DEGREE PROJECT IN MATHEMATICS,
SECOND CYCLE, 30 CREDITS
STOCKHOLM, SWEDEN 2017

An exploration of topological properties of high-frequency one-dimensional financial time series data using TDA

PATRICK TRUONG

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ENGINEERING SCIENCES

An exploration of topological properties of high-frequency one-dimensional financial time series data using TDA

PATRICK TRUONG

Degree Projects in Financial Mathematics (30 ECTS credits)
Degree Programme in Industrial Engineering and Management
KTH Royal Institute of Technology year 2017
Supervisor at KTH: Danica Kragic, Florian Pokorny, Jimmy Olsson
Examiner at KTH: Jimmy Olsson

TRITA-MAT-E 2017:80
ISRN-KTH/MAT/E--17/80--SE

Royal Institute of Technology
School of Engineering Sciences
KTH SCI
SE-100 44 Stockholm, Sweden
URL: www.kth.se/sci

Abstract

Topological data analysis has been shown to provide novel insight in many natural sciences. To our knowledge, the area is however relatively unstudied on financial data. This thesis explores the use of topological data analysis on one dimensional financial time series. Takens embedding theorem is used to transform a one dimensional time series to an m-dimensional point cloud, where m is the embedding dimension. The point cloud of the time series represents the states of the dynamical system of the one dimensional time series. To see how the topology of the states differs in different partitions of the time series, sliding window technique is used. The point cloud of the partitions is then reduced to three dimensions by PCA to allow for computationally feasible persistent homology calculation. Synthetic examples are shown to illustrate the process. Lastly, persistence landscapes are used to allow for statistical analysis of the topological features. The topological properties of financial data are compared with quantum noise data to see if the properties differ from noise. Complexity calculations are performed on both datasets to further investigate the differences between high-frequency FX data and noise. The results suggest that high-frequency FX data differs from the quantum noise data and that there might be some property other than mutual information of financial data which topological data analysis uncovers.

Sammanfattning

Topologisk dataanalys har visat sig kunna ge ny insikt i många naturvetenskapliga discipliner. Till vår kännedom är tillämpningar av metoden på finansiell data relativt ostuderad. Uppsatsen utforskar topologisk dataanalys på en endimensionell finanstidsserie. Takens inbäddningsteorem används för att transformera en endimensionell tidsserie till ett m-dimensionellt punktmoln, där m är inbäddningsdimensionen. Tidsseriens punktmoln representerar tillstånd hos det dynamiska systemet som associeras med den endimensionella tidsserien. För att undersöka hur topologiska tillstånd varierar inom tidsserien används fönsterbaserad teknik för att segmentera den endimensionella tidsserien. Segmentens punktmoln reduceras till 3D med PCA för att göra ihållande homologi beräkningsmässigt möjligt. Syntetiska exempel används för att illustrera processen. En jämförelse mellan topologiska egenskaper hos finansiell tidseries och kvantbrus utförs för att se skillnader mellan dessa. Även komplexitetsberäkningar utförs på dessa dataset för att vidare utforska skillnaderna mellan kvantbrus och högfrekventa FX-data. Resultatet visar på att högfrekvent FX-data skiljer sig från kvantbrus och att det finns egenskaper förutom gemensam information hos finansiella tidsserier som topologisk dataanalys visar på.

Acknowledgements

I would like to thank my mentors and supervisors Florian Pokorny and Danica Kragic for their patience, guidance and time. Their encouragement, knowledge, and support have been of utmost importance in bringing this thesis together. In particular, they have provided me with many insightful discussion, as well as contacts to discuss with, in which many new ideas came to life. Special credit needs to be given to Florian Pokorny for all the extra effort and late office hours he has put into my supervision. Besides this, Florian Pokorny kept a great sense of humor throughout and was also great at balancing informal and formal conversation, which kept the thesis supervision very pleasant.

I further wish to thank my formal supervisor Jimmy Olsson for providing formal supervision and setting up the thesis; Marcello Paris at UniCredit for contributing in many insightful discussion about TDA and financial markets, as well as providing data for this thesis; Fredrik Giertz at AP3 for conversations about quantitative topics in financial markets; Mikael Vejdemo-Johansson at the city university of New York for discussions about TDA and how it has been used in other fields; Wojtek Chacholski at KTH for discussions about how current research in TDA; Paul Rosen at University of South Florida for conversations about persistent homology on one-dimensional time series inference and Danijela Damjanovic at KTH for discussions about dynamical systems. All these people have contributed to my understanding of the field of TDA, as well as how well it connects to time series analysis and financial markets.

Contents

Chapter 1

Introduction

1.1 Background

Topological data analysis (TDA) is an emerging field in which topological properties of data are analyzed. These topological properties have been shown to be able to provide novel insights in data, which traditional statistics cannot. Traditional techniques of data analysis have not always been able to keep up with the increasing quantity and complexity of data since they may at times apply to many simplistic assumptions [1]. TDA is an attempt to address this problem by the idea that data have shape which could have meaning. The field has century-old mathematical foundation stemming from topological and computation geometry. Early contributions to the field of TDA were made by Edelbrunner et al. [2]. Zomorodian and Carlsson used the foundation laid by Edelbrunner et al. to develop the early TDA technique: Persistent homology [3]. The area was then made popular by an overview paper by Carlsson in 2009 [4].

TDA analyzes point clouds in metric spaces (often Euclidean spaces). It has been successfully applied to give new insight to complex problems related to neuroscience, biology, medicine and social sciences amongst others [5–19]. Combining topological methods with statistical methods have been proven to be a valuable method for understanding and visualizing data. TDA has been made considerably more accessible to the general data scientists public recent years by open source software and library packages as `Dionysus`, `GUDHI` [20], `PHAT`

[21, 22] as well as R TDA interface bindings to these efficient C++ libraries provided by Fasy et al. [23].

Analyzing the quantitative properties of financial data has long been studied by both financial professionals as well as the academical community. Researchers have applied all kind of different mathematical modeling, machine learning, artificial intelligence and data analysis methods to a myriad of different areas in finance [24–77]. Furthermore, much of the current academic interest in mathematical finance still lies in quantitative approach in analyzing financial data [78]. Traditional techniques for data analysis of financial data are therefore a well-studied area. Meanwhile, the emerging subfield of TDA provides an exceptional opportunity for a fresh approach to financial data mining. While the existing studies concerning topological aspects of financial data. The area of TDA in finance has to our knowledge received limited attention by the academic community. Studies focusing on topological aspects of financial data, but does not directly use TDA, use other methods which could contain information in the topology, such as network reconstruction or geometry-based methods. For example, Vandewalle et al. studied the topology exhibited by minimum spanning trees to detect correlation structures between stocks [79] and Phoa used diffusion maps to study the geometry of stock comovements [80]. To our knowledge only Gidea and Gidea et al. has provided studies in this area to this date. Gidea used persistent homology to detect early signs of critical transition in financial data [81] and Gidea et al. studied return point clouds between indices using persistent homology [82]. Gidea et al. claim that certain persistence patterns in the homology groups give an early indication of a financial crisis. Although, the area of TDA applied to financial markets has received limited attention, relevant areas such as TDA for time series and signals have been previously studied. Kasawneh et al. have proposed the use of Takens' embedding to reconstruct a time series into a point cloud [13, 83–85]. They used Takens' embedding in combination with maximum persistence to measure the stability of stochastic delay systems. Lastly, Perea and Harer suggested that maximum persistence in combination with a sliding window technique could be used to quantify periodicity of a signal [12]. These studies will be further explained in the literature review and previous studies chapter 2.

Financial markets are information-driven and a highly competitive en-

vironment where any additional information could be of value. In addition, alpha return opportunities are only prevalent to those seeking unique and unexploited strategies and methods. TDA is to our knowledge relatively unstudied as a tool for financial analysis and has shown to be able to uncover useful information in other areas of science [5–19]. Therefore, an investigation of how TDA could be used for extracting knowledge from financial data is highly relevant. Takens' embedding has been shown to be able to convert time series data to meaningful point clouds for persistent homology computation. In addition, the use of sliding window technique allows for segmentation of a long time-series into chunks, which makes the topological features more comparable within and between datasets. Also, as both methods have proven to be useful in conjunction with TDA in other areas we believe that they are good starting points to investigate.

1.2 Problem

Noise in data has been shown to pose a challenge for the research community [86, 87]. Many of the scientific communities contributions to quantitative forecasting models have very little practical utility because often the improvements made to models would have been dwarfed by the variance in real data [86].This indicates the need for a method that shows other aspects of data.

Financial data have complicated variance and dependencies. However, it is not completely random [88–91]. Researchers have found that traditional financial analytics which utilizes low-level price data as an analytical basis are not reliable due to the complex character of the data. However, using higher level representation models of the data can reduce the noise in the data and thus make it more appropriate for traditional financial analytics [92]. By using such representation two things are done; 1) certain characteristics of the higher level representation are predetermined, and 2) certain aspects of the information contained in noise is disregarded. As TDA has shown potential to uncover novel insight about data in other areas of natural science [8], it is relevant to investigate whether or not it is possible to use TDA to extract information from financial data.

1.3 Preliminary Aim

This thesis aims to use topological data analysis to investigate if there exist distinguishable topological features in different segments of a financial time series.

1.4 Preliminary Research Question

This thesis aims to investigate the following questions:

- Is it possible to use topological data analysis to infer knowledge about one-dimensional financial time series?
- What kind of insight does topological data analysis provide?

1.5 Limitations

This thesis is intended to investigate the use of topological data analysis for analyzing one-dimensional financial data. It is solely done for academic purposes and not intended to be viewed as any financial or investment advice. Further, the thesis is limited by the availability of open source topological data analysis packages and libraries.

1.6 Contributions to Science

To our knowledge, the only published works on analyzing financial data with TDA are Gidea and Gidea et al. [81, 82]. This thesis directly addresses the lack of research conducted in using topological data analysis for one-dimensional financial data. As such the thesis can be viewed as an attempt to apply theoretical knowledge about topological data analysis to a real-world problem in the financial markets and thus generalize the use area of the method. It also shows how to approach the problem of analyzing the topological properties of one-dimensional time series using TDA.

Chapter 2

Literature Review and Previous Studies

2.1 Topology and Financial Markets

2.1.1 Topology to analyze groups of assets

There exists previous work studying the topology of financial markets without using TDA methods. These studies often analyze relationships of groups of stocks or assets. For example analysis of the topology of minimal spanning trees constructed using stock correlations [79]. This section outlines studies conducted in this manner and is presented to give the reader a brief overview of non-TDA related methods where topological analysis can be used. However, this thesis has a significantly different approach than these studies, as it is using TDA to analyze financial data. In addition, this thesis focuses on analyzing the topological features of one dimensional financial time series as opposed to multidimensional objects.

Vandewalle et al. researched the topology of stock markets as early as 2000s' [79]. They analyzed the cross-correlation of daily fluctuations for all US stocks during the year of 1999 by using a minimum spanning tree and looking at the topology exhibited by the minimal spanning tree. The main features observed by was the nodes, links and dangling endpoints. It was emphasized that these features had differ-

ent qualitative meanings and they seemed stable over time.

Phoa studied the geometry of co-movement in a set of stocks [80]. More specifically Phoa analyzed monthly total return for January 2002 to April 2012 for index constituents of the S&P 100 and S&P 500. Although, this study did not directly use topological data analysis on financial data, it did highlight the fact that geometry can be used efficiently at looking at the correlation structure of the stock market. Phoa used diffusion maps to project high-dimensional stock correlation matrix ($100x100$ and $500x500$ matrix) to a 3D hyperplane. The closer two assets were in the hyperplane, the higher their correlation. In other words, the diffusion map contains information in the distances. Phoa further motivated that diffusion map was a suitable method for stock data because it was robust to noise, i.e. small perturbations in the data did not have a large effect on the results, unlike some other dimensionality reduction methods. The property of robustness was very helpful when dealing with real financial data, which often were noisy. However, Phoa highlighted that a disadvantage of this methods was that the coordinates did not have an intuitive economic meaning. Another aspect that Phoa highlighted was that projection to 2D or 3D hyperplane allowed for good and intuitive visualization. However, the eigenvalues indicated that there was relevant geometric information in the fourth and fifth coordinate. In addition, Phoa noted that while the diffusion map contained information in distances, they were quite hard to read and thus it would have been beneficial with additional quantitative information that measured the assets' global tendency to move together - i.e. the size or compactness of the cloud as a whole - as well as the ability to identify the most significant local concentration within the cloud. In the study Phoa suggested using a quantitative summary called *global concentration* measure, which was defined as $(\mathrm{tr}\ \Sigma)^{-\frac{1}{2}}$, to measure the concentrations. However, the *global concentration* measure did not capture information about how the overall concentration changed, which in this case had to be visually read from the diffusion maps. The benefit of the geometric approach was twofold 1) that it could compare portfolio concentration against a benchmark and 2) that it could identify local concentrations that were of interest. Such local concentration could be relevant in the case of idiosyncratic shocks, which affect only localized regions in the abstract asset space.

2.2 Topological data analysis on financial data

This section outlines studies using TDA methods on financial data. The studies in this section use a similar methodology as used in this thesis.

Gidea has recently researched the use of TDA of critical transitions in financial networks [81]. In this study TDA was used as a method to detect early signs for *critical transition* in financial data. By *critical transition* the author referred to an abrupt change in the behaviour of a complex system, which arose due to small perturbations in the external conditions. This effects of this *critical transistion* caused the system to switch from one steady state to some other steady state. The author stated that examples of critical transitions were market crashes, abrupt shifts in ocean circulation and climate, regime changes in ecosystems, asthma attacks and epileptic seizures etc. As such, this study was an attempt at using TDA for *change point detection* in time-series data. Gidea used price time-series of multiple stocks to build time-dependent correlation networks, which exhibit topological structures. *Persistent homology* was then used to analyze these structures in order to track changes in topology when approaching a critical transition. The information of the topological structure was encoded in *persistence diagrams*, which provide a robust summary of the topological information on the network.

As a case study, Gidea used a portfolio of stocks consisting of the DJIA stocks listed as of February 19, 2008. The data was restricted to the time period between January 2004 to September 2008 (when Lehman Brothers filed for bankruptcy). The focus of the case study was the *critical transition* during a period prior to the financial crisis of 2007-2008. A weighted network was constructed using correlation distances. For correlation distance calculations Gidea used arithmetic return as opposed to standard log return. The use of arithmetic return was motivated by [93]. Gidea used *persistent homology* in these correlation networks to quantify changes when approaching a *critical transition*. Gidea chose not to consider higher-dimensional homology groups because the correlation network was small and therefore the presence of higher dimensional structures would likely be accidental. The findings of this study were that there were significant topological

changes of the correlation network in the period prior to the onset of the 2007-2008 financial crisis. The changes could be characterized by an increase in the cross-correlation between various stocks, as well as by the emergence of sub-networks of cross-correlated stocks. Lastly, the authors stated that the findings were coherent with other studies [93–96]. The studies by Nobi et al. [94, 95] focused on the analysis correlation network topology during crises without the use of TDA. The studies used correlation network constructed using the standard log return $r_i(t) = \frac{\ln(r(t))-\ln(r(t-1)}{\sigma}$ as opposed to Gideas arithmetic return. The study by Scheffer et al. [96] focuses on early-warning signals.

Another recent research is a study on using TDA on financial time series during financial crash periods by Gidea and Katz [82]. This study focuses on the technology crash of 2000 as well as the great financial crisis 2007-2009. The method was similar to the previous study i.e. It used persistent Homology to detect and quantify topological patterns in multidimensional time series, limiting to 1-dimensional homology. The authors used sliding window technique and extracted time-dependent point cloud datasets to associate a topological space. The topological features was encoded in *persistence landscapes* and the temporal changes in the *persistence landscapes* was quantified via L^p-norms. The findings was that in the vicinity of financial crashes the L^p-norm exhibit strong growth prior to primary peak, which ascended during a crash. More specifically, the L^p-norm of the *persistence landscapes* exhibited a strong rising trend 250 trading days prior to both the dotcom-boom 03-10-2000 and the Lehman-bankruptcy 09-15-2008. This study proved that TDA provides a new type of econometric analysis, which could complement other statistical measures. In this study four major US stock indices; S&P 500, DJIA, NASDAQ, and Russel 2000 between 23-12-1998 and 08-12-2016 was analyzed, using daily log return as data points. The point cloud to be analyzed thus became a wxd-matrix where $d = 4$ and w was the size of a sliding window. Each dimension was analyzed individually to form a 4-dimensional point cloud.

The first study by Gidea [81] focused on groups of stocks. The difference from the studies in section 2.1.1 is that Gidea used persistent homology to identify topological features as opposed to visual inspection. The study viewed the financial market as a complex system with

different states similar to this thesis. This showed that TDA for low dimensional topological analysis could potentially be used to obtain useful information about dynamical systems. The first study used network reconstructions of the time series. The second study by Gidea and Katz [82] worked with time series similar to this thesis. It also studied low-dimensional topological features with persistent homology, similar to this thesis. One interesting aspect of this study was the construction of 2D point cloud by plotting return data of two different indices against each other. The fact that this study only investigated low dimensional topological features means that it was essentially looking at return spreads across assets. Holes in these point clouds typically represents that the assets do not move similarly, and thus the finding of this study essentially is a strong divergence in correlation 250 trading days prior to the financial booms.

2.3 Topological Data Analysis for time series and signals

Time series do not have immediately obvious point cloud representation. Therefore, using topology to analyze it is not straightforward. Previous studies on applying topological methods for analyzing time series data will be presented in this section.

2.3.1 Takens embedding and persistence for Time-delay systems

Fourier and power spectrum analysis have been used when time series and signals are periodic. When the time series are non-periodic however the methods often yield faulty results [97]. Also, these methods do not manage to appropriately account for systems evolution through time [98].

Kasawneh et al. used a combination of Takens' embedding and TDA (maximal persistence and persistent homology) to analyze stochastic delay equations [13, 83–85]. More specifically, in [85] they used maximal persistence to analyze Hayes equation and stochastic version

Mathieu's equation i.e. equations wherein states evolve through time. Point clouds of these equations were obtained via Takens' embedding. These point clouds were then analyzed with TDA. Their results indicated that using Takens' embedding in combination with TDA was a valid tool for analyzing the stability of stochastic delay equations. More specifically, it has been shown to be able to analyze the stability of stochastic delay systems. In [83] datasets were simulated from Euler-Maryuama method and the dataset was converted to a point cloud via Takens embedding. The points cloud was then used to study the equilibrium and periodic solutions using persistent homology. The study was very similar to the previously mentioned study. However, using persistent homology instead of maximal persistent did not allow for multidimensional analysis. The other studies conducted by Khasawneh et al. are also similar [13, 84].

These studies show TDA can be used for analyzing dynamical systems associated with time series by using Takens embedding. Both studies are conducted on simulated data. The time series are processed in a similar manner in this thesis. However, it is conducted on real data as opposed to simulated data.

2.3.2 Sliding windows of time series for persistent homology

When analyzing time series it is often relevant whether or not analysis is conducted on segments or the whole time series. Looking at segments is interesting for financial data because it is often taught that financial markets move in regimes. A clear example of regime change in financial data is when important financial news impacts assets [99]. This section outlines studies that have used TDA on sliding windows technique on time series data to draw conclusions about both the segments and the whole time series by looking at continuous segments of it.

Perea and Harer developed a method for topological study of time series data using sliding window and time-delay embedding [12]. Time-delay embedding was used to transform windowed time series into point clouds. They suggested that maximum persistence of these point-clouds could be used to quantify periodicity at the signal. In other

words, they used maximum persistence to measure "roundness" of the point cloud. In the paper, they further pointed out that periodicity, in this case, was defined as repetitions of patterns and quantified the recurrence as the degree of circularity or roundness of the point-cloud.

Berwald et al. claimed that detailed descriptions of complex high-dimensional and chaotic systems were difficult or impossible to obtain in many cases. They suggested that a more reasonable approach to analyzing this kind of system was to recognize and mark transitions of a system between qualitatively different regimes of behavior [11]. In this paper, they developed a framework with a high degree of success in picking out a cyclically orbiting regime from a stationary equilibrium regime in high-dimensional stochastic dynamical systems. This was done by combining persistent homology with machine learning techniques. To obtain the dynamical system description from observational time series Berwald et al. used the same sliding window method as Perea and Harer. The point of interest in this paper was to detect if the system underwent a bifurcation process with the use of persistent homology. Lastly, classification algorithms were implemented to check whether or not the system actually underwent bifurcation from the persistence barcode constructed.

The first study of this section showed the possibility to find recurrences of a time series using time-delay embedding on sliding windows in combination with persistent homology, showcasing the possibility to find structure in time series. However, Berwald et al. [11] pointed out the difficulty in finding a good structure for complex systems. In the best case this thesis could hope to find clear structures as described in Perea and Harer [12], but due to the complexity of financial time series, this thesis instead investigates if it is possible to use TDA to infer some knowledge about the property of financial time series. The study by Berwald et al. [11] also showed the possibility of combining machine learning and quantitative models with TDA. This fact does not directly relate to the work in this thesis. However, it is interesting to point out to highlight the added value of TDA.

Chapter 3

Theory Section

3.1 Topological Data Analysis for time series analysis

Topological data analysis (TDA) uses topology to find structure in data. The methods include mapper and persistent homology [100, 101]. They are often used to extract information from noisy and complex datasets and for comprehension of high dimensional data without loss of information.

Many methods of dimensionality reductions also allow for comprehension of high dimensional data. These methods often reduce the dimension by feature extraction, meaning that information not incorporated in the extracted features is lost in the process. TDA, on the other hand, uses the topological abstractions to get a complete view of the qualitative aspect of the data.

3.1.1 Homology

The geometry presented by data in a metric space is not always relevant, sometimes more basic properties such as the number of components, holes or voids are of interest. Algebraic topology captures these properties by counting them or associating vector spaces or algebraic structures to them. Homology of field coefficients associates a vec-

tor space $H_i(X)$ to space X for each natural number $i \in \{0, 1, 2, \dots\}$ such that $\dim(H_0(X))$ is the number of connected components in X, $\dim(H_1(X))$ is the number of holes in X, $\dim(H_2(X))$ is the number of voids in X and $\dim(H_k(X))$ is the k-th homology group in X. The k-th homology group describes the k-dimensional holes in X.

3.1.2 Persistent Homology

Persistent Homology is a method commonly associated with TDA. It studies the qualitative aspects of data by computing its topological features. It is robust to perturbations, independent of embedding dimensions and coordinates and can thus provide a compact representation of qualitative features of data [101]. As it based on homology it uses algebraic topology, which has a well established theoretical foundation for studying qualitative aspects of data with complex structure. As input a point cloud on a metric space is used, such as $X = \{x_1, \dots, x_n\}$ in an Euclidean Space $\mathbb{R}^d$. To associate a topological space, simplicial complexes for filtration values $\varepsilon \in \mathbb{R}$ (which for alpha complexes are distances $\varepsilon > 0$) are constructed.

3.1.3 Simplicial Complexes

A simplex is a n-dimensional counterpart to a triangle or tetrahedron. The n-simplex is the n-dimensional polytope created by the convex hull of its $n + 1$ vertices. Let σ be an *n-simplex*. The *vertex* of σ is each of the $n + 1$ points used to define σ and the face of σ is the convex hull of any subset of the vertices of σ. The definition of *simplicial complex* is:

Definition 3.1.1. *A simplicial complex is a topological space realized as a union of any collection of simplices Σ which has the following two properties:*

- *Any face of a simplex Σ is also in Σ.*
- *The intersection of any two simplices of Σ is also a simplex.*

A Voronoi decomposition can be used to define a simplicial complex. Let S be a finite set of points in $\mathbb{R}^d$, $\varepsilon > 0$ and let S_ε denote the union of balls $\bigcup_{s \in S} B(s, \varepsilon)$, where B are balls. Given the Voronoi diagram of $s \in S$, the Delaunay triangulation is obtained by connecting points at

the intersection of the balls and Voronoi regions around s: $V_s \bigcap B(S, \varepsilon)$. Two points are connected using edges and three points are connected using triangles etc. The resulting complex created is called the *alpha complex* of S at scale ε, and is denoted $A(S_\varepsilon)$

After computing the simplicial complexes the features are prevalent in the space S_ϵ composed of vertices, edges, and other higher dimensional polytopes. Using homology it is then possible to measure features such as components, holes, voids and other higher dimensional equivalent features. The persistence of these features are presented in *persistence Diagrams* or *persistence barcodes*. However, the interpretation of results is not straight-forward from a statistical point of view. The space in which the persistence diagrams and barcodes resides in lacks the geometric properties that would otherwise make it easy to define basic concepts such as mean, median etc. [101].

A more detailed explanation of the methods is given by [102]. The figures below show the construction of an alpha complex.

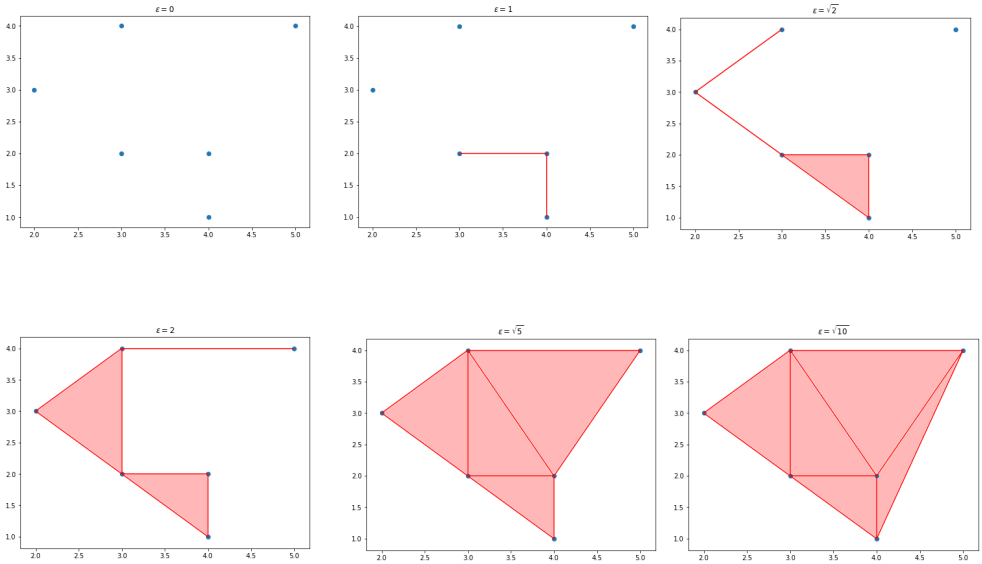


Figure 3.1: Construction of an alpha complex for random data points.

