256 Shades of Grey: Topological Analysis of the Gray-Scott Model Using Persistence Landscapes

---

A Thesis
Presented to
The Established Interdisciplinary Committee for Mathematics and Physics
Reed College

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Arts

Joseph Joe

May 2016

Approved for the Committee
(Mathematics and Physics)

Daniel Borrero

Kyle Ormsby

Acknowledgements

I have really enjoyed my time at Reed and am glad that I had the opportunity to attend this school. I'd like to thank my parents who have made it possible. Next, I'd like to thank my officemate Aaron, who has been a source of encouragement and delightful to be around. Also, to the Reed Basketball Team and to our coach Ried. We've come a long way from a winless team to a team that can at least hold the opponent under a hundred. Finally, to my thesis advisors Kyle and Daniel. Thank you for being patient and helpful throughout the thesis process.

Table of Contents

List of Figures

Abstract

The Gray-Scott reaction-diffusion model is analyzed using topological data. Various patterns are simulated, and persistent homology is used to distinguish patterns. The persistence landscape is shown to be sensitive to small changes in reaction parameters and robust in classifying the patterns.

Introduction

Nonlinear phenomena are ubiquitous in nature. For instance, air resistance is a force that often acts proportionally to the square of a body's velocity. So, air resistance increases quadratically rather than linearly with velocity. Another nonlinear phenomenon is weather, which is highly variable, changing drastically due to differences in air pressure, temperature, and moisture between one place and another. The complex and pervasive nature of nonlinearity makes it a fascinating area of study.

While linear systems, described by equations like

$$\frac{d^n y}{dt^n} + A_1(t)\frac{d^{n-1}y}{dt^{n-1}} + \cdots + A_{n-1}(t)\frac{dy}{dt} + A_n(t)y = f(t), \tag{1}$$

have solutions that can be added together to form new solutions, nonlinear systems do not have this property and the whole is not the sum of its parts, making nonlinear systems notoriously difficult to solve.

One of the most famous nonlinear systems is the Navier-Stokes equation, which describes the motion of a fluid in two or three dimensions. Assuming incompressible and homogeneous fluid flow, the equation reads

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} = -\frac{\nabla p}{\rho} + \nu \nabla^2 \mathbf{u} + \mathbf{f}, \tag{2}$$

where $\mathbf{u}$ is the velocity of the fluid, ν is the viscosity, p is the pressure of the fluid, and $\mathbf{f}$ is the external volumetric force (e.g., gravity).

The Navier-Stokes equation is used to model water flow in pipes, weather, and air flow around airplane wings. Although approximate solutions can be obtained with computational methods, the existence and smoothness of an exact solution in 3 dimensions has been an open problem in mathematics and physics since the 19th century [15].

In this thesis, we study reaction-diffusion (RD) systems, which are nonlinear systems that model how chemical concentrations change in time and space due to chemical reactions and spatial diffusion. We restrict our attention to the Gray-Scott (GS) model [2], which will be discussed in Chapter 1. RD systems are governed by partial differential equations, which generally take the form

$$\frac{\partial \mathbf{u}}{\partial t} = \mathbf{D}\nabla^2 \mathbf{u} + r(\mathbf{u}), \tag{3}$$

Figure 1: The Belousov Zhabotinsky reaction is performed in a Petri dish. Metal ions (cerium in this case) catalyze oxidation of the reductants by bromic acid in acidic water solution. The waves of oxidation are blue-gray [1].

where $\mathbf{u}$ are generic chemical species, $\mathbf{D}$ is a diagonal matrix of diffusion coefficients, and $r(\mathbf{u})$ is a general reaction term.

RD systems are particularly interesting because solutions often display complex patterns that do not resemble the initial state they started from. The fact that nonlinear systems can develop into these stable patterns is surprising and also requires new techniques in analyzing nonlinearity, and pattern formation. Alan Turing [16] proposed RD systems as the mechanism in chemical morphogenesis, the process by which chemical patterns form, because patterns in RD systems resemble the spots and stripes found on animals. Experiments have verified that Turing's patterns exist [17], but there is no concrete evidence that patterns on animals form via reaction-diffusion.

The analysis performed in this thesis is conducted on image data, snapshots of the reaction-diffusion system in time. As an example, Fig. 1 displays the Belousov-Zhabotinsky reaction. By oxidation of a metal catalyst, the solution changes in color. The spontaneous growth of the spiraling patterns is suprising, and was initially thought to violate the Second Law of Thermodynamics. However, the system is far from equilibrium, nonlinear effects are not negligable and normal thermodynamic rules do not apply without modification. The work of Field, Koros, and Noyes [18] along with theoretical advances in non-equilibrium thermodynamics provided the mechanism for the reaction [19].

We are particularly interested in the time evolution of these images, which represent how these patterns form from a bland initial state. Looking at the pictures, the goal is to calculate quantities that can differentiate or identity similarities between the patterns.

Common image analysis techniques include image segmentation (partitioning an image into segments that are easier to analyze), single particle tracking, image en-

hancement (to remove noise), and the discrete Fourier transform (frequency-domain representation of the image). However these techniques are largely ineffective in our scenario because we are interested in 3 dimensional structures from our images.

Therefore, we use a new framework in our analysis. We use topological information that is robust under noise and captures the essential features of the reaction images. In the first chapter, we discuss the specific reaction diffusion system under consideration, the Gray-Scott model. The second chapter deals with homology theory, which condenses our images into algebraic objects that can be compared quantitatively. The third chapter outlines how we use homology to analyze the images along with some other statistical methods. The fourth chapter explains our methodology and results.

Chapter 1

The Gray-Scott Model

In his study of morphogenesis, Turing demonstrated that an initially uniform state can evolve into a spatially heterogeneous pattern in response to small perturbations [16]. His model consisted of two chemicals, an activator and an inhibitor. The activator stimulates the production of itself and the other chemical. The inhibitor depletes the formation of the activator. He showed that if the diffusion of the inhibitor was greater than that of the activator, then diffusion would cause instability, and the chemical concentrations would form patterns.

This result is strange, considering that diffusion usually makes systems homogeneous. When sugar is placed in coffee, the sugar molecules diffuse over time until they are well mixed in with the rest of the liquid. Stirring speeds up the process, but the molecules always diffuse down the concentration gradient, toward areas of lower concentration.

To gain an intuitive understanding, we consider the chemical reaction $A+B \rightarrow 2B$. If A is spread throughout, but there is none of B, of course there will not be a reaction. But, if we place B at various sites and only allow A to diffuse, then the reaction occurs at the sites of B. At these points, the concentration of B builds up. Eventually, A would disappear and we would be left with spots of B. If, somehow A were supplied throughout and B removed from the system, the reaction could possibly achieve a balance: the supplying and diffusion of A, which creates B through the reaction, counteracts the removal of B. Therefore, a steady, long-lived pattern is obtained, with high concentrations of A between the spots of high B. Turing predicted that this structure develops spontaneously from an almost homogeneous distribution of A and B, provided that A diffuses more rapidly than B [20].

Although the Gray-Scott (GS) model [2] is not an activator-inhibitor system, it also displays spontaneous pattern formation through diffusion. The Gray-Scott model as a RD system is simpler, given that only one of the chemicals is active. The two

reactions it describes are

$$\begin{aligned} U + 2V &\to 3V \\ V &\to P \end{aligned} \tag{1.1}$$

where U, V and P are generic chemicals. Specifically, we see that V is both a reactant and a product, as it appears in both sides of the reaction in Eq. 1.1. P is an inert product and does not participate in the reaction after formation. The patterns that develop in the GS model show great variety; they can be time-dependent or time-independent, and differ on how long they take to stabilize.

Equation 1.1 can be translated into the following equations:

$$\begin{aligned} \frac{\partial u}{\partial t} &= d_u \nabla^2 u - uv^2 + F(1-u) \\ \frac{\partial v}{\partial t} &= d_v \nabla^2 v + uv^2 - (F+k)v \end{aligned} \tag{1.2}$$

The concentrations of U and V are u and v respectively and they are functions of time t, and of the spatial variables $(x, y) \in [0, 1] \times [0, 1] \subset \mathbb{R}^2$. We choose a square domain for our simulations of these reactions and enforce periodic boundary conditions so that the chemicals do not travel outside the square.

The terms $\pm uv^2, F(1-u)$, and $(F+k)v$ in Eq. 1.2 are part of the general reaction terms in Eq. 3. Specifically, the law of mass action [21] states that the rate of a chemical reaction is proportional to the product of each chemical concentration raised to the reaction coefficient, explaining the terms $\pm uv^2$.

The remaining terms in Eq. 1.2 and their interpretation will now be explained.

Diffusion Coefficients d_u, d_v

These are called the diffusion coefficients and they multiply the Laplacian operator ∇^2, which causes the chemicals to "diffuse" or spread out. ∇^2 captures the fact that chemical concentrations (or heat, smoke, etc) move from areas of high concentration to areas of low concentration.

Qualitatively, $\nabla^2 u(x, y)$ is positive if the surrounding regions of (x, y) have a greater value of u, causing u to diffuse into (x, y) from the surroundings. This is reflected by the fact that the term adds a positive contribution to $\frac{\partial u}{\partial t}$ to increase $u(x, t)$. Similarly, $\nabla^2 u(x, y)$ is negative if the surrounding regions of (x, y) have a lesser value of u, adding a negative contribution to $\frac{\partial u}{\partial t}$ to decrease $u(x, t)$. As the constants d_u and d_v multiply the Laplacian, they specify the rate of diffusion.

Replenishment and Diminishment F, k

$F(1-u)$ is the *replenishment term* and F is called the *feed rate*, representing the rate of replenishment. The reaction stops when all of the chemical U has been used up in the system. Therefore, to perpetually fuel the reaction, a positive contribution to $\frac{\partial u}{\partial t}$ is added, proportional to $1-u$.

This models a membrane surrounding the region where the reaction is taking place. Outside the membrane, there is a large supply of U and it diffuses through the membrane into the region. The membrane has a permeability F and the rate of diffusion is greater when the region has a low concentration of U. The region also diffuses the chemical V out of the region through the membrane proportional to the amount of V present in the region, hence contributing the term $-Fv$ to $\frac{\partial v}{\partial t}$ in Eq. 1.2

The parameter k specifies the rate at which V is converted to P (the second reaction in Equation 1.1), adding the term $-kv$ to Eq. 1.2. The term $-(F+k)v$ is called the *diminishment term* because it removes V from the system, so that V does not increase without bound.

1.1 Characterization of the Gray-Scott Patterns

Given the simplicity of the GS model, the variety of patterns that emerge is fascinating. In a paper categorizing the emergent patterns, Pearson [2] identified 12 different non-trivial patterns through computer simulation and nonlinear analysis with $d_u = 2d_v$. According to Pearson, this ratio of diffusion constants show spontaneous pattern formation for a wide range of parameter values and no patterns occur when the ratio is 1. Fig. 1.1 shows maps of the concentration of u for varying values of the parameters F and k, where red denotes $u = 1$ and blue denotes $u \approx 0.2$. The Greek letters identifying the 12 patterns are also shown in Fig. 1.2, in the $F-k$ parameter space. These trivial states undergo instabilities on the bifurcation lines displayed in Fig. 1.2. These are derived in [22] and [23].

Pearson's simulations were conducted on a 256×256 pixel computational domain. There are two trivial states, which are spatially uniform in color, all red or blue. The red state R has $U = 1$, $V = 0$ and the blue state B has roughly $U = 0.3$, $V = 0.25$ (the exact values depends on parameters F, k [22]).

1.2 Numerical simulation

Our simulations follow Pearson's procedure [2] and use the code written by Hawkins [24]. Equations 1.2 are solved numerically by forward Euler integration of the discrete Laplacian.

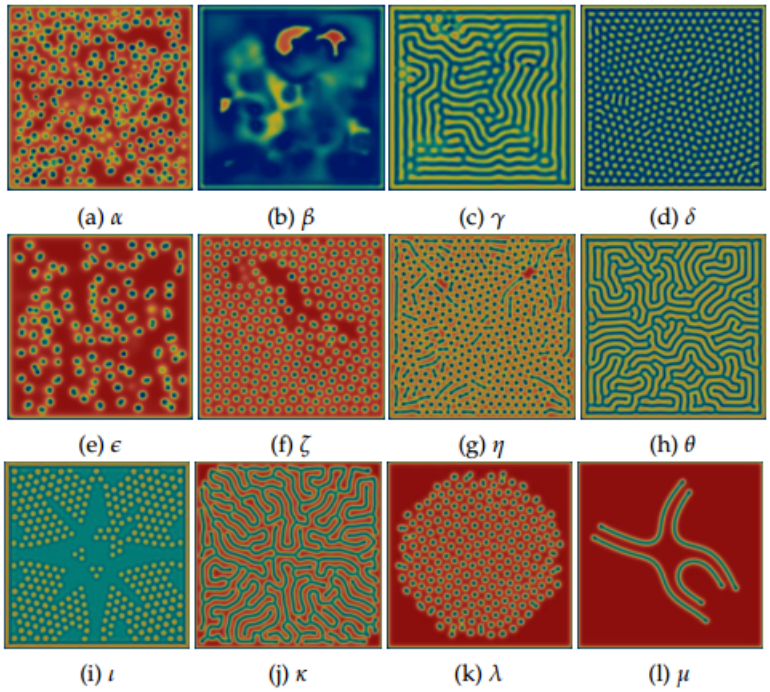

Figure 1.1: Adapted from [2]. Pearson's categorization of the patterns in the Gray-Scott model. The concentration of the chemical U is displayed on the 256×256 pixel grid. Each pattern is designated by a Greek letter which corresponds to the plot in Fig. 1.2. Red and blue indicate $u = 1$ and $u \approx 0.2$ respectively.

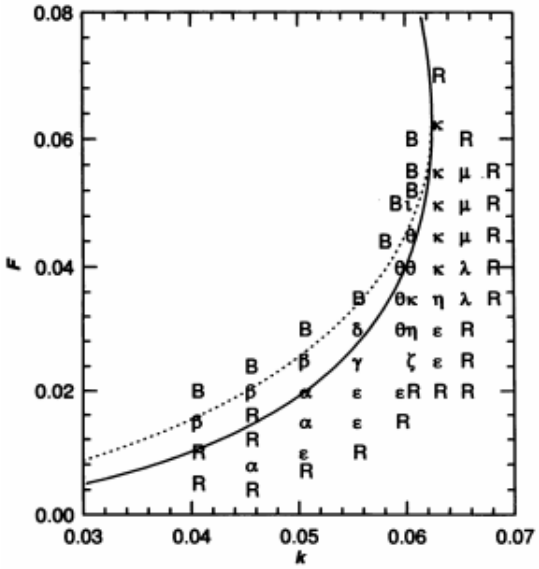

Figure 1.2: Adapted from [2]. The plot displays the parameter space of the Gray-Scott patterns. The Greek letters indicate the location in parameter space where the patterns in Fig. 1.1 were found; B and R indicate that the system evolved to uniform blue and red states, respectively.

