Bayesian Topological Machine Learning

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Christopher Anthony Oballe
August 2020

Dedicated to my friends and family, whose unyielding support and encouragement have made this possible.

Acknowledgments

I would like to express my utmost gratitude to my advisor, Professor Vasileios Maroulas, whose guidance and support have sharpened my professional skill set, broadened my research interests, and led me to abundant opportunities for success.

I would also like to thank all of my committee members, Dr. Piotr Franaszczuk, Professor Mathew Langford, and Professor Michael Berry, as well as my close research collaborators at Army Research Lab, David Boothe and Scott Kerick, for their useful insights and willingness to help me grow as a researcher.

I would like to extend a special thanks to two of my co-authors, Dr. Farzana Nasrin and Dr. Joshua Mike for their collaboration and contributions to my published works. In particular, Dr. Mike (alongside my advisor) developed the theoretical foundations of the kernel density estimator for persistence diagrams described in Chapter 4, which allowed me to apply it in my research.

Of course, I am deeply grateful for the love of my family, in particular my mother and stepfather, Regina and David Waldo, who have incessantly encouraged my interests and intensely supported me throughout the good and bad times throughout my life, my father and grandfather, Chris Oballe (Sr.) and James Harrison, whose collective wisdom has steered me toward a life of integrity, my grandmother, Paula Davis, and mother again, whose warm-heartedness and compassion have always made me feel accepted and provided a blueprint for how to treat others. My little brother, Matthew Oballe, only gets an honorable mention because he called me a nerd once.

I would also like to thank my close friends, Noah Brewer, Joel Alvis, Zach Givens, Sean Haugh, Ray Nathanson, Eric Olberding, and Alan Cherne for keeping me sane and exchanging ideas while we mutually figure out how to navigate life.

My PhD thesis was partially sponsored by the US Army Research Office Grant number W911NF-17-1-0313, Thor Industries/Army Research lab with Grant number W911NF-17-2-0141, and the US Army Research Lab under Cooperative Agreement Number W911NF-19-2-0302.

Abstract

Topological data analysis encompasses a broad set of ideas and techniques that address 1) how to rigorously define and summarize the shape of data, and 2) use these constructs for inference. This dissertation addresses the second problem by developing new inferential tools for topological data analysis and applying them to solve real-world data problems. First, a Bayesian framework to approximate probability distributions of persistence diagrams is established. The key insight underpinning this framework is that persistence diagrams may be viewed as Poisson point processes with prior intensities. With this assumption in hand, one may compute posterior intensities by adopting techniques from the theory of marked point processes. After defining a Bayesian model in generality, a conjugate family of prior intensities is introduced via Gaussian mixtures to obtain a closed form of the posterior intensity. This enables efficient computation of posterior distributions for persistence diagrams. The utility of this Bayesian framework is demonstrated on classification problems with materials science and electroencephalography data. Viewing persistence diagrams as point processes, one may also define a kernel density estimator to approximate probability distributions of persistence diagrams in a nonparametric fashion. This dissertation uses the kernel density estimator to create a novel hypothesis test to detect specific time series dynamics in noisy measurements. Finally, the problem of data augmentation, the overarching goal of which is to increase training set diversity by generating additional training examples from existing ones by preserving large scale structures in elements of the training set, is considered. Herein, a novel data augmentation framework that considers the topology of data is introduced. Intuitively speaking, this new method 'adds noise' to training examples through controlled topological perturbations, which preserve large scale structure in data. The effectiveness of the novel data augmentation pipeline in training deep learners to classify atomic probe tomography and image data in the cases of balanced and unbalanced training examples is examined.

Contents

1 Introduction **3**

2 Background **6**
2.1 Persistent Homology 6
2.1.1 Vietoris-Rips Filtrations 6
2.1.2 Sublevel Set Filtrations 8
2.2 Point Processes 10
2.2.1 Intuition 10
2.2.2 A Formal Viewpoint 12
2.2.3 Finite Set Statistics 15

3 A Bayesian Framework for Persistent Homology **19**
3.1 Bayesian Inference 20
3.1.1 Model 20
3.1.2 A Conjugate Family of Prior Intensities: Gaussian Mixtures 25
3.1.3 Sensitivity Analysis 26
3.2 A Note on Bandwidth Selection 34
3.3 Classification 36
3.3.1 Atom Probe Tomography Data 37
3.4 Bayesian Topological Signal Processing 39
3.4.1 Signal Processing Preliminaries 40
3.4.2 Topological Uncertainty Quantification for Random Signals 40
3.4.3 Signal Processing Results 44

4 Nonparametric Density Estimation for Persistence Diagrams **51**
4.1 Kernel Density Estimation 51
4.1.1 Construction 52
4.1.2 A Measure of Dispersion 53
4.1.3 Application to Neuroimaging Data 54

5 Data Augmentation with Topological Noise **59**
5.1 Background 59
5.1.1 Computational Topology Preliminaries 59
5.1.2 Point Cloud Classification with PointNet 61
5.2 Data Augmentation 61
5.3 Experimental Results 64
5.3.1 Point Cloud Data 64
5.3.2 Image Data 66

6 Conclusion **69**

References **71**

Vita **77**

List of Tables

3.1 The parallels between the Bayesian framework for RVs and its counterpart for random PDs. 21

3.2 Case-IV: The first, second and third columns match the parameters of observed persistence diagram $\mathcal{D}_Y$ used in computing the posteriors depicted in the third column of Figure 3.2 and Figure 3.3, and second column of Figure 3.4, respectively with $\alpha = 0.5$. The parameters are presented in Table 3.4. The color maps represent scaled intensities. A variation in the level of intensity is observed for all of them compared to their respective cases due to the added term in the posterior intensity. The posterior intensities in first and second columns exhibit the estimation of the hole with higher variability when compared to their respective figures in Case-I and Case-II. The posteriors in the third column demonstrate dominance of the prior relative to their corresponding figures in Case-III, especially when one examines those for informative, weakly informative and bimodal uninformative priors. 31

3.3 List of Gaussian mixture parameters of the prior intensities in Equation (3.28). The means $\mu_i^{D_X}$ are 2×1 vectors and the rest are scalars 32

3.4 Parameters for (M3′) in Equation (3.29) and (3.30). We set the weight and mean of the Gaussian component, $c^{D_{Y_S}} = 1$ and $\mu^{D_{Y_S}} = (0.5, 0)$ respectively for all of the cases. The first row corresponds to parameters in the functions characterizing D_Y that are used in computing the posterior depicted in the first column of Figure 3.2. The second row corresponds to analogous parameters that are used in computing the posterior depicted in the second columns of Figures 3.2–3.2. Similarly, the third row corresponds to parameters in the functions characterizing D_Y used for computing the posterior presented in the third columns of Figures 3.2–3.4. 33

3.5 Parameters for the prior intensities used in cross-validation of materials science data. Each prior λ is indexed by its corresponding class for Prior-1 or U in the case of the Prior-2. The summary of AUCs across 10-folds for materials science data after scoring with Algorithm 1 is presented in the last three columns. 39

3.6 Parameters from fitting autoregressive models to real EEG epochs. 45

3.7 Precisions and recalls for each feature and classifier. Results are reported as mean ± standard error across each class. 47

4.1 The p-values and run times for each method (KDE, PI, and PL) used for the hypothesis test of Eq. (4.1.3). 56

4.2 The sample MADs for SNR_1 and SNR_5 computed by taking the means of the distributions in Fig 4.3(e) and Fig 4.3(f), respectively. 56

5.1 Choices for the energy function in Equation 5.7, E. 64

5.2 This table shows the maximum validation accuracy over every epoch obtained by training PointNet with augmented datasets built from a balanced training set. Results are shown using for different choices of energy function E in Algorithm 2; see Table 5.1. 64
5.3 This table shows the maximum validation accuracy over every epoch obtained by training PointNet with augmented datasets built from an unbalanced training set. results are shown for different choices of energy function E in Algorithm 2; see Table 5.1. 66
5.4 Maximum validation accuracy for each augmentation method over 24 training epoches. 67

List of Figures

2.1 Two shapes illustrating different types of homological features. Each object is a connected component and thus a 0-dimensional feature. The red triangles in the shape on the right are 1-dimensional features, while the empty region inside the tetrahedron on the left is a 2-dimensional feature. 8

2.2 Vietoris Rips complexes. 9

2.3 (a) An example of a dataset, (b) its persistence diagram, and (c) its tilted representation. 9

2.4 Show above (a) are the sublevel sets $C_{-0.5}$, C_0, $C_{0.25}$, and C_1 for a damped cosine $e^{-2t}\cos(8\pi t)$. (b) shows the persistence diagram of the sublevel set filtration. The points in (b) are colored to match the connected components their birth coordinates correspond to. The transition from C_0 to $C_{0.25}$ depicts the Elder rule; notice that in C_0, there are light blue and purple connected components, which merge together in $C_{0.25}$. A similar merging happens in the transition from $C_{0.25}$ to $C_{0.5}$. Since the purple component has a later birth value, it disappears into the light blue component, which persists until it merges into the green component by the same line of reasoning. 11

3.1 (a) A sample from the prior point process $\mathcal{D}_X$ and an observed persistence diagram $\mathcal{D}_Y$. (b) Decomposition of $\mathcal{D}_X$ into $\mathcal{D}_{X_O}$,$\mathcal{D}_{X_V}$ and $\mathcal{D}_Y$ into $\mathcal{D}_{Y_O}$,$\mathcal{D}_{Y_S}$. The points in $\mathcal{D}_{X_V}$ have no relationship to those in $\mathcal{D}_Y$, while those in $\mathcal{D}_{X_O}$ only generate observed points in $\mathcal{D}_{Y_O}$. The remaining observed points in $\mathcal{D}_{Y_S}$ model unanticipated features that one may obtain due to uncertainty/noise. 22

3.2 Case-I: Posterior intensities obtained by using Proposition 3.1.1. We consider informative (a), weakly informative (d), unimodal uninformative (g), and bimodal uninformative (j) prior intensities . The color maps represent scaled intensities. The list of associated parameters of observed persistence diagram used for this case are in Table 3.4. Posteriors computed from all of these priors estimate the $1-$dimensional hole accurately for two choice of variance in the observed persistence diagram ($\sigma^{D_{Y_O}} = 0.01$ and $\sigma^{D_{Y_S}} = 0.1$), which are presented in (b), (e), (h) and (k). After increasing the variance to $\sigma^{D_{Y_O}} = 0.1$, we observe the posteriors still estimate hole but with higher variance as presented in (c), (f), (i) and (k). 28

3.3 Case-II: We consider informative (a), weakly informative (d), unimodal uninformative (g), and bimodal uninformative (j) prior intensities to estimate posterior intensities using Proposition 3.1.1. The color maps represent scaled intensities. The parameters we use for estimating the posterior intensity are listed in Table 3.4. The posterior intensities estimated from the informative prior in (b) and (c) estimate the $1-$dimensional hole with high certainty. Also, the posterior intensities estimated from the weakly informative and uninformative priors in (e), (h), and (k) imply existence of a hole with lower certainty. (c), (f), (i) and (l) represent the posterior with higher variance in observed PD component D_{Y_S}. As this makes the assumption that every observed point is associated to D_X, we observe increased intensity skewed towards the spurious point in (f). Furthermore in (i) and (l), we observe bimodality in the posterior intensity. 29

3.4 Case-III: Posterior intensities obtained by using Proposition 3.1.1. We consider informative (a), weakly informative (d), unimodal uninformative (g), and bimodal uninformative (j) prior intensities. The color maps represent scaled intensities. Parameters of the observed PD used to estimate the posterior intensity are listed in Table 3.4. With choices of $\sigma^{D_{Y_O}} = 0.01$ and $\sigma^{D_{Y_S}} = 0.1$, we observe the posteriors can deduce existence of the prominent feature, as presented in (b), (e), (h), and (k) as we have more confidence on the component of observed data associated to prior. Otherwise, with an increased variance $\sigma^{D_{Y_O}} = 0.1$, only the posterior intensity from the informative prior is able to detect the hole with high certainty, as observed in (c). For the weakly informative and uninformative priors, the posteriors in (f), (i), and (l) may not detect the hole directly, but the mode of (f) with higher variance and the tail towards the prominent point in (i) and (l) imply the existence of a hole in the underlying persistence diagram. 30

3.5 The observed datasets generated for Case-I, Case-II, and Case-III by sampling the unit circle and perturbing with Gaussian noise having variances $0.001I_2$, $0.01I_2$, and $0.1I_2$, respectively. 32

3.6 (a) Image of APT data with atomic neighborhoods shown in detail on the left. Each pixel represents a different atom, the neighborhood of which is considered. Certain patterns with distinct crystal structures exist, e.g., the orange region is copper-rich (left), but overall no pattern is identified. Putting a single atomic cubic unit cell under a microscope, the true crystal structure of the material, which could be either body-centered cubic (BCC) (b) or face-centered cubic (FCC) (c), is not revealed. This distinction is obscured due to further experimental noise. Notice there is an essential topological difference between the two structures in (b) and (c): The BCC structure has one atom at its center, whereas the FCC is hollow in its center, but has one atom in the center of each of its faces. 37

3.7 Persistence diagrams for members of the BCC and FCC classes. 38

3.8 This figure illustrates sources of uncertainty in persistence diagrams. Shown above are signals with additive noise (a) $\mathcal{N}(0, 0.01)$, and (b) $\mathcal{N}(0, 0.1)$ along with their persistence diagrams. The persistence diagram for the true underlying signal is shown in red. Several spurious features arise due to noise. Additionally, true features also shift around. 41

3.9 (a) The damped cosine $e^{-2t}\cos(8\pi t)$ with additive noise $\mathcal{N}(0, 0.01)$ and (b) its persistence diagram. (b) shows an uninformative prior intensity with a single component at $(1, 1)$ with covariance matrix $10I$. Using the model from Equation (3.31) with the prior in (c) and the observed diagram in (b) results in the posterior intensity shown in (d). To account for spurious points, which we suspected to be low persistence in this example, we placed components at $(0.5, 0.1), (1, 0.1), (0.75, 0.1)$ and $(1.75, 0.1)$. . 45
3.10 This figure demonstrates the effect of greater low frequency power on the persistence diagram of a signal. Notice in (a) that elements of the persistence diagram show greater spread along the diagonal. This phenomenon arises since the low frequency signal scatters the higher frequency peaks in along the Amplitude axis. 47
3.11 This plot depicts the relationship between the cardinality of persistence diagrams and the frequency of the dominant oscillation for one second autoregressive signals across various damping factors. For each included frequency and damping factor, we simulated thirty signals (each had a component fixed at zero to give the $1/f$ PSD commonly seen in EEG), computed their persistence diagrams, then recorded their average cardinality. We see a strong positive correlation between this average cardinality and the frequency of the dominant oscillation (i.e., PSD Peak Frequency) consistent with the idealized deterministic sinusoid case. 48
3.12 The peak frequency f for $\mathcal{A}_f^\beta$ plotted against the average birth variance for its persistence diagrams. Colors depict the damping factor β. 48
3.13 Here are the average (log) power spectral densities along with examples of signals and persistence diagrams from each class for damping factors of 32. 49

4.1 *Left:* A persistence diagram split according to Eq. (4.1). The dashed black line, $d = b + \sigma$, separates the diagram into the red upper points of $\mathcal{D}^u$ and the yellow lower points of $\mathcal{D}^\ell$. *Right:* The red and black gradients represent the upper singleton densities $p^{(1)}$ and $p^{(2)}$ given by Eq. (4.2). The green gradient represents the lower density p^ℓ defined in Eq 4.4. While each of these densities is defined on the wedge $W \subset \mathbb{R}^2$, the global kernel in Eq. (4.6) is defined on $\bigcup_N W^N$ for each input-cardinality N. 53
4.2 A segment of an EEG signal with (a) SNR_1, and (b) SNR_5, respectively, along with (c) and (d): their corresponding periodograms (estimates of the power spectral densities). The associated point clouds are given in (e) and (f), respectively, and (g) and (h) their resulting persistence diagrams. 57
4.3 This figure shows the distribution for each statistic we considered when comparing SNR_1 to SNR_5. Each column represents a class, either SNR_1 or SNR_5, and each row a particular statistic. 58

5.1 Shown above in blue are 2-dimensional point clouds created with Algorithm 2 with learning rate β fixed at 0.01, E = Wass from Table 5.1, and the number of epochs set to (a) 100, (b) 200, (c) 400, and (d) 800, respectively. As each epoch corresponds to a step in gradient descent, we expect point clouds created with a larger amount of epochs to bear a higher degree of topological similarity to the target point cloud, shown in orange. Indeed, we observe that the blue point clouds resemble circles with radii increasing to that of the target point cloud as the number of epochs increases. . 63

5.2 Shown above are training accuracies for (a) balanced and (c) unbalanced training sets. Similarly, (b) and (d) display the corresponding validation accuracies. To prevent overcluttering, validation accuracies are only shown for traditional and was-L2 data augmentation, since was-L2 attained the highest validation accuracy (taken over every epoch) for the unbalanced training cases. 65
5.3 Original images (a,c) from the MNIST dataset and the resulting images (b,d) after topological augmentation with P. Here, lighter colored pixels have smaller values. . 67
5.4 The distribution of images in the training and validation sets are shown in (a) and (b), respectively. 68
5.5 (a) Validation accuracies and (b) validation losses shown for the unaugmented training set (None), the training set augmented with random shifts (Trad.), and the training set augmented with our topological method (Topo.), respectively. 68

Chapter 1

Introduction

In recent years, technological advances, like improvements to sensor networks [33] and state-of-the-art physical simulations [58], as well as the increasing digitization of the world [76] have led to an abundance of data, paving the way for new scientific inquiries. For example, Atom Probe Tomography [34] uses highly curved electronic fields and computer reconstruction to obtain local snapshots of atomic environments, allowing researchers to distinguish crystal structure that predicts the chemical and physical properties of a crystalline material. In neuroscience, measurements of electrical activity in the brain (electroencephalography, EEG) obtained from electrodes placed on a subject's head can provide information about brain function and serve as important clinical diagnostic tools [7]. With the recent abundance of data come challenging problems, often arising from measurement noise, high dimensionality, or class imbalances, and their resolutions demand novel viewpoints for data analysis and machine learning (i.e., automated pattern detection [65]). Through the lens of topological data analysis (TDA), data is viewed as a shape whose local and global structure are tied to aspects of data arising at different scales [27].

In particular, TDA encompasses a broad set of ideas and techniques that address 1) how to rigorously define and summarize the shape of data, and 2) use these constructs for inference [13, 74]. Solutions to the former problem were addressed with the advent of methods for computational topology [20], two notable subcategories being those reliant on computational persistent homology and Mapper, respectively. The Mapper algorithm, introduced in [39], uses multi-scale clustering to create topological networks of high-dimensional data. This allowed for, in addition to novel visualization techniques, new insights to be drawn from large datasets, the most noteworthy perhaps being the discovery of a subset of breast cancer patients with 100% survival rate and no metastasis [55]. On the other hand, the ubiquitous family of methods for computational persistent homology were introduced across several works, a comprehensive survey of which may be found in [57]. Computational persistent homology uses data as building blocks to construct nested sequences of geometrical (e.g., simplicial [21, 78] or cubical [72]) complexes, indexed by an increasing real-valued parameter. These increasing sequences, known as filtrations, are viewed as multi-scale approximations of a latent geometric object. Persistent homology records the filtration values at which each homological feature appears and disappears in a filtration of data, collects both of these values into a pair, then aggregates all pairs into a collection called a *persistence diagram* to serve as a topological summary of data.

A large body of inferential methods within TDA involves persistence diagrams. For example, the work [10] introduces persistence landscapes, Hilbert space representations of persistence diagrams built with collections of tent functions; [2] introduces another Hilbert space representation, the persistence image, which is a pixellated version of a persistence diagram created by convolving

its points with a Gaussian kernel then binning. Both of these vectorizations enjoy useful statistical properties [11, 14] and can be used as features in machine learning architectures. Class imbalances in training data (i.e. having a dramatically greater number of examples from one class versus another) and a lack of diversity in training sets present a challenge for learning algorithms, in particular those based on deep learning, [30] that topological summaries alone cannot overcome. To train deep learners and help their ability to generalize from limited data, practitioners rely on data augmentation [66, 73, 75], the overarching goal of which is to create additional training examples from existing ones by preserving large scale structures in members of the training set. For example, common data augmentation methods rely on geometric transformations (for example, rotations and translations) or the addition of noise to training examples to produce new data. The core justification for using data augmentation is that augmented datasets prevent overfitting to the training set and encourage the extraction of meaningful global features by including greater small scale variations, which are characteristic of many datasets. Chapter 5 proposes a novel data augmentation framework by incorporating the topology of data to create new training examples. Intuitively speaking, the new method 'adds noise' to training examples through controlled topological perturbations, which preserve large scale structure in data. Chapter 5 also examines the effectiveness of the novel data augmentation pipeline in training deep learners to classify APT and image data in the cases of balanced and unbalanced training examples.

Another body of inferential methods within TDA relate to statistical distributions directly involving persistence diagrams. The work [22] provides a notion of confidence sets for persistence diagrams by combining kernel density estimators from bootstrapped samples of data with stability results [15] for widely-used persistence diagram metrics. The end result of this work defines a confidence region in a persistence diagram distinguishing true topological features from those that are likely the result of noise. The first Bayesian considerations in a persistent homology context take place in [49] when the authors discuss a conditional probability setting on persistence diagrams where the likelihood for the observed point cloud has been substituted by that for its associated topological summary. Defining probability distributions of persistence diagrams themselves is a challenging problem since they have no intrinsic order, implying they are random sets as opposed to random vectors. To this end, Chapters 3 and 4 treat persistence diagrams as point processes to rigorously define probability distributions of persistence diagrams.

First, Chapter 3, a large portion of which was published in the *SIAM Journal on Mathematics of Data Science* [48], develops a Bayesian framework for persistent homology by modelling persistence diagrams as Poisson point processes (see Sections 2.2.1 and 2.2.2).The defining feature of these point processes is that they are solely characterized by a single parameter known as the intensity. Utilizing the theory of marked point processes, we obtain a method for computing posterior intensities that does not require us to consider explicit maps between input diagrams and underlying parameters, alleviating the computational burden associated with deriving the posterior intensity from Bayes' rule alone. In particular, for a given collection of observed persistence diagrams, we treat random persistence diagrams as Poisson point processes with prior uncertainty captured in presupposed intensities. In applications, one may select an informative prior by choosing an intensity based on expert opinion, or alternatively choose an uninformative prior intensity when information is not available. The likelihood surrogates in our model account for epistemological uncertainty, and create posterior intensities by weighing prior assumptions against evidence. We build our analog of standard Bayesian inference for persistence diagrams using the theory of marked Poisson point processes. A central idea throughout Chapter 3 is to use the topological summaries of data in place of the actual data. Another key contribution in Chapter 3 is the derivation of a *closed form* of the posterior intensity, which relies on conjugate families of Gaussian mixtures. An advantage of this Gaussian mixture representation is that it allows us to perform Bayesian inference in an efficient

and reliable manner. Indeed, this model can be viewed as an analog of the ubiquitous example in standard Bayesian inference where a Gaussian prior and likelihood yield a Gaussian posterior. We present a detailed example of our closed form implementation to demonstrate computational tractability and showcase its applicability by using it to build a Bayes factor classification algorithm; we test the latter in a classification problem for materials science data.as Poisson point processes (see Sections 2.2.1 and 2.2.2). The goal is to view point clouds through their topological descriptors as this can reveal essential shape peculiarities latent in the point clouds.

The later part of Chapter 3 investigates the capabilities of the Bayesian topological model for signal classification by testing its capability to classify autoregressive time series. This method is compared to other model-free approaches for signal classification, specifically those based on feature extraction and machine learning architectures in [29] and [5]. As this portion of Chapter 3 involves topological signal processing, links between persistence diagrams of signals to characteristics like frequency content and instantaneous amplitude are also examined. Ideas in this later portion of Chapter 3 appear in two works I co-authored, one of which was published in the *2019 18th IEEE International Conference On Machine Learning And Applications* [54] and another which has been submitted to *IEEE Transactions on Signal Processing*

Chapter 4 (published in the *Journal of Machine Learning Research* [46]) defines a kernel density estimator that can be used to approximate probability distributions directly on the space persistence diagrams. This chapter compliments Chapter 3 by outlining a tool for persistence diagram density estimation from a nonparametric, frequentist viewpoint. The primary focus of Chapter 4 is on an application of the kernel density estimator to a statistical problem arising in electroencephalography data, which was my major contribution in [46]. In particular, the kernel density estimator is used to construct and perform a novel hypothesis test to detect the presence of certain time series dynamics in noisy measurements.

The chapters are organized as follows. Chapter 2 contains theoretical overviews of persistent homology and point processes, which contain the primary mathematical machinery used to develop new probabilistic methods for persistence diagrams in Chapters 3 and 4. Chapter 3 describes a framework for Bayesian inference with persistence diagrams and moreover 3.4 shows how persistence diagrams created from continuous signals may be interpreted to reveal information about frequency content and instantaneous amplitude. Chapter 4 uses a kernel density estimator for persistence diagrams to define a new statistical test, then applies the new test to detect time series dynamics in noisy autoregressive signals. Finally, Chapter 5 introduces a new topological data augmentation method for training deep neural networks.

Chapter 2

Background

We begin with preliminary discussions of the frequently used mathematical machinery underpinning our ideas. We assume some familiarity with basic notions in topology and abstract algebra, comprehensive overviews of which may be found in [53, 28, 18]. In Section 2.1, we review persistent homology theory and define persistence diagrams for point clouds as well as continuous signals. Persistence diagrams, whose characteristics we are interested in modelling, are the main topological descriptors of data we adopt in the forthcoming chapters. In Chapters 3 and 4, we achieve stochastic models for persistence diagrams by treating them as point processes, thus pertinent definitions and theorems from point processes are discussed in Section 2.2.

2.1 Persistent Homology

Traditionally, topological data analysis views data as noisy samples from random processes defined on underlying geometrical objects living in some ambient space. To compute persistent homology of data, one builds a sequence of topological approximations from a sample in the hope that certain features of the underlying geometrical object may be deduced. Depending on the exact nature of data at hand, there are generally several methods for constructing these sequences. This work employs two popular methods, Vietoris-Rips and sublevel set filtrations. Both of these constructs are generalized by the notion of filtration functions, which are discussed in Section 5.1.1, but we do not make us of this abstract viewpoint until Chapter 5; hence our preliminary discussions frame Vietoris-Rips and sublevel set filtrations in concrete, geometric contexts.

2.1.1 Vietoris-Rips Filtrations

We start by discussing simplices and simplicial complexes, intermediary geometric structures for constructing persistence diagrams from Vietoris-Rips filtrations.

Definition 2.1.1 ([53]). *A d-dimensional collection of data $\{v_0, \dots, v_n\} \subset \mathbb{R}^d \setminus \{0\}$, referred to as a point cloud and whose elements are vertices, is said to be geometrically independent if for any set $t_i \in \mathbb{R}$ with $\sum_{i=0}^n t_i = 0$, the equation $\sum_{i=0}^n t_i v_i = 0$ implies that $t_i = 0$ for all $i \in \{0, \dots, n\}$.*

Definition 2.1.2 ([53]). *A $k-$simplex is a collection of $k+1$ geometrically independent elements along with the convex hull: $[v_0, \dots, v_k] = \left\{ \sum_{i=0}^k \alpha_i v_i : \sum_{i=0}^k \alpha_i = 1 \right\}$. We say that the vertices $v_0, \dots, v_n$ span the $k-$dimensional simplex, $[v_0, \dots, v_k]$. The faces of a $k-$simplex $[v_0, \dots, v_k]$, are the $(k-1)-$simplices spanned by subsets of $\{v_0, \dots, v_k\}$.*

Definition 2.1.3 ([53]). *A simplicial complex $\mathcal{K}$ is a collection of simplices satisfying two conditions: (i) if $\xi \in \mathcal{K}$, then all faces of ξ are also in $\mathcal{K}$, and (ii) the intersection of two simplices in $\mathcal{K}$ is either empty or contained in $\mathcal{K}$. We denote the subset of k-simplices in $\mathcal{K}$ by $\mathcal{K}^k$.*

Examples of simplicial complexes are shown in Figures 2.1 and 2.2. To avoid unnecessary complications when we introduce the boundary map (Definition 2.1.5), we adopt the convention that any collection of vertices $\mathcal{K}^0$ has a fixed ordering, which is inherited by vertices who span elements of $\mathcal{K}^k$. We may freely make this assumption since it can be shown that the persistent homology of a simplicial complex does not depend on any ordering of its vertices [53, 28]. Henceforth, $[v_0, \ldots, v_k] \in \mathcal{K}^k$ denotes a k-simplex with $v_0 < v_1 < \cdots < v_k$.

Definition 2.1.4 ([20]). *The kth chain group $C_k(\mathcal{K})$ of a simplicial complex $\mathcal{K}$ is the collection of all formal sums $\{\sum \alpha_i \sigma_i : \alpha_i \in \mathbb{Z}, \sigma_i \in \mathcal{K}^k\}$*

To avoid cluttering notation, we henceforth omit $\mathcal{K}$ from $C_k(\mathcal{K})$.

Definition 2.1.5 ([20]). *The kth boundary map $\partial_k : \mathcal{K}^k \to \mathcal{K}^{k-1}$ is defined for $k \geq 1$ by:*

$$\partial([v_0, \ldots, v_k]) = \sum_i (-1)^i [v_0, \ldots, \hat{v}_i, \ldots, v_k], \tag{2.1}$$

where $[v_0, \ldots, \hat{v}_i, \ldots, v_k]$ in Equation (2.1) refers to the face of $[v_0, \ldots, v_k]$ with v_i omitted. The map ∂_k extends linearly to a map $\partial_k : C_k \to C_{k-1}$. We take ∂_0 to be the zero map.

Theorem 2.1.1 ([20]). *$\partial_{k-1}\partial_k = 0$ for all $k \geq 1$.*

Definition 2.1.6 ([20]). *The kth-homology group of a simplicial complex, $H_k(\mathcal{K})$, is defined as $H_k(\mathcal{K}) := \ker \partial_k / \mathrm{Im}\, \partial_{k+1}$.*

Theorem 2.1.1 ensures that $\mathrm{Im}\, \partial_{k+1} \subset \ker \partial_k$, so Definition 2.1.6 is well-defined. Moreover, by the Structure Theorem for Finitely Generated Abelian Groups [18], $H_k(\mathcal{K})$ is isomorphic to a direct sum $\mathbb{Z}^n \oplus \mathbb{Z}_{q_1} \oplus \cdots \oplus \mathbb{Z}_{q_T}$ for $q_i \in \mathbb{N}$. We refer to the generators of $H_k(\mathcal{K})$ as *k-dimensional homological features*. Informally, k-dimensional homological features correspond to k-dimensional encasings of $(k+1)-$dimensional empty space. For example, 0-dimensional, 1-dimensional, and 2-dimensional homological features are generally thought of as connected components, loops, and voids, respectively; see Figure 2.1.

Given a point cloud of data, $\mathcal{V}$, to serve as a vertex set, our goal is to build a sequence of simplicial complexes that reasonably approximates an underlying shape characterizing the data. We accomplish this by using the Vietoris-Rips complexes.

Definition 2.1.7. *Let $\mathcal{V} = \{v_i\}_{i=0}^L$ be a point cloud in $\mathbb{R}^{d_\mathcal{V}}$ and $r > 0$. The Vietoris-Rips complex of $\mathcal{V}$ at scale r is defined to be the simplicial complex $\mathcal{R}_r(\mathcal{V})$ satisfying $[v_{i_1}, \ldots, v_{i_l}] \in \mathcal{R}_r(\mathcal{V})$ if and only if $\mathrm{diam}(v_{i_1}, \ldots, v_{i_l}) < 2r$.*

Suppressing $\mathcal{V}$ in Definition 2.1.7, notice as r increases that $\mathcal{R}_r$ only changes at finitely many values of r, which we denote by $\{r_i\}_{i=1}^N$. The increasing family of simplicial complexes $\mathcal{R}_{r_1} \subset \mathcal{R}_{r_2} \subset \ldots \mathcal{R}_{r_N}$ is called a Vietoris-Rips filtration. As this family evolves, homological features appear and disappear in R_{r_i}. A Rips persistence diagram $\mathcal{D}$ is a multiset of points in $\mathcal{W} := \mathbb{W} \times \{0, 1, \ldots, d_\mathcal{V} - 1\}$, where $\mathbb{W} := \{(b, d) \in \mathbb{R}^2 : d \geq b \geq 0\}$ and each element (b, d, k) represents a k-dimensional homological feature appearing then disappearing at scales b and d, respectively, during a Vietoris-Rips filtration. Intuitively speaking, the feature (b, d, k) is a $k-$dimensional hole lasting for duration

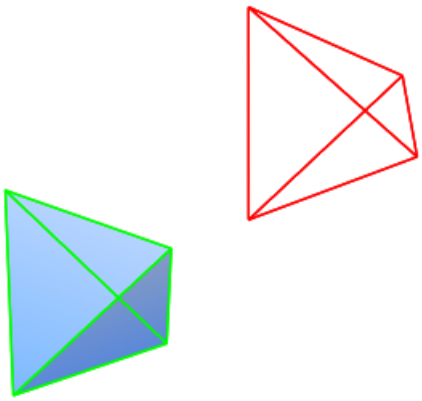


Figure 2.1: Two shapes illustrating different types of homological features. Each object is a connected component and thus a 0-dimensional feature. The red triangles in the shape on the right are 1-dimensional features, while the empty region inside the tetrahedron on the left is a 2-dimensional feature.