3.1.4 Persistence Diagram

Persistent homology captures how long topological features persists. The ranks of the persistent homology groups are presented in persis-

tence diagrams. It is a multiset of points in $\mathbb{R}^2$ and is defined as [101]:

Definition 3.1.2. *A persistence diagram is a multiset that is the union of a finite multiset of points in $\mathbb{R}^2$ with the multiset of points on the diagonal $\Delta = \{(x, y) \in \mathbb{R}^2 | x = y\}$, where each point on the diagonal has infinite multiplicity.*

A finite persistence diagram is a set of real intervals $\{(b_i, d_i)\}_{i \in I}$, where I is a finite set and b_i is the birth of the i-th feature and d_i is the death of the i-th feature. An example of a birth-death diagram is shown in fig. 3.2

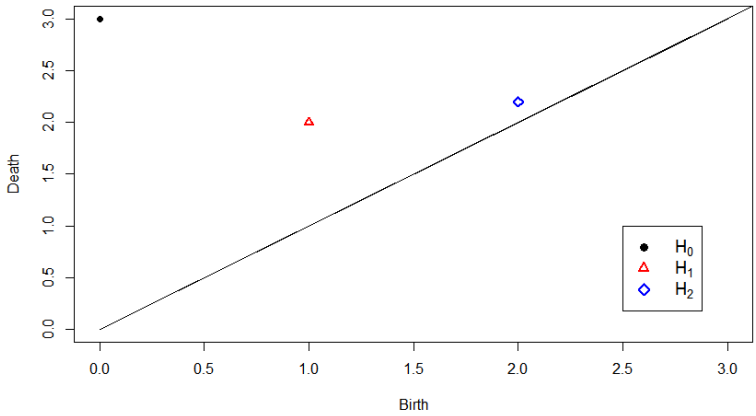


Figure 3.2: Illustration of a birth-death diagram.

3.1.5 Maximum Persistence

The maximum persistence gives an indication of circularity and non-circularity in a point cloud for i-th homology. It is the radius of the most persistent homology group defined as:

$$\text{maxPers}(D_i) = \max_{(\text{birth,death}) \in D_i} (\text{death} - \text{birth}).$$

D_i is the persistence diagram for i-th homology. As a point cloud become more circular, the persistence diagram has a more prominent off-diagonal point [85].

3.1.6 Persistence Landscape

Persistence Landscape is a piecewise linear function which is a summarization of a persistence diagram. It is introduced by Bubenik and is a useful vectorization for statistical analysis of persistence diagrams [103, 104]. In essence, the persistence landscape rotates the persistence diagram so that the diagonal becomes the new x-axis. The i-th order of persistence landscapes creates a piecewise linear function from the i-th largest value of the points in the persistence diagram after the rotation. For a birth-death pair $p = (b, d) \in D$, where D is the persistence diagram, the piecewise linear functions, $\Lambda_p(t) : \mathbb{R} \to [0, \infty]$, are

$$\Lambda_p(t) = \begin{cases} t - b, t \in [b, \frac{b+d}{2}], \\ d - t, t \in [\frac{b+d}{2}, d], \\ 0 \qquad otherwise. \end{cases}$$

The persistence landscape is then $F : \mathbb{R} \to \mathbb{R}$

$$\{F(t) = \sup_{p \in D}(\Lambda_p(t))\}.$$

Figures presenting persistence landscapes will be presented in the method section.

3.2 Dynamical Systems

Dynamical systems are constructed from an abstract phase state or state space. The coordinates of the space represent the states available. The system is considered dynamical because states can change depending on time. Dynamical systems can be both deterministic and stochastic. A dynamical system can therefore formally be described as a phase or state space, S, a temporal space, T and an evolutionary function Φ, where $\Phi : S \times T \to S$. In other words the state x_{t+1} is given by $\phi(x_t)$, where time $t = 0, 1, 2, \ldots$. When the variables are discrete it is called a state space, whereas when the variables are continuous the equivalent space is called phase space.

3.2.1 Takens embedding

To understand Takens embedding it is vital to understand what dynamical systems manifolds and embeddings are.

Dynamical systems are mathematical objects used to model phenomena with states that vary over time. These systems are often used to predict, explain or understand phenomena. The state at time t is a description of the system and the evolution of the system is a trajectory through the space of possible system states. Attractors are points in the space that the trajectory is drawn towards. These possible system states are called the state space or phase space of the dynamical system. A time series can be projections of observed states from such a dynamical system. The manifold of these dynamical systems can, therefore, contain information which is useful for understanding the underlying phenomena [105]. An underlying assumption in this thesis is that financial time series are dynamical systems.

An n-dimensional manifold is a topological space, $\mathcal{M}$, for which every point $x \in \mathcal{M}$ has a neighborhood homeomorphic to Euclidean space $\mathbb{R}^n$ [106]. I.e. it is a space that is locally Euclidean, but globally might be complicated topological structures. A smooth map $\Phi : M_1 \to M_2$, where M_1 and M_2 are smooth manifolds, is an embedding of M_1 in M_2 if Φ is a diffeomorphism from M_1 to a smooth submanifold of M_2. M_2 is then the embedding space with embedding dimension $\dim(M_2)$. Another way to express this is that $\Phi(M_1)$ is a realization of M_1 as a submanifold of M_2.

Takens delay coordinate embedding makes it possible to reconstruct a time series into a higher dimensional space so that the topology of the original manifold which generates the time series values are preserved. The point cloud reconstructed from a time-series has the same topology as the attractor of the dynamical system. *Whitney's embedding theorem* states that all n-dimensional manifolds can be embedded in $2d+1$-dimensional Euclidean space [107]. Takens extended this theorem by proposing that an d-dimensional manifold which contains the attractor A could be embedded in R^{2n+1} [108]. Takens theorem finds the function Φ which maps $M_1 \to M_2$, where $\dim(M_2)$ is the embedding dimension which can be $\mathbb{R}^{2n+1}$.

So the Takens embedding gives the possibility to obtain a continuous

transformation from the original manifold M to $X \in \mathbb{R}^d$ where d is the embedding dimension and X is the trajectory matrix defined as

Definition 3.2.1. *Let $x = \{x_1, x_2, \ldots, x_N\}$ be a time series and X be a trajectory matrix consisting of sequence of state variable observations with d-dimensions and τ time lag i.e.*

$$X = \begin{bmatrix} X_{1+(d-1)\tau} \\ X_{2+(d-1)\tau} \\ \vdots \\ X_N \end{bmatrix} = \begin{bmatrix} x_{1+(d-1)\tau} & \cdots & x_{1+\tau} & x_1 \\ x_{2+(d-1)\tau} & \cdots & x_{2+\tau} & x_2 \\ \vdots & \vdots & \vdots & \vdots \\ x_N & \cdots & x_{N-(d-2)\tau} & x_{N-(d-1)\tau} \end{bmatrix}.$$

where each point in space is represented by a row. This is our state space reconstruction.

An attractor is then the pattern created by the points X in space. A more formal definition is given by [109] as:

Definition 3.2.2. *Suppose $x(t) = v_j(y)$ for some $j = 1, \ldots, n$ where $v(t) = (v_1(y), \ldots, v_n(t))$ is a curve on a manifold Ω. Suppose $v(t)$ visits each part of Ω which means that $v(t)$ is dense in Ω under its topology. Then there exists $\tau > 0$, $K \in Z$, where Z denotes the real numbers, such that the corresponding vectors $(x(t), x(t+\tau), \ldots, x(t+K_\tau))$ are on a manifold topologically equivalent to Ω.*

Takens embedding assumes that the time series data is not contaminated by noise [19], as such noise get amplified according to the largest Lyapunov exponent in the process and can greatly affect the reconstructed attractor [110]. Takens embedding requires the choice of embedding dimension, m, and time delay, τ. There is no generic optimal method for choosing embedding parameters [111]. The parameter choices are important for a good quality attractor reconstruction when time series have finite length and are noisy. Below some methods for choosing parameters are presented.

Determination of dimension

A d-dimensional topological space can be embedded in $2d + 1$ Euclidean space [107]. The problem with this approach is that the orig-

inal attractor dimension d is not always known. A tighter boundary is given by Sauer who showed that the required dimension could be $d > 2d_0$, where d_0 is the box-counting dimension of the attractor of the underlying system [112]. Another approach is the False nearest neighbors approach proposed by Kennel et al. [113]. A property when embedding is that when m embedding dimensions are too low, distant points in the original phase space are close points in the reconstructed phase space. These points are called false neighbors. When calculating the false nearest neighbor for each point x_i look for the nearest neighbor x_j in an m-dimensional space. Then a ratio

$$R_i = \frac{|x_{i+1} - x_{j+1}|}{|x_i - x_j|}$$

is calculated. If the ratio R_i exceeds a given threshold R, then the point is marked as a false neighbor. If the embedding dimension is high enough the ratio R_i is zero. One way to calculate this is to embed the time series x with lag τ on a range of different embedding dimensions m. Find all nearest neighbors and compute the percentage of neighbors that remain when additional dimensions are unfolded [114].

Another method for determining m is to use singular value decomposition as used in [109]. A sufficient m should be given by the same number of linearly independent vectors derived from a trajectory matrix [115, 116].

Determination of time-delay

Two criteria are important when estimating time delay τ. 1) τ has to be large enough so that the information from the value of x at time $n + \tau$ is significantly different from information already known from observing values of x at time n. 2) τ should not be large enough so that the system loses memory of its initial state [117]. In the case that the dataset is infinite and noise free the time delay τ is not relevant, and any value chosen should suffice. As most data does not follow these properties choosing a good τ is important in most cases. If τ is too small the attractor becomes only a diagonal in the reconstructed space because of high correlation among coordinates. If τ is too large then components will be uncorrelated, which means that the reconstructed

attractor does not represent the true dynamics of the system. Further, τ should not be close to an integer multiple of a periodicity of the system. There is currently no general way of determining optimal τ [118]. The methods often used to determine τ is based on autocorrelation or mutual information. Two common autocorrelation approaches are when the autocorrelation first approaches 0 or $1/e$. Lastly, estimations of correlation dimension have also been used to determine τ [111].

3.3 Properties of Financial Time Series

Financial time series can be viewed through different resolutions. Common data resolutions are 1-min, 3-min, 5-min, 10-min, 15-min, 30-min, 60-min, 1-hour, 2-hour, daily, week, month, quarter time series. Financial time series are results of complex interactions caused by supply and demand of assets and capital. Relative to other economic time series the financial time series have some characteristic properties and shapes caused by the micro structure of the financial market [119]. The complex underlying dynamics causes these time series to have high volatility which change through time. Systematic factors can cause these time series to have trend and cycle part. However, any seasonal part often does not play any significant role [119]. It is often assumed that financial time series are *martingales*, meaning that only the latest price influence the current price [119]. This is mathematically expressed as:

$$E[P_{t+1}|P_t, P_{t-1}, \dots] = P_t,$$

i.e. The conditional expectation of the next price, given all the past prices, is equal to the most recent price. It assumes that all non-overlapping price changes are linearly independent. Another way to express this is

$$P_t = P_{t+1} + a_t,$$

where a_t is called the martingale difference and is typically assumed to be $a_t \sim \mathbb{N}(0, \sigma^2)$.

The asset price cannot be smaller than zero. Therefore, the minimal asset net return is

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}} = -1.$$

Conventionally it is assumed that the asset distribution is normally distributed. The gross return for k period's from time $t-k$ to time t can be expressed as the products of the periods returns:

$$R_t(k) + 1 = (R_t + 1) \cdot (R_{t-1} + 1) \cdot \cdots \cdot (R_{t-k+1} + 1) = \\ = \frac{P_t}{P_{t-1}} \cdot \frac{P_t}{P_{t-1}} \cdots \frac{P_{t-k+1}}{P_{t-k}} = \frac{P_t}{P_{t-k}}.$$

These returns terms are normally distributed, but the product of them is not. To overcome this a logarithmic transform is used so that log-normal distribution is obtained. The logarithmic transform of random variable with log-normal distribution is normally distributed,

$$X \sim \text{Lognormal}(\mu, \sigma^2),$$
$$Y = \ln X \sim \mathbb{N}(\mu, \sigma^2).$$

Therefore, by applying logarithmic transformation to the log-normally distributed gross returns one obtains normally distributed log-normal returns, which we can take the sum of,

$$R_t + 1 = \frac{P_t}{P_{t+1}} \sim \text{lognormal}(\mu, \sigma^2),$$
$$r_t = \ln R_t + 1 = \ln P_t - \ln P_{t-1} \sim \mathbb{N}(\mu, \sigma^2).$$

The return for k periods from $t-k$ to time t is expressed

$$r_t(k) = r_t + r_{t-1} + r_{t-2} + \cdots + r_{t-k+1} = \sum_{i=t-k+1}^{t} r_i.$$

An example is shown of financial time series and corresponding log-normal return is also shown in figure 3.3 and figure 3.4:

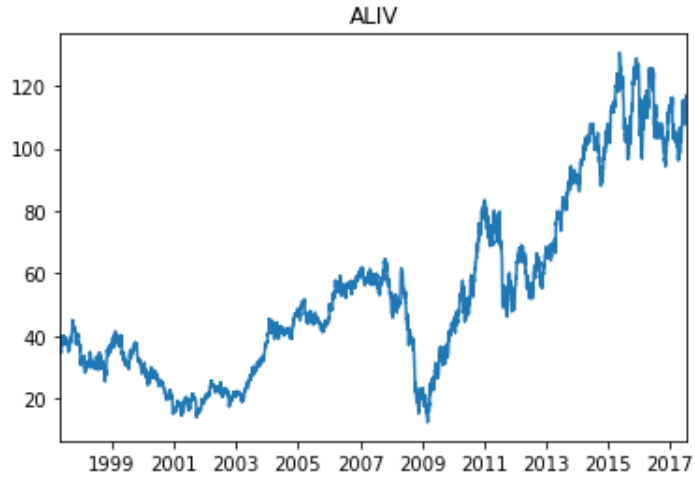


Figure 3.3: Financial time series of the Swedish Autoliv stock in OMXS30 between 1997 - 2017.

.

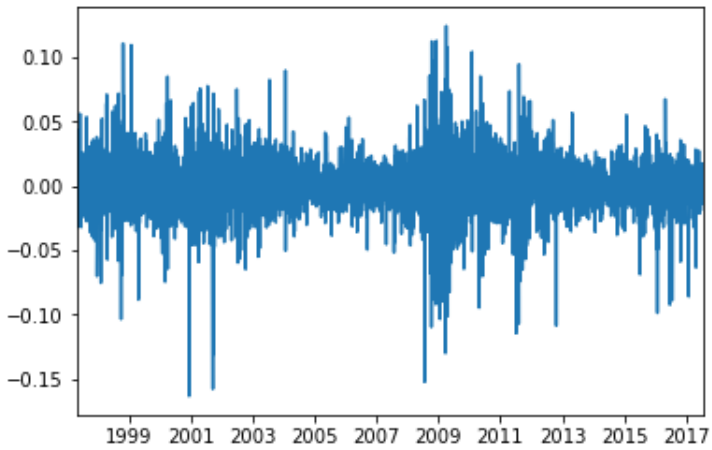


Figure 3.4: Log-normal return plot corresponding to figure 3.3.

.

Normality of log-return is a common assumption in quantitative financial studies [119]. The distribution is symmetric so the skewness and kurtosis are expressed as:

$$\mathrm{SK}_r = E\left[\frac{(r_t - \mu)^3}{\sigma^3}\right] = 0,$$

$$\mathrm{K}_r = E\left[\frac{(r_t - \mu)^4}{\sigma^4}\right] = 3.$$

However, empirical studies have shown that market estimates of skewness are negative and the point estimates of return means are close to zero, which means that the return distribution is skewed to make big negative returns more probable than big positive returns. The kurtosis has been empirically shown to be consistently bigger than 3, indicating that empirical distributions are more peaked than a theoretical normal distribution. This means that low positive and negative returns are more probable than suggested by a theoretical normal distribution. Fig. 3.5 shows the theoretical and empirical log-normal return distribution of Autoliv stock.

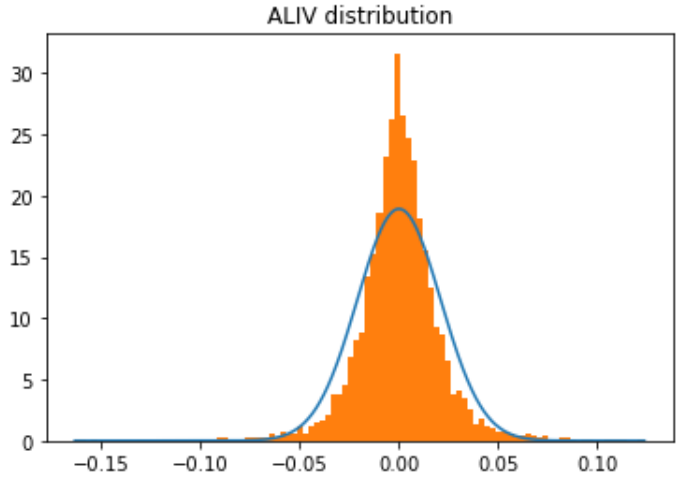


Figure 3.5: Theoretical normal distribution and Empirical Log-normal return distribution of Autoliv.

.

The fact that the empirical distributions are skewed and more peaked than theoretical distribution has been well known for a long time and have been described as far back as the 1960s by Mandelbrot and Fama [120, 121]. Some studies suggest that the Laplace distribution is a more suitable distribution for financial returns [122].

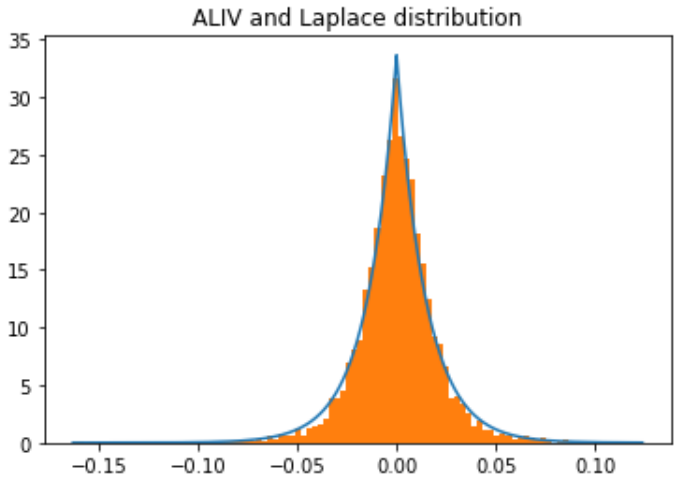


Figure 3.6: Theoretical Laplace distribution and Empirical Log-normal return distribution of Autoliv

.

Fig. 3.6 show that Laplace distribution does seem to fit the empirical log-return distribution better. QQ-plots show their fitness to respective distributions.

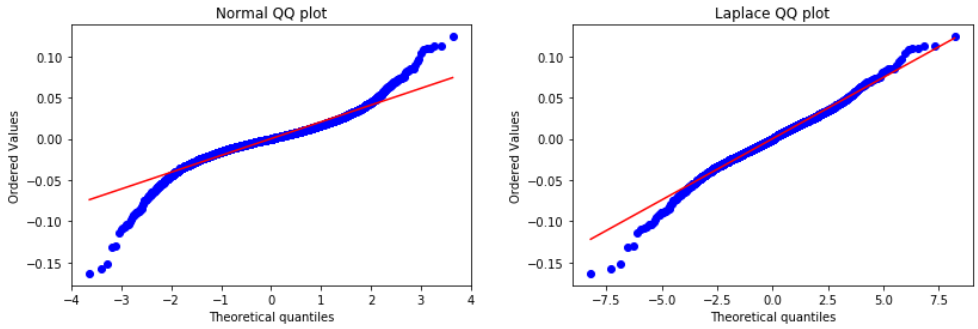


Figure 3.7: (Left) ALIV normal QQ plot, sum of squared error $SSE = 0.1645$, (Right) and Laplace QQ plot, $SSE = 0.0268$.

The parameters for the QQ-plot is found using least-squares regression. The fitted Laplace distribution is La$(\mu = 0, b = 0.15)$. The SSE for laplacian QQ-plot is lower than the normal QQ-plot indicating that Laplace distribution has a better fit for financial returns and is a viable alternative to a normal distribution. Further, the QQ-plot shows that the empirical distributions have heavy tails in comparison to a

normal distribution and only heavy left tail in relation to Laplace distribution.

Lastly, it is often assumed that log-returns are independent, identically distributed with zero mean and constant variance i.e. financial time series are often assumed to be strict white noise processes. However, empirical studies have shown that these time series often are more complex than this [119]. None of the conditions are fulfilled in reality. In fact, the volatility has been shown to be constantly changing over time. This phenomenon studied as early as the 1960s by Mandelbrot [120].

3.4 Time Series and Signal De-noising

Financial time series inherently are quite jittery, which might affect Takens state space reconstruction. Smoothing might remove some of the jitters and make Takens state space reconstruction more efficient. Below are some basic smoothing methods.

3.4.1 Moving Average

The moving average (or rolling average) is a smoothing method for time series. It is created by averaging different subsets of fixed size of the data. The moving average is created by shifting forward the subset window along the time series. I.e. given a data sequence $\{a_i\}_{i=1}^{N}$ an n-moving average is a sequence $\{s_i\}_{i=1}^{N-n+1}$ defined from a_i by taking the arithmetic mean of subsequences of b terms.

$$s_i = \frac{1}{n} \sum_{j=1}^{i+n-1} a_j.$$

The sequences of S_n giving n-moving averages are

$$s_2 = \frac{1}{2}(a_1 + a_2, a_2 + a_3, \ldots, a_{n-1} + a_n),$$

$$s_3 = \frac{1}{3}(a_1 + a_2 + a_3,\ a_2 + a_3 + a_4, \ldots,\ a_{n-2} + a_{n_1} + a_n).$$

The method is often used as a technical analysis indicator for financial data.

3.5 Time Series Point Cloud Representation

Many different approaches can be used to represent a financial time series as a point cloud. This section will go through some of the methods available.

3.5.1 Sliding Window

Sliding window technique can be used to get different sets of point clouds from a single time series. Using this method time series data $f(T)$ are segmented into $SW_{M,tau} = \{f(t), f(t+\tau), \ldots, f(t+M_\tau\}$ i.e. $M+1$ partitions, where M depends on our time series length T, window size M_τ and step size τ. An illustration of the procedure is shown in figure 3.8.

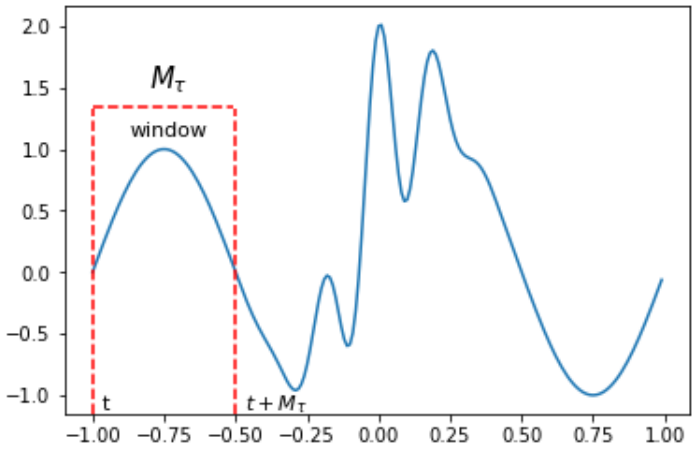


Figure 3.8: Illustration of Sliding window procedure, see also Parea and Harer [12].

.

3.6 Principal component analysis

The computational time for construction of alpha complexes on high dimensions can be prohibitively high because of the complexity of the Delaunay triangulation. For n points in $\mathbb{R}^d$ the complexity for Delaunay triangulation can be $\mathcal{O}(n^{\frac{\lceil d \rceil}{2}})$ [123, 124]. In practice the complexity is much lower in $\mathbb{R}^3$, as the complexity is bound to $\mathcal{O}(n \log n))$ for points distributed on generic smooth surfaces in $\mathbb{R}^3$ [125]. Therefore, dimensionality reduction can be performed to reduce the dimensions, which makes computations for large datasets more feasible. Principal component analysis makes it possible to summarize variables with a smaller number of components. These components collectively account for the most of variance of the original data. The principal components are normalized linear combinations of the original data features that are uncorrelated to each other [126],

$$Z_k = \phi_{1k} X_1 + \phi_{2k} X_2 + \cdots + \phi_{pk} X_p,$$

where Z_k is the k-th principal component, $X_1, \ldots, X_p$ are p different features of the data and $\phi_{1k}, \ldots, \phi_{pk}$ are the loadings or weights for Z_k, where $\sum_{j=1}^{p} \phi_{jk}^2 = 1$.

The variance or proportion of variance of the PCA can be used as a diagnostics tool for PCA. The variance for k-th principal component is

$$\frac{1}{n} \sum_{i=1}^{n} Z_{ik}^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} (\phi_{jm} x_{ij})^2,$$

and the proportion of variance explained by k-th principal component is obtained by dividing the k-th principal component by the number of features,

$$\frac{\frac{1}{n} \sum_{i=1}^{n} Z_{ik}^2}{\sum_{j=1}^{p} Var(X_j)} = \frac{\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} (\phi_{jm} x_{ij})^2}{\frac{1}{n} \sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{p} (\phi_{jm} x_{ij})^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2}.$$

A bar chart representing the variations or proportion of variance of each principal component is called a scree plot. These PCA variations

are proportional to the eigenvalues and can be used as diagnostics tools for PCA. It is desirable that the first few principal components account for most of the variation of the data.

