Topological Portfolio Evaluation

Alejandro Ortega, Andre Wang, Fred Xu

Duke University

May 1, 2018

Abstract

Portfolio optimization concerns the reallocation of resources while maximizing returns. Topological Data Analysis via Persistent Homology on time series of securities returns provide an interesting option to evaluate the performance of portfolios, by computing topological features that reveal an underlying structure to the data. We construct filtrations, persistence diagrams and a sliding window analysis to evaluate the topology of a risk-resilient and a non risk-resilient fund.

I. Introduction

Portfolio analysis plays a critical role in determining investment strategies by evaluating certain properties of a collection of financial assets over a period of time. Traditionally, investors consider various risk-return models based on historical returns and volatility in analyzing portfolios. Such models usually provide a reliable guideline of portfolio performance under stable market conditions. [4] [3] However, these models become far less dependable in times of significant changes in market conditions (e.g. market crash). Therefore, a challenging problem of practical interest is to analyze portfolios in critical market conditions. Specifically, it would be meaningful to evaluate risk resilience in major market declines. [2]

In this paper, we propose a new method to evaluate the risk resilience of portfolios by measuring changes in the topological structures of financial time-series data. Specifically, we are interested in finding out how evolutions of topological structures differ between risk resilient portfolios and non-risk resilient portfolios during market decline. For this purpose, we selected two comparable mutual funds offered by The Vanguard Group: Vanguard High Dividend Yield Index Fund (VHDYX) and Vanguard Selected Value Fund (VASVX). We chose the timeframe of analysis to be Jan/01/2008-Mar/31/2008 (61 trading days) when the market continued its major decline from the 2007 financial crisis. Although both funds focus on U.S. large-cap stocks, they demonstrated different risk tolerance during that period. Compared to the general market loss of 8.63%, VHDYX demonstrated risk resilience by only showing a loss of 6.65%, whereas VASVX showed a loss of 10.36% and thus was non-risk resilient. A preliminary comparison of key financial statistics such as Sharpe Ratio (0.64 vs. 0.67), however, does not reveal any significant distinctions that might explain the different risk tolerances of the two funds.

In our topological data analysis approach, we first constructed a weighted network of portfolio holdings for each of the two portfolios, where vertices are individual stocks and edge weight measure pairwise correlation between them. We then tracked the changes in topological features of this network during market decline.

II. Methods

Our goal is to create a filtration for the Rips complex so that we can find out more about the topological differences between the two mutual funds we are analyzing. In order to create this filtration, we employ a technique called the

sliding window technique. In this section we will explore more about the methodology we used in this paper.
We will first introduce some basic definitions.

Definition 1. *We denote t_1 as January 2nd 2008, and we denote t_{61} as March 31st 2008.*

The data we gathered consists of stock prices between January 2nd 2008 and March 31st 2008, and note that the stock prices are only available at weekdays.

Definition 2. *For a specific stock A, we define $p(t_i, A)$ to be its stock price at time t_i*

From Definition 1 and 2, if we use a sliding window with time horizon i.e window size equals to s, then the first window we have for stock A is define to be:

Definition 3.

$$\overrightarrow{win(t_1, A)} = (p(t_1, A), p(t_2, A), ..., p(t_s, A))$$

From Definition 3, we see that the second window for stock A is:

$$\overrightarrow{win(t_2, A)} = (p(t_2, A), p(t_3, A), ..., p(t_{s+1}, A))$$

In our later analysis that uses the sliding window technique, we choose the time horizon to be 5, this translates to as the data for a work week. The reason why we choose such a small time horizon is because there is empirical evidence against using large window size when non-stationary behavior is present. [1]. After we have developed the notion of what a window is in stock price data. We now proceed to define our network which is represented as simple undirected graph. Since we are only analyzing two mutual funds, VHDYX - Vanguard High Dividend Yield Index Fund and VASVX - Vanguard Selected Value Fund. We define the following:

Definition 4. *We denote the network of VHDYX as $\mathbf{G}_{VHDYX}$($\mathbf{V}$,$\mathbf{E}$), and analogously the network of VASVX= $\mathbf{G}_{VASVX}$($\mathbf{V}$,$\mathbf{E}$). The vertices $\mathbf{V}$ of the network of VHDYX correspond to each individual stock holdings of the mutual fund VHDYX. Each pair of distinct vertices A,B $\in$ $\mathbf{V}$ is connected by an edge $\overrightarrow{AB}$, and is assigned a weight $w(\overrightarrow{AB}, t_i)$ at time t_i.*

