Inference on a Time Series of Images Using Topological Data Analysis

by

Susan Glenn

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2024

Date of final oral examination: 08/19/2024

The dissertation is approved by the following members of the Final Oral Committee:

Jun Zhu, Professor, Statistics
Jessi Cisewski-Kehe, Assistant Professor, Statistics
William M. Bement, Professor, Integrative Biology
Hyunseung Kang, Associate Professor, Statistics
Keith Levin, Assistant Professor, Statistics

To my parents, Kirsta and Andy

To my sisters, Abby, Josie, and Sammy

To my grandparents, Gram, Bobob, Grandma, and Grandpa

ACKNOWLEDGMENTS

I would like to thank my two advisors, Jun Zhu and Jessi Cisewski-Kehe, although words cannot describe how much I truly love and appreciate both their kindness and dedication to helping me finish my PhD. You both taught me how to do research with your wisdom and patience. Thank you for helping me and for being in my life! I hope that one day I can be as amazing advisors as you were.

Thank you to Hyunseung Kang and Keith Levin for serving on my committee and always offering me the best advice and encouragement. You both helped me well before you joined my committee. Thank you to William Bement for bringing me this data which changed the course of my thesis and for being such an amazing collaborator.

Next, I would like to thank Dr. West, Michelle Hornung, Lisa Dussault, Dr. Anderson, Dr. Bozzuto, and all the numerous health care workers at the Carbone Cancer Center for saving my life. You showed me that even in darkness we are never alone and there is always room to hope. Thank you Michelle Hornung for your warmth and insight. I would especially like to thank Dr Malinda West for her dedication and compassion; you truly were a light in the darkness.

And finally, I would like to thank all of my friends for their kindness and unwavering support. I especially would like to thank Margaret Turner, Ajinkya Kokandakar, Tanaya Purohit, Bethany Van Houten, and Hailey

Bruzzone for keeping me sane and always making me laugh. Thank you to Manjusha Kancharla for bringing so much happiness and joy into my life. Thank you to all of my family members who have loved and supported me throughout my whole life. And thank you to my parents for always being there for me. I couldn't ask for better parents, and I love you both more than words can say. Thank you to my sisters, forever and always my best friends and three parts of my soul.

The person who walked into the PhD is not the same person who walked out of the PhD, and isn't that the most remarkable thing of all.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

INFERENCE ON A TIME SERIES OF IMAGES USING TOPOLOGICAL DATA ANALYSIS

Susan Glenn

Under the supervision of Professors Jun Zhu and Jessi Cisewski-Kehe
At the University of Wisconsin-Madison

Topological data analysis (TDA) uses persistent homology to quantify loops and higher-dimensional holes in data, making it particularly relevant for examining the characteristics of cell images in the field of cell biology. In the context of a cell injury, as time progresses, a wound in the form of a ring emerges in the cell image and then gradually vanishes. Performing statistical inference on this ring-like pattern in a single image is challenging due to the absence of repeated samples. This dissertation contributes to the literature on topological inference, with a specific focus on the analysis of image data.

First, we introduce a novel TDA-based framework to estimate underlying structures within individual images and quantify associated uncertainties through confidence regions. Our proposed method partitions the image into the background and the damaged cell regions. Then pixels within the affected cell region are used to establish confidence regions in the space of persistence diagrams (topological summary statistics). This method corrects biases in the estimates of topological features on persistence diagrams that are common in traditional TDA approaches. A simulation study is conducted to evaluate the coverage probabilities of the proposed confidence regions in comparison to an alternative approach also presented in this dissertation.

Additionally, we develop two hypothesis testing frameworks to distinguish signal from noise on a persistence diagram of an image. Both frameworks adapt the traditional permutation test to accommodate a persistence diagram as a summary statistic. Another key contribution of this dissertation is extending the methodology for analyzing patterns in a single image to an image evolving throughout time. This extension uses higher-dimensional topological features to establish temporal connections among lower-dimensional features, enabling the tracking of evolving topological features over time.

Keywords: Confidence Regions, Image Processing, Pattern Detection, Permutation Test, Topological Data Analysis, Uncertainty Quantification

Jun Zhu and Jessi Cisewski-Kehe

1.1 Introduction

Ring-like patterns are ubiquitous in biology, being evident during cell division (Pollard and O'Shaughnessy, 2019), development (Haglund et al., 2019), and the response of immune cells to challenges (Herron et al., 2022), to name a few examples. Of particular interest here are the rings of proteins that form around wounds made in single cells as part of the healing response (Mandato and Bement, 2001); an example of these patterns can be seen in Figure 2.5. Such rings close over the wound site, healing the damage, and manipulations that disrupt healing typically alter the organization of the rings (Burkel et al., 2012). Currently, assessments of wound ring disorganization are largely subjective, or are based on simple comparisons of features like aspect ratios, rather than any metric of underlying ring pattern quality. The purpose of this document is to develop a statistical method to objectively identify rings and quantify their associated uncertainty.

