MINIMUM EMBEDDING DIMENSION FROM THE PERSPECTIVE OF PERSISTENT HOMOLOGY

By

Christopher Lloyd Sukhu

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Computational Mathematics, Science and Engineering – Master of Science

2019

ABSTRACT

MINIMUM EMBEDDING DIMENSION FROM THE PERSPECTIVE OF PERSISTENT HOMOLOGY

By

Christopher Lloyd Sukhu

We investigate the use of 1-dimensional persistence diagrams to determine minimum embedding dimension. In particular, we test the claim that persistence diagrams look qualitatively the same once the correct dimension is reached. In some cases, this appears to not be true so we turn to a quantitative measure, the bottleneck distance, to see if the persistence diagrams are close once the minimum embedding dimension is attained. In some instances, we see that the persistence diagrams fail to converge experimentally under the bottleneck distance. The main issue appears to be that it is difficult to explicitly characterize the persistent homology of delay embeddings of arbitrary time series. Instead we restrict to periodic time series where there exists such an explicit characterization. We apply Fourier analysis to see that that number of peaks in the frequency spectrum of a delay embedded time series is related to the minimum embedding dimension. Moreover, we give a method to filter out less significant peaks while not altering the persistent homology much, with respect to the bottleneck distance.

ACKNOWLEDGEMENTS

Thank you to Drs. Elizabeth Munch and Jose Perea for their respective roles in helping me complete this thesis. The images used in Figures 2.1 and 2.3 are credited to Dr. Elizabeth Munch. Thank you to Luis Polanco for helping me generate Figure 4.3.

Software used include: Ripser and Hera libraries for computing persistence and bottleneck distances, respectively. Julia libraries: DifferentialEquations.jl, DynamicalSystems.jl, DataFrames.jl, RecurrenceAnalysis.jl, and Plots.jl. Python libraries: scikit-tda and NumPy/SciPy.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Studying physical processes or phenomena through experimental time series is ubiquitous in science. To understand the underlying dynamical behavior from these time series we turn to the Takens' embedding theorem [17] which allows us to reconstruct the underlying dynamics using delay embeddings. However, the delay embeddings require parameter choices, namely the delay and embedding dimension. The choice of embedding dimension is what this thesis chiefly addresses.

Tracking topological invariants like homology for determining the embedding dimension has been done before [11]. We build upon this by considering multi-scale topological quantities, namely persistent homology. We take a different perspective in this thesis by considering the geometry of the embedded signal to be important as well. Periodic signals correspond to circular embeddings so we exploit the connection to 1-dimensional persistence to be able to choose the embedding dimension. This is quite different but still has some of the same flavor as the traditional methods involving nearest neighbors [2, 8].

This thesis is divided into three main chapters. In Chapter 2, we develop some background pertaining to delay embeddings and persistent homology.

In Chapter 3, we investigate the use of 1-dimensional persistence diagrams to determine minimum embedding dimension. This was first considered in [6, 9] where it was claimed that persistence diagrams look qualitatively the same once the minimum embedding dimension is reached. In some cases, this appears to not be true so we turn to a quantitative measure, the bottleneck distance, to see if the persistence diagrams are close once the minimum embedding dimension is attained. In some instances, we see that the persistence diagrams fail to converge experimentally under the bottleneck distance. The main issue appears to be that it is difficult to explicitly characterize the persistent homology of delay embeddings of arbitrary time series.

Therefore in Chapter 4, we restrict to periodic time series where there exists such an explicit characterization [14]. We apply Fourier analysis to see that that number of peaks in the frequency

spectrum of a delay embedded time series is related to the minimum embedding dimension. Moreover, we give a method to filter out less significant peaks while not altering the persistent homology much, with respect to the bottleneck distance.

CHAPTER 2

BACKGROUND

2.1 Delay embeddings

We first state Takens' embedding theorem [7, 17] and explain its interpretation and consequences in the context of time series analysis and signals processing. We will often reformulate the notation and terminology involved in delay embeddings to maintain consistency with the primary sources while making certain things clearer and more convenient in context. The only disadvantage is a small amount of redundancy which stems from the abundance of domain-specific applications motivated by Takens' embedding theorem. Then we discuss the minimum embedding dimension problem.

2.1.1 Takens' embedding theorem

First, we explain some of the notation and terminology. The functions y are called measurement functions. The space $\text{Diff}^2(M)$ is the subspace of diffeomorphisms in $C^r(M, M)$, also called the diffeomorphism group (which underpins the theorem's terse notation). And "generic" means open and dense with respect to the C^1 topology. We use the term "delay embedding" to refer to the process of taking repeated measurements implied by Takens' embedding theorem. See Figure 2.1.

Theorem 2.1.1 ***(Takens)*** *Let M be a compact manifold of dimension m. For pairs (ϕ, y), with $\phi \in Diff^2(M)$, $y \in C^2(M, \mathbb{R})$, it is a generic property that the map $\Phi_{(\phi,y)} : M \to \mathbb{R}^{2m+1}$, defined by*

$$\Phi_{(\phi,y)}(x) = (y(x), y(\phi(x)), \ldots, y(\phi^{2m}(x)))$$

is an embedding.

We often make the assumption that real-world data lies on a lower dimensional manifold [3]. This assumption combined with Takens' embedding theorem means that we can take a sequence

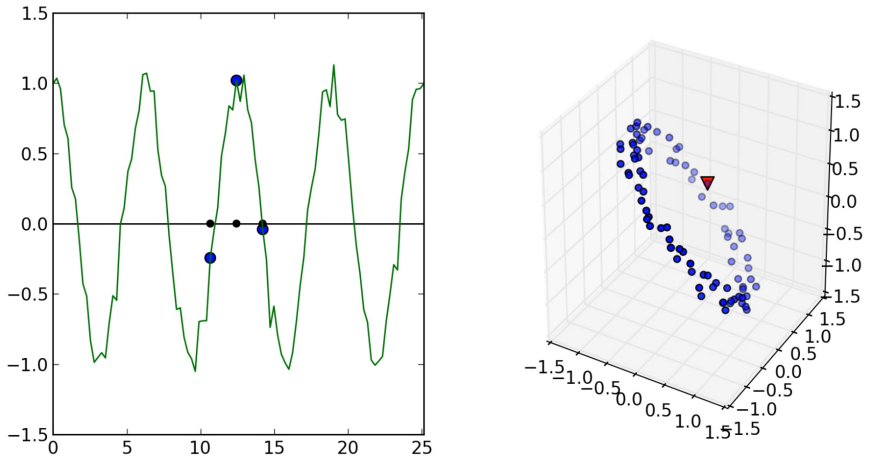

Figure 2.1: Delay embedding of a noisy cosine wave into $\mathbb{R}^3$ which captures the underlying periodic structure.

of measurements, often indexed by time, to reconstruct a dynamical system, i.e. any deterministic process that evolves with time. All we need is a sufficiently nice measurement function of some observable of the dynamical system and enough measurements. The number of measurements required is twice the dimension of the dynamical system's true state space plus one [7]. We note that this dimension is usually not known in practice, but we take consolation that it is at least finite. To sum up, we can study any deterministic process by a finite time series of measurements of a single observable. This is truly remarkable, so we take care to state some caveats and hidden assumptions.

Takens' embedding theorem assumes no experimental noise and the ability to make measurements up to arbitrary precision. Moreover, while the number of measurements needed is finite, in practice it might still be very large for dynamical systems with high-dimensional state spaces, especially if there is noise involved. Another hidden assumption is that we need evenly spaced measurements which is often not possible in some applications. We note that the time interval, or delay, between measurements is not specified by the theorem as well. Despite these limitations, delay embedding has been widely employed as a tool in time series analysis and signals processing. See [1] for a survey of delay embeddings as well as specific applications to EEG analysis.

For a real-world application, there are a number of parameter choices that go into delay

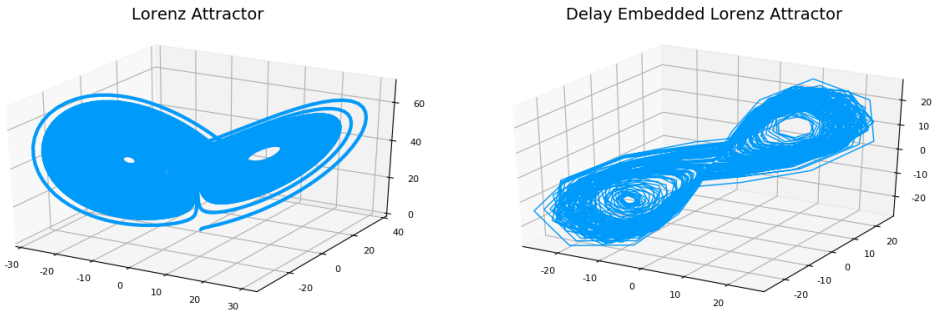

Figure 2.2: Lorenz attractor (left) and reconstructed attractor (right).

embeddings. One is the observable to measure and yet another is the measurement function to use. These are typically chosen with regards to whatever is available and convenient. The main two parameters that must be specified are the number of measurements and the delay. We refer to the number of measurements as the "embedding dimension" and endeavor to find the smallest embedding dimension; this is the minimum embedding dimension problem which is the central problem addressed by this thesis.

Before we proceed, we mention that there are a number of reformulations and extensions to Takens' embedding theorem. See [16] for the following (paraphrased) version involving fractal sets: A dynamical system with an underlying attractor A of dimension d_A needs an embedding dimension greater than twice d_A to be reconstructed faithfully. See Figure 2.2 for the Lorenz attractor with parameters $\sigma = 16$, $\rho = 40$, and $\beta = 4$ for which $d_A = 2.06 \pm 0.01$ [10]. We can see that the minimum embedding dimension is 3. According to the embedding theorem, an embedding dimension of 5 is sufficient for a good reconstruction, larger than the minimum embedding dimension.

