Scientific Research Publishing

Journal of Mathematical Finance, 2020, 10, 648-678
https://www.scirp.org/journal/jmf
ISSN Online: 2162-2442
ISSN Print: 2162-2434

Topology Data Analysis Using Mean Persistence Landscapes in Financial Crashes

Alejandro Aguilar, Katherine Ensor

Department of Statistics, Rice University, Houston, USA
Email: aa109@rice.edu

How to cite this paper: Aguilar, A. and Ensor, K. (2020) Topology Data Analysis Using Mean Persistence Landscapes in Financial Crashes. *Journal of Mathematical Finance*, **10**, 648-678.
https://doi.org/10.4236/jmf.2020.104038

Received: September 26, 2020
Accepted: November 3, 2020
Published: November 6, 2020

Abstract

Topological features in high dimensional time series are used to characterize changes in stock market dynamics over time. We explored the daily log returns of four major US stock market indices and 10 *ETF* sectors between January 2010-June 2020. Topological data analysis and persistence homology were used on two sequences of point cloud data sets the stock indices and the *ETF* sectors, respectively. Using these sequences, the daily log returns, persistence diagrams, persistence landscapes, and mean landscapes were used to quantify topological patterns in the multidimensional time series. For example, norms of the persistence landscapes were generated to detect critical transitions in the daily log returns. To measure statistical significance, we implemented three permutation tests with a significance level $\alpha = 0.05$ to determine if topological features change within a particular time frame by comparing sliding windows in the sequence of point cloud data sets. We found that between July 1, 2019 and July 1, 2020, there is evidence of changing structure in the US stock market. Critical transitions are identified by the statistical properties of the norms of the persistence landscape between contiguous daily sliding windows of the stock indices and *ETF* sector series.

Keywords

Topological Data Analysis, Topological Time Series, Persistent Homology, Mathematical Finance

1. Introduction

Topological data analysis (TDA) extracts topological features by examining the shape of the data through persistent homology to produce topological summaries. Two topological summaries, the persistent barcode [1] [2], and the persistent diagram [3], provide visual representation of persistent topological features.

However, these topological summaries lack geometric properties and do not have a unique (Fréchet) mean [4], which makes it difficult to conduct statistical analysis and machine learning. In fact, Bubenik [5] states effective algorithms do not exist for computing means for a wide variety of examples, but he notes that [6] and [7] have made noteworthy advancement in this direction.

In Bubenik [5], the persistence landscape is provided as an alternative topological summary. The computation time is less for the persistence landscape than the persistence barcode and persistence diagram, because the persistence landscape is a sequence of piece-wise linear functions. Yet, the main advantage of using a persistence landscape is that they are situated in a separable Banach space, which means that we may use probability theory and random variables. After defining the persistence landscape and the norms of persistence landscapes, Bubenik [5] further develops his work by introducing the mean or average persistence landscape, which may be used for statistical analysis and again is something the persistence diagram and persistence barcode do not have. Furthermore, he proves many statistical properties for using persistence landscapes, such as convergence, stability, the Central Limit Theorem, and the Strong Law of Large Numbers (SLLN), which is important, so that one may conduct statistical inference. In particular, Bubenik [5] conducts a permutation test using multiple persistence landscapes to obtain a p-value (see Section 2.5, Section 2.6, Section 3.4), which is something that also cannot be done with persistence diagrams and persistence barcodes. This permutation test and the persistence landscape can be found in [8]. Other notable topological data analysis applications include the discovery of a subgroup of breast cancers [9], an understanding of the topology of the space of natural images [10], brain signals [11], and pre-clinical spinal cord injury [12].

With this alternative topological summary and the ability to conduct statistical inference, we bring our focus to critical transitions in complex dynamical systems, in particular, the financial market. Scheffer *et al.* [13] asserted that predicting critical transitions in a complex dynamical system prior to occurring is unreliable and very challenging, because the state of the complex system may not fluctuate substantially before reaching a critical threshold. However, in their work, they found for vast classes of systems, early warning signals may exist to indicate when a critical transition is imminent. Even though Scheffer *et al.* [13] presented examples of early warning signals in ecosystems, time series, climate dynamics, and epileptic seizures, there is not an example of an early warning signal for a financial crash. While Scheffer *et al.* [13] clarified that some predictability may be employed by experts, in general financial markets are complicated to predict. Although Scheffer *et al.* [13] referenced excellent works with financial early warning indicators, such as the VIX (volatility based index), systematic relationships in the variance and first order auto-correlation, and correlation increases across returns in falling markets, it was not relevant to our study, which leads us to research more about financial market dynamics.

Ensor and Koev [14] focused on the multivariate GARCH (MGARCH) and the hierarchical regime switching dynamic covariance (HRSDC) models, in which both models examined the co-variance structure within and between market sectors for the time period January 2, 1998 to December 2001. The HRSDC model provided early detection of several anomalous behaviors, such as the decline of Enron, the unusual returns for Silicon Graphics, and the fall of Lehmann Brothers. The detection of these anomalous behaviors is based on price movement of individual securities when viewed as a system of securities with the correlation within and between sectors. Therefore, Ensor and Koev [14] demonstrated a nested model is useful for understanding the correlation structure between different market sectors and how these sectors interacted as the market changes between regimes.

While Ensor and Koev [14] study prompted us to use *ETF* sectors in our study and is effective in identifying anomalous behaviors, such as the decline of Enron and the fall of Lehnman Brothers, our interest is to use TDA to detect early warning signals for financial crashes and examine topological features changing within time with statistical significance. A number of recent studies have explored the use of TDA on financial time series data to detect early warning signals of financial crashes. Gidea [2] analyzed the cross correlation network of the daily returns (adjusted closing prices) of the Dow Jones Industrial Average (DJIA) stocks listed as February 19, 2008 from January 2004 to September 2008. They tracked the topological changes when approaching a critical transition and showed some presence of early signs of a critical transition. On the other hand, Gidea *et al.* [15] analyzed four major cryptocurrencies (Bitcoin, Ethereum, Litecoin, and Ripple) before the beginning of 2018 and showed these cryptocurriences exhibiting highly erratic behavior. The paper introduced a method that combines TDA with machine learning to understand what happens before a critical transition. Moreover, they use Takens' theorem, the time delay embedding theorem, and C^1-norms of persistence landscapes. While the paper has valid analysis, our interest is in stocks and *ETF* sectors rather than cyprtocurrencies.

Alternatively, Gidea and Katz [16] investigated the daily log-returns of four stock indices (DJIA, S&P500, NASDAQ, and Russell 2000) from December 23, 1987 and December 08, 2016, where the topological properties of these stock indices were examined. This paper uses a sequence of a point cloud data set with a sliding window. Gidea and Katz [16] provided an excellent framework using persistence diagrams, persistence landscapes, and norms for persistence landscapes and we were able to replicate all of their results for 2000 and 2008 crashes. They demonstrated that the variance as defined in [17] shows rising trends, we are not convinced about the average spectral densities and auto-correlation function (ACF) with their associated Kendall-Mau tests demonstrated trends.

While these papers provide insightful groundwork for TDA in financial markets and cryptocurrencies, such as showing how to use cross correlation networks to track topological changes, using TDA with machine learning to understand what happens before a critical transition, and using the norms of persistence landscapes to indicate an approaching critical transition, these financial papers lack statistical inference. We are motivated to explore how the topological features change within a given time period for stocks and *ETF* sectors and find any statistical significant using a permutation test [5] [8], which we discuss in detail in Section 2.5, Section 2.6, and Section 3.4.

While we acknowledge the previous cited authors, we deem our contributions as an empirical framework that adapts their analytical models to new data sets and expand by conducting statistical inference. Similar to Gidea and Katz [16], we investigate the same four major indices (DJIA, S&P500, NASDAQ, and Russell 2000), but we extend our data set to include 10 *ETF* sectors (Consumer Discretionary, Consumer Staples, Energy, Financials, Health Care, Industrials, Materials, Information Technology, Utilities, and Index) for January 4, 2010-July 1, 2020 to examine their topological features to detect a critical transition or transitions. Moreover, we generate several topological summaries, norms for persistence landscapes $p=1$ and $p=2$, and conduct statistical inference on how these topological features change over time. In particular, we want to compare only sliding windows within a sliding step of one day from each other, which will be done separately for all the stock indices and for all the *ETF* sectors. We also compare all the stock indices against *ETF* sectors within the same sliding window. Our hypotheses tests will distinguish for two groups at a time if the means of topological features are the same either within a sliding step of one day in their respective sliding windows or within the same sliding window. The statistical tests of interest have not been seen before in any financial papers and will be our main contribution. The remainder of this paper is organized as follows.

In Section 2, we provide background information on algebraic topology, homology, constructing the Vietoris-Rips complex, persistent homology, topological summaries, norms of persistent landscapes, and statistical inference. In Section 3, we outline our methods for obtaining the data, constructing a sequence of a point cloud data, using persistent homology on a sequence of a point cloud data set, generating topological summaries, and performing statistical inference. In Section 4, we present our findings from our data. In Section 5, we discuss and provide an interpretation of our results. In Section 6, we conclude the paper.

2. Background

This study presents a topological data analysis of financial time series data. Here we provide background material about four relevant areas: algebraic topology, homology, topological summaries, and norms for persistent landscapes. We apply topological data analysis to a sequence of point cloud data sets to examine their topological properties within a point cloud matrix of d 1-dimensional time series. For our analysis, a sequence of point cloud data sets denoted X_n is shown below:

$$X_1 = \begin{bmatrix} x(t_1) \\ x(t_2) \\ \vdots \\ x(t_w) \end{bmatrix} = \begin{bmatrix} x_1^1 & \cdots & x_1^d \\ x_2^1 & \cdots & x_2^d \\ \vdots & \ddots & \vdots \\ x_w^1 & \cdots & x_w^d \end{bmatrix}$$

$$\vdots \tag{1}$$

$$X_q = \begin{bmatrix} x(t_q) \\ x(t_{q+1}) \\ \vdots \\ x(t_{q+w-1}) \end{bmatrix} = \begin{bmatrix} x_q^1 & \cdots & x_q^d \\ x_{q+1}^1 & \cdots & x_{q+1}^d \\ \vdots & \ddots & \vdots \\ x_{q+w-1}^1 & \cdots & x_{q+w-1}^d \end{bmatrix},$$

where each point in the sequence is expressed as $x(t_n) = (x_q^1, x_q^2, \cdots, x_q^d) \in \mathbb{R}^d$, d is the column number from a 1-dimensional time series, w is the sliding window size for a certain number of trading days (n_{td}) with a sliding step of one day, and $n = 1, 2, \cdots, q$. To obtain q, the difference is taken between the total number of days of the daily log returns (n_{dlr}) and one less than the sliding window size $w-1$, so that q becomes $q = n_{dlr} - (w-1)$ or $q = n_{dlr} - w + 1$. The total number of days of the daily log returns (n_{dlr}) is the total number of trading days (n_{td}) minus 1 or $n_{dlr} = n_{td} - 1$. To approximate the daily log returns, the formula is discussed in Section 3.1. So, every point cloud is compromised of a $d \times w$ matrix, where $w > d$ [16]. Note our method uses a sliding window w as seen in [16] and it does not apply the sliding window embedding theorem or Takens' theorem. In the next two subsections, we provide background information on algebraic topology and persistent homology, so that for every point cloud, we generate topological summaries and compute their L^p norms based on their corresponding persistence landscapes to conduct statistical inference. For a more in depth background, we refer readers to [3] [5] [18] [19] [20] [21].

2.1. Algebraic Topology

To produce topological summaries, we must first construct a Vietoris-Rips filtration for each point cloud in a sequence of point cloud data sets, which requires understanding simplices and simplicial complexes and are defined below [19]:

Definition 2.1 *Let* $\{a_0, \cdots, a_n\}$ *be a geometrically independent set in* $\mathbb{R}^N$. *We define the* **n-simplex** σ *spanned by* $a_0, \cdots, a_n$ *to be the set of all points x of* $\mathbb{R}^N$ *such that*:

$$x = \sum_{i=0}^{n} t_i a_i, \quad \text{where } t = \sum_{i=0}^{n} t_i = 1, \tag{2}$$

and $t_i \geq 0$ for all *i*.

Definition 2.2 *A* **simplicial complex** *K in* $\mathbb{R}^N$ *in a collection of simplices in* $\mathbb{R}^N$ *such that*:

- Every face of a simplex of *K* is in *K*.
- The intersection of any two simplexes of *K* is a face.

2.2. Homology

In homology, we are interested in a vector space $H_i(X)$ to a space *X* for each natural number $i \in \{0,1,2,\cdots\}$, because $H_i(X)$ counts the number of *k*-dimensional holes in *X*. For example, $H_0(X)$ counts the number of 0-dimensional holes or the number of connected components in *X*, while $H_1(X)$ counts number of 1-dimensional holes or the number of loops in *X*. Furthermore, the algebraic structures must be homotopy invariant, meaning they must not change through deformations. Yet, it is very challenging to determine the homology of arbitrary topological spaces, because it is computationally inefficient, so instead we approximate using simplicial complexes.