3.7 Entropy

3.7.1 Shannon Entropy

Shannon entropy H is defined as

$$H = -\sum_i p_i \log_b p_i,$$

where p_i is the probability of a certain occurrence. It is an estimate of the average minimum number of bits required to encode a piece of information.

3.7.2 Gzip compress-to-ratio

Gzip compress-to-ratio is the ratio of a file compressed with gzip against the original file i.e. how much entropy there is in a piece of information in practice.

$$\text{Gzip compress-to-ratio} = \frac{\text{Original file size}}{\text{Gzip compressed file size}}.$$

Chapter 4

Method

This section outlines the methodology used in this thesis.

4.1 Data pre-processing

The data used consisted of a financial time series of nanosecond FX data and quantum noise, QN, reference data. The datasets were provided by Marcello Paris from the investment bank UniCredit. For this thesis, the ask price was used simply because it is the price used for spot purchases. To make the FX data stationary log-return transformation was used i.e.

$$r_i = \ln P_t - \ln P_{t-1}.$$

The FX dataset was then standardized to get it to unit variance by setting

$$X_{standardized} = \frac{X_{raw} - \mu}{\sigma}.$$

The unit variance was required to make it comparable with other datasets. Standardization was used instead of normalization because the procedure was unbounded. This was necessary because extreme values could contain important information in financial data.

An investigation of the probability distribution of the FX dataset was then performed to know what type of distribution on the random data would make the fairest reference. The investigation was conducted using empirical distributions and QQ-plots. The quantum noise data were normalized to the open interval $(0, 1)$ with

$$X_{normalized} = \frac{X_{raw} - X_{min}}{X_{max} - X_{min}}.$$

The normalization was required to make it more practical as a tool for random variable generation from different distributions. As theory section 3.3 has stated it is often assumed that financial data returns are normally distributed. Also, there are studies claiming that a Laplacian distribution is a better fit than a normal distribution[122]. To obtain normally distributed $N(0, 1)$ random variables from $U(0, 1)$ distributed data inverse transform sampling was used. Inverse transform sampling is defined as:

$$Y = \mu + \sqrt{2}\sigma \text{erf}^{-1}(2 * X - 1), \qquad X \in U(0, 1),\ Y \in N(\mu, \sigma),$$

where the right side of the equation is the inverse CFD of $N(\mu, \sigma)$. If $N(\mu, \sigma) = N(0, 1)$ then normally distributed random variables can be used get Laplace distributed $L(0, b)$ random variables. The inverse transform sampling was used to sample $N(0, 1)$ distributed random variables $Z_k, k \in \{1, \cdots, 4\}$. Then following formula gives $L(0, b)$ random variables from $N(0, 1)$ random variables:

$$V = \frac{Z_1 \cdot Z_2 - Z_3 \cdot Z_4}{b}, \qquad Z_1, Z_2, Z_3, Z_4 \in N(0, 1),\ V \in La(0, b),$$

where setting the scaling factor $b = 1$ gives $La(0, 1)$ samples from the $N(0, 1)$ samples. All random variables were standardized.

Lastly because of truncation error and the fine granularity of the nanosecond FX data, the return-values from the financial time series were discrete. To make the datasets comparable in with respect to complexity, the QN data was quantized. To quantize the QN data, the number of unique log-returns was calculated. The QN was then multiplied by scaling factor s, rounded to nearest integer after scaling, and rescaled

to original scale by dividing by the scaling factor s to keep the standardization properties $\mu = 0$ and $\sigma = 1$ as good as possible. The formula for quantization is presented below:

$$QN_{discrete} = \frac{\|QN \cdot s\|}{s}.$$

4.2 Analysis process description

This section gives an overview of the analysis process.

4.2.1 Sliding window

To analyze if different segments of the time series have different topological features sliding window first used to partition the time series into different windows. The sliding window was presented above in theory section 3.5.1. There are two parameters which need to be chosen; window size w and the step or gap size g. The choice of parameters should be viewed as looking at the data with different scaling. Choices were made for computational reasons and different parameters were chosen to verify results experimentally.

4.3 Point cloud representation of time series using Takens embedding

A State space reconstruction was then performed using Takens embedding on each sliding window partition. The method was presented in the theory section 3.2.1. It constructs a state space from time series values and requires two parameters; time delay τ and embedding dimension m. This transforms a time series $X = \{x_1, x_2, \ldots, x_N\}$ to a trajectory matrix $f(X) = \{X_{1+(d-1)\tau}, X_{2+(d-1)\tau}, \ldots, X_N\}$, where each $X_{n+(d-1)\tau}$ are windows of the original time series X containing m data points. Each window $X_{n+(d-1)\tau}$ then represents a point in the state space reconstruction and the points reside in a m dimensional space. When Takens embedding did not yield any successful reconstructions,

the specific window was discarded. These cases are prevalent when windows contain only single value i.e. $W = \{0, 0, \ldots, 0\}$. As mentioned in the theory section 3.2.1 there are no universal method for selecting optimal τ and m. However, there are some standards for parameter selection. For the sake of comparability and computational resources, same parameters were used throughout. The choice of parameters should be seen in this case purely as motivated heuristics.

τ has to be large enough so that the information from values of the time series, X, at time $n + \tau$ is significantly different from what is already contained in X_n and τ should not be large enough to lose memory of its initial state[117]. It should also not be large be an integer multiple of a periodicity of the system [118]. The periodicity can be detected as peaks in the spectral density [127]. For the selection of τ a qualitative analysis of the data based on the properties of financial time series was used in conjunction with the more formal methods of first zero and first $1/e$ decay of the autocorrelation function [111]. To check for periodicity in the system power spectral density estimation by Welch method was used.

The embedding dimension of a d-dimensional topological space can be $2d + 1$ in Euclidean space [107]. However, the original dimension d is not known in the FX dataset. A common problem when having a low embedding dimension m is that distant points in the original state space are close in the reconstructed space. The false nearest neighbors (FNN) approach addresses this problem and is therefore used to find the embedding dimension m [113]. Details of the method are found in theory section 3.2.1. Ideally, zero FNNs would be preferred. However, the dataset had FNN with very long convergence towards zero or asymptotic convergence above zero which would make it either impossible or require unfeasible computational power to reach zero FNN. To make the computations feasible a the embedding dimension m was selected to be the mean of the derivative of the FNN lower than an arbitrary set threshold ϵ,

$$m = E[\text{dFNN}], \quad \text{dFNN}_i \leq \epsilon, \quad i \in 1, 2, 3, \ldots, N,$$

where dFNN is the derivative of FNN and N is the number of embedding dimensions in FNN.

It should be noted to the reader that Takens embedding is not the only available method for point cloud representation of time series. Gidea et al. use a return point-cloud, whereby a point cloud is created by having different return time series as features [81, 82]. Using this method means that an analysis of topology in volatility is conducted. The method does not allow for topological data analysis of one dimensional time series. Other methods that can be used for include circular coordinate representation of time series, network representations (such as recurrence network [128] and complex networks [129]) and visibility graphs [130]. Takens embedding was chosen because it shows properties of the dynamical system of time series.

4.4 Dimensionality reduction of Reconstructed state space

The choice of embedding dimension $m \gg 3$ made the reconstructed state space high dimensional. To make the extraction of topological features computationally feasible for m dimensions PCA was used to reduce the dimensions from $\mathbb{R}^m \rightarrow \mathbb{R}^3$. The reason PCA was chosen was that it represents the dimensional directions with most variations and thus contains most useful information. PCA spree plots are used as diagnostics tools for the PCA. A drawback of this method or any other dimensionality reduction method is that information is lost in the reduction of dimensions.

4.5 Topological data analysis of dimensionality reduced reconstructed state space

It is deceivingly hard to detect topological features by visual inspection even in low dimensions. To extract the topological features persistent homology was employed. The point cloud resided in $\mathbb{R}^3$ and had a large amount of data points. Alpha complexes were fastest to construct and were therefore used to represent the topological features.

The birth-death diagrams resulting from persistent homology was then used to construct persistence landscapes. The use of persistence landscape was two-fold. Firstly the birth-death diagrams can be hard to interpret when there is a lot of features. More importantly, it does not reside in a vector space, but rather in a Polish space and therefore common statistical procedures are not efficient at analyzing the outputs [131]. The persistence landscape, on the other hand, resides in a vector space and are easily combined with common statistical tools [103]. One way to make it possible to use statistics on the persistence diagrams is to use Wasserstein distance [131]. However, the Wasserstein distance was computationally unfeasible for this thesis. The construction of persistence landscapes can also be quite computationally expensive if there are a lot of topological features in the birth-death diagram. As a speedup noisy topological features can be eliminated from the birth-death diagram before constructing the persistence landscapes. This can be done by specifying a cut-off value ϵ and removing all topological feature below this radius threshold but was not needed in this thesis.

Since the persistence landscape resides in the vector space statistical procedures can aid in their interpretation [103]. In this thesis, the mean landscape was used to summarize the persistence landscape. Integral of the persistence landscapes and maximum persistence was used for window-by-window comparison with complexity calculations. The persistence landscape integrals and maximum persistence were compared against Shannon entropy and gzip compress-to-ratio. Lastly, comparisons of the distribution of the persistence landscape integrals for the FX data and reference data was also performed.

Chapter 5

Synthetic examples of topological data analysis of reconstructed state spaces

This section provides synthetic examples of topological data analysis of reconstructed state spaces to give the reader an intuitive understanding of the process used in this thesis. Takens embedding allows for reconstructing a time series into a m-dimensional point cloud. The topological features in the point cloud then resemble some property of a time series. To give an understanding of what these topological features represented in a time series this section will use simulated data and their corresponding state space reconstruction to demonstrate. Further, the effect of noise and quantized data on the reconstructed state space will also be shown.

5.1 Pure models

The first example presented is a simple sine-wave simulated with 1000 data points

$$y = sin(x), \quad 0 \leq x \leq 16\pi.$$

Using $m = 2$ following state space reconstructions are recreated using

different τ.

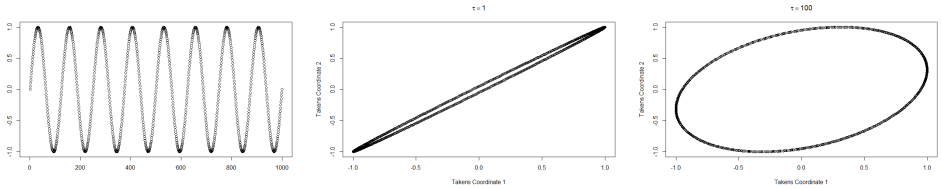


Figure 5.1: (Left) The sin-plot, (middle) reconstructed state space $\tau = 1$ (right) and $\tau = 100$.

A smaller τ yield a more collapsed representation almost becoming a diagonal. However, both figures are homotopy equivalent as both form loops. Their topological features have different persistence in the persistence diagram.

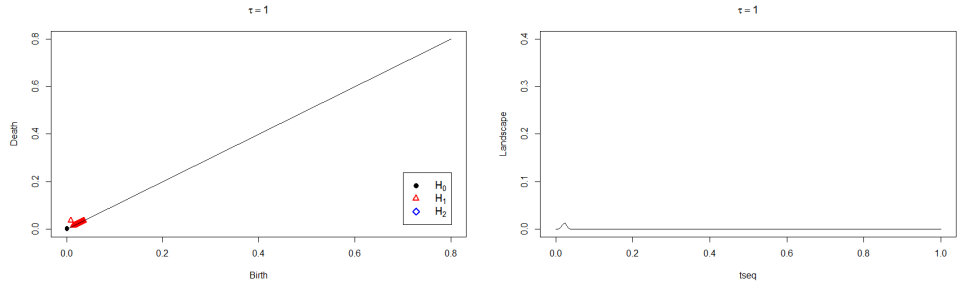


Figure 5.2: (Left) Persistence diagram (right) and landscape for $\tau = 1$.

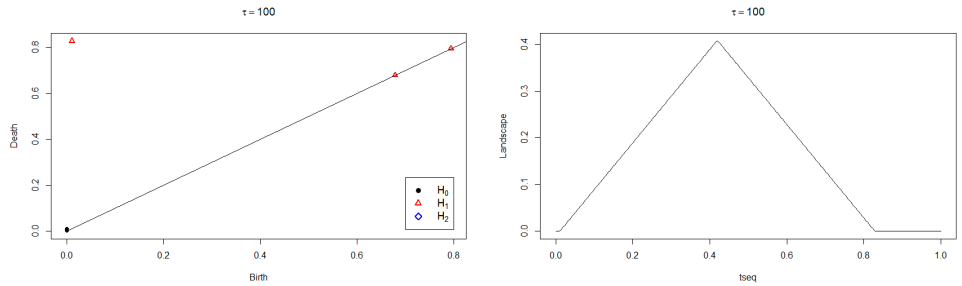


Figure 5.3: (Left) Persistence diagram (right) and landscape for H_1 $\tau = 100$.

Fig 5.2 and 5.3 show the persistence diagrams and landscapes for the sin-wave simulated values. The landscape summarizes the H_1 com-

ponents (red components). Notice that they indicate the same homology. The homological persistence differs when changing τ. A smaller τ gives a smaller persistence, meaning that noise could more easily "hide" the true topology in the case of smaller τ. This is because a smaller τ incorporates less information to the state space reconstruction. This phenomenon will be further investigated further down in section 5.2.

The second model is composed of high and low-frequency part and a linear component. The example is simulated with 1000 data points.

$$y = k \cdot sin(x) \cdot sin(ax) + a \cdot x, \quad 0 \leq x \leq \pi, \quad k = 4, \quad a = 32.$$

Using $m = 3$ allows each of the three components gets an own axis representation on the phase state reconstruction.

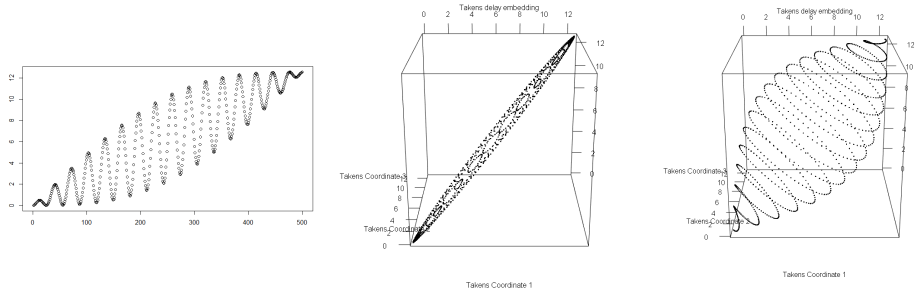


Figure 5.4: (Left) Plot of the second equation, (middle) reconstructed state space $\tau = 1$ (right) and $\tau = 20$.

The case when $\tau = 20$ yields an oval with a void. The spiraling loops are composed by the high-frequency part $sin(ax)$, the radius component is composed by the low-frequency part $sin(X)$ and the length is composed of the linear component x in the equation. k is only a scaling component. Since $\tau = 1$ yielded a collapse result only $\tau = 20$ persistence diagram will be presented.

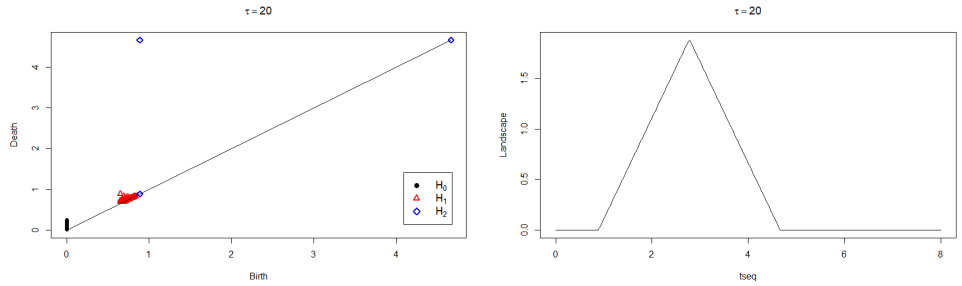


Figure 5.5: Persistence for $\tau = 20$(Left) Birth-death diagram, (right) landscape for H_2.

The landscape in fig 5.6 shows the summary of H_2 components instead (the blue components).

The phenomenon of state space reconstruction collapsing to the diagonal due to low τ is shown in the case when $\tau = 1$ in fig 5.4[117]. Interestingly applying PCA to the collapsed reconstruction state space with $\tau = 1$ in 5.4, gives an "enhanced" representation of the topology of the figure.

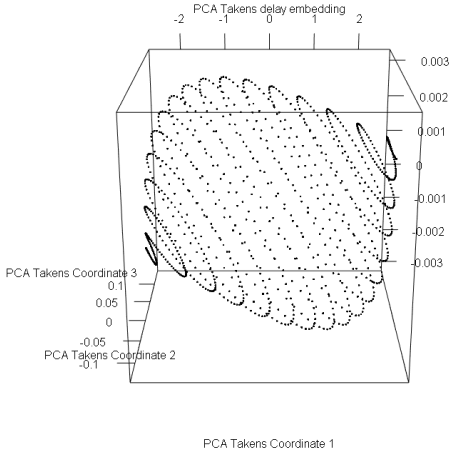


Figure 5.6: True topology of the collapsed state space reconstruction when $\tau = 1$ is spanned up by PCA.

This property can be attributed to the fact that the principal components span up the basis that accounts for most of the variation. The collapsed representation is not completely collapsed and only visually

obscuring the topological properties in this case. However, it is possible that other cases can completely obscure the topological properties. Therefore, PCA should only be seen as an enhancement of topological properties in an environment when the topology of the point cloud is more discernible than the noise in the data. More importantly, the PCA does not change the underlying topology in the case when the dimensions $\mathbb{R}^k \to \mathbb{R}^k$. The same cannot be necessarily be said for when $\mathbb{R}^k \to \mathbb{R}^n$, where $n < k$.

5.2 Noisy models

Now noise is added to

$$y = k \cdot sin(x) \cdot sin(ax) + a \cdot x + \epsilon, \quad 0 \leq x \leq \pi, \quad k = 4, \quad a = 32.$$

The noise component is

$$\epsilon = f \cdot \frac{(max(x) - min(x))}{50},$$

where f is a scaling factor. A low noise example $f = 1$ and high noise example $f = 10$ is presented.

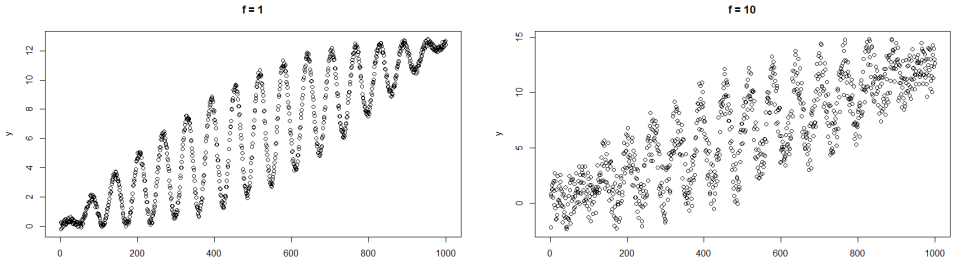


Figure 5.7: Model with noise added (left) $f = 1$ and (right) $f = 10$.

When $\tau = 1$ following state space reconstruction of low noise model and PCA for it is given

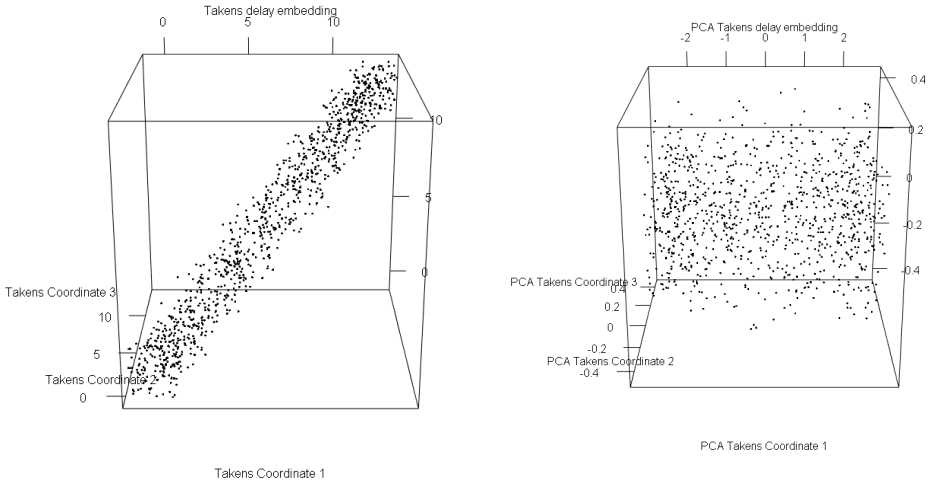


Figure 5.8: (left) State space reconstruction $\tau = 1$ and (right) PCA of the results.

When the noise is larger than the small variation caused by a collapsed state space reconstruction, the PCA in combination with persistent homology is no longer available to recover the true topology. Now the dominating factor becomes the noise which hides the true topology of the data. The persistence diagram shows that the same as mentioned and is therefore left out. As it did not manage to uncover the low noise model. The high noise model for $\tau = 1$ is omitted.

Now using the low-noise for $\tau = 20$ gives the following result.

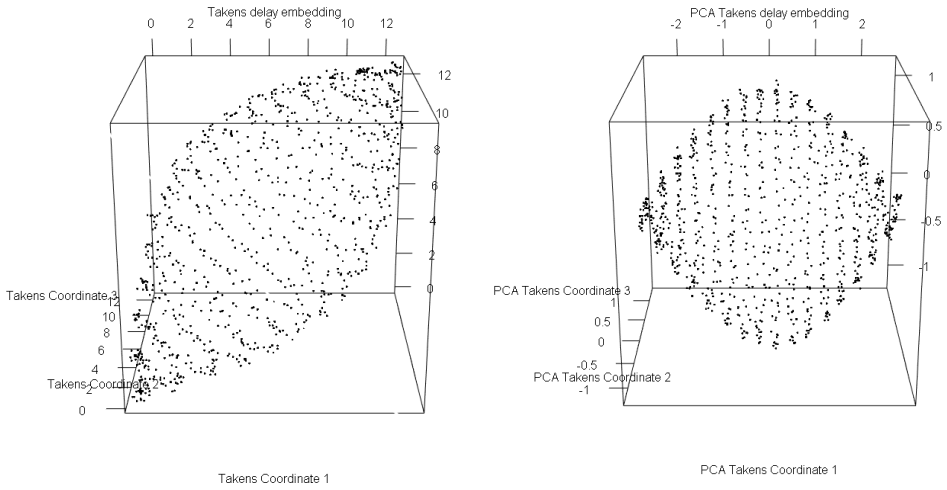


Figure 5.9: (left) State space reconstruction $(\tau = 20, f = 1)$ and (right) PCA of the results.

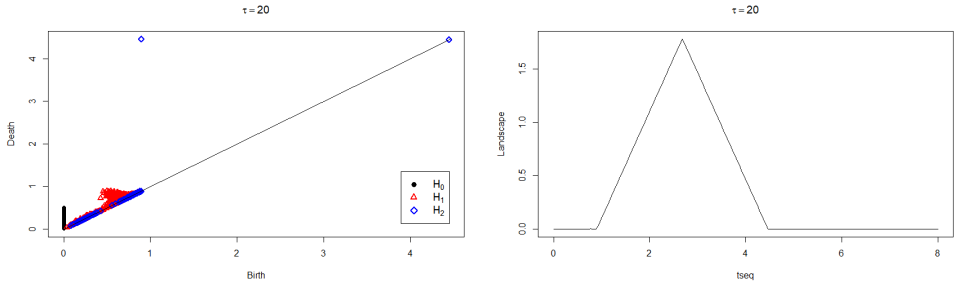


Figure 5.10: (left) Persistence diagram of state space reconstruction $\tau = 20$ and (right) its corresponding landscape of H_2 groups.

The results in fig 5.6 and 5.10 are similar. This indicates that adding a small amount of noise to a reconstructed state space do not significantly impact the topological properties when the reconstruction is not collapsed to the diagonal. Now the following results are obtained for the high noise case $f = 10$ with $\tau = 20$.

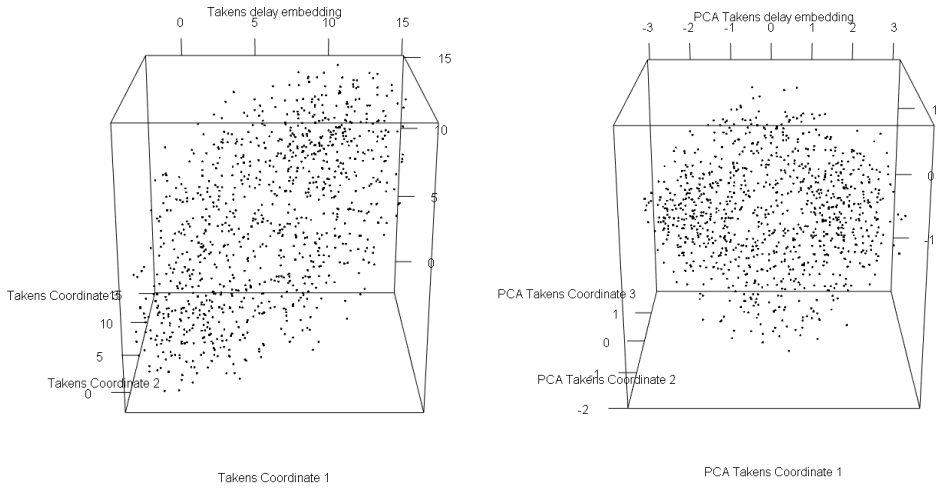


Figure 5.11: (left) State space reconstruction ($\tau = 20, f = 10$) and (right) PCA of the results.

