8-2017

Data Analysis Methods using Persistence Diagrams

Andrew Marchese
University of Tennessee, Knoxville, amarche1@vols.utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Andrew Marchese entitled "Data Analysis Methods using Persistence Diagrams." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Mathematics.

Vasileios Maroulas, Major Professor

We have read this dissertation and recommend its acceptance:

Jan Rosinski, Michael Langston, Haileab Hilafu

Accepted for the Council:

Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Data Analysis Methods using Persistence Diagrams

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Andrew Marchese

August 2017

To my fiancée Megan, whose support and inspiration drove me to succeed.

Acknowledgments

I would like to express my utmost gratitude to my mentor Professor Vasileios Maroulas. Your insight, encouragement, and direction have been invaluable in helping me grow as a researcher and a person. I would also like to express gratitude to each of my committee members. Professor Jan Rosinski has provided valuable insight and feedback during my entire graduate career. Professor Haileab Hilafu has been very supportive and I appreciate our useful conversations and his feedback. Professor Michael Langston helped me to view problems from a different computational and algorithmic point of view and has provided helpful support. Additionally, I want to express gratitude to Dr. Chris Symons for both academic and financial support. Your support and feedback have been instrumental in my success.

Of course, I would not be here without the constant support and motivation of my parents, Terry and Charlie, my brother and sister, Nicholas and Liana, my sister-in-law, Emily, my future mother-in-law, Debbie, my grandmothers, Joan and Angela, and my nephew, Artie. Your encouragement during this endeavor has been unmatched. Moreover, I want to thank my colleagues and friends, including but certainly not limited to Mark Clark, Kai Kang, Brian Allen, Ernest Jum, Eddie Tu, Chase Worley, Gregory Schmidt, Grace McClurkin, Kevin Sonnanburg and many more. A special thank you to Joshua Mike for his collaboration, insight, and friendship over the past few years. I would also like to extend my thanks to both Gary Kulik and Professor Lisa Berger, whose mentorship led me to pursue a career in Mathematics.

Lastly but certainly not least, I would not be where I am today without the seemingly endless support, encouragement, and patience of my fiancée Megan Margino. You have been

a constant source of motivation and optimism, and I am deeply grateful to have you in my life.

Abstract

In recent years, persistent homology techniques have been used to study data and dynamical systems. Using these techniques, information about the shape and geometry of the data and systems leads to important information regarding the periodicity, bistability, and chaos of the underlying systems. In this thesis, we study all aspects of the application of persistent homology to data analysis. In particular, we introduce a new distance on the space of persistence diagrams, and show that it is useful in detecting changes in geometry and topology, which is essential for the supervised learning problem. Moreover, we introduce a clustering framework directly on the space of persistence diagrams, leveraging the notion of Fréchet means. Finally, we engage persistent homology with stochastic filtering techniques. In doing so, we prove that there is a notion of stability between the topologies of the optimal particle filter path and the expected particle filter path, which demonstrates that this approach is well posed. In addition to these theoretical contributions, we provide benchmarks and simulations of the proposed techniques, demonstrating their usefulness to the field of data analysis.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In recent years, many signal analysis problems in defense, medicine, computer vision, and signal processing have seen machine learning and data centric approaches to solutions Garrett et al. [40], Sherwin and Sajda [78], Azimi-Sadjadi et al. [7], Dhanalakshmi et al. [28]. Garrett et al. [40] use support vector machines, neural networks, and linear discriminant analysis alongside with features derived from Fourier analysis to classify electroencephalogram and electrocardiogram (EE and EKG) signals. In Sherwin and Sajda [78], the researchers use anomaly detections techniques on EEG signals to detect a difference between musical expects and nonexperts. Azimi-Sadjadi et al. [7] and Dhanalakshmi et al. [28] attempt to classify audio signals using support vector machines and other machine learning techniques.

These types of supervised and unsupervised learning techniques are helpful for specific tasks such as threat detection, disease detection, object recognition, and pattern recognition Srinivas et al. [79], Zhang et al. [89]. Though these classical machine learning approaches produce sufficient results on certain types of data, they rely on extracting features from the data, usually in the form of Fourier coefficients Xu et al. [88], Sahidullah and Saha [72]. While this type of feature extraction is commonplace, this type of method doesn't take into account the *shape* and *geometry* of the underlying data, properties that can be important to the identification of important signal features such periodicity and chaos Venkataraman et al. [84], Perea and Harer [65]. Due to this, even when an algorithm based on these techniques produces good accuracy, the resulting algorithm can be difficult to interpret.

In order to solve this issue, researchers have taken to utilizing topology to study data in a direction commonly known as topological data analysis Carlsson [21], Carlsson et al. [20]. This type of technique offers a new way to look at data. By treating the underlying dataset as a point cloud and analyzing its topology, researchers are able to draw conclusions about the data set. This type of analysis has seen successful use in the context of supervised machine learning in the fields of medicine, social sciences, and text recognition Nicolau et al. [63], Adcock et al. [2], Lum et al. [53]. For example, in Adcock et al. [2], the researchers extract relevant topological features in conjunction with support vector machines to classify written digits. Due to the intuitive nature of topological data analysis, it is possible to detect properties such as periodicity directly for classification purposes, even when the periodicity may differ widely with respect to period and shape Perea et al. [66].

In terms of signal analysis, the signal is first embedded into a higher dimension using delay embedding theory Takens [81]. The parameters and robustness of this type of delay are well studied. In particular, it has been shown that there are optimal parameters for reconstructing the phase space of the signal for certain types of signals Perea and Harer [65]. Moreover, many studies have empirically shown that the appropriate delay parameter can be estimated by minimizing the autocorrelation of the signal Venkataraman et al. [84], Emrani et al. [35].

Once the signal has been embedded into a k-dimensional point cloud, a topological object is created via one of several different techniques Edelsbrunner and Harer [32]. These techniques, such as Čech Complex and Vietoris-Rips Complex, are built through constructing relevant open sets (in our case, open spheres) around each data point, and using properties regarding their intersection to derive a simplicial complex. This simplicial complex has topological properties regarding the amount of connected components, holes, and voids it contains. By varying a scale parameter of the open sets in these constructions, a sequence of simplicial complexes are obtained for a grid of growing scale, leading to information about how long (with respect to scale) objects such as holes and connected components persist. This persistence information is vital for classification and clustering, and contains information regarding the shape of the underlying data Venkataraman et al. [84], Pereira and

de Mello [67], Marchese and Maroulas [55]. This topological information is then summarized in a persistence diagram.

These persistence diagrams, which can be thought of as a collection of points in $\mathbb{R}^2$, capture important topological and geometric information about the underlying signal Perea et al. [66], Venkataraman et al. [84], Emrani et al. [35]. A point (x, y) in the persistence diagram indicates a *topological feature* born at scale x and persisting until scale y. For example, a 0-dimensional topological feature is a connected component or cluster, a 1-dimensional topological feature is a hole, and so on. The scale y indicates when this hole fills in. The persistence of these topological features is crucial in determining the behavior of the underlying signal and the dynamic generating it. For example, a very persistent 1-dimensional hole corresponds to periodicity in the dynamic, a very persistent 2-dimensional hole may correspond to chaos in the system Emrani et al. [35], Venkataraman et al. [84], Edelsbrunner and Harer [32], Perea and Harer [65]. Small persistence, on the other hand, can indicate difference in the geometry of the underlying system. An example of a persistence diagram can be seen in Fig. 1.1.

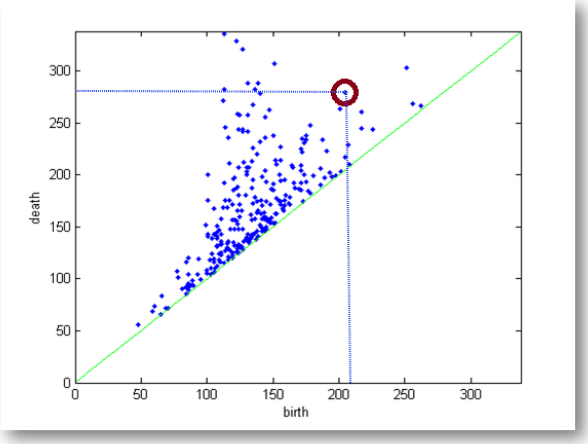


Figure 1.1: An example of a persistence diagram Mischaikow and Nanda [61].

Persistence diagrams have recently seen intense active research, including significant successful effort toward facilitating previously challenging computations with them; for example, Wasserstein distance in Kerber et al. [50] or the creation of persistence diagrams with packages such as Dionysus Morozov [62] and Ripser Bauer [10] take advantage of certain

properties of simplicial complexes Chen and Kerber [24]. Moreover, various approaches to define analogues of traditional statistics have been presented. For example, the relationship between persistence diagrams and the noise of the underlying system is studied in Adler et al. [3]. This work demonstrates the lack of noise-induced topological features for large enough feature death value (scale) under Gaussian noise as the number of sample points goes to infinity, indicating the importance of small persistence points in handling noise in the system. The studies Mileyko et al. [59], Turner et al. [83] introduce notions of mean and variance for finite collections of persistence diagrams. The work in Emmett et al. [34] further summarizes the data by extracting kernel density estimates for the distribution of birth and death times, which in turn produces a very rough marginal distribution of a certain homological dimension. This method loses information about individual features in persistence diagrams. A notion of confidence set for persistence diagrams is introduced in Fasy et al. [37], and in Bobrowski et al. [15], kernel densities are introduced for the purposes of data analysis, but the distribution of the underlying persistence diagrams is not explored.

Some researchers use these persistence diagrams for classification by first extracting features from them and subsequently using these features in classical machine learning algorithms such as support vector machines and random forests Pereira and de Mello [67], Emrani et al. [35], Zhu [91]. For example, Emrani et al. [35] extracts information regarding the most persistent hole in the data, and uses this alone to classify breathing signals. However, this type of approach necessarily loses information in the persistence diagram and must make assumptions about which type of features are important. To address this issue, it is beneficial to use the entire persistence diagram in the learning process. Venkataraman et al. [84] introduces a 1-Nearest Neighbor classifier using the Wasserstein distance to analyze the classification problem on an action recognition dataset. While this may be viable for classification, the clustering problem in this framework remains unexplored. In Chapter 3, we present a direct solution to this problem by introducing a novel clustering algorithm on the space of persistence diagrams utilizing the notion of Fréchet means, a generalization of centroids. Moreover, we present results for this clustering scheme and compare to signal feature and signal distance based methods in Chapter 4.

While the Wasserstein distance may be ideal in some situations, it is important to note that it loses information about the cardinality of the persistence diagrams and the behavior of small persistence points, that contain relevant information regarding the geometry of the underlying data. We propose a new distance on the space of persistence diagrams taking this into account. This distance is inspired by the study of point processes and characterizes the two important notions of persistence diagram distance - set membership and point location Schuhmacher et al. [75]. Moreover, we explore the notion of means of persistence diagrams under this new distance. Because we cannot simply compute means of persistence diagrams in the traditional sense, we consider a more general Fréchet mean. We show that this mean is well defined and that it is guaranteed to exist in most real data situations. Theoretical underpinnings for this distance and its corresponding Fréchet mean are presented in Chapter 3. Moreover, classification benchmark results on real and synthetic datasets are presented in Chapter 4.

Finally, we will draw a connection between these topology based data analysis schemes and stochastic filtering techniques. In some situations, data is observed in a lower dimension than that of the true dynamics, yet we would still like to perform inference on the original space. In order to take this Hidden Markov Mode problem, we must first estimate the true state of the system using the observed state. This is known as the filtering (or smoothing) problem Cappé et al. [19], Xiong [87], Maroulas and Xiong [58]. We will consider the particle filter, a certain type of filtering algorithm that makes very loose assumptions on the underlying systems dynamics and noise Gordon et al. [42], Ren et al. [70], Doucet et al. [29]. This filtering technique works through drawing many samples, called particles, from the prior probability distribution of the true state, and propagating these particles through the dynamics in order to estimate the posterior distribution of the underlying data, given the available observations. This type of filtering technique has seen applications in many fields such as target tracking and biology Maroulas and Nebenführ [57]. In particular, we show in Chapter 5 that the topological information contained in the optimal particle filter path and the expected particle filter path are similar. This shows that observing the persistent homology of the mean particle filter path is well posed, in the sense that this persistent homology is close to the persistent homology of the optimal path. Moreover, we provide

simulations demonstrating the connection between persistent homology and particle filtering techniques.

Via the techniques and results explained above, this dissertation provides a holistic analysis of the learning problem on the data space of persistence diagrams. We consider both supervised and unsupervised learning, providing a framework to researchers in many different areas and applications. In particular, our unsupervised learning framework in Chapter 3 is the first of its kind on the space of persistence diagrams, and opens the way for a new way to study signals and dynamical systems. Additionally, through the new distance introduced in Chapter 3, we provide a concrete way to measure small changes in the geometry of dynamical systems that have been hinted at in previous research, but not properly formalized. We show that these small changes in geometry are essential to the classification problem. These theoretical algorithms are implemented and benchmarked in Chapter 4 against real and synthetic datasets. Moreover, we draw a connections between analyzing samples of a dynamical system and analyzing an estimate for the ground truth of the system itself, through the stability result and simulations in Chapter 5.

Chapter 2

Computational Topology Preliminaries

This chapter will introduce computational topology preliminaries that will be needed for the contributions discussed in later chapters. An interested reader can find a more detailed introduction to the field of computational topology in Edelsbrunner and Harer [32]. In order to perform topological analysis on a signal, it is first necessary to transform the signal into a higher dimension through Takens' delay embedding Takens [81]. First, we give some topological preliminaries.

Definition 1. *Two topological spaces T and H are homeomorphic if there is a continuous bijection $f : T \to H$ such that $f^{-1} : H \to T$ is also continuous.*

Definition 2. *A $d-$dimensional manifold M is a topological space T such that for for every point $x \in T$, there is a neighborhood around x homeomorphic to the open $d-$ball in Euclidean space.*

Definition 3. *A map $f : M \to N$ between two manifolds is said to be a diffeomorphism if it is differentiable and its inverse f^{-1} is differentiable. In general, if f is ℓ times differentiable it is called a C^{ℓ} diffeomorphism.*

Definition 4. *A map $f : T \to H$ between topological spaces in an embedding if f gives a homeomorphism between T and $f(T)$.*

Theorem 2.1 ([81]). *Let M be a compact manifold of dimension d and let C^2 denote the space of functions which are twice differentiable and have continuous first and second derivatives. For a C^2 diffeomorphism $\phi : M \to M$ and C^2 function $f : M \to \mathbb{R}$, it is a generic property that $\Phi_\phi : M \to \mathbb{R}^{2m+1}$ given by*

$$\Phi_\phi(x) = (f(x), f(\phi(x)), ..., f(\phi^{2m}(x))) \tag{2.1}$$

is an embedding.

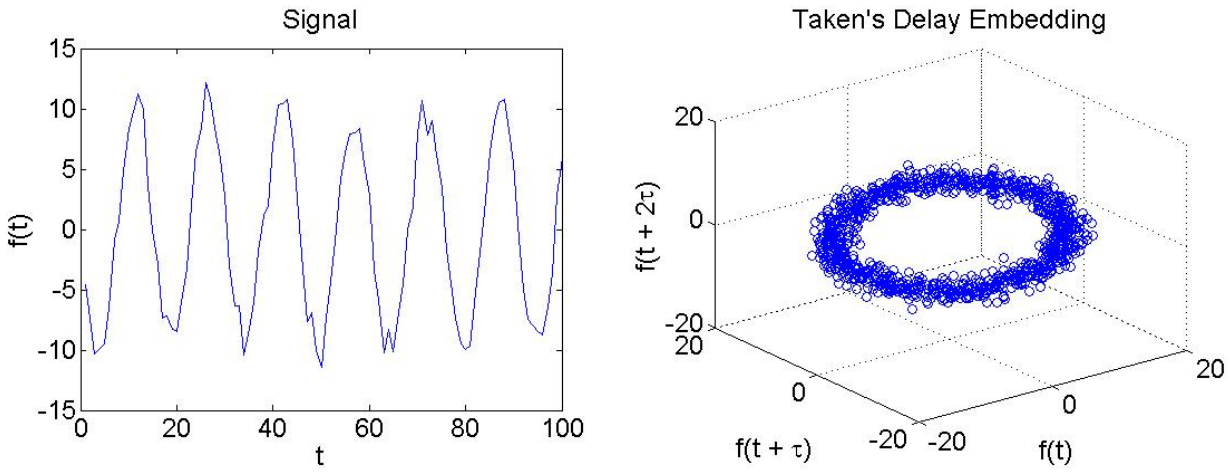


Figure 2.1: Above we depict the signal (left) and its corresponding Taken's delay embedding (right) for $\tau = 10$ and $d = 3$.

This result guarantees that for a correctly chosen Φ, we can recover the topology of the underlying dynamical system by performing the embedding in Eq. (2.1). We choose Φ to be a delay embedding. In particular, for a time-series $f(t)$, given delay dimension d and delay parameter τ we consider the map

$$f(t) \to (f(t), f(t+\tau), ..., f(t+(d-1)\tau)) \tag{2.2}$$

This type of delay embedding is widely used in the field of topological data analysis Venkataraman et al. [84], Anirudh et al. [5], Zhang et al. [89], Perea et al. [66]. Moreover, the parameters τ and d have been well studied in order to guarantee the underlying topology will be recovered Perea and Harer [65]. An example of this embedding performed on a

signal for $d = 3$ and $\tau = 10$ can be seen in Figure 2.1. Note that the resulting object is a $d-$dimensional point cloud $\mathcal{P}$. At a glance, it is obvious that this object has topological and geometric properties, but it is not immediately clear how these can be quantified. For example, examining the topology of the points directly leads to a trivial topology on the space, as they are a set of separate discrete points. However, there is an obvious torus shape that we wish to recover, including a prominent hole in the center. Notice that if we keep track of the points in our delay embedding, we see that this large hole is a result of the periodicity of the signal; this important observation will be expanded upon later. To fully understand this point cloud's topology, we require persistent homology theory, which will allow us to transform this set of discrete points into a topological summary capturing information about the perceived torus shape and hole in the data.

We now give a brief synopsis of the topological theory required in order to build a family of topological objects from a point cloud of data. Once this family is created, a persistence diagram is created to summarize the information.

Definition 5. *A $k-$**dimensional simplex** σ is a set of $(k+1)$ vertex points, $\{p_i\}_{i=1}^{k+1}$, and all convex combinations of these points:*

$$\sigma = \left\{ \sum_{i=1}^{k+1} \alpha_i p_i : \sum_{i=1}^{k+1} \alpha_i = 1 \text{ and } 0 \leq \alpha_i \leq 1 \right\}.$$

*The convex combination of any subset of $\{p_i\}_{i=1}^{k+1}$ is also a simplex, and is called a **face** of σ.*

Intuitively, a $0-$simplex is a point, a $1-$simplex is a line, a $2-$simplex is a triangle, *et cetera*.

Definition 6. *An (abstract) simplicial complex is a collection of simplices G such that whenever G contains a simplex σ, G also contains all the faces of σ.*

When considering an embedding of an abstract simplicial complex we also require that the intersection of any two simplices is a face of both simplices. For example, two triangles must intersect at a common vertex, a common edge, or not at all. This simplicial complex can be thought of as a higher dimensional generalization of a graph. We observe that this

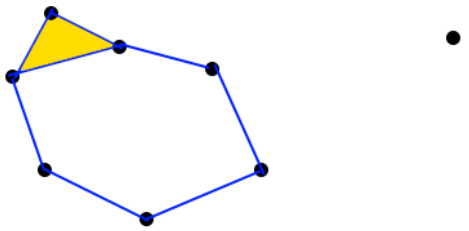


Figure 2.2: An example of a 2−dimensional simplicial complex.

simplicial complex in Figure 2.2 has certain properties. For example, notice that is has two distinct connected components - one containing all of the points in the "cycle", and one containing just one point off to the side. Moreover, notice that the "cycle" forms a one dimensional hole. Our goal is to capture this topological information. In order to do this we must understand the relationship between boundaries and edges in a simplicial complex by considering objects called chain groups. We fix a simplicial complex S and an integral domain I.

Definition 7. *For a field A and integer $j \geq 0$, a formal sum of $j-$dimensional simplices is an object of the form $\sum a_j \sigma_j$, where $a_j \in A$ and σ_j is a $j-$simplex.*

Definition 8. *The j^{th} chain group $C_j(K)$ is the module of formal sums of the $j-$dimensional simplices in S with coefficients in I.*

When the simplicial complex K is obvious from context or not needed, we may simply refer to $C_j(K)$ as C_j. By considering a sequence of maps between these chain groups, we obtain information about the structure of S.

Definition 9. *The j^{th} boundary map $\partial j : C_j \to C_{j-1}$ defined by*

$$\partial j(p_0, ..., p_j) = \sum_{i=0}^{j} (-1)^i (p_0, ..., p_{i-1}, p_{i+1}, ..., p_j),$$

where a j-simplex is mapped to the alternating sum of its $j-1-$dimensional faces.

We call formal sums of simplicial complexes $j-$chains for homological dimension j. Moreover, define the boundary of a $j-$chain to be the formal sum of its $j-1$-dimensional faces. The following result holds Edelsbrunner and Harer [32].

Lemma 2.1.1. *$\partial_{j+1} \circ \partial_j = 0$ for all integers $j \geq 0$.*

Thus we have a chain complex $C_0 \xleftarrow{\partial_1} C_1 \xleftarrow{\partial_2} ... \xleftarrow{\partial_\ell} C_\ell$ such that $\text{Im}(\partial_{j+1}) \subseteq \ker(\partial_j)$, where ℓ is the maximum dimension simplex in S, and Im and ker are the image and kernel of the maps.

Definition 10. *A subgroup H of a group G is said to be normal if for all $h \in H$ and $g \in G$, $ghg^{-1} \in H$.*

Definition 11. *Let H be a subgroup of a group G. The left cosets of H in G is the set $gH = \{gh : h \in H\}$.*

Definition 12. *For a group G and a normal subgroup H, define the quotient group G/H to be the set of cosets of H in G.*

Given this chain complex and Lemma 2.1.1, it is natural to consider the behavior of the quotient group $\ker(\partial_j)/\text{Im}(\partial_{j+1})$.

Definition 13. *The j^{th} homology module is defined as the quotient group $H_j(S) = \ker(\partial_j)/Im(\partial_{j+1})$. The j^{th} Betti number β_j is the rank of the free part of $H_j(K)$.*

We now examine each of these groups in order to understand them in more depth. Fix a homological dimension k and consider $\ker(\partial_k)$. This group contains all chains q in C_k such that $\partial_k q = 0$. We call such a chain a k-cycle. Now consider $\text{Im}(\partial_k)$. These are all chains that are $k-1$-boundaries of k-chains. We finally come to $H_j(S)$, the group of cycles modded out by boundaries. This group can be thought of as objects that cycles (in the kernel), but are not boundaries (modding out by the kernel). Thus, this corresponds to holes that are not filled in. This contains information regarding structure of S. In particular, the dimension of $H_j(S)$ over I is the number of $j-$dimensional holes in S. For example, for $j = 0$, the dimension of $H_0(S), dim(H_0(S)) = \beta_0$, is the number of connected components in S. For $j = 1$, it is the number of $1-$dimensional holes. In Figure 2.2, the dimension of $H_0(S), dim(H_0(S)) = \beta_0$, is two (two connected components), the dimension of $H_1(S), dim(H_1(S) = \beta_1$, is one (one 1-dimensional hole), and β_j is 0 for all $j \geq 1$. By constructing these boundary maps for a given simplicial complex, we are able to compute the Betti numbers directly.

