Let Topology Tell Stories: Prediction of Chinese Stock Market Crashes Based on Dynamic-sized Sliding Window

Research Progress Report *

Gehuiming Zhu, Dake Zhang

Latest version: 2022-07-02

Contents

1 Introduction

Since the occurrence of some global stock market crashes, a large literature in academia has focused on evaluating the volatility of stock markets and identify potential stock market crashes in the near future, because a robust and accurate Early Warning System (EWS) can reduce the negative impact on society and help people make optimal

*Note: **1.This report is only available to a select group of people, such as advisors, interviewers, etc. Please do not distribute or retain without authorization; 2.This is just a draft PDF of our research progress, currently not an informal draft paper (although it does look like a prototype paper). Therefore, the accuracy of the method needs further discussion, and we will make major changes. At the same time, in order to explain some methods and concepts, we quoted and simply rewrote some mathematical definitions in textbooks or papers. We are also thinking about how to make these concepts more innovative and more relevant to our research.**

portfolios.A stock market crash refers to an abnormal economic phenomenon of a sudden plummeting of stock prices in a short time, on the condition that the market's internal contradictions gradually accumulate and finally it enters an extremely unstable state where the crash can be elicited by any small accidental factor [Sornette, 2009]. The crash can quickly spread to other related sectors and disrupt the financial market infrastructure in the end. Over the past three decades, there have been several crashes in the Chinese A-share market, resulting in the reduction of the wealth of investors and impediments of national economic development.

Traditional approaches for prediction of stock market crashes can be roughly divided into three categories: technical analysis, time series analysis, machine learning-based methods. With the advances in machine learning algorithms, many researchers have utilized them to obtain a relatively comprehensive understanding of stock markets with handcrafted features. Common methods choose to incorporate appropriate indicators to a specific model, like the logit model, to make it able to predict crashes. In recent years, neural networks (NNs) have demonstrated their ability to recognize patterns in time series and predict the stock market crisis in various time frames.

In the above-mentioned methods, modeling with handcrafted features relies heavily on prior statistical assumptions of the market, which could be inaccurate and incomplete in some cases, and they fail to consider the strong relationship among different stock indices [Madaleno and Pinho, 2012]. Meanwhile, NNs require a sufficient quantity and quality of training data which is usually not available when it comes to the stock market crashes. Additionally, NNs-based methods are time-consuming and lack of interpretability.

With applications of modern mathematics on data analysis becoming increasingly widespread, we tend to introduce Topological Data Analysis (TDA), a method that combines topology, statistics, and computational geometry. TDA aims at excavating and extracting the valuable shape information hidden in high-dimensional and noisy data sets [Lum et al., 2013] without any prior statistical assumption. Over the past few years, TDA has demonstrated its remarkable capability in various fields: Nielson et al. [2015] use TDA to discover novel relationship between traumatic brain injury (TBI) and spinal cord injury (SCI); Zhu [2013] proposes Similarity Filtration with Time Skeleton (SIFTS) algorithm based on TDA as a new structure representation of text; Jassim and Asaad [2018] utilize TDA to detect image tampering with the example of face images.

One of the most powerful tools in TDA is persistent homology. What topology studies are some special geometric features that remain the same when the shape changes continuously, called "topological features". And persistent homology provides us with a way to look at the full picture of the high-dimensional data set without the need for dimensionality reduction. In other words, it puts our data in its original and high-dimensional space but can provide insights about the underlying topological features. Additionally, persistent homology is a dynamic procedure and thereby tolerant to noise within the data set since the main focus is on the noticeable features. Furthermore, unlike other mathematical models, persistent homology does not rely on any statistical assumptions, which normally are the bottlenecks of the general applicability of those models.

In this paper, we propose a modified approach based on TDA, using different stock market indices, to analyze financial time series so as to forecast stock market crashes in an accurate and robust way. Our approach pays attention to the relationship among different stock market indices from the perspective of topological properties instead of specific economic indicators. The contributions of our paper are as follows: (1) We summarize a pipeline for the application of TDA on the prediction of stock market crashes. It can provide support for future TDA research in finance. (2) Our method demonstrates remarkable performance in predicting upcoming stock market crashes. Especially, we apply Dynamic-sized Sliding Window to improve the time efficiency while ensuring that the captured topological features are prominent enough to be detected, which is more precise than using the existing fixed sliding window method without reasonable explanation regarding the choice of the window size [Gidea and Katz, 2018, Gidea et al., 2020].

2 Methods

2.1 Background Knowledge of TDA

2.1.1 Data Input

As a first step, we focus on the data input. Suppose we have a set of d-dimensional points $\{x_1, \ldots, x_n\}$, in the form of $x_i = \left(x_i^1, \ldots, x_i^d\right) \in \mathbb{R}^d$, $i \in [1, n]$. Then we can construct a point cloud $X_1 = [x_1, \ldots, x_n]^T$ in Euclidean space $\mathbb{R}^d$ that consists of n points. Therefore, a point cloud is a $d \times n$ matrix. The following procedures are based on this data format.

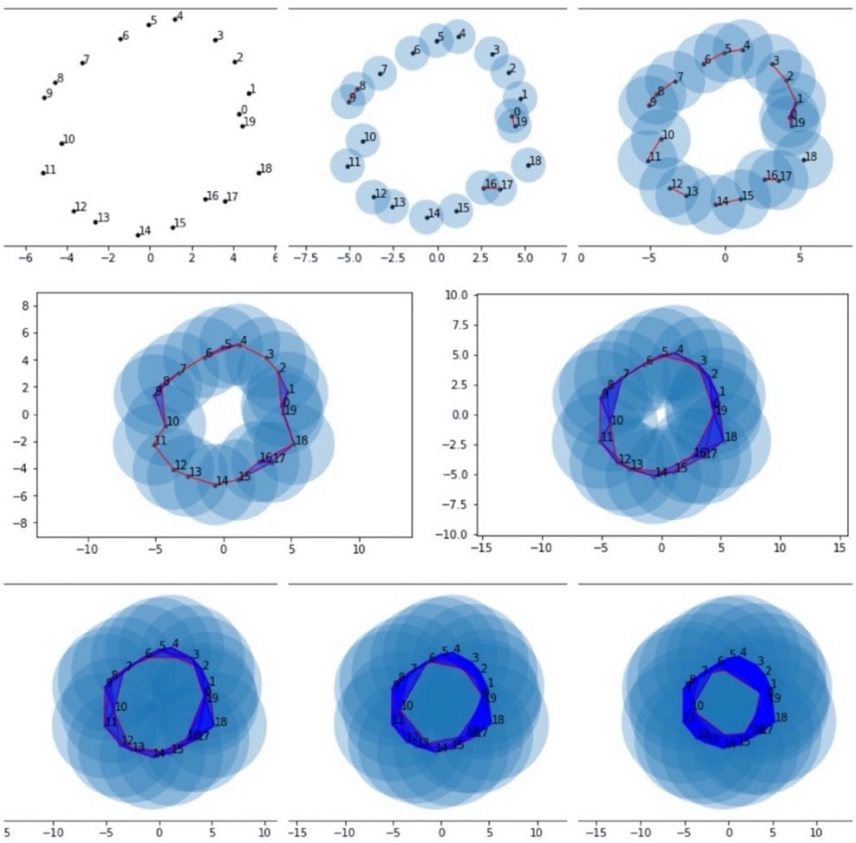

Figure 1: Rips filtration of simplicial complexes with ε from 0 to 7.