Now that simplicial complexes have been defined, we are introducing the p^{th} homology of a simplicial complex *K*. First, we denote the field with two elements as $\mathbb{F}_2$. Second, for a given simplicial complex *K*, we let $C_p(K)$ denote the $\mathbb{F}_2$-vector space with basis given by the *p*-simplices of *K*. Third, for any $p \in \{1,2,\cdots\}$, we define the linear map:

$$\partial_p : C_p(K) \to C_{p-1}(K) : \sigma \mapsto \sum_{\tau \subset \sigma, \tau \in K_{p-1}} \tau, \tag{3}$$

The kernel of $\partial_p : C_p(K) \to C_{p-1}(K)$ is the subgroup $\partial_p^{-1}(0)$ of $C_p(K)$ and is called the group of ***p*-cycles**. The image of $\partial_{p+1} : C_{p+1}(K) \to C_p(K)$ is the image ∂_{p+1} is the subgroup of $\partial_{p+1}(C_{p+1}(K))$ of $C_p(K)$ and is called the group of ***p*-boundaries** [19].

Definition 2.3 *For any* $p \in \{0,1,2,\cdots\}$, *the* $\boldsymbol{p^{th}}$ **homology of a simplicial complex** *K is the quotient vector space is defined as*:

$$H_p(K) = \text{kernel}(\partial_p) / \text{image}(\partial_{p+1}). \tag{4}$$

Its dimension is defined by:

$$\beta_p(K) := \dim H_p(K) = \dim \text{ kernel}(\partial_p) - \dim \text{ image}(\partial_{p+1}), \tag{5}$$

which is called the $\boldsymbol{p^{\text{th}}}$ **Betti number** of *K* [22].

The *p*-cycles that are not boundaries represent *p*-dimensional holes, which the p^{th} Betti number counts. For the p^{th} homology of a filtered simplicial complex *K*, we apply definition 2.3 and define as:

Definition 2.4 *Let K be a finite simplicial complex, and let* $K_1 \subset K_2 \subset K_3 \ldots \subset K_l = K$ *be a finite sequence of nested subcomplexes of K. The simplicial complex K with such a sequence of subcomplexes is called a* **filtered simplicial complex**. *The* p^{th} *persistent homology of K is the pair*

$$\left(\left\{H_p(K_i)\right\}_{1 \le i \le l}, \left\{f_p^{i,j}\right\}_{1 \le i \le j \le l}\right),$$

where $i, j \in \{1,\cdots,l\}$ for all $i \le j$, $f_p^{i,j} : H_p(K_i) \to H_p(K_j)$ are the linear maps induced by the inclusion maps $K_i \to K_j$ [22].

The p^{th} persistent homology of a filtered simplicial complex provides more information about the maps between each subcomplex than the homologies of single subcomplexes, which is explained further in Section 2.2.2. While there are several filtered simplicial complexes, such as the Cech, Alpha, and Delaunay, we chose the Vietoris-Rips complex, because it is computationally efficient [22].

2.2.1. Vietoris-Rips Construction

Definition 2.5 *Let* $X = \{x_1, \cdots, x_n\}$ *be a collection of points in* $\mathbb{R}^d$. *Given a distance* $\epsilon > 0$, $\mathcal{R}(X,\epsilon)$ *denotes the simplicial complex on n vertices* $x_1, \cdots, x_n$, *where an edge between the vertices* x_i *and* x_j *with* $i \neq j$ *is included if and only if* $d(x_i, x_j) \leq \epsilon$ *or generally the k-simplex are included with vertices* $x_{i_0}, \cdots, x_{i_k}$ *if and only if all of the pairwise distances are at most* ϵ. *This type of simplicial complex is called a* **Vietoris-Rips complex** [8] [16].

When $\epsilon < \epsilon'$, the Vietoris-Rips complex forms a filtration, $\mathcal{R}(X,\epsilon) \subseteq \mathcal{R}(X,\epsilon')$, which by definition 2.4 is a filtered simplicial complex. While there is no clear criteria for select ϵ', [23] used $\epsilon' = 0.05$ in their study. In this study, the Vietoris-Rips complex of X_n is denoted as $\mathcal{R}(X_n,\epsilon)$ and follows definition 2.5, where X_n is a sequence of point cloud data sets as given by Equation (1). Moreover, the filtration of $\mathcal{R}(X_n,\epsilon) \subseteq \mathcal{R}(X,\epsilon')$ is shown below:

$$\begin{gathered} \mathcal{R}(X_1,\epsilon) \subseteq \mathcal{R}(X_1,\epsilon') \\ \vdots \\ \mathcal{R}(X_q,\epsilon) \subseteq \mathcal{R}(X_q,\epsilon'), \end{gathered} \tag{6}$$

where q is the difference between the number of the daily log returns and the sliding window $(w+1)$ or $q = n_{dlr} - w + 1$. By definition 2.4, $\mathcal{R}(X_n,\epsilon)$ is a filtered simplicial complex.

2.2.2. Persistent Homology

Using definition 2.4 and definition 2.5, it is possible to find the p-dimensional homology of the Vietoris-Rips complex of X_n labelled as $H_p(\mathcal{R}(X_n,\epsilon))$ with coefficients in field $\mathbb{Z}/2\mathbb{Z}$ for small values of p and for different values of ϵ [8]. Recall from section 2.2, $H_p(K_i)$ is a vector space and $\beta_p(K_i)$ counts the number of p-dimensional holes. When $\epsilon < \epsilon'$, we apply definition 2.4 to the filtration $\mathcal{R}(X_n,\epsilon) \subseteq \mathcal{R}(X_n,\epsilon')$, which induces linear maps $f_p^{i,j} : H_p(\mathcal{R}(X_n,\epsilon)) \to H_p(\mathcal{R}(X_n,\epsilon'))$ as seen below:

$$\begin{gathered} H_p(\mathcal{R}(X_1,\epsilon)) \to H_p(\mathcal{R}(X_1,\epsilon')) \\ \vdots \\ H_p(\mathcal{R}(X_q,\epsilon)) \to H_p(\mathcal{R}(X_q,\epsilon')), \end{gathered} \tag{7}$$

where $q = n_{dlr} - w + 1$. Each $H_p(\mathcal{R}(X_n,\epsilon))$ is a vector space whose generators correspond to holes in $\mathcal{R}(X_n,\epsilon)$, and the linear maps $f_p^{i,j}$ allow us to track the generators from $H_p(\mathcal{R}(X_n,\epsilon)) \to H_p(\mathcal{R}(X_n,\epsilon'))$. A suitable basis is selected by applying the Fundamental Theorem of Persistence Homology.

Theorem 2.1 (Fundamental Theorem of Persistent Homology) *The Fundamental Theorem of Persistent Homology states there is a choice of basis vectors* $H_p(K_i)$ *for each* $i \in \{1, \cdots, l\}$ *such that each map is determined by a bipartite matching of basis vectors* [22].

Given Theorem 2.1, there is a choice of basis vectors of $H_p(\mathcal{R}(X_n,\epsilon))$, such that one may construct a well-defined and unique collection of disjoint half-open intervals, where a generator $x \in H_p(\mathcal{R}(X_n,\epsilon))$ corresponds to a half-open interval $[b_i,d_i)$, which represents the lifetime of *x*. The endpoints b_i and d_i refer to *x* first appearing and finally disappearing respectively in $\mathcal{R}(X_n,\epsilon)$. Specifically, if $x \neq 0$ is not in the image of $f_p^{b_{i-1},b_i}$, then *x* is born in $H_p(\mathcal{R}(X_n,\epsilon))$. Conversely, if $d_i > b_i$ is the smallest index for which $f_p^{b_i,d_i}(x) = 0$, then *x* dies in $H_p(\mathcal{R}(X_n,\epsilon))$. Persistence is determined by a generator's lifetime in the half-open interval, where a generator is considered more persistent the longer it appears in the half-open interval. If $f_p^{b_i,d_i}(x) = 0$ for all $b_i > d_i$ in I_j, then *x* lives forever, and its lifetime is represented by the interval $[b_i,\infty)$ [22]. Then, the set of vector spaces $H_p(\mathcal{R}(X_n,\epsilon))$ together with the corresponding linear maps is referred to as a persistence module, which is the foundation for constructing topological summaries.

2.3. Topological Summaries

To visualize, construct, and produce topological summaries, Theorem 2.1 is used to select the choice of basis vectors from $H_p(\mathcal{R}(X_n,\epsilon))$ and the corresponding linear maps $f_p^{b_i,d_i}$, in which all topological summaries are derived from the persistent modules.

2.3.1. Persistence Module, Persistence Barcode, Persistence Diagram

Definition 2.6 *A* **persistence module** *is defined as a vector space* M_α *for all* $a \in \mathbb{R}$ *and linear maps* $M(a \le b): M_a \to M_b$ *for all* $a \le b$ *such that*:

1) $M(a \le a)$ is the identity map;

2) For all $a \le b \le c, M(b \le c) \circ M(a \le b) = M(a \le c)$.

For additional information about the construction of a persistence module, see [5]. There are three main types of topological summaries associated with a persistence module. The first type of topological summary is a called a **barcode**. It represents a finite collection of disjoint half-open intervals I_j, in which each interval's endpoints are a birth-death pairs, (*b*) and (*d*) respectively. In particular, an interval starts with the time of birth (*b*) and ends with the time of death (*d*) of a topological feature. The ***p*th barcode** is denoted by $B_p = \{I_j\}$. A topological feature's survival or persistence is represented by the interval's length. The second type of topological summary is the ***p*th persistence diagram**, which is denoted as $D_p = \{(b_i,d_i)\}_{i \in I_j}$, where b_i and d_i are the bar codes intervals' end points and $-\infty < b_i < d_i < \infty$.

Unfortunately, the geometric properties of the barcodes and persistence diagrams present a difficult challenge for the calculation of means and variances, since two barcodes or two persistence diagrams may not have the same unique Friechet mean, which means statistical inference cannot be done. While the barcode and the persistence diagram are conventional topological summaries, Bubenik [5] showed how the persistence landscape is a better alternative.

2.3.2. Persistence Landscapes and Mean Landscape

Bubenik and Dlotko [18] proved numerous statistical properties of persistence landscapes that we may use for statistical inference, such as stability, convergence, central limit theorem, and strong law of large numbers. The persistent landscape and mean landscape are also used as topological summaries to indicate how persistence changes by examining the number of peaks. First, given a pair of numbers (b,d) with $b<d$, the piecewise linear (PL) function $f_{(b,d)}:\mathbb{R}\to[0,\infty]$ is defined by [18]:

$$f_{(b,d)}=\begin{cases}0 & \text{if } x\notin(b,d)\\ x-b & \text{if } x\in\left(b,\dfrac{b+d}{2}\right]\\ -x+d & \text{if } x\in\left(\dfrac{b+d}{2},d\right)\end{cases} \tag{8}$$

Second, given a persistence module, *M*, the persistence landscape may be defined as the function $\lambda:\mathbb{N}\times\mathbb{R}\to R$ given by:

$$\lambda(k,t)=\sup\left(h>0\mid \text{rank}M\,(t-h\le t+h)>k\right). \tag{9}$$

Third, given a persistence diagram $D_p=\{(b_i,d_i)\}_{i\in I}$ for $b<d$, $f(b,d)(t)=\max(0,\min(b+t,d-t))$, and the persistence landscape is defined as follows:

$$\lambda(k,t)=\text{kmax}\left\{f_p^{b_i,d_i}(t)\mid(b_i,d_i)\in D_p(t)\right\}_{i\in I}, \tag{10}$$

where kmax denotes the K^{th} largest element. Using Equation (10) for X_n, the persistence landscape of X_n denoted by $\lambda(X_n)$ is the following:

$$\begin{gathered}\lambda(X_1)=\text{k-max}\left\{f_p^{b_i,d_i}(X_1)\mid(b_i,d_i)\in D_p(X_1)\right\}_{i\in I}\\ \vdots\\ \lambda(X_q)=\text{k-max}\left\{f_p^{b_i,d_i}(X_q)\mid(b_i,d_i)\in D_p(X_q)\right\}_{i\in I},\end{gathered} \tag{11}$$

where $q=n_{dlr}-w+1$. This results in the following lemma from [5]:

Lemma 2.2

The persistence landscape has the following properties:

1) $\lambda_k(t)\ge 0$,

2) $\lambda_k(t)\ge\lambda_{k+1}(t)$, and

3) $\lambda_k(t)0$ is 1-Lipschitz.

From Equation (10), the persistence landscape is obtained and used to calculate the mean landscape, which is defined below:

Definition 2.7 *Let* $Y_1,\cdots,Y_n$ *be independent and identically distributed copies of Y, and let* $\Lambda^1,\cdots,\Lambda^n$ *be corresponding persistence landscapes. The* **mean landscape** $\bar{\Lambda}^n$ *is given by the point wise mean, in particular,* $\bar{\Lambda}^n(\omega)=\bar{\Lambda}^n$, *where*

$$\bar{\lambda}^n(k,t)=\frac{1}{n}\sum_{i=1}^{n}\lambda^i(k,t). \tag{12}$$

Using Equation (12) for X_n, we have the following:

$$\begin{gathered}\bar{\lambda}^n\left(X_1\right)=\frac{1}{n}\sum_{i=1}^{n}\lambda^i\left(X_1\right)\\ \vdots \\ \bar{\lambda}^n\left(X_q\right)=\frac{1}{n}\sum_{i=1}^{n}\lambda^i\left(X_q\right),\end{gathered}\tag{13}$$

where $q=n_{dlr}-w+1$. The mean landscape is used in section 2.5 and section 2.6.