$d-b$. An illustration of increasing Vietoris-Rips complexes is shown in Figure 2.2, and an example of a Rips persistence diagram is shown in Figure 2.3.

For a Rips persistence diagram $\mathcal{D}$, the map $T : \mathbb{W} \to T(\mathbb{W})$ given by $T(b,d) = (b, d-b)$ defines a tilted representation of $\mathcal{D}$, $T(\mathcal{D}) = \cup_{(b,d,k)\in\mathcal{D}}(T(b,d),k)$. The coordinate system for $\mathcal{D}$ and $T(\mathcal{D})$ are referred to as birth-death and birth-persistence, respectively; see Figure 2.3(c). From our viewpoint, the range of values over which a homological feature exists in a Vietoris-Rips filtration is a proxy of its topological significance, so the tilted representation preserves all topologically relevant information in a Rips persistence diagram, and we find the tilted coordinate system more convenient to work with in our computational applications. Because of this, we often abuse notation by writing $\mathbb{W}$ and $\mathcal{D}$ for $T(\mathbb{W})$ and $T(\mathcal{D})$, respectively. This is no cause for concern, however, as our results do not depend on either choice of persistence diagram coordinate system. We also fix the homological dimension of features in a Rips persistence diagram by defining $\mathcal{D}^k := \{(b,d) \in \mathbb{W} \ : \ (b,d,k) \in \mathcal{D}\}$.

2.1.2 Sublevel Set Filtrations

In this section, we briefly discuss how to employ sublevel set filtrations to convert signals, which we subsequently assume are continuous functions of time $x : \mathbb{R} \to \mathbb{R}$, to persistence diagrams. For the time being, we refer to these as sublevel set persistence diagrams to distinguish them from Rips persistence diagrams discussed in Section 2.1.1. Sublevel set persistence diagrams usually contain fewer elements than their corresponding signals, so in a sense they compress information about continuous functions into lower dimensional representations.

For a fixed real number V, the sublevel set C_V of a signal x is defined by $C_V := \left\{x^{-1}\Big((-\infty, V]\Big)\right\}$, where $x^{-1}\Big((-\infty, V]\Big) := \{t \in \mathbb{R} : x(t) \leq V\}$. $x^{-1}\Big((-\infty, V]\Big)$ is the set of times where $x(t)$ is less

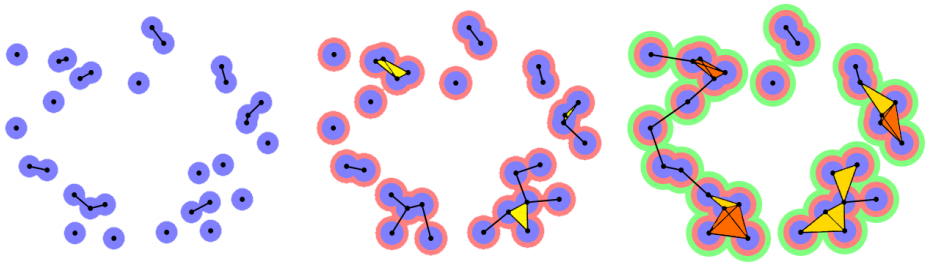
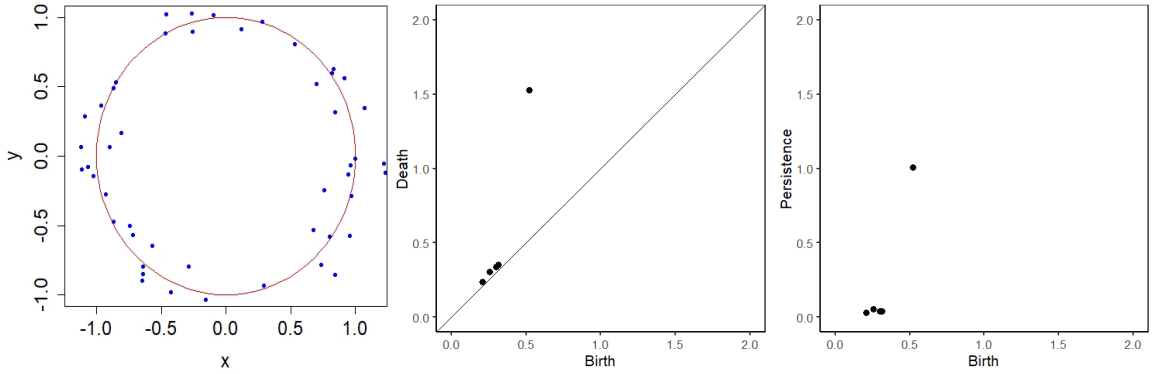


Figure 2.2: Vietoris Rips complexes.



(a) (b) (c)

Figure 2.3: (a) An example of a dataset, (b) its persistence diagram, and (c) its tilted representation.

than or equal to V. We refer to the collection $C := \{C_V\}_{V \in \mathbb{R}}$ as the sublevel set filtration of x. From the sublevel filtration C, we obtain a persistence diagram for x by tracking the birth and death values of connected components (which are in this case disjoint contiguous intervals) in C_V as V increases. Specifically, as V increases, connected components appear or merge in C_V. To create a persistence diagram, we record the values, b and d, respectively, at which a connected component appears and disappears by merging into one that appeared earlier (a convention known as the Elder rule [20]). For most continuous signals encountered in applications, it can be shown b is always a local minimum and d is always a local maximum [20]. In the end, we aggregate each connected component's birth-death pair (b, d) to create a collection known as a sublevel set persistence diagram. We sometimes refer to the elements (b, d) as homological features. For a visual summary of persistence diagram generation through sublevel set filtrations, see Figure 2.4, which shows a damped cosine $s(t; f, \beta) := e^{-\beta t} \cos(2\pi f t)$ with $f = 4$ and $\beta = 2$. Notice in general

that $s(t; f, \beta)$ has f local minima (ignoring units of f) on the interval $[0, 1]$. As the number of points sublevel in a persistence diagram of a signal is equal to the number of its local minima, we immediately conclude that the diagram's cardinality is f. To compute the coordinates of points in the sublevel set persistence diagram, notice that $e^{-\beta t}$ is monotonically decreasing while each oscillation of $\cos(2\pi f t)$ starts at a local maximum and has a unique local minimum. The latter implies that each oscillation gives rise to a connected component during the sublevel set filtration of s and that this connected component merges into another when the filtration reaches the value of the local maximum where the oscillation begins. The former implies that connected components for oscillations occurring later in time disappear sooner than those born earlier due to the Elder Rule.

Henceforth, we use the term persistence diagram to refer interchangeably to Rips and sublevel set persistence diagrams, only making distinctions between the two when it is not obvious from context which type we are working with. We also denote both types of persistence diagram by $\mathcal{D}$. Our methods in Chapters 3 and 4 all abstract away from the spaces in which data live, working solely with persistence diagrams, so they work in the same fashion with persistence diagrams created with any filtration. As we will see, the choice to work in persistence diagram space is advantageous when stochastic models in data space are difficult to specify.

2.2 Point Processes

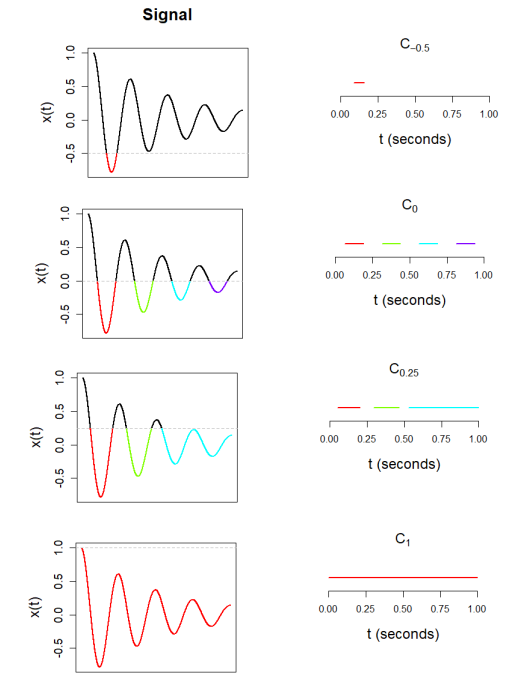
This section contains an overview of the theory of point processes, primarily Poisson point processes. Detailed treatments of Poisson point processes can be found in [36, 16] and references therein. To build intuition, we introduce point processes in Section 2.2.1 without relying on sophisticated mathematical notions. Section 2.2.2 builds on this intuition by defining point processes in a more mathematically useful fashion in addition to introducing theorems and tools for rigorously working with point processes. Section 2.2.3 presents ideas from a computationally-focused formulation of point process theory known as finite set statistics, which are vital to the kernel density estimator presented in Chapter 4.

2.2.1 Intuition

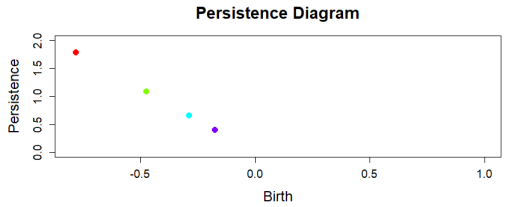
Intuitively, a random persistence diagram is a random collection of points in a subset of the plane. Elements of a persistence diagram lack an intrinsic ordering, and persistence diagrams generated by independent samples of data from the same distribution can have different cardinalities. The theory of random variables is ill-equipped to handle such objects directly since it concerns itself with random elements that take values in a Hilbert space. However, the theory of point processes rigorously treats random collections like random persistence diagrams and provides machinery to model them. A point process is a set-valued random element characterized by a random variable K that prescribes a cardinality distribution and a set of spatial distributions conditioned on K that dictate where to place points in samples once cardinality is known. Formally,

Definition 2.2.1. *For a discrete random variable K, a (finite) point process $P := \{x_1, \ldots, x_K\}$ is a random collection of elements in a Polish Space $\mathbb{X}$ characterized by a collection of symmetric conditional probability distributions $P|K = k \sim \mathbb{P}_k$.*

In general, an exact description for K and the spatial distributions of a random persistence diagram is computationally intractable, even in scenarios where a reliable stochastic model for data is obtainable. Thus, studying random persistence diagrams requires a well-grounded framework for



(a)



(b)

Figure 2.4: Show above (a) are the sublevel sets $C_{-0.5}$, C_0, $C_{0.25}$, and C_1 for a damped cosine $e^{-2t}\cos(8\pi t)$. (b) shows the persistence diagram of the sublevel set filtration. The points in (b) are colored to match the connected components their birth coordinates correspond to. The transition from C_0 to $C_{0.25}$ depicts the Elder rule; notice that in C_0, there are light blue and purple connected components, which merge together in $C_{0.25}$. A similar merging happens in the transition from $C_{0.25}$ to $C_{0.5}$. Since the purple component has a later birth value, it disappears into the light blue component, which persists until it merges into the green component by the same line of reasoning.

estimating their distributions. We describe two such frameworks from a Bayesian and nonparametric viewpoint in Chapters 3 and 4, respectively; central to former are Poisson point processes, which are defined below.

Definition 2.2.2. *Let* $\lambda : \mathbb{R}^2 \to \mathbb{R}$ *be a non-negative function satisfying* $0 < \Lambda < \infty$ *where* $\Lambda := \int_{\mathbb{R}^2} \lambda(\mathbf{x})\, d\mathbf{x}$. *The function* λ *is called the intensity function or simply the intensity. A Poisson*

point process, $\mathcal{P}$ is a point process satisfying:

$$K \sim \text{Poisson}(\Lambda) \tag{2.2}$$

$$\mathcal{P}|K = k \sim \frac{1}{\Lambda^n} \prod_{i=1}^{n} \lambda(\mathbf{x}_i). \tag{2.3}$$

Definition 2.2.2 prescribes a natural recipe to sample from a Poisson point process. First, one determines the cardinality of $\mathcal{P}$ by drawing from Poisson(Λ). With this number in hand, say n, one then makes n independent draws from the probability density $\frac{\lambda}{\Lambda}$ to spatially distribute the points in $\mathbb{R}^2$. Poisson point processes are completely characterized by their intensities. This makes them a convenient tool for Bayesian inference on the space of persistence diagrams since they can be used to specify prior distributions using a single, real-valued function on $\mathbb{R}^2$. With prior distributions in hand, the last ingredient one needs for Bayesian inference is a likelihood, which is obtained via marked point processes (defined below).

Definition 2.2.3. *Let $\ell(\mathbf{y}|\mathbf{x})$ be a probability density parameterized by $\mathbf{x}$. A marked point processes is a collection (P, M) where*

$$M|P = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\} \sim \prod_{i=1}^{k} \ell(\mathbf{y}_i|\mathbf{x}_i). \tag{2.4}$$

Elements of M in Definition 2.2.3 are determined by independent draws from a stochastic kernel $\ell(\cdot|\mathbf{p})$, which is a probability density parameterized by the elements of samples from P.

2.2.2 A Formal Viewpoint

We now build on the intuitions from Section 2.2.1. For the remainder of this section, we take $\mathbb{X}$ and $\mathcal{X}$ to be a Polish space and its Borel σ-algebra, respectively.

Definition 2.2.4. *A finite point process P is a pair $(\{p_n\}, \{\mathbb{P}_n\})$ where $\sum_{n=0}^{\infty} p_n = 1$ and $\mathbb{P}_n$ is a symmetric probability measure on $\mathcal{X}^n$, where $\mathcal{X}^0$ is understood to be the trivial σ-algebra.*

The sequence $\{p_n\}$ defines a cardinality distribution and the measures $\{\mathbb{P}_n\}$ give spatial distributions of vectors $(x_1, \ldots, x_n)$ for fixed n (in other words, $\mathbb{P}_n$ are probability distributions conditioned on cardinality). From a statistical point of view, Definition 2.2.4 naturally prescribes a method for sampling a finite point process: (i) determine the number of points n by drawing from $\{p_n\}$ then, (ii) spatially distribute $(x_1, \ldots, x_n)$ according to a draw from $\mathbb{P}_n$. Point processes are meant to give a recipe for drawing random collections of elements $\{x_1, \ldots, x_n\} \subset \mathbb{X}$ whose order is irrelevant, thus any sensible definition of P that leads to sampling vector-valued elements $(x_1, \ldots, x_n)$ should assign equal weight to all permutations. This is ensured by the symmetry requirement in Definition 2.2.4. We may suggestively write

$$\{x_1, \ldots, x_n\} := \mathcal{S}_n(x_1, \ldots, x_n), \tag{2.5}$$

where the right hand side of Equation (2.5) denotes the orbit of $(x_1, \ldots, x_n)$ under the permutation group $\mathcal{S}_n$ (once again see [18]), then ponder if there is a way to reformulate Definition 2.2.4 to create a set of probability distributions on collections of $\{x_1, \ldots, x_n\}$. The answer is in the affirmative, as the next definition shows.

Definition 2.2.5. *Let $(\{p_n\}, \{\mathbb{P}_n\})$ be a finite point process. The Janossy measures $\{\mathbb{J}_n\}$ are defined as the set of measures satisfying $\mathbb{J}_n(A) = n! p_n \mathbb{P}_n(A)$, for all $n \in \mathbb{N}$ and $A \in \mathcal{X}^n$.*

Given a collection of disjoint rectangles $A_1, \ldots, A_n \subset \mathbb{X}$, the value $\mathbb{J}_n(A_1 \times \cdots \times A_n)$ is the probability of observing exactly one element in each of $A_1, \ldots, A_n$ and none in the complement of their union if we restrict $\mathbb{X}^n$ to the subspace $\mathbb{X}_S^n := \{(x_1, x_2, \ldots, x_n) \in \mathbb{X}^n : x_1 < x_2 < \cdots < x_n\}$ [16]. As the elements in $\mathbb{X}_S^n$ are all representative elements of disjoint orbits $\mathcal{S}_n(x_1, \ldots, x_n)$, samples from $\mathbb{J}_n$ in this restricted space may be understood as set-valued elements in the sense of Equation (2.5). Embracing this viewpoint, we subsequently consider samples of P, say $\mathcal{P}$, to be set-valued.

For applications, we are primarily interested in Janossy measures $\mathbb{J}_n$ that admit densities j_n with respect to a reference measure on $\mathbb{X}$. We are now ready to describe the class of finite point processes that model persistence diagrams in our Bayesian formulation.

Definition 2.2.6. *Let Λ be a finite measure on $\mathbb{X}$ and define $\mu := \Lambda(\mathbb{X})$. The finite point process Π is Poisson if, for all $n \in \mathbb{N}$ and disjoint measurable rectangles $A_1 \times \cdots \times A_n \in \mathcal{X}^n$, $p_n = e^{-\mu}\frac{\mu^n}{n!}$, and $\mathbb{P}_n(A_1 \times \cdots \times A_n) = \prod_{i=1}^n \left(\frac{\Lambda(A_i)}{\mu}\right)$. We call Λ an intensity measure.*

Equivalently, a Poisson point process is a finite point process with Janossy measures $\mathbb{J}_n(A_1 \times \cdots \times A_n) = e^{-\mu}\prod_{i=1}^n \Lambda(A_i)$. We assume the intensity measure in Definition 2.2.6 admits a density, λ, with respect to some reference measure on $\mathbb{X}$. Notice that for all $A \in \mathcal{X}$, $\mathbb{E}(|\Pi \cap A|) = \sum_{n=0}^{\infty} p_n \mathbb{E}_{\mathbb{P}_n}\left(\sum_{k=0}^n k\binom{n}{k}\mathbb{1}_{A^k \times (A^c)^{n-k}}\right)$. Elementary calculations then show $\mathbb{E}(|\Pi \cap A|) = \Lambda(A)$. Thus, we interpret the intensity measure of a region A, $\Lambda(A)$ as the expected number of elements in Π that land in A. The intensity measure serves as an analog to the first order moment for a random variable.

The next two definitions involve a joint point process wherein points from one space parameterize distributions for the points living in another. Consequently, we introduce another Polish space $\mathbb{M}$ along with its Borel σ-algebra $\mathcal{M}$ to serve as the mark space in a marked Poisson point process.

Definition 2.2.7. *Suppose $\ell : \mathbb{X} \times \mathbb{M} \to \mathbb{R}^+ \cup \{0\}$ is a function satisfying: 1) for all $x \in \mathbb{X}$, $\ell(x, \bullet)$ is a probability measure on $\mathbb{M}$, and 2) for all $B \in \mathcal{M}$, $\ell(\bullet, B)$ is a measurable function on $\mathbb{X}$. Then, ℓ is a stochastic kernel from $\mathbb{X}$ to $\mathbb{M}$.*

Definition 2.2.8. *A marked Poisson point process Π_M is a finite point process on $\mathbb{X} \times \mathbb{M}$ such that: (i) $(\{p_n\}, \{\mathbb{P}_n(\bullet \times \mathbb{M})\})$ is a Poisson PP on $\mathbb{X}$, and (ii) for all $(x_1, \ldots, x_n) \in \mathbb{X}^n$, measurable rectangles $B_1 \times \cdots \times B_n \in \mathcal{M}^n$, $\mathbb{P}_n((x_1, \ldots, x_n) \times B_1 \times \cdots \times B_n) = \frac{1}{n!}\sum_{\pi \in \mathcal{S}_n}\prod_{i=1}^n \ell(x_{\pi(i)}, B_i)$, where $\mathcal{S}_n$ is the set of all permutations of $(1, \ldots, n)$ and ℓ is a stochastic kernel.*

Given a set of observed marks $M = \{y_1, \ldots, y_m\}$, after adopting Definition 2.2.5 , it can be shown [68] that the Janossy densities for the point process induced by Π_M on $\mathbb{X}$ given M are

$$j_{n|M}(x_1, \ldots, x_n) = \begin{cases} \mathbb{1}_{\emptyset}, & n = m = 0 \\ \sum_{\pi \in \mathcal{S}_n}\prod_{i=1}^n p(x_i|y_{\pi(i)}), & n = m > 0, \\ 0, & \text{otherwise}, \end{cases} \tag{2.6}$$

where p is the stochastic kernel for Π_M evaluated in $\mathbb{X}$ for a fixed value of $y \in \mathbb{M}$.
The following theorems allow us to construct new Poisson point processes from existing ones. Their proofs can be found in [36].

Theorem 2.2.1 (The Superposition Theorem, [36])**.** *Let $\{\Pi_n\}_{n\in\mathbb{N}}$ be a collection of independent Poisson point processes each having intensity measure Λ_n. Then their superposition Π given by $\Pi := \bigcup_{n\in\mathbb{N}} \Pi_n$ is a Poisson point process with intensity measure $\Lambda = \sum_{n\in\mathbb{N}} \Lambda_n$.*

Theorem 2.2.2 (The Mapping Theorem, [36]). *Let Π be a Poisson point process on $\mathbb{X}$ with σ-finite intensity measure Λ and let $(\mathbb{T}, \mathcal{T})$ be a measurable space. Suppose $f : \mathbb{X} \to \mathbb{T}$ is a measurable function. Write Λ^* for the induced measure on T given by $\Lambda^*(B) := \Lambda(f^{-1}(B))$ for all $B \in \mathcal{T}$. If Λ^* has no atoms, then $f \circ \Pi$ is a Poisson point process on T with intensity measure Λ^*.*

Theorem 2.2.3 (The Marking Theorem, [36]). *The marked Poisson point process in Definition 2.2.8 has the intensity measure given by $\Lambda_M(C) = \iint_C \Lambda(dx)\ell(x, dm)$ for all $C \in \mathcal{X} \times \mathcal{M}$, where Λ is the intensity measure for the Poisson point process that Π_M induces on $\mathbb{X}$, and ℓ is a stochastic kernel.*

The final tool we need is the probability generating functional as it enables us to recover intensity measures using a notion of differentiation. The probability generating functional can be interpreted as the point process analog of the probability generating function.

Definition 2.2.9 ([52]). *Let $\mathcal{P}$ be a finite point process on a Polish space $\mathbb{X}$. Denote by $\mathcal{B}(\mathbb{C})$ the set of all functions $h : \mathbb{X} \to \mathbb{C}$ with $||h||_\infty < 1$. The probability generating functional of P denoted $G : \mathcal{B}(\mathbb{C}) \to \mathbb{R}$ is given by*

$$G(h) = J_0 + \sum_{n=1}^{\infty} \frac{1}{n!} \int_{\mathbb{X}^n} \left(\prod_{j=1}^{n} h(x_j) \right) \mathbb{J}_n(dx_1 \dots dx_n) \tag{2.7}$$

Definition 2.2.10 ([52]). *Let G be the probability generating functional given in Equation (2.7). The functional derivative of G in the direction of $\eta \in \mathcal{B}(C)$ evaluated at h, when it exists, is given by $G'(h; \eta) = \lim_{\epsilon \to 0} \frac{G(h+\epsilon\eta) - G(h)}{\epsilon}$.*

It can be shown that the functional derivative satisfies the familiar product rule [40], namely for $G = \prod_{i=1}^{m} G_i$,

$$G'(h; \gamma) = \sum_{i=1}^{m} G_i'(h; \gamma) \prod_{j \neq 1} G_j(h). \tag{2.8}$$

As is proved in [52], the intensity measure Λ of the Poisson PP in Definition 2.2.6 can be obtained by differentiating G, i.e.,

$$\Lambda(A) = G'(1; \mathbb{1}_A), \tag{2.9}$$

where $\mathbb{1}_A$ is the indicator function for any $A \in \mathcal{X}$. Generally speaking, one obtains the intensity measure for a general point process through $\Lambda(A) = \lim_{h \to 1} G'(h; \mathbb{1}_A)$, but the preceding identity suffices for our purposes since we only consider point processes for which Equation (2.7) is defined for all bounded h.

Corollary 2.2.1. *The intensity function for the point process whose Janossy densities are listed in Equation (2.6) is $\sum_{i=1}^{m} p(x|y_i)$ if $m > 0$ and 0 otherwise.*

Proof. We substitute the Janossy measures $\mathbb{J}_n$ for the point process (Equation (2.6)) into Equation (2.7). This yields $G(h) = \int_{\mathbb{X}^m} \frac{1}{m!} \sum_{\pi \in \mathcal{S}_n} \prod_{i=1}^{n} p(x_i|y_{\pi(i)})\, dx_i$ if $m > 0$, and $G(h) = \mathbb{1}_\emptyset$ otherwise. It is easy to see that the functional derivative (Definition 2.2.10) of a constant is zero, so our claim follows immediately in the latter case by appealing to Equation (2.9). In the former case, we have

$$
\begin{aligned}
G(h) &= \int_{\mathbb{X}^m} \frac{1}{m!} \sum_{\pi \in \mathcal{S}_m} \prod_{i=1}^{n} h(x_i) p(x_i|y_{\pi(i)})\, dx_i \\
&= \frac{1}{m!} \sum_{\pi \in \mathcal{S}_m} \int_{\mathbb{X}^m} \prod_{i=1}^{m} h(x_i) p(x_i|y_{\pi(i)})\, dx_i \quad \text{(by linearity of the integral)} \\
&= \frac{1}{m!} \sum_{\pi \in \mathcal{S}_m} \left(\prod_{i=1}^{m} \int_{\mathbb{X}} h(x) p(x|y_{\pi(i)})\, dx \right) \quad \text{(by Fubini's Theorem)} \\
&= \prod_{i=1}^{m} \int_{\mathbb{X}} h(x) p(x|y_i)\, dx. \quad \text{(by symmetry)}
\end{aligned}
$$

Write $G_i(h)$ for the probability generating functional of the point process with $j_1 = p(x|y_i)$ and $j_n = 0$ for $n \neq 1$. We can summarize the preceding string of equalities by $G(h) = \prod_{i=1}^{m} G_i(h)$. We apply the product rule for functional derivatives (Equation (2.8)), then substitute $h = 1$ and $\gamma = \mathbb{1}_{x'}$ (the indicator of a singleton) to obtain

$$
G'(1; \mathbb{1}_{x'}) = \sum_{i=1}^{m} p(x'|y_i) \prod_{i \neq j} \int_{\mathbb{X}} p(x|y_j)\, dx = \sum_{i=1}^{m} p(x|y_i)
$$

Appealing to Equation (2.9) establishes the claim. □

2.2.3 Finite Set Statistics

In Chapter 4, we consider the notion of probability density functions for random persistence diagrams that, unlike those in the formulation of Chapter 3, are described by arbitrary point processes. In this section, we therefore establish background to make the notion of probability density for a random persistence diagram explicit and well-defined. Although this endeavor may be achieved by appealing to traditional point process theory, we elect to use machinery from finite set statistics [40], which refashions point process theory into a computationally friendly framework. This choice was essential to derive results in Chapter 4, because while marked Poisson point processes (Section 2.2.2) are natural candidates for Bayesian models and amenable to techniques of calculus, arbitrary point processes are unwieldy from a computational standpoint.

As before, we consider a random persistence diagram D as a random multiset of features $D = \{\xi_i\} \subset \mathbb{W} \times \{0, \ldots, k-1\}$. For underlying datasets sampled from $\mathbb{R}^n$ with bounded cardinality, the affiliated persistence diagrams also have bounded feature cardinality and degree of homology. Thus, we assume that the cardinality of a random persistence diagram is bounded above by some value $|D| \leq M \in \mathbb{N}$, and so consider the space $\mathcal{C}_{\leq M}(\mathbb{W}) = \{D \text{ is a multiset in } \mathbb{W} : |D| \leq M\}$.

We view $\mathcal{C}_{\leq M}(\mathbb{W})$ through a list of functions h_N which each map the appropriate dimension of Euclidean space into its corresponding cardinality component, $\mathcal{C}_N(\mathbb{W})$. This viewpoint facilitates the definition of probability densities.

Definition 2.2.11 ([46])**.** *For each $N \in \{0, ..., M\}$, consider the space of N topological features, denoted $\mathcal{C}_N(\mathbb{W}) = \{D$ multiset in $\mathbb{W} : |D| = N\}$, and the associated map $h_N : \mathbb{W}^N \to \mathcal{C}_N(\mathbb{W})$ defined by*

$$
h_N(\xi_1, ..., \xi_N) = \{\xi_1, ..., \xi_N\}. \tag{2.10}
$$

The map h_N creates equivalence classes on $\mathbb{W}^N$ according to the action of the permutations Π_N; specifically, $[Z] = [(\xi_1, ..., \xi_N)]_{h_N} = \left\{\left(\xi_{\pi(1)}, ..., \xi_{\pi(N)}\right) : \pi \in \Pi_N\right\}$ for each $Z = (\xi_1, ..., \xi_N) \in \mathbb{W}^N$.

These equivalence classes yield the space

$$\mathbb{W}^N/\Pi_N = \left\{ [\boldsymbol{\xi}]_{h_N} : \boldsymbol{\xi} \in \mathbb{W}^N \right\}, \tag{2.11}$$

equipped with the quotient topology. The topology on $\mathcal{C}_{\leq M}(\mathbb{W})$ is defined so that each h_N lifts to a homeomorphism between $\mathbb{W}^N/\Pi_N$ and $\mathcal{C}_N(\mathbb{W})$, and we write $\mathbb{W}^N/\Pi_N \cong \mathcal{C}_N(\mathbb{W})$.

With a topology in hand, one can define probability measures on the associated Borel σ-algebra. Thus, we define a random persistence diagram D to be a random element distributed according to some probability measure on $\mathcal{C}_{\leq M}(\mathbb{W})$ for a fixed maximal cardinality $M \in \mathbb{N}$. We denote associated probabilities by $\mathbb{P}[\cdot]$ and expected values by $\mathbb{E}[\cdot]$. Since $\mathbb{W}^N/\Pi_N \cong \mathcal{C}_N(\mathbb{W})$, we work toward defining probability densities on the collection of Euclidean spaces $\cup_{N=0}^M \mathbb{W}^N$.

Definition 2.2.12 ([46])**.** *For a given random persistence diagram D and any Borel subset A of $\mathbb{W}$, the belief function β_D is defined as*

$$\beta_D(A) = \mathbb{P}[D \subset A]. \tag{2.12}$$

Since A is a Borel subset of $\mathbb{W}$, the collection $O_A = \{D \in \mathcal{C}_{\leq M}(\mathbb{W}) : D \subset A\}$ is the quotient of $\cup_{N=0}^M A^N \subset \cup_{N=0}^M \mathbb{W}^N$ under h_N; moreover, A^N is clearly Borel in the Euclidean topology of $\cup_{N=0}^M \mathbb{W}^N$. Therefore, since h_N induces a homeomorphism (see Defn 2.2.11), O_A is a Borel subset of $\mathcal{C}_{\leq M}(\mathbb{W})$. The belief function of a random persistence diagram is similar to the joint cumulative distribution function for a random vector, in particular by yielding a probability density function through Radon-Nikodým type derivatives.

Definition 2.2.13 ([46])**.** *Fix ϕ defined on Borel subsets of $\mathcal{C}_{\leq M}(\mathbb{W})$ into $\mathbb{R}$. For an element $\xi \in \mathbb{W}$ or a multiset $Z \subset \mathbb{W}$ with $Z = \{\xi_1, ..., \xi_N\}$, the set derivative (evaluated at the empty set $\emptyset$) is respectively given by*

$$\begin{aligned} \frac{\delta \phi}{\delta \xi}(\emptyset) &= \lim_{n \to \infty} \frac{\phi(B(\xi, 1/n))}{\lambda(B(\xi, 1/n))}, \\ \frac{\delta \phi}{\delta Z}(\emptyset) &= \frac{\delta^N \phi}{\delta \xi_1 ... \delta \xi_N} = \left[\frac{\delta}{\delta \xi_1} \cdots \frac{\delta}{\delta \xi_N} \phi \right](\emptyset), \end{aligned} \tag{2.13}$$

where $B(\xi, 1/n)$ are Euclidean balls and λ indicates Lebesgue measure on $\mathbb{W}$.

As with typical derivatives, there is a complementary set integration operation for set derivatives. Set derivatives (at $\emptyset$) are essentially Radon-Nikodým derivatives with order tied to cardinality, and so the corresponding set integral acts like Lebesgue integration summed over each cardinality.

Definition 2.2.14 ([46])**.** *Consider a Borel subset A of $\mathbb{W}$ and a Borel subset O of $\mathcal{C}_{\leq M}(\mathbb{W})$. For a set function $f : \mathcal{C}_{\leq M}(\mathbb{W}) \to \mathbb{R}$, its set integrals over A and O are respectively defined according to the following sums of Lebesgue integrals:*

$$\int_A f(Z) \delta Z = \sum_{N=0}^M \frac{1}{N!} \int_{A^N} f(h_N(\xi_1, ..., \xi_N)) d\xi_1 ... d\xi_N, \tag{2.14a}$$

$$\int_O f(Z) \delta Z = \sum_{N=0}^M \frac{1}{N!} \int_{h_N^{-1}(O)} f(h_N(\xi_1, ..., \xi_N)) d\xi_1 ... d\xi_N, \tag{2.14b}$$

where $Z = \{\xi_1, ..., \xi_N\} \subset \mathbb{W}$ is a persistence diagram.

Dividing by $N!$ in Equations (2.14a) and (2.14b) accounts for integrating over $\mathbb{W}^N$ instead of $\mathbb{W}^N/\Pi_N \cong \mathcal{C}_N(\mathbb{W})$. It has been shown that set derivatives and integrals are inverse operations [40]; specifically, the set derivative of a belief function yields a probability density for a random diagram D such that

$$\beta_D(A) = \int_A \frac{\delta \beta_D}{\delta Z}(\emptyset)\delta Z. \tag{2.15}$$

Indeed, $A^N = h_N^{-1}(\{D \subset A\})$ so that Eq. (2.14a) also holds as an integral over $O_A = \{D \in \mathcal{C}_{\leq M} : D \subset A\}$ in the sense of Eq. (2.14b).