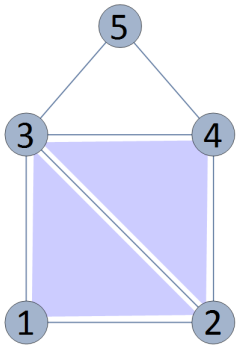


Figure 2.3: A cartoon of the simplicial complex K considered for this example.

Example 1. *Consider the simplicial complex depicted in Fig. 2.3. Notice we only have up to* $2-$*simplicies, yielding the chain complex*

$$\{0\} \xleftarrow{0} C_0 \xleftarrow{\partial_1} C_1 \xleftarrow{\partial_2} C_2 \xleftarrow{\partial_3} C_3 = \{0\}.$$

We write the simplicial complex as

$$\begin{aligned} K &= \{(1),(2),(3),(4),(5),(1,2),(1,3),(2,3),(2,4),(3,4),(3,5),(4,5),(1,2,3),(2,3,4)\} \\ &= \{v_1, v_2, v_3, v_4, v_5, e_1, e_2, e_3, e_4, e_5, e_6, e_7, f_1, f_2\} \end{aligned}$$

and use this simplex ordering to choose bases for each chain group in order to describe each boundary map as a matrix, as shown below:

$$\partial_1 = \begin{bmatrix} -1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & -1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & -1 & -1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}, \quad \partial_2 = \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 1 & -1 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

This organized presentation of the boundary maps enables direct computation of Betti numbers. In this case, $\ker(\partial_1)$ *is 3-D,* $\mathrm{im}(\partial_2)$ *is 2-D, and so* H_1 *is 1-D and* $\beta_1 = 1$*. Here,* H_1 *is generated by the cycle* $(3,4)+(4,5)+(5,3) = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & -1 & 1 \end{bmatrix}^T$*. Similarly,* H_0

is 1-D and $\beta_0 = 1$*, generated by the lone connected component* $(1) + (2) + (3) + (4) + (5) = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \end{bmatrix}^T$.

With the computational topology theory established, we are now ready to analyze the point cloud object in Figure 2.1. We need a method for transforming a cloud of points to a family of simplicial complexes. There are several constructions to this end.

Definition 14. *For some topological space* T*, and some covering of* T*,* $U = \{U_\alpha\}_{\alpha \in A}$ *for some indexing set* A*, the Čech complex* $C(U)$ *is the simplicial complex whose vertex set is* A *and a set of points* $\{\alpha_0, ..., \alpha_k\}$ *is a simplex if and only if* $\bigcap_{i=0}^{k} U_{\alpha_i} \neq \emptyset$*.*

For example, in Definition 14, consider T to be a collection of points in $\mathbb{R}^n$ and the covering U to be spheres of radius $\epsilon/2$ around each point. In this case, we use the notation $C_\epsilon(T)$ for the Čech complex. With this construction, we simply put a k-simplex between $k+1$ points if their $\epsilon/2$ radius spheres have a common nonempty intersection. Figure 2.4 shows an example of a Čech complex construction.

Definition 15. *Given open sets* O_i *contained in* T*, the nerve* N *of* T *is an abstract simplicial complex defined as* $N = \{O_i \subseteq T | \bigcap O_i \neq \emptyset\}$

Definition 16. *A homotopy between continuous functions* $f : T \to H$ *and* $g : T \to H$ *is a continuous map* $\gamma : T \times [0, 1] \to H$ *such that* $H(x, 0) = f$ *and* $H(x, 1) = g$*. If this holds,* f *and* g *are said to be homotopic.*

Definition 17. *Two topological spaces* T *and* H *are homotopy equivalent if there are continuous maps* $f : T \to H$ *and* $g : H \to T$ *such that* $g \circ f$ *and* $f \circ g$ *are homotopic to the identity.*

This complex forms the nerve of T for a fixed radius ϵ. Precisely, the Nerve Theorem conveys the connection between the Čech complex and the topology of T Hatcher [44], Edelsbrunner and Harer [31], i.e. that the Čech complex and T have the similar topology in some sense (homotopy equivalent). Thus, if we want to understand the space T, one needs to study the topology of the corresponding Čech complex. However, due to the computational limitations, sometimes we consider an approximation to the Čech complex, the Vietoris-Rips complex.

Definition 18. *The Vietoris-Rips complex of a finite subset S of a metric space T for some ϵ is the simplicial complex with vertices S such that $\{p_0, ..., p_j\}$ is a simplex if and only if $\rho(p_i, p_v) < \epsilon$ for $0 \leq i, v \leq k$, where $\rho(\cdot, \cdot)$ is the metric on T.*

In contrast to the Čech complex, in the Vietoris-Rips complex a $k-$simplex is created on points $p_0, ..., p_{k-1}$ when for each pair of vertices the balls around each pair of points has nonempty intersection. These two constructions are related, in the sense that the Čech complex can be approximated with the Vietoris-Rips complex. Consider a Čech complex for a fixed radius $\epsilon/2$ denoted C_ϵ and a Vietoris-Rips complex for a fixed radius ϵ denoted VR_ϵ. Then the following approximation holds.

Lemma 2.1.2 (Edelsbrunner and Harer [32])**.** *For a fixed radius ϵ, we have that $C_\epsilon \subseteq VR_\epsilon \subseteq C_{2\epsilon}$*

Lemma 2.1.2 says that the Vietoris-Rips complex serves as an approximation to the Čech complex, and so we may consider the Vietoris-Rips complex in practice, as it is easier to compute. Observing Definitions 14 and 18 requires a fixed value of ϵ. However, it is relevant for the sake of classification to consider what these complexes look like for an increasing sequence of such ϵ values. In particular, we are able to track the presence of $n-$dimensional voids and connected components as we vary the radius ϵ, yielding invaluable information about the topology and geometry of the underlying data. In particular, consider an increasing sequence $\epsilon_1 < ... < \epsilon_j$ such that we have an associated sequence of nested simplicial complexes, $C_{\epsilon_1} \subseteq ... \subseteq C_{\epsilon_j}$. This sequence of complexes is called a filtration Edelsbrunner and Harer [31]. For each simplicial complex C_{ϵ_i}, there are associated homology groups $H_n(C_{\epsilon_i})$. We consider these homology groups over the field $\mathbb{Z}_2 = \mathbb{Z}/2\mathbb{Z}$. In particular, it is possible to keep track of topological features (such as connected components and k-dimensional holes) as the radius ϵ changes.

Pertinent topological features such as connected components and holes, corresponding to the Betti numbers for each simplex in the chosen filtration, are kept track of and it is possible to see how long these topological features persist as ϵ grows. Of course, eventually when the radius ϵ is large enough, all pairs of points will be within ϵ and there will be one connected component and no voids. However, what is important is to monitor how

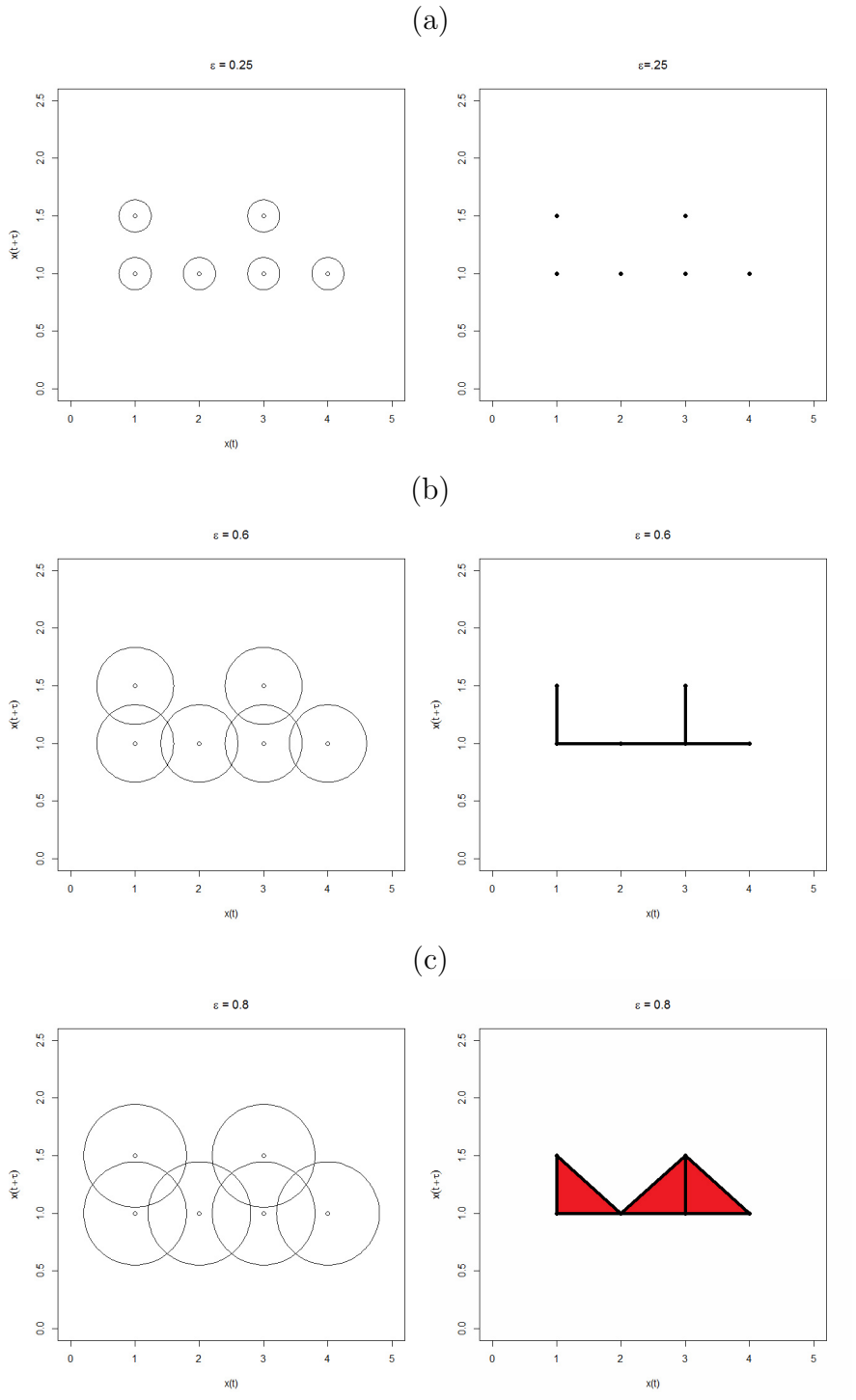


Figure 2.4: The construction of Čech complexes (right) from the point cloud (left) for varying epsilon: (a) $\epsilon = .25$; (b) $\epsilon = .6$; (c) $\epsilon = .8$.

these topological features change over increasing ϵ. We represent the changing topological features in a chosen complex with a persistence diagram. Consider the complex and fix a homological dimension to consider. For example, setting $k = 0$ corresponds to keeping track of the connected components of the complex, i.e. β_0. Suppose a connected component η is "born" at time b_η, and persists until time d_η. The persistence diagram corresponding to a Vietoris-Rips complex consists of birth-death points (b_η, d_η) for all η that exists for $\epsilon \in G$.

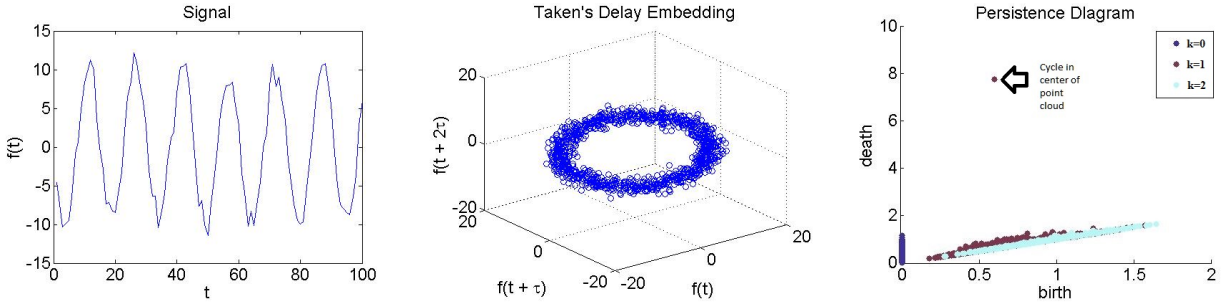


Figure 2.5: Above we depict the signal (left), it's Taken's delay embedding (center), and it's persistent homology (right). Notice the long persistent element corresponding to the cycle in the center of the point cloud.

Thus, a persistence diagram is a set of points $\{(b, d)|b, d \in \mathbb{R}^2 \text{ and } d > b\}$ where each point corresponds to a topological feature in a corresponding family of simplicial complexes. In particular, each point (b, d) denoted a topological features being born at radius b and "dying" at radius d. Here, "dying" can be thought of as a homological feature getting filled in with a lower dimensional simplex. The persistence of a feature is defined as $d - b$, and refers to how "long" (with respect to radius) a topological feature persists before it is filled in. For example, in Figure 2.5, by examining the point cloud in the center image, we expect a hole to form at a relatively small radius and for this hole to persist for a long time. This can be seen in the persistence diagram on the right at the point $(.65, 7.9)$. Note that there is a persistence diagram for each homological dimension k, which we combine onto one image.

Chapter 3

Clustering and Classification

The problem of analyzing signals and implementing learning algorithms to perform inference on them is a well studied field with many applications. These range from medical applications involving EEG data Garrett et al. [40], Sherwin and Sajda [78] to acoustic signals with military Azimi-Sadjadi et al. [7] and non-military applications Dhanalakshmi et al. [28].

Recently, topological techniques have been developed to offer a new way to analyze data e.g. see Carlsson [21], Bubenik [16], Bampasidou and Gentimis [9], Xia and Wei [86], Robins and Turner [71]. These techniques have been used in the field of machine learning, with applications in medicine, social sciences, and text recognition Nicolau et al. [63], Adcock et al. [2], Lum et al. [53]. In particular, topological data analysis techniques excel at analyzing point clouds in $\mathbb{R}^d$ for some dimension d. Through the use of time-delay embeddings, it is possible to transform the signal classification problem into a point cloud classification problem. Inspired by their original use in reconstructing the phase-space of a dynamical system Takens [81], this delay technique has been effectively used to combine topology and statistics Seversky et al. [76]. Leveraging topological theory, the underlying topological properties of the point cloud produced by this delay embedding are analyzed. Precisely, the persistent homology of the point cloud is represented by a persistence diagram, which can be thought of as a multiset of points in $\mathbb{R}^2$, where the (x, y) coordinates signify how long topological properties of the data persist. In this chapter, we will discuss how this persistence diagram can be used for classification and clustering.

3.1 Clustering on the space of Persistence Diagrams

Clustering is the problem of dividing a dataset into some K number of groups based on some measure of similarity [77]. In this section, we will assume that K is known a priori, but we make no assumptions about the labels of persistence diagrams in the training set. Using the underlying topology of the signal and the dynamic generating the signal can give valuable insight into features vital for class differentiation. Although there has been much work focused on extracting features from persistence diagrams for signal classification or classifying directly on the space of persistence diagrams the same cannot be said for clustering. In Pereira and de Mello [67], the authors transform signals into persistence diagrams and subsequently extract features such as maximum persistence hole ($\max_i(d_i - b_i)$), number of holes, and average lifetime ($\frac{\sum d_i - b_i}{n}$) from persistence diagrams and proceed to use these features to train a traditional clustering algorithm; while this is a first step to using persistence diagrams in the clustering of time-series, these features give a very coarse estimate of the persistence diagrams, and in turn give a summary of the persistence diagram, which is already a topological summary. Though there have been studies focusing on clustering and describing a high-dimensional dataset's shape and topology using persistent homology Lum et al. [54], Carlsson [21], the problem of time-series clustering directly on the space of persistence diagrams remains unexplored.

For this reason, we instead focus on developing a clustering technique directly on the space of persistence diagrams. In particular, our goal is to introduce a $K-$means type algorithm on the space of persistence diagrams. In order to achieve this, we need to establish a metric on the space of persistence diagrams, and a notion of "mean" or centers for the persistence diagrams. These concepts are very general and require few assumptions. We start by describing the space of persistence diagrams under a certain metric.

Definition 19. *Given two persistence diagrams* $\mathbb{D}_1$ *and* $\mathbb{D}_2$ *and* $p > 1$*, define the p-Wasserstein distance* $W_p(\mathbb{D}_1, \mathbb{D}_2)$ *by*

$$W_p(\mathbb{D}_1, \mathbb{D}_2) = (\inf_{\gamma} \sum_{x \in \mathbb{D}_1} ||x - \gamma(x)||_\infty{}^p)^{1/p}, \tag{3.1}$$

where γ ranges over all bijections from $\mathbb{D}_1$ to $\mathbb{D}_2$, p is a fixed parameter, and $||z||_\infty = max(|z_1|, ..., |z_m|)$ for $z \in \mathbb{R}^m$.

Definition 20. *Given two persistence diagrams $\mathbb{D}_1$ and $\mathbb{D}_2$, define the Bottleneck distance $W_\infty(\mathbb{D}_1, \mathbb{D}_2)$ by*

$$W_\infty(\mathbb{D}_1, \mathbb{D}_2) = \inf_{\gamma} \sup_{x \in \mathbb{D}_1} ||x - \gamma(x)||_\infty, \tag{3.2}$$

where γ ranges over all bijections from $\mathbb{D}_1$ to $\mathbb{D}_2$, and $||z||_\infty = max(|z_1|, ..., |z_m|)$ for $z \in \mathbb{R}^m$.

In the limit as p goes to infinity, the Wasserstein distance becomes the Bottleneck distance. The Wasserstein distance measures how far off the best possible mapping from $\mathbb{D}_1$ to $\mathbb{D}_2$ is over all points. In order to account for cardinality differences in the persistence diagrams, ad hoc matching to the diagonal is allowed in order to ensure bijections γ between $\mathbb{D}_1$ and $\mathbb{D}_2$ exist (in other words, any number of features with 0 persistence can be added as needed).

It is important to note that the space of all finite persistence diagrams under the Wasserstein distance is not complete Mileyko et al. [59]. For example, consider t_n to be the topological feature with $t_n = (b_n, d_n) = (0, 2^{-n})$ for $n \in \mathbb{N}$, and $\mathbb{D}_n$ the persistence diagram containing $t_1, ..., t_n$. Then $W_p(\mathbb{D}_n, \mathbb{D}_{n+\ell}) < \frac{1}{2^{n+\ell}}$, and so this sequence is cauchy, but the number of points will grow to infinitely as $n \to \infty$, so this sequence does not have a limit in the space. Due to this, when considering the space of persistence diagrams under the Wasserstein distance, we must restrict the space as $P_{wass} = \{\mathbb{D} | W_p(\mathbb{D}, \mathbb{D}_0) < \infty\}$ where $\mathbb{D}_0$ is the persistence diagram only containing the diagonal.

When introducing an algorithm of clustering, it is important to have a notion of center. While it is not obvious how a set of persistence diagrams can be "averaged," we consider the Fréchet mean, which is a generalization of a centroid. Consider a probability measure $\mathcal{D}$ on the space of $(P_{wass}, \mathcal{B}(P_{wass}))$ where $\mathcal{B}(P_{wass})$ is the Borel $\sigma-$algebra on P_{wass} such that

$$F^{W_p}_{P_{wass}}(\mathbb{D}_1) = \int_{P_{wass}} W_p(\mathbb{D}_1, \mathbb{D}_2)^2 d\mathcal{D}(\mathbb{D}_2) < \infty, \tag{3.3}$$

for all $\mathbb{D}_1 \in P_{wass}$.

Definition 21. *Given a probability space* $(P_{wass}, \mathcal{B}(P_{wass}), \mathcal{D})$*, the Fréchet variance of* $\mathcal{D}$ *is*

$$Var_{\mathcal{D}}^{W_p} = \inf_{\mathbb{D} \in P_{wass}} [F_{P_{wass}}^{W_p}(\mathbb{D}) = \int_{P_{wass}} W_p(\mathbb{D}, \mathbb{D}_2)^2 d\mathcal{D}(\mathbb{D}_2)], \tag{3.4}$$

and the Fréchet expectation or Fréchet mean of $\mathcal{D}$ *is*

$$\mathbb{E}^{W_p}(\mathcal{D}) = \{\mathbb{D} | F_{P_{wass}}^{W_p}(\mathbb{D}) = Var_{\mathcal{D}}^{W_p}\}. \tag{3.5}$$

In [59], it was shown that the space of persistence diagrams P_{wass} admits Fréchet means under certain types of probability distributions. In particular, this holds for the empirical distribution over a finite set of persistence diagrams; we will be interested in discrete probability distributions for a set $\{\mathbb{D}_i\}_{i=1}^T$ of persistence diagrams. Consider the uniform distribution $\frac{1}{T}\sum_{i=1}^T \delta_{\mathbb{D}_i}$ over this set. Leveraging this theory, we propose the following $K-means$ type clustering algorithm. The set of Fréchet means for a sample diagram is shown in Fig. 3.1.

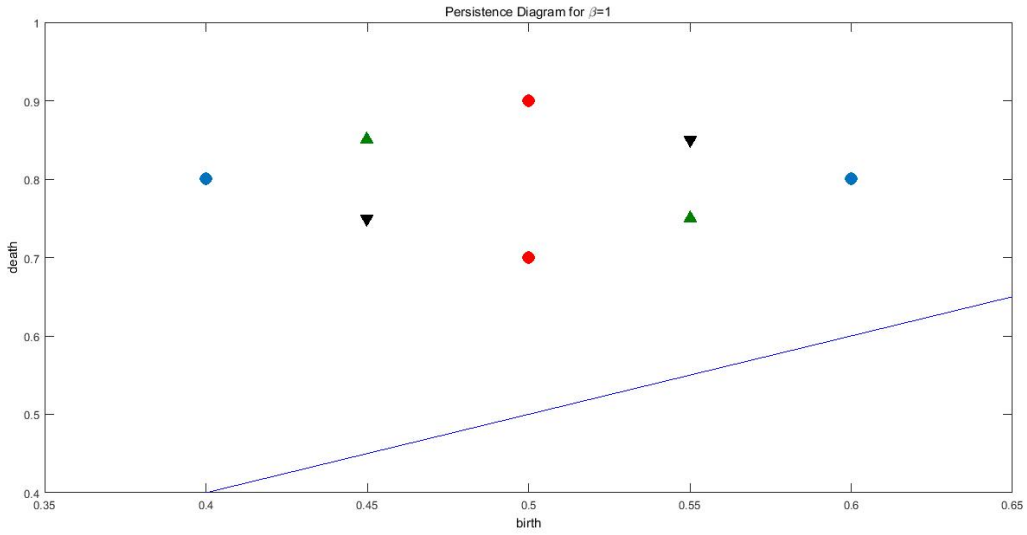


Figure 3.1: Consider four persistence diagrams, with red points, black points, blue points, and green points. We are interested in the Fréchet mean of the red persistence diagram and the blue persistence diagram. Notice that both the green persistence diagram and the black persistence diagram minimize Eq.(3.4).