2.4. Norms for Persistence Landscapes

Gidea and Katz [16] applied L^p norms of the persistence landscapes to identify the signs of a financial crash, which usually occurs within a time of high variance and cross-correlations among stocks or *ETFs*, and demonstrated that L^1 and L^2 norms of the persistence landscapes of four stock indices exhibited significant rising trends before the financial crashes. We adopt their approach in our study.

Therefore, for real valued functions on $\mathbb{R}\times\mathbb{R}$, for $1\le p<\infty$, *p*-norms of persistence landscapes are defined as:

$$\|\lambda\|_p=\sum_{i=1}^{\infty}\left[\int_{-\infty}^{\infty}\lambda_k\left(t\right)^p\,\mathrm{d}t\right]^{\frac{1}{p}},\tag{14}$$

and for $p=\infty$,

$$\|\lambda\|_\infty=\sup_{k,t}\lambda_k\left(t\right).\tag{15}$$

Applying Equation (14) to our sequence of point cloud data sets X_n results in:

$$\begin{gathered}\left\|\lambda\left(X_1\right)\right\|_p=\sum_{i=1}^{\infty}\left[\int_{-\infty}^{\infty}\lambda\left(X_1\right)^p\,\mathrm{d}t\right]^{\frac{1}{p}}\\ \left\|\lambda\left(X_q\right)\right\|_p=\sum_{i=1}^{\infty}\left[\int_{-\infty}^{\infty}\lambda\left(X_q\right)^p\,\mathrm{d}t\right]^{\frac{1}{p}},\end{gathered}\tag{16}$$

where $q=n_{dlr}-w+1$.

2.5. Statistical Inference: Part I

To compare the topological features between two groups, the persistence landscape is used to conduct a hypothesis test and statistical inference, which require several assumptions provided by [5]. First, the persistence landscapes lie in a separable Banach space $\mathcal{L}^p\left(\mathcal{S}\right)$ for $1\le p\le\infty$, where $\mathcal{S}=\mathbb{N}\times\mathbb{R}$. Second, Y is to be a random variable on some underlying probability space $\left(\Omega,\mathcal{F},P\right)$ with a corresponding landscape Λ. Third, if we have $\omega\in\Omega$, then $Y\left(\omega\right)$ is the random variable and $\Lambda\left(\omega\right)=\lambda\left(Y\left(\omega\right)\right):=\lambda$ is the corresponding topological summary statistic. To avoid confusion, we use Y instead of X as a random variable, because our sequence of point cloud data sets uses the variable X_n. In addition, Bubenik [5] proved the convergence of persistence landscapes using the

Strong Law of Large Numbers and the Central Limit Theorem, which is extremely important for setting up our random variables and hypothesis test. Our random variable Y is defined as:

$$Y = f\left(\lambda\left(k,t\right)\right) = \sum_k \int_{\mathbb{R}} \lambda_k\left(t\right)\mathrm{d}t, \tag{17}$$

where $f \in \mathcal{L}^b\left(\mathcal{S}\right)$ is a continuous linear functional, $\frac{1}{a}+\frac{1}{b}=1$, and Y satisfies the (SLLN) and (CLT) as seen in [5], which implies Y has an adequate sample size and follows an approximately normal distribution.

The statistical properties and definitions above are utilized to a conduct hypothesis tests with corresponding p-value based on a permutation test. To compare the topological features of two groups, Y_1 and Y_2, where k_1 and k_2 are samples taken from these groups respectively, and Λ_1 and Λ_2 are the corresponding landscapes respectively. The associate sample values of Y_1 and Y_2 are denoted as $y_1^1,\cdots,y_1^{k_1}$ and $y_2^1,\cdots,y_2^{k_2}$ and the corresponding landscapes of these sample values are labelled as $\lambda_1^1,\cdots,\lambda_1^{k_1}$ and $\lambda_2^1,\cdots,\lambda_2^{k_2}$. We apply Equation (17) to Y_1 and Y_2, so the functional of Y_1 and Y_2 are as follows:

$$\begin{aligned} Y_1 &= f\left(y_1^1\right),\cdots,f\left(y_1^{k_1}\right) = f\left(\lambda_1^1\left(k,t\right)\right),\cdots,f\left(\lambda_1^{k_1}\left(k,t\right)\right) = \sum_{i=1}^{k_1}\int_{\mathbb{R}}\lambda_1^i\left(k,t\right)\mathrm{d}t \\ Y_2 &= f\left(y_2^1\right),\cdots,f\left(y_2^{k_2}\right) = f\left(\lambda_2^1\left(k,t\right)\right),\cdots,f\left(\lambda_2^{k_2}\left(k,t\right)\right) = \sum_{i=1}^{k_2}\int_{\mathbb{R}}\lambda_2^i\left(k,t\right)\mathrm{d}t. \end{aligned} \tag{18}$$

Recall the sample mean is $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$, so the sample means of the Y_1 and Y_2 are the following:

$$\begin{aligned} \bar{Y}_1 &= \frac{1}{k_1}\sum_{i=1}^{k_1} f\left(y_1^i\right) = \frac{1}{k_1}\sum_{i=1}^{k_1} f\left(\lambda_1^i\left(k,t\right)\right) \\ \bar{Y}_2 &= \frac{1}{k_2}\sum_{i=1}^{k_2} f\left(y_2^i\right) = \frac{1}{k_2}\sum_{i=1}^{k_2} f\left(\lambda_2^i\left(k,t\right)\right), \end{aligned} \tag{19}$$

where again k_1 and k_2 are the samples taken from Y_1 and Y_2. We assume that μ_1 and μ_2 are the expectations of Y_1 and Y_2. So, μ_1 and μ_2 are assumed to be the population means of Y_1 and Y_2. Therefore, the statistical hypothesis is:

$$H_0 : \mu_1 = \mu_2 \quad H_a : \mu_1 \neq \mu_2. \tag{20}$$

To test the null-hypothesis, we use a two sample permutation test. Let

$$t = \frac{\left|\bar{Y}_1 - \bar{Y}_2\right|}{\sqrt{\frac{Var\left(Y_1\right)}{k_1}+\frac{Var\left(Y_2\right)}{k_2}}}. \tag{21}$$

Using Equation (21), $t_1,\cdots,t_m$ of the test statistic are calculated for permutations $s=1,\cdots,m$. The observed value of the test statistic is expressed as t_{observed}. The p-value is calculated by comparing t_{observed} with t_s and averaging the number of times $t_{\text{observed}} \le t_s$. Thus, Equation (21) becomes:

$$\begin{aligned} t_{\{1,Y_1,Y_2\}} &= \frac{\left|\bar{Y}_1 - \bar{Y}_2\right|}{\sqrt{\frac{Var(Y_1)}{k_1} + \frac{Var(Y_2)}{k_2}}} \\ &\vdots \\ t_{\{m,Y_1,Y_2\}} &= \frac{\left|\bar{Y}_1 - \bar{Y}_2\right|}{\sqrt{\frac{Var(Y_1)}{k_1} + \frac{Var(Y_2)}{k_2}}}. \end{aligned} \tag{22}$$

A general form of Equation (22) is:

$$t_{\{s,Y_1,Y_2\}} = \frac{\left|\bar{Y}_1 - \bar{Y}_2\right|}{\sqrt{\frac{Var(Y_1)}{k_1} + \frac{Var(Y_2)}{k_2}}}. \tag{23}$$

Hence, using Equation (22) and every instance where $t_{\text{observed}} \le t_s$, the *p*-value is obtained as:

$$p\text{-value}_{\{Y_1,Y_2\}} = \frac{1}{m}\sum_{i=1}^{m} t_{\{i,Y_1,Y_2\}}. \tag{24}$$

To measure the statistical significance, [8] used a significance level $\alpha = 0.05$ in their study, which we incorporate in our study. We may apply the above assumptions, equations, and definitions to compare the topological features of more groups.

2.6. Statistical Inference: Part II

Instead of conducting one hypothesis test, multiple hypotheses tests are conducted to determine how the topological features in our sequence of point cloud data sets X_n change within a particular time frame. The hypotheses tests are done on all the sliding window matrices within X_n. In particular, two adjacent sliding window matrices are compared, where adjacent means the sliding window matrices differ by a sliding step of one day. For example, the sliding window matrices X_1 and X_2 would be compared, while the sliding window matrices X_1 and X_3 would not be compared. Therefore, the assumptions, equations, and definitions from section 2.5 are applied to X_n. When hypotheses tests are performed, there are $q = n_{dlr} - w + 1$ random variables (see Equation (1)), which is also the size of the sequence of the point cloud data set X_n.

So, we let $Y_1, Y_2, \cdots, Y_q$ be random variables, where $k_1, k_2, \cdots, k_q$ are taken as samples from these groups respectively, and $\Lambda_1, \Lambda_2, \cdots, \Lambda_q$ are the corresponding landscapes respectively. The associate sample values of $Y_1, Y_2, \cdots, Y_q$ are denoted as $y_1^1, \cdots, y_1^{k_1}$, $y_2^1, \cdots, y_2^{k_2}$, $\cdots$, $y_q^1, \cdots, y_q^{k_q}$, and the corresponding landscapes of these sample values are labelled as $\lambda_1^1, \cdots, \lambda_1^{k_1}$, $\lambda_2^1, \cdots, \lambda_2^{k_2}$, $\cdots$, $\lambda_q^1, \cdots, \lambda_q^{k_q}$. The functional in Equation (17) is used to define the following for $Y_1, Y_2, \cdots, Y_q$:

$$\begin{gathered} Y_1 = \sum_{i=1}^{k_1} \int_{\mathbb{R}} \lambda_1^i \left(X_1 \right) \mathrm{d}t \\ \vdots \\ Y_q = \sum_{i=1}^{k_q} \int_{\mathbb{R}} \lambda_j^i \left(X_q \right) \mathrm{d}t, \end{gathered} \tag{25}$$

where $q = n_{dlr} - w + 1$. Recall the sample mean is $\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$, so the sample means of the $Y_1, Y_2, \cdots, Y_q$ as follows:

$$\begin{gathered} \overline{Y}_1 = \frac{1}{k_1} \sum_{i=1}^{k_1} f\left(\lambda^i \left(X_1 \right) \right) \\ \overline{Y}_q = \frac{1}{k_q} \sum_{i=1}^{k_q} f\left(\lambda^i \left(X_q \right) \right), \end{gathered} \tag{26}$$

where $q = n_{dlr} - w + 1$. We assume that $\mu_1, \mu_2, \cdots, \mu_q$ are the expectations of $Y_1, Y_2, \cdots, Y_q$. So, $\mu_1, \mu_2, \cdots, \mu_q$ are assumed to be population means of $Y_1, Y_2, \cdots, Y_q$, and the statistical hypotheses are:

$$\begin{gathered} H_0 : \mu_1 = \mu_2 \quad H_a : \mu_1 \neq \mu_2 \\ \vdots \\ H_0 : \mu_{q-1} = \mu_q \quad H_a : \mu_{q-1} \neq \mu_q, \end{gathered} \tag{27}$$

where $q = n_{dlr} - w + 1$. To test the null-hypothesis, we use a two sample permutation test with statistics,

$$\begin{gathered} t_{\{Y_1, Y_2\}} = \frac{\left| \overline{Y}_1 - \overline{Y}_2 \right|}{\sqrt{\frac{Var\left(Y_1\right)}{k_1} + \frac{Var\left(Y_2\right)}{k_2}}} \\ \vdots \\ t_{\{Y_{q-1}, Y_q\}} = \frac{\left| \overline{Y}_{q-1} - \overline{Y}_q \right|}{\sqrt{\frac{Var\left(Y_{q-1}\right)}{k_{q-1}} + \frac{Var\left(Y_q\right)}{k_q}}}. \end{gathered} \tag{28}$$

where $q = n_{dlr} - w + 1$. Using Equation (28), $t_1, \cdots, t_m$ of the test statistic are calculated for permutations $s = 1, \cdots, m$. The observed value of the test statistic is expressed as t_{observed}. The *p*-value is calculated by comparing t_{observed} with t_s and averaging the number of times $t_{\text{observed}} \leq t_s$. Using Equation (23), Equation (28) becomes:

$$\begin{gathered} t_{\{s, Y_1, Y_2\}} = \frac{\left| \overline{Y}_1 - \overline{Y}_2 \right|}{\sqrt{\frac{Var\left(Y_1\right)}{k_1} + \frac{Var\left(Y_2\right)}{k_2}}} \\ \vdots \\ t_{\{s, Y_{q-1}, Y_q\}} = \frac{\left| \overline{Y}_{q-1} - \overline{Y}_q \right|}{\sqrt{\frac{Var\left(Y_{q-1}\right)}{k_{q-1}} + \frac{Var\left(Y_q\right)}{k_q}}}, \end{gathered} \tag{29}$$

where $q = n_{dlr} - w + 1$. Hence, using Equation (29) and every instance where $t_{\text{observed}} \le t_s$, the p-value is obtained as:

$$p\text{-value}_{\{Y_1,Y_2\}} = \frac{1}{m}\sum_{i=1}^{m} t_{\{i,Y_1,Y_2\}}$$
$$p\text{-value}_{\{Y_{q-1},Y_q\}} = \frac{1}{m}\sum_{i=1}^{m} t_{\{i,Y_{q-1},Y_q\}}, \tag{30}$$

where $q = n_{dlr} - w + 1$. In our study, we also conduct hypotheses tests between two sequences of point cloud data sets, X_n^1 and X_n^2, within the same sliding window, so using the same assumptions, definitions, and results from this section. The only difference is a change in subscripts and superscripts. This case is presented in Section 3.4.