2.1.2 Simplicial Complex, Rips filtration and Homology

We simply review some basic concepts of Topological Data Analysis. In topology, a simplicial complex K is a set of simplices that satisfies the following conditions:

- if τ is a face of σ, and σ is a simplex from K, then τ is also in K. To express in mathematics, if $\tau \subseteq \sigma$ and $\sigma \in K$, then $\tau \in K$.
- if τ is the intersection of any two simplices $\sigma_1, \sigma_2 \in K$, then τ is a face to both σ_1 and σ_2. To express in mathematics, if $\sigma_1, \sigma_2 \in K$, $\tau = \sigma_1 \cap \sigma_2$, then $\tau \subseteq \sigma_1, \tau \subseteq \sigma_2$.

For one of the point clouds we input, $Z = (z_1, \ldots, z_w)$ in gspace $\mathbb{R}^d$, we associate it to a topological space as follows. A scale $\varepsilon > 0$ is introduced, and we define the Vietoris-Rips simplicial complex $R(Z, \varepsilon)$, or simply Rips complex as follows:

$$R(Z, \varepsilon) = \{\sigma \subseteq Z \mid d(z_i, z_j) \leq \varepsilon, \forall z_i \neq z_j \in \sigma\} \tag{1}$$

where $d(z_i, z_j)$ is the Euclidean metric between z_i and z_j.

In mathematics, a filtration $\mathscr{F}$ is an indexed set S_i of sub-objects of a given algebraic structure S, with the index i running over some index set I that is a totally ordered set, subject to the condition that: if $i \leq j$ in I, then $S_i \subset S_j$. The Rips simplicial complexes $R(Z, \varepsilon)$ form a filtration (see figure 1), in the condition that:

$$\text{if } \varepsilon < \varepsilon', \text{ then } R(Z, \varepsilon) \subseteq R(Z, \varepsilon'). \tag{2}$$

Informally, the homology groups: $H_0(R(Z, \varepsilon))$, $H_1(R(Z, \varepsilon))$, $H_2(R(Z, \varepsilon)), \ldots, H_n(R(Z, \varepsilon))$, represent the homology of a Rips complex $R(Z, \varepsilon))$, a set of topological invariants of $R(Z, \varepsilon)$. The n-th homology group $H_n(R(Z, \varepsilon))$ describes the n-dimensional holes in $R(Z, \varepsilon)$. In this case, we are only interested in the 1-dimensional cycles existing during the filtration. According to Zhu [2013], the 1-dimensional loops are informative since they can provide insights into periodic and repetitive patterns often found in time-series data and therefore can serve as important discriminating

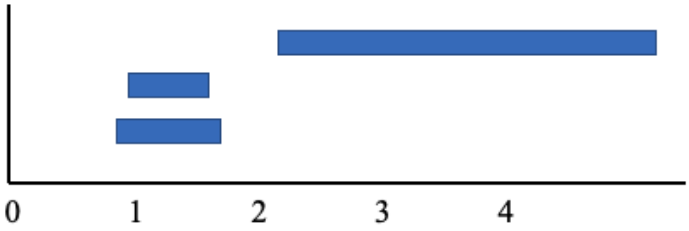

Figure 2: The persistence barcode, the length of the barcode represents the persistence of the topological feature. The longer barcode represents a signal while the shorter one represents a noise.

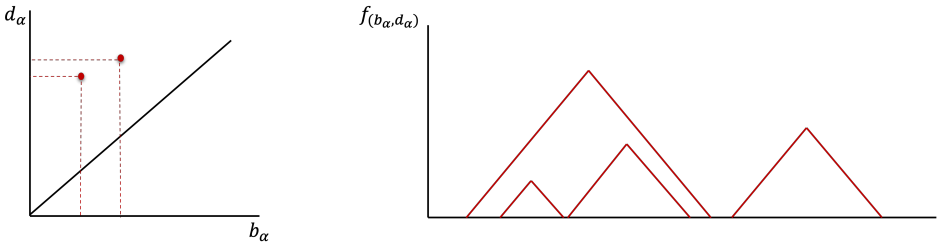

Figure 3: The persistence diagram and persistence landscapes.

descriptors. Therefore, we focus our efforts on computing and analyzing 1-dimensional persistence diagrams for the characterization of periodic and repetitive patterns.

Similar to the Rips complexes, the corresponding homologies also has the filtration property, that is, if $\varepsilon < \varepsilon'$, then $H_n(R(Z,\varepsilon)) \subseteq H_n\left(R\left(Z,\varepsilon'\right)\right)$. These inclusions determine canonical homomorphisms $H_n(R(Z,\varepsilon)) \hookrightarrow H_n\left(R\left(Z,\varepsilon'\right)\right)$, for $\varepsilon < \varepsilon'$.

One of the extraordinary features of TDA is the tolerance of small turbulence, it means small deformation and distortion of data set only results in trivial changes in the persistence diagram and persistence landscapes [Gidea and Katz, 2018]. This feature is regarded as "robustness", and that is the reason why persistence homology can be used to extract the valuable information hidden in noisy data sets.

2.1.3 Representations of Topological Changes: Persistence Barcode, Persistence Diagrams and Persistence Landscapes

We hypothesized there exists two values $\varepsilon_1 < \varepsilon_2$, and there is a non-zero k-dimensional homology class $\alpha \in H_n\left(R\left(Z,\varepsilon_1\right)\right.$), but $\alpha \notin H_n\left(R\left(Z,\varepsilon_1-\delta\right)\right)$, for $\delta > 0$, even if δ is indefinitely small. Also, for all $\varepsilon_1 < \varepsilon' < \varepsilon_2$, $\alpha \in H_n\left(R\left(Z,\varepsilon'\right)\right)$, but $\alpha \notin H_n\left(R\left(Z,\varepsilon_2\right)\right)$. In this case, ε_1 can be viewed as the 'birth' value (b_α) of the class α, and ε_2 can be viewed as the 'death' value (d_α) of it. The persistence of class α, or the time α lasts, is measured by $t_\alpha\left(b_\alpha,d_\alpha\right) = d_\alpha - b_\alpha$. The α that persists for a longer time is called a significant one, or a 'signal', while the α that persists for a shorter time is called a noisy one, or a 'noise'. The persistence of α can be reflected directly from the persistence barcode (see figure 2). The length of the barcode represents the persistence of the topological feature.

The information on the n-dimensional homology generators at all scales can be encoded in a persistence diagram P_n. The horizontal axis of it corresponds to the birth value (b_α) of the class α, and the vertical axis corresponds to the death value (d_α). The clinodiagonal is generally added to the coordinate axis, where the further away the point $\rho_\alpha\left(b_\alpha,d_\alpha\right)$ is from this axis, the longer persistence it has (see figure 3).