Start with a dataset of persistence diagrams $D = \{\mathbb{D}_i\}_{i=1}^N$ to be clustered, with a known K a priori number of clusters. First, we will uniformly at random select persistence diagrams

$M_1^{(0)}, ..., M_K^{(0)}$ from D without replacement. These persistence diagrams will act as initial centroids for the algorithm. Then, for each persistence diagram $\mathbb{D}_i \in D$, assign a label $l_i^{(0)}$ according to which centroid persistence diagram it is closest too. At this point, for each cluster $1 \leq J \leq K$, there is a set of persistence diagrams $Diags_J^{(0)}$ associated to it. Now, in order to update the centers $M_1^{(0)}, ..., M_K^{(0)}$, define $M_J^{(1)}$ to be the Fréchet mean of the diagrams in $Diags_J^{(0)}$. This process continues until the labels $l_i^{(t)} = l_i^{(t+1)}$ for some iteration t, for all $1 \leq i \leq N$. Pseudo-code for this algorithm is presented in Alg. 1, and a single iteration of the algorithm is shown in Fig. 3.2.

Algorithm 1 Clustering on the space of persistence diagrams using Fréchet means.

Input: Persistence Diagrams $D = \{\mathbb{D}_i\}_{i=1}^n$, number of clusters K, maxiter.
Training Phase
Initialize Centers
for j=1 to K **do**
    Randomly initialize centroid diagram $M_j^{(0)}$ for cluster j
count = 0
while Not Converged **do**
    for i=1 to n **do** Assign $\mathbb{D}_{x_i}$ label $\ell_i^{(t)}$ where $\ell_i^{(t)} = \text{argmin}_{1 \leq j \leq K} d(\mathbb{D}_{x_i}, M_j^{(t-1)})$
    for j=1 to K **do** Compute $M_j^{(t)} = \text{FréchetMean}(Diags_J^{(t-1)})$ where $Diags_J^{(t-1)} = \{\mathbb{D}_{x_i} | \ell_i^{(t-1)} = j\}$
    count = count + 1
    if $\ell^{(t-1)} = \ell^{(t)}$ for all $1 \leq i \leq n$ OR count==maxiter **then**
        Converged
Return $\{M_j^{(T)}\}_{j=1}^K$, $\{\ell_i^{(T)}\}_{i=1}^N$

In order to quantify the effectiveness of the clustering, we consider the associated cost function to Algorithm 1:

$$G(Diags_1, ..., Diags_K) = \min_{M_1,...,M_K} \sum_{i=1}^{K} \sum_{\mathbb{D}_j \in Diags_i} (W_p(\mathbb{D}_j, M_i))^2 \tag{3.6}$$

This cost function is known as the within-cluster sum of squares Shalev-Shwartz and Ben-David [77], Hastie et al. [43]. In particular, it measures the total sum of squared distances from each persistence diagram to its associated cluster centroid diagram. Of course, it is

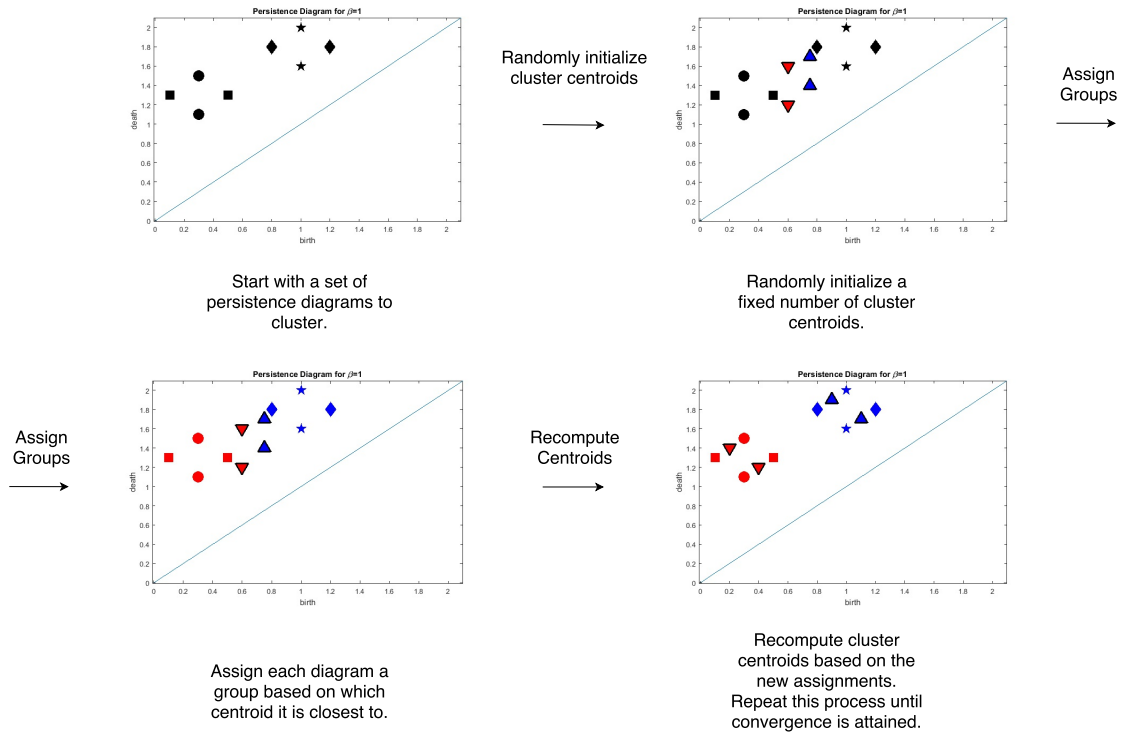


Figure 3.2: A cartoon demonstrating how Algorithm 1 works on a small dataset of persistence diagrams.

not obvious that the labels will stop changing at some iteration t. The following result guarantees that the algorithm converges to a (local) minimum.

Theorem 3.1. *Algorithm 1 converges to a local minimum.*

Proof. The algorithm only changes the value of G defined in Eqn. (3.6) at two steps - the update of the Fréchet Means M and the update of the cluster assignments $Diags_J$. We show that in each of these updates, G cannot increase. First fix an iteration $t \in \mathbb{N}$ and Fréchet Means $M_1^{(t)}, ..., M_K^{(t)}$. We must show that

$$\sum_{i=1}^{K} \sum_{\mathbb{D}_j \in Diags_i^{(t)}} W_p(\mathbb{D}_j, M_i^{(t)})^2 \leq \sum_{i=1}^{K} \sum_{\mathbb{D}_j \in Diags_i^{(t-1)}} W_p(\mathbb{D}_j, M_i^{(t)})^2 \tag{3.7}$$

Eqn. (3.7) clearly holds given that the clusters $Diags_i^{(t)}$ are assigned based on distance to the closest mean $M^{(t)}$. It remains to show the objective function G of Eqn. (3.6) cannot increase when we fix the labels and update the Fréchet Means. Let $M_i^{(t)}$ be defined as

the Fréchet mean of the elements of $Diags_i^{(t-1)}$. Precisely, $M_i^{(t)}$ is a persistence diagram minimizing

$$\sum_{\mathbb{D}_j \in Diags_i^{(t-1)}} W_p(\mathbb{D}_j, M_i^{(t)})^2$$

Thus, by this definition,

$$\sum_{i=1}^{K} \sum_{\mathbb{D}_j \in Diags_i^{(t-1)}} W_p(\mathbb{D}_j, M_i^{(t)})^2 \leq \sum_{i=1}^{K} \sum_{\mathbb{D}_j \in Diags_i^{(t-1)}} W_p(x, M_i^{(t-1)})^2$$

holds. Since the objective function of the algorithm cannot increase at either of the two update steps, and there are only a finite number of possible labellings, the objective function decreases to a local minimum. □

Theorem 3.1 guarantees that the algorithm does not increase the cost of the objective function at every iteration. Because there are only a finite number of permutations of the labels, this guarantees convergence to some labeling which will not change under further runs of the algorithm, which we denote as the local minimum.

For the above clustering scheme, we consider the space of persistence diagrams under the Wasserstein distance. Yet, in order to account for cardinality differences in the persistence diagrams, ad hoc matching to the diagonal is allowed in order to ensure bijections γ between $\mathbb{D}_1$ and $\mathbb{D}_2$ exist (in other words, any number of features with 0 persistence can be added as needed). It is assumed that there are infinitely many points along the diagonal of each persistence diagram with infinite multiplicity Edelsbrunner and Harer [32]. Note that the Wasserstein distance does not explicitly penalize for cardinality differences between persistence diagrams. Differences in cardinality may play an important role in the classification problem. Additionally, allowing mapping to the diagonal can lead to situations where two very different persistence diagrams are considered "close" under the Wasserstein distance. Moreover, through this assumption, the Wasserstein distance necessarily puts a smaller weight on low persistence points. We note some researchers contend that the small persistence points are considered "topological noise" Edelsbrunner et al. [33], Balakrishnan et al. [8], Edelsbrunner and Harer [31], Atienza et al. [6], Pokorny et al. [68]. However, in

line with recent studies, we argue that small persistence points are useful for data analysis Xia and Wei [86], Robins and Turner [71], Bubenik [18]. In the next section, we consider the classification problem, and define a new metric to address these issues.

3.2 Classification on the space of Persistence Diagrams

Whereas in the last section we assumed a dataset of persistence diagrams without any associated labels, in this section we will consider the case of supervised learning. That is, we start with the notion of a dataset of persistence diagrams with a priori knowledge of their associated class labels. Similarly to the clustering framework, several approaches have been taken in order to solve the classification problem using persistence diagrams.

One technique researchers have used to is extract a feature vector $\alpha \in \mathbb{R}^d$ from the persistence diagrams, and use this in conjunction with classical machine learning algorithms (such as neural networks, support vector machines, etc.) Adcock et al. [1], Zhang et al. [89]. Unfortunately, this technique takes a persistence diagram, which is already a topological summary, and further summarizes it into features such as average length of persistence, number of features, etc. This technique necessarily loses information about the underlying topology.

Instead, classification should be done directly on the space of persistence diagrams. Several researchers have tackled the classification problem in this way using the Wasserstein or Bottleneck distance. In Venkataraman et al. [84], the authors implement a nearest neighbor classification algorithm using the Wasserstein distance with $p = 1$ to classify signals arising from motion sensors on the human body. Similarly, Emrani et. al. Emrani et al. [35] classify acoustic breathing patterns using a distance based technique, allowing for the distinction of wheezing and non-wheezing in patients. The authors of Seversky et al. [76] use both the Wasserstein distance and the $p = \infty$ Bottleneck distance to classify signals from a wide variety of applications. In addition, the authors propose treating the persistence diagrams as images and training an SVM classifier using these images. Similarly, the researchers in Anirudh et al. [5] treat persistence diagrams as points on a hypersphere using

a probability density function formulation. This representation is then used for classification of signals on a stroke rehabilitation dataset.

Typically, two main distances have been used in the study of persistence diagrams: the Wasserstein distance Kerber et al. [49], Mileyko et al. [59] and the Bottleneck distance Kerber et al. [49]. In terms of theory, the Wasserstein distance is well studied, as shown in the previous section; however, the Wasserstein distance has several drawbacks that make its use questionable in certain applications. Both of these distances require the construction of a bijection between the two persistence diagrams, but this isn't always possible due to cardinality differences. To get around this, these distances assume infinitely many points with infinite cardinality along the diagonal on each persistence diagram, as discussed in the previous section. This can lead to situations where two very different persistence diagrams are considered close. For example, in Figure 3.3, we compare two persistence diagrams. One has a feature born at time $b = .1$, while the other has a feature born at $b = .8$. While their persistence is the same, the *scale* of the features is extremely different, and in turn these persistence diagrams may be representing two very different point clouds arising from two distinct dynamics. The Wasserstein distance would consider these two topological features to be very similar. Consider the example in Figure 3.4. Notice that the persistence diagrams look very similar in terms of large persistence features. However, taking into consideration the differences in small persistence points, these persistence diagrams are very different.

Second, the Wasserstein distance does not explicitly penalize for cardinality differences between persistence diagrams. While this may not inherently seen like a problem, it can be an insurmountable problem in the classification problem. Consider the example in Fig. 4.7, where two very different dynamics may produce persistence diagrams with very different cardinality. This is extremely important when we look at the generating dynamics. The left diagram is generated by a random walk type dynamic, while the right diagram is generated by a multi-stable dynamic. In the context of classification, it is very important to differentiate between these types of dynamics.

In this section, we propose a distance that addresses this issue by directly mapping all points in a persistence diagram. In particular, the proposed distance takes into consideration the geometry of the underlying point cloud by matching small persistence points to each

other. This distance can be thought of as a point matching (similar to the Wasserstein, but we forbid matches to the diagonal), plus a regularization term considering the cardinalities between persistence diagrams. For lower weight on the cardinality difference, the proposed distance will focus more on small persistence points, related to small geometry changes in the underlying data (see Fig. 4.9). On the other hand, for a larger regularization term, the distance will focus on cardinality differences between the persistence diagrams, related to larger changes in the underlying geometry and topology, potentially caused by dynamics and/or noise differences (see Fig. 4.7) Adler et al. [3].

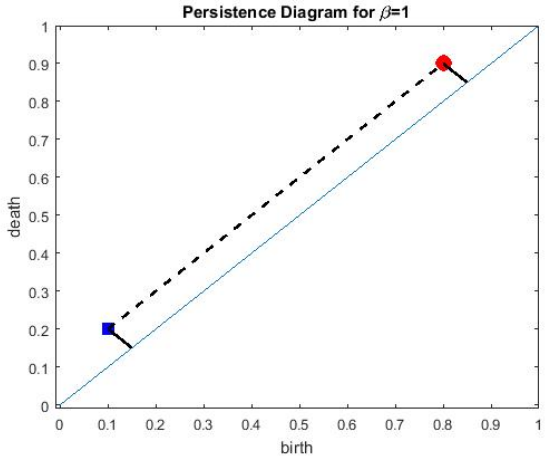


Figure 3.3: Consider two persistence diagrams, one with a point at (.1,.2), indicated by the blue square, and one with a point at (.8,.9), indicated by the red circle. Because these points are so far away, the Wasserstein Distance (indicated by the solid black line) finds the optimal bijection as the one mapping both to the diagonal. This results in a distance of .1 between the two diagrams. However, mapping these points to each other (indicated by the dashed line) may be more useful for classification purposes. This mapping of points to each other results in a distance of .7, and this better represents the differences in the persistence diagrams.

We start by defining the new metric d_p^c, to the space of persistence diagrams (Def. 22). Relying on this distance we propose a classification scheme and we show that under mild conditions, the metric space of persistence diagrams is a Polish space and it admits statistical structure, i.e. Fréchet means and variance.

Definition 22. *Consider two persistence diagrams* $\mathbb{D}_1$ *and* $\mathbb{D}_2$*, with cardinalities* n *and* m *respectively such that* $m \leq n$*. Denote the points in* $\mathbb{D}_1$ *by* $(x_1, ..., x_n)$ *and points in* $\mathbb{D}_2$ *by*

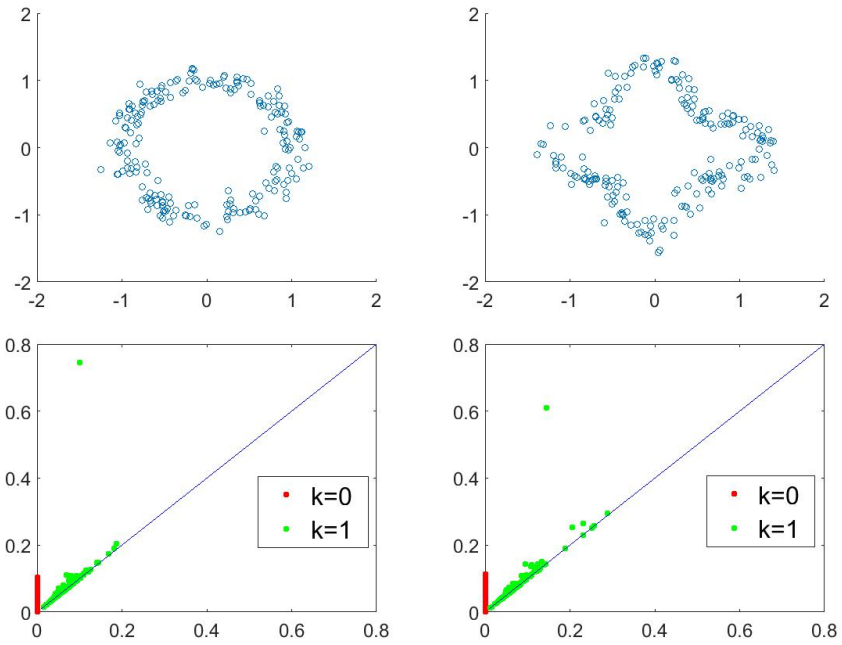


Figure 3.4: *Left:* Noisy small scale ϵ features are paired with a single persistent cycle. *Right:* As left, but now we see several low persistence points at larger ϵ scale, these features alone, which represent the shape of the signal, tell the difference between the two dynamics.

$(y_1, ..., y_m)$. *Let* $c > 0$ *and* $1 < p < \infty$ *be fixed parameters. We define the* d_p^c *distance as*

$$d_p^c(\mathbb{D}_1, \mathbb{D}_2) = (\frac{1}{m}(\min_{\pi \in \Pi_m} \sum_{i=1}^{n} \min(c, ||x_i - y_{\pi(i)}||_\infty)^p + c^p|n-m|))^{\frac{1}{p}} \tag{3.8}$$

where Π_m *is the set of permutations of* $(1, ..., m)$. *If* $|\mathbb{D}_\mathbb{1}| > |\mathbb{D}_2|$, *define* $d_p^c(\mathbb{D}_1, \mathbb{D}_2) := d_p^c(\mathbb{D}_2, \mathbb{D}_1)$.

Proposition 1. *The* d_p^c *in Eq. (3.8) is a metric on the space of persistence diagrams.*

Proof. We adapt the proof from Schuhmacher et al. [75] to the space P_W. According to Definition 22, it is clear we have that $d_p^c \geq 0$ and that d_p^c is symmetric and satisfies the identity. It remains to show the triangle inequality. We consider three persistence diagrams $\mathbb{D}_1 = (t_1, ..., t_\ell), \mathbb{D}_2 = (u_1, ..., u_n), \mathbb{D}_3 = (v_1, ..., v_m)$. Assume that $\ell \leq n$ and that at most one of the cardinalities is zero. Since W is a closed and bounded subset of $\mathbb{R}^2$, we consider some dummy points $(a_i)_{i \in \mathbb{N}}$ and $(b_i)_{i \in \mathbb{N}}$ at least distance c from W and each other. The two cases we must consider are $\ell \leq n \leq m$ and $\ell, m \leq n$.

We first treat the case when $\ell \leq n \leq m$. Extend the persistence diagram $\mathbb{D}_1$ with points $t_{\ell+j} = a_j$ for $1 \leq j \leq m-\ell$ and similarly for $\mathbb{D}_2$ with $u_{n+j} = b_j$ for $1 \leq j \leq m-n$. This way the cardinality difference is equal to zero in Eq. (3.8). Moreover, after the dummy points have been added in, let η and ν be the minimum permutations from $\mathbb{D}_1$ to $\mathbb{D}_3$ and from $\mathbb{D}_3$ to $\mathbb{D}_2$ respectively. Then, according to Eq. (3.8) and $a \leq c^p m$ implying $\frac{a}{m} \leq \frac{a+c^p(n-m)}{n}$, we have

$$d_p^c(\mathbb{D}_1, \mathbb{D}_2) = (\frac{1}{n} \min_{\pi \in \Pi_n} \sum_{i=1}^{n} \min(c, ||t_i - u_{\pi(i)}||_\infty)^p)^{\frac{1}{p}}$$

$$\leq (\frac{1}{m} \min_{\pi \in \Pi_m} \sum_{i=1}^{m} \min(c, ||t_i - u_{\pi(i)}||_\infty)^p)^{\frac{1}{p}} \tag{3.9}$$

The right hand side of Eq. (3.9) can further be bounded by

$$(\frac{1}{m} \sum_{i=1}^{m} \min(c, ||t_i - v_{\eta(i)}||_\infty)^p + \min(c, ||v_i - u_{\nu(i)}||_\infty)^p)^{\frac{1}{p}} \tag{3.10}$$

$$\leq (\frac{1}{m} \sum_{i=1}^{m} \min(c, ||t_i - v_{\eta(i)}||_\infty)^p)^{\frac{1}{p}} + (\frac{1}{m} \sum_{i=1}^{m} \min(c, ||v_i - u_{\nu(i)}||_\infty)^p)^{\frac{1}{p}}$$

$$= d_p^c(\mathbb{D}_1, \mathbb{D}_3) + d_p^c(\mathbb{D}_3, \mathbb{D}_2)$$

Note that in Eq. 3.10, we are mapping from $\mathbb{D}_1$ to $\mathbb{D}_3$ in the most optimal way (via permutation η) and then from $\mathbb{D}_3$ to $\mathbb{D}_2$ in the most optimal way (via permutation ν).

The second case is when $\ell, m \leq n$. Take η and ν to be the minimum permutations from $\mathbb{D}_1$ to $\mathbb{D}_3$ and from $\mathbb{D}_3$ to $\mathbb{D}_2$ respectively as above. Then, similarly, we have that

$$d_p^c(\mathbb{D}_1, \mathbb{D}_2) = (\frac{1}{n} \min_{\pi \in \Pi_n} \sum_{i=1}^{n} \min(c, ||t_i - u_{\pi(i)}||_\infty)^p)^{\frac{1}{p}}$$

$$\leq (\frac{1}{m} \sum_{i=1}^{m} \min(c, ||t_i - v_{\eta(i)}||_\infty)^p + \min(c, ||v_i - u_{\nu(i)}||_\infty)^p)^{\frac{1}{p}}$$

$$\leq (\frac{1}{m} \sum_{i=1}^{m} \min(c, ||t_i - v_{\eta(i)}||_\infty)^p)^{\frac{1}{p}} + (\frac{1}{m} \sum_{i=1}^{m} \min(c, ||v_i - u_{\nu(i)}||_\infty)^p)^{\frac{1}{p}}$$

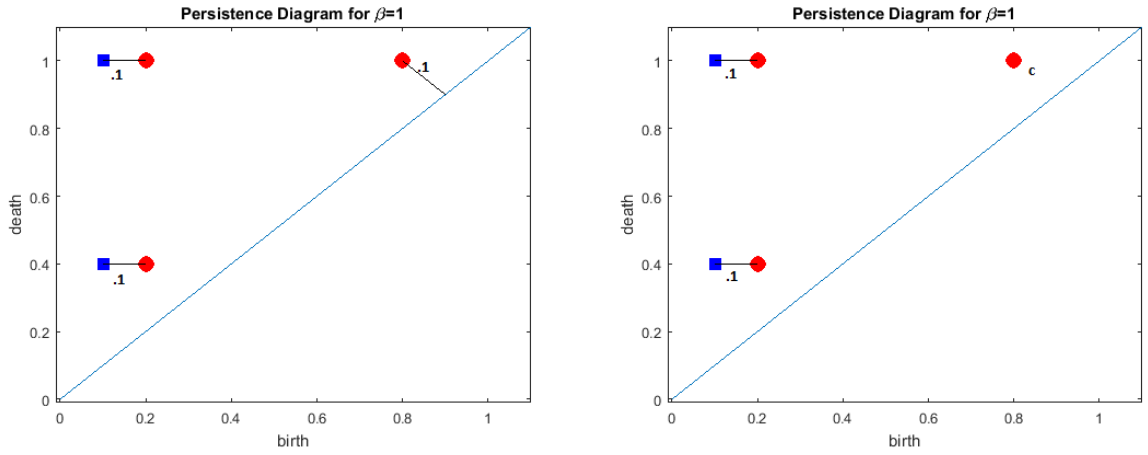
$$= d_p^c(\mathbb{D}_1, \mathbb{D}_3) + d_p^c(\mathbb{D}_3, \mathbb{D}_2)$$

□

Intuitively, all points in $\mathbb{D}_1$ are matched to points in $\mathbb{D}_2$ as best as possible, and then a regularization term that depends on the remaining difference in cardinality and term c is penalized. Note that for this distance, it is not needed to assume infinitely many points on the diagonal, as in the Wasserstein distance, and thus situations as in Fig. 3.3 are avoided. This distance provides a meaningful distance between two persistence diagrams that also takes into consideration the cardinality difference between them. The parameter c in Eq. (3.8) serves as a penalty on the cardinality difference between persistence diagrams. In particular, a smaller c contributes more weight on the matching between small persistence points, which is important for small geometric differences between signals (see Fig. 4.9). On the other hand, a larger c will largely weight on cardinality differences, which is vital for differentiating between large geometry difference in the point clouds (see Fig. 4.7). These large changes in cardinality can be caused by large differences in dynamic behavior leading to large geometry and topology difference in the point clouds (see Fig. 4.7). In particular, higher scale noise leads to a larger cardinality in persistence diagrams, due to a number of small scale holes appearing Adler et al. [3].

