Data Science in Finance and Economics

http://www.aimspress.com/journal/dsfe

DSFE, 3(3): 208–229.
DOI:10.3934/DSFE.2023013
Received: 27 April 2023
Revised: 24 June 2023
Accepted: 03 July 2023
Published: 11 July 2023

Research article

A topological based feature extraction method for the stock market

Chen Chang[1], **Hongwei Lin**[2,*]

[1] Polytechnic Institute, Zhejiang University, China
[2] School of Mathematical Science, Zhejiang University, China

* **Correspondence:** Email: hwlin@zju.edu.cn.

Abstract: We proposed a topology-based method for pre-processed time series data extracted from stock market data. The topology features are extracted from data after denoising and normalization by using a version of weighted Vietoris-Rips complex. We compare the features from bullish, bearish and normal periods of the Chinese stock market and found significant differences between the features extracted from the groups. Based on the previous research mentioned in the context, we proposed a topology-based stock market index which has the ability to distinguish different stages of the stock market and forewarn stock market crashes.

Keywords: topological data analysis; stock market; feature extraction

JEL Codes: C65

1. Introduction

Since the inception of the stock market, investors have actively participated in the pursuit of profits. Consequently, as the stock market has evolved numerous tools have been developed to analyze and predict market trends. The analysis and prediction of the stock market have become prominent research areas due to the multitude of factors influencing it.

Topological data analysis (TDA) (Edelsbrunner et al., 2000) is an emerging tool that extracts topological features from high-dimensional datasets. Recently, there has been a surge in studies applying TDA to the stock market. For instance, TDA was employed in 2008 to predict the market crash and subsequent improvements were made (Prabowo et al., 2021). Additionally, Yen et al. developed topology-based methods utilizing fusion models to analyze the market process and its intricacies (Yen et al., 2021). The extracted features from TDA were used as inputs in traditional machine learning techniques to study market crashes and other market phenomena (Basu and Li, 2019).

In this paper, we propose a novel method for feature extraction combining manifold learning-based stock selection with TDA. This approach simplifies computation while capturing the distinctions between different market periods. Furthermore, our method incorporates the adoption of distance-to-measure (DTM) based Vietoris-Rips filtration (Anai et al., 2020) which enhances robustness against outliers.

2. Related work

In this section, we mainly review works from two aspects, i.e., works on predicting the stock market or other derivative markets by various techniques, and works on TDA applications.

2.1. Stock market and technical analysis

Technical analysis was born with the early stock market's rapid growth. Joseph de la Vega analyzed the market data of Dutch financial markets in the 17th century (Corzo et al., 2014) which is considered the budding of modern technical analysis. Tools for technical analysis also sprouted. Japanese started using candlestick chart to represent the rise and fall of the price of rice in the 17th century and Charles Dow proposed a similar version in the US in 1900. Candlestick chart can easily represent opening, high, low and closing price for each time period the user wants to display and it has become the most used chart in different markets for its simplicity and certainty (Nison, 1994).

Investors and researchers proposed many other technical indicators besides candlestick charts. G. Appel proposed moving average convergence divergence in 1970s (Appel, 1985) which is designed to predict the market by observing the movement of the index which reflects the strength, direction, momentum and duration of a trend (Appel, 2005). J. W. Wilder developed the relative strength index (RSI) in 1978 (Wilder, 1978), explaining that the RSI can speculate whether a stock is under overbought or oversold territory. Similarly, J. Bolliner proposed Bollinger Bands (BOLL)in 1980s which aims to show whether the current price is high or low by using Lower Band, 20SMA and Upper Band with a belt region (Bollinger, 2002).

The above methods share common features and also called the principles of technical analysis. First, market action discounts everything which means that all information that may affect the market is already reflected by the price and volume. Second, history tends to repeat itself making the prediction of the market using old data possible. Third, prices move in trends indicating underlying patterns (Kirkpatrick II and Dahlquist, 2010; Deng, 2008; Teixeira and De Oliveira, 2010).

In the second half of the 20th century, with the development of the computer and the internet, using the computer for systematic trading has become a new trend. For most strategies only using market data as input, computer based trading system can be seen as a branch of technical analysis. Various methods of data analysis have been used on market predicting and systematic trading.

2.2. Topological data analysis

Topology originated in the 18th century. Although topology was originally designed for studying shapes and surfaces recent studies have adopted computational topology for studying large and high-dimensional data sets (Carlsson, 2009). The fundamental idea is to recognize shapes, discover insights in the data and identify meaningful sub-groups, which can then be studied by using standard statistical techniques (Lum et al., 2013). With the development of computational topology, interdisciplinary studies related with data mining, pattern recognition, machine learning and topology emerged. Edelsbrunner et

al. proposed persistent homology group first (Edelsbrunner et al., 2000) which can be illustrated by persistence diagram (Cohen-Steiner et al., 2007). Owing to its stability against small perturbation or missing data, persistence diagram as input of maching learning based calculation in metric space is widely accepted (Zeng et al., 2021). TDA has been applied on many different fields including 3D shape matching (Carrière et al., 2015), recurrent system modeling (Skraba et al., 2012) and signal analysis (Perea and Harer, 2015). In medical data analysis, TDA is also an widely used technique in various areas including cancer classification(Wu et al., 2017) and cardiac trabeculae restoration.(Wang et al., 2021). As for combining with deep learning, TDA also has considerable potential(Hu et al., 2019; Hu et al., 2021; Wang et al., 2020).

Persistent homology (Edelsbrunner and Harer, 2010) is the flagship tool of TDA (Hensel et al., 2021). In the analysis of real-world data, topological features obtained from databases are unclear. Using a filtration (connected to the scale parameter) persistent homology is able to capture topological changes across the whole range of scales and stores this information in persistence diagrams.

Persistence diagram has its limitations. The length of persistence diagram is not stable and can not be directly regarded as input of machine learning or deep learning. Researchers proposed several discretized method based on persistence diagram including persistence landscape (Bubenik and Dłotko, 2017), persistence image (Adams et al., 2017) and so on.

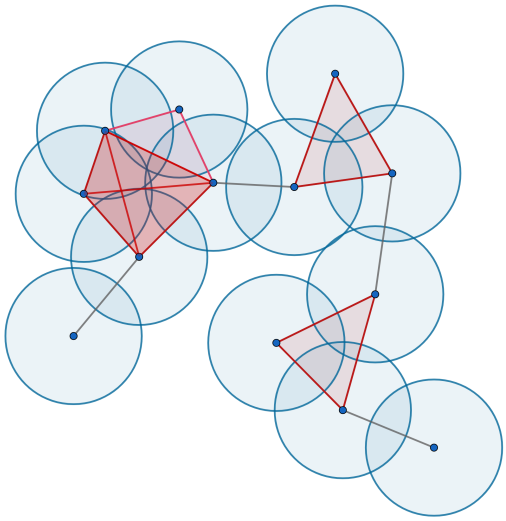

Figure 1. A certain state from the process of Vietoris Rips filteration. The 13 points are 0-simplices. Two 0-simplices form a 1-simplex (an edge) if their $\varepsilon/2$-neighborhoods (yellow circles) intersect. Three vertices form a 2-simplex (a triangle) if they are pairwise connected by edges. Four vertices form a 3-simplex (a tetrahedron) if they are pairwise connected by edges (Topaz et al., 2015).

2.3. TDA with stock market

Many researches focus on combining TDA with stock market these years. Guo et al. analyzed Chinese stock market from 2013 to 2020 and constructed planar maximal filtered graphs in 3 turbulent periods to discover systematically important companies(Guo et al., 2022). Westlin used TDA to capture the connection between credit expansion and stock market crashes(Westlin et al., 2022). Yen et al. built

Laplacian spectra to discover the possibility of understanding market crashes in four different crashes which happened in Taiwan, Singapore and US stock markets(Yen and Cheong, 2021; Yen et al., 2023). Prabowo et al. used TDA to predict early warning signs of market declines that are inevitable(Prabowo et al., 2021). Katz et al. utilized TDA to examine market instabilities (Katz and Biem, 2021). These works were influenced to varying extents by the earlier and subsequent research by Gidea and Katz (Gidea, 2017; Gidea et al., 2020), where they found the potential of TDA in classifying and predicting market crashes. It is worth noting that the aforementioned works predominantly focus on specific periods of the market, specifically prior to the occurrence of crashes. In contrast, our work emphasizes the performance of TDA across the entire market cycle, particularly during the transitions between different stages.

