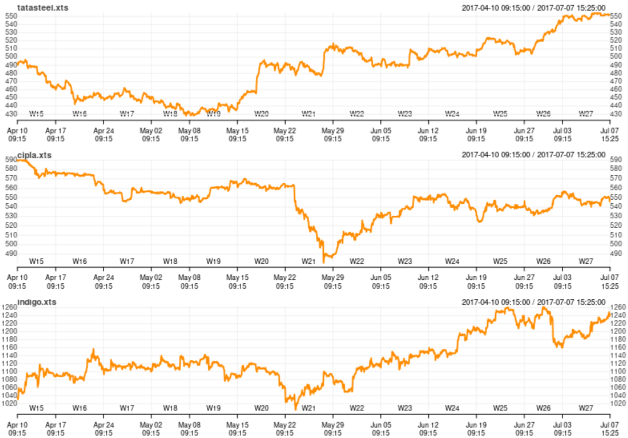


High-Frequency TDA

- High frequency data is time series of stock price data with intervals of a few minutes.
- Time Series Analysis is difficult and usually bears little resemblance to lower frequency data.
- Does TDA tell us anything for high frequency data?

High-Frequency TDA

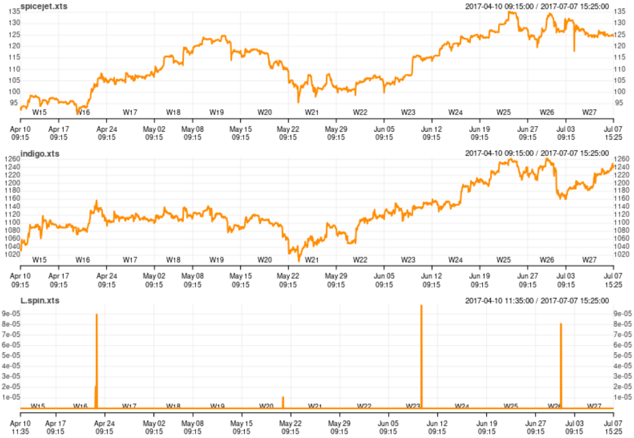
Point cloud data consists of 10 minute stock prices of five companies listed on the Bombay Stock Exchange: CIPLA, TATA STEEL, RELIANCE, INDIGO, SPICEJET



High-Frequency TDA

Results

Took the sliding window to be $b = 5$ days. This chart shows results for SPICEJET and INDIGO.



No conclusive findings!

Summary

- TDA for time series shows promise, however, robust justification for findings is needed to rule out correlation-causation fallacies.
- Further work
  - Does volatility in the markets cause topological patterns in the returns data? Do known models show this?
  - Can these empirical findings be explained by theory?