Statistical Computational Topology and Geometry for Understanding Data

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Joshua Lee Mike

August 2017

This work is dedicated to my wife, Gracie, my mom, Rebecca, and my advisor, Vasileios. Without these three, none of it is possible.

Acknowledgments

I would like to thank my advisor Prof. Vasileios Maroulas for his intense support and endless feedback. I would lilke to thank my committee members, Professors Conrad Plaut, Ken Stephenson, and Michael Berry, for their feedback and support. I would like to thank my academic brother (by now, Dr.) Andrew Marchese for working hard alongside me. I would like to thank Dr. Fernando Schwartz for introducing me to topological data analysis and starting my academic career. I would like to thank Prof. Colin Sumrall for showing me how fun and interesting interdisciplinary work can be. I would like to thank Grace McClurkin for being wonderful, balancing me out, and keeping my pace reasonable. I would like to thank Russel Nalepa for showing me that mathematics is about far more than numbers. I would like to thank Prof. Brian Allen for many discussions about geometry and mathematics. I would also like to thank the many mathematicians at UTK who have blessed me with their insight, instruction, friendliness, camaraderie, and more.

Abstract

Here we describe three projects involving data analysis which focus on engaging statistics with the geometry and/or topology of the data.

The first project involves the development and implementation of kernel density estimation for persistence diagrams. These kernel densities consider neighborhoods for every feature in the center diagram and gives to each feature an independent, orthogonal direction. The creation of kernel densities in this realm yields a (previously unavailable) full characterization of the (random) geometry of a dataspace or data distribution.

In the second project, cohomology is used to guide a search for kidney exchange cycles within a kidney paired donation pool. The same technique also produces a score function that helps to predict a patient-donor pair's a priori advantage within a donation pool. The resulting allocation of cycles is determined to be equitable according to a strict analysis of the allocation distribution.

In the last project, a previously formulated metric between surfaces called continuous Procrustes distance (CPD) is applied to species discrimination in fossils. This project involves both the application and a rigorous comparison of the metric with its primary competitor, discrete Procrustes distance. Besides comparing the separation power of discrete and continuous Procrustes distances, the effect of surface resolution on CPD is investigated in this study.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

There is a common theme to the projects contained within this dissertation which needs to be discussed at the forefront. The overarching goal is to engage statistics with geometry in data analysis, an innately interdisciplinary task. The applications presented here are primarily biological, although one can apply the techniques in much broader context. Specifically, we aim to bring the ideas of geometry and shape, which live in a continuous or even smooth realm, into the subject of data analysis wherein the objects are by their very nature discrete and finite. Indeed, all the projects discussed here try to bridge this gap in various ways.

There are a couple important themes to keep in mind for this work. First, in the world of data analysis, geometry and topology often cannot separate cleanly. In our attempt to introduce either subject into data analysis, we must assume that the data comes from a space with a certain level of structure, such as the vague structure of a topological space, the middle ground of a simplicial complex, or the rigid structure of a Riemannian manifold. Despite the fact that one makes an assumption, it is often unclear how much structure the data actually possesses. In trying to discover unassailable topological features, one may also find signs of geometric structure; indeed, this scenario will be discussed explicitly for persistent homology. Relationships between topological type and coarse geometric restrictions are classically exemplified by theorems in geometry such as Gauss's Theorem Egregium [82], many pinching theorems [18], and the monotonicity of geometric flows [46].

Second, probability and statistics are a powerful tool for bridging the gap between classic continuity and discrete datasets. For example, the notion of stability allows us to label

persistent features as important data characterisitics in persistent homology. Introducing probability densities allows one to discern important features from unimportant ones in a *reliable* fashion regardless of their persistence and also opens the way for describing dependencies between the scale and/or persistence of various features in a sampled dataset.

Altogether, the results divulged here are better motivated from the viewpoint of data analysis. Though a technique may be topologically motivated, the ability to retrieve geometric information about the distribution of datapoints can be of tremendous help for machine learning problems such as classification or data visualization. In the end, we need not fully separate the topological from the geometric; instead, we must be able to draw conclusions about *shape* from these methods. Since the goal is to develop tools for data analysis, a lot of statistical analysis is done for the applications and the theorems require a probabilistic point of view.

The history of these ideas begins with topological data analysis (or TDA), which extends notions of homology and other topological invariants for discrete data through their ambient metric. Since persistent homology and other methods in TDA often depend heavily on the ambient metric, it is also important to discuss to a lesser extent some history in computational and metric geometry. Metric geometry also applies directly to the last work presented, on continuous Procrustes distance [6], which defines a metric between surfaces embedded in $\mathbb{R}^3$ and offers a stronger characterization of shape for data with more structure.

Because of the often broad scope within these projects, the relevent mathematical background will be discussed immediately prior to the results of each chapter. We hope that this introduction gives the reader a feel for the overarching ideas among these projects.

Computational Topology is a relatively new field of study wherein various topological invariants are used to describe a dataset. While knot theory has seen extensive use in fields such as statistical mechanics [48] and cell biology [45], the application of homology to discrete data sets has exploded in just the past decade [29]. The broad use of homology has taken up the term Topological Data Analysis (TDA), and applies principally to a technique called persistent homology. Nevertheless, the realm of TDA has branched out into dimensionality reduction [87], big data analysis [80], and ranking [47]. Mapper [80] is commonly used as a data reduction technique in order to obtain some information about the structure of

a massive dataset; for example, various kinds of breast cancer patients are delineated by mapper in [64].

Persistent homology in particular lent a new point of view to data analysis. In short, the technique yields a collection of shape traits $\{b_i, d_i, k_i\}$ which come from homology group generators of dimension k_i and persist from scale b_i to scale d_i. The application of persistent homology only requires a metric between the data points and can therefore be used in an exploratory fashion without stringent assumptions. Persistent homology has been used commonly in machine learning contexts to define specific features for classification. This is particularly common in signal analysis wherein, according to Takens' theorem [85], the delay embedding of a signal yields a high dimensional reflection of its underlying dynamics. The Euclidean metric is naturally apropos for the delay embedding and thus it easily yields a meaningful persistence diagram.

Applications of TDA range from classification and clustering in fields such as action recognition [90], handwriting analysis [4], and biology [33, 66, 64, 59, 60] to the analysis of dynamical systems [65, 72] and complex systems such as sensor networks [24, 26, 28, 17, 16]. Specifically, techniques including persistent homology [25, 12, 94, 42, 11] and manifold learning [87, 73, 10] have helped to compress nonlinear point cloud data from a new geometrically faithful point of view. In the realm of signal analysis, classification and clustering based on geometric and topological features of the phase-space of the signal detects features that traditional signal analysis techniques fail to detect [56, 57, 90, 77, 33]. In many cases, the application of TDA lacks a measure of reliability, and research in the area has quickly begun to implement various statistical measures for persistence diagrams and other related topological summaries.

The transformation of data into a persistence diagram is exceptionally nonlinear, and precise definitions can be elusive. So, while many good results have been presented, there is still much work required to fully integrate TDA with statistics. Much of the work presented here is part of this effort to introduce formal statistics into the realm of TDA. In particular, we introduce a framework for creating probability density functions which describe distributions of persistence diagrams along with a method for estimating these densities. The proposed

method is nonparametric since the transition (of a distribution) from underlying dataset to persistence diagram is still poorly understood.

From a broad perspective, the subject matter of computational geometry involves either (i) applying ideas from geometry to create new algorithms or (ii) trying to describe geometric structures through discrete and algorithmic reasoning. We explore both sides of computational geometry within the field of metric geometry. Metric geometry is the branch of geometry which aims to study metric spaces. Metrics are very important for data analysis, especially if one wishes to delineate between groups or measure dispersion. The definition of an ambient metric is essential for TDA: the closeness of various data points relies on having a notion of distance between points. The results which pertain to persistent homology also make use of metrics defined between persistence diagrams. Lastly, the third project analyses and applies a metric defined between surfaces.

Overview of Thesis

In the first chapter, we build a kernel density on the space of persistence diagrams. We being by discussing the topological background needed to define a persistence diagram. This result also requires the introduction of random set theory, which is the lens through which we view random pesistence diagrams. We demonstrate convergence on compact sets of global probability density function estimates built from the persistence diagram kernel density to the true distribution of the sampled diagrams. We also present convergence of mean absolute deviation with respect to the bottleneck metric.

The second chapter applies combinatorial Hodge theory to kidney paired donation. This chapter begins with background for the kidney paired donation as well as cohomology and Helmholtz decomposition. This mathematics motivates the introduction of an organ allocation algorithm as well as a score which describes a priori allocation preference. Our results show that the new algorithm yields a more equitable allocation than its competitors, while minimizing loss of efficiency.

The third and final chapter describes an application of metric geometry to fossil speciation. In particular, we discuss and compare discrete and continuous variants of Procrustes distance within this context. We test both the efficacy and reliability of

continuous Procrustes distance by its power to separate species from fossils described by triangulations at various levels of resolution.

Chapter 2

Nonparametric Estimation of Probability Density Functions of Random Persistence Diagrams

Summary

We introduce a nonparametric way to estimate the global probability density function for a random persistence diagram. Precisely, a kernel density function centered at a given persistence diagram and a given bandwidth is constructed. Our approach encapsulates the number of topological features and considers the appearance or disappearance of features near the diagonal in a stable fashion. In particular, the structure of our kernel individually tracks long persistence features, while considering features near the diagonal as a collective unit. The choice to describe short persistence features as a group saves computational time and reduces scaling while simultaneously retaining accuracy. Indeed, we prove that the associated kernel density estimate converges to the true distribution as the number of persistence diagrams increases and the bandwidth shrinks accordingly. We also establish the convergence of the mean absolute deviation estimate, defined according to the bottleneck metric. Lastly, examples of kernel density estimation are presented for typical underlying datasets.

2.1 Introduction

Topological data analysis (TDA) encapsulates a range of data analysis methods which investigate the topological structure of a dataset [29]. One such method, persistent homology, describes the connectivity structure of a given dataset and summarizes this information as a persistence diagram. TDA, and in particular persistence diagrams, have been employed in several studies with topics ranging from classification and clustering [90, 4, 66, 56] to the analysis of dynamical systems [65, 78, 42, 77] and complex systems such as sensor networks [25, 94, 11]. In this work, we establish the probability density function (pdf) for a random persistence diagram.

Persistence diagrams offer a topological summary for a collection of a-dimensional data, say $\{x_i\} \subset \mathbb{R}^a$, which focuses on the global geometric structure of the data. A persistence diagram is a collection of homological features $\{(b_i, d_i, k_i)\}$, each representing a k_i-dimenional hole which appears at scale $b_i \in \mathbb{R}^+$ and is filled at scale $d_i \in (b_i, \infty)$. In general, the dataset arises from any metric space, though restricting to $\{x_i\} \subset \mathbb{R}^a$ guarantees $k_i \in \{0, ..., a-1\}$. For example, if the data form a time series trajectory $x_i = f(t_i)$, the associated persistence diagram describes multistability through a corresponding number of persistent 0-dimensional features or periodicity through a single persistent 1-dimensional feature. In a typical persistence diagram, few features exhibit long persistence (range of scales $d_i - b_i$), and such features describe important topological characteristics of the underlying dataset. Moreover, persistent features are stable under perturbation of the underlying dataset [21].

Persistence diagrams have recently seen intense active research, including significant successful effort toward facilitating previously challenging computations with them; these efforts impact evaluation of Wasserstein distance in [49] and the creation of persistence diagrams with packages such as Dionysus [34] and Ripser [9] which take advantage of certain properties of simplicial complexes [19]. Recently, various approaches have defined specific summary statistics such as center and variance [13, 61, 89, 57], birth and death estimates [32], and confidence sets [35]. Here we introcuce a nonparametric method to construct density functions for a distribution of persistence diagrams. The development of these densities

offers a consistent framework to understand the above summary statistic results through a single viewpoint.

We naturally think of a (random) persistence diagram as a random element which depends upon a stochastic procedure which is used to generate the underlying dataset that it summarizes. Thus, sample datasets yield sample persistence diagrams without direct access to the distribution of persistence diagrams. In this sense, a distribution of persistence diagrams is defined by transforming the distribution of underlying data under the process used to create a persistence diagram, as discussed in [61]. The persistence diagrams are created through a complex and nonlinear process which relies on the global geometric arrangement of datapoints (see Section 2.2); thus, the structure of a persistence diagram distribution remains unclear even for underlying data with a well-understood distribution. Indeed, known results for the persistent homology of noise alone, such as [5], primarily concern the asymptotics of feature cardinality at coarse scale. With little previous knowledge, we study these distributions through nonparametric means. Kernel density estimation is a well known nonparametric technique for random vectors in $\mathbb{R}^a$ [75]; however, persistence diagrams lack a vector space structure and thus these techniques cannot be applied directly here.

The studies [13] and [35] work with kernel density estimation on the underlying data to estimate a target diagram as the number of underlying datapoints goes to infinity. In both cases, the target diagram is directly associated to the probability density function (pdf) of the underlying data (via the superlevel sets of the pdf). The first work constructs an estimator for the target diagram, while the second defines a confidence set. In either case, kernel density estimation is used to approximate the pdf of the underlying datapoints, assuming the data are independent and identically distibuted (i.i.d.). In contrast, our work considers a different kind of kernel density to estimate probability densities for a random persistence diagram, with convergence as the number of persistence diagrams goes to infinity. In addition, since we treat the entire underlying dataset as a random element, we need not consider the underlying datapoints to be sampled i.i.d. from a fixed distribution.

The features (b_i, d_i, k_i) in a persistence diagram come without an ordering and their cardinality is variable, being bounded but not defined by the cardinality of the underlying

dataset. Thus, any notion of persistence diagram density must be (i) invariant to the ordering of features and (ii) account for variability in their cardinality. Indeed, the approach used to analyse a collection of persistence diagrams in [11] is a good step toward understanding a random persistence diagram, but requires a choice of order and considers only a fixed number of features and is therefore unsuitable for creating probability densities. In this work, we offer a kernel density with the desireable properties (i) and (ii) which also calls attention to the persistence of each feature, focusing on features with long persistence in order to combat the curse of dimensionality. Simultaneously, the kernel density still considers features of short persistence, but simplifies their treatment in order to facilitate computation. The kernel density is defined on a pertinent space of finite random sets which is equipped to describe pdfs for random persistence diagrams generated from associated data with bounded cardinality of topological features. In this sense, our kernel density (and subsequent kernel density esimation) does not relate directly to the underlying data, but rather the distribution of persistence diagrams which describes the geometry of the random underlying dataset. The requirement of bounded feature cardinality is trivially satisfied for datasets with bounded cardinality, which is reasonable for application and theory. Indeed, the creation of a persistence diagram from an infinite collection of data is often nonsensical (e.g., for anything with unbounded noise), and a scaling limit should be considered instead.

We establish this kernel density estimation problem through the lens of finite set statistics and we consequently begin with relevant background in topological data analysis in Section 2.2 and finite set statistics in Section 2.3. For further details about these two subjects, the reader may refer repectively to [29] and [58]. Our results are presented in Section 2.4. In Subsection 2.4.1, we construct the kernel density associated to a center persistence diagram and kernel bandwidth parameter. This consists of decomposing the center persistence diagram into lower and upper halves, finding pdfs associated to each half, and lastly determining the pdf for their union. After the kernel density is introduced and an explicit pdf is delivered, its convergence is presented in Theorem 2.1 along with a detailed proof. In Subsection 2.4.3, we define the mean absolute deviation as a measure of dispersion, and show the convergence of its kernel density estimator. Next, Section 2.5 presents in detail a simple example of the kernel density. Additionally, an example of persistence diagram kernel

density estimation is demonstrated for underlying data with a typical annular distribution. Finally, we end with conclusions and a discussion in Section 2.6.

2.2 Topological Data Analysis Background

The topological background discussed here builds toward the definition of persistence diagrams, the pertinent objects in this work. We begin by briefly discussing simplicial complexes and homology, an algebraic descriptor for coarse shape in topological spaces. In turn, persistent homology, and its summary, persistence diagrams, are techniques for bringing the power and convenience of homology to describe subspace filtrations of topological spaces.

We first consider topological spaces of discernable dimension, called manifolds.

Definition 2.2.1. *A topological space X is called a k-dimensional manifold if every point $x \in X$ has a neighborhood which is homeomorphic to an open neighborhood in $\mathbb{R}^k$.*

We generalize the fixed-dimension notion of a manifold in order to define simplicial homology for simplicial complexes. We then discuss the Čech construction which is used to associate simplicial complexes to datasets.

Definition 2.2.2. *A k-simplex is a collection of $k+1$ linearly independent vertices along with all convex combinations of these vertices:*

$$(v_0, ..., v_k) = \left\{ \sum_{i=0}^{k} \alpha_i v_i \Big| \sum_{i=0}^{k} \alpha_i = 1 \text{ and } \alpha_i \geq 1 \, \forall i \right\}. \tag{2.2.1}$$

Topologically, a k-simplex is treated as a k-dimensional manifold (with boundary). An oriented simplex is typically described by a list of its vertices, such as (v_0, v_1, v_2). The faces of a simplex consist of all the simplices built from a subset of its vertex set; for example, the edge (v_1, v_2) and vertex (v_2) are both faces of the triangle (v_0, v_1, v_2).

Definition 2.2.3. *A simplicial complex $\mathcal{K}$ is a collections of simplices wherein*
(i) if $\sigma \in \mathcal{K}$, then all its faces are also in $\mathcal{K}$, and
(ii) the intersection of any pair of simplexes in $\mathcal{K}$ is another simplex in $\mathcal{K}$.
We denote the collection of k-simplices within $\mathcal{K}$ by $\mathcal{K}^{[k]}$.

A simplicial complex is realized by the union of all its simplicies; an example is shown in Fig. 2.1. Conditions (i) and (ii) in Defn. 2.2.3 establish a unique topology on the realization of a simplicial complex which restricts to the subspace topology on each open simplex. This topology is also consistent with the Euclidean subspace topology if the vertices are in $\mathbb{R}^a$.

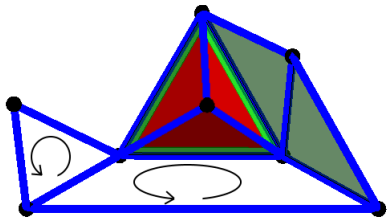

Figure 2.1: An example of a simplicial complex realized in $\mathbb{R}^3$. This particular complex has one connected component and two cycles, which generate the 0-homology and 1-homology groups, respectively. The other homology groups are trivial.

Here we define the homology groups for a simplicial complex through purely combinatorial means, which allows for automated computation.

Definition 2.2.4. *The chain group (over $\mathbb{Z}$) on a simplicial complex $\mathcal{K}$ of dimension k is denoted by $C_k(\mathcal{K})$ and is defined as formal sums of k-simplices in $\mathcal{K}$:*

$$C_k(\mathcal{K}) = \left\{ \sum_{\sigma \in \mathcal{K}^{[k]}} n_\sigma \sigma \middle| n_\sigma \in \mathbb{Z} \right\}. \tag{2.2.2}$$

Definition 2.2.5. *The k-th boundary map is a homomorphism $\partial_k : C_k(\mathcal{K}) \to C_{k-1}(\mathcal{K})$ defined on each simplex as an alternating sum over the faces of one dimension less:*

$$\partial_k(v_0, ..., v_k) = \sum_{n=0}^{k} (-1)^n (v_0, ..., v_{n-1}, v_{n+1}, ..., v_k). \tag{2.2.3}$$

Remark 2.2.1. *Chain groups give an algebraic way to describe subsets of simplices as a formal sum. Toward this viewpoint, the chain group is often defined over $\mathbb{Z}_2 = \{0, 1\}$ instead of $\mathbb{Z}$. In this case, the boundary maps can be understood classically; e.g., the boundary of a triangle yields (the sum of) its three edges and the boundary of an edge yields (the sum of) its endpoints. When viewed over $\mathbb{Z}$, the presence of sign specifies simplex orientation.*

Putting chain groups of every dimension together along with the boundary maps successively defined between them, we obtain a chain complex:

$$\{0\} \xleftarrow{\mathbf{0}} C_0(\mathcal{K}) \xleftarrow{\partial_1} C_1(\mathcal{K}) \xleftarrow{\partial_2} C_2(\mathcal{K}) \xleftarrow{\partial_3} ... \tag{2.2.4}$$

The composition of subsequent boundary maps yields the trivial map [29]; this property is typically rephrased as $\text{im}(\partial_{k+1}) \subset \ker(\partial_k)$ which enables definition of the following modular groups.

Definition 2.2.6. *The homology group of dimension k is given by*

$$H_k(\mathcal{K}) = \ker(\partial_k)/\,\text{im}(\partial_{k+1}) = \big\{[x] = x + \text{im}(\partial_{k+1}) \big| x \in \ker(\partial_k)\big\}, \tag{2.2.5}$$

where $[x] = \{x + y | y \in \text{im}(\partial_{k+1})\}$ defines the coset equivalence class of x.

The generators of the homology group correspond to topological features of the complex $\mathcal{K}$; for example, generators for the 0-homology group correspond to connected components, generators of 1-homology group correspond to holes in $\mathcal{K}$, etc. The interpretation of these features is exemplified by taking the topological boundary of a $k+1$ ball (that is, a k-sphere); for example, the boundary of an interval is two (disconnected) points while the boundary of a disc is a loop.

We wish to extend the notion of homology for a discrete set of data $\boldsymbol{x} = \{x_i\}_{i=1}^N$ within a metric space (X, d_X). Treating the set itseslf as a simplicial complex, its homology yields only the cardinality of the data points. So, we utilize the metric to obtain more information. Here we denote by $B(x_0, r_0)$ a metric ball centered at x_0 of radius r_0. Fix a radius $r > 0$ and consider the collection of neighborhoods $U = \{U_i\} = \{B(x_i, r)\}$ along with its union $\mathcal{U}_r$. The filtration of sets $\{\mathcal{U}_r\}_{r\in\mathbb{R}^+}$ naturally yields information about the arrangement within X of the dataset $\boldsymbol{x}$ at various scales. To make homology computations more tractable for $\mathcal{U}_r$, we instead consider the associated nerve complexes.

Definition 2.2.7. *The nerve $\mathcal{N}(U)$ of a collection of open sets U is the simplicial complex where a k-simplex $(v_{i_0}, ..., v_{i_k})$ is in $\mathcal{N}(U)$ if and only if $\cap_{j=0}^k U_{i_j} \neq \emptyset$. The nerve of the*

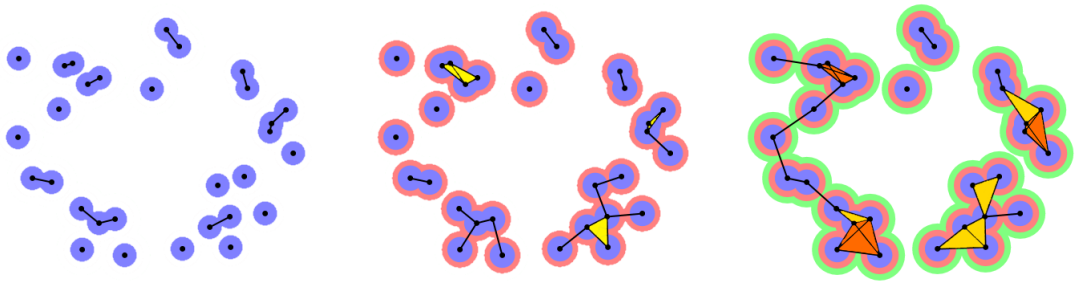

Figure 2.2: The neighborhood space and Čech complex of matching radius plotted at three different radii. Yellow indicates a triangle while orange indicates a tetrahedron. This family of simplicial complexes is the filtration utilized to compute and define persistent homology.

neighborhoods $U = \{B(x_i, r)\}$ *is called the Čech complex on the data* $\{x_i\}$ *at radius* r *and is denoted by* $\check{C}ech(\boldsymbol{x}, r)$.

Examples of the Čech complex for the same data at different radii are depicted in Fig. 2.2, where they are superimposed with the associated neighborhood space. Any nerve complex trivially satisfies the requirements for a simplicial complex [29]. Moreover, the nerve theorem states that the nerve and union of a collection of convex sets have similar topology (they are homotopy equivalent) [43]; specifically, the Čech complex and neighborhood space $\mathcal{U}$ have identical homology for any given radius.

A priori, it is unclear which choice of scale (radius), best describes the data; and oftentimes different scales reveal different information. Thus, to investigate the topology of our data, we consider the appearance and disappearance of homological features at growing scale. This multiscale viewpoint, called persistent homology, is introduced in [30] and yields a topological summary of the data called a persistence diagram. This is possible because we have a growing filtration of complexes, so each complex is included in the next (see Fig. 2.2). These inclusion maps induce inclusion maps at the chain group level and in turn induce maps (though not typically inclusions) at the level of homology groups. These induced maps $f_{r_1,r_2} : H_k(\check{C}ech(\boldsymbol{x}, r_1)) \to H_k(\check{C}ech(\boldsymbol{x}, r_2))$ are referred to here as the persistence maps, and take features to features (i.e., generators to generators) or to zero [20]. Thus, each feature is tracked by how far the persistence maps preserve it. In turn, tracking features is boiled down to a very specific algorithm for obtaining the birth and death radii for each homological feature (e.g., see [29]). Features which persist over a large range of scale are

typically considered more important, and their presence is stable under small perturbations of the underlying data [21].

Persistent homology yields a collection of features, each described by its birth and death radii in the filtration of Čech complexes along with its topological dimension; or, in short it yields a persistence diagram $\mathcal{D} = \{\xi_i\}_{i=1}^M = \{(b_i, d_i, k_i)\}_{i=1}^M$. Here, we interpret the birth-death values as coordinate points with homological dimension as labels. Example persistence diagrams are shown in Fig. 2.3. Specifcally, for data in $\mathbb{R}^a$, we consider each feature as an element of

$$\mathcal{W}_{0:a-1} = W \times \{0, ..., a-1\}, \tag{2.2.6}$$

where $W = \{(b, d) \in \mathbb{R}^2 : d > b \geq 0\}$ is the infinite wedge. As a topological space, the a-fold multiwedge $\mathcal{W}_{0:a-1}$ is treated as a-disconnected copies of W.

It is desireable to define a metric between persistence diagrams with which to measure topological similarity. One may consider Hausdorff distance, which is commonly used to compare sets of points in a metric space.

Definition 2.2.8. *Consider two subsets U and V within a metric space (X, d_X). The Hausdorff distance between them is $d_H(U, V) = \max\{H_+(U, V), H_+(V, U)\}$ where*

$$H_+(U, V) = \inf\{\epsilon \geq 0 | V \subset (U + \epsilon)\}$$

and $U + \epsilon = \cup_{x \in U} B(x, \epsilon)$ is the ϵ-neighborhood of U.

Hausdorff distance is well defined between persistence diagrams, but it does not properly account for cardinality nor the unstable presence of features near the diagonal $\Delta = \{(b, d) \in \mathbb{R}^2 | b = d\}$. For example, extra features may be produced at the diagonal under very small perturbations in the underlying dataset, as seen in Fig. 2.3. Nevertheless, we use Hausdorff distance to directly compare datasets. We consider the following metric to compare persistence diagrams.

Definition 2.2.9. *The bottleneck distance between two persistence diagrams D_1 and D_2 is given by*

$$W_\infty(D_1, D_2) = \min_\gamma \max_{x \in D_1} \|x - \gamma(x)\|_\infty. \tag{2.2.7}$$

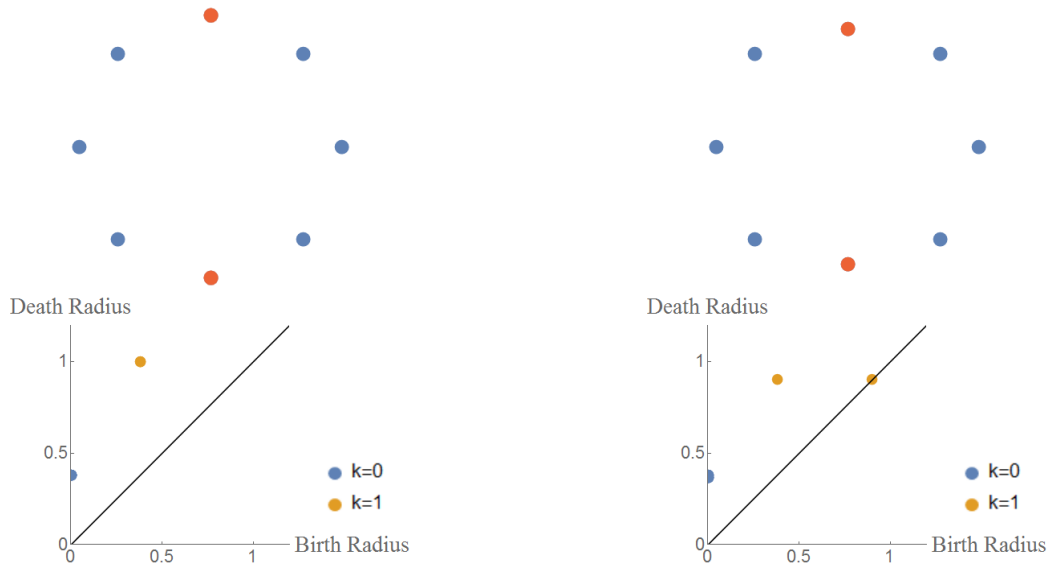

Figure 2.3: Top: Two datasets which are very close in Hausdorff distance. Bottom: The corresponding persistence diagrams, which are far in Hausdorff distance (Defn. 2.2.8) but are very close in the bottleneck distance (Defn. 2.2.9).

where γ ranges over all possible bijections between D_1 and D_2 which match homological dimension. The diagonal $\{b = d\}$ is included in both persistence diagrams with infinitely multiplicity so that any feature may be matched to the diagonal.