Topological data analysis (TDA) provides a framework for the quantification of the global shape of data. For the wounded cell example, TDA can quantify the pattern of an image by representing each detected ring as a loop on a two-dimensional persistence diagram. However, statistical inference requires addressing the uncertainty of these estimates. Direct inference on persistence diagrams is challenging due to their complex multivariate, multidimensional structure, where even averages are not necessarily unique (Mileyko et al., 2011; Turner et al., 2014). Existing techniques in topological inference include resampling procedures, randomization tests, distance functions, density estimation, and spatial point processes; however, many of these statistical methods are applied to other types of data outside of a single image or are designed to compare multiple groups (Robinson and Turner, 2017; Rabadan and Blumberg, 2019; Boissonnat et al., 2018; Abdallah et al., 2023).

TDA has been applied to analyze a wide range of image processing problems. Much of the current literature is dedicated to machine learning tasks, such as classification or prediction, typically involving multiple images (e.g., Singh et al. 2023; Skaf and Laubenbacher 2022; Bukkuri et al. 2021). Applications of TDA for inference in image analysis typically involve either multiple images of a single subject or comparisons between two distinct groups (e.g., Chung et al. 2009; Wang et al. 2023; Singh et al. 2023). When a single image is examined, the focus is often on extracting topological features without addressing statistical inference (e.g., Singh et al. 2023; Gupta et al. 2023). Notably, there is an existing method for inference on topological features extracted from point cloud data (Fasy et al., 2014). There is even less literature focusing on TDA applied to a time series of a single image, with most time series TDA tasks, again, centered on prediction, signal processing, and classification of point cloud data or networks (Gholizadeh and Zadrozny, 2018; Ravishanker and Chen, 2019).

Overall, there is a dearth of TDA methodology in the context of an image of a ring in a living system, as many existing methods are either designed for data types other than images, focus on multiple images, or address tasks other than inference.

A primary challenge—and a unique contribution of this dissertation—is in accurately quantifying the uncertainty associated with the ring's topological features. To address this central challenge, we develop a new method for constructing confidence regions for the persistence diagram of a single image in Chapter 2. Our focus is specifically on persistence diagrams due to their capability to discriminate and perform inference on individual topological features. The proposed method uses segmentation, dividing the image into contiguous regions, which are subsequently matched to corresponding loops identified in the persistence diagram. These matched loops serve as the basis for estimating the shapes within the underlying pattern such as rings in the case of the current application. The confidence regions built for each matched loop are derived by analyzing the pixel distribution within each partition. This method provides unbiased estimates and asymptotic confidence regions with accurate coverage probabilities. In addition, we extend the method in Fasy et al. (2014) from point clouds to images as an alternative to compare against our method. The proposed method allows for inference on the persistence diagram of a single image which yields a simple intuitive interpretation and is computationally efficient, whereas traditional methods in TDA are limited in this setting. While motivated by the wounded cell application, this proposed method generalizes to settings with a single image characterized by one or more loops.

In Chapter 3, we introduce significance testing on a persistence diagram. Two separate tests are introduced in order to find statistically significant topological features on a persistence diagram of a single image, which notably the confidence regions in Section 2.1 are unable to do. Building on similar approaches by Robinson and Turner (2017); Abdallah et al. (2023), we use a variation of the permutation test to perform topological inference which focuses on the case of a single persistence diagram of an image, or M-dimensional array. One of the hypothesis tests helps identify statistically significant features for the method described in Section 2.1, thereby providing a more rigorous framework for inference. The second test has a broader scope outside of the method in Section 2.1 and will be applied, as an example, to the methodology in Chapter 4. In Section 4.2, we extend the analysis to incorporate time more directly; such that, a method is created to track a pattern in a single image as the pattern evolves throughout time. In the next section, Section 1.2, we provide a background on TDA and explain how TDA can be applied to analyze the shape of images.

1.2 Topological Data Analysis and Persistence Diagrams

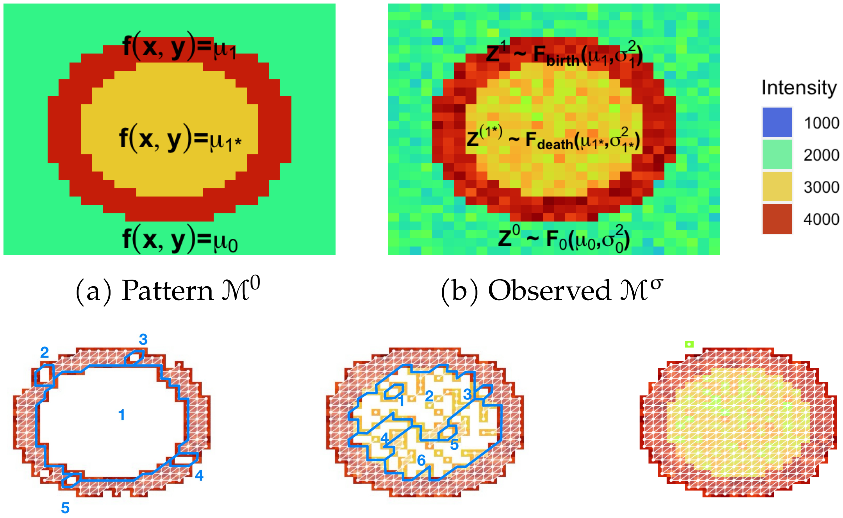
This section introduces key principles used in TDA and their application to data in the context of images. First, concepts in algebraic topology, such as persistent homology, are described. Then the focus is on how to characterize the intrinsic shape and structure of an image and represent this information on a persistence diagram.