After introducing the notion of this weighted network, we want to define the weight function for each edge $\overrightarrow{AB}$ in the network. **Note:**The weight function is time dependent. Therefore, for a sliding window of time horizon equals to 5, we calculate the Pearson correlation at time t_i between

$$\overrightarrow{win(t_1, A)} = (p(t_i, A), \ldots, p(t_{\mathrm{t+4}}, A))$$

and

$$\overrightarrow{win(t_1, B)} = (p(t_i, B), \ldots, p(t_{\mathrm{t+4}}, B))$$

. Formally we have the following definition:

Definition 5. $C_{t_i}(A, B) =$

$$\frac{\sum_{k=i}^{k=i+4}[p(t_k,A)-\mu(\overrightarrow{win(t_i,A)}))][p(t_k,B)-\mu(\overrightarrow{win(t_1,B)}]}{\sqrt{\sum_{k=i}^{k=i+4}[p(t_k,A)-\mu(\overrightarrow{win(t_1,A)})]^2}\sqrt{\sum_{k=i}^{k=i+4}[p(t_k,B)-\mu(\overrightarrow{win(t_1,B)})]^2}}$$

where $\mu(\overrightarrow{window(t_i, A)})$ is the average stock price of stock A.

After calculating the Pearson correlation of each two vertices, we can define the weight function for each edge $\overrightarrow{AB}$ as following:

Definition 6.

$$\begin{cases} w(\overrightarrow{AB}, t_i) = C_{t_i}(A, B) & \textit{if } C_{t_i}(A, B) > 0 \\ w(\overrightarrow{AB}, t_i) = 0 & \textit{otherwise} \end{cases}$$

By assigning all negative Pearson correlation to have zero weight we are losing possibly useful information of anti-correlation. However, more than 99.5% of our data has positive correlation, and for those with negative correlation, their correlation has negligible magnitude, i.e. smaller than 0.05. Ideally, having the information about anti-correlation is useful, but one of the financial characteristics of mutual funds is that most of its stocks don't move in completely opposite directions. Therefore we choose to focus on stocks that move in the same direction.

By far we have defined the weight between 2 nodes, so we can move forward to calculate the distance matrix of a network at time t_i, we use $\mathbf{G}_{\mathrm{VHDYX}}$(**V**,**E**) as an example, the distance matrix can be similarly calculated for the other mutual fund.

Definition 7. *We denote the distance matrix of* $\mathbf{G}_{\text{VHDYX}}(\mathbf{V},\mathbf{E})$ *at time* t_i *to be* $\mathrm{D}_{t_i}(\mathrm{VHDYX})$.

$$\begin{cases} [D_{t_i}(VHDYX)]_{(i,j)} = w(\overrightarrow{ij}, t_i) & \textit{if } i \neq j \\ [D_{t_i}(VHDYX)]_{(i,j)} = 0 & \textit{otherwise} \end{cases}$$

From Definition 7 we see that the *(i,j)* entry of D_{t_i}(VHDYX) is the weight at time t_i between node i and node j in $\mathbf{G}_{\text{VHDYX}}(\mathbf{V},\mathbf{E})$, and we also set the diagonal to be zero to prevent self-loop. By definition $\mathrm{D}_{t_i}(\mathrm{VHDYX}) \in \mathbb{R}^{m,m}$ where m is *card(**V**)*, which is the number of stocks in mutual fund VHDYX.

Now the value of all entries of D_{t_i}(VHDYX) are between 0 and 1. We define the adjacency matrix A_{t_i}(VHDYX) from D_{t_i}(VHDYX) at time t_i and threshold x to be the following:

Definition 8.

$$\begin{cases} [A_{t_i}(VHDYX)]_{(i,j)} = 1 & [D_{t_i}(VHDYX)]_{(i,j)} \geq x \\ [A_{t_i}(VHDYX)]_{(i,j)} = 0 & \textit{otherwise} \end{cases}$$

This adjacency matrix A_{t_i}(VHDYX) encodes a graph with all the same vertex as $\mathbf{G}_{\text{VHDYX}}(\mathbf{V},\mathbf{E})$, but only keeps the edges that have a weight $\geq$ the threshold distance x. With the definition of distance and threshold, we can build a Rips complex from this adjacency matrix A_{t_i}(VHDYX). Note that this Rips complex is formed at time t_i, therefore, if we lower the value of the threshold x from 1 continuously to 0, we have a filtration $\mathcal{F}$ on the Rips complex of the entire network $\mathbf{G}_{\text{VHDYX}}(\mathbf{V},\mathbf{E})$.