2.1.2 Minimum embedding dimension

The minimum dimension required for an embedding is often smaller than the bound given in Takens' embedding theorem. Being able to determine the minimum embedding dimension for a particular dynamical system is interesting in its own right. However, the main practical benefit is that having a smaller dimension facilitates further computations on a reconstructed dynamical

system.

Before we proceed, we acknowledge that the terms "minimum embedding dimension" and "embedding dimension" are potentially ambiguous or misleading. When performing a delay embedding with a particular "embedding dimension", the result may fail to be an embedding. A better term, perhaps, would be "reconstruction dimension" or "attempted embedding dimension", however we stick to the terms commonly used in the literature with the hope that reader will keep this warning in mind [2, 7, 14]. A more serious issue is what constitutes a "minimum embedding dimension." The Takens' embedding theorem and its related variants [16] give theoretically sufficient (minimum) embedding dimensions. Since we seek even smaller dimensions for practical benefits, which is often possible for specific dynamical systems, we are content to call whichever dimension we settle on according to our analysis the "minimum embedding dimension" regardless of whether we can theoretically justify this or not. And when designing a method which chooses or optimizes the embedding dimension according to a particular heuristic or quantity of interest, we call the result the "minimum embedding" dimension.

2.1.3 Average False Nearest Neighbors

There exist many methods in the literature attempting to determine the minimum embedding dimension. One survey of these methods can be found in [1]. We now discuss the popular method of Average False Nearest Neighbors (AFNN) [2] which we will use for comparison purposes later. AFNN builds upon the method of False Nearest Neighbors (FNN) [8] which rests on the idea that points close together in a dimension too low to be an embedding will move further apart if the dimension increases.

Let $\{x_i : i = 1, \ldots, N\}$ be a collection of sequential measurements, i.e. a time series. The reconstructed time-delay vectors are $\{y_i(d) = (x_i, x_{i+\tau}, \ldots, x_{i+(d-1)\tau} : i = 1, \ldots, N - (d-1)\tau\}$, for a fixed delay τ, determined in advance. Next we define a quantity that relates distances in one

embedding dimension to the next

$$a(i,d) = \frac{||y_i(d+1) - y_{n(i,d)}(d+1)||_\infty}{||y_i(d) - y_{n(i,d)}(d)||_\infty}$$

where $n(i,d)$ is the index of the nearest neighbor of $y_i(d)$ in dimension d (if the denominator were to be zero, we choose the next nearest neighbor). AFNN builds on FNN by considering the arithmetic mean

$$E(d) = \frac{1}{N - d\tau} \sum_{i=1}^{N-d\tau} a(i,d)$$

and defining

$$E1(d) = \frac{E(d+1)}{E(d)}$$

to track changes as the embedding dimension increases. If $E1(d)$ stops changing for a sequence of dimensions $d = 1, 2, \ldots$ after d_k, we say $d_k + 1$ is the minimum embedding dimension.

In practice, it may be difficult to determine if $E1(d)$ has stopped changing or is merely increasing slowly. Indeed, when the time series is random, $E1(d)$ might fail to converge in any meaningful sense. So we define

$$E^*(d) = \frac{1}{N - d\tau} \sum_{i=1}^{N-d\tau} |x_{i+d\tau} - x_{n(i,d)+d\tau}|$$

and

$$E2(d) = \frac{E^*(d+1)}{E^*(d)}$$

where we expect for random data $E2(d)$ will always equal 1, since the value of the data is independent across time. For a time series which is deterministic, this will not be the case, so we can use $E2(d)$ as heuristic for distinguishing time series arising from random versus deterministic processes.

The AFNN method consists of computing both $E1(d)$ and $E2(d)$ across a range of dimensions d and analyzing the results qualitatively.

2.2 Persistent homology

We develop the basics of simplicial homology and persistent homology needed for the remainder of this thesis, and therefore not in full generality or detail. The selection of topics closely matches

the background found in [14]. For more on simplicial homology and persistent homology see [12] and [5, 13], respectively.

We fix a prime p and the finite field $\mathbb{F}_p$ with p elements henceforth.

2.2.1 Simplicial Homology

We begin with the notion of simplices, which can be thought of as the appropriate generalization of triangles or tetrahedrons to arbitrary dimensions. Let $\{v_0, \ldots, v_k\}$ be a collection of $k+1$ points in $\mathbb{R}^k$ which are affinely independent, i.e. $v_1 - v_0, \ldots, v_k - v_0$ are linearly independent.

Definition 2.2.1 *(Geometric k-simplex) A k-simplex σ spanned by $\{v_0, \ldots, v_k\}$ is the set of points $x \in \mathbb{R}^k$ such that*

$$x = \sum_{i=0}^{k} t_i v_i \text{ where } \sum_{i=0}^{k} t_i = 1 \text{ and } t_i \geq 0 \ \forall i.$$

That is, σ is the convex hull of $\{v_0, \ldots, v_k\}$.

We often use the notation $\sigma[v_0, \ldots, v_k]$ to denote a simplex spanned by those points. Examples of k-simplices include 0-simplices which are points, 1-simplices which are line segments, 2-simplices which are triangles, and 3-simplices which are tetrahedrons.

The points $\{v_0, \ldots, v_k\}$ are called the vertices of a k-simplex σ. The dimension of a k-simplex σ is simply k. And a simplex spanned by a subset of $\{v_0, \ldots, v_k\}$ is called a face of σ.

Definition 2.2.2 *(Geometric simplicial complex) K is called a simplicial complex if it is a set of simplices such that*

(i) Every face of a simplex in K is also in K.

(ii) The intersection of any two simplices in K is disjoint or a face of both.

A subcomplex is then a subset of the simplices of a simplicial complex K that contains the faces of all its elements.

Typically, we are not interested in the particular embedding of a simplicial complex into Euclidean space and just want a purely combinatorial structure for computations. Hence, we consider an abstract simplicial complex that has analogous properties to the definitions above.

Definition 2.2.3 *(Abstract simplicial complex) An abstract simplicial complex K is a finite collection of sets Σ such that $\sigma \in \Sigma$ and $\nu \subseteq \sigma$ implies $\nu \in \Sigma$.*

The terminology is analogous to before: The sets σ in Σ are simplices. The dimension of a simplex σ is the cardinality of σ minus one. And a non-empty subset ν of σ is called a face of σ. A subcomplex is then a subset of K which is also an abstract simplicial complex.

We are content to use the same notation and terminology because of the following. Starting with a geometric simplicial complex K', we can obtain an abstract simplicial complex K by only keeping the vertices of K'. We then call K' a geometric realization of K. See [5] for details in the proof of the following theorem.

Theorem 2.2.4 *(Geometric realization theorem) Every abstract simplicial complex of dimension n has a geometric realization in $\mathbb{R}^{2n+1}$.*

Hence, we will just say simplicial complex to refer to an abstract simplicial complex. When considering sums of simplices for abstract simplicial complexes we do so only formally so that the operations make sense algebraically, without regard to the underlying geometric meaning.

Definition 2.2.5 *(k-chains) We say c is a k-chain if it is a finite formal sum*

$$c = \sum_j \gamma_j \sigma_j \text{ with } \gamma_j \in \mathbb{F}_p$$

and each σ_j is a k-simplex in K.

We let $C_k(K)$ be the vector space over the field $\mathbb{F}_p$ generated by the k-dimensional simplices of K, i.e. $C_k(K)$ is a vector space consisting of k-chains.

Definition 2.2.6 *(Boundary of k-simplex) The boundary ∂ of a k-simplex σ is the alternating formal sum of k-1 dimensional faces of σ. We denote this as follows*

$$\partial(\sigma) = \sum_{i=0}^{k} (-1)^i \sigma[v_0, \ldots, \widehat{v_i}, \ldots, v_k]$$

where the hat symbol over a vertex means that vertex is not present in the spanning set.

Definition 2.2.7 *(Boundary of k-chain) The boundary ∂ of a k-chain c is defined by linearly extending ∂ as follows*

$$\delta(c) = \sum_j \gamma_j \partial(\sigma_j).$$

Remark 2.2.8 *One can verify that $\partial \circ \partial = \partial^2 = 0$.*

Definition 2.2.9 *(k-cycles) A k-chain c with $\partial(c) = 0$ is called a k-cycle.*

We denote the space of all k-cycles as Z_k which is a subspace of C_k.

Definition 2.2.10 *(k-boundaries) If a k-chain c is the boundary of a (k+1)-chain then it is called a k-boundary.*

We denote the space of all k-boundaries as B_k which is a subspace of C_k. By Fact 2.2.8 we have that $B_k \subset Z_k$ so we define the following.

Definition 2.2.11 *(Simplicial homology) The k-th simplicial homology group of K with $\mathbb{F}_p$-coefficients is defined as the quotient $H_k(K) = Z_k / B_k$.*

The ranks of the simplicial homology groups, known as Betti numbers, are of particular interest to us.

Definition 2.2.12 *(Betti numbers) The rank of $H_k(K)$ is called the k-th mod p Betti number of K which we denote $\beta_k(K)$.*

The prime p is usually suppressed from the notation because it will be clear in context.

2.2.2 Persistent homology

We now turn to persistent homology which we develop on top of simplicial homology.