TDA uses ideas from algebraic topology and computational geometry to extract meaningful insights and patterns from data. In particular, persistent homology is used to quantify the shape of a dataset through identifying holes in the space and determining their number, strength (through persistence), and dimension. Viewing shape through this perspective of connectivity and continuity, topological features are used to characterize a space.

Homology associates algebraic structures, called homology groups, with topological spaces. These groups $H_m(X)$, where m represents the homology group dimension, can be thought of as characterizing a topological space X by the number of connected components (the number of zero-dimensional homology group generators, $H_0(X)$), the number of loops (the number of one-dimensional homology group generators, $H_1(X)$), and the number of voids (the number of two-dimensional homology group generators, $H_2(X)$) in X (Chazal and Michel, 2021; Edelsbrunner and Harer, 2010). When $m \geqslant 3$, $H_m(X)$ corresponds to higher dimensional holes in X. In Chapter 2, we restrict our focus to the first homology group (H_1) since the interest is in the loops, or rings, found in images such as in Figure 2.5. Chapters 3 and 4 expand to focusing on higher dimensional homology groups, along with loops and connected components. Persistent homology tracks the evolution of these homology groups across various scales. (Otter et al., 2017; Edelsbrunner and Harer, 2010).

When the topological space is an image $\mathcal{M}$, the scales can refer to the intensity values of pixels $Z(x,y)$ where the (x,y) coordinates represent the locations of the center of the pixels in the image. Homology groups at different intensities are computed from a triangulation on the upper-level sets of the image, defined as $\mathcal{M}^{-1}(\delta,\infty) = \{(x,y) \in \mathbb{R}^2 | Z(x,y) > \delta\}$ where δ is the threshold for intensity values (Chazal and Michel, 2021). This triangulation breaks down the space into simplices—geometric elements on which the computations are carried out. A simplicial complex $\mathcal{K}$ is a set composed of zero-simplices (points), one-simplices (line segments), and two-simplices (triangles), and three-simplices (tetrahedral), such that (i) any face of a simplex of $\mathcal{K}$ is also a simplex in $\mathcal{K}$, and (ii) the intersection of any two simplices in $\mathcal{K}$ is a face of both simplices or empty. Let V be the set of points ((x,y)-coordinates) and K be the set of line segments and triangles which make up $\mathcal{K}$. When a pixel is in $\mathcal{M}^{-1}(\delta,\infty)$ the triangulation puts a zero-simplex at the pixel center and connects each zero-simplex to neighboring zero-simplices by one-simplices. The pairwise connection of three zero-simplices form a two-simplex (Chazal and Michel, 2021; Otter et al., 2017).

Figure 1.1 shows several examples of simplicial complexes built on upper-level sets of the data along with the correct segmentation of the data and the underlying pattern from which the data were generated (e.g., partitions an image into background and manifold(s), details are discussed in Section 2.1). As the threshold parameter δ decreases from positive infinity to zero, the space becomes more connected, capturing the homology of each simplicial complex. While δ varies, a *filtration* is formed by a finite sequence of nested sub-complexes $\mathcal{K}_{\delta_1} \subset \mathcal{K}_{\delta_2} \subset \ldots \subset \mathcal{K}_{\delta_l} = \mathcal{K}$. Figures 1.1c-1.1e illustrate different $\mathcal{K}_\delta$ on the upper-level sets in a filtration of $\mathcal{M}$. The 'birth time' b of a loop, is the value of δ when it first appears in the filtration (e.g., Figure 1.1c), and its 'death time' d is the value at which it merges with another feature (e.g., Figure 1.1e). Persistence, defined as the feature's lifetime (persistence $= b - d$), can be interpreted as longer lifetimes indicate topological signal and and shorter lifetimes indicate topological noise (Fasy et al., 2014).



$f(x, y)=\mu_1$
$f(x, y)=\mu_{1*}$
$f(x, y)=\mu_0$
$Z^1 \sim F_{birth}(\mu_1,\sigma_1^2)$
$Z^{(1*)} \sim F_{death}(\mu_{1*},\sigma_{1*}^2)$
$Z^0 \sim F_0(\mu_0,\sigma_0^2)$
Intensity
1000
2000
3000
4000
1
2
3
4
5
1
2
3
4
5
6

(a) Pattern $\mathcal{M}^0$

(b) Observed $\mathcal{M}^\sigma$

(c) δ=3778 (birth)

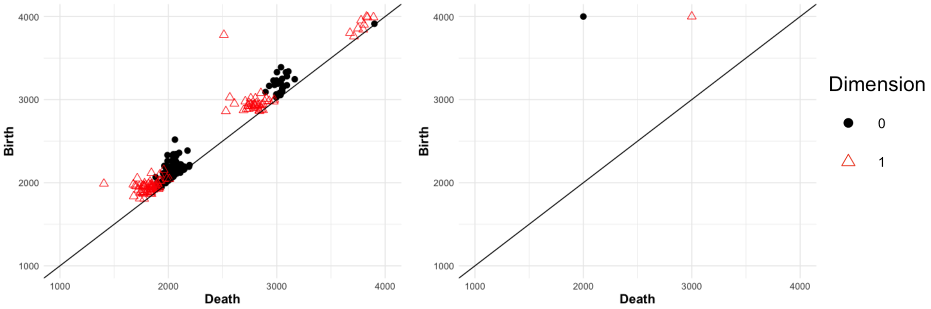
(d) δ=3000

(e) δ=2512 (death)