Definition 2.2.15 ([46]). *For a random persistence diagram D, a global probability density function (global pdf) $f_D : \cup_{N\in\mathbb{N}} \mathbb{W}^N \to \mathbb{R}$ must satisfy*

$$\sum_{\pi\in\Pi_N} f_D(\xi_{\pi(1)}, ..., \xi_{\pi(N)}) = \frac{\delta^N \beta_D}{\delta\xi_1 \cdot ... \cdot \delta\xi_N}(\emptyset). \tag{2.16}$$

and is described by its layered restrictions $f_N = f_D\big|_{\mathbb{W}^N} : \mathbb{W}^N \to \mathbb{R}$ for each N.

The following proposition is critical to determine the global pdf for (i) the union of independent singleton diagrams (i.e., $\left|D^j\right| \leq 1$), (ii) a randomly chosen cardinality, N, followed by N i.i.d. draws from a fixed distribution, and (iii) a random persistence diagram kernel density function. The proof of this proposition follows similar arguments to [40] (Theorem 17, pp. 155–156).

Proposition 2.2.1 ([46]). *Let D be a random persistence diagram with cardinality bounded by M and let $\beta_D(S) = \mathbb{P}(D \subset S)$ be the belief function for D. Then β_D expands as $\beta_D(S) = a_0 + \sum_{m=1}^M a_m q_m(S)$, where $a_m = \mathbb{P}(|D| = m)$ and $q_m(S) = \mathbb{P}[D \subset S \big| \, |D| = m]$.*

Lastly, we encounter a computationally convenient summary for a random persistence diagram called the probability hypothesis density (PHD). The integral of the PHD over a subset U in $\mathbb{W}$ gives the expected number of points in the region U; moreover, any other function on $\mathbb{W}$ with this property is a.e. equal to the PHD [40].

Definition 2.2.16 ([46]). *The probability hypothesis density (PHD) for a random persistence diagram D is defined as the set function $F_D(a) = \frac{\delta\beta_D}{\delta Z}(\{a\})$ and is expressed as a set integral as*

$$F_D(a) = \int_{\{Z:\{a\}\subset Z\}} \frac{\delta\beta}{\delta Z}(\emptyset)\delta Z. \tag{2.17}$$

In particular, $\mathbb{E}(|D \cap U|) = \int_U F_D(u)\, du$ for any region U.

Definition 2.2.16 is equivalent to an intensity function of a point process. In general, the intensity function induced by a given global pdf may be undefined, but under mild conditions Equation (2.17) is finite. Since D is a random persistence diagram, the PHD is always defined as a distribution and can always be integrated to obtain the identity $\mathbb{E}(|D \cap U|) = \int_U F_D(u)\, du$ for any region U.

Proposition 2.2.1 leads to the following lemma which is crucial for determining the kernel density. We refer to a random persistence diagram D with $|D| \leq 1$ as a singleton diagram, and such singletons are indexed by superscripts.

Lemma 2.2.1 ([46]). *Consider a multiset of independent singleton random persistence diagrams $\left\{D^j\right\}_{j=1}^M$. If each singleton D^j is described by the value $q^{(j)} = \mathbb{P}[D^j \neq \emptyset]$ and the subsequent conditional pdf, $p^{(j)}(\xi)$, given $\left|D^j\right| = 1$, then the global pdf for $D = \cup_{j=1}^M D^j$ is given by*

$$f_D(\xi_1, ..., \xi_N) = \sum_{\gamma\in I(N,M)} \mathcal{Q}(\gamma) \prod_{k=1}^N p^{(\gamma(k))}(\xi_k), \tag{2.18}$$

for each $N \in \{0, ..., M\}$ *where*

$$\mathcal{Q}(\gamma) = \mathcal{Q}^*(\gamma) \prod_{k=1}^{N} q^{(\gamma(k))}, \tag{2.19}$$

$I(N, M)$ *consists of all (strictly) increasing injections* $\gamma : \{1, ..., N\} \to \{1, ..., M\}$*, which enumerate (unordered) correspondences between the input features* $(\xi_1, \ldots, \xi_N)$ *and a subset of the* M *random singletons, and*

$$\mathcal{Q}^*(\gamma) = \frac{\prod_{j=1}^{M}(1 - q^{(j)})}{\prod_{k=1}^{N}(1 - q^{(\gamma(k))})}. \tag{2.20}$$

Proof. See [46]. □

Chapter 3

A Bayesian Framework for Persistent Homology

The first Bayesian considerations in a persistent homology context take place in [49] when the authors discuss a conditional probability setting on persistence diagrams where the likelihood for the observed point cloud has been substituted by that for its associated topological summary. In this chapter, we describe a full Bayesian treatment for persistent homology predicated upon creating posterior distributions of persistence diagrams.

The homological features in persistence diagrams have no intrinsic order implying they are random sets as opposed to random vectors. To this end, we model random persistence diagrams as Poisson point processes (see Sections 2.2.1 and 2.2.2). The defining feature of these point processes is that they are solely characterized by a single parameter known as the intensity. Utilizing the theory of marked point processes, we obtain a method for computing posterior intensities that does not require us to consider explicit maps between input diagrams and underlying parameters, alleviating the computational burden associated with deriving the posterior intensity from Bayes' rule alone.

In particular, for a given collection of observed persistence diagrams, we treat random persistence diagrams as Poisson point processes with prior uncertainty captured in presupposed intensities. In applications, one may select an informative prior by choosing an intensity based on expert opinion, or alternatively choose an uninformative prior intensity when information is not available. The likelihood surrogates in our model account for epistemological uncertainty, and create posterior intensities by weighing prior assumptions against evidence. We build our analog of standard Bayesian inference for persistence diagrams using the theory of marked Poisson point processes. A central idea throughout Chapter 3 is to use the topological summaries of data in place of the actual data.

Another key contribution in Chapter 3 is the derivation of a *closed form* of the posterior intensity, which relies on conjugate families of Gaussian mixtures. An advantage of this Gaussian mixture representation is that it allows us to perform Bayesian inference in an efficient and reliable manner. Indeed, this model can be viewed as an analog of the ubiquitous example in standard Bayesian inference where a Gaussian prior and likelihood yield a Gaussian posterior. We present a detailed example of our closed form implementation to demonstrate computational tractability and showcase its applicability by using it to build a Bayes factor classification algorithm; we test the latter in a classification problem for materials science data. Specifically, in Section 3.3 we use our classification algorithm to distinguish crystal structure, which allows scientists to predict the properties of a crystalline material. Our goal is to view point clouds through their topological descriptors as this

can reveal essential shape peculiarities latent in the point clouds.

To summarize, the main ideas in Chapter 3 are:

1. Theorem 3.1.1, which provides the generalized Bayesian framework for computing the posterior distribution of persistence diagrams.

2. Proposition 3.1.1, which yields a conjugate family of priors based on Gaussian mixtures for the proposed Bayesian framework.

3. A classification scheme using Bayes factors considering the posteriors of persistence diagrams and its application to a materials science problem.

4. A demonstration that our Bayesian model is competitive with well-established nonparametric methods of signal processing at the task of signal classification.

Chapter 3 is organized as follows. Our methods are presented in Section 3.1. In particular, Section 3.1.1 establishes the Bayesian framework for persistence diagrams, while Section 3.1.2 contains the derivation of a closed form for a posterior distribution based on a Gaussian mixture model. A classification algorithm with Bayes factors is discussed in Section 3.3. To assess the capability of our algorithm, we investigate its performance on materials data in Section 3.3.1. In Section 3.4, we explore our Bayesian model in the context of signal processing.

3.1 Bayesian Inference

In this section, we construct a framework for Bayesian inference with persistence diagrams by modelling them as Poisson point processes. First, we derive a closed form for the posterior intensity given a persistence diagram drawn from a finite point process, and then we present a family of conjugate priors.

3.1.1 Model

According to Bayes' theorem, posterior density is proportional to the product of a likelihood function and a prior. Thus, to adapt the Bayesian framework to persistence diagrams, we must utilize notions analogous to prior distributions and likelihood functions for point processes. In particular, our Bayesian framework views a random persistence diagram as a Poisson point process equipped with a prior intensity while observed persistence diagrams $\mathcal{D}_Y$ are considered to be marks from a marked Poisson point process. This enables modification of the prior intensity by incorporating observed persistence diagrams, yielding a posterior intensity based on data. Some parallels between our Bayesian framework and that for random variables are illustrated in Table 3.1.

Let $(D_X^k, D_Y^k) \in \mathbb{W} \times \mathbb{W}$ be a finite point process and that satisfies the following assumptions:

(M1) For $k_1 \neq k_2$, $(D_X^{k_1}, D_Y^{k_1})$ and $(D_X^{k_2}, D_Y^{k_2})$ are independent.

(M2) For k fixed, $D_X^k = D_{X_O}^k \cup D_{X_V}^k$ and some $\alpha : \mathbb{W} \to [0,1]$, $D_{X_O}^k$ and $D_{X_V}^k$ are independent Poisson point processes having intensity functions $\alpha(x)\lambda_{D_X^k}(x)$ and $(1-\alpha(x))\,\lambda_{D_X^k}(x)$, respectively.

(M3) For k fixed, $D_Y^k = D_{Y_O}^k \cup D_{Y_S}^k$ where

(i) $(D_{X_O}^k, D_{Y_O}^k)$ is a marked Poisson point process with a stochastic kernel density $\ell(y|x)$.

Table 3.1: The parallels between the Bayesian framework for RVs and its counterpart for random PDs.

	Bayesian Framework for RVs	Bayesian Framework for Random PDs
Prior	Modeled by a prior density f	Modeled by a Poisson PP with prior intensity λ
Likelihood	Depends on observed data	Stochastic kernel that depends on observed PDs
Posterior	Compute the posterior density	A Poisson PP with posterior intensity

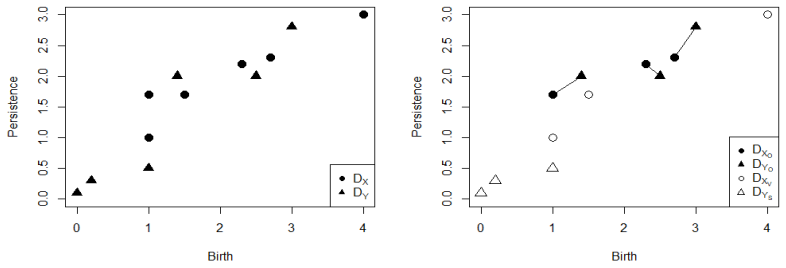
(ii) $D_{Y_O}^k$ and $D_{Y_S}^k$ are independent finite Poisson point processes where $D_{Y_S}^k$ has intensity function $\lambda_{D_{Y_S}^k}$.

Hereafter we abuse notation by writing D_X for D_X^k. The modeling assumption (M1) allows us to develop results independently for each homological dimension k then combine them using independence. In (M2), the random persistence diagram D_X is modelled as a Poisson point process with prior intensity λ_{D_X}. There are two cases we may encounter for any point x from the prior intensity due to the nature of persistence diagrams. First, depending upon the noise level in data, any feature x in $\mathcal{D}_X$ may not be represented in observations; in our model this scenario happens with probability $1-\alpha(x)$, and we account for this case with D_{X_V} in (M2). Otherwise, a point x spawns a mark with a probability of $\alpha(x)$, a scenario we model with D_{X_O} in (M2). These considerations explain why the intensities of D_{X_O} and D_{X_V} are proportional to the intensity λ_{D_X} weighted by $\alpha(x)$ and $1-\alpha(x)$, respectively. The total prior intensity for D_X is given by the sum of the intensities for D_{X_O} and D_{X_V} by Theorem 2.2.1. (M3) concerns observed persistence diagrams $\mathcal{D}_Y$, and decomposes the point process from which they are drawn into two independent point processes, D_{Y_O} and D_{Y_S}. D_{Y_O} is linked to D_{X_O} via a marked point process with stochastic kernel $\ell(y|x)$ defined in Equation (2.6), whereas the component D_{Y_S} accounts for any point y that arises from noise or unanticipated geometry. See Figure 3.1 for a graphical representation of these ideas.

Theorem 3.1.1 (Bayesian Theorem for Persistence Diagrams)**.** *Let D_X be a persistence diagram modeled by a Poisson point process as in (M2). Suppose D_{X_O} and D_{X_V} have prior intensities $\alpha(x)\lambda_{D_X}$ and $(1-\alpha(x))\lambda_{D_X}$, respectively. Consider $\mathcal{D}_{Y^1},\ldots,\mathcal{D}_{Y^m}$ independent samples from the point process that characterizes the persistence diagram D_Y of (M3) and denote $\mathcal{D}_{Y^{1:m}} := \cup_{i=1}^m \mathcal{D}_{Y^i}$ where $\mathcal{D}_{Y^i} = \mathcal{D}_{Y_O^i} \cup \mathcal{D}_{Y_S^i}$ for all $i = 1,\cdots,m$. Moreover, $\ell(y|x)$ is the likelihood associated for the stochastic kernel between D_{X_O} and D_{Y_O}, and $\lambda_{\mathcal{D}_{Y_S}}$ is the intensity of D_{Y_S} as defined in (M3). Then, a pointwise unbiased estimator for the posterior intensity of D_X is given by*

$$\lambda_{D_X|\mathcal{D}_{Y^{1:m}}}(x) = (1-\alpha(x))\,\lambda_{D_X}(x) + \frac{1}{m}\alpha(x)\sum_{i=1}^{m}\sum_{y\in\mathcal{D}_{Y^i}} \frac{\ell(y|x)\lambda_{D_X}(x)}{\lambda_{D_{Y_S}}(y) + \int_{\mathbb{W}} \ell(y|u)\alpha(u)\lambda_{D_X}(u)du} \quad . \tag{3.1}$$

Before we give an entirely rigorous proof of Theorem 3.1.1, we will make a plausibility argument for Equation (3.1) for the case $\alpha(x) = \alpha$.



(a) (b)

Figure 3.1: (a) A sample from the prior point process $\mathcal{D}_X$ and an observed persistence diagram $\mathcal{D}_Y$. (b) Decomposition of $\mathcal{D}_X$ into $\mathcal{D}_{X_O}$,$\mathcal{D}_{X_V}$ and $\mathcal{D}_Y$ into $\mathcal{D}_{Y_O}$,$\mathcal{D}_{Y_S}$. The points in $\mathcal{D}_{X_V}$ have no relationship to those in $\mathcal{D}_Y$, while those in $\mathcal{D}_{X_O}$ only generate observed points in $\mathcal{D}_{Y_O}$. The remaining observed points in $\mathcal{D}_{Y_S}$ model unanticipated features that one may obtain due to uncertainty/noise.

Plausibility Argument. By definition, $\lambda_{D_X|\mathcal{D}_Y}(x)$ is uniquely characterized by the identity

$$\int_A \lambda_{D_X|\mathcal{D}_Y}(x)\,dx = \mathbb{E}\left(|(D_X|\mathcal{D}_Y) \cap A|\right) \tag{3.2}$$

for any $A \subset \mathbb{R}^2$, where the expectation on the right hand side of Equation (3.2) is taken with respect to the probability measure of $D_X|\mathcal{D}_Y$ (see [16] for the construction of probability measures of point processes). We show that Equation (3.1) satisfies Equation (3.2) for a single diagram (m = 1). Notice that

$$\begin{aligned} D_X|\mathcal{D}_Y &= (D_{X_V} \cup D_{X_O})|\mathcal{D}_Y && (3.3)\\ &= D_{X_V} \cup (D_{X_O}|\mathcal{D}_Y) && (3.4)\end{aligned}$$

since D_{X_V} is independent of D_Y. Thus, by Theorem 2.2.1,

$$\begin{aligned} \mathbb{E}\left(|(D_X|\mathcal{D}_Y) \cap A|\right) &= \mathbb{E}\left(|D_{X_V} \cup (D_{X_O}|\mathcal{D}_Y) \cap A\right) && (3.5)\\ &= \mathbb{E}\left(|D_{X_V} \cap A|\right) + \mathbb{E}\left(|(D_{X_O}|\mathcal{D}_Y) \cap A\right) && (3.6)\end{aligned}$$

By assumption, the left term in Equation (3.6) is

$$\mathbb{E}\left(|D_{X_V} \cap A|\right) = \int_A (1-\alpha)\lambda_{D_X}(x)\,dx. \tag{3.7}$$

It remains to compute the right term in Equation (3.6). To this end, let $\mathbb{P}(y \in \mathcal{D}_{Y_O}, x \in A)$ be the joint probability that y is drawn from D_{Y_O} and that its corresponding element in the draw from D_{X_O}, x, lies in A. Notice,

$$\begin{aligned} \mathbb{E}\left(|(D_{X_O}|\mathcal{D}_Y) \cap A\right) &= \sum_{y \in \mathcal{D}_Y} \mathbb{P}(y \in \mathcal{D}_{Y_O}, x \in A) && (3.8)\\ &= \sum_{y \in \mathcal{D}_Y} \mathbb{P}(y \in \mathcal{D}_{Y_O})\mathbb{P}(x \in A | y \in \mathcal{D}_{Y_O}) && (3.9)\end{aligned}$$

where Equation 3.8 follows since elements of $\mathcal{D}_Y$ are drawn from D_{Y_S} or D_{Y_O}; in the former case, the elements bear no association to those in draws from D_{X_O} while in the former, each element corresponds to exactly one element in a draw from D_{X_O}. Moreover, each element in $\mathcal{D}_Y$ is an independent draw from the scaled intensity of D_Y since D_Y is a Poisson point process by Theorems 2.2.3 and 2.2.1. Since D_Y is an independent mixture of Poisson processes with intensities $\lambda_{D_{Y_O}}$ and $\lambda_{D_{Y_S}}$, it follows

$$\mathbb{P}(y \in \mathcal{D}_{Y_O}) = \frac{\lambda_{D_{Y_O}}(y)}{\lambda_{D_{Y_O}}(y) + \lambda_{D_{Y_S}}(y)} \tag{3.10}$$

$$= \frac{\int \alpha \lambda_{D_X}(x)\ell(y|x)\,dx}{\lambda_{D_{Y_S}}(y) + \int \alpha \lambda_{D_X}(x)\ell(y|x)\,dx} \tag{3.11}$$

where the identity $\lambda_{D_{Y_O}}(y) = \int \lambda_{D_X}(x)\ell(y|x)\,dx$ follows by Theorem 2.2.3. To compute the other component in Equation (3.9), we rely on Bayes rule to compute a posterior probability:

$$\mathbb{P}(x|y) = \frac{\mathbb{P}(x)\mathbb{P}(y|x)}{\int \mathbb{P}(x)\mathbb{P}(y)|x)\,dx}. \tag{3.12}$$

Using the fact that D_{X_O} is a marked Poisson point process, we substitute for each of the components in Equation (3.12):

$$\mathbb{P}(x|y) = \frac{\mathbb{P}(x)\mathbb{P}(y|x)}{\int \mathbb{P}(x)\mathbb{P}(y|x)\,dx} \tag{3.13}$$

$$= \frac{\lambda_{D_X}(x)\ell(y|x)}{\int \lambda_{D_X}(x)\ell(y|x)\,dx}. \tag{3.14}$$

Next, we integrate Equation (3.14) over A:

$$\mathbb{P}(x \in A|y \in \mathcal{D}_{Y_O}) = \frac{1}{\int \lambda_{D_X}(x)\ell(y|u)\,du} \int_A \lambda_{D_X}(x)\ell(y|x)\,dx \tag{3.15}$$

Substituting Equations (3.11) and (3.15) in Equation (3.9) yields

$$\mathbb{E}\left(|(D_{X_O}|\mathcal{D}_Y) \cap A\right) = \sum_{y \in \mathcal{D}_Y} \int_A \frac{\alpha \lambda_{D_X}(x)\ell(y|x)}{\lambda_{D_{Y_S}}(y) + \int \alpha \lambda_{D_X}(x)\ell(y|x)\,dx}) \,dx. \tag{3.16}$$

Further substitution of Equations (3.16) and (3.7) into Equation 3.2 reveals:

$$\mathbb{E}\left(|(D_X|\mathcal{D}_Y) \cap A|\right) = \int_A \left((1-\alpha)\lambda_{D_X}(x) + \sum_{y \in \mathcal{D}_Y} \frac{\alpha \lambda_{D_X}(x)\ell(y|x)}{\lambda_{D_{Y_S}}(y) + \int \alpha \lambda_{D_X}(x)\ell(y|x)\,dx}\right) dx. \tag{3.17}$$

By Theorem 2.2.1, the intensity for the point process given by the superposition $\cup_{i=1}^m D_X|\mathcal{D}_{Y_i}$ is $m\lambda_{D_X|\mathcal{D}_Y}(x)$, establishing Equation (3.1) after substituting the integrand from Equation (3.17). □

Now we prove Theorem 3.1.1 using tools discussed in Section (2.2.2).

Proof of Theorem 3.1.1. By Theorem 2.2.1, we decompose $\lambda_{D_X|\mathcal{D}_{Y^{1:m}}}$ to write

$$\lambda_{D_X|\mathcal{D}_{Y^{1:m}}} = \lambda_{D_{X_Y}|\mathcal{D}_{Y^{1:m}}} + \lambda_{D_{X_O}|\mathcal{D}_{Y^{1:m}}} = (1-\alpha(x))\lambda_{D_X} + \lambda_{D_{X_O}|\mathcal{D}_{Y^{1:m}}}, \tag{3.18}$$

where the second equality follows because D_{X_V} is independent of D_Y. Theorem 2.2.1 allows us to express $\lambda_{D_{X_O}}$ as the average of intensity functions $\lambda_{D_{X_O^i}}$ for $i = 1, \cdots, m$, where the $D_{X_O^i}$ are independent and equal in distribution to D_{X_O}. That is, $\lambda_{D_{X_O}} = \frac{1}{m}\sum_{i=1}^{m} \lambda_{D_{X_O^i}}$, and by conditioning we have,

$$\lambda_{D_{X_O}|\mathcal{D}_{Y^{1:m}}} = \frac{1}{m}\sum_{i=1}^{m} \lambda_{D_{X_O^i}|\mathcal{D}_{Y^i}}. \tag{3.19}$$

So to expand Equation (3.18) it suffices to compute $\lambda_{D_{X_O^i}|D_{Y^i}}$ for fixed i. First, we express the finite point process (D_X, D_Y) as a marked Poisson point process. To this end, we adopt a construction from [68], the augmented space $\mathbb{W}' := \mathbb{W} \cup \{\Delta\}$, where Δ is a dummy set used for labeling points in D_{Y_S}. Next, we define the random set, $\mathcal{H} = \mathcal{H}_{\mathbb{W}} \cup \mathcal{H}_{\Delta}$ such that

$$\mathcal{H} := \Big\{(x, y) \in (\mathcal{D}_{X_O}, \mathcal{D}_{Y_O})\Big\} \bigcup \Big\{(\Delta, y) : y \in \mathcal{D}_{Y_S}\Big\}. \tag{3.20}$$

One can observe that $\mathcal{H}$ is the superposition of two marked Poisson point processes $\mathcal{H}_{\mathbb{W}}$ and $\mathcal{H}_{\Delta}$, taking values in $\mathbb{W} \times \mathbb{W}$ and $\Delta \times \mathbb{W}$, respectively. Moreover, it directly follows from $(M2)$ and $(M3)(i)$ that $\mathcal{H}_{\mathbb{W}}$ has marginal intensity function $\alpha(x)\lambda_{D_X}(x)$ on $\mathbb{W}$ and stochastic kernel density $\ell(y|x)$ while $(M3)(ii)$ shows that $\mathcal{H}_{\Delta}$ has marginal intensity function $\lambda_{D_{Y_S}}(\mathbb{W})$ on $\{\Delta\}$ with stochastic kernel density $\frac{\lambda_{D_{Y_S}}(y)}{\lambda_{D_{Y_S}}(\mathbb{W})}$. By Theorem 2.2.3, the intensity functions for $\mathcal{H}_{\mathbb{W}}$ and $\mathcal{H}_{\Delta}$ are $\alpha(x)\lambda_{D_X}(x)\ell(y|x)$ and $\lambda_{D_{Y_S}}(y)$, respectively. Hence, applying Theorem 2.2.1 to Equation (3.20) reveals that the intensity function for $\mathcal{H}$, $\lambda_{\mathcal{H}}$, is given by

$$\lambda_{\mathcal{H}}(x, y) = \alpha(x)\lambda_{D_X}(x)\ell(y|x)\mathbb{1}_{x\in\mathbb{W}} + \lambda_{D_{Y_S}}(y)\mathbb{1}_{x\in\Delta}. \tag{3.21}$$

Let $\mathcal{H}_Y := \{y : (x, y) \in \mathcal{H}\}$, $\mathcal{H}_X := \{x : (x, y) \in \mathcal{H}\}$ be the projections of $\mathcal{H}$ onto its first and second coordinates, respectively. It immediately follows from Theorem 2.2.2 that $\mathcal{H}_Y$ is a Poisson point process on $\mathbb{W}$ since it is the image of $\mathcal{H}$ under a projection. Therefore, by treating the first coordinates of $\mathcal{H}$ as marks, we may express $\mathcal{H}$ as a marked Poisson point process having intensity function $\lambda_{\mathcal{H}_Y}$ on $\mathbb{W}$ and stochastic kernel density $p(x|y)$ from $\mathbb{W}$ to $\mathbb{W}'$. Another application of Theorem 2.2.3 then implies

$$\lambda_{\mathcal{H}}(x, y) = \lambda_{\mathcal{H}_Y}(y)p(x|y). \tag{3.22}$$

From Equations (3.21) and (3.22), we obtain the identity

$$p(x|y) = \frac{\alpha(x)\lambda_{D_X}(x)\ell(y|x)\mathbb{1}_{x\in\mathbb{W}} + \lambda_{D_{Y_S}}(y)\mathbb{1}_{x\in\Delta}}{\lambda_{\mathcal{H}_Y}(y)}, \;\; \lambda_{\mathcal{H}_Y}(y) \neq 0. \tag{3.23}$$

Equation (3.23) describes the probability density of $\mathcal{H}$ at $x \in \mathbb{W}'$ for $y \in \mathbb{W}$ fixed. Substituting Equation (3.23) for the Janossy density in Equation (2.6) and applying Corollary 2.2.1 gives the intensity function for the point process $\mathcal{H}_X|\mathcal{D}_{Y^i}$ whenever $\lambda_{\mathcal{H}_Y}(y) \neq 0$ for any $y \in D_{Y^i}$,

$$\lambda_{\mathcal{H}_X|\mathcal{D}_{Y^i}}(x) = \sum_{y\in D_{Y^i}} \frac{\alpha(x)\lambda_{D_X}(x)\ell(y|x)\mathbb{1}_{x\in\mathbb{W}} + \lambda_{D_{Y_S}}(y)\mathbb{1}_{x\in\Delta}}{\lambda_{\mathcal{H}_Y}(y)}, \;\; \lambda_{\mathcal{H}_Y}(y) \neq 0 \tag{3.24}$$

Restricting Equations (3.21) and (3.22) to $\mathbb{W} \times \mathbb{W}$, we obtain $p(x|y)\lambda_{\mathcal{H}_Y}(y) = \ell(y|x)\alpha(x)\lambda_{D_X}(x) = 0$, whenever $\lambda_{\mathcal{H}_Y}(y) = 0$, from which we conclude $\lambda_{\mathcal{H}_Y}(y) \neq 0$ a.s. Hence, restricting Equation (3.24) to $\mathbb{W} \times \mathbb{W}$ yields

$$\lambda_{D_{X_O}|D_{Y^i}}(x) = \sum_{y\in D_{Y^i}} \frac{\alpha(x)\lambda_{D_X}(x)\ell(y|x)}{\lambda_{\mathcal{H}_Y}(y)}, \quad a.s. \tag{3.25}$$

Notice that $\mathcal{H}_Y$ is the same point process as $D_{Y_O} \cup D_{Y_S}$. Theorem 2.2.2 implies that D_{Y_O} is a Poisson point process, and D_{Y_S} is a Poisson point process by (M3), so by Theorem 2.2.1, $\lambda_{\mathcal{H}_Y} = \lambda_{D_{Y_O}} + \lambda_{D_{Y_S}}$, where $\lambda_{D_{Y_O}}(y) = \lambda_{(D_{X_O}, D_{Y_O})}(\mathbb{W} \times y) = \int_{\mathbb{W}} \alpha(u)\lambda_{D_{X_O}}(u)\ell(y|u)du$ by Theorem 2.2.3. Employing Equation (3.25) one gets that

$$\lambda_{D_{X_O}|D_{Y^i}}(x) = \alpha(x) \sum_{y \in D_{Y^i}} \frac{\ell(y|x)\lambda_{D_X}(x)}{\lambda_{D_{Y_S}}(y) + \int_{\mathbb{W}} \ell(y|u)\alpha(u)\lambda_{D_X}(u)du}, \tag{3.26}$$

which proves Theorem 3.1.1 after substituting into Equation (3.18). □

One important point about Theorem 3.1.1 is that, instead of relying on a likelihood function for the point cloud data, our Bayesian model considers the likelihood for the persistence diagram generated by the observed point cloud data at hand. This is analogous to the idea of substitution likelihood by Jeffreys in [31].

3.1.2 A Conjugate Family of Prior Intensities: Gaussian Mixtures

This section focuses on constructing a family of conjugate prior intensities, i.e., a collection of priors that yield posterior intensities of the same form when used in Equation (3.1). Exploiting Theorem 3.1.1 with Gaussian mixture prior intensities, we obtain Gaussian mixture posterior intensities. As persistence diagrams are stochastic point processes on the space $\mathbb{W}$, not $\mathbb{R}^2$, we consider a Gaussian density restricted to $\mathbb{W}$. Namely, for a Gaussian density on $\mathbb{R}^2$, $\mathcal{N}(z; \upsilon, \sigma I)$, with mean υ and covariance matrix σI , we use

$$\mathcal{N}^*(z; \upsilon, \sigma I) := \mathcal{N}(z; \upsilon, \sigma I)\mathbb{1}_{\mathbb{W}}(z) \tag{3.27}$$

to denote its restriction to $\mathbb{W}$, where $\mathbb{1}_{\mathbb{W}}$ is the indicator function of the wedge $\mathbb{W}$.

Consider a random persistence diagram D_X as in (M2) and a collection of observed persistence diagrams $\{\mathcal{D}_{Y^1}, \cdots, \mathcal{D}_{Y^m}\}$ that are independent samples from the Poisson point process characterizing the persistence diagram D_Y in (M3). We denote $\mathcal{D}_{Y^{1:m}} := \cup_{i=1}^m \mathcal{D}_{Y^i}$. To create a conjugate family (in the sense that applying Theorem 3.1.1 to a prior intensity from the family of functions yields a posterior intensity in the same family) we specialize (M2) and (M3) below:

(M2′) $D_X = D_{X_O} \cup D_{X_V}$, where D_{X_O} and D_{X_V} are independent Poisson point processes with intensities $\alpha\lambda_{D_X}(x)$ and $(1-\alpha)\lambda_{D_X}(x)$, respectively, with

$$\lambda_{D_X}(x) = \sum_{j=1}^{N} c_j^{D_X} \mathcal{N}^*(x; \mu_j^{D_X}, \sigma_j^{D_X} I), \tag{3.28}$$

where N is the number of mixture components.

(M3′) $D_Y = D_{Y_O} \cup D_{Y_S}$ where

(i) the marked Poisson point process (D_{X_O}, D_{Y_O}) has density $\ell(y|x)$ given by

$$\ell(y|x) = \mathcal{N}^*(y; x, \sigma^{D_{Y_O}} I). \tag{3.29}$$

(ii) D_{Y_O} and D_{Y_S} are independent finite Poisson point processes and D_{Y_S} has intensity function given below.

$$\lambda_{D_{Y_S}}(y) = \sum_{k=1}^{M} c_k^{D_{Y_S}} \mathcal{N}^*(y; \mu_k^{D_{Y_S}}, \sigma_k^{D_{Y_S}} I), \tag{3.30}$$

where M is the number of mixture components.

Proposition 3.1.1. *Suppose that the assumptions (M1),(M2′), and (M3′) hold; then, the posterior intensity of Equation* (3.1) *in Theorem 3.1.1 is a Gaussian mixture of the form*

$$\lambda_{D_X|D_{Y^{1:m}}}(x) = (1-\alpha)\lambda_{D_X}(x) + \frac{\alpha}{m}\sum_{i=1}^{m}\sum_{y\in\mathcal{D}_{Y^i}}\sum_{j=1}^{N} C_j^y \mathcal{N}^*(x;\mu_j^y,\sigma_j^y I), \tag{3.31}$$

$$\begin{aligned} \textit{where} \quad C_j^y &= \frac{w_j^y}{\lambda_{D_{Y_S}}(y) + \alpha\sum_{j=1}^{N} w_j^y Q_j^y};\ Q_j^y = \int_{\mathbb{W}} \mathcal{N}(u;\mu_j^y,\sigma_j^y I)du; \\ w_j^y &= c_j^{D_X}\mathcal{N}(y;\mu_j^{D_X},(\sigma^{D_{Y_O}}+\sigma_j^{D_X})I); \\ \textit{and} \quad \mu_j^y &= \frac{\sigma_j^{D_X} y + \sigma^{D_{Y_O}}\mu_j^{D_X}}{\sigma_j^{D_X}+\sigma^{D_{Y_O}}}; \sigma_j^y = \frac{\sigma^{D_{Y_O}}\,\sigma_j^{D_X}}{\sigma_j^{D_X}+\sigma^{D_{Y_O}}}. \end{aligned}$$

The proof of Proposition 3.1.1 follows from well known results about products of Gaussian densities given below; for more details, the reader may refer to [45] and references therein.