The d_p^c distance differs from the Wasserstein distance in the following way. The Wasserstein distance provided in Definition 3.2 finds the best possible mapping between two persistence diagrams (where matches to the diagonal are allowed), and then the difference in cardinality is penalized according to the distance to the diagonal. In the d_p^c metric, points are matched as best as possible from the smaller persistence diagram (in cardinality) to the larger persistence diagram, and then the remaining points are penalized each by c. An example of this can be seen in Fig. 3.5. Note that if two persistence diagrams have the same cardinality and if the optimal bijection from the Wasserstein distance in Eq. (3.2) does not map any points to the diagonal, the Wasserstein distance is equal to the d_p^c distance (up to a constant), assuming c is large enough.

We first examine the space of persistence diagrams under this new metric and examine the properties it exhibits. Define the space of persistence diagrams in $W \subset \mathbb{R}^2$ as $P_{W,k} = \{\{x_1, ..., x_l\} | l \leq k, x_i \in W, 1 \leq i \leq l\}$, where W is a closed and bounded subset of $\mathbb{R}^2$ and $k \in \mathbb{N} \bigcup \{0\}$ and define $P_W = \bigcup_K P_{W,k}$. In real data applications, the point cloud lies in a



(a) Wasserstein Distance **(b)** d_p^c Distance

Figure 3.5: Consider two persistence diagrams, one with points represented by red circles, the other with points represented by blue squares. We compare the Wasserstein distance (on the left) to the proposed d_p^c metric (on the right). Note that the distance between points is computed via the sup norm. Notice how the Wasserstein metric imposes a penalty of .1 to the extra point (the minimal distance to the diagonal), while the d_p^c imposes a penalty of c, which will usually be larger.

bounded subset of $\mathbb{R}^d$ and as such the associated persistence diagram is bounded in space. Moreover, the number of points in the point clouds considered is bounded depending upon the application (due to sampling rate and observation time). Due to this, it is appropriate to consider the space $P_{W,k}$ for some W and k in real data situations. We show in the next lemma that (P_W, d_p^c) is a complete and separable space.

Lemma 3.1.1. *P_W under the d_p^c distance is a complete, separable metric space.*

Proof. We first show completeness. Let $\{\mathbb{D}_n\}_{i=1}^k$ be a Cauchy sequence of persistence diagrams. It is clear that for some k_0, we have that $j, l \geq k_0$ implies $|\mathbb{D}_j| = |\mathbb{D}_l| = k$, so we may assume without loss of generality that the associated cardinalities are equal. Fix an $\epsilon > 0$. Note there is N such that for $n, m > N$, $d_p^c(\mathbb{D}_n, \mathbb{D}_m) < \epsilon$. In particular, since their cardinalities are the same, we have that

$$d_p^c(\mathbb{D}_n, \mathbb{D}_m) = (\frac{1}{k} \min_{\pi \in \Pi_k} \sum_{i=1}^{k} ||x_i^n - x_{\pi(i)}^m||_\infty^p)^{\frac{1}{p}} < \epsilon$$

and so we have that, for a given point $x_i^n \in \mathbb{D}_n$,

$$||x_i^n - x_{\pi(i)}^m||_\infty < (k)^{\frac{1}{p}}\epsilon$$

where $\pi(i)$ is the minimal permutation.

Thus, there is a sequence of points $x_i^n, x_{\pi_{n+1}(i)}^{n+1}, x_{\pi_{n+2}(i)}^{n+2}, ...$ such that the distance between any two points in this sequence is less than $2(k)^{\frac{1}{p}}\epsilon$ via the triangle inequality, where $\pi_{n+\alpha}$ is the minimal permutation between persistence diagrams $\mathbb{D}_n$ and $\mathbb{D}_{n+\alpha}$. This is a Cauchy sequence in W under the inf $-$norm. Since W is complete, this sequence converges to some limit $x_i \in S$. Repeating this for each element in $\mathbb{D}_n$, we generate a persistence diagram $\mathbb{D}^*$ consisting of points $(x_1, ..., x_k)$ chosen as the limits above.

Therefore, for any fixed ϵ^p, since each sequence above converges to the corresponding limits, there is some N such that for $j > N$ we have $||x_i^j - x_i||_\infty < \epsilon$ This implies that

$$d_p^c(\mathbb{D}_j, \mathbb{D}^*) = (\frac{1}{k}\min_{\pi\in\Pi_k}\sum_{i=1}^{k}||x_i^n - x_{\pi(i)}||_\infty^p)^{\frac{1}{p}} \leq (\frac{1}{k}\sum_{i=1}^{k}||x_i^n - x_i||_\infty^p)^{\frac{1}{p}} < (\frac{1}{k}k\epsilon^p)^{\frac{1}{p}} = \epsilon$$

Since this sequence converges to a limit in this space, this space is complete.

Finally, it remains to show separability. Consider the space $P_{\mathbb{Q}\bigcap W}$ of all persistence diagrams with points in $\mathbb{Q}\bigcap W$ with finitely many points. Then for any persistence diagram $\mathbb{D}$, find $\mathbb{D}_q \in P_{\mathbb{Q}\bigcap W}$ such that $|\mathbb{D}| = |\mathbb{D}_q| = k$ and for all $x_i \in \mathbb{D}$, there is a corresponding $y_{x_i} \in \mathbb{D}_q$ such that $||x_i - y_{x_i}||_\infty^p \leq \epsilon$. Then

$$d_p^c(\mathbb{D}, \mathbb{D}_q) = \frac{1}{k}(\min_{\pi\in\Pi_k}\sum_{i=1}^{k}||x_i - y_{\pi(i)}||_\infty^p) \leq \frac{1}{k}\sum_{i=1}^{k}||x_i - y_{x_i}||_\infty^p) \leq \frac{1}{k}\sum_{i=1}^{k}\epsilon = \epsilon$$

□

We now study probability objects on the space of persistence diagrams under the d_p^c metric through Fréchet means and variances. We recall these definitions from the previous section in light of our new distance. Fix a closed and bounded subset W of $\mathbb{R}^2$ and a positive integer k. Consider a probability measure $\mathcal{D}$ on the space of $(P_{W,k}, \mathcal{B}(P_{W,k}))$ where $\mathcal{B}(P_{W,k})$

is the Borel $\sigma-$algebra on $P_{W,k}$ such that

$$F_{P_{W,k}}(\mathbb{D}_1) = \int_{P_{W,k}} d_p^c(\mathbb{D}_1, \mathbb{D}_2)^2 d\mathcal{D}(\mathbb{D}_2) < \infty, \tag{3.11}$$

for all $\mathbb{D}_1 \in P_{W,k}$.

Definition 23. *Given a probability space $(P_{W,k}, \mathcal{B}(P_{W,k}), \mathcal{D})$, the Fréchet variance of $\mathcal{D}$ is*

$$Var_{\mathcal{D}} = \inf_{\mathbb{D} \in P_{W,k}} [F_{P_{W,k}}(\mathbb{D}) = \int_{P_{W,k}} d_p^c(\mathbb{D}, \mathbb{D}_2)^2 d\mathcal{D}(\mathbb{D}_2)], \tag{3.12}$$

and the Fréchet expectation or Fréchet mean of $\mathcal{D}$ is

$$\mathbb{E}(\mathcal{D}) = \{\mathbb{D} | F_{P_{W,k}}(\mathbb{D}) = Var_{\mathcal{D}}\}. \tag{3.13}$$

In other words, the Fréchet mean is any persistence diagram minimizing the Fréchet variance. Next, we show that the Fréchet mean exists for a probability distribution over $P_{W,k}$.

Lemma 3.1.2. *$F_{P_{W,K}}(\mathbb{D})$ is a continuous function.*

Proof. Fix some $\epsilon > 0$. Consider persistence diagrams $\mathbb{D}$ and $\mathbb{E}$ such that $d_p^c(\mathbb{D}, \mathbb{E}) < \min(\frac{\sqrt{\epsilon}}{2}, \max_{\mathbb{D}_1, \mathbb{D}_2}(d_p^c(\mathbb{D}_1, \mathbb{D}_2))$. Then we have that

$$|F_{P_{W,K}}(\mathbb{D}) - F_{P_{W,K}}(\mathbb{E})| = |\int_{P_{W,k}} d_p^c(\mathbb{D}, \mathbb{D}_1)^2 d\mathcal{D}(\mathbb{D}_1) - \int_{P_{W,k}} d_p^c(\mathbb{E}, \mathbb{D}_1)^2 d\mathcal{D}(\mathbb{D}_1)|$$

$$= |\int_{P_{W,k}} d_p^c(\mathbb{D}, \mathbb{D}_1)^2 - d_p^c(\mathbb{E}, \mathbb{D}_1)^2 d\mathcal{D}(\mathbb{D}_1)| \leq \int_{P_{W,k}} |d_p^c(\mathbb{D}, \mathbb{D}_1)^2 - d_p^c(\mathbb{E}, \mathbb{D}_1)^2| d\mathcal{D}(\mathbb{D}_1)$$

$$\leq \int_{P_{W,k}} |(d_p^c(\mathbb{D}, \mathbb{E}) + d_p^c(\mathbb{E}, \mathbb{D}_1))^2 - d_p^c(\mathbb{E}, \mathbb{D}_1)^2| d\mathcal{D}(\mathbb{D}_1)$$

$$\leq \int_{P_{W,k}} |d_p^c(\mathbb{D}, \mathbb{E})^2 + d_p^c(\mathbb{E}, \mathbb{D}_1)^2 + 2 d_p^c(\mathbb{D}, E) d_p^c(\mathbb{E}, \mathbb{D}_1) - d_p^c(\mathbb{E}, \mathbb{D}_1)^2| d\mathcal{D}(\mathbb{D}_1)$$

$$= \int_{P_{W,k}} |d_p^c(\mathbb{D}, \mathbb{E})^2 + 2 d_p^c(\mathbb{D}, \mathbb{E}) d_p^c(\mathbb{E}, \mathbb{D}_1)| d\mathcal{D}(\mathbb{D}_1) \leq \epsilon$$

□

Theorem 3.2. *Let $\mathcal{D}$ be a probability measure on $(P_{W,k}, \mathcal{B}(P_{W,k}))$ satisfying Eq. (3.11). Then $\mathbb{E}(\mathcal{D}) \neq \emptyset$.*

Proof. Let $\{\mathbb{D}_i\}_{i=1}^{\infty}$ be a sequence of persistence diagrams such that $F_{P_{W,k}}(\mathbb{D}_i) \to Var_{\mathcal{D}}$. We claim that the space $P_{W,k}$ is totally bounded under the d_p^c metric. Fix $\epsilon > 0$. For each $1 \leq i \leq k$, consider the set of persistence diagrams D_i such that D_i contains all persistence diagrams with points on a ϵ grid of W. Let $D = \bigcup_{i=1}^{k} D_i$. Then, for any persistence diagram $\mathbb{D} \in P_{W,k}$, there is $\mathbb{D}^* \in D$ such that $|\mathbb{D}| = |\mathbb{D}^*|$ and

$$d_p^c(\mathbb{D}, \mathbb{D}^*) = (\frac{1}{m}(\min_{\pi \in \Pi_m} \sum_{i=1}^{m} \min(c, ||x_i - y_{\pi(i)}||_\infty)^p))^{\frac{1}{p}} \leq (\frac{1}{m}(\sum_{i=1}^{m} \epsilon^p))^{\frac{1}{p}} = \epsilon$$

Thus, this space is totally bounded and by Lemma 3.1.1, this space is complete. Thus it is compact, and since $F_{P_{W,K}}$ is continuous by Lemma 3.1.2, $Var_{\mathcal{D}}$ is attained.

□

Note that the Fréchet mean may not be unique, as in Fig. 3.1. The Fréchet mean can be thought of as a centroid on the data metric space of persistence diagrams. The framework of Theorem 3.2 basically guarantees the mean of a set of persistence diagrams exists in the space $P_{W,k}$. For a finite set of persistence diagrams $D_n = \{\mathbb{D}_i\}_{i=1}^{n}$, consider the empirical distribution $\mathcal{D}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbb{D}_i}$, that is, the uniform discrete distribution on the finite set of persistence diagrams. Given this distribution, the Fréchet mean of persistence diagrams D_n is given by $\mathbb{E}(\mathcal{D}_n)$.

Using this d_p^c distance, we consider a classification algorithm for signals via the data space of persistence diagrams. Suppose for each class $C_1, ..., C_L$ there are corresponding training sets $T_{C_1}^{\beta_l}, ..., T_{C_L}^{\beta_l}$ containing persistence diagrams corresponding to Betti number β_l, for $l = 0, ..., B_M$, where B_M is the largest Betti number considered. Then, for a new signal $x(t)$ with corresponding β_l persistence diagram $\mathbb{D}_x^{\beta_l}$, define the average distance to a class C_k, $1 \leq k \leq L$, associated with the Betti number β_l by

$$d_{\beta_l}(x, C_k) = \frac{1}{|T_{C_k}^{\beta_l}|} \sum_{\mathbb{D} \in T_{C_k}^{\beta_l}} d_p^c(\mathbb{D}_x^{\beta_l}, \mathbb{D}) \tag{3.14}$$

where $|T_{C_k}^{\beta_l}|$ is the cardinality of the training set of persistence diagrams corresponding to class k and Betti number l. We assign the signal x a label $\hat{C}$ defined as

$$\hat{C} = \underset{1 \leq k \leq L}{\operatorname{argmin}} \sum_{l=0}^{B_M} r_l d_{\beta_l}(x, C_k) \tag{3.15}$$

where $\sum_{l=0}^{B_M} r_l = 1$. The r_l's are weights which determine how much each Betti number β_l is considered. Taking all r_i's to be equal gives equal weight to each Betti number. In some situations, prior knowledge may lead to setting some Betti numbers to higher values.

Algorithm 2 Signal classification using persistence diagrams under the d_p^c metric.

Input 1: New signal x, parameters $r_1, .., r_{B_M}, c, p$
Input 2: Signals $S_{C_k} = \{x_i^k\}_{i=1}^{|S_{C_k}|}$ for each class C_k, $1 \leq k \leq L$
Training Phase
for k = 1 to L **do**
    for i = 1 to $|S_{C_k}|$ **do**
        Compute point cloud $\mathcal{P}_i^k$ for x_i^k using delay embedding.
        for l=1 to B_M **do**
            Store Persistence Diagram $\mathbb{D}_{x_i^k}^{\beta_l}$
Prediction Phase
Compute point cloud $\mathcal{P}_x$ for new signal x
for l=1 to B_M **do**
    Store Persistence Diagram $\mathbb{D}_x^{\beta_l}$
$\Sigma_{C_1}, ..., \Sigma_{C_L} \leftarrow 0$
for k = 1 to L **do**
    for i = 1 to $|S_{C_k}|$ **do**
        for l = 1 to B_M **do**
            $\Sigma_{C_k} \leftarrow \Sigma_{C_k} + r_l(d_p^c(\mathbb{D}_x^{\beta_l}, \mathbb{D}_{x_i^k}^{\beta_l}))$
    $\Sigma_{C_k} \leftarrow \frac{\Sigma_{C_k}}{|S_{C_k}|}$
$\hat{C} \leftarrow \operatorname{argmin}(\Sigma_{C_1}, ..., \Sigma_{C_L})$

It is important to consider the several different tuning parameters in the model. The parameter p indicates how sensitive the model is to "outliers" on persistence diagrams, or points that are hard to match to others. As p tends to ∞, only the most extreme point match is taken into consideration into the model. The parameter c determines how much the cardinality difference between persistence diagrams will be penalized. This can be tuned for specific applications, as the cardinality difference is important for classification when

dealing with large geometry difference between classes in the dataset (potentially caused by noise Adler et al. [3]). Finally, the parameters r_l determine how much each homological dimension should be weighted in the class distance computation. For example, it is well known that when using the delay embedding technique to analyze the persistent homology of signals, periodicity can be captured in the β_1 dimension Perea et al. [66] and chaos can be captured in the higher dimensions Venkataraman et al. [84]. Moreover, later we will show that bi-stability can be captured in the β_0 dimension for certain types of dynamics. These parameters give the opportunity for researchers to bring prior knowledge into the classification problem.

Note that the algorithm requires comparing the distance between a new persistence diagram and every persistence diagram in the training set. In some applications where the classes are sufficiently well separated in the data space of persistence diagrams, it may be sufficient to compare a new persistence diagram with the "mean" of a set of persistence diagrams. With the guarantee that the mean of persistence diagrams exists, a new classifier can be implemented where only the mean persistence diagram of each classes training set needs to be stored and compared to, reducing the runtime of Alg. 2. The algorithm is as follows (also see Alg. 3): For each class $C_1, ..., C_L$ with corresponding persistence diagram training sets $T^{\beta_l}_{C_1}, ..., T^{\beta_l}_{C_L}$, compute the Fréchet mean $\mathbb{D}^{\beta_l}_{C_k}$ of the distribution

$$\mathcal{D}^l_{C_k} = \frac{1}{|T^{\beta_l}_{C_k}|} \sum_{\mathbb{D} \in T^{\beta_l}_{C_k}} \delta_{\mathbb{D}}$$

where δ is the Dirac function. Once these means are computed, a new signal x will be classified according to

$$\hat{C} = \underset{1 \leq k \leq L}{\operatorname{argmin}} \sum_{l=0}^{B_M} r_l d^c_p(\mathbb{D}^{\beta_l}_x, \mathcal{D}^l_{C_k}) \tag{3.16}$$

where $\sum_{l=0}^{B_M} r_l = 1$ and B_M is the maximum Betti number considered. Now, the computational cost is passed onto the computation of the Fréchet mean, which only must be computed once. In particular, the runtime of the original algorithm is proportional to the number of signals in the training set, whereas in the proposed algorithm, the runtime is only proportional to the number of classes. This algorithm is similar to Alg. 2 above, except

for the storing of the Fréchet means. In the case where the classes are sufficiently separated in a topological sense, we anticipate that this algorithm will provide similar results to the original algorithm at a fraction of the computational cost. However, when classes are not well separated, we the original Alg. 2 may be better suited.

Algorithm 3 Signal classification using persistence diagrams under the d_p^c metric using Fréchet means.

Input 1: New signal x, parameters $r_1, .., r_{B_M}, c, p$
Input 2: Signals $S_{C_k} = \{x_i^k\}_{i=1}^{|S_{C_k}|}$ for each class C_k, $1 \leq k \leq L$
Training Phase
for k = 1 to L **do**
    for i = 1 to $|S_{C_k}|$ **do**
        Compute point cloud $\mathcal{P}_i^k$ for x_i^k using delay embedding.
        for l=1 to B_M **do**
            Store Persistence Diagram $\mathbb{D}_{x_i^k}^{\beta_l}$
for k=1 to L **do**
    for l=1 to B_M **do**
        Compute the Fréchet mean $\mathbb{D}_k^{\beta_l}$ of class C_k in dimension l
Prediction Phase
Compute point cloud $\mathcal{P}_x$ for new signal x
for l=1 to B_M **do**
    Store Persistence Diagram $\mathbb{D}_x^{\beta_l}$
$\Sigma_{C_1}, ..., \Sigma_{C_L} \leftarrow 0$
for k = 1 to L **do**
    for l = 1 to B_M **do**
        $\Sigma_{C_k} \leftarrow \Sigma_{C_k} + r_l(d_p^c(\mathbb{D}_x^{\beta_l}, \mathbb{D}_k^{\beta_l}))$
    $\Sigma_{C_k} \leftarrow \frac{\Sigma_{C_k}}{|S_{C_k}|}$
$\hat{C} \leftarrow \text{argmin}(\Sigma_{C_1}, ..., \Sigma_{C_L})$

3.3 Discussion and Future Directions

We propose the first clustering algorithm directly on the space of persistence diagrams. By clustering directly on the space of persistence diagrams, no information from the original persistence diagrams is lost, in contrast with feature-based techniques. Moreover, we prove convergence of our proposed algorithm. In order to improve our clustering algorithm, in the future we will consider different types of distances on the space of persistence diagrams. This

is in line with similar work which suggests that the Wasserstein distance excels at detecting large differences in topology (i.e. differences in periodicity and multi-stability), but struggles in detecting small differences in the geometry of the phase space Marchese and Maroulas [55]. In future work, we will develop an Algorithm for computing the Fréchet mean with the proposed d_p^c distance, which is able to detect these differences, and leverage this to create a new clustering scheme based on these distances.

Further, a new classification scheme for signals has been introduced on the data space of persistence diagrams. A major advantage of this algorithm is the tuning that can be applied to these parameters to fit the data. As c increases, the penalty for cardinality difference is more severe, and as p increases, the penalty for matching points that are far away from each other is higher. In particular, when we expect a large difference in the underlying geometry and topology of the datasets, a larger parameter c is important for good classification. On the other hand, when we care more about the geometry differences in the signals, a smaller c allows more of a focus to be placed on the mapping of the small persistence features to each other and less focus to be placed on the cardinality difference.

We note that the d_p^c distance may not be the most adequate choice for all datasets. In particular, in the case of a classification problem in which one class is expected to have a large, prominent topological feature while the other class is not expected to have this feature, the Wasserstein or Bottleneck distance may be good choices (see Emrani et al. [35]). However, this prior information about prominent features is often not available.