Figure 1.1: (a) Partitions of the underlying pattern into background μ_0, loop μ_1, and interior of the loop μ_{1*}. (b) Partitions of the data into into background, the loop, and the interior of the loop. Z^k where $k = \{0, 1, 1*\}$ represents a pixel intensity value and each F_k is a distribution from which the pixel was drawn. (c) The simplicial complex $\mathcal{K}_{3778}$, built on $(\mathcal{M}^\sigma)^{-1}(3778, \infty)$, contains one connected component and five loops. The birth time of the true loop (1) is 3778. (d) $\mathcal{K}_{3000}$ contains five connected component and six loops. (e) $\mathcal{K}_{2512}$ contains two connected component and no loops, where 2512 is the death time of the large loop (1) born at 3778.

The evolution of the homology groups of $\mathcal{M}$ over the course of the filtration is graphically represented on a persistence diagram $\mathcal{P}(\mathcal{M})$. Figure 1.2 shows an example of a persistence diagram of the data (Figure 1.1b) compared to the persistence diagram of the underlying pattern (Figure 1.1a) from which the data were generated. Features of each dimension, such as connected components and loops are represented in the diagram by displaying the death and birth times as (x,y) coordinates. Each homology group, is represented by a shape and color: connected components are black dots and loops are red triangles. The number of red triangles in each diagram is the number of loops detected in the upper-level set filtration for an image. The more persistent loops are farther from the diagonal line $y = x$ (i.e. birth=death).

In the persistence diagram for the data (Figure 1.2a), the birth time of the most persistent loop is 3778 and the death time is 2512, both of these are estimates of the birth and death time of the corresponding loop in the underlying pattern. All the other loops which are closer to the diagonal are small loops which are just due to noise. In the persistence diagram of the underlying pattern (Figure 1.2b) there is only one loop detected with a birth time of 4000 and a death time of 3000.



(a) Observed $\mathcal{P}(\mathcal{M}^\sigma)$ (b) Pattern $\mathcal{P}(\mathcal{M}^0)$

Figure 1.2: (a) The persistence diagram of the data in Figure 1.1b with loops (red triangles) and connected components (black dots). (b) The persistence diagram of the underlying pattern in Figure 1.1a which has only one loop and one connected component.

In the context of our application, a persistence diagram may be viewed as an estimate of the underlying pattern of $\mathcal{M}$, where a different realization of the image for the same data-generating process generally results in a different persistence diagram. The number of loops, and their corresponding birth and death times, can be an estimate of the pattern of the ring structure. In the next Chapter, we outline a method to get uncertainty estimates for the birth and death times of the loops found in the data which allows for inference on the true persistence diagram Figure 1.2b from the observed persistence diagram Figure 1.2a.

2 CONFIDENCE REGIONS FOR A PERSISTENCE DIAGRAM OF A SINGLE IMAGE

2.1 Setup

In this section, we develop a method to assess the uncertainty in the estimated persistence diagram by constructing confidence regions around the birth and death times of the elements in $H_1(\mathcal{M}^\sigma)$, the generators of the one-dimensional homology groups (i.e., loops). These confidence regions should cover the one-dimensional homology group generators of the persistence diagram of the noiseless true manifold $\mathcal{M}^0$. However, as is demonstrated in Section 2.3, there is considerable bias in the estimated birth and death times of loops using upper-level set filtrations for a raw image, which we refer to as the *traditional TDA* estimates.

An approach for reducing the influence of outliers when estimating persistence diagrams for point-cloud data uses upper-level set filtrations on kernel density estimates or regression models of the data, rather than a different type of filtration (e.g., a Vietoris-Rips filtration) on the point-cloud data directly (Chazal and Michel, 2021; Fasy et al., 2014). This technique is used in Fasy et al. (2014) to construct confidence regions on persistence diagrams for point-cloud data.

Since the confidence regions are centered around the estimated birth and death times, we need to obtain unbiased estimates of the birth and death times of loops in images. One possible approach, outlined in Section 2.7, is to estimate a smoother function of the image and doing an upper-level set filtration extending the inference approach in Fasy et al. (2014) from point-cloud data to a single image. We refer to this proposed extension as *smooth TDA*. This is used as a comparison to our primary proposed approach which we refer to as *partitioned TDA*. The partitioned TDA method provides unbiased estimates without smoothing, and is presented in detail below in Section 2.2.

Let the image $\mathcal{M}$ be defined by some function $f(x, y)$ discretized onto a 2D grid $\mathcal{G}_{d_1 \times d_2}$, where each (x,y) coordinate represents the grid columns $x = \{1, 2, \ldots, d_1\}$ and grid rows $y = \{1, 2, \ldots, d_2\}$. The true pattern is the noiseless image $\mathcal{M}^0 = \{f(x, y) : (x, y) \in \mathcal{G}\}$. However, in practice there is some zero-centered noise $\varepsilon(x, y)$ drawn from distribution $\mathbf{F}(0, \sigma^2(x, y))$ added to the function so that $\mathcal{M}^\sigma = \{f(x, y) + \varepsilon(x, y) : (x, y) \in \mathcal{G}\}$ with the σ as the exponent of $\mathcal{M}^\sigma$ indicating there is noise in the image. Each grid value, or pixel, in $\mathcal{M}^\sigma$ has intensity $Z(x, y)$ drawn from:

$$Z(x, y) \sim \mathbf{F}(f(x, y), \sigma^2(x, y)), \tag{2.1}$$

where the mean is defined by $\mathcal{M}^0$ and the error is defined by ε.