Using the finite difference method, the Laplacian is (for fixed t and grid size 1 in both x and y dimensions) approximately

$$\nabla^2 u(x,y) \approx u(x-1,y) + u(x+1,y) + u(x,y-1) + u(x,y+1) - 4u(x,y)$$

Moreover, using the Euler method [25] we have for fixed (x,y)

$$\begin{aligned} u(t+1) &\approx u(t) + \nabla^2 u(t) - u(t)v(t)^2 + F(1-u(t)) \\ v(t+1) &\approx v(t) + \nabla^2 v(t) + u(t)v(t)^2 - (F+k)v(t) \end{aligned}$$

Our domain is a spatial grid of 128×128 pixels and the time step is 1. From his computational study, Pearson noted that there is no qualitative differences for domain sizes up to 1024×1024 and time steps as low as 0.01 [2]. The system is initially put in the red state R ($U = 1$, $V = 0$), with a 20×20 pixel square in the center with $U = 0.5$, $V = 0.25$. The entire grid is perturbed with 1% random noise to break up the initial symmetry of the central square.

The GS model displays spots, stripes, and spiral patterns from a simple PDE. Different patterns are formed, depending on the choice of parameters F and k which represent the rates at which the chemical U is resupplied and V removed. The patterns are certainly qualitatively distinguishable by inspection, so we hope that the topological information present in the patterns provide a quantitative classification that can distinguish between them.

Chapter 2

Homology

Topology is the study of shapes. Two shapes are considered equivalent if they can be continuously deformed into one another. In order to determine whether two shapes are equal, topological invariants are used. These are properties that do not change under continuous deformations. For example, the number of connected components is an invariant. We could stretch or wiggle a piece of string all we want and as long as we do not tear it (a discontinuous deformation), the string will always consist of one piece.

In algebraic topology, homology is another such invariant. Homology assigns groups to topological spaces and, as an invariant, it provides a test for whether two spaces are not equal. These groups also have a geometric interpretation. Homology captures the number of "holes" in a space. For a space X, its homology is the set of groups

$$H_0(X), H_1(X), H_2(X), \ldots$$

where the kth homology group $H_k(X)$ describes the k-dimensional holes in X. A 0-dimensional hole is a gap between 2 connected components (a maximal connected subset of X), so $H_0(X)$ describes the number of connected components. The dimension of these groups are the number of the "holes" in the space. A 1-dimensional hole is enclosed by a loop and a 2-dimensional hole is the cavity or hollow within a solid shell. For examples of these "holes", see the torus in Fig. 2.1.

Homology groups are calculable for a space of any dimension and describe higher dimensional holes that cannot be easily visualized, so they are useful for analyzing high dimensional data. The data from the Gray-Scott reaction with its two spatial and one temporal dimensions restricts our analysis to homology groups in dimensions $0, 1, 2$. By describing the "holes" in a space, homology will allow us to analyze the shape of our data. Before presenting homology theory, we present some examples.

In general, the homology of a space does not depend on the larger space it is embedded in, and in the next section we will clarify what we really mean by a hole. The space S^1 is the circle (see Fig. 2.2). It consists of one connected component and in the plane, it encloses a 1-dimensional hole. The circle's homology groups are

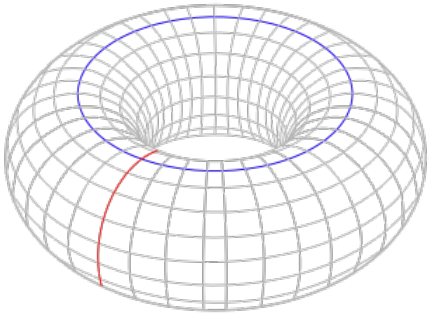

Figure 2.1: Adapted from [3]. The torus T is the surface around a donut. The red path on the surface encloses a loop and a cavity. The blue path encloses a loop. The 2D hole is described by the homology group $H_2(T) = \mathbb{Z}$. The 1D holes are described by $H_1(T) = \mathbb{Z} \oplus \mathbb{Z}$.

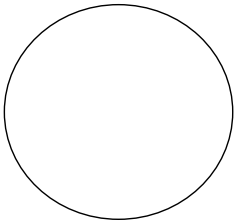

Figure 2.2: The circle S^1.

$$H_k(S^1) = \begin{cases} \mathbb{Z} & k = 0, 1 \\ 0 & \text{otherwise} \end{cases}$$

A 2D sphere S^2 has one connected component, no 1D holes, and one 2D hole (See Fig. 2.3). Its homology groups are

$$H_k(S^2) = \begin{cases} \mathbb{Z} & k = 0, 2 \\ 0 & \text{otherwise} \end{cases}$$

In general for an n-dimensional sphere S^n, the homology groups are

$$H_k(S^n) = \begin{cases} \mathbb{Z} & k = 0, n \\ 0 & \text{otherwise} \end{cases}$$

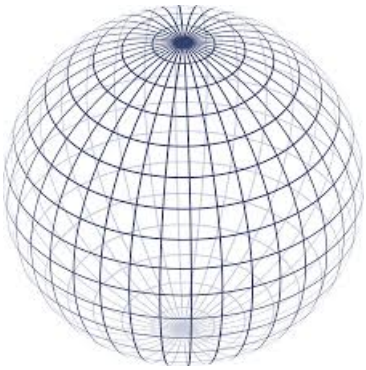

Figure 2.3: Adapted from [4]. The sphere S^2.

Figure 2.4: The ball B^2.

The solid 2D ball, B^2 is the filled-in circle (See Fig. 2.4). It has one connected component, but no holes. Filling in the center has removed the 1D hole. Its homology groups are

$$H_k(B^2) = \begin{cases} \mathbb{Z} & k = 0 \\ 0 & \text{otherwise} \end{cases}$$

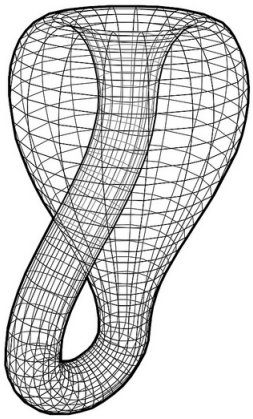

Figure 2.5: Adapted from [5]. The Klein bottle K is a nonorientable surface that intersects itself in 3D.

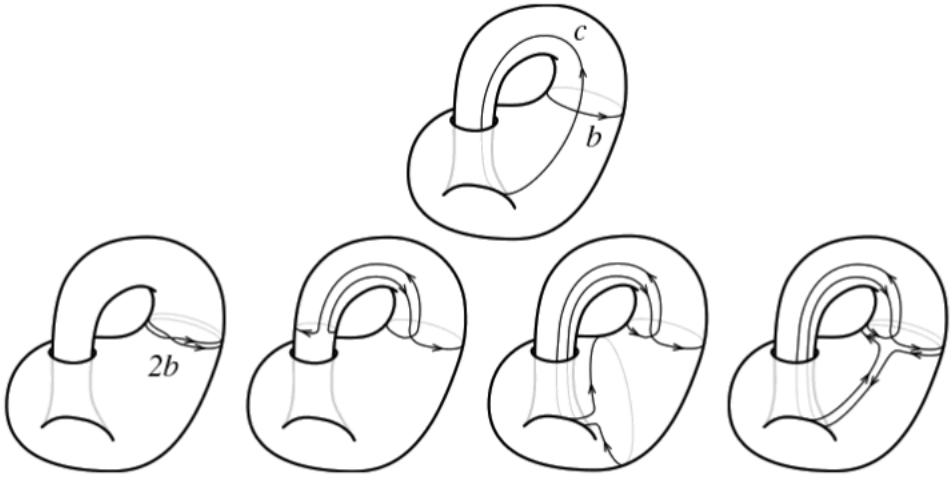

Figure 2.6: Figure adapted from [6, p.258]. The Klein bottle and two paths b, c that enclose 1D holes. The diagram demonstrates how $2b = b + c + (-b) + (-c) = 0$, explaining the torsion coefficient 2.

In these previous examples, the homology groups are the direct product of p copies of $\mathbb{Z}$, where p is the number of "holes". However, the Klein bottle (shown in Fig. 2.5) is

an example of a space with torsion, and its homology includes groups other than $\mathbb{Z}$.

The Klein bottle is a nonorientable surface. It only has one "side" because you can move from the "outside" to the "inside" without crossing an edge, so there is not a consistent choice of a normal vector. As a consequence of its nonorientability, a path may enclose a hole, but multiple traversals can result in a path that does not enclose a hole. In Fig. 2.6, the path b seems to enclose a 1D hole. But, traversing that path twice $2b$, can be decomposed into going around b, then around c, around b in the opposite direction, and finally around c in the opposite direction. So, $2b = b + c + (-b) + (-c)$. The right side equals 0, but we know that b is not 0 (it does not represent a loop that can be deformed into a point, keeping the loop on the Klein bottle). This property is captured by the torsion coefficient 2, as the homology groups are

$$H_k(K) = \begin{cases} \mathbb{Z} & k = 0 \\ \mathbb{Z} \oplus \mathbb{Z}_2 & k = 1 \\ 0 & \text{otherwise} \end{cases}$$

The group $\mathbb{Z}_2$ consists of the elements $0, 1$ and in this number system, $1 + 1 = 0$, providing an algebraic explanation of why $2b = b + b = 0$.

The fundamental theorem of finitely generated abelian groups states that a finitely generated abelian group is isomorphic to group of the form

$$\mathbb{Z}^n \oplus \mathbb{Z}_{k_1} \oplus \cdots \oplus \mathbb{Z}_{k_n}$$

where k_1 divides k_2, which divides k_3, and so on up to k_n. Our homology groups will take this general form, where $\mathbb{Z}_{k_i}$ are the torsion groups and k_i are the invariant factors. However, as persistent homology requires field coefficients [26], the software we use to calculate homology [27] uses $\mathbb{Z}_2$ coefficients. Therefore, our homology groups will be direct sums of $\mathbb{Z}_2$ (a $\mathbb{Z}_2$ vector space).

2.1 Simplicial Homology

Our digital images are composed of pixels, 2-dimensional cubes, and our reaction cubes (the images stacked on top of each other) are composed of voxels, the 3-dimensional analog of pixels. As such, spaces made up of intervals, squares, cubes, etc. (i.e. cubical complexes) are the natural setting. Cubical homology is treated in [24, 28, 29].

Here, our homology theory will be built on simplicial complexes, which are spaces constructed from higher dimensional triangles. Simplicial homology is easily illustrated and is also generalized in singular homology, where the topological invariance of homology groups is easily proven. We stress that all these homology theories are

equivalent (the homology groups are isomorphic).

The equivalence of simplicial and cubical homology is proven in [30]. The following treatment of simplicial homology follows [31].

2.1.1 Simplexes

Defininition 1. A set $A = \{a_0, a_1, ..., a_k\}$ of $k+1$ points in $\mathbb{R}^n$ is *geometrically independent* if the vectors $a_1 - a_0, a_2 - a_0, ..., a_k - a_0$ are linearly independent.

Equivalently, A is *geometrically independent* if there is no hyperplane of dimension $k-1$ that contains all the points in A.

Defininition 2. Let $A = \{a_0, a_1, ..., a_k\}$ be a set of geometrically independent points in $\mathbb{R}^n$ where $k < n$. The *k-simplex*, σ^k, spanned by A is the set of all points $x \in \mathbb{R}^n$ for which there exist nonnegative real numbers $\lambda_0, \lambda_1, ..., \lambda_k$ such that

$$x = \sum_{i=0}^{k} \lambda_i a_i \quad \text{and} \quad \sum_{i=0}^{k} \lambda_i = 1.$$

Lemma 1 will show that this expression is unique. The set $\{\lambda_i\}$ are the *barycentric coordinates* of the point x and the set $\{a_i\}$ are the *vertices* of the k-simplex σ^k. For a simplex σ^n with vertices $a_0, ..., a_n$ we write $\sigma^n = \langle a_0, \ldots, a_n \rangle$.

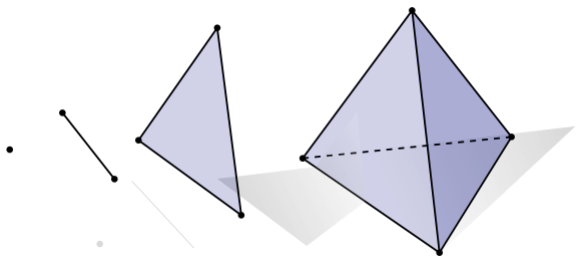

Figure 2.7: Adapted from [7]. $0, 1, 2$ simplexes

From these definitions, we see that a 0-simplex is a single point, a 1-simplex is a closed line segment, a 2-simplex is the interior and boundary of a triangle, and a 3-simplex is the interior and boundary of a tetrahedron. Moreover, a k-simplex is the smallest convex set with k vertices.

Now, we prove the result that simplexes are independent of their ambient space. A k simplex in $\mathbb{R}^n$ is equivalent to a k simplex in $\mathbb{R}^m$ where $m \neq n$.

Lemma 1. Let σ be a k simplex in $\mathbb{R}^n$ and let τ be a k simplex in $\mathbb{R}^m$, where $m, n \geq k$. Then, σ and τ are homeomorphic.

Proof. Let $\sigma = \langle v_0, \ldots, v_k \rangle$ and $\tau = \langle w_0, \ldots, w_k \rangle$. Let $h : \sigma \to \tau$ be given by

$$h\left(\sum_{j=0}^{k} \lambda_j v_j\right) = \sum_{j=0}^{k} \lambda_j w_j$$

for all nonnegative $\lambda_0, ..., \lambda_k$ such that $\sum_{j=0}^{k} \lambda_j = 1$. The function h is continuous because it is a linear mapping. The inverse $h^{-1}\left(\sum_{j=0}^{k} \lambda_j w_j\right) = \sum_{j=0}^{k} \lambda_j v_j$ is also continuous, so h is a homeomorphism. □

2.1.2 Simplicial Complexes

A triangle (2 simplex) is composed of sides (1 simplexes) and points (0 simplicies). In the same manner, a k simplex is composed of all lower dimensional simplexes. So, we define the faces of a simplex.

Defininition 3. For $k \leq n$, a simplex σ^k is a *face* of a simplex σ^n, if each vertex of σ^k is a vertex of σ^n.

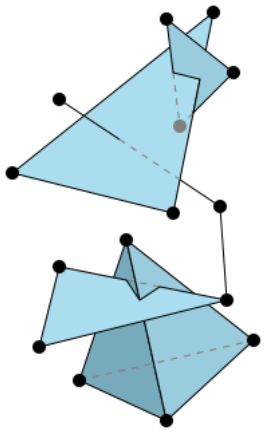

Figure 2.8: Adapted from [8]. Simplexes that are not properly joined. The intersection of the simplexes are not faces of either simplex.

Defininition 4. Two simplexes are *properly joined* if they do not intersect or if their intersection is a face of both simplexes.

With the previous definitions, we may now define simplicial complexes.

Defininition 5. A *simplicial complex* is a finite collection K of simplexes which are properly joined with the additional property that each face of a simplex in K must also be in K. The *dimension* of K is the largest positive integer r such that K has an r-simplex. The *geometric carrier* of K, denoted $|K|$, is $\bigcup_{\sigma \in K} \sigma$ with the Euclidean subspace topology.

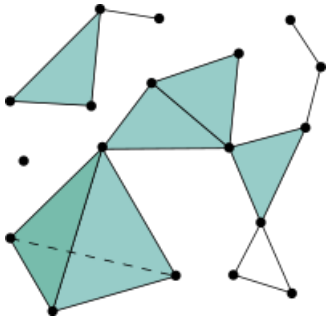

Figure 2.9: Adapted from [9]. A simplicial complex

Defininition 6. A topological space X is *triangulable* if there exists a geometric complex K whose geometric carrier $|K|$ is homeomorphic to X. The space $|K|$ is called a *triangulation* of X.