Chapter 4

Data Analysis Results

4.1 Computational Preliminaries

4.1.1 Hungarian Algorithm

Both the Wasserstein distance and the proposed d_p^c distance can be computed using the Hungarian algorithm. The Hungarian algorithm is a combinatorial algorithm solving the optimal assignment problem for points in polynomial time Kuhn [51], Golin [41]. In particular, consider two persistence diagrams. The goal is to find the optimal bijection between the points in one persistence diagram to the points in the other persistence diagram, such as the one demonstrated in Fig. 4.1. In order to do this, consider the points in persistence diagrams to be $X = \{x_i\}_{i=1}^N$ and $Y = \{y_i\}_{i=1}^M$. Moreover, define the cost of mapping x_i to y_j to be the distance $d(x_i, y_j)$ for the chosen metric d.

For the proposed d_p^c distance, we use this algorithm to find the optimal matching of points between the two persistence diagrams, and then penalize the extra points as described in Definition 22. Computing the d_p^c distance using the Hungarian algorithm can be thought of as adding in "dummy" points to the smaller persistence diagram from $\mathbb{R}^2 \backslash W$, that are at least distance c from all points in the larger persistence diagram. Then, an optimal point assignment can be found under the Hungarian algorithm between the two "equal" size persistence diagrams.

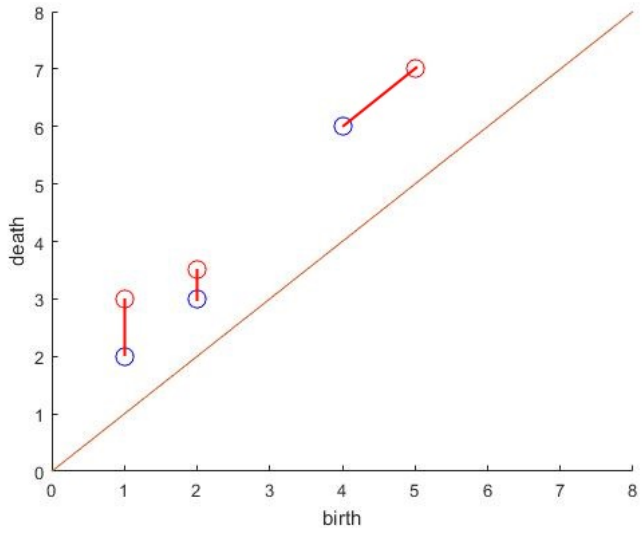


Figure 4.1: An optimal bijection between two persistence diagrams, one with points in blue, one with points in red.

For the Wasserstein distance, it is a bit more challenging as considerations have to be made to allow for matches to the diagonal. In this case, let X_0 be the off-diagonal points in the first diagram, let X_0' be the orthogonal mapping to the diagonal, and similarly for Y_0 and Y_0'. The goal is to find the optimal mapping between $X_0 \bigcup Y_0'$ and $Y_0 \bigcup X_0'$. Intuitively, this is because points in the first diagram can be mapped to either points in Y or the diagonal of X, but they would only ever be mapped to their closed point on the diagonal, which is given by X_0'.

4.1.2 $k-$fold cross-validation

When training a supervised learning model, it is often a bad idea to directly measure the error using the data that was used to create the model. This is due to the bias introduced into the model by the training examples, leading to an underestimation of the true error of the model. In order to alleviate this problem, we will implement k-fold cross validation Hastie et al. [43] for validating our supervised learning scheme. The process works as follows. First, the dataset of persistence diagrams is partitioned into k parts. Then, the Alg. 2 is trained k distinct times - each time, one of the partitions is left out of the model. Then, the error is measured on the held out data, and the error is averaged over all k models. This

averaged error is then used as a measure of the true error of the model. This procedure is depicted in Fig. 4.2. The full pipeline for the signals considered id depicted in Fig. 4.3.

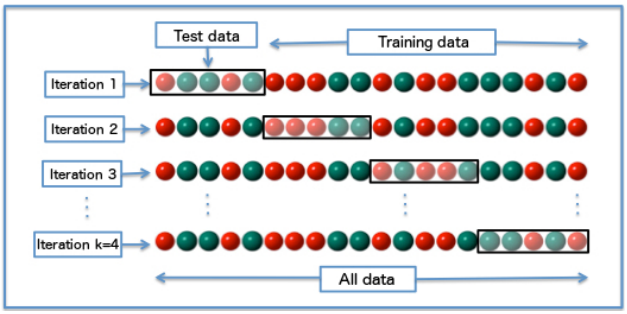


Figure 4.2: 4-fold cross-validation for estimating true error Flock [38].

4.2 Competing Methods

A common feature extraction technique is through the use of wavelets Chan and Fu [22], Agrawal et al. [4], Kawagoe and Ueda [48]. Through the use of a Discrete Wavelet Transform (DWT), time and frequency properties of the signals can be maintained and extracted. The technique works through the use of wavelets, which serve as basis functions. In particular, it is common to use Haar wavelets. The Haar basis is constructed by starting with the piecewise continuous function

$$\phi(x) = \begin{cases} 1 & \text{if } 0 \leq x < \frac{1}{2} \\ -1 & \text{if } \frac{1}{2} < x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Then, define $\phi_j k(x) = \phi(2^j x - k)$ where $j \geq 0$ and $0 \leq k < 2^j$ Strang [80], Chui [25]. In this case, the features extracted would be related to the coefficients of the basis functions. Similarly, features can be extracted via Fourier techniques. In particular, we will consider features related to the power Cepstrum, given by $|\mathcal{F}^{-1}(log(|\mathcal{F}(f(t))|^2))|^2$ where $\mathcal{F}$ is the Fourier Transform and $f(t)$ is the time-series. The power Cepstrum has long been used to analyze acoustic signals and has been used in similar acoustic signal classification applications [79] [64].

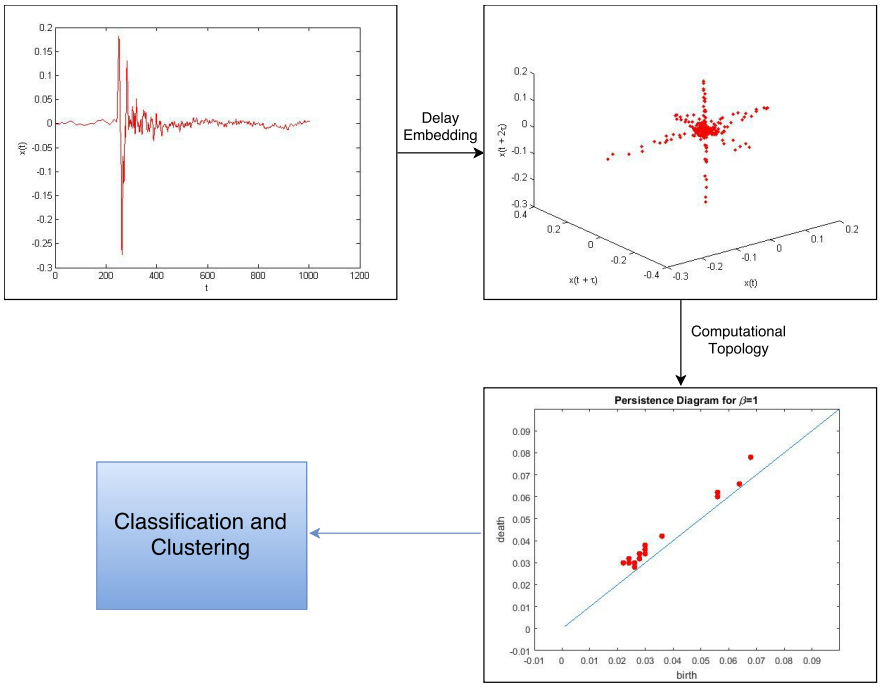


Figure 4.3: The methodology through which the signals are processed is visualized above.

It is intuitive to also consider foregoing the extraction of features, and instead defining a dissimilarity measure directly between signals and use a hierarchal clustering technique based on this metric. One such technique is through Dynamic Time Warping (DTW) Sakoe and Chiba [73, 74]. DTW attempts to match these signals by constructing a matrix corresponding to the squared distance between samples of the signals. Using this matrix, the algorithm attempts to minimize the warping cost

$$DTW(f_i(x), f_j(x)) = \min(\sqrt{\sum_{\ell=1}^{L} w_\ell})$$

where w_ℓ is the matrix element $(i,j)_k$ that also belongs to the kth element of a warping path - an ordered set of matrix elements representing a mapping from f_i to f_j Ratanamahatana and Keogh [69]. In other words, each $w_l = (d(f_i(x_k), f_j(x_\ell)))$ is the cost associated with mapping point $f_i(x_k)$ from the first time-series to point $f_j(x_\ell)$ in the second time-series.

In both the DTW and feature extraction approach, clustering can then be implemented through a scheme known as as hierarchical clustering. In particular, we will focus on

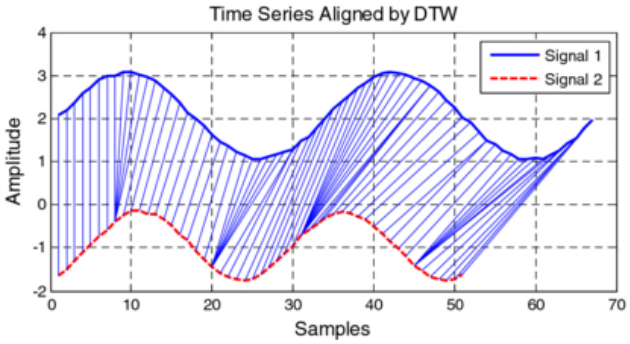


Figure 4.4: An example of a DTW mapping between two signals Zhen et al. [90].

agglomerative hierarchical clustering, or "bottom-up" clustering Hastie et al. [43], Ward Jr [85].

The method works as follows. Assume that we start with a dataset of points $(x_1, ..., x_n)$ and a distance matrix D between them. When $x_i \in \mathbb{R}^d$ for some dimension d, this distance is usually the Euclidean metric. However, for points $x_i \in \mathbb{X}$ for some generic space $\mathbb{X}$, we just need some distance $d(\cdot, \cdot)$ between them. We start by treating each data-point as its own cluster. Then, using some predetermined linkage criteria, we combine two of the clusters. Common linkage criteria between clusters X_1, X_2 are:

Type	Formula
Complete	$\max\{d(x_i, x_j) \mid x \in X_1, x \in X_2\}$
Single	$\min\{d(x_i, x_j) \mid x \in X_1, x \in X_2\}$
Ward	$\frac{\lvert X_1 \rvert \lvert X_2 \rvert}{\lvert X_1 \rvert + \lvert X_2 \rvert} \lvert\lvert \bar{X}_1 - \bar{X}_2 \rvert\rvert^2$

This procedure continues until all data-points are merged into one cluster or a stopping criteria is reached. The result is a dendrogram, showing when each cluster was merged. If we know the number of clusters K a priori, we can cut off the dendrogram at K clusters.

4.3 Clustering Results

We now benchmark our proposed Algorithm 1 on several datasets. In order to assess the efficacy of our clustering algorithm and comparison algorithms, we find the best labeling of classes with respect to misclassification rate of the ground truth labels, and we denote this misclassification rate as "error". We benchmark Algorithm 1 against a Dynamic Time

Warping (DTW) hierarchical clustering algorithm, and a wavelet feature-based hierarchical clustering algorithm. For both of these cases, the Ward criteria is used. We stress that our algorithm is not a classification algorithm, but we show error with respect to the ground truth only in order to benchmark against current algorithms.

For the computation of Fréchet means, we implement an algorithm introduced by [82] that estimates a local minimum of the Fréchet function. To implement this algorithm, we initialize by randomly selecting a persistence diagram from our dataset. Then, we find the optimal matching from this persistence diagram to every other persistence diagram, and average the mapped points. This algorithm continues until convergence is reached in the sense of the center diagram not changing. The Vietoris-Rips complexes are computed using Ripser Bauer [10].

4.3.1 Synthetic Dataset

We benchmark Algorithm 1 on a synthetic dataset with 4 different types of signals.

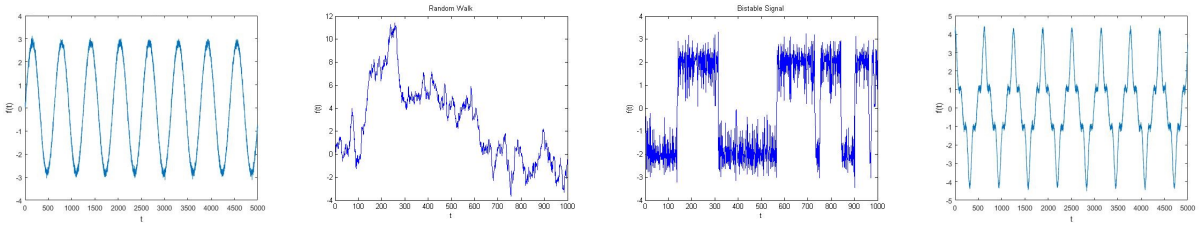
$$w_i = \omega \sin(o_i) + \eta_i, \tag{4.1}$$

$$v_{i+1} = v_i + \eta_i, \tag{4.2}$$

$$u_{i+1} = \alpha \sin(u_i) + \eta_i, \tag{4.3}$$

$$z_i = \omega(1 + .5\cos(o_i))(\cos(o_i)) + \eta_i, \tag{4.4}$$

where $\eta_i \sim N(0, .1)$, $\omega \sim Unif(1, 3)$, $\alpha = 2.5$ and $o_i \in O$ is a grid of values (we choose $O = [0, 50]$ with increments of .01). A depiction of these different is shown in Fig. 4.5. Notice that the different classes of signals vary widely in terms of periodicity, multi-stability, and noise. We generate a dataset of 10000 signals of each class, giving a population of 40000 signals. We then sample 100 signals from each class and test the algorithm on subsets of 400 signals each.



(Class 0) (Class 1) (Class 2) (Class 3)

Figure 4.5: The four classes of signals used for benchmarking the clustering algorithm.

Table 4.1: Results for clustering of signals from Eqs. (4.7)-(4.8).

Method	Error
DTW	.6188
Wavelet	.6406
Fréchet Mean Clustering	**.3183**

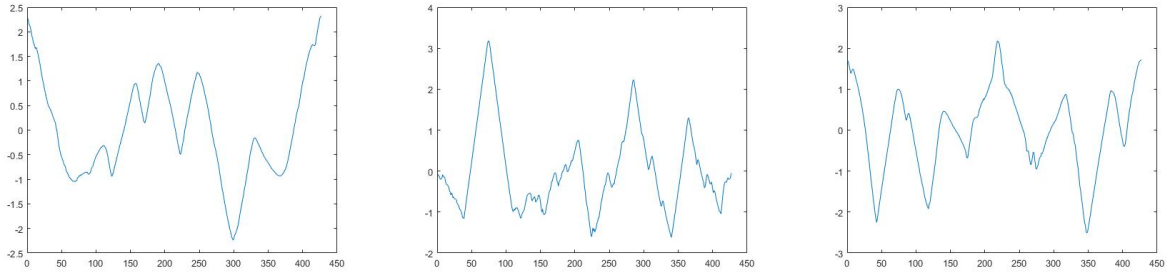
Note that the proposed method outperforms the DTW and wavelet clustering methods, as seen in Table 4.1.

4.3.2 Real Dataset

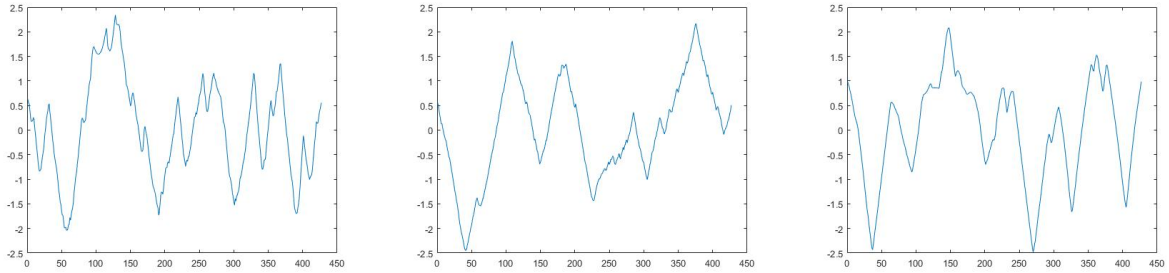
We now benchmark Algorithm 1 on the OSU Leaf dataset (available at http://www.cs.ucr.edu/~eamonn/time_series_data/). This dataset contains 6 classes and 442 time-series, where each class is representative of a different species of leaf Gandhi [39]. The goal is to stress our proposed algorithm and competing algorithms on a dataset with more clusters with more similar types of signals (as seen in Fig. 4.6). It is important to note that this is a very challenging signal classification dataset due to the similarity of the signals. As shown in Table 4.2, our proposed algorithm outperforms the competing DTW and wavelet based clustering algorithms, though all algorithms have relatively low accuracy due to the difficulty of classification on this dataset.

Table 4.2: Results for clustering of signals from OSU Leaf dataset.

Method	Error
DTW	.6448
Wavelet	.6991
Fréchet Mean Clustering	**.5769**



(Class 0) (Class 1) (Class 2)



(Class 3) (Class 4) (Class 5)

Figure 4.6: Examples of signals from the six classes used for benchmarking the clustering algorithm in the OSU dataset.

4.4 Classification Results

The Vietoris-Rips complexes are computed using Ripser Bauer [10]. We consider Betti numbers β_0 and β_1 in Eq.(3.15), with $r_0 = r_1 = .5$, assigning equal weight to β_0 and β_1 features.

4.4.1 Synthetic Dataset 1

We now aim to benchmark the classifier introduced in Eq. (3.15) on a synthetic dataset, where signals are produced from two different stochastic dynamics:

$$v_{i+1} = v_i + \eta_i, \tag{4.5}$$

$$u_{i+1} = \alpha \sin(u_i) + \zeta_i, \tag{4.6}$$

where $\eta_i \sim N(0, \sigma)$, $\alpha = 2.5$, $\zeta_i \sim N(0, \sigma)$, and the hyper-parameter $\sigma \sim Unif(0, .5)$. Note that for a given signal $v = \{v_i\}_{i=1}^N$ the noise $\eta = \{\eta_i\}_{i=1}^N$ is generated independently of each other, and independent of the signal. We set $v(0) = u(0) = 0$, and fifty signals of each dynamic are generated for the synthetic dataset.

These dynamics produce two very different types of signals. Signals generated from Eq.(4.5) are random walks with Gaussian noise, as in Fig. 4.7 (a). Notice that random walk signals exhibit no form of periodicity or consistency.

The second dynamic, Eq.(5.4), generates signals that tend towards a bistable solution Law et al. [52]. The reason for this is due to the symmetry of the equation under the transformation $u \to -u$, leading to two solutions. Under no noise, the system would fall into one of these solutions depending on the initial parameters. Once noise is introduced into the system, the solution to the dynamic randomly transitions between the two solutions at a rate depending on the magnitude of the noise, as can be seen in Fig. 4.7(b). Note that while in each of the two stable solutions, the signal exhibit a local periodicity. Due to the different behavior of the two dynamics, it is reasonable to expect a classifier to distinguish between them. In particular, this difference of the underlying geometry can be seen in Figs. 4.7 (c) and (d). Note that in the point clouds of bistable signals the points are attracted to four difference centers, corresponding to the two stable states in the original signal.

The classifier in Eq.(3.15) is now implemented in order to differentiate between signals from the two dynamics above using the proposed d_p^c distance as described in Section **??**. We test our proposed classifier against one using the Wasserstein distance, and standard Fourier and wavelet methods.

Results are presented in Fig. 4.8. We note that the proposed d_p^c distance outperforms the Wasserstein distance over all values of p and c as well as the Fourier and wavelet classifiers. The d_p^c classifier is consistent with respect to changes in p, and increasing c from 1 to 5 gives a slight increase in accuracy of the d_p^c classifier. This implies that the classifier is learning some important features of the cardinality of the persistence diagrams, related to the large differences in the geometry and topology of the underlying systems. This is reinforced by the cardinality statistics seen in Fig. 4.7 (g) and (f). In addition, we see that the proposed

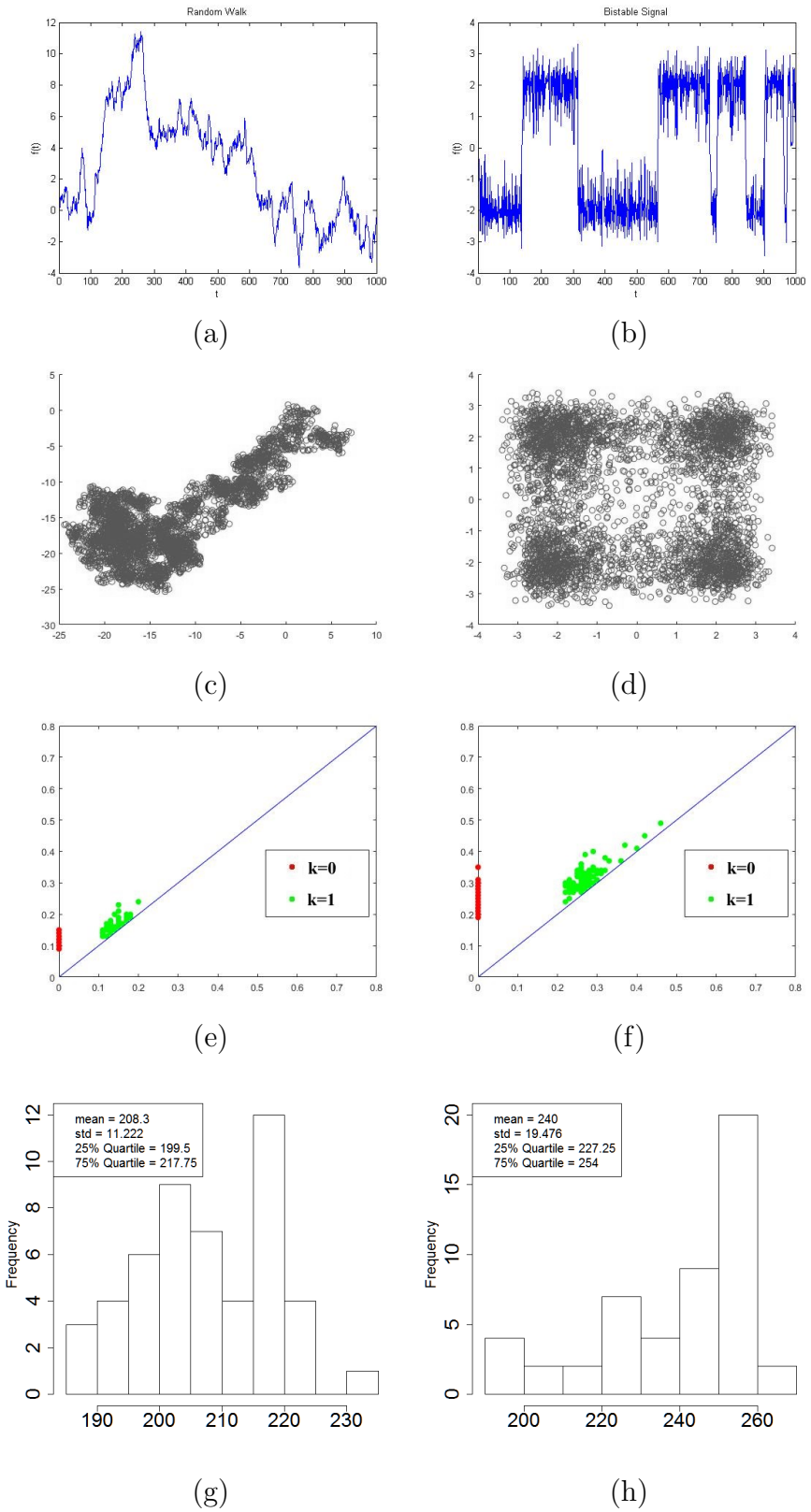


Figure 4.7: Sample signals, point clouds, persistence diagrams, and cardinality statistics for (Left) Random walk signals as in Eq. (4.5) and (Right) Bistable signals as in Eq. (5.4).