However, the persistence barcode and persistence diagram we introduced are the relatively conventional topological summaries. Both of them have difficulties with statistical computing due to the complex geometry of the space. Bubenik et al. [2015] define a new topological summary for data called 'persistence landscapes' (see figure 3). This tool embeds the space of persistence diagram into a function in Banach space $L^P(\mathbb{N}\times\mathbb{R})$. For each point in P_n, the linear function is as follows:

$$f_{(b_\alpha,d_\alpha)} = \begin{cases} x - b_\alpha, & \text{if } x \in \left(b_\alpha, \frac{b_\alpha+d_\alpha}{2}\right) \\ -x + d_\alpha, & \text{if } x \in \left(\frac{b_\alpha+d_\alpha}{2}, d_\alpha\right) \\ 0, & \text{if } x \notin (b_\alpha, d_\alpha) \end{cases} \tag{3}$$

Table 1: Pseudo Code for a Persistent Homology Process

Time Series Featurization via TDA
Input: A d-dimensional time series: $\Sigma = (t_1, \ldots, t_N), t_i = \left(t_i^1, \ldots, t_i^d\right) \in \mathbb{R}^d, i \in [1,\ \mathrm{N}]$.
1. Construct a point cloud $T_i = \left[t_i, \ldots, t_{i+(w-1)}\right]^T$ in Euclidean space $\mathbb{R}^d$.
2. Construct the Rips complex, and compute persistence diagram, as well as b_α and d_α.
3. Compute the persistence landscape $f_{(b_\alpha, d_\alpha)}$.
4. Compute the L^1 and L^2 norms.
Output: The L^1 and L^2 norms of all point clouds of Σ.

Every $f_{(b_\alpha, d_\alpha)}$ is linearly piecewise. In order to combine those pairs, we utilize a sequence of functions $\lambda = (\lambda_i)_{i \in \mathbb{N}}$, where $\lambda_i \in \mathbb{R}$ is given by

$$\lambda_i(x) = \begin{cases} i \to \max\left\{f_{(b_\alpha,d_\alpha)}(x) \mid (b_\alpha, d_\alpha) \in P_n\right\}, & f_{(b_\alpha,d_\alpha)}(x) \text{ exits} \\ 0, & f_{(b_\alpha,d_\alpha)}(x) \text{ does not exit} \end{cases} \tag{4}$$

where $i \to \max$ denotes the i-th largest value of $f_{(b_\alpha,d_\alpha)}(x)$.

Therefore, the persistent landscape form a Banach space $L^P(\mathbb{N} \times \mathbb{R})$, the norm of λ is given by:

$$\|\lambda\|_p = \left(\int_\Omega |\lambda|^p du\right)^{1/p} \tag{5}$$

where $p \geq 1$, du is with respect to the Lebesgue measure on $\mathbb{R}$. In this paper we will only use the L^1 and L^2 norms.

2.1.4 Time Series Featurization via TDA

In the previous introduction to the method of topology data analysis, we base on the angle of a given set of points, means that the point in the point cloud is a given group of stationary points, without asking us to consider how these points come. Here we will use the principles described above to briefly clarify how topological data analysis is applied over time series. The application of TDA in time series mainly uses Taken's Theorem, which means suppose we have a d-dimensional time series of length $\mathrm{N} : \Sigma = (t_1, \ldots, t_N)$, $t_i = \left(t_i^1, \ldots, t_i^d\right) \in \mathbb{R}^d$, $i \in [1,\ \mathrm{N}]$. Then for each t_i in Σ, we supply a sliding window of length w. Therefore, we have a point cloud T_i in Euclidean space $\mathbb{R}^d$ that consists of w points:

$$T_i = \left[t_{i,\ldots,}t_{i+(w-1)}\right]^T = \begin{bmatrix} t_i^1 & t_i^2 & \cdots & t_i^d \\ t_{i+1}^1 & t_{i+1}^2 & \cdots & t_{i+1}^d \\ \vdots & \vdots & \ddots & \vdots \\ t_{i+(w-1)}^1 & t_{i+(w-1)}^2 & \cdots & t_{i+(w-1)}^d \end{bmatrix} \tag{6}$$

It is a $d \times w$ matrix.

2.1.5 Summary: Pseudo Code for a Persistent Homology Process

The following persistent homology process of a time series can be summarized in table 1.

In this paper, we will use a 4-dimensional time series consisting of four stock market indices: SHANGHAI COMPOSITE INDEX, SZSE COMPONENT INDEX, CSI300, and SSE SME COMPOSITE between January 24, 2006 and January 18, 2019. We use this time series to perform a persistent homology dynamic process introduced above and calculate the norms, in order to test the prediction effect of TDA on stock market crashes.

2.2 TDA Feasibility Verification: Testing on Synthetic Time Series

In order to validate persistent homology's ability to detect the increased volatility of given time series, we will test it with a synthetic time series. We utilize a discrete-time 2-dimensional dynamical system called the Duffing map (also called as 'Holmes map'). It is an example of a dynamical system that exhibits different behaviors with different values of parameters a and b. The Duffing map takes a point (x_n, y_n) in the plane and maps it to a new point given by:

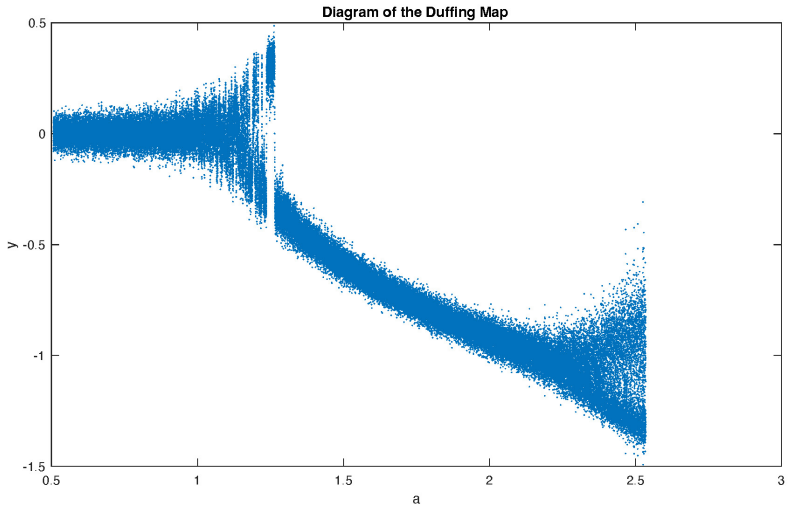

Figure 4: The diagram of the Duffing map with a noise ($b = 0.15$), in (a, y)-coordinates.

$$\begin{aligned} x_{n+1} &= y_n \\ y_{n+1} &= -bx_n + ay_n - y_n^3 \end{aligned} \tag{7}$$

We modify the system by making the parameter a change slowly at each step of the iteration to simulate the elapsing time. Specifically, we add Δt to parameter a_n to obtain a_{n+1} for the next iteration. In this case, Δt is set to be 1.0×10^{-4}. Therefore, the series of y_n can be viewed as a time series. Small Gaussian noise (τ) is also added to the system to test the robustness of persistent homology. We set the SIGNAL-NOISE RATIO(SNR) to be 30 , and the noise is automatically generated by MATLAB. The current system is like this:

$$\begin{aligned} x_{n+1} &= y_n + \tau \\ y_{n+1} &= -bx_n + a_n y_n - y_n^3 + \tau \\ a_{n+1} &= a_n + \Delta t \end{aligned} \tag{8}$$

With different values of the parameter b, we can get different time series of y_n with different patterns. Figure 4 shows a time series with a fixed value $b = 0.15$. Clearly, when $0.5 < a < 1$, the series seems very regular. The series' fluctuations become intensified when about $1 < a < 1.2$. But it soon goes back to the regular mode. When $a \approx 2.5$, it transits to be chaotic. (We must emphasis that the generation of noise is greatly stochastic but it can influence the time when the y_n gets into chaotic. What we focus is only on the time trend of y_n).