3. Background

Considering matrix $A = [v_1, v_2, \ldots, v_k]$ with k as the number of vectors in A, a simplicial complex is a kind of abstract triangulated structure built on A to represent the intrinsic topological space consisting of vectors in A.

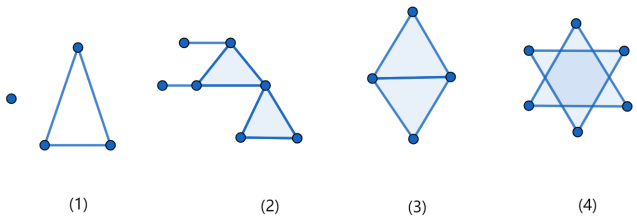

Figure 2. Simplicial complexes and a counter example

A single simplicial complex $\mathscr{K}$ consists of n-dimensional simplexes while satisfying:

- Every face of a simplex from $\mathscr{K}$ is also in $\mathscr{K}$.
- The non-empty intersection of any two simplices $\sigma_1, \sigma_2 \in \mathscr{K}$ is a face of both σ_1 and σ_2.

A simplex is a generalization of triangle to arbitrary dimensions.

- 0-dimensional simplex is a vertex.
- 1-dimensional simplex is an edge.
- 2-dimensional simplex is a triangular face.
- 3-dimensional simplex is a tetrahedron.
- ...

The dimension of a simplicial complex is the highest dimension of simplexes it contains. In Figure 2, (1) is a 1-dimensional simplicial complex, (2) and (3) are 2-dimensional simplicial complex; (4) is not a simplicial complex.

Vietoris-Rips filtration is a common filtering method applied to point cloud data and scalar data (Edelsbrunner and Harer, 2010). It is induced by the function

$$f(v_0, v_1, \ldots, v_n) = \frac{1}{2} \max_{i \neq j} |v_i - v_j| \tag{1}$$

This process can be seen as constructing a sphere with each point in the point cloud as the center. The radii of all spheres are the same at a given "time" and the intersection of different spheres determines the structure formed on the point cloud. In this study, the filtration built on A are constructed by using Vietoris-Rips complex. It follows the following steps: (Sheehy, 2012; Gromov, 1987)

- All vectors from A is regarded as a vertex and we denote $X = v_i, i = 1,2,\ldots,k$.
- Set a parameter ε and initialize $\varepsilon = 0$ then change ε across a set of multiple ascending order, i.e. $\varepsilon = \varepsilon_0, \varepsilon_1, \ldots, \varepsilon_{max}$ while $\varepsilon_0 < \varepsilon_1 < \cdots < \varepsilon_{max}$.
- For ε_i, corresponding k-dimensional simplexes denoted by $\sigma = \{v_1, v_2, \ldots, v_k + 1\}$ would belong to the Rips complex if and only if for every edge (v_i, v_j), $|v_i - v_j| \leqslant \varepsilon_i$ where $1 \leqslant i \leqslant j \leqslant k+1$.

ε can also be thought as radius of balls centered at v_i in Euclidean n-dimensional space $E = \mathbb{R}^n$ which is denoted as $\overline{B}(v_i, \varepsilon)$. When two balls have a common intersection, an edge is formed. When three balls intersect one another, a triangular face is formed. The same holds for higher-dimensional simplexes.

As illustrated in Figure 1, the invariant topological features in the filtration of Vietoris-Rips complexes are connected components (0-dimensional topological features), holes (1-dimensional topological features) and voids (2-dimensional topological features). With algebraic topology as its theoretical foundation, persistent homology uses kth homology to track the emergency and vanishing of k-dimensional topological features that persist in the filtration process.

With the increasing of ε, k-th homology classes emerge and disappear and by using a persistence diagram the information of how the homology changes across the filtration can be captured. The form of persistence diagram (Cohen-Steiner et al., 2010; Stolz, 2014) is $\overline{\mathbb{R}^2} := \mathbb{R}^2 \cup \mathbb{R} \times \{\infty\}$ and the multiplicity $\mu_k^{i,j} \in \overline{\mathbb{R}^2}$ counts the number of k-th homology classes that are born at ε_i and die at ε_j.

4. Data

In this study, we collect stock data from Shanghai stock exchange (SSE) and Shenzhen stock exchange (SZSE) from December 1990 to December 2020 including A-shares and Sci-Tech innovation board (STAR Market). For simplicity without loss of generality, we use the daily closing and split-adjusted share prices. For comparison with the ground truth of the change in the market, we collected SSE Composite Index and SZSE Component Index. The detailed data is obtained by using tushare data interface package.

The closing price of each stock can be represented by an n-dimensional vector $v_i = (v^1, v^2, \ldots, v^n)$ and let matrix $A = [v_1, v_2, \ldots, v_k]$ with k as the number of vectors in A. Considering suspension, the closing price may be absent which is unacceptable in latter data processing. Thus, data cleaning is necessary to make sure all the vectors are complete.

4.1. Preprocessing

4.1.1. Data cleaning

The closing price of data we gathered may be absent due to the suspension of the stock so data must be completed. The most common way of completing the missing data caused by suspension is to copy the previous day's data. Additionally, several different methods can be used such as treating missing attribute values as special values, simply deleting the stock from data, using techniques including but not limited to regression, machine learning and deep learning to predict the missing value (Chu et al.,

2016), (Rossmanith, 2003). We choose to consider the absent price to be the same as that of the previous trading day and if a stock is kept suspended during the study period, then we remove the stock from our data.

4.1.2. Data normalization

To examine and analyze the relationship between stocks that exhibit variations in market value and closing price across different orders of magnitude as well as to facilitate subsequent processing it is necessary to normalize the vectors during the data preprocessing stage. Here, we simply normalize data to 0–1 range as

$$\tilde{v}^i = \frac{v^i - v_{min}}{v_{max} - v_{min}} \tag{2}$$

where v_{min} and v_{max} are the minimum and maximum closing prices respectively of the stock we choose.

5. Method

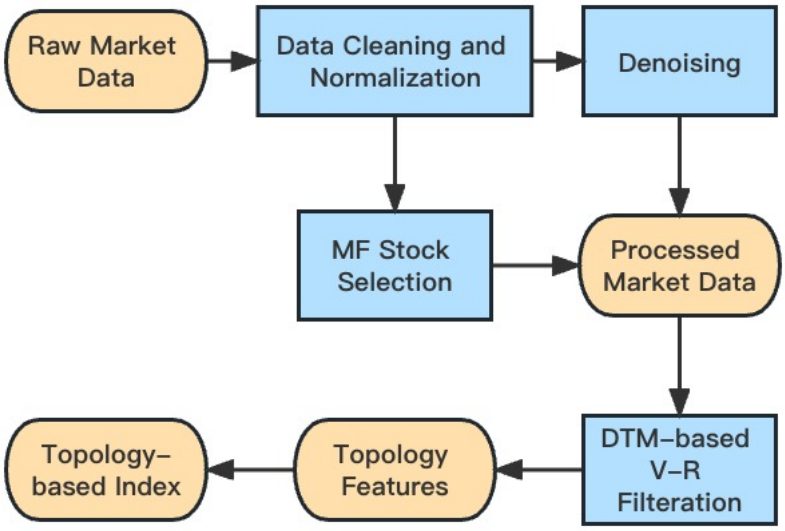

Figure 3. Pipeline of our work.

After preprocessing, we first use manifold technique to select stocks from the market then denoise the data by using wavelet transform, and combine the selected code and denoised data together as a new dataset. Finally, after DTM-based Vietoris-Rips filtration we obtain persistent diagram as topological features of the stock market. Figure 3 shows the scheme of our method.