In this work, the following assumptions are made regarding the topological features of the noise-free image, $\mathcal{M}^0$, which are estimated from the topological features of its noisy counterpart, $\mathcal{M}^\sigma$. The proposed method, partitioned TDA, involves partitioning the image in a way that distinguishes the background and n_p other topological structures (e.g., loops and the interior of loops).

Assumption 1. *The true image $\mathcal{M}^0$ can be segmented into contiguous regions with constant functional values: $f(x,y) = \mu_k \ \forall \ (x,y)$ within partition $\mathcal{G}_k$. Image $\mathcal{M}^\sigma$ can be segmented into $n_p + 1$ contiguous regions where each region is defined as $\mathcal{M}_k^\sigma = \{f(x,y) + \varepsilon(x,y) : (x,y) \in \mathcal{G}_k\}$ for $k = \{0, \dots, n_p\}$ where $\mathcal{G}_k = \{(x,y) \in \mathcal{G} : f(x,y) = \mu_k\}$.*

Assumption 2. *If the true image, $\mathcal{M}^0$, has at least one feature that is homeomorphic to a one-sphere (loop), let n_1 denote the number of one-spheres. Any partition of $\mathcal{M}^0$ that is homeomorphic to a one-sphere has pixel intensities fixed at $f(x,y) = \mu_i$ for $i = \{1, \dots, n_1\}$ where $2n_1 \leqslant n_p$, and the partition interior to this one-sphere has pixel intensities fixed at $f(x,y) = \mu_{i*}$. Let μ_0 be designated as the mean of the background noise partition (if it exists).*

Assumption 3. *For an upper-level set filtration assume for the majority of $i = \{1, \dots, n_1\}$ that $\mu_i \geqslant \mu_{i*}$ and $\mu_i \geqslant \mu_0$.*

If all the inequalities from Assumption 3 are $\geqslant$, for a given setting, then an upper-level set filtration is sufficient. However, depending on how many $\mu_i \leqslant \mu_{i*}$, a lower-level set filtration may capture the topological features more effectively.

In Section 2.2, we explain how partitioned TDA constructs confidence regions for an image with a single H_1 feature (i.e., loop) so that $n_p =$

3 (background, H_1 feature, and the region interior to the H_1 feature). Section 2.5 generalizes this method to handle multiple H_1 features.

2.2 Confidence Regions for a Single Image with a Single H_1 Feature

Here we consider the setting with a single loop in $\mathcal{M}^0$. Assumptions 1 and 2 imply that $\mathcal{M}^0$ can be segmented into three contiguous regions where the background region is defined as $\mathcal{M}^0_0 = \{\mu_0 : (x, y) \in \mathcal{G}_0\}$, the part of the image homeomorphic to a one-sphere is defined as $\mathcal{M}^0_1 = \{\mu_1 : (x, y) \in \mathcal{G}_1\}$, and part of the image that is interior to this one-sphere is defined as $\mathcal{M}^0_{1*} = \{\mu_{1*} : (x, y) \in \mathcal{G}_{1*}\}$. For Sections 2.2-2.5, we assume the true partitions $\mathcal{G}_0$, $\mathcal{G}_1$, and $\mathcal{G}_{1*}$ are known. However, in practice the true partitions are unknown and segmentation is used to estimate each $\mathcal{G}_k$. Section 2.6 proposes an algorithm for reducing the bias in the confidence region coverage due to the misclassification of pixels in an estimated segmentation.

Using the known partitions, the data $\mathcal{M}^\sigma$ can be separated into three distributions from which pixels are drawn ($\mathcal{M}^\sigma_0$, $\mathcal{M}^\sigma_1$, $\mathcal{M}^\sigma_{1*}$) as defined in

Assumption 1:

$$\begin{aligned}\mathcal{M}_0^{\sigma} &\text{ is the background partition where } Z^0 \sim \mathbf{F_0}(\mu_0, \sigma_0^2)\\ \mathcal{M}_1^{\sigma} &\text{ is the part homeomorphic to a one-sphere where } Z^1 \sim \mathbf{F_{birth}}(\mu_1, \sigma_1^2)\\ \mathcal{M}_{1*}^{\sigma} &\text{ is the part interior to the one-sphere where } Z^{1*} \sim \mathbf{F_{death}}(\mu_{1*}, \sigma_{1*}^2)\end{aligned} \tag{2.2}$$