Then we let the parameter b take 4 different specific values ($b = 0.15$, $b = 0.16$, $b = 0.17$ and $b = 0.18$). Accordingly, we obtain four different time series of y_n. We intercepted a short segment in each of the four series (about from $a = 2.1$ to $a = 2.8$, the total length of the time series in 6000 points). All the series we obtain demonstrates an obvious trend of evolving from ordered to chaotic (see figure 5). For each a_{n-1}, we have a point $y_n = \left(y_n^1, y_n^2, y_n^3, y_n^4\right) \in \mathbb{R}^4$. We choose a sliding window of length $w = 50$, therefore a point cloud is constructed $Y_n = (y_n,\ y_{n+1}, \ldots, y_{n+w-1})$. Each point cloud has w points in $\mathbb{R}^4$. A filtration of simplicial complexes is constructed, and we compute the L^p-norms ($p = 1$ and $p = 2$) of the persistence landscapes of 1D persistent homology of each 4D-point cloud. The L^1-norms ($\|\lambda(Y_n)\|_1$) is marked blue and the L^2-norms ($\|\lambda(Y_n)\|_2$) is marked red in figure 6.

It's obvious that the L^p-norms ($p = 1$ and $p = 2$) of the persistence landscapes show great signs of increasing when the series of y_n goes into the chaotic mode (at about when $a = 2.6$). Therefore, we can draw the conclusion that our method of persistent homology is able to detect the increased volatility of the data set.

3 Rethinking of TDA: A Modified Version

3.1 Some Thoughts on Traditional Persistence Homology

Generally, for each t_i in $\sum = (t_1, \ldots, t_N)$, the sliding window w takes a fixed value. However, we observe that a major problem with this kind of application is the selection of sliding window greatly influences the L^p-norms ($p = 1$

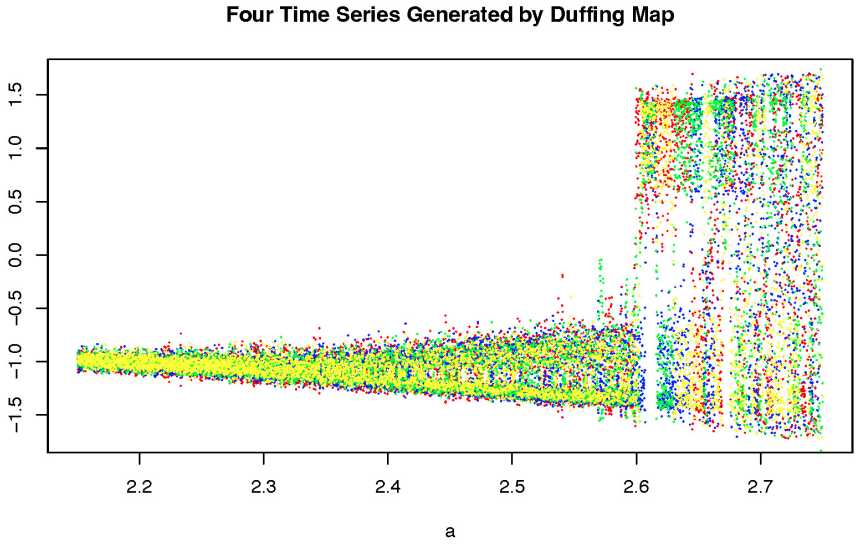

Figure 5: Four time series generated by Duffing Map ($b = 0.15$, $b = 0.16$, $b = 0.17$ and $b = 0.18$), the horizontal axis corresponds to the parameter a and the vertical axis corresponds to y. All the time series shows an obvious trend of evolving from ordered to chaotic.

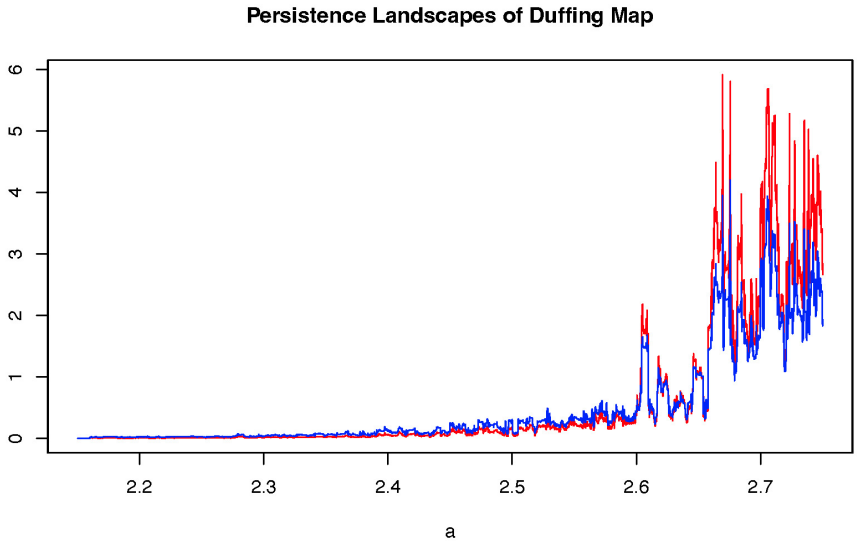

Figure 6: The L^p-norms ($p = 1$ and $p = 2$) of the persistence landscapes of 1D persistent homology of Duffing Map. The blue line corresponds to the time series of L^1-norms while the red line corresponds to the time series of L^2-norms. Both lines show a clear sign of a sharp rise during the transition of the series of y from ordered to chaotic.