3. Methods

In this section, we describe the methods to obtain the data and analyze the financial time series using topological data analysis, statistical inference, and RStudio [23]. The data, which were obtained from Yahoo Finance, consisted of daily adjusted closing prices (amended for corporate actions such as stocks and dividends) for four major US stock indices: S&P 500, DJIA, NASDAQ, and Russell 2000 and 10 *ETF* sectors between January 4, 2010 and July 1, 2020 (2641 trading days). During this time period, a decline in the daily log returns happened on March 16, 2020. In order to examine this date of interest, we limited our data sets to 1001 trading days (n_{td}) before March 16, 2020 to observe any patterns in the L^p norms and determine any critical thresholds. To analyze the data, we first approximated the daily log returns of the adjusted closing prices. A return is defined as $r_t = \left(\frac{x_t - x_{t-1}}{x_{t-1}}\right)$, where x_t is the actual value (adjusted closing price) of the desired stock index or *ETF* sector. The daily log returns are defined as:

$$\log_{10}\left(\frac{x_t}{x_{t-1}}\right) = \log_{10}(x_t) - \log_{10}(x_{t-1}) \approx r_t,$$

which is an approximation of a return [24]. Since the daily log returns are forward daily changes, then the time frame of the daily log returns is from January 5, 2010 to June 30, 2020.

3.1. Point Cloud Data

After approximating the daily log returns, we designed two sequences of point cloud data sets, each with a sliding window of $w = 50$ and a sliding step of one day, which is based on the same method found in [16]. The first sequence of point cloud data set denoted by X_n^{SI} examined the four major US stock indices ($d = 4$), which resulted in a 4 × 50 matrix for each individual point cloud for a total of $q = n_{dlr} - w + 1 = (1001 - 1) - (50 - 1) = 951$ point clouds as seen below from using Equation (1):

$$X_1^{SI}=\begin{bmatrix} x(t_1)\\ x(t_2)\\ \vdots \\ x(t_{50}) \end{bmatrix}=\begin{bmatrix} x_1^1 & \cdots & x_1^4\\ x_2^1 & \cdots & x_2^4\\ \vdots & \ddots & \vdots\\ x_{50}^1 & \cdots & x_{50}^4 \end{bmatrix}$$
$$\vdots \tag{31}$$
$$X_{951}^{SI}=\begin{bmatrix} x(t_{951})\\ x(t_{952})\\ \vdots \\ x(t_{1000}) \end{bmatrix}=\begin{bmatrix} x_{951}^1 & \cdots & x_{951}^4\\ x_{952}^1 & \cdots & x_{952}^4\\ \vdots & \ddots & \vdots\\ x_{1000}^1 & \cdots & x_{1000}^4 \end{bmatrix}.$$

The second sequence of point cloud data set denoted by X_n^{ETF} examined the 10 *ETF* sectors ($d=10$), which yielded a 10 × 50 matrix for each single point cloud for a total of $q=n_{dlr}-w+1=951$ point clouds as seen below from using Equation (1):

$$X_1^{ETF}=\begin{bmatrix} x(t_1)\\ x(t_2)\\ \vdots \\ x(t_{50}) \end{bmatrix}=\begin{bmatrix} x_1^1 & \cdots & x_1^{10}\\ x_2^1 & \cdots & x_2^{10}\\ \vdots & \ddots & \vdots\\ x_{50}^1 & \cdots & x_{50}^{10} \end{bmatrix}$$
$$\vdots \tag{32}$$
$$X_{951}^{ETF}=\begin{bmatrix} x(t_{951})\\ x(t_{952})\\ \vdots \\ x(t_{1000}) \end{bmatrix}=\begin{bmatrix} x_{951}^1 & \cdots & x_{951}^{10}\\ x_{952}^1 & \cdots & x_{952}^{10}\\ \vdots & \ddots & \vdots\\ x_{1000}^1 & \cdots & x_{1000}^{10} \end{bmatrix}.$$

3.2. Vietoris-Rips Complex and Persistent Homology

Next, we constructed Vietoris-Rips complexes and filtration for each point cloud in X_n^{SI} and X_n^{ETF} from definition 2.4, definition 2.5, and Equation (6) and *R*-package "TDA" [25]. The Rips filtration for all the stock indices and all the *ETFs* are denoted by $R\left(X_n^{SI},\epsilon\right)$ and $R\left(X_n^{ETF},\epsilon\right)$ respectively for $\epsilon>0$. For the maximum filtration, we used $\epsilon'^{SI}=0.055$ and $\epsilon'^{ETF}=0.08$, which are based on similar methods found in [16]. Therefore, we obtained the following Rips filtration:

$$R\left(X_n^{SI},\epsilon\right)=R\left(X_n^{SI},0\right)\subset\cdots\subset R\left(X_n^{SI},0.055\right), \tag{33}$$

$$R\left(X_n^{ETF},\epsilon\right)=R\left(X_n^{ETF},0\right)\subset\cdots\subset R\left(X_n^{ETF},0.08\right), \tag{34}$$

where $n=1,\cdots,951$. Based on the Equations (6), (33), and (34), we computed only the $p=1$ dimensional homology $H_1\left(R\left(X_n,\epsilon\right)\right)$ with coefficients in the field $\mathbb{Z}/2\mathbb{Z}$ from Equation (7) as follows:

$$H_1\left(R\left(X_n^{SI},0\right)\right)\to H_1\left(R\left(X_n^{SI},0.055\right)\right), \tag{35}$$

$$H_1\left(R\left(X_n^{ETF},0\right)\right)\to H_1\left(R\left(X_n^{ETF},0.08\right)\right), \tag{36}$$

where $n=1,\cdots,951$. Also, we are only interested in the persistence of loops in as they appear in each point cloud during the transition states of the market, which is why we did the first dimensional homology. From definition 2.4, the filtration from Equations (33) and (34) induced a sequence of linear maps $f_1^{b_i,d_i,SI}: H_1\left(R\left(X_n^{SI},0\right)\right) \to H_1\left(R\left(X_n^{SI},0.055\right)\right)$ and $f_1^{b_i,d_i,ETF}: H_1\left(R\left(X_n^{SI},0\right)\right) \to H_1\left(R\left(X_n^{SI},0.08\right)\right)$. The images of these maps are the persistent homology groups. The collection of vector spaces $H_1\left(R\left(X_n^{SI}\right)\right)$ and $H_1\left(R\left(X_n^{ETF}\right)\right)$ along with the corresponding linear maps is a persistent module, which leads us to the topological summaries.

3.3. Topological Summaries

By modifying the *R* script in [5], the first dimensional persistence diagrams denoted by $D_1\left(X_n^{SI}\right)=\left\{b_i,d_i\right\}_{i\in I}$ and $D_1\left(X_n^{ETF}\right)=\left\{b_i,d_i\right\}_{i\in I}$ for each point cloud data set were used along with Equations (10) and (11) to produce the analogous first dimensional persistent landscapes $\lambda\left(X_n^{SI}\right)$ and $\lambda\left(X_n^{ETF}\right)$ as seen below:

$$\lambda\left(X_n^{SI}\right)=\text{k-max}\left\{f_1^{b_i,d_i,SI}\left(X_n^{SI}\right)\mid\left(b_i,d_i\right)\in D_1\left(X_n^{SI}\right)\right\}, \tag{37}$$

$$\lambda\left(X_n^{ETF}\right)=\text{k-max}\left\{f_1^{b_i,d_i,ETF}\left(X_n^{ETF}\right)\mid\left(b_i,d_i\right)\in D_1\left(X_n^{ETF}\right)\right\}, \tag{38}$$

where $n=1,\cdots,951$. Next, the norms of the persistence landscapes $\left\|\lambda\left(X_n^{SI}\right)\right\|_p$ and $\left\|\lambda\left(X_n^{ETF}\right)\right\|_p$ were computed for $p=1$ and $p=2$ using Equations (14)-(16). The norms of the persistence landscapes and the daily log returns were plotted in juxtaposition, where it is important to remember that a point in the norms of persistence landscapes refers to a sliding window of 50 trading days in the daily log returns. After generating the topological summaries, the mean landscape is constructed using definition 2.7 and Equations (12)-(13) for the time period between July 1, 2019 and July 1, 2020 (253 trading days) or the time frame between July 2, 2019 to June 30, 2019 (252 days for the daily log returns). For this reason, the sequences of point cloud data sets will go from $q=951$ to $q=203$. Also, recall that the daily log returns are forward daily changes, so the time frame of the daily log returns are from July 2, 2019 to June 30, 2020. We assigned $\lambda_1^{1^{SI}},\cdots,\lambda_1^{k_1^{SI}},\cdots,\lambda_q^{1^{SI}},\cdots,\lambda_q^{k_q^{SI}}$ and $\lambda_1^{1^{ETF}},\cdots,\lambda_1^{k_1^{ETF}},\cdots,\lambda_q^{1^{ETF}},\cdots,\lambda_q^{k_q^{ETF}}$ to be the corresponding landscapes for all the point clouds in X_n^{SI} and X_n^{ETF} to obtain the mean landscapes as seen below:

$$\bar{\lambda}\left(X_n^{SI}\right)=\frac{1}{k^{SI}}\sum_{i=1}^{k^{SI}}\lambda^i\left(X_n^{SI}\right), \tag{39}$$

$$\bar{\lambda}\left(X_n^{ETF}\right)=\frac{1}{k^{ETF}}\sum_{i=1}^{k^{ETF}}\lambda^i\left(X_n^{ETF}\right), \tag{40}$$

where $n=1,\cdots,203$ for $1,\cdots,k^{SI}$ and $1,\cdots,k^{ETF}$ samples [5]. We are interested in time period between July 1, 2019 and July 1, 2020, which has 253 trading days, because we wanted to observe market conditions prior to our market decline of interest and see if we are able to detect any critical transitions. Therefore, we provide summary statistics for this time period for all the stock indices and all the *ETF* sectors. The daily log returns, persistent diagrams, persistent landscapes, and the mean landscapes for sliding windows of 50 trading days were generated and plotted together for July 2, 2019 and June 30, 2020, but we only highlighted specific date ranges near our market fall of interest and peaks in the norms in the persistence landscape for all the stock indices and *ETF* sectors, which is discussed in Section 4.

3.4. Statistical Inference

While the topological summaries were useful for examining topological features, we were also interested in finding statistical significant for any changes of these topological features within time. The time period of interest is July 1, 2019 to July 1, 2020, which has $n_{td} = 253$ trading days.

We make the same assumptions from Section 2.5 and Section 2.6. Our random variables will derive from our two sequences of point cloud data sets, X_n^{SI} and X_n^{ETF}. Since our time period of interest has 253 trading days, our sequence of point cloud data sets are size $q = n_{dlr} - w + 1 = 203$. For this reason, we have $q = n_{dlr} - w + 1 = 203$ random variables in each sequence of point cloud data sets.

For all the stock indices and all the *ETF* sectors, we have Y_n^{SI} and Y_n^{ETF} be random variables respectively for $1, \cdots, k^{SI}$ and $1, \cdots, k^{ETF}$ samples for these groups respectively and Λ_n^{SI} and Λ_n^{ETF} are the corresponding landscapes respectively for $n = 1, \cdots, 203$. The associate sample values of Y_n^{SI} and Y_n^{ETF} are denoted as $y_n^{1^{SI}}, \cdots, y_n^{k^{SI}}$ and $y_n^{1^{ETF}}, \cdots, y_n^{k^{ETF}}$ respectively and the corresponding landscapes of these sample values are labelled as $\lambda_n^{1^{SI}}, \cdots, \lambda_n^{k^{SI}}$ and $\lambda_n^{1^{ETF}}, \cdots, \lambda_n^{k^{ETF}}$ respectively.