Definition 2.2.13 *(Filtration) A filtration of a simplicial complex K is a nested sequence of subcomplexes*

$$\emptyset = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_m = K.$$

Inclusions $K_i \subseteq K_j$ for $i \leq j$ in the filtration induce homomorphisms $f_k^{i,j} : H_k(K_i) \to H_k(K_j)$ for each dimension k. We obtain the following sequence of homorphisms of homology groups

$$0 = H_k(K_0) \to H_k(K_1) \to \cdots \to H_k(K_m) = H_k(K)$$

where the arrows are $f_k^{i,j}$ for $0 \leq i \leq j \leq m$,

Persistent homology refers to the images of these induced homomorphisms.

Definition 2.2.14 *(Persistent homology) The k-th persistent homology groups of a simplicial complex K are the images of the homomorphisms induced by a filtration of K, $H_k^{i,j} = im f_k^{i,j}$. And the k-th persistent Betti numbers are the ranks of the homology groups, $\beta_k^{i,j} = rank(H_k^{i,j})$.*

We say a homology class α is born at K_b if it is not in the image of $f_k^{b-1,b} : H_k(K_{b-1}) \to H_k(K_b)$. If α is born at K_b, then we say it dies entering K_d if the image of $f_k^{b-1,d-1} : H_k(K_{b-1}) \to H_k(K_{d-1})$ does not contain the image of α but the image of $f_k^{b-1,d} : H_k(K_{b-1}) \to H_k(K_d)$ does.

Definition 2.2.15 *(Persistence diagrams) A persistence diagram of homological dimension k, or k dimensional persistence diagram, denoted dgm(k), is a multiset of points (b,d) for every k-dimensional homology class that is born at K_b and dies entering K_d. We also adjoin the points on the diagonal $\Delta = \{(x, x) : x \geq 0\}$ to dgm(k) each with countably infinite multiplicity.*

If k is clear from context we just say dgm. We refer to $d - b$ as the lifetime or persistence of α.

Definition 2.2.16 *(Maximum persistence) Let $(b, d) \in dgm$. Define $pers(b, d) = d - b$ if $(b, d) \in \mathbb{R}^2$ and as ∞ otherwise. Maximum persistence, denoted $mp(dgm)$, is defined to be*

$$mp(dgm) = max_{(b,d)\in dgm} pers(b, d).$$

When comparing two diagrams, we use the notation dgm_1 and dgm_2 for convenience as in the following definition.

Definition 2.2.17 *(Bottleneck distance) The bottleneck distance denoted d_B is a metric on persistence diagrams defined as follows*

$$d_B(dgm_1, dgm_2) = \inf_{\phi} \sup_{x \in dgm_1} ||x - \phi(x)||_\infty$$

where ϕ is a bijection $dgm_1 \to dgm_2$.

We always have these bijections, despite potentially unequal numbers of off-diagonal points, because we can identify points with the diagonal.

We now turn to a particular kind of filtration/complex called the Vietoris-Rips filtration/complex or simply the Rips filtration/complex which we will use later. First we let $X \subset \mathbb{R}^n$ be a compact set, such as a finite point cloud.

Definition 2.2.18 *(Rips complex) Fix $r \geq 0$. The Rips complex $R_r(X)$ is the simplicial complex whose vertices are the points of X and whose k-simplices consist of the k+1-tuples of points of X $\{x_0, \ldots, x_k\}$ with pairwise distances $||x_i - x_j|| \leq r$ for all i, j with $0 \leq i < j \leq k$.*

Note that higher dimensional simplices are added to the Rips filtration if and only if all its edges (1-simplices) are. Also, we can adapt this definition for any metric space.

Definition 2.2.19 *(Rips filtration) Let $0 = r_0 \leq r_1 \leq \cdots \leq r_m$. If $r \leq s$, then $R_r(X) \subseteq R_s(X)$ so we obtain the following filtration, called the Rips filtration, from the Rips complexes*

$$X = R_0 \subseteq R_1 \subseteq \cdots \subseteq R_m$$

where $R_j = R_{r_j}(X)$ and R_m is the largest simplicial complex with X as its vertex set.

It is sufficient to consider only a finite set of values of r_j to capture all homological changes in our particular case.

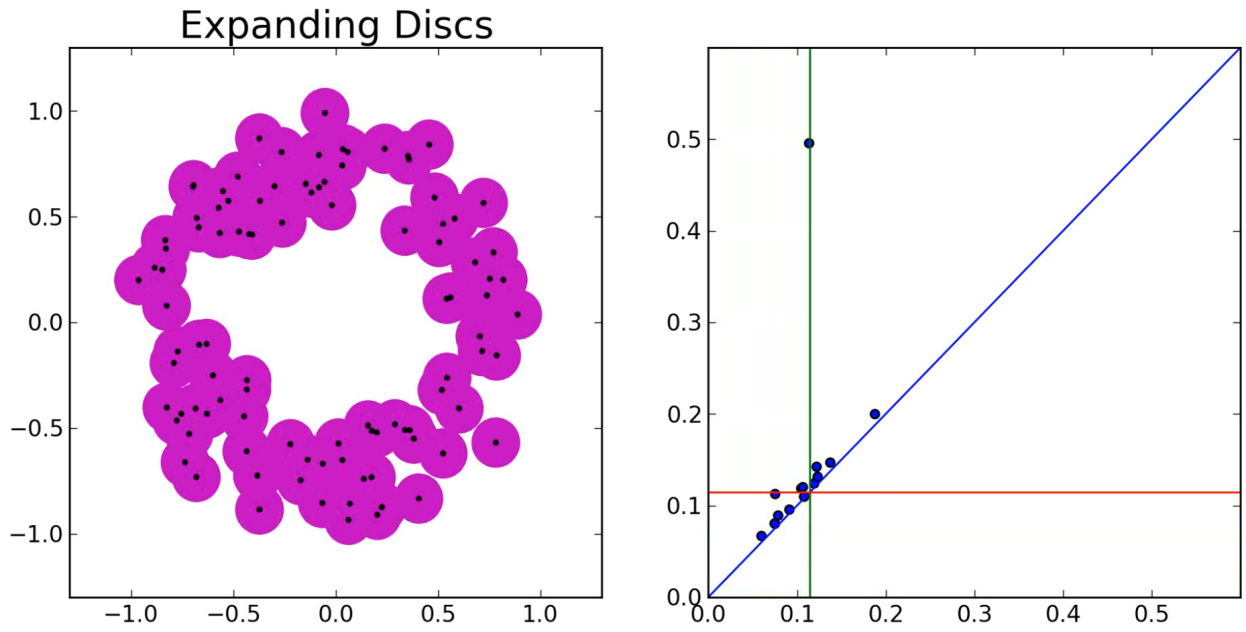

Figure 2.3: Pictured (left) is a point cloud in $\mathbb{R}^2$ at a particular distance r_j just over the value of 0.1 in the Rips filtration, with purple shaded disks of radius r showing which edges will be included in the Rips complex. Pictured (right) is the one dimensional persistence diagram showing the 1-cycle that is born at this particular point in the filtration, which we see will die at filtration value around 0.5, when the purple disks will be large enough to fill in the center of the point cloud and therefore connect the corresponding vertices, killing the 1-cycle.

The persistence diagram of a point cloud X will be denoted $dgm(X)$ where we consider the persistent homology induced by the Rips filtration on X. See Figure 2.3 for an illustration.

Since we intend to apply persistence to point clouds from experimental data, we accept that there will be measurement error and therefore greatly rely on the following stability theorem. See [4, 13] for more details. First, we define the Hausdorff distance for point clouds.

Definition 2.2.20 *(Hausdorff distance) The Hausdorff distance d_H between two point $X, Y \subseteq \mathbb{R}^n$ clouds is*

$$d_H(X,Y) = max\{\sup_{x\in X}\inf_{y\in Y}||x-y||, \sup_{y\in Y}\inf_{x\in X}||x-y||\}.$$

Theorem 2.2.21 *(Stability of persistence diagrams)*

$$d_B(dgm(X), dgm(Y) \leq 2d_H(X,Y)).$$

Let X be the true point cloud of some experimental process and Y be the measured point cloud with some measurement error so that $d_H(X,Y) \leq \epsilon$. Then, Theorem 2.2.21 says that the bottleneck distance is bounded above by 2ϵ, so the error in persistence is not much worse than the measurement error.

CHAPTER 3

PERSISTENCE FOR MINIMUM EMBEDDING DIMENSION

3.1 Introduction

We explore the idea of tracking the persistent homology of delay embedded systems across a range of embedding dimensions to see if this has any utility in the minimum embedding dimension problem. Tracking topological features for delay embeddings has been done before for homology [11]. More recently, this has been tried using persistent homology [6, 9] for the minimum embedding dimension problem. The claim is that, intuitively, one might expect that for increasing embedding dimension the persistent homology of a reconstructed system might stabilize past the minimum embedding dimension, and therefore the persistence diagrams should converge in some sense. Or at the very least the diagrams should look qualitatively similar. To push this idea further, we compute persistence diagrams from some actual examples and test for convergence experimentally using the bottleneck metric.

3.2 Method

We compute delay embeddings for a range of increasing dimensions as in [2] where AFNN is used to determine the minimum embedding dimension. The zero and one dimensional persistence diagrams are computed with $\mathbb{F}_2$ coefficients and we plot the bottleneck distances between adjacent dimensions to check for any stabilizing behavior. We compare against AFNN as the ground truth. For the E1 and E2 plots, if the dimension stops changing at index d we consider d to be the minimum embedding dimension. For the bottleneck distance plots, the x-axis is the higher of the two adjacent dimensions and the value on the y-axis represents the change in persistence going up a dimension. If the values are small after index d, we consider this to be an indication that d is the minimum embedding dimension.

3.3 Results

We consider four data sets taken from [2]. These are data sets from the first coordinate of a Henon attractor, torus (i.e. sum of two non-commensurate sines), sum of four iterated sine maps, and the Santa Fe competition [18]. We reproduce the AFNN E1 and E2 plots for each example so there may be small differences with the original plots in [2]. See Figures 3.1-3.8 for the relevant figures.