Visually inspection does not show any clear H_2 groups in the high noise model. Applying persistent homology to analyze the data following was obtained.

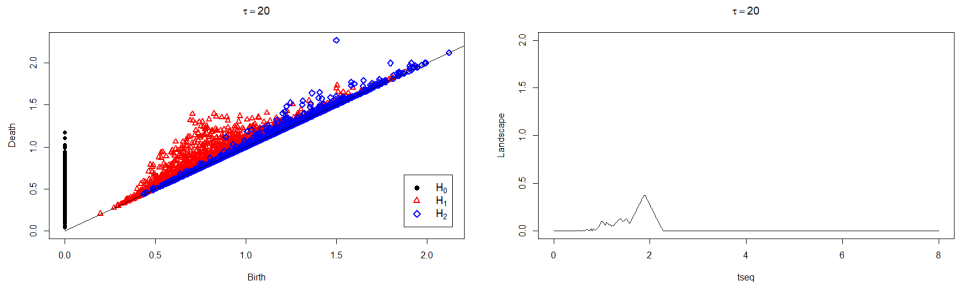


Figure 5.12: (left) Persistence diagram of state space reconstruction ($\tau = 20, f = 10$) and (right) its corresponding landscape of H_2 groups.

H_2 is considerably less prominent but persistent homology still manages to detect it. The noisy features are also much more prominent in this case as seen in fig 5.12.

5.3 Smoothing noisy data

Smoothing the noise makes the values contain less jitter. By removing this the topology of the manifold generated by state space reconstruction becomes much clearer. To show this the high noise model with $f = 10$ is reconstructed with $\tau = 20$ and then smoothed using moving averages with window size $M = 20$. The following results are obtained

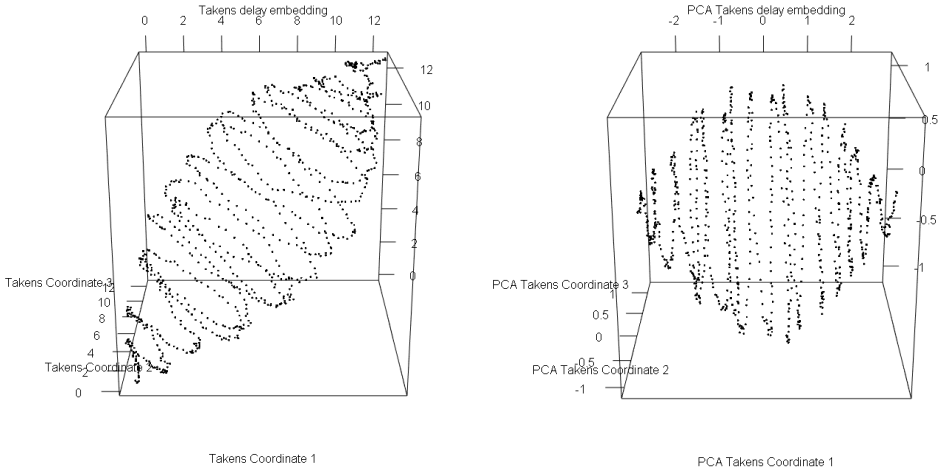


Figure 5.13: (left) State space reconstruction of smoothed model with $(\tau = 20, f = 10)$ and (right) corresponding PCA.

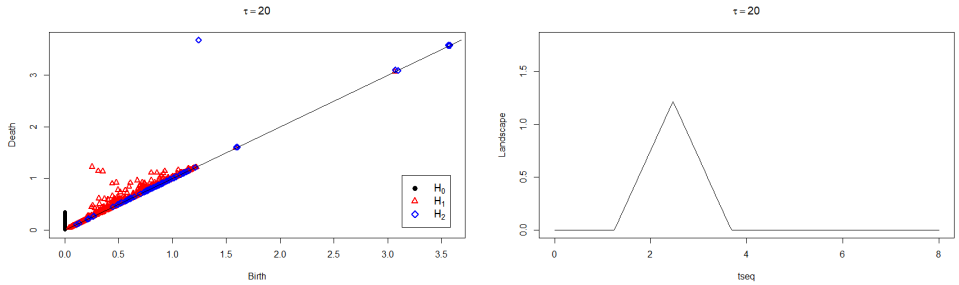


Figure 5.14: (left) Persistence diagram of state space reconstruction of smoothed model with $(\tau = 20, f = 10)$ and (right) its corresponding landscape of H_2 groups.

Now comparing fig 5.14, 5.12 and 5.5 it is evident that smoothing

data can improve topological features prominence. Smoothing did not manage to uncover the void when $\tau = 1$.

5.4 Effect of quantization of data

The data is quantized using

$$Y_{discrete} = \frac{\|Y \cdot s\|}{s},$$

where $s = 0.5$ is chosen, to get quantization fewer steps than rounding to integers. The following pure model is quantized:

$$y = k \cdot sin(x) \cdot sin(ax) + a \cdot x, \quad 0 \leq x \leq \pi, \quad k = 4, \quad a = 32,$$

and the noisy model. The models are presented below.

$$y = k \cdot sin(x) \cdot sin(ax) + a \cdot x + \epsilon, \quad 0 \leq x \leq \pi, \quad k = 4, \quad a = 32.$$

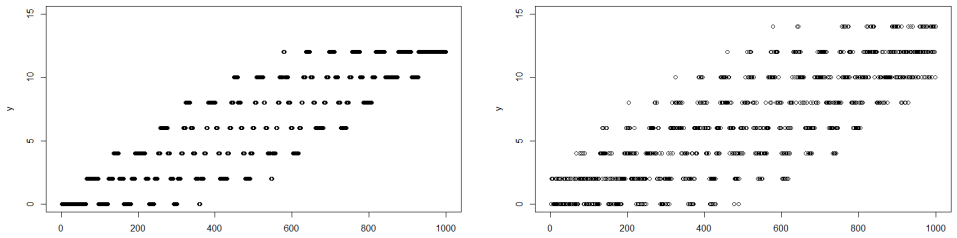


Figure 5.15: (left) Quantized pure model and (right) quantized noise model with $f = 10$.

First looking at the pure model, the topological properties of the manifold reconstructed are clear without smoothing techniques.

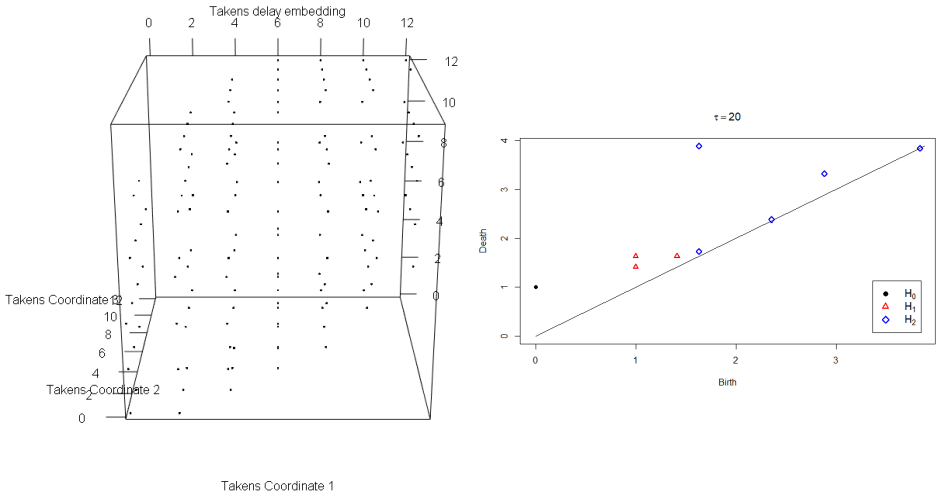


Figure 5.16: (left) Reconstructed state space of the quantized pure model $\tau = 20$ and (right) corresponding persistence diagram.

Adding low noise to the model does not significantly affect the results and figures of them are therefore omitted. Adding high noise $f = 10$ makes the topological features hidden in the reconstructed state space.

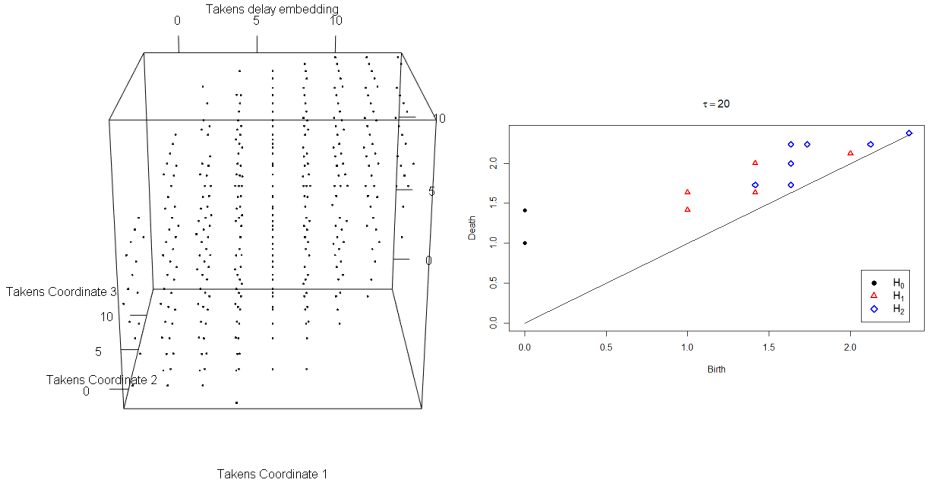


Figure 5.17: (left) Reconstructed state space of the quantized noisy model $\tau = 20$ and (right) corresponding persistence diagram.

Now by smoothing the quantized data, we can again recover the topology.

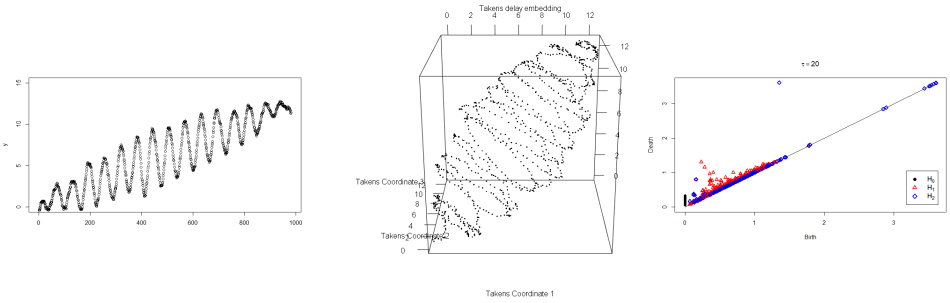


Figure 5.18: (left) Smoothed quantized noisy model $(M = 20, \tau = 20)$, (middle) its reconstructed state space and (right) corresponding persistence diagram.

The noisy model $f = 10$ was smoothed with window size $M = 20$, and as fig 5.18 show, the reconstructed state space manages to recover the same topology as the pure model. While topological features can be detected in noisy data they are much less persistent. When this low persistent is coupled with quantized data, the topological features can disappear. To counteract the effect of quantization, smoothing can be used.

5.5 Higher dimension

Previous sections presented the models that could be reconstructed perfectly using 3 dimensions. This section presents an example of a model requiring 4 dimensions to be presented using only 3 dimensions. The following model used is

$$y = (k \cdot sin(x) \cdot sin(ax) + a \cdot x) \cdot sin(4x), \quad 0 \leq x \leq \pi, \quad k = 4, \quad a = 32.$$

This is the same model as above model containing the H_2 group but multiplied by another sinus function. This sinus function should be represented by an additional dimension. As it is a sinus function with

two periods, it should be represented as a high dimensional loop. However, it is not possible to visualize such a case. Performing PCA on a state space reconstructed using the models with $\tau = 20$ and $m = 4$ to get the dimension down to 3 yields the following result.

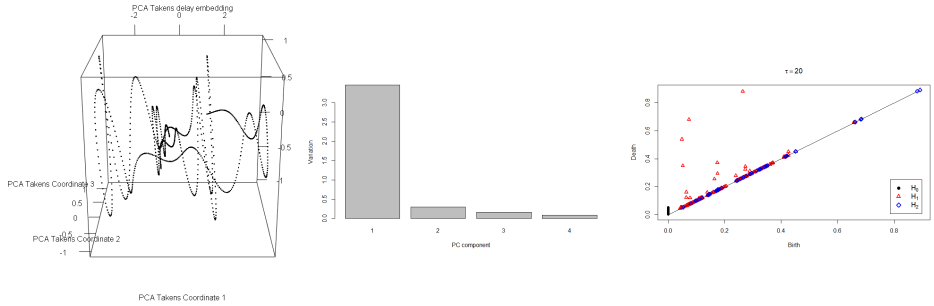


Figure 5.19: (left) PCA of Reconstructed state space of the 4D model $\tau = 20$, (middle) corresponding PCA scree plot and (right) persistence diagram.

The PCA of a higher dimensional structure does not necessarily retrieve the topology of the higher dimensional structure. Instead, it shows the topological feature of the principal components. Now adding noise with noise factor $f = 10$ and quantizing the data with scaling factor $s = 0.5$ the same procedure yields

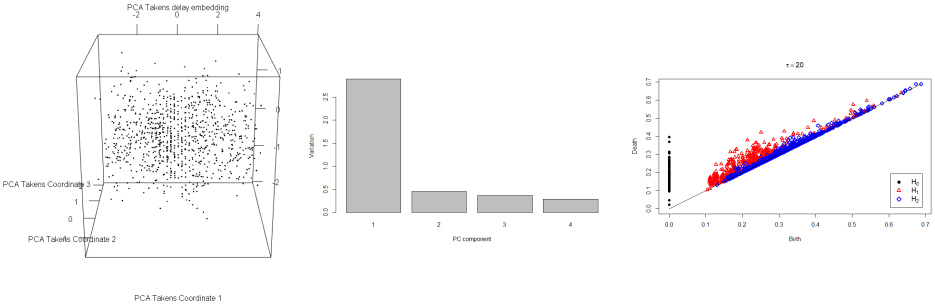


Figure 5.20: (left) PCA of Reconstructed state space of the noisy quantized 4D model $\tau = 20$, (middle) corresponding PCA scree plot and (right) persistence diagram.

It is evident that having quantized noisy data the topological features easily become obscured. Now using the moving average with window size $M = 20$. The following results are obtained.

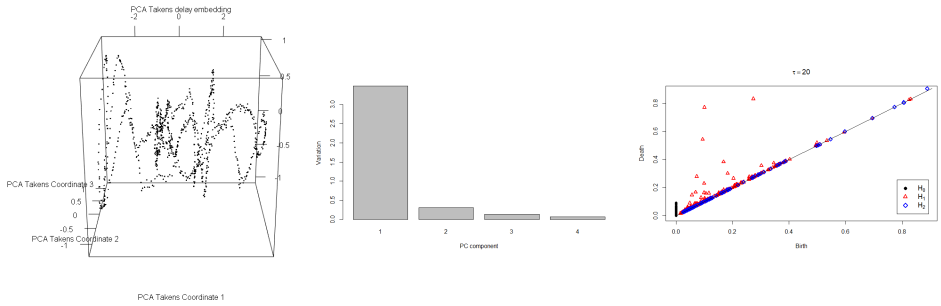


Figure 5.21: (left) PCA of Reconstructed state space of the smoothed noisy quantized 4D model $\tau = 20$, (middle) corresponding PCA scree plot and (right) persistence diagram.

Using the moving average with window size $M = 20$, can completely recover the topological features obscured by quantizing noisy data.

Interestingly when taking the PCA the topological features are mainly H_1 in this case.

Chapter 6

Results

6.1 Data and pre-processing

The datasets consisted of nanosecond EURUSD and quantum noise, QN, provided by UniCredit. The nanosecond EURUSD had approximately 8.26 million data points between 2017-08-14 and 2017-08-18. The dataset was composed of Unix time stamp, bid and ask data. All values were denoted to the fifth decimal point. The data is presented below:

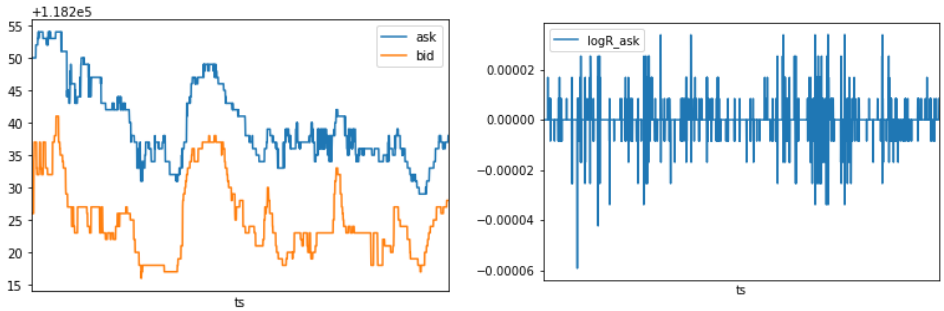


Figure 6.1: Sample raw data of 2000 data points with bid, ask (left) and corresponding log-returns for ask (right).

The data is then standardized to get it to unit variance and the resulting log-return plot becomes:

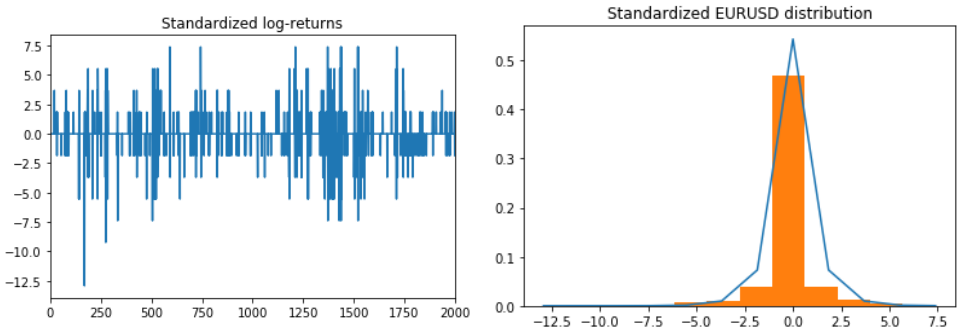


Figure 6.2: (Left) standardized log-return ask prices with $\mu = 0$ and $\sigma = 1$. (Right) Empirical and best fitted laplace distribution $L(0, 0.92)$ of standardized log-return ask prices.

The QN data is used as a reference of randomness. It is provided in binary format but is converted to 4-byte integers to get integer representation of the randomness. The data is normalized to the open interval $(0, 1)$. This made the QN data uniformly distributed $U(0, 1)$. The plots below show the normalized QN data.

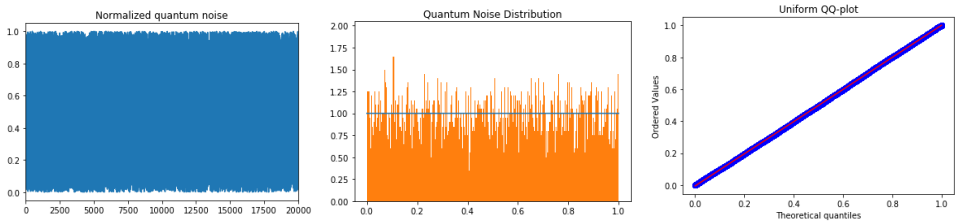


Figure 6.3: (Left) Sample of 20000 QN data points, (middle) distribution of the data, (right) Uniform QQ-plot showing $U(0, 1)$ fit.

It was desirable to have the reference data properties as close to the EURUSD data as possible, therefore an investigation of the properties of EURUSD data was conducted. QQ-plot was performed with normal, Laplace and uniform distribution. The plots of the results are shown below.

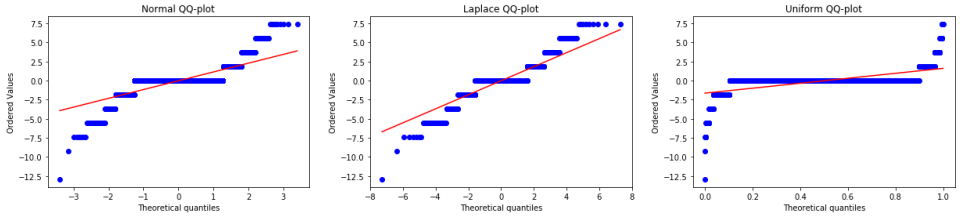


Figure 6.4: Best fitting QQ-plot (Left) Sample of 2000 of EURUSD QQ-plot for $N(0, 1.15) \approx N(0, 1)$, $SSE = 1824.2428$, (middle) QQ-plot for $La(0, 0.92) \approx La(0, 1)$, $SSE = 1124.2149$ (right) and Uniform QQ-plot for $U(-1.6, 1.6)$, $SSE = 8078.1710$.

From the QQ-plot it was evident that the empirical distribution had heavier tails than both the normal and Laplace distribution. The left tail was heavier than the right tail, which indicated that negative drawdowns were more likely than positive gains as extreme events. The results $SSE_{uniform} > SSE_{normal} > SSE_{laplace}$ indicated that Laplace distribution, $La(0, 1)$, was a more suitable distribution than $N(0, 1)$ for standardized EURUSD log-return data.

As the EURUSD data was shown to be $La(0, 1)$, the QN data was used to sample random variables from $La(0, 1)$ distribution. This was done by first sampling $N(0, 1)$ random variables from $U(0, 1)$ distributed QN data by means of inverse transform sampling. Then Laplace random variables was sampled with scaling factor $b = 1$ to obtain $La(0, 1)$ distributed random variables. The Laplace QN data was then standardized to get it to the same order of magnitude as EURUSD data for comparability. The standardized Laplace QN data is shown below.

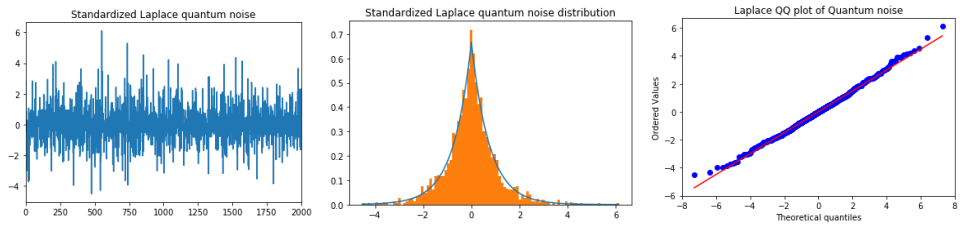


Figure 6.5: 2000 standardized Laplace samples generated with $U(0, 1)$ normalized QN data.

The distribution of the standardized Laplace samples got slightly changed

scaling factor b from the standardization. However, it was $La(0, 0.8) \approx La(0, 1)$-distributed. As the standardized EURUSD data were best fitted with $La(0, 0.92) \approx La(0, 1)$, the standardized Laplace QN were now in both same order of magnitude and from a similar distribution as the standardized EURUSD data.

The EURUSD data had discrete values, therefore quantization was performed on the QN data. The EURUSD data had 77 unique log-returns. Scaling factor $s = 4.22$ was chosen in the quantization procedure $QN_{discrete} = \frac{\|QN \cdot s\|}{s}$, so that the standardized Laplace QN data also had 77 unique values. The resulting data is shown below.

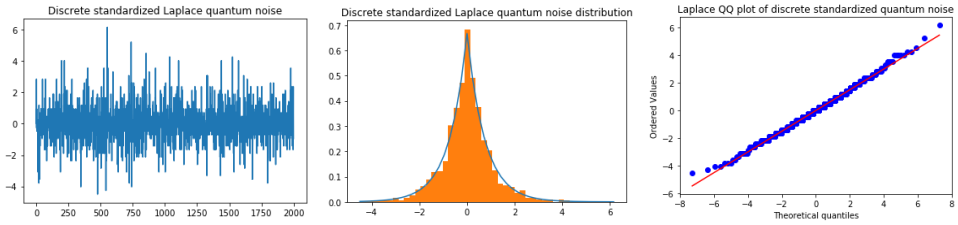


Figure 6.6: 2000 discrete standardized Laplace samples with 77 unique values generated with $U(0, 1)$ normalized QN data.

6.2 Takens Embedding

This section shows the results and motivations for parameter selections in the Takens embedding. The same parameter choice is made for both EURUSD and QN data to make both datasets reconstructed to a state space in a similar manner.

6.2.1 Selection of time delay

The choice of $\tau = 1$ was made based on qualitative properties of the dataset as well as ACF calculations. It is commonly assumed that financial time series follow the Martingale property $E[X_{n+1}|X_t, X_{t-1}, \dots] = X_t$ meaning that it loses memory after $\tau = 1$. To check the validity of the choice $\tau = 1$, autocorrelation function method with the constraints first zero and first $1/e$ decay was also used. The autocorrelation function of 5 randomly selected windows are presented below.

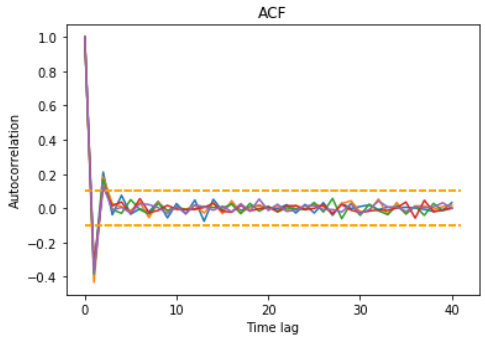


Figure 6.7: ACF plot of 5 different windows with 2000 dp.

Fig 6.7 shows the quick drop-off of ACF below 1 and $1/e$ at time $t = 1$, also suggesting a choice of $\tau = 1$. Fig 6.7 shows only ACF calculations for five windows. However, iterative calculations through all sliding windows show that the ACF behaved roughly the same on all windows. Moreover, Zaldivar et al. have pointed out that τ should not be an integer multiple of a periodicity of the system [118]. As $\tau = 1$ is a multiple integers of all periodic systems, it was important to check that the system was non-periodic. This was done with power spectral density estimation using Welch method. The results for the power spectral density estimation is shown below.