The functional in Equation (25) is used to define the random variables for all the stock indices and all the *ETFs* as follows:

$$Y_n^{SI} = \sum_{i=1}^{k^{SI}} \int_{\mathbb{R}} \lambda_n^{i^{SI}} \left(X_n^{SI} \right) \mathrm{d}t, \tag{41}$$

$$Y_n^{ETF} = \sum_{i=1}^{k^{ETF}} \int_{\mathbb{R}} \lambda_n^{i^{ETF}} \left(X_n^{ETF} \right) \mathrm{d}t, \tag{42}$$

where $n = 1, \cdots, 203$ for $1, \cdots, k^{SI}$ and $1, \cdots, k^{ETF}$ samples. We recall the sample mean $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$, so using Equation (26) for the sample means for the random variables of all the stock indices and *ETF* sectors, we have the following:

$$\bar{Y}_n^{SI} = \frac{1}{k^{SI}} \sum_{i=1}^{k^{SI}} f\left(\lambda_n^{i^{SI}} \left(X_n^{SI} \right) \right), \tag{43}$$

$$\bar{Y}_n^{ETF} = \frac{1}{k^{ETF}} \sum_{i=1}^{k^{ETF}} f\left(\lambda_n^{i^{ETF}} \left(X_n^{ETF} \right) \right), \tag{44}$$

where $n = 1, \cdots, 203$ for $1, \cdots, k^{SI}$ and $1, \cdots, k^{ETF}$ samples. We assume that μ_n^{SI} and μ_n^{ETF} are the expectations and population means of Y_n^{SI} and Y_n^{ETF} respectively for $n = 1, \cdots, 203$. We set up three sets of hypotheses test and an analogous *p*-value based on a permutation test. For our first two sets of statistical hypotheses, we desire separate hypotheses tests for all the stock indices and for all the *ETF* sectors within a one day lag in their respective sliding windows. Our statistical hypotheses will distinguish for two groups at a time if the means of topological features are the same within a one day lag in their respective sliding windows and point cloud data sets as seen below:

$$H_0:\mu_{n-1}^{SI}=\mu_n^{SI}\quad vs.\quad H_a:\mu_{n-1}^{SI}\neq\mu_n^{SI}. \tag{45}$$

$$H_0:\mu_{n-1}^{ETF}=\mu_n^{ETF}\quad vs.\quad H_a:\mu_{n-1}^{ETF}\neq\mu_n^{ETF}. \tag{46}$$

For our third set of statistical hypotheses, we also wish to compare all the stock indices against all the *ETF* sectors within the same sliding windows. Our statistical hypotheses will determine for two groups at a time if the means of topological features are the same within the same sliding window as shown below:

$$H_0:\mu_n^{SI}=\mu_n^{ETF}\quad vs.\quad H_a:\mu_n^{SI}\neq\mu_n^{ETF}, \tag{47}$$

where in Equations (45)-(47), $n=1,\cdots,203$. To test the null hypotheses found in Equations (45) and (46), we used a two-sample permutation test from Equation (28) to obtain:

$$t=\frac{\left|\bar{Y}_{n-1}^{SI}-\bar{Y}_n^{SI}\right|}{\sqrt{\frac{Var\left(Y_{n-1}^{SI}\right)}{k^{SI}}+\frac{Var\left(Y_n^{SI}\right)}{k^{SI}}}}, \tag{48}$$

$$t=\frac{\left|\bar{Y}_{n-1}^{ETF}-\bar{Y}_n^{ETF}\right|}{\sqrt{\frac{Var\left(Y_{n-1}^{ETF}\right)}{k^{ETF}}+\frac{Var\left(Y_n^{SI}\right)}{k^{ETF}}}}, \tag{49}$$

where $n=1,\cdots,203$ for $1,\cdots,k^{SI}$ and $1,\cdots,k^{ETF}$ samples. To test the null hypotheses found in Equation (47), we used a two-sample permutation test from Equations (28) to obtain:

$$t=\frac{\left|\bar{Y}_n^{SI}-\bar{Y}_n^{ETF}\right|}{\sqrt{\frac{Var\left(Y_n^{SI}\right)}{k^{SI}}+\frac{Var\left(Y_n^{ETF}\right)}{k^{ETF}}}}, \tag{50}$$

where $n=1,\cdots,203$ for $1,\cdots,k^{SI}$ and $1,\cdots,k^{ETF}$ samples. Using Equations (48), (49), and (50), $t_1,\cdots,t_m$ of the test statistic were calculated for permutations $s=1,\cdots,m$. The observed value of the test statistic is expressed as t_{observed}. The *p*-value is calculated by comparing t_{observed} with t_s and averaging the number of times $t_{\text{observed}}\leq t_s$. Using Equation (23), Equations (48) and (49) become:

$$t_{\left\{s,Y_{n-1}^{SI},Y_n^{SI}\right\}}=\frac{\left|\bar{Y}_{n-1}^{SI}-\bar{Y}_n^{SI}\right|}{\sqrt{\frac{Var\left(Y_{n-1}^{SI}\right)}{k^{SI}}+\frac{Var\left(Y_n^{SI}\right)}{k^{SI}}}}, \tag{51}$$

$$t_{\left\{s,Y_{n-1}^{ETF},Y_n^{ETF}\right\}}=\frac{\left|\bar{Y}_{n-1}^{ETF}-\bar{Y}_n^{ETF}\right|}{\sqrt{\frac{Var\left(Y_{n-1}^{ETF}\right)}{k^{ETF}}+\frac{Var\left(Y_n^{ETF}\right)}{k^{ETF}}}}. \tag{52}$$

Similarly, Equation (50) becomes:

$$t_{\left\{s,Y_n^{SI},Y_n^{ETF}\right\}} = \frac{\left|\overline{Y}_n^{SI} - \overline{Y}_n^{ETF}\right|}{\sqrt{\frac{Var\left(Y_n^{SI}\right)}{k^{SI}} + \frac{Var\left(Y_n^{ETF}\right)}{k^{ETF}}}}, \tag{53}$$

where in Equations (51)-(53), where $n = 1,\cdots,203$ for $1,\cdots,k^{SI}$ and $1,\cdots,k^{ETF}$ samples. Hence, using Equations (48) and (52) and every instance where $t_{\text{observed}} \le t_s$, the *p*-values were obtained as:

$$p\text{-value}_{\left\{Y_{n-1}^{SI},Y_n^{SI}\right\}} = \frac{1}{m}\sum_{i=1}^{m} t_{\left\{i,Y_{n-1}^{SI},Y_n^{SI}\right\}}, \tag{54}$$

$$p\text{-value}_{\left\{Y_{n-1}^{ETF},Y_n^{ETF}\right\}} = \frac{1}{m}\sum_{i=1}^{m} t_{\left\{i,Y_{n-1}^{ETF},Y_n^{ETF}\right\}}. \tag{55}$$

Similarly, using Equation (53) and every instance where $t_{\text{observed}} \le t_s$, the *p*-value was obtained as:

$$p\text{-value}_{\left\{Y_n^{SI},Y_n^{ETF}\right\}} = \frac{1}{m}\sum_{i=1}^{m} t_{\left\{i,Y_n^{SI},Y_n^{ETF}\right\}}, \tag{56}$$

where in Equations (54)-(56), $n = 1,\cdots,203$. To evaluate statistical significance, using Equations (41)-(56), a permutation is completed at a significance level of $\alpha = 0.05$ for homology in degree 1 for all our hypothesis tests. Since we are only interested in the number of loops, we will look at homology in degree 1. All these hypothesis testing methods were modified from the *R* script in [5]. After finding the *p*-values, we plotted the daily log returns with the *p*-values that were less than or greater than or equal to our significant level α for either all the stock indices or all the *ETF* sectors along a sliding window of 50 trading days.

4. Results

The goal of this study is to detect a statistically relevant critical transition and characterize any changes in topological features over time. To assess the statistical significance of observed differences in the topological features that change over time, we used a permutation test. For degree 1, we obtained ten sample values of the random variables Y_n^{SI} and Y_n^{ETF} as in Equation (41). Using Equations (43), (45), (48), (51), and (54), the permutation test is implemented with a significance level $\alpha = 0.05$ when comparing all stock indices in different sliding windows between July 1, 2019 and July 1, 2020. The permutation test yields 164 *p*-values of 0.0000, 2 *p*-values of 0.001, and 36 *p*-values of 1 for homology in degree 1.

Using Equations (44), (46), (49), (52), and (55), the permutation test is conducted with a significance level $\alpha = 0.05$ when comparing all the *ETF* sectors in different sliding windows between July 1, 2019 and July 1, 2020. The permutation test returns 164 *p*-values of 0.0000, 4 *p*-values of 0.001, and 33 *p*-values of 1 for homology in degree 1. Using Equations (43), (44), (47), (50), (53), and (56), the permutation is performed with a significance level $\alpha = 0.05$ when comparing all the stock indices and all the *ETF* sectors in the same sliding windows between July 1, 2019 and July 1, 2020, which results in 199 p-values of 0.0000 and 2 p-values of 0.001 for homology in degree 1.

In order to understand these results, we will review the daily log returns, the norms of the persistence landscapes, and the topological summaries of all the stock indices and all the *ETF* sectors. When reviewing the daily log returns for DJIA, the S&P 500, NASDAQ, and Russell 2000 between January 5, 2010 and June 30, 2020 (see **Figure 1**), the stock indices range from −0.05 and 0.05 from 2010 to mid 2011, with some positive and negative spikes that appear leading up to 2012. From 2012 to March 2020, the daily log returns once again fall between −0.05 and 0.05. However, from March 2020 to June 2020, the market is highly volatile. Similar patterns are observed for the *ETF* sectors, but there is a notable spike around 2017 and from March 2020 to June 2020, the *ETF* sectors are more volatile than the stock indices as shown in **Figure 2**.

When we examine the daily log returns of all of the stock indices between January 5, 2010-June 1, 2020, the minimum daily log return occur on March 16, 2020, where Russell 2000 had a return of −0.154, the S&P 500 had a return at −0.1277, and the other stock indices were in between these values. When reviewing the daily log returns for all of the *ETFs* sectors for the same time period, the minimum daily log return also occurs on March 16, 2020, where Information Technology (XLK) had a return of −0.1487, Consumer Staples (XLP) had a return of −0.0702, and the other *ETF* sectors were in between these values. While March 16, 2020 is not recognized as an official financial crash or meltdown, this

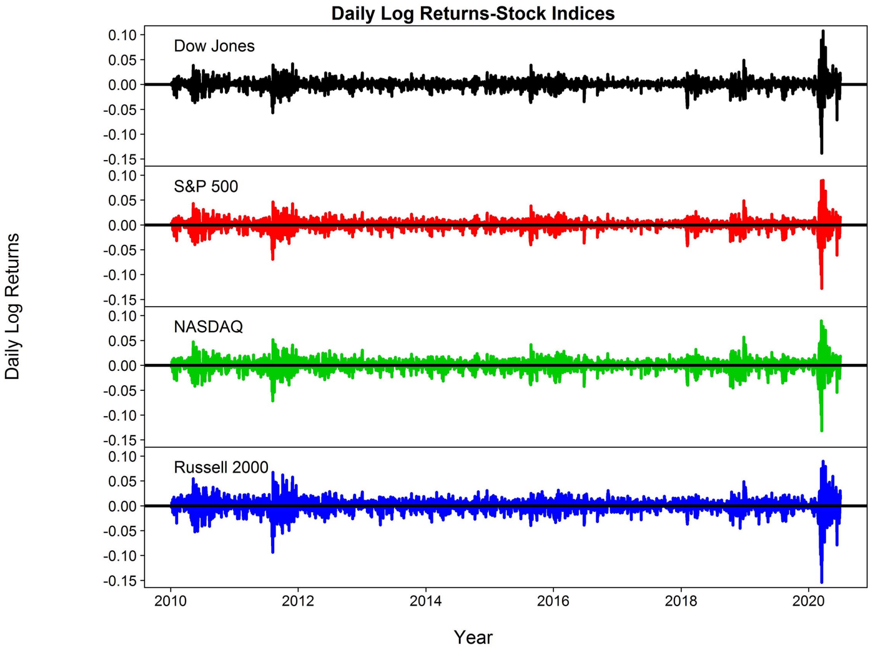

Figure 1. The figures are the daily log returns for all the stock indices from January 5, 2010 to June 30, 2020. The reporting period of this figure contains 2641 trading days from January 4, 2010 to July 1, 2020.

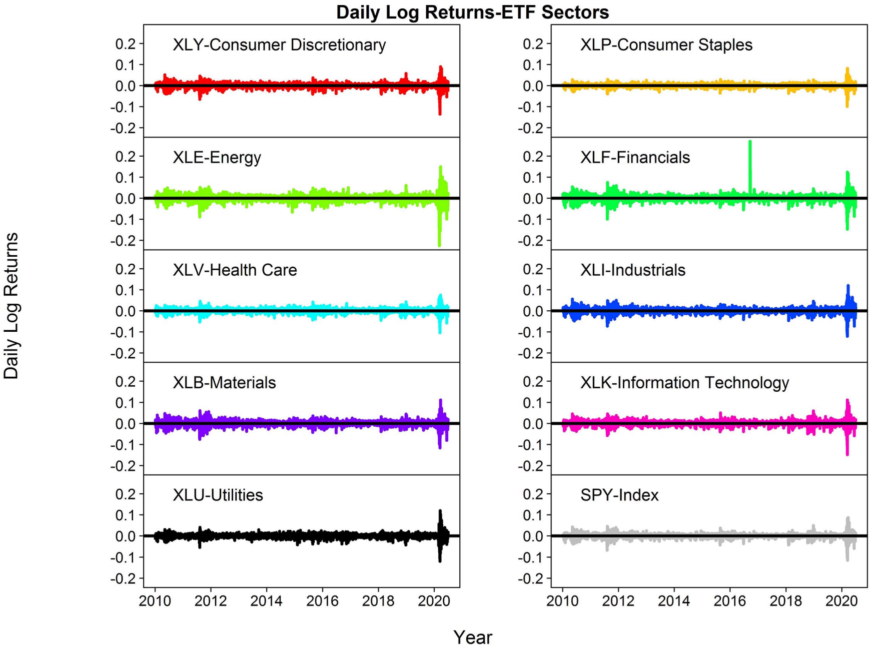

Figure 2. The figures are the daily log returns for all the *ETF* sectors from January 5, 2010 to June 30, 2020. The reporting period of this figure contains 2641 trading days from January 4, 2010 to July 1, 2020.

date is noteworthy, and warrants closer examination for potential critical transitions prior to this date. Focusing on when the peaks occur, we include summary statistics for July 1, 2019 to July 1, 2020 for all the stock indices and all the *ETF* sectors in Table 1 and Table 2, respectively.