Remark 2.2.2. *Due to the unstable presence of features near the diagonal, typical metrics on persistence diagrams such as the bottleneck distance treat the diagonal as part of every persistence diagram [61] in order to achieve stability with respect to perturbations of the underlying dataset. Morally, one considers the diagonal as representing vacuous features which are born and die simultaneously. For convenient computation, the definition of bottleneck distance can be applied to each homological dimension separately.*

2.3 Random Persistence Diagrams

A persistence diagram changes its feature cardinality under small perturbation of the underlying dataset (see Fig. 2.3), and these features have no intrinsic order. Consequently, we cannot treat persistence diagrams as elements of a vector space. Instead, we consider a random persistence diagram D as a random multiset of features $D = \{\xi_i\} \subset \mathcal{W}_{0:a-1}$ in the

multiwedge defined in Eq. (2.2.6). For underlying datasets sampled from $\mathbb{R}^a$ with bounded cardinality, the affiliated Čech persistence diagrams also have bounded feature cardinality and homological dimension. Thus, we assume that the cardinality of a random persistence diagram is bounded above by some value $|D| \leq M \in \mathbb{N}$. Sets of this type come equipped with a natural topology

Definition 2.3.1. *Consider* $\mathcal{C}_{\leq M}(\mathcal{W}_{0:a-1}) = \left\{D \subset \mathcal{W}_{0:a-1} \middle| |D| \leq M\right\}$*, the space of at most* M *topological features. The hit-or-miss topology on* $\mathcal{C}_{\leq M}$ *is defined according to the sub-basis consisting of collections of the form*

$$\mathcal{W}_{0:a-1}^{F} = \{D \in \mathcal{C}_{\leq M} | D \cap F \neq \emptyset\}$$

for each compact set $F \subset \mathcal{W}_{0:a-1}$*.*

We view $\mathcal{C}_{\leq M}(\mathcal{W}_{0:a-1})$ through a list of functions h_N which each map the appropriate dimension of Euclidean space into its corresponding cardinality component, $\mathcal{C}_N(\mathcal{W}_{0:a-1})$. This viewpoint facilitates the definition of probability densities.

Definition 2.3.2. *For each* $N \in \{0, ..., M\}$*, consider the space of* N *topological features,* $C_N(\mathcal{W}_{0:a-1}) = \left\{D \subset \mathcal{W}_{0:a-1} \middle| |D| = N\right\}$*, and the associated map* $h_N : \mathcal{W}_{0:a-1}^N \to \mathcal{C}_N(\mathcal{W}_{0:a-1})$ *defined by*

$$h_N(\xi_1, ..., \xi_N) = \{\xi_1, ..., \xi_N\}. \tag{2.3.1}$$

The hit-or-miss topology on $C_{\leq M}(\mathcal{W}_{0:a-1})$ *is defined so that each* h_N *is continuous [41]. Moreover,* h_N *creates equivalence classes on* $\mathcal{W}_{0:a-1}^N$ *according to the action of the permutations* Π_N*; specifically,* $[Z] = [(\xi_1, ..., \xi_N)]_{h_N} = \left\{\left(\xi_{\pi(1)}, ..., \xi_{\pi(N)}\right) \middle| \pi \in \Pi_N\right\}$ *for each* $Z = (\xi_1, ..., \xi_N) \in \mathcal{W}_{0:a-1}^N$*. These equivalence classes yield the space*

$$\mathcal{W}_{0:a-1}^N / \Pi_N = \left\{[\boldsymbol{\xi}]_{h_N} \middle| \boldsymbol{\xi} \in \mathcal{W}_{0:a-1}^N\right\}, \tag{2.3.2}$$

wherein the topology inherited by $\mathcal{W}_{0:a-1}^N / \Pi_N$ *is such that* h_N *lifts to a homeomorphism between* $\mathcal{W}_{0:a-1}^N / \Pi_N$ *and* $C_N(\mathcal{W}_{0:a-1})$*.*

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 $C_{\leq M}(\mathcal{W}_{0:a-1})$ 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 $\mathcal{W}_{0:a-1}^N/\Pi_N \cong C_N(\mathcal{W}_{0:a-1})$, we work toward defining probability densities on the collection $\cup_{N=0}^M \mathcal{W}_{0:a-1}^N$.

Definition 2.3.3. *For a given random persistence diagram D, the belief function β_D is given by*

$$\beta_D(A) = \mathbb{P}[D \subset A], \tag{2.3.3}$$

for any Borel set $A \subset \mathcal{W}_{0:a-1}$.

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 2.3.4. *[58] Fix ϕ defined on Borel subsets of $C_{\leq M}(\mathcal{W}_{0:a-1})$ into $\mathbb{R}$. For an element $\xi \in \mathcal{W}_{0:a-1}$ or a multiset $Z \subset \mathcal{W}_{0:a-1}$ with $Z = \{\xi_1, ..., \xi_N\}$, the set derivative (evaluated at $\emptyset$) is respectively given by*

$$\begin{aligned} \frac{\delta \phi}{\delta \xi}(\emptyset) &= \lim_{n \to \infty} \frac{\phi(B(\xi, 1/n))}{\lambda(B(\xi, 1/n))}, \\ \frac{\delta \phi}{\delta Z}(\emptyset) &= \frac{\delta^N \phi}{\delta \xi_1 ... \delta \xi_N} = \left[\frac{\delta}{\delta \xi_1} \cdots \frac{\delta}{\delta \xi_N} \phi \right](\emptyset), \end{aligned} \tag{2.3.4}$$

where λ indicates Lebesgue measure.

Remark 2.3.1. *The definition of a set derivative evaluated on a nonempty set is more involved, and is found in [58]. Here we are primarily concerned with evaluation at $\emptyset$, since this suffices for the definition of a probability density function. Also, note that set derivatives satisfy the product rule.*

Remark 2.3.2. *Restricting to a particular cardinality N, consider $\phi_N = [\phi \circ h_N]$, a function on Euclidean space which is invariant under the action of Π_N. The viewpoint of ϕ_N elucidates the relationship between set derivatives and Radon-Nikodým derivatives with*

respect to Lebesgue measure. This viewpoint also shows that the iterated derivative given in Eq. (2.3.4) is independent of order and thus is well-defined for a multiset Z.

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

Definition 2.3.5. *Consider a subset $A \subset \mathcal{W}_{0:a-1}$ or a collection $O \subset C_{\leq M}(\mathcal{W}_{0:a-1})$ of Borel sets. For a set function $f : C_{\leq M}(\mathcal{W}_{0:a-1}) \to \mathbb{R}$, its set integrals over A and O are respectively given by*

$$\int_A f(Z)\delta Z = \sum_{N=0}^{M} \frac{1}{N!} \int_{A^N} f(h_N(\xi_1, ..., \xi_N)) d\xi_1 ... d\xi_N, \tag{2.3.5a}$$

$$\int_O f(Z)\delta Z = \sum_{N=0}^{M} \frac{1}{N!} \int_{h_N^{-1}(O)} f(h_N(\xi_1, ..., \xi_N)) d\xi_1 ... d\xi_N, \tag{2.3.5b}$$

where $Z = \{\xi_1, ..., \xi_N\} \subset \mathcal{W}_{0:a-1}$ is a persistence diagram.

Dividing by $N!$ accounts for integrating over $\mathcal{W}_{0:a-1}^N$ instead of $\mathcal{W}_{0:a-1}^N / \Pi_N \cong C_N(\mathcal{W}_{0:a-1})$.

It has been shown that set derivatives and integrals are inverse operations [58]; specifically, the set derivative of a belief function yields a probability density for a random diagram D such that

$$\mathbb{P}[D \subset A] = \beta_D(A) = \int_A \frac{\delta \beta_D}{\delta Z}(\emptyset)\delta Z. \tag{2.3.6}$$

Indeed, $A^N = h_N^{-1}(\{D \subset A\})$ so that Eq. (2.3.6) also holds as an integral over $O_A = \{D \subset A\}$ as in Eq. (2.3.5b). The sets O_A form a basis for the hit-or-miss Borel sets and consequently extend Eq. (2.3.6) to express $\mathbb{P}[D \in O]$ in terms of $\frac{\delta \beta_D}{\delta Z}(\emptyset)$ for any Borel measureable O.

Definition 2.3.6. *For a random persistence diagram D, a global probability density function (global pdf) $f_D : \cup_{N \in \mathbb{N}} \mathcal{W}_{0:a-1}^N \to \mathbb{R}$ is given by layered restrictions $f_N = f_D\big|_{\mathcal{W}_{0:a-1}^N} : \mathcal{W}_{0:a-1}^N \to \mathbb{R}$ for each N which satisfy*

$$\frac{\delta^N \beta_D}{\delta \xi_1 \cdot ... \cdot \delta \xi_N}(\emptyset) = \sum_{\pi \in \Pi_N} f_D(\xi_{\pi(1)}, ..., \xi_{\pi(N)}). \tag{2.3.7}$$

Remark 2.3.3. *It is necessary to make a distinction between local and global densities because the global pdf is not defined on a single Euclidean space, and is instead expressed as a collection of densities over a range of dimensions. Specifically, while each local density f_N (for input cardinlity N) is defined on $\mathcal{W}^N_{0:a-1}$, the global pdf f_D is defined on $\cup_{N=1}^{M}\mathcal{W}^N_{0:a-1}$ and restricts to a local density on each input dimension.*

Each local density $f_N(Z) = f_D\big|_{\mathcal{W}^N_{0:a-1}}(Z)$ decomposes into the product of the probability density associated with $\mathbb{P}[D = Z\,||D| = N]$ and the cardinality probability $\mathbb{P}[|Z| = N]$ (this follows from Prop. 2.3.7). Thus, each local density does not integrate to one, but instead to the associated probability $\mathbb{P}[|Z| = N]$. Also, the global pdf is not a set function and does not require division by $N!$, leading to the following relation: $\int_{A^N} f_D(\xi_1, ..., \xi_N)d\xi_1...d\xi_N = \frac{1}{N!}\int_{A^N}\frac{\delta^N\beta_D}{\delta^N Z}(\emptyset)d\xi_1...d\xi_N$.

Remark 2.3.4. *While the global pdf and its local constituents need not be symmetric with respect to Π_N, there is a unique choice of global pdf (up to sets of Lebesgue measure 0) which satisfies Eq. (2.3.7) and is symmetric under the action of Π_N. In this case, we safely abuse notation by denoting $f_D(\{\xi_1, ..., \xi_N\}) := N!f_D(\xi_1, ..., \xi_N)$ and often write $f_D(Z)$ and allow context to determine whether Z denotes a set or a vector.*

The following proposition is critical to determine the global pdf for (i) the union of independent singleton diagrams (i.e., $|D^j| \leq 1$), (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 [55] (Theorem 17, pp 155–156).

Proposition 2.3.7. *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*

$$\beta_D(S) = a_0 + \sum_{m=1}^{M} a_m q_m(S^m),$$

where $a_m = \mathbb{P}(|D| = m)$ and $q_m(S^m) = \mathbb{P}(D \subset S\big|\,|D| = m)$.

Remark 2.3.5. *The decomposition in Prop. 2.3.7 is often applied as a first step toward finding the local density constituents of the global pdf. In particular,* $f_N = f_D\big|_{\mathcal{W}^N_{0:a-1}} = 0$ *for* $N > M$.

Lastly, we encounter a computationally convenient summary for a random persistence diagram called probability hypothesis density:

Definition 2.3.8. *[58] 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*

$$F_D(a) = \int_{\{Z|\{a\}\subset Z\}} \frac{\delta\beta}{\delta Z}(\emptyset)\delta Z. \tag{2.3.8}$$

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

2.4 Kernel Density Estimation

2.4.1 Construction

To characterize distributions of persistence diagrams, our first goal is the creation of a kernel density function about a center persistence diagram $\mathcal{D}$ with a kernel bandwidth parameter $\sigma > 0$. Prop. 2.3.7 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. Singletons are indexed by superscripts because subscripts are used in Theorem 2.1 to index the sample diagrams used to build a kernel density estimate.

Lemma 2.4.1. *Consider a collection of independent singleton random persistence diagrams* $\{D^j\}_{j=1}^M$. *If each singleton* D^j *is described by the value* $q^j = \mathbb{P}[D^j \neq \emptyset]$ *and the subsequent pdf,* $p^j(\xi) = \mathbb{P}\left[D^j = \{\xi\} \,\middle|\, |D^j| = 1\right]$, *then the global pdf for* $D = \cup_{j=1}^M D^j$ *is given by*

$$f_D(\xi_1, ..., \xi_N) = \sum_{\gamma \in I(N,M)} \mathcal{Q}(\gamma) \prod_{k=1}^{N} p^{\gamma(k)}(\xi_k), \tag{2.4.1}$$

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

$$\mathcal{Q}(\gamma) = \mathcal{Q}^*(\gamma) \prod_{k=1}^{N} q^{\gamma(k)}, \tag{2.4.2}$$

$I(N, M)$ *consists of all increasing injections* $\gamma : \{1, ..., N\} \to \{1, ..., M\}$, *and*

$$\mathcal{Q}^*(\gamma) = \frac{\prod_{j=1}^{M}(1 - q^j)}{\prod_{k=1}^{N}(1 - q^{\gamma(k)})}. \tag{2.4.3}$$

Proof. Since the singleton events D^j are independent, the belief function for $D = \cup_j D^j$ decomposes into $\beta_D(S) = \prod_{j=1}^{M} \beta_{D^j}(S)$. Next, we employ the product rule for the set derivative (see Defn. 2.3.4) to obtain the global pdf for D in terms of the singleton belief functions and their first derivatives. Higher derivatives of β_{D^j} are zero since D^j are singletons (see Remark 2.3.5). Thus, the product rule yields first derivatives on all (ordered) subsets of the singleton belief functions:

$$\frac{\delta^N \beta_D}{\delta\xi_1 ... \delta\xi_N}(\emptyset) = \sum_{1 \le j_1 \ne, ..., \ne j_N \le M} \frac{\beta_{D^1}(\emptyset) \cdots \beta_{D^M}(\emptyset)}{\beta_{D^{j_1}}(\emptyset) \cdots \beta_{D^{j_N}}(\emptyset)} \left[\frac{\delta \beta_{D^{j_1}}}{\delta \xi_1}(\emptyset) \cdots \frac{\delta \beta_{D^{j_N}}}{\delta \xi_N}(\emptyset) \right].$$

By Prop. 2.3.7, we have that $\beta_{D^j}(\emptyset) = (1 - q^j)$ and $\frac{\delta \beta_{D^{j_i}}}{\delta \xi_i}(\emptyset) = q_{j_i} p^{j_i}(\xi_i)$ and so

$$\frac{\delta^N \beta_D}{\delta\xi_1 ... \delta\xi_N}(\emptyset) = \sum_{1 \le j_1 \ne, ..., \ne j_N \le M} \left[\frac{\prod_{j=1}^{M}(1 - q^j)}{\prod_{j=1}^{N}(1 - q^{j_k})} \prod_{k=1}^{N} q^{j_k} \right] \prod_{k=1}^{N} p^{j_k}(\xi_k),$$

which nearly resembles Eq. (2.4.1). To bridge the gap, we describe the choice of indices j_i by an injective function from $\{1, ..., N\}$ into $\{1, ..., M\}$. In turn, each such injective function is uniquely determined by the composition of an increasing injection $\gamma \in I(N, M)$ which decides the range of the function and permutations on the domain, Π_N. These permutations take into account the order of the range. The value of $\mathcal{Q}$ is independent of order, and thus is determined by γ as in Eq. (2.4.2). We reorder the product in order to shift these

permutations onto the input variables, obtaining

$$\frac{\delta^N \beta_D}{\delta\xi_1 ... \delta\xi_N}(\emptyset) = \sum_{\pi\in\Pi_N} \sum_{\gamma\in I(N,M)} \mathcal{Q}(\gamma) \prod_{k=1}^{N} p^{\gamma(k)}(\xi_{\pi(k)}). \tag{2.4.4}$$

Finally, the global pdf in Eq. (2.4.1) follows directly from applying Eq. (2.3.7) to Eq.(2.4.4).

■

Remark 2.4.1. *The global pdf in Eq. (2.4.1), and in particular the sum over $\gamma \in I(N, M)$, accounts for each possible combination of singleton presence. Moreover, summing over permutations as in Eq. (2.4.4) and dividing by $N!$ yields a symmetric pdf with terms for every possible assignment between singletons and inputs. The weights $\mathcal{Q}(\gamma)$ indicate the probability of each assignment occuring, and is the product of the appropriate probability for each singleton to be either present, q^j, or absent, $1 - q^j$, for each j.*

Example 2.4.1. Consider two 1-dimensional singleton diagrams, D^1 and D^2, with probabilities of being nonempty $q^1 = 0.6$ and $q^2 = 0.8$, respectively. The corresponding local densities when nonempty are given by $p^1(x) = \frac{1}{\sqrt{2\pi}} e^{-(x+1)^2/2}$ and $p^2(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-1)^2/2}$. Lemma 2.4.1 yields the global pdf for $D = D^1 \cup D^2$ through a collection of local densities $\{f_0, f_1(x), f_2(x, y)\}$ with $f_0 = \mathbb{P}[D = \emptyset] = (1-q^1)(1-q^2) = 0.08$, $f_1 = f_D\big|_{\mathbb{R}}$, and $f_2 = f_D\big|_{\mathbb{R}^2}$. We sum over permutations and divide by $N!$ ($N = 1, 2$ is the input cardinality) to obtain a symmetric global pdf.

$$\begin{aligned} f_1(x) &= (1-q^2)q^1 p^1(x) + (1-q^1)q^2 p^2(x) = \frac{0.12}{\sqrt{2\pi}} e^{-(x+1)^2/2} + \frac{0.32}{\sqrt{2\pi}} e^{-(x-1)^2/2}, \\ f_2(x,y) &= \frac{q^1 q^2}{2} \left[p^1(x)p^2(y) + p^1(y)p^2(x)\right] = \frac{0.24}{2\pi} \left(e^{-((x-1)^2+(y+1)^2)/2} + e^{-((x+1)^2+(y-1)^2)/2}\right). \end{aligned} \tag{2.4.5}$$

Accounting for each cardinality and following Eq. (2.4.5), the total probability adds up to

$$\begin{aligned} \mathbb{P}[|D| = 0] + \mathbb{P}[|D| = 1] + \mathbb{P}[|D| = 2] &= f_0 + \int_{\mathbb{R}} f_1(x)dx + \int_{\mathbb{R}^2} f_2(x,y)dxdy \\ &= (0.08) + (0.12 + 0.32) + (0.24 + 0.24) = 1, \end{aligned}$$

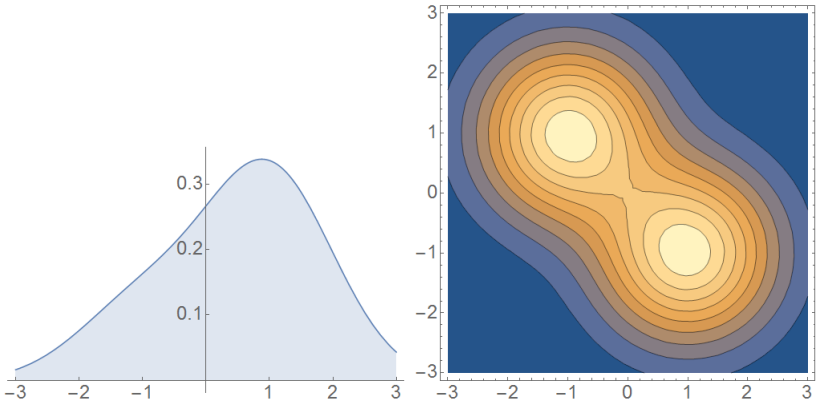

Figure 2.4: Plots for the local densities shown in Eq. (2.4.5). These pdfs cover the different possible input dimensions and are symmetric under permutations of the input.

as desired. The local densities in Eq. (2.4.5) are plotted in Fig. 2.4.

Now we turn toward defining the kernel density. We first define a random persistence diagram as a union of simpler constituents, and then determine its global pdf by combination in a fashion similar to Lemma 2.4.1. Indeed, we define the desired kernel density as the global pdf for this composite random diagram. To start, we fix a homological dimension k and consider a center diagram $\mathcal{D} \subset \mathcal{W}_k$. Since k is fixed, we treat $\mathcal{D} = \{\xi_i\}_{i=1}^M = \{(b_i, d_i)\}_{i=1}^M$ within $W = \{(b, d) \in \mathbb{R}^2 : d > b \geq 0\}$.

Long persistence points in a persistence diagram represent prominent topological features which are stable under perturbation of underlying data, and so it is important to track each independently. In contrast, we leverage the point of view that the small persistence features near the diagonal are considered together as a single geometric signature as opposed to individually important topological signatures. Toward this end, features with short persistence are grouped together and interpreted through i.i.d. draws near the diagonal. Treating these points collectively greatly speeds computation (see Fig. 2.5), while simultaneously retaining crucial geometric information for applications such as classification. Thus, we split $\mathcal{D}$ into upper and lower portions according to the bandwidth σ as

$$\mathcal{D}^u = \{(b_i, d_i, k) \in \mathcal{D} : d_i - b_i \geq \sigma\} \text{ and } \mathcal{D}^\ell = \{(b_i, d_i, k) \in \mathcal{D} : d_i - b_i < \sigma\}. \tag{2.4.6}$$

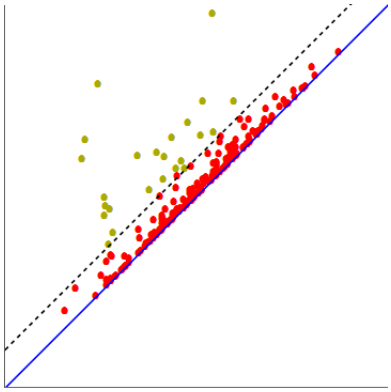

Figure 2.5: A persistence diagram with points below and above the cutoff (dashed black line $d = b + \sigma$). In a typical diagram, there are many points close to the diagonal and tracking these points individually requires a lot of computation.

Now define random diagrams D^u centered at $\mathcal{D}^u$ and D^ℓ centered at $\mathcal{D}^\ell$ and subsequently define the random persistence diagram $D = D^u \cup D^\ell$ according to their union. Ultimately, the global pdf of D centered at $\mathcal{D}$ is our kernel density.

Definition 2.4.2. *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. (2.4.8)) along with its potential position (b, d) sampled according to a modified Gaussian distribution $\mathcal{N}^*((b_j, d_j), \sigma I)$. The global pdf for D^u is then determined by Lemma 2.4.1, where each p^j is given by the pdf associated with $\mathcal{N}^*((b_j, d_j), \sigma I)$,*

$$p^j(b, d) = \frac{\varphi_j(b, d)}{\int_W \varphi_j(u, v) du\, dv} \mathbb{1}_W(b, d), \tag{2.4.7}$$

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

The global pdf for each D^j is readily obtained by a pair of restrictions. First, we restrict the usual Gaussian distribution to the halfspace $T = \{(b, d) \in \mathbb{R}^2 : b < d\}$. Features sampled below the diagonal are considered to disappear from the diagram and thus we define the chance to be nonempty by

$$q^j = \mathbb{P}(D^j \neq \emptyset) = \int_{\{v>u\}} \varphi_j(u, v)\, du\, dv. \tag{2.4.8}$$

Afterward, the Gaussian restricted to T is further restricted to W and renormalized to obtain a probability measure as in Eq. (2.4.7). This double restriction to both T and W is necessary for proper restriction of the Gaussian pdf and definition of $q^j = \mathbb{P}(\mathcal{D}_j \neq \emptyset)$. Indeed, restriction to W alone causes points with small birth time to have an artificially high chance to disappear; while restriction to T alone nonsensically yields features born before the end of time ($b < 0$). In kernel density estimation, the effects of this distinction become negligible as the bandwidth goes to zero. In practice, this distinction is imporant for features with small birth time relative to the bandwidth.

Remark 2.4.2. *In the Čech construction of a persistence diagram, a feature lies on the line $b = 0$ if and only if it has homological dimension $k = 0$. Consequently, for a feature $(0, d_j)$ with $k = 0$, we instead take*

$$p^j(d) = \frac{\phi_j(d)}{\int_{\mathbb{R}^+} \phi_j(u)\, du} \mathbb{1}_{\mathbb{R}^+}(d) \text{ and } q^j = \int_{\mathbb{R}^+} \phi_j(u)\, du$$

where ϕ_j is the 1-dimensional Gaussian centered at d_j with standard deviation σ.

Whereas the large persistence features in D^u have small chance to fall below the diagonal and disappear, the existence of the small persistence features in D^ℓ is volatile: these features disappear and appear fluidly under small changes in the underlying data. The distribution of D^ℓ is described by a probability mass function (pmf) ν and lower density p^ℓ.

Definition 2.4.3. *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 $\mathcal{D}^\ell$ onto the diagonal $b = d$, then creating a restricted Gaussian kernel density estimation for these features; specifically,*

$$p^\ell(b, d) = \frac{1}{N_\ell} \sum_{(b_i, d_i) \in \mathcal{D}^\ell} \frac{1}{\pi\sigma^2} e^{-\left(\left(\frac{b_i + d_i}{2} - b\right)^2 + \left(\frac{b_i + d_i}{2} - d\right)^2\right)/2\sigma^2}. \tag{2.4.9}$$

By Prop. 2.3.7 and Eq. (2.3.7), global pdfs of random persistence diagrams are described by a random vector pdf for each cardinality layer, resulting in the following global pdf for

D^ℓ:

$$f_{D^\ell}(\xi_1, ..., \xi_N) = \nu(N) \prod_{j=1}^{N} p^\ell(\xi_j). \tag{2.4.10}$$

Combining the expressions for D^ℓ and D^u, we arrive at the following proposition.

Proposition 2.4.4. *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. (2.4.6). Define D^ℓ with global pdf from Eq. (2.4.10), and D^u with global pdf from Eq. (2.4.1). Treating the random persistence diagrams D^u and D^ℓ as independent, the kernel density centered at $\mathcal{D}$ with bandwidth σ (i.e., the global pdf for the random persistence diagram $D = D^u \cup D^\ell$) is given by*

$$K_\sigma(Z, \mathcal{D}) = \sum_{j=0}^{N_u} \nu(N-j) \sum_{\gamma \in I(j, N_u)} \mathcal{Q}(\gamma) \prod_{k=1}^{j} p^{\gamma(k)}(\xi_k) \prod_{k=j+1}^{N} p^\ell(\xi_k) \tag{2.4.11}$$

where $Z = (\xi_1, ..., \xi_N)$ is the input and $N_u = |\mathcal{D}^u|$ depends on both $\mathcal{D}$ and σ. Here $\mathcal{Q}(\gamma)$ is given by Eq. (2.4.2), each p^j refers to the modified Gaussian pdf as shown in Eq. (2.4.7) for its matching feature ξ_j in D^u, and p^ℓ is given by Eq. (2.4.9).

Proof. Since D^u and D^ℓ are independent random persistence diagrams, the belief function decomposes into $\beta_D(S) = \beta_{D^u}(S)\beta_{D^\ell}(S)$. Moreover, since derivatives above order N_u vanish for β_{D^u} (see Remark 2.3.5), the product rule and binomial-type counting yield

$$\begin{aligned} \frac{\delta^N \beta_D}{\delta\xi_1 ... \delta\xi_N}(\emptyset) &= \sum_{j=0}^{N_u} \sum_{1 \le i_1 \ne ... \ne i_j \le N} \frac{\delta^j \beta_{D^u}}{\delta\xi_{i_1} ... \delta\xi_{i_j}}(\emptyset) \frac{\delta^{N-j}\beta_{D^\ell}}{\delta\xi_1 ... \delta\hat{\xi}_{i_1} ... \delta\hat{\xi}_{i_j} ... \delta\xi_N}(\emptyset) \\ &= \sum_{\pi \in \Pi_N} \sum_{j=0}^{N_u} \frac{1}{j!(N-j)!} \frac{\delta^j \beta_{D^u}}{\delta\xi_{\pi(1)} ... \delta\xi_{\pi(j)}}(\emptyset) \frac{\delta^{N-j}\beta_{D^\ell}}{\delta\xi_{\pi(j+1)} ... \delta\xi_{\pi(N)}}(\emptyset) \end{aligned} \tag{2.4.12}$$

where $\delta\hat{\xi}_i$ indicates that the given index is skipped in the set derivative (having been allocated to the other factor). Similar to the proof of Lemma 2.4.1, the choice of indices i_j is replaced with a permutation $\pi \in \Pi_N$; however, the ordering within each derivative is unrelated the choice of i_j, leading to $j!$-fold and $(N-j)!$-fold redundancy within each term.

Taking Eq. (2.4.10) together with Eq. (2.3.7) yields

$$\frac{\delta \beta_{D^\ell}}{\delta \xi_{\pi(j+1)} \ldots \delta \xi_{\pi(N)}}(\emptyset) = (N-j)! \nu(N-j) \prod_{j=1}^{N-j} p^\ell(\xi_j).$$

Also, Eq. (2.4.1) and Eq. (2.3.7) yield

$$\frac{\delta \beta_{D^u}}{\delta \xi_{\pi(1)} \ldots \delta \xi_{\pi(j)}}(\emptyset) = \sum_{\pi^* \in \Pi_j} \sum_{\gamma \in I(j, N_u)} \mathcal{Q}(\gamma) \prod_{k=1}^{j} p^{\gamma(k)}(\xi_{\pi^*(k)}).$$

We subsitute these relations into the final expression of Eq. (2.4.12). The first of these substitutions is straightforward, while the second has $j!$-fold redundant permutations overtop the existing permutations in Π_N. These substitutions yield that $\frac{\delta^N \beta_D}{\delta \xi_1 \ldots \delta \xi_N}(\emptyset) = \sum_{\pi \in \Pi_N} K_\sigma(Z, \mathcal{D})$ as described in Eq. (2.4.11) and shows that the kernel $K_\sigma(Z, \mathcal{D})$ satisfies the definition of a global pdf for D (Defn. 2.3.6). Finally, the sum over permutations is removed according to Eq. (2.3.7) to obtain the expression for $f_D(Z) = K_\sigma(Z, \mathcal{D})$. ■

A specific example of the component distributions provided for the kernel in Prop. 2.4.4 is presented in Fig. 2.6. The dashed black line separates the center persistence diagram $\mathcal{D}$ into an upper portion $\mathcal{D}^u$ and a lower portion $\mathcal{D}^\ell$ as in Eq. (2.4.6). In the upper region, the red and blue gradients represent the upper distributions p^1 and p^2 given by Eq. (2.4.7). In the lower region, the green gradient represents the restricted Gaussian kernel density estimate p^ℓ defined in Eq 2.4.9.

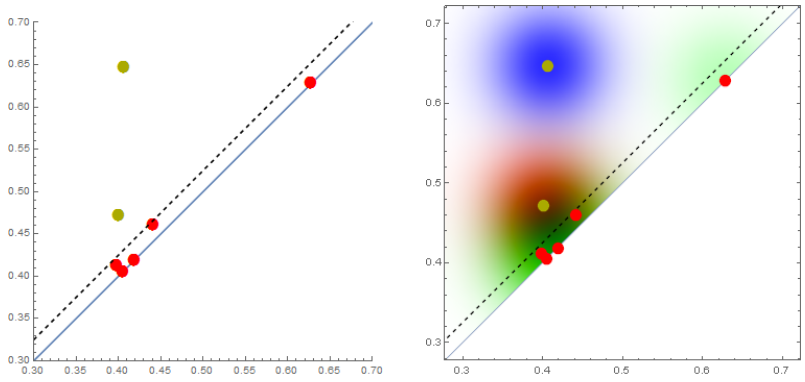

Figure 2.6: Left: a persistence diagram. Right: a depiction of its corresponding singleton (red and blue) and lower (green) distributions.