Lemma 3.1.1. *For $p\times p$ matrices H, R, P, with R and P positive definite ,and a $p\times 1$ vector s, $\mathcal{N}(y;Hx,R)\mathcal{N}(x;s,P) = q(y)\mathcal{N}(x;\hat{s},\hat{P})$, where $q(y) = \mathcal{N}(y;Hs,R+HPH^T)$, $\hat{s} = s + K(y - Hs)$, $\hat{P} = (I-KH)P$ and $K = PH^T(HPH^T+R)^{-1}$.*

Proof of Proposition 3.1.1. Using Lemma 3.1.1, we first derive $\ell(y|x)\,\lambda_{D_X}(x)$ by observing that, in our model, $H = I, R = \sigma^{D_{Y_O}}I, s = \mu_j^{D_X}$ and $P = \sigma_j^{D_X}I$. By typical matrix operations we obtain, $K = \frac{\sigma_j^{D_X}}{\sigma_j^{D_X}+\sigma^{D_{Y_O}}}$, $\hat{s} = \frac{\sigma_j^{D_X} y + \sigma^{D_{Y_O}}\mu_j^{D_X}}{\sigma_j^{X}+\sigma^{D_{Y_O}}}$, and $\hat{P} = \frac{\sigma^{D_{Y_O}}\sigma_j^{D_X}}{\sigma_j^{D_X}+\sigma^{D_{Y_O}}}$. Hence the numerator and denominator of the second term in Equation (3.1), $\sum_{j=1}^{N} c_j^{D_X}\mathcal{N}(y;\mu_j^{D_X},(\sigma^{D_{Y_O}}+\sigma_j^{D_X})I)\mathcal{N}^*(x;\mu_j^y,\sigma_j^y I)$, and $\lambda_{D_{Y_S}}(y) + \alpha\sum_{j=1}^{N} c_j^{D_X}\mathcal{N}(y;\mu_j^{D_X},(\sigma^{D_{Y_O}}+\sigma_j^{D_X})I)\int_{\mathbb{W}}\mathcal{N}(u;mu_j^y,\sigma_j^y I)du$, respectively, yield

$$\sum_{j=1}^{N}\left[\frac{c_j^{D_X}\mathcal{N}(y;\mu_j^{D_X},(\sigma^{D_{Y_O}}+\sigma_j^{D_X})I)}{\lambda_{D_{Y_S}}(y)+\alpha\sum_{j=1}^{N} c_j^{D_X}\mathcal{N}(y;\mu_j^{D_X},(\sigma^{D_{Y_O}}+\sigma_j^{D_X})I)\int_{\mathbb{W}}\mathcal{N}(u;\mu_j^{x|y},\sigma_j^{x|y}I)du}\right]\mathcal{N}^*(x;\mu_j^{x|y},\sigma_j^{x|y}I),$$

where the bracketed expression is the definition of C_j^y. □

3.1.3 Sensitivity Analysis

In this section, we present a detailed example of computing the posterior intensity according to Equation (3.31) for a range of parametric choices. Any reader interested in reproducing these results may download our R-package BayesTDA or Python package bayes-tda. We consider circular point clouds often associated with periodicity in signals [43] and focus on estimating homological features with $k = 1$ as they correspond to 1-dimensional holes. Precisely our goals are to: (i) illustrate posterior intensities and draw analogies to standard Bayesian inference; (ii) determine the relative contributions of the prior and observed data to the posterior; and (iii) perform sensitivity analysis. We consider 4 cases, where Case I corresponds to an ideal choice of model parameters with respect to the true distribution of data, and Cases II - IV are purposefully disadvantageous.

We start by considering a Poisson point process with prior intensity λ_{D_X} that has the Gaussian mixture form given in (M2′). We take into account four types of prior intensities: (i) informative, (ii) weakly informative, (iii) unimodal uninformative, and (iv) bimodal uninformative; see Figures 3.2–3.4 (a), (d), (g), (j), respectively. We use one Gaussian component in each of the first three priors, as the underlying shape has single 1−dimensional feature, and two for the last one to include a case where we have no information about the cardinality of the underlying true diagram.

The parameters of the Gaussian mixture density in Equation (3.28) used to compute these prior intensities are listed in Table 3.3. To present the intensity maps uniformly throughout this example while preserving their shapes, we divide the intensities by their corresponding maxima. This ensures all intensities are on a scale from 0 to 1. The observed persistence diagrams are generated from point clouds that are sampled uniformly from the unit circle and then perturbed by varying levels of Gaussian noise; see Figure 3.5. These point clouds provide persistence diagrams $\mathcal{D}_{Y^i}$ for $i = 1, 2, 3$, which are considered as independent samples from Poisson point process D_Y, exhibiting distinctive characteristics such as only one prominent feature with high persistence and no spurious features (*Case-I*), one prominent feature with high persistence and very few spurious features (*Case-II*), and one prominent feature with medium persistence and more spurious features (*Case-III*).

For each observed persistence diagram, 1-dimensional persistence features are presented as green dots overlaid on their corresponding posterior intensity plots. For *Cases-I-III*, we set the probability α of the event that a feature in $\mathcal{D}_X$ spawns a mark in $\mathcal{D}_Y$ to 1 and later in *Case-IV*, we decrease α to 0.5 while keeping all other parameters the same for the sake of comparison. The choice of $\alpha = 0.5$ anticipates that any feature has equal probability to appear or disappear in the observation and in turn provides further intuition about the contribution of prior intensities to the estimated posteriors. We observe that in all cases, the posterior detects the $1-$dimensional hole; however, with different levels of uncertainty each time. For example, for the cases where the data are trustworthy, expressed by a likelihood with tight variance, or in the case of an informative prior, the posterior accurately estimates the $1-$dimensional hole. In contrast, when the data suffer from high uncertainty and the prior is uninformative, then the posterior offers a general idea that the true underlying shape is a circle, but the exact estimation of the 1-dimensional hole is not accurate. We examine the cases in detail below.

Case-I: To compute posterior intensities, we consider informative, weakly informative, unimodal uninformative, and bimodal uninformative prior intensities, which are depicted in Figures 3.2 (a), (d), (g) and (j), respectively. The prior intensities parameters are listed in Table 3.3. The observed persistence diagram is obtained from the point cloud in Figure 3.5 (left) via a Vieotoris-Rips filtration; see Section 2.1.1. The parameters associated to the observed persistence diagram are listed in Table 3.4. For the observed persistence diagram arising from data with very low noise, we observe that the posterior computed from any of the priors predicts the existence of a one dimensional hole accurately. First, with a low variability in the observed persistence diagram ($\sigma^{D_{Y_O}} = 0.01$ and $\sigma^{D_{Y_S}} = 0.1$), the posterior intensities estimate the hole with high certainty (Figure 3.2 (b), (e), (h) and (k) respectively). Next, to determine the effect of observed data on the posterior, we increase the variance of the observed persistence diagram component D_{Y_O}, which consists of features in observed persistence diagrams that are associated to the underlying prior. Here, we observe that the posterior intensities still estimate the hole accurately due to the trustworthy data; this is evident in Figure 3.2 (c), (f), (i) and (l). In Figure 3.2, the posteriors in (b), (e), (h), and (k) have lower variance around the 1-dimensional feature in comparison to those in (c), (f), (i), and (l), respectively.

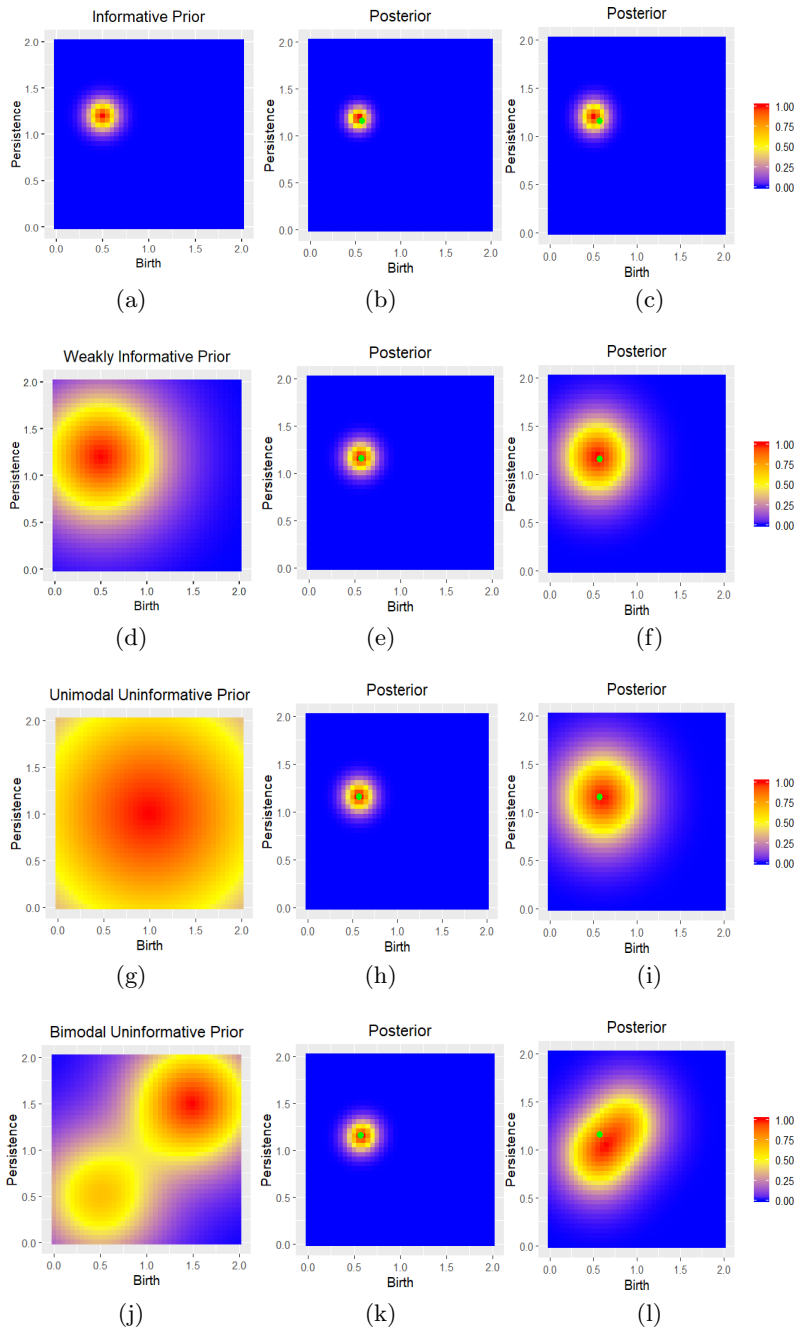


Figure 3.2: Case-I: Posterior intensities obtained by using Proposition 3.1.1. We consider informative (a), weakly informative (d), unimodal uninformative (g), and bimodal uninformative (j) prior intensities . The color maps represent scaled intensities. The list of associated parameters of observed persistence diagram used for this case are in Table 3.4. Posteriors computed from all of these priors estimate the 1−dimensional hole accurately for two choice of variance in the observed persistence diagram ($\sigma^{D_{YO}} = 0.01$ and $\sigma^{D_{YS}} = 0.1$), which are presented in (b), (e), (h) and (k). After increasing the variance to $\sigma^{D_{YO}} = 0.1$, we observe the posteriors still estimate hole but with higher variance as presented in (c), (f), (i) and (k).

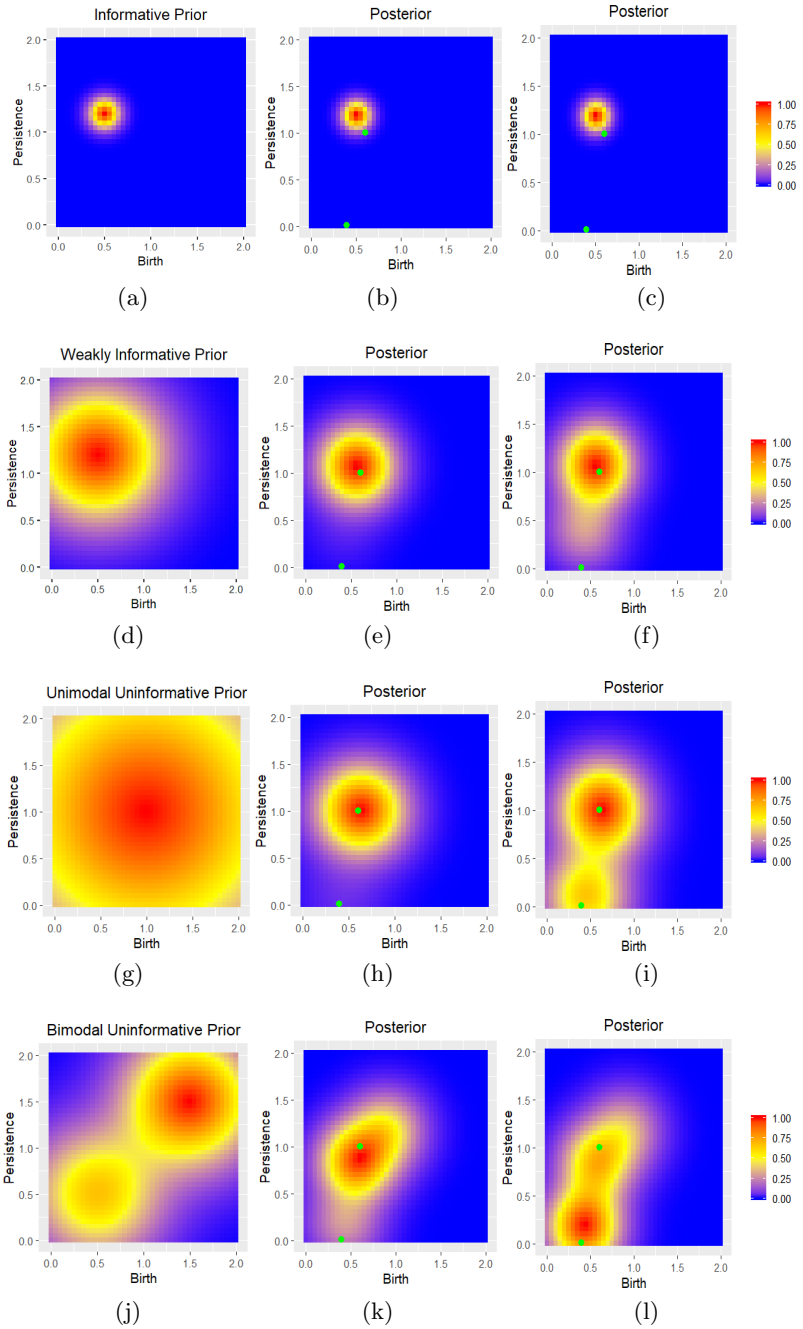


Figure 3.3: Case-II: We consider informative (a), weakly informative (d), unimodal uninformative (g), and bimodal uninformative (j) prior intensities to estimate posterior intensities using Proposition 3.1.1. The color maps represent scaled intensities. The parameters we use for estimating the posterior intensity are listed in Table 3.4. The posterior intensities estimated from the informative prior in (b) and (c) estimate the $1-$dimensional hole with high certainty. Also, the posterior intensities estimated from the weakly informative and uninformative priors in (e), (h), and (k) imply existence of a hole with lower certainty. (c), (f), (i) and (l) represent the posterior with higher variance in observed PD component D_{Y_S}. As this makes the assumption that every observed point is associated to D_X, we observe increased intensity skewed towards the spurious point in (f). Furthermore in (i) and (l), we observe bimodality in the posterior intensity.

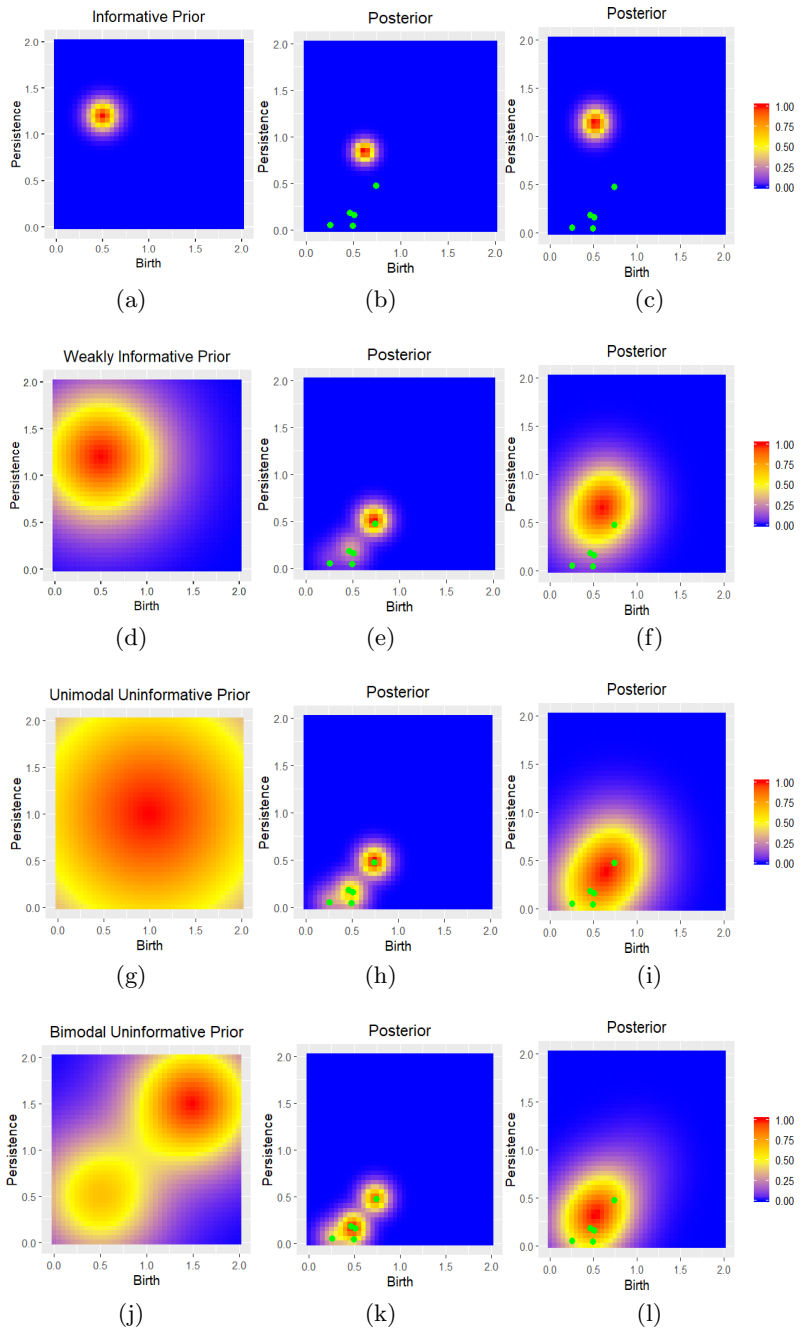


Figure 3.4: Case-III: Posterior intensities obtained by using Proposition 3.1.1. We consider informative (a), weakly informative (d), unimodal uninformative (g), and bimodal uninformative (j) prior intensities. The color maps represent scaled intensities. Parameters of the observed PD used to estimate the posterior intensity are listed in Table 3.4. With choices of $\sigma^{D_{Y_O}} = 0.01$ and $\sigma^{D_{Y_S}} = 0.1$, we observe the posteriors can deduce existence of the prominent feature, as presented in (b), (e), (h), and (k) as we have more confidence on the component of observed data associated to prior. Otherwise, with an increased variance $\sigma^{D_{Y_O}} = 0.1$, only the posterior intensity from the informative prior is able to detect the hole with high certainty, as observed in (c). For the weakly informative and uninformative priors, the posteriors in (f), (i), and (l) may not detect the hole directly, but the mode of (f) with higher variance and the tail towards the prominent point in (i) and (l) imply the existence of a hole in the underlying persistence diagram.

Table 3.2: Case-IV: The first, second and third columns match the parameters of observed persistence diagram $\mathcal{D}_Y$ used in computing the posteriors depicted in the third column of Figure 3.2 and Figure 3.3, and second column of Figure 3.4, respectively with $\alpha = 0.5$. The parameters are presented in Table 3.4. The color maps represent scaled intensities. A variation in the level of intensity is observed for all of them compared to their respective cases due to the added term in the posterior intensity. The posterior intensities in first and second columns exhibit the estimation of the hole with higher variability when compared to their respective figures in Case-I and Case-II. The posteriors in the third column demonstrate dominance of the prior relative to their corresponding figures in Case-III, especially when one examines those for informative, weakly informative and bimodal uninformative priors.

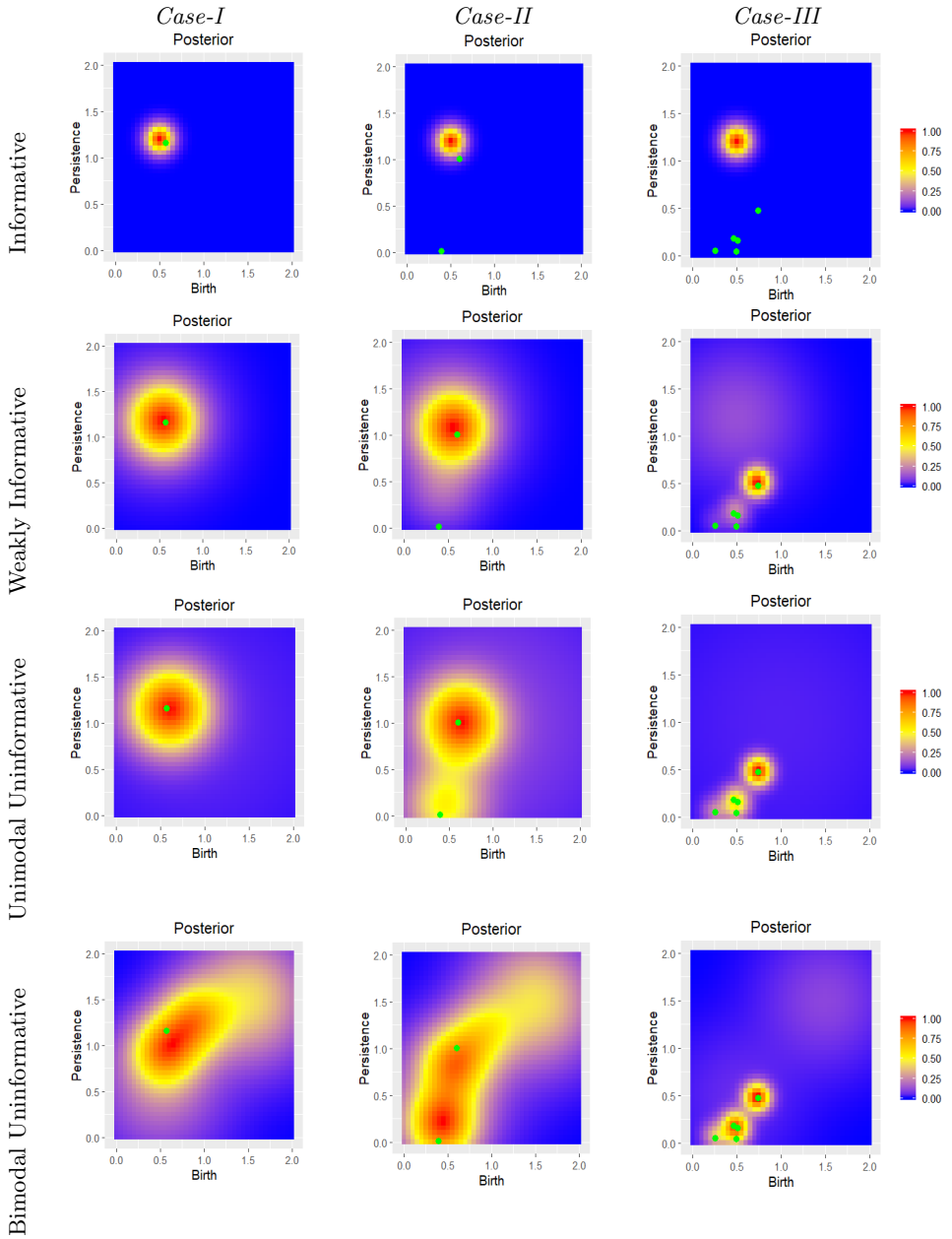


Table 3.3: List of Gaussian mixture parameters of the prior intensities in Equation (3.28). The means $\mu_i^{D_X}$ are 2×1 vectors and the rest are scalars

.

	$\mu_i^{D_X}$	$\sigma_i^{D_X}$	$c_i^{D_X}$
Informative Prior	$(0.5, 1.2)$	0.01	1
Weakly informative Prior	$(0.5, 1.2)$	0.2	1
Unimodal Uninformative Prior	$(1, 1)$	1	1
Bimodal Uninformative Prior	$(0.5, 0.5)$ $(1.5, 1.5)$	0.2 0.2	1 2

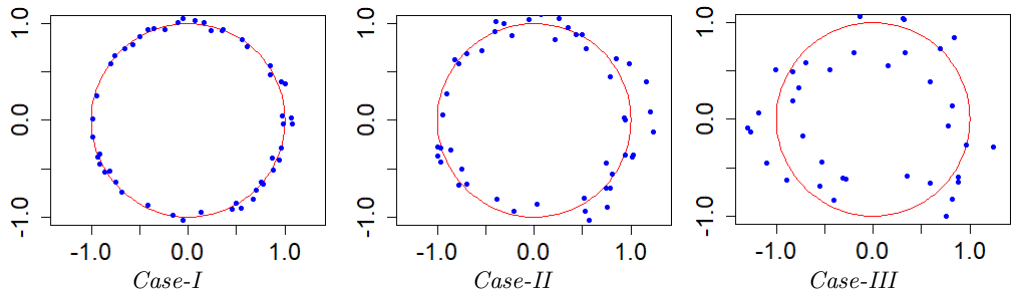


Figure 3.5: The observed datasets generated for Case-I, Case-II, and Case-III by sampling the unit circle and perturbing with Gaussian noise having variances $0.001I_2$, $0.01I_2$, and $0.1I_2$, respectively.

Case-II: Here, we consider the same four priors as in *Case-I* (see Figure 3.3 (a), (d), (g) and (j)). The point cloud in Figure 3.5 (center) is more perturbed around the unit circle than that of *Case-I* (Gaussian noise with variance $0.01I_2$). Due to this, the associated persistence diagram exhibits spurious features. The parameters used for this case are listed in Table 3.4. We compute the posterior intensities for each type of prior. First, to illustrate the posterior intensity and check the capability of detecting the $1-$dimensional feature, we use moderate noise for the observed persistence diagram ($\sigma^{D_{Y_O}} = 0.1$ and $\sigma^{D_{Y_S}} = 0.1$). The results are presented in Figure 3.3 (b), (e), (h), and (k); overall, the posteriors estimate the prominent feature with different variances in their respective posteriors. Next, to illustrate the effect of observed data on the posterior, we increase the variance $\sigma^{D_{Y_S}}$ of D_{Y_S}. According to our Bayesian model, the persistence diagram component D_{Y_S} contains features that are not associated with D_X, so increasing $\sigma^{D_{Y_S}}$ anticipates that observed points are more likely linked to D_X, and therefore one may expect to observe increased intensity skewed towards the spurious points that arise from noise. Indeed, posterior intensities with weakly informative, unimodal uninformative, and bimodal uninformative priors exhibit the skewness toward the spurious point in Figure 3.3 (f), (i) and (l) respectively, but this is not the case when an informative prior is used. In (f), we observe increased intensity skewing towards the

Table 3.4: Parameters for (M3′) in Equation (3.29) and (3.30). We set the weight and mean of the Gaussian component, $c^{D_{Y_S}} = 1$ and $\mu^{D_{Y_S}} = (0.5, 0)$ respectively for all of the cases. The first row corresponds to parameters in the functions characterizing D_Y that are used in computing the posterior depicted in the first column of Figure 3.2. The second row corresponds to analogous parameters that are used in computing the posterior depicted in the second columns of Figures 3.2–3.2. Similarly, the third row corresponds to parameters in the functions characterizing D_Y used for computing the posterior presented in the third columns of Figures 3.2–3.4.

Case-I	*Case-II*	*Case-III*	*Case-IV*
			$\sigma^{D_{Y_O}} = 0.1$ $\sigma^{D_{Y_S}} = 0.1$
$\sigma^{D_{Y_O}} = 0.01$ $\sigma^{D_{Y_S}} = 0.1$	$\sigma^{D_{Y_O}} = 0.1$ $\sigma^{D_{Y_S}} = 0.1$	$\sigma^{D_{Y_O}} = 0.01$ $\sigma^{D_{Y_S}} = 0.1$	$\sigma^{D_{Y_O}} = 0.1$ $\sigma^{D_{Y_S}} = 1$
$\sigma^{D_{Y_O}} = 0.1$ $\sigma^{D_{Y_S}} = 0.1$	$\sigma^{D_{Y_O}} = 0.1$ $\sigma^{D_{Y_S}} = 1$	$\sigma^{D_{Y_O}} = 0.1$ $\sigma^{D_{Y_S}} = 0.1$	$\sigma^{D_{Y_O}} = 0.01$ $\sigma^{D_{Y_S}} = 0.1$

spurious points, and in (i) and (l) the intensity appears to be bimodal with two modes – one at the prominent and other at the spurious point. For the bimodal uninformative prior since one mode is located close to the spurious point in the observed persistence diagram, we observe higher intensity for that mode in the posterior (Figure 3.3 (l)) with another mode estimating the prominent feature.

Case-III: The same four priors are also considered here. The observed persistence diagram is constructed from the point cloud in Figure 3.5 (right) via a Vietoris-Rips filtration. The point cloud has Gaussian noise with variance $0.1I_2$, and due to the high noise level in sampling relative to the unit circle, the associated persistence diagram exhibits one prominent feature and several spurious features. We repeat the parameter choices as in *Case-I* for the variances of the observed persistence diagram. For the choices of $\sigma^{D_{Y_O}} = 0.01$ and $\sigma^{D_{Y_S}} = 0.1$, the posteriors computed from all of the four priors are able to detect the difference between the one prominent and other spurious points. We increase the variance of D_{Y_O} to determine the effect of the observed persistence diagram on the posterior and we observe that only the posterior intensity from informative prior has evidence of the hole (Figure 3.4(c)). For the weakly informative and uninformative priors, while the posteriors in (f), (i) and (l) may not detect the hole clearly, in (f) we observe a mode with higher variance and in (i) and (l), a tail towards the high persistence point implying presence of a hole. It should be noted that with the informative prior the posterior intensity identifies the hole closer to the mode of the prior as we increase the variance in $\sigma^{D_{Y_O}}$.

Case-IV: Lastly, in this case we concentrate on the effect of α. The rest of the parameters used for this case remain the same and are listed in Table 3.4. We decrease α to 0.5 to model the scenario that a feature in D_X has equal probability to spawn a mark or not in the observed persistence diagram $\mathcal{D}_Y$. The columns of Table 3.2 correspond to the parameters of the observed persistence diagram $\mathcal{D}_Y$ used in computing the posteriors depicted in the third column of Figure

3.2, third column of Figure 3.3, and second column of Figure 3.4, respectively. By comparing them with their corresponding scenarios in previous cases, we notice a change in the intensity level in all of these due to the first term of the posterior intensity on the right hand side of Equation (3.31). Indeed, comparing with the corresponding figures in *Case-I*, we observe that the posterior intensities estimate the hole with higher variability for the weakly informative and unimodal uninformative priors. For the bimodal prior, we observe bimodality in the posterior. Next for *Case-II*, the existence of a hole is evident for the informative and weakly informative priors, but with higher uncertainty when compared to their previous cases. The unimodal and bimodal uninformative priors lead to bimodal and trimodal posteriors, respectively. We observe that the posterior resembles the prior intensity more closely when we compare them to their respective figures in *Case-III*. One can especially see this with the informative, weakly informative, and bimodal uninformative priors, which have significantly increased intensities at the locations of the prior modes.

3.2 A Note on Bandwidth Selection

We briefly remark on a method for selection of $\sigma^{D_{Y_O}}$ in the stochastic kernel given in Equation (3.29). Take $m_1, \ldots, m_O$ to be i.i.d random measures (or equivalently, point processes; see [16] or [52]) on $\mathbb{R}^2$ such that there exists a non-negative constant C with $|m_1| < C$. For our purposes, it suffices to consider the m_i as finite discrete measures. Suppose that $\mathbb{E}(m_1)$ has a bounded probability density p with respect to Lebesgue measure (if we think of m_1 as a point process, p is its intensity). Let $K_H : \mathbb{R}^2 \to \mathbb{R}$ be a kernel density with bandwidth matrix H. Define

$$\hat{p}_H(x) := \frac{1}{O}\sum_{i=1}^{O}\int K_H(x-y)m_i(dy), \tag{3.32}$$

and

$$\hat{p}_{iH}(x) := \frac{1}{O-1}\sum_{i\neq j}\int K_H(x-y)m_j(dy). \tag{3.33}$$

Let $MISE$ denote the mean integrated squared error:

$$MISE(H) := \mathbb{E}\left(\int (p(x)-\hat{p}_H(x))^2 dx\right). \tag{3.34}$$

We are interested in choosing H that minimizes Eq. (3.34). Equivalently, we may minimize $J(H) := MISE(H) - \|p\|^2$. The quantity

$$\hat{J}(H) := \frac{1}{O^2}\sum_{i,j}\iint K_H^{(2)}(x-y)m_i(dx)m_j(dy) - \frac{2}{O}\sum_i\int \hat{p}_{iH}(x)m_i(dx) \tag{3.35}$$

where $K_H^{(2)} : x-y \mapsto \int K_H((x-y)-z)K_H(z)dz$ is the convolution of K_H with itself, is an unbiased estimator of $J(H)$ [14].