The norms of the persistence landscapes in homology degree 1 presented in Figure 3 and Figure 4 display all of the stock indices and all of the *ETFs* respectively for $p=1$ and $p=2$ for 1001 trading days prior to March 16, 2020. For the stock indices, the L^1 distances are less than 0.01 between 2017 and 2018, less than 0.02 between 2018 and 2020, but the greatest L^1 distance occurs in 2020 at approximately 0.08 as seen in Figure 3. The L^1 distance for all of the *ETFs*, have more spikes than the L^1 distances of the stock indices, especially between 2018 and 2020, but the greatest L^1 distance occurs in March 2020 at approximately 0.14 as seen in Figure 3. While the L^2 norms for all of the stock indices and all of the *ETFs* have similar distances, there is a noticeable spike in 2020. However, the distances in L^2 are not as great as in L^1, as shown in Figure 3 and Figure 4.

Figure 3 and Figure 4 highlight the peak in more detail for the time period between January 3, 2020 to June 30, 2020. While critical points are discernible in the month of February 2020, the peaks occurred on February 21, 2020 and March 3, 2020 for all of the stock indices and for all of the *ETF* sectors respectively as seen in Figure 4. Recall that a point on the norms of the persistence landscapes coincides with a sliding window of 50 trading days in the daily log returns, which means the peaks are from February 21, 2020 to May 1, 2020 and March 3, 2020 to May 12, 2020 for all of the stock indices and for all of the *ETF*

Table 1. Summary statistics for stock indices.

Stock Name	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\sigma}$	$\hat{\gamma}$	$\hat{\kappa}$
Dow Jones	−1e−04	5e−04	0.0228	−0.8479	13.1333
S&P 500	2e−04	5e−04	0.0213	−0.8691	12.6087
NASDAQ	9e−04	5e−04	0.0213	−1.0494	12.6029
Russell 2000	−3e−04	7e−04	0.0263	−1.3226	11.2358

Table 2. Summary statistics for *ETF* sectors.

Stock Symbol	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\sigma}$	$\hat{\gamma}$	$\hat{\kappa}$
XLY	0.0003	0.0004	0.0210	−1.3141	14.1179
XLP	0.0001	0.0003	0.0173	−0.2487	12.6118
XLE	−0.0017	0.0012	0.0348	−1.3392	13.4664
XLF	−0.0006	0.0008	0.0276	−0.6256	10.4510
XLV	0.0004	0.0004	0.0187	−0.4299	10.1362
XLI	−0.0004	0.0006	0.0243	−0.5575	10.0397
XLB	−0.0001	0.0005	0.0232	−0.7113	10.0980
XLK	0.0012	0.0006	0.0242	−0.6869	12.5942
XLU	−0.0001	0.0006	0.0235	−0.0751	11.9571
SPY	0.0002	0.0004	0.0207	−0.8911	11.7189

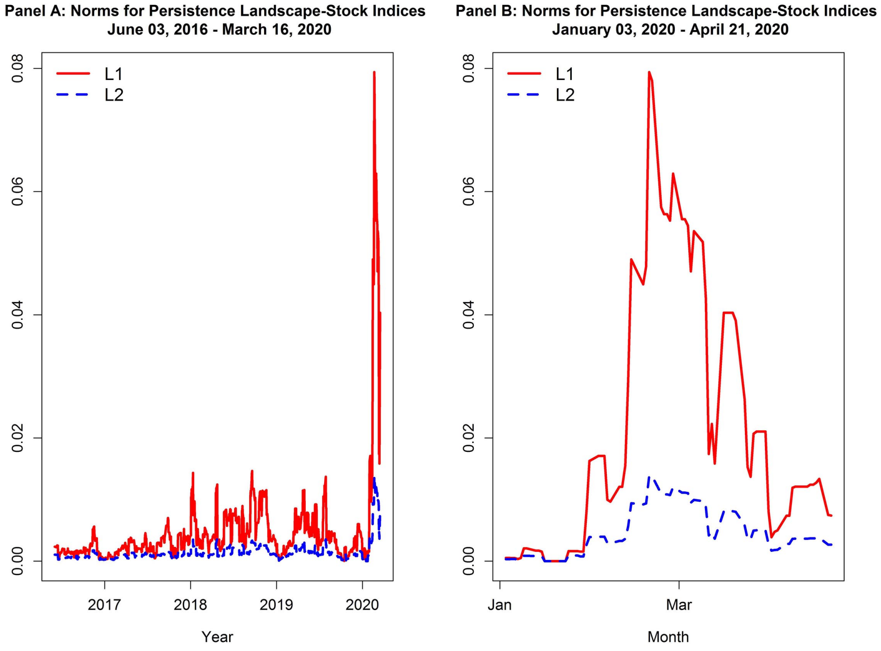

Figure 3. The figures are the norms of the persistence landscapes of all the stock indices, where $p=1$ (solid line) and $p=2$ (dashed line)) and each point in the figure represents a sliding window of 50 trading days. Panel A plots the time frame June 3, 2016 to March 16, 2020, where the last sliding window is from March 16, 2020 to May 26, 2020 and the reporting period of this figure contains 1001 trading days from June 2, 2016 to May 27, 2020. Panel B plots the time frame between January 3, 2020 to June 30, 2020, where the last sliding window is from April 21, 2020 to June 30, 2020 and the reporting period of this figure contains 76 trading days from January 2, 2020 to July 1, 2020.

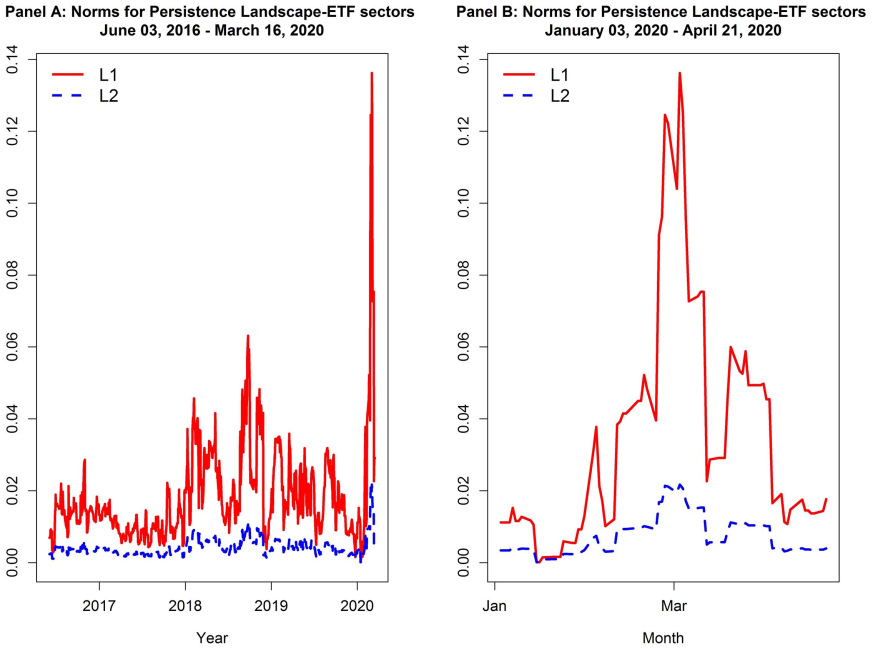

Figure 4. The figures are the norms of the persistence landscapes of all the *ETF* sectors, where $p=1$ (solid line) and $p=2$ (dashed line)) and each point in the figure represents a sliding window of 50 trading days. Panel A plots the time frame between June 3, 2016 to March 16, 2020, where the last sliding window is from March 16, 2020 to May 26, 2020 and the reporting period of this figure contains 1001 trading days from June 2, 2016 to May 27, 2020. Panel B plots the time frame between January 3, 2020 to June 30, 2020, where the last sliding window is from April 21, 2020 to June 30, 2020 and the reporting period of this figure contains 76 trading days from January 2, 2020 to July 1, 2020.

sectors respectively. In particular, Figure 5 and Figure 6 emphasize this point, where the norms of the persistence landscapes and the daily log returns of either all of the stock indices or all of the *ETF* sectors are next to each other. The sliding windows of 50 trading days of the daily log returns synchronize to the first point in the norms of the persistence landscape and to the maximum values of the norms of the persistence landscapes as indicated by Figure 5 and Figure 6.

Aside from the norms of the persistence landscapes, we produce topological summaries to represent the persistence of topological features for all the stock indices and for all the *ETF* sectors between January 3, 2020 and June 30, 2020. Along with these topological summaries (the persistence diagram, the persistence landscape, the mean landscape), we plotted the daily log returns for the corresponding sliding window of 50 trading days shown in Figures 7-12.

Figure 7 and Figure 8 indicate that the daily log returns are centered around zero from January 3, 2020 to February 21, 2020 for all of the stock indices and from January 3, 2020 to February 26, 2020 for all of the *ETF* sectors. Not much persistence is evident in the persistence diagram and few spikes appear in persistence landscape and mean landscape. Figure 9 and Figure 10 illustrate more variability in the daily log returns from March 1, 2020 to April 16, 2020 for all of

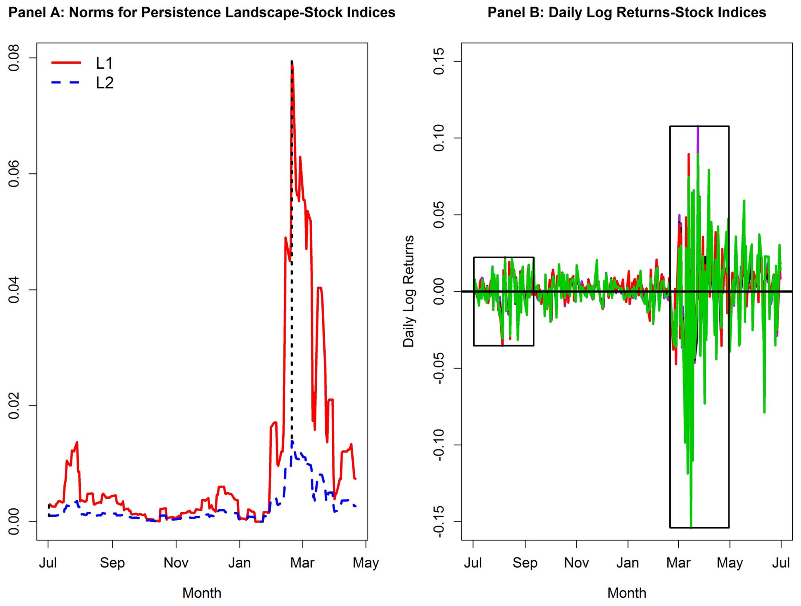

Figure 5. The figures are the norms of the persistence landscapes and the daily log returns of all the stock indices from for July 2, 2019-June 30, 2020. Panel A is the norms of the persistence landscape, where $p=1$ (solid line) and $p=2$ (dashed line)), each point in the figure represents a sliding window of 50 trading days, and two dashed lines depicting the first and maximum points in this figure. Panel B plots the daily log returns with two sliding windows of 50 trading days (depicted as rectangles) corresponding to the first and max points in the Panel A. The first sliding window is from July 2, 2019 to September 11, 2019, while the second sliding window is from February 21, 2020 to May 1, 2020. The reporting period of this figure contains 253 trading days from July 1, 2019 to July 1, 2020.

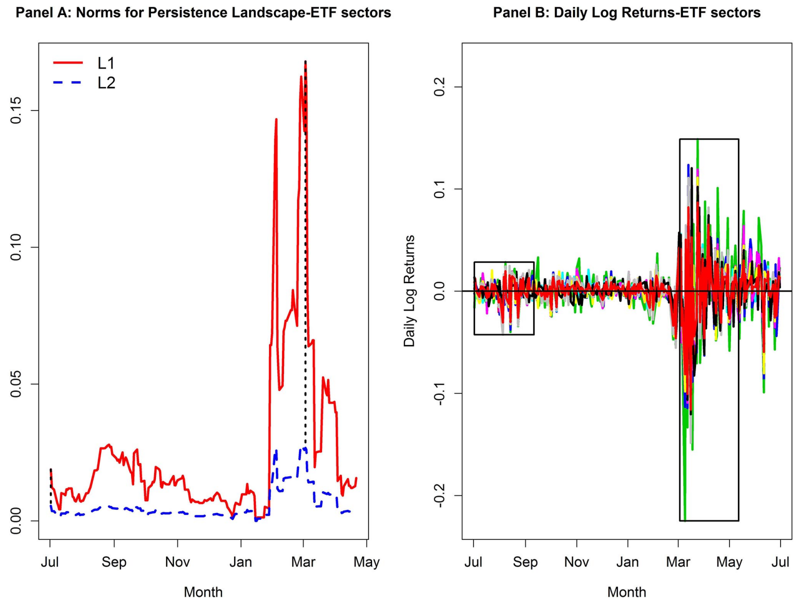

Figure 6. The figures are the norms of the persistence landscapes and the daily log returns of all the *ETF* sectors from for July 2, 2019-June 30, 2020. Panel A is the norms of the persistence landscape, where $p=1$ (solid line) and $p=2$ (dashed line)), each point in the figure represents a sliding window of 50 trading days, and two dashed lines depicting the first and maximum points in this figure. Panel B plots the daily log returns with two sliding windows of 50 trading days (depicted as rectangles) corresponding to the first and max points in the Panel A. The first sliding window is from July 2, 2019 to September 11, 2019, while the second sliding window is from March 3, 2020 to May 12, 2020. The reporting period of this figure contains 253 trading days from July 1, 2019 to July 1, 2020.