Remark 2.4.3. *Fig. 2.6 does not depict the full kernel density; a detailed example is found in Subsection 2.5. Specifically, while each of the densities p^1, p^2, and p^ℓ in shown in Fig. 2.6 is defined on W (2-dimensional), the global pdf is defined on W^N for each input-cardinality N, as in Eq. (2.4.11). Intuitively, the global pdf places priority to assign input features to the upper densities, and extra features are delegated to the lower density. The specific weight of this preference is $\dfrac{q^k}{1-q^k}$ and depends on the ratio of persistence to bandwidth.*

Prop. 2.4.4 extends to the analogue result for a center persistence diagram with features of varied homological dimension.

Corollary 2.4.5. *Consider a persistence diagram $\mathcal{D} = \bigcup_{k=0}^{a-1} \mathcal{D}_k \times \{k\}$ split according to homological dimensions with associated random persistence diagrams D_k defined according to Eq. (2.4.11) for each center diagram $\mathcal{D}_k$. Treating each D_k as independent, the full global pdf for $D = \bigcup D_k$ centered at $\mathcal{D}$ with bandwidth σ is given by*

$$K_\sigma(Z, \mathcal{D}) = \Lambda(N) \prod_{k=0}^{a-1} K_\sigma(Z_k, \mathcal{D}_k), \tag{2.4.13}$$

where $Z = \bigcup_{k=0}^{a-1} Z_k \times \{k\} \subset \mathcal{W}_{0:a-1}$ with each $Z_k \subset W$ of cardinality $|Z_k| = N_k$ within the multi-index $N = (N_0, ..., N_{a-1})$ and

$$\Lambda(N) = \frac{N!}{|N|!} := \frac{\prod N_k!}{(\sum N_k)!}.$$

Proof. The result follows immediately from taking set derivatives of the full belief function $\beta_D(S) = \prod \beta_{D_k}(S)$. In particular, the set derivatives $\frac{\delta \beta_{D_k}}{\delta Z}(\emptyset)$ are zero unless $Z \subset \mathcal{W}_k$. Thus, the product rule leaves only the single term $\frac{\delta \beta_D}{\delta Z}(\emptyset) = \prod_{k=0}^{a-1} \frac{\delta \beta_{D_k}}{\delta Z_k}(\emptyset)$. In turn, each kernel global pdf $K_\sigma(Z_k, \mathcal{D}_k)$ is related to the associated belief function derivative by a sum over permutations Π_{N_k} (see Eq. (2.3.7)). Compositions of these permutations are $N_k!$-fold redundant against the $|N|!$ permutations in $\Pi_{|N|}$, yielding the coefficient $\Lambda(N)$. ■

2.4.2 Covergence

Our primary goal is to prove the convergence (to the target distribution) of the kernel density estimate defined according to the kernel established in Prop. 2.4.4. Toward this end, consider persistence diagrams $\{D_i\}_{i=1}^n$ which are i.i.d. sampled according to a target global pdf f. To prove convergence, we require the following assumptions on f:

$(A1)$ $f(Z) = 0$ for $|Z| > M \in \mathbb{N}$ (bounded cardinality).

$(A2)$ The local density $f_N : \mathcal{W}_k^N \to \mathbb{R}$ is bounded for each $N \in \{1, ..., M\}$.

$(A3)$ There exists $C_N > 0$ so that $f(\xi_1, ..., \xi_N) \leq C_N \|(\xi_1, ..., \xi_N)\|^{-2N}$ for each $N \in \{1, ..., M\}$.

Remark 2.4.4. *These assumptions hold for typical (random) underlying datasets. For example,* $(A1)$ *holds for underlying data in* $\mathbb{R}^a$ *of bounded cardinality. The conditions* $(A2)$ *and* $(A3)$ *hold for random datasets with Gaussian noise.*

The following theorem shows that the kernel density estimate converges to the true global pdf of a persistence diagram as the number of sampled diagrams increases. The target pdf describes the distribution of geometry attained by the underlying random dataset. The pdf tracks not only the birth and death of features, but also their prevalence. In particular, the persistence diagram pdf tied to a random dataset can determine which geometric features are stable regardless of their persistence.

Theorem 2.1. *Consider a random persistence diagram global pdf* f *satisfying assumptions* $(A1)$-$(A3)$. *Define the kernel* $K_\sigma(Z, \mathcal{D})$ *according to Prop. 2.4.4 and consider the kernel density estimate* $\hat{f}(Z) = \sum_{i=1}^n K_\sigma(Z, D_i)$, *with centers* D_i *sampled i.i.d. according to* f *and bandwidth* $\sigma = O(n^\alpha)$ *chosen with* $0 < \alpha < \alpha_{2M}$. *Then, as* $n \to \infty$, $\hat{f} \to f$ *uniformly on compact subsets of* W.

Remark 2.4.5. *The value of* α_{2M} *is inherited from bandwidth selection for* $2M$*-dimensional kernel density estimates [75]. The proof for the theorem is given for* $k > 0$. *The case for* $k = 0$ *is obtained by a slight modification of the proof and the full result follows by an application of Corollary 2.4.5.*

Remark 2.4.6. *In summary, we present a direct proof for Theorem 2.1 which involves controlling the portions of Eq. (2.4.11) which change with input cardinality, such as ν, $\mathcal{Q}$, and the lower distribution p^ℓ. In particular, we show that in the limit, diagrams contribute only to the matching cardinality layer of the global pdf; that is, only persistence diagram samples with $|D_i| = m$ contribute to $\hat{f}(\xi_1, ..., \xi_m)$ approaching $f(\xi_1, ..., \xi_m)$ for each $m \in \{0, ..., M\}$. Indeed, this approach intuitively fits with the decomposition of a global pdf into local constituents via input cardinality.*

Recall Prop. 2.4.4, which defines the pertinent kernel density $K_\sigma(Z, \mathcal{D})$ evaluated at $Z = (\xi_1, ..., \xi_N)$ according to center diagram $\mathcal{D}$ and bandwidth σ by

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

where $\mathcal{Q}(\gamma)$ is given by Eq. (2.4.2), each p^j refers to the modified Gaussian pdf shown in Eq. (2.4.7) for its matching feature ξ_j in D^u, $N_u = |\mathcal{D}^u|$, and p^ℓ is given by Eq. (2.4.9). Also recall that $\mathcal{D}$ is split into $\mathcal{D}^\ell$ and $\mathcal{D}^u$ according to Eq. (2.4.6), D^ℓ is defined with global pdf from Eq. (2.4.10), and D^u is defined with global pdf from Eq. (2.4.1).

Throughout the proof we use ξ_i to denote input features and $Z = \{\xi_1, ..., \xi_N\}$ or $Z = (\xi_1, ..., \xi_N)$ to denote an input persistence diagram as a set or vector of features. Several preliminary lemmas are presented before the main body of the proof. We begin with a critical lemma which controls the number of features sampled in the band diagonal $\Delta_a^b = \{(b, d) \in W : a < d - b < b\}$.

Lemma 2.4.1. *Consider a random persistence diagram D distributed according to f satisfying assumptions (A1)-(A3). Then there exists $C > 0$ so that $\mathbb{E}^f(|\Delta_0^\sigma \cap D|) \leq C\sigma$.*

Proof. Consider a region $A \subset W$ and a counting function $\kappa_A(Z) = |Z \cap A|$ such that $\kappa_A(\{\xi_1, ..., \xi_N\}) = \sum_{i=1}^N \mathbb{1}_A(\xi_i)$. It is clear that this set function is well defined and measureable if A is measureable. Using set integration (Defn. 2.3.5),

$$\mathbb{E}(|\Delta_0^\sigma \cap D|) = \int_W \kappa_{D_0^\sigma}(Z) f(Z) \delta Z = \sum_{N=0}^{M} \frac{N}{N!} \int_W \mathbb{1}_{\Delta_0^\sigma}(\xi_1) \left[\int f(\xi_1, ...\xi_N) d\xi_2 ... d\xi_N \right] d\xi_1 \tag{2.4.14}$$

The expressions in Eq. (2.4.14) can be phrased in terms of the probability hypothesis density from Eq. (2.3.8), and are bounded by

$$\int_{\Delta_0^\sigma} F_D(\xi)d\xi \leq \int_0^L \int_{y-\sigma}^y F_D(x,y)\,dx\,dy + \int_L^\infty \int_{y-\sigma}^y C_3 y^{-2}\,dx\,dy$$
$$\leq LC_2\sigma + 3C_3\sigma/L = (LC_2 + C_3/L)\sigma$$

where assumptions (A2) and (A3) respectively yield the bounds C_2 and $C_3 y^{-2}$ on the probability hypothesis density, F_D. ■

Lemma 2.4.1 yields control over the counting measure ν_i defined in Defn. 2.4.3 and the coefficients $\mathcal{Q}_i^*(\cdot)$ of Eq. (2.4.3) which respectively determine the distribution of lower and upper cardinalities for a persistence diagram sampled according to the kernel density $K_\sigma(Z, D_i)$.

Corollary 2.4.2. *Consider a random persistence diagram D distributed according to f satisfying assumptions (A1)-(A3). Take ν to be the lower cardinality probability mass function for the kernel density $K_\sigma(Z, D)$ shown in Eq. (2.4.11). Then, there exists $C > 0$ so that $\mathbb{E}^f \nu(j_0) \leq C\sigma$ whenever $j_0 \neq 0$.*

Proof. Since D is random with respect to f, ν is random with respect to f as well. Recall that ν is defined so that $\mathbb{E}^\nu(\boldsymbol{a}) = \left|D^\ell\right|$ for $\boldsymbol{a}$ distributed according to ν and thus $\mathbb{E}^f[\mathbb{E}^\nu(\boldsymbol{a})] \leq C\sigma$ for some $C > 0$ by Lemma 2.4.1. Subsequently, the value $\mathbb{E}^f \nu(j_0)$ is controlled by this double expectation so long as $j_0 \neq 0$. Indeed,

$$\mathbb{E}(x_i) = \sum_{j=0}^\infty j\nu_i(j) = \sum_{j=1}^\infty j\nu_i(j) \geq \sum_{j=1}^\infty \nu_i(j) \geq \nu_i(j_0)$$

for any $j_0 > 0$ and $\nu_i(j_0) = 0$ for $j_0 < 0$ since it represents a cardinality distribution. ■

In th following lemma, the result of Lemma 2.4.1 is used to control the expressions $\mathcal{Q}(\gamma)$ or $\mathcal{Q}^*(\gamma)$, of Eq. (2.4.2) and Eq. (2.4.3) respectively, in the kernel density estimate.

Lemma 2.4.3. *Consider a random persistence diagram D distributed according to f satisfying assumptions (A1)-(A3). Take $\mathcal{Q}$ of Eq. (2.4.2) and $\mathcal{Q}^*$ of Eq. (2.4.3) to be the*

upper singleton probabilities for the kernel density $K_\sigma(Z, D)$ *shown in Eq.* (2.4.11)*. Then, there exists* $C > 0$ *so that* $\mathbb{E}^f[\mathcal{Q}(\gamma)] \leq \mathbb{E}^f[\mathcal{Q}^*(\gamma)] \leq C\sigma$ *for any* $\gamma \in I(j, N)$ *with* $j < N$*.*

Proof. Since every $q^k \in (0, 1)$, we have that $\mathcal{Q}(\gamma) \leq \mathcal{Q}^*(\gamma)$; and furthermore, since $\gamma \in I(j, N)$ are not onto when $j < N$, each product $\mathcal{Q}^*$ is bounded by one of the terms of the $(1 - q_i^k)$ type. By construction, these terms depend monotonically upon a feature's persistence, and the maximum (over all indices $j < N$ and functions γ) is tied to the least persistent feature of D_i^u.

For a feature (b, d) of persistence $p = d - b$, we define $q(p) := \int_{-p/(\sqrt{2}\sigma)}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$ in concordance with Eq. (2.4.8); or in terms of the error function Φ, $q(p) = \frac{1}{2}\left(1 + \Phi\left(\frac{p}{2\sigma}\right)\right)$. Define the minimal persistence as $p_{\min}(Z) = \sup\{p || \Delta_0^p \cap Z| = \emptyset\}$ which satisfies $p_{\min}(Z) \geq p$ if and only if $|\Delta_0^p \cap Z| = \emptyset$. In turn, we may bound $\mathcal{Q}^*(\gamma) \leq (1 - q(p_{\min}(D))$ independently of γ. By Lemma 2.4.1, there is $C > 0$ such that $\mathbb{P}^f[|\Delta_0^\sigma \cap D| \neq \emptyset] \leq \mathbb{E}^f[|\Delta_0^\sigma \cap D|] \leq C\sigma$, which controls the distribution of the minimal persistence.

In particular, $q'(p) = \frac{1}{2\sigma\sqrt{\pi}} e^{-p^2/4\sigma^2}$ by the fundamental theorem of calculus. The control of Lemma 2.4.1 also allows us to utilize integration via the probability of matching superlevel sets. Specifically, since $\mathcal{Q}^*(\gamma) \leq (1 - q(p_{\min}(D))$, we have

$$\mathbb{E}^f[\mathcal{Q}^*(\gamma)] \leq \int_{\mathcal{W}_{0:a-1}} (1 - q(p_{\min}(Z)))f(Z)\delta Z = \int_{\infty}^{0} \mathbb{P}^f[p_{\min} \geq p] \frac{d}{dp}\Big[1 - q(p)\Big] dp, \quad (2.4.15)$$

where integration starts at $p = \infty$ because $(1 - q(p))$ is monotonically decreasing to zero at infinity and ends at 0 since every persistence diagram satisfies $p_{\min} \geq 0$.

The drastic change of variables shown in Eq. (2.4.15) follows from rephrasing Lebesgue integration as Riemann integration. Consider the integral $\int_{\mathcal{W}_{0:a-1}} g(p_{\min}(Z))f(Z)\delta Z$ where $g : \mathbb{R}^+ \to \mathbb{R}$ is monotonically decreasing and satisfies $\lim_{p\to\infty} g(p) = 0$ ($g(p) = (1 - q(p))$ satisfies these conditions). Here, the set integral (against the pdf) is treated as integration with respect to a probability measure, wherein the value is given by the limit of integrals of simple functions. We take simple functions $s_n = \sum_{i=0}^{n} [g(p_i) - g(p_{i-1})] \mathbb{1}_{\{p_{\min} \geq p_i\}}$ according to the partition $\infty = p_0 > p_1 > ... > p_n = 0$. Choosing that $p_1 \to \infty$ and partition size going to zero, these simple functions pointwise approach g in the limit from below (since g is monotone decreasing), and thus their integrals converge to an integral of g by monotone

convergence theorem. This limit is rephrased as a Riemann integral as follows:

$$\begin{aligned}\lim_{n\to\infty}\int_{\mathcal{W}_{0:a-1}} s_n &= \lim_{n\to\infty}\sum_{i=1}^{n}[g(p_i)-g(p_{i-1})]\mathbb{P}^f[p_{\min}(Z)\geq p_i]\\ &= \lim_{n\to\infty}\sum_{i=1}^{\infty} g'(p_i)[p_i-p_{i-1}]\mathbb{P}^f[p_{\min}(Z)\geq p_i]\\ &= \int_{\infty}^{0} g'(p)\mathbb{P}^f[p_{\min}(Z)\geq p]\,dp\end{aligned}$$

where the bounds on the Riemann integral follow the partition bounds. Taking $g(p) = 1-q(p)$ yields the change of variables shown in Eq. (2.4.15).

We now further bound the expectation in Eq. (2.4.15). Replacing terms with their definitions and using the bound control from Lemma 2.4.1 we obtain:

$$\begin{aligned}\mathbb{E}^f\left[\mathcal{Q}^*(\gamma)\right] &\leq \int_0^{\infty}\mathbb{P}^f(\Delta_0^p\cap D\neq\emptyset)\frac{1}{2\sigma\sqrt{\pi}}e^{-p^2/4\sigma^2}dp\\ &\leq \frac{C}{2\sigma\sqrt{\pi}}\int_0^{\infty} pe^{-(p/2\sigma)^2}dp = \frac{C}{2\sigma\sqrt{\pi}}\left[-2\sigma^2e^{-p^2/4\sigma^2}\right]_{p=0}^{\infty} = \frac{C}{\sqrt{\pi}}\sigma.\end{aligned}$$

■

Proof of Theorem 2.1. For convenience, we denote the upper cardinalities by $N_i = |D_i^u|$ and total cardinalities by $M_i = |D_i|$ for the sample persistence diagrams. Denote the set of strictly increasing functions from $\{1,...,j\}$ into $\{1,...,N_i\}$ by $I(j,N_i)$. Here we use 'id' to denote the identity map, where $I(N_i,N_i) = \{\text{id}\}$. The proof is organized by splitting the kernel densities into several pieces and then controlling each piece separately.

First, we separate the kernel $K_\sigma(Z, D_i)$, defined in Eq. (2.4.11), into three portions, A_i, B_i, and C_i, according to the upper cardinality j:

$$\begin{aligned} K_\sigma(Z, D_i) &= \sum_{j=0}^{N_i} \nu_i(N-j) \sum_{\gamma \in I(j,N_i)} \mathcal{Q}_i(\gamma) \prod_{k=1}^{j} p_i^{\gamma(k)}(\xi_k) \prod_{k=j+1}^{N} p_i^{\ell}(\xi_k) \\ &= \nu_i(N-N_i)\mathcal{Q}_i(\mathrm{id}) \prod_{k=1}^{N_i} p_i^{k}(\xi_k) \prod_{k=N_i+1}^{N} p_i^{\ell}(\xi_k) \\ &+ \sum_{j=0, j\neq N}^{N_i-1} \nu_i(N-j) \sum_{\gamma \in I(j,N_i)} \mathcal{Q}_i(\gamma) \prod_{k=1}^{j} p_i^{\gamma(k)}(\xi_k) \prod_{k=j+1}^{N} p_i^{\ell}(\xi_k) \\ &+ \mathbb{1}_{\{n\in\mathbb{N} | n<N_i\}}(N)\nu_i(0) \sum_{\gamma \in I(N,N_i)} \mathcal{Q}_i(\gamma) \prod_{k=1}^{N} p_i^{\gamma(k)}(\xi_k) \\ &= A_i + B_i + C_i, \end{aligned} \tag{2.4.16}$$

where A_i follows from $j = N_i$, C_i follows from $j = N$ ($C_i = 0$ if $N_i \leq N$), and B_i consists of all remaining terms.

The terms B_i in Eq. (2.4.16) are controlled by the lower product $\left[\prod_{k=j+1}^{N} p_i^{\ell}(\xi_k)\right]$. Since $(1-q_i^j) \leq 1$ and $\nu_i(N-j) \leq 1$ for any choice of γ and j, we have that B_i is bounded above by

$$\sum_{j=0, j\neq N}^{N_i-1} \sum_{\gamma \in I(j,N_i)} \left[\prod_{k=1}^{j} q_i^{\gamma(k)} p_i^{\gamma(k)}(\xi_k) \prod_{k=j+1}^{N} p_i^{\ell}(\xi_k) \right]. \tag{2.4.17}$$

The bounding sum of Eq. (2.4.17) consists of restricted $2N$-dimensional Gaussians, with the weights q_i^j dominating the restriction rescaling in Eq. (2.4.7). Fix $\pi \in \Pi_N$ and $j \in \{0, ..., M-1\} \setminus \{N\}$. Without loss of generality, we treat the case when the permutation π is the identity. Since our ultimate goal is to control the kernel density estimate $\hat{f}$, consider the portion of $\sum_{i=1}^{n} \frac{1}{n} B_i$ for which the cardinalities $M_i = |D_i|$ are fixed at level $M_i = m \in \{0, ..., M\}$. Now, $m = |D_i| \geq N_i > j$, so there is some extension for every γ within the sum, $\gamma^* \in \Pi_m$. Recall that this collection is random because each D_i is randomly distributed according to f, therefore we consider the expectation with respect to this randomness:

$$\mathbb{E}^f \left[\sum_{\{i:M_i=m\}} \frac{1}{|\{i : M_i = m\}|} \prod_{k=1}^{M_i} q_i^{\gamma^*(k)} p_i^{\gamma^*(k)}(\xi_k) \right] \to f(\xi_1, ..., \xi_m),$$

for any point $(\xi_1, ..., \xi_m)$ as a $2m$-dimensional Gaussian kernel density estimate with a proper choice of $\sigma = O(n^{-\alpha})$ appropriate for $2M$ (and hence $2m$) dimensions [75]. Integrating both sides against the extra coordinates, Assumptions (A2) and (A3) along with the dominated convergence theorem yield

$$\mathbb{E}^f \left[\sum_{\{i:M_i=m\}} \frac{1}{|\{i : M_i = m\}|} \prod_{k=1}^{j} q_i^{\gamma(k)} p_i^{\gamma(k)}(\xi_k) \right] \to \int_{W^{m-j}} f(\xi_1, ..., \xi_m) d\xi_{j+1} ... d\xi_m, \quad (2.4.18)$$

which is again bounded via (A2) and (A3). Of course, $|\{i : M_i = m\}| \leq n$, so taking Eq. (2.4.18) into account for every m bounds the averaging sum of the upper product: $\frac{1}{n} \sum_{i=1}^{n} \prod_{k=1}^{j} q_i^{\gamma(k)} p_i^{\gamma(k)}(\xi_k)$.

Relying on Eq. (2.4.17), we must also consider the lower product $\prod_{k=j+1}^{N} p_i^{\ell}(\xi_k)$. Since the points ξ_i are fixed, we focus on their minimal persistence $p_{\min} = \min_i(d_i - b_i)$. Thus,

$$p_i^{\ell}(\xi_i) \leq \frac{1}{2\pi\sigma^2} e^{-(b-d)^2/4\sigma^2} \leq \frac{1}{2\pi\sigma^2} e^{-p_{\min}^2/4\sigma^2},$$

and subsequently,

$$\left[\prod_{k=j+1}^{N} p_i^{\ell}(\xi_k) \right] \leq \frac{1}{(2\pi\sigma^2)^N} e^{-N p_{\min}^2/4\sigma^2} \to 0, \quad (2.4.19)$$

as $\sigma \to 0$, uniformly on any compact subset of W (or $\mathcal{W}_{0:a-1}$). Altogether, Eqs. (2.4.18) and (2.4.19) gaurantee that the term $\sum_{i=1}^{n} \frac{1}{n} B_i \to 0$ as $n \to \infty$ in the kernel density estimation.

Next we focus on the terms A_i in Eq. (2.4.16). We split the sum $\frac{1}{n} \sum_{i=1}^{n} A_i$ according to the cardinality of D_i. Specifically, separate A_i into the cases where $M_i \neq N_i$ or $M_i = N_i$. First consider the associated set of indices $\{i : M_i \neq N_i\}$ and define the mismatch number $\mathrm{MM}(n)$ to be its cardinality. Critical to our argument, the mismatch number is random with respect to f because it is defined according to the features in D_i. We obtain the following mismatched term:

$$\frac{1}{n} \sum_{\{i:N_i \neq M_i\}} A_i \leq \left(\frac{MM(n)}{n} \right) \frac{1}{MM(n)} \sum_{\{i:N_i \neq M_i\}} \left[\mathcal{Q}_i(\mathrm{id}) \prod_{k=1}^{N_i} p_i^k(\xi_k) \prod_{k=N_i+1}^{N} p_i^{\ell}(\xi_k) \right] \quad (2.4.20)$$

The bounding sum in Eq. (2.4.20) is split into pieces where $M_i = m$ for each m between 0 and M. Using the same strategy yielding Eq. (2.4.18), with $MM(n)$ in place of n, the sum of the upper product converges to layered integrals of f for each level m and each $N_i < m$ by extending $\gamma = \mathrm{id}$. Using the same approach leading to Eq. (2.4.19), the lower product vanishes in the limit if $N_i \neq N$, or is an empty product if $N_i = N$; in either case, this factor is bounded. Now, according to Lemma 2.4.1, $\mathbb{P}^f(M_i \neq N_i) = \mathbb{P}^f(D_i \cap \Delta_0^{\epsilon\sigma} \neq \emptyset) \leq C_5\sigma$; consequently, $\mathbb{E}^f[MM(n)/n] \to 0$ and the mismatch terms on left hand side of Eq. (2.4.20) follow.

Now consider the indices for which $N_i = M_i$. In this case, since D_i^ℓ are empty, $\nu_i = \delta_0$, and the only values which contribute to the sum are for $N_i = N$. The remaining portion of the kernel density estimate is given by

$$\frac{1}{n}\mathbb{E}^f \sum_{\{i:N_i=M_i\}} A_i = \frac{1}{n}\mathbb{E}^f \left[\sum_{\{i:N_i=M_i\}} \left(\mathcal{Q}_i(\mathrm{id}) \prod_{k=1}^{N} p_i^k(\xi_k) \right) \right] = \frac{1}{n}\mathbb{E}^f \left[\sum_{\{i:N_i=M_i\}} \left(\prod_{k=1}^{N} q_i^k p_i^k(\xi_k) \right) \right]. \tag{2.4.21}$$

As shown, the terms in Eq. (2.4.21) are restricted $2N$ dimensional Gaussians. It is known [75] that restricted Gaussian kernel density estimates like $\left[\prod_{k=1}^N q_i^k p_i^k(\xi_k)\right]$ converge (uniformly on compactly contained sets) to the true value of the chosen draws D_i for suitable choice of α in $\sigma = O(n^{-\alpha})$ as restricted by $N \leq M$. After correcting for the samples with $N_i < M_i = N$, the samples D_i are treated as random draws from $f(D\,||D| = N)$. Consequently, we may conclude that the target distribution associated with $\left[\prod_{k=1}^N q_i^k p_i^k(\xi_k)\right]$ is the rescaled $\frac{1}{f(N)} f(\xi_1, ..., \xi_N)$, where $f(N) := \mathbb{P}^f(|D| = N)$. This rescaling for the conditional pdf $f(D\,||D| = N)$ is necessary to reweight according to Prop. 2.3.7.

Application of classical kernel density estimate results require division by the cardinality of the draw, when in context n is generally larger than this cardinality. Thus, we must again consider the cases wherein $N_i \neq M_i$. Consequently, we find that the expectation for the ratio between the true draw cardinality and n is given by $\mathbb{P}^f(|D| = N) + O(\sigma)$ according to Lemma 2.4.1. Indeed, this ratio converges to $f(N) := \mathbb{P}^f(|D| = N)$. After this final correction, we have shown that $\frac{1}{n}\sum_{i=1}^n A_i$ approach the true pdf $f(\xi_1, ..., \xi_N)$.

Lastly, we need only to control the terms C_i from Eq. (2.4.16). We begin by bounding the probability mass functions ν_i by 1 and considering only terms for which the characteristic function is nonzero:

$$\frac{1}{n}\sum_{i=1}^{n} C_i = \frac{1}{n}\sum_{\{i:N<N\}} \nu_i(0) \sum_{\gamma\in I(N,N_i)} \mathcal{Q}_i(\gamma)\prod_{k=1}^{N} p_i^{\gamma(k)}(\xi_k) \leq \frac{1}{n}\sum_{\{i:N<N_i\}} \sum_{\gamma\in I(N,N_i)} \mathcal{Q}_i(\gamma)\prod_{k=1}^{N} p_i^{\gamma(k)}(\xi_k). \tag{2.4.22}$$

Next, we split the term $\mathcal{Q}(\gamma)$ according to Eq. (2.4.2) and apply Lemma 2.4.3 to the upper bound in Eq. (2.4.22) to obtain the larger upper bound

$$\frac{1}{n}\sum_{\{i:N<N_i\}} \sum_{\gamma\in I(N,N_i)} \mathcal{Q}^*(\gamma)\prod_{k=1}^{N} q_i^{\gamma(k)} p_i^{\gamma(k)}(\xi_k) \leq C\left[\frac{1}{n}\sum_{\{i:N<N_i\}} \sum_{\gamma\in I(N,N_i)} \prod_{k=1}^{N} q_i^{\gamma(k)} p_i^{\gamma(k)}(\xi_k)\right]\sigma. \tag{2.4.23}$$

The expectation of the bracketed terms in Eq. (2.4.23) converges in a fashion identical to the terms $\frac{1}{n}\sum_{i=1}^{n} A_i$. Since these terms are multiplied by σ, altogether $\left[\frac{1}{n}\sum_{i=1}^{n} C_i\right]$ vanishes in the limit as $n \to \infty$. Putting together the limits of each portion built from $K_\sigma(Z, D_i) = A_i + B_i + C_i$, the theorem follows. ■

2.4.3 A Measure of Dispersion

Theorem 2.1 has established the convergence of a kernel density estimator. Despite lacking a vector space structure on the space of persistence diagrams, it is equipped with the bottleneck metric (Defn. 2.2.9). Thus, we aim to measure dispersion with respect to a distribution of persistence diagrams through mean absolute deviation in this metric.

Definition 2.4.4. *The mean absolute bottleneck deviation (MAD) from origin diagram $\mathcal{D}$ with respect to a global pdf f is given by*

$$MAD_f(\mathcal{D}) = \int_{\mathcal{W}_{0:a-1}} W_\infty(\mathcal{D}, Z) f(Z) \delta Z \tag{2.4.24}$$

The following proposition aids in proving convergence of MAD kernel estimates.

Proposition 2.4.5. *Consider D distributed according to the kernel density $K_\sigma(\cdot, \mathcal{D})$ with center diagram $\mathcal{D}$ and bandwidth σ. Fix $\delta \geq 1$. Then,*

$$\mathbb{P}\left[W_\infty(D, \mathcal{D}) < \delta\sigma\right] \geq \left(\int_{B(\mathbf{0},\delta)} \frac{1}{2\pi} e^{-(x^2+y^2)/2}\, dx\, dy\right)^M \tag{2.4.25}$$

where M is the maximal cardinality of D (a multiple of $|\mathcal{D}|$). Here $B(x, r)$ refers to a ball with respect to the infinity metric (as is used in bottleneck distance).