For the Henon attractor, E1 sharply stops increasing after dimension 2, which is considered to be the minimum embedding dimension. E2 is not always approximately 1 so we consider the embedded time series to be deterministic. The bottleneck distances are all fairly small which is possibly consistent with the minimum embedding dimension being 2.

For the torus, the E1 values quickly stop increasing after dimension 3, which is exactly what we would expect for a standard torus. E2 is not always approximately 1 so we consider the embedded time series to be deterministic. We see that the bottleneck distance for the one dimensional diagrams have the largest jump when going to embedding dimension 3 and afterwards being fairly small, suggesting a minimum embedding dimension of 3.

The minimum embedding dimension appears to be 4 for the sum of four iterated sine maps according to E1 and E2 suggests it is deterministic as well. NaN values fail to appear in the E1 plot which is why it may look different from the corresponding figure in [2].

NaN values also fail to appear in E1 plot for the Santa Fe time series. For this experimental time series, the results are less clear. The minimum embedding dimension is suggested to be 7 in [2]. The bottleneck distances are possibly starting to show interesting behavior around dimension 7 but this is far from clear.

3.4 Discussion

Unfortunately, we do not really see any consistent convergent behavior with respect to the bottleneck distances. So it is not clear that tracking the persistence diagrams across embedding dimensions will be useful beyond the Lorenz attractor examples in [6, 9].

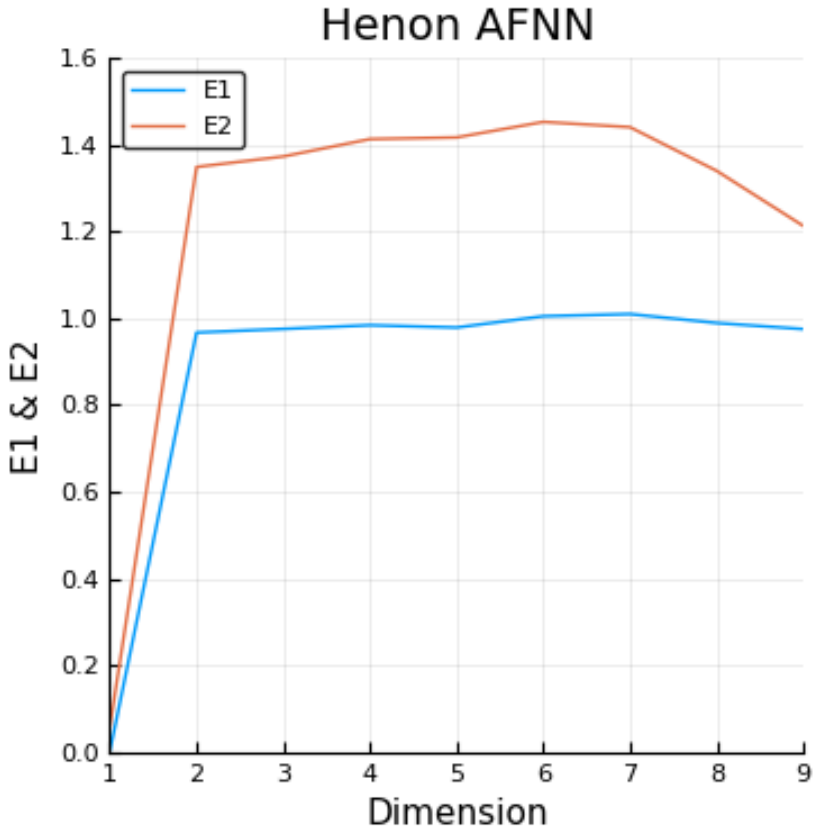

Figure 3.1: E1 sharply stops increasing after dimension 2, which is considered to be the minimum embedding dimension. E2 is not always approximately 1 so we consider the embedded time series to be deterministic.

There are two potential issues here: One is that since the average distance between points in $\mathbb{R}^m$ increases as the embedding dimension m increases, we actually expect the bottleneck distances to diverge as the dimension goes to infinity. The other is that not all systems will have a few prominent off-diagonal points in their zero or one dimensional persistence diagrams that will be reflected well in the bottleneck distance, so they might not even look qualitatively similar.

For the first issue, we can try re-scaling the persistence diagrams by multiplying the birth and death points by $1/\sqrt{m}$. We try this with the Lorenz attractor from Figure 2.2. See Figure 3.9. We expect two 1-dimensional features from the reconstructed Lorenz attractor, which corresponds to what we computed. The persistence diagrams look qualitatively similar from embedding dimension 3 to 4, indicating a minimum embedding dimension of 3, which is what we expect. Normalizing shows clearer convergence as the persistence points are artificially getting larger due to increased

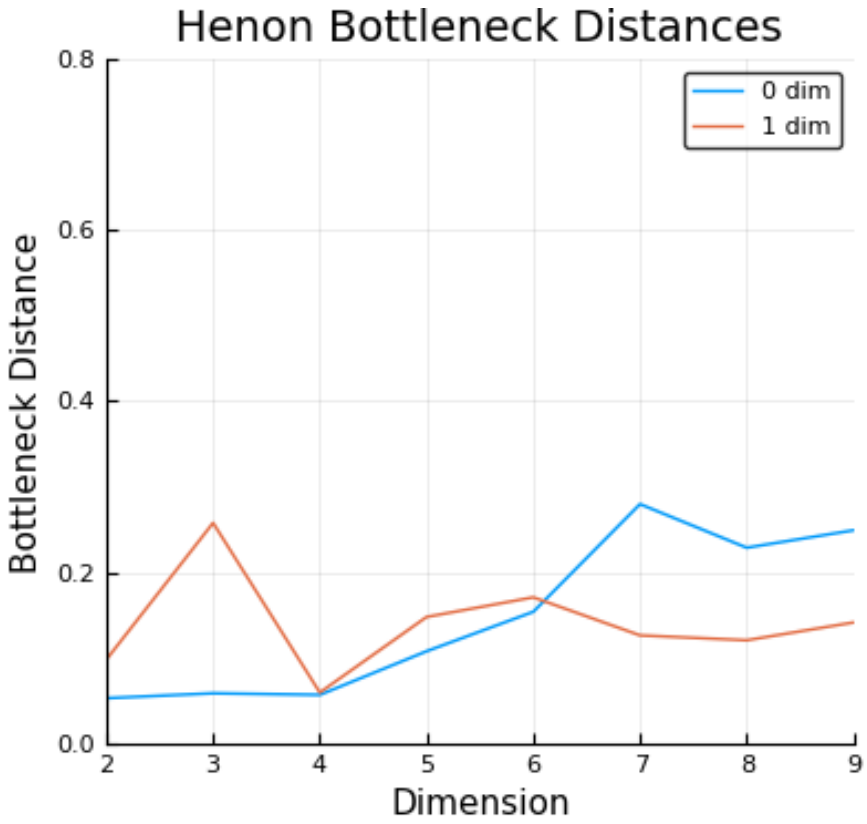

Figure 3.2: All of these bottleneck distance values are fairly small which is possibly consistent with the minimum embedding dimension being 2.

average distance between points in higher dimensions. However, there is seemingly no evidence that this generalizes in a way that might be useful for determining the minimum embedding dimension of other systems.

The second issue is the main problem. It is not clear what we can say theoretically about the persistent homology of delay embeddings of arbitrary time series or signals. This motivates the next part of this thesis where we restrict to the periodic case and focus on maximizing the connection to persistence.

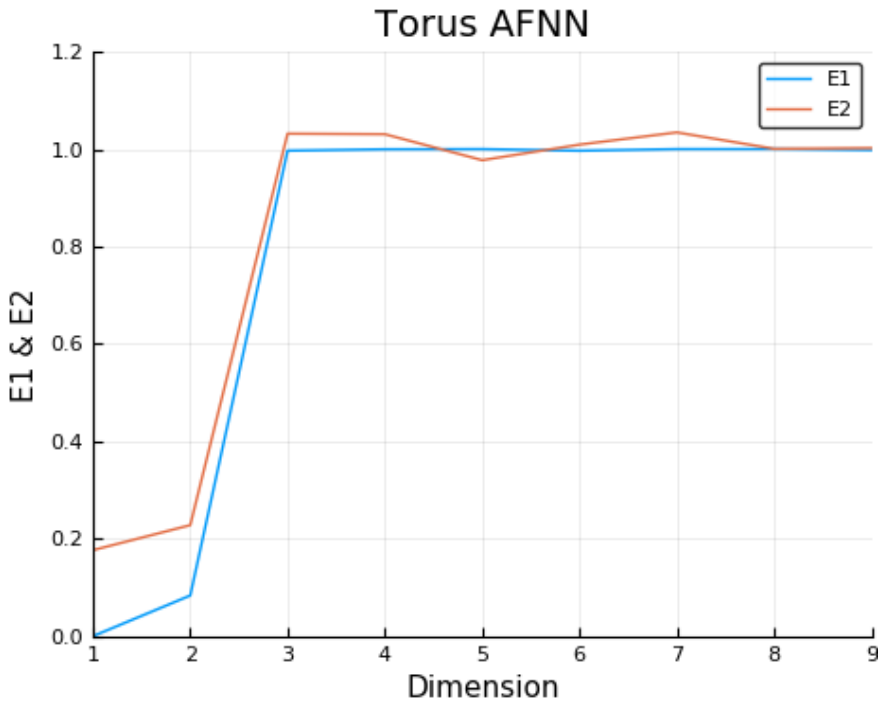

Figure 3.3: E1 values quickly stop increasing after dimension 3, which is exactly what we would expect for a standard torus. E2 is not always approximately 1 so we consider the embedded time series to be deterministic.