Defininition 7. If K is a simplicial complex and r is a positive integer, the *r-skeleton* of K is the simplicial complex consisting of all simplicies of K of dimension less than or equal to r.

Defininition 8. An *oriented n-simplex* assigns an orientation to an n-simplex. Given an ordering for the vertices of an n-simplex $\sigma^n = \langle a_0, \ldots, a_n \rangle$, the *positively oriented simplex* $+\sigma^n$ is the equivalence class of even permutations of the given ordering of vertices. The *negatively oriented simplex* $-\sigma^n$ is the equivalence class of odd permutations of the given ordering of vertices. Assigning an orientation to an n-simplex induces orientations on its faces.

An *orientated simplicial complex* is a simplicial complex with an orientation assigned to each of its simplexes. Moreover, if two simplexes share a face, they must induce opposite orientations on that face.

Triangulations and orientations are convenient methods in visualizing a polyhedron and cataloguing the arrangement of the simplexes. The actual topological structure (described by its homology groups) does not depend on these, as different triangulations and orientations will give the same answers (See Theorem 2).

2.2 Simplicial Homology Groups

Recall that the homology groups should describe the holes in a space. We introduce the notion of "cycles" which are simplicial complexes that traverse a region, but end up back at their starting point. In 2D, we can see that a cycle can enclose a hole. If the hole is filled in and part of the complex, then $H_1(X)$ should be 0 because there are no 1D holes. But, if the hole is not filled in, then $H_1(X)$ should reflect the fact that there is a hole.

In the case of the filled-in complex, the complex has a boundary. But, the hollow complex has no boundary (because it is itself a boundary, see Theorem 1). So, in a sense, holes are cycles that do not have boundaries.

Defininition 9. The *n-dimensional chain group* $C_n(K)$ is the free abelian group on the set of n-simplexes in a simplicial complex K.

Defininition 10. The *boundary map*, $\partial_p : C_p(K) \to C_{p-1}(K)$ is defined for each p-simplex $\langle a_0...a_p\rangle$ by

$$\partial_p(\sigma) = \sum_{i=0}^{p}(-1)^i\langle a_0...\hat{a_i}...a_p\rangle$$

where $\langle a_0...\hat{a_i}...a_p\rangle$ is the (p-1)-face of σ obtained by removing the vertex a_i.
The boundary map is extended linearly and it is a homomorphism between chain groups.

Theorem 1. The boundary map applied successively gives the zero map, $\partial_{p-1}\circ\partial_p = 0$

Proof. It suffices to show that $\partial_{p-1} \circ \partial_p(\sigma) = 0$ for an arbitrary p-simplex σ. By definition, $\partial_p(\sigma) = \sum_{i=0}^{p}(-1)^i\langle a_0...\hat{a_i}...a_p\rangle$, so

$$\partial_{p-1}\partial_p(\sigma) = \sum_{j<i}(-1)^i(-1)^j\langle a_0...\hat{a_j}...\hat{a_i}...a_p\rangle + \sum_{j>i}(-1)^i(-1)^{j-1}\langle a_0...\hat{a_i}...\hat{a_j}...a_p\rangle.$$

In the second sum, we relabel $i \leftrightarrow j$ to obtain the negative of the first sum. □

By the previous theorem we have that Im $\partial_{p+1} \subset$ Ker ∂_p [1], so we define the simplicial homology groups.

Defininition 11. The p^{th} simplicial homology group of K is the quotient group

$$H_p(K) = \text{Ker } \partial_p \text{ / Im } \partial_{p+1}.$$

We also call $Z_p = \text{Ker } \partial_p$ the *p-dimensional cycle group* and $B_p = \text{Im } \partial_{p+1}$ the *p-dimensional boundary group.* The rank of $H_p(K)$ is the p^{th} Betti number β_p.

In this previous section, simplicial homology defined homology groups on simplicial complexes. The homology groups were constructed with coefficients over $\mathbb{Z}$ because elements in the chain group were sums of simplexes. As commented previously, our computations use homology with coefficients in $\mathbb{Z}_2$. We briefly remark on the connection between the homology over $\mathbb{Z}$, $H_n(X)$, and homology over $\mathbb{Z}_2$, $H_n(X;\mathbb{Z}_2)$. To construct a homology theory over $\mathbb{Z}_2$, one can take the homology of the chain complex $C_n(X) \underset{\mathbb{Z}}{\otimes} \mathbb{Z}_2$. With this tensor product, we obtain a chain group with $\mathbb{Z}_2$ coefficients. By the Universal Coefficient Theorem [32, p.261], $H_n(X;\mathbb{Z}_2) \simeq H_n(X) \underset{\mathbb{Z}}{\otimes} \mathbb{Z}_2 \oplus \text{Tor}(H_{n-1}(X), \mathbb{Z}_2)$. Therefore, $H_n(X)$ and $H_{n-1}(X)$ determine $H_n(X;\mathbb{Z}_2)$.

[1] The Image of a map is its range. The Kernel of a map is the subset of its domain that gets sent to zero.

2.3 Singular Homology

In this section, we consider a more general formation of homology because simplicial homology is too restrictive, simplicial complexes being only a small subset of the spaces that mathematicians are interested in. Singular homology generalizes many results of simplicial homology and is built on *singular simplexes*, which are maps from simplexes to the space under consideration.

Singular homology has advantages over simplicial homology in that it can deal with non-simplicial structures, and furthermore the focus is placed on the space itself rather than specific triangulations of the space, so the theory does not need to check whether different triangulations result in equivalent homology groups. Also, singular and simplicial homology are equivalent (see Theorem 2 below).

The main result in this section is that homeomorphic spaces have isomorphic homology groups. Spaces with different homologies are not homeomorphic, so homology proves to be effective in identifying different spaces.

Defininition 12. A singular k-simplex in a space X is a map $\sigma : \sigma^k \to X$.

Analogous to simplicial homology, we define the singular chain group, singular boundary homomorphism, and singular homology groups.

Defininition 13. The n^{th} singular chain group of X, $C_n(X)$ is the free abelian group with basis the set of singular n-simplexes in X.

Defininition 14. The boundary map $\partial_n : C_n(X) \to C_{n-1}(X)$ is defined by

$$\partial_n(\sigma) = \sum_i (-1)^i \sigma|\langle a_0, \ldots, \hat{a_i}, \ldots, a_p\rangle,$$

where $\sigma|\langle a_0, \ldots, \hat{a_i}, \ldots, a_p\rangle$ is the restriction of the map σ to the $p-1$ simplex $\langle a_0, \ldots, \hat{a_i}, \ldots, a_p\rangle$.

The same proof in Theorem 1 can be used to show that $\partial_n \circ \partial_{n+1} = 0$. So, as before the singular homology groups are defined to be $H_n(X) = \operatorname{Ker} \partial_n / \operatorname{Im} \partial_{n+1}$.

Theorem 2. The simplicial and singular homology groups of a space X are isomorphic.

Proof. See "The Equivalence of Simplicial and Singular Homology" [32, p.128]. □

Theorem 3. If X is non-empty and path-connected, then $H_0(X) \approx \mathbb{Z}$. For any space X, $H_0(X)$ is a direct sum of $\mathbb{Z}$'s, one for each path-component of X.

Proof. See "Singular Homology" [32, p.109]. □

We will now prove that homotopy equivalent spaces have isomorphic homology groups. This is done by showing that a map $f : X \to Y$ induces a homomorphism $f_* : H_n(X) \to H_n(Y)$ for each dimension n and that f_* is an isomorphism if f is a homotopy equivalence.

For a map $f : X \to Y$, an induced homomorphism $f_\# : C_n(X) \to C_n(Y)$ is defined by composing each singular n-simplex $\sigma : \Delta^n \to X$ with f to get a singular n-simplex $f_\#(\sigma) = f\sigma : \Delta^n \to Y$.

Then $f_\#$ is extended linearly via $f_\# \left(\sum_i n_i \sigma_i \right) = \sum_i n_i f_\#(\sigma_i) = \sum_i n_i f\sigma_i$.

The maps $f_\# : C_n(X) \to C_n(Y)$ satisfy $f_\# \partial = \partial f_\#$ since

$$\begin{aligned} f_\# \partial(\sigma) &= f_\# \left(\sum_i (-1)^i \sigma | \langle a_0 \cdots \hat{a}_i \cdots a_n \rangle \right) \\ &= \sum_i (-1)^i f\sigma | \langle a_0 \cdots \hat{a}_i \cdots a_n \rangle \\ &= \partial f_\#(\sigma) \end{aligned}$$

The fact that the maps $f_\# : C_n(X) \to C_n(Y)$ satisfy $f_\# \partial = \partial f_\#$ is also expressed by saying that the $f_\#$'s define a *chain map* from the singular chain complex of X to that of Y.

The relation $f_\# \partial = \partial f_\#$ implies that $f_\#$ takes cycles to cycles since $\partial\alpha = 0$ implies $\partial(f_\#\alpha) = f_\#(\partial\alpha) = 0$. Also $f_\#$ takes boundaries to boundaries since $f_\#(\partial\beta) = \partial(f_\#\beta)$. Hence $f_\#$ induces a homomorphism $f_* : H_n(X) \to H_n(Y)$.

An algebraic statement of what we have just proved is:

Theorem 4. A chain map between chain complexes induces homomorphisms between the homology groups of the two complexes

Two basic properties of induced homomorphisms are:
1. $(fg)_* = f_* g_*$. This follows from the associativity of function composition.
2. $1_* = 1$ where 1 denotes the identity map of a space or a group.

From these properties, we have that homeomorphisms between spaces induce isomorphisms between their homology groups.

2.4 Persistent Homology

2.4.1 Motivation

The previous sections have described how to compute homology groups and Betti numbers for topological spaces. Topological spaces are idealized objects and even

Figure 2.10: Adapted from [10]. A noisy 2 dimensional data cloud. The data resembles a circle.

simple spaces such as a circle or a sphere consists of infinitely many points. We would like to use topological methods to analyze real world data sets which are finite in size and often noisy, so it may be difficult to determine whether the data points actually live in a meaningful topological space. Moreover, even if we knew the space the data is being sampled from, perhaps we would like a procedure to reconstruct spaces from the data to compare with the actual space.

For example, Fig. 2.10 displays a data set in the form of a point cloud. Although noisy, it is clear that the data resembles an annulus. So, if we could calculate Betti numbers for the data set, we would like $\beta_0 = 1$ and $\beta_1 = 1$ just as they would be for an annulus. However, as a discrete space, the data consists of a thousand separate points, so $\beta_0 = 1000$ and $\beta_1 = 0$.

Persistent homology is a tool in topological data analysis that deals with these problems. It identifies the shape of the data and quantitatively measures how well the data fits that shape. Fundamental to persistent homology is a *filtration*, which is a parameterized family of spaces constructed from the data. For point cloud data illustrated in Fig. 2.11, a common filtration is the *Vietoris-Rips complex*. The Vietoris-Rips complex grows balls around each of the points and adds simplexes as the balls intersect. Specifically, balls of radius ϵ are placed around each of the points.

In our analysis, we do not use the Vietoris-Rips complex to analyze our reaction images. We form a filtration using threshold (chemical concentration) as the persistence parameter and describe this procedure in Section 2.6. However, the Vietoris-Rips complex is useful in visualizing what persistent homology is and we describe it in detail for illustration purposes.

In the Vietoris-Rips Complex, each radius value ϵ has an associated simplicial complex constructed by the following:

- A 0-simplex is added for each point

- A 1-simplex is added between points x_1 and x_2 if their balls include the other

point, i.e. the distance between x_1 and x_2 is less than ϵ

- A 2-simplex is added between points x_1, x_2, and x_3 if the distance between each pair of points is less than ϵ

- ...

Fig. 2.11 shows simplicial complexes for select values of radius ϵ. When $\epsilon = 0$, the complex is just the points themselves. For a large enough value of ϵ, all the points are within ϵ of each other (the balls overlap by a wide enough margin) and the complex is a single k-simplex, where k is the dimension of the data set (if the points are in $\mathbb{R}^3$, then $k = 3$).

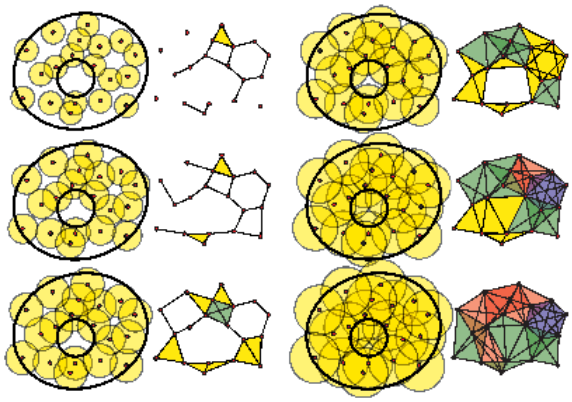

Figure 2.11: Adapted from [11]. A sequence of Vietoris-Rips complexes for a point cloud data set representing an annulus. Upon increasing the radius of the balls, holes appear and disappear.

For each value of ϵ, we have a simplicial complex with its own set of holes. As ϵ is increased, all the holes will eventually be filled, but some holes are more persistent in the filtration than others. Some holes may only last for a short radius interval if the points are clustered closely together. Certainly, the "real" hole is the central gap in the annulus. Other holes may form due to how the data points are spaced apart and they are essentially noise. When a hole appears in a filtration step, the filtration value is called the birth time. When a hole disappears, either by being filled in or by merging with another component in the case of H_0, the filtration value is called the death time.

Therefore, of utmost importance to persistent homology is the lifetime, birth, and death of holes. This information can be encoded in a barcode diagram. The barcode diagram for Fig. 2.11 is displayed in Fig. 2.12. As described in the previous section, each k-dimensional hole is a generator for the k-dimensional homology group H_k. The

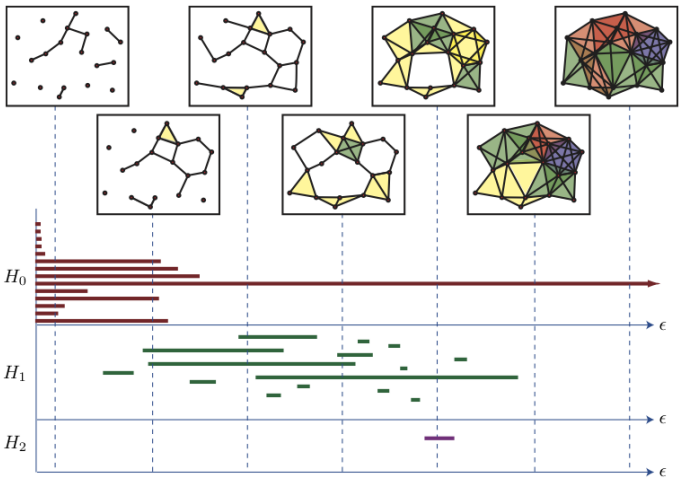

Figure 2.12: Adapted from [11]. The barcodes for the filtration in Fig. 2.11. The horizontal axis represents the filtration parameter, the radius value ϵ. A bar starts at the radius value for which the generator of homology (hole) appears. A bar ends ends at the radius value for which the generator disappears.

barcode diagram plots a bar for all generators of homology in each dimension. The bar starts at the birth time and ends at the death time. The length of the bars denote the persistence of the generator. Longer bars denote features that were present for a wide range of filtration values. The most persistent features are characteristic of the space the points are sampled from, while the short lived features are noise.