Proof. Note that the lower bound integral is the probability for a pair $z = (x, y)$ of independent standard normal variables to lie in $B((0, 0), \delta)$. In order to bound the bottleneck distance $W_\infty(D, \mathcal{D}) < \delta\sigma$, it is sufficient that each constituent feature does not stray too far from either its corresponding center or the diagonal (see Fig. 2.6 for reference). Specifically, we follow Defn. 2.2.9 to build a correspondence between D and $\mathcal{D}$ so that the maximal distance undercuts $\delta\sigma$, and thus the (potentially smaller) bottleneck distance is also bounded by $\delta\sigma$. For clarity, features in D are denoted using ζ while features in $\mathcal{D}$ are denoted using ξ.

Consider each feature $\xi^j \in \mathcal{D}^u = \mathcal{D} \cap \{d - b \geq \sigma\}$ and its associated random singleton diagram $D^j = \{\zeta^j\}$ or $\emptyset$ as in Defn. 2.4.2. Assuming the disc neighborhood $B_j = B(\xi^j, \delta\sigma)$ is contained in the wedge $W = \{(b, d) \in \mathbb{R}^2 | d > b \geq 0\}$, $z^j = \frac{\zeta^j - \xi^j}{\sigma}$ has identical density to the Gaussian random variable $z \sim N((0, 0), I_2)$ whenever $\zeta^j \in B_j$ (or equivalently $z^j \in B((0, 0), \delta)$). Thus, we obtain $\mathbb{P}\left[\zeta^j \in B(\zeta^j, \delta\sigma)\right] = \mathbb{P}\left[|z| \leq \delta\right]$ for the probability that ζ^j can be mapped to ξ^j in a bounding correspondence. If $B_j \not\subseteq W$, this probability is higher than required because ξ^j can be mapped to the diagonal and thus the case $D^j = \emptyset$ is included.

Now take into account the features in $\mathcal{D}^\ell$ and the associated random features D^ℓ as in Defn. 2.4.3. Although the features in D^ℓ are not necessarily independent, we may assume without loss of generality the worst case, in which the maximal cardinality is drawn. Given a fixed cardinality, the draws of D^ℓ are independent. Since any feature may be mapped to the diagonal in the bottleneck distance, a bounding correspondence can be obtained whenever the draws in D^ℓ and features in $\mathcal{D}^\ell$ are close enough to the diagonal (within $\delta\sigma$). Indeed, the features in $\mathcal{D}^\ell$ are by definition distance $\sigma \leq \delta\sigma$ from the diagonal. Restricting to W, the pdf for the draws of $D^\ell = \{(b_j, d_j)\}_{j=1}^{|N_\ell|}$ is given by $p^\ell(b, d) =$

$\frac{1}{\pi N_\ell \sigma^2} \sum_{j=1}^{N_\ell} e^{-\left(\left(x-\frac{b_j+d_j}{2}\right)^2+\left(y-\frac{b_j+d_j}{2}\right)^2\right)/2\sigma^2}$. Consider the sets $U_j = B\left(\left(\frac{b_j+d_j}{2}, \frac{b_j+d_j}{2}\right), \delta\sigma\right)$ and $\mathcal{U} = \bigcup_{j=1}^{N_\ell} U_j$. For each lower feature $(b,d) \in D^\ell$, mapping to the diagonal yields a bounding correspondence and the associated probability is bounded below by $\mathbb{P}[d-b \leq \delta\sigma] = \int_{\Delta_0^{\delta\sigma}} p^\ell(x,y)\,dx\,dy \geq \int_{W\cap\mathcal{U}} p^\ell(x,y)\,dx\,dy$ since $W \cap U \subset \Delta_0^{\delta\sigma} = \{(b,d) \in W | d-b \leq \delta\sigma\}$. Next, we restrict the lower bounding integral for each term of p^ℓ to its matching subset U_j and change variables to attain the desired form:

$$\begin{aligned}\int_{W\cap\mathcal{U}} p^\ell(x,y)\,dx\,dy &\geq \sum_{j=1}^{N_\ell} \int_{U_j} \frac{1}{2\pi N_\ell \sigma^2} e^{-\left(\left(x-\frac{b_j+d_j}{2}\right)^2+\left(y-\frac{b_j+d_j}{2}\right)^2\right)/2\sigma^2}\,dx\,dy \\ &= \int_{B((0,0),\delta)} \frac{1}{2\pi} e^{-(x^2+y^2)/2}\,dx\,dy.\end{aligned}$$

Overall, this argument shows that with probability at least $\mathbb{P}(|\boldsymbol{z}| \leq \delta)^M$ there is a correspondence which bounds the bottleneck distance by $\delta\sigma$ and the result follows. ■

Next, we relax assumption $(A2)$ by considering the entire multi-wedge $\mathcal{W}_{0:a-1}$ and tighten the decay control from assumption $(A3)$. Formally,

$(A2)^*$ The local density $f_N : \mathcal{W}_{0:a-1}^N \to \mathbb{R}$ is bounded for each $N \in \{1,...,M\}$.

$(A3)^*$ There exists $C > 0$ so that $f(\xi_1,...,\xi_N) \leq C \left\|(\xi_1,...,\xi_N)\right\|^{-2N-2}$ for $N \in \{1,...,M\}$.

These assumptions (and $(A1)$) are required for the subsequent lemma, which ensures that the mean absolute bottleneck deviation (MAD) is finite.

Lemma 2.4.6. *Consider a random persistence diagram D distributed according to a global pdf f satisfying assumptions $(A1)$, $(A2)^*$, and $(A3)^*$. Then D has finite MAD for any choice of origin diagram $\mathcal{D}$.*

Proof. Choose an arbitrary persistence diagram $\mathcal{D}$. Since bottleneck distance is defined according to the sup-norm (see Eq. (2.2.7)), the bottleneck distance to the null persistence diagram (i.e., without any features) is precisely half the maximal persistence. Thus, we begin by showing that the maximal persistence moment is finite. Taking $Z = \{\xi_1,...,\xi_N\}$

with $\xi_i = (b_i, d_i, k_i)$, we have:

$$\int_{\mathcal{W}_{0:a-1}} \max(d_i - b_i)\delta Z \leq \int_{\mathcal{W}_{0:a-1}} \|Z\| \, f(Z)\delta Z \tag{2.4.26}$$

since $\max(d_i - b_i) \leq \max(\|(b_i, d_i)\|) \leq \|Z\|$. Consider a compact set $K \subset \mathcal{W}_{0:a-1}$ which contains a neighborhood of the origin. Given assumptions $(A2)^*$ and $(A3)^*$, Eq. (2.4.26) is bounded by the following finite expression.

$$\int_{\mathcal{W}_{0:a-1}} \|Z\| \, f(Z)\delta Z \leq \int_K C_2 \|Z\| \, \delta Z + \sum_{N=1}^{M} \int_{h_N^{-1}(h_N(K)^c)} C_3 \|Z\|^{-2N-1} \, d\xi_1 ... d\xi_N. \tag{2.4.27}$$

Lastly, we take advantage of the Minkowski inequality, which holds trivially for set integration since it is a linear combination of Lebesgue integrals. Indeed,the MAD centered at $\mathcal{D}$ is bounded as follows.

$$\int_{\mathcal{W}_{0:a-1}} W_\infty(\mathcal{D}, Z) f(Z)\delta Z \leq \int_{\mathcal{W}_{0:a-1}} W_\infty(\mathcal{D}, \emptyset) f(Z)\delta Z + \int_{\mathcal{W}_{0:a-1}} W_\infty(\emptyset, Z)\delta Z \tag{2.4.28}$$

where $\emptyset$ represents the null persistence diagram and the distance to the null persistence diagram is precisely half the maximal persistence. Since f integrates to 1, the first integral simplifies to the finite distance $W_\infty(\mathcal{D}, \emptyset)$, while the second integral is finite according to Eq. (2.4.27). ■

Remark 2.4.7. *The results of [5], which yield many long persistence points for heavy tailed distributions, show that a condition of the type* $(A3)^*$ *is necessary. If the underlying datapoints are not heavy-tailed, this condition is easily satisfied. One may replace Lemma 2.4.6 and its assumptions by directly assuming that the maximal persistence moment is bounded; with this, the results of the lemma follow immediately from Eq.* (2.4.28)*. This direct assumption is weaker (implied by* $(A1)$*,* $(A2)^*$*, and* $(A3)^*$*), but may be difficult to show in practice.*

Theorem 2.2. *Consider a distribution of persistence diagrams with bounded global pdf,* f*, satisfying assumptions* $(A1)$*,* $(A2)^*$*, and* $(A3)^*$*. Let* $\hat{f}(Z) = \frac{1}{n}\sum_{i=1}^{n} K_\sigma(Z, D_i)$ *be a kernel density estimate with centers* D_i *sampled i.i.d. according to* f *and bandwidth* $\sigma = O(n^\alpha)$

chosen with $0 < \alpha < \alpha_{2M}$*. Then, the mean absolute bottleneck deviation estimate converges; in other words,*

$$\int_{\mathcal{W}_{0:a-1}} W_\infty(\mathcal{D}, Z)\hat{f}(Z)\delta Z \to \int_{\mathcal{W}_{0:a-1}} W_\infty(\mathcal{D}, Z)f(Z)\delta Z \tag{2.4.29}$$

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

Proof. The MAD of f with origin $\mathcal{D}$ is finite by Lemma 2.4.6. To show convergence of the estimate, we begin by adding and subtracting the integral of the sample estimator for the MAD. Then, we split the sum into $n+1$ terms via the triangle inequality to obtain

$$\begin{aligned}
&\left|\int_{\mathcal{W}_{0:a-1}} W_\infty(\mathcal{D}, Z)f(Z)\delta Z - \int_{\mathcal{W}_{0:a-1}} W_\infty(\mathcal{D}, Z)\hat{f}(Z)\delta Z\right| \\
&\quad\le \left|\int_{\mathcal{W}_{0:a-1}} W_\infty(\mathcal{D}, Z)f(Z)\delta Z - \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{W}_{0:a-1}} W_\infty(\mathcal{D}, D_i)K_\sigma(Z, D_i)\delta Z\right| \\
&\quad+ \frac{1}{n}\sum_{i=1}^{n}\left|\int_{\mathcal{W}_{0:a-1}} W_\infty(\mathcal{D}, Z)K_\sigma(Z, D_i)\delta Z - \int_{\mathcal{W}_{0:a-1}} W_\infty(\mathcal{D}, D_i)K_\sigma(Z, D_i)\delta Z\right|.
\end{aligned} \tag{2.4.30}$$

The term of the upper bound in Eq. (2.4.30) trivially simplfies to obtain the sample estimator for the MAD:

$$\begin{aligned}
&\left|\int_{\mathcal{W}_{0:a-1}} W_\infty(\mathcal{D}, Z)f(Z)\delta Z - \sum_{i=1}^{n}\frac{1}{n}\int_{\mathcal{W}_{0:a-1}} W_\infty(\mathcal{D}, D_i)K_\sigma(Z, D_i)\delta Z\right| \\
&\qquad= \left|\int_{\mathcal{W}_{0:a-1}} W_\infty(\mathcal{D}, Z)f(Z)\delta Z - \frac{1}{n}\sum_{i=1}^{n} W_\infty(\mathcal{D}, D_i)\right|.
\end{aligned} \tag{2.4.31}$$

The MAD sample estimator converges since the MAD is finite, and thus this term vanishes as $n \to \infty$. The remaining term of the upper bound in Eq. (2.4.30) is further bounded via the reverse triangle inequality; specifically,

$$\begin{aligned}
&\sum_{i=1}^{n}\frac{1}{n}\left|\int_{\mathcal{W}_{0:a-1}} W_\infty(\mathcal{D}, Z)K_\sigma(Z, D_i)\delta Z - \int_{\mathcal{W}_{0:a-1}} W_\infty(\mathcal{D}, D_i)K_\sigma(Z, D_i)\delta Z\right| \\
&\qquad\le \sum_{i=1}^{n}\frac{1}{n}\left|\int_{\mathcal{W}_{0:a-1}} W_\infty(D_i, Z)K_\sigma(Z, D_i)\delta Z\right|.
\end{aligned} \tag{2.4.32}$$

Toward bounding Eq. (2.4.32), choose a threshold parameter $a = O(\sigma^\beta)$ for some $\beta \in (0,1)$, so that $a \to 0$ but $a/\sigma \to \infty$ in the sample size (and bandwidth) limit. Next, take $A_i = \{Z \subset W : W_\infty(Z, D_i) \leq a\}$ and split the integral between A_i and its complement as

$$\int_{\mathcal{W}_{0:a-1}} W_\infty(D_i, Z) K_\sigma(Z, D_i) \delta Z = \int_{A_i} W_\infty(D_i, Z) K_\sigma(Z, D_i) \delta Z + \int_{A_i^c} W_\infty(D_i, Z) K_\sigma(Z, D_i) \delta Z.$$

The integral over A_i is trivially bounded by a. Integration over the complementary events is controlled via layered integration along with Prop. 2.4.5. For $a/\sigma > 1$, which occurs when n is large enough, we obtain

$$\begin{aligned} \int_{A_i^c} W_\infty(D_i, Z) K_\sigma(Z, D_i) \delta Z &= a\, \mathbb{P}^i \left[W_\infty(D_i, Z) > a \right] + \int_a^\infty \mathbb{P}^i \left[W_\infty(D_i, Z) > b \right] db \\ &\leq a \left(1 - \mathbb{P}[|z| \leq a/\sigma]^M \right) + \int_a^\infty \left(1 - \mathbb{P}[|z| \leq b/\sigma]^M \right) db, \end{aligned} \tag{2.4.33}$$

where $z = (x, y)$ is distributed as a pair of independent standard normals. We chose $a/\sigma = O(\sigma^{\beta-1}) \to \infty$ and so $\mathbb{P}(|z| \leq a/\sigma) \to 1$ exponentially fast and the last term vanishes quickly as $\sigma \to 0$.

The layered integral used in Eq. (2.4.33) is similar to the argument used in Lemma 2.4.3. For brevity, denote $g(Z) = W_\infty(D_i, Z)$ so that $A_i = \{Z : g(Z) \leq a\}$ and $A_i^c = \{Z : g(Z) > a\}$. The function $g(Z)$ is approximated from below by the simple functions $g_n(Z) = \sum_{i=1}^n (b_i - b_{i-1}) \mathbb{1}_{\{Z | b_i \leq g(Z)\}}$ for and appropriate choice of partitions $\{0 = b_0^n < b_1^n < \ldots < b_n^n\}_{n=1}^\infty$ with $b_n^n \to \infty$. By monotone convergence theorem the integrals of $g_n(Z)$ converge to the integral of $g(Z)$. In particular, rephrase $\int_{A_i^c} g(Z) K_\sigma(Z, D_i)$ to obtain

$$\lim_{n \to \infty} \sum_{i=1}^n \mathbb{P}^i \left[\{Z | b_i \leq g(Z)\} \cap A_i^c \right] (b_i - b_{i-1}) \to \int_0^\infty \mathbb{P}^i [\{Z | b \leq g(Z)\} \cap A_i^c] \, db.$$

Moreover,

$$\mathbb{P}^i[\{Z | b \leq g(Z)\} \cap A_i^c] = \mathbb{P}^i[\{Z | b \leq g(Z)\} \, | A_i^c] \cdot \mathbb{P}^i[A_i^c] = \begin{cases} \mathbb{P}^i[A_i^c] & b \leq a \\ \mathbb{P}^i[\{Z | b \leq g(Z)\}] & b > a \end{cases}.$$

The first line of Eq. (2.4.33) follows from the definitions of $g(Z)$ and A_i^c, while the second line follows from application of Prop. 2.4.5. Thus, both bounding terms in Eq. (2.4.30) converge to zero and thus the kernel estimate converges to the true mean absolute deviation. ■

2.5 Examples

Here we provide detailed examples of the kernel density and kernel density estimation of an unknown pdf. For simplicity, we restrict to a single homological dimension, say $k = 1$. Due to the intrinsic high dimension of the kernel, we present contour plots for slices of the kernel density. Specifically, for inputs $((b_1, d_1), ..., (b_N, d_N))$, we consider the kernel density evaluated at $(b_1, d_1) \in W$ with (b_i, d_i) fixed for $i \geq 2$. For clarity, the unique symmetric pdf $f_{sym}(\xi_1, ..., \xi_N) = \frac{1}{N!} \sum_{\pi \in \Pi_N} f(\xi_{\pi(1)}, ..., \xi_{\pi(N)})$ is used in the contour plots (see Remark 2.3.4). For explicit computation, we choose the probability mass function

$$\nu(j) = \max \left\{ \frac{N_\ell + 1 - |N_\ell - j|}{(N_\ell + 1)^2}, 0 \right\} \tag{2.5.1}$$

when evaluating the lower density in Eq. (2.4.10), where $N_\ell = \left|\mathcal{D}^\ell\right|$ is the lower cardinality of the center diagram. This probability mass function describes the sum of two independent uniform random integers in $\{0, ..., N_\ell\}$.

Example 2.5.1. Consider the center persistence diagram $\mathcal{D} = \{(1, 3), (2, 4), (1, 1.3), (3, 3.2)\} \subset W$ and bandwidth $\sigma = 1/2$. We construct the associated kernel density via Eq. (2.4.11) and follow with some plots and analysis of the kernel density.

First, we decribe the random diagram associated to the lower center diagram $\mathcal{D}^\ell = \{(1, 1.3), (3, 3.2)\}$. Recall that the lower random diagram D^ℓ is described in Defn. 2.4.3 according to a probability mass function (pmf) ν for the cardinality of D^ℓ and a single probability density $p^\ell(b, d)$ for the subsequent features' locations in the wedge W. The pmf ν is defined according to Eq. (2.5.1) with $N_\ell = 2$; that is, $\nu(\{0, 1, 2, 3, 4\}) = \{1/9, 2/9, 3/9, 2/9, 1/9\}$ respectively, and zero otherwise. Following Defn. 2.4.3, we project $\mathcal{D}^\ell$ onto the diagonal to obtain $\{(1.15, 1.15), (3.1, 3.1)\}$. Relying on Eq. (2.4.9), the resulting

lower density is given by

$$p^{\ell}(b,d) = \frac{2}{\pi}\left[e^{-\left((b-1.15)^2+(d-1.15)^2\right)} + e^{-\left((b-3.1)^2+(d-3.1)^2\right)}\right]$$

restricted above the diagonal . The coefficient $\frac{2}{\pi}$ is propogated by a direct substitution into Eq. (2.4.9), yielding $\frac{1}{N_\ell \pi \sigma^2} = \frac{1}{2\pi(0.5^2)} = \frac{2}{\pi}$.

The global pdf for the kernel is defined on $\bigcup_{N=0}^{6} W^N$ which lacks a fixed dimension, so we consider a local portion of the kernel when the input is restricted to 3 features. Specifically, we take $Z = (\xi_1, \xi_2, \xi_3)$ with each $\xi_i = (b_i, d_i)$. The local kernel is given by

$$\begin{aligned} K_\sigma(Z,\mathcal{D}) = \; & 9.01 \times 10^{-2} p^{\ell}(b_3,d_3) e^{-2\left((b_1-1)^2+(d_1-3)^2\right)} e^{-2\left((b_2-2)^2+(d_2-4)^2\right)} \\ & + 4.96 \times 10^{-4} p^{\ell}(b_2,d_2) p^{\ell}(b_3,d_3) e^{-2\left((b_1-2)^2+(d_1-4)^2\right)} \\ & + 4.96 \times 10^{-4} p^{\ell}(b_2,d_2) p^{\ell}(b_3,d_3) e^{-2\left((b_1-1)^2+(d_1-3)^2\right)} \\ & + 1.22 \times 10^{-6} p^{\ell}(b_1,d_1) p^{\ell}(b_2,d_2) p^{\ell}(b_3,d_3). \end{aligned} \tag{2.5.2}$$

The exponentials shown are $\frac{\pi}{2} q^1 p^1$ and $\frac{\pi}{2} q^2 p^2$, where the removed coefficient, $\frac{2}{\pi} = \frac{1}{2\pi(0.5)^2}$, is tied to a 2-dimensional Gaussian of variance $\sigma = 0.5$. Similarly, the coefficient on the exponents, $-2 = -1/(2(0.5)^2)$, is tied to a 2-dimensional Gaussian of variance $\sigma = 0.5$. The coefficients for each line of Eq. (2.5.2) are built from the lower and upper cardinality probabilities given by ν and $\mathcal{Q}^*$ along with the coefficient $\frac{2}{\pi}$ of each upper Gaussian. For example, $0.0901 = \nu(1)\left(\frac{2}{\pi}\right)^2 = \frac{8}{9\pi^2}$ and $4.96 \times 10^{-4} = \nu(2)(1-q^2)\frac{2}{\pi} = \frac{2}{3\pi}(0.00234)$.

Remark 2.5.1. *The terms $(1-q^k)$ within the $\mathcal{Q}^*$ expression (see Eq. (2.4.3)) are very small and appear in terms for which the corresponding upper feature is unassigned. These terms are so small because the upper features have very long persistence in this example (four times the bandwidth), and so the terms in Eq. (2.5.2) which do not include both upper Guassians p^1 and p^2 have much smaller contribution to the overall pdf. Consequently, the kernel places much higher probability density near input diagrams with features nearby each upper feature in the center diagram. This behavior is seen in Fig. 2.8, 2.9, 2.10, and their respective analyses, and is directly correlated to the ratio of persistence to bandwidth for each feature.*

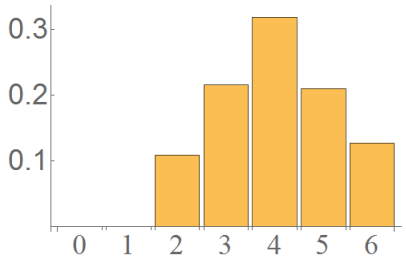

Figure 2.7: Cardinality probabilities $f(j) = \mathbb{P}[|D| = j]$ for D distributed according to global pdf $K_\sigma(\cdot, \mathcal{D})$. For the given choice of ν, $|D|$ takes on values between 0 and $6 = |\mathcal{D}^u| + 2\left|\mathcal{D}^\ell\right|$.

Having constructed the expression for a portion of the kernel density, we now turn to plotting and analyzing it. As described in Remark 2.3.3, the integral of the local density for input cardinality $|Z| = 3$ given in Eq. (2.5.2) yields the probability that a random draw has cardinality three: $(\mathbb{P}(|D| = 3) \approx 0.22)$. As expected, the random persistence diagram is most likely to have cardinality 4 $(\mathbb{P}(|D| = 4) \approx 0.32)$, matching the cardinality of the center diagram. The plot of the probability mass function for the cardinality of D (where D is distributed according to the kernel) primarily inherits its pattern from ν, shifted up by $2 = |\mathcal{D}^u|$, and is shown in Fig. 2.7.

This kernel density is quite complex to visualize, since the input space has variable dimension, up to 12 (for the maximum of $6 = |\mathcal{D}^u| + 2\left|\mathcal{D}^\ell\right|$ input features). First, we consider the probability hypothesis density (or PHD, as defined in Eq. (2.3.8)) along with the kernel density evaluated at a single input feature in Fig. 2.8. Recall that the PHD indicates the expected number of features in a region. Due to the construction of our kernel density, the PHD for the kernel is a more typical kernel density estimate in W where the centers consist of the individual features of $\mathcal{D}$ as opposed to the entire diagram. Consequently, one may instead consider Fig. 2.8 (along with Fig. 2.9 and Fig. 2.10) as a comparison between the information obtained from a typical (in $\mathbb{R}^2$) kernel density estimate and a persistence diagram kernel density. Even though the PHD indicates a very high chance for features near the diagonal, the first feature is far more likely to be near one of the high persistent features in the center diagram.

As described in Remark 2.5.1, the kernel has a tendency to prioritize upper features; this preference is seen in Fig. 2.9 for two input features and in Fig. 2.10 for three input features.

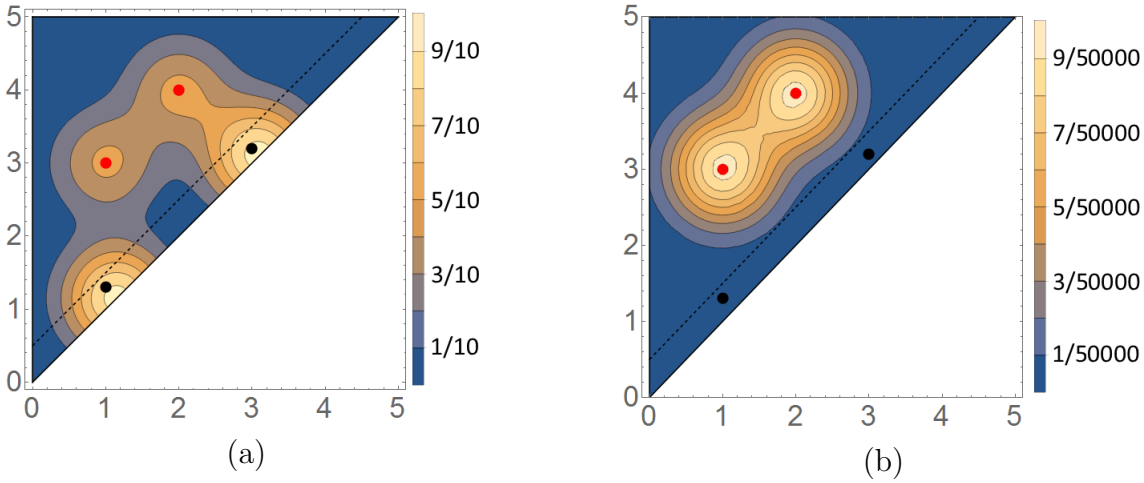

Figure 2.8: Contour maps for (a) the probability hypothesis density associated to the kernel density and (b) the density $K_\sigma(\{(b,d)\},\mathcal{D})$ restricted to a single input feature $Z = \{(b,d)\}$. The center diagram is indicated by red (upper) and black (lower) points. Scale bars at the right of each plot indicate the range of probability density in each shaded region.

If we consider the density evaluated at $f_{\mathcal{D}}((b,d),(1,3))$ or $f_{\mathcal{D}}((b,d),(2,4))$ (Fig. 2.9 (a) or (b), respectively), the remaining slice is very nearly a Gaussian centered at the other upper feature. If the known feature is instead close to the diagonal, as in Fig. 2.9 (c), the density slice is close to a Gaussian mixture between the two upper Gaussians p^1 and p^2.

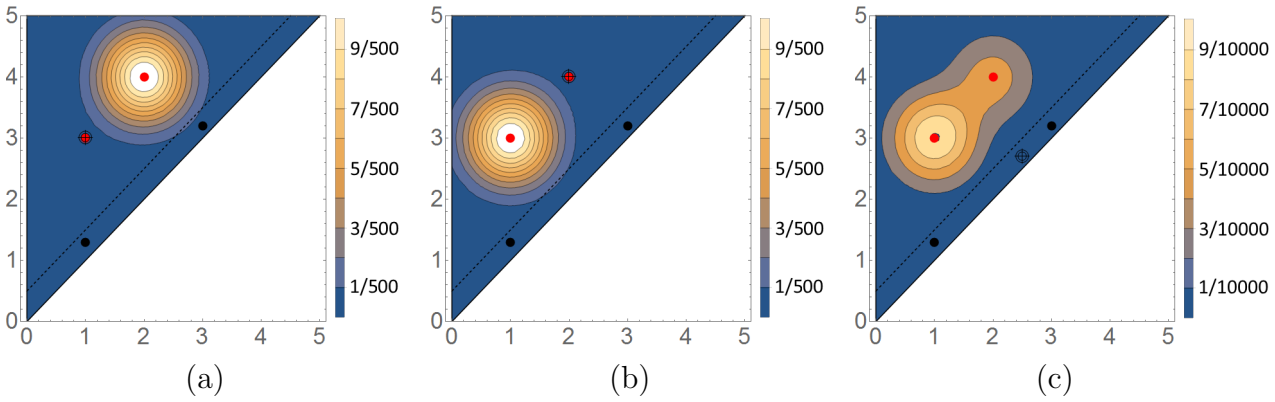

Figure 2.9: Contour maps for slices of the kernel density with certain features already chosen, indicated by crosshairs. Since the symmetric version of the density is used, the order of the features is irrelevant. The center diagram is indicated by red (upper) and black (lower) points. Scale bars at the right of each plot indicate the range of probability density in each shaded region.

It is not until the input diagram has at least three features; i.e., $Z = \{\xi_1, \xi_2, \xi_3\}$, that any slice of the density is concentrated near the diagonal. In this case, we fix two input features and leave the third free, such as in the plot of $f(\xi) := K_\sigma(\{\xi\} \cup \mathcal{D}^u, \mathcal{D})$ shown in Fig. 2.10 (a); since the entire upper center diagram $\mathcal{D}^u$ is given, this slice of the kernel is proportional to the lower distribution p^ℓ. In fact, the modes in Fig. 2.10 (a) correspond to the modes of the kernel (restricted to W^3). Center features of both long and short persistence are of similar importance when an input feature lies between the upper and lower portions of the center diagram, such as in Fig. 2.10 (b).

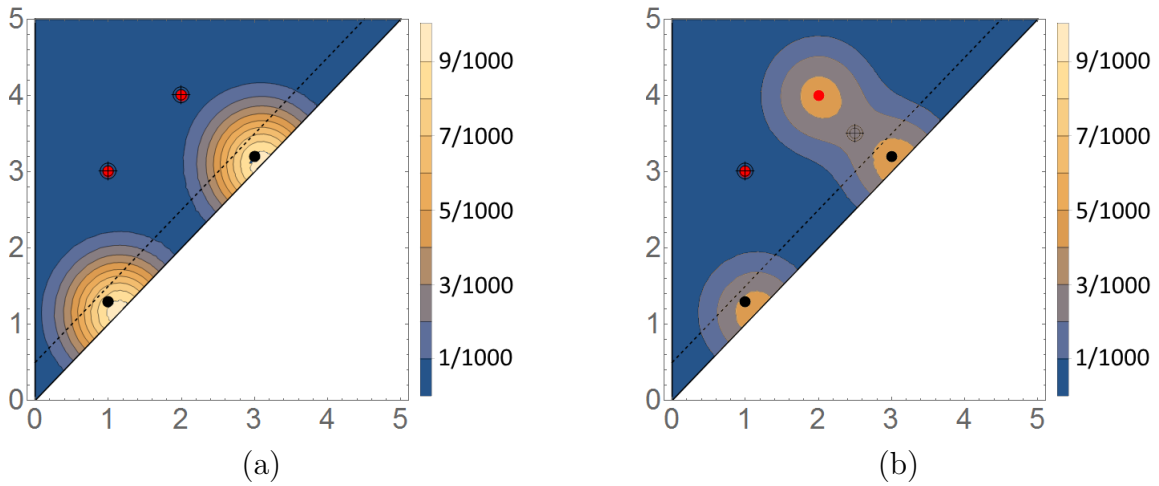

Figure 2.10: Contour maps for slices of the kernel density with certain features already chosen, indicated by crosshairs. Since the symmetric version of the density is used, the order of the features is irrelevant. The center diagram is indicated by red (upper) and black (lower) points. Scale bars at the right of each plot indicate the range of probability density in each shaded region.