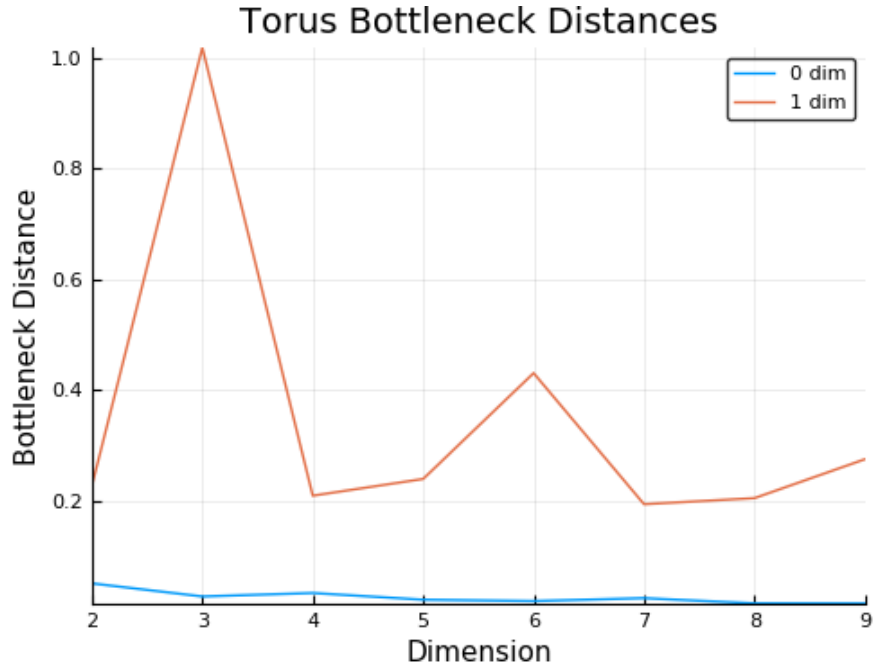

Figure 3.4: We see that the bottleneck distance for the one dimensional diagrams have the largest jump when going to embedding dimension 3 and afterwards being fairly small, suggesting a minimum embedding dimension of 3.

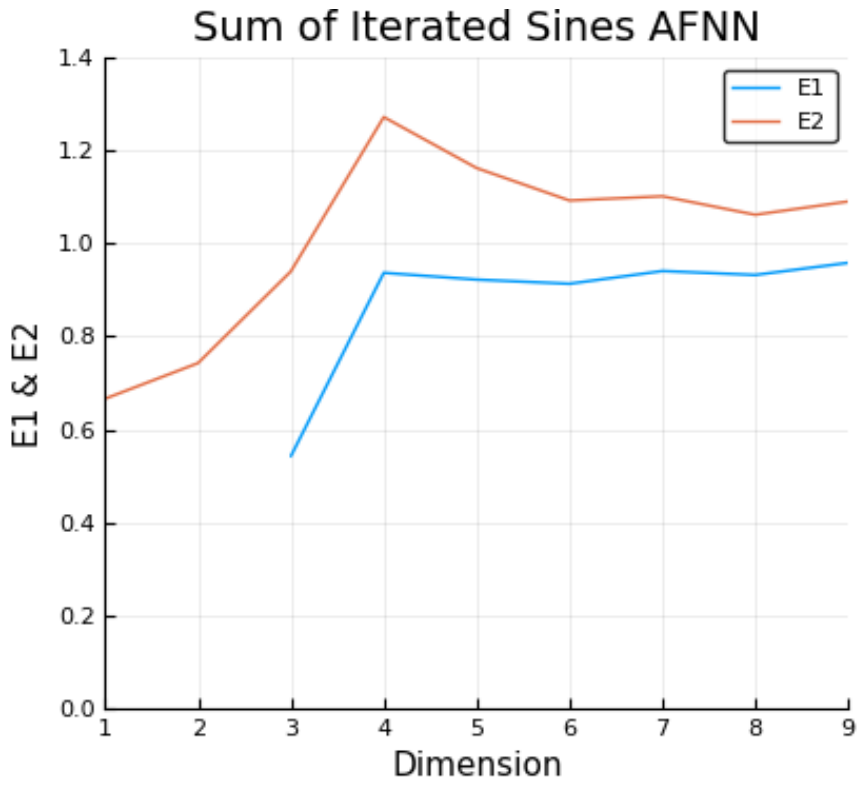

Figure 3.5: NaN values fail to appear in E1 plot. The minimum embedding dimension appears to be 4 for this deterministic time series.

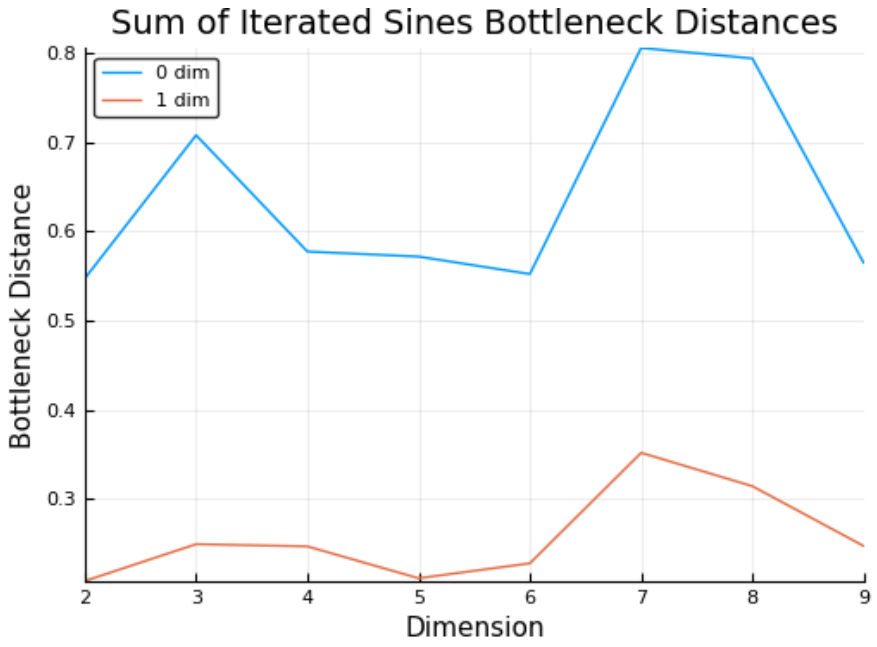

Figure 3.6: We do not observe any indication that the bottleneck distances are stabilizing after embedding dimension 4.

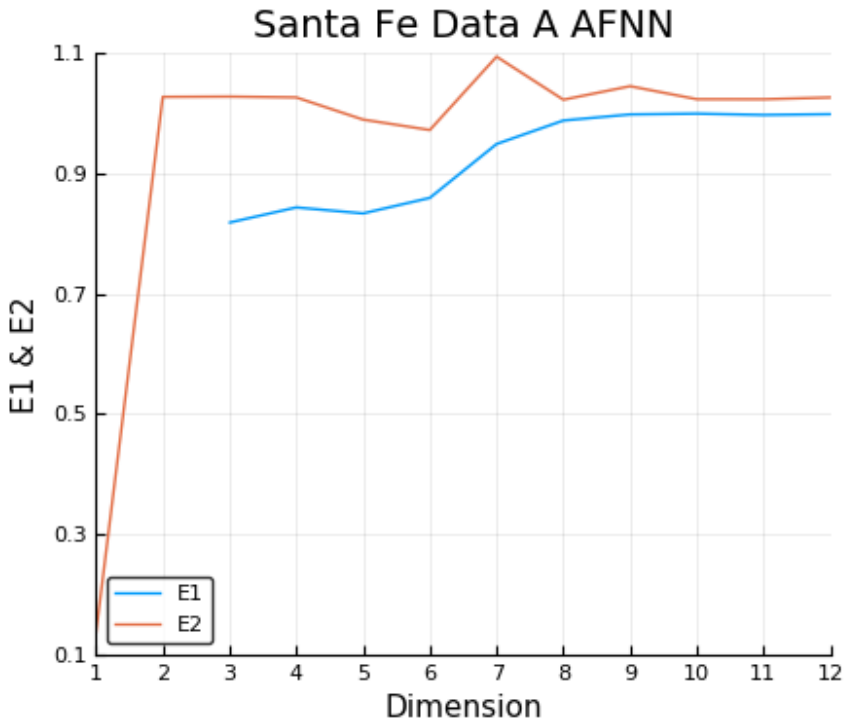

Figure 3.7: NaN values fail to appear in E1 plot. For this experimental time series, the results are less clear. The minimum embedding dimension is suggested to be 7 in [2].

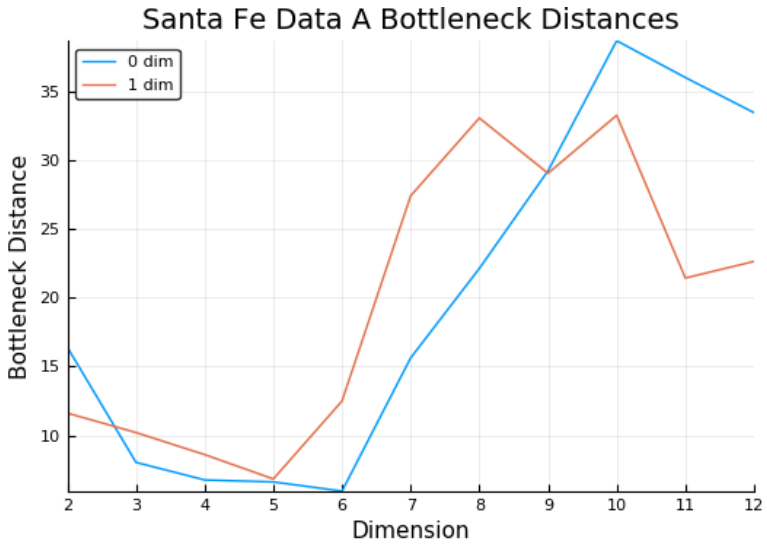

Figure 3.8: The bottleneck distances are possibly starting to show interesting behavior around dimension 7 but this is far from clear.