In sum, homologies can be calculated on spaces, but assigning a space to real data can be difficult. So, persistent homology considers multiple spaces and keeps track of the long-lasting features with respect to an adjustable parameter. These long-lasting, or persistent features are representative of the space the data is representing and from this information the shape of the data can be inferred.

2.4.2 Prior Work

Persistent homology has been recently used to study biological structures. In [13] Kovacev-Nikolic et al. study the maltose-binding protein (MBP), which is a bacterial protein found in Escherichia coli. They analyze multiple 3-dimensional structures of MBP, obtained by X-ray crystallography available from the Protein Data Bank. Using the persistence landscape (see Section 2.8), they are able to distinguish closed and open structures of MBP. In [33], Chung et al. apply persistent homology on digital images of the brain. They examine the thickness of gray matter and discover that for patients with autism, their brains display extra homological structures not found for patients without autism.

In soft condensed matter and atomic structures of materials, Xia et al. in [34] use

persistent homology to predict the energy of stability of fullerene molecules. They investigate the ground-state structures of small fullerene molecules with the Vietoris-Rips complex and identify local and global β_2 bars with the heat of formation and the total curvature energies respectively. In [35], Nakamura et al use persistent homology to extract a hierarchical structure of amorphous materials and distinguish them from perfect crystalline and random structures. They analyze the medium-range order in amorphous silica glass. In [36], Yamamoto et al. examine rough surfaces and numerically confirm that the lifespan of 0^{th} and 1^{st} homology generators follow power-law distributions, whose scaling exponents vary according to the roughness of the surface.

In [37], the 1D Gray-Scott model has been analyzed using time series of Betti numbers. Gameiro et al. calculate the maximal Lyapunov exponent for the time series and confirm the existence of spatio-temporal chaos in the Gray-Scott model. In [24], Hawkins analyzes the 2D Gray-Scott model and he calculates the Shannon entropy from the time series of Betti numbers. As an extension of [24], we employ persistent homology to solve the thresholding problem that Hawkins encountered.

2.4.3 Theory

In this section, we develop the mathematical theory behind persistent homology and make precise the ideas listed in the previous sections. We begin by defining a filtration.

Defininition 15. A filtration is a finite, nested sequence of simplicial complexes $\{X^i\}$ such that $i \leq j$ implies $X^i \subset X^j$.

$$\emptyset \subset X^0 \subset X^1 \subset ... \subset X^N = X.$$

Filtrations can be be thought of as sublevel sets X_i of a function f. For the sublevel sets to be nested, we require $f : X \to \mathbb{R}$ to be non-decreasing, meaning if σ is a face of τ, then $f(\sigma) \leq f(\tau)$.

The *sublevel subcomplexes* are $X^i = f^{-1}(-\infty, a_i]$, where $\{a_i\}_{i \in I}$, $a_i \in \mathbb{R}$ is an increasing sequence and I is an indexing set.

Each filtration step, X^i has its own homology groups. To look at which generators of homology persist, we compare the homology groups of two different filtration steps, X^i and X^j. Consider the step from $H_p(X^i)$ to $H_p(X^j)$. A homology class in $H_p(X^i)$ may be a boundary in $H_p(X^j)$, so it may become trivial and the hole is filled in the new filtration step. It may merge with an existing homology class in $H_p(X^j)$, in which case it becomes a nontrivial element of $H_p(X^j)$ by quotienting out by the new boundary group.

This process is displayed in Fig. 2.13. The purple region are the homology classes that have merged with existing homology classes in the new filtration step. The homology class γ is born at X_i because it is not in the purple region, which represents

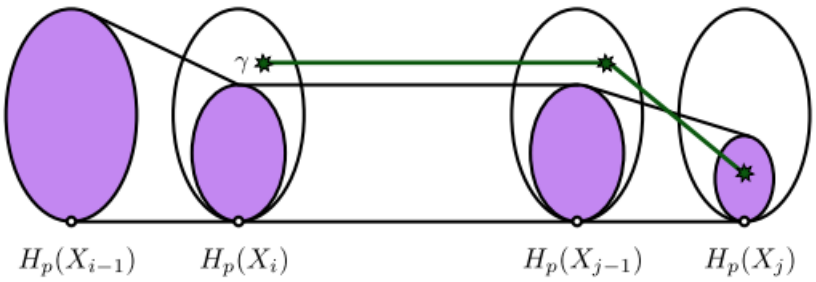

Figure 2.13: Adapted from [12]. This figure illustrates when a homology class is born and when it dies. Each oval represents the p^{th} homology group at a step in the filtration. The homology class γ is born at the i^{th} filtration step and dies entering the j^{th} filtration step by merging with an existing homology class.

the image of map between the homology groups $\varphi : H_p(X_{i-1}) \to H_p(X_i)$ induced by the inclusion map $i : X_{i-1} \to X_i$. This simply means that the "hole" appears for the first time in the i^{th} filtration step.

The class γ dies entering X_j because it was not in the image of the map from $H_p(X_{i-1})$ to $H_p(X_{i-1})$ and it merges with an older class upon entering $H_p(X_j)$. The "hole" has merged with a pre-existing "hole" in the j^{th} filtration step. Moreover, to have died at the j^{th} filtration step, the "hole" must be alive in the previous filtration step, the $j-1^{\text{th}}$. The white dot at the bottom of each oval represents the identity, the trivial element (p chains that are boundaries).

Now, the purpose of persistent homology is to keep track of those homology classes which are present throughout many steps in the filtration. This is accomplished by defining persistent homology groups, which consider the cycles in X^i that do not become boundaries in X^j.

Defininition 16. The p-persistent k-th homology group of K^l, denoted $H_k^{l,p}$ is

$$H_k^{l,p} = Z_k^l / (B_k^{l+p} \cap Z_k^l),$$

where Z_k^l is the k-dimensional cycle group of the l-th filtration step K^l and B_k^{l+p} is the k-dimensional boundary group of the $l+p$-th filtration step K^{l+p}.

The p-persistent k-th Betti number of the l-th filtration step is $\beta_k^{l,p}$, the rank of $H_k^{l,p}$.

For each $i \leq j$, we have the continuous inclusion map from $|X^i|$ to $|X^j|$. This map induces a homomorphism $f_p^{i,j} : H_p(X^i) \to H_p(X^j)$ for every dimension p of X.

Let $\gamma \in H_p(X^i)$. In terms of the persistent homology groups, γ is:

- *born* at X^i if $\gamma \notin H_p^{i-1,i}$, and

- *dies* entering X^j if $f_p^{i,j-1}(\gamma) \notin H_p^{i-1,j-1}$ but $f_p^{i,j}(\gamma) \in H_p^{i-1,j}$.

The *persistence* of γ is $j - i$

2.4.4 Barcodes and Persistence Diagrams (PDs)

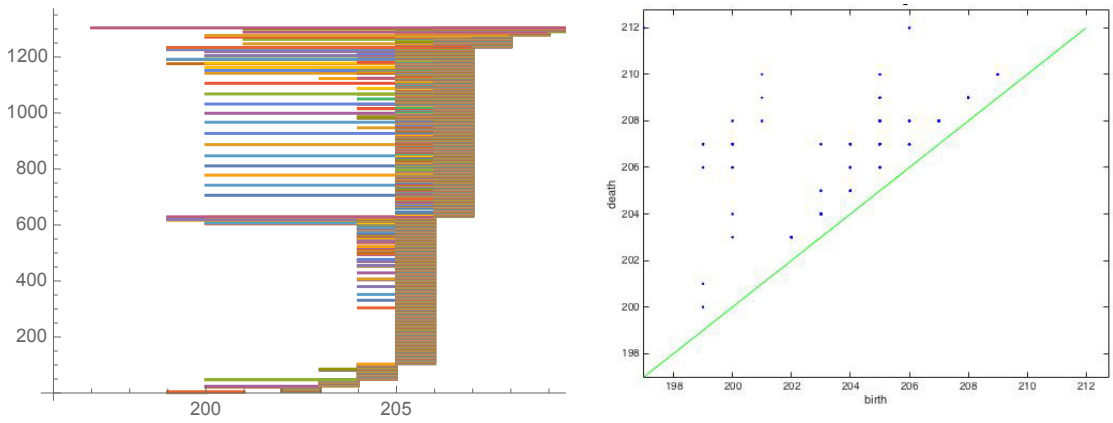

Figure 2.14: The barcodes and persistence diagram are shown side by side. The diagrams show the persistence information for α in the second dimension.

As described previously, barcodes are one way to visualize persistence. All generators of homology are plotted as bars starting at their birth times and ending at their death times. The persistence of the homology classes can also be visualized by a 2-dimensional scatter plot on $(\mathbb{R} \cup \{\pm\infty\})^2$. Let $\mu_p^{i,j}$ be the number of p-dimensional homology classes that are born at X_i and die entering X_j. Then,

$$\mu_p^{i,j} = (\beta_p^{i,j-1} - \beta_p^{i,j}) - (\beta_p^{i-1,j-1} - \beta_p^{i-1,j})$$

for all $i < j$ and for all p. The first difference is the number of classes born at or before X_i and die entering X_j. The second difference is the number of classes that are born at or before X_{i-1} and die entering X_j.

We draw each point (a_i, a_j) with multiplicity $\mu_p^{i,j}$ to obtain the *pth persistence diagram*. Points along the diagonal are also included with infinite multiplicity. This is important for Theorem 5. Fig. 2.14 shows a side by side comparison of a barcode and persistence diagram (PD). Persistent features in PDs are shown by points whose distance away from the diagonal is large (their death time is large compared to their birth time). It is easy to go from one representation to the other. Suppose we have barcodes. The start and end points of a bar make up the birth-death pair that is plotted in a scatter plot. Given a PD, for each point a bar is drawn beginning at its birth time. Its length is the vertical distance of the point to the diagonal.

Suppose we have two functions $f, g : X \to \mathbb{R}$ and the persistence diagrams arising from their sublevel sets. If f is similar to g, then we expect their filtrations to be

similar and therefore the diagrams as well. This is an important result in the stability of persistence diagrams and their susceptibility to perturbations and noise [12, 38].

Definiinition 17. Suppose given $X, Y \subset (M, \|\cdot\|)$, a normed space. The p^{th} Wasserstein distance between X and Y using the L^q distance is

$$W_p[L_q](X,Y) = \inf_{\varphi: X \to Y} \left(\sum_{x \in X} (\|x - \varphi(x)\|_q)^p \right)^{1/p}, \tag{2.1}$$

where $\|y\|_q = \left(\sum_n \|y_n\|^q \right)^{1/q}$ and $\varphi : X \to Y$ is a bijection between the two spaces.

When $p, q = \infty$, we have the bottleneck metric.

$$d_B(X,Y) = W_\infty[L^\infty](X,Y) = \inf_{\varphi: X \to Y} \sup_{x \in X} \|x - \varphi(x)\|_\infty, \tag{2.2}$$

where $\|y\|_\infty = \sup_n \|y_n\|$.

Theorem 5. If $D(f), D(g)$ are the persistence diagrams from the sublevel sets of f and g, then

$$d_B(D(f), D(g)) \leq \|f - g\|_\infty.$$

Proof. See "The Stability of Persistence Diagrams" by Cohen-Steiner et. al. [38]. □

So, small changes in a function imply only small changes in the diagram, making persistence diagrams stable.

2.5 Persistence Landscapes

Persistent homology provides a comprehensive survey of the many geometric features found in the space. Depending on the choice of barcodes or persistence plots, the persistent features are represented by long bars or points far away from the diagonal. However, barcode diagrams can contain thousands of bars and it is difficult to compare multiple diagrams. Also, when considering diagrams that are describing different samplings of the same space, we would like to consider the typical behavior of the space with averages of topological information.

In the case of non-persistent homology, we can simply average the Betti numbers, but for persistent homology, the procedure is not as clear. One attempt at defining the mean of two persistence diagrams was the Frechet mean [39]. For two persistence diagrams D_1, D_2 and a matching $b : D_1 \to D_2$ which realizes the p^{th} Wasserstein distance $W_p[L^q](D_1, D_2)$, the Frechet mean is the collection of midpoints of the line segments between x and $b(x)$ for all $x \in D_1$. However, the Frechet mean is not

unique. Consider a diagram D_1 which consists of the points (j, k) and $(j+1, k+1)$ (along with the diagonal) and a diagram D_2 with the points $(j, k+1)$ and $(j+1, k)$. Then, there are two minimal matchings, one matching the points horizontally, the second vertically. The Frechet mean in the first case consists of the points $(j+1/2, k)$ and $(j+1/2, k+1)$, while in the second case, $(j, k+1/2)$ and $(j+1, k+1/2)$. The non-uniqueness of Frechet means makes a statistical analysis of persistence diagrams difficult. There have been attempts to resolve these issues, but these ideas have not been realized computationally yet [39, 40].

In this section, we present persistence landscapes. As motivation, we return to the Frechet mean. The Frechet mean is not unique because it operates on a discrete set of points. This situation resembles computing the mean of two integers within the integers $\mathbb{Z}$ (Is the mean of five and eight 6 or 7?). The problem is solved introducing rational numbers $\mathbb{Q}$ (6.5 is a rational number). The same reasoning is used to introduce persistence landscapes. The idea is to embed persistence diagrams into the larger space of piecewise linear functions, where the mean is uniquely defined.

Persistence landscapes can distill the information in a persistence diagram diagrams into a number by an averaging procedure, where more weight is placed on the long bars rather than the short bars. This average can then be used as a measure to compare multiple barcode diagrams. From this comparison, we will be able to make assertions about the underlying topological spaces the barcode diagrams are representing.

The persistence landscape is treated formally in [41], but we will describe it intuitively following [13, 39].

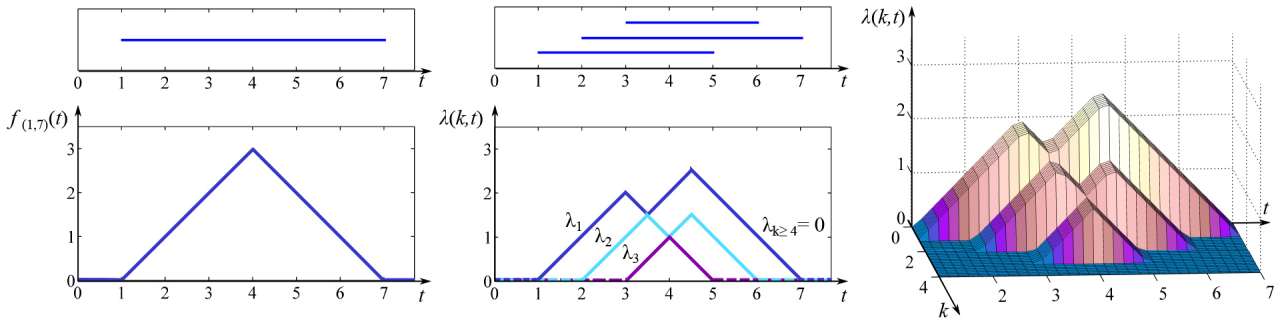

Figure 2.15: Adapted from [13]. Construction of a persistence landscape: (left) from an interval to the auxiliary function f; (middle) from a barcode diagram to a persistence landscape; (right) 3-D visualization of the persistence landscape.