Example 2.5.2. Here we consider the random persistence diagram generated from a specific random dataset in $\mathbb{R}^2$. Our goal in this example is to build and demonstrate a kernel density estimate for the random persistence diagram determined by this random dataset. Specifically, we generate 100 sample datasets which each consist of 15 points sampled uniformly from the unit circle with additive Gaussian noise, $N((0,0), \frac{1}{9}I_2)$. This toy dataset is prototypical for signal analysis (corresponding to the circular dynamics of a noisy sine curve), wherein the high dimensional point cloud is obtained through delay-embedding of the signal. An in-depth analysis of using delay embedding alongside persistent homology is found in [65].

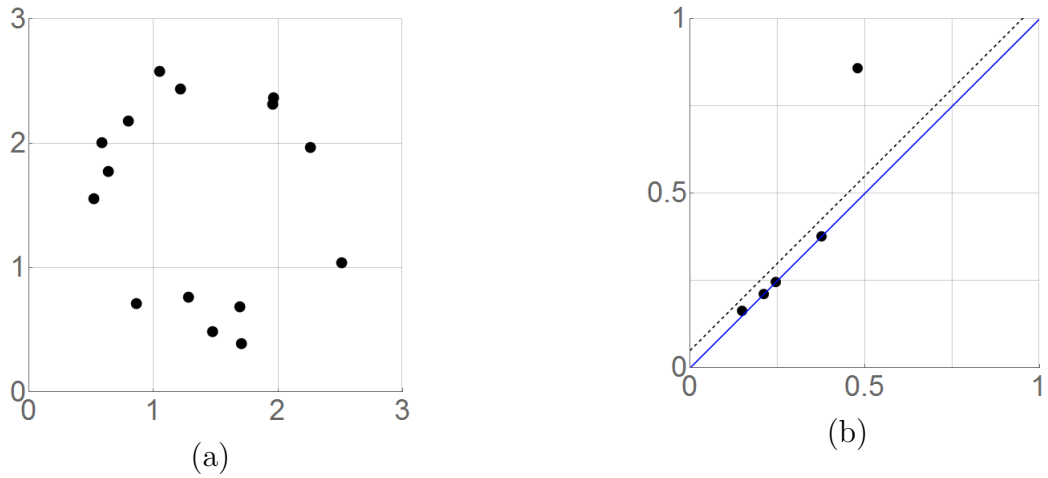

Figure 2.11: An example underlying dataset and its associated persistence diagram (separated into upper and lower halves by the dashed line). The persistence diagrams are used as the centers for the kernel density estimate.

These datasets each yield a Čech persistence diagram as described in Section 2.2 for homological dimension $k = 1$. A sample dataset and its associated $k = 1$ persistence diagram are shown in Fig. 2.11. Since these datasets are sampled from a perfect circle perturbed by small noise, one expects the associated 1-homology to have a single persistent feature with smaller features caused by noise.

From the collection of persistence diagrams, we build a kernel density estimate with a bandwidth of $\sigma = 0.05$. The kernel density estimate (KDE) with input cardinality 1 ($\hat{f}(Z)$ with $Z = (b, d)$) is shown in Fig. 2.12 (a); this local (i.e. fixed input cardinality) pdf is roughly Gaussian but does not inherit rotational symmetry from the Gaussian noise. Additionally, there is a second, much smaller mode, close to the origin (marked in Fig. 2.12 (a)). Next, we find the highest mode of the KDE with input cardinality 1, $(b_0, d_0) \approx (0.6, 0.9)$. Then, consider the slice of the KDE, $\hat{f}((b, d), (b_0, d_0))$, shown in Fig. 2.12 (b). This slice is similar to Fig. 2.10 (a), wherein all the center's upper features are given in the input; both instances are concentrated near the diagonal and show that volatile, short-persistence features are typically present at only certain scales.

By directly observing the collection of persistence diagrams, one can tell that the higher scale, low persistence features tend to appear in addition to the nearby high-persistence feature. In this sense, the large-scale features shown in Fig. 2.12 (b) yield information about

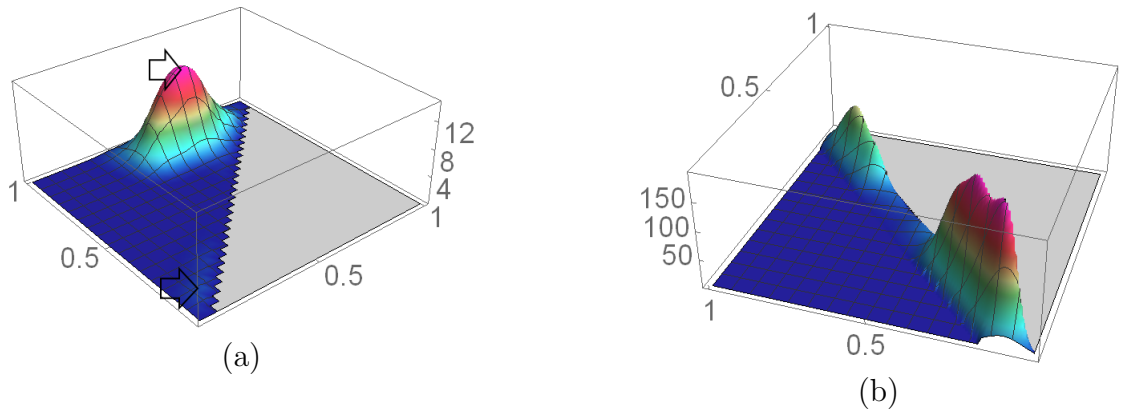

Figure 2.12: The kernel density estimate built from 100 dataset samples such as the one shown in Fig. 2.11. (a): The cardinality-1 estimate $\hat{f}(\xi)$ with its two modes circled and marked. (b): A slice of the cardinality-2 estimate $\hat{f}(\xi, \xi_0)$ where ξ_0 is fixed to be the larger mode on the left.

the distribution of persistence diagrams that is ignored by naïve 2-dimensional kernel density estimation, for which the long- and short-persistence features are indistinguishable due to their adjacency.

2.6 Discussion and Conclusions

We've presented a nonparametric approach to approximating density functions of finite random persistence diagrams. This includes the introduction of a kernel density function, as well as proof that the kernel density itself and its mean absolute deviation converge to those of the target distribution. Future work will investigate the convergence of powers of the absolute deviation (e.g., bottleneck variance) and deviations involving the Wasserstein metric (an L^p generalization of bottleneck metric, see [29]).

Our approach is fully data-driven, a necessary step since distributions of persistence diagrams were previously poorly understood. The assumptions $(A1)$-$(A3)$, $(A2)^*$, and $(A3)^*$ are typical for kernel density estimators [75]. Similar assumptions on the underlying data are inherited by the random persistence diagram, because variation in Čech persistent homology is controlled by interpoint distances. In particular, probability density decay follows the same trends as noise in the underlying data; this is seen in Fig. 2.12 (a) for Gaussian noise. Thus, the kernel density estimates defined here can be reliably used for data analysis,

adding a detailed tool to the methods used in topological data analysis. In particular, this is the first result yielding probability density functions which fully characterize a random persistence diagram. For applications in machine learning such as classification, the kernel density estimates carry information for generating more sophisticated features than previously available; e.g., the value of the global pdf at a specific input or list of inputs or the integral of the global pdf over a specified region. Access to a pdf also provides a means to check for classification robustness in terms of likelihood or Bayes factors, providing a measure of the confidence in a particular outcome.

Lending credence to applicability in data analysis, an example of kernel density estimation is presented in Subsection 2.5. In this example, underlying datasets are generated to lie on the unit circle with additive noise, a prototypical example for topological data analysis. Our analysis yields detailed information about the distribution of diagrams, even though only two 2-dimensional slices of the kernel density estimate are shown. We demonstrate that the kernel density is useful in practice by determining stable feature-containing regions in the random persistence diagram, regardless of their persistence.

In the context of Fig. 2.6, it is clear that sampling from the kernel density is straightforward, and in fact computation time scales linearly in the number of features in the center diagram $\mathcal{D}$. In contrast, precise evaluation of the kernel global pdf at a diagram requires the more thorough computations shown in Eq. (2.4.11). Evaluation of individual feature pdfs on the multi-wedge $\mathcal{W}_{0:a-1}$ only scales quadratically on the cardinality $|\mathcal{D}|$ and higher degree calculations are required only for combinatorics on the large persistence features in D^u. Consequently, these calculations are tractable so long as D^u does not grow too much in cardinality, while an increased cardinality for D^ℓ has a lesser effect on computation time.

Since the associations dictated by $\gamma \in I(j, N_i)$ (defined in Lemma 2.4.1) are known a priori, the calculation is parallellizable, and computation can be made rapid even for the evaluation of the global density function associated with diagram D shown in Fig. 2.5. Nevertheless, the density described in Eq. (2.4.11) is well organized for approximate evaluation. While Eq. (2.4.11) is sufficient for set integration, it is not symmetric under permutations of the inputs ξ_i, and consequently does not represent the density at a set

$\{\xi_1, ..., \xi_N\}$. A symmetric version is desireable for methods such as maximum likelihood or mode estimation [40]. Indeed, a symmetric pdf is available by summing over Π_N as per Eq. (2.3.7) to obtain the set derivative of the belief function. At no loss of accuracy, the large sum over Π_N need not range over all permutations and one may instead sort over compositions $\beta \circ \pi$ for $\beta \in I(j, N)$ and $\pi \in \Pi_j$ for each $j \in \{0, ..., |D^u|\}$. Since one expects that $|D^u| = N_u << N$, this reorganization significantly diminishes the number of computations.

The kernel density presented here treats the small persistent features in D^ℓ as a single group, since the proof of Theorem 2.1 uses very little information about the lower random diagram. Consequently, it may be helpful in practice to cluster the lower portion of the center diagram, followed by defining a random diagram centered at each cluster. This approach complicates the expression and evaluation of the kernel density, but does not complicate sampling from the kernel density. The goal of this approach is to more carefully capture the geometric features of the underlying random dataset, since such geometric features often correspond to briefly persistent homological features. For example, geometric features are of paramount importance for classifying periodic signals through their delay embeddings, wherein the large persistent feature indicates periodicity and thus is expected to appear in every class.

Chapter 3

Combinatorial Hodge Theory for Equitable Kidney Paired Donation

Summary

Kidney Paired Donation (KPD) is a system whereby incompatible patient-donor pairs (PD pairs) are entered into a pool to find compatible cyclic kidney exchanges where each pair gives and receives a kidney. The donation allocation decision problem for a KPD pool has traditionally been viewed within an economic theory and integer-programming framework. While previous allocation schema work well to donate the maximum number of kidneys at a specific time, certain groups of patients are rarely matched in such an exchange. Consequently, these methods lead to inequity in the exchange, and the same patients are often left unmatched repeatedly. Here we instead utilize computational topology methods to find kidney exchange cycles. Our topological methods naturally find the cyclic structure in a kidney exchange, and we use this structure in our search for an equitable kidney allocation. Another key result of our approach is a score function defined on PD pairs which measures disparity within a KPD pool; i.e., this function measures the variable chance for each PD pair to take part in the kidney exchange. Specifically, we show that PD pairs with underdemanded donors or highly sensitized patients have lower scores than typical PD pairs. Furthermore, our results demonstrate that PD pair score and the chance to obtain a kidney are positively correlated when using top trading cycles and chains and an integer

programming implementation (rCM). In contrast, the chance to obtain a kidney through our method is independent of score, and thus unbiased in this regard.

3.1 Introduction

While the medical procedure of organ donation has improved in past decades, the availability of organs is still quite scarce [2]. As a result, the waiting list for kidney donations is growing [2]. Live kidney donation increases the number of possible transplants and is often superior to a cadaveric kidney [84]. Many ideas have been discussed about how to encourage and optimize live organ donation in an ethical fashion to help alleviate the shortage of kidneys [70, 1, 50, 23].

Even if incompatible with their patient, a donor can still represent the patient in a kidney exchange. In the simplest case of kidney exchange, one patient-donor (PD) pair is matched up with another so that each donor can give to the other pair's patient. Larger *exchange cycles* (so-called because they occur in a loop), such as the triple in Fig. 3.1 are also possible. There are several reasons for incompatibility between a donor and patient, most commonly blood-type incompatibiliy and HLA (Human Leukocyte Antigens) sensitivity [84]. In particular, blood type AB donors can only donate to rare type AB patients, leading to a lower chance for such *underdemanded pairs* to take part in a kidney exchange. HLA sensitivity can arise from receiving blood transfusions, pregnancy, and transplanted organs [84], whereas blood type is permanent and inborn. Sensitivity is measured through a calculated panel reactive antibody (CPRA) percentage, with $CPRA > 80\%$ patients labeled as *highly sensitized*. Precisely, CPRA is the expected chance of HLA senstivity to a randomly chosen donor nationwide based on the results from a panel of HLA markers.

To determine the allocation of organs in a kidney paired donation (KPD) pool, most existing models seek to maximize a utility, with values given for each possible kidney donation. Donation utilities could include a fixed value, the quality adjusted life years (QALY) expected from each donation, and/or the probability of a successful transplant. Organ Procurement and Transplant Network (OPTN) has three main ethical concerns for its operations: utility, justice, and respect for persons. In [1], they describe a just allocation

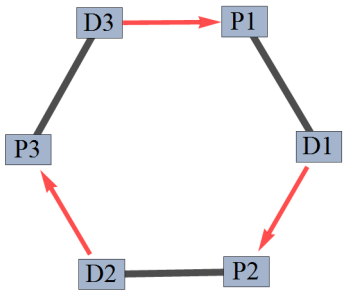

Figure 3.1: Depiction of a three-way kidney exchange cycle. The PD pairs are numbered 1 through 3 with P for patient and D for donor indicating each individual's role. Donations are indicated by red arrows from the donor to patient. For example, D1 represents P1, but is not compatible with P1, instead giving their kidney to P2. In a kidney exchange cycle, each PD pair both donates and receives a kidney.

method as "fairness in the pattern of distribution of the benefits and burdens of an organ procurement and allocation program"; we refer to this quality as *equity* in this paper. Ideally, an equitable distribution of organs does not depend on the difficulty in finding a compatible kidney. Despite the difficulty in such a task, an important equity goal is to mitigate powerful bias against any group, such as highly sensitized patients, minorities, or underdemanded pairs.

Equity and utility are in competition, as shown in [27] which boosts the number of kidneys allocated to highly sensitized patients and investigates the consequential loss of utility. The approach used in [27] requires the *a priori* specification of a disadvantaged group of PD-pairs considered less likely to be allocated in the kidney exchange. Then, the chosen group is given either bonus utility per allocation or a quota to fulfill. In contrast, we present a new algorithm that individually corrects for each PD pair and possible donation. Our algorithm considers each donation's potential exchange cycles together, creating a new description of the KPD pool with such global information ultimately summarized as a single value for every donation. Complementary, our algorithm *a posteriori* produces a score for each patient and donor quantitatively describing their relative advantage within the original description. Such a continuous score is more versatile than the discrete description of a PD pair as underdemanded or highly sensitized.

To rigorously define inequity, we consider a probabilistic notion of decision bias based on conditional allocation chance, parameterized by a function defined on PD pairs g. Consider

X to be a randomly chosen PD pair in a random kidney exchange and let f indicate kidney allocation. Then, the function $h(t) = \mathbb{P}[f(X) = 1 | g(X) = t]$ describes the expected dependence of kidney allocation on the parameter g; in particular, linear trends in the graph of $h(t)$ describe systematic allocation bias according to the parameter g.

Typical models describe a KPD pool as a kidney exchange graph (KEG) with PD pair vertices and edges drawn from donors to compatible patients. Optimized match [76] finds the highest utility pairwise matching of PD pairs using the blossom algorithm first described in [31]. While it is very fast, optimized match fails to include any exchange cycles with more than 2 patients, limiting its applicability. Indeed, [74] and [7] investigate the effects of allowing larger exchange cycles, with substantial differences in smaller KPD pools (e.g., with hundreds of PD pairs) and even more so in pools with many highly sensitized patients. The top trading cycles and chains (TTCC) algorithm [71] describes a kidney exchange as analogous to house allocation, wherein one decides who receives what space in college houses or offices. The simplicity of TTCC allows the description of kidney donation chains that start with an altruistic donor and end at any patient. The TTCC algorithm is fast and allocates larger exchange cycles. Indeed, [71] and [7] showed that the addition of donation chains is beneficial as a whole by enabling more donations, especially in sparse KEGs. In [88], optimal kidney allocation is solved as an integer programming problem, utilizing the random maximal matching (rCM) algorithm to find cycle allocations of length 2 or 3 with maximal cardinality, potentially under extra constraints.

On the other hand, *none* of the aforementioned algorithms directly analyze the equity in a kidney exchange. While some studies make strides to help particular disadvantaged groups, none seeks to quantify or eliminate general disparity in a kidney exchange as ours does. Moreover, our algorithm is shown to be competitively efficient, consistently outperforming TTCC and competitive with the maximal cardinality solution provided by rCM. The speed and scaling of our algorithm is faster than the traditional integer programming description of the KEG problem.

Overview

The rest of the paper is organized as follows. We begin by introducing, motivating, and describing our algorithm (Hodge Cycle) in Section 3.2. Precisely, we delve into some of the technical details behind simplicial cohomology with definitions, examples, and concluding with the Helmholtz decomposition theorem (Theorem 3.1). We then explain how Theorem 3.1 applies to finding exchange cycles in a graph. Finally, we outline and discuss the Hodge Cycle algorithm in detail. In Section 3.3 we discuss our results, starting with a couple motivating examples. The first of these gives a concrete and simple example of the decomposition in Theorem 3.1, while the second shows some specific instances of Hodge Cycle's behavior. Next we lay out the methods used to generate our synthetic kidney exchange graphs. Our first primary result uses Theorem 3.1 to produce a meaningful score which measures advantage within our KEGs and connects low scores to previously known disadvantaged groups (High CPRA and underdemanded pairs). Our second primary result demonstrates correlation between combined scores and the chance to obtain a kidney from TTCC and rCM, with little correlation between the scores and the chance to obtain a kidney from Hodge Cycle. We finish the results by testing the speed and efficacy of Hodge Cycle and comparing our method's allocation cardinality and overall utility with TTCC and rCM directly. Finally, we conclude and discuss our method in Section 3.4.

3.2 Methodology

Consider a kidney paired donation (KPD) pool as a directed graph, $G = (V, E)$ called a kidney exchange graph (KEG). Each vertex in G is either a patient or a donor. Representation edges are drawn from patient to donor, indicating that the donor represents the patient in the exchange as a friend or family member. Donation edges are drawn from donor to patient, indicating that the donor can give their kidney to the patient. Donation edges are given variable weights describing the utility of the donation, while a constant weight is given to all representation edges. Specifically, donation edges have utilities between 0.5 and 1.0 and representation edges' utilities are always equal to 1. There are no edges from

donor to donor or patient to patient, so any KEG is bipartite, which allows for extra runtime optimization.

In this formulation *every* kidney exchange cycle is described by an oriented cycle in the KEG; i.e., a cycle where the end of every edge touches the beginning of the next. Exchange cycles that share a vertex cannot be simultaneously allocated because every PD pair can take part in at most one exchange. Thus, the initial goal is to find a collection of vertex-disjoint exchange cycles maximizing the sum utility of all constituent edges; however, it is of paramount importance to achieve equity; i.e., to give disadvantaged PD pairs equal chance of obtaining a kidney as compared to non-disadvantaged PD pairs. As a result of this alternative goal, our method must achieve suboptimal utility, though it is desireable to minimize the loss of efficiency.

Combinatorial Hodge Theory

The primary mathematical tool used in the algorithm is 1-cohomology on a graph, for which we introduce the combinatorial gradient and divergence defined between spaces of functions on the vertices and edges of a graph. Preliminary results related to our work are briefly discussed; however, the interested reader may refer to [29] and [47] for further details.

We consider a graph G with vertices V and edges E and denote the space of (linear) functions on the vertices by $C^0(G)$; we will refer to such functions as scoring functions. Similarly, we denote the space of functions on oriented edges of G by $C^1(G)$ and call such functions edge flows. We will consider an edge flow $h \in C^1(G)$ as being defined on $V \times V$ with antisymmetry; for example, if $h(c, a) = 2$ on the edge from vertex c to vertex a, then $h(a, c) = -2$ on the oppositely oriented edge. Moreover, for (c, d) not an edge on the graph we define $h(c, d) = h(d, c) = 0$. Denoting the combinatorial gradient ∂_0, we obtain the following setup, called a chain complex for the graph G:

$$\{0\} \to C^0(G) \xrightarrow{\partial_0} C^1(G) \xrightarrow{\partial_1} \{0\}.$$

In particular, the adjoint $\partial_0^* : C^1(G) \to C^0(G)$ is the combinatorial divergence. Additionally, $\partial_1 : C^1(G) \to C^2(G)$ is analogous to the curl operator, although ∂_1 is trivial

on a graph. For functions $f \in C^0(G)$ on vertices and $h \in C^1(G)$ on edges, consider the following formulas:

$$[\partial_0 f](v, w) = f(w) - f(v) \tag{3.2.1}$$

$$[\partial_0^* h](v) = \sum_{w \in V} h(v, w). \tag{3.2.2}$$

According to Eq. (3.2.2) and antisymmetry, ∂_0^* sums outgoing edges and subtracts incoming edges on an edge flow, measuring the outward flow of h at the input vertex. Taking $f = \partial_0^*(h)$ defines a new function on the vertices. For example, consider the edge flow in Fig. 3.2, then $f(a) = 5 - 2 - 7 = -4$ and $f(b) = 11 + 3 - 5 = 9$.

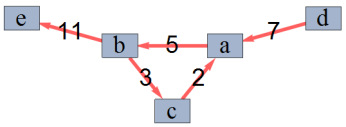

Figure 3.2: A graph with vertices a through e and an edge flow h defined on it via weights. For example, $h(a, b) = 5$ and $h(b, e) = 11$.

This setup enables us to define the 1-cohomology group of the graph G via group modulus by $H^1(G) = \ker(\partial_1)/\operatorname{im}(\partial_0) = \{x + \operatorname{im}(\partial_0) : x \in \ker(\partial_1)\}$. The elements of $H^1(G)$ are equivalence classes, where two elements of $\ker(\partial_1)$ are equivalent in H^1 when their difference lies in $\operatorname{im}(\partial_0)$. Each generator of the 1-cohomology group describes a hole or a cycle present in G and so $H^1(G)$ reflects the oriented cycle structure of the graph G. While this description shows promise for kidney exchange cycle allocation, the modular description of $H^1(G)$ poses a problem for direct application, so we appeal to Helmholtz decomposition [47]:

Theorem 3.1. *(Helmholtz Decomposition) Define the Helmholtzian (or 1-Laplacian)* $\Delta_1 : C^1(G) \to C^1(G)$ *as* $\Delta_1 = \partial_0 \circ \partial_0^* + \partial_1^* \circ \partial_1$. *The set of edge flows* $C^1(G)$ *on a graph* G *can be expressed in the following orthogonal sum:*

$$C^1(G) = \operatorname{im}(\partial_0) \oplus \ker(\Delta_1)$$

Moreover,

$$\ker(\Delta_1) = \ker(\partial_1) \cap \ker(\partial_0^*) \cong H^1(G)$$

via the canonical surjection $\phi(x) = x + \text{im}(\partial_0)$.

Helmholtz decomposition as presented above is a special case of Hodge decomposition for a graph [44]. As stated, Theorem 3.1 enables the decomposition of an edge flow $U \in C^1(G)$ into two orthogonal pieces

$$U = V + W, \tag{3.2.3}$$

where $V \in \ker(\Delta_1) \cong H^1(G)$ captures the globally cyclic portion of U and $W \in \text{im}(\partial_0)$ corresponds to a consistent scoring $R \in C^0(G)$ on the vertices (differences do not depend on a path) with

$$W = \partial_0 R. \tag{3.2.4}$$

The orthogonality of these two portions of the utilities U is critical for obtaining equity in a kidney exchange.

Hodge Cycle Algorithm

We describe the utility on a KEG as an edge flow U. Theorem 3.1 allows us to decompose the utilities as per Eq. (3.2.3), and Algorithm 1 begins by projecting onto $\ker(\Delta_1)$ to find V; because Theorem 3.1 canonically equates $Ker(\Delta_1)$ with the 1-cohomology of our KEG, V represents the tendency of the kidney donations to occur in (oriented) cycles. Moreover, Theorem 3.1 states that V and W are orthogonal; therefore, the new cyclic KEG ignores the inequity expressed by the scoring R, where R is defined implicitly through Eq. (3.2.4). Due to the bipartite nature of KEGs, each PD pair has a representation edge with clear 'incoming' and 'outgoing' donation edges. Consequently, PD pairs with smaller incoming and outgoing total utility are impacted with larger representation edge values in W and correspond to PD pairs belonging to fewer or smaller utility cycles. In particular, these PD pairs have donation utilities involving low-score patients and high-score donors which are higher in V than in U, and thus these edges will be more negative in $W = U - V$. This relationship allows us to use the scoring R to quantify and investigate inequity in KPD pools. Specifically, we use

the difference of patient and donor scores (i.e., the value of $-W$) to quantify each PD pair's relative advantage within the KEG.

In this fashion, the new KEG with utilities from V represents a kidney exchange which is indifferent to PD pair score. It is important, however, that as an *a priori* assigned score, R *cannot* measure inequity. We interpret inequity as correlation between allocation chance and a given parameter value; since inequity is a global and non-additive function of the allocation distribution, no locally defined score can be used to minimize inequity.

Since $\ker(\Delta_1)$ is isomorphic to the 1-cohomology, the new utility values in V can be expressed as a sum of weighted representing cycles C_i with weights α_i indicating the relative utility in each cycle: $V = \sum \alpha_i C_i$. Thus, the utility of each edge $e \in E$ represents the total preference of its resident cycles as opposed to the preference of the individual donation or PD pair: $V(e) = \sum_{\{i:e\in W_i\}} \alpha_i$. While this interpretation is non-unique, depending on the choice of representatives W_i, the resulting summary statistic is well-defined and offers an organization of cycle weights without requiring the identification all possible cycles. Such a cycle count is a daunting computational task, with the total number of unique cycles scaling at least exponentially with the number of edges in a graph. In some sense, our algorithm does not compare single donations, but rather the potential exchange cycles tied to each donation. Moreover, due to the particular structure of a KEG, we obtain a concrete interpretation for the patient donor scores, which are precisely the values of $-W = V - U$ evaluated on PD pair edges, wherein $U(e) = 1$. Since V and $-W$ differ by a constant when restricted to PD pair edges, one can interpret the PD pair score as an estimation for the relative number (and utility) of potential exchange cycles to which a PD pair belongs, compared to the average.

Despite the drastic change of viewpoint, the cyclic sum utility of any cycle is equal to its original sum utility. This follows because the sum of a gradient over any cycle is zero, much like integrating a conservative function over a closed curve. This means that any cycle-wide optimization on both U and V, such as integer programming methods, will not yield any new results. For these reasons, we use a greedy algorithm that *locally* seeks to optimize the new utilities on V by comparing individual donation edges, keeping in mind that edges in V contain non-local information about the original utilities U. In our results,

we ultimately verify the expectation that using V to locally optimize yields an equitable donation allocation.

The greedy algorithm typically begins with the highest utility edge in V. Since V is divergence-free, there is always a positive utility outward edge for any vertex belonging to at least one cycle. As a result, the algorithm *always* closes a cycle. After a vertex is visited twice, a cycle is found and it must be tested that the cycle is oriented in U; this is necessary to guarantee that the cycle represents a kidney exchange. If a cycle is *not* oriented in U, this means it contains edges from donor to donor and patient to patient and thus includes a donor who does not donate and a patient who does not receive a kidney; consequently, we refer to such cycles as bad cycles. In the event of a bad cycle, the starting edge for the next iteration is randomized to prevent a repeat. In practice, most cycles found are oriented in U; moreover, the costly projection step need not be redone until an oriented cycle is found and so this repetition is not a stumbling block for efficacy or computation speed. After an oriented cycle is found, the kidneys are allocated and all the cycle's vertices and incident edges are removed from the KEG and U, since their donations are now in use. The algorithm restarts with a new projection V for the smaller U, iteratively finding a collection of exchange cycles.

For a graph, we solve the graph Laplacian equation to find a scoring R with W = $\partial_0 R$ and then take V = U - W. This step uses most of the memory and computation of the algorithm, and primarily scales with the size of the matrix describing the Laplacian, or $O(n^2)$ where n is the number of vertices (patients and donors). Since the projection is done once for each cycle recorded and cycle length is typically small, the time to find a complete kidney allocation scales as $O(n^3)$. Our implementation is summarized in Algorithm 1. MATLAB code is available on the author's website at https://sites.google.com/site/joshmikemath/code.

3.3 Results

In all of our results we implement simple improvements to the greedy algorithm which require negligible local computation. First, we increase the depth of the greedy algorithm,

Data: A kidney exchange graph, G, with utility edge flow, U.
Result: A collection of cycles (representing kidney allocation).
begin
  $HM \leftarrow \Delta_1$ on G `// Helmholtzian matrix`
  $V \leftarrow \text{Proj}_{\ker(HM)}U$ `// Globally cyclic portion`
  BadCycle $\leftarrow$ **False**
  while $V \neq 0$ **or** not max iteration **do**
    if *BadCycle* **then**
      $v(1) \leftarrow$ a random vertex in V
    else
      $v(1) \leftarrow$ source of highest utility edge in V
    end
    Cycle $\leftarrow \emptyset$, $k \leftarrow 1$
    while *Cycle is empty* **do**
      $e \leftarrow$ highest utility edge with source $v(k)$
      $v(k+1) \leftarrow$ target vertex of e
      if $v(k+1) = v(j)$ *for some* $j \leq k$ **then**
        Cycle $\leftarrow v(j)$ through $v(k)$
      end
      $k \leftarrow k+1$
    end
    if *Cycle is oriented in KEG* **then**
      BadCycle $\leftarrow$ **False**
      Record Cycle
      Remove Cycle from G and U
      $HM \leftarrow \Delta_1$ on G
      $V \leftarrow \text{Proj}_{\ker(HM)}U$
    else
      BadCyle $\leftarrow$ **True**
    end
  end
end

Algorithm 1: Hodge Cycle

weighting the next donation utilities by half and ignoring representation utilities. Second, we change utility for donations that finish a cycle to discourage large exchange cycles, which are challenging or unrealistic to actually perform.