Define $\hat{H} := \arg\min_H \hat{J}(H)$ and suppose $\mathscr{H}$ minimizes Eq. (3.34). By Stone's Theorem [70], for non-negative, Holder continuous K_H attaining their maximums at $(0,0)$, and bounded p, we have

$$\frac{\|p-\hat{p}_{\hat{H}}\|}{\|p-\hat{p}_{\mathscr{H}}\|} \to 1 \ \ a.s. \tag{3.36}$$

as $O \to \infty$. In Proposition 3.1.1, we have that λ_{D_X} is given by a modified mixed Gaussian,

$$\lambda_{D_X}(x) = \sum_{j=1}^{N} C_j^{D_X} \mathcal{N}^*(x; \mu_j^{D_X}, \sigma_j^{D_X}), \tag{3.37}$$

and the stochastic kernel $\ell(y|x)$ is given by

$$\ell(y|x) = \mathcal{N}^*(y; x, \sigma^{D_{Y_O}}). \tag{3.38}$$

To use results Equation (3.35) in the selection of the bandwidth of $\ell(y|x)$, we must express Equation (3.31) in a form reminiscent of Equation (3.32). A useful setting in which this is possible is when $\alpha = 1$, the prior intensity λ_{D_X} is unimodal, the noise intensity $\lambda_{D_{Y_S}}$ is identically zero, and the covariance magnitude σ^{D_X} of the prior is much greater than the covariance magnitude of the stochastic kernel. Such a scenario arises when one uses unimodal uninformative priors and considers all points in observed persistence diagrams to encode relevant geometric information, i.e. when one computes posteriors solely from an empirical standpoint.

Henceforth, assume that in Equation (3.31), $\alpha = 1$, the prior intensity λ_{D_X} is unimodal, namely $\lambda_{D_X} = c^{D_X}\mathcal{N}^*(x; \mu^{D_X}, \sigma^{D_X})$,the noise intensity $\lambda_{D_{Y_S}}$ is identically zero, and $\sigma^{D_X} >> \sigma^{D_{Y_O}}$. This simplifies Equation (3.31) to

$$\lambda_{D_X|\mathcal{D}_Y^{1:m}}(x) = \frac{1}{m}\sum_{i=1}^{m}\sum_{y\in\mathcal{D}_{Y^i}} \frac{\mathcal{N}^*\left(x; y, \frac{\sigma^{D_{Y_O}}\sigma^{D_X}}{\sigma^{D_{Y_O}}+\sigma^{D_X}}\right)}{Q^y} \tag{3.39}$$

where $Q_j^y = \int_{\mathbb{W}} \mathcal{N}(u; \mu_j^y, \sigma_j^y)du$. After restricting each kernel in Equation (3.32) to $\mathbb{W}$ then rescaling to ensure unit mass, Equation (3.39) is equivalent to Equation (3.32) with a Gaussian kernel having bandwidth matrix $\frac{\sigma^{D_{Y_O}}\sigma^{D_X}}{\sigma^{D_{Y_O}}+\sigma^{D_X}}I$ and discrete measures $m_i(dx) = \sum_{y\in D_{Y^i}} \delta_y(dx)$ where δ_y is the Dirac delta function for y. Hence, an optimal choice for $\sigma^{D_{Y_O}}$, say $\sigma_{opt}^{D_{Y_O}}$, follows from requiring H in Equation (3.34) to be of the form hI for some positive constant h, then solving $\frac{\sigma^{D_{Y_O}}\sigma^{D_X}}{\sigma^{D_{Y_O}}+\sigma^{D_X}}I = \hat{H}$. In particular,

$$\sigma_{opt}^{D_{Y_O}} = \frac{\hat{h}\sigma^{D_{X_O}}}{\sigma^{D_{X_O}} - \hat{h}} \tag{3.40}$$

where $\hat{h}$ is unique diagonal entry of $\hat{H}$. Our assumptions about K_H and m_i allow us to express Eqn. (3.35) in a computational tractable form. Namely, for a set of observed diagrams $\mathcal{D}_{Y^1}, \ldots, \mathcal{D}_{Y^m}$ and $H = hI$, write

$$\hat{p}_{D_{Y^i}}^H(x) = \frac{1}{m-1}\sum_{D_{Y^j}\neq D_{Y^i}} \left(\sum_{y\in D_{Y^j}} \mathcal{N}(x; y, H)\right). \tag{3.41}$$

Then,

$$\hat{J}(H) = \frac{1}{m^2}\sum_{D_{Y^i}, D_{Y^j}} \left[\sum_{y\in\mathcal{D}_{Y^i}, y'\in\mathcal{D}_{Y^j}} \mathcal{N}\left(y - y'; 0, \sqrt{2}H\right)\right] - \frac{2}{m}\sum_{\mathcal{D}_{Y^i}} \left[\sum_{x\in\mathcal{D}_{Y^i}} \hat{p}_{D_{Y^i}}^H(x)\right] \tag{3.42}$$

In practice, one can compute $\sigma_{opt}^{D_{Y_O}}$ using Equation (3.42).

3.3 Classification

The Bayesian framework introduced in this paper allows us to explicitly compute the posterior intensity of a persistence diagram given data and prior knowledge. This lays the foundation for supervised statistical learning methods in classification. In this section, we build a Bayes factor classification algorithm based on notions discussed in Section 3.1 and then apply it on materials science data, in particular, on measurements for spatial configurations of atoms.

We commence our classification scheme with a persistence diagram $\mathcal{D}$ belonging to an unknown class. We assume that $\mathcal{D}$ is sampled from a Poisson point process D in $\mathbb{W}$, with the prior intensity λ_D having the form in (M2′). Consequently, its probability density has the form

$$p_D(\mathcal{D}) = \frac{e^{-\lambda}}{|\mathcal{D}|!} \prod_{d \in \mathcal{D}} \lambda_D(d) = \frac{e^{-\lambda}}{|\mathcal{D}|!} \prod_{d \in \mathcal{D}} \sum_{i=1}^{N} c_i^D \mathcal{N}^*(d; \mu_i^D, \sigma_i^D I), \tag{3.43}$$

where $\lambda = \int_{\mathbb{W}} \lambda_D = \mathbb{E}(|D|)$. Next, suppose we have two training sets $T_Y := \mathcal{D}_{Y_{1:n}}$ and $T_{Y'} := \mathcal{D}_{Y'_{1:m}}$ from two classes of random diagrams D_Y and $D_{Y'}$, respectively. The likelihood densities of respective classes take the form of Equation (3.29).

We then follow Equation (3.31) to obtain the posterior intensities of D given the training sets T_Y and $T_{Y'}$ from the prior intensities and likelihood densities.

In particular, the corresponding posterior probability density of D given the training set T_Y is

$$\begin{aligned} p_{D|\mathcal{D}_Y}(\mathcal{D}|T_Y) &= \frac{e^{-\lambda}}{|\mathcal{D}|!} \prod_{d \in \mathcal{D}} \lambda_{D|T_Y}(d) \\ &= \frac{e^{-\lambda}}{|\mathcal{D}|!} \prod_{d \in \mathcal{D}} \Big[(1-\alpha)\lambda_{\mathcal{D}}(d) + \frac{\alpha}{n} \sum_{y_j \in T_Y} \sum_{i=1}^{N} C_i^{d|y_j} \mathcal{N}(d; \mu_i^{d|y_j}, \sigma_i^{d|y_j} I) \Big], \end{aligned} \tag{3.44}$$

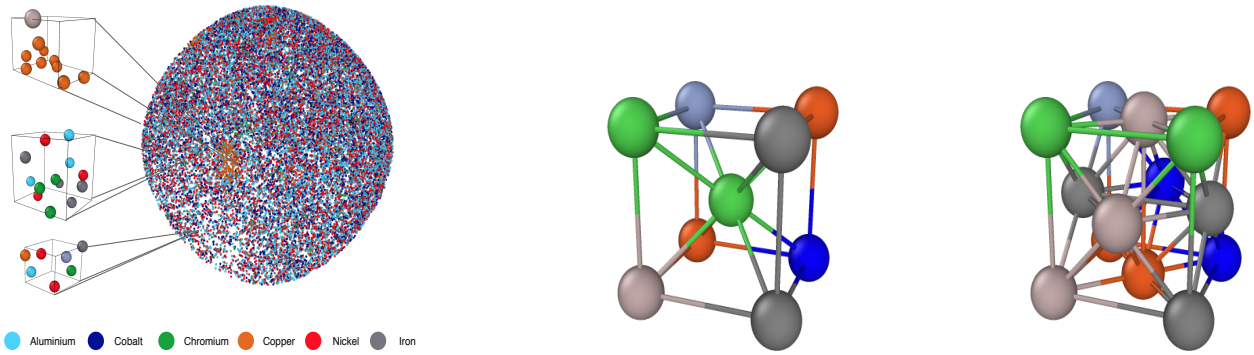
and the posterior probability density given T'_Y is given by an analogous expression. The Bayes factor defined by

$$BF(\mathcal{D}) = \frac{p_{D|\mathcal{D}_Y}(\mathcal{D}|T_Y)}{p_{D|\mathcal{D}_{Y'}}(\mathcal{D}|T_{Y'})} \tag{3.45}$$

provides the decision criterion for assigning $\mathcal{D}$ to either D_Y or $D_{Y'}$. More specifically, for a threshold c, $BF(\mathcal{D}) > c$ implies that $\mathcal{D}$ belongs to D_Y and $BF(\mathcal{D}) < c$ implies otherwise. We summarize this scheme in Algorithm 1 .

Algorithm 1 Bayes Factor Classification of Persistence Diagrams

1: **Input 1**: Prior intensities λ_{D_Y}, and $\lambda_{D_{Y'}}$ for two classes of diagrams D_Y and $D_{Y'}$ respectively; a threshold $c > 0$.
2: **Input 2**: Two training sets T_Y and $T_{Y'}$ sampled from D_Y and $D_{Y'}$, respectively.
3: **for** D_Y and, $D_{Y'}$ **do**
4: Compute $p_{D|\mathcal{D}_Y}(\mathcal{D}|T_Y)$ and $p_{D|\mathcal{D}_{Y'}}(\mathcal{D}|T_{Y'})$.
5: **end for**
6: Compute $BF(\mathcal{D})$ as in Equation (3.45)
7: **if** $BF(\mathcal{D}) > c$ **then**
8: assign $\mathcal{D}$ to D_Y.
9: **else**
10: assign $\mathcal{D}$ to $D_{Y'}$.
11: **end if**



(a) (b) (c)

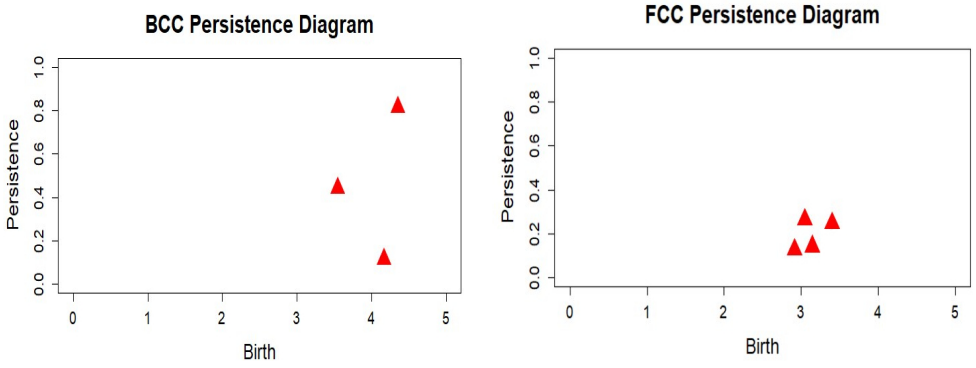
Figure 3.6: (a) Image of APT data with atomic neighborhoods shown in detail on the left. Each pixel represents a different atom, the neighborhood of which is considered. Certain patterns with distinct crystal structures exist, e.g., the orange region is copper-rich (left), but overall no pattern is identified. Putting a single atomic cubic unit cell under a microscope, the true crystal structure of the material, which could be either body-centered cubic (BCC) (b) or face-centered cubic (FCC) (c), is not revealed. This distinction is obscured due to further experimental noise. Notice there is an essential topological difference between the two structures in (b) and (c): The BCC structure has one atom at its center, whereas the FCC is hollow in its center, but has one atom in the center of each of its faces.

3.3.1 Atom Probe Tomography Data

A crucial first step in understanding properties of a material is determining its crystal structure. For highly disordered metallic alloys, such as High Entropy Alloys (HEAs), Atomic Probe Tomography (APT) gives a snapshot of the local atomic environment. APT has two main drawbacks: experimental noise and missing data. Approximately 67% of the atoms in a sample are not registered in a typical experiment, and those atoms that are captured have their spatial coordinates corrupted by experimental noise. Our goal in this section is to use Algorithm 1 to classify the crystal lattice of a noisy and sparse materials dataset, where the unit cells are either Body centered cubic (BCC) or Face centered cubic (FCC); recall Figure 3.6. The BCC structure has a single atom in the center of the cube, while the FCC has a void in its center but has atoms on the centers of the cubes' faces (Figure 3.6 (b-c)). Despite notable differences in the physical configurations of each class, sparsity and noise do not allow the crystal structure to be revealed.

For high-entropy alloys, our object of interest, APT, provides the best atomic level characterization possible. Due to the sparsity and noise in the resulting data, there are only a few algorithms for successfully determining the crystal structure; see [25, 50]. These algorithms, designed for APT data, rely on knowing the global structure *a priori* (which is not the case for High Entropy Alloys (HEAs)) and seek to discover small-scale structure within a sample.

To bypass this restriction, the neural network architecture of [77] provides a way to classify the crystal structure of a noisy or sparse dataset by looking at a diffraction image. In particular, the authors therein employ a convolutional neural network for classifying the crystal structure by examining a computer-generated diffraction pattern. The authors suggest their method could be used to determine the crystal structure of APT data. However, the synthetic data considered in



(a) (b)

Figure 3.7: Persistence diagrams for members of the BCC and FCC classes.

[77] is not a realistic representation of experimental APT data, where about 65% of the data is missing and furthermore corrupted by observational noise. Most importantly, their synthetic data is either sparse or noisy, not a combination of both. The algorithm is also not publicly available, so a side by side comparison of our method with theirs using HEAs is not feasible.

It is natural to consider persistence diagrams in this setting because they distill salient information about the materials patterns with respect to connectedness and empty space (holes) within cubic unit cells, i.e we can differentiate between atomic unit cells by examining their homological features. In particular, after storing both types of spatial configurations as point clouds, we compute their Vietoris-Rips filtrations (see Section 2.1.1), collecting resultant 1-dimensional homological features into persistence diagrams; see Figure 3.7. The data set had 200 diagrams from each class. To perform classification with Algorithm 1, we started by specifying priors for each class, $\lambda_{D_{BCC}}$ and $\lambda_{D_{FCC}}$. Two scenarios were considered, namely using separate priors (Prior-1 in Table 3.5) and the same prior (Prior-2 in Table 3.5) for both BCC and FCC classes. In particular, for Prior-1 we superimpose 50 persistence diagrams from each class and find the highly clustered areas by using K-means clustering. The centers of the clusters from K-means are then used as the means in Gaussian mixture priors; see Equation (3.31). In this manner, we produce different priors for BCC and FCC classes. On the other hand for Prior-2 we choose a flat prior with higher variance level than that of Prior-1 for both of the classes. The parameters for these two prior intensities are in Table 3.5. For all cases, we set $\sigma^{D_{Y_O}} = 0.1$ and $\lambda_{D_{Y_S}}(x) = 5\mathcal{N}^*(x; (0,0), 0.2I)$. We chose a relatively high weight for $\lambda_{D_{Y_S}}$ because the nature of the data implied that extremely low persistence holes were rare events arising from noise. To perform 10-fold cross validation, we partitioned persistence diagrams from both classes into training and test sets. During each fold, we took the training sets from each class, T_{BCC} and T_{FCC}, and input them into Algorithm 1 as T_Y and $T_{Y'}$, respectively. Next, we computed the Bayes factor $BF(\mathcal{D}) = \frac{p_{D_{BCC}}(\mathcal{D}|T_{BCC})}{p_{D_{FCC}}(\mathcal{D}|T_{FCC})}$ for each diagram $\mathcal{D}$ in the test sets. After this, we used the Bayes factors to construct receiver operating characteristic (ROC) curves and computed the resulting areas under the ROC curves (AUCs). Finally, we used the AUCs from 10-fold cross validation to build a bootstrapped distribution by resampling 2000 times. Information about these bootstrapped distributions is summarized in Table 3.5, which shows our scoring method almost perfectly distinguishes between the BCC and FCC classes using

the Bayesian framework of Section 3.1. Also, it exemplifies the robustness of our algorithm as two different types of priors produce near perfect accuracy.

Table 3.5: Parameters for the prior intensities used in cross-validation of materials science data. Each prior λ is indexed by its corresponding class for Prior-1 or U in the case of the Prior-2. The summary of AUCs across 10-folds for materials science data after scoring with Algorithm 1 is presented in the last three columns.

Priors	**Parameters for Prior Intensities**				**Summary of AUC**		
		μ_i^D	σ_i^D	c_i^D	5th percentile	Mean	95th percentile
Prior-1	λ_{BCC}	(0.5,0.24) (3.6,3.6) (3.7,0.65)	2 2 2	1 1 1	0.931	0.941	0.958
	λ_{BCC}	(0.4,0.27) (2.8,1.2) (2.9,3)	2 2 2	1 1 1			
Prior-2	λ_U	(1,1)	20	1	0.928	0.94	0.951

3.4 Bayesian Topological Signal Processing

Data from measurements sampled in time create time series, which are useful descriptors of natural phenomena and artificially-created signals alike. Time series analysis is performed to extract meaningful statistics and other characteristics from data [56]. Frequency domain methods are particularly suited for data exhibiting oscillatory behavior. A commonly used tool for nonparametric spectral analysis is the discrete Fourier transform [63]. Alternatively, parametric methods are frequently used to estimate the power spectral density of stochastic, stationary signals [60]. For non-stationary signals, time-frequency methods, e.g. wavelets [61] or matching pursuits [19, 41], provide useful characterizations of different time series. One of the important applications of time series analysis is in neuroscience and neurology where measurements of electrical activity in the brain (electroencephalography, EEG) can provide information about brain function and serves as an important clinical diagnostic tool. In Section 3.4, we present a novel approach to time series analysis using persistent homology within a Bayesian framework. We illustrate the method on data generated by an autoregressive model simulated frequency characteristics of EEG signals [23].

Persistent homology has been used for signal processing in many applications, including detection of periodic behavior, change point monitoring in genetic regulatory systems, detection of bifurcation in stochastic delayed differential systems, acoustic signal identification, and prediction of financial crashes from time series of stock returns [26, 59, 32, 35, 64, 44]. These applications are bolstered by numerous theoretical results concerning the homology of simplicial complexes built on point clouds obtained from delay embeddings of time series [69, 59, 35]. One may also create persistence diagrams for signals by using sublevel set filtrations; see Section 2.1.2. This method is attractive for signal processing as it may be used directly on signals to summarize their shape without selection of a delay parameter; however, a framework for inference about qualities such as frequency and instantaneous amplitude of a signal from its sublevel set persistence diagram remains

wanting. In Section 3.4, we examine links between frequency content and instantaneous amplitude of signals and their sublevel set persistence diagrams, then explore our Bayesian method for inference with these topological objects. We also investigate the capabilities of the Bayesian topological model 3.1.1 to describe signals by testing it to classify autoregressive signals. We compare this method to other model-free approaches for signal classification, specifically those based on feature extraction and machine learning architectures in [29] and [5].

3.4.1 Signal Processing Preliminaries

Throughout this section, $x(t)$ denotes a real-valued, continuous time signal. We suppose x is continuous to avoid complications when we introduce topological concepts. Unless otherwise noted, time t is given in seconds and units of frequency are in Hz. In general, x may be a sample from a stochastic process, which we denote by X. For applications, one often uses a discretized approximation to x obtained by fixing a sampling frequency $\Delta t > 0$ then measuring x at regular intervals. We denote elements of the discretized collection $\{x(n\Delta t)\}_{n=1}^{N}$ by x_n where N is the total number of grid points.

We now recall concepts and tools from signal processing employed in Section 3.4.3. During our analyses, we investigate autoregressive signals satisfying the $AR(p)$ model $x_n = \sum_{i=1}^{p} a_i x_{n-i} + w_n$ where p is the model order, a_i are real-valued coefficients, and w_n are independent, identically distributed samples from a Gaussian random variable $\mathcal{N}(0, \sigma)$ with standard deviation σ. This autoregressive model is used in various applications across many fields with an easily computable power spectral density that is described by finitely many parameters β_k and f_k,

$$\log p_X(f) = 2\left(-\sum_{k=1}^{P} \log \left| 1 - e^{-\beta_k + 2\pi i (f_k - f)\Delta t} \right| + \log \sigma \right), \tag{3.46}$$

where σ is the standard deviation of w_n in $A(p)$. A derivation of Equation (3.46) may be found in [23]. Through inspection of Equation (3.46), one surmises that the f_k are locations of local maxima in the log PSD (and therefore also local maxima of the PSD) and the β_k are corresponding damping factors controlling the width of each peak. Namely, small and large β_k are associated with narrow and broad peaks, respectively. As a peak broadens, the presence of oscillations of its corresponding frequency diminish in the average signal. Therefore, we expect power spectral densities with narrow peaks to yield signals with stronger oscillations.

3.4.2 Topological Uncertainty Quantification for Random Signals

In practice, we observe random signals. Uncertainty in signals arises independently from stochasticity in underlying data-generating phenomena and measurement noise due to limitations in data-collection methods. The latter source of uncertainty is pictured in Figure 3.8 where a signal is embedded in different levels of white Gaussian noise. The persistent diagrams created from stochastic signals (subsequently referred to as random persistence diagrams) inherit randomness, necessitating a probabilisitic description. Intuitively, a random persistence diagram is a random collection of points in a subset of the plane. Elements of a persistence diagram lack an intrinsic ordering, and moreover persistence diagrams sampled from the same random signal can have different cardinalities. The theory of random variables is ill-equipped to handle such objects directly since it concerns itself with random elements that take values in a Hilbert space. However, the theory of point processes rigorously treats random collections like random persistence diagrams and provides machinery to model them.

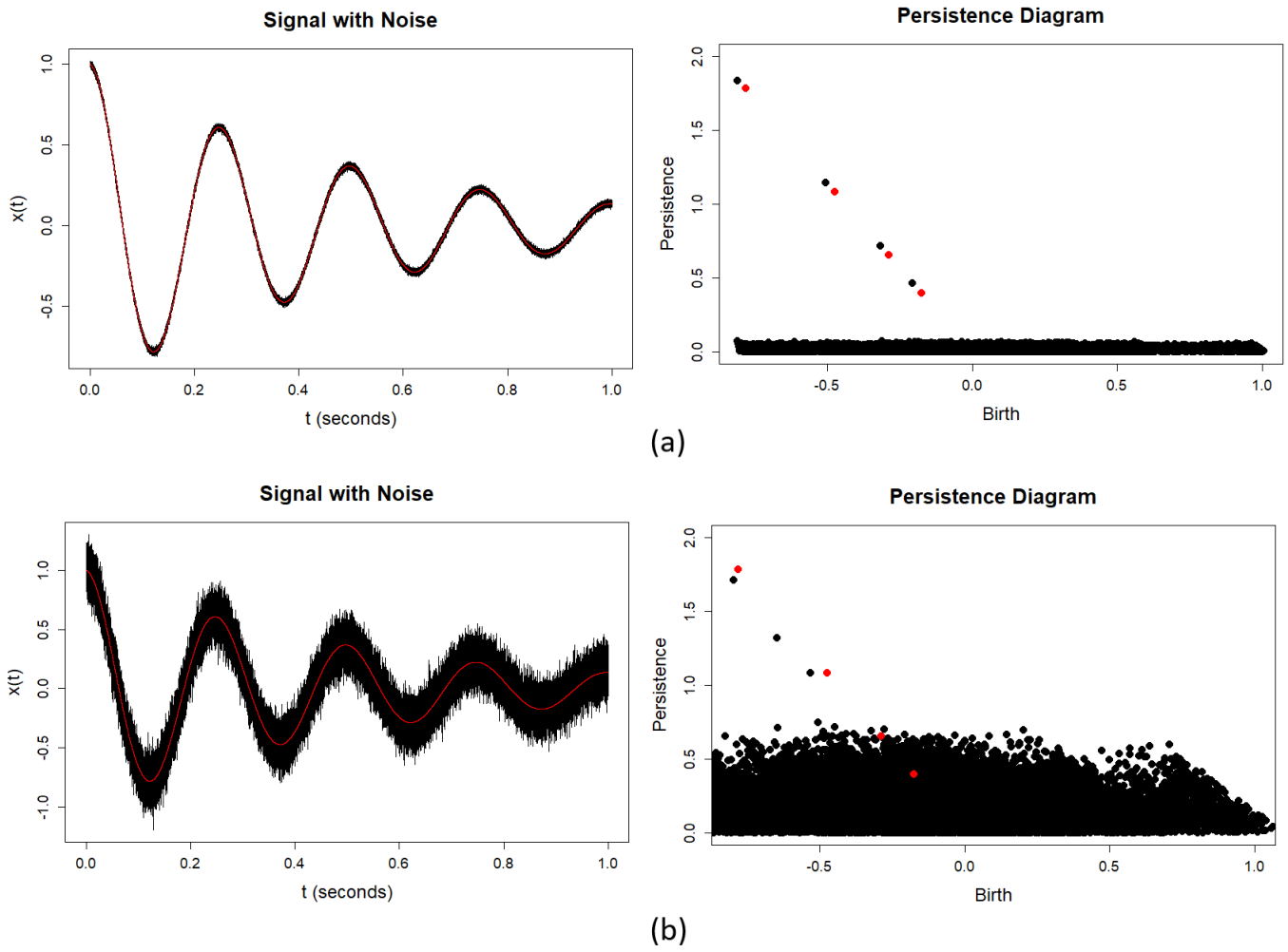


Figure 3.8: This figure illustrates sources of uncertainty in persistence diagrams. Shown above are signals with additive noise (a) $\mathcal{N}(0, 0.01)$, and (b) $\mathcal{N}(0, 0.1)$ along with their persistence diagrams. The persistence diagram for the true underlying signal is shown in red. Several spurious features arise due to noise. Additionally, true features also shift around.

The persistence diagram of a signal encodes information about its frequency and peak-to-peak amplitude. This is exemplified in Proposition 3.4.1 and Corollary 3.4.1 below.

Proposition 3.4.1. *Let $a_\beta(t)$ be a monotonically non-increasing function of time parameterized by a real number β and suppose that $\phi(2\pi ft)$ is a periodic signal parameterized by a positive frequency f such that $\phi(0)$ is a local maximum and each cycle of ϕ has unique local minimum. The persistence diagram $\mathcal{D}_{f,\beta}$ of $a_\beta(t)\phi(2\pi ft)$ on the interval [0,1] is given by $\mathcal{D}_{f,\beta} = \left\{\left(a_\beta(t_i + t_m)\phi(2\pi f(t_i + t_m)), a_\beta(t_i)\phi(2\pi f t_i)\right)\right\}_{i=1}^{\lfloor f \rfloor}$ where, $\lfloor f \rfloor$ denotes the integer part of f, $t_i = \frac{i-1}{f}$ and t_m is the time of the first local minimum of $a_\beta(t)\phi(2\pi ft)$ in $[0, 1]$.*

Proof. Without loss of generality, suppose f is an integer. The case for real-valued f is similar. By

definition, $a_\beta(t)\phi(2\pi ft)$ has f cycles in $[0,1]$. Let M_i and m_i respectively denote the maximum and local minimum of the ith cycle of $a_\beta(t)\phi(2\pi ft)$ in $[0,1]$. Since each cycle has a unique local minimum, the only connected components that arise during the sublevel set filtration of $a_\beta(t)\phi(2\pi ft)$ are C_{m_i}, which are clearly born at values m_i of the sublevel set filtration. It remains to deduce the value at which each C_{m_i} disappears. From the assumption that $\phi(0)$ is a local maximum, each cycle begins at a local maximum and monotonicity of $a_\beta(t)$ ensures that M_i occurs at the beginning of the ith cycle and at the end of the $(i-1)$th cycle. Recall that connected components merge at values M_i during the sublevel set filtration. By monotonicity of a_β, the smallest of the M_i (and hence the first one encountered in the sublevel set filtration) is M_f. This merges connected components C_{m_f} and $C_{m_{f-1}}$. Once again by monotonicity of $a_\beta(t)$, $m_f \leq m_{f-1}$, so by the Elder Rule, C_{m_f} disappears by merging into $C_{m_{f-1}}$. Thus, the first point in $D^{f,\beta}$ is (m_f, M_f). Using the monotonicity of $a_\beta(t)$ to continue in this fashion, we see that $\mathcal{D}^{f,\beta} = \{(m_i, M_i)\}_{i=1}^{f}$. The result now follows by explicitly computing m_i and M_i, which is done by exploiting the periodicity of ϕ. □

Example 3.4.1. Consider the family of damped cosines $s(t; f, \beta) := e^{-\beta t}\cos(2\pi ft)$ parameterized by frequency f and a damping factor β on the interval $[0,1]$. The parameter β controls the rate at which the instantaneous amplitude of s decreases. For simplicity, we suppose without loss of generality that f is a positive integer. Let $\mathcal{D}^{f,\beta}$ be the persistence diagram for $s(t; f, \beta)$. Figure 2.4 shows $s(t; 4, 2)$ and $\mathcal{D}^{4,2}$. Notice in general that $s(t; f, \beta)$ has f local minima (ignoring units of f) on the interval $[0,1]$. As the number of points in a persistence diagram of signal is equal to the number of its local minima, we immediately conclude that $|\mathcal{D}^{f,\beta}| = f$, where $|\mathcal{D}^{f,\beta}|$ means the cardinality of $\mathcal{D}^{f,\beta}$. To compute the coordinates of points in $\mathcal{D}^{f,\beta}$, notice that $e^{-\beta t}$ is monotonically decreasing while each oscillation of $\cos(2\pi ft)$ starts at a local maximum and has a unique local minimum. The latter implies that each oscillation gives rise to a connected component during the sublevel set filtration of s and that this connected component merges into another when the filtration reaches the value of the local maximum where the oscillation begins. The former implies that connected components for oscillations occurring later in time disappear sooner than those born earlier due to the Elder Rule. Hence, the coordinates of the persistence diagram are given by $\{(m_i, M_i)\}_{i=1}^{f}$, where m_i and M_i denote the local minimum and maximum of the ith oscillation. Explicitly, these are given by $(m_i, M_i) = \left(e^{-\beta(t_i+\frac{1}{2f})}\cos\left(2\pi f(t_i+\frac{1}{2f})\right), e^{-\beta t_i}\cos\left(2\pi f t_i\right)\right)$ where $t_i := \frac{i-1}{f}$.

Proposition 3.4.1 establishes direct links between frequencies and instantaneous amplitudes to the persistence diagrams for a large class of deterministic signals that naturally occur in applications. One may suspect that these relationships manifest in stochastic settings. Figure 3.8(a) shows the signal $s(t; 4, 2)$ with additive Gaussian white noise and its persistence diagram. Although the presence of noise introduces several spurious low persistence features in the persistence diagram, the relationship unveiled in Proposition 3.4.1 is evident in the higher persistence points. In this case, noticeable relationships still exist between the frequency and instantaneous amplitude of the deterministic signal embedded in noise and its persistence diagram. The next proposition and corollary concern signals that are entirely stochastic. They serve to illustrate the complex relationship between frequency content and instantaneous amplitude to the persistence diagrams of random signals. To motivate the importance of the next proposition, suppose x is sampled from a stochastic process and consider its discretization $\{x_n\}_{n=0}^{N}$, which we momentarily refer to as x by a slight abuse of notation. Ideally, the probability distribution $\mathbb{P}(D^x)$ for the random persistence diagram of D^x, would be expressed as $\mathbb{P}(D^x) = \cup_{i,j\in\mathcal{I}}\{\mathbb{P}(x_i, x_j)\}$ for a fixed set of index pairs $\mathcal{I}$. In this situation, D^x is a fixed union of joint random variables. Unfortunately, the aforementioned decomposition of $\mathbb{P}(D^x)$ does not hold in general, specifically because distinct orderings of values of

elements in x are associated to different minimum-maximum pairings. The next lemma establishes events $\mathcal{E}$ in which $\mathbb{P}(D^x|\mathcal{E}) = \cup_{i,j\in\mathcal{I}}\{\mathbb{P}(x_i, x_j)\}$ holds.