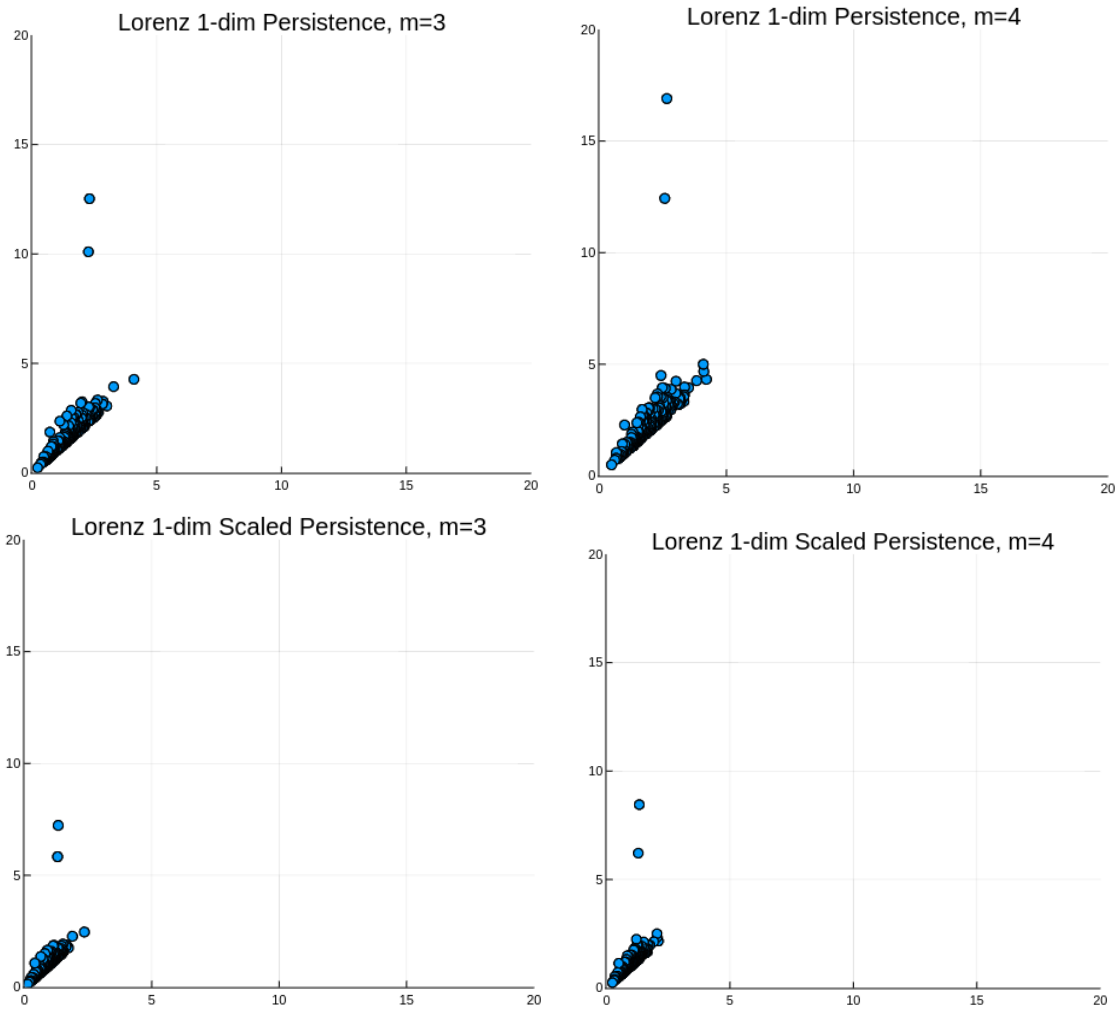

Figure 3.9: Lorenz attractor 1-dimensional persistence diagrams for embedding dimensions 3 and 4. Note that the diagonal is not pictured. The main focus should be on the two high persistence points, corresponding to the figure 8 shape of the Lorenz attractor.

CHAPTER 4

MINIMUM EMBEDDING DIMENSION OF PERIODIC SIGNALS

4.1 Sliding windows and persistence

We now summarize some of the main points of [14] which forms the basis of the next section. In this chapter, we only consider 1-dimensional persistence. Maximum persistence of the 1-dimensional persistence diagram is used as a measure of roundness corresponding to the shape of a delay embedding of a periodic signal. The following notation and terminology for delay embeddings will be used for the remainder of the chapter.

Definition 4.1.1 *(Sliding window embedding) Suppose that f is a function defined on an interval of $\mathbb{R}$. Then choose an embedding dimension $M \in \mathbb{Z}_{\geq 0}$ and delay $\tau \in \mathbb{R}_{>0}$. The sliding window embedding of f based at $t \in \mathbb{R}$ into $\mathbb{R}^{M+1}$ is the point*

$$SW_{M,\tau} f(t) = \begin{bmatrix} f(t) \\ f(t+\tau) \\ \vdots \\ f(t+M\tau) \end{bmatrix}.$$

For a range of values t we get the sliding window point cloud for f. The quantity $M\tau$ is called the window size.

Let $C(X,Y)$ denote the set of continuous functions from X to Y equipped with the sup norm. Let $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$. The sliding window embedding induces a mapping

$$SW_{M,\tau} : C(\mathbb{T},\mathbb{R}) \to C(\mathbb{T},\mathbb{R}^{M+1}).$$

Proposition 4.1.2 *$SW_{M,\tau} : C(\mathbb{T},\mathbb{R}) \to C(\mathbb{T},\mathbb{R}^{M+1})$ is a bounded linear operator with norm $||SW_{M,\tau}|| \leq \sqrt{M+1}$.*

Since we are considering periodic functions f it makes sense to approximate f using Fourier series. See [15] for an accessible introduction to Fourier series. Let $f(t) = S_N f(t) + R_N(f(t))$. The first term is the N-truncated Fourier series

$$S_N f(t) = \sum_{n=-N}^{N} = \widehat{f}(n)e^{int}$$

and $R_N f$ is the remainder. The n-th Fourier coefficient is

$$\widehat{f}(n) = \frac{1}{2\pi}\int_0^{2\pi} f(t)e^{-int}dt.$$

The following theorem, reprinted from [14] with a minor revision, tells us how $SW_{M,\tau}$ behaves with respect to Fourier series approximations of functions $f \in L^2(\mathbb{T})$.

Theorem 4.1.3 *(Approximation) Let $T \subset \mathbb{T}$, $f \in C^k(\mathbb{T}, \mathbb{R})$, $X = SW_{M,\tau}f(T)$, and $Y = SW_{M,\tau}S_N f(T)$. Then,*

(i)

$$d_H(X,Y) \leq \sqrt{2} \cdot ||R_N f^{(k)}||_2 \cdot \frac{(N+1)^{1-2k}}{2k-1} \cdot \sqrt{M+1},$$

(ii)

$$|mp(dgm(X)) - mp(dgm(Y))| \leq 2d_B(dgm(X), dgm(Y)),$$

(iii)

$$d_B(dgm(X), dgm(Y)) \leq 2\sqrt{2} \cdot ||R_N f^{(k)}||_2 \cdot \frac{(N+1)^{1-2k}}{2k-1} \cdot \sqrt{M+1}.$$

As we take more terms in the Fourier series approximation, the remainder goes to zero, and the persistent homology of the approximation approaches that of the true sliding window point cloud.

Another result is that the sliding window point cloud has maximum persistence when the window size $M\tau$ is proportional to the underlying frequency $\frac{2\pi}{L}$, with proportionality constant $\frac{M}{M+1}$. Here L is the period of f, specifically $f(t + \frac{2\pi}{L}) = f(t)$. A lower bound on maximum persistence is also derived and is shown to depend on the field of coefficients used to compute persistent homology.

The approach to delay embedding in [14] is markedly different from what is normally done in the literature. Delay embeddings of periodic signals f have a clear geometric interpretation which has an explicit connection to 1-dimensional persistent homology. $SW_{M,\tau}f$ lives on an M-dimensional torus embedded in $\mathbb{R}^{M+1}$, so we expect at least one prominent off-diagonal point in the 1-dimensional persistence diagram and choose parameters to maximize persistence.

4.2 Minimum embedding dimension and persistence

We now describe a method for choosing the minimum embedding dimension for periodic signals in a way that emphasizes maximum persistence.

Let $f \in L^2(\mathbb{T})$ and $f \in C^k(\mathbb{T})$. Then

$$f(t) = \sum_{n=-\infty}^{\infty} \widehat{f}(n) e^{int}$$

since the Fourier series converges by the Riemann-Lebesgue Lemma [15]. Now consider the power spectrum $S_{ff}(n) = |\widehat{f}(n)|^2$. For a periodic signal, we can expect a few significant peaks in the power spectrum. Say there are d peaks, then the minimum embedding dimension should be $M = 2d$ to lose no information [14].

What if we wish to discard some of the peaks in the power spectrum? Say we regard smaller values as noise. Fix $\epsilon > 0$. We want to construct a new function g that only has peaks in the power spectrum above ϵ

$$g(t) = \sum_{n=-\infty}^{\infty} g_n e^{int}$$

with $g_n = \widehat{f}(n)$ if $|\widehat{f}(n)|^2 > \epsilon$ and $g_n = 0$ otherwise.