5.1. Denoising

For financial data, due to the influence of various incidental factors in the market, financial data is characterized by considerable noise especially financial time series data. Pan suggested that chaotic dynamics imposed with some fractional order noise as well as conventional Gaussian noise comprises the noise in the market (Pan et al., 2012). Hence we will find a simple but effective method for denoising

in advance. However, the financial time series itself is characterized by non-stationary, non-linear and high signal-to-noise ratio. Traditional denoising processing methods often have many defects. Wavelet-based method is developed according to the requirements of time-frequency localization. It has the properties of self-adaptation and mathematical microscope, and is especially suitable for the processing of non-stationary and non-linear signals.

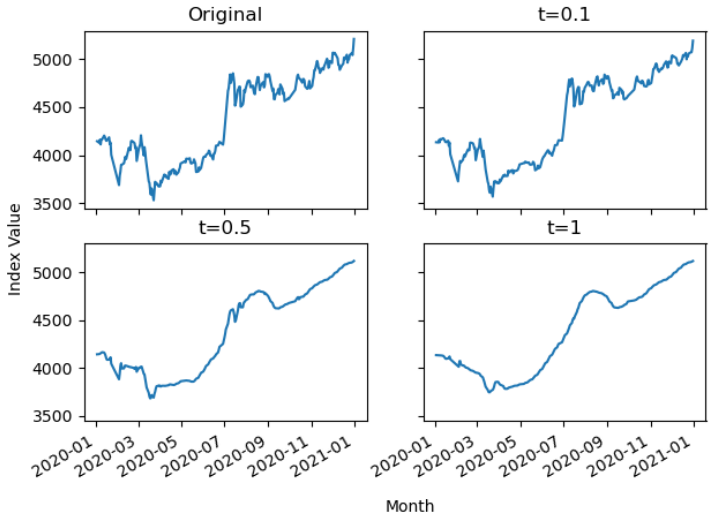

Figure 4. Different threshold for CSI300 in 2020, where t is the hard threshold in DWT.

The wavelet transform is similar to fourier transform but preserves some differences. Instead of decomposing the signal into sines and cosines, the wavelet transform uses functions that are localized in both time and frequency. Wavelets are small waves located in different times and can be stretched and shifted to capture features in both time and frequency. Thus, it provides information from both domains. Wavelet transform has two types, the continuous wavelet transform (CWT) and the discrete wavelet transform (DWT). We use the latter technique in our study. DWT is computed as

$$\begin{aligned} DWT_{\chi}^{\psi}(k,s) &= phi_{\chi}^{\psi}(\frac{k}{2^s},\frac{1}{2^s}) \\ &= \int_{\infty}^{\infty} \chi(t)\psi^*(\frac{t-k/2^s}{1/2^s})\,dt. \end{aligned} \tag{3}$$

DWT has been used for denoising in many fields in the last decade (Chen et al., 2003). Supposing the original data are in the form $y(n) = x(n) + e(n)$ where $y(n)$ is the observed data, $x(n)$is the original data and $e(n)$ is Gaussian white noise in the form $e(n) = N(0,\sigma^2)$. The main purpose of denoising is to reduce $e(n)$ as much as possible and not to disturb $x(n)$ simultaneously. Many important features of original signal are captured by a subset of DWT coefficients that is much smaller than the original signal itself. By choosing suitable threshold properly, coefficients can be kept after DWT. Shrinkage process which was proposed in 1995 provides hard thresholding and soft thresholding for performing filtration (Donoho et al., 1995). All the coefficients smaller than the threshold are set to 0 in hard thresholding. Both coefficients larger or smaller than the threshold are shrunken to 0 in soft thresholding. We will use hard thresholding in our later process.

5.2. Manifold learning based stock selection

In order to compare the persistence diagram derived from different stages of the stock market cycles without interference of the number of stock and the length of the time span, we introduce manifold learning. The number of stocks in stock exchange changes rapidly due to listing and delisting of companies. Components of current indexes also change periodically which prevent us from customizing the periods we would like to study. In this section, we apply a manifold learning based high-dimensional point cloud simplification method to the stock market data (Xu et al., 2022) while the stock dataset is regarded as point cloud in the following.

5.2.1. LBO on manifold

The Laplace-Beltrami operator (LBO) Δ is the divergence of the gradient. Suppose a differentiable manifold M with its Riemannian metric g and $f \in C^2$ is a real-valued function defined on M. The LBO Δ on M is defined as following:

$$\Delta f = div(grad\ f) \tag{4}$$

where *grad* means gradient operator and *div* represents divergence operator.

The eigenvectors and eigenvalues of Equation 4 can be calculated by solving the following equation:

$$\Delta f = -\lambda f \tag{5}$$

where λ is the eigenvalues of Equation 5 and the solution is

$$0 \leq \lambda_0 \leq \lambda_1 \leq \lambda_2 \cdots \leq +\infty \tag{6}$$

with corresponding eigenfunctions:

$$\phi_0(x), \phi_1(x), \phi_2(x), \ldots, . \tag{7}$$

When the manifold is closed, the smallest eigenvalue $\lambda_0 \equiv 0$.

The eigenvalues λ_i specify the discrete frequence domain of an LBO and the eigenfunctions are the extensions of the basis functions in the Fourier analysis to a manifold (Vallet and Lévy, 2008) and eigenfunctions corresponding to larger eigenvalues contain higher-frequency information from data (Dong et al., 2006). Hence, we use these eigenfunctions to choose components of the new simplified dataset and the details are introduced in the next section.

5.2.2. Discrete LBO on market data

We use normalized but non-denoised market data as the input, which has been introduced in the last section and the discretization method in (Belkin and Niyogi, 2003) to construct the discrete LBO. The method contains two steps:

1. Adjacency graph construction
2. Weight computation

Adjacency graph construction. We choose k-nearest neighbors (KNN) to construct an adjacency graph on market data. For any normalized $\tilde{v}_i$, its KNN set $N_{\tilde{v}_i} = \{\tilde{v}_j, j = 1, 2, \ldots, k_n\}$ contains k_n nearest

points of $\tilde{v}_i$ in Euclidean space. Thus, we can build a connection between isolated data points which can be presented in a directed graph.

Weight computation. Based on the adjacency graph, the weight w_{ij} between $\tilde{v}_i$ and $\tilde{v}_j$ can be calculated as follows:

$$w_{ij} = \begin{cases} -e^{-\frac{\|v_i - v_j\|_2^2}{t}}, & \text{if } i, j \text{ are adjacent,} \\ \Sigma_{k_n \neq i} - w_{ik_n}, & \text{if } i = j, \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

where t is an parameter for adjustment.

However, the KNN algorithm is not symmetric which means that supposing $\tilde{v}_m \in N_{\tilde{v}_n}$ but not $\tilde{v}_n \in N_{\tilde{v}_m}$, and the weighted matrix $\tilde{W} = [w_{ij}]$ constructed by Equation 8 is not symmetric. So we use

$$W = \frac{\tilde{W} + \tilde{W}^T}{2} \tag{9}$$

as the weight matrix, which is symmetric. Additionally, we construct

$$A = diag(w_{11}, w_{22}, \ldots, w_{nn}) \tag{10}$$

which is extracted from diagonal of W. The LBO on manifold can be discretized into the matrix

$$L = A^{-1}W \tag{11}$$

and Equation 5 is discretized as

$$L\phi = A^{-1}W\phi = \lambda\phi \tag{12}$$

which is equal to

$$W\phi = \lambda A\phi \tag{13}$$

where λ is the eigenvalues of L, and ϕ is the eigenvectors. It has been shown that λ satisfies (Belkin and Niyogi, 2003)

$$0 = \lambda_1 \leq \lambda_2 \leq \ldots \lambda_n, \tag{14}$$

and the eigenvectors are pairwise orthogonal. As mentioned above, the larger λ is, the greater is number of feature point in the corresponding ϕ. We identify the feature points of the eigenvector ϕ_1 and incorporate them into the simplified set. Subsequently, we repeat this process for ϕ_2, ϕ_3 and so on until the desired number of stocks is achieved. For each point x, we use its KNN neighbors N_x for detecting the local maximum and minimum of $\phi(x)$. The method is illustrated as follows:

Local extreme point detection: For a data point x in the data set, if $\forall y \in N_x, \phi(y) > \phi(x)$ then x is labeled as a local maximum point and vise versa for $\forall y \in N_x, \phi(y) < \phi(x)$, x is a local minimum point.