Proposition 3.4.2. *Let x be a signal randomly sampled from a stochastic process X. Consider the time series $\{x_n\}_{n=1}^L$ created by measuring x at regular intervals on $[0,1]$. Denote the persistence diagram for $\{x_n\}_{n=1}^L$ by $\mathcal{D}^x$. Let Π_L be the set of permutations of $\{1,2,\ldots,L\}$ and suppose for a fixed element π of Π, say $\pi = \{i_1, i_2, \ldots, i_L\}$, O_π is the event that $\{i_1 < i_2 < \cdots < i_L\}$. $D^x|O_\pi$ is a fixed union of joint random variables; we denote the probability distribution for this collection by $\mathbb{P}(D^x|O_\pi)$.*

Proof. The event O_π corresponds to an exact ordering of the values of elements in x. The exact ordering preserves the minimum-maximum associations used to construct sublevel sets to create D^x. In particular, $D^x = \cup_{i,j\in\mathcal{I}_\pi}\{(x_i, x_j)\}$ for a fixed set of index pairs $\mathcal{I}_\pi$ given O_π. This establishes the claim. □

Corollary 3.4.1. *Let x be a signal randomly sampled from a stochastic process X. Consider the time series $\{x_n\}_{n=1}^L$ created by measuring x at regular intervals on $[0,1]$. Denote the persistence diagram for $\{x_n\}_{n=1}^L$ by $\mathcal{D}^x$. Let Π_L be the set of permutations of $\{1,2,\ldots,L\}$ and suppose for a fixed element π of Π, say $\pi = \{i_1, i_2, \ldots, i_L\}$, O_π is the event that $\{i_1 < i_2 < \cdots < i_L\}$. The probability distribution for $\mathcal{D}^x$ is given by*

$$\mathbb{P}(\mathcal{D}^x) = \sum_{\pi\in\Pi_L} \mathbb{P}(\mathcal{D}^x|O_\pi)\mathbb{1}_{O_\pi}(\mathcal{D}^x). \tag{3.47}$$

where $\mathbb{P}(\mathcal{D}^x|O_\pi)$ is the distribution for a random collection of points given the ordering O_π.

Let Π^k be the subset of Π such that each element of Π^k is an event with k local minima. Then, the cardinality distribution of $\mathcal{D}^x$ is given by

$$\mathbb{P}(|\mathcal{D}^x| = k) = \sum_{\pi\in\Pi^k} \mathbb{P}(\pi). \tag{3.48}$$

Proof. Equation (3.47) follows from Proposition 3.4.2 by the law of total probability and Equation (3.50) is established by a straightforward counting argument. □

Example 3.4.2. Let x be a signal randomly sampled from Gaussian white noise X. Consider the time series $\{x_n\}_{n=1}^4$ created by measuring x at regular intervals on $[0,1]$. We are interested in the cardinality of $\mathcal{D}^x$ as well as the coordinates of its points. Since $\mathcal{D}^x$ inherits stochasticity from x, these quantities are specified by probability distributions. By Proposition 3.4.1, the probability distribution for $\mathcal{D}^x$ is given by

$$\mathbb{P}(\mathcal{D}^x) = \sum_{\pi_{i,j,k,l}\in\Pi_4} \mathbb{P}(\mathcal{D}^x|O_\pi)\mathbb{1}_{O_\pi}(\mathcal{D}^x). \tag{3.49}$$

It is helpful to write one of the terms in the summation of Equation (3.49). Consider the term that corresponds to the event $O_{1,3,2,4} = \{x_1 < x_3 < x_2 < x_4\}$. For this event, the sublevel set filtration of $\{x_n\}_{n=1}^4$ has connected components born at x_1 and x_3. These merge when the filtration reaches x_2, at which point x_3 disappears due to the Elder Rule. The filtration terminates at x_4. Hence, the persistence diagram for this event is $\{(x_1, x_4), (x_2, x_3)\}$. Each of these elements are random samples from joint random variables satisfying $(x_1, x_4) \sim \mathbb{P}(x_1, x_4|x_1 < x_4)$ and $(x_2, x_3) \sim \mathbb{P}(x_2, x_3|x_2 < x_3)$. We denote this by $S(\mathcal{D}^x|O_{1,3,2,4}) = \{\mathbb{P}(x_1, x_4|x_1 < x_4), \mathbb{P}(x_2, x_3|x_2 < x_3)\}$. For the case of white

noise, we have $\mathbb{P}(x_1, x_4 | x_1 < x_4) = Ke^{\frac{-x_1^2}{2}} e^{\frac{-x_4^2}{2}} \mathbb{1}_{x_1<x_4}(x_1, x_4)$ where K is a normalizing constant; the equation for $\mathbb{P}(x_2, x_3 | x_2 < x_3)$ is similar.

Notice that the events in Π_L each correspond to signals that have a fixed number of local minima. Let Π^k be the subset of Π such that each element of Π^k is an event with k local minima. Then, the cardinality distribution of $\mathcal{D}^x$ is given by

$$\mathbb{P}(|\mathcal{D}^x| = k) = \sum_{\pi \in \Pi^k} \mathbb{P}(\pi). \tag{3.50}$$

Intuitively, the frequency content and distribution for instantaneous amplitudes of a random signal x influence the probability of events in Π in Corollary 3.4.1. In general, describing the persistence diagram in its entirety for a time series of length L arising from a random signal requires one to construct $L!$ different probability distributions. We circumvent this problem by using the flexible Bayesian framework described in Section 3.1.1 to approximate distributions for persistence diagrams. With this tool in hand, we can model the distributions of persistence diagrams in a computational expedient manner. Figure 3.9 visually depicts the intensity in Equation (3.31) using a persistence diagram from a simulated signal.

3.4.3 Signal Processing Results

Electroencephalography is a neuroimaging technique wherein electrodes are placed on a subject's head to measure localized changes in voltage over time, which are reported as a collection of time series. In our experiments, we examine synthetic EEG signals generated according to Equation (3.46). We select EEG signals because they have well-studied power spectra. In particular, the log power spectral density for EEG is approximately inversely related to frequency, a phenomena we subsequently refer to as $1/f$ behavior. We can easily simulate this behavior with Equation (3.46) by including a frequency component f_1 equal to zero. Moreover, EEG signals associated with different brain states often exhibit a prominent peak in their power spectral density at a nonzero frequency, indicating the discernible presence of oscillations at that frequency in the signals. For example EEG signals with peaks in the $4-7$ Hz, $8-12$ Hz, or $14-32$ Hz ranges, referred to as theta, alpha, and beta signals, respectively, may indicate heightened emotional states, eyes open, or drowsiness [1, 6]. Oscillatory behavior can also be accounted for with Equation (3.46) by including a particular nonzero frequency component f_2. Selection of the damping factor β_1 for the zero frequency component f_1, as well as ranges for f_2 and β_2 (the location of the oscillatory component in the PSD and its accompanying damping factor) was done by using the Burg method to fit autoregressive models to real EEG signals (model order was determined using the methodology from [23]). Two one second and two five second epochs were selected for their visually apparent oscillations. The exact location of spectral peaks along with their corresponding damping factors as determined after fitting autoregressive models are shown in Table 3.6.

Informed by the parameters in Table 3.6, we created 29×29 sets of synthetic EEG signals $\mathcal{A}_f^\beta$ where $f \in \{4, 5, \ldots, 32\}$ and $\beta \in \{4, 5, \ldots, 32\}$. A fixed set $\mathcal{A}_f^\beta$ contained 30 signals simulated by our autoregressive filter with $f_1 = 0$, $f_2 = f$, $\beta_1 = 200$, and $\beta_2 = \beta$. The signals in $\mathcal{A}_f^\beta$ are draws from a stochastic process whose PSD has a peak at f with a damping factor of β, so we expect these signals to resemble an oscillator of frequency f; the strength of this resemblance diminishes as β increases. Since the PSD for $\mathcal{A}_f^\beta$ more closely resembles $1/f$ as β increases, we expect more low-power, high frequency signals riding on high-power, low frequency signals as β decreases. Moreover, the ratio of the peak at zero to that of the peak at f increases as β increases.

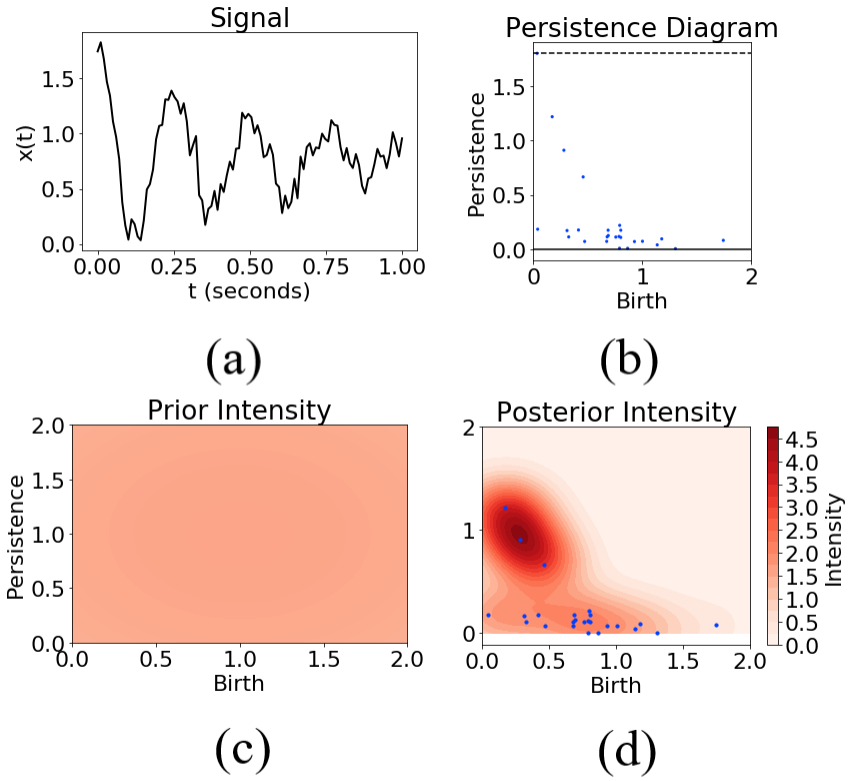


Figure 3.9: (a) The damped cosine $e^{-2t}\cos(8\pi t)$ with additive noise $\mathcal{N}(0, 0.01)$ and (b) its persistence diagram. (b) shows an uninformative prior intensity with a single component at $(1, 1)$ with covariance matrix $10I$. Using the model from Equation (3.31) with the prior in (c) and the observed diagram in (b) results in the posterior intensity shown in (d). To account for spurious points, which we suspected to be low persistence in this example, we placed components at $(0.5, 0.1), (1, 0.1), (0.75, 0.1)$ and $(1.75, 0.1)$.

Table 3.6: Parameters from fitting autoregressive models to real EEG epochs.

Signal Length	1 Second								5 Seconds							
	f_1	f_2	f_3	f_4	β_1	β_2	β_3	β_4	f_1	f_2	f_3	f_4	β_1	β_2	β_3	β_4
Signal 1	0	5.87	18.59	-	344.80	5.37	16.6	-	0	6.00	14.4	20.85	24.98	10.54	31.64	26.97
Signal 2	0	10.70	-	-	202.78	7.41	-	-	0	10.16	23.02	-	17.24	4.06	20.37	-

Relationships between the Frequency and Diagram Domains

In this section, we examine relationships between f and β and the persistence diagrams of $\mathcal{A}_f^\beta$. Specifically, we look at how f and β relate to the average cardinality and variance in birth values for persistence diagrams of signals in $\mathcal{A}_f^\beta$. As in the deterministic setting (Proposition 3.4.1), we expect cardinality to show a strong association with the peak frequency f for $\mathcal{A}_f^\beta$. Our choice to inspect birth time variance is motivated by the following thought experiment.

Figure 3.10 (a) and (b) show the persistence diagrams for $10\cos(4\pi t)+\cos(58\pi t)$ and $\cos(4\pi t)+\cos(58\pi t)$ on the interval $[0,1]$, respectively. In both subfigures, the points near the diagonal correspond to the small peaks that ride along the low frequency wave. Alternatively, the two points farthest from the diagonal correspond to the global maxima and minima, which correspond to those of the low frequency signal. In this fashion, the persistence diagram spatially decomposes frequencies where high and low persistence points correspond to features of low and high frequency components, respectively. As can be seen in Figure 3.10 (a), greater low frequency power causes a wider spread in the birth times of the low persistence points. Moreover, the high persistence points have a smaller birth coordinate and a larger death coordinate, owing to the fact that greater low frequency power makes the global minima and maxima of the aggregate signal smaller and larger, respectively. Based on these observations, we expect a greater birth time variance for signals with power spectral densities that have a greater ratio of low-frequency to high-frequency power. Stated differently, signals that have a higher damping factor for their oscillatory component are expected to have larger birth time variance.

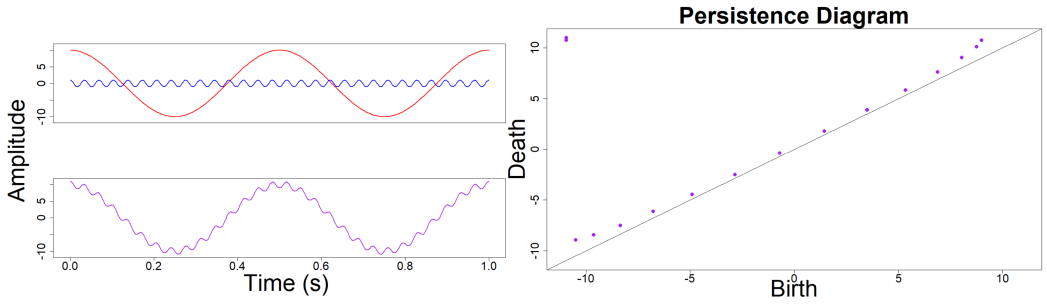
Figure 3.11 shows the average cardinality of persistence diagrams in $\mathcal{A}_f^\beta$ against the location of peak frequencies f with colors showing the damping factors β. A strong monotonic trend is apparent for all damping factors. For smaller damping factors, the relationship more heavily resembles that in the deterministic setting (see Section 3.4.2) where persistence diagram cardinality from a one second epoch is in fact equal to the frequency of the signal. Notice the effect of β diminishes as f increases, which implies cardinality is more sensitive to higher frequencies.

Figure 3.12 shows the relationship between f, β, and the birth value variance for $\mathcal{A}_f^\beta$. Birth variance increased as the damping factors increased. This is consistent with the idea that higher low-frequency-power-to-high-frequency-power ratios result in more variation for low persistence birth times. Interestingly, birth variance decreased as frequency increased, suggesting this trend is less notable at higher frequencies.

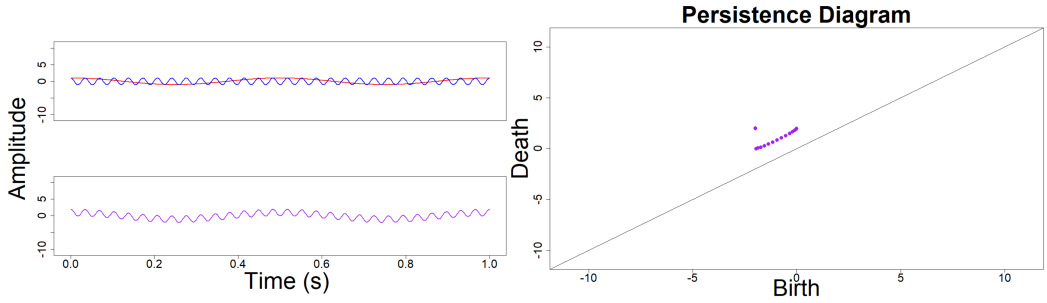
Signal Classification

As a point of comparison to traditional signal processing techniques and to showcase the utility of topological methods, we used features derived from Bayesian persistent homology (Section 3.4.2) and discrete Fourier transforms to classify signals with different dynamics. Specifically, we considered five classes of signals in total, each with a distinct rhythm quantified by a nonzero f_k parameter in our autoregressive filter (additionally, all classes had a peak at zero with a fixed damping factor to simulate the $1/f$ behavior commonly present in the PSD of EEG). Four classes we considered were: alpha (a), high beta (hb), low beta (lb), and theta (t), which had (autoregressive) spectral peaks at 10, 21, 14, and 6 Hz, respectively. We also included a null (n) class that had no peaks aside from one at zero. These spectral peaks were chosen for their prevalence in EEG data. We considered signals with damping factors of thirty-two to mimic weak oscillatory behavior in order to tackle a challenging problem. Visual descriptions of our data are provided in Fig. 3.13. Results from our classification are shown in Table 3.7.

We used three different classifiers in our experiment to assign a label to each of our signals via



(a)



(b)

Figure 3.10: This figure demonstrates the effect of greater low frequency power on the persistence diagram of a signal. Notice in (a) that elements of the persistence diagram show greater spread along the diagonal. This phenomenon arises since the low frequency signal scatters the higher frequency peaks in along the Amplitude axis.

Table 3.7: Precisions and recalls for each feature and classifier. Results are reported as mean ± standard error across each class.

	Bayesian		**PSD**		**CWT**	
Classifier	Precision	Recall	Precision	Recall	Precision	Recall
LR	0.84 ± 0.06	0.85 ± 0.07	0.90 ± 0.04	0.90 ± 0.04	0.91 ± 0.03	0.90 ± 0.04
SVM - Lin.	0.92 ± 0.05	0.91 ± 0.04	0.91 ± 0.03	0.90 ± 0.05	0.91 ± 0.04	0.91 ± 0.03
MLP	0.89 ± 0.05	0.88 ± 0.04	0.90 ± 0.02	0.89 ± 0.02	0.92 ± 0.03	0.93 ± 0.02

leave-one-out-crossvalidation (LOOCV). The first classifier, logistic regression (LR), is the most basic and models the probability that a signal belongs to class i as a function of a linear combination of input features. Another classifier we used was a support vector machine with a linear kernel (SVM-Lin) , which works by deciding optimal decision boundaries between the classes in feature space, then assigning new observations to classes based on these decision boundaries. A hyperparameter in SVM-Lin is a regularization term that penalizes mis-classification. To tune this hyperparameter,

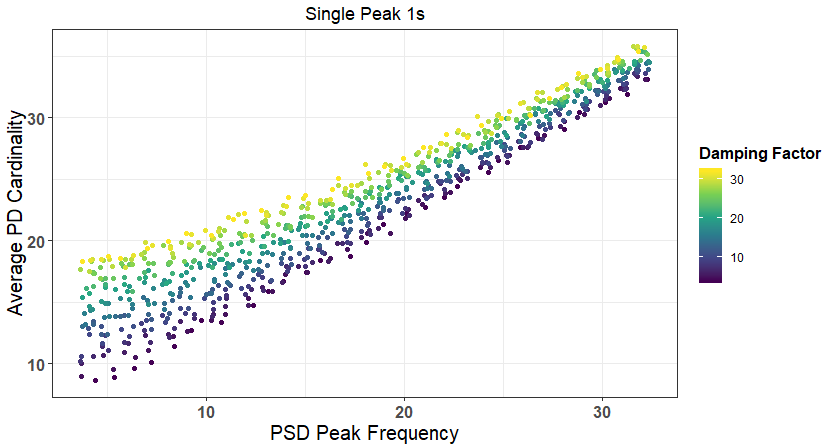


Figure 3.11: This plot depicts the relationship between the cardinality of persistence diagrams and the frequency of the dominant oscillation for one second autoregressive signals across various damping factors. For each included frequency and damping factor, we simulated thirty signals (each had a component fixed at zero to give the $1/f$ PSD commonly seen in EEG), computed their persistence diagrams, then recorded their average cardinality. We see a strong positive correlation between this average cardinality and the frequency of the dominant oscillation (i.e., PSD Peak Frequency) consistent with the idealized deterministic sinusoid case.

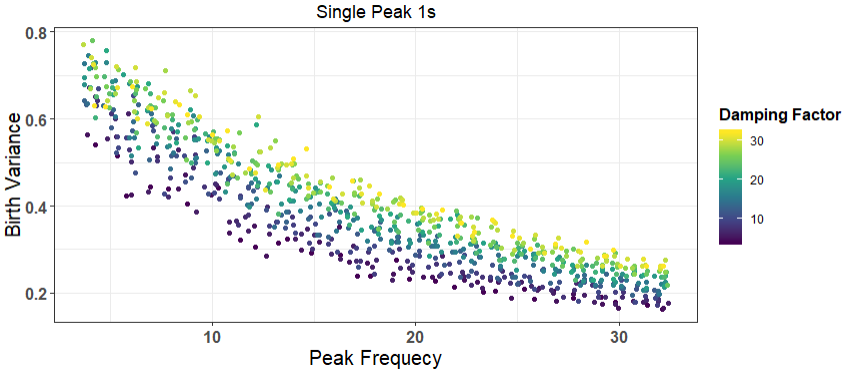


Figure 3.12: The peak frequency f for $\mathcal{A}_f^\beta$ plotted against the average birth variance for its persistence diagrams. Colors depict the damping factor β.

we relied on a grid search. The last classifier we considered was a multilayer perceptron (MLP), also known as a feedforward neural network. We chose a neural network architecture analogous to one used in [29]. Namely, we used 1 hidden layer of 4 neurons with a saturating linear activation function, followed by a softmax layer for classification. For each feature, we performed a grid search

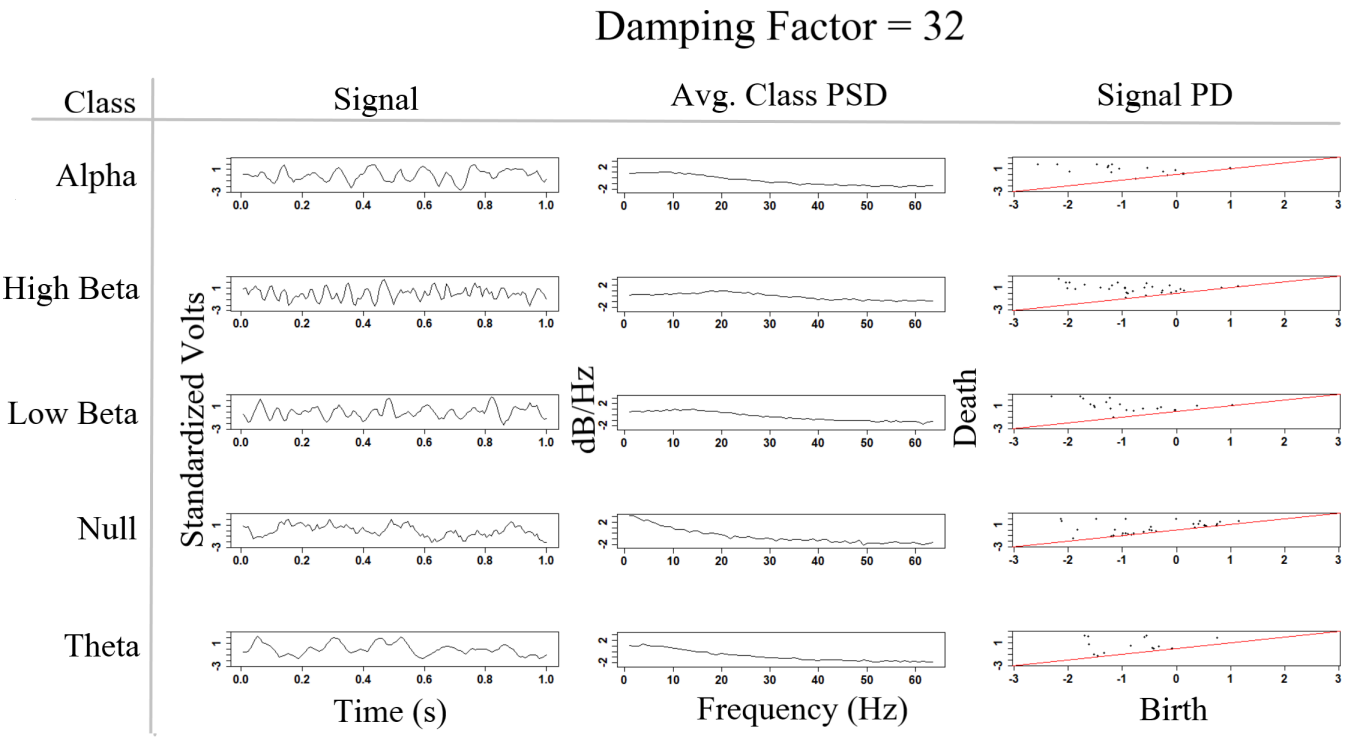


Figure 3.13: Here are the average (log) power spectral densities along with examples of signals and persistence diagrams from each class for damping factors of 32.

to determine an optimal number of epochs for training the neural network.

We used three sets of features for each classifier. The first two sets were derived from standard signal processing methods. The first set was based upon PSDs estimated via discrete Fourier transforms. In particular, for each signal, we computed its DFT then recovered an estimate to the PSD. We then binned the total power in frequency bands of interest (0.5-4 Hz, 4-7Hz, 7-12Hz, 12-21 Hz, 21-32Hz, $>$ 32 Hz) to obtain a 6-dimensional feature vector. The second set of features were created from continuous wavelet transforms. Specifically, we used the Mexican hat mother wavelet to obtain time-frequency plots for each signal. These time-frequency plots were then binned in the same frequency bands as those we considered in the PSD estimate to obtain 6 dimensional vectorizations. The final set of features we considered were derived from the Bayesian method outlined in Section 3.4.2. For each signal, we computed its persistence diagram, then used Equation (3.31) to estimate a posterior intensity. We then took 6 features of the posterior intensity to obtain 6-dimensional vectorizations. As the estimated posterior intensities were sensitive to the prior used to fit them, we tuned parameters in our prior λ_D using a grid search and LR as LR had no hyperparamaters to consider, unlike the other two classifiers. The 6 features of the posterior for a diagram $\mathcal{D}$, $\lambda_{D|\mathcal{D}}$, were as follows:

1. The total intensity, $\int_{\mathbb{W}} \lambda_{D|\mathcal{D}}(\mathbf{u})\, d\mathbf{u}$, which gives the expected cardinality.
2. The birth coordinate of the center of mass of $\lambda_{D|\mathcal{D}}$, $\int_{\mathbb{W}} \mathbf{u}\lambda_{D|\mathcal{D}}(\mathbf{u})\, d\mathbf{u}$.
3. The persistence coordinate of the center of mass.
4. The variance in birth coordinates for the posterior means of $\lambda_{D|\mathcal{D}}$.
5. The variance in persistence coordinates for the posterior means of $\lambda_{D|\mathcal{D}}$.
6. The covariance of birth and persistence coordinates of posterior means in $\lambda_{D|\mathcal{D}}$.

Examining Table 3.7 shows competitive performance of Bayesian-derived persistence diagram features for signal classification to that of well-established features for signal processing.

Chapter 4

Nonparametric Density Estimation for Persistence Diagrams

4.1 Kernel Density Estimation

Appealing to Lemma 2.2.1, one may define a kernel density for a random persistence diagram that considers all features individually; however, the computation of Equation 2.18 may be intractible as the number of terms in the summation scales exponentially with the cardinality of the evaluation diagram. To that end, kernel density, centered at a persistence diagram $\mathcal{D}$, has a bandwidth $\sigma > 0$ that reduces computational burden by treating some features individually and others collectively.

Typically, persistence diagrams have the majority of their points concentrated close to the diagonal. Consequently, the bandwidth σ is responsible for splitting a persistence diagram into upper and lower portions; see Equation 4.1 and Figure 4.1 (*Left*). The upper portion models the most topologically prominent points, which encompass topological information about the data, and its distribution reflects uncertainty in the precise location for prominent topological features in a persistence diagram. The lower portion models the majority of points in a persistence diagram. These points arise as a result of local noise in the underlying data, and in this fashion its distribution prescribes a noise likelihood model. Moreover, one can evaluate diagrams of any cardinality in the kernel (in this sense, the kernel is a global density). On the other hand, if one fixes the cardinality, one obtains the local kernel.

The construction of the kernel density proceeds by treating the upper and lower parts as independent, which necessitates the establishment of two density functions, one for each portion. The density for the upper part follows the recipe of Lemma 2.2.1 with a modified Gaussian chosen for $p^j(\xi)$ in Equation 2.18. To construct the kernel for the lower portion, we utilize (i) the number of points in $\mathcal{D}^\ell$ to create a pertinent counting measure, and (ii) a modified Gaussian mixture with mean the projection of each point in $\mathcal{D}^\ell$ to the diagonal. When evaluating a persistence diagram in the composite kernel, some of the points are evaluated in the density for the lower while others are evaluated in the density for the upper part. For a particular allocation of points to the upper and lower portions and by independence, the total evaluation follows from multiplying the results of these two evaluations together. However, since it is unknown a priori which input points should be used in each kernel, one must account for every possible partitioning of input points.

Section 4.1.1 gives a precise construction of our kernel density estimator. In [46], we establish that our kernel density estimator converges to the true probability density of random persistence diagrams (see Section 2.2.3 for a precise definition of this notion) under mild assumptions as the bandwidth of the kernel goes to zero; other results, like the convergence of the intensity of the kernel

density estimator to that of the true random diagram are also established. A particularly useful contribution in [46] is the introduction of a new statistic for persistence diagrams, the mean absolute bottleneck deviation (MAD), and proof that the sample MAD computed with the persistence diagram kernel density estimator converges to the true MAD. We define the MAD in Section 4.1.2, and finally show its utility in Section 4.1.3 with an application involving neuroimaging data.

4.1.1 Construction

To alleviate computational burden, our kernel density relies on a bandwidth parameter σ that partitions its center diagram into upper and lower portions.

$$\mathcal{D}^u = \{(b_i, d_i, k) \in \mathcal{D} : d_i - b_i \geq \sigma\} \text{ and } \mathcal{D}^\ell = \{(b_i, d_i, k) \in \mathcal{D} : d_i - b_i < \sigma\}. \tag{4.1}$$

Definition 4.1.1 ([46])**.** *Each feature $\xi_j = (b_j, d_j) \in \mathcal{D}^u$ yields an independent random singleton diagram D^j defined by its chance to be nonempty $q^{(j)}$ (via Eq. (4.3)) along with its potential position (b, d) sampled according to a modified Gaussian distribution, denoted by $N^*((b_j, d_j), \sigma I)$. The global pdf for D^u is then determined by Lemma 2.2.1, where each $p^{(j)}$ is given by the pdf associated with $N^*((b_j, d_j), \sigma I)$, which is given by*

$$p^{(j)}(b, d) = \frac{\varphi_j(b, d)}{\int_W \varphi_j(u, v) du\, dv} \mathbb{1}_W(b, d), \tag{4.2}$$

where φ_j is the pdf of the (unmodified) normal $N((b_j, d_j), \sigma I)$, and $\mathbb{1}_W(\cdot)$ is the indicator function for the wedge.

$$q^{(j)} = \mathbb{P}(D^j \neq \emptyset) = \int_{\{v>u\}} \varphi_j(u, v)\, du\, dv. \tag{4.3}$$

Definition 4.1.2 ([46])**.** *The lower random diagram D^ℓ is defined by choosing a cardinality N according to a pmf ν followed by N i.i.d. draws according to a fixed density p^ℓ. First, take $N_\ell = \left|\mathcal{D}^\ell\right|$ and define $\nu(\cdot)$ with mean N_ℓ and so that $\nu(n) = 0$ for $n > mN_\ell$ for some $m > 0$ independent of N_ℓ. The subsequent density $p^\ell(b, d)$ is given by projecting the lower features $\mathcal{D}^\ell$ of the center diagram $\mathcal{D}$ onto the diagonal $b = d$, then creating a restricted Gaussian kernel density estimation for these features; specifically,*

$$p^\ell(b, d) = \frac{1}{N_\ell} \sum_{(b_i, d_i) \in \mathcal{D}^\ell} \frac{1}{\pi\sigma^2} e^{-\left(\left(b - \frac{b_i + d_i}{2}\right)^2 + \left(d - \frac{b_i + d_i}{2}\right)^2\right)/2\sigma^2}. \tag{4.4}$$

In Section 4.1.3, we choose the probability mass function

$$\nu(N) = \max\left\{\frac{N_\ell + 1 - |N_\ell - N|}{(N_\ell + 1)^2}, 0\right\} \tag{4.5}$$

for ν in the lower portions of our kernels.

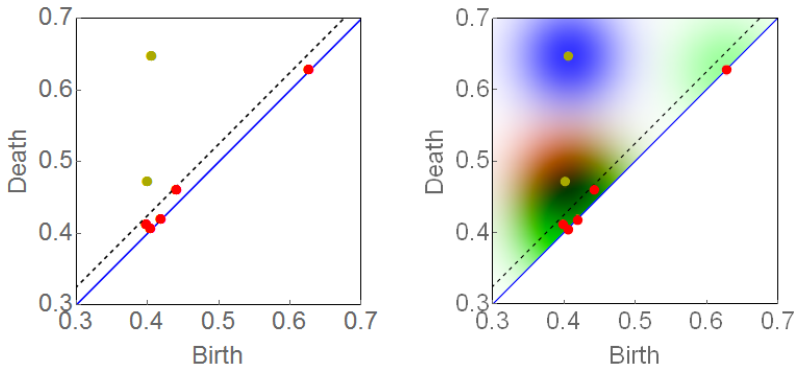


Figure 4.1: *Left:* A persistence diagram split according to Eq. (4.1). The dashed black line, $d = b + \sigma$, separates the diagram into the red upper points of $\mathcal{D}^u$ and the yellow lower points of $\mathcal{D}^\ell$. *Right:* The red and black gradients represent the upper singleton densities $p^{(1)}$ and $p^{(2)}$ given by Eq. (4.2). The green gradient represents the lower density p^ℓ defined in Eq 4.4. While each of these densities is defined on the wedge $W \subset \mathbb{R}^2$, the global kernel in Eq. (4.6) is defined on $\bigcup_N W^N$ for each input-cardinality N.