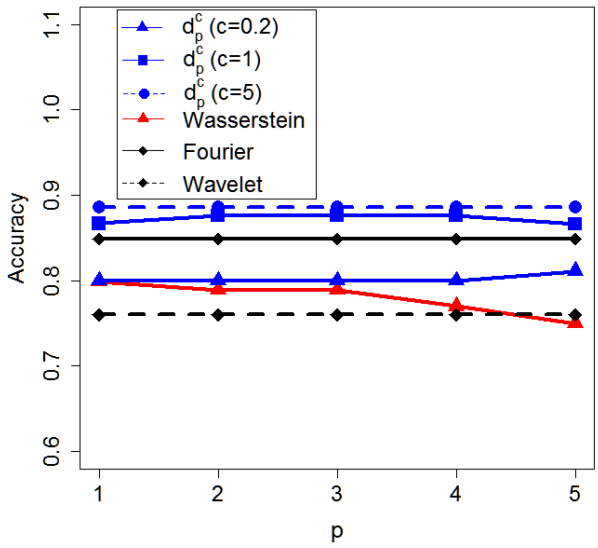


Figure 4.8: Synthetic data results for $p = 1, ..., 5$ and $c = 0.2, 1, 5$ for Eqs. (4.5) and (5.4).

d_p^c classifier has very good accuracy with respect to larger values of p, while the Wasserstein distance starts to deteriorate.

4.4.2 Synthetic Dataset 2 - A Parameter Sensitivity Test

The sensitivity of our method with respect to the two tuning parameters of the d_p^c distance is examined in this section. We consider the signals as in Fig. 4.9 (a) and (b), which produce point clouds as in Fig. 4.9 (c) and (d). These signals are given by the formulas below:

$$v_i = \omega sin(o_i) + \eta_i, \tag{4.7}$$

$$u_i = \omega(1 + .5cos(o_i))(cos(o_i)) + \eta_i, \tag{4.8}$$

where $\eta_i \sim N(0, \sigma)$, $\omega \sim Unif(1, 3)$ and $o_i \in O$ is a grid of values (we choose $O = [0, 50]$ with increments of .01). For testing purposes, we generate twenty five of each signal. For this particular dataset, the small persistence points in the persistence diagrams is the main discriminating factor between the two classes of signals since both types of signals have similar topology as can be seen by the prominent features in Fig. 4.9 (e) and (f). Thus, the d_p^c algorithm should perform well on it, due to its explicit mapping of small points to

each other, and not to the diagonal, in contrast to the Wasserstein distance. However, the inherent noise properties of the underlying signals and point clouds will not be useful for classification, since the noise and total variation is on the same scale for both dynamics. In particular, it is expected that cardinality is not an important factor for this dataset, and in fact may hinder classification, since both cases consider similar noise and variation and only have small differences in geometry.

The results are depicted in Fig. 4.12 for $c = 0.4, 1, 3$ and 5 and for $p = 1, ..., 5$. As expected, we have very good classification results for a small c, when the algorithm is prominently classifying via mapping the small persistence points to each other. As c increases, the penalty for differing cardinality grows and our classifier starts to break down on this dataset. This demonstrates that the cardinality alone is not enough for good classification, but having a scalable c parameter allows a trade-off between cardinality and mapping between features.

In general, when the underlying systems generating the data have small differences in their geometry, we expect results to deteriorate as c increases. This can also be seen in the cardinality statistics of the persistence diagrams in Fig. 4.9 (g) and (h). On the other hand, we expect results to improve as c increases for datasets where there is a large geometry or topology difference in the underlying point cloud; this is potentially caused by either differences in underlying noise in the system, or large differences in dynamic behavior. For systems with a very persistent topological feature present in one class that is absent in the other class, we expect results to improve as p increases. Intuitively, this is because as p increases, the algorithm focuses more on the most extreme matching. When p is set to infinity, only the most extreme matching will be considered. We also present results for the Wasserstein distance, as well as Fourier and wavelet based methods. Note that the d_p^c distance outperforms all other methods, followed by the Wasserstein distance.

4.4.3 Real Acoustic Signal Dataset

We now benchmark the classifier on a real acoustic dataset. The dataset has been provided by the US Army Research Laboratory (ARL). The dataset contains digital recordings of various explosions. In total, 102 signals are used from 2 classes, focusing on signals from

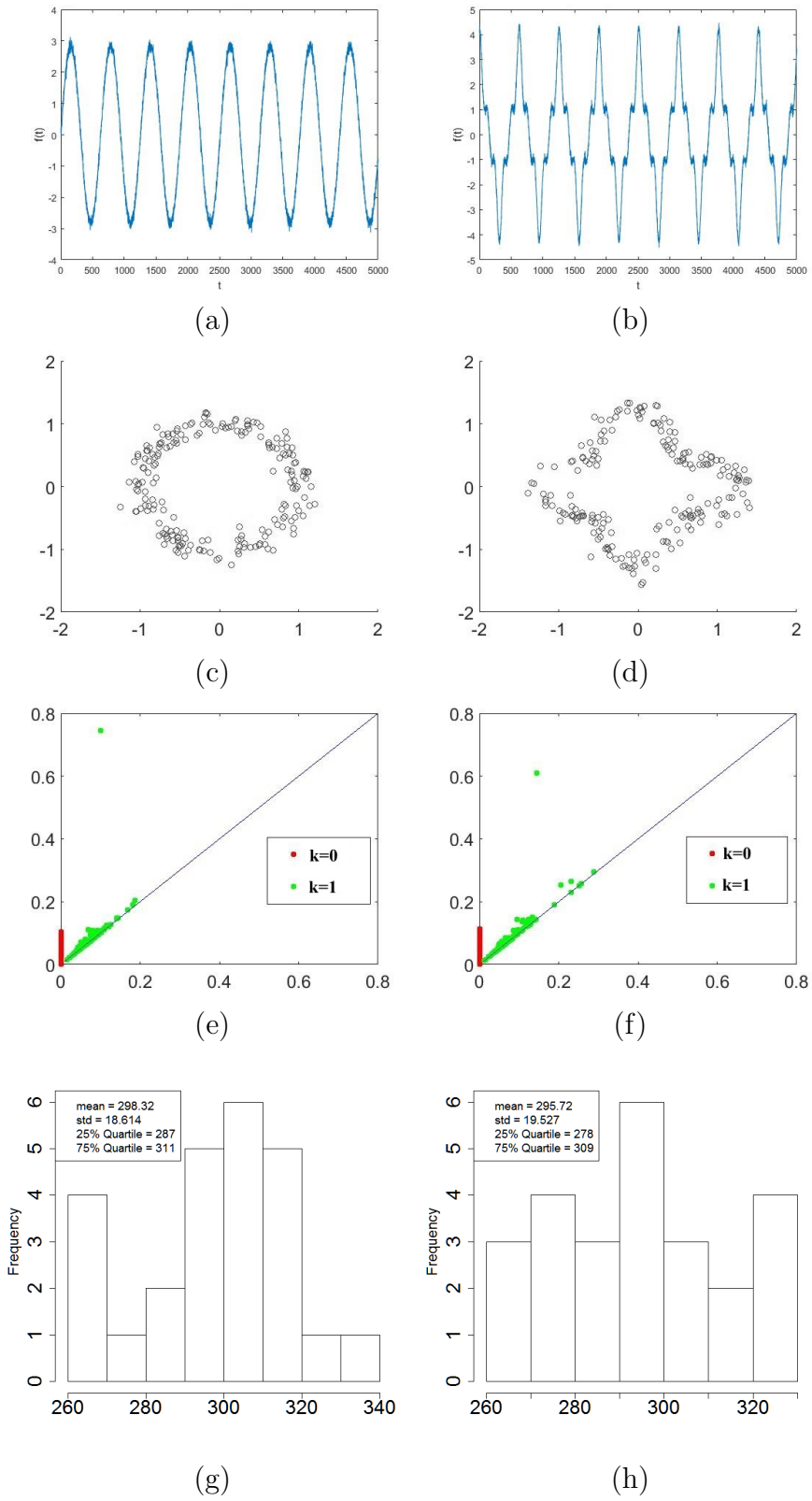


Figure 4.9: Sample signals, point clouds, persistence diagrams, and cardinality statistics for (Left) Periodic signals as in Eq. (4.7) and (Right) Doubly Periodic signals as in Eq. (4.8).

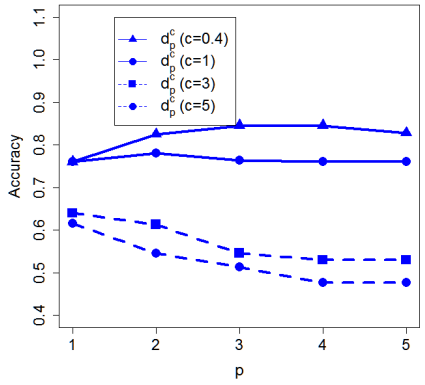


Figure 4.10: Results for $p = 1, ..., 5$ and $c = .4, 1, 3, 5$.

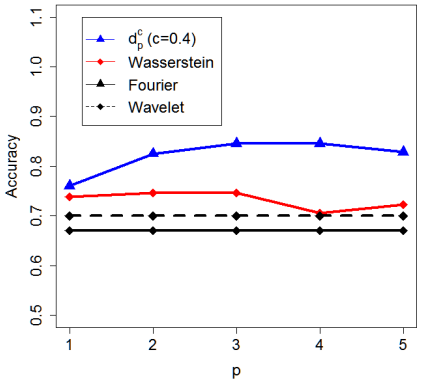


Figure 4.11: Results for competing methods.

Figure 4.12: Results on dataset generated by Eqs. (4.7) and (4.8).

tube explosions for different caliber sizes. The class to predict is the weapon type. The dataset contains 66 signals from class 0 and 36 signals from class 1. Each signal contains measurements from four microphones. The data is recorded at a sampling rate of about 1000 Hz, and consists of 6 to 7 seconds of data per signal. Due to the highly concentrated nature of the signals, the signals are first truncated around to about one second of data around the explosion.

The data contains information from four microphones (channels) placed close together. First, a single channel is selected and used for all signals. Results for differing p and c values are shown in Fig. 4.16(a). We note that under optimal parameters for each classifier, the d_p^c distance outperforms all other classifiers. As p increases, the d_p^c classifier increases in accuracy and outperforms the Wasserstein distance. In contrast, the Wasserstein distance experiences a significant drop-off as p increases to 4 and 5. We note that for $p = 3$, the Wasserstein classifier spikes in accuracy and is competitive with the d_p^c classifier at this specific p value. Moreover, as c increases, the d_p^c accuracy starts deteriorating. This indicates that the small persistence points are important for classification on this dataset. This is also demonstrated in the cardinality statistics presented in Fig. 4.13 (g) and (h), which shows a similar cardinality between the classes.

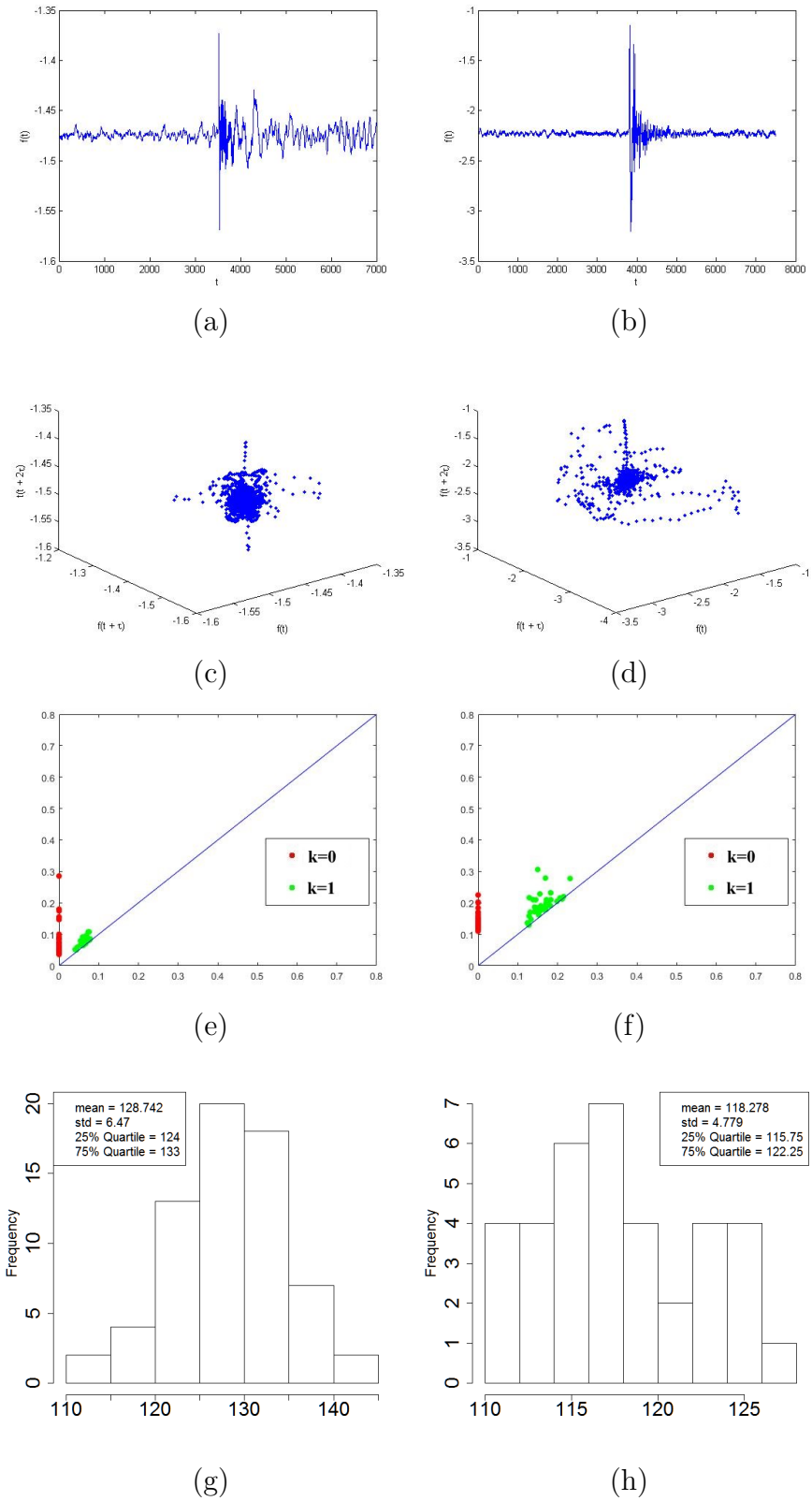


Figure 4.13: Sample signals, point clouds, persistence diagrams, and cardinality statistics for (Left) Class 1 and (Right) Class 2 signals.

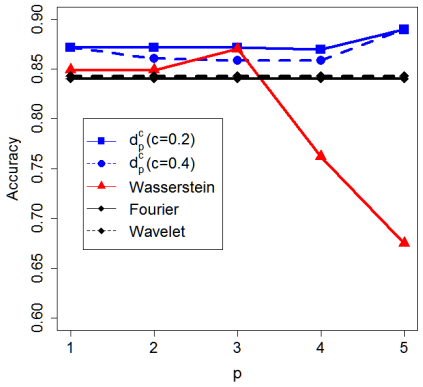


Figure 4.14: Single channel results for $p = 1, ..., 5$ and $c = .2, .4$.

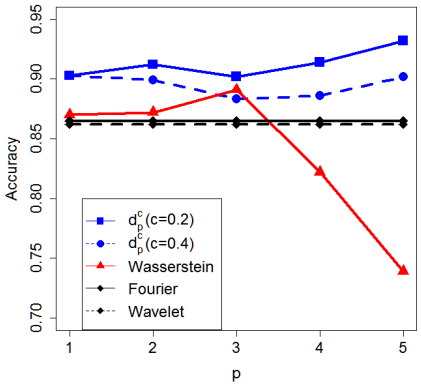


Figure 4.15: Multichannel results for $p = 1, ..., 5$ and $c = .2, .4$.

Figure 4.16: Results on acoustic dataset.

Table 4.3: Confusion matrix for the d_p^c classifier on the one channel dataset for $p = 5, c = .2$.

		Predicted Labels	
		0	**1**
True Labels	**0**	59	7
	1	4	32

Table 4.4: Confusion matrix for the d_p^c classifier on the multichannel dataset for $p = 5, c = .2$.

		Predicted Labels	
		0	**1**
True Labels	**0**	61	5
	1	2	34

Next, we train a classifier for each channel and classify the four-channel signals via a majority vote scheme. Under the scheme, each channel reports a class prediction, and then the class with the most "votes" is chosen as the prediction. The prediction is chosen uniformly at random if there is a tie between the four channels. These results are shown for various c and p values in Fig. 4.16(b).

As in the single channel case, the proposed d_p^c distance outperforms the other classifiers under optimal parameters. We again notice that for p increasing to 4 and 5, the Wasserstein classifier starts to deteriorate as the d_p^c accuracy is consistent. Moreover, as in the single channel case, the d_p^c classifier accuracy deteriorates as c increases, indicating that large geometry and topology differences are not prominent between the two classes. The confusion matrices for the best performing d_p^c classifier is displayed in Tables 4.3 and 4.4. In both the single and multi channel cases, around the same proportion of each class was misclassified.

4.5 Discussion and Future Directions

One main observation on the results of the clustering scheme on the synthetic dataset is that the algorithm performs well when the classes of signals are well separated in terms of features that persistent homology is sensitive too. In particular, persistent homology excels at detecting differences in the shape of signals, such as periodicity and bi-stability Perea et al. [66], Venkataraman et al. [84]. We observe this on our results, as our proposed algorithm successfully characterizes the periodic signals and the bistable signals as distance classes, and furthermore separated them from the random walk signals, which should experience no patterns such as these. We compare this to the competing algorithms, that try to find structure in the random walk signals instead of grouping them based on their lack of structure.

With regard to the proposed classification scheme, we notice that the scheme outperforms similar techniques using the Wasserstein distance as well as classifiers using signal processing Fourier techniques and wavelet techniques. Employing this new distance that depends on the tuning parameters p and c, we offer a comparison of the d_p^c distance (Definition 22) to the Wasserstein distance under different parameters and a sensitivity testing. A major advantage of this algorithm is the tuning which can be applied to these parameters to fit the data. As c increases, the penalty for cardinality difference is more severe, and as p increases, the penalty for matching points that are far away from each other is higher. In particular, when we expect a large difference in the underlying geometry and topology of the datasets, a larger parameter c is important for good classification. On the other hand, when we care more about the geometry differences in the signals, a smaller c allows more of a focus to be placed on the mapping of the small persistence features to each other and less focus to be placed on the cardinality difference. For example, a common trend in the real dataset benchmarked in Fig. 4.9 is the improvement of the d_p^c classifier as c decreases, suggesting that under the conditions of this dataset, matching of small persistence points is more important than cardinality differences.

A future direction we will pursue is increasing the efficiency of the runtime of our algorithm. We compute the Wasserstein and d_p^c distance via utilizing the Hungarian matching algorithm Kuhn [51], which runs in $O(n^3)$ time. Recent work has shown that this runtime can be significantly reduced by matching the points in a persistence diagram in a way which leverages their geometry Kerber et al. [50]. In future work, this type of matching algorithm will be incorporated into our proposed algorithms, in both distance computation and in computing the Fréchet means.

Another future direction of research is further quantifying the cardinality of persistence diagrams as a function of the type and scale of the noise in the system. As a preliminary result on this topic, consider Table 4.5. In this example, we consider periodic signals as in Eq. (4.7), but generate 10 signals for each level of changing variance. We then compute the average cardinality of the persistence diagrams at each level. We notice that, at least for Gaussian noise, the cardinality increases as the scale of the variance increases. Intuitively,

this is due to the number of small persistence points (in particular, 1−dimensional holes) increases in the topology of the point cloud.

Table 4.5: Differing cardinalities of persistence diagrams corresponding to the differing scales of noise.

Variance of Noise	Average Cardinality of Persistence Diagram
0	351.4
.1	470.1
.2	493.5
.3	497
.4	505.5
.5	506.4
.6	512.9

Chapter 5

Topological Stability in State Estimation

In the previous chapters, we have considered the time-series data as observations of the true space that we are interested in. However, it is often the case in data analysis where a low dimensional signal is observed from a high dimensional system. In this case, there is a Hidden Markov Model system in place where instead of performing inference on the lower dimensional space, it would make more sense to reconstruct the original dynamical system and performance inference in this space Crisan and Doucet [27], Kantas et al. [47]. To this end, we study the relationship between persistent homology and filtering and smoothing techniques. These filtering techniques allow us to estimate the path of a dynamical system in some higher dimension given the observations in a lower dimension Beskos et al. [11]. Moreover, persistent homology can reveal the underlying topology of a filtered dynamical system, which can be utilized for classification Venkataraman et al. [84], Marchese and Maroulas [56]. Topological features such as connected components and 1-dimensional holes are indications of specific long-time behavior in a dynamical system, such as bistability or periodicity. We are interested in how persistent homology can be used to study dynamical systems in the context of filtering and smoothing. In particular, we will present a result on how much the persistent homology of a expected filtered path can differ from the persistent homology of the optimal path.

5.1 Topological Stability

Before considering how persistent homology interacts with the filtering of dynamical systems, we discuss some well-known stability results of the Wasserstein and Bottleneck distances Cohen-Steiner et al. [26]. We begin by discussing how to construct a persistence diagram by from a function instead of a point cloud. We follow the presentation in Edelsbrunner and Harer [32]. Let $\mathbb{X}$ be a triangulable space (in our case we will consider subsets of $\mathbb{R}^d$) and $f : \mathbb{X} \to \mathbb{R}$ be a continuous function. Given a threshold $a \in \mathbb{R}$, define the sublevel set of f at a to be $\mathbb{X}_a = f^{-1}(-\infty, a]$. As with persistent homology of point clouds, considering the sublevel sets for an increasing series of values of a leads to a nested sequence of homology groups. For example, for $a_1 < a_2 < ... < a_j$ we have corresponding sublevel sets

$$\mathbb{X}_{a_1} \subseteq \mathbb{X}_{a_2} \subseteq ... \subseteq \mathbb{X}_{a_j}$$

Each of these groups has a corresponding homology group for each topological dimension k, which are naturally included in each other, leading to a chain

$$H_k(\mathbb{X}_{a_1}) \to H_k(\mathbb{X}_{a_2}) \to ...H_k(\mathbb{X}_{a_j})$$

At certain values of a, the homology group will change. These values of a are called homological critical values, corresponding to a $k-$dimensional homological feature being born or dying.