5.3. Distance-to-Measure Vietoris-Rips Filtration

The classical Vietoris-Rips filtration is highly susceptible to noise and outliers which hinders its direct applicability in practice (Anai et al., 2020). To mitigate the impact of noise and outliers, a Vietoris-Rips filtration based on Distance-to-Measure (DTM) is introduced in (Anai et al., 2020) leveraging the concepts from (Chazal et al., 2011). This modified approach aims to alleviate the adverse effects of

noise and outliers encountered in the original filtration method. Given a measure f, for every $v \in X$ and $\varepsilon \in \mathbb{R}^+$ we define

$$r_v(\varepsilon) = \begin{cases} -\infty & \text{if } \varepsilon < f(v), \\ (\varepsilon^p - f(v)^p)^{\frac{1}{p}} & \text{otherwise} \end{cases} \tag{15}$$

where the fixed real number $p \geqslant 1$. We denote the ball which has the radius $r_v(\varepsilon)$ by $\overline{B}_f(v,\varepsilon) = \overline{B}(v, r_v(\varepsilon))$. Some examples of $r_v(\varepsilon)$ are represented in Figure 5.

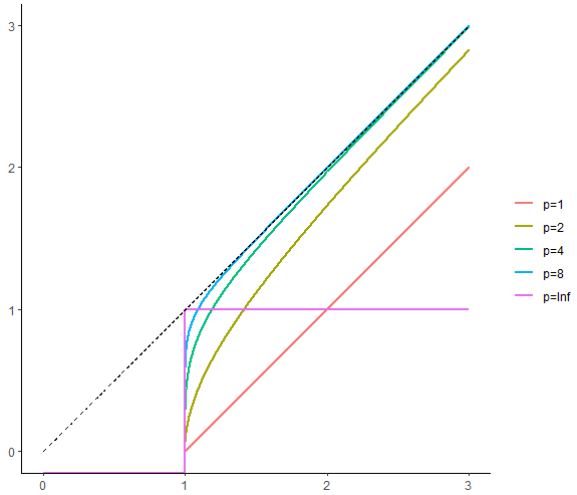

Figure 5. Graph of $r_v(\varepsilon)$ while $f(v) = 1$ for different values of p.

Equation 15 provides a weighted-based method to optimize the process of filtration. Balls on vertices with larger value of $r_v(\varepsilon)$ will latter grow than those with smaller values which may lead to less contribution to the form of key topological features. The core of DTM is the measure $f(v)$. For more detail on DTM please refer to (Anai et al., 2020).

The discrete form of DTM is

$$d^2(v) = \frac{1}{k_0} \sum_{k=1}^{k_0} \|v - p_k(v)\|^2 = f^2(v) \tag{16}$$

where $p_1(v), \ldots, p_{k_0}(x)$ are k_0 nearest neighbors of v. The DTM enables vertices that in higher-density area tend to start growing earlier, and filter more significant features than the original Vietoris-Rips filtration. Moreover, m is defined as $\frac{k_0}{k}$. An example of persistence diagram is shown in Figure 6.

6. Implementation and results

6.1. Implementation

We choose the daily data of SSE and SZSE from 01/01/2006 to 12/31/2017 and cut the time span at January 1 and July 1 in every year thus creating 30 slices. Detailed information is shown in Table 1. Slices from 07/01/2006 to 06/30/2009 and from 07/01/2014 to 12/31/2015 are chosen as bullish and bearish periods, respectively. The Chinese stock market suffered violent volatility during the two periods as shown in Figure 7.

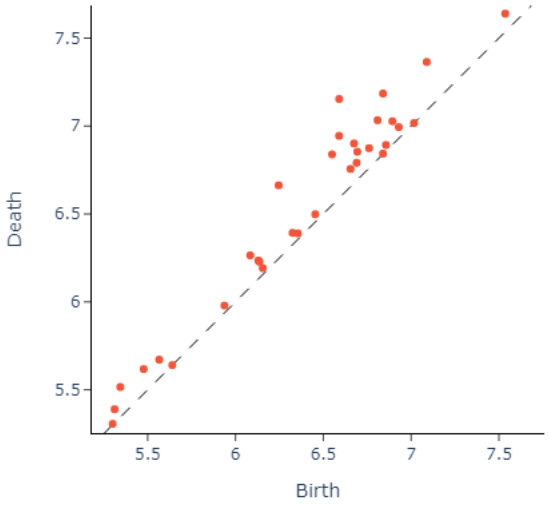

Figure 6. Example of persistence diagram generated by DTM based V-R filtration where number k of stocks is 100, m, p of DTM is set as $0.05, 1$.

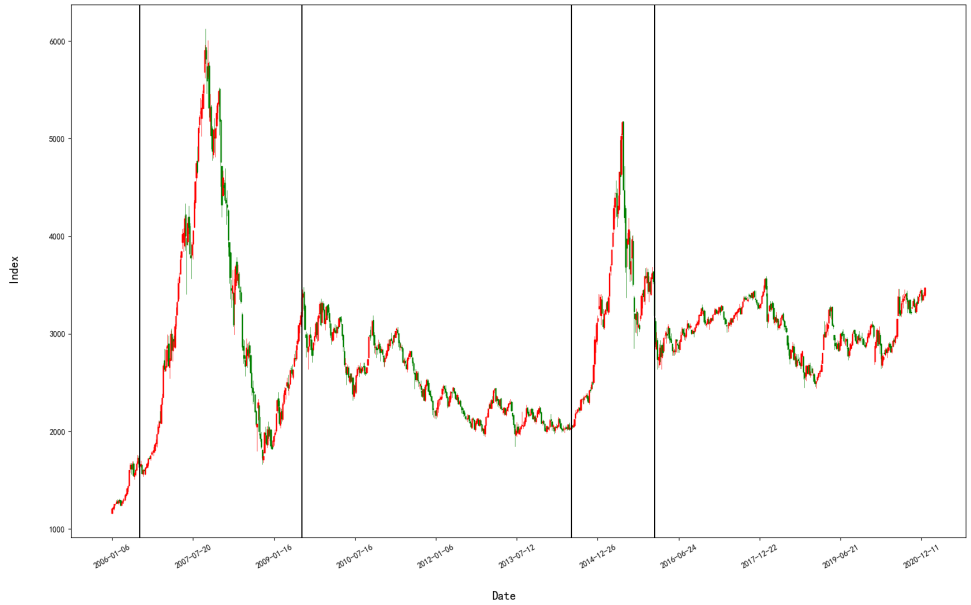

Figure 7. Weekly candlesticks chart of SSE index from 01/01/2006 to 12/31/2020 where the split line of normal times and bullish and bearish times are drawn by black solid line.

The datasets are denoised using db8 of Daubechies Wavelets (Daubechies, 1992) with threshold set to 0.01. We choose $k_n = 20$ for building the matrix of LBO on manifold. After selecting stocks using the method presented in Section 5.2, a persistence diagram is built by DTM based Vietoris-Rips filtration. We draw the persistence diagram first on the Birth-Death coordinate and then transform it to birth-duration for conspicuous results because almost all the feature points lie near $y = x$ in original persistence diagram.

Table 1. Time spans from 01/01/2006 to 12/31/2020.