Motivating Examples

First, we present the process used to decompose (as in Theorem 3.1) the utilities of an edge flow U on a small graph with easily identifiable canonical cycles (Fig. 3.3a). We readily identify two basis cycles $C1$ and $C2$ and thus present $\ker(\Delta_1)$ as $\{M(C_1) + N(C_2) : M, N \in \mathbb{R}\}$ (Fig. 3.3b). To find the globally cyclic portion, V, of U, we can use Gramm-Schmidt to find an orthonormal basis for $\ker(\Delta_1)$ and project onto this basis (Fig. 3.3c). Alternatively, one can minimize the Euclidean distance $\|W\| = \|U - V\|$, which in general leads to the graph Laplacian equation used in practice, or use singular value decomposition of Δ_1 to find a different orthonormal basis. Theorem 3.1 tells us that there are only 2 pieces in our decomposition ($U = V + W$) and therefore $W = U - V$ is the gradient portion of U (Fig. 3.3d). The gradient portion is used to find the scoring R so that $W = \partial_0 R$. One arbitrarily defines the score for a single vertex and subsequently follows the graph $W = U - V$ to find the other score values (Fig. 3.3e). The scoring function is then shifted to obtain a mean of 0 to allow better comparison among KEGs (Fig. 3.3f).

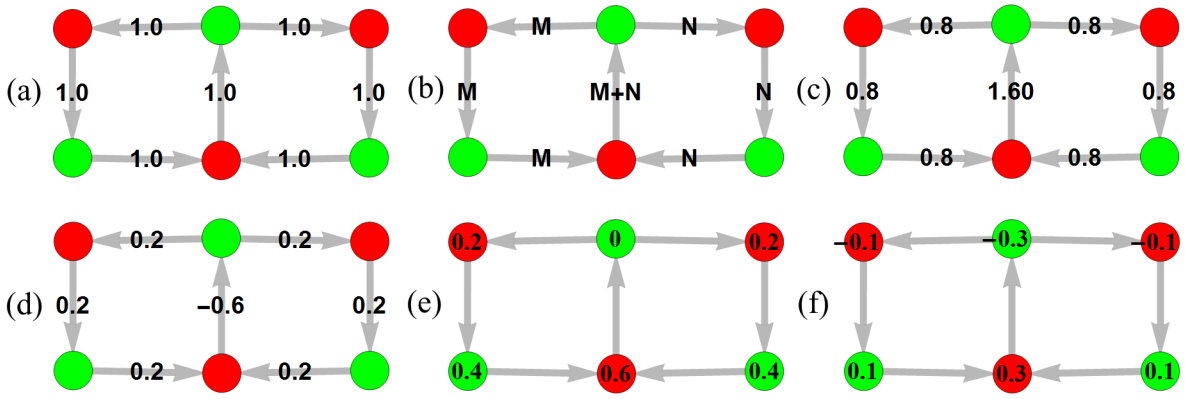

Figure 3.3: An edge flow and the process used to Hodge decompose it. (a) Original edge flow U. (b) Space of cyclic flows, $\ker(\Delta_1)$. (c) The globally cyclic portion V minimizing $\|U - V\|$. (d) The gradient portion, $W = U - V$. (e) A scoring made by assigning 0 to one vertex value. (f) A new scoring with mean 0.

Next, Fig. 3.4 presents a more complicated graph where the canonical cycles are not as clearly identifiable. In this case, the edge flow includes different utility values. This enables the example to exhibit differences in greedy algorithm behavior caused by the cyclic projection. For example, the cyclic portion in Fig. 3.4b has lesser utilities for shortcut donation edges. Allocating such donations prevents a PD pair from belonging to *any* cycles in the future, which would lead to a strictly suboptimal allocation. For example, two edges in particular (with weights circled in Fig. 3.4a and 3.4b) change from high utilities of 1 to very low utilities of 0.21 and 0.30 after projecting, causing the greedy algorithm to allocate different (and fewer) donations than it would without the projection.

Fig. 3.4c depicts the cyclic portion of the new KEG after the first exchange cycle is allocated by our algorithm. The new cyclic portion separates into two cyclic exchanges. One cycle is weighted 0.66 and the other is weighted 0.38; the donations belonging to both cycles are weighted as the sum, 1.04. Without the cyclic projection step, the greedy algorithm prefers the smaller cycle even though the larger cycle includes strictly more PD pairs. This choice hinges on one pair of donations (with weights boxed in Fig. 3.4a and 3.4c), comparing utilities of 1 and 0.8 in U; after the cyclic projection these utilities are respectively changed to 0.38 and 0.66 in V_1, reversing the decision and allocating the larger cyclic exchange.

Recall that the greedy algorithm has been enhanced to improve performance and keep cycles short; indeed, by eliminating the depth parameter in the example, Hodge Cycle allocates only 7 of 8 kidneys while eliminating the cycle-closing multiplier yields a single unwieldy 16-cycle. These examples demonstrate the improvement gleaned from these additions.

Generated Data

Here refer to Table 3.1 for our choice of parameters. The KEGs are created with two primary descriptors from kidney donation taken into account: blood type and HLA (Human Leukocyte Antigen) sensitivity. The data are initialized by creating an array of patient donor pairs, where typical KEGs have between 50 and 200 PD pairs. Every patient and donor is then given a randomly assigned blood type according to two fixed multinomial distributions, one for patients and one for donors (Table 3.1, left). Each patient is also given a random

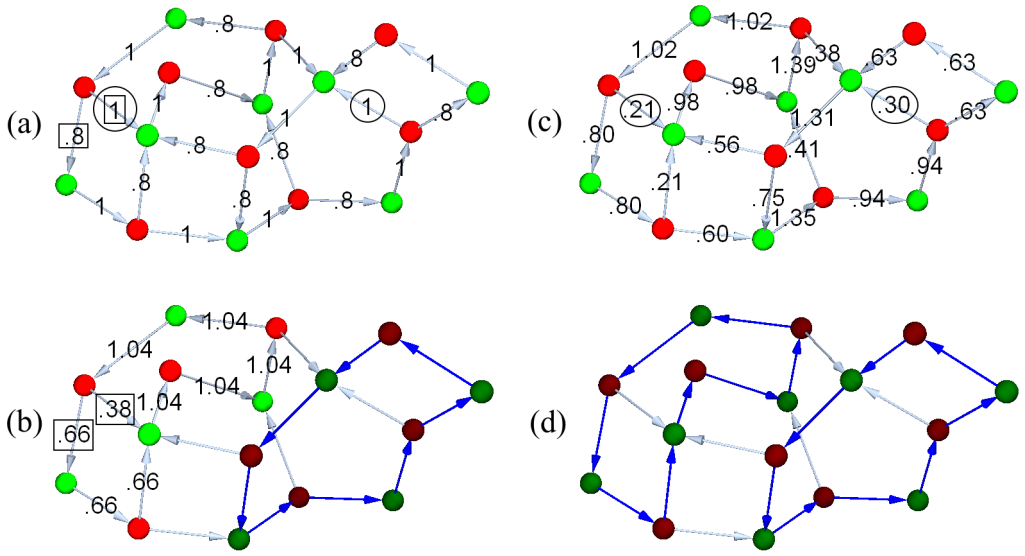

Figure 3.4: The progression of Hodge Cycle on a tiny KEG. Green nodes are patients while red nodes are donors. Recorded cycles are dark and blue. Removed edges are unlabeled. (a) Initial KEG, U. (b) Initial cyclic KEG, V_0. (c) One cycle allocated, V_1. (d) Two cycles allocated (empty).

calculated panel reactive antibody (CPRA) value. First, the patient is labeled as either low (0-10%), medium (10-80%), or high (80-100%) CPRA according to a fixed multinomial distribution (Table 3.1, right); then, the patient is given a specific CPRA value from a continuous uniform distribution appropriate to their CPRA level (low, medium, or high);

Most simulations herein are paired: one using all the US proportions and the other using all the uniform proportions (Table 3.1). The US proportions reflect the US average blood type proportions (Table 3.1, left) for waiting list patients and the whole population for patients and donors, respectively [3]. Using both proportions allows us to investigate our algorithm's performance in a realistic setting as well as its sensitivity and robustness. Since rarity and assymetry of blood types are confounded, assigning blood groups with equal probability allows us to investigate the disparity without the effects of rarity. In other simulations, we will vary the proportions of CPRA to investigate our algorithm's effectiveness on highly sensitized populations.

After the PD pairs have been initialized, we draw edges between the nodes. First, an edge is drawn from patient to donor for every PD pair and given a utility of 1; then, each other donor-patient combination is checked to potentially draw an edge from donor to patient. An

Table 3.1: (*(Left)* Proportions of Blood type for the generated KEGs. (*Right*) Proportions of CPRA level for the generated KEGs. Specifically, these are the probabilities used in our multinomial distributions. All US averages are taken from [3]. CPRA level is subsequently used to choose a specific value uniformly within the specified range (0-10, 10-80, or 80-100%).

Blood Type	US wait-list	US whole	Uniform
O	48.6%	44%	25%
A	32.7%	42%	25%
B	14.9%	10%	25%
AB	3.8 %	4%	25%

CPRA level	US wait-list	Uniform
Low	81.3%	10%
Med	11%	70%
High	7.7%	20%

edge is drawn from a donor to a patient when the following three conditions are satisfied: (i) they must not constitute a PD pair because there is already an edge going the other direction; (ii) they must be blood type compatible, which means that the patient must have all blood markers the donor has; (iii) a randomly generated percentage value must be greater than the patient's CPRA value, otherwise the patient is considered HLA sensitive. If an edge is successfully drawn from donor to patient, the utility is assigned a uniformly random value between 1/2 and 1. For scale, 200 PD pair KEGs using US average blood type and CPRA values have about 21,500 edges on average.

Measuring Equity

We examine that the gradient portion W in Theorem 3.1 creates a score which in context measures a PD pair's advantage (higher chance) of obtaining a kidney from a utilitarian allocation method, such as the methods in [71, 88, 27]. Recall that the score is a function on the vertices of the KEG (see Eq. (3.2.4)), giving a value for each patient and each donor. As discussed in the description of Hodge Cycle, low patient scores or high donor scores for a PD pair cause fewer and lower utility donations in the original KEG as compared to the globally cyclic portion used by our algorithm; indeed, here we demonstrate that highly sensitized patients (CPRA > 80) have low patient scores and that underdemanded (type AB donor) pairs have high donor scores; both patterns lead to a smaller PD pair score in $-W$. Recall

that a person with type AB blood can only donate a kidney to a patient with rare AB blood. Highly sensitized patients cannot receive most people's kidneys, reacting to many of their HLA markers. As a result, PD pairs of either group can take part in fewer exchanges, leaving them less likely to receive a kidney in a purely utilitarian allocation procedure.

We begin by generating KEGs with CPRA levels and blood types distributed according to US averages. Now we directly compare a patient's assigned CPRA value to their score in Fig. 3.5. These results show a negative, loosely logarithmic correlation between patient scores and CPRA, as well as an increase in score variability as CPRA increases. In particular, the average patient scores plummet at high CPRA values, expressing the extreme difficulty for highly sensitized patients to find a matching donor. Since AB-patients are rare in the US population (about 4% of the population), we expect AB donor pairs to have a disadvantage in obtaining an exchange cycle. Indeed, Fig. 3.5 (right) shows AB donors' (shown as red x's) restricted donations reflected with substantially higher donor scores.

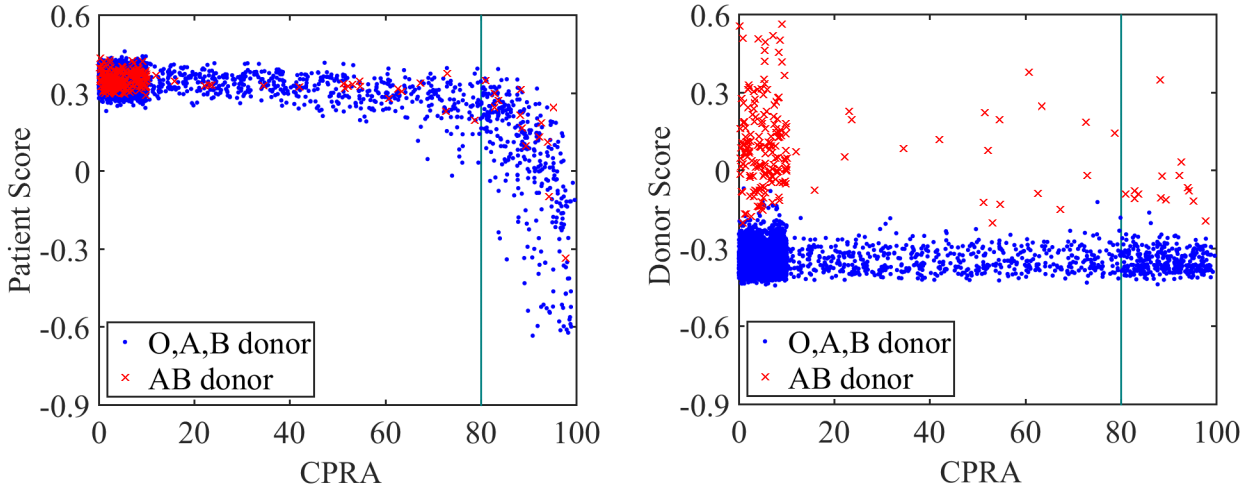

Figure 3.5: (*Left*) Patient score and CPRA. (*Right*) Donor score and representing patient CPRA. A comparison of vertex scores with generated patient CPRA values in KEGs with US proportioned blood types and CPRA values (as in Table 1). Both plots represent 50 random KEGs with 100 PD pairs each and show *every* PD pair. Additionally, the underdemanded PD pairs who have AB donors have been marked with red Xs.

In similar fashion, we compare scores with CPRAs in KEGs with completely uniform distribution of CPRA value and blood type (Fig. 3.6). With this very different KPD pool composition, we still see the same logarithmic shaped correlation between CPRA and patient

score, demonstrating the robustness of the scoring function. Moreover, by making blood type uniform, here we have reduced the effect of blood type disparity; this is reflected in Fig. 3.6 (right) by the difference in donor scores between AB donor pairs and the rest: much smaller than the effect of CPRA or the same variation with US average proportions shown in Fig. 3.5 (right). Lastly, note that blood type differences have little effect in Fig. 3.6 (left), while Fig. 3.6 (right) shows no correlation between donor scores and CPRA, demonstrating that patient scores are largely independent of donor characteristics and conversely.

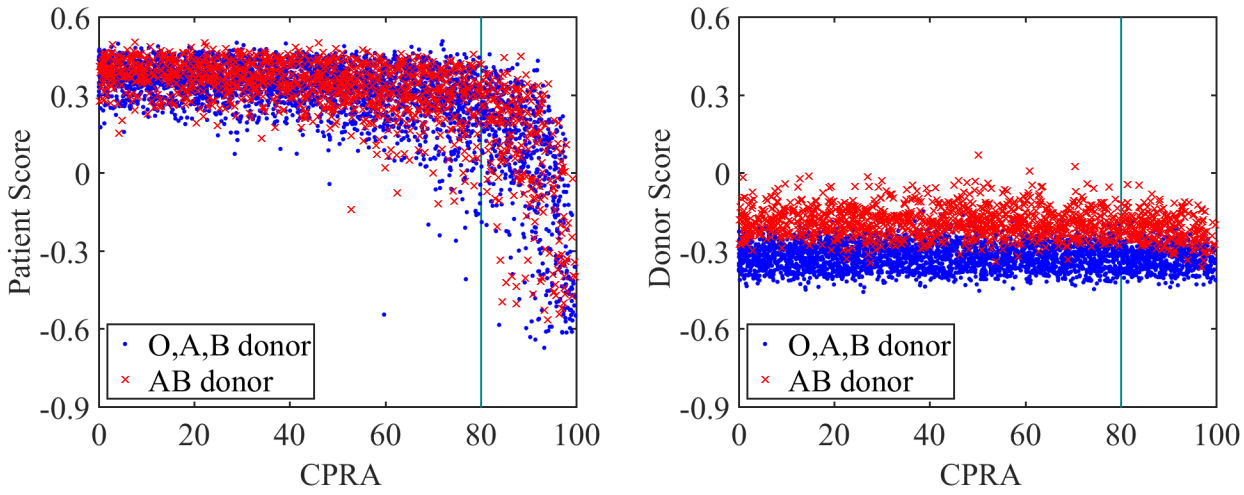

Figure 3.6: (*Left*) Patient score and CPRA. (*Right*) Donor score and representing patient CPRA. A comparison of vertex scores with generated patient CPRA values in KEGs with uniformly proportioned blood types and CPRA values (Table 1). Both plots represent 10 random KEGs with 100 PD pairs each, and show *every* PD pair. Additionally, the underdemanded PD pairs who have AB donors have been marked with red Xs.

To further understand the relationship between scores and equity, we directly investigate the chance of obtaining a kidney from Hodge Cycle (HC), Top Trading Cycles and Chains (TTCC), and random maximal matching (rCM) as a function of score. TTCC and rCM are presented here as our benchmark and as utilitarian allocation methods. We begin by assigning a combined score to each PD pair defined as the difference of the patient score and donor score, so that higher scores indicate an advantage while lower scores indicate a disadvantage in obtaining a kidney from a utilitarian allocation method. From these scorings we create probability density functions (PDFs) for (i) all PD pairs, (ii)-(iv) PD pairs allocated by HC, TTCC, and rCM (Fig. 3.7). The score PDFs allow us to compare each method's

treatment of equity within the KEGs. Moreover, including the score distribution for the whole population helps to indicate the scale of any differences. For example, all methods consistently show very high scoring PD pairs (say ≥ 0.75) always obtaining kidneys. The lower plots in Fig. 3.7 zoom into the left tail (representing disadvantaged PD pairs) because these probability densities are much smaller. The tails show HC allocating proportionally far more kidneys to low scoring (disadvantaged) PD pairs.

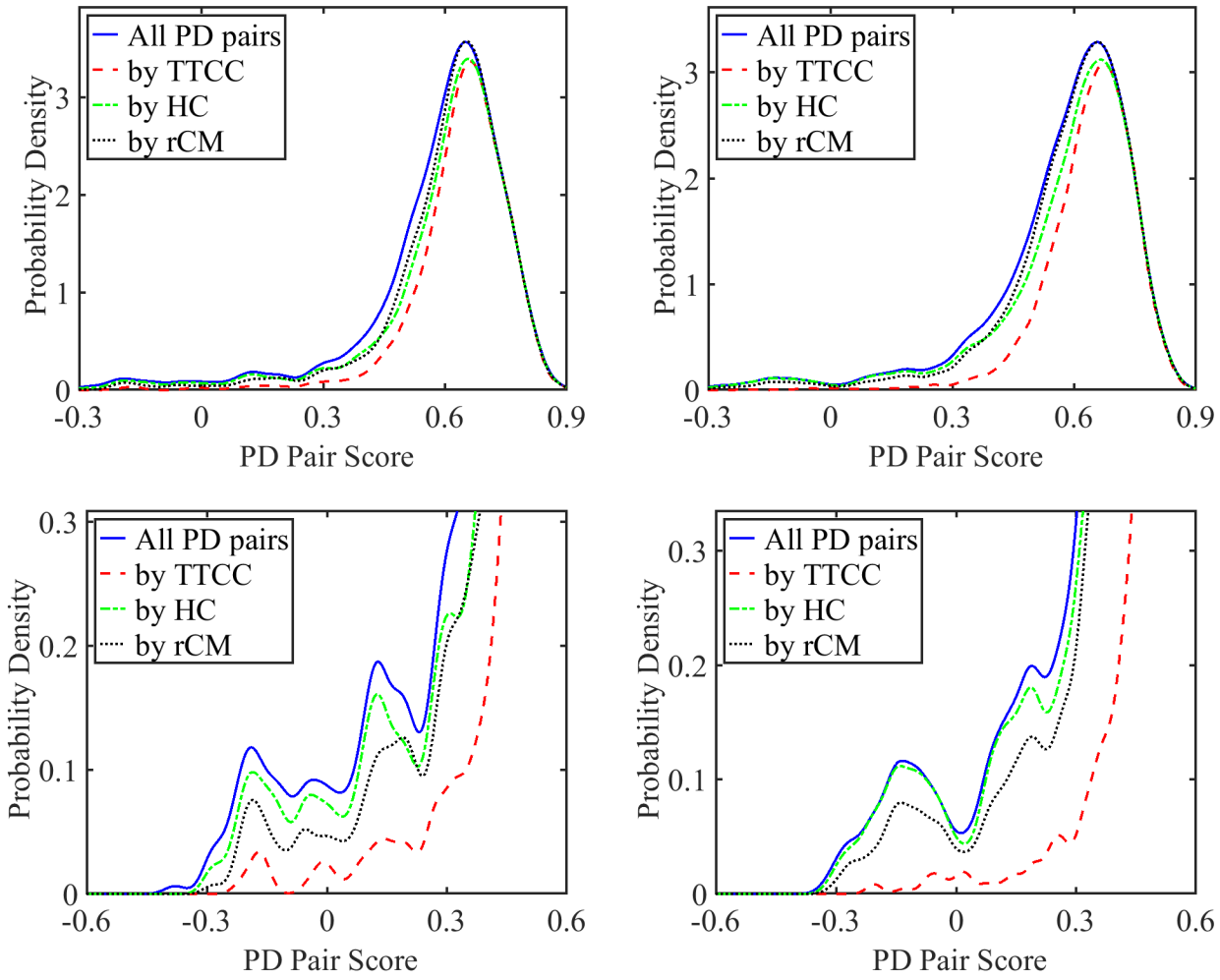

Figure 3.7: (*Left*) Scoring probability densities for US average blood type and CPRA. (*Right*) Scoring probability densities for uniform blood type and CPRA. Probability density functions for the score of a random PD pair, given allocation by HC, TTCC, or rCM. The bottom plots are zooms of the above plots' left tails. Each plot represents 50 random KEGs with (left) 50 PD pairs each or (right) 100 PD pairs each. Pairs without an obtainable cycle have been removed from the score determination.

Finally, we use these probability densities to plot the proportion of PD pairs which are allocated a kidney as a function of their combined PD pair score, shown in Fig. 3.8. Here we see that TTCC and rCM strongly prefer allocating kidneys to higher scoring pairs. This trend clearly depicts the quantitative aspect of the scoring, as the chance of obtaining a kidney smoothly increases with the score. This correlation, along with the trends in Figs 3.5 and 3.6, demonstrate inequity in kidney allocation with specific bias against highly sensitized and underrepresented PD pairs. On the other hand, Hodge Cycle does not show particular bias to any PD pairs. Indeed, Fig. 3.8 exhibits the primary objective of the Hodge Cycle algorithm: independence between kidney allocation and PD pair score. Note that many disadvantaged patients are not shown in these allocations because they had no possible exchange cycle in their entire KEG. In particular, 1.72% of all pairs are missing in Figs 3.7 and 3.8 (left) and 2.01% in Figs 3.7 and 3.8 (right).

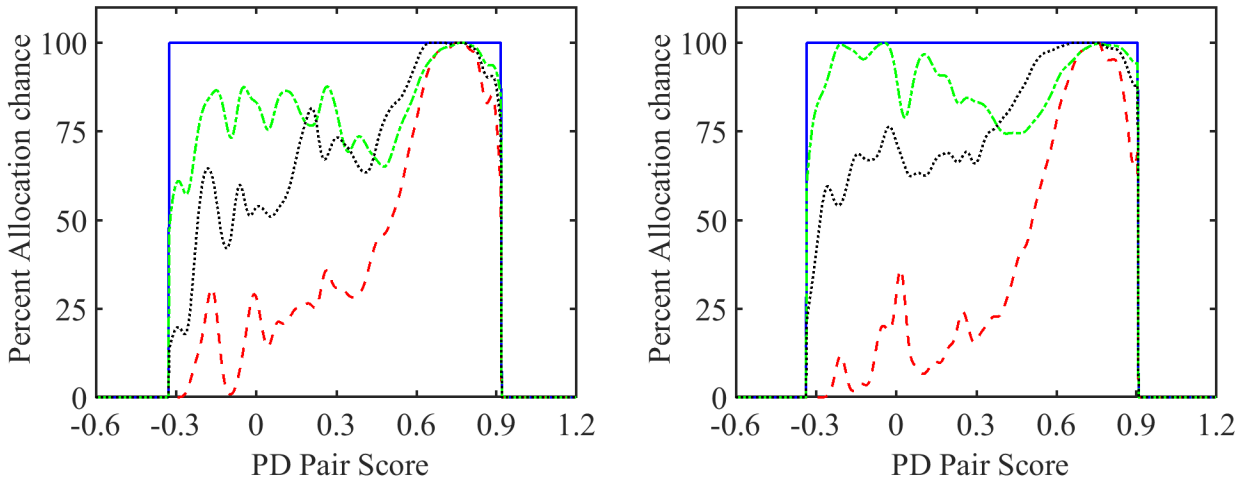

Figure 3.8: Chance to obtain a kidney via HC, TTCC, or rCM as a function of PD pair scores (patient score minus donor score). The solid blue line indicates the range of occuring PD pair scores. These plots are likelihood ratios of the densities seen in Fig. 3.7; specifically, these values are the conditional probabilities of being allocated a kidney given a particular score.

Hodge Cycle Performance

Now we analyze the efficiency and efficacy of Hodge Cycle (HC) by measuring the time ellapsed, number of donations allocated, and the average utility in various KEGs. As before,

we use Top Trading Cycles and Chains (TTCC) and random maximal matching (rCM) as benchmarks for the performance of our algorithm.

Timed trials of HC yield the average times seen in Fig. 3.9. This graph demonstrates that Hodge Cycle is very efficient on sparse KEGs, but runs slower on dense KEGs. As expected from the description of our algorithm, the time taken to find a full cycle configuration scales at $O(n^3)$ in the number of PD pairs, n. Computation time also scales linearly with the density of the graph. Decreasing patient sensitivity increases the graph density by increasing the chance of drawing edges. Computation time does not seem to be a severe problem for regional scale KEGs; for example, a KPD pool with 1000 PD pairs should take about an hour to complete at the given speed.

All timed simulations in Fig. 3.9 were performed on a Sager laptop with an Intel Core i7 and 8 GB RAM. Due to the memory requirements of Hodge Cycle, the timed KEG sizes are typically under 15,000 edges.

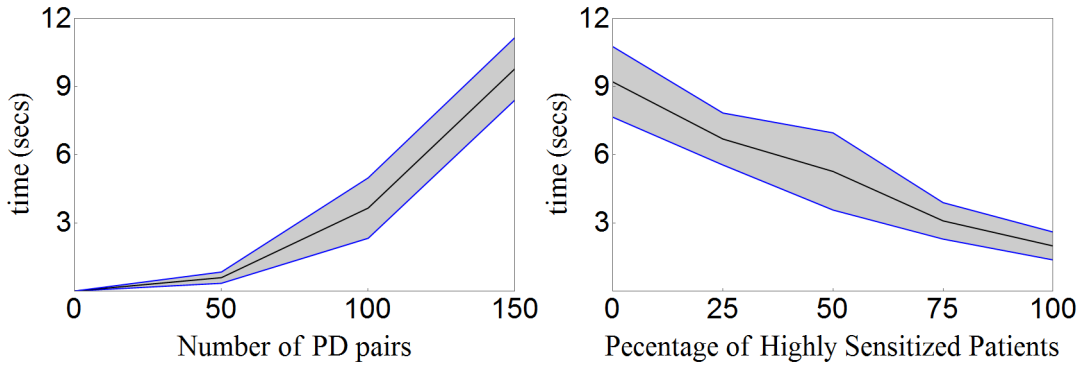

Figure 3.9: (*Left*) KEGs with US proportions and varying numbers of PD pairs. (*Right*) KEGs with 150 PD pairs and varying patient sensitivity. Both plots describe the average time for Hodge Cycle to find an allocation with one standard deviation error bars. Non-high CPRA patients are split evenly between low (0-10%) and medium (10-80%) CPRA.

Fig. 3.10 compares the proportions of patients who obtain a kidney from HC, TTCC, and rCM. In particular, we demonstrate that HC consistently outperforms TTCC while HC performs similarly to the cardinality optimal rCM. This small discrepancy shows HC to be competitive, especially since rCM directly maximizes cardinality wihle ignoring equity. Additionally, Hodge Cycle is shown to be persistently effective when solving KEGs containing primarily highly sensitized patients.

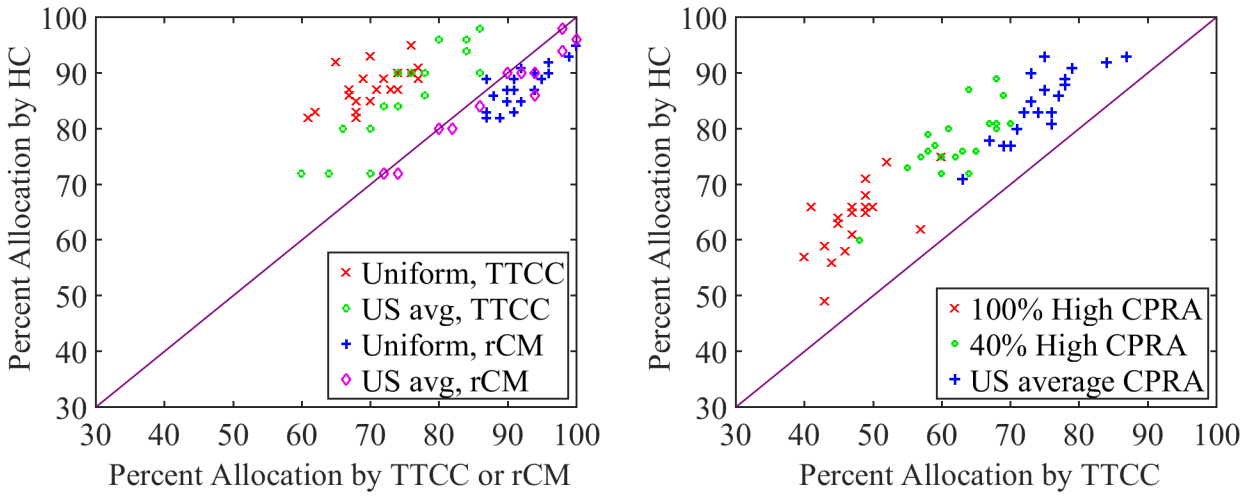

Figure 3.10: (*Left*) KEGs with US proportions and 50 PD pairs each and KEGs with uniform proportions and 100 PD pairs each. (*Right*) KEGs with 100 PD pairs and varying patient sensitivity. Cross comparison of the percentage of patients who were allocated kidneys by HC, TTCC, and rCM algorithms. Each point represents a single randomly generated KEG.