The loop in the true pattern, of which we are trying to estimate its birth and death times, has a birth time of μ_1 determined by $\mathcal{M}_1^0$ and a death time of μ_{1*} determined by $\mathcal{M}_{1*}^0$, as shown in Figure 1.1a. In order to make the confidence regions, we define the joint distribution of the sample means of the pixel intensities associated with the birth and death times, $\bar{Z}^1$ and $\bar{Z}^{1*}$, respectively, as follows:

$$\mathbf{X} = \begin{pmatrix} \bar{Z}^{1*} \\ \bar{Z}^1 \end{pmatrix} \overset{\text{approx}}{\sim} \left(\begin{pmatrix} \mu_{1*} \\ \mu_1 \end{pmatrix}, \begin{pmatrix} \frac{\sigma_{1*}^2}{n_d} & 0 \\ 0 & \frac{\sigma_1^2}{n_b} \end{pmatrix} \right) \tag{2.3}$$

where n_d and n_b are the number of pixels in $\mathcal{M}_{1*}^{\sigma}$ and $\mathcal{M}_1^{\sigma}$, respectively. By the Central Limit Theorem, $\mathbf{X}$ approximately follows a bivariate normal distribution allowing for a confidence region to be created based on: $(\bar{\mathbf{X}} - \boldsymbol{\mu})^{\mathbf{T}} \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}) \sim \chi_2^2$. The asymptotic confidence region for the birth and death times of $\mathcal{M}^0$ is as follows:

$$\mu(\theta) = \bar{\mathbf{X}} + \sqrt{\chi^2_{2,\alpha}}\sqrt{\hat{\boldsymbol{\Sigma}}}\begin{bmatrix}\cos(\theta)\\ \sin(\theta)\end{bmatrix} \text{ for } 0 < \theta < 2\pi \tag{2.4}$$

where the variance can be estimated by sample variance.

The segmentation of $\mathcal{M}^\sigma_k$ for $k = \{0, 1, 1*\}$ creates the confidence regions in Equation (2.4) and the unbiased estimators for $(\mu_{1*}, \mu_1) : (\bar{Z}_{1*}, \bar{Z}_1)$. However, these unbiased estimates are not derived from an upper-level set filtration on $\mathcal{M}^\sigma$. This approach for generating confidence regions is called partitioned TDA; next we describe the bias in traditional TDA methods.

2.3 Bias in Traditional TDA Birth and Death Times

The level of bias in traditional TDA birth times depends on the proportion of vertices in the simplicial complex that comprise the birth of the loop relative to the set of pixels associated with the corresponding true loop pattern. A similar bias is found with traditional TDA death times and the relationship between the structure of the simplicial complex and the interior of the true pattern. A more technical explanation is provided next.

Assumption 3 states that $\mu_1 \geqslant \mu_0$ and $\mu_1 \geqslant \mu_{1*}$. When applying an upper-level set filtration to $\mathcal{M}^\sigma$, a number of loops can be identified along with the associated birth and death times $\{(d_1, b_1), \ldots, (d_j, b_j), \ldots (d_{\beta_1}, b_{\beta_1})\}$. Let β_1 be the total number of loops detected and (d_j, b_j) be the traditional TDA birth and death times for the loop $\mathcal{M}_1^{\sigma}$ (topological signal). All other birth and death times are topological noise and not a part of the true pattern $\mathcal{M}^0$. The birth time, b_j, is the largest δ value in the filtration when the loop in $\mathcal{M}_1^{\sigma}$ first appears in the simplicial complex $\mathcal{K}_{b_j} = \{V_{b_j}, K_{b_j}\}$. The part of the simplicial complex, $\mathcal{K}_{b_j}$, that comprises the birth of the loop is defined as follows:

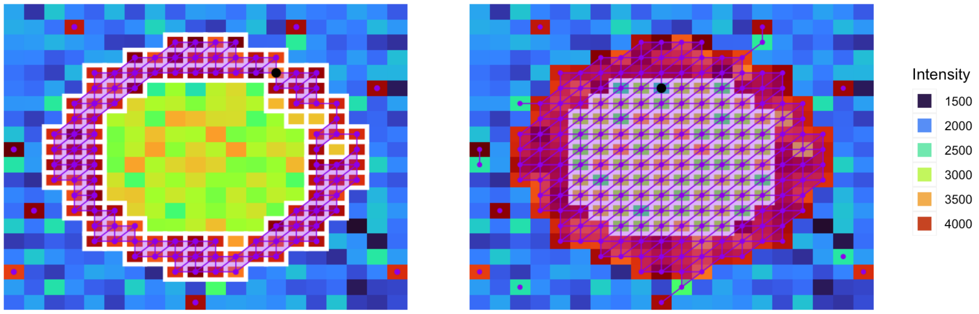
$$\mathcal{K}_{\text{birth}} = \{V_{\text{birth}}, K_{\text{birth}}\} \subseteq \mathcal{K}_{b_j} \text{ and } V_{\text{birth}} \subseteq \mathcal{G}_1. \tag{2.5}$$

Similarly, the death time, d_j, is the largest δ value in the filtration when the loop in $\mathcal{M}_1^{\sigma}$ disappears in the simplicial complex $\mathcal{K}_{d_j} = \{V_{d_j}, K_{d_j}\}$. The part of the simplicial complex, $\mathcal{K}_{d_j}$, that makes up the interior of the loop is defined as follows:

$$\mathcal{K}_{\text{death}} = \{V_{\text{death}}, K_{\text{death}}\} \subseteq \mathcal{K}_{d_j} \text{ and } V_{\text{death}} = \mathcal{G}_{1*}. \tag{2.6}$$