Intervals	Length(trading day)[1]	Intervals	Length(trading day)
01/01/2006-06/30/2006	116	07/01/2013-12/31/2013	126
07/01/2006-12/31/2006[2]	128	01/01/2014-06/30/2014	120
01/01/2007-06/30/2007	122	07/01/2014-12/31/2014	126
07/01/2007-12/31/2007	130	01/01/2015-06/30/2015	120
01/01/2008-06/30/2008	122	07/01/2015-12/31/2015	126
07/01/2008-12/31/2008	128	01/01/2016-06/30/2016	122
01/01/2009-06/30/2009	126	07/01/2016-12/31/2016	126
07/01/2009-12/31/2009	132	01/01/2017-06/30/2017	120
01/01/2010-06/30/2010	120	07/01/2017-12/31/2017	128
07/01/2010-12/31/2010	124	01/01/2018-06/30/2018	122
01/01/2011-06/30/2011	120	07/01/2018-12/31/2018	126
07/01/2011-12/31/2011	128	01/01/2019-06/30/2019	124
01/01/2012-06/30/2012	120	07/01/2019-12/31/2019	125
07/01/2012-12/31/2012	126	01/01/2020-06/30/2020	118
01/01/2013-06/30/2013	118	07/01/2020-12/31/2020	126

[1] Both Shanghai and Shenzhen stock exchange would close in national statutory holidays. Since most of the national statutory holidays are in the first half of the year, the first half of the years always have less trading days than the rest.

[2] Bold time spans are chose as bullish and bearish times in the history of Chinese stock market while others are regarded as normal times. The candlesticks chart of SSE index is shown in Figure 7 where the split lines are drawn.

6.2. Results

The transferred persistence diagrams of DTM based filtration are shown in Figure 8 where the persistence diagrams of bullish and bearish periods are illustrated in the transferred coordinates. The number of stock k is set as 100, 300, respectively. The parameter m of DTM based Vietoris-Rips filtration can be selected from $0.05, 0.07$ and p from $1, 2$ m, p. Combining the three parameters in groups, we get eight results. In every group of comparison, the result significantly shows that the persistence diagram from normal period tends to have later birth time than that of bullish and bearish peroids and the parameters of k, m, p have little impact on the results. Thus, the persistence diagram of DTM based filtration is insensitive to parameter changes.

We also choose nine periods corresponding to bullish, bearish and normal periods and each label has three periods under it. The detailed information of the nine periods are shown in Table 2 and Figure 9.

By comparing the chosen periods, we found differences between the bullish and bearish periods. Figure 10 shows that the trending of three groups is totally different. By comparing the corresponding persistent diagrams, we can see that the persistence diagrams of bearish periods has earlier birth times

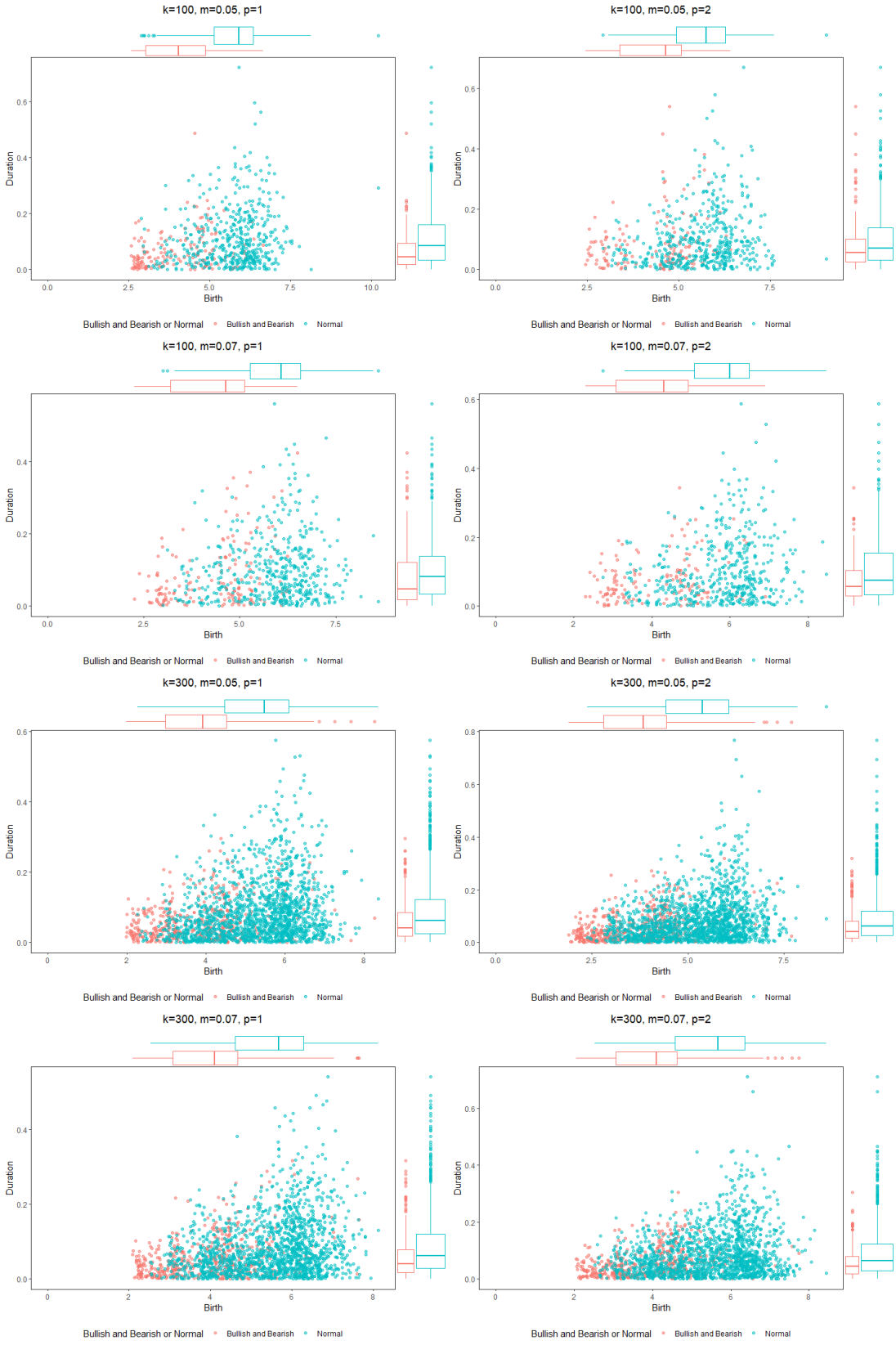

Figure 8. Transferred persistence diagram of DTM based filtration of the 30 slices where k is the number of stock selected by discrete LBO, $m = \frac{k_0}{k}$ and p is the parameter in $r_v(\varepsilon)$.

Table 2. Detailed information of bearish, bullish and normal periods.

Label	Interval	Time span(Trading day)	Change of SSE index
Bear	04/13/2010-07/01/2010	54	3129.69 to 2373.79
Bear	01/15/2008-04/04/2008	54	5503.93 to 3446.24
Bear	06/15/2015-08/27/2015	54	5174.42 to 3083.59
Bull	08/01/2007-10/18/2007	54	4488.77 to 5825.28
Bull	03/15/2015-05/30/2015	54	3391.16 to 4611.74
Bull	01/04/2019-03/30/2019	56	2446.02 to 3090.76
Normal	11/28/2009-02/14/2010	54	3114.29 to 3018.13
Normal	01/30/2014-04/24/2014	56	2045.93 to 2057.03
Normal	09/15/2019-12/04/2019	54	3041.92 to 2878.12

Table 3. p of pairwise Wilcoxon signed-rank test of every group of data.