The decay of the Fourier coefficients is related to the smoothness of the function. See [15] for more details. In particular we have,

$$|\widehat{f}(n)| \leq \frac{\sup_t |f^{(k)}(t)|}{|n|^k}.$$

So if we choose

$$N = \left\lceil \frac{\sup_t |f^{(k)}(t)|^{1/k}}{\epsilon^{1/2k}} + 1 \right\rceil$$

where $[\cdot]$ is the integer part of the number, then g is supported on the interval $[-N, N]$. Define the set $J \subset [-N, N] \cap \mathbb{Z}$ where $j \in J$ if and only if $|\widehat{f}(j)|^2 > \epsilon$. Then we can rewrite g as

$$g(t) = \sum_{j \in J} \widehat{f}(j) e^{ijt}.$$

We can think of g as a truncation of f. Fixing ϵ determines how large N should be and which coefficients should be included. And we also get the embedding dimension $M = 2|J|$.

We now want to evaluate how much of the energy, i.e. the L^2 norm , of the signal is maintained when constructing g. The following two identities will be useful to this end.

Proposition 4.2.1 *If $f \in L^2(\mathbb{T})$,*

$$||f||_2^2 = ||S_N f||_2^2 + ||R_N||_2^2.$$

Proposition 4.2.2 *(Parseval's identity) If $f \in L^2(\mathbb{T})$,*

$$||f||_2^2 = \sum_{n \in \mathbb{Z}} |\widehat{f}(n)|^2.$$

Using Propositions 4.2.1 and 4.2.2 we can compute all of the terms in

$$\frac{||g||_2}{||f||_2} = \frac{||g||_2}{(||S_N f||_2^2 + ||R_N f||_2^2)^{1/2}}$$

except $||R_N f||_2^2$ which we must estimate.

Proposition 4.2.3 *(Remainder estimate) The L^2 norm of the remainder is bounded as follows*

$$||R_N f||_2^2 \leq 2(\sup_t |f^{(k)}(t)|)^2 \frac{(N+1)^{1-2k}}{2k-1}$$

.

Proof.

$$\begin{aligned}
||R_N f||_2^2 &= \sum_{n>|N|} |\widehat{f}(n)|^2 \\
&\leq \sum_{n>|N|} \frac{(\sup_t |f^{(k)}(t)|)^2}{|n|^{2k}} \\
&\leq (\sup_t |f^{(k)}(t)|)^2 \cdot 2 \sum_{n=N+1}^{\infty} \frac{1}{n^{2k}} \\
&\leq 2(\sup_t |f^{(k)}(t)|)^2 \int_{N+1}^{\infty} \frac{1}{x^{2k}} dx \\
&\leq 2(\sup_t |f^{(k)}(t)|)^2 \frac{(N+1)^{1-2k}}{2k-1}.
\end{aligned}$$

□

Let $B = 2(\sup_t |f^{(k)}(t)|)^2 \frac{(N+1)^{1-2k}}{2k-1}$. We can now estimate the percentage of the energy is at least

$$\frac{||g||_2}{||f||_2} = \frac{||g||_2}{(||S_N f||_2^2 + ||R_N f||_2^2)^{1/2}} \geq \frac{||g||_2}{(||S_N f||_2^2 + B)^{1/2}} \times 100\%.$$

If we wish to maintain a certain percentage of the energy of the signal, we can tune ϵ accordingly.

Since the goal was to maximize persistence we want to bound the bottleneck distance between the persistence diagrams associated to f and g.

Theorem 4.2.4 *Let $T \subset \mathbb{T}$, $f \in C^k(\mathbb{T}, \mathbb{R})$, $X = SW_{M,\tau} f(T)$, and $Y = SW_{M,\tau} g(T)$. Then,*

$$d_B(dgm(X), dgm(Y)) \leq 2\sqrt{M+1} \left(\sqrt{\epsilon}(2N - |J|) + \sqrt{2}||R_N f^{(k)}||_2 \frac{(N+1)^{1-2k}}{2k-1} \right).$$

Proof.

$$\begin{aligned}
|f(t) - g(t)| &= |\sum_{n \notin J} \widehat{f}(n) e^{int}| \\
&\leq |\sum_{n \in [-N,N],\, n \notin J} \sqrt{\epsilon}| + |R_N f(t)| \\
&\leq \sqrt{\epsilon}(2N - |J|) + \sqrt{2}||R_N f^{(k)}||_2 \frac{(N+1)^{1-2k}}{2k-1}
\end{aligned}$$

By Proposition 4.1.2,

$$||SW_{M,\tau}f(t) - SW_{M,\tau}g(t)|| \leq \sqrt{M+1}||f(t) - g(t)||_\infty$$
$$\leq \sqrt{M+1}\left(\sqrt{\epsilon}(2N - |J|) + \sqrt{2}||R_N f^{(k)}||_2 \frac{(N+1)^{1-2k}}{2k-1}\right)$$

Choosing $\delta > \sqrt{M+1}(\sqrt{\epsilon}(2N - |J|) + \sqrt{2}||R_N f^{(k)}||_2 \frac{(N+1)^{1-2k}}{2k-1}) \implies d_H(X,Y) \leq \delta$. *Letting* δ *approach its lower bound and using Theorem 2.2.21 we get the desired result.* □

We now have all the ingredients to illustrate the method on numeric data which is covered in the next section.

4.3 Application and discussion

We illustrate the method on a synthetic data set generated from $Re(\sum_{n=1}^{5} \widehat{f}(n)e^{2int})$ which is 2-periodic on the interval $t \in [0, 2\pi)$. The coefficients $\widehat{f}(n)$ are chosen uniformly randomly from the unit disk in $\mathbb{C}$. We sample the signal at 50 evenly spaced time points on this interval. Gaussian noise centered at 0 with standard deviation 25% of signal amplitude is added to the sampled signal. This synthetic data set is similar to one found in [14].

Cubic spline interpolation is then used on the signal to get a continuous function with two continuous derivatives. This allows us to match the theory of the previous section with the application. In particular, $f \in C^2$ so we use $k = 2$ in the appropriate bounds. Also, when we are performing the delay embedding we may want to choose the delay τ small enough to require evaluating the function at time points not present in the sampling. Cubic spline interpolation allows us to sidestep this problem.

Fixing $\epsilon = 10$, a user defined parameter, we compute that we need $N = 12$ terms in the Fourier series approximation. See Figure 4.1 and 4.2 for the time series and power spectrum of the sampled signal and truncated signal, respectively. The truncated signal is estimated to maintain at least 93% of the energy of the original signal. And the embedding dimension is determined to be $M = 10$. We

choose $\tau = \frac{2\pi}{120(10+1)}$ for a small enough delay. The point clouds are then centered and normalized as in [14]. The persistence diagrams are then computed for $\mathbb{F}_{11}$ coefficients along with the bottleneck distance $d_B(dgm1, dgm2) = 0.112062$. See Figure 4.3. The bound on $d_B(dgm1, dgm2)$ implied by Theorem 4.2.4 is rather generous since $2\sqrt{10+1}$ is already much larger than 0.112062. In any case, what we have shown is that it is possible to filter out less significant peaks from a signal's power spectrum without altering the persistent homology much, with respect to the bottleneck distance. In doing so, we have also chosen a minimum embedding dimension.

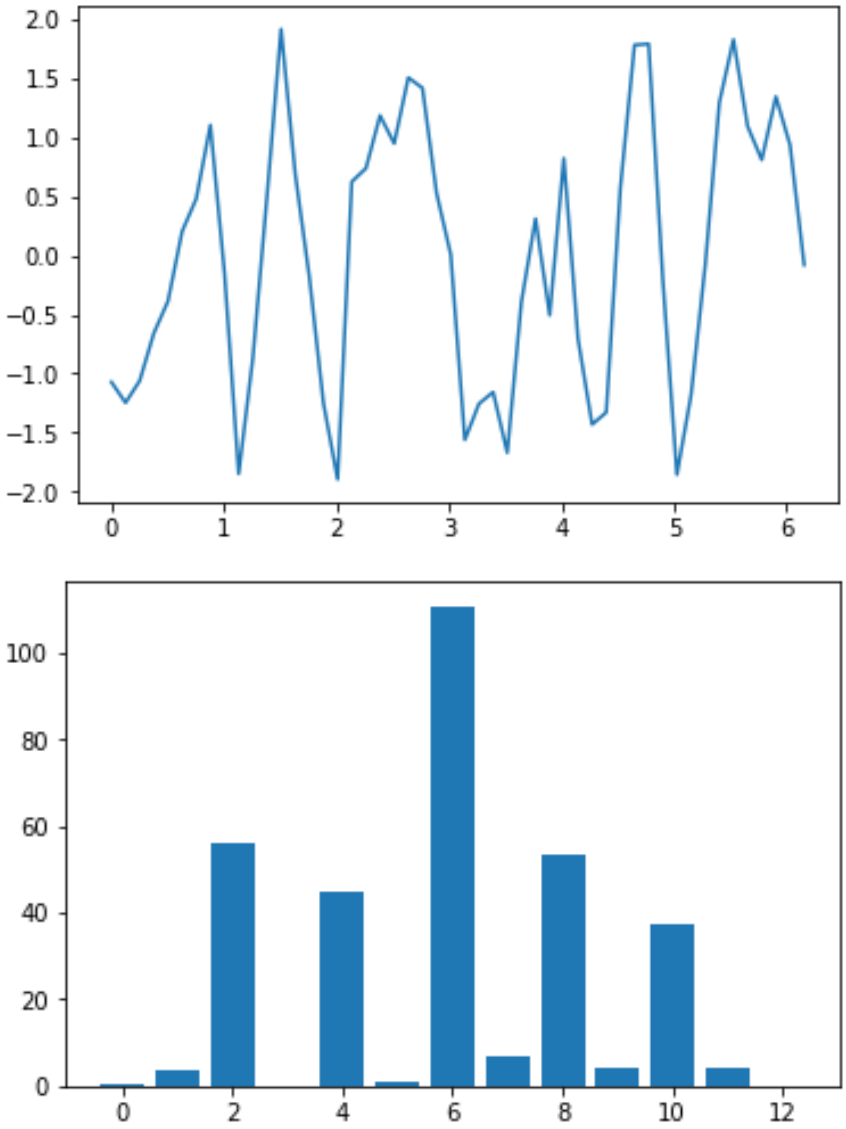

Figure 4.1: Pictured (above) is the sampled noisy signal. The power spectrum (below) appears to have many small peaks.