Defininition 18. The persistence landscape is a function $\lambda : \mathbb{N} \times \mathbb{R} \to \overline{\mathbb{R}}$ and it is defined for each dimension p, where $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ Alternatively, it may be thought

of as a sequence of functions $\lambda_k : \mathbb{R} \to \overline{\mathbb{R}}$. Define

$$\lambda_k(t) = \sup(m \geq 0 | \beta_p^{t-m,t+m} \geq k).$$

To help visualize the graph of $\lambda : \mathbb{N} \times \mathbb{R} \to \overline{\mathbb{R}}$, we can extend it to a function $\overline{\lambda} : \mathbb{R} \times \mathbb{R} \to \overline{\mathbb{R}}$ by setting

$$\overline{\lambda}(x,t) = \begin{cases} \lambda(\lceil x \rceil, t), & \text{if } x > 0 \\ 0, & \text{if } x \leq 0. \end{cases}$$

For a persistence diagram with m off diagonal points $\{(a_i, b_i)\}$, $\lambda_k(t)$ is the kth largest value of $\{f_{(a_i,b_i)}(t)\}_{i=1}^m$, where $f_{(a,b)}(t) = \max(\min(t-a, b-t), 0)$. $\lambda_k(t) = 0$ whenever $k > m$.

For each bar, we first calculate the shortest distance from t to its endpoints. If t is not within the ends of the bar, the distance is 0. Then, $\lambda_k(t)$ is simply the k-th largest distance.

Informally, this procedure stacks isosceles triangles, whose bases are the intervals in the barcode diagram. Fig. 2.15 shows the construction of a persistence landscape from a barcode diagram. There are 3 bars, so there are 3 functions $\lambda_1, \lambda_2, \lambda_3$.

The persistence landscape encodes the same information as the persistence diagram, but is presented in a more accessible manner. For one, it creates a filtered representation of the persistence diagram premised on the assumption that homology classes with large lifespans are more important than those with shorter lifespans. This is reflected by the monotonicity of the sequence, $\lambda_k \geq \lambda_{k+1}$ for all $k \in \mathbb{N}$. Second, the data is now in the form of piecewise linear functions which can be averaged to obtain another piecewise linear function. Given persistence landscapes $\lambda_1, \ldots, \lambda_n$, the averaged persistence landscape is determined pointwise $\sum_{i=1}^{n} \lambda_i / n$. This average is uniquely defined and as a real-valued function on $\mathbb{N} \times \mathbb{R}$, we can use the L^p metric to quantify how different two landscapes are, where $\|\lambda\|_p = \left(\sum_{k=1}^{\infty} \int_{\mathbb{R}} |\lambda_k|^p \right)^{1/p}$.

When the metric L^1 is used, we consider the total area underneath the persistence landscape. We call this the persistent landscape area:

$$\|\lambda\|_1 = \sum_{k=1}^{\infty} \int_{\mathbb{R}} |\lambda_k| \tag{2.3}$$

The points that have small lifetimes contribute little to the persistent landscape area and are thus essentially regarded as noise.

Beyond the scope of this thesis, persistent landscapes have other useful properties. Defining a distance between landscapes, this distance sets a lower bound for the bottleneck distance between the associated persistence diagrams. Landscapes are also Banach space random variables, and satisfy analogous results from probability theory (Strong Law of Large Numbers, Central Value Theorem) [41].

In this chapter, we have seen how homology captures connectivity and shape information. It is particularly adaptable to images, which are already in the form of a discrete cubical complex. Persistent homology allows analysis of the geometric features at all resolutions, from transitory phenomena to the robust. Persistence landscapes provide a method to average persistence diagrams and to compare them. Overall, these topological summaries compress the intricacies of the data and simplify analysis by highlighting the important attributes. In the next chapter, we will describe statistical methods that use the information gathered from persistent homology.

Chapter 3

Statistical Methods

In this chapter, we describe the statistical methods that we used to analyze the persistent homological data. The methods used are Principal Component Analysis (PCA) [42], Multidimensional Scaling (MDS) [43] and a classification test [39]. PCA is useful because it reduces the dimensionality of the data and converts the correlated variables into linearly uncorrelated variables called principal components. The data here are the persistent areas are calculated for various reactions. As we will show in Chapter 4, this 3 dimensional data is shown to be well-represented by just 2 variables. MDS reconstructs possible coordinates given pairwise distances of a set of objects.

The following treatment follows [42] and [43].

3.1 Principal Component Analysis

PCA is a statistical procedure that converts a set of observations of possibly correlated variables to a set of values of linearly uncorrelated variables called principal components. Fig. 3.1 demonstrates PCA for 3 dimensional data. The points nearly lie along a 2D plane. PCA determines the axes of this plane and expresses the data in terms of these new axes, ultimately rotating the data to highlight any redundancy in the data.

The data is represented as a matrix $\mathbf{D}$.

$$\mathbf{D} = \begin{bmatrix} \mathbf{y_1}^\intercal \\ \mathbf{y_2}^\intercal \\ \vdots \\ \mathbf{y_n}^\intercal \end{bmatrix}, \tag{3.1}$$

where $\mathbf{y_i}^\intercal$ is the row vector representing the i^{th} observation. The elements of $\mathbf{y_i}$ are the different variables recorded from the i^{th} observation.
Essentially, we look for the dimensions along which the observations are most widely spread out or separated. The first principal component is the dimension with the largest spread, the second principal component is the dimension with the next largest

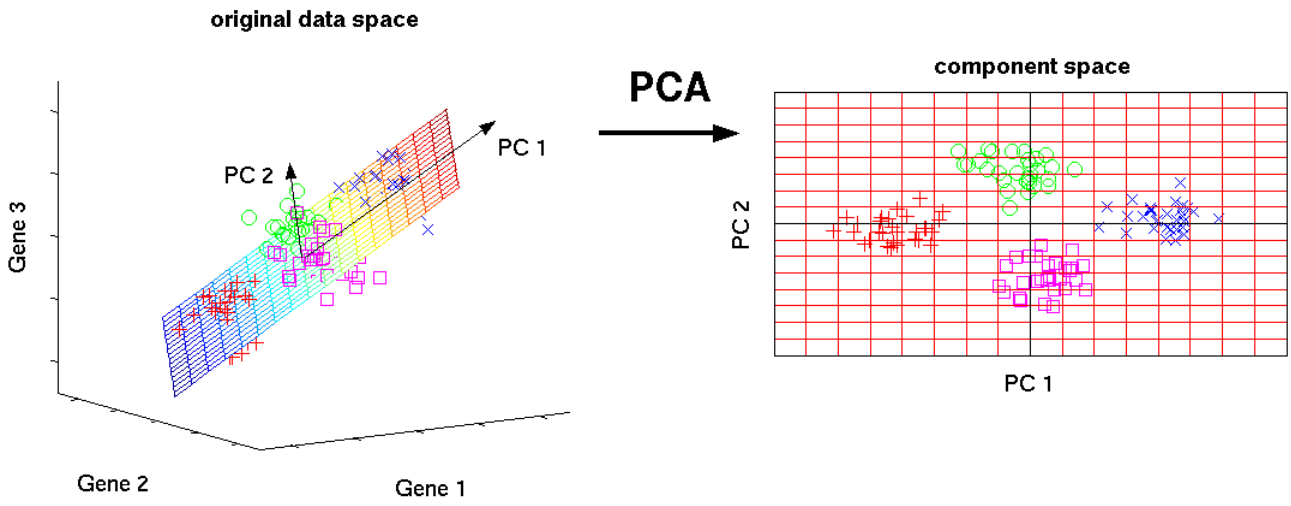

Figure 3.1: Adapted from [14]. PCA is used to reduce a large number of variables (genes) to a lower number of new variables called principal components. 3D gene expression samples under 4 experimental conditions are projected onto a 2D principal component space that maintains the largest variance in the data. This visualization allows qualitative conclusions to be made about the separability of the 4 experimental conditions. PCA simply rotates the original data such that the principal components are the axes of a new coordinate system.

spread orthogonal to the first, and so on. When our variables are highly correlated, the first two principal components will explain most of the variance in the data, and we may then visualize the p-dimensional data in 2 dimensions.

Figure 3.1 shows why we can reduce the dimensionality of the data. Looking at the data, the points mostly lie in a plane. In this plane, the points resemble an ellipse and the semi major and semi minor axes show the spread of the data. Thus, PCA is akin to fitting an n dimensional ellipsoid to the data, where each principal axis of the ellipsoid represents a principal component. If some axis is small, then the variance along that axis is also small. So, by projecting the data only the principal components with the highest variances, we reduce the dimensionality of the data but retain most of the essential information.

3.1.1 The Procedure

We have a sample of n vectors $\mathbf{y_i}^\intercal$ in a p-dimensional space. We would like to find the axes that best describe these points. So, we translate the origin to the mean vector

$$\overline{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y_i}. \tag{3.2}$$

From here on we will write $\mathbf{y_i}$ but we will assume the vector has been centered to $\mathbf{y_i} - \overline{\mathbf{y}}$.

Next, we need to rotate the axes so that in these new axes, the variables are uncorrelated. We rotate the axes by multiplying each $\mathbf{y_i}$ by an orthogonal matrix $\mathbf{A}$:

$$\mathbf{z_i} = \mathbf{A}\mathbf{y_i} \tag{3.3}$$

In these new axes, the new variables (principal components), which are the components $z_{\mathbf{i}j}$ in the transformed observation vector $\mathbf{z_i} = \mathbf{A}\mathbf{y_i}$, are uncorrelated. The principal components are uncorrelated when their covariance is 0. Therefore, finding these axes is equivalent to diagonalizing the covariance matrix of $\mathbf{z}$, $\mathbf{S}_z$, whose off diagonal elements are the covariance between the principal components.

$$\mathbf{S}_z = \mathbf{A}\mathbf{S}\mathbf{A}^\intercal = \begin{bmatrix} s_{z_1}^2 & 0 & \dots & 0 \\ 0 & s_{z_2}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & s_{z_p}^2 \end{bmatrix}, \tag{3.4}$$

where $\mathbf{S}$ is the covariance matrix of $\mathbf{y_i}$ and $s_{z_i}^2 \geq s_{z_{i+1}}^2$. So, the orthogonal matrix $\mathbf{A}$ is

$$\mathbf{A} = \begin{bmatrix} \mathbf{a_1}^\intercal \\ \mathbf{a_2}^\intercal \\ \vdots \\ \mathbf{a_p}^\intercal \end{bmatrix}, \tag{3.5}$$

where $\mathbf{a_i}$ is the i^{th} normalized eigenvector of $\mathbf{S}$. Then the variances $s_{z_i}^2$ are the eigenvalues λ_i of $\mathbf{S}$ and the principal components of the vector $\mathbf{y_i}$ are the variables

$$\mathbf{z}_{\mathbf{i}j} = \mathbf{a_j}^\intercal \mathbf{y_i}. \tag{3.6}$$

If the first k variances of the principal components make up a high proportion of the total variance, i.e. $\frac{\lambda_1+\lambda_2+\cdots+\lambda_k}{\sum_i \lambda_i} \approx 1$, then the first k dimensions represent the data well. If k is 2 or 3, we may graph the data.

3.2 Multidimensional Scaling

Unlike PCA, which identifies the ideal axes to display a set of points, MDS obtains a hypothetical set of coordinates given distances or differences between the points. Particularly, this method allows us to visualize the information contained in a distance matrix.

3.2.1 The Procedure

Given the matrix $\mathbf{D}^2$, whose elements $\mathbf{D}_{ij}^2$ are the squared distance from item i to item j, we can rewrite its elements in terms of the hypothetical coordinates of the

items, $\mathbf{x_i}$ and $\mathbf{x_j}$. The elements of are the matrix $\mathbf{D}^2$ are

$$\mathbf{D}^2_{ij} = (\mathbf{x_i} - \mathbf{x_j})^\intercal(\mathbf{x_i} - \mathbf{x_j}) = |\mathbf{x_i}|^2 - 2\mathbf{x_i}^\intercal\mathbf{x_j} + \left|\mathbf{x_j}\right|^2. \tag{3.7}$$

We rewrite the above as

$$\mathbf{D}^2 = \mathbf{Z} - 2\mathbf{X}^\intercal\mathbf{X} + \mathbf{Z}^\intercal$$

$$\mathbf{Z} = \begin{bmatrix} |\mathbf{x_1}|^2 & |\mathbf{x_1}|^2 & \dots & |\mathbf{x_1}|^2 \\ |\mathbf{x_2}|^2 & |\mathbf{x_2}|^2 & \dots & |\mathbf{x_2}|^2 \\ \vdots & \vdots & \ddots & \vdots \\ |\mathbf{x_n}|^2 & |\mathbf{x_n}|^2 & \dots & |\mathbf{x_n}|^2 \end{bmatrix} \quad , \quad \mathbf{X} = \begin{bmatrix} \mathbf{x_1} & \mathbf{x_2} & \dots & \mathbf{x_n} \end{bmatrix}$$

Now, we will solve for the matrix $\mathbf{X}$, the matrix with the coordinates of our items. We define the $n \times n$ centering matrix as

$$\mathbf{H} = \frac{1}{n}\begin{bmatrix} n-1 & & & & \\ & n-1 & & -1 & \\ & & \ddots & & \\ & -1 & & n-1 & \\ & & & & n-1 \end{bmatrix} \tag{3.8}$$

It is easy to show that

$$\mathbf{B} = -\frac{1}{2}\mathbf{H}\mathbf{D}^2\mathbf{H} = \widetilde{\mathbf{X}}^\intercal\widetilde{\mathbf{X}} \tag{3.9}$$

where

$$\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{H}.$$

Since $\mathbf{B}$ is a positive definite, symmetric matrix, it is diagonalizable and can be written as

$$\mathbf{B} = \mathbf{U}\Sigma\mathbf{U}^\intercal \tag{3.10}$$

$$= \mathbf{U}\Sigma^{\frac{1}{2}}\,(\mathbf{U}\Sigma^{\frac{1}{2}})^\intercal, \tag{3.11}$$

where $\mathbf{U}$ is the $n \times n$ matrix with the normalized eigenvectors of $\mathbf{B}$ as its columns, Σ is a $n \times n$ real diagonal matrix, and the entries in $\Sigma^{\frac{1}{2}}$ are the square root of the entries in Σ.

Finally, one solution for $\mathbf{X}$ is

$$\mathbf{X} = \mathbf{U}\Sigma^{\frac{1}{2}} \tag{3.12}$$

From a matrix of distances ($\mathbf{D}^2$), MDS has constructed a hypothetical set of coordinates ($\mathbf{X}$, whose columns $\mathbf{x}_i$ represent the ith observation) that match these distances.

3.3 Classification through Persistence Landscapes

Although homology can test whether spaces are different, we would like to use homology to measure how similar they are. Specifically, after recording certain observations, we want a classification method that is able to determine the factors that produced them. We will use this method to recover the parameters of the Gray-Scott reaction from the homological data. This classification is described in [39] and is implemented in [44].