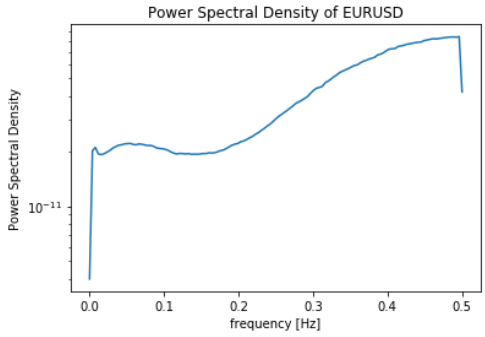


Figure 6.8: Welch estimate of power spectral density of EURUSD data. Spikes at a certain frequency indicates the periodicity $p = \frac{1}{\text{Hz}}$. The Power spectral density shows no spikes.

Fig 6.8 shows no peaks indicating that there is no periodicity in the signal. Therefore, the choice of $\tau = 1$ was supported by both martingale property and ACF method and did not violate the multiple integers of periodicity constraint.

6.2.2 Selection of embedding dimension

The false nearest neighbor computations five random windows of the EURUSD data is shown below.

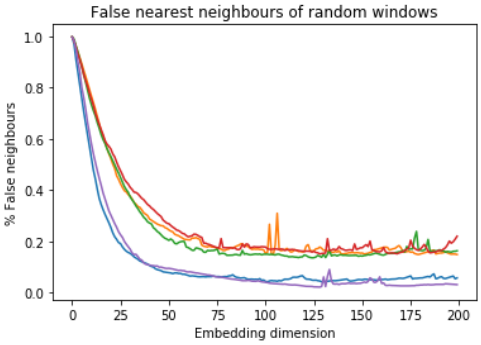


Figure 6.9: False nearest neighbors plot of 5 different samples for 2000 dp.

The result is presented to highlight the drop-off and convergence above zero of the false nearest neighbors. For the actual results sliding windows of the EURUSD with a window size of 2000 data points and a gap size of 200 000 was constructed yielding $\{X_1, X_2, \ldots, X_{41}\}$ window items. The selection of embedding dimension was then based on when the mean of the derivative of the false nearest neighbors lower than an arbitrary set threshold of 0.002 i.e. when an additional embedding dimension makes very little difference to the amount of false nearest neighbors. This embedding dimension was found to be $m = 35$. The summary results for the sliding windows are shown below.

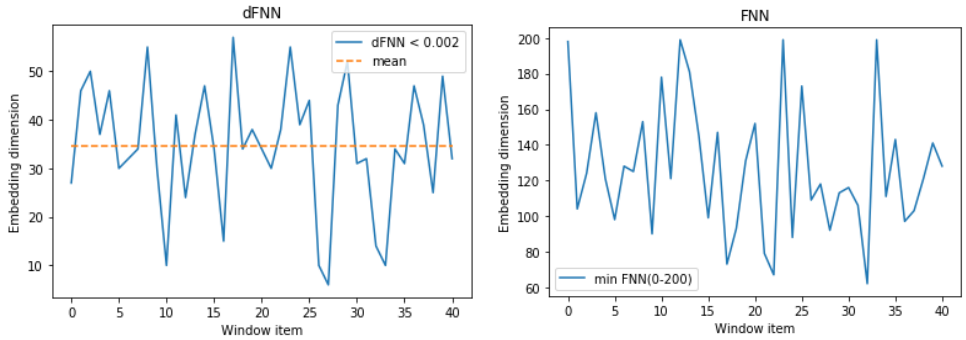


Figure 6.10: (left) derivative of FNN less than 0.002. (right) lowest FNN between 0 and 200 embedding dimensions of sliding window with 2000 dp window size and 200 000 dp gap size. X-axis indicating window item $i = \{1, \ldots, 41\}$.

6.3 Examples of TDA on state space reconstructions

In this section TDA of reconstructed state space examples are shown to provide an understanding of the result summaries. First examples of how the non-PCA Takens embedding looked geometrically is provided. Four windows will be shown; EURUSD data random window, EURUSD window with low complexity, EURUSD window with the high complexity, and QN random window. The gzip-compress-to-ratio and Shannon entropy is provided for each window. Secondly, PCA results of above window are shown. Lastly, persistence diagrams and landscapes of the windows are also provided.

6.3.1 Non-PCA State space reconstruction

The state space reconstruction was constructed using embedding dimension $m = 35$ as it was found to be the dimension where an additional dimension did not add much to reducing false neighbors. As visual inspection of high dimensions is restricted, a 3D plot of the first three embedding dimensions is provided.

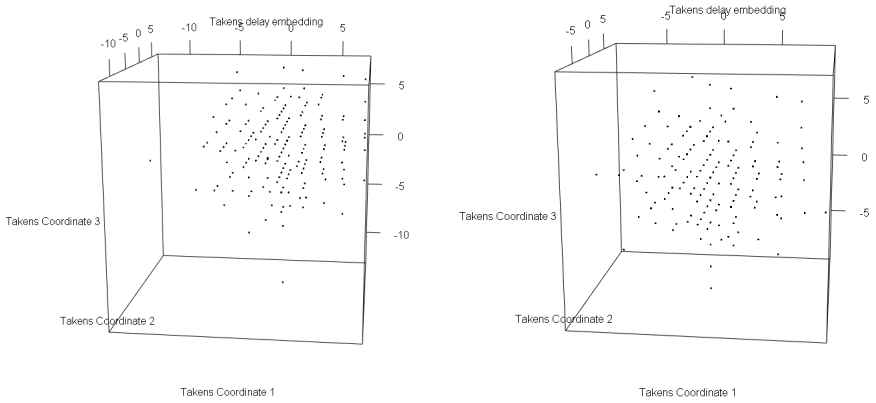


Figure 6.11: Takens embedding with $m = 35$ and $\tau = 1$ (left) EURUSD 2000 dp sample (gzip $G = 0.0637$, Shannon entropy, $S = 1.9800$). (right) EURUSD 2000 dp window of minimum complexity ($G = 0.0103$, $S = 0.0486$).

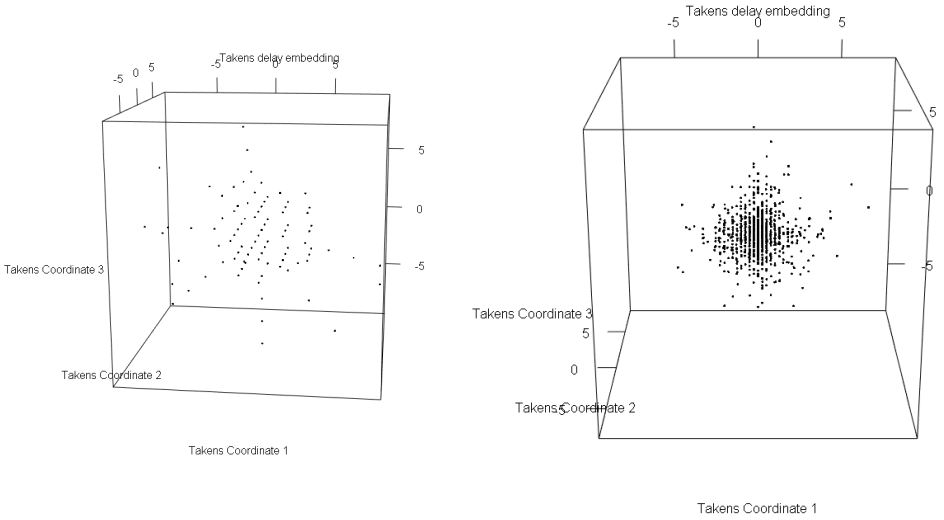


Figure 6.12: Takens embedding with $m = 35$ and $\tau = 1$ (left) EURUSD 2000 dp window of high complexity ($G = 0.1166$ $S = 3.6322$). (right) standardized Laplace QN 2000 dp sample ($G = 0.1508$, $S = 4.0728$).

It is possible to see that the point cloud of the QN data is spanned over a smaller volume (data between $[-6, 6]$)than the EURUSD data (data between $[-8, 8]$), however, it is much denser. The embedding of the windows in fig 6.11 are quite similar. The (left) window show

some points further away from the main point cloud than the (right) window with lowest gzip compress-to-ratio and Shannon entropy. In fig 6.12 The window with highest gzip compress-to-ratio and high entropy have flairs coming out of the main point cloud. The QN data embedding has subtle flairs coming out of the main point cloud.

6.3.2 PCA state space reconstruction

The embedding dimension $m = 35$ was used for state space reconstruction. PCA was used so that $\mathbb{R}^{35} \to \mathbb{R}^3$. The PCA of state space reconstruction is shown below.

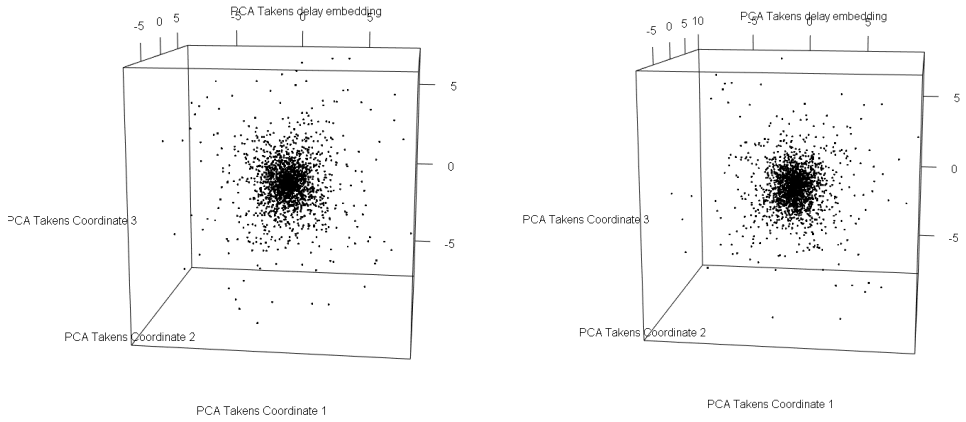


Figure 6.13: PCA Takens embedding with $m = 35$ and $\tau = 1$ (left) EURUSD 2000 dp sample ($G = 0.0637$, $S = 1.9800$). (right) EURUSD 2000 dp window of minimum complexity ($G = 0.0103$, $S = 0.0486$).

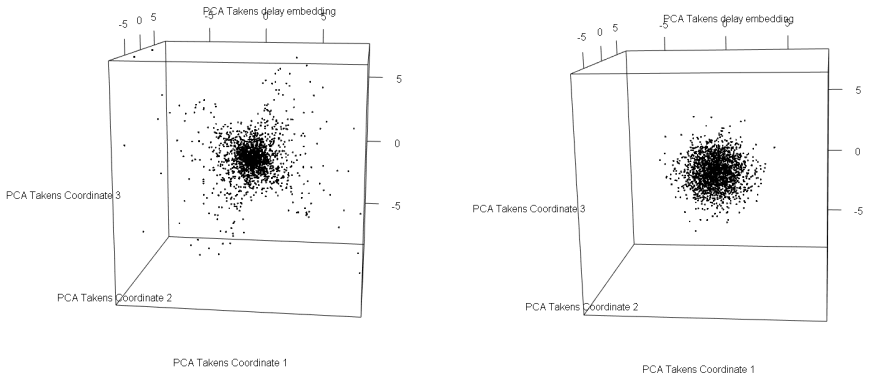


Figure 6.14: PCA Takens embedding with $m = 35$ and $\tau = 1$ (left) EURUSD 2000 dp window of high complexity ($G = 0.1166$, $S = 3.6322$) (right) standardized Laplace QN 2000 dp sample ($G = 0.1508$, $S = 4.0728$).

The EURUSD spans a larger volume than PCA of QN data similar to above non-PCA 3D point clouds. The point clouds in fig 6.13 are quite similar. They have a large point cloud mass in the middle and some sparse points on the outskirts. The EURUSD point cloud in (left) fig 6.14 have much more distinct patterns of points extending towards the outskirts than the point clouds in fig 6.13. The QN point cloud in (right) fig 6.13 is much more concentrated than the other point clouds. Below PCA spree plots are presented.

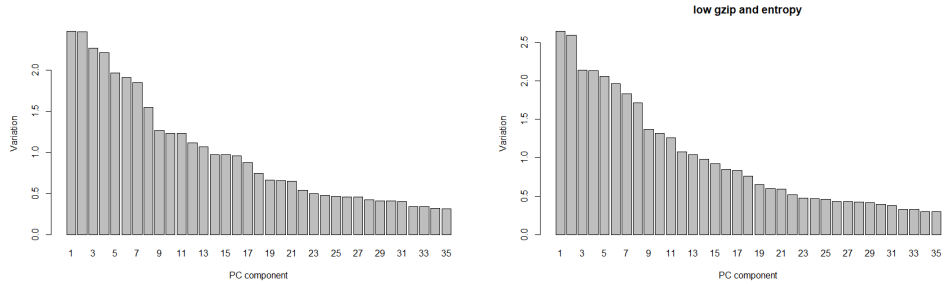


Figure 6.15: PCA scree plot for PCA Takens embedding with $m = 35$ and $\tau = 1$ (left) EURUSD 2000 dp sample ($G = 0.0637$, $S = 1.9800$). (right) EURUSD 2000 dp window of minimum complexity ($G = 0.0103$, $S = 0.0486$).

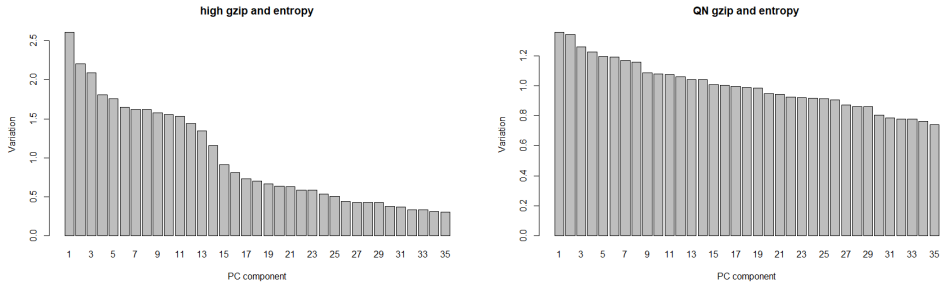


Figure 6.16: PCA scree plot for PCA Takens embedding with $m = 35$ and $\tau = 1$ (left) EURUSD 2000 dp window of high complexity ($G = 0.1166$, $S = 3.6322$) (right) standardized Laplace QN 2000 dp sample ($G = 0.1508$, $S = 4.0728$).

The PCA spree plots for the EURUSD value have a quick drop-off of variation. However, it does also indicate that a significant amount of variation is beyond the three first principal component. The slow drop-off on the QN-data shows that the principal components account for approximately the same amount of variation. As the variations should be quite uniform among the dimensions, random data should be expected to have principal components with approximately equal variation.

6.3.3 Topological Data Analysis

Persistent Homology

Persistent homology results of windows are presented as birth-death diagrams below.

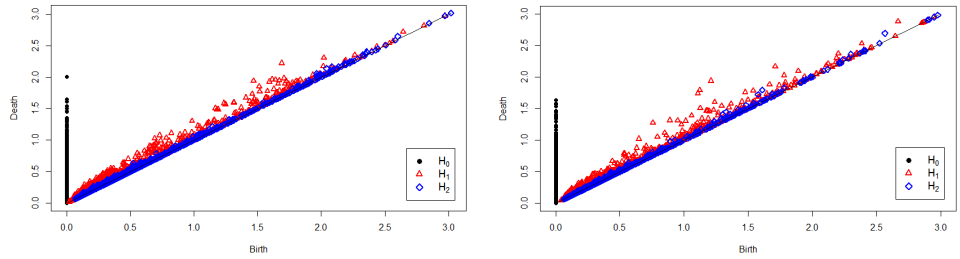


Figure 6.17: Birth-Death diagrams of PCA Takens results (left) EURUSD 2000 dp sample ($G = 0.0637$, $S = 1.9800$). (right) EURUSD 2000 dp window of minimum complexity ($G = 0.0103$, $S = 0.0486$).

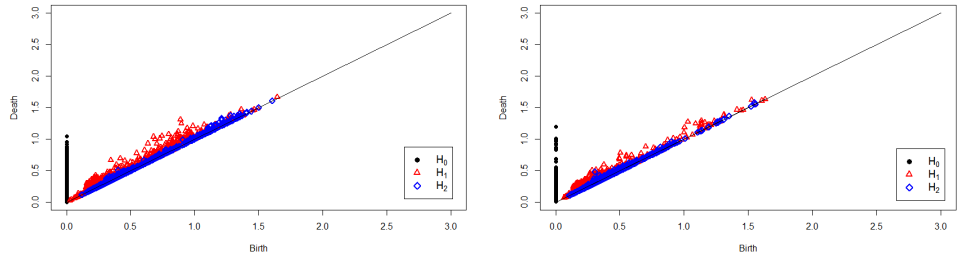


Figure 6.18: Birth-Death diagrams of PCA Takens results (left) EURUSD 2000 dp window of high complexity ($G = 0.1166$, $S = 3.6322$) (right) standardized Laplace QN 2000 dp sample ($G = 0.1508$, $S = 4.0728$).

Fig. 6.18 show that the topological features for the high entropy window are more similar to the QN features than the low entropy and random window.

Persistence Landscape

Persistence landscapes summaries of the persistence diagrams are shown below.

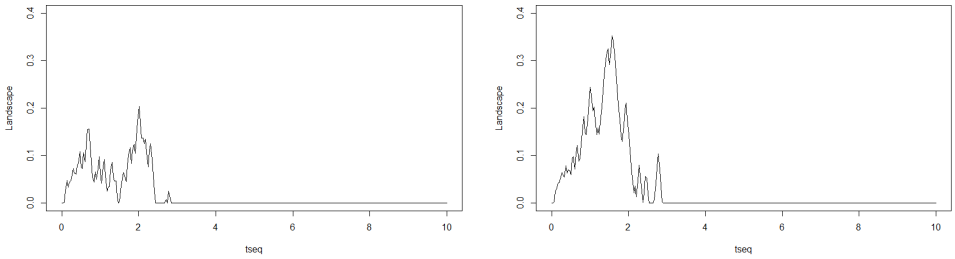


Figure 6.19: Persistence landscapes of H_1 (left) EURUSD 2000 dp sample with integral $I = 8.231$. (right) EURUSD low complexity $I = 10.595$.

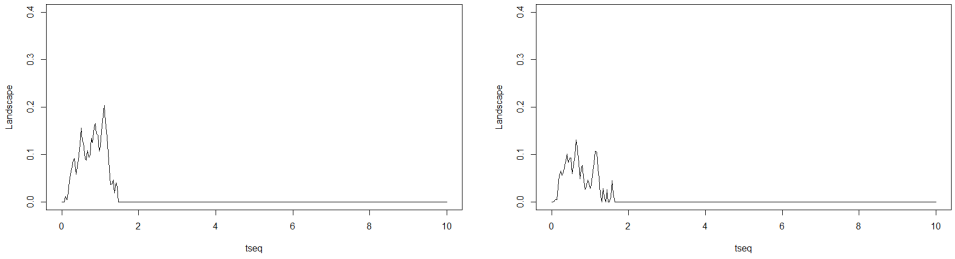


Figure 6.20: Persistence landscapes of H_1 (left) EURUSD high complexity $I = 3.951$. (right) Quantum noise 2000 dp sample $I = 4.434$.

The H_1 landscapes are summaries of the H_1 groups in the persistence diagrams. Below is also one example of noise reduced landscape, to show that H_2 features are mostly noise feature.

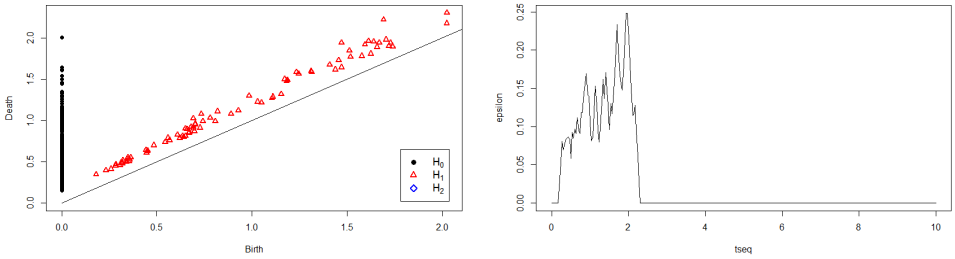


Figure 6.21: $\epsilon = 0.15$ cut-off (left) Birth-death diagram of EURUSD 2000 dp sample (right) corresponding persistence landscape $I = 7.901$.

In this case, $\epsilon = 0.15$ cut-off value is chosen to be the maximum of the persistence of quantum noise diagrams, because any topological features exhibited by the quantum noise are noise features. As seen in fig 6.19 and 6.21 the integral of the persistence landscapes do not significantly differ. Also, comparing the EURUSD birth-death diagram on 6.17 and 6.21 notice the lack of H_2 groups in the ϵ noise adjusted birth-death diagram. However, in the results below it is set to $\epsilon = 0$, so that statistics can be performed without inducing any bias.

6.4 Statistical analysis of Topological features

This section presents the main findings of the thesis. First, a more thorough results description of one window size is presented, then figures for all experimental cases are provided in the next section.

6.4.1 Mean landscapes

Below a mean landscape for sliding window with window size $M = 2000$ and gap size $G = 2000$ is shown.

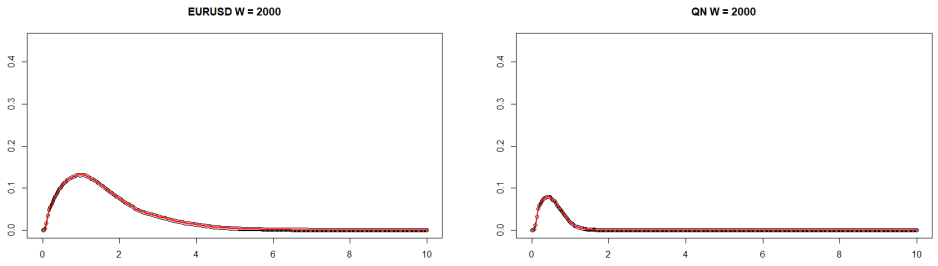


Figure 6.22: (left) mean of H_1 landscape of EURUSD and (right) laplace QN.

Above the mean landscapes for H_1 are shown. It was constructed with 95 % bootstrap confidence band. The mean persistence landscapes showed that EURUSD data have a lot more persistent H_1 than the Laplace QN. This indicated that the EURUSD data have some properties that differ from randomly generated variables.

6.4.2 Persistence and complexity

The persistence landscape integrals and maximum persistence were also generated for the same windows. This method allowed for comparison between persistence of each window. The results are shown below.

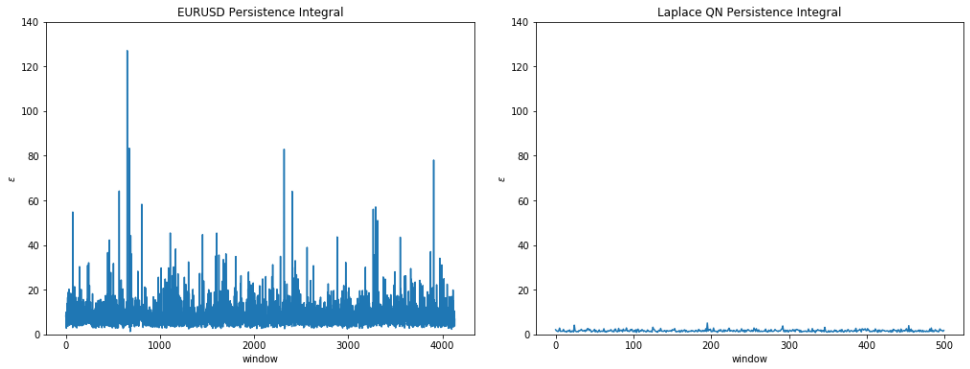


Figure 6.23: (left) EURUSD integrals of persistence landscapes of H_1 and (right) laplace QN. Note that laplace QN only have 500 windows.

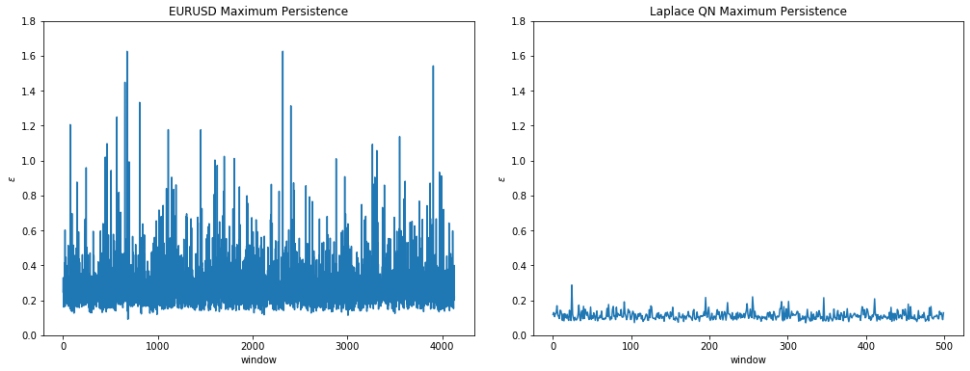


Figure 6.24: (left) EURUSD maximum persistence of H_1 and (right) laplace QN. Note that laplace QN only have 500 windows.

The persistence landscapes integrals and maximum persistence for EURUSD was an order of magnitude larger than the Laplace QN data. The theory is that persistence in topological features indicates the presence of some global property[132]. It was, therefore, relevant to compare these values with the complexity of the corresponding windows. The complexity of each window was therefore calculated.