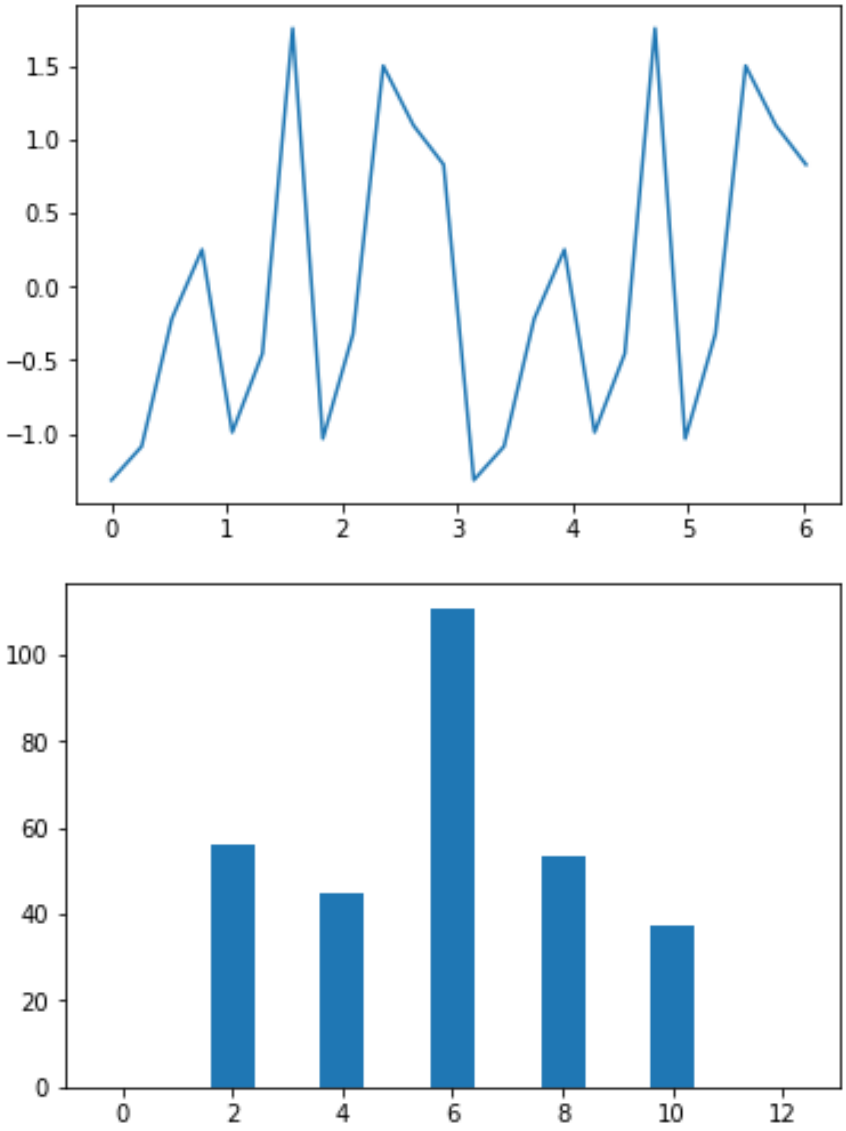

Figure 4.2: The truncated power spectrum (below) corresponds to a cleaner signal pictured (above).

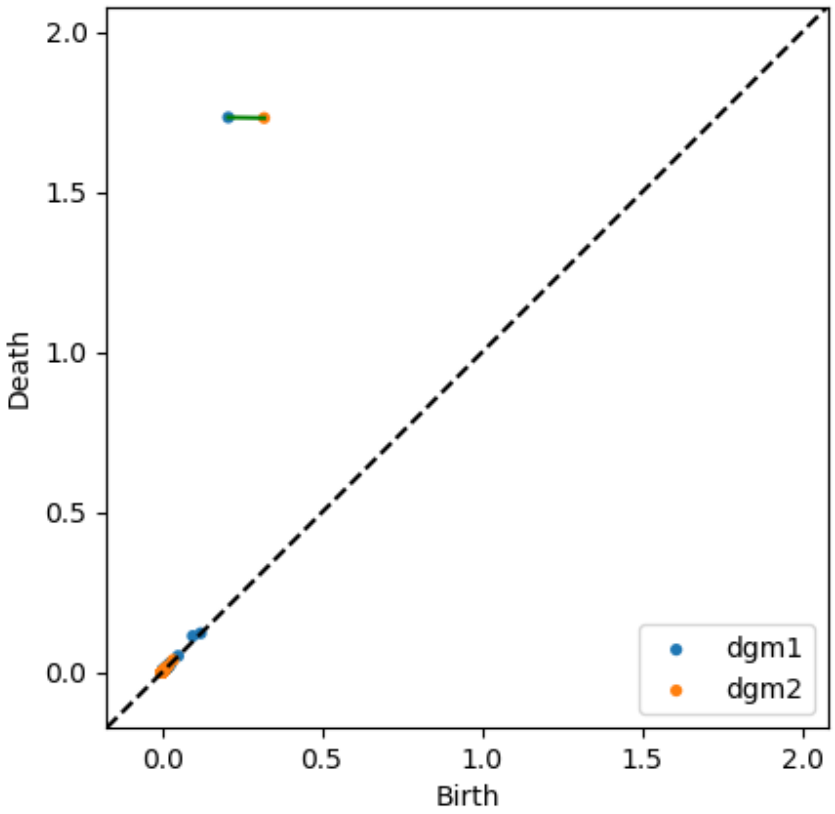

Figure 4.3: The persistence diagrams are superimposed to show the matching done to compute the bottleneck distance. Here dgm1 corresponds to the original signal and dgm2 corresponds to the truncated signal.

CHAPTER 5

CONCLUSIONS, DISCUSSION, AND FUTURE WORK

The main conclusion of this thesis is that we can use persistent homology to study time series if we have a geometric understanding of their delay embeddings. Takens' theorem gives us a topological guarantee which, while helpful, is not sufficient for every purpose. Periodic time series have a nice, geometric form when embedded so we can use persistent homology successfully.

The theoretical results in Chapter 4 can be applied to more data sets for testing. While the theory is interesting in its own right, we would like further evidence that the bounds derived are useful in practice. Additionally, we could consider alternative ways of determining peaks in the power spectrum such as peak-finding algorithms or considering statistical properties of the power spectrum.

It is still unknown what we can say theoretically about the persistent homology of time series which are not periodic or quasi-periodic. This is perhaps the most interesting idea for future work.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Galka Andreas. *Topics in nonlinear time series analysis, with implications for EEG analysis*, volume 14. World Scientific, 2000.

[2] Liangyue Cao. Practical method for determining the minimum embedding dimension of a scalar time series. *Physica D: Nonlinear Phenomena*, 110(1-2):43–50, 1997.

[3] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[4] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, Jan 2007.

[5] Herbert Edelsbrunner and John Harer. *Computational topology: an introduction*. American Mathematical Soc., 2010.

[6] Joshua Garland, Elizabeth Bradley, and James D. Meiss. Exploring the topology of dynamical reconstructions. *Physica D: Nonlinear Phenomena*, 334:49 – 59, 2016. Topology in Dynamics, Differential Equations, and Data.

[7] JP Huke. Embedding nonlinear dynamical systems: A guide to takens' theorem. 2006.

[8] Matthew B. Kennel, Reggie Brown, and Henry D. I. Abarbanel. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Phys. Rev. A*, 45:3403–3411, Mar 1992.

[9] Slobodan Maletić, Yi Zhao, and Milan Rajković. Persistent topological features of dynamical systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 26(5):053105, 2016.

[10] Mark J. McGuinness. The fractal dimension of the lorenz attractor. *Physics Letters A*, 99(1):5 – 9, 1983.

[11] MR Muldoon, RS MacKay, JP Huke, and DS Broomhead. Topology from time series. *Physica D: Nonlinear Phenomena*, 65(1-2):1–16, 1993.

[12] James R. Munkres. *Elements of algebraic topology*. Addison-Wesley, 1984.

[13] Steve Y Oudot. *Persistence theory: from quiver representations to data analysis*, volume 209. American Mathematical Society Providence, RI, 2015.

[14] Jose A Perea and John Harer. Sliding windows and persistence: An application of topological methods to signal analysis. *Foundations of Computational Mathematics*, 15(3):799–838, 2015.

[15] María Cristina Pereyra and Lesley A Ward. *Harmonic analysis: from Fourier to wavelets*, volume 63. American Mathematical Soc., 2012.

[16] Tim Sauer, James A. Yorke, and Martin Casdagli. Embedology. *Journal of Statistical Physics*, 65(3):579–616, Nov 1991.

[17] Floris Takens. Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980*, pages 366–381. Springer, 1981.

[18] Andreas S Weigend. *Time series prediction: forecasting the future and understanding the past*. Routledge, 2018.