The next proposition gives the recipe for evaluating an input diagram in the kernel density constructed with Definitions 4.1.1 and 4.1.2.

Proposition 4.1.1 ([46])**.** *Fix a center persistence diagram $\mathcal{D}$ and bandwidth $\sigma > 0$. Split $\mathcal{D}$ into $\mathcal{D}^\ell$ and $\mathcal{D}^u$ according to Eq.* (4.1)*. Define D^ℓ with global pdf from Equation* (4.4)*, and D^u with global pdf from Equation* (2.18)*. Treating the random persistence diagrams D^u and D^ℓ as independent, define their union D. The following kernel density satisfies Definition 2.2.15 as the global pdf of D:*

$$K_\sigma(Z, \mathcal{D}) = \sum_{j=0}^{N_u} \nu(N-j) \sum_{\gamma \in I(j, N_u)} \mathcal{Q}(\gamma) \prod_{k=1}^{j} p^{(\gamma(k))}(\xi_k) \prod_{k=j+1}^{N} p^\ell(\xi_k), \tag{4.6}$$

where $Z = (\xi_1, ..., \xi_N)$ is the input, $\xi_i = (b_i, d_i)$ for $i = 1, ..., N$ are the features, and $N_u = |\mathcal{D}^u|$ depends on both $\mathcal{D}$ and σ. Here $\mathcal{Q}(\gamma)$ is given by Eq. (2.19)*, each $p^{(j)}$ refers to the modified Gaussian pdf as shown in Eq.* (4.2) *for its matching feature ξ_j in D^u, and p^ℓ is given by Equation* (4.4)*.*

Proof. See [46]. □

4.1.2 A Measure of Dispersion

Under mild assumptions, the kernel density estimator in Proposition 4.1.1 converges to the true pdf as the bandwidth parameter σ goes to zero [46]. Along with density function estimation, the kernel density estimator accurately estimates statistics of persistence diagrams like average spread. In the absence of vector space structure on the space of persistence diagrams, we turn to the bottleneck metric (Definition 4.1.3) to define a notion of spread. Specifically, we measure dispersion with respect to a distribution of persistence diagrams through its mean absolute deviation in this metric.

Definition 4.1.3 ([46]). *The mean absolute bottleneck deviation (MAD) from origin diagram $\mathcal{D}$ with respect to a global pdf f is given by*

$$MAD_f(\mathcal{D}) = \int_{\mathbb{W}} W_\infty(\mathcal{D}, Z) f(Z) \delta Z \tag{4.7}$$

Theorem 4.1.1 ([46]). *Consider a distribution of persistence diagrams with bounded global pdf, f, satisfying assumptions $(A1)$, $(A2)^*$, and $(A3)^*$. Let $\hat{f}(Z) = \frac{1}{n}\sum_{i=1}^{n} K_\sigma(Z, \mathcal{D}_i)$ be a kernel density estimate with centers $\mathcal{D}_i$ sampled i.i.d. according to global pdf f and bandwidth $\sigma = O(n^{-\alpha})$ chosen with $0 < \alpha < \alpha_{2M}$. Then, the mean absolute bottleneck deviation estimate converges; in other words,*

$$\int_{\mathbb{W}} W_\infty(\mathcal{D}_0, Z) \hat{f}(Z) \delta Z \to \int_{\mathbb{W}} W_\infty(\mathcal{D}_0, Z) f(Z) \delta Z \tag{4.8}$$

as $n \to \infty$ for any origin diagram $\mathcal{D}_0$.

4.1.3 Application to Neuroimaging Data

Electroencephalography (EEG) monitors electrical activity in the brain by measuring changes in voltage over time at particular locations on the scalp. To obtain data, researchers place arrays of electrodes on subjects' heads that record fluctuations in voltage as time series. It is known that EEG in the 1-100 Hz range is heavily involved in cognition [7]; moreover, specific bands in the 1-100 Hz range are hypothesized to be associated with certain tasks or brain states. For example, alpha range EEG, which has a spectral peak in the 8-12 Hz range, is thought to be important in inhibition and excitation during decision problems [37]. Collecting EEG is a noninvasive procedure, however, measurements are often obscured by noise arising from electrical activity in the environment, movement, or other physiological processes like heartbeat. Consequently, a critical problem is the need to detect EEG with the same underlying dynamics, e.g. EEG that is predominantly composed of 8-12 Hz oscillations, in the presence of varying levels of noise [38].

In this example, we consider the widely-used autoregressive EEG model introduced in [23], and we employ the KDE established in Proposition 4.1.1 to statistically analyze EEG. The authors in [23] model an EEG time series $(x_{t_i})_{i=0}^{L}$ of time length L as a convolution of white noise with a linear filter function given in Eq. (4.9),

$$h(t_i) = \sum_{j=1}^{p} e^{-\beta_j t_i} \cos(\omega_j t_i), \tag{4.9}$$

where ω_j correspond to centers of peaks in the power spectral density of $(x_{t_i})_{i=0}^{L}$ while the parameters β_j are approximately equal to $1/2$ of their respective widths (both ω_j and β_j are given in Hz). Recall that the power spectral density describes the contribution of each frequency to the total power of $(x_{t_i})_{i=0}^{L}$ after decomposing $(x_{t_i})_{i=0}^{L}$ into a series of oscillatory functions. For example, a power spectral density with a narrow peak at 10 Hz corresponds to a time series that heavily resembles a function oscillating at 10 Hz. A broader peak at 10 Hz in essence means that the time series has a greater contribution from more frequencies surrounding 10 Hz, diminishing the resemblance (for comparison, the power spectral density of white noise is completely flat). This view of the power spectral density means one can effectively simulate EEG comprised of oscillations in a desired band of frequencies by selecting appropriate parameters in Eq. (4.9).

We focus on alpha range (8-12 Hz) EEG. Specifically, we simulate 200 EEG signals in the alpha range by first generating 200 white noise vectors of length $L = 1,024$ through independent draws from $\mathcal{N}(0, 1)$ then convolving them with the linear filter described by Eq. (4.9) with $p = 1, \beta_1 = 3.7$,

and $\omega_1 = 10.5$. We corrupt 100 signals by additive noise $\mathcal{N}(0, 10^{-1/20})$, while the rest are corrupted by $\mathcal{N}(0, 10^{-5/20})$. This yields two collections of EEG signals with signal-to-noise ratios (SNRs) of 1 and 5, denoted by SNR_1 and SNR_5, respectively.

Next, we convert each EEG signal into a persistence diagram using the methodology of [59]. Namely, we transform EEG signals to point clouds in $\mathbb{R}^2$ using delay embeddings where the delay parameter was determined by the sampling rate (100 Hz) along with the dominant underlying frequency of the signals (10 Hz); we then center and scale the point clouds by their variances along the vertical and horizontal axes; see Figure 4.2 (e) and (f). Once we obtain point clouds, we compute persistence diagrams for 1-dimensional homological features using Rips filtrations; see Section 2.1.1. We choose to focus solely on 1-dimensional homological features since they relate to periodicity in the underlying time series, which is the defining characteristic of our signals.

Denote the family of persistence diagrams created from SNR_i for $i = 1, 5$ by D_{SNR_i}, and let f_{SNR_i} be their global probability densities. Our goal is to verify that SNR_1 and SNR_5 EEG have the same underlying dynamics. A sensible strategy is to select a quantity created from persistence diagrams that is robust to noise, approximate its distribution for SNR_i using D_{SNR_i}, and then compare the two empirical distributions. To this end, we start by approximating f_{SNR_i} with the kernel density estimators $\hat{f}_{SNR_i}(Z) := 10^{-2} \sum_{\mathcal{D} \in D_{SNR_i}} K_\sigma(Z, \mathcal{D})$. For each fixed $i = 1, 5$ the persistence diagrams $\mathcal{D} \in D_{SNR_i}$ are the 100 diagrams of each SNR_i case. For the noise likelihood model related to the lower part of a persistence diagram, $\mathcal{D}^\ell$, we use Eq. (4.5) as the cardinality distribution. Given that features with higher persistence generally describe global topology that is more resilient to noise, and relying on these kernels $\hat{f}_{SNR_i}$, we take $S = 1,000$ sample persistence diagrams and compute their bottleneck distance $W_\infty(\emptyset, S_i^j) = \max_{(b,d) \in S_i^j} d - b$, where S_i^j is the jth sample persistence ($j = 1, \ldots, S$) diagram distributed according to $\hat{f}_{SNR_i}$, $i = 1, 5$. These distances create empirical distributions, one for each SNR_i EEG denoted by F_{SNR_i}. We formally proceed with hypothesis testing

$$H_0 : F_{SNR_1} = F_{SNR_5} \text{ vs } H_1 : F_{SNR_1} \neq F_{SNR_5}.$$

Failure to reject H_0 in this case is evidence that D_{SNR_1} and D_{SNR_5} have similar behavior for the features less affected by noise, which in turn implies that SNR_1 and SNR_5 have similar underlying dynamics. Finally, we compare these distributions with a two-sided Kolmogorov-Smirnov (KS) Test [67] that yields a $p-$value=0.72.

For the sake of comparison to other TDA methods, we also compute persistence images (PIs) with resolution 50×50 and spread 0.2 using the ramp function to produce weights, [2], and persistence landscapes (PLs) from D_{SNR_i}, [10]. We examine the L_∞-norm as a summary for each of these vectorizations (the L_∞-norm of the first landscape in particular for PLs) since this measurement is also associated with high persistence features. After computing L_∞-norms for each of the PIs and PLs obtained from D_{SNR_i}, we resample each L_∞ empirical distribution 1,000 times to create bootstrapped distributions with size matching those of the W_∞ distributions obtained from the kernel density estimators; see Fig. 4.3 (c),(d),(e), and (f). In the end, we also compare the bootstrapped distributions with a two-sided KS-test. Table 4.1 shows the KS-test p-values and a standardized run time for each method.

Notice the kernel density max persistence and landscape L_∞ correctly fail to reject H_0 at the most commonly used significance levels ($p-$value $= 0.79$). In particular, our method is competitive with landscapes (with a slight edge on computational time). On the other hand, the persistence image L_∞ incorrectly rejects H_0 ($p-$value close to 0). Failure of PIs to recognize different dynamics may be a result of the fact that in addition to accounting for the max persistence (through the use of the ramp function for weights), the PI L_∞ also considers the cardinality of each diagram,

although the contribution of cardinality diminishes for higher resolutions and smaller spreads.

Finally, we report estimates for $MAD_{f_{SNR_1}}(\emptyset)$ and $MAD_{f_{SNR_5}}(\emptyset)$ by taking the means of our empirical distributions for the max persistences; see Table 4.2. Notice the estimates are very close numerically and by appealing to Theorem 4.1.1, one could argue they are close to their true values. Hence, the MAD offers more evidence that SNR_1 and SNR_5 are statistically indistinguishable.

Table 4.1: The p-values and run times for each method (KDE, PI, and PL) used for the hypothesis test of Eq. (4.1.3).

	KS-Test P-value	Time (s)
KDE MP	0.72	0.047
PI L_∞	6.15×10^{-9}	0.042
PL L_∞	0.79	0.048

Table 4.2: The sample MADs for SNR_1 and SNR_5 computed by taking the means of the distributions in Fig 4.3(e) and Fig 4.3(f), respectively.

	Sample MAD
SNR_1	1.040
SNR_5	1.035

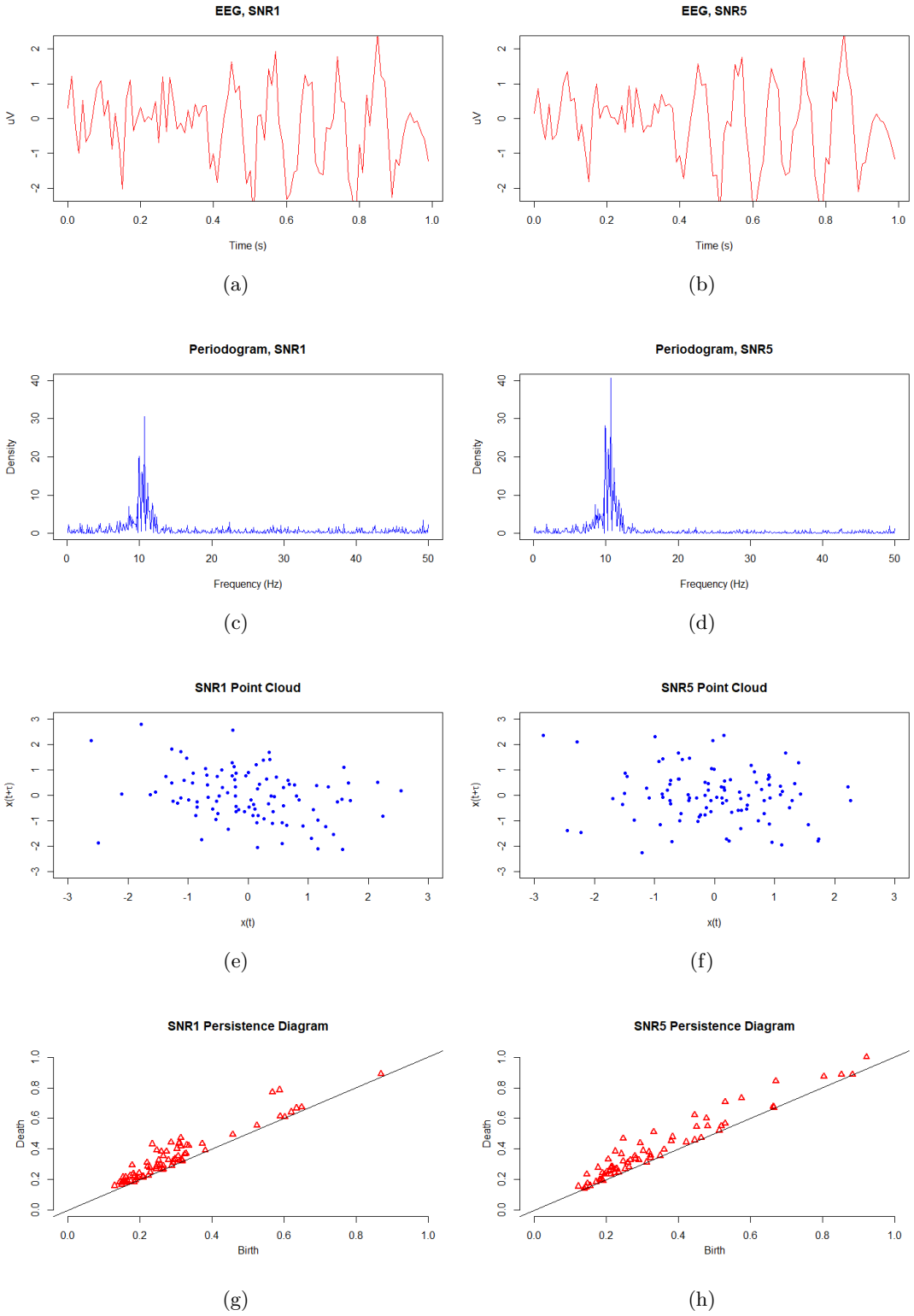


Figure 4.2: A segment of an EEG signal with (a) SNR_1, and (b) SNR_5, respectively, along with (c) and (d): their corresponding periodograms (estimates of the power spectral densities). The associated point clouds are given in (e) and (f), respectively, and (g) and (h) their resulting persistence diagrams.

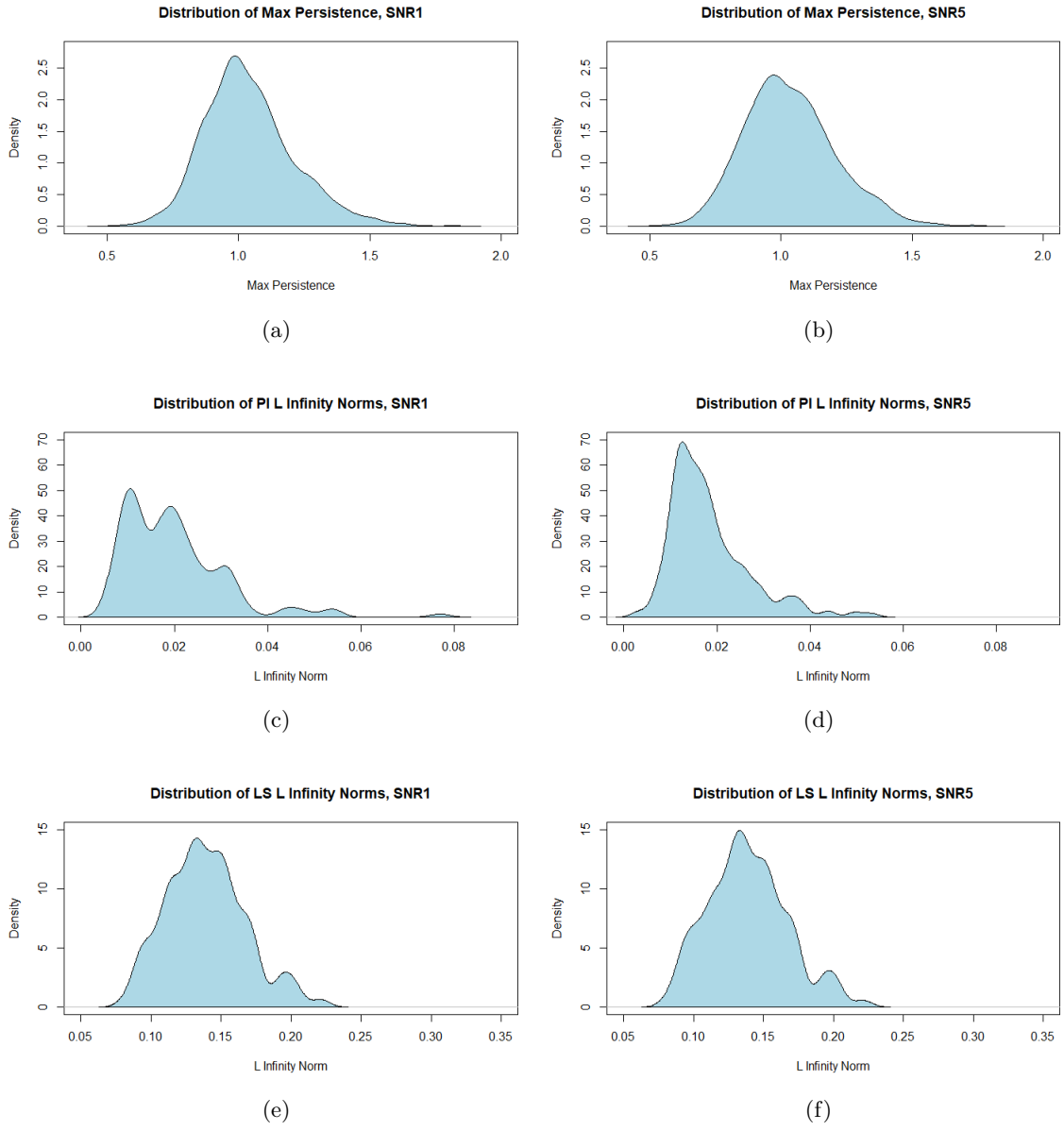


Figure 4.3: This figure shows the distribution for each statistic we considered when comparing SNR_1 to SNR_5. Each column represents a class, either SNR_1 or SNR_5, and each row a particular statistic.

Chapter 5

Data Augmentation with Topological Noise

Algorithms for creating persistence diagrams (from filtrations defined on data) yield a pair of inverse mappings from samples in a dataset to points in a persistence diagram, and vice versa [21, 78]. These inverse mappings may be used to optimize functions of persistence diagrams with respect to data via the chain rule and gradient descent, thereby allowing one to manipulate the topology of data with aptly defined persistence diagram functions. In [24], this inverse map is shown to be differentiable under mild conditions, and a Newton-Raphson method for continuation of one to point cloud to another based on their persistence diagrams is introduced. The work [62] more explicitly shows how the inverse map allows for gradient-based optimization of persistence diagram functions via the chain rule. More recently, [9] uses the persistence inverse map for gradient-based optimization of loss functions in deep learning architectures that utilize persistence diagrams to control the topology of data. Our work differs from these previous in that we use the persistence inverse map within a probabilistic setting for data augmentation. Our framework enables one to generate new training examples with varying levels of topological noise. In summary, we define a Boltzmann distribution whose energy term measures the topological similarity between a sampled dataset Z and a target dataset X using their respective filtered complexes. Gradient-based optimization of this energy term then allows us to explore the state space of datasets, where the samples we obtain bear different levels of topological similarity to X.

Chapter 5 is structured as follows. In Section 5.1, we summarize the computational persistent homology background necessary to conceptually grasp our approach. Next, we present our topological data augmentation method in Section 5.2. Section 5.3 follows up with applications of our method to classification problems with materials science point cloud data arising and the MNIST dataset.

5.1 Background

5.1.1 Computational Topology Preliminaries

Definition 5.1.1 (Filtration, [20])**.** *Let $\mathcal{K}$ be a simplicial complex, and suppose $f : \mathcal{K} \to \mathbb{R}$ satisfies (i): $f(\sigma) \leq f(\tau)$ whenever σ is a face of τ. Define $\mathcal{K}(a) := f^{-1}(-\infty, a]$ and notice that (i) implies $\mathcal{K}(a)$ is a subcomplex of $\mathcal{K}$ for every $a \in \mathbb{R}$. Taking $a_1 < a_2 < \cdots < a_n$ to be the values of f for every simplex in $\mathcal{K}$ and denoting $\mathcal{K}(a_i) := \mathcal{K}_i$, we obtain an increasing sequence of subcomplexes $\emptyset = \mathcal{K}_0 \subset \mathcal{K}_1 \subset \cdots \subset \mathcal{K}_n = \mathcal{K}$, which we call the filtration of f.*

As an example of Definition 5.1.1, the function f that corresponds to the Vietoris Rips filtration (Definition 2.1.7) is given by $f : [v_{i_0}, v_{i_1}, \ldots, v_{i_k}] \mapsto \max_{\alpha,\beta} ||v_{i_\alpha} - v_{i_\beta}||$ for every simplex $[v_{i_0}, v_{i_1}, \ldots, v_{i_k}] \in \mathcal{K}$. A filtration induces k-th persistent homology through the maps

$$\begin{array}{ccccccccccc} \emptyset & = & \mathcal{K}_0 & \subset & \mathcal{K}_1 & \subset \ldots \subset & \mathcal{K}_n & = & \mathcal{K} \\ \downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow \\ 0 & \longrightarrow & H_k(\mathcal{K}_0) & \longrightarrow & H_k(\mathcal{K}_1) & \longrightarrow \ldots \longrightarrow & H_k(\mathcal{K}_n) & = & H_k(\mathcal{K}). \end{array}$$

Namely, the birth and death times of a homological feature ξ of degree k are the values a_i and a_j (respectively) of f that correspond to $H_k(\mathcal{K}_i)$, the homology group in which ξ first appears, and $H_k(\mathcal{K}_j)$, the earliest homology group where ξ is killed by an element of $\operatorname{Im} \partial_{k+1}$. One encounters a problem when two homological features, say ξ and ξ', merge at $H_k(\mathcal{K}_j)$ because at that point in the filtration, $\xi - \xi' \in \operatorname{Im} \partial_k$. In this event, one must decide which of the homological features lives or dies. This ambiguity is resolved by a convention known as the Elder Rule [20], which dictates that the feature appearing earlier in the filtration persists. A filtration can be visualized mentally as building $\mathcal{K}$ by gradual addition of subcomplexes. To simplify matters, we only consider filtrations where a single simplex is added during each step (this can be ensured for simplicial complexes built on data by the use of jitter as necessary). With the addition of simplices during a filtration, homological features are created or destroyed. Those whose addition spawns a homological feature are called positive simplices, while their counterparts that kill features are called negative simplices. It can be shown [21] that each homological feature which occurs during a filtration maps to a simplex pair, (σ, τ), where σ and τ are the positive and negative simplices that create and destroy the feature, respectively. The collection $\{(f(\sigma), f(\tau)\}_{(\sigma,\tau)\in\mathcal{P}}$, where $\mathcal{P}$ is the set of all positive-to-negative simplex pairs for the filtration of f, is known as a persistence diagram, which we henceforth denote by $\mathcal{D}$. We also write b_σ and d_τ for the values $f(\sigma)$ and $f(\tau)$, respectively. This can be summarized by a map,

$$\pi : (b_\sigma, d_\tau) \mapsto (\sigma, \tau) \tag{5.1}$$

sending points of a persistence diagram to their corresponding positive negative simplex pairs. Furthermore, we have a map

$$\gamma : \sigma \mapsto [v_{i_1}, v_{i_2}, \ldots, v_{i_k}] \tag{5.2}$$

which sends a simplex to its vertex set. Composition of these maps, which we denote by

$$\rho := \gamma \circ \pi, \tag{5.3}$$

yields a relation between $\mathcal{D}$ and the set of vertices of $\mathcal{K}$.

Now, suppose $E : \mathcal{D} \mapsto \epsilon_{\mathcal{D}}$ is a map that sends $\mathcal{D}$ to a real number, $\epsilon_{\mathcal{D}}$. The map ρ enables differentiation of E with respect to the vertex set of $\mathcal{D}$ through the chain rule [62]. This result allows for gradient calculation, and hence optimization, of real-valued functions of persistence diagrams with respect to the data on which they are built.

Let $\mathcal{D}_X := \{\mathbf{p}_j^X\}_{j=1}^N$ and $\mathcal{D}_Z := \{\mathbf{p}_j^X\}_{j=1}^M$ be the persistence diagrams of a target point cloud and a sampled point cloud, respectively. In Section 5.2, we use 2 functions to measure similarity between these two persistence diagrams, the first of which is the Wasserstein distance, defined as

$$W_p^q(\mathcal{D}_X, \mathcal{D}_Z) := \inf_{\iota:\mathcal{D}_Z\to\mathcal{D}_X} \sum_{\mathbf{p}\in\mathcal{D}_Z} ||\mathbf{p} - \iota(\mathbf{p})||_p^q, \tag{5.4}$$

where ι denotes a bijection between $\mathcal{D}_Z$ and $\mathcal{D}_X$. To ensure Equation (5.4) is well-defined, both $\mathcal{D}_Z$ and $\mathcal{D}_X$ are considered to contain the diagonal, i.e. the set of points $\{(b, d) \in \mathbb{R}^2 : b = d\}$, when

working with the Wasserstein metric. The second function we consider, the d_p^c distance [42], does not impose this constraint, and is defined by

$$D_p^c(\mathcal{D}_X, \mathcal{D}_Z) := \inf_{\text{inj}:\mathcal{D}_Z \to \mathcal{D}_X} \sum_{\mathbf{p} \in \mathcal{D}_Z} \min(||\mathbf{p} - \text{inj}(\mathbf{p})||_p^p, c) + c|M - N|, \tag{5.5}$$

where inj denotes an injection and we have supposed without loss of generality that $M < N$. The d_p^c distance penalizes cardinality differences between persistence diagrams and does not allow matching to the diagonal. The final persistence diagram function we consider for point cloud data measures differences $\mathcal{D}_Z$ and the empty diagram. It can still be used to promote desired topological qualities in sampled point clouds. The third function was introduced in [9], and is defined below:

$$\mathcal{E}_2^2(D_Z) = \sum_{\mathbf{p} \in \mathcal{D}_Z} ||d_{\mathbf{p}} - b_{\mathbf{p}}||_2^2 ||\frac{d_{\mathbf{p}} - b_{\mathbf{p}}}{2}||_2^2. \tag{5.6}$$

In Equation (5.6), $b_{\mathbf{p}}$ and $d_{\mathbf{p}}$ denote the birth and death coordinates of $\mathbf{p}$.

5.1.2 Point Cloud Classification with PointNet

In Section 5.3, we consider a classification problem involving 3-dimensional point cloud data. Due to the nature of point cloud data, which represent unordered collections of elements in a metric space, deep learning architectures for point cloud classification commonly rely on feature engineering to extract signatures from point clouds that are invariant to orderings of elements and large scale geometric transformations like rotations [4, 8, 71]. To circumvent this preprocessing step in our application, we use the deep learning architecture PointNet [12], which takes point cloud data directly as input then learns features to extract during training. The key components of the PointNet architecture are two layers that learn symmetric functions of inputs for data classification. These symmetric functions respect the fact that point clouds are unordered sets and thus encourages the network to learn features that are invariant under permutations. PointNet is one of pioneering architectures for point set classification and segmentation, outperforming traditional architectures on a variety of benchmark datasets [12]. Additionally, it has been implemented in Keras, a high-level API for TensorFlow, and is publically available at https://github.com/garyli1019/pointnet-keras.

5.2 Data Augmentation

For the remainder of this section, $X = \{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^{d_X}$ is a point cloud for a positive dimension d_X, $\mathcal{K}_X$ is a filtered simplicial complex whose vertex set and filtration function are X and $f : \mathcal{K}_X \to \mathbb{R}$, respectively, and $\mathcal{D}_X$ is the persistence diagram associated to f and $\mathcal{K}_X$ of a fixed homological dimension.

In this section, we propose our topological data augmentation framework, which builds augmented datasets through the introduction of topological noise to existing training examples. Intuitively, injecting topological noise into X amounts to perturbing the elements $\{x_1, x_2, \ldots, x_n\}$ in a manner that retains a portion (contingent on noise level) of their topological structure. This can be achieved by mapping X to a topological signature, in this case a filtered simplicial complex, manipulating this element in a stable fashion, then mapping the newly obtained topological signature back to a point cloud in $\mathbb{R}^{d_X}$. For example, a concrete implementation of the aforementioned data transformation pipeline (i) sends X to its filtered simplicial complex $\mathcal{K}_X$, (ii) shifts points in the persistence diagram $\mathcal{D}_X$ to create a new diagram $\mathcal{D}'_X$ within a fixed Wasserstein distance, (iii) updates the persistence pairs in $\mathcal{K}_X$ via the map π to reflect movement of elements in $\mathcal{D}_X$,

then (iii) sends $\mathcal{D}'_X$ back to a point cloud X' using the map γ and the updated persistence pairs. Unfortunately, (ii) presents a major difficulty since elements in distinct persistence pairs are generally related as they have vertices in common, and all of these relationships must be accounted for and tracked when shifting points in $\mathcal{D}_X$ to induce geometrically consistent changes in X. Our proposed method avoids the problem in (ii) as it does not rely on direct manipulation of persistence diagrams. Rather than perturbing X to obtain a noisy sample, we start with a random point cloud Z then, at a rate determined by the desired noise level, adjust elements in Z to maximize their topological similarity to those in X.

Specifically, we consider a Boltzmann-like distribution for point clouds that incorporates topology,

$$p\Big([Z,\mathcal{K}_Z]\Big|[X,\mathcal{K}_X]\Big) \propto e^{-\beta E\Big([Z,\mathcal{K}_Z],[X,\mathcal{K}_X]\Big)}, \tag{5.7}$$

where β is an inverse temperature parameter controlling the level of topological noise in our samples, and E is an energy function measuring the topological similarity between X and Z using their respective filtered simplicial complexes, $\mathcal{K}_X$ and $\mathcal{K}_Z$. High and low likelihoods for samples Z in Equation (5.7) correspond to low and high amounts of topological noise, respectively. We may obtain samples of varying likelihoods from Equation (5.7) by minimizing E through backpropogation and gradient descent, adjusting our training time and controlling our learning rate (which we set to equal β) to accommodate the desired noise level in our samples. Generically, small scale topology greatly varies in data sets since it is the result of local noise. By augmenting data using Equation (5.7), we can increase small scale topological variation in training sets, which assists deep learners in their capability to prioritize topologically relevant structure in point cloud data. We can also address issues of class imbalance by using Equation (5.7) to generate low noise samples from underrepresented classes. Our topological noise sampling method is summarized in Algorithm 2 and a demonstration is depicted in Figure 5.1. A critical hyperparameter in our data augmentation pipeline is the energy function E. It is essential that E both measures topological similarity and is differentiable with respect to elements in the sample Z. To account for the sensitivity of our data augmentation pipeline to this hyperparameter, we choose several different versions of E for data augmentation with point clouds; these are summarized in Table 5.1.

Algorithm 2 Sample with Topological Noise.

Input: A point cloud from the training set X, a filtration function f, an inverse temperature parameter β, an energy function E, and the number of gradient updates M.
$\mathcal{K}_X, \mathcal{D}_X := \text{Filtration}(X, f)$
for $i = 1, 2, \ldots, M$ **do**
  $\mathcal{K}_Z, \mathcal{D}_Z := \text{Filtration}(Z, f)$
  $G := \text{Grad}(Z, \mathcal{K}_Z, \mathcal{D}_Z, X, \mathcal{K}_X, \mathcal{D}_X)$
  $Z := Z - \beta \cdot G$
end for
Return Z

The energy functions we use decompose to

$$E\Big([Z,\mathcal{K}_Z],[X,\mathcal{K}_X]\Big) = E_{\text{spat}}(Z,X) + E_{\text{top}}(\mathcal{K}_Z,\mathcal{K}_X) \tag{5.8}$$

where E_{spat} and E_{top} measure similarities between the coordinates and topological arrangements of Z and X, respectively. In all of our choices for E, we set $E_{\text{spat}} = 0$ or $E_{\text{spat}}(X,Z) = ||X - Z||_2^2$, where the former selection is used to augment training sets solely with data that is topologically

similar, neglecting specific coordinates in the training examples entirely. We set $E_{\text{spat}} = 0$ in some versions of E to investigate whether topology alone is sufficient to augment training sets. Our choices for $E_{\text{top}}(\mathcal{K}_Z, \mathcal{K}_X)$ are $\mathcal{E}_2^2(\mathcal{K}_Z, \mathcal{K}_X)$, $D_2^2(\mathcal{K}_Z, \mathcal{K}_X)$, and $W_2^2(\mathcal{K}_Z, \mathcal{K}_X)$. Table 5.1 summarizes different versions of E we use for data augmentation.