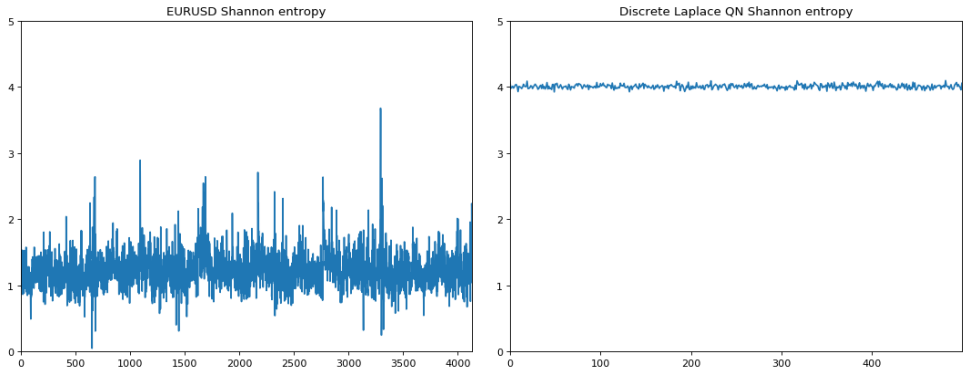


Figure 6.25: (left) Shannon entropy of EURUSD data and (right) laplace QN. Note that laplace QN only have 500 windows.

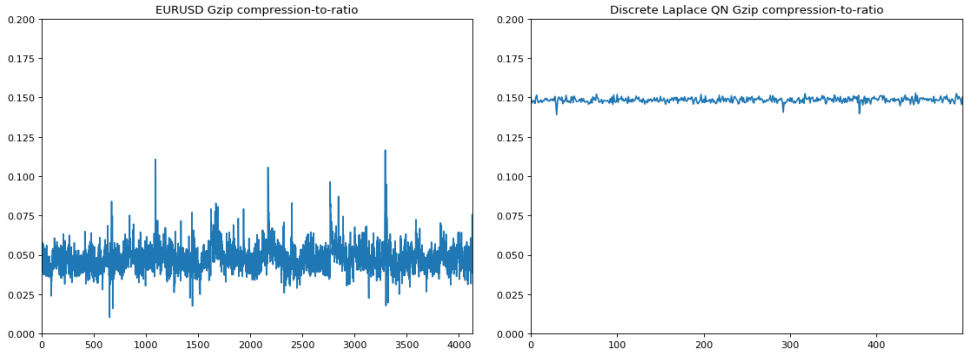


Figure 6.26: (left) Gzip-compress-to-ratio of EURUSD data and (right) laplace QN. Note that laplace QN only have 500 windows.

Fig. 6.22 and 6.24 show that the QN topological features are less persistent than EURUSD indicating the presence of some property in the EURUSD data. At the same time fig. 6.25 and 6.26 show that QN has more entropy than EURUSD indicating that there is more order in EURUSD data than QN. By taking the correlation of the values it was possible to compare the relation among the entropy's and persistence landscape integrals.

	PI	MP	Gzip	Shannon
PI	1			
MP	0.8984	1		
Gzip	-0.0501	-0.368	1	
Shannon	-0.0639	-0.0484	0.9390	1

Table 6.1: Correlation matrix of the persistence landscape integrals (PI), maximum persistence (MP), gzip compress-to-ratio and Shannon entropy of EURUSD.

The correlation matrix shows high correlation between persistence integral and maximum persistence, and between Shannon entropy and Gzip compress-to-ratio. It also shows low correlation between persistence and entropy calculations. This indicated that the persistence in homology groups accounts for other features in the data than entropy.

6.5 Empirical distribution of topological features

Lastly, distributions of persistence integrals were calculated to be able to understand its probabilistic properties.

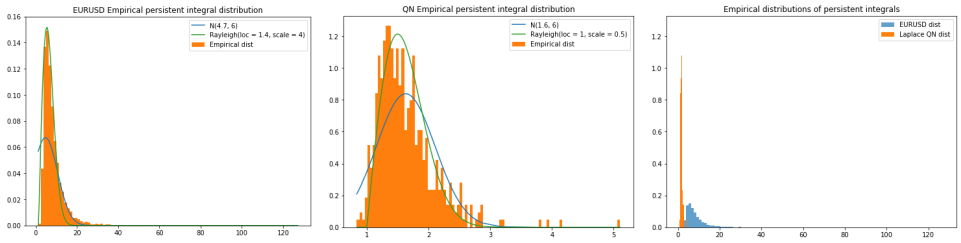


Figure 6.27: (left) EURUSD persistence landscape integral distribution, (middle) Laplace QN distributions and (right) empirical distribution of EURUSD and QN for comparison.

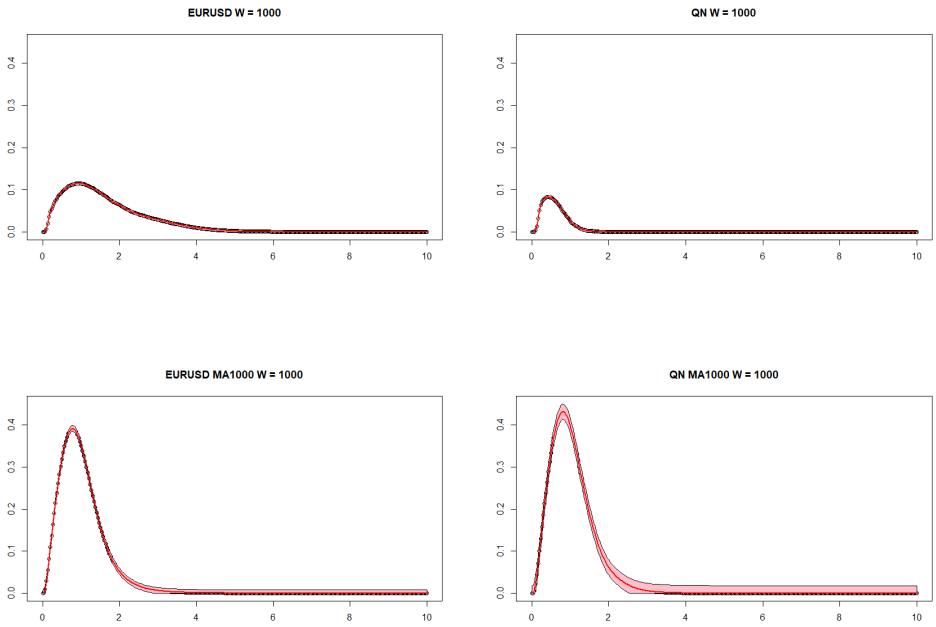
Fig. 6.27 show that the persistence integrals are distributed by some right-skewed distribution. For reference theoretical normal- and Rayleigh

distribution is fitted. The right-skew can be explained by noise features.

6.6 Results from other windows

This section shows the results obtained from performing calculations on window size 1000 and 2000. All gap size used are the same as the window size $G = W$. The gap size was chosen so that there was as much disjunct information as possible. In addition, because theoretical models showed that moving average helps to uncover the underlying topology, moving average with window size 1000 is also calculated for both sliding window sizes. Only a small subset of relevant figures will be presented in this section. A full disclosure of the figures is presented in the appendix 9.1.

6.6.1 Mean Landscapes

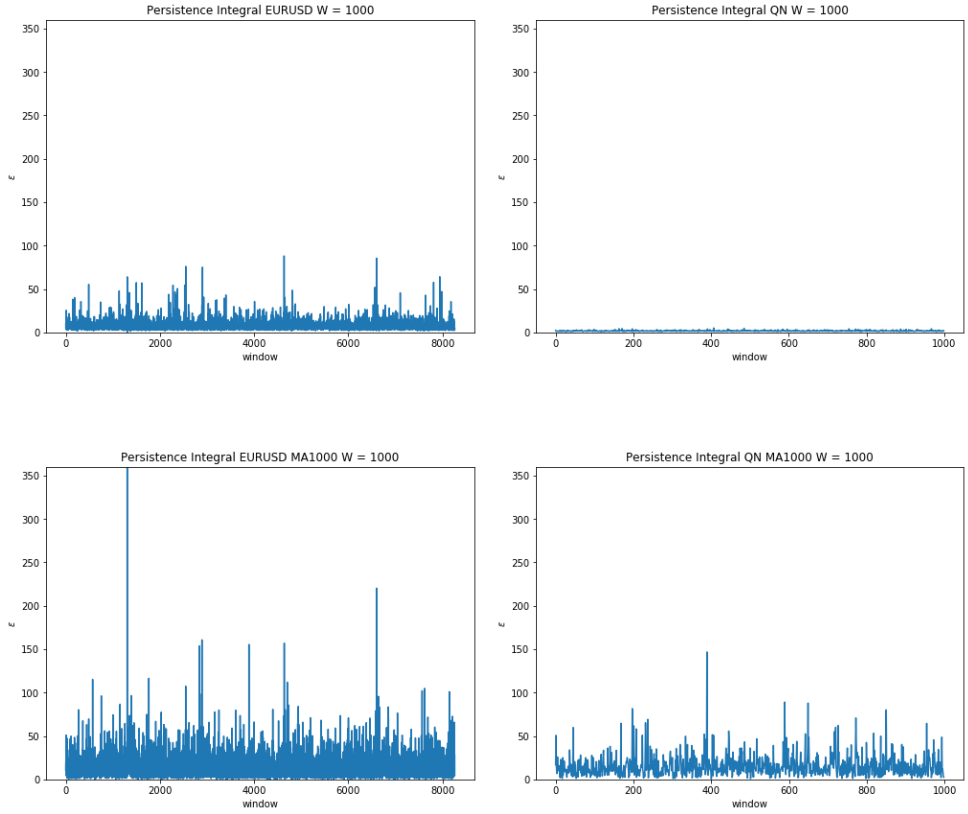


Figures for non-smoothed and smoothed window size $W = 1000$ is presented. For the non-smoothed landscapes, the EURUSD data showed

more persistence in both $W = 1000$ and $W = 2000$ cases. However, when the data was smoothed prior to constructing the landscapes the persistence became very similar.

6.6.2 Persistence integral

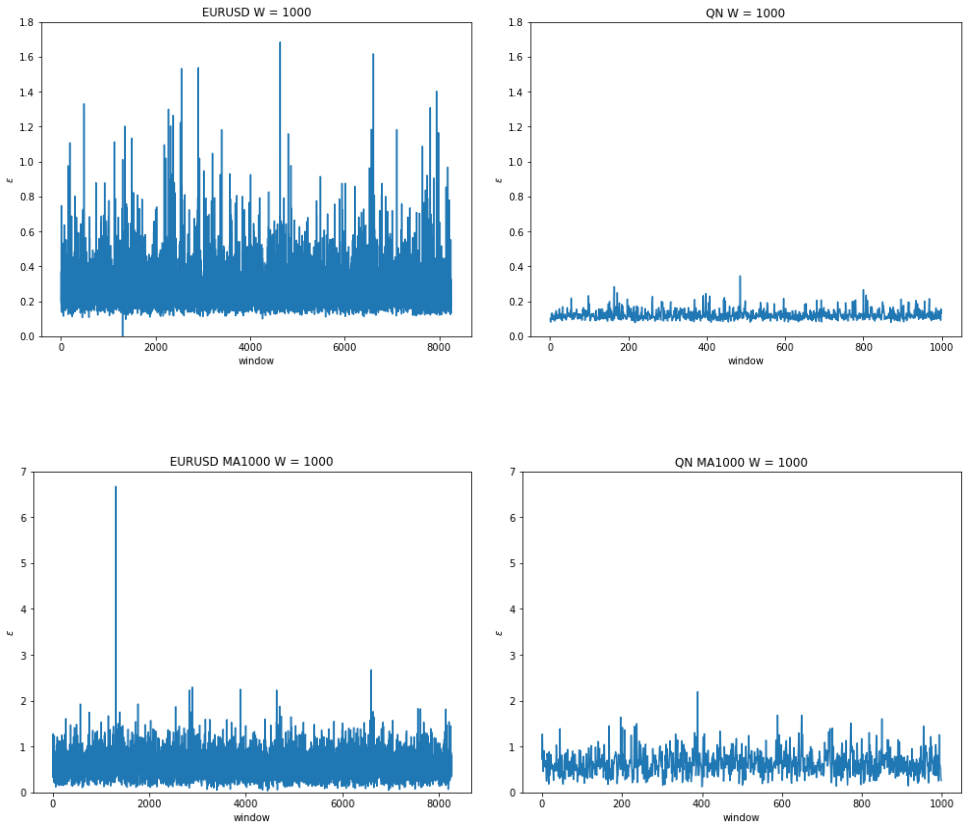
This section shows the calculated persistence integrals.



Figures for non-smoothed and smoothed window size $W = 1000$ is presented. In EURUSD $W_{ma1000} = 1000$ the highest persistence was $\epsilon = 994.4698$, but for scaling and overview reasons the y-axis was kept smaller. The persistence integrals for EURUSD was higher for all windows. For $W_{ma1000} = 1000$ QN had higher persistence integral than the other QN windows and remarkably smoothing the QN data gave higher persistence integral.

6.6.3 Maximum persistence

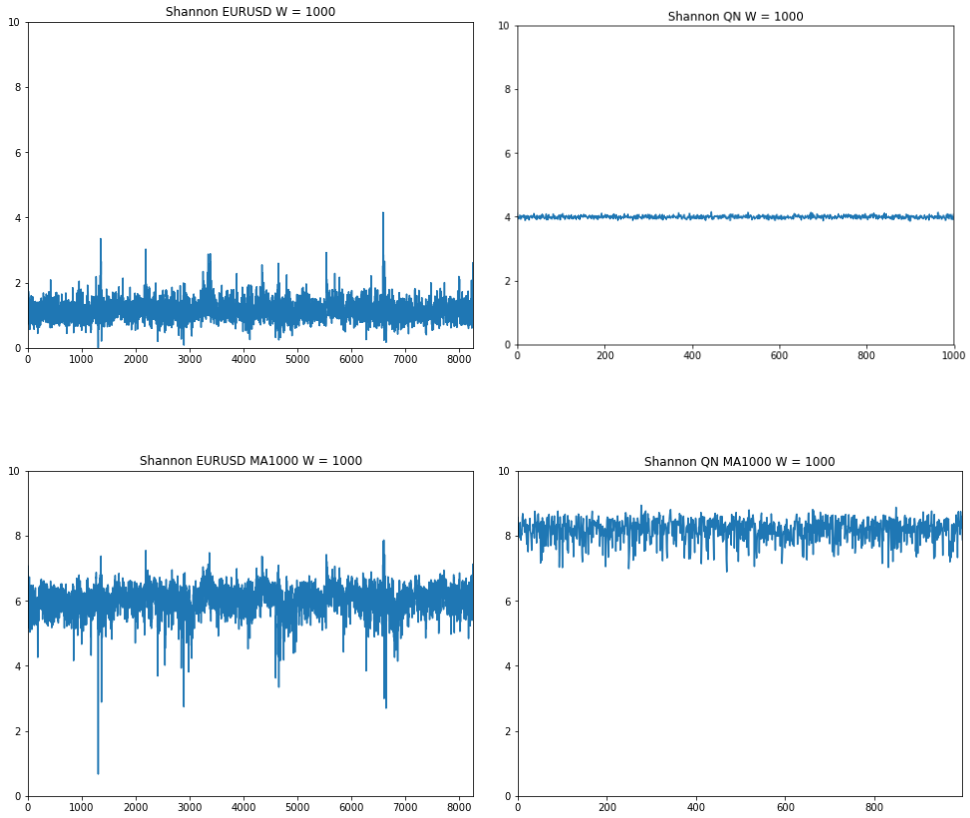
This section shows the calculated maximum persistence.



Figures for non-smoothed and smoothed window size $W = 1000$ is presented. The maximum persistence increased when taking the moving average but the difference in the relationship between EURUSD and QN is preserved. EURUSD has higher maximum persistence for all windows.

6.6.4 Shannon Entropy

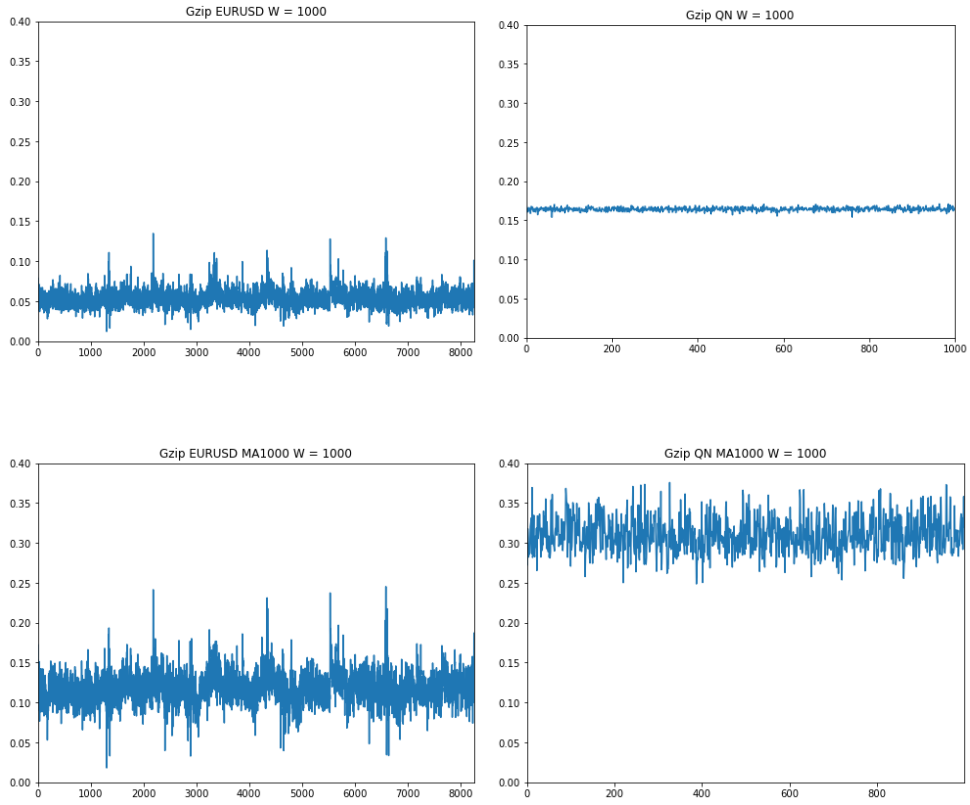
This section shows the calculated Shannon entropy.



Figures for non-smoothed and smoothed window size $W = 1000$ is presented. EURUSD Shannon entropy was lower for all windows. Smoothing the data gave higher entropy for all cases. Smoothing the QN data also made the variance for the entropy increase.

6.6.5 Gzip Compress-to-ratio

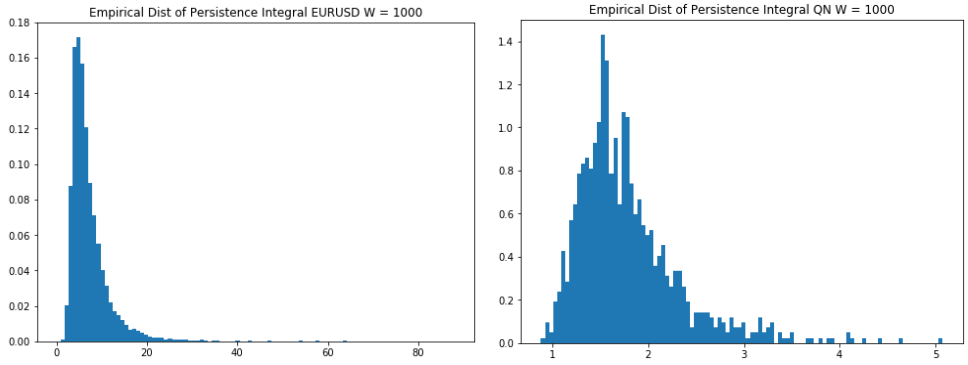
This section shows the calculated Gzip-compress-to-ratios for various window sizes.



Figures for non-smoothed and smoothed window size $W = 1000$ is presented. The gzip-compress-to-ratio had a very high correlation with Shannon entropy. The results for gzip-compress-to-ratio indicate same results as above presented Shannon entropy results.

6.6.6 Empirical Distribution of Persistence Integral

This section shows the calculated empirical distributions of the persistence integrals.



Figures for non-smoothed window size $W = 1000$ is presented. All empirical distributions were right-skewed. The EURUSD data showed distributions with higher kurtosis than QN data.

Chapter 7

Discussion

To the best of our knowledge, the area of applied topological data analysis of financial data is relatively unstudied in the academic community. Contribution in this field has been made by Gidea et al. [81, 82]. The studies by Gidea focus on topological structures prevalent in point clouds representing interrelationships between assets with the purpose to achieve an early indicator of a crash event. This thesis provides a practical investigation of topological data analysis of one-dimensional financial time series. The investigation is conducted by looking at the persistence of H_1 groups in the dimensionality-reduced reconstructed state space of the time series. In essence, the manifold generating the one-dimensional financial time series is embedded in a $\mathbb{R}^{35}$ embedding space. The embedded manifold is then projected to a PCA feature space in $\mathbb{R}^3$. This is done in an attempt to detect some property of a one-dimensional time series rather than attempting to detect regime shifts as Gidea at al. [81, 82].

The procedure used is quite extensive and include many areas which could be further investigated. To use persistent homology on one-dimensional financial time series, it needed to be represented as a point cloud. To do this Takens embedding was used. This means that the persistent homology was essentially used to analyze the dynamical system of the time series. The parameter choices used for Takens embedding, in this case, was motivated using the properties of financial time series as well as quantitative methods commonly associated with them. It must be noted that there is no fixed theorem for how these

parameters should be chosen, therefore the choices made in this thesis should be viewed as heuristics. The choice of point cloud representation of one-dimensional time series impacts what is actually studied with TDA. Therefore the use of any other point cloud representations would also implicate studying other properties of the data. Investigating other point cloud representations of one-dimensional time series could therefore also be interesting.

Another interesting aspect of the process was the impact of dimensionality reduction method on the topological structures. PCA was used to get $\mathbb{R}^m \to \mathbb{R}^3$ for feasible computational time. In the synthetic examples chapter 5 the topological features of the PCA are shown to differ significantly from the topology of the embedded space in higher dimensions. As such persistent homology analyze the topological features of a dimensionality reduced embedding rather than the embedded time series. Different methods of dimensionality reduction are likely to exhibit different topological features as they use different forms of feature extractions.

Furthermore, to our knowledge, the effects of quantization of data on persistent homology is a challenge not yet addressed in the research community. As evident from the synthetic examples chapter 5 quantization in combination with noise can significantly alter the persistence calculation.

The results showed that EURUSD had both lower entropy and higher persistence in topological features than QN data. This suggests that EURUSD data have properties that differ from random noise. The low correlation among the entropy calculations and persistence of topological features suggests that the persistence of H_1 features tells us about a different feature than mutual information. The difference in persistence of H_1 features between EURUSD and QN suggest that there is some additional useful information in the topological features. An investigation of what these topological features actually implies could be useful to further understand this topic.

Interestingly, the maximum persistence and persistence integral is higher for QN when taking the moving average of the datasets, while the other statistics keep the same relations as the non-moving average counterpart. A possible explanation for this is the quantization effects. While the EURUSD data has many equivalent values in succession, the

QN data has considerably more variance in the quantized data. Taking the moving average of these datasets thus makes the EURUSD contain a broader range of values than the QN data.

Another interesting aspect is that persistence integrals exhibit peaks at roughly the same areas irrespective of window size, whereas maximum persistence differs. The maximum persistence has been shown by Khasawne et al. to give an indication of the stability of a stochastic system [13, 83–85]. Low maximum persistence has been shown to indicate stable regions, as the Takens' point cloud gets centered as a mass. This means that high persistence could indicate instability and unboundedness in the time series system, as the point clouds are more scattered. The figures for maximum persistence $W = 1000$ exhibits more tops than $W = 2000$ indicating that systems are more unstable when looking at shorter windows.

Further, from the empirical distribution of the results, it is possible to see that the persistence integral follow similar distributions. The QN has much higher kurtosis, which means that the persistence integrals of the QN data are much more homogenous. When moving average is applied both distributions kurtosis increases, indicating that the persistence integrals are more varying in this case. This effect again can be attributed to broader value range for EURUSD data and narrower value range of QN data.

Chapter 8

Conclusion

In conclusion, this thesis has investigated the use of topological data analysis on one-dimensional time series and shown that TDA might be able to uncover some properties that warrant further research. The process used in this thesis is extensive with many alternation possibilities in parameter choices and sub-method choices. Using this process to analyze one-dimensional time series it can be shown that EURUSD nanosecond data differs from quantum noise data. Also, the fact that the persistence of topological features has low correlation with entropy calculations indicate that topological data analysis manages to uncover other property of non-randomness than mutual information theory.

Lastly, a brief summary of the some of the research question posed in section 1.4, as well as some interesting points are presented:

Is it possible to use topological data analysis to infer knowledge about the financial markets?

Persistent homology can show some property of financial time series. This property differs from mutual information.

What constraints and implications does the processing and pre-processing of the datasets impose?

The process used in this thesis converted the financial time series to a point cloud representing the states of its dynamical system. To do this Takens embedding was used. To make the computational cost feasible PCA was used on the high dimensional embedded time series

yielded by Takens embedding. PCA on high dimensional embedding space essentially means that we are looking at the topology of embedding space on the PCA feature space. For a more topological analysis of the embedding space rather than the PCA feature space other complexes than alpha complex would need to be used.

What are the benefits of topological data analysis for financial markets?

As the persistent homology statistics differs from entropy calculations they could potentially be used as an alternative metric for any other machine learning algorithm, such as clustering.

What are the limitations, pros, and cons of using the data and method suggested?

Pros of the extensive methodology used is that it is computationally feasible even for large datasets. However, cons include that the precise nature of what is explained is not entirely clear.