	Parameters	$k=100\ m=0.05\ p=1$	$k=100\ m=0.05\ p=2$	$k=100\ m=0.07\ p=1$	$k=100\ m=0.07\ p=2$	$k=300\ m=0.05\ p=1$	$k=300\ m=0.05\ p=2$	$k=300\ m=0.07\ p=1$	$k=300\ m=0.07\ p=2$
Birth	Normal-Abnormal	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$
	Bearish-Bullish	2.804×10^{-15}	8.468×10^{-6}	2.052×10^{-6}	6.549×10^{-8}	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$
	Bearish-Normal	1.775×10^{-14}	5.475×10^{-15}	5.197×10^{-13}	4.832×10^{-13}	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$
	Bullish-Normal	1.775×10^{-14}	$< 2.2\times10^{-16}$	1.085×10^{-14}	4.631×10^{-11}	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$
Duration	Normal-Abnormal	1.929×10^{-8}	3.080×10^{-2}	1.084^{-3}	1.680×10^{-2}	3.415×10^{-10}	5.147×10^{-15}	4.606×10^{-15}	1.326×10^{-12}
	Bearish-Bullish	0.3258	7.072×10^{-2}	0.4263	7.699×10^{-2}	3.654×10^{-4}	5.477×10^{-4}	2.834×10^{-4}	9.751×10^{-3}
	Bearish-Normal	1.115×10^{-4}	9.846×10^{-3}	9.103×10^{-2}	2.555×10^{-2}	9.684×10^{-13}	1.102×10^{-8}	7.577×10^{-11}	1.316×10^{-8}
	Bullish-Normal	1.618×10^{-3}	5.630×10^{-2}	0.4263	0.7235	9.982×10^{-9}	2.187×10^{-4}	2.834×10^{-4}	7.410×10^{-7}

Figure 9. Candlestick charts of nine custom periods. Three rows correspond to bearish, bullish and normal periods, respectively.

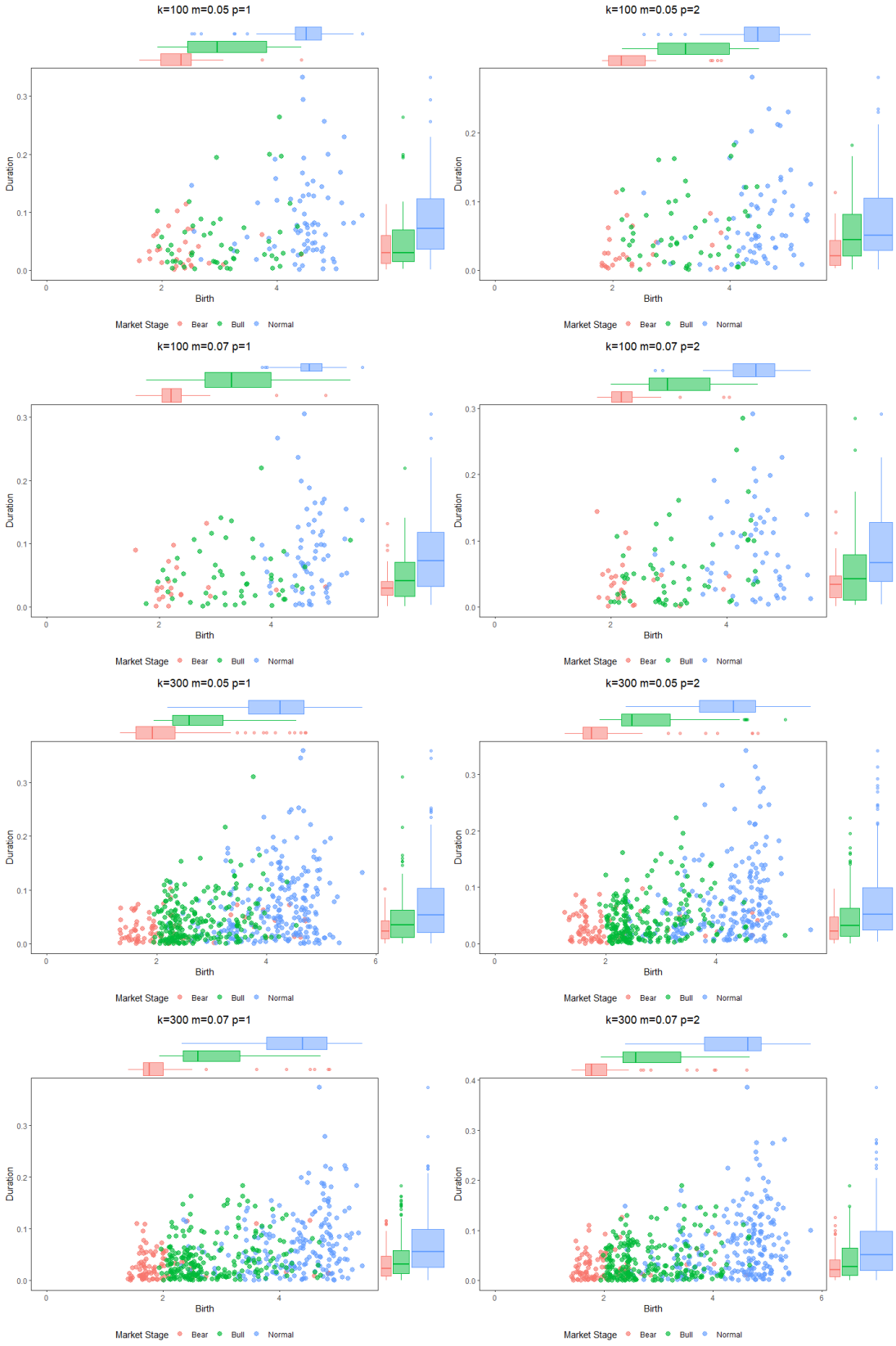

Figure 10. Transferred persistence diagram of DTM based filtration of 9 custom periods. Box-plot is drawn with respect to two axes.

than bullish periods and no other significant difference is found among two groups. However, the conclusion may be not so convincing because the percentage of gains and losses is different and the stock market tends to fall quicker than rise. In (Gidea, 2017), Gidea inferred that bullish period is similar to bearish period. The trends among different stock are close to each other and the other circumstance is the opposite conclusion.

Furthermore, from the comparison of the two groups we can infer that the more stocks are chosen, the more features could be obtained from filtration whereas the length of the time span has little effect on the number of features. Besides birth time, features from normal period have longer duration than that of bullish and bearish but the difference is less significant than birth time.

Table 3 summarizes the results of pairwise Wilcoxon test within each group of features of our study (Wilcoxon, 1992). Wilcoxon signed-rank test is a non-parametric statistical hypothesis test which does not assume samples for testing are normally distributed which is the property of persistent diagram feature points. From Table 3, we conclude that in every group of comparison of birth time there lies significant difference between compared features. However, in terms of duration using 100 or less stocks for extracting features does not show significant difference in some groups.

However, using 300 stocks we can highlight the differences which means the more stocks are selected for comparing, the more significant differences are shown. Moreover, the difference from the group of 14-years history of Chinese market is more significant than that from selected periods which may be attributed to the length of time span or fluctuation levels of market and needs more analysis.

7. Index computation

Based on previous results, we show the potential of applying TDA to extract features from stock market and summarizing the features into an index. With the index, we can easily compare the performance of market in different periods and detect the change of trend of market.

Currently, there is not a widely accepted standard for classifying stock market indices. Achelis roughly divided them into two categories (Achelis, 2001) market sentiment indicators and market strength indicators. The former predict market movements, i.e. bullish and bearish periods and latter measure the strength of the movements. Sentiment indicators tend to be stable but delaying and strength indicators usually are sensitive but responsive.

In this section, we propose a topology-based stock market index referred to as *topo index* which characterizes the similarity of movements of different stocks of the market.

We choose to use sliding window based method to compute topo index. The value of the index of each day is computed from a w days period ending on the day. But on certain days, too many stocks suspended so there are not enough stocks remaining after data cleaning for stock selection. So we change the strategy of completing missing data. Whole table would be complete first and a single data can only represent absent data **5** times, that is, 6 or more continuous absence won't be filled. After that, in each window we directly remove the stocks that remain uncompleted. In every window, we use the pipeline in Figure 3 with slightly different parameters to filter the stocks and compute topology features. After capturing the features, we compute the L^1 norm of features' births times which is topo index's value T of the certain window:

$$T = \frac{\sum b_i}{n} \tag{17}$$

where b_i is the birth time of ith feature.