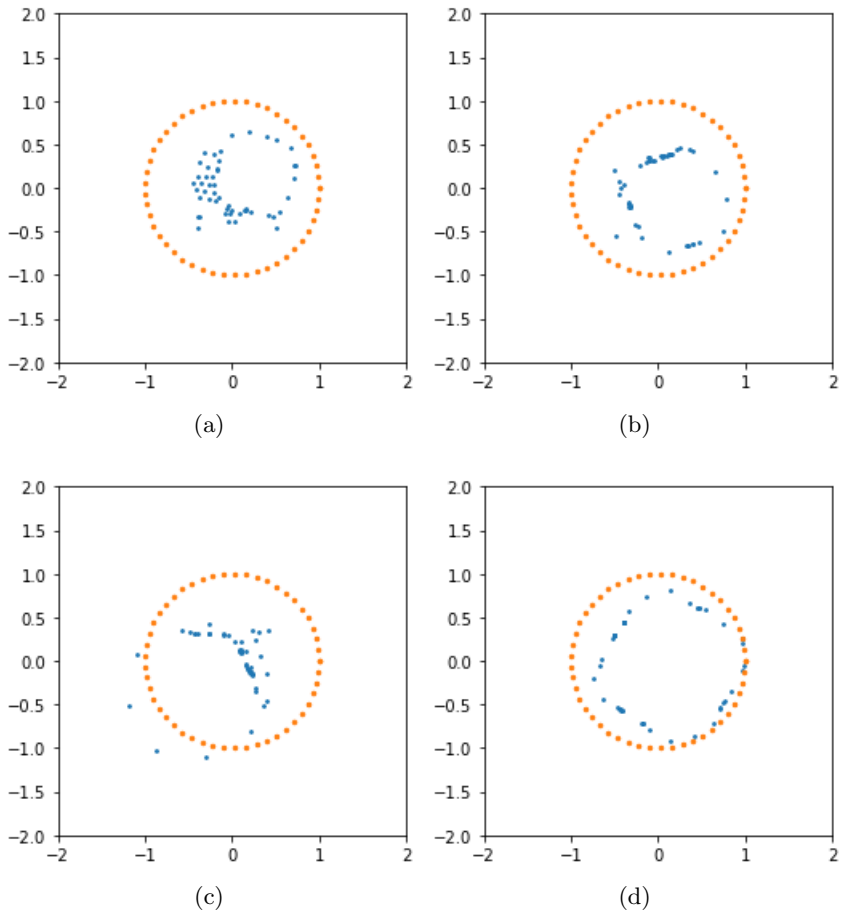


Figure 5.1: Shown above in blue are 2-dimensional point clouds created with Algorithm 2 with learning rate β fixed at 0.01, E = Wass from Table 5.1, and the number of epochs set to (a) 100, (b) 200, (c) 400, and (d) 800, respectively. As each epoch corresponds to a step in gradient descent, we expect point clouds created with a larger amount of epochs to bear a higher degree of topological similarity to the target point cloud, shown in orange. Indeed, we observe that the blue point clouds resemble circles with radii increasing to that of the target point cloud as the number of epochs increases.

Table 5.1: Choices for the energy function in Equation 5.7, E.

E	E_{spat}	E_{top}
Wass	0	W_2^2
Wass-L2	$\mathcal{L}_2$	D_2^2
dpc	0	D_2^2
dpc-L2	$\mathcal{L}_2$	D_2^2
poly	0	$\mathcal{E}_2^2$
poly-L2	$\mathcal{L}_2$	$\mathcal{E}_2^2$

5.3 Experimental Results

5.3.1 Point Cloud Data

To determine the efficacy of topological data augmentation in training deep neural networks, we train PointNet to distinguish the two types of atomic neighborhoods from Chapter 3 (FCC and BCC) using augmented datasets. We use the most realistic noise and sparsity parameters (1 and 67%, respectively) for the APT data. We independently consider balanced and unbalanced distributions for training. In the former, our training set consisted of 1000 points clouds from each class; in the latter, the FCC and BCC classes had 1000 and 100 examples, respectively. We created distinct training sets for each energy function in Table 5.1 by generating a new training example for each element in the training set with Algorithm 2, fixing the number of epoches and learning rate at 20 and 0.01, respectively. For comparison, we also generate augmented datasets using random rotations and white noise perturbations of point clouds. In both the unbalanced and balanced cases, the test set consisted of 1000 examples from each class. Tables 5.2 and 5.3 show the maximum validation accuracies obtained during training of PointNet with each augmented dataset in the balanced and unbalanced cases, respectively, while Figure 5.2 visually depicts training and validation accuracies.

Table 5.2: This table shows the maximum validation accuracy over every epoch obtained by training PointNet with augmented datasets built from a balanced training set. Results are shown using for different choices of energy function E in Algorithm 2; see Table 5.1.

Method	Max Val. Acc.
Traditional	0.75
poly	0.74
poly-L2	0.74
was	0.74
was-L2	0.74
dpc	0.74
dpc-L2	0.75

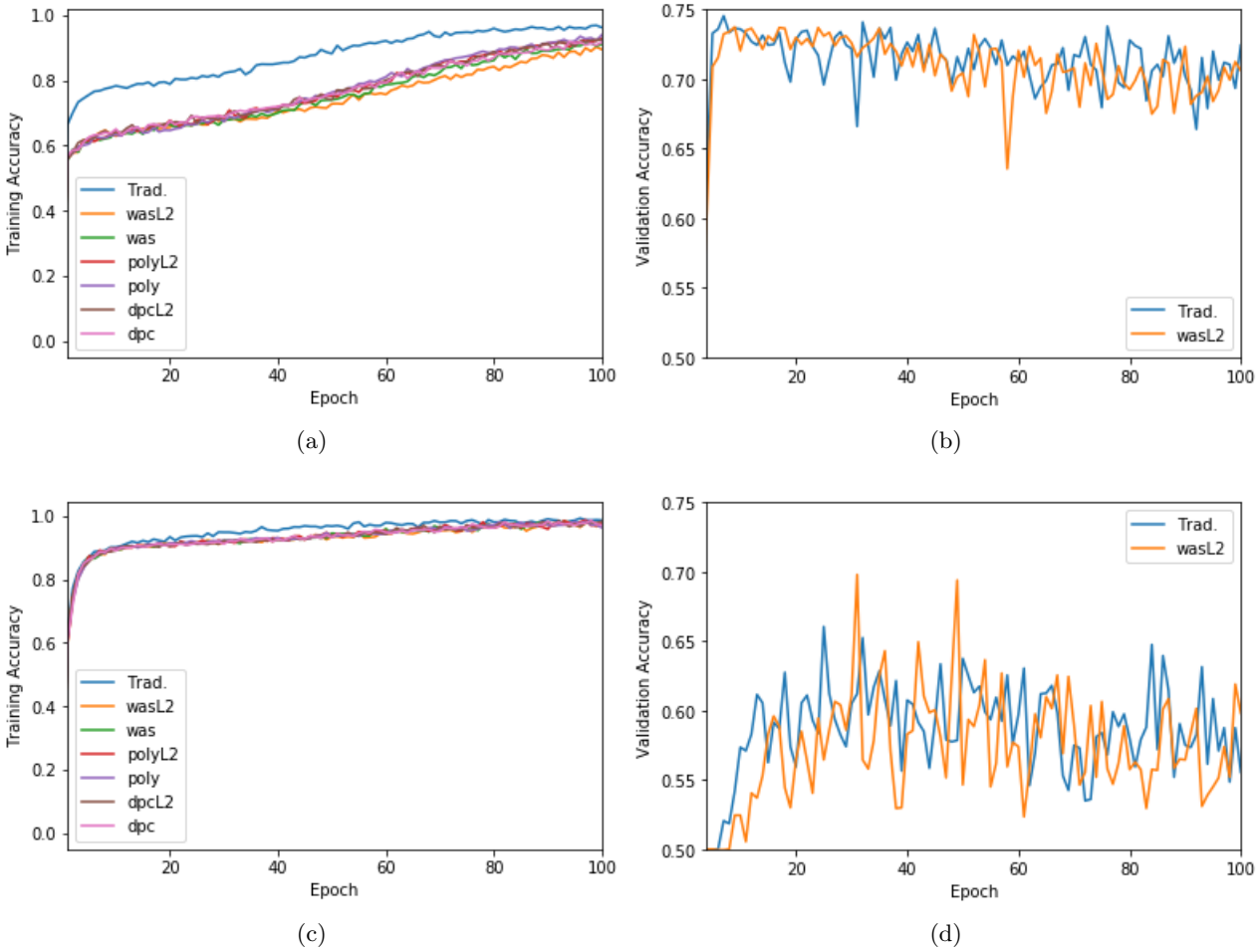


Figure 5.2: Shown above are training accuracies for (a) balanced and (c) unbalanced training sets. Similarly, (b) and (d) display the corresponding validation accuracies. To prevent overcluttering, validation accuracies are only shown for traditional and was-L2 data augmentation, since was-L2 attained the highest validation accuracy (taken over every epoch) for the unbalanced training cases.

In the case of balanced training data, we saw similar performance in terms of validation accuracy across all augmented datasets. Notably, the training accuracy in this case was consistently higher for the traditionally augmented training set than the accuracies for the topologically augmented training sets. This suggests that topological data augmentation introduced greater variation in the training set than could be obtained from rotating and perturbing point clouds alone. In the case of unbalanced training data, topological data augmentation slightly improved maximum validation accuracy when compared to traditional data augmentation, with the poly-L2 and was-L2 set showing the most dramatic increases.

Table 5.3: This table shows the maximum validation accuracy over every epoch obtained by training PointNet with augmented datasets built from an unbalanced training set. results are shown for different choices of energy function E in Algorithm 2; see Table 5.1.

Method	Max Val. Acc.
Traditional	0.66
poly	0.65
poly-L2	0.68
was	0.63
was-L2	0.70
dpc	0.66
dpc-L2	0.66

5.3.2 Image Data

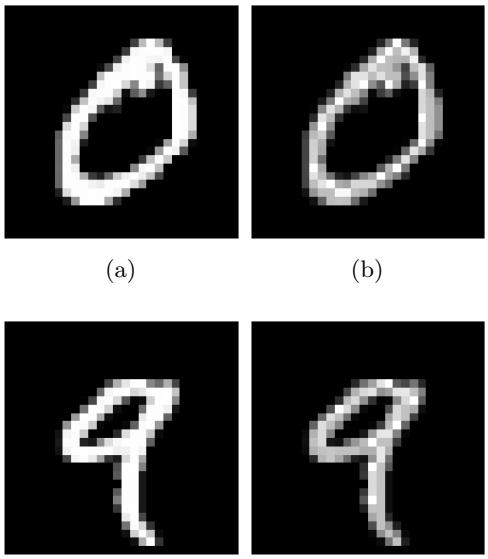
We also apply our topological data augmentation pipeline in classification of the MNIST dataset [17], which consists of 28-by-28 greyscale images of handwritten digits. For each image in our dataset, we compute persistence diagrams for 0-dimensional features using lower star filtrations (i.e., sublevel set filtrations) of cubical complexes, which are analogous to simplicial complexes and better suited for image data; see [72]. An inverse mapping between points in a persistence diagram and the pixels in its corresponding image akin to the map 5.3 exists in the setting of lower star filtrations as well, enabling optimization of persistence diagram functions with respect to the data. The energy function we use for image data is

$$P(\mathcal{D}) = \sum_{\mathbf{p} \in \mathcal{D}} b_{\mathbf{p}}^2 + d_{\mathbf{p}}^2 \tag{5.9}$$

Unlike the case for point cloud data, we do not begin with a random image then steer it toward a target image with desired topological characteristics. Rather, we directly manipulate the topology of an image in a controlled manner (via gradient descent with Equation 5.9) to introduce variation in the training set. In Equation 5.9, $\mathcal{D}$ is a 0-dimensional persistence diagram, so minimizing -P increases the 0-dimensional persistence of the data. For the image data, this manifests as a decrease in the number of local minima, leading to images that have a few isolated local minima; see Figure 5.3.

The neural net architecture we use for classification of the MNIST digits is shown below in 1-8:

1. Convolutional layer, 16 channels, 3-by-3 filter, 1-by-1 stride.
2. Convolutional layer, 8 channels, 3-by-3 filter, 1-by-1 stride.
3. Max pooling, 2-by-2 filter.
4. Dropout layer, 0.25.
5. Flatten layer.
6. Dense layer, 32 neurons.



(a) (b)

(c) (d)

Figure 5.3: Original images (a,c) from the MNIST dataset and the resulting images (b,d) after topological augmentation with P. Here, lighter colored pixels have smaller values.

7. Dropout, 0.5.

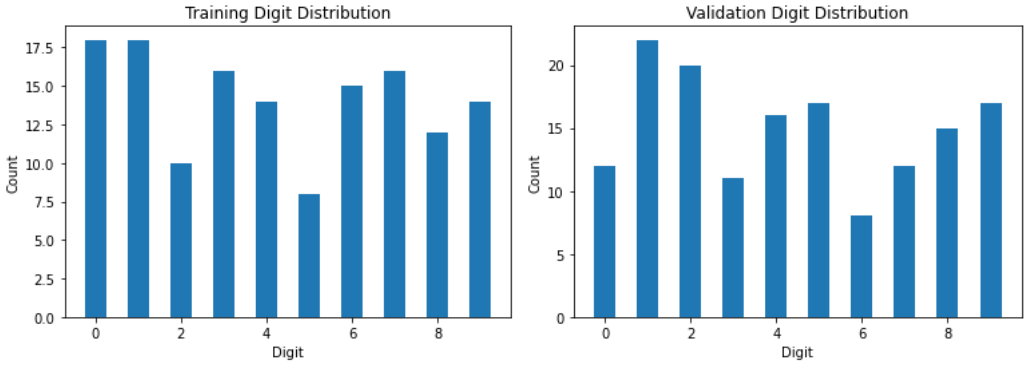
8. Softmax activation layer, 9 neurons.

In particular, we take 150 MNIST digits for training and validation, respectively. The class distributions of the the training and validation sets are shown in Figure 5.4.

For comparison, we also consider a training set that is augmented by applying a random horizontal and vertical shift to the pixels in an image, and a training set with no augmentation. The augmented training sets are created by generating an additional training example by applying our augmentation method to each element of the training set. Results for classification of the MNIST data with each augmentation method are shown in Figure 5.5. Maximum validation accuracies for each method are shown in Table 5.4.

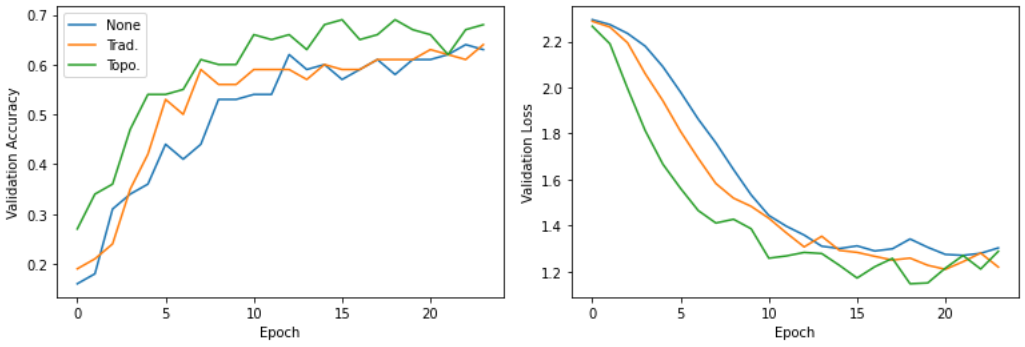
Table 5.4: Maximum validation accuracy for each augmentation method over 24 training epoches.

Method	Max Val. Acc.
None	0.64
Traditional	0.64
Topological	0.69



(a) (b)

Figure 5.4: The distribution of images in the training and validation sets are shown in (a) and (b), respectively.



(a) (b)

Figure 5.5: (a) Validation accuracies and (b) validation losses shown for the unaugmented training set (None), the training set augmented with random shifts (Trad.), and the training set augmented with our topological method (Topo.), respectively.

Chapter 6

Conclusion

Herein, we have detailed new inferential tools for topological data analysis and provided several applications demonstrating their utility in real-world data problems.

Chapter 3 describes the first approach to introduce a generalized Bayesian framework for persistent homology. This toolbox gives experts the opportunity to incorporate their prior belief about data in conjunction with topological data analysis notions when faced with research questions. The framework is entirely predicated upon modelling random persistence diagrams with Poisson point processes because of their nice qualities that behave well with Bayesian formulations. Specifically, since they are characterized entirely by their intensity measures, they allow us to quantify prior uncertainty with presupposed intensity functions, and allow for efficient computation of posterior intensities if we regard observed persistence diagrams as noisy observations described by marks in a marked Poisson point process. Interestingly, recent works [3, 47] could also be used to devise an alternative, parallel point-process-based Bayesian framework for persistent homology. In particular, one could directly use Bayes rule with prior distributions constructed from [47] then obtain posteriors with the methodology from [3], which outlines a procedure for Monte Carlo estimation of Choquet integrals. This is a worthwhile future direction for research of this nature.

It should be noted that the Bayesian model considers persistence diagrams, which are summaries of the data at hand, for defining a substitution likelihood rather than using the underlying point cloud data. This does not adhere to a strict Bayesian viewpoint, as we model the behavior of the persistence diagrams without considering the underlying data (materials data in our example) used to create it; however, our paradigm incorporates prior knowledge and observed data summaries to create posterior probabilities, analogous to the notion of substitution likelihood detailed in [31]. The general relationship between the likelihood models related to point cloud data and those of their corresponding persistence diagrams remains an important open problem.

Chapter 3 also introduces a conjugate-like family of prior intensities and stochastic kernels (our likelihood analog), which can be used to obtain a closed form for posterior intensities. A detailed example is presented to illustrate the qualities of posterior intensities arising under several interesting parameter choices in our model. This example establishes evidence that our Bayesian framework updates prior uncertainty with new observations in a manner similar to that for standard random variables. Thus, the Bayesian inference developed herein can be reliably used for machine learning and data analysis techniques directly on the space of persistence diagrams. Indeed, a classification algorithm is derived and successfully applied on materials science data to assess the capability of our Bayesian framework.

Chapter 3 also gives an interpretable framework for signal processing via sublevel persistent homology. Explicit representations for persistence diagrams of signals are provided in Propositions

3.4.1 and 3.4.1, which aid interpretability of persistence diagrams in a signal processing context. In future work, one may expand on Proposition 3.4.1 by finding closed forms for more general family of signals, and investigate families of stochastic signals for which Proposition 3.4.1 yields tractable probability densities. This would further enhance interpretation of persistence diagrams, making them more attractive objects for the scientific community. Results in Chapter 3 for signal classification provide evidence that posterior distributions created via our Bayesian framework are probabilistic descriptors competitive with well-established descriptors like the power spectral density and continuous wavelet transform for distinguishing time series dynamics. In the future, one may examine the effectiveness of higher resolution summaries of posterior intensities for use in advanced deep learning architectures. This is necessary for the incorporation of topological methods into state-of-the-art signal processing methods. The main component of Chapter 4 is also an application of persistent homology to signal processing, which compliments Chapter 3 by adopting a nonparametric viewpoint for persistence diagram probability density estimation. In particular, Chapter 4 creates a new statistical test with a persistence diagram kernel density estimator, and shows how it can be used to detect time series dynamics in noisy measurements.

Chapter 5 provides a method for data augmentation based on introducing sensible topological variation to training sets. The effectiveness of our method is measured by monitoring the performance of a sophisticated point-cloud-classification deep learning architecture on a classification problem with realistic APT data. Topological data augmentation saw similar performance, measured by validation accuracy, to that of traditional data augmentation. Moreover, topological data augmentation appeared to introduce more variation in the training set than did its traditional counterpart. In the future, traditional and topological data augmentation may be combined to produce more diverse training sets from limited data. Additionally, although we focused on point cloud data and persistence diagrams created via Vietoris Rips filtrations, the map ρ exists for any filtered simplicial complex and thus our data augmentation method can be applied for any filtrations that are easily differentiable with respect to their vertices. We briefly investigated this differentiable map for lower-star filtrations of image data by examining the MNIST dataset. Backpropogation for persistence diagram functions built from a variety of filtration function was explored in [9] in the context of loss functions for a neural network. Finally, we remark that Equation (5.7) may be used for topological data analysis in broader Bayesian contexts, particularly as likelihoods in posterior computations and variational autoencoders, and to construct empirical distributions for data based on topological similarity to a target dataset. Regarding variational autoencoders, the work [51] explores the use of topological loss functions, practically similar to those in Table 5.1, as regularization terms to encourage topological similarity between data and latent space representations. Alternatively, our formulation provides a generative distribution given data and a filtered simplicial complex.

References

[1] *Swaiman's Pediatric Neurology*. Elsevier, 2017.

[2] Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *The Journal of Machine Learning Research*, 18(1):218–252, 2017.

[3] Hamzeh Agahi, Hossein Mehri-Dehnavi, and Radko Mesiar. Monte carlo integration for choquet integral. *International Journal of Intelligent Systems*, 34(6):1348–1358, 2019.

[4] M. Aubry, U. Schlickewei, and D. Cremers. The wave kernel signature: A quantum mechanical approach to shape analysis. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1626–1633, 2011.

[5] Mojtaba Bandarabadi, António Dourado, César Alexandre Teixeira, Theoden I. Netoff, and Keshab K. Parhi. Seizure prediction with bipolar spectral power features using adaboost and svm classifiers. *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6305–6308, 2013.

[6] Robert J. Barry, Adam R. Clarke, Stuart J. Johnstone, Christopher A. Magee, and Jacqueline A. Rushby. EEG differences between eyes-closed and eyes-open resting conditions. *Clinical Neurophysiology*, 118(12):2765–2773, December 2007.

[7] David L. Boothe, Alfred B. Yu, Pawel Kudela, William S. Anderson, Jean M. Vettel, and Piotr J. Franaszczuk. Impact of neuronal membrane damage on the local field potential in a large-scale simulation of cerebral cortex. *Frontiers in Neurology*, 8:236, 2017.

[8] M. M. Bronstein and I. Kokkinos. Scale-invariant heat kernel signatures for non-rigid shape recognition. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1704–1711, 2010.

[9] Rickard Brüel-Gabrielsson, Bradley J. Nelson, Anjan Dwaraknath, Primoz Skraba, Leonidas J. Guibas, and Gunnar Carlsson. A topology layer for machine learning, 2019.

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

[11] Peter Bubenik. The persistence landscape and some of its properties, 2018.

[12] R. Qi Charles, Hao Su, Mo Kaichun, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[13] F. Chazal and B. Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. arXiv:1710.04019, 2017.

[14] Frédéric Chazal and Vincent Divol. The density of expected persistence diagrams and its kernel based estimation, 2018.

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

[16] D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes. Vol. I.* Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2003.

[17] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[18] David Steven Dummit and Richard M Foote. *Abstract Algebra*, volume 3. Wiley Hoboken, 2004.

[19] P. J. Durka, D. Ircha, and K. J. Blinowska. Stochastic time-frequency dictionaries for matching pursuit. *IEEE Transactions on Signal Processing*, 49(3):507–510, 2001.

[20] Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction*. American Mathematical Society, Providence, RI, 2010.

[21] 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.

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

[23] P. J. Franaszczuk and K. J. Blinowska. Linear model of brain electrical activity?EEG as a superposition of damped oscillatory modes. *Biological Cybernetics*, 53(1):19–25, November 1985.

[24] Marcio Gameiro, Yasuaki Hiraoka, and Ippei Obayashi. Continuation of point clouds via persistence diagrams. *Physica D: Nonlinear Phenomena*, 334:118–132, Nov 2016.

[25] Baptiste Gault, Michael P Moody, Julie M Cairney, and Simon P Ringer. Atom probe crystallography. *Materials Today*, 15(9):378–386, 2012.

[26] Shafie Gholizadeh and Wlodek Zadrozny. A short survey of topological data analysis in time series and systems analysis, 2018.

[27] R. Ghrist. Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society*, 45:61–75, 2008.

[28] Allen Hatcher. *Algebraic Topology*. Cambridge Univ. Press, Cambridge, 2000.

[29] C. Ieracitano, N. Mammone, A. Bramanti, S. Marino, A. Hussain, and F. C. Morabito. A time-frequency based machine learning system for brain states classification via eeg signal processing. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2019.

[30] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.

[31] H. Jeffreys. *Theory of Probability*. Clarendon Press, 1961.

[32] Marian Gidea Jesse Berwald. Critical transitions in a model of a genetic regulatory system. *Mathematical Biosciences and Engineering*, 11(1551-0018_2014_4_723):723, 2014.

[33] Olfa Kanoun and H-R Trankler. Sensor technology advances and future trends. *IEEE Transactions on Instrumentation and Measurement*, 53(6):1497–1501, 2004.

[34] Thomas F. Kelly and Michael K. Miller. Atom probe tomography. *Review of Scientific Instruments*, 78(3):031101, 2007.

[35] Firas Khasawneh and Elizabeth Munch. Exploring equilibria in stochastic delay differential equations using persistent homology. volume 8, 08 2014.

[36] J. F. C. Kingman. *Poisson processes*. Clarendon Press, Oxford, 1993.

[37] Wolfgang Klimesch. Alpha-band oscillations, attention, and controlled access to stored information. *Trends in Cognitive Sciences*, 16(12):606 – 617, 2012.

[38] Wolfgang Klimesch. Alpha-band oscillations, attention, and controlled access to stored information. *Trends in Cognitive Sciences*, 16(12):606 – 617, 2012.

[39] P. Y. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Vejdemo-Johansson, M. Alagappan, J. Carlsson, and G. E. Carlsson. Extracting insights from the shape of complex data using topology. In *Scientific reports*, volume 3, 2013.

[40] R. Mahler. *Statistical multisource-multitarget information fusion*. Artech House, Boston, 2007.

[41] S. G. Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[42] A. Marchese and V. Maroulas. Topological learning for acoustic signal identification. In *2016 19th International Conference on Information Fusion (FUSION)*, pages 1377–1381, July 2016.

[43] Andrew Marchese and Vasileios Maroulas. Signal classification with a point process distance on the space of persistence diagrams. *Advances in Data Analysis and Classification*, 12(3):657–682, 2018.

[44] Andrew Marchese and Vasileios Maroulas. Signal classification with a point process distance on the space of persistence diagrams. *Advances in Data Analysis and Classification*, 12(3):657–682, September 2018.

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

[46] Vasileios Maroulas, Joshua L Mike, and Christopher Oballe. Nonparametric estimation of probability density functions of random persistence diagrams. *Journal of Machine Learning Research*, 20(151):1–49, 2019.

[47] Vasileios Maroulas, Joshua L Mike, and Christopher Oballe. Nonparametric estimation of probability density functions of random persistence diagrams. *Journal of Machine Learning Research*, 20(151):1–49, 2019.

[48] Vasileios Maroulas, Farzana Nasrin, and Christopher Oballe. A bayesian framework for persistent homology. *SIAM Journal on Mathematics of Data Science*, 2(1):48–74, 2019.

[49] Y. Mileyko, S. Mukherjee, and J. Harer. Probability measures on the space of persistence diagrams. *Inverse Problems*, 27(12):124007, 2011.

[50] Michael P Moody, Baptiste Gault, Leigh T Stephenson, Ross KW Marceau, Rebecca C Powles, Anna V Ceguerra, Andrew J Breen, and Simon P Ringer. Lattice rectification in atom probe tomography: Toward true three-dimensional atomic microscopy. *Microscopy and Microanalysis*, 17(2):226–239, 2011.

[51] Michael Moor, Max Horn, Bastian Rieck, and Karsten Borgwardt. Topological autoencoders, 2019.

[52] J.E. Moyal. The General Theory of Stochastic Population Processes. *Acta Mathematica*, 108:1–31, 1962.

[53] J. R. Munkres. *Elements Of Algebraic Topology*. CRC Press, 1984.

[54] F. Nasrin, C. Oballe, D. Boothe, and V. Maroulas. Bayesian topological learning for brain state classification. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1247–1252, 2019.

[55] Monica Nicolau, Arnold J Levine, and Gunnar 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.

[56] Alan V Oppenheim, John R Buck, and Ronald W Schafer. *Discrete-Time Signal Processing; 2nd ed.* Prentice-Hall signal processing. Prentice-Hall, Upper Saddle River, NJ, 1999.

[57] Nina Otter, Mason A Porter, Ulrike Tillmann, Peter Grindrod, and Heather A Harrington. A roadmap for the computation of persistent homology. *EPJ Data Science*, 6(1), Aug 2017.

[58] Abby L Parrill and Kenny B Lipkowitz. *Reviews in Computational Chemistry, Volume 29*. Wiley Online Library, 2016.

[59] Jose A. Perea and John Harer. Sliding windows and persistence: An application of topological methods to signal analysis. *Found. Comput. Math.*, 15(3), June 2015.

[60] R. Pintelon and J. Schoukens. Time series analysis in the frequency domain. *IEEE Transactions on Signal Processing*, 47(1):206–210, 1999.

[61] R. Pintelon and J. Schoukens. Time series analysis in the frequency domain. *IEEE Transactions on Signal Processing*, 47(1):206–210, 1999.

[62] Adrien Poulenard, Primoz Skraba, and Maks Ovsjanikov. Topological function optimization for continuous shape matching. In *Computer Graphics Forum*, volume 37, pages 13–25. Wiley Online Library, 2018.

[63] M. D. Sacchi, T. J. Ulrych, and C. J. Walker. Interpolation and extrapolation using a high-resolution discrete fourier transform. *IEEE Transactions on Signal Processing*, 46(1):31–38, 1998.

[64] Nicole Sanderson, Elliott Shugerman, Samantha Molnar, James D. Meiss, and Elizabeth Bradley. Computational topology techniques for characterizing time-series data. In *Advances in Intelligent Data Analysis XVI*, pages 284–296. Springer International Publishing, 2017.

[65] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[66] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6:1–48, 2019.

[67] Richard Simard and Pierre L'Ecuyer. Computing the two-sided kolmogorov-smirnov distribution. *Journal of Statistical Software, Articles*, 39(11):1–18, 2011.

[68] S. S. Singh, B. Vo, A. Baddeley, and S. Zuyev. Filters for spatial point processes. *SIAM Journal on Control and Optimization*, 48(4):2275—-2295, 2009.

[69] Primoz Skraba, Vin de Silva, and Mikael Vejdemo-Johansson. Topological analysis of recurrent systems. In *NIPS 2012*, 2012.

[70] Charles J. Stone. An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.*, 12(4):1285–1297, 12 1984.

[71] Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Computer graphics forum*, volume 28, pages 1383–1392. Wiley Online Library, 2009.

[72] Hubert Wagner, Chao Chen, and Erald Vuçini. Efficient computation of persistent homology for cubical data. In *Topological methods in data analysis and visualization II*, pages 91–106. Springer, 2012.

[73] Xiang Wang, Kai Wang, and Shiguo Lian. A survey on face data augmentation for the training of deep neural networks. *Neural Computing and Applications*, Mar 2020.

[74] L. Wasserman. Topological data analysis. *Annual Review of Statistics and Its Application*, 5(1):501–532, 2018.

[75] Qingsong Wen, Liang Sun, Xiaomin Song, Jingkun Gao, Xue Wang, and Huan Xu. Time series data augmentation for deep learning: A survey, 2020.

[76] John B Williams. *The electronics revolution: inventing the future*. Springer, 2017.

[77] Angelo Ziletti, Devinder Kumar, Matthias Scheffler, and Luca M Ghiringhelli. Insightful classification of crystal structures using deep learning. *Nature communications*, 9(1):2775, 2018.

[78] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete & Computational Geometry*, 33(2):249–274, 2005.

Vita

Christopher Oballe was born in Austin, Texas on August 28th, 1992. He attended Vista Ridge High School, and he developed interests in mathematics and science after participating in advanced courses and science fairs. He graduated at the top of his class in 2011.

In the summer of 2011, he started attending the University of Texas at Austin. He was first exposed to proof-based mathematics when he participated in a research initiative during his first year, which included an advanced linear algebra course taught by Professor Ronny Hadani. Chris' interest in mathematics research grew immensely when he participated in an REU at NIMBios, located at the University of Tennessee, in the summer of 2013. Chris graduated with High Honors in 2015 with his Bachelor of Science in Mathematics.

Chris began his PhD in the summer of 2015 at the University of Tennessee, Knoxville. He passed preliminary examinations in analysis and topology after his first year. In the summer of 2017, he joined Professor Vasileios Maroulas' research group, and his research became focused on the synthesis of computational statistics, machine learning, and the topology of data. In 2018, he started working at Army Research Lab in Aberdeen, Maryland as a part of a cooperative agreement between UTK and ARL. He became an ORAU Pre-doctoral Fellow at ARL in 2019. During his time at ARL, Chris collaborated with neuroscientists to create novel topological methods for EEG analysis.

As of the summer of 2020, Chris has accepted a post-doctoral position at the University of Notre Dame, working under Professor Nicholas Zabaras.