Now we investigate the utility of the exchanges that have been allocated; in order to do this in a careful fashion we measure the average utility of allocated donations. For each KEG, the average utility is the sum allocated donation utility divided by the number of patients who obtained kidneys. Fig. 3.11 compares the proportion of kidneys allocated to the average donation utility on a series of KEGs. Note that the average donation utility of a kidney exchange often suffers when many highly sensitized patients are involved in the allocation. Moreover, the average utility depends on the sparsity of the graph (as a result of patient CPRA) and not necessarily the percentage of patients saved, as the measurements *within* each sesitivity group are uncorrelated. The average donation utility in the worst case scenario of entirely highly sensitized patients is about 0.82, which is still higher than the expected average over all possible exchanges of 0.75. rCM is not included in this analysis because the method does not use utilities and instead maximizes cardinality, and as expected its average donation utility was found to be approximately 0.75 in any case.

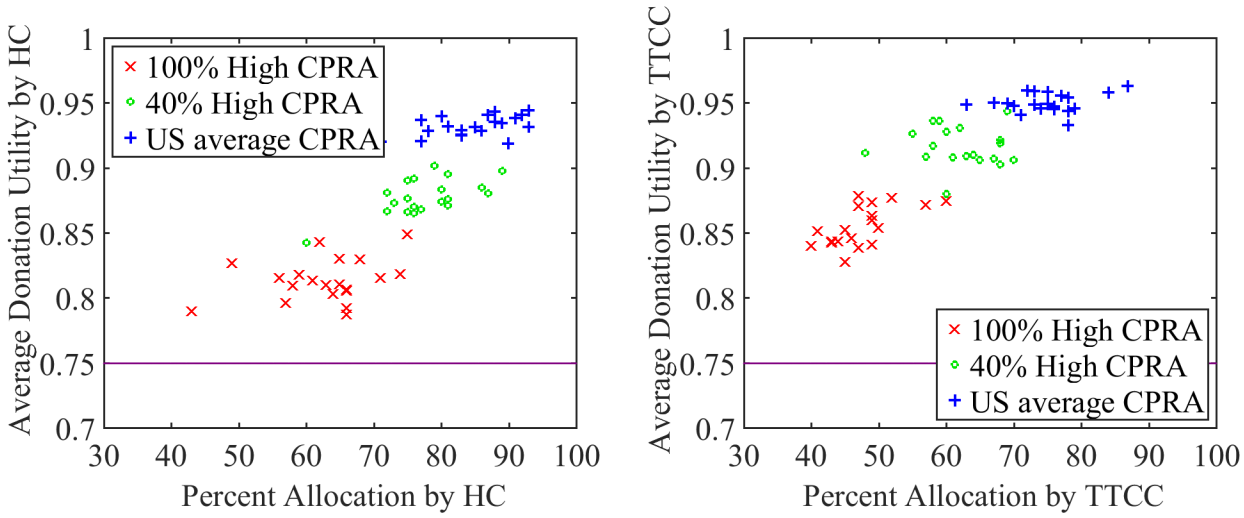

Figure 3.11: (*Left*) Average utilities allocated by Hodge Cycle compared to the percent of patients who were allocated kidneys. (*Right*) Average utilities allocated by TTCC compared to the percent of patients who were allocated kidneys. Each point represents a single randomly generated KEG with 100 PD pairs. 0.75 is the expected average of all donation utilities in a KEG, since they are chosen uniformly between 0.5 and 1. The same KEGs are used to create both plots, as well as Fig. 3.10 (right).

3.4 Discussion

We presented an algorithm focused on measuring and eliminating inequity on a KEG while simultaneously achieving near-optimal utility. When information is known about the relationships between the PD pairs in a KPD pool to create a KEG, Hodge Cycle measures each patient's disadvantage and then systematically eliminates inequity in the KPD allocation; therefore, the scoring distributions in Figs 3.5, 3.6, and 3.8 demonstrated the importance of building the entire KEG before making allocation decisions. In particular, this lack of bias could shorten wait times for sensitized patients who otherwise remain on dialysis while their chances deteriorate further. Ultimately, this supports and utilizes hospital protocols for obtaining HLA marker information from donors and HLA sensitivity information from patients. In general, an equitable allocation is desired by OPTN (Organ Procurement and Transplantation Network) and our algorithm can be used to explore and eliminate all types of bias, e.g., due to location, ethnicity, or blood type, so long as associated variables are included in the creation of the KEG and its utilities. In particular, African

Americans are known to have difficulties in obtaining kidneys [83] and liver transplantation suffers from serious geographical disparity [51].

Future study with Hodge Cycle will investigate the effects of an ongoing KPD pool with newcomers entering the pool over time. In particular, we will study the ongoing effects upon the population of highly sensitized patients. It is important to note that many of the highly sensitized patients were removed from Figs 3.5 and 3.6 since they were not a part of *any* cycles in their initial KEG, and consequently had no potential to obtain a kidney regardless of the allocation method. In typical time dependent models (such as in [96]) and in actual scenarios, these highly sensitized patients tend to collect in KPD pools due to their challenge in obtaining a kidney. We have already shown that Hodge Cycle has little to no bias against highly sensitized patients and therefore we expect that our algorithm will mitigate the growing collection of sensitization in KPDs.

Theoretically and experimentally, our algorithm scales at $O(n^3)$ on the number of PD pairs, n. We wish to speed the algorithm to enable use in larger KPD pools, since a national KPD pool could have upwards of 10,000 or more PD pairs and would be difficult for our algorithm as-is. Specifically, we will investigate using the 1-cohomology group to more efficiently construct a basis. Such a change in approach will also simplify parallelization and allow for more efficient memory usage. The local methods used, including the use of greedy algorithm for starting and subsequent edge choices, need to be investigated further. It is unclear whether a better method exists, especially when trying to balance equity with utility or cardinality. One possibility is to randomly choose edges according with probabilities proportional to the (local) weights of V.

The results of this paper were performed on data generated through populations averages from the literature (see Table 1 and [3]) due to the challenge in obtaining sufficiently detailed potential donation data, which is private and therefore has restricted access. While our data is built to reflect actual KPD pools, particularly the disadvataged groups studied herein, such considerations are still limited. Besides using only a couple of the most important medical indicators, the different characteristics and relationships in an actual KEG may be highly interdependent. Our method has been shown to level the field for both underdemanded and highly senstized PD pairs, and may be helpful to other groups as well.

Chapter 4

Species discrimination in paleobiology through computational geometry

Summary

One important and sometimes contentious challenge in paleobiology is discriminating between species, which is increasingly accomplished by comparing specimen shape. While lengths and proportions are needed to achieve this task, finer geometric information, such as concavity, convexity, and curvature, plays a crucial role in the undertaking. Nonetheless, standard morphometric methodologies such as landmark analysis are not able to capture in a quantitative way these features and other important fine-scale geometric notions.

Here we develop and implement state-of-the-art techniques from the emerging field of computational geometry to tackle this problem with the Mississippian blastoid Pentremites. We adapt a previously known computational framework to produce a measure of dissimilarity between shapes. More precisely, we compute "distances" between pairs of 3D surface scans of specimens by comparing a mix of global and fine-scale geometric measurements. This process uses the 3D scan of a specimen as a whole piece of data incorporating complete geometric information about the shape; as a result, scans used must accurately reflect the geometry of whole, undamaged, undeformed specimens. Using this information we are able to represent these data in clusters and ultimately reproduce and refine results obtained in

previous work on species discrimination. Our methodology is landmark free, and therefore faster and less prone to human error than previous landmark-based methodologies.

4.1 Introduction

Shape has often been used along with discrete morphologies to investigate many questions in biology and paleobiology, including species discrimination [15, 91, 54, 8], ontogeny [69, 14, 79] ecophenotypic variation [68, 93, 67], evolution via heterochrony [63, 62, 52], functional morphology [97], phylogeography [38], and many others [95]. Recent advances involving 3D landmark-based analysis have made marked improvements but nevertheless still rely on an expert handpicking a set of landmark points that represent the geometry of an entire object. The representation of shape by a relatively few landmarks is subjective because user-selected points are chosen for the ease of consistent identification by the user and not necessarily because they represent points with the greatest variance among groups. Additionally, important shape variation outside the landmarks, such as concavity versus convexity, will not be captured by these methods, limiting their usefulness in many situations.

To mitigate these problems, we apply a recent computational-geometry technique called continuous Procrustes distance, or CP-distance (see [6]). This methodology determines dissimilarity, or distance, between pairs of 3D surface scans of specimens drawn from mixed populations of species. Form taxa are best separated by incorporating a particular combination of geometric features of the 3D scans (such as curvature and area density) into the CP-distance algorithm. The statistical validity of our process is established by contrasting our results with those obtained by [8]. More precisely, our benchmarking process shows that the data set contains two clusters and that this classification is equivalent–with a high degree of confidence–to the ones formed by empirically assigned groups of [8].

Computational geometry allows us to consider the entire surface 3D scan as data for comparing specimen surfaces, opening the door to incorporating information such as curvature into the analysis. Our results provide strong experimental evidence supporting the use of these techniques for similar studies in different taxa. Our procedure is still timeconsuming because of the scanning and mesh-generation phases, which usually take

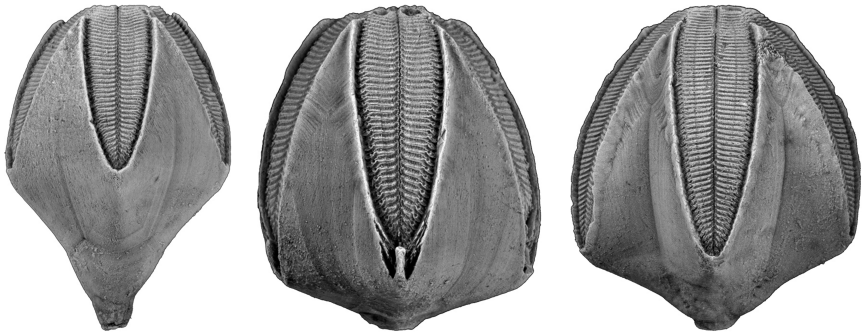

Figure 4.1: *(Left) Pentremites pyriformis. (Middle) P. tulipiformis. (Right) P. fredericki. (All) P. pyriformis* and *P. symmetricus* are the pyriform samples in this study, while *P. tulipiformis*, *P. fredericki*, and *P. spicatus* are the gadoniform samples.

about 1 hour per specimen. This motivated us to explore the effect of lowering scan resolution, which can vastly improve overall speed. To test this concept, we replicate our analysis using artificially produced lower-resolution scenarios. Our proposed methodology is considerably foreshortened because it does not require an expert to choose and record landmarks on the 3D scans. Rather, only a single small hole is made in the mesh at an easily identified homologous point: in this case the circular stem facet at the base of the theca (see Fig. 4.1 or Fig. 4.2). Our techniques also reduce the potential error resulting from user bias and open the door to incorporating fine-structure, curvature-based features into the analysis.

For this study we investigate species discrimination in the Mississippian blastoid echinoderm Pentremites. The bud-shaped theca of blastoids houses the viscera and is commonly well preserved three dimensionally. Much confusion exists concerning species delimitation in Pentremites because: (1) most populations, especially in the Late Mississippian, show high morphological variation; (2) several species cooccur in most localities; and (3) methodology used to describe shape of species lack the power to differentiate specimens into species groups [92, 8]. Pentremites species can be divided into two morphological groups based on proportions of the theca. The pyriform group includes taxa with an elongated pelvis (the lower part of the theca below the ambulacra) that is similar in size to the vault (the upper part of the theca bearing the ambulacra). The godoniform group includes taxa with a foreshortened pelvis that is much smaller than the vault (see Fig. 4.1). Species delineation within these groups has traditionally relied on the

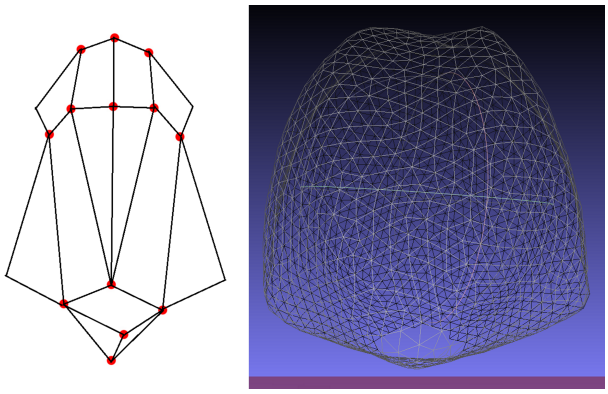

Figure 4.2: A visual comparison of the data used for computing discrete Procrustes distance (left image) versus the data used for continuous Procrustes distance (right image). (See the Approach and Algorithms section for precise definitions). The images above represent samples in the same orientation for clarity of reader comparison, though both methods are unaffected by orientation. Although thirteen 3D landmark points are chosen (by hand) for computing discrete Procrustes distance [8], the entire 3D scan is used for computing continuous Procrustes distance in this work.

presence of concavities and convexities of ambulacra and interabulacra, vault:pelvis ratios, identification of discrete morphologies, and stratigraphic arguments [39, 53, 92, 8]. Thecae of the blastoid Pentremites, the main focus of the comparison paper by [8], are ideal for the use of advanced comparison techniques. While three plate junctions between thecal plates offer obvious type 1 landmarks [37, 8], much of the shape variation in Pentremites occurs in the geometry of the plates themselves; since these lie between landmarks, such details cannot be easily quantified by geometric morphometrics. Moreover, there is great morphological variance within populations of each species, including genetic differences, ecophenotypic variation, and allometry [53, 92, 8]. Although this pilot study focuses on species discrimination, understanding the continuum of shapes for each species using this new technique can ultimately result in more insight on a diversity of issues that will be explored in a forthcoming work.

4.2 Materials and Methods

Specimens and Scans

In this study we use two data sets:

- A set of landmark configurations obtained from [8] via Dryad. These data consist of a set of 13 ordered, handpicked landmark points for each of 52 specimens. The data set is considered as the benchmark group in our study, and we shall use it to validate our algorithm.

- A new set of 3D scans of 20 Pentremites, chosen to closely resemble those used by [8] to better compare the effectiveness of our proposed methodology. Many of the [8] specimens had to be eliminated a priori because they were incompletely preserved or were heavily encrusted by adhering matrix. This is a limitation of the CP-distance algorithm which naturally arises when taking the entire surface scan as input data. All specimens were collected from the Upper Mississippian (Chesterian) Glen Dean Formation near Hopkinsville, Kentucky (see [8] for details). All specimens of Pentremites used in this study and in [8] are large, presumably mature individuals chosen to limit allometric effects.

Scans in the new data set were obtained using the Next Engine 3D scanner to capture and preprocess 3D meshes of each blastoid specimen. We used MeshLab to artificially cut out a very small hole in the bottom of each 3D scan, a technical requirement of the CP-distance algorithm. The region removed, the circular stem facet at the base of the theca, was consistent among the specimens and easily identified in the scans. The cutout removes about 0.5% of the points in each scan. This process ensures that the surface has the topological type of a disk; while this is required for the algorithm as is, it is possible to define and compute CP-distance for sphere-type surfaces as well.

Approach and Algorithms

Discrete Procrustes distance, or DP-distance, is a standard process for comparing a pair of shapes using a predetermined set of landmark points demarked on each surface (i.e., a landmark configuration). This well-known technique gives a quantitative measure of dissimilarity between configurations and was used in [8] to verify the classification of species. Specifically, DP-distance finds the minimum value that the sum of all distances between corresponding points within the two landmark sets achieves as the configurations are "aligned" in space (see [8] for details). In contrast, the CP-distance method developed

in [6] is an extension of the DP-distance. It incorporates full geometric information (readily available in 3D scans) such as curvature into the analysis. DP-distance is computed via minimizing a sum of distances over a discrete set of landmarks, whereas CP-distance is computed via minimizing integrals of geometric quantities. Both methods are scale and rigid motion invariant. This is to say, the outcomes of the calculations are independent of the size and position/orientation of the scans and resulting landmarks. For specific details, see [6].

In this study we use DP-distance and CP-distance to generate dissimilarity matrices associated to each data set. Each matrixs (ij) entry corresponds to the distance (or dissimilarity) between specimen i and specimen j. These matrices are symmetric, nonnegative, and have only zeros along the diagonal. We first use the DP-distance on pairs of landmark configurations from [8] to generate the matrix D^{At}, in an attempt to reproduce their results in the dissimilarity-matrix language of this study. We then use the CP-distance algorithm on pairs of new 3D scans to generate the dissimilarity matrix D^{CP}. Finally, we devise a methodology for comparing these matrices and determine whether they carry the same "clustering information" within them.

Our specimen scans were performed at the (medium) resolution of 4400 points per square inch. At this resolution, a mesh of about 5000 points represents a typical specimen (Fig. 4.2). To determine the breaking point of our method, we reran the whole matrix comparison process using 3D scans artificially downsampled to resolutions of 50%, 20%, 10%, and 5% of the original scans. Downsampling was carried out using quadric edge collapse in MeshLab.

In this work we deploy three methods for comparing dissimilarity matrices and their cluster information:

1. Mantel's Test (performed on two dissimilarity matrices of equal size) is used to check for correlations between each distance's interspecimen measurements.

2. A multidimensional scaling algorithm is used for visualizing cluster information.

3. Aggregate clustering is applied to understand groupings at different scales.

1. Mantels Test Methodology. – Mantels test is a standard statistical tool used to compare distance matrices, often used in a biological setting with actual distances [81]. Here we use

Mantels test to compare the distance matrices D^{CP} and D^{At}. It is important to note that both matrices represent mathematical metrics on the same kinds of objects, making this comparison a classical and direct use of Mantels test. The dimensions of these matrices must be equal for Mantels test to be used. Because we were not able to obtain full 3D scans of all the specimens of [8], the dissimilarity matrices DAt and DCP are of different sizes. To overcome this and at the same time use the most information possible, we proceed as follows. We randomly choose 20 specimens from within Atwoods landmark data (52 specimens total) and produce a "submatrix" of D^{At} corresponding to the intersample DP-distances between these 20 specimens; moreover, we make sure that the number of specimens from each species matches the amount in D^{CP}. This sampling method means that the results of our test will address how similar the metrics are when considering the comparison of different species. We then carry out Mantels test between D^{CP} and the randomly sampled submatrix of D^{At}. The sampling process is repeated 5000 times, 1000 times for each one of the four possible resolutions: 100%, 50%, 20%, and 10%, and once to cross-compare D^{At} with itself to see the effect of the sampling method on correlation values.

In addition, Mantels test was also performed pairwise between different resolutions of D^{CP}. This process was repeated 10,000 times, 1000 times for each possible pairing. We note that at 5% resolution, three of the samples could not be processed by the CP-distance due to lack of mesh points. (We observed a threshold of about 80 mesh points for the algorithm to work.) Because of this, our Mantels test with 5% resolution always involves sampling of *P. tulipiformis*.

2. Multidimensional Scaling (MDS) Algorithm. – MDS is a tool used for dimensional reduction or presentation of complicated high-dimensional data [22]. We use this procedure to obtain a "visual realization" of the cluster structures that lie within the dissimilarity matrices. The MDS methodology differs little from the principal components analysis method (PCA) used in [8] but presents some graphical advantages. For example, it allows us to remove the requirement of a global landmark alignment. For our aggregate clustering, MDS was used with six dimensions. This number is chosen based on a drop in the eigenvalues of the resulting covariance matrix, in a manner similar to what is done for PCA (see Fig. 4.3). The dimension is fixed on all runs to obtain consistency across the clustering procedure.

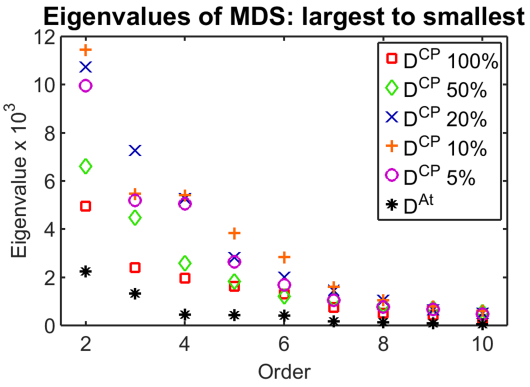

Figure 4.3: Diagram depicting the second through tenth eigenvalues of the MDS procedure for each dissimilarity matrix. For most matrices, there is a noticeable decline between eigenvalues 6 and 7, and so 6 eigenvalues were used for all clustering algorithms at all resolutions, for consistency. The first eigenvalue is always much larger than the rest, and so is omitted for scale purposes.

3. Aggregate Clustering (AC). – AC is a well established method [86]. It is particularly useful for determining clusters within data sets. It produces dendrograms, which convey grouping information at different scales (Fig. 4.6). Within AC, we use the complete algorithm for all resolutions of D^{CP}, because it is appropriate for high-dimensional data sets but does not assume our clusters have any particular shape [36]. The Ward algorithm was used for D^{At}, because it is the best general agglomerative method, we expect elliptical clusters from our MDS graph [36], and the Ward algorithm yields results most similar to those in [8].

To test for robustness, we performed a cross-validation. The Ward AC was repeated on random samples of 90% of the (D^{CP} or D^{At}) specimens. The resulting clustering is given labels and used to classify the remaining 10% by the cluster of each specimens nearest neighbor. This classification is then compared with the results of [8] for benchmarking purposes; we aim to reproduce the number of correct identifications. For this process, the data was always split into three clusters, with labels of (1) *P. fredericki*/ *P. spicatus*, (2) *P. tulipiformis*, and (3) *P. pyriformis*. The same labels are used to describe the benchmark group, although *P. spicatus* is not present in these data.

Table 4.1: *(Left)* Average correlations between D^{At} and different resolutions of D^{CP} via Mantels test. These simulations all involve the same sampling scheme, so they can be directly compared. *(Right)* Table of average correlations between different resolutions of D^{CP} via Mantels test. No sampling occurs except in the rightmost column.

	D^{At}
D^{At}	0.95
$D^{CP}100\%$	0.81
$D^{CP}50\%$	0.82
$D^{CP}20\%$	0.86
$D^{CP}10\%$	0.73

	$D^{CP}100\%$	$D^{CP}50\%$	$D^{CP}20\%$	$D^{CP}10\%$	$D^{CP}5\%$
$D^{CP}100\%$	1	0.97	0.92	0.88	0.90
$D^{CP}50\%$		1	0.92	0.89	0.89
$D^{CP}20\%$			1	0.90	0.89
$D^{CP}10\%$				1	0.98
$D^{CP}5\%$					1

4.3 Results

For each level of resolution, we performed Mantels test with 1000 iterations over 1000 random samples of submatrices of D^{At}. Our results (Table 4.1, left) show that the correlation between the submatrices of D^{At} and D^{CP} is consistently above 70%, with a very low p-value (p-value ¡10-6) at all levels of resolution. Considering all possible multiple-way comparisons, we can express an overall p-value of at most 0.00072.

Mantels test was also performed between different resolutions of CP-distance. Our results (Table 4.1, right) show correlation between 100% and 50% resolution is above 97%. Similarly, 10% and 5% resolution are highly correlated. The correlations between higher and lower resolutions are lower, suggesting a qualitative change as resolution drops.

The next step in the comparison process was to implement the MDS algorithm on the dissimilarity matrices D^{CP} and D^{At} at all five resolutions. (Note that this method does not require the matrices to be of equal size.) The output of this algorithm consists of the plots displayed in Fig. 4.4. An interesting observation is that the variance observed in the plot within each species' cluster increases as the resolution decreases. Careful attention to individual specimens also reveals that the horizontal component in the plot captures how elongated specimens are and thus divides the pyriform and godoniform groups quite naturally. On the other hand, it appears that the vertical component in the plot captures the concavity between ambulacra. However, since the MDS algorithm does not embed the data set directly onto coordinate axes (as does the alternative method, PCA), these interpretations are rather hard to prove.

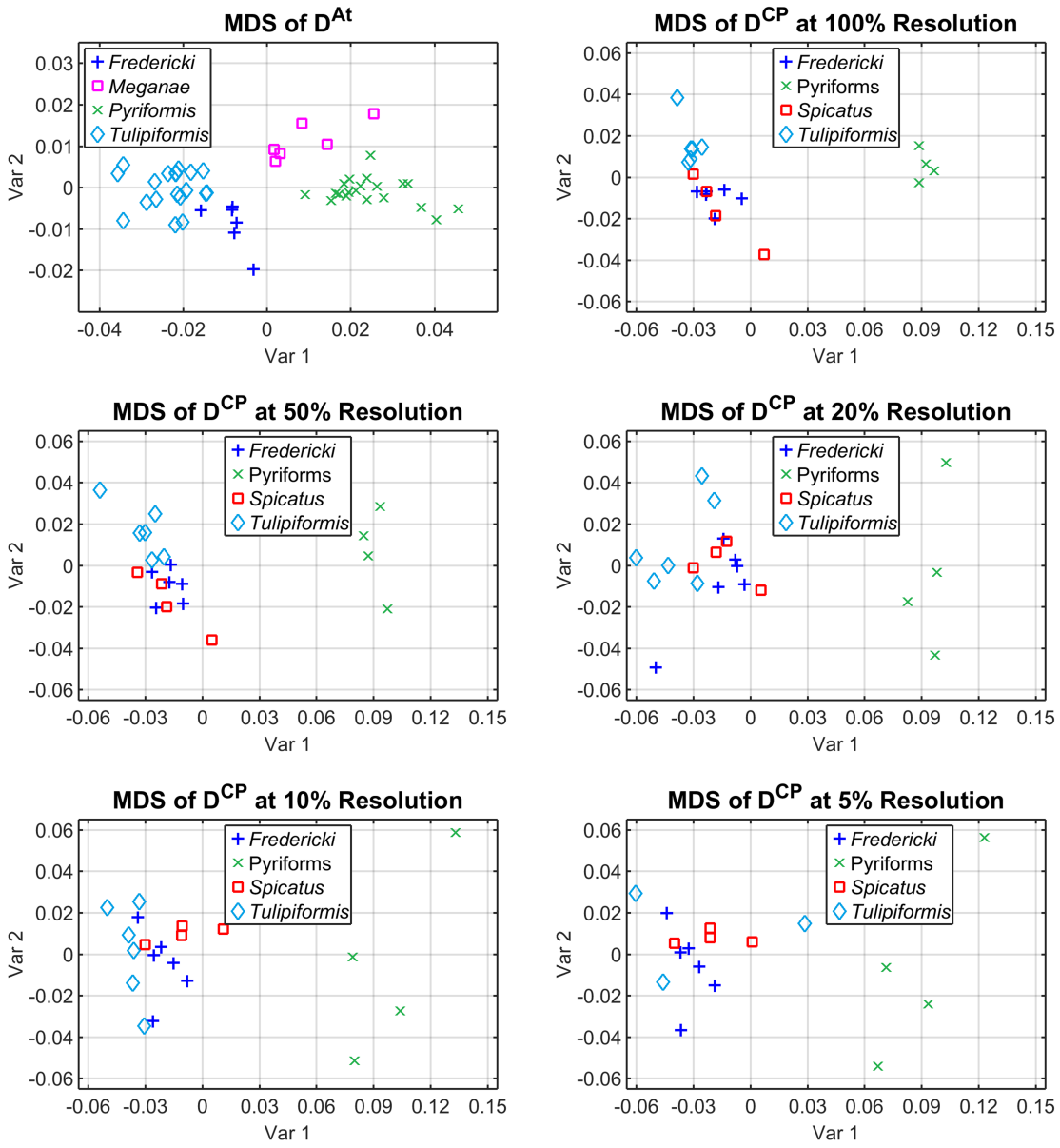

Figure 4.4: Each plot conveys a particular multidimensional scaling, where individual titles indicate which dissimilarity matrix was used as input. Similar to principle components, Var 1 represents the direction with highest variance and Var 2 represents the direction with second highest variance.

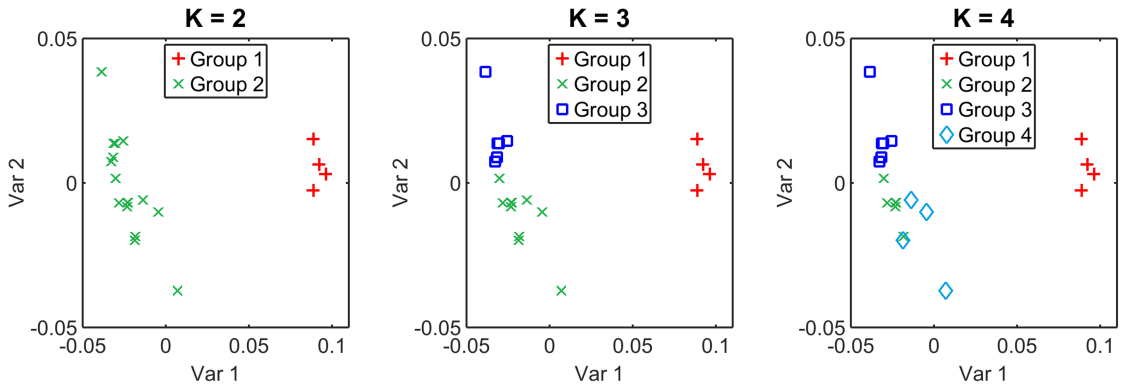

Figure 4.5: K-means was performed with the MDS embedding of 100% Resolution D^{CP} resulting in the above figures. *(Left)* $k = 2$ *(Center)* $k = 3$ and *(Right)* $k = 4$. These divisions help support the results of our aggregate clustering.

The final step in the analysis consists of deploying AC on the dissimilarity matrices D^{CP} and D^{At} at all five resolutions. Like MDS, this method also does not require the matrices to be of equal size. Dendrograms of these clusters can be seen in Fig. 4.6. We can also see the progression of variance as clusters are aggregated in Fig. 4.7. This progression of variance helps us to determine that there are three clusters present in our data. Fig. 4.5, which depicts k-means as performed on our MDS embedding, also supports our clustering choice. k-means with $k = 3$ splits our data into the same three groups as the AC does.