Definition 24. *A continuous function f is said to be tame if it only has a finite number of homological critical values.*

First we give a stability result for the Bottleneck distance d_B,

Theorem 5.1 (Chazal et al. [23])**.** *Given a pair of tame Lipschitz functions $f, g : \Omega \to \mathbb{R}$ on a triangulable space Ω, we have that*

$$W_\infty(dgm(f), dgm(g)) \leq \|f - g\|_\infty$$

In particular, we are interested in a special case of this result. Consider finite sets of points U and V in $\mathbb{R}^d$. Define $f(z) = \min_{u \in U} \|u - z\|_\infty$ and $g(z) = \min_{v \in V} \|v - z\|_\infty$. With f and g as defined above, we note that $dgm(f)$ and $dgm(g)$ are exactly the persistence diagrams created from the Čech complex over U and V. Our goal is to bound the difference between $dgm(U)$ and $dgm(V)$ given that U and V are "close" as quantified by the Hausdorff distance:

Definition 25. *The Hausdorff distance between sets U and V is given by*

$$d_H(U, V) = \max(\max_{u \in U} \min_{v \in V} |u - v|, \max_{v \in V} \min_{u \in U} |u - v|)$$

In particular, Theorem 5.1 can instead be stated as:

Theorem 5.2 (Chazal et al. [23])**.** *Given totally bounded metric spaces X and Y,*

$$W_\infty(dgm(U), dgm(V)) \leq 2d_H(U, V)$$

The idea of the proof for Theorem 5.2 is as follows:

Assume $d_H(X, Y) < \epsilon$, where $U = \{u_i\}_{i=1}^N$ and $V = \{v_i\}_{i=1}^M$. Then for all $v \in V$, there exists $u \in U$ such that $|v - u| < \epsilon$, and similarly for $u \in U$, there is $v \in V$ such that $|u - v| < \epsilon$. Thus, there is some correspondence $\phi : U \to V$ such that $|u - \phi(u)| < \epsilon$, and similar some correspondence from V to U. However, it is important to note that there is not necessarily a bijection such that this is true. In order to address this, consider $\hat{U}$ and $\hat{V}$. $\hat{U}$ contains M copies of each point in U, and similarly $\hat{V}$ contains N copies of each point in V. Note that there is a bijection $\psi : \hat{U} \to \hat{V}$ such that $|u - \psi u| < \epsilon$. Also, it is important to note that the homology groups of U and $\hat{U}$ are the same, and so we lose no information with the extension of U and V. In particular, it is clear that $VR_\epsilon(U) \hookrightarrow VR_\epsilon(\hat{U})$, and in fact, the underlying homology groups are isomorphic.

Using this bijection, for a fixed scale r we are able to define a injection $\psi_\epsilon : VR_{r-\epsilon}(\hat{U}) \to VR_r(\hat{V})$ where a simplex $\sigma = (u_1^\sigma, ..., u_\ell^\sigma)$ is mapped onto $\psi_\epsilon(\psi(u_1^\sigma), ..., \psi(u_\ell^\sigma))$. Similarly, the map ${\psi_r}^{-1} : VR_r(\hat{V}) \to VR_{r+\epsilon}(\hat{U})$ is injective. Thus, in some sense for any fixed scale r we

have "squeezed" $VR_r(\hat{V})$ between $VR_{r-\epsilon}(\hat{U})$ and $VR_{r+\epsilon}(\hat{U})$. Moreover, the maps ϕ_ϵ above induce injections onto the chain groups of the associated Rips complexes, as seen below.

$$\begin{array}{ccccccc}
C_0^{r-\epsilon}(\hat{U}) & \longleftarrow & C_1^{r-\epsilon}(\hat{U}) & \longleftarrow & \cdots & \longleftarrow & C_k^{r-\epsilon}(\hat{U}) \\
\downarrow{\scriptstyle \phi_{r-\epsilon}} & & \downarrow{\scriptstyle \phi_{r-\epsilon}} & & \downarrow{\scriptstyle \phi_{r-\epsilon}} & & \downarrow{\scriptstyle \phi_{r-\epsilon}} \\
C_0^{r}(\hat{V}) & \longleftarrow & C_1^{r}(\hat{V}) & \longleftarrow & \cdots & \longleftarrow & C_k^{r}(\hat{V}) \\
\downarrow{\scriptstyle \phi_{r}^{-1}} & & \downarrow{\scriptstyle \phi_{r}^{-1}} & & \downarrow{\scriptstyle \phi_{r}^{-1}} & & \downarrow{\scriptstyle \phi_{r-}^{-1}} \\
C_0^{r+\epsilon}(\hat{U}) & \longleftarrow & C_1^{r+\epsilon}(\hat{U}) & \longleftarrow & \cdots & \longleftarrow & C_k^{r+\epsilon}(\hat{U})
\end{array}$$

This technique is known as interleaving of persistence diagrams. In particular, using the maps defined above and proceeding with these homological algebra techniques, it can be shown that the persistence diagrams $dgm(U)$ and $dgm(V)$ are close together under the bottleneck distance.

Using this above result, we obtain a bound for the p-Wasserstein distance.

Theorem 5.3. *Given totally bounded metric spaces X and Y,*

$$W_p(dgm(U), dgm(V)) \leq 2d_H(U,V)(T_f)^{\frac{1}{p}},$$

where T_f is the total number of possible features (which depends on the number of points in U and V).

Proof. The proof follows directly from Theorem 5.2. We have that

$$W_p(dgm(U), dgm(V))^p = \inf_{\gamma} \sum_{u \in dgm(U)} \|u - \gamma(u)\|_\infty^p \leq 2^p \sum_{u \in dgm(U)} d_H(U,V)^p = 2^p T_f d_H(U,V)^p.$$

□

5.2 Preliminaries of the Particle Filter

Here we give a synopsis of the filtering problem following Law et al. [52], Doucet et al. [30]. For the rest of this section, we will assume the following dynamic/observation system:

$$v_{j+1} = \phi(v_j) + \xi_j \tag{5.1}$$

$$z_{j+1} = h(v_{j+1}) + \eta_{j+1} \tag{5.2}$$

where ξ_j, η_j are noise distributed according to some distribution i.i.d., $h : \mathbb{R}^n \to \mathbb{R}^m$ is the observation function and $\phi : \mathbb{R}^n \to \mathbb{R}^n$ drives the dynamics. Intuitively, we are interested in a dynamic v_{j+1} with hidden states which we estimate through the observation operator $h(v_{j+1})$, leading to known states of the observations z_{j+1}. We first define a Markov kernel.

Definition 26. *A function $p : \mathbb{R}^l \times B(\mathbb{R}^l) \to \mathbb{R}^+$ is a Markov kernel if*

- *For each $x \in \mathbb{R}^l$, $p(x, \cdot)$ is a probability measure on $(\mathbb{R}, \mathbb{R}^l)$*
- *$x \to p(x, A)$ is $B(\mathbb{R}^l)$ measurable for all $A \in B(\mathbb{R}^l)$.*

In particular, for the filtering problem we assume that at any time j, we have access to $z_{1:j} = \{z_1, ..., z_j\}$, and our goal is to estimate the filtering distribution of the true state $\pi(v_j|z_{1:j})$. Since $\pi(v_j|z_{1:j}) = \int_{\mathbb{R}^n} \pi(v_{j+1}|v_j)\pi(v_j|z_{1:j})dv_j$, we must estimate $\pi(v_{j+1}|v_j)$ and $\pi(v_j|z_{1:j})$. $\pi(v_{j+1}|v_j)$ can be obtained through the Markov Kernel generated through the dynamics Φ, so it remains to estimate $\pi(v_j|z_{1:j})$. By Baye's formula, we see that $\pi(v_j|z_{1:j}) = \frac{\pi(z_j|v_j)\pi(v_j|z_{1:j-1})}{\pi(z_j|z_{1:j-1})}$. In the case when the observation and dynamic functions are linear (when $\phi(v) = Mv$ and $h(v) = Hv$ for $M \in \mathbb{R}^{n\times n}$ and $H \in \mathbb{R}^{m\times n}$) and the noise is Gaussian, it is well known that the Kalman filter provides a closed form for the optimal estimate of the filtering distribution Kalman et al. [45]. However, this is often not the case. In particular, we will consider situations in which the dynamics are highly nonlinear with Gaussian noise. In this case, we will make use of the particle filter, which has been used in many applications in which the dynamics or observation is non-linear Evangelou and Maroulas [36], Kang et al. [46]. Intuitively, the particle filter algorithm proceeds as follows. First, in the prediction step, particles from the previous posterior distribution are propagated through the dynamic. Then, in the update step, the particles are corrected using the information from the observation.

In particular, we will use a Markov Kernel to handle the propagation of particles through the dynamic ϕ. Denote by $\mu_j = \pi(v_j|z_{1:j})$ the actual distribution of the true state given the observations and $\hat{\mu_j} = \pi(v_{j+1}|z_{1:j})$. At each step j, μ_j is approximated by a collection of P particles $\{v_j^{(i)}\}_{i=1}^N$ via a distribution ${\mu_j}^N = \sum_{i=1}^N w_j^{(i)} \delta_{v_j^{(i)}}$ for weights $w_j^{(i)}$ summing to 1, and similarly, $\hat{\mu_j}$ is approximated by ${\hat{\mu_j}}^N = \sum_{i=1}^N \hat{w}_{j+1}^{(i)} \delta_{\hat{v}_{j+1}^{(i)}}$. The particle filter proceeds in two

steps: predicting the next location of the true state, and correcting this prediction using the observation.

For the prediction step, the particles $\{v_j^{(i)}\}_{i=1}^N$ are propagated through the dynamics through the relevant kernel; we have that $\hat{v}_{j+1}^{(i)} \sim p(v_j^{(i)}, \cdot)$. For the correction step, the data is considered through $g_j(v_{j+1}) \propto \pi(z_{j+1}|v_{j+1})$. Thus, each of the weights $w_j^{(i)}$ is updated according to $\hat{w}_{j+1}^{(i)} = g_j(\hat{v}_{j+1}^{(i)}) w_j^{(n)}$. However, if this technique is followed it will lead to a situation where one particle has weight 1 and the rest tend to 0. To alleviate this problem, a re-sampling step is added, in which the particles are re-sampled according to their weights $\hat{w}_{j+1}^{(i)}$, and each of these re-sampled particles are given equal weight. A cartoon demonstrating the particle filter can be seen in Fig. 5.1.

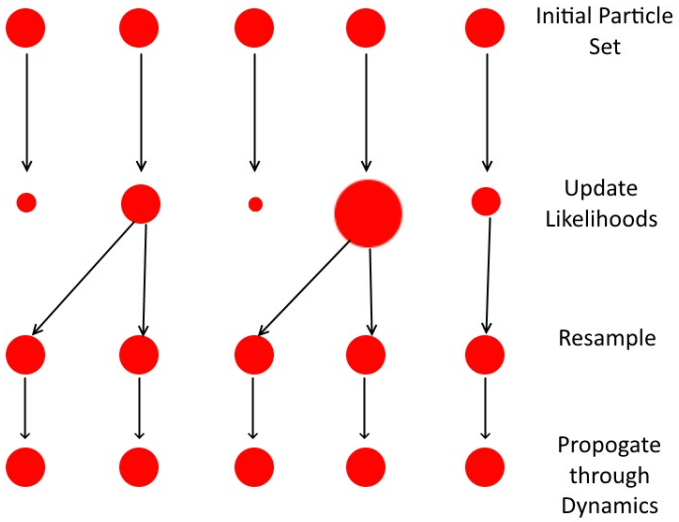


Figure 5.1: Cartoon demonstrating how the particles are updated through the particle filter.

5.3 Topological Stability for the Particle Filter

We now present a stability result for persistence diagrams of the particle filter smoothed path compared to the optimal path. We first must establish the three lemmas in line with Law et al. [52]. First, we establish the following operators on the space of probability measure on $\mathbb{R}^n$:

$$(L_j\mu)(dv) = \frac{g_j(v)\mu(dv)}{\int_{\mathbb{R}^n} g_v(v)\mu(dv)}$$

$$(P(\mu)(dv) = \int_{\mathbb{R}^n} p(v', dv)\mu(dv')$$

$$(S^N\mu)(dv) = \frac{1}{N}\sum_{i=1}^{N} \delta_{v^{(n)}}(dv), v^{(i)} \sim \mu i.i.d.$$

These represent the likelihood, dynamics, and sampling operators. In particular, it is important to note that $\mu_{j+1}^N = L_j S^N P \mu_j^N$ and that $\mu_{j+1} = L_j P \mu_j$. For the below lemmas, let $\mu, \nu, \nu' \in \pi(\mathbb{R}^n)$, the space of probability measures on $\mathbb{R}^n$.

Lemma 5.3.1. $sup_{\mu \in P(\mathbb{R}^n)} sup_{|f|_\infty \leq 1} \sqrt{\mathbb{E}|S^N\mu(f) - \mu(f)|^2} \leq \frac{1}{\sqrt{N}}$

Lemma 5.3.2. $sup_{|f|_\infty \leq 1} \sqrt{\mathbb{E}|P\nu(f) - P\nu'(f)|^2} \leq sup_{|f|_\infty \leq 1} \sqrt{\mathbb{E}|\nu(f) - \nu'(f)|^2}$

Lemma 5.3.3. *Assume there exists a* $\kappa \in (0, 1]$ *such that for all* $v \in \mathbb{R}^n$ *and* $j \in \mathbb{N}$,

$$\kappa \leq g_j(v) \leq \kappa^{-1},$$

then we have that $sup_{|f|_\infty \leq 1} \sqrt{\mathbb{E}|L_j\nu - L_j\mu|} \leq 2\kappa^{-2}\sqrt{\mathbb{E}|\nu - \mu|}$

Theorem 5.4 summarizes the stability results and basically shows that we can get close to the homology of the optimal smoothed path with the mean particle filter path. Generically, this path is reflective of the dynamics' general behavior and may be used to present bifurcations as parameters are changed in the system.

Theorem 5.4. *Assume there is some probability* $\delta \in [0, 1]$ *such that discrete dynamics as in Eqs.* (5.1) *and* (5.2) *occur in a region* $\Omega \in \mathbb{R}^n$ *bounded by* R. *Assume there exists a* $\kappa \in (0, 1]$ *such that for all* $v \in \mathbb{R}^n$ *and* $j \in \mathbb{N}$,

$$\kappa \leq g_j(v) \leq \kappa^{-1},$$

and then we have that with probability δ,

$$W_p(dgm(\overline{\mu_J}), dgm(\mathbb{E}^\omega \overline{\mu_J^N}) < \epsilon$$

where $\overline{\mu_J}$ is the optimal smoothed path and $\mathbb{E}^\omega \overline{\mu_J^N}$ is the expected particle filter path with N particles up to time J. Here $\epsilon = O(N^{-1/2})$, with constant depending on R, κ, and J, and δ depending on $sup_{x\in\Omega}|\psi(x) - x|$ and the distribution of the dynamic noise ξ_j.

Proof. We have that

$$sup_{|f|_\infty \leq 1} \sqrt{\mathbb{E}|\mu_{j+1}^N - \mu_{j+1}|^2} \leq sup_{|f|_\infty \leq 1} \sqrt{\mathbb{E}|L_j S^N P\mu_j^N - L_j P\mu_j|^2}.$$

By the triangle inequality, the above is bounded by

$$sup_{|f|_\infty \leq 1} \sqrt{\mathbb{E}|L_j S^N P\mu_j^N - L_j P\mu_j^N|^2} + sup_{|f|_\infty \leq 1} \sqrt{\mathbb{E}|L_j P\mu_j^N - L_j P\mu_j|^2}$$

By Lemmas 5.3.2 and 5.3.3, this is bounded by

$$2\kappa^{-2}(sup_{|f|_\infty \leq 1} \sqrt{\mathbb{E}|S^N P\mu_j^N - P\mu_j^N|^2} + sup_{|f|_\infty \leq 1} \sqrt{\mathbb{E}|\mu_j^N - \mu_j|^2})$$

Applying Lemma 5.3.1, we bound this by

$$2\kappa^{-2}(\frac{1}{\sqrt{N}} + sup_{|f|_\infty \leq 1} \sqrt{\mathbb{E}|\mu_j^N - \mu_j|^2})$$

Thus for each $1 \leq j$, we have that

$$sup_{|f|_\infty \leq 1} \sqrt{\mathbb{E}|\mu_{j+1}^N - \mu_{j+1}|^2} \leq 2\kappa^{-2}(\frac{1}{\sqrt{N}} + sup_{|f|_\infty \leq 1} \sqrt{\mathbb{E}|\mu_j^N - \mu_j|^2})$$

By Jensen's inequality with $\phi(x) = |x|^2$,

$$sup_{|f|_\infty \leq 1} |\mathbb{E}^\omega \mu_J^N(f) - \mu_J(f)| \leq \sup_{|f|_\infty \leq 1} \sqrt{\mathbb{E}^\omega |\mu_J^N(f) - \mu_J(f)|^2} \leq \sum_{j=1}^{J} (2\kappa^{-2})^j \frac{1}{\sqrt{N}}$$

Though the above is for $\|f\|_\infty \leq 1$, we can still obtain

$$|\mathbb{E}^\omega \mu_J^N(q) - \mu_J(q)| \leq \|q\|_\infty \sum_{j=1}^{J} (2\kappa^{-2})^j \frac{1}{\sqrt{N}}$$

for any continuous q.

Since the dynamics occur within a bounded region $\Omega \subset \mathbb{R}^n$, we can find a value so that $\Omega \subset B(0, R)$; taking q to be the identity gives

$$|\mathbb{E}^{\omega}\overline{\mu_J^N} - \overline{\mu_J}| \leq R\sum_{j=1}^{J}(2\kappa^{-2})^j \frac{1}{\sqrt{N}}$$

with probability δ. Note that $\mathbb{E}^{\omega}\overline{\mu_J^N}$ is the expectation of the mean smoothed particle filter estimate, and $\overline{\mu_J}$ is the optimal estimate given the data. As we increase the number of particles $\overline{\mu_J^N}$ becomes closer to $\mathbb{E}^{\omega}\overline{\mu_J^N}$.

Denote $\tilde{v_j} = \mathbb{E}\overline{\mu_j^N}$ and $v_j = \overline{\mu_j}$. We now apply Theorem 5.3. Notice that $d_H(\mathbb{E}\overline{\mu_{1:J}}, \overline{\mu_{1:J}^N}) \leq \frac{C}{\sqrt{N}}$, and so we have that

$$W_p(dgm(v), dgm(\tilde{v})) \leq 2^p T_f \frac{C}{N^{\frac{1}{2}}}$$

for all $p \geq d$, where T_f is the bound on the total number of features in the persistence diagrams of f and g. Note that T_f depends only on the total path length J. □

5.4 Simulations

In order to experimentally illustrate the importance of above result, we consider several dynamics under various noise conditions. For each of these experimental trials, we present the ground truth path and compare it to the particle filter path. Though our Theorem 5.4 establishes a stability relationship between the optimal particle filter path and the expected particle filter path, it is impossible in practice to observe the optimal particle filter path. The optimal particle filter path is representative of the ground truth in the case of the identify observation function. Due to this, we compare the expected particle filter path to the ground truth location of our simulated dynamics. As a result, the noise of our system plays an important role in these simulations by gauging the expected difference between the optimal path and the ground truth. We suspect that the homology of filtered paths will typically be cleaner than that of the ground truth; however, there are cases when the

particle filter performs poorly to ascertain the homology of the system when transitions are categorized by rare events.

We first examine the following system:

$$x_t = \begin{bmatrix} cos(.1) & sin(.1) \\ -sin(.3) & cos(.3) \end{bmatrix} x_{t-1}(.5 + ((x_{t-1}^1)^2 + (x_{t-1}^2)^2)/100)^{-1} + \xi_t \tag{5.3}$$

$$y_t = x_t^1 + \eta_t \tag{5.4}$$

where $x_t \in \mathbb{R}^2, y_t \in \mathbb{R}$, x_t^i is the i^{th} component of x_t and

$$\xi_t \sim N((0,0), I) \tag{5.5}$$

$$\eta_t \sim N((0,0), I) \tag{5.6}$$

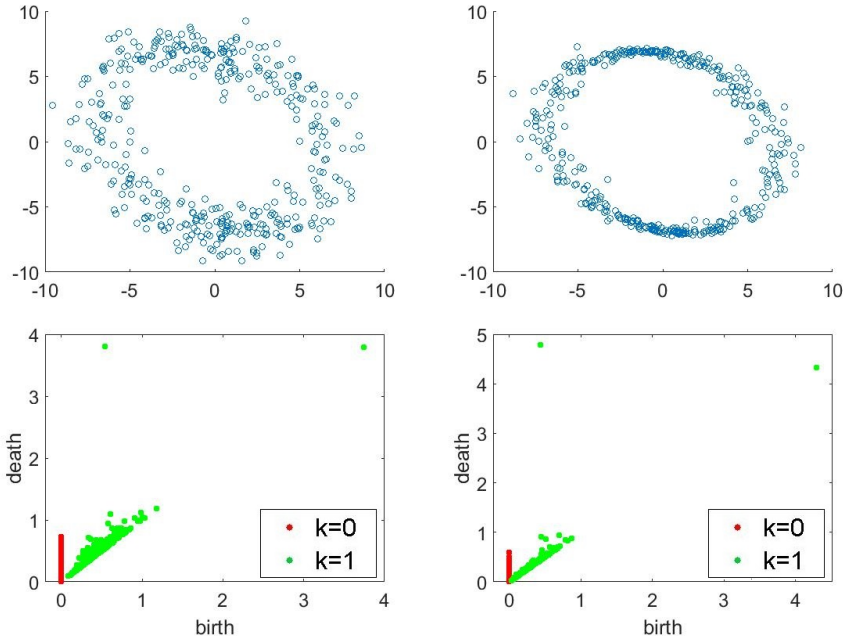


Figure 5.2: Ground truth path (top left), particle filter path (top right), ground truth persistence diagram (bottom left), and particle filter path persistence diagram (bottom right) for Equations (1-4) with dynamic noise variance I.

These dynamics produce an ellipse in $\mathbb{R}^2$, as seen in Figure 5.2. Notice that the persistent homology of the particle filter path is very similar to the persistent homology of the ground

truth. The Wasserstein Distance between the diagrams is on the scale of the noise and both diagrams exhibit a prominent cycle.

We now run the same system as above with less noise. In particular, we take the variance of our dynamic noise to be $0.1I$. We demonstrate this in Figure 5.3. The Wasserstein distance between the diagrams is still on the scale of the noise, which is much smaller in this case.