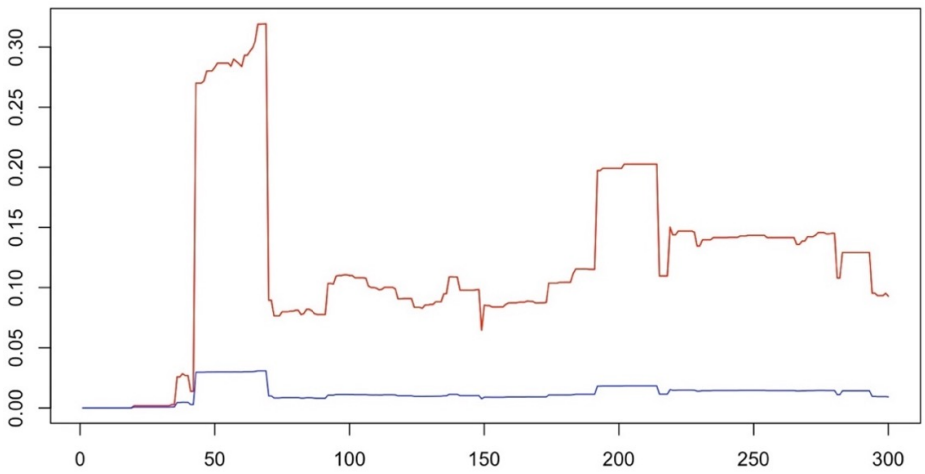

Figure 7: The L^p-norms ($p = 1$ and $p = 2$) of different sliding windows on September 17, 2009. The horizontal axis corresponds to different w, and the blue line and the red line correspond to the L^1 and L^2 norms.

and $p = 2$) we compute. We perform an experiment on the financial time series that we will use in Section 4. The experiment consists of three steps, which are described in: first, we randomly choose a fixed time point, e.g. at September 17, 2009, and then we apply a sliding window w of dynamic size from 0 to 300 and compute the norms of each size, at last, we graph them (see figure 7). The horizontal axis corresponds to different w and the blue and red lines correspond to the L^1 and L^2 norms for each size of the sliding window. As shown in figure 7, the norms undergo huge fluctuation according to the change of w. These results suggest that the approach of taking a fixed sliding window has a huge room for improvement.

3.2 Dynamic-sized Sliding Window

As we know, the L^1 and L^2 norms represent the topological features of a point cloud. However, in order to get accurate early warning signals (EWS), we must make sure that the point cloud with a specific sliding window contains the most significant topological features. Nevertheless, the topological features calculated with a fixed sliding window may fail to prove they are significant, since the structure formed by the points in this fixed window may not represent the complete topological structure. For tackling the above problem, in this paper we propose a new method called 'Dynamic-sized Sliding Window', which means that sliding window w has a dynamic size, in order to make the topological features more significant.

We still choose the day September 17, 2009, and graph the scatter diagram of 3D-point cloud with a sliding window of 300 days. We just take a 3D-point cloud with three stock market indices as an example because it looks more intuitive in the 3D-coordinate system, and since changes in stock market indices are always synchronized, it is more than enough to use three indices to represent the market changes. We mark each point in this sliding window with a serial number. The point closest to our specified date in the time series is marked as No.1, and the point farthest from our specified date in the time series is marked as No.300. The colors of these points represent their serial number, see figure 8. Through figure 8 we can clearly observe that the distribution of points becomes most discrete and chaotic when their serial numbers are near 50 and 200. Also, from figure 7 we see the maximum value of L^1 and L^2 norms when the size of the sliding window is around 50 and 200 . We believe this is not a coincidence. As TDA is able to detect the critical transition, when the points turn from the aggregate distribution to the chaotic distribution, the values of the norms are very likely to increase, which represent more significant topological features.

In order to be able to capture this change from aggregation to chaos in time, we first take a large window alternative range Γ, e.g. w from 30 to 100 . for each point τ in this range Γ, we apply another small fixed embedding dimension (sliding window) w'. We compute the Euclidean distance in w' for τ in Γ, which is defined as the mean of the sum of distances from all points in w' to the calculated average center point. Then we draw the curve of the Euclidean distance $\Omega(\ )$. The peaks μ of the curve are determined by the parameter σ, $\mu = \{\varphi_i \,|\Omega\,(\varphi_i) > \Omega(\phi), |\,\varphi_i - \phi\,| \leq \sigma\}$. We compute the L^1/L^2 norms for each sliding window of size φ_i and choose the biggest L^1/L^2 norms as the representative norms for the point τ, and the corresponding φ_i is determined to be the suitable sliding window size for the time point τ (The parameters w' and σ used to compute L^1 and L^2 norms may be different and we will explain how to determine them in practical applications in the section 4). In this way, we can get more significant topological features

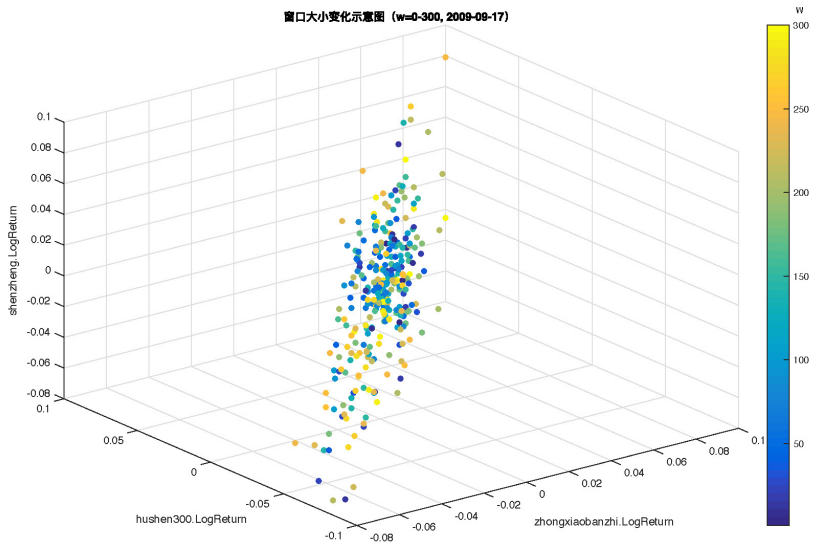

Figure 8: The scatter diagram of a 3D-point cloud with a sliding window of 300 days on September 17, 2009 to show the dispersion of these points. The three axes correspond to three stock market indices and the colors of the points correspond to the serial numbers of points in this window.

with fewer calculations. However, what we must emphasize is that we get the topological features that are as obvious as possible, but not necessarily the most significant one. Because TDA is very sensitive to the degree of chaos, but our method is to measure the degree of dispersion of points. The discrete distribution of points does not necessarily cause confusion. Our solution is to take the form of looking at the peaks of the Euclidean distance curve. The peaks represent a process in which the point shifts from a clustered distribution to a discrete distribution and we believe this process is likely to indicate that the distribution of points is gradually becoming confusing. At the same time, we took the corresponding larger norms of several peaks. In this way, we can get more significant topological features with fewer calculations.

4 Empirical Application: Analysis of China's A-share Market data

4.1 Data

We analyze the time series of four major China's A-share Market indices: SHANGHAI COMPOSITE INDEX, SZSE COMPONENT INDEX, CSI300, and SSE SME COMPOSITE between January 24, 2006 and January 18, 2019. Overall, they are relatively precise and objective representations of the A-share Market of China. Figure 9 demonstrates the trend of these four indices when two devastating stock market crashes took place from November 2007 to October 2008, and from June 2015 to January 2016. As is shown in table 2, we identify several significant peaks of each index under consideration, where the dates of peaks of four indices manifest conspicuous consistency.

Table 2: Dates of peaks of four indices.