Suppose we are given N different classes $C_1, \ldots, C_N$ of persistence landscapes. All the members in a class share certain similarities. The classes will represent the persistence landscapes of different Gray-Scott patterns or of different values of the feed rate F when the reaction rate k is fixed. Each class C_n also consists of S different persistence landscapes $\lambda_{nj}^k, j = 1, \ldots, S$ in dimension k where $k \in \{0, 1, 2\}$. Our classification scheme is based on L^p distances (see Eq. 2.1) between persistence landscapes, so we choose and fix a real number $p \in [1, \infty]$.

If we are given a new persistence landscape L and would like to decide which of the classes $C_1, \ldots, C_N$ it belongs to, we choose a dimension k and use the following classification scheme based on the persistence information.

Given N classes of persistence landscapes, for each $n = 1, ..., N$ and S samples of landscapes from each class, we define the average classifier $\overline{C}_n^k$ of the n^{th} class in dimension k via

$$\overline{C}_n^k = \frac{1}{S} \cdot \sum_{j=1}^{S} \lambda_{nj}^k \tag{3.13}$$

Now let L be any other persistence landscape we would like to classify. Then we say that L has been classified to belong to class C_n using dimension k, if there exists a unique index $n \in 1, ..., N$ such that

$$||L^k - \overline{C}_n^k||_{L^p} \leq ||L^k - \overline{C}_j^k||_{L^p}. \tag{3.14}$$

If no such unique index n exists, then we say that the classifier fails to classify λ.

Given different persistence landscapes for a class, this procedure simply computes their average to determine the typical persistence behavior. Then, given a new persistence landscape L, we use the average classifier that is closest to L to determine the class which most likely contains L.

Chapter 4

Methodology and Results

Images of the GS patterns are obtained by numerically solving [24] the PDEs in Eq. 1.2 for a range of parameter values F and k. We choose to perform our simulations on a 128×128 pixel grid and record the chemical concentration of V. For our experiments, we run the simulation for $15,000$ time steps record the values of V every 50^{th} time step, resulting in an image. We stack the images on top of one another to obtain a 3D representation of the reaction growing in time. We call this a time tube.

Each point in the time tube has a grayscale value representing the concentration of the chemical at a given location and time, so our 3 dimensional time tube is like a murky chemical cloud. To obtain a topological space, we include only the parts in the time tube whose grayscale value is below a certain threshold. The chemical concentration (from 0 to 1) is converted to a greyscale value between 0 (black) and 255 (white). We represent the highest chemical concentration as black and low chemical concentrations as white. Increasing the threshold value includes whiter pixels, points with smaller concentrations of V. We can obtain a filtration by varying the threshold value as increasing the threshold produces a space that contains those produced from smaller thresholds. Note that when the threshold is 0, the space is empty and when the threshold is 255, the space includes the entire time tube.

We format our time tube data appropriately into a file as input to the program **Perseus** [27]. Perseus computes persistent homology efficiently with an algorithm based on discrete Morse theory [45]. For a time tube, the corresponding input file lists the grayscale values of each 3D cube in the larger cubical domain. The grayscale values are the birth times of the 3D cubes because we have 255 spaces in our filtration, one for each grayscale value. Once a 3D cube is born in a space, spaces associated with a later filtration step always contain that 3D cube. Perseus computes the birth and death values for all generators of the 0^{th}, 1^{st}, and 2^{nd} homology groups (Our space is 3 dimensional, so no $4, 5, \ldots$ simplexes exist in our space. We then have that the n^{th} homology groups, where $n > 2$, are trivial groups). The birth and death pairs are represented as barcodes or persistence diagrams.

We analyze persistence in three ways. First, the bottleneck distance in Eq. 2.2 is

computed between the persistence diagrams of all of Pearson's 12 nontrivial patterns. Using MDS on these distances, we show that the patterns are distinguishable. Second and in a similar vein, we compute the persistent landscape area for all of Pearson's 12 nontrivial patterns and use principal component analysis to again distinguish the patterns. This affirms our use of persistent landscape area as a sort of metric. Third, for the entire F-k parameter space, the persistent landscape area is computed and plotted in contour maps. Finally, we use the classification method in Section 3.3 to recover F and k.

4.1 Analysis with Multidimensional Scaling

We simulate each of the 12 Pearson patterns for 10 different initial conditions. Recall that the system is put in the state $U = 1, V = 0$ with a 20×20 pixel square in the center with $U = 0.5, V = 0.25$. Each initial condition perturbs the grid with 1% random noise. The persistence diagram is calculated for each, and using the **R** package **TDA** [46], the bottleneck distance between each pair of patterns is calculated. We create a distance matrix where the element in the ith row and jth column is the bottleneck distance between the ith and jth pattern (the patterns are ordered first by their alphabetical names and then by the trial number i.e. $\alpha_1, \alpha_2, \cdots, \alpha_{10}, \beta_1, \cdots, \zeta_{10}$).

R's MDS function **cmdscale** is used to produce a set of coordinates for the patterns in 2 dimensions. Due to computational limitations, bars whose persistence was less than 5 were disregarded for bottleneck distances in dimension 1 and 2.

The results for the $0^{\text{th}}, 1^{\text{st}}, 2^{\text{nd}}$ bottleneck distances are plotted in Fig. 4.1. In these plots, we see distinct clustering of different patterns. For example, for the 0^{th} dimension, the patterns $\kappa, \iota, \delta, \beta$ are clearly distinguished from the other patterns. The 1^{st} and 2^{nd} dimensions separate more patterns, but there is always overlap (e.g. α and ϵ). We will see that PCA is able to distinguish these patterns.

4.2 Analysis using Principal Component Analysis

The same 10 trials of each of the 12 Pearson patterns from the previous section are used. The persistent landscape areas in dimensions $0, 1, 2$ are calculated for each pattern and form the i^{th} row $\mathbf{y}_i^\intercal = [y_{i0}, y_{i1}, y_{i2}]$ for each pattern and trial in our data matrix $\mathbf{Y}$. So, the first row of $\mathbf{Y}$ contains the $0^{\text{th}}, 1^{\text{st}}$ and 2^{nd} persistent areas of the first trial of α. The last row of $\mathbf{Y}$ contains the persistent areas of the 10^{th} trial of ζ.

Using **R**'s PCA function **prcomp**, the 3 eigenvectors of the covariance matrix of the principal components are

$$\begin{bmatrix} 0.4948 \\ -0.7241 \\ 0.4805 \end{bmatrix}, \begin{bmatrix} -0.5170 \\ -0.6897 \\ -0.5069 \end{bmatrix}, \begin{bmatrix} -0.6984 \\ -0.0024 \\ 0.7157 \end{bmatrix}$$

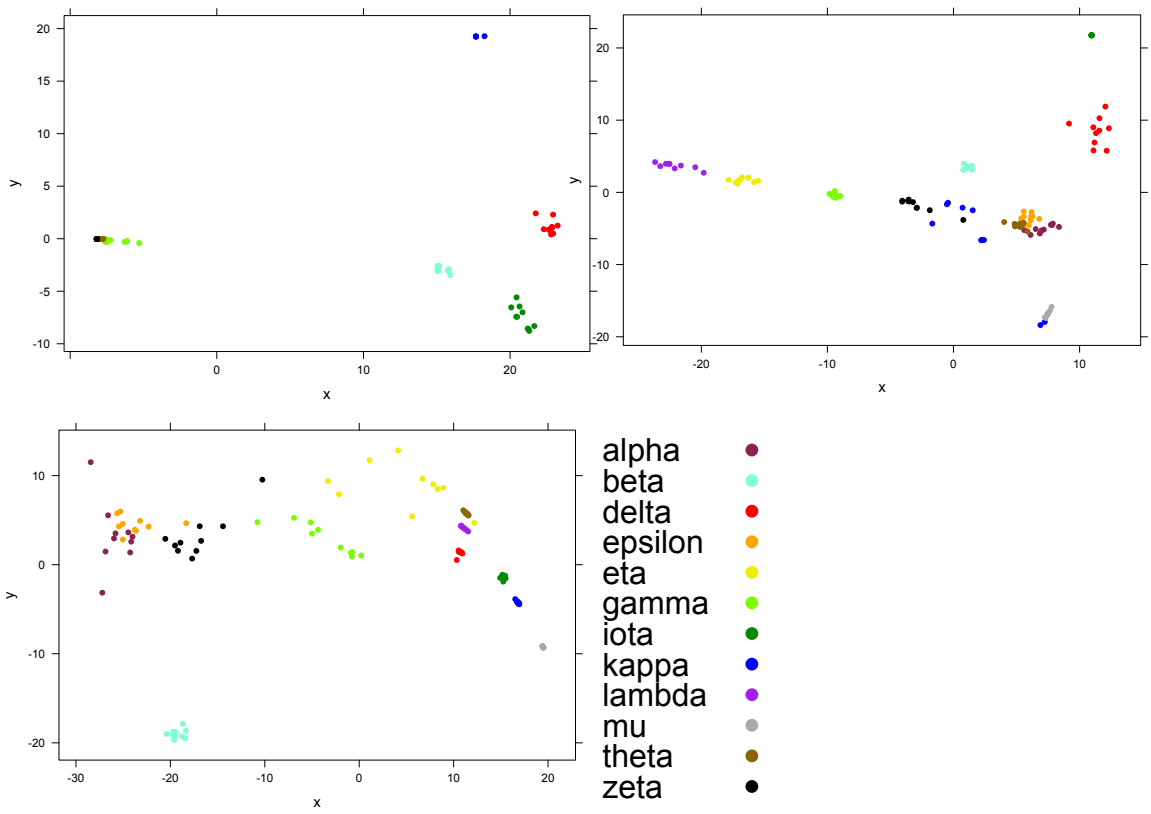

Figure 4.1: MDS using the bottleneck distances between the persistence diagrams (PDs). The top left panel shows an MDS plot for the 0^{th} dimensional PDs and the top right panel, for the 1^{st} dimension. The bottom left panel shows the plot for the 2^{nd} dimensional PDs.

The standard deviations of the principal components are $1.1347, 0.9931$, and 0.8521. So, the first two principal components capture 0.758 of the total variance. In Fig. 4.2, the persistent areas are projected onto the first 2 principal components. The data is well grouped by the pattern types, as compared to MDS. Therefore, the persistent areas show great utility in distinguishing different patterns.

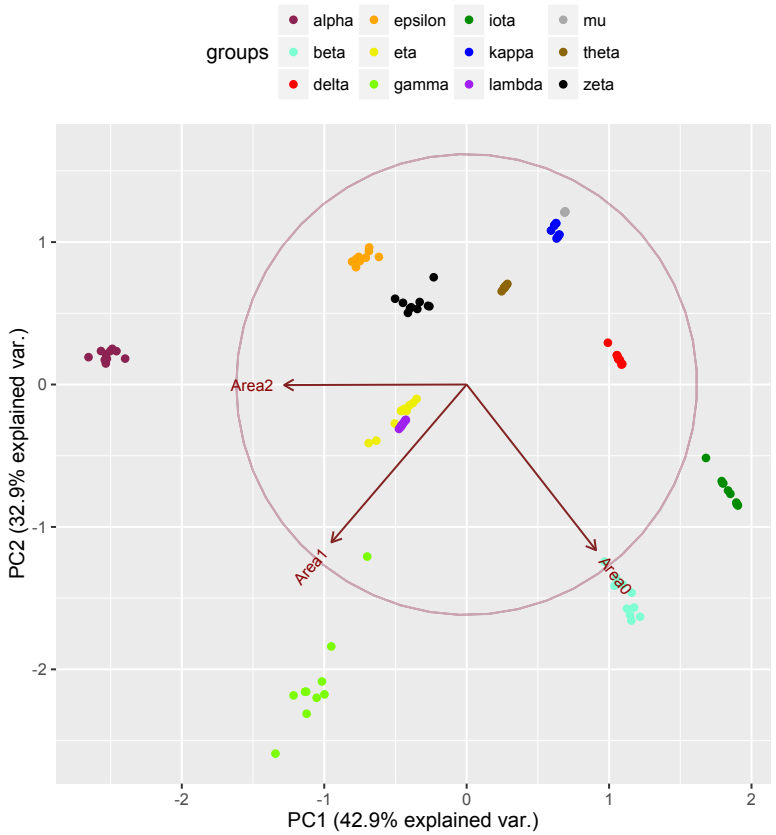

Figure 4.2: PCA plot of the persistent areas projected onto the first 2 principal components. Although some patterns are spread wider than others, PCA analysis on the persistent areas do group up the trials by their pattern type. The eigenvectors of the covariance matrix are shown as red arrows inside the unit circle. We see some overlap between λ, η, so in distinguishing patterns, our persistent areas are liable to error. We shall test how well persistent areas distinguish patterns in Section 4.4.

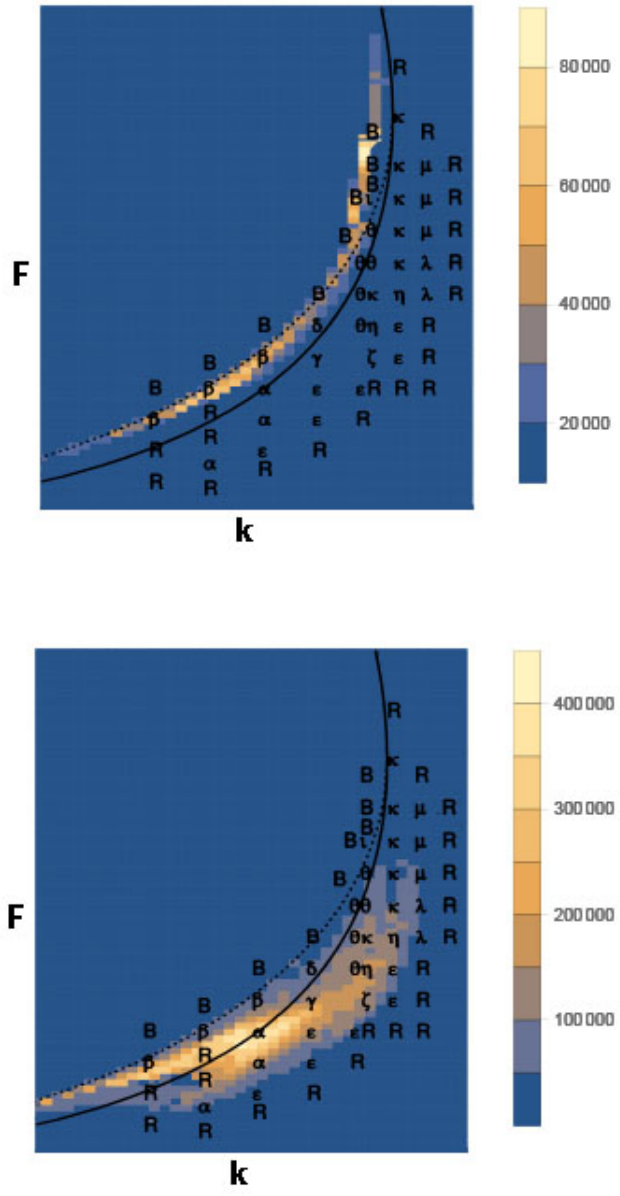

Figure 4.3: Top and Bottom: The plots of the 0^{th} and 1^{st} persistent landscape areas.