We encode the information of filtration $\mathcal{F}$ in persistence diagrams $\mathbf{P}_{\text{VHDYX}}(t_i)$ where $\mathbf{P^j}_{\text{VHDYX}}(\mathbf{t_i})$ is the j^{th} dimension persistence diagram of this filtration $\mathcal{F}$ at time t_i. As it turns out looking at the persistent diagram alone won't give us too much information due to the noise cycles near the diagonal. In order to measure how the topological features of $\mathbf{G}_{\text{VHDYX}}(\mathbf{V},\mathbf{E})$ changes when time goes by, we endow a metric space structure upon the space of the diagrams at a specific time t_i. We choose to use the degree 1 Wasserstein distance to measure the distance between two diagrams. Specifically, in a specific dimension j we calculate the Wasserstein distance between the diagrams generated at each sequential time step, we then assign this distance to the smaller time step. Formally, we have:

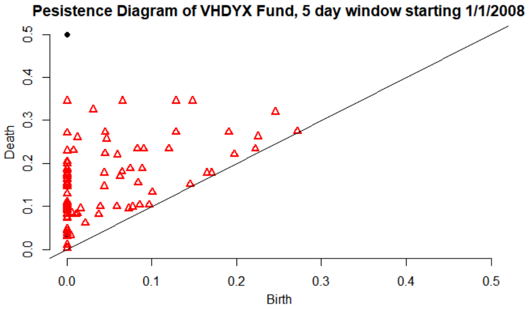


Figure 1: *Example of a Persistence Diagram.*
Black dot (●): 0 dimensional cycles
Red Triangle (△): 1 dimensional cycles

Definition 9.

$$f^j(t_i) = inf_{\phi}[\textstyle\sum_{q \in \mathbf{P^j}_{VHDYX}(\mathbf{t_i})} ||q - \phi(q)||_1]$$

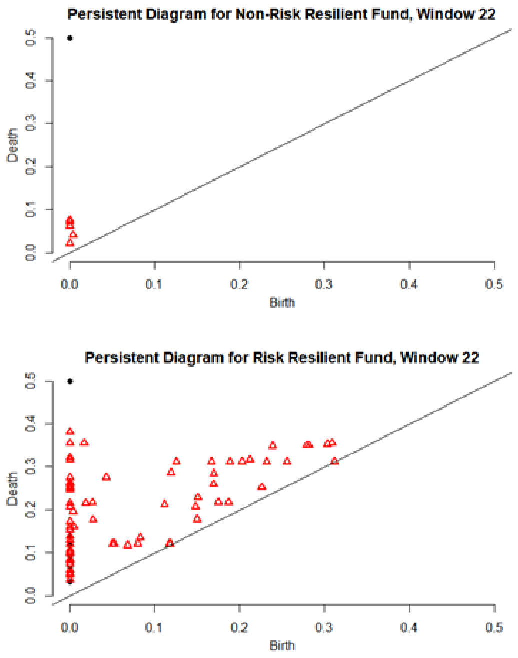
where the summation over all bijections ϕ : $\mathbf{P^j}_{VHDYX}(\mathbf{t_i}) \rightarrow \mathbf{P^j}_{VHDYX}(\mathbf{t}_{i+1})$

Note that this function f^j has domain $\in \mathbb{Z}$ and $\in [1,60]$. We can generate a similar function for mutual fund VASVX. Now that we have this time-series of Wasserstein distance for both mutual funds, we can plot them against time from January to March, and see if we can observe some difference in how the topological features change in each mutual fund.

III. Results and Discussion

We generated persistent diagrams of the two correlation networks in dimensions 0 and 1 through the filtration of Vietoris-Rips complex. Higher dimensional persistent homology is ignored as higher dimensional cycles are probably accidental. [4]

Figure 2 shows the persistent diagrams for both funds at a specific time window of our sliding window analysis when market experienced significant fluctuation. 0-dimensional cycles (connected components) are represented by black dots, while 1-dimensional cycles are represented by red triangles. Points along the



Persistent Diagram for Non-Risk Resilient Fund, Window 22

Persistent Diagram for Risk Resilient Fund, Window 22

Figure 2: *Persistence Diagrams for the Risk Resilient vs. Non-Risk Resilient Fund for Sliding Window 22*
Black dot (●): 0 dimensional cycles
Red Triangle (△): 1 dimensional cycles

identity line are likely due to noises in the dataset. In comparison, the persistent diagram for the non-risk resilient fund exhibits a lack of cycle activity, whereas the persistent diagram for the risk resilient fund shows high cycle activity in this particular window. However, this result is not consistent over all the time windows, so individual persistent diagrams do not provide as much insight.