Indices	Starting Date	1st Peak	2nd Peak	3rd Peak
SHANGHAI COMPOSITE INDEX	2006-01-24	2007-10-16	2008-01-14	2015-06-12
SZSE COMPONENT INDEX	2006-01-24	2007-10-16	2008-01-14	2015-06-12
CSI300	2006-01-24	2007-10-16	2008-01-14	2015-06-12
SSE SME COMPOSITE	2006-01-24	2008-01-15	2015-06-12	

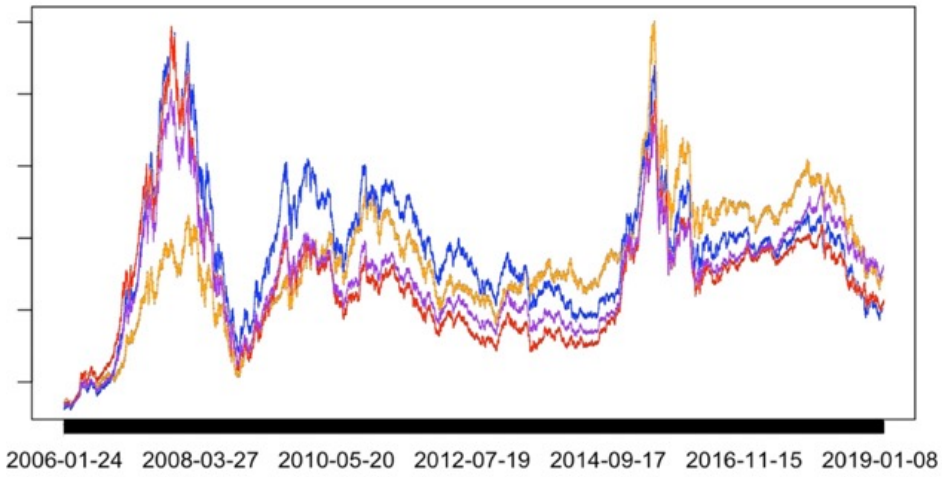

Figure 9: The price trend of these four indices.

4.2 Modified TDA on China's A-share Market

4.2.1 Data Preprocessing and Parameter Optimization

We choose closing prices of all the trading days between January 24, 2006 and January 18, 2019 of the four indices to create time series, and apply to them the log-returns $r_{ij} = \ln(P_{i,j}/P_{i-1,j})$, where $P_{i,j}$ represents the closing price of index j at day i. Therefore, we obtain four new time series of daily log-returns of the indices, and they form a 4×3158 matrix.

We utilize the Dynamic-sized Sliding Window mentioned in section 3.2 to acquire the best sliding window size for each trading day, as well as the prophetic L^1 and L^2 norms. In order to get the most suitable parameters w' and σ, we perform several experiments to test the average error rate and the corresponding computing complexity. We use System Sampling method to select a sample set from the original matrix, and then adopt different parameters w' and σ to obtain computing complexity defined as the number of times of filtration performed on a corresponding point cloud, as well as the average error rate defined as the average of the difference between our representative norm and the actual maximum norm divided by the later. The testing results are shown in the Appendix A and B. On the basis of those results, we finally choose $w' = 12$, $\sigma = 5$, which take into account both the accuracy of the results and the computing complexity.

4.2.2 TDA on Typical Dates

Figure 10 shows the Rips persistence diagrams and the corresponding persistence landscapes, calculated with a fixed sliding window of 50 trading days at 10/16/2007, 06/12/2015, and the dates six months ago. It can be clearly noticed that as the stock market becomes more volatile, loops in the relevant point clouds become much more prominent.

4.2.3 Modified TDA on Holistic Time Series

Figure 11 manifests the normalized time series of the L^1-norms (blue line) and L^2-norms (red line) of persistence landscapes calculated with the Dynamic-sized Sliding Window. Note that those time series of norms are generated under the condition where $\Gamma = 30$ to 100, $w' = 12$ and $\sigma = 5$. We can clearly see that the L^1 and L^2 norms exhibit a clear sign of great increases before the stock market crashes took places in 2008 and 2015. Nevertheless, we must emphasize that we cannot rely on the L^1 or L^2 norms alone to predict approaching major crashes [Gidea and Katz, 2018], since the L^1 and L^2 norms may exhibit great increment while there is no crash in the near future. L^1 and L^2 norms also help explain the difference between the crash in 2008 and the crash in 2015. The 2008 stock market crash in China's A-share market was mainly caused by the global financial crisis that time, while the 2015 crash was mainly the result of domestic economic factors. The co-movements between different countries [Mobarek et al., 2016] and different crises (stock market crisis, currency crisis and debt crisis) made the volatility more frequent and violent than that in 2015 stock market crash.

To sum up, these experimental results support the original hypothesis that TDA is sensitive to regime transition and further has the ability to forecast the stock market crashes.

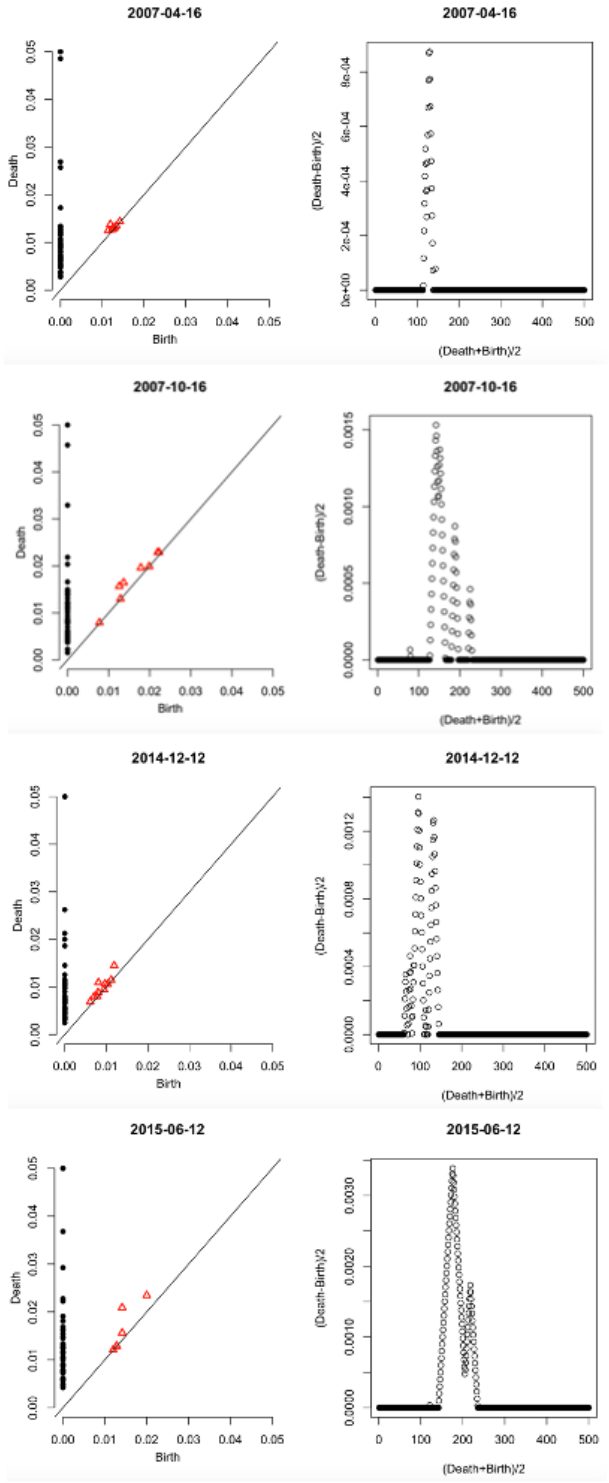

Figure 10: The Rips persistence diagrams and the corresponding persistence landscapes calculated with the sliding window of 50-days intervals ending at selected dates. The solid black dots represent connected components, red triangles represent loops. The top 2 rows: the date when stock indices reach their peaks before the financial crisis in 2008 and half a year ago. The bottom 2 rows: the date when stock indices reach their peaks before the financial crisis in 2015 and half a year ago.