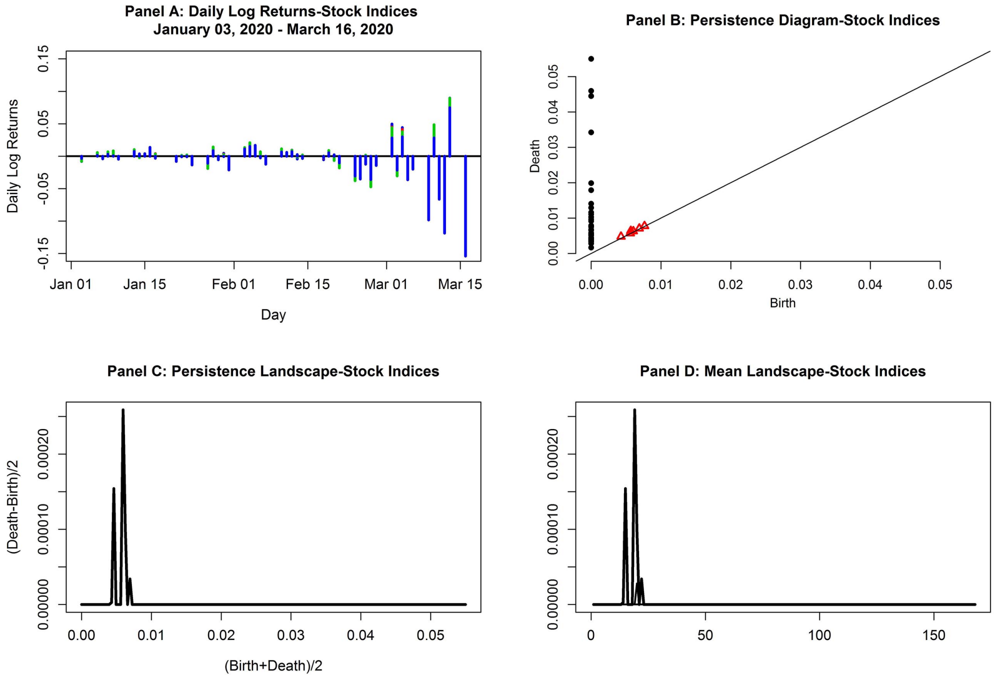

Figure 7. These figures are the daily log returns and topological summaries of all the stocks indices from January 3, 2020-March 16, 2020. Panel A plots the daily log returns for all the stock indices with a sliding window of 50 trading days. Panel B plots the first dimension of the Vietoris-Rips persistence diagram, where the solid black dots represent connected components and the red triangles represent loops. Panel C plots the first dimension of corresponding persistence landscape. Panel D plots the corresponding the mean landscape.

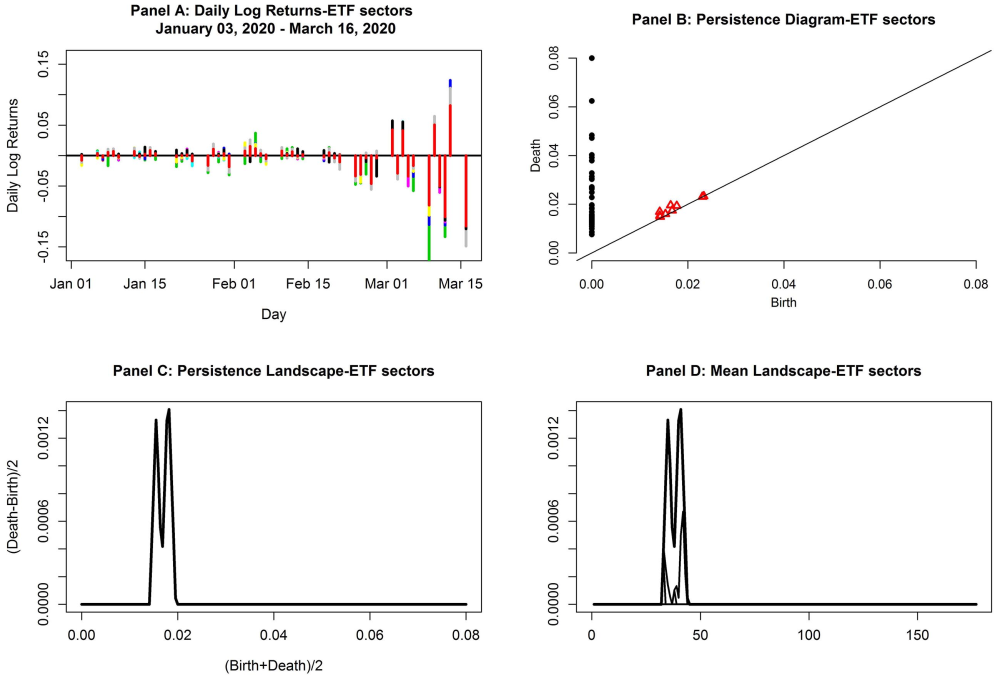

Figure 8. These figures are the daily log returns and topological summaries of all the *ETF* sectors from January 3, 2020-March 16, 2020. Panel A plots the daily log returns for all the *ETF* sectors with a sliding window of 50 trading days. Panel B plots the first dimension of the Vietoris-Rips persistence diagram, where the solid black dots represent connected components and the red triangles represent loops. Panel C plots the first dimension of corresponding persistence landscape. Panel D plots the corresponding the mean landscape.

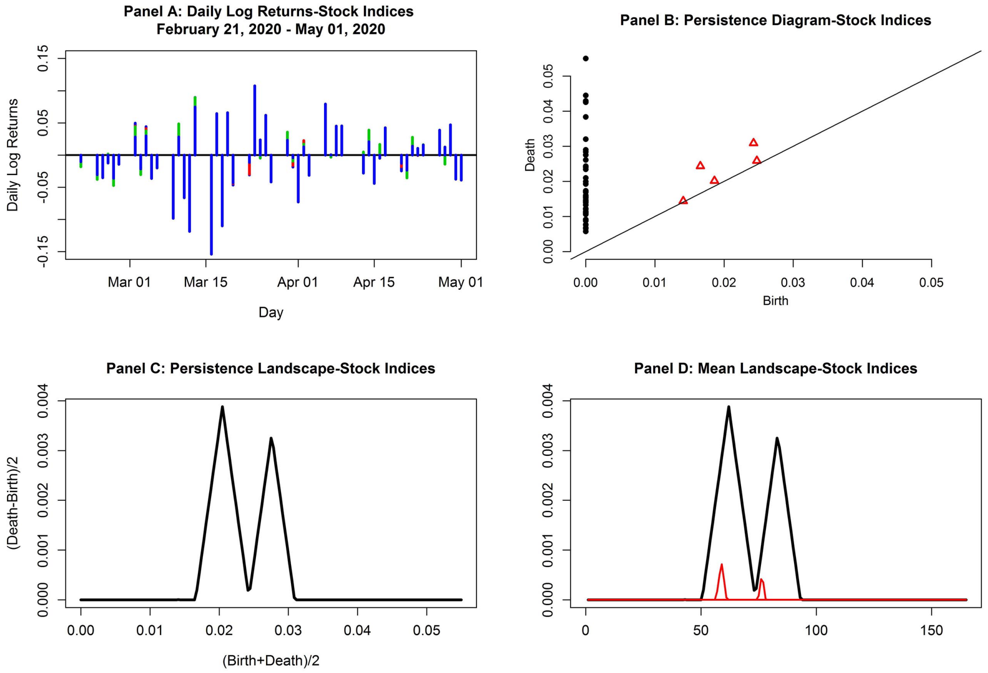

Figure 9. These figures are the daily log returns and topological summaries of all the stocks indices from February 21, 2020 to May 1, 2020. Panel A plots the daily log returns for all the stock indices with a sliding window of 50 trading days. Panel B plots the first dimension of the Vietoris-Rips persistence diagram, where the solid black dots represent connected components and the red triangles represent loops. Panel C plots the first dimension of corresponding persistence landscape. Panel D plots the corresponding the mean landscape.

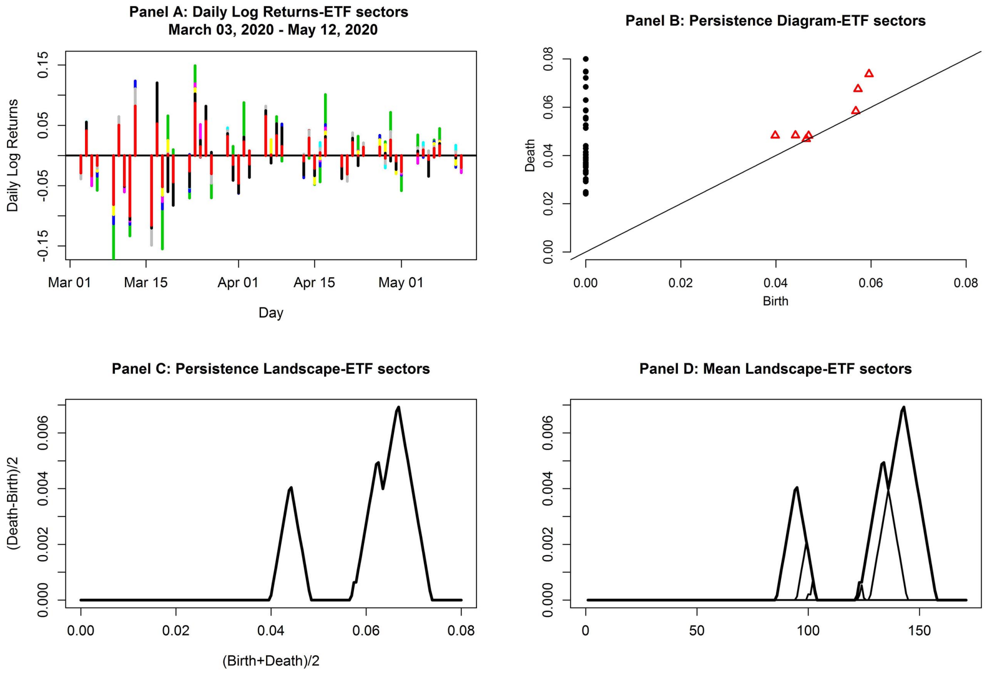

Figure 10. These figures are the daily log returns and topological summaries of all the *ETF* sectors from March 3, 2020-May 12, 2020. Panel A plots the daily log returns for all the *ETF* sectors with a sliding window of 50 trading days. Panel B plots the first dimension of the Vietoris-Rips persistence diagram, where the solid black dots represent connected components and the red triangles represent loops. Panel C plots the first dimension of corresponding persistence landscape. Panel D plots the corresponding the mean landscape.

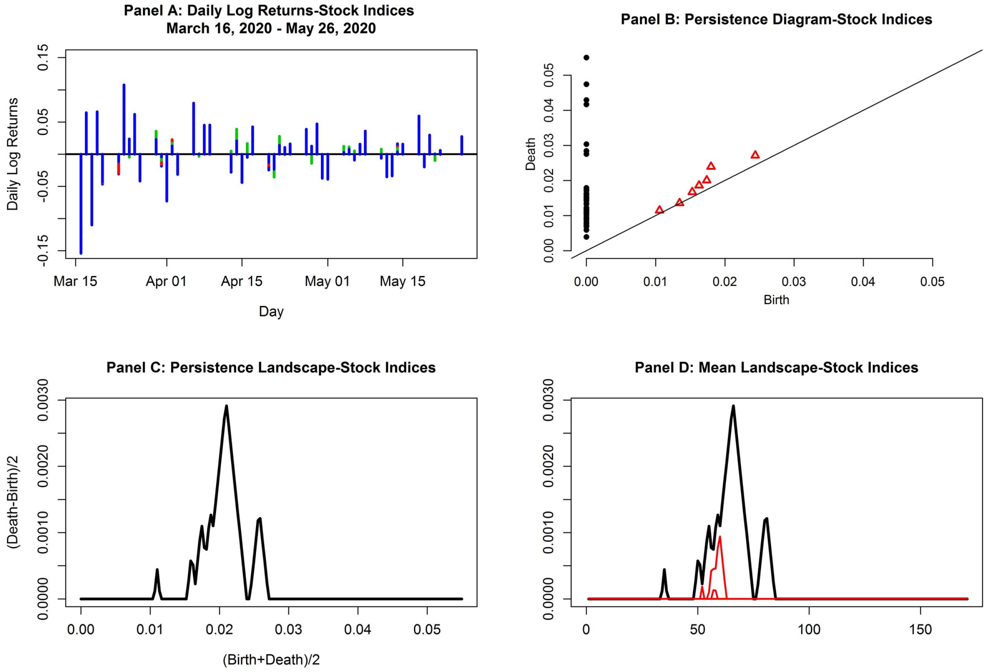

Figure 11. These figures are the daily log returns and topological summaries of all the stocks indices from March 16, 2020-May 26, 2020. Panel A plots the daily log returns for all the stock indices with a sliding window of 50 trading days. Panel B plots the first dimension of the Vietoris-Rips persistence diagram, where the solid black dots represent connected components and the red triangles represent loops. Panel C plots the first dimension of corresponding persistence landscape. Panel D plots the corresponding the mean landscape.