Figure 2.1 illustrates the difference between b_j, d_j, $\mathcal{K}_{b_j}$, $\mathcal{K}_{\text{birth}}$, $\mathcal{K}_{d_j}$, $\mathcal{K}_{\text{death}}$, $\mathcal{M}_1^{\sigma}$, and $\mathcal{M}_{1*}^{\sigma}$. The white rectangles in each subfigure outline pixels located in $\mathcal{G}_1$ with intensity values in $\mathcal{M}_1^{\sigma}$ (Figure 2.1a) or pixels located in $\mathcal{G}_{1*}$ with intensity values in $\mathcal{M}_{1*}^{\sigma}$ (Figure 2.1b), the total purple simplicial complexes are either $\mathcal{K}_{b_j}$ (Figure 2.1a) or $\mathcal{K}_{d_j}$ (Figure 2.1b), while the part of the purple simplicial complexes within the white rectangles are either $\mathcal{K}_{\text{birth}}$

(Figure 2.1a) or $\mathcal{K}_{\text{death}}$ (Figure 2.1b). The black zero-simplex is the location of the pixel which has intensity b_j (Figure 2.1a) or d_j (Figure 2.1b). Note that any of the white rectangles beneath the purple simplicial complex appear light purple.



(a) $\mathcal{K}_{b_j}$ on $(\mathcal{M}^{\sigma})^{-1}(b_j, \infty)$ (b) $\mathcal{K}_{d_j}$ on $(\mathcal{M}^{\sigma})^{-1}(d_j, \infty)$

Figure 2.1: Illustration of simplicial complexes (purple) built on the upper-level sets at the birth time and death times of an image with a loop, with $\mu_1 = 4000$ and $\mu_{1*} = 3000$. (a) The simplicial complex $\mathcal{K}_{b_j} = \mathcal{K}_{3597}$ at the birth of the loop where the black dot is the pixel with intensity value equal to 3597, which is the upper-level set threshold associated with the birth of the loop. The white rectangles indicate the pixels of $\mathcal{M}_1^{\sigma}$. (b) The simplicial complex $\mathcal{K}_{d_j} = \mathcal{K}_{2593}$ at the death of the loop where the black dot is the pixel with intensity value equal to 2593, which is the upper-level set threshold associated with the death of the loop. The white rectangles, which indicate the pixels of $\mathcal{M}_{1*}^{\sigma}$, appear light purple due to the overlaying two-simplices in $\mathcal{K}_{d_j}$.

The level of bias in the estimate of b_j using traditional TDA depends on the proportion between the number of vertices in the set $\mathcal{K}_{\text{birth}}$ and the number of elements in the set $\mathcal{G}_1$, represented by p_b. According to Equation (2.5), where $Z(V_{\text{birth}}) \subseteq \mathcal{M}_1^{\sigma}$ and $Z^1 \sim \mathbf{F_{birth}}(\mu_1, \sigma_1^2)$, the proportion p_b

is defined as follows:

$$p_b = 1 - \frac{|V_{\text{birth}}|}{|\mathcal{M}_1^{\sigma}|} = 1 - \frac{|V_{\text{birth}}|}{n_b}, \tag{2.7}$$

where $|X|$ is the cardinality of the set X.

Since the birth time is the minimum intensity values of all the pixels $Z^1(x, y)$ where $(x, y) \in V_{\text{birth}}$, then $\mathbf{F}^{-1}_{\mathbf{birth}}(p_b) = b_j$. The bias in the birth time is:

$$\text{Bias}(\mu_1, b_j) = \mu_1 - E\{\mathbf{F}^{-1}_{\mathbf{birth}}(p_b)\}. \tag{2.8}$$

The birth time is unbiased if b_j falls within the 50th percentile of all pixels comprising loop $\mathcal{M}^1$, given that $\mathbf{F_{birth}}$ is a symmetric distribution.

As an example, the number of vertices in $\mathcal{K}_{b_j} = \mathcal{K}_{3597}$ in Figure 2.1a is $V_{b_j} = 109$ and the number of vertices in $\mathcal{K}_{\text{birth}}$ is $V_{\text{birth}} = 101$. The number of pixels which make up the loop $\mathcal{G}_1$ is $n_b = 106$. Therefore, $p_b \approx 0.05$ and the birth time $b_j = 3597$ is in the 5th percentile of all the pixels which make up the loop. The bias of this estimate with a true birth time $\mu_1 = 4000$ is $4000 - 3597 = 403$.