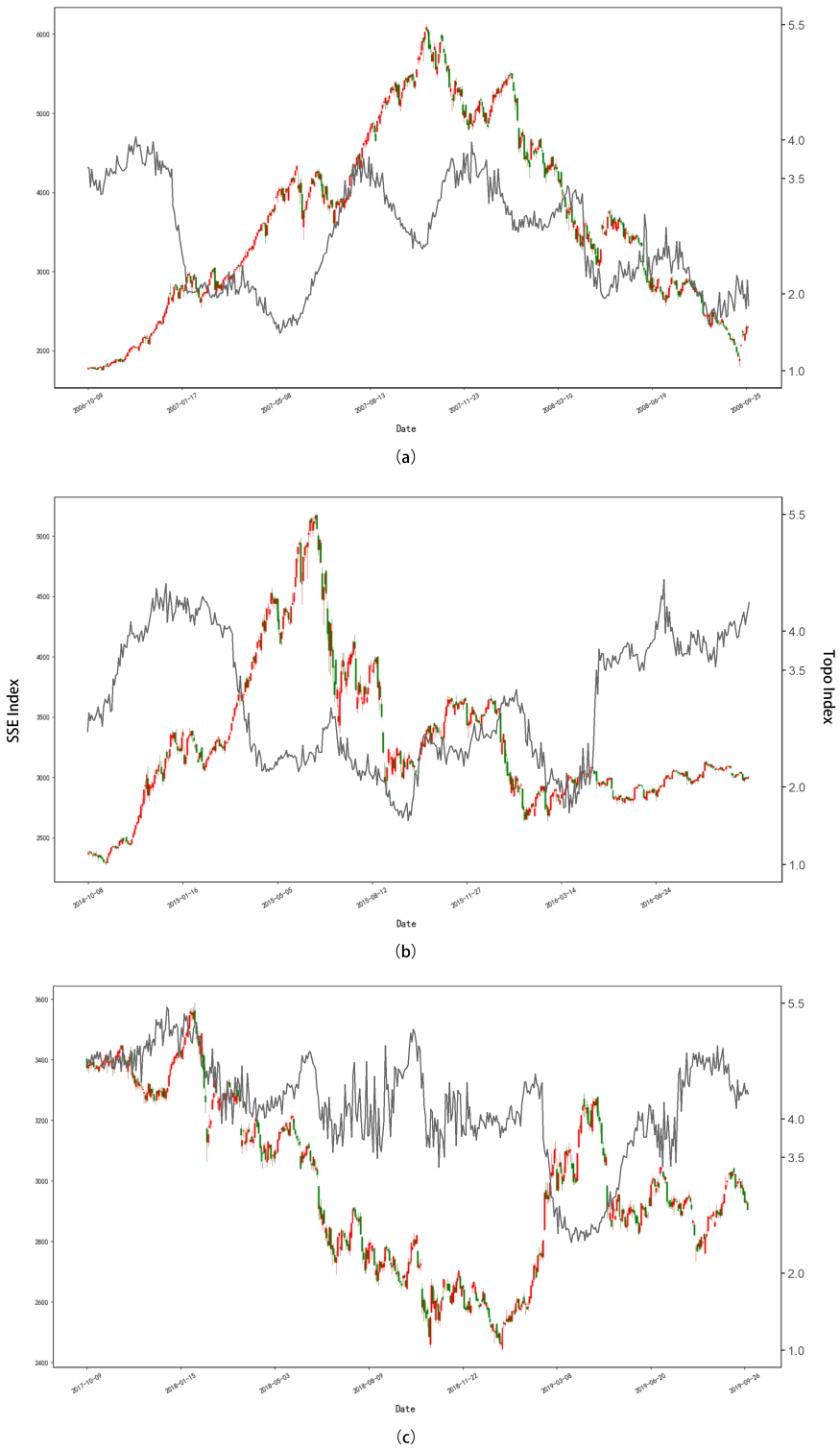

Figure 11. Topo-index of three different periods of Chinese stock market. (a) and (b) correspond to two major crashes in 2008 and 2015, and (c) is a relatively stable period around 2019.

We conduct a series of experiments by choosing different window size, number of stocks and other parameters in the original methods. Finally, we choose 300 as the number of stocks and 70 as the window size. Too few stocks leads to severe turbulence and too many leads to decrease of efficiency. 70 trading days is roughly three months or a quarter in real time which is not only a widely used length of cycle in economical researches but also the unit for measuring market cycle (Kusewitt Jr, 1985). As for other parameters, we choose $m = 0.07$ and $p = 1$ for least turbulence. The results we get are shown in Figure 11.

Topo index contains much potential information. First of all, the result mostly complies to the former conclusion. In normal times, the birth time of topology features is later than that of bullish times and bearish times. In Figure 11(b), we can see a complete cycle of Chinese stock market in 2015. Topo index keeps high and then falls with the rise of SSE index and falls again with the fall of SSE index and finally returns to a high level after the cycle.

Also, topo index has the potential for assisting in forewarning market crash. In Figure 11(a)(b), topo index stays at the platform at 2.0 without exceeding for few weeks which we infer that 2.0 may be the boundary between the birth time of topology feature of bullish and bearish periods.

Topo index is also very sensitive to the change of trend of the market. In Figure 11(b), after the crash in 04/2016 the topo index rises to normal level in few days rather than grows slowly like MA does. In some days when SSE has over 3% rise or fall, topo index also reacts significantly such as 19/04/2007 in which SSE fall by 4.52%.

At last, as control group topo index usually keeps at a high level with high fluctuation in normal period as shown in Figure 11. Although there is misleading around 02/2019, topo index recovers rapidly after the SSE get back to normal, and the boundary of normal period is around 3.5. If we want to use hard threshold to distinguish normal, bullish and bearish periods based on the above results, we can draw a rough conclusion: when topo index is above 3.5, the market is in normal period; between 2.0 and 3.5, bullish period; under 2.0, bearish period. The conclusion has a certain degree of uncertainty and may not always be applicable in some situations as other indices or signals but it works at most times.

8. Conclusions and discussion

In this paper, we proposed a topology-based feature extraction method. Not all the stocks have to be included in computation The users can customize the number of stocks by using manifold learning based stock selection without loss of generalization. The result shows significant difference between bullish, bearish and normal periods which can visualize and be used for further study such as deep learning.

We also developed a topology-based stock market index, i.e., topo index which possesses the advantage of low latency compared to other similar indices as it selects a quarter as a sliding window to evaluate the recent market trends.

However, the method has limitations. The method can only extract features of entire market, but fails to do so to sub-sectors of the market. Moreover, the value of the features being input of machine learning and deep learning for further analysis is still unproven.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

Achelis S (2001) Technical analysis from a to z: covers every trading tool from the absolute breath index to the zig zag.

Adams H, Emerson T, Kirby M, et al. (2017) Persistence images: A stable vector representation of persistent homology. *J Mach Learn Res* 18.

Anai H, Chazal F, Glisse M, et al. (2020) Dtm-based filtrations. In *Topological Data Analysis*, 2020: 33–66. Springer.

Appel G (1985) *The moving average convergence-divergence trading method: advanced version*. Scientific Investment Systems.

Appel G (2005) *Technical analysis: power tools for active investors*. FT Press.

Basu D, Li T (2019) A machine-learning-based early warning system boosted by topological data analysis. *Available at SSRN*.

Belkin M, Niyogi P (2003) Laplacian eigenmaps for dimensionality reduction and data representation. *Neural comput* 15: 1373–1396. https://doi.org/10.1162/089976603321780317

Bollinger J (2002) *Bollinger on Bollinger bands*. McGraw-Hill New York.