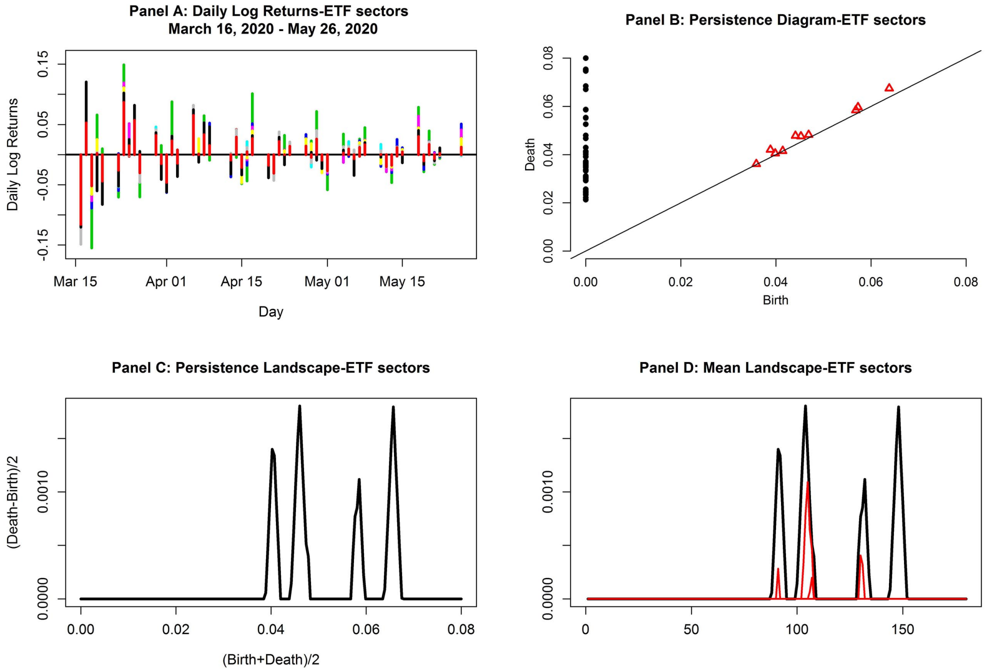

Figure 12. These figures are the daily log returns and topological summaries of all the *ETF* sectors from March 16, 2020-May 26, 2020. Panel A plots the daily log returns for all the *ETF* sectors with a sliding window of 50 trading days. Panel B plots the first dimension of the Vietoris-Rips persistence diagram, where the solid black dots represent connected components and the red triangles represent loops. Panel C plots the first dimension of corresponding persistence landscape. Panel D plots the corresponding the mean landscape.

the stock indices and from March 1, 2020 to May 1, 2020 for all of the *ETF* sectors. Significant persistence is apparent in the persistence diagram and more spikes appear in the persistence landscape and mean landscape.

5. Discussion

From reviewing the norms of the persistence landscape, the daily log returns, persistence diagrams, persistence landscapes, and mean landscapes for all of the selected dates, it is clear that the number of the loops in the relevant point clouds are more pronounced resulting in more persistence, which signifies that the stock market is transitioning from a stable state to a more unpredictable, volatile state. Moreover, the *ETF* sectors demonstrate more volatility than the stock indices. These stock indices' findings coincide with the 2000 and 2008 market crashes findings found in [16]. Similar to Gidea and Katz [16], we observe L^1 distances that confirm the critical thresholds prior to the 2020 peak and exhibit more than the L^2 norm. In other words, the L^p-norms exhibit strong growth around the emergence of the primary peak.

While the highest peak occurred on February 21, 2020 for all of the stock indices and March 3, 2020 for all of the *ETF* sectors in the L^p norms, the Coronavirus (COVID-19) broke out in 2019 in Wuhan, China, but on January 21, 2020, the first US case was confirmed. The most important dates are March 13, 2020 when President Trump declares national emergency, March 15, 2020 when the Center of Disease Control and Prevention warns against large gatherings, and March 17, 2020 when COVID is present in all 50 states. The daily log returns for all of the stock indices and for all of the *ETF* sectors do not include negative values. Yet, there are other dates that could have lead to a market decline in March 16, 2020. For example, on January 30, 2020 when World Health Organization (WHO) declares a global health emergency or between February 5, 2020 and February 29, 2020 when the outbreak becomes an epidemic. While we acknowledge that it is quite difficult to predict a market crash, the norms of the persistence landscape performed really well as indicator in detecting critical transitions and the topological summaries authenticated volatility by of the number of loops increasing.

Our hypotheses tests aimed to find how topological features change within time, notably between July 1, 2019 and July 1, 2020. Our hypotheses tests for all of the stock indices found evidence of difference in topological features when comparing adjacent sliding windows of a sliding step of one day. In particular, we found for the chosen time frame that the daily log returns of all the stock indices significantly differ in the number of loops. Equivalently, our hypotheses tests for all of the *ETF* sectors found evidence of difference in topological features when comparing adjacent sliding windows of a sliding step of one day. Specifically, we found for the selected time frame that the daily log returns of all the *ETF* sectors significantly differ in the number of loops. Our last hypotheses tests between all of the stock indices and all of the *ETF* sectors within the same sliding window found inconclusive evidence of difference in topological features for the entire time frame.

6. Conclusions

In this paper, we investigated the topological features of four major indices and 10 *ETF* sectors for January 4, 2010-July 1, 2020. We used two sequences of point cloud data sets, one for all the stock indices and the other for all the *ETFs* with a sliding window $w = 50$. Both sequences were used to perform TDA through algebraic topology and persistent homology. From there, topological summaries are generated to determine persistence and the norms for persistence landscapes are used to detect a critical transition by adapting methods found in [16]. Our goal is to determine how the statistical significance of topological features of stock indices and *ETF* sectors change for a specific time frame. We found that between July 1, 2019 and July 1, 2020, there is evidence of difference of topological features for all the stock indices and all the *ETFs*. As a result, critical transitions are determined using the norms of the persistence landscape and topological features of stock indices and *ETF* sectors change within time when comparing two sliding windows of a sliding step of one day.

We conclude with possible future research goals. Further work could be done analyzing persistence landscapes for homology in degree two. It would be interesting to study topological features based on higher degree persistence. Furthermore, it would be fascinating to expand to commodities, futures, and other financial time series. Moreover, it would be more resourceful to expand topological data analysis to statistics beyond statistical inference and use for predictive modeling with machine learning.

This table presents summary statistics for all the stock indices. We estimated the mean ($\hat{\mu}$), standard deviation ($\hat{\sigma}^2$), variance ($\hat{\sigma}$), skewness ($\hat{\gamma}$), and kurtosis ($\hat{\kappa}$) of the daily log returns from July 2, 2019 to June 30, 2020. The reporting period of this table contains 253 trading days from July 1, 2019 to July 1, 2020.

This table presents summary statistics for all the *ETF* sectors. We estimated the mean ($\hat{\mu}$), standard deviation ($\hat{\sigma}^2$), variance ($\hat{\sigma}$), skewness ($\hat{\gamma}$), and kurtosis ($\hat{\kappa}$) of the daily log returns from July 2, 2019 to June 30, 2020. The reporting period of this table contains 253 trading days from July 1, 2019 to July 1, 2020.

Acknowledgements

The authors would like to thank Tracy Volz for helpful discussions in the editing process. The authors would also like to thank the Center of Computational Finance and Economic Systems (https://cofes.rice.edu).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Collins, A., Zomorodian, A., Carlsson, G. and Guibas, L.J. (2004) A Barcode Shape Descriptor for Curve Point Cloud Data. *Computers & Graphics*, **28**, 881-894. https://doi.org/10.1016/j.cag.2004.08.015

[2] Mileyko, Y., Mukherjee, S. and Harer, J. (2011) Probability Measures on the Space of Persistence Diagrams. *Inverse Problems*, **27**, Article ID: 124007. https://doi.org/10.1088/0266-5611/27/12/124007

[3] Turner, K., Mileyko, Y., Mukherjee, S. and Harer, J. (2014) Fréchet Means for Distributions of Persistence Diagrams. *Discrete & Computational Geometry*, **52**, 44-70. https://doi.org/10.1007/s00454-014-9604-7

[4] Munch, E., Bendich, P., Turner, K., Mukherjee, S., Mattingly, J. and Harer, J. (2013) Probabilistic Fréchet Means and Statistics on Vineyards.

[5] Nicolau, M., Levine, A.J. and Carlsson, G. (2011) Topology Based Data Analysis Identifies a Subgroup of Breast Cancers with a Unique Mutational Profile and Excellent Survival. *Proceedings of the National Academy of Sciences*, **108**, 7265-7270. https://doi.org/10.1073/pnas.1102826108

[6] Carlsson, G., Ishkhanov, T., De Silva, V. and Zomorodian, A. (2008) On the Local Behavior of Spaces of Natural Images. *International Journal of Computer Vision*, **76**, 1-12. https://doi.org/10.1007/s11263-007-0056-x

[7] Wang, Y., Ombao, H. and Chung, M.K. (2019) Statistical Persistent Homology of Brain Signals. *IEEE International Conference on Acoustics, Speech and Signal Processing*, Brighton, 12-17 May 2019, 1125-1129. https://doi.org/10.1109/ICASSP.2019.8682978

[8] Nielson, J.L., Paquette, J., Liu, A.W., Guandique, C.F., Tovar, C.A., Inoue, T., Irvine, K.-A., Gensel, J.C., Kloke, J., Petrossian, T.C., *et al.* (2015) Topological Data Analysis for Discovery in Preclinical Spinal Cord Injury and Traumatic Brain Injury. *Nature Communications*, **6**, 8581. https://doi.org/10.1038/ncomms9581

[9] Scheffer, M., Bascompte, J., Brock, W.A., Brovkin, V., Carpenter, S.R., Dakos, V., Held, H., Van Nes, E.H., Rietkerk, M. and Sugihara, G. (2009) Early-Warning Signals for Critical Transitions. *Nature*, **461**, 53-59. https://doi.org/10.1038/nature08227

[10] Ensor, K.B. and Koev, G.M. (2014) Computational Finance: Correlation, Volatility, and Markets. *Wiley Interdisciplinary Reviews: Computational Statistics*, **6**, 326-340. https://doi.org/10.1002/wics.1323

[11] Gidea, M. (2017) Topological Data Analysis of Critical Transitions in Financial Networks. In: *International Conference and School on Network Science*, Springer, Berlin, 47-59. https://doi.org/10.1007/978-3-319-55471-6_5

[12] Gidea, M., Goldsmith, D., Katz, Y., Roldan, P., Shmalo, Y., *et al.* (2020) Topological Recognition of Critical Transitions in Time Series of Cryptocurrencies. *Physica A: Statistical Mechanics and Its Applications*, **548**, Article ID: 123843. https://doi.org/10.1016/j.physa.2019.123843

[13] Guttal, V., Raghavendra, S., Goel, N. and Hoarau, Q. (2016) Lack of Critical Slowing Down Suggests That Financial Meltdowns Are Not Critical Transitions, Yet Rising Variability Could Signal Systemic Risk. *PLoS ONE*, **11**, e0144198. https://doi.org/10.1371/journal.pone.0144198

[14] Edelsbrunner, H., Letscher, D. and Zomorodian, A. (2002) Topological Persistence and Simplification. *Discrete & Computational Geometry*, **28**, 511-533. https://doi.org/10.1007/s00454-002-2885-2

[15] Hatcher, A. (2002) Algebraic Topology. Cambridge University Press, Cambridge.

[16] Munkres, J.R. (1984) Elements of Algebraic Topology. The Benjamin/Cummings Publishing Company, Meno Park, CA.

[17] Dundas, B.I. (2013) Differential Topology.

[18] Bubenik, P. and Dłotko, P. (2017) A Persistence Landscapes Toolbox for Topological Statistics. *Journal of Symbolic Computation*, **78**, 91-114. https://doi.org/10.1016/j.jsc.2016.03.009

[19] Otter, N., Porter, M.A., Tillmann, U., Grindrod, P. and Harrington, H.A. (2017) A Roadmap for the Computation of Persistent Homology. *EPJ Data Science*, **6**, 17. https://doi.org/10.1140/epjds/s13688-017-0109-5

[20] Kovacev-Nikolic, V., Bubenik, P., Nikolić, D. and Heo, G. (2016) Using Persistent Homology and Dynamical Distances to Analyze Protein Binding. *Statistical Applications in Genetics and Molecular Biology*, **15**, 19-38. https://doi.org/10.1515/sagmb-2015-0057

[21] RStudio Team (2020) RStudio: Integrated Development Environment for R. RStudio, PBC, Boston.

[22] Shumway, R.H. and Stoffer, D.S. (2016) Time Series Analysis and Its Applications. Springer, Berlin. https://doi.org/10.1007/978-3-319-52452-8

[23] Gidea, M. and Katz, Y. (2018) Topological Data Analysis of Financial Time Series: Landscapes of Crashes. *Physica A*: *Statistical Mechanics and Its Applications*, **491**, 820-834. https://doi.org/10.1016/j.physa.2017.09.028

[24] Bubenik, P. (2015) Statistical Topological Data Analysis Using Persistence Landscapes. *The Journal of Machine Learning Research*, **16**, 77-102.

[25] Fasy, B.T., Kim, J., Lecci, F., Maria, C. and Rouvreau, V. (2019) TDA: Statistical Tools for Topological Data Analysis. RStudio, PBC, Boston. https://cran.r-project.org/web/packages/TDA/index.html