We then sought out to capture the evolutions of persistent diagrams by computing the time-series representing the distances between the diagrams of a mutual fund against the general market. For each of the two mutual funds, we computed the 1-dimensional Wasserstein distances between the persistent diagram of that fund and a baseline fund at each time window and obtained the first plot of Figure 3. The Wasserstein distance between two persistence diagrams is the cost of the optimal matching between points of the two diagrams, and can be used to detect changes in the topology. The black line in the plot represents the risk resilient fund, and the red line represents the non-risk resilient fund. As shown, they mostly follow the same trend but do show differences at some points. However, the significance of these differences is limited, since the diagram distance in this case is only measuring how differently a fund is behaving relative to the market. But it does not indicate the direction of the difference. For example, on the time interval 0 to 10, the non-risk resilient fund's persistent diagrams are much further away from the market compared to those of the risk resilient fund, but that observation alone does not tell us anything about the relative performance of each portfolio.

Instead of computing the distances between persistent diagrams of different funds, we then computed the distance between t to $t+1$ increments of the sliding window to find changes in the topology within each fund as time progresses. As shown in the second plot of Figure 3, there is noticeably more volatility in the topology of the risk resilient fund. This observation suggests that a possible explanation for the higher performing stock lies in a volatile

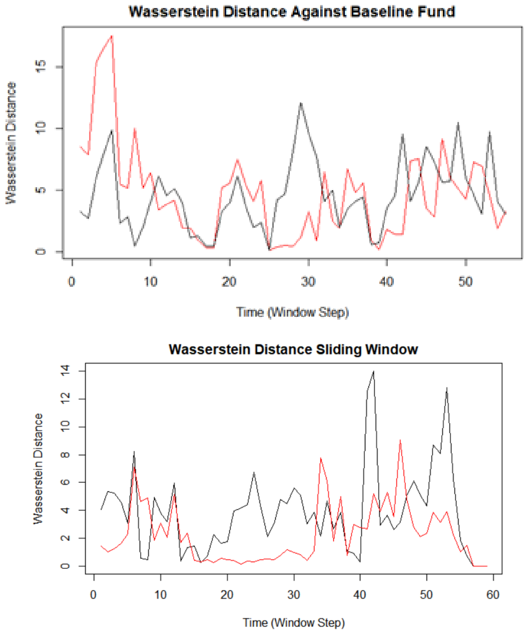


Figure 3: ***Wasserstein Distance Diagrams***
Black Line: Risk Resilient Fund
Red Line: Non-risk Resilient Fund

topology that can better adapt to the changes in the market. The analysis of the distances between persistent diagrams shows that a risk resilient portfolio experiences more predominant changes in the topological structure (especially connected components and 1-dimentional cycle) in times of drastic market downturn compared to a non-risk resilient portfolio. The lack of topological stability of a portfolio could be the very reason for a portfolio to successfully defend itself against poor market condition.

In the future, we hope to compute the persistent homology among more portfolios to test the statistical significance of the above finding. The most logical step in future work is in applying our analysis to a larger dataset that will enable a more thorough evaluation of portfolio performance using statistical methods and a more nuanced evaluation of portfolios.

Addressing portfolio selection, neural networks have been used to optimize the decision process that goes into reallocating resources while maximizing return. [5] Providing a mathematical formulation of the portfolio management problem, we can use this technique to better understand the selection of a portfolio from a machine learning perspective and better set up the portfolio-selection problem for a TDA approach. This paper also defines several metrics that we can use to evaluate our TDA approach to the stock selection problem, such modified Accumulated Portfolio Value that takes into account different starting times. Evaluating a portfolio via machine learning and then by TDA is a goal of future research.

References

[1] Peter Bubenik. Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.*, 16(1):77–102, January 2015.

[2] Marian Gidea. Topological data analysis of critical transitions in financial networks. In Erez Shmueli, Baruch Barzel, and Rami Puzis, editors, *3rd International Winter School and Conference on Network Science*, pages 47–59, Cham, 2017. Springer International Publishing.

[3] Mark Grinblatt and Sheridan Titman. Portfolio performance evaluation: Old issues and new insights. *The Review of Financial Studies*, 2(3):393–421, 1989.

[4] Michael C. Jensen. Risk, the pricing of capital assets, and the evaluation of investment portfolios. *The Journal of Business*, 42(2):167–247, 1969.

[5] Zhengyao Jiang, Dixing Xu, and Jinjun Liang. A deep reinforcement learning framework for the financial portfolio management problem. *CoRR*, abs/1706.10059, 2017.