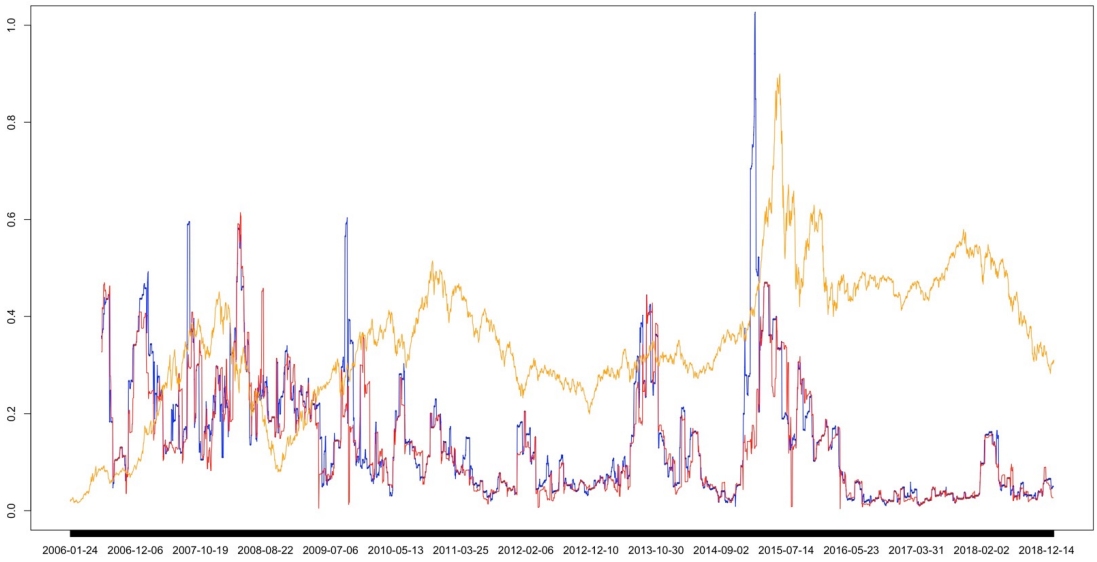

Figure 11: The time series of normalized L^1 (blue line) and L^2 (red line) norms of persistence landscapes calculated with a dynamic-sized sliding window.

5 What's Next...

- We tend to formalize the Prediction of Stock Market Crashes task from the perspective of computer science, including benchmark, evaluation criteria, and our state-of-art approach. Though great progress has been made in the applications of TDA on financial market crashes [Gidea and Katz, 2018], both their and our experiments lack quantitative evaluation and extensive comparisons with common methods.

- Although we tried to calculate the accuracy and recall rate of our method, China's stock market started late and has only had two stock market crashes since 2005. The lack of data forced us to think of other solutions. In one way, we use our modified TDA to predict sharp declines in stock returns, but not limited to crashes. We are also evaluating the feasibility of this approach.

- Like machine learning methods, TDA method lacks theoretical explanation to some extent. Why can our holes capture market signals? Why does these signals represent the possibility of a stock market crash? There are still many problems worthy of combining mathematical theory with financial theory.

- The four dimensions of the time series we currently use are the four indices of the Chinese stock market. However, the high temporal correlation exhibited by these four dimensions makes us doubt whether information and computing resources are wasted. We are considering using the indices of different sectors whose changes are not so synchronized as a proxy to examine the experimental results.

References

Didier Sornette. Why stock markets crash. In *Why Stock Markets Crash*. Princeton university press, 2009.

Mara Madaleno and Carlos Pinho. International stock market indices comovements: A new look. *International Journal of Finance & Economics*, 17(1):89–102, 2012.

Pek Y Lum, Gurjeet Singh, Alan Lehman, Tigran Ishkanov, Mikael Vejdemo-Johansson, Muthu Alagappan, John Carlsson, and Gunnar Carlsson. Extracting insights from the shape of complex data using topology. *Scientific reports*, 3(1):1–8, 2013.

Jessica L Nielson, Jesse Paquette, Aiwen W Liu, Cristian F Guandique, C Amy Tovar, Tomoo Inoue, Karen-Amanda Irvine, John C Gensel, Jennifer Kloke, Tanya C Petrossian, et al. Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury. *Nature communications*, 6(1):1–12, 2015.

Xiaojin Zhu. Persistent homology: An introduction and a new text representation for natural language processing. In *IJCAI*, pages 1953–1959, 2013.

Sabah Jassim and Aras Asaad. Automatic detection of image morphing by topology-based analysis. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1007–1011. IEEE, 2018.

Marian Gidea and Yuri Katz. Topological data analysis of financial time series: Landscapes of crashes. *Physica A: Statistical Mechanics and its Applications*, 491:820–834, 2018.

Marian Gidea, Daniel Goldsmith, Yuri Katz, Pablo Roldan, and Yonah Shmalo. Topological recognition of critical transitions in time series of cryptocurrencies. *Physica A: Statistical mechanics and its applications*, 548:123843, 2020.

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

Asma Mobarek, Gulnur Muradoglu, Sabur Mollah, and Ai Jun Hou. Determinants of time varying co-movements among international stock markets during crisis and non-crisis periods. *Journal of Financial Stability*, 24:1–11, 2016.