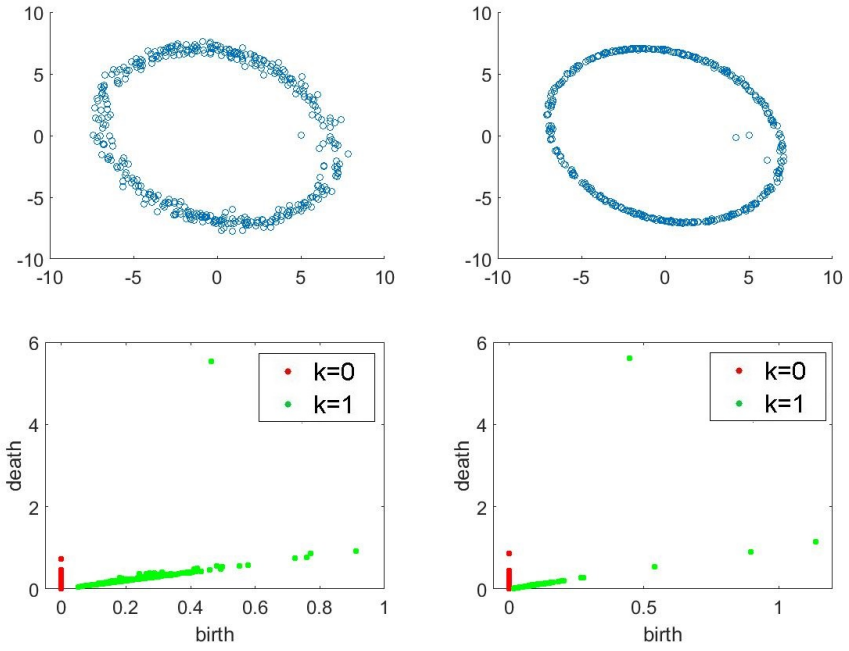


Figure 5.3: Ground truth path (top left), particle filter path (top right), ground truth persistence diagram (bottom left), and particle filter path persistence diagram (bottom right) for Equations (1-4) with dynamic noise variance $0.1I$.

Notice in the above examples the initial conditions can be clearly seen inside the circle. Depending on the choice of initial condition, this can drastically affect the persistent homology. In practice, it may be a good idea to let the process "burn-in", discarding some first number of points in both paths. Finally, we consider a different type of dynamic. In this case, we consider a one-dimensional bi-stable dynamic given by

$$x_t = 2.5(sin(x_{t-1}))^{\frac{1}{27}} + \xi_t \tag{5.7}$$

$$y_t = x_t + \eta_t \tag{5.8}$$

where $x_t \in \mathbb{R}, y_t \in \mathbb{R}$, and

$$\xi_t \sim N(0, .4) \tag{5.9}$$

$$\eta_t \sim N(0, 1) \tag{5.10}$$

The large root term is chosen to better separate the two stable states. This is demonstrated in Figure 5.4. Notice that in this case, the ground truth persistent homology seems to do a better job of picking up the bi-stability in the system than the smoothed path, as opposed to the above examples.

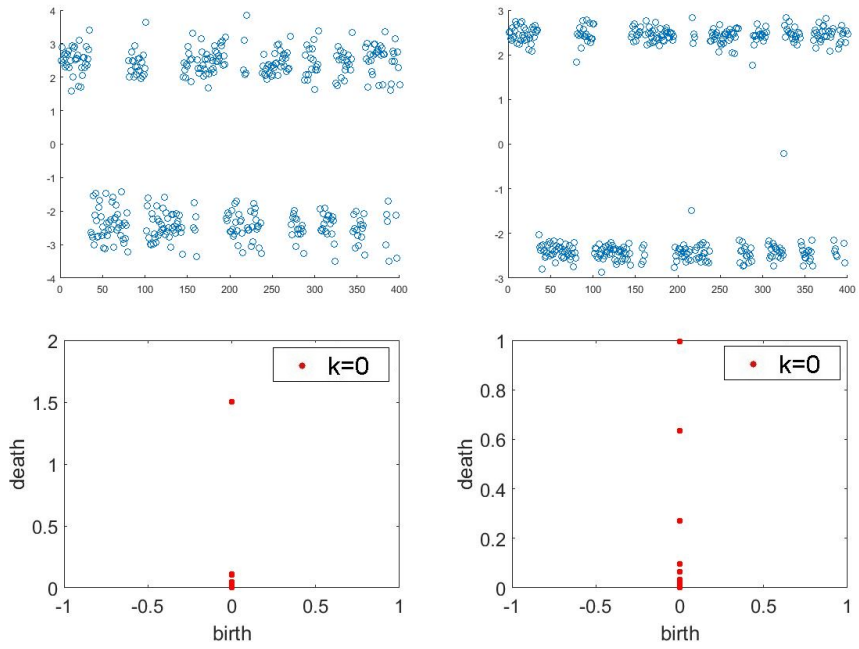


Figure 5.4: Ground truth path (top left), particle filter path (top right), ground truth persistence diagram (bottom left), and particle filter path persistence diagram (bottom right) for Equations (5-8).

5.5 Discussion and Conclusion

We have demonstrated the ability to capture the homology of a dataset and its applicability to filtering. Such a preliminary investigation opens the way for more sophisticated questions about the nature and importance of persistent homology. For example, a complex dynamic

can exhibit different speed at different points along its trajectory. Consequently, the persistence of features where the dynamic's speed and noise are small will be emphasized over features where the dynamic's speed and noise are large. While the effects of noise are difficult to manage, the speed of the dynamics are essentially expressed as a relative point density in the overall path.

Moreover, in future work we will work to establish a stability between the ground truth and the observed particle filter path. However, we note that in real data this will heavily depend on the observation function, which often maps the high dimensional dynamic into a lower dimensional observation space. Due to this, such a bound may be intractable in all but very simple situations (such as identity or linear observation functions, which are already well studied Kalman et al. [45]).

In recent works Bubenik [17], Bobrowski and Mukherjee [12], Bobrowski and Weinberger [13], Bobrowski et al. [14], a concerted effort has been made to apply statistics to the realm of persistent homology. While Mileyko et al. [60] has defined a probability measure on the set of persistence diagrams by pushing forward the probabilities of sample sets, this description offers no insight into the structure or computation of such probability spaces. In order to truly utilize persistence diagrams, we need some probability distribution that offers several competing qualities:

- Close to the actual distribution (push-forward).
- Representation as a probability density.
- Computability of random samples.
- Flexibility in the number of topological features present in each sample.

In future work, we will develop a probabilistic framework on the space of persistence diagrams. Previous attempts to impose structure on the statistics of persistence homology generally utilize the bottleneck distance. This distance is less sensitive to smaller or multiple features and can be overly sensitive to outliers. In order to improve the implementation of topological methods in data analysis, this suggests the need for a more general theory for defining probabilistic framework from a general metric on the space of persistence diagrams.

Moreover, defining probability distributions on the space of persistence diagrams would allow us to study hidden markov model problems where the observations themselves are persistence diagrams. For example, consider a sensor tracking network in which sensors are only able to communicate with their local neighbors. This local information is enough to produce a Vietoris-Rips complex, and thus a persistence diagram, which can be seen as a summary of the current system state. This information provides insight into the coverage gaps in the network, giving valuable insight into weaknesses for threat detection.

Bibliography

[1] Aaron Adcock, Erik Carlsson, and Gunnar Carlsson. The ring of algebraic functions on persistence bar codes. *Homology, Homotopy and Applications*, 18(1):381–402, 2016. 24

[2] Aaron Adcock, Erik Carlsson, and Gunnar Carlsson. The ring of algebraic functions on persistence bar codes. *Homology, Homotopy and Applications*, 18(1):381–402, 2016. 2, 17

[3] Robert J Adler, Omer Bobrowski, and Shmuel Weinberger. Crackle: The homology of noise. *Discrete & Computational Geometry*, 52(4):680–704, 2014. 4, 26, 29, 35

[4] Rakesh Agrawal, Christos Faloutsos, and Arun Swami. Efficient similarity search in sequence databases. In *International Conference on Foundations of Data Organization and Algorithms*, pages 69–84. Springer, 1993. 40

[5] Rushil Anirudh, Vinay Venkataraman, Karthikeyan Natesan Ramamurthy, and Pavan Turaga. A Riemannian framework for statistical analysis of topological persistence diagrams. *arXiv:1605.08912*, 2016. 8, 24

[6] Nieves Atienza, Rocio Gonzalez-Diaz, and Matteo Rucco. Separating topological noise from features using persistent entropy. In *Federation of International Conferences on Software Technologies: Applications and Foundations*, pages 3–12. Springer, 2016. 23

[7] M. R. Azimi-Sadjadi, Y. Yang, and S. Srinivasan. Acoustic classification of battlefield transient events using wavelet subband features. *Proc. SPIE Defense and Security Symposium*, page 6562, 2007. 1, 17

[8] Sivaraman Balakrishnan, Brittany Fasy, Fabrizio Lecci, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman. Statistical inference for persistent homology. 2013. 23

[9] Maria Bampasidou and Thanos Gentimis. Modeling collaborations with persistent homology. *arXiv preprint arXiv:1403.5346*, 2014. 17

[10] Ulrich Bauer. Ripser, 2015. https://github.com/Ripser/ripser. 3, 43, 45

[11] Alex Beskos, Dan Crisan, Ajay Jasra, Kengo Kamatani, and Yan Zhou. A stable particle filter in high-dimensions. *arXiv preprint arXiv:1412.3501*, 2014. 57

[12] Omer Bobrowski and Sayan Mukherjee. The topology of probability distributions on manifolds. *Probability Theory and Related Fields*, 161(3-4):651–686, 2015. 69

[13] Omer Bobrowski and Shmuel Weinberger. On the vanishing of homology in random Čech complexes. *arXiv:1507.06945*, 2015. 69

[14] Omer Bobrowski, Sayan Mukherjee, and Jonathan E Taylor. Topological consistency via kernel estimation. 69

[15] Omer Bobrowski, Sayan Mukherjee, and Jonathan E Taylor. Topological consistency via kernel estimation. *arXiv:1407.5272*, 2014. 4

[16] Peter Bubenik. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, 16(1):77–102, 2015. 17

[17] Peter Bubenik. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, 16(1):77–102, 2015. 69

[18] Peter Bubenik. Discovering geometry using topological data analysis. Talk at Joint Mathematics Meetings, 2016. https://jointmathematicsmeetings.org/amsmtgs/2180_abstracts/1125-55-1287.pdf. 24

[19] Olivier Cappé, Eric Moulines, and Tobias Rydén. Inference in hidden markov models. In *Proceedings of EUSFLAT Conference*, pages 14–16, 2009. 5

[20] Erik Carlsson, Gunnar Carlsson, and Vin De Silva. An algebraic topological method for feature identification. *International Journal of Computational Geometry*, 16(4):291–314, 2006. 2

[21] G. Carlsson. Topology and data. *Bull. Am. Math. Soc.*, 46(2):255–308, 2009. 2, 17, 18

[22] Kin-Pong Chan and Ada Wai-Chee Fu. Efficient time series matching by wavelets. In *Data Engineering, 1999. Proceedings., 15th International Conference on*, pages 126–133. IEEE, 1999. 40

[23] Frédéric Chazal, Vin De Silva, and Steve Oudot. Persistence stability for geometric complexes. *Geometriae Dedicata*, 173(1):193–214, 2014. 58, 59

[24] Chao Chen and Michael Kerber. Persistent homology computation with a twist. In *Proceedings 27th European Workshop on Computational Geometry*, volume 11, 2011. 4

[25] Charles K Chui. Wavelets: a tutorial in theory and applications. *Wavelet Analysis and its Applications, San Diego, CA: Academic Press,— c1992, edited by Chui, Charles K.*, 1, 1992. 40

[26] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. *Discrete Comput. Geom*, 37:103–120, 2007. 58

[27] Dan Crisan and Arnaud Doucet. A survey of convergence results on particle filtering methods for practitioners. *IEEE Transactions on signal processing*, 50(3):736–746, 2002. 57

[28] P. Dhanalakshmi, S. Palanivel, and V. Ramalingam. Classification of audio signals using svm and rbfnn. *Expert Systems with Applications*, 36(3):6069–6075, 2009. 1, 17

[29] Arnaud Doucet, Simon Godsill, and Christophe Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and computing*, 10(3):197–208, 2000. 5

[30] Arnaud Doucet, Nando De Freitas, and Neil Gordon. An introduction to sequential monte carlo methods. In *Sequential Monte Carlo methods in practice*, pages 3–14. Springer, 2001. 60

[31] H. Edelsbrunner and J. Harer. *Computational Topology: An Introduction*. American Mathematical Society, 2010. 13, 14, 23

[32] Herbert Edelsbrunner and John Harer. *Computational topology: an introduction*. American Mathematical Society, 2010. 2, 3, 7, 10, 14, 23, 58

[33] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, pages 454–463. IEEE, 2000. 23

[34] Kevin Emmett, Daniel Rosenbloom, Pablo Camara, and Raul Rabadan. Parametric inference using persistence diagrams: A case study in population genetics. *arXiv:1406.4582*, 2014. 4

[35] S. Emrani, T. Gentimis, and H. Krim. Persistent homology of delay embeddings and its application to wheeze detection. *Signal Processing Letters, IEEE*, 21(4):459–463, 2015. 2, 3, 4, 24, 37

[36] Evangelos Evangelou and Vasileios Maroulas. Online filtering and estimation for dynamic spatiotemporal processes with data from an exponential family. *Tech. Rep., University of Bath*, 2015. 61

[37] Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, and Aarti Singh. Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6):2301–2339, 2014. 4

[38] Fabian Flock, 2016. xi, 40

[39] Ashit Gandhi. Content-based image retrieval: Plant species identification, 2002. 44

[40] D. Garrett, D. A. Peterson, C. W. Anderson, and M. H. Thaut. Comparison of linear, nonlinear, and feature selection methods for eeg signal classification. *IEEE Transactions on Neural System and Rehabilitation Engineering*, 11:141–166, 2003. 1, 17

[41] Mordecai J. Golin. Bipartite matching & the hungarian method. `http://www.cse.ust.hk/~golin/COMP572/Notes/Matching.pdf`. 38

[42] Neil Gordon, B Ristic, and S Arulampalam. Beyond the kalman filter: Particle filters for tracking applications. *Artech House, London*, 830, 2004. 5

[43] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2009. 21, 39, 42

[44] Allen Hatcher. *Algebraic topology*. Cambridge University Press, 2002. 13

[45] Rudolph Emil Kalman et al. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960. 61, 69

[46] Kai Kang, Vasileios Maroulas, and Ioannis D Schizas. Drift homotopy particle filter for non-gaussian multi-target tracking. In *Information Fusion (FUSION), 2014 17th International Conference on*, pages 1–7. IEEE, 2014. 61

[47] Nikolas Kantas, Arnaud Doucet, Sumeetpal S Singh, Jan Maciejowski, Nicolas Chopin, et al. On particle methods for parameter estimation in state-space models. *Statistical science*, 30(3):328–351, 2015. 57

[48] Kyoji Kawagoe and Tomohiro Ueda. A similarity search method of time series data with combination of fourier and wavelet transforms. In *Temporal Representation and Reasoning, 2002. TIME 2002. Proceedings. Ninth International Symposium on*, pages 86–92. IEEE, 2002. 40

[49] M. Kerber, D. Morozov, and A. Nigmetov. Geometry helps to compare persistence diagrams. *Proceedings of the Eighteenth Workshop on Algorithm Engineering and Experiments*, pages 103–112, 2016. 25

[50] Michael Kerber, Dmitriy Morozov, and Arnur Nigmetov. Geometry helps to compare persistence diagrams. In *Proceedings of the Eighteenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, pages 103–112. SIAM, 2016. 3, 55

[51] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–87, 1955. 38, 55

[52] Kody Law, Andrew Stewart, and Konstantinos Zygalakis. *Data Assimilation A Mathematical Introduction*. Springer, 2015. 46, 60, 62

[53] P. Y. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Vejdemo-Johansson, M. Alagappan, J. Carlsson, and G. Carlsson. Extracting insights from the shape of complex data using topology. *Scientific Reports 3*, 3(1236), 2013. 2, 17

[54] PY Lum, G Singh, A Lehman, T Ishkanov, Mikael Vejdemo-Johansson, M Alagappan, J Carlsson, and G Carlsson. Extracting insights from the shape of complex data using topology. *Scientific reports*, 3:1236, 2013. 18

[55] Andrew Marchese and Vasileios Maroulas. A point process distance on the space of persistence diagrams. *Submitted to Advances in Data Analysis and Classification, Sept. 2016*, 2016. https://sites.google.com/site/andrewmarchesemath/journal_paper_pre.pdf?attredirects=0&d=1. 3, 37

[56] Andrew Marchese and Vasileois Maroulas. Topological learning for acoustic signal identification. In *2016 19th International Conference on Information Fusion (FUSION)*, pages 1377–1381, 2016. 57

[57] V. Maroulas and A. Nebenführ. Tracking rapid intracellular movements: A bayesian random set approach. *The Annals of Applied Statistics*, 9(2):926–949, 2015. 5

[58] Vasileios Maroulas and Jie Xiong. Large deviations for optimal filtering with fractional brownian motion. *Stochastic processes and their applications*, 123(6):2340–2352, 2013. 5

[59] Y. Mileyko, S. Mukherjee, and J. Harer. Probability measures on the space of persistence diagrams. *Inverse Problems*, 27(12), 2011. 4, 19, 20, 25

[60] Yuriy Mileyko, Sayan Mukherjee, and John Harer. Probability measures on the space of persistence diagrams. *Inverse Problems*, 27(12):124007, 2011. 69

[61] Konstantin Mischaikow and Vidit Nanda. Morse theory for filtrations and efficient computation of persistent homology. *Discrete & Computational Geometry*, 50(2):330–353, 2013. x, 3

[62] Dmitriy Morozov. Dionysus, a C++ library for computing persistent homology, 2007. 3

[63] M. Nicolau, A. Levine, and G. Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences*, 108(17):7265–7270, 2011. 2, 17

[64] A. V Oppenheim and R. W. Schafer. From frequency to quefrency: A history of the cepstrum. *IEEE Signal Processing Magazine*, pages 95–106, 2004. 40

[65] 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. 1, 2, 3, 8

[66] Jose A Perea, Anastasia Deckard, Steve B Haase, and John Harer. Sw1pers: Sliding windows and 1-persistence scoring; discovering periodicity in gene expression time series data. *BMC bioinformatics*, 16(1):257, 2015. 2, 3, 8, 35, 54

[67] Cássio MM Pereira and Rodrigo F de Mello. Persistent homology for time series and spatial data clustering. *Expert Systems with Applications*, 42(15):6026–6038, 2015. 3, 4, 18

[68] Florian T Pokorny, Danica Kragic, Lydia E Kavraki, and Ken Goldberg. High-dimensional winding-augmented motion planning with 2d topological task projections and persistent homology. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pages 24–31. IEEE, 2016. 23

[69] Chotirat Ann Ratanamahatana and Eamonn Keogh. Everything you know about dynamic time warping is wrong. In *Third Workshop on Mining Temporal and Sequential Data*. Citeseer, 2004. 41

[70] Guohua Ren, Vasileios Maroulas, and Ioannis Schizas. Distributed spatio-temporal association and tracking of multiple targets using multiple sensors. *IEEE Transactions on Aerospace and Electronic Systems*, 51(4):2570–2589, 2015. 5

[71] Vanessa Robins and Katharine Turner. Principal component analysis of persistent homology rank functions with case studies of spatial point patterns, sphere packing and colloids. *Physica D: Nonlinear Phenomena*, 334:99–117, 2016. 17, 24

[72] Md Sahidullah and Goutam Saha. Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition. *Speech Communication*, 54(4):543–565, 2012. 1

[73] Hiroaki Sakoe and Seibi Chiba. A dynamic programming approach to continuous speech recognition. In *Proceedings of the seventh international congress on acoustics*, volume 3, pages 65–69. Budapest, Hungary, 1971. 41

[74] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978. 41

[75] D. Schuhmacher, B. Vo, and B. Vo. A consistent metric for performance evaluation of multi-object filters. *IEEE Trans. Signal Process.*, pages 3447–3457, 2008. 5, 27

[76] Lee M. Seversky, Shelby Davis, and Matthew Berger. On time-series topological data analysis: New data and opportunities. *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 59–67, 2016. 17, 24

[77] S Shalev-Shwartz and S Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. 18, 21

[78] J. Sherwin and P. Sajda. Musical experts recruit action-related neural structures in harmonic anomaly detection: Evidence for embodied cognition in expertise. *Brain and Cognition*, 83:190–202, 2013. 1, 17

[79] U. Srinivas, N. M. Nasrabadi, and V. Monga. Graph-based multi-sensor fusion for acoustic signal classification. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages pp. 261 – 265, 2013. 1, 40

[80] Gilbert Strang. Wavelet transforms versus fourier transforms. *Bulletin of the American Mathematical Society*, 28(2):288–305, 1993. 40

[81] F. Takens. Detecting strange attractors in turbulence. *Dynamical Systems and Turbulence, Warwick 1980, Lecture Notes in Mathematics*, 898:366–381, 1980. 2, 7, 8, 17

[82] K. Turner, Y. Mileyko, S. Mukherjee, and J. Harer. Fréchet means for distribution of persistence diagrams. *Discrete Computational Geometry*, 52:44–70, 2014. 43

[83] Katharine Turner, Yuriy Mileyko, Sayan Mukherjee, and John Harer. Fréchet means for distributions of persistence diagrams. *Discrete & Computational Geometry*, 52(1): 44–70, 2014. 4

[84] V. Venkataraman, K. N. Ramamurthy, and P. Turaga. Persistent homology of attractors for action recognition. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 4150–4154, 2016. 1, 2, 3, 4, 8, 24, 35, 54, 57

[85] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963. 42

[86] Kelin Xia and Guo-Wei Wei. Persistent homology analysis of protein structure, flexibility, and folding. *International Journal for Numerical Methods in Biomedical Engineering*, 30(8):814–844, 2014. 17, 24

[87] Jie Xiong. *An introduction to stochastic filtering theory*, volume 18. Oxford University Press on Demand, 2008. 5

[88] Min Xu, Ling-Yu Duan, Jianfei Cai, Liang-Tien Chia, Changsheng Xu, and Qi Tian. Hmm-based audio keyword generation. In *Pacific-Rim Conference on Multimedia*, pages 566–574. Springer, 2004. 1

[89] Z. Zhang, Y. Song, H. Cui, J. Wu, F. Schwartz, and H. Qi. Early mastitis diagnosis through topological analysis of biosignals from low-voltage alternate current electrokinetics. *Conf Proc IEEE Eng Med Biol Soc*, pages 542–545, 2015. 1, 8, 24

[90] Dong Zhen, HL Zhao, Fengshou Gu, and AD Ball. Phase-compensation-based dynamic time warping for fault diagnosis using the motor current signal. *Measurement Science and Technology*, 23(5):055601, 2012. xi, 42

[91] Xiaojin Zhu. Persistent homology: An introduction and a new text representation for natural language processing. In *IJCAI*, pages 1953–1959, 2013. 4

Vita

Andrew Marchese was born in Long Island, NY on September 16th, 1989. He attended Mount Sinai High School where his interest in mathematics was fostered by his parents and his math teacher, Gary Kulik. In 2007, he started attending Stony Brook University. After a brief hiatus from studying mathematics, Andrew returned to the subject under the generous mentorship of Professor Lisa Berger. Shortly thereafter, Andrew attended the University of Arizona for an REU program, which helped develop his interest in mathematics research. Andrew graduated with a B.S. in Mathematics from Stony Brook in 2011.

In 2012, Andrew started to pursue his Ph.D. at the University of Tennessee, Knoxville. After passing his preliminary exams in Analysis and Algebra, Andrew moved towards Data Methods and Machine Learning using Probability, Statistics, and Topology. In 2014, Andrew was able to pursue these interests working with Professor Vasileios Maroulas in collaboration with Dr. Chris Symons at Oak Ridge National Lab. Shortly thereafter, Andrew started working on techniques focused on combining topology with data analysis, which led to the research contained in his dissertation.