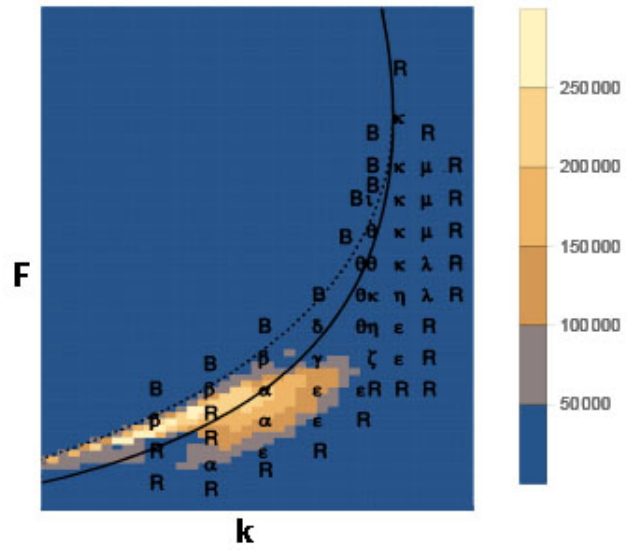

Figure 4.4: The plot of the 2^{nd} persistent landscape area

4.3 Persistent Landscape Areas

The previous analyses were able to group and distinguish patterns based on their persistent homology. However, the plots are unrelated to the F-k parameter space. Rather, the previous methods assign hypothetical coordinates to the patterns given metrics between patterns such as distance. Here, we consider the persistent homology of all patterns in the F-k parameter space. In Fig. 4.3 and 4.4, the 0^{th}, 1^{st}, and 2^{nd} persistent landscape areas are plotted in contour maps. As an example, the persistent landscapes for α are shown in Fig. 4.5. We vary the parameters F from $[0, 0.8]$ and k from $[0.03, 0.07]$ in increments of 0.001.

Far away from the bifurcation curves in Pearson's parameter space (See Fig. 1.2), the contour maps in all dimensions have small persistent areas because these patterns are the completely uniform states. Outside the solid line is the red state R, which evolves to a completely black picture. Outside the dotted line is the blue state B, which evolves to a completely white picture. The time tubes from the states R and B resemble respectively the whole region of the cubical domain and the empty space. These spaces do not have very many changes in homology as the threshold is varied, so we expect the areas to be small.

Qualitatively, the contour maps have maxima and high values near the dotted and solid lines because these patterns have many homological generators and geometric features arising from the loss of stability of the uniform states at the bifurcation curves. Our contour plots in Figs. 4.3 and 4.4 suggest that transitions between pat-

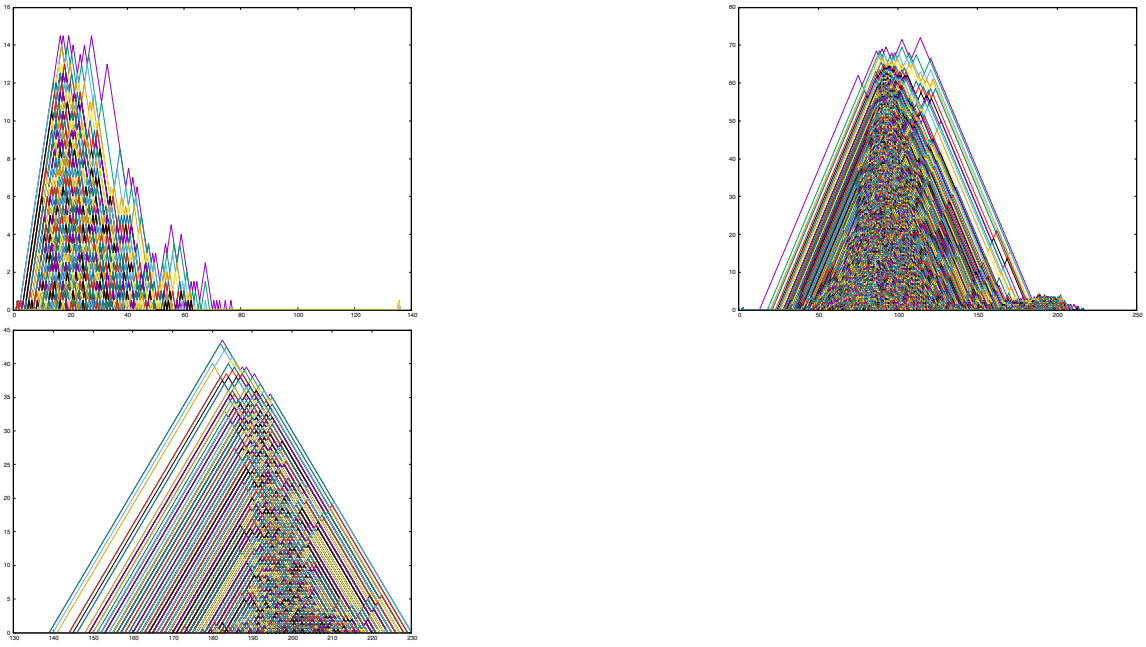

Figure 4.5: Example persistence landscapes for α. The top left and right plots display the 0^{th} and 1^{st} dimensional persistence landscapes respectively. The bottom plot displays the 2^{nd} dimensional persistence landscape.

terns can be detected with homology especially as each dimension highlights different aspects of the bifurcation curves. The 0^{th} dimensional case captures the dotted line, which denotes the loss of stability of the state B via a Hopf bifurcation. It makes sense that β lies along the peaks of the 0 dimensional contour plot because β generates many components which contribute to the 0^{th} area.

The 1^{st} dimensional case captures the lower region of the solid line, which is the boundary of the saddle node bifurcation, peaking at the pattern α. The plot tapers off but shows high activity around γ, ϵ, and η. At first glance, the 2^{nd} dimensional case seems to not give much more information than the 1^{st} dimensional case. It peaks around β rather than at α.

Figures 4.6 and 4.7 show the magnitude of the gradient of the contour plots in Figs. 4.3 and 4.4. We are interested in the gradient because because it shows the changes in homology. High gradient values between patterns would suggest sharp changes in homology. Since homology is a topological invariant, the different homologies would show that the patterns form different topological spaces. The gradient plots show this behavior in the lower left region of the $F - k$ parameter space.

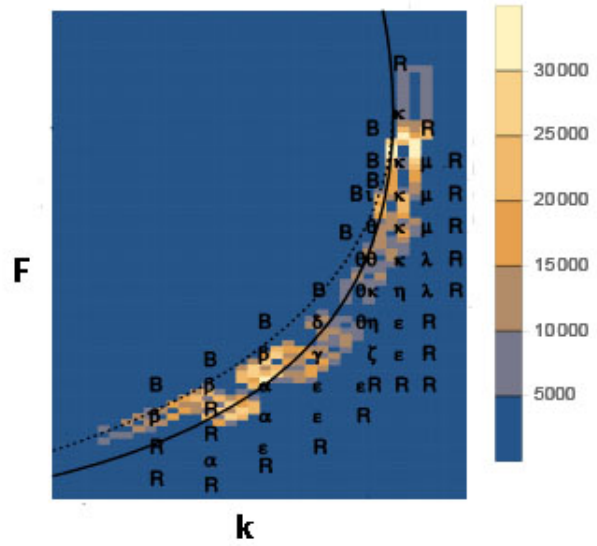

Figure 4.6: Plot of the gradient magnitude of the 0^{th} area contour plot in Fig. 4.3

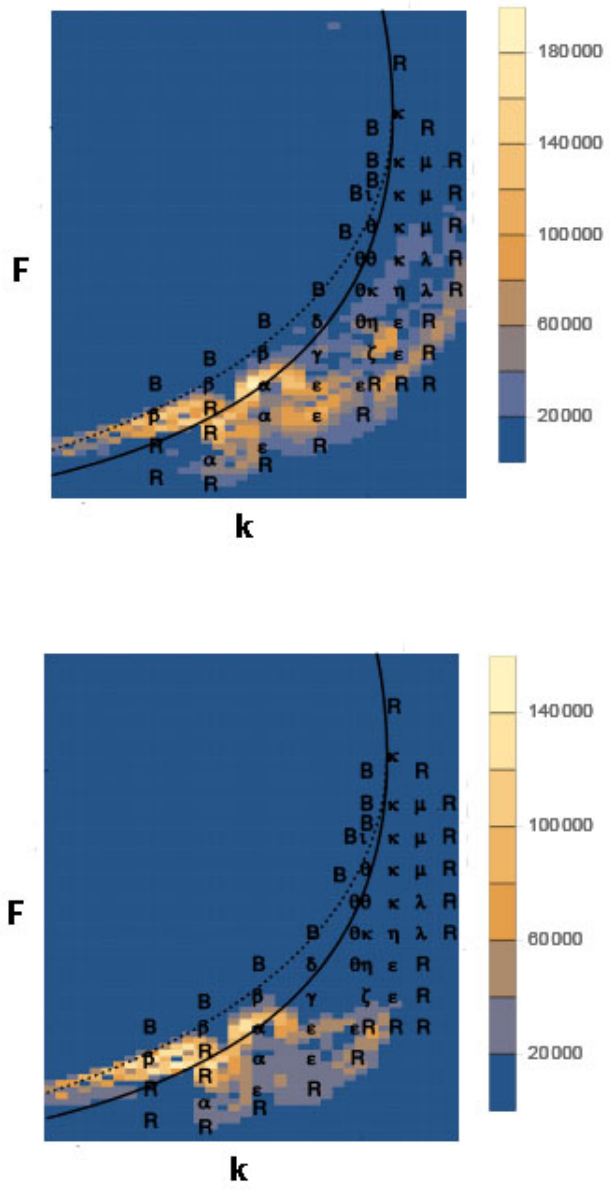

Figure 4.7: Top & Bottom: Plots of the gradient magnitude of the 1st and 2nd area contour plots in Figs. 4.3 and 4.4

4.4 Classifying Patterns

The persistence landscape area contour maps seems to trace out parts of the bifurcation curves, but they do not give very much insight into how the spaces formed by the reactions change when the parameters F, k are varied. In this section, we use statistical classification [39] to identify the reaction parameters given the topological information for a given data set.

First, we vary the parameter F for a fixed k. In each collection, we vary F by 0.002 and have 9 different trials for each F, k pair. Each trial begins the simulation with a different initial condition as described in Section 4.1. We consider the following 4 data sets with N different F values. Since F is varied by 0.002, for the data sets $N = 10, 8, 8, 8$ respectively.

(F1) $k = 0.0610$ and F is varied from $0.020 - 0.038$

(F2) $k = 0.0635$ and F is varied from $0.024 - 0.038$

(F3) $k = 0.0560$ and F is varied from $0.012 - 0.026$.

(F4) $k = 0.0505$ and F is varied from $0.008 - 0.022$

Second, we vary the parameter k for a fixed F. Again we vary k by 0.002. We consider the following 3 data sets with N different k values.

(K1) $F = 0.023$ and k is varied from $0.050 - 0.062$

(K2) $F = 0.029$ and k is varied from $0.056 - 0.064$

(K3) $F = 0.409$ and k is varied from $0.060 - 0.066$

Third, we consider the patterns that appear more than once in Pearson's categorization, $\alpha, \beta, \epsilon, \kappa, \lambda, \theta$. The patterns have N different (F, k) values $3, 3, 6, 5, 2, 5$ respectively, meaning they appear N times in Pearson's categorization (See Fig. 1.2). For each collection Pα, Pβ, ..., we perform 9 trials. So, for each of the 3 patterns of α, we run the simulation for 9 different initial conditions.

Based on the topological data collected from the above reactions, we perform a statistical study to reconstruct the reaction, i.e. determine the parameters F, k. We use the statistical classification implemented in [44] and described in Section 3.3. For each collection, we choose two trials as a training set. With this training set, we calculate the average persistence landscape of the N classes (i.e. F, k pairs). We then use these to classify the rest of the data. We repeat this procedure for all possible permutations of training sets (picking 2 trials out of the 9), and average the results. These averages are displayed in Tables 4.1, 4.2, 4.3, and 4.4. We discuss the classification in more detail before explaining the results.

Figure 4.8: An illustration of the classification method. The data set varies the parameter F. The top row shows the results from the classification method from the best match - zeroth to the worst - fourth class. The bottom row shows that the actual class was the 2nd box from the left. The numbers in the box show how far away these classes are from the actual class. In this example, the zeroth class misses by 1, being one box to the left of the "hit" box. The first class misses by 2 because it is two boxes away from the box to the right of the "hit" box, i.e. the box labeled 1. The second class is a hit because it is at the same location as where the box labeled 2 is.

For the K and F collections, we can measure the accuracy of the classification method. Figure 4.8 illustrates the method. Suppose we are trying to classify a landscape into one of five classes. Since the collections are columns or rows in the $F - k$ parameter space, the classes can be visualized as a row of boxes. Recall from Section 3.3 that the classification finds the class whose average peristence landscape is closest to the given persistence landscape, i.e. the distance between the given persistence landscape and that of the chosen class is the smallest. We can rank all the classes by these distances, from smallest to largest. Since the classes are like a row of boxes, where the boxes next to each other differ in the parameter k or F by 0.002, we expect the classes further from the actual class to have a large distance. Of course, the zeroth class (the best matching class) should be the actual class of the pattern. The first class (next best matching class) should be a class next to the actual class in the row of boxes. The second class should be 2 boxes away from the actual class. This is made clear by in the example in Fig. 4.8.

The bottom row is labeled "actual" because it highlights the actual class, "hit", that the given landscape should be classified in. Also this row shows the expected ranking for the classification method. We expect the classification to rank the classes in order from their distance away from the box "hit". It is unclear whether the box to the left or to the right should be the first class, so we consider them both and label them 1 because they are one box away from the "hit" box. The boxes labeled 2 are two boxes away from the "hit" box, and so on.

The top row labeled "classification" shows an example of a ranking from a classification. We can determine the success of the classification by examining its rankings. We call the result of the classification a hit, miss by 1, miss by 2, or wrong depending on how close the result is to the expected result. In Fig. 4.8, our zeroth class misses

by 1, since it is one box away from the "hit" box, which is the actual class. Next, we consider whether our first class from the classification matches the next best matching class. Our first class misses by 2 because it is 2 boxes away from a box labeled 1. Our third best guess is a hit because it is at the same location as the box labeled 2. So, if the parameter value of the chosen class differs from that of the actual class by $0.002 \times n$, the result misses by n. We will call misses by more than 2, wrong.

For third set of data, the P collections representing the patterns $\alpha, \beta, \ldots$, we simply identify whether the classification is a hit or a miss. A hit identifies the exact parameters F, k, while a miss chooses the wrong parameters.

For the F and K collections, Tables 4.1, 4.2, and 4.3 display the classification results for the persistence landscapes from the 0^{th}, 1^{st}, and 2^{nd} dimensions respectively. This classification method is very robust. The zeroth class is correct almost all the time and the first class is correct over half the time.

Table 4.4 displays the classification results for the P collections and shows that persistence landscapes are able to group patterns with different parameters. The number of "hits" greatly exceed the number of "wrongs", so most of the time, a pattern is matched to its correct parameters. So, even though a pattern such as α may appear at different locations in parameter space (Fig. 1.2), persistence landscapes are group them by their parameters.

Table 4.1: Persistence landscapes are obtained for the 0^{th} dimension. The left tables utilize the metric L^1 and the right tables, the metric L^2. From top to bottom, the tables show the results from analyzing the zeroth, first, and second classes.