Bubenik P, Dłotko P (2017) A persistence landscapes toolbox for topological statistics. *J Symb Comput* 78: 91–114. https://doi.org/10.1016/j.jsc.2016.03.009

Carlsson G (2009) Topology and data. *B Am Math Soc* 46: 255–308. https://doi.org/10.1090/S0273-0979-09-01249-X

Carrière M, Oudot SY, Ovsjanikov M (2015) Stable topological signatures for points on 3d shapes. In *Computer graphics forum*, 34: 1–12. Wiley Online Library. https://doi.org/10.1111/cgf.12692

Chazal F, Cohen-Steiner D, Mérigot Q (2011) Geometric inference for probability measures. *Found Comput Math* 11: 733–751. https://doi.org/10.1007/s10208-011-9098-0

Chen AS, Leung MT, Daouk H (2003) Application of neural networks to an emerging financial market: forecasting and trading the taiwan stock index. *Comput Oper Res* 30: 901–923. https://doi.org/10.1016/S0305-0548(02)00037-0

Chu X, Ilyas IF, Krishnan S, et al. (2016) Data cleaning: Overview and emerging challenges. In *Proceedings of the 2016 international conference on management of data*, 2201–2206. https://doi.org/10.1145/2882903.2912574

Cohen-Steiner D, Edelsbrunner H, Harer J (2007) Stability of persistence diagrams. *Discrete computat geom*, 37: 103–120. https://doi.org/10.1007/s00454-006-1276-5

Cohen-Steiner D, Edelsbrunner H, Harer J, et al. (2010) Lipschitz functions have l p-stable persistence. *Found comput math* 10: 127–139. https://doi.org/10.1007/s10208-010-9060-6

Corzo T, Prat M, Vaquero E (2014) Behavioral Finance in Joseph de la Vega's Confusion de Confusiones. *J Behav Financ* 15: 341–350. https://doi.org/10.1080/15427560.2014.968722

Daubechies I (1992) *Ten lectures on wavelets*. SIAM. https://doi.org/10.1137/1.9781611970104

Deng M (2008) On the theoretical foundation of technical analysis: Market action discounts everything. *Available at SSRN*. https://doi.org/10.2139/ssrn.1259164

Dong S, Bremer PT, Garland M, et al. (2006) Spectral surface quadrangulation. In *ACM SIGGRAPH 2006 Papers*, 1057–1066. https://doi.org/10.1145/1179352.1141993

Donoho DL, Johnstone IM, Kerkyacharian G, et al. (1995) Wavelet shrinkage: asymptopia? *J R Stat Soc B* 57: 301–337. https://doi.org/10.1111/j.2517-6161.1995.tb02032.x

Edelsbrunner H, Harer J (2010) *Computational topology: an introduction*. American Mathematical Society. https://doi.org/10.1090/mbk/069

Edelsbrunner H, Letscher D, Zomorodian A (2000) Topological persistence and simplification. In *Proceedings 41st annual symposium on foundations of computer science*, 454–463. IEEE.

Gidea M (2017) Topological data analysis of critical transitions in financial networks. In *International conference and school on network science*, 47–59. Springer. https://doi.org/10.2139/ssrn.2903278

Gidea M, Goldsmith D, Katz Y, et al. (2020) Topological recognition of critical transitions in time series of cryptocurrencies. *Physica A* 548: 123843. https://doi.org/10.1016/j.physa.2019.123843

Gromov M (1987) Hyperbolic groups. In *Essays in group theory*, 75–263. Springer.

Guo H, Yu H, An Q, et al. (2022), Risk analysis of china's stock markets based on topological data structures. *Procedia Comput Sci* 202: 203–216. https://doi.org/10.1016/j.procs.2022.04.028

Hensel F, Moor M, Rieck B (2021) A survey of topological machine learning methods. *Front Artif Intell* 4: 52. https://doi.org/10.3389/frai.2021.681108

Hu X, Li F, Samaras D, et al. (2019) Topology-preserving deep image segmentation. *Advances in neural information processing systems*, 32.

Hu X, Wang Y, Fuxin L, et al. (2021) Topology-aware segmentation using discrete morse theory. *arXiv preprint arXiv*.

Katz YA, Biem A (2021) Time-resolved topological data analysis of market instabilities. *Physica A* 571: 125816. https://doi.org/10.2139/ssrn.3581869

Kirkpatrick II CD, Dahlquist JA (2010) *Technical analysis: the complete resource for financial market technicians*. FT press.

Kusewitt Jr JB (1985) An exploratory study of strategic acquisition factors relating to performance. *Strategic Manage J* 6: 151–169. https://doi.org/10.1002/smj.4250060205

Lum PY, Singh G, Lehman A, et al. (2013) Extracting insights from the shape of complex data using topology. *Sci rep* 3: 1236. https://doi.org/10.1038/srep01236

Nison S (1994) *Beyond Candlesticks: New Japanese Charting Techniques Revealed*. John Wiley & Sons. ISBN 9780471007203.

Pan I, Korre A, Das S, et al. (2012) Chaos suppression in a fractional order financial system using intelligent regrouping pso based fractional fuzzy control policy in the presence of fractional gaussian noise. *Nonlinear Dynam* 70: 2445–2461. https://doi.org/10.1007/s11071-012-0632-7

Perea JA, Harer J (2015) Sliding windows and persistence: An application of topological methods to signal analysis. *Found Comput Math* 15: 799–838. https://doi.org/10.1007/s10208-014-9206-z

Prabowo NA, Widyanto RA, Hanafi M, et al. (2021) With topological data analysis, predicting stock market crashes. *Int J Informatics Inf Syst* 4: 63–70. https://doi.org/10.47738/ijiis.v4i1.78

Rossmanith R (2003) Completion of market data. *Math Financ*.

Sheehy DR (2012) Linear-size approximations to the vietoris–rips filtration. In *Proceedings of the twenty-eighth annual symposium on Computational geometry*, 2012: 239–248. https://doi.org/10.1145/2261250.2261286

Skraba P, De Silva V, Vejdemo-Johansson M (2012) Topological analysis of recurrent systems. In *NIPS 2012 Workshop on Algebraic Topology and Machine Learning, December 8th, Lake Tahoe, Nevada*, 1–5.

Stolz B (2014) Computational topology in neuroscience. *Master's thesis* (University of Oxford, 2014).

Teixeira LA, De Oliveira ALI (2010) A method for automatic stock trading combining technical analysis and nearest neighbor classification. *Expert Syst Appl* 37: 6885–6890. https://doi.org/10.1016/j.eswa.2010.03.033

Topaz CM, Ziegelmeier L, Halverson T (2015) Topological data analysis of biological aggregation models. *PloS One* 10: e0126383. https://doi.org/10.1371/journal.pone.0126383

Vallet B, Lévy B (2008) Spectral geometry processing with manifold harmonics. In *Computer Graphics Forum*, 27: 251–260. Wiley Online Library. https://doi.org/10.1111/j.1467-8659.2008.01122.x

Wang F, Liu H, Samaras D, et al. (2020) Topogan: A topology-aware generative adversarial network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*,118–136. Springer. https://doi.org/10.1007/978-3-030-58580-8_8

Wang F, Kapse S, Liu S, et al. (2021) Topotxr: A topological biomarker for predicting treatment response in breast cancer. In *Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event*, 386–397. Springer. https://doi.org/10.1007/978-3-030-78191-0_30

Westlin E (2022) Using topological data analysis on credit data to predict stock market crashes.

Wilcoxon F (1992) Individual comparisons by ranking methods. In *Breakthroughs in statistics*, 196–202. Springer. https://doi.org/10.1007/978-1-4612-4380-9_16

Wilder JW (1978) *New concepts in technical trading systems*. Trend Research.

Wu P, Chen C, Wang Y, et al. (2017) Optimal topological cycles and their application in cardiac trabeculae restoration. In *Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA,* 25: 80–92. Springer.

Xu C, Lin H, Fang X (2022) Manifold feature index: A novel index based on high-dimensional data simplification. *Expert Syst Appl* 200: 116957. https://doi.org/10.1016/j.eswa.2022.116957

Yen PTW, Cheong SA (2021) Using topological data analysis (tda) and persistent homology to analyze the stock markets in singapore and taiwan. *Front Phys* 20.

Yen PTW, Xia K, Cheong SA (2021) Understanding changes in the topology and geometry of financial market correlations during a market crash. *Entropy* 23: 1211. https://doi.org/10.3390/e23091211

Yen PTW, Xia K, Cheong SA (2023) Laplacian spectra of persistent structures in taiwan, singapore, and us stock markets. *Entropy* 25: 846. https://doi.org/10.3390/e25060846

Zeng S, Graf F, Hofer C, et al. (2021) Topological attention for time series forecasting. *Adv Neur Inf Processing Syst* 34: 24871–24882.

AIMS Press