A visual analysis of the AC outcome of Fig. 4.6 shows that the CP-distance dissimilarity matrix manages to clearly separate *P. tulipiformis* from the cluster containing *P. fredericki* and *P. spicatus* samples. This improves the result of the comparison study by [8], in which the separation of *P. tulipiformis* and *P. fredericki* is incomplete, and a few *P. fredericki* samples are (incorrectly) categorized as *P. tulipiformis*. Furthermore, while both old and new methods separate *P. pyriformis* from the other species, CP-distance-based analysis does a better job. Indeed, it separates the data into two distinct, large clusters that correspond to the pyriform and godoniform groups; moreover, it is evident that this clustering happens at a coarser level than the separation of species. Finally, we see that as the resolution of scans decreases, the advantages of CP-distance method over the traditional techniques disappear: at 20% resolution, CP-distance no longer separates *P. tulipiformis* and *P. fredericki* better

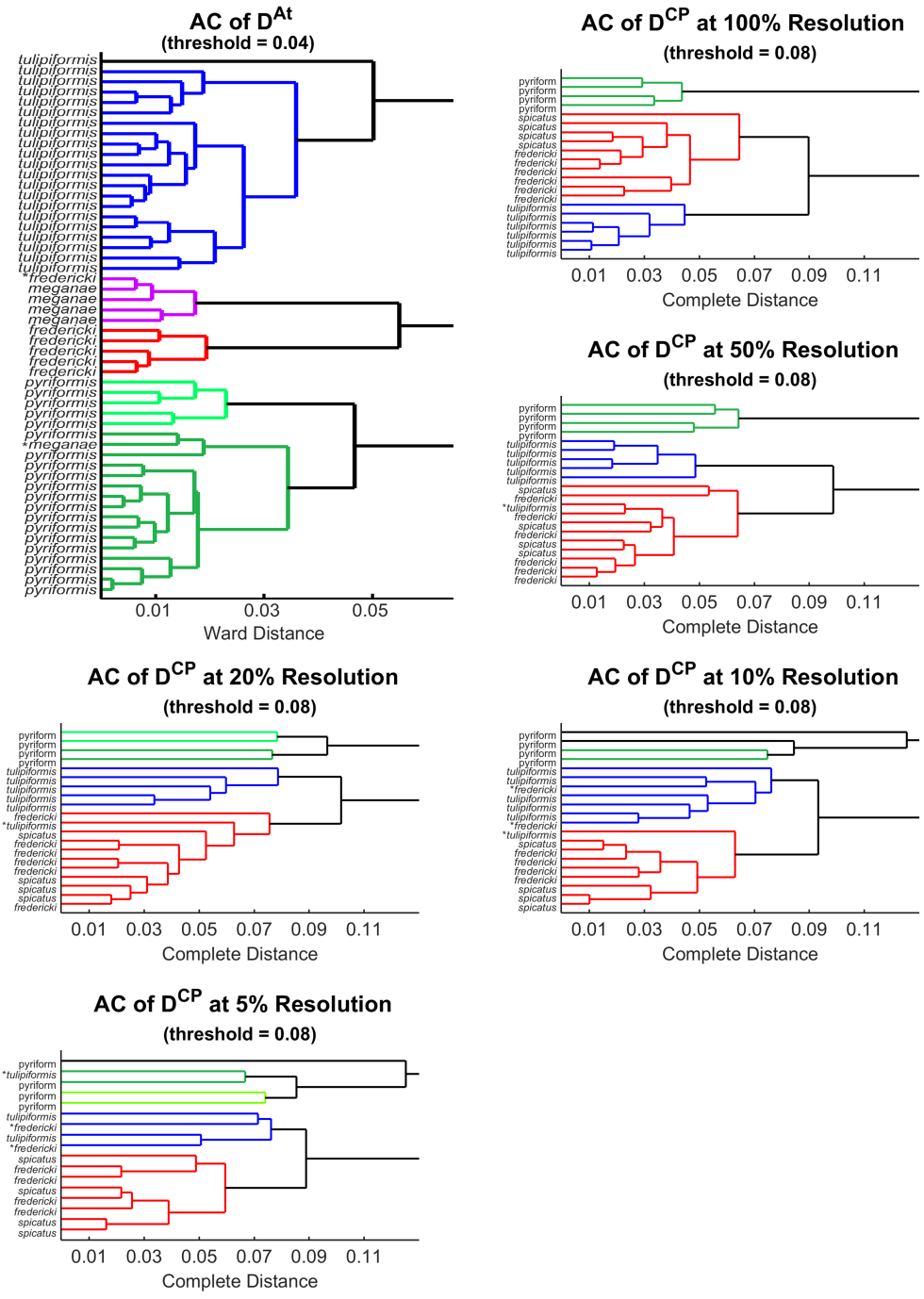

Figure 4.6: Each dendrogram depicts hierarchical aggregate clustering on a different dissimilarity matrix. All D^{CP} matrices used complete distance due to small sample numbers (so cluster shape is unclear), while D^{At} uses the Ward algorithm to compute distance between clusters. Distance thresholds were chosen so that the resulting clusters are preserved under larger perturbations. Errors in classification are marked with an asterisk. One primary shortcoming can be seen readily in these dendrograms: *Pentremites spicatus* and *P. fredericki* are not well separated.

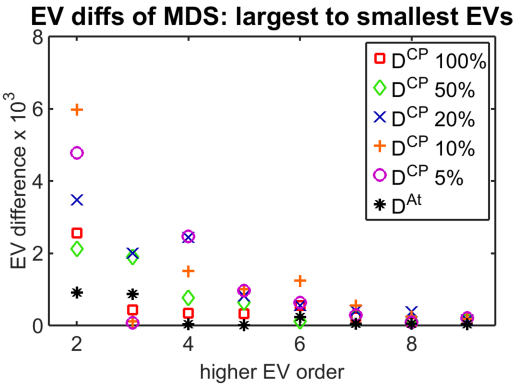

Figure 4.7: Depiction of the amount of variance needed to agglomerate each cluster (as indicated in the horizontal axis) for each dataset. We can consider this as the variance accounted for by each particular transition (e.g., from 2 to 3 clusters). The first datum is absent because it is much larger than the rest. After the transition from 2 to 3 clusters, the rest account for about the same amount of variance, so we conclude that there should be 3 clusters.

than DP-distance. However, even at 5% resolution, the larger clusters of the pyriform and godoniform groups still clearly separate, and some (coarse) information still can be obtained.

Table 4.2 depicts the results of our cross-validation procedure. For each trial, Ward AC was used to cluster 90% of our data. These clusters were used to classify the remaining 10%, and these were compared to the benchmark group.

4.4 Discussion

One of the most evident advantages of the CP-distance-based methodology over the DP-distance procedure is the speed at which samples can be analyzed. Although both analyses require specimens to be scanned, the proposed methodology nearly eliminates the tedious (and human error-prone) landmark picking phase. An evident advantage of the traditional method in the blastoids example is that it can be deployed using only a portion of the surface scan, because the full set of landmarks lies entirely on one side of the specimens due to strong (pentagonal) symmetry; an extension of this work consists in taking specimen symmetry into consideration when applying our framework. Nevertheless, this gain does not translate into an overall advantage in speed, as landmarks have to be marked under a microscope a priori, so

Table 4.2: Reliability of our methods ability to cluster and ultimately classify species of Pentremites. *(Left)* Each column depicts the results of our cross-validation for a particular resolution of data using D^{CP}. Ninety percent of our data was sampled, and there are only two remaining specimens to be classified. The percentage of trials with each number of correct classifications is listed. The bottom row depicts the overall proportion of classifications that were correct. *(Right)* Analogous table using D^{At} on five remaining samples (10% of the total), no variation in resolution. Each cross-validation experiment was done with 10,000 sampling trials.

D^{CP} Reliability					D^{At} Reliability	
	$D^{CP}100\%$	$D^{CP}50\%$	$D^{CP}20\%$	$D^{CP}10\%$	5 correct	34.39%
2 correct	61.21%	30.83%	33.28%	25.58%	4 correct	32.27%
1 correct	32.40%	49.31%	50.04%	50.99%	3 correct	20.15%
0 correct	6.39%	19.86%	16.68%	23.43%	2 correct	9.98%
Overall	77.41%	55.49%	58.30%	51.08%	1 correct	2.76%
					0 correct	0.45%
					Overall	76.84%

that they can be identified, and their position recorded a posteriori, because their position can be documented visually but is not evident in the scans [8]. Consequently, the identification of landmarks on the thecal surface requires knowledge of blastoid morphology to orient the specimens and identify the particular three plate junctions that serve as landmarks and manual dexterity to mark the landmarks properly. The process provides two points for user error to enter into the analysis–the initial marking of the landmarks and the recording of their position on the scans. In contrast, our CP-distance-based method only requires marking the position of the stem facet, which is easily seen on the scans. Therefore, our method nearly eliminates sources of human error because its continuous description of shape is independent of how the observer records the data.

The proposed CP-distance methodology also provides advantages when compared with traditional methods by taking into consideration a more complete representation of specimen morphology. This is achieved by incorporating fine-scale shape measurements such as curvature. Indeed, a natural progression is seen when going from the vault:pelvis and height:width ratios to summarize shape (encoded as a 2D rectangle) to the DP-distance landmark analysis based on 3D scans, in which 13 points represent the specimens shape. In this context, our methodology now uses the full geometric information of each specimens

shape. Advances in computer imagery allow us to deploy this sophisticated technique on very high resolution scans, which represent an objects shape accurately with thousands or even tens of thousands of points.

CP-distance-based analysis could potentially have a tremendous impact in the study of organisms that have a distinct shape but lack homologous landmarks, which makes morphometric analyses impossible. Such studies may include growth of colonial organisms or investigating ecophenotypic variation in organisms such as corals, sponges, and stromatolites.

Traditional landmark analysis has advantages in two areas over CP-analysis. First, type 1 landmarks, as used in [8], are points that have anatomical homology between specimens. Whereas calculated geometric "feature points" may change position between species, the type 1 landmarks represent identical points, allowing for different types of questions to be addressed; however, as mentioned above, landmarks could be used as the feature points for CP-distance in a future study.

Second, landmarks used for [8] require preservation of only the A ambulacrum and adjacent interambulacra for any given specimen. Specimens could be analyzed with confidence even if other portions of the theca were missing or covered with matrix. The CP-distance method, as presently used, requires complete preservation of thecal surfaces preserved three dimensionally and without significant adhering matrix. Our forthcoming work will address both of these issues.

4.5 Conclusion

We compare a novel computational geometry methodology based on continuous Procrustes distance with the standard method of discrete Procrustes distance for tackling the problem of species discrimination on Pentremites. Our results show that the CP-distance methodology is not only (slightly) superior at resolving the issue but also significantly reduces processing time and vulnerability to human error.

Specifically, advantages appear in the separation of *P. tulipiformis* species from the cluster containing the *P. fredericki* and *P. spicatus* species. In contrast, [8] clusters a few *P. fredericki* as *P. tulipiformis*, and the respective clusters are visually mixed. We attribute this

advantage to the CP-distance calculation taking into consideration curvature measurements, which reflect the concavity of the specimens plates. This differs significantly from the output that can be obtained using DP-distance alone; in that methodology, concavity is measured only as a subtle change in the position of a couple of landmarks. This phenomenon is vastly amplified when large regions of the sample are concave and eventually adds up to a large difference in the resulting CP-distance. As a consequence, we are effectively able to separate species that are concave from those that are not.

In addition, we determined the sensitivity of the CP-distance methodology to varying resolutions of scans. We found a threshold of approximately 1000 points per scan, under which the reliability of the method begins to wither. Nevertheless, we showed that even at the lowest resolution levels there is information to be gained. Because of the relatively low cost and short time involved in performing low-resolution scans, our technique opens the door for future researchers in the area to produce several low-cost, preliminary studies; this would allow them to choose where to invest their in-depth efforts more effectively.

Bibliography

[1] (2010). Ethical principles to be considered in the allocation of human organs. Organ Procurement and Transplantation Network. 53

[2] (2014). Spring 2014 regional meeting data. United Network for Organ Sharing. 53

[3] (2015). Tn and us kidney transplant summary. Scientific Registry of Transplant Recipients. viii, 65, 66, 74

[4] Adcock, A., Carlsson, E., and Carlsson, G. (2016). The ring of algebraic functions on persistence bar codes. *Homology, Homotopy and Applications*, 18(1):381–402. 3, 7

[5] Adler, R. J., Bobrowski, O., and Weinberger, S. (2014). Crackle: The homology of noise. *Discrete & Computational Geometry*, 52(4):680–704. 8, 40

[6] Al-Aifari, R., Daubechies, I., and Lipman, Y. (2013). Continuous procrustes distance between two surfaces. *Communications on Pure and Applied Mathematics*, 66(6):934–964. 2, 76, 80

[7] Ashlagi, I., Gamarnik, D., Rees, M. A., and Roth, A. E. (2012). The need for (long) chains in kidney exchange. Technical report, National Bureau of Economic Research. 55

[8] Atwood, J. W. and Sumrall, C. D. (2012). Morphometric investigation of the pentremites fauna from the glen dean formation, kentucky. *Journal of Paleontology*, 86(5):813–828. xiii, 76, 77, 78, 79, 80, 81, 82, 85, 88, 89

[9] Bauer, U. (2015). Ripser. https://github.com/Ripser/ripser. 7

[10] Belkin, M. and Niyogi, P. (2001). Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, volume 14, pages 585–591. 3

[11] Bendich, P., Marron, J. S., Miller, E., Pieloch, A., and Skwerer, S. (2016). Persistent homology analysis of brain artery trees. *The Annals of Applied Statistics*, 10(1):198–218. 3, 7, 9

[12] Binchi, J., Merelli, E., Rucco, M., Petri, G., and Vaccarino, F. (2014). jholes: A tool for understanding biological complex networks via clique weight rank persistent homology. *Electronic Notes in Theoretical Computer Science*, 306:5–18. 3

[13] Bobrowski, O., Mukherjee, S., and Taylor, J. E. (2014). Topological consistency via kernel estimation. *arXiv:1407.5272*. 7, 8

[14] Bookstein, F. L., Gunz, P., Mitterœcker, P., Prossinger, H., Schæfer, K., and Seidler, H. (2003). Cranial integration in homo: singular warps analysis of the midsagittal plane in ontogeny and evolution. *Journal of Human Evolution*, 44(2):167–187. 76

[15] Budd, A. F., Johnson, K. G., and Potts, D. C. (1994). Recognizing morphospecies in colonial reef corals: I. landmark-based methods. *Paleobiology*, 20(04):484–505. 76

[16] Carlsson, G. and De Silva, V. (2010). Zigzag persistence. *Foundations of computational mathematics*, 10(4):367–405. 3

[17] Carlsson, G., De Silva, V., and Morozov, D. (2009). Zigzag persistent homology and real-valued functions. In *Proceedings of the twenty-fifth annual symposium on Computational geometry*, pages 247–256. ACM. 3

[18] Cheeger, J. (1969). Pinching theorems for a certain class of riemannian manifolds. *American Journal of Mathematics*, 91(3):807–834. 1

[19] Chen, C. and Kerber, M. (2011). Persistent homology computation with a twist. In *Proceedings 27th European Workshop on Computational Geometry*, volume 11. 7

[20] Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. (2007). Stability of persistence diagrams. *Discrete & Computational Geometry*, 37:103–120. 13

[21] Cohen-Steiner, D., Edelsbrunner, H., Harer, J., and Mileyko, Y. (2010). Lipschitz functions have l p-stable persistence. *Foundations of computational mathematics*, 10(2):127–139. 7, 14

[22] Cox, T. F. and Cox, M. A. (2000). *Multidimensional scaling*. CRC press. 81

[23] Danovitch, G. M. (2014). The high cost of organ transplant commercialism. *Kidney international*, 85(2):248–250. 53

[24] De Silva, V. and Ghrist, R. (2006). Coordinate-free coverage in sensor networks with controlled boundaries via homology. *The International Journal of Robotics Research*, 25(12):1205–1222. 3

[25] De Silva, V. and Ghrist, R. (2007a). Coverage in sensor networks via persistent homology. *Algebraic & Geometric Topology*, 7(1):339–358. 3, 7

[26] De Silva, V. and Ghrist, R. (2007b). Homological sensor networks. *Notices of the American mathematical society*, 54(1). 3

[27] Dickerson, J. P., Procaccia, A. D., and Sandholm, T. (2014). Price of fairness in kidney exchange. In *Proceedings of the 2014 International Conference on Automous Agents and Multiagent Systems*, pages 1013–1020. International Foundation for Autonomous Agents and Multiagent Systems. 54, 66

[28] Dłotko, P., Ghrist, R., Juda, M., and Mrozek, M. (2012). Distributed computation of coverage in sensor networks by homological methods. *Applicable Algebra in Engineering, Communication and Computing*, 23(1-2):29–58. 3

[29] Edelsbrunner, H. and Harer, J. (2010). *Computational topology: an introduction*. American Mathematical Society. 2, 7, 9, 12, 13, 49, 57

[30] Edelsbrunner, H., Letscher, D., and Zomorodian, A. (2002). Topological persistence and simplification. *Discrete & Computational Geometry*, 28(4):511–533. 13

[31] Edmonds, J. (1965). Paths, trees, and flowers. *Canadian Journal of Mathematics*, 17(3):449–467. 55

[32] Emmett, K., Rosenbloom, D., Camara, P., and Rabadan, R. (2014). Parametric inference using persistence diagrams: A case study in population genetics. *arXiv:1406.4582*. 7

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

[34] Fasy, B. T., Kim, J., Lecci, F., Maria, C., Rouvreau., V., The included GUDHI is authored by Clement Maria, Dionysus by Dmitriy Morozov, P. b. U. B. M. K., and Reininghaus., J. (2015). Tda: Statistical tools for topological data analysis r package version 1.4.1. 7

[35] Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. (2014). Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6):2301–2339. 7, 8

[36] Ferreira, L. and Hitchcock, D. B. (2009). A comparison of hierarchical methods for clustering functional data. *Communications in Statistics-Simulation and Computation*, 38(9):1925–1949. 82

[37] Foote, M. (1991). *Morphological and taxonomic diversity in a clade's history: the blastoid record and stochastic simulations*, volume 28. Museum of Paleontology, University of Michigan Ann Arbor. 78

[38] Frost, S. R., Marcus, L. F., Bookstein, F. L., Reddy, D. P., and Delson, E. (2003). Cranial allometry, phylogeography, and systematics of large-bodied papionins (primates: Cercopithecinae) inferred from geometric morphometric analysis of landmark data. *The Anatomical Record*, 275(2):1048–1072. 76

[39] Galloway, J. J. and Kaska, H. Y. (1957). Genus pentremites and its species. *Geological Society of America Memoirs*, 69:1–114. 78

[40] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA. 51

[41] Goodman, I. R., Mahler, R. P., and Nguyen, H. T. (2013). *Mathematics of data fusion*, volume 37. Springer Science & Business Media. 16

[42] Guillemard, M. and Iske, A. (2011). Signal filtering and persistent homology: an illustrative example. *Proc. Sampling Theory and Applications (SampTA11)*. 3, 7

[43] Hatcher, A. (2002). Algebraic topology. 2002. *Cambridge UP, Cambridge*, 606(9). 13

[44] Hodge, W. V. D. (1989). *The theory and applications of harmonic integrals*. CUP Archive. 59

[45] Hua, X., Nguyen, D., Raghavan, B., Arsuaga, J., and Vazquez, M. (2007). Random state transitions of knots: a first step towards modeling unknotting by type ii topoisomerases. *Topology and its Applications*, 154(7):1381–1397. 2

[46] Huisken, G. (1990). Asymptotic-behavior for singularities of the mean-curvature flow. *Journal of Differential Geometry*, 31(1):285–299. 1

[47] Jiang, X., Lim, L.-H., Yao, Y., and Ye, Y. (2011). Statistical ranking and combinatorial hodge theory. *Mathematical Programming*, 127(1):203–244. 2, 57, 58

[48] Kauffman, L. H. (1988). Statistical mechanics and the jones polynomial. *New problems, methods and techniques in quantum field theory and statistical mechanics*, pages 175–222. 2

[49] Kerber, M., Morozov, D., and Nigmetov, A. (2016). Geometry helps to compare persistence diagrams. *Proceedings of the Eighteenth Workshop on Algorithm Engineering and Experiments*, pages 103–112. 7

[50] Kessler, J. B. and Roth, A. E. (2012). Organ allocation policy and the decision to donate. *The American Economic Review*, 102(5):2018–2047. 53

[51] Ladner, D. P. and Mehrotra, S. (2016). Methodological challenges in solving geographic disparity in liver allocation. *JAMA surgery*, 151(2):109–110. 74

[52] Lieberman, D. E., Carlo, J., de León, M. P., and Zollikofer, C. P. (2007). A geometric morphometric analysis of heterochrony in the cranium of chimpanzees and bonobos. *Journal of Human Evolution*, 52(6):647–662. 76

[53] Macurda Jr, D. B. (1966). The ontogeny of the mississippian blastoid orophocrinus. *Journal of Paleontology*, pages 92–124. 78

[54] Maderbacher, M., Bauer, C., Herler, J., Postl, L., Makasa, L., and Sturmbauer, C. (2008). Assessment of traditional versus geometric morphometrics for discriminating

populations of the tropheus moorii species complex (teleostei: Cichlidae), a lake tanganyika model for allopatric speciation. *Journal of Zoological Systematics and Evolutionary Research*, 46(2):153–161. 76

[55] Mahler, R. P. (1995). Unified nonparametric data fusion. In *SPIE's 1995 Symposium on OE/Aerospace Sensing and Dual Use Photonics*, pages 66–74. International Society for Optics and Photonics. 19

[56] Marchese, A. and Maroulas, V. (2016). Topological learning for acoustic signal identification. In *2016 19th International Conference on Information Fusion (FUSION)*, pages 1377–1381. 3, 7

[57] Marchese, A. and Maroulas, V. (2017). A point process distance on the space of persistence diagrams. *In Submission*. 3, 7

[58] Matheron, G. (1975). *Random Sets and Integral Geometry*. John Wiley & Sons. 9, 17, 18, 20

[59] Mike, J. and Maroulas, V. (2017). Combinatorial hodge theory for equitable kidney paired donation. *In Submission*. 3

[60] Mike, J., Sumrall, C. D., Maroulas, V., and Schwartz, F. (2016). Nonlandmark classification in paleobiology: computational geometry as a tool for species discrimination. *Paleobiology*, pages 1–11. 3

[61] Mileyko, Y., Mukherjee, S., and Harer, J. (2011). Probability measures on the space of persistence diagrams. *Inverse Problems*, 27(12). 7, 8, 15

[62] Mitteroecker, P., Gunz, P., and Bookstein, F. L. (2005). Heterochrony and geometric morphometrics: a comparison of cranial growth in pan paniscus versus pan troglodytes. *Evolution & development*, 7(3):244–258. 76

[63] Mitteroecker, P., Gunz, P., Weber, G., and Bookstein, F. L. (2004). Regional dissociated heterochrony in multivariate analysis. *Annals of Anatomy-Anatomischer Anzeiger*, 186(5-6):463–470. 76

[64] Nicolau, M., Levine, A., and Carlsson, G. (2011). 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. 3

[65] Perea, J. A. and Harer, J. (2015). Sliding windows and persistence: An application of topological methods to signal analysis. *Foundations of Computational Mathematics*, 15(3):799–838. 3, 7, 47

[66] Pereira, C. M. and de Mello, R. F. (2015). Persistent homology for time series and spatial data clustering. *Expert Systems with Applications*, 42(15):6026–6038. 3, 7

[67] Piras, P., Marcolini, F., Raia, P., Curcio, M., and Kotsakis, T. (2010). Ecophenotypic variation and phylogenetic inheritance in first lower molar shape of extant italian populations of microtus (terricola) savii (rodentia). *Biological Journal of the Linnean Society*, 99(3):632–647. 76

[68] Reyment, R., Bookstein, F., McKenzie, K., and Majoran, S. (1988). Ecophenotypic variation in mutilus pumilus (ostracoda) from australia, studied by canonical variate analysis and tensor biometrics. *Journal of Micropalaeontology*, 7(1):11–20. 76

[69] Rohlf, F. J. (1998). On applications of geometric morphometrics to studies of ontogeny and phylogeny. *Systematic Biology*, 47(1):147–158. 76

[70] Ross, L. F. and Woodle, E. (1998). Kidney exchange programs: An expanded view of the ethical issues. In *Organ Allocation*, pages 285–295. Springer. 53

[71] Roth, A. E., Sönmez, T., and Uktu, Ü. M. (2003). Kidney exchange. Technical report, National Bureau of Economic Research. 55, 66

[72] Rouse, D., Watkins, A., Porter, D., Harer, J., Bendich, P., Strawn, N., Munch, E., DeSena, J., Clarke, J., and Gilbert, J. (2015). Feature-aided multiple hypothesis tracking using topological and statistical behavior classifiers. In *SPIE Defense+ Security*, pages 94740L–94740L. International Society for Optics and Photonics. 3

[73] Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326. 3

[74] Saidman, S. L., Roth, A. E., Sönmez, T., Ünver, M. U., and Delmonico, F. L. (2006). Increasing the opportunity of live kidney donation by matching for two-and three-way exchanges. *Transplantation*, 81(5):773–782. 55

[75] Scott, D. W. (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons. 8, 29, 35, 36, 49

[76] Segev, D. L., Gentry, S. E., Warren, D. S., Reeb, B., and Montgomery, R. A. (2005). Kidney paired donation and optimizing the use of live donor organs. *Journal of the American Medical Association*, 293(15):1883–1890. 55

[77] Seversky, L. M., Davis, S., and Berger, M. (2016). On time-series topological data analysis: New data and opportunities. *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 59–67. 3, 7

[78] Sgouralis, I., Nebenführ, A., and Maroulas, V. (To Appear 2017). A Bayesian topological framework for the identification and reconstruction of subcellular motion. *SIAM Journal on Imaging Sciences*. 7

[79] Sheets, H. D., Kim, K., and Mitchell, C. E. (2004). A combined landmark and outline-based approach to ontogenetic shape change in the ordovician trilobite triarthrus becki. In *Morphometrics*, pages 67–82. Springer. 76

[80] Singh, G., Mémoli, F., and Carlsson, G. E. (2007). Topological methods for the analysis of high dimensional data sets and 3d object recognition. In *SPBG*, pages 91–100. 2

[81] Sokal, R. R. and Rohlf, F. J. (2012). *Biometry: The principles and practice of statistics in biological research*. W. H. Freeman and Co.: New York. 80

[82] Spivak, M. (1981). *Comprehensive Introduction to Differential Geometry Volume, IV[A]*. Publish or Perish, Inc., University of Tokyo Press. 1

[83] Taber, D. J., Gebregziabher, M., Hunt, K. J., Srinivas, T., Chavin, K. D., Baliga, P. K., and Egede, L. E. (2016). Twenty years of evolving trends in racial disparities for adult kidney transplant recipients. *Kidney International*. 74

[84] Takemoto, S., Port, F. K., Claas, F. H., and Duquesnoy, R. J. (2004). HLA matching for kidney transplantation. *Human Immunology*, 65(12):1489–1505. 53

[85] Takens, F. (1980). Detecting strange attractors in turbulence. *Dynamical Systems and Turbulence, Warwick 1980, Lecture Notes in Mathematics*, 898:366–381. 3

[86] Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to data mining. 1st.* Boston: Pearson Addison Wesley. xxi. 82

[87] Tenenbaum, J. B., De Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323. 2, 3

[88] Toulis, P. and Parkes, D. C. (2011). A random graph model of kidney exchanges: efficiency, individual-rationality and incentives. In *Proceedings of the 12th ACM conference on Electric commerce*, pages 323–332. ACM. 55, 66

[89] Turner, K., Mileyko, Y., Mukherjee, S., and Harer, J. (2014). Fréchet means for distribution of persistence diagrams. *Discrete & Computational Geometry*, 52:44–70. 7

[90] Venkataraman, V., Ramamurthy, K. N., and Turaga, P. (2016). Persistent homology of attractors for action recognition. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 4150–4154. 3, 7

[91] Villemant, C., Simbolotti, G., and Kenis, M. (2007). Discrimination of eubazus (hymenoptera, braconidae) sibling species using geometric morphometrics analysis of wing venation. *Systematic Entomology*, 32(4):625–634. 76

[92] Waters, J. A., Horowitz, A. S., and Macurda Jr, D. B. (1985). Ontogeny and phylogeny of the carboniferous blastoid pentremites. *Journal of Paleontology*, pages 701–712. 77, 78

[93] Wilk, J. and Bieler, R. (2009). Ecophenotypic variation in the flat tree oyster, isognomon alatus (bivalvia: Isognomonidae), across a tidal microhabitat gradient. *Marine Biology Research*, 5(2):155–163. 76

[94] Xia, K., Feng, X., Tong, Y., and Wei, G. W. (2015). Persistent homology for the quantitative prediction of fullerene stability. *Journal of computational chemistry*, 36(6):408–422. 3, 7

[95] Zelditch, M. L., Swiderski, D. L., and Sheets, H. D. (2012). *Geometric morphometrics for biologists: a primer*. Academic Press. 76

[96] Zenios, S. A. (2002). Optimal control of a paired-kidney exchange program. *Management Science*, 48(3):328–342. 74

[97] Zollikofer, C. P. E. and Ponce De León, M. S. (2004). Kinematics of cranial ontogeny: heterotopy, heterochrony, and geometric morphometric analysis of growth models. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 302(3):322–340. 76

Vita

Joshua Mike was born in Youngstown, OH on May 6th, 1989. His interest in mathematics was fostered by his Mamaw and mother from an early age. He attended Ursuline High School where he gained further interest in both Mathematics and Science. His high school math teacher Russel Nalepa introduced Joshua and fellow students to proofs and higher mathematics, giving the subject a new light.

Joshua Mike completed his undergraduate education at Youngstown State University in northeast Ohio. There he obtained Bachelor's degrees in both mathematics and chemistry in 2011, hoping to study mathematics for the sake of data analysis. Besides his coursework, Joshua tutored undergraduate mathematics and performed undergraduate research involving biophysics ODE models.

Immediately after graduation, Joshua began to pursue his mathematics PhD at the University of Tennessee while working as a mathematics teaching assistant. Joshua started studying analysis and geometry, passing his preliminary exams in Analysis and Partial Differential Equations. His oral exam was presented under the direction of Dr. Fernando Schwartz and involved continuous Procrustes distance. After his oral exam, Joshua began to work under the direction of Prof. Vasileios Maroulas, now engaging statistics with geometry and leading to the results detailed in his dissertation.