Appendix A Parameter Optimization for L^1 norms

w'	σ	Complexity	Deviation	w'	σ	Complexity	Deviation	w'	σ	Complexity	Deviation
5	2	2884	0.123984239	10	5	1115	0.219173603	15	8	NA	NA
5	3	2096	0.13811594	10	6	876	0.25559566	15	9	NA	NA
5	4	1580	0.162775739	10	7	787	0.265609087	15	10	NA	NA
5	5	1318	0.199388568	10	8	NA	NA	16	2	2826	0.110144446
5	6	1105	0.22058527	10	9	NA	NA	16	3	1895	0.157604323
5	7	972	0.254969737	10	10	NA	NA	16	4	1361	0.212027397
5	8	867	0.273185118	11	2	2839	0.091341937	16	5	1047	0.242225192
5	9	737	0.314331403	11	3	1994	0.157911908	16	6	836	0.284812829
5	10	NA	NA	11	4	1475	0.162200462	16	7	689	0.314842384
6	2	2807	0.111105618	11	5	1054	0.227786188	16	8	NA	NA
6	3	1944	0.149429121	11	6	862	0.252648401	16	9	NA	NA
6	4	1509	0.176836531	11	7	NA	NA	16	10	NA	NA
6	5	1251	0.210485425	11	8	640	0.314356242	17	2	2814	0.112066897
6	6	1117	0.233251426	11	9	NA	NA	17	3	1779	0.159759458
6	7	963	0.251013925	11	10	NA	NA	17	4	1314	0.193141626
6	8	810	0.29194106	12	2	2844	0.099219791	17	5	1029	0.252091889
6	9	697	0.32494953	12	3	1942	0.139611807	17	6	855	0.286071852
6	10	NA	NA	12	4	1357	0.18983173	17	7	NA	NA
7	2	2845	0.104136551	12	5	1068	0.233080996	17	8	NA	NA
7	3	1961	0.143732209	12	6	843	0.274175163	17	9	NA	NA
7	4	1460	0.17814222	12	7	716	0.303621891	17	10	NA	NA
7	5	1236	0.205256829	12	8	NA	NA	18	2	2713	0.106477566
7	6	1006	0.233704827	12	9	NA	NA	18	3	1792	0.153861919
7	7	918	0.250423341	12	10	NA	NA	18	4	1303	0.208754499
7	8	NA	NA	13	2	2814	0.108707784	18	5	1033	0.245235919
7	9	NA	NA	13	3	1807	0.16191627	18	6	835	0.27762138
7	10	NA	NA	13	4	1350	0.202681151	18	7	NA	NA
8	2	2856	0.10824136	13	5	1017	0.234792029	18	8	NA	NA
8	3	1931	0.137222655	13	6	855	0.269514102	18	9	NA	NA
8	4	1467	0.165935918	13	7	NA	NA	18	10	NA	NA
8	5	1144	0.211979929	13	8	NA	NA	19	2	2776	0.102501907
8	6	972	0.237587003	13	9	NA	NA	19	3	1761	0.164719638
8	7	835	0.281241772	13	10	NA	NA	19	4	1313	0.203716047
8	8	705	0.300087885	14	2	2828	0.114883565	19	5	971	0.253410304
8	9	NA	NA	14	3	1853	0.145875888	19	6	805	0.291696769
8	10	NA	NA	14	4	1371	0.195122357	19	7	NA	NA
9	2	2832	0.100329085	14	5	1039	0.250687664	19	8	NA	NA
9	3	1914	0.141405865	14	6	889	0.278596639	19	9	NA	NA
9	4	1473	0.177637423	14	7	NA	NA	19	10	NA	NA
9	5	1138	0.222067386	14	8	NA	NA	20	2	2557	0.111695071
9	6	917	0.251249858	14	9	NA	NA	20	3	1745	0.139685539
9	7	NA	NA	14	10	NA	NA	20	4	1284	0.185780806
9	8	NA	NA	15	2	2879	0.103524595	20	5	1017	0.222129097
9	9	NA	NA	15	3	1809	0.153392695	20	6	835	0.261368946
9	10	NA	NA	15	4	1260	0.228003723	20	7	NA	NA
10	2	2784	0.10704327	15	5	1040	0.246046766	20	8	NA	NA
10	3	1936	0.14457965	15	6	813	0.30744049	20	9	NA	NA
10	4	1425	0.192966053	15	7	NA	NA	20	10	NA	NA

Appendix B Parameter Optimization for L^2 norms

w'	σ	Complexity	Deviation	w'	σ	Complexity	Deviation	w'	σ	Complexity	Deviation
5	2	2887	0.086697014	10	5	1122	0.173466485	15	8	NA	NA
5	3	2114	0.095497466	10	6	873	0.194439895	15	9	NA	NA
5	4	1578	0.134338819	10	7	NA	NA	15	10	NA	NA
5	5	1333	0.142781189	10	8	NA	NA	16	2	2823	0.088613562
5	6	1096	0.18405138	10	9	NA	NA	16	3	1909	0.112273123
5	7	979	0.194648435	10	10	NA	NA	16	4	1360	0.165856363
5	8	NA	NA	11	2	2835	0.072690832	16	5	1037	0.194996628
5	9	NA	NA	11	3	1994	0.114667308	16	6	834	0.232514242
5	10	NA	NA	11	4	1466	0.132829284	16	7	NA	NA
6	2	2814	0.085402924	11	5	1063	0.178209783	16	8	NA	NA
6	3	1959	0.11754947	11	6	878	0.203330527	16	9	NA	NA
6	4	1515	0.136760406	11	7	NA	NA	16	10	NA	NA
6	5	1264	0.16462255	11	8	640	0.264337551	17	2	2801	0.077856129
6	6	1116	0.178776619	11	9	NA	NA	17	3	1775	0.125056384
6	7	961	0.207042388	11	10	NA	NA	17	4	1302	0.149839142
6	8	816	0.229693594	12	2	2860	0.069568135	17	5	1012	0.203847603
6	9	NA	NA	12	3	1932	0.109765185	17	6	857	0.220103461
6	10	NA	NA	12	4	1364	0.138070097	17	7	NA	NA
7	2	2849	0.083255621	12	5	1072	0.183855653	17	8	NA	NA
7	3	1958	0.117403181	12	6	853	0.214057464	17	9	NA	NA
7	4	1462	0.144299267	12	7	724	0.233824511	17	10	NA	NA
7	5	1238	0.156088374	12	8	624	0.258128599	18	2	2714	0.083345628
7	6	984	0.184382943	12	9	NA	NA	18	3	1789	0.115090303
7	7	916	0.200579112	12	10	NA	NA	18	4	1311	0.155597387
7	8	NA	NA	13	2	2830	0.07255144	18	5	1037	0.183267578
7	9	NA	NA	13	3	1815	0.120817642	18	6	844	0.213895337
7	10	NA	NA	13	4	1358	0.156288838	18	7	NA	NA
8	2	2847	0.079302342	13	5	1006	0.18715842	18	8	NA	NA
8	3	1926	0.106781273	13	6	NA	NA	18	9	NA	NA
8	4	1461	0.132454115	13	7	721	0.220145147	18	10	NA	NA
8	5	1146	0.16879686	13	8	NA	NA	19	2	2788	0.074607239
8	6	980	0.194664445	13	9	NA	NA	19	3	1799	0.110549899
8	7	816	0.235491997	13	10	NA	NA	19	4	1299	0.155828528
8	8	705	0.234522193	14	2	2832	0.083387602	19	5	978	0.195365019
8	9	NA	NA	14	3	1836	0.113851752	19	6	807	0.220153762
8	10	NA	NA	14	4	1359	0.149385889	19	7	NA	NA
9	2	2821	0.07925396	14	5	1062	0.186096173	19	8	NA	NA
9	3	1926	0.111570118	14	6	NA	NA	19	9	NA	NA
9	4	1487	0.134699211	14	7	NA	NA	19	10	NA	NA
9	5	1147	0.169374998	14	8	NA	NA	20	2	2546	0.073683906
9	6	903	0.1998112	14	9	NA	NA	20	3	1742	0.109160914
9	7	NA	NA	14	10	NA	NA	20	4	1275	0.147534508
9	8	NA	NA	15	2	2878	0.072164922	20	5	1016	0.191310899
9	9	NA	NA	15	3	1814	0.11342969	20	6	833	0.208768721
9	10	NA	NA	15	4	1254	0.173687746	20	7	NA	NA
10	2	2784	0.084611711	15	5	1035	0.199501111	20	8	NA	NA
10	3	1939	0.111473747	15	6	824	0.231648104	20	9	NA	NA
10	4	1424	0.138796144	15	7	NA	NA	20	10	NA	NA