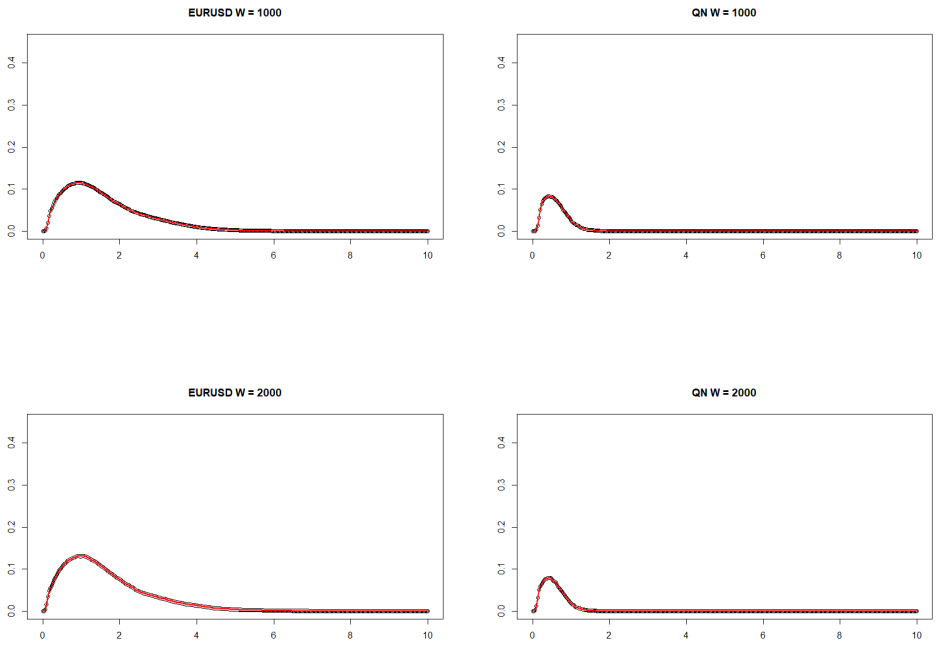
Chapter 9

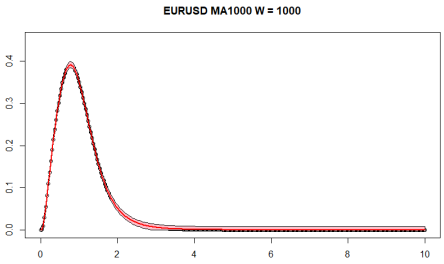
Appendices

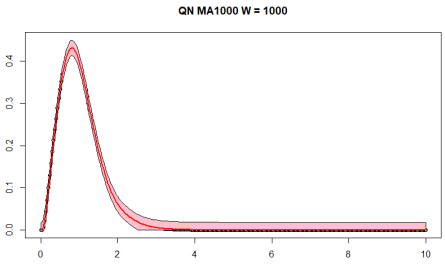
9.1 Results from other windows

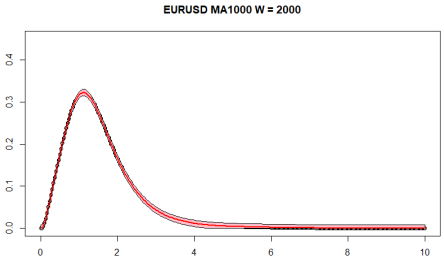
This appendix section will present all the results from section 6.6.

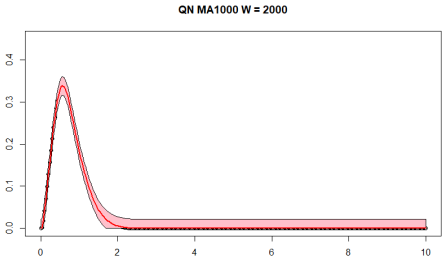
9.1.1 Mean Landscapes



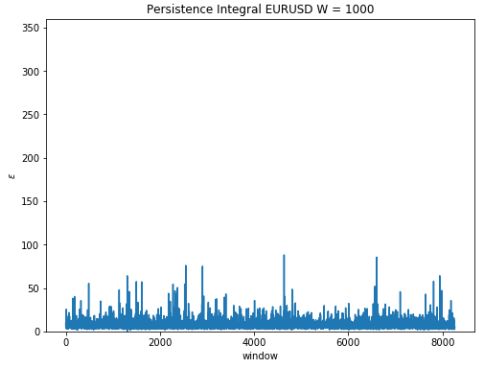

EURUSD MA1000 W = 1000
0.0
0.1
0.2
0.3
0.4
0
2
4
6
8
10

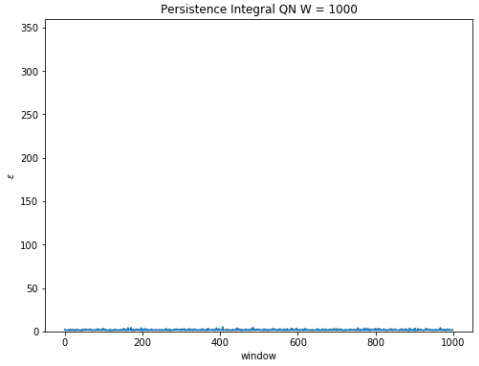

QN MA1000 W = 1000
0.0
0.1
0.2
0.3
0.4
0
2
4
6
8
10

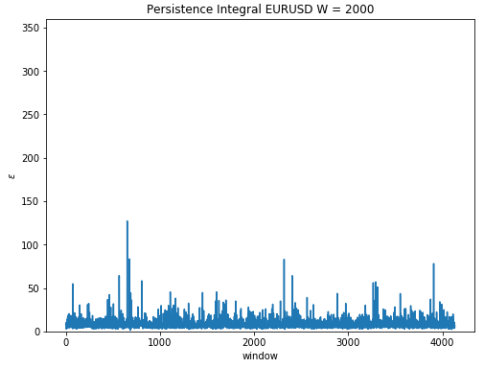

EURUSD MA1000 W = 2000
0.0
0.1
0.2
0.3
0.4
0
2
4
6
8
10

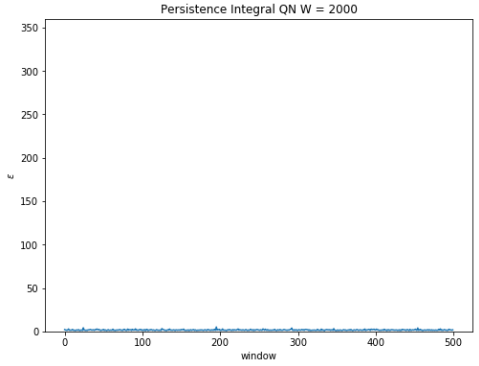

QN MA1000 W = 2000
0.0
0.1
0.2
0.3
0.4
0
2
4
6
8
10

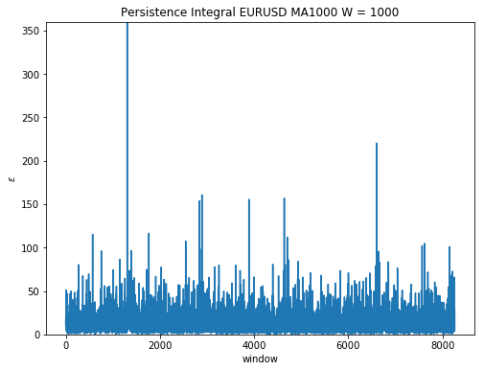
9.1.2 Persistence Integrals

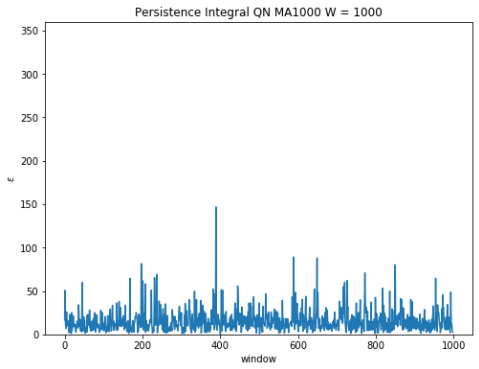


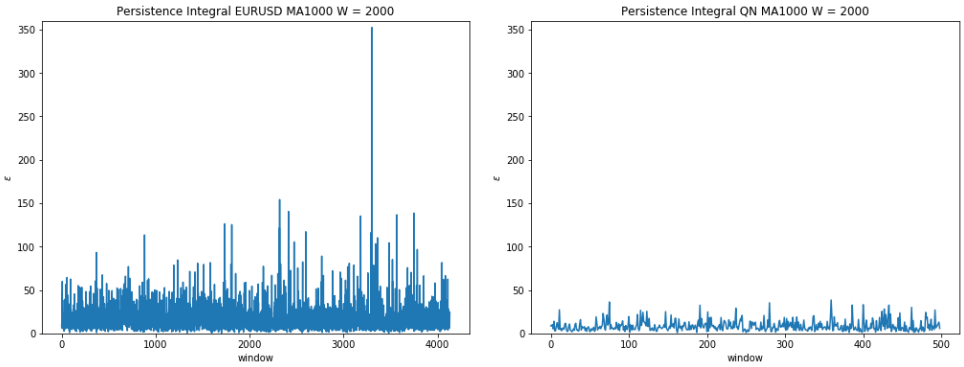




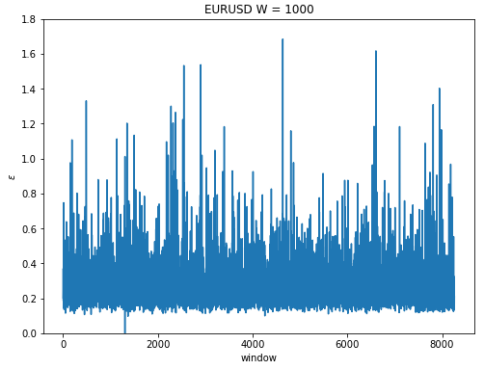


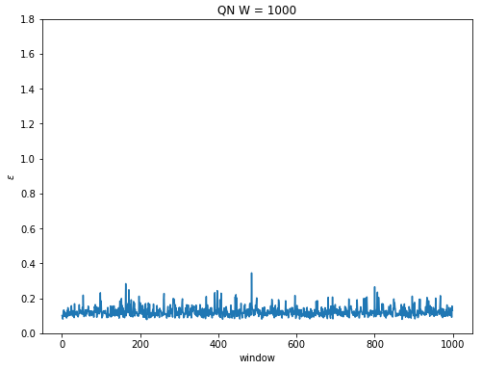


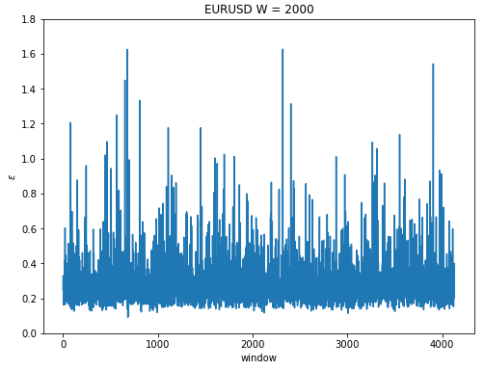


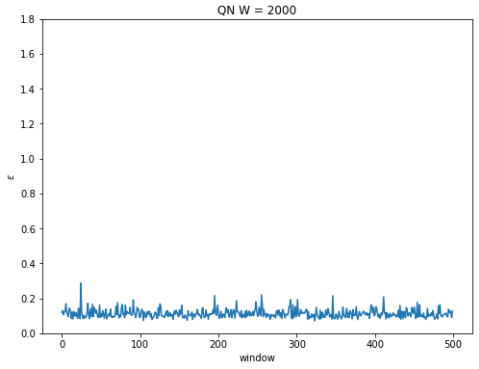

Persistence Integral EURUSD MA1000 W = 2000
ε
window
0
50
100
150
200
250
300
350
0
1000
2000
3000
4000
Persistence Integral QN MA1000 W = 2000
ε
window
0
50
100
150
200
250
300
350
0
100
200
300
400
500

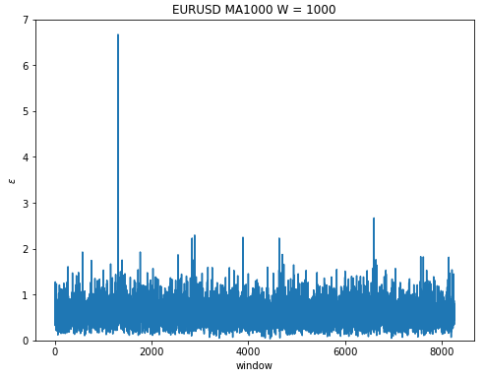
9.1.3 Maximum persistence

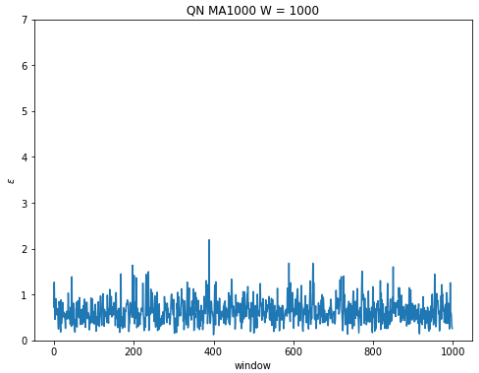


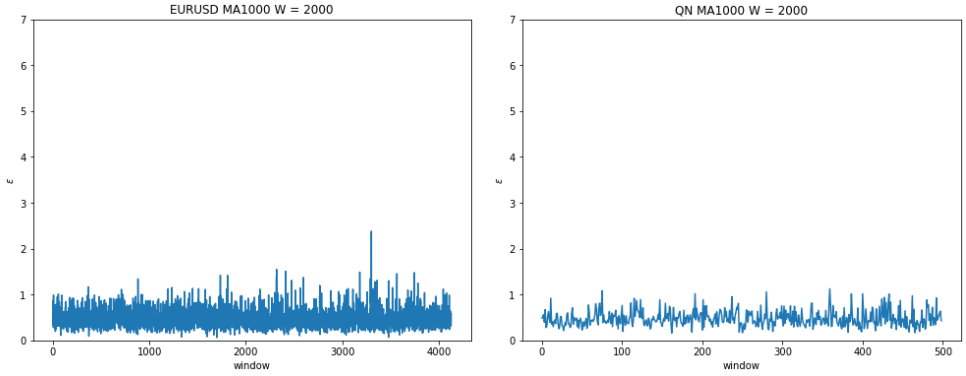




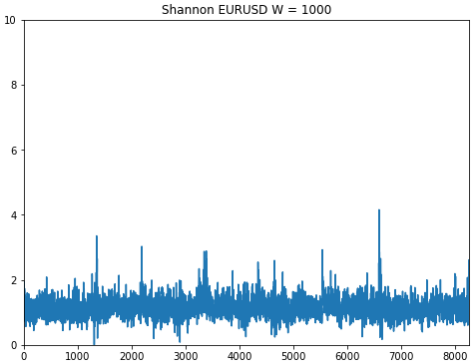


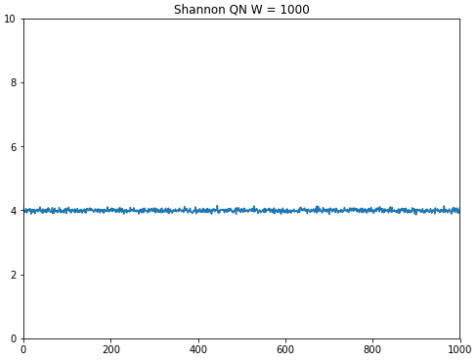


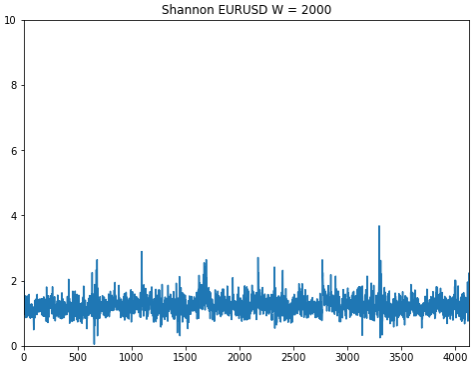


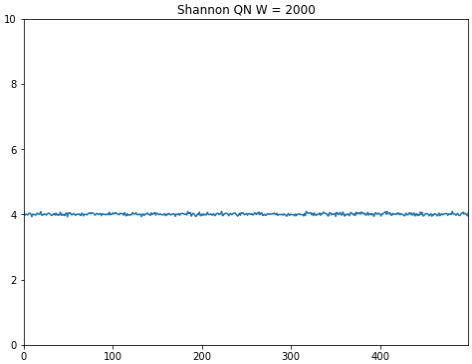

EURUSD MA1000 W = 2000
ε
window
0
1000
2000
3000
4000
QN MA1000 W = 2000
ε
window
0
100
200
300
400
500

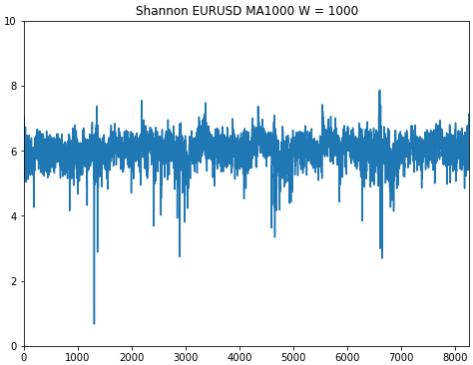
9.1.4 Shannon Entropy

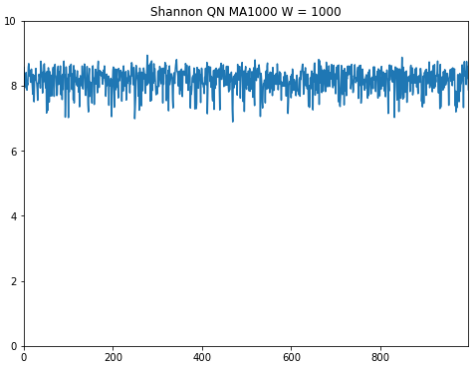


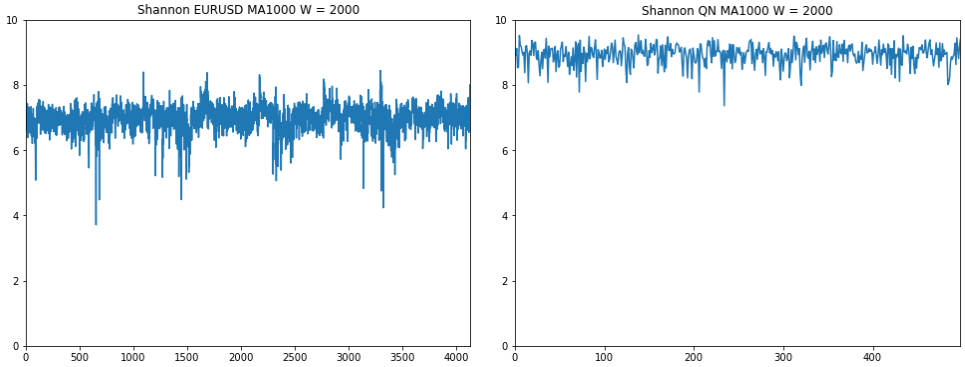




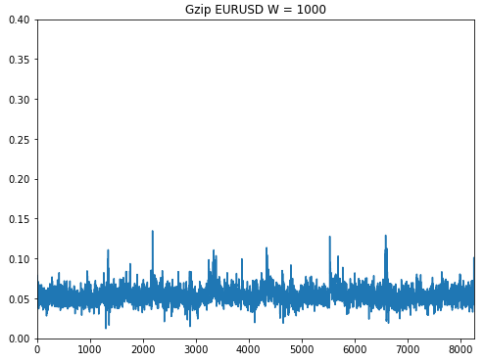


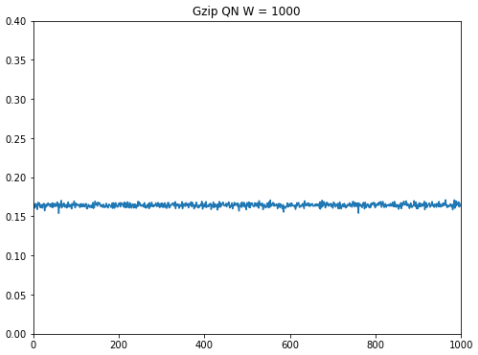


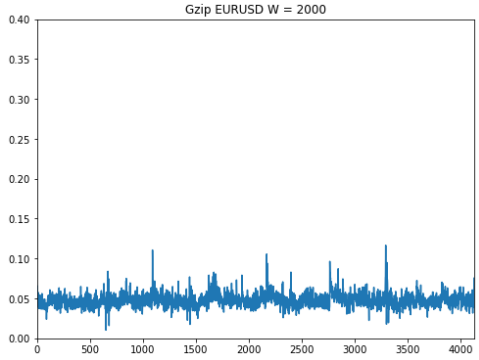


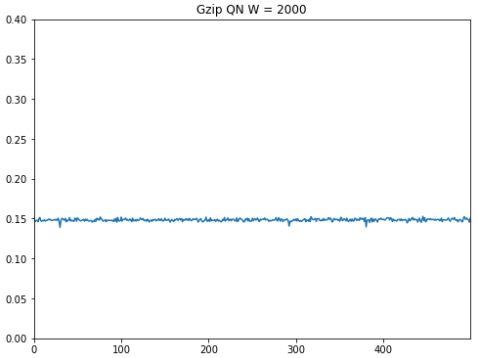

Shannon EURUSD MA1000 W = 2000
Shannon QN MA1000 W = 2000

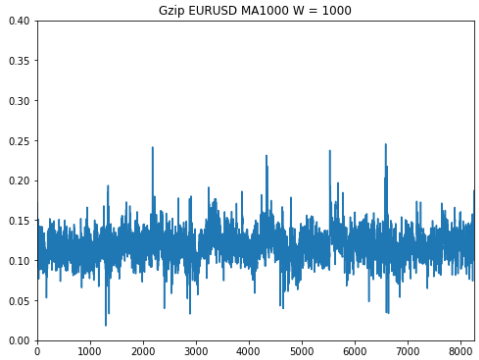
9.1.5 Gzip Compress-to-ratio

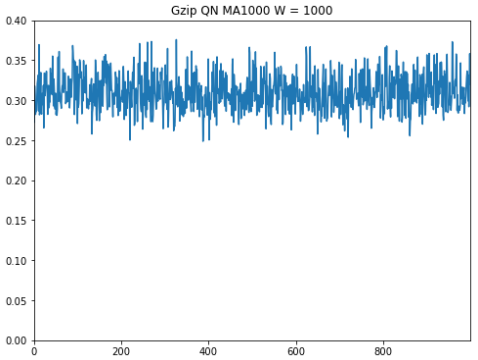


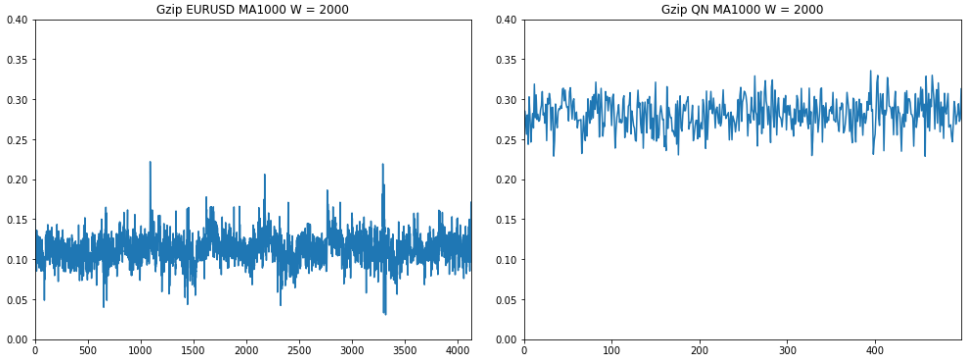




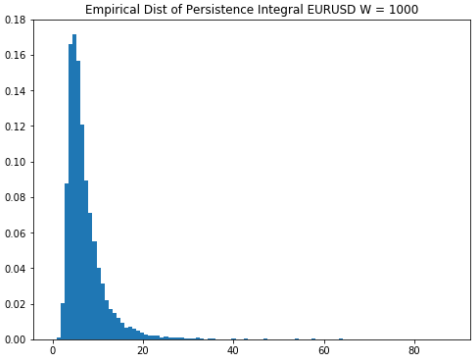


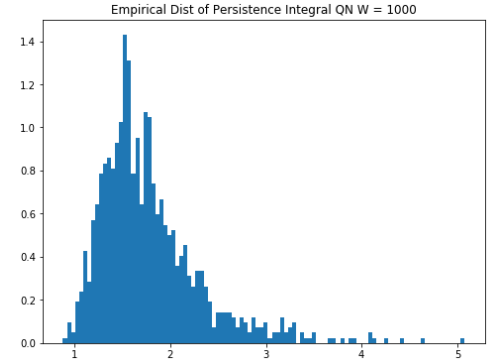


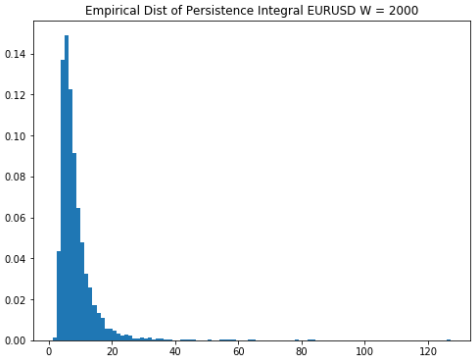


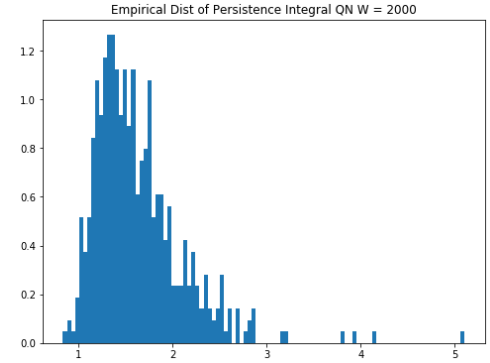

Gzip EURUSD MA1000 W = 2000
Gzip QN MA1000 W = 2000

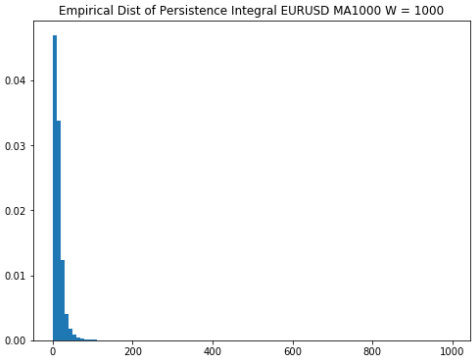
9.1.6 Empirical Distribution of Persistence Integral

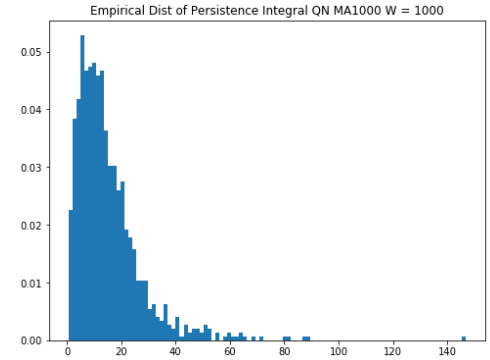


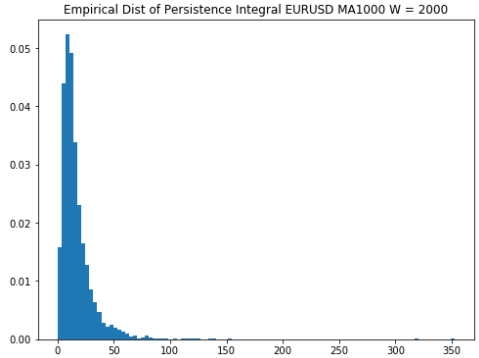


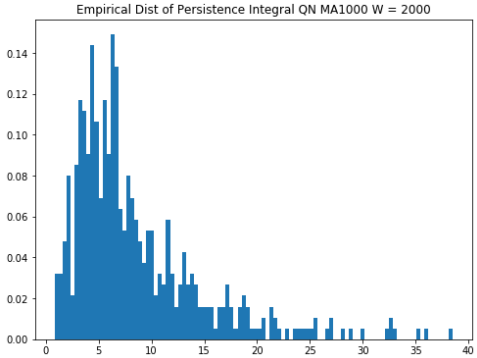









Empirical Dist of Persistence Integral EURUSD MA1000 W = 2000
0.05
0.04
0.03
0.02
0.01
0.00
0
50
100
150
200
250
300
350


Empirical Dist of Persistence Integral QN MA1000 W = 2000
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
5
10
15
20
25
30
35
40

Bibliography

[1] Munch E. A User's Guide to Topological Data Analysis. *Journal of Learning Analytics*, 4(2):47–61, 2017.