L^1

pattern	hit	miss 1	miss 2	wrong
F1	66.8	3.25	0	0
F2	54.2	1.8	0	0
F3	54.9	1.1	0	0
F4	56	0	0	0
K1	49	0	0	0
K2	34.9	0.1	0	0
K3	28	0	0	0
pattern	hit	miss 1	miss 2	wrong
F1	64.9	5.1	0	0
F2	54.2	1.8	0	0
F3	47.9	1.1	0	7
F4	49	0	0	7
K1	35	0	0	0
K2	28.8	6.2	0	0
K3	21	0	7	0
pattern	hit	miss 1	miss 2	wrong
F1	29.6	2.9	0	37.5
F2	21.8	0.2	0	34
F3	19.7	5.1	0.8	30.4
F4	33.6	0	0	22.4
K1	16.2	2.6	7	23.2
K2	7.8	18.1	0	9.1
K3	21	0	0	7

L^2

pattern	hit	miss 1	miss 2	wrong
F1	67.2	2.8	0	0
F2	54.2	1.8	0	0
F3	53.6	1	0	1.4
F4	56	0	0	0
K1	49	0	0	0
K2	35	0	0	0
K3	28	0	0	0
pattern	hit	miss 1	miss 2	wrong
F1	66.2	3.8	0	0
F2	54.2	1.8	0	0
F3	45.6	2.5	0	7.9
F4	50.9	0	2	3.1
K1	37.9	0	5.5	5.6
K2	35	0	0	0
K3	21	0	7	0
pattern	hit	miss 1	miss 2	wrong
F1	29.9	0.2	0	39.9
F2	15.2	0	0	40.8
F3	19.1	0.4	0	36.5
F4	28.9	1.4	0.3	25.4
K1	22.5	0.9	3.6	22
K2	19.7	1.8	0	13.5
K3	21	0	0	7

Table 4.2: Persistence landscapes are obtained for the 1st dimension. The left tables utilize the metric L^1 and the right tables, the metric L^2. From top to bottom, the tables show the results from analyzing the zeroth, first, and second classes.

L^1

pattern	hit	miss 1	miss 2	wrong
F1	69.4	0.6	0	0
F2	54.5	1.5	0	0
F3	56	0	0	0
F4	56	0	0	0
K1	49	0	0	0
K2	35	0	0	0
K3	27	1	0	0
pattern	hit	miss 1	miss 2	wrong
F1	69.4	0.6	0	0
F2	47.9	2.6	0	5.5
F3	49	0	0	7
F4	49	0	0	7
K1	24.7	0	0	24.3
K2	35	0	0	0
K3	21	7	0	0
pattern	hit	miss 1	miss 2	wrong
F1	27.3	0	7	35.7
F2	7.5	0.7	3.9	43.9
F3	7.1	0	2.3	46.6
F4	7	0	0	49
K1	7	7	3.7	31.3
K2	2.4	0.8	13.4	18.4
K3	14	0	0	14

L^2

pattern	hit	miss 1	miss 2	wrong
F1	67	3	0	0
F2	49.6	3.5	1.4	1.5
F3	56	0	0	0
F4	55.6	0	0	0.4
K1	49	0	0	0
K2	35	0	0	0
K3	27	1	0	0
pattern	hit	miss 1	miss 2	wrong
F1	61.6	8.4	0	0
F2	30.4	9.9	1	14.7
F3	49	0	0	7
F4	47.5	0	0	8.5
K1	28	0	0	21
K2	35	0	0	0
K3	20.9	7.1	0	0
pattern	hit	miss 1	miss 2	wrong
F1	22.3	0.2	7	40.5
F2	7.2	2.5	4.3	42
F3	2.4	0	0	53.6
F4	4.3	0	0	51.7
K1	3.8	4.4	0	40.8
K2	11.6	5.2	0	18.2
K3	14.9	0	0	13.1

Table 4.3: Persistence landscapes are obtained for the 2^{nd} dimension. The left tables utilize the metric L^1 and the right tables, the metric L^2. From top to bottom, the tables show the results from analyzing the zeroth, first, and second classes.

L^1

pattern	hit	miss 1	miss 2	wrong
F1	62.5	6.2	1.3	0
F2	56	0	0	0
F3	56	0	0	0
F4	56	0	0	0
K1	48.9	0	0	0.1
K2	32.9	0.7	1.4	0
K3	28	0	0	0
pattern	hit	miss 1	miss 2	0
F1	48.6	14.4	3.6	3.4
F2	56	0	0	0
F3	41.3	0	0	14.7
F4	41.3	0	0	14.7
K1	23.5	6.5	7.5	0.5
K2	23.7	4.2	7.1	0
K3	7	12.3	8.7	0
pattern	hit	miss 1	miss 2	wrong
F1	24.4	12.6	1.4	31.6
F2	27.9	0	0	28.1
F3	14	0	6.3	35.7
F4	14	0	0	42
K1	10.7	1.3	0	37
K2	8.4	8.1	14	4.5
K3	8.6	5.3	0	14.1

L^2

pattern	hit	miss 1	miss 2	wrong
F1	63	5	1.8	0.2
F2	56	0	0	0
F3	56	0	0	0
F4	56	0	0	0
K1	49	0	0	0
K2	32.9	1.4	0.7	0
K3	28	0	0	0
pattern	hit	miss 1	miss 2	wrong
F1	50.1	11.2	2.4	6.3
F2	56	0	0	0
F3	42.2	0	0	13.8
F4	42	0	0	14
K1	27.8	10.3	3.9	7
K2	20.8	3.1	11.1	0
K3	13	12.6	2.4	0
pattern	hit	miss 1	miss 2	wrong
F1	27.9	12	1.2	28.9
F2	21	0	0	35
F3	15.4	0	5.9	34.7
F4	16.9	0	0	39.1
K1	13.3	3.4	0	32.3
K2	13	5.2	6.2	10.6
K3	13.3	5.2	0	9.5

Table 4.4: Persistence landscapes for the P collections are obtained for all three dimensions. The top table utilize the metric L^1 and the bottom table, the metric L^2.

L^1

	dim0		dim1		dim2	
pattern	hit	wrong	hit	wrong	hit	wrong
α	21	0	21	0	21	0
β	21	0	21	0	21	0
ϵ	42	0	42	0	42	0
κ	27.4	7.6	31.7	3.3	35	0
λ	14	0	14	0	14	0
θ	33.5	2.5	34.6	0.4	33.4	1.6

L^2

	dim0		dim1		dim2	
pattern	hit	wrong	hit	wrong	hit	wrong
α	21	0	21	0	21	0
β	21	0	21	0	21	0
ϵ	40.9	1.1	42	0	42	0
κ	26.7	8.3	29.8	5.2	34.8	0.2
λ	14	0	14	0	14	0
θ	33.2	2.8	35	0	33.4	1.6

Conclusion

In this thesis, we have explored how topology can be used in data analysis. We analyzed a simple pattern forming system, the Gray-Scott model, with homology. Moreover, persistent homology proves to be particularly robust, providing a mechanism to discriminate between noise and the actual topological characteristics in data sets. Even slight changes in the reaction parameters made a noticeable difference in the topology. Further work could include recovering the time from a slice of the reaction. Instead of considering the reaction as a whole, parts of the reaction could be considered independently. It would be interesting to determine whether persistent homology could detect what stage the reaction is in, and if there is an optimal way to split up the reaction.

Persistent homology captures the topology of a filtration. In order to produce a filtration, we have varied threshold values in our analysis. However, there are many other parameters that we could vary including the domain size, boundary conditions, and diffusion rates. Multidimensional persistence extends persistent homology to a multi-filtration, i.e., to spaces filtered by multiple parameters. It would be interesting to determine how these parameters affect the topology of the reaction in addition to the threshold.

References

[1] "Belousov Zhabotinsky reaction 8 x normal speed." `https://www.youtube.com/watch?v=3JAqrRnKFHo`, January 2010.

[2] J. E. Pearson, "Complex patterns in a simple system," *Science*, vol. 261, pp. 189–192, 1993.

[3] Y. Mrabet, "Simple torus with cycles." https://commons.wikimedia.org/wiki/File

[4] Geek3, "Sphere wireframe 10deg 6r." https://commons.wikimedia.org/wiki/File2009.

[5] Lhoste, "31 août klein bottle." http://www.lhostemusic.com/klein-bottle/.

[6] D. S. Richeson, *Euler's Gem: The Polyhedron Formula and the Birth of Topology*. Princeton University Press, Princeton, 2012.

[7] K. Crane, "Discrete exterior calculus." http://brickisland.net/cs177fa12/?p=307, 2012.

[8] Trevorgoodchild, "Simplicial complex nonexample," 2009.

[9] cflm, "Simplicial complex example." https://commons.wikimedia.org/wiki/File:Simplicial_complex_example.svg, 2009.

[10] G. Ellis, "About hap: Persistent homology." http://hamilton.nuigalway.ie/Hap/www/SideLinks/About/aboutPersistent.html.

[11] R. Ghrist, "Barcodes: The persistent topology of data," *Bulletin of the American Mathematical Society*, vol. 45, no. 1, pp. 61–75, 2008.

[12] E. Munch, *Applications of Persistent Homology to Time Varying Systems*. PhD thesis, Department of Mathematics, Duke University, 2013.

[13] V. Kovacev-Nikolic, P. Bubenik, and D. Nikolić, "Using persistent homology and dynamical distances to analyze protein binding," *Statistical Applications in Genetics and Molecular Biology*, vol. 15, no. 1, pp. 19–38, 2016.

[14] M. Scholz, *Approaches to Analyse and Interpret Biological Profile Data*. PhD thesis, Max Planck Institute of Molecular Plant Physiology, Potsdam University, January 2006.

[15] "Millenium problems." `http://www.claymath.org/millennium-problems`, April 2016.

[16] A. M. Turing, "The chemical basis of morphogenesis," *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, vol. 237, no. 641, pp. 33–72, 1952.

[17] N. Tompkins, N. Li, C. Girabawe, M. Heymann, G. B. Ermentrout, I. R. Epstein, and S. Fraden, "Testing Turing's theory of morphogenisis in chemical cells," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 12, pp. 4397–4402, 2014.

[18] R. Field, E. Körös, and R. Noyes, "Oscillations in chemical systems. II. thorough analysis of temporal oscillation in the bromate-cerium-malonic acid system," *Journal of the American Chemical Society*, vol. 94, no. 25, pp. 8649–8664, 1972.

[19] F. Sagués and I. R. Epstein, "Nonlinear chemical dynamics," *Dalton Trans.*, pp. 1201–1217, 2003.

[20] P. K. Maini, K. J. Painter, and H. Nguyen Phong Chau, "Spatial pattern formation in chemical and biological systems," *J. Chem. Soc., Faraday Trans.*, vol. 93, pp. 3601–3610, 1997.

[21] C. Guldberg and P.Waage, "Studies concerning affinity," *Journal of Chemical Education*, vol. 63, no. 12, pp. 1044–1047, 1986.

[22] W. Mazin, K. Rasmussen, E. Mosekilde, P. Borckmans, and G. Dewel, "Pattern formation in the bistable Gray-Scott model," *Mathematics and Computers in Simulation*, vol. 40, no. 3-4, pp. 371–396, 1996.

[23] Y. Kyrychko, K. Blyuss, S. J. Hogan, and E. Schöll, "Control of spatiotemporal patterns in the Gray-Scott model," *Chaos*, vol. 19, no. 4, 2009.

[24] J. Hawkins, *Looking at Pictures: Topological Analysis of Complex Reaction-Diffusion Patterns*. Undergraduate thesis, Reed College, 2015.

[25] J. Butcher, *Numerical Methods for Ordinary Differential Equations*. Wiley, New York City, 2nd ed., 2008.

[26] A. Zomorodian and G. Carlsson, "Computing persistent homology," *Discrete and Computational Geometry*, vol. 33, no. 2, pp. 249–274, 2005.

[27] V. Nanda, "Perseus, the persistent homology software." `http://www.sas.upenn.edu/~vnanda/perseus/`.

[28] T. Kaczynski, K. Mischaikow, and M. Mrozek, *Computational Homology*, vol. 157 of *Applied Mathematical Sciences*. Springe-Verlag, New York, 2004.

[29] M. Gameiro, *Topological Analysis of Patterns*. PhD thesis, Department of Mathematics, Georgia Institute of Technology, 2005.

[30] S. Eilenberg and S. MacLane, "Acyclic models," *American Journal of Mathematics*, vol. 75, no. 1, pp. 189–199, 1953.

[31] F. H. Croom, *Basic concepts of algebraic topology*. Undergraduate Texts in Mathematics, Springer-Verlag, New York, 1978.

[32] A. Hatcher, *Algebraic Topology*. Cambridge University Press, Cambridge, 2001.

[33] M. K. Chung, P. Bubenik, and P. T. Kim, "Persistence diagrams of cortical surface data," *Information Processing in Medical Imaging*, vol. 21, pp. 386–397, 2009.

[34] K. Xia, X. Feng, Y. Tong, and G. W. Wei, "Persistent homology for the quantitative prediction of fullerene stability," *Journal of Computational Chemistry*, vol. 36, no. 6, pp. 408–422, 2014.

[35] T. Nakamura, Y. Hiraoka, A. Hirata, E. G. Escolar, and Y. Nishiura, "Persistent homology and many-body atomic structure for medium-range order in the glass," *Nanotechnology*, vol. 26, no. 30, 2015.

[36] K. Yamamoto, "Topological analysis of rough surfaces using persistent homology," *Journal of the Physical Society of Japan*, vol. 84, no. 11, pp. 113001–1–4, 2015.

[37] M. Gameiro, K. Mischaikow, and W. Kalies, "Topological characterization of spatial-temporal chaos," *Physical Review E.*, vol. 70, no. 3, 2004.

[38] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer, "Stability of persistence diagrams," *Discrete and Computational Geometry*, vol. 37, no. 1, pp. 103–120, 2007.

[39] P. Dlotko and T. Wanner, "Topological microstructure analysis using persistence landscapes." `https://hal.archives-ouvertes.fr/hal-01259127/document`, 2016.

[40] E. Munch, P. Bendich, K. Turner, S. Mukherjee, J. Mattingly, and J. Harer, "Probabilistic frechet means and statistics on vineyards." arXiv: 1307.6530.

[41] P. Bubenik, "Statistical topological data analysis using persistence landscapes," *Journal of Machine Learning Research*, vol. 16, pp. 77–102, 2015.

[42] A. C. Rencher and W. F. Christensen, *Methods of Multivariate Analysis*. Wiley Series in Probability and Statistics, Wiley, New York City, 3rd ed., 2012.

[43] T. F. Cox and M. A. A. Cox, *Multidimensional Scaling*, vol. 88 of *Chapman and Hall/CRC Monographs on Statistics and Applied Probability*. Chapman and Hall/CRC, London, 2nd ed., 2000.

[44] U. Bauer, M. Kerber, J. Reininghaus, and H. Wagner, *Mathematical Software – ICMS 2014: 4th International Congress, Seoul, South Korea, August 5-9, 2014. Proceedings*, ch. PHAT – Persistent Homology Algorithms Toolbox, pp. 137–143. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014.

[45] K. Mischaikow and V. Nanda, "Morse theory for filtrations and efficient computation of persistent homology," *Discrete and Computational Geometry*, vol. 50, no. 2, pp. 330–353, 2013.

[46] B. T. Fasy, J. Kim, F. Lecci, C. Maria, and V. Rouvreau, "TDA: Statistical tools for topological data analysis." https://cran.r-project.org/web/packages/TDA/index.html, 2015.