[2] Edelsbrunner H. Letcher D. and Zomorodian A. Topological Persistence and Simplification. *Discrete & Computational Geometry*, 28(4):511–533, 2002.

[3] Zomorodian A. and Carlsson G. Computing Persistent Homology. *Discrete Computational Geometry*, 33(2):249–274, 2005.

[4] Carlsson G. Topology and Data. *Bulletin of the American Mathematical Society*, 46:255–308, 2009.

[5] Ferri M. Persistent Topology for Natural Data Analysis - A Survey. *arXiv*, pages 1–16, 2017.

[6] Goldfarb D. An Application of Topological Data Analysis to Hockey Analytics. *ArXiv*, pages 1–19, 2014.

[7] Pereda M. Battiston F. Patania A. Poledna S. Hedblom D. Oztan B. T. Herzog A. John P. Gurciullo S., Smallegan M. and Mikhaylov S. Complex Politics: A Quantitative Semantic and Topological Analysis of UK House of Commons Debates. *ArXiv*, pages 1–24, 2015.

[8] Offroy M. and Duponchel L. Topological Data Analysis: A Promising Big Data Exploration Tool in Biology, Analytical Chemistry and Physical Chemistry. *Analytica Chimica Acta*, 910:1–11, 2016.

[9] Levine A. J. Nicolau M. and Carlsson G. Topology based Data Analysis Identifies a Subgroup of Breast Cancers with a

Unique Mutational Profile and Excellent Survival. *Proceedings of the National Academy of Sciences of the United States of America*, 108(17):7275–7270, 2011.

[10] Berwald J. and Gidea M. Critical Transitions in a Model of a Genetic Regulatory System. *Mathematical Biology and Engineering*, 11:723, 2014.

[11] Gidea M. Berwald J. and Vejdemo-Johansson M. Automatic Recognition and Tagging of Topologically different regimes in Dynamical Systems. *Discontinuity, Nonlinearity and Complexity*, 3:413, 2015.

[12] Perea J. A. and Harer J. Sliding Windows and Persistence: An Application of Topological Methods to Signal Analysis. *Foundation of Computational Mathematics*, 15(3):799–838, 2015.

[13] Khasawneh F. A. and Munch E. Chatter Detection in Turning using Persisten Homology. *Mechanical Systems and Signal Processing*, 70-71:527–541, 2016.

[14] Pereira C. M. M. and de Mello R. F. Persistent Homology for Time Series and spatial Data Clustering. *Expert Systems with Applications*, 42:6062–6038, 2015.

[15] Rucco M. Marco P. and Merelli E. Topological Classifier for Detecting the Emergence of Epileptic Seizures. *ArXiv*, pages 1–21, 2016.

[16] Davis S. Seversky L.M. and Berger M. On Time-Series Topological Data Analysis: New Data and Opportunities. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 59:1–9, 2016.

[17] Gentimis T. Ermani S. and Krim H. Persisten Homology of Delay Embeddings and its Application to Wheeze Detection. *IEEE Signal Processing Letters*, 21(4):459–463, 2014.

[18] Haase S.B. Perea J., Deckard A. and Harer J. SW1PerS: Sliding Windows and 1-Persistence Scoring; Discovering Periodicity in Gene Expression Time Series Data. *BMC Bioinformatics*, 16, 2014.

[19] Rajković M. Maletić S., Zhao Y. Persistent Topological Features of Dynamical Systems. *Chaos*, 26(5):1–15, 2016.

[20] J.-D. Glisse M. Maria C., Boissonnat and M. Yvinec. The GUDHI library: Simplicial Complexes and Persistent Homology. *International Congress on Mathematical Software*, pages 167–174, 2014.

[21] Kerber M. Bauer U. and Reininghaus J. Clear and Compress: Computing Persistent Homology in Chunks. *ArXiv:1303:0477v1*, 2013.

[22] Reininghaus J. Bauer U., Kerber M. and Wagner H. Phat - Persistent Homology Algorithms Toolbox. *Journal of Symbolic Computation*, 78:76–90, 2017.

[23] Lecci F. Fasy B. T., Kim J. and Maria C. Introduction to the R package TDA. *ArXiv:1411.1830*, 2015.

[24] Chikazawa T. The Prediction of Stock Price with Regression Analysis and Its Application to Investment. *Journal of the Japanese Society of Computational Statistics*, 6(2):65–70, 1993.

[25] Siew and Nordin. Regression Techniques for the Prediction of Stock Price Trend. *International Conference on Statistics in Science, Business and Engineering*, pages 1–5, 2012.

[26] Grauer M. Enke D. and Mehdiyev N. Stock Market Prediction with Multiple Regression, Fuzzy Type-2 Clustering and Neural Networks. *Procedia Computer Science*, 6:201–206, 2011.

[27] Gong J. and Sun S. A New Approach of Stock Price Trend Prediction Based on Logistic Regression Model. *International Conference on New Trends in Information Service Science*, pages 1–6, 2009.

[28] Kayakutlu G. Guresen E. and Daim T. U. Using Artificial Neural Network models in Stock Market Index Prediction. *Expert Systems with Applications*, 38:10389–10397, 2011.

[29] Boyacioglu M. A. Kara Y. and Baykan Ö. K. Predicting Direction of Stock Price Index Movement using Artificial Neural Networks and Support Vector Machines: The Sample of the Istanbul Stock Exchange. *Expert Systems with Applications*, 38:5311–5319, 2011.

[30] Shavandi H. Hadavandi E. and Ghanbari A. Integration of Genetic Fuzzy Systems and Artificial Neural Networks for Stock Price Forecasting. *Knowledge-Based Systems*, 23:800–808, 2010.

[31] Zhang Z-G. Wang J-Z., Wang J-J and Guo S-P. Forecasting Stock Indices with Back Propagation Neural Network. *Expert Systems with Applications*, 38:14346–14355, 2011.

[32] Ticknor J. L. A Bayesian Regularized Artificial Neural Network for Stock Market Forecasting. *Expert Systems with Applications*, 40:5501–5506, 2013.

[33] Iacomin R. Stock Market Prediction. *19th International Conference on System Theory, Control and Computing (ICSTCC), October 14-16*, pages 1–6, 2015.

[34] Cakra Y. E. and Trisedya B. Stock Price Prediction using Linear Regression based on Sentiment Analysis. *International Conference on Advanced Computer Science and Information Systems*, pages 147–154, 2015.

[35] Cakra Y. E. and Trisedya B. Sentiment Analysis on Social Media for Stock Movement Prediction. *Expert Systems with Applications*, 42:9603–9611, 2015.

[36] Mackinnon R. Leung C. and Wang Y. A Machine Learning approach for Stock Market Prediction. *Proceedings of the 18th International Database Engineering Applications Symposium*, pages 274–277, 2014.

[37] Leung C. and Mackinnon R. Stock Price Prediction in Undirected Graphs Using a Structural Support Vector Machine. *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 1:548–555, 2015.

[38] Phayung M. and Islam R. Predicting Stock Market Price using Support Vector Regression. *2013 International Conference on Informatics, Electronics and Vision*, pages 1–6, 2013.

[39] Liu Y. Xia Y. and Chen Z. Support Vector Regression for Prediction of Stock Trend. *2013 6th International Conference on Information Management, Innovation Management and Industrial Engineering*, 2:123–126, 2013.

[40] Han I. Kim M. J. and Lee K. C. Hybrid Knowledge Integration using the Fuzzy Genetic Algorithm: Prediction of the Korea Stock Price Index. *Intelligent Systems in Accounting, Finance and Management*, 12:43–60, 2004.

[41] Shiba T. and Takeji Y. Asset Price Prediction using Seasonal Decomposition. *Financial Engineering and the Japanese Markets*, 1:37–53, 1994.

[42] Wichaidit S. and Kittitornkun S. Predicting SET50 Stock Price using Carima (Cross Correlation ARIMA). *2015 International Computer Science and Engineering Conference*, pages 1–4, 2015.

[43] Kom S. Wijaya Y. B. and Napitupulu T. A. Stock Price Prediction: Comparison of ARIMA and Artificial Neural Network Methods - An Indonesia Stock's Case. *2010 Second International Conference on Advances in Computing, Control and Telecommunication*, pages 176–179, 2010.

[44] Chen T-L. Cheng C-H. and Wei L-Y. A Hybrid Model based on Rough Sets Theory and Genetic Algorithms for Stock Price Forecasting. *Information Sciences*, 180:1610–1629, 2010.

[45] Zhang Z-G. Wang J-J., Wang J-Z. and Guo S-P. Stock Index Forecasting based on a Hybrid Model. *Omega*, 40:758–766, 2012.

[46] Chai Y. Gao T., Li X. and Tang Y. Deep Learning with Stock Indicators and Two-Dimensional Principal Component Analysis for Closing Price Prediction System. *2016 7th IEEE International Conference on Software Engineering and Service Science*, pages 166–169, 2016.

[47] Dagli C.H. Li H. and Enke D. Forecasting series-based Stock Price data using Direct Reinforcement Learning. *2004 IEEE International Joint Conference on Neural Networks*, 2:1103–1108, 2004.

[48] Lee J. W. Stock Price Prediction using Reinforcement Learning. *ISIE*, pages 690–695, 2001.

[49] Lee J. W. Stock Price Prediction using Reinforcement Learning. *ISIE*, pages 690–695, 2001.

[50] Sapankevych N. I. and Sankar R. Time Series Prediction using Support Vector Machine - A Survey. *IEEE Computational Intelligence Magazine*, pages 24–38, 2009.

[51] Singh P. A Brief Review of Modeling Approaches based on Fuzzy Time Series. *International Journal of Machine Learning and Cybernetics*, 1(1):1–24, 2015.

[52] Wang R. Stock Selection based on Data Clustering Method. *2011 7th International Conference on Computational Intelligence and Security*, pages 1542–1545, 2011.

[53] De Luca G. and Zuccolotto P. A Tail Dependence-based Dissimilarity measure for Financial Time Series Clustering. *Advances in Data Analysis and Classification*, 5(4):323–340, 2011.

[54] De Luca G. and Zuccolotto P. Dynamic Tail Dependence Clustering of Financial Time Series. *Statistical Papers*, pages 1–17, 2015.

[55] Durante F. Pappadà R. and Torelli N. Clustering of Financial Time Series in Risky Scenarios. *Advances in Data Analysis and Classification*, 8(4):359–376, 2014.

[56] Bastos J. A. and Caiado J. Clustering Financial Time Series with Variance Ratio Statistics. *Quantitative Finance*, 14(12):2121–2133, 2014.

[57] Gao Z. and Yang J. J. Financial Time Series Forecasting with Grouped Predictors using Hierarchical Clustering and Support Vector Regression. *International Journal of Grid Distribution Computing*, 7(5):53–64, 2014.

[58] Dose C. and Cincotti S. Clustering of Financial Time Series with Application to Index and Enhanced Index Tracking Portfolio. *Physica A*, 355:145–151, 2005.

[59] De Carlo F. Facchi P. Pantaleo E. Basalto N., Bellotti R. and Pascazio S. Hausdorff Clustering of Financial Time Series. *Physica A*, 379:635–644, 2007.

[60] Paterlini S. Pattarin F. and Minerva T. Clustering Financial Time Series: an Application to Mutual Funds Style Analysis. *Computational Statistics & Data Analysis*, 47:353–372, 2004.

[61] Saâdaoui F. A Probabilistic Clustering Method for US Interest Rate Analysis. *Quantitative Finance*, 12(1):135–148, 2012.

[62] Li Z. and Tian M. A New Method for Dynamic Stock Clustering Based on Spectral Analysis. *Computational Economics*, pages 1–20, 2016.

[63] Bentes S. R. Forecasting Volatility in Gold Returns under the GARCH, IGARCH and FIGARCH frameworks: New Evidence. *Physica A*, 438:355–364, 2015.

[64] Molnar P. High-low Range in GARCH models of Stock Return Volatility. *Applied Economics*, 48(51):4977–4991, 2016.

[65] Sharma P. and Vipul. Forecasting Stock Market Volatility using Realized GARCH model: International Evidence. *The Quarterly Review of Economics and Finance*, 59:222–230, 2016.

[66] Pinto J. C. Curto J. D. and Tavares G. N. Modeling Stock Markets' Volatility using GARCH models with Normal, Student's t and Stable Paretian Distribution. *Statistical Papers*, 50(2):311–321, 2009.

[67] Awartani B. M. A. and Corradi V. Predicting the Volatility of the S&P-500 Stock Index via GARCH models: the Role of Asymmetries. *International Journal of Forecasting*, 21:167–183, 2005.

[68] Krzemienowski A. and Szymcszyk S. Portfolio Optimization with a Copula-based extension of Condidtional Value-at-Risk. *Annals of Operations Research*, 237(1):219–236, 2016.

[69] Janke O. and Li Q. Portfolio Optimization under Shortfall Risk Constraint. *A Journal of Mathematical Programming and Operations Research*, 65(9):1733–1755, 2016.

[70] Urosevic B Rankovic V., Drenovak M. and Jelic R. Mean-Univariate GARCH VaR Portfolio Optimization: Actual Portfolio Approach. *Computers & Operations Research*, 72:83–92, 2016.

[71] Fulga C. Portfolio Optimization with Disutility-based Risk Measure. *European Journal of Operational Research*, 251:541–533, 2016.

[72] Fulga C. Portfolio Optimization under Loss Aversion. *European Journal of Operational Research*, 251:310–322, 2016.

[73] Cong F. and Oosterlee C.W. Multi-period Mean-Variance Portfolio Optimization based on Monte-Carlo Simulation. *Journal of Economics Dynamics & Control*, 64:23–38, 2016.

[74] Dillo M. J. and Tangman D. Y. A High-order Finite Difference method for Options Valuation. *Computers and Mathematics with Applications*, 2017.

[75] Nguyen D. Lo C.C. and Skindilias K. A Unified Tree Approach for Options Pricing under Stochastic Volatility Models. *Finance Research Letters*, 20:260–268, 2017.

[76] Johnson F. M. Derivative Pricing with Non-Linear Fokker-Planck Dynamics. *Physica A*, 324:359–365, 2003.

[77] Gourieroux C. and Sufana R. Derivative Pricing with Wishart Multivariate Stochastic Volatility. *Journal of Business & Economics Statistics*, 28(3):438–451, 2010.

[78] Cesa M. A Brief History of Quantitative Finance. *Probability, Uncertainty and Quantitative Risk*, 2(1):1–16, 2017.

[79] Brisbois F. Vandewalle N. and Tordoir X. Non-random Topology of Stock Markets. *Quantitative Finance*, 1:372–374, 2001.

[80] Phoa W. Portfolio Concentration and the Geometry of Co-Movement. *The Journal of Portfolio Management*, pages 142–151, 2013.

[81] Gidea M. Topology Data Analysis of Critical Transition in Financial Networks. *arXiv*, pages 1–13, 2017.

[82] Gidea M. Topological Data Analysis of Financial Time Series: Landscapes of Crashes. *arXiv*, pages 1–28, 2017.

[83] Khasawne F.A. and Munch E. Exploring Equilibria in Stochastic Delay Differential Equations using Persistent Homology. *Proceedings of the ASME 2014 International Design Engineering Technical Conferences Computers and Information in Engineering Conference*, 2014.

[84] Khasawne F.A. and Munch E. Stability Determination in Turning using Persistent Homology and Time Series Analysis. *Proceedings of the ASME 2014 International Mechanical Engineering Congress Exposition*, 2014.

[85] Khasawne F.A. and Munch E. Utilizing Topological Data Analysis for Studying Signals of Time-Delay. *Time Delay Systems: Advances in Delays and Dynamics*, 7:93–106, 2017.

[86] Keogh E. and Kasetty S. On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration. *Data Mining and Knowledge Discovery*, 7:349–371, 2003.

[87] Yang Q. and Wu X. 10 Challenging Problems in Data Mining Research. *International Journal of Information Technology Decision Making*, 5(4):597–604, 2006.

[88] Preis T. Econophysics - Complex Correlations and Trend Switching in Financial Time Series. *The European Physical Journal Special Topics*, 194(1):5–86, 2011.

[89] BenSaïda A. and Latimi H. High Level Chaos in the Exchange and Index Markets. *Chaos, Solitons & Fractals*, 54:90–95, 2013.

[90] Zhuang X-F. and Chan L-W. Volatility Forecasts in Financial Time Series with HMM-GARCH Models. *IDEAL*, pages 807–812, 2004.

[91] Han J. and Zhang X-P. Financial Time Series Volatility Analysis Using Gaussian Process State-Space Models. *IEEE Global Conference on Signal and Information Processing*, pages 1–5, 2015.

[92] Bao D. A Generalized model for Financial Time Series representation and Prediction. *Applied Intelligence*, 29(1):1–11, 2008.

[93] Schäfer R. Leyvraz F. Seligman T. H. Guhr R. Münnix M. C., Shimada T. and Stanley H. E. Identifying States of a Financial Market. *Scientific Reports*, 2:1–6, 2012.

[94] Kim D. H. Nobi A., Lee S. and Lee J. W. Correlation and Network Topologies in Global and Local Stock Indices. *Physics Letters A*, 378:2482–2489, 2014.

[95] Ha G. G. Nobi A., Maeng S. E. and Lee J. W. Effects of Global Financial Crisis on Network Structure in a Local Stock Market. *Physica A*, 407:135–143, 2014.

[96] Brock W. A. Brovkin . Carpenter S. R. Dakos V. Held H. Nes E. H. V. Rietkerk M. Scheffer M., Bascompte J. and Sugihara G. Early-Warning Signals for Critical Transitions. *Nature*, 461(7260):53–59, 2009.

[97] McKelvey T. and Guerin G. Non-parametric Frequency Response Estimation using a Local Rational Model. *16th IFAC Symposium on System Identification the International Federation of Automatic Control Brussels*, 16(1):49–54, 2012.

[98] Baba N. Gelfert K. Kantz H., Just W. and Riegert K. Fast Chaos versus White Noise: Entropy Analysis and a Fokker-Planck model for the Slow Dynamics. *Physica D: Nonlinear Phenomena*, 187(1-4):200–213, 2004.

[99] Ang. A. and Timmermann A. Regime Changes and Financial Markets. *National Buerau of Economic Research*, pages 1–34, 2011.

[100] Wasserman L. Topological Data Analysis. *arXiv:1609.08227v1*, 2016.

[101] Tillman U. Grindrod P. Otter N., Porter M. A. and Harrinton H. A. A Roadmap for the Computation of Persistent Homology. *arXiv:1506:08903v7*, 2017.

[102] Edelsbrunner H. and Harer J. Computational Topology: An Introduction. *American Math Society*, 2010.

[103] Bubenik P. Statistical Topological Data Analysis using Persistence Landscapes. *arXiv:1207.6437v4*, 2015.

[104] Bubenik P. and Dlotko. P. A Persistence Landscapes Toolbox for Topological Statistics. *Journal of Symbolic Computation*, 78:91–114, 2017.

[105] Dean T. and Wellman M. Planning and Control. *Morgan Kaufmann Publishing Company*, 1991.

[106] Wilson J. Manifolds. *WOMP 2012*, pages 1–11, 2012.

[107] Whitney H. Differentiable Manifolds. *Annals of Mathematics*, 37(3):645–680, 1936.

[108] Takens F. Detecting Strange Attractors in Turbulence. *Dynamical Systems and Turbulence*, 898:366, 1981.

[109] Korku T. K. Taken Theorem with Singular Spectrum Analysis Applied to Noisy Time Series. *Electronic Theses and Dissertations*, 3013, 2016.

[110] Farmer J. D. Casdagli M., Eubank S. and Gibson J. State Space Reconstruction in the Presence of Noise. *Physica D: Nonlinear Phenomena*, 51(1-3):52–98, 1991.

[111] Small M. Optimal Time Delay Embedding for Nonlinear Time Series Modeling. *arXiv:nlim/0312011v1*, 2003.

[112] Yorke J. A. Sauer T. and Casdagli M. Embedology. *Journal of Statistical Physics*, 65(3-4):579–616, 1991.

[113] Brown R. Kennel M. B. and Abarbanel H.D. Determining Embedding Dimension for Phase-space Reconstruction using a Geometrical Construction. *Physica Review A*, 45(6):3403–3411, 1992.

[114] Abarbanel H.D.I. Analysis of Observed Chaotic Data. *Institute for Nonlinear Science Book Series*, pages 40–43, 1996.

[115] Jones R. Broomhead D. and King G. P. Comment on Singular-Value Decomposition and Embedding Dimension. *Physical Review A*, 37:5004, 1988.

[116] Paluš M. and Dvořák I. Singular-value Decomposition in Attractor Reconstruction: Pitfalls and Precaustions. *Physica D: Nonlinear Phenomena*, 55:221–234, 1992.

[117] Perc M. Kodba S. and Marhl M. Detecting Chaos from a Time Series. *European Journal of Physics*, 26(1):205–215, 2005.

[118] Zaldivar J.M. Strozzi F. and Zbilut J.P. Application of Nonlinear Time Series Analysis Techniques to High-Frequency Currency Exchange Data. *Physica A: Statistical Mechanics and its Applications*, 312:520–538, 2002.

[119] Arlt J. and Arltová M. Financial Time Series and Thier Features. *Acta Oeconomica Pragensia*, 9(4):7–20, 2001.

[120] Mandelbrot B. The Variation of Certain Speculative Prices. *The Journal of Business*, 36(4):394–419, 1963.

[121] Fama E. F. The Behaviour of Stock-Market Prices. *The Journal of Business*, 38(1):34–105, 1965.

[122] Linden M. A Model for Stock Return Distribution. *International Journal of Finance and Economics*, 6:159–169, 2001.

[123] Attali D. Amenta N. and Devillers O. Complexity of Delaunay Triangulation for Points on Lower-Dimensional Polyhedra. *Proceedings of the 18th annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1106–1113, 2007.

[124] Attali D. and Boissonnat J-D. Complexity of the Delaunay Triangulation of points on Polyhedral Surfaces. *Discrete Computational Geometry*, 30(3):437–452, 2001.

[125] Boissonnat J-D. Attali D. and Lieutier A. Complexity of the Delaunay Triangulation of Points on Surfaces: the Smooth Case. *Proceedings of the 19th Annual Symposium on Computational Geometry*, pages 201–210, 2003.

[126] Hastie T James G., Witten D. and Tibshirani R. An Introduction to Statistical Learning with Applications in R. *Springer*, pages 374–385.

[127] McElroy T. and Holan S. Using Spectral Peaks to Detect Seasonality. *U.S. Census Bureau and University of Missouri-Columbia*, pages 1–12.

[128] Cai S.M. Hong L. Lang P., Liu D.B. and Zhou P.L. Recurrence Network Analysis of the Synchronous EEG Time Series in Normal and Epileptic brains. *Cell Biochem Biophys*, 66(2):331–336, 2013.

[129] Donges J.F. Marwan N. Zou Y. Xiang R. Donner R. B., Michael S. and Kurths J. Recurrence-based Time Series Analysis by means of Complex Networks Methods. *arXiv:1010.6032*, 2010.

[130] Gu C. Stephen M. and Yang H. Visibility Graph Based Time Series Analysis. *PLoS One*, 10(11), 2015.

[131] Mukherjee S. Mileyko Y. and Harer J. Probability Measures on the Space of Persistence Diagrams. *Inverse Problems*, 27(12):1–25, 2011.

[132] Ziegelmeier L. Topaz C.M. and Halverson T. Topological Data Analysis of Biological Aggregation Models. *PLoS One*, 10(5), 2015.

TRITA -MAT-E 2017:80
ISRN -KTH/MAT/E--17/80--SE