The level of bias of d_j (using traditional TDA) depends on the proportion between the number of vertices in the set $\mathcal{K}_{\text{death}}$ and the number of elements in the set $\mathcal{G}_{1*}$, denoted by p_d. Based on the Assumptions in Section 2.1, all the pixels which make up the interior of the loop are a part of the simplicial complex at the death of the loop. From Equation (2.6), $Z(V_{\text{death}}) \subseteq \mathcal{M}^{\sigma}_{1*}$ and $Z^{1*} \sim \mathbf{F_{death}}(\mu_{1*}, \sigma^2_{1*})$ and consequently, the proportion p_d is:

$$p_d = 1 - \frac{|V_{\text{death}}|}{|\mathcal{M}^{\sigma}_{1*}|} = 1 - \frac{n_d}{n_d} = 0. \tag{2.9}$$

Then $\mathbf{F}^{-1}_{\mathbf{death}}(0) = \min(Z^{1*}) = d_j$ where the bias in the estimate is:

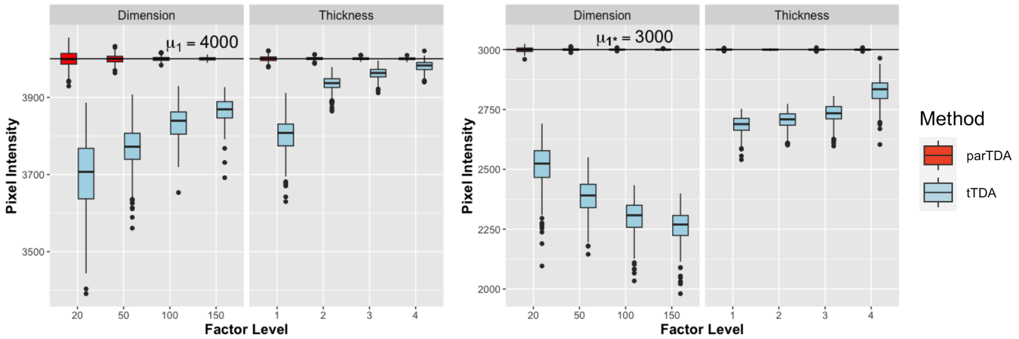
$$\text{Bias}(\mu_{1*}, d_j) = \mu_{1*} - E\{\min(Z^{1*})\}. \tag{2.10}$$

Therefore, the death time is an unbiased estimator of μ_{1*} when there is only one pixel which makes up $\mathcal{M}^{\sigma}_{1*}$ since $E(\min(Z^{1*})) = E(Z^{1*}) = \mu_{1*}$.

Simulations Testing Bias in Birth and Death times

Figure 2.2 displays results of an exploration of the relationship between bias in traditional TDA estimates of the birth and death times and the construction of the image, compared to the unbiased partitioned TDA estimates. Differences in the image dimension and the area of the partitions ($\mathcal{G}^{\sigma}_1$ and $\mathcal{G}^{\sigma}_{1*}$) change the amount of bias in the traditional TDA estimates of the birth and death times of the loop. Two simulations studies are carried out: (1) considers four different loop thickness levels and (2) considers four different different image dimensions levels. Each factor level for both simulations has 100 iid images generated with one loop ($(\mu_{1*}, \mu_1) = (3000, 4000)$). At each of the loop thickness level, $\{1, 2, 3, 4\}$, the birth and death times of the loop (d_j, b_j) is calculated for each image. Level 1 is for a very thin loop (two pixels thick), level 2 is a medium thin loop (seven pixels thick), level 3 is a medium thick loop (11 pixels thick), and level 4 is for a thick loop (16 pixels thick). Similarly, at each image dimension level, $\{20 \times 20, 50 \times 50, 100 \times 100, 150 \times 150\}$, the birth and death times of the loop (d_j, b_j) are calculated for each image. These results are shown in boxplots in Figure 2.2, where the light blue boxplots are the traditional TDA birth and death times, denoted as tTDA, while the red boxplots are the partitioned TDA birth and death times, denoted as parTDA.

As seen in Figure 2.2, estimates of the birth (Figure 2.2a) and death (Figure 2.2b) times across all different factor levels (dimension and thickness) using partitioned TDA are unbiased. Whereas, estimates of the birth and death times using traditional TDA are biased and this bias changes depending on different factor levels.



(a) Birth times (b_j) (b) Death Times (d_j)

Figure 2.2: Boxplots illustrating estimated birth (a) and death (b) times of loops using partitioned TDA (red) and traditional TDA (blue), based on 100 iid images within each factor level of loop thickness or dimensions. The true birth and death times are indicated by the horizontal solid black lines. The traditional TDA (tTDA) estimates have a strong negative bias with higher variability, while the proposed partitioned TDA (parTDA) estimates appear to be unbiased with lower variability.

When interpreting Figure 2.2, the dimension of the image serves as a proxy for the pixel sample size of the partitions, with higher dimensions indicating larger sample sizes in both $\mathcal{M}_1^\sigma$ and $\mathcal{M}_{1*}^\sigma$. As image dimension increases, the bias in the birth time estimates using traditional TDA decreases as well as the variance. However, for the death time, the bias increases as the image dimensions increase. This result is consistent with the discussion of p_d in Equation (2.9). Loop thickness, which serves as a proxy for examining the area of $\mathcal{M}_1^\sigma$ and $\mathcal{M}_{1*}^\sigma$, shows less bias in both the birth and death times. In general, thicker loops or larger image dimensions (more pixels making up the loop) lead to less biased estimates of the birth time. Thicker loops or smaller image dimensions (fewer pixels making up the inside of the loop) lead to less biased estimates of the death time. In certain situations the traditional TDA estimate, (d_j, b_j) is an unbiased estimator for (μ_{1*}, μ_1) whereas, $(\bar{Z}_{1*}, \bar{Z}_1)$ is unbiased regardless of the way the loop or image is constructed.

2.4 Matching Loops Between Traditional TDA and Partitioned TDA

The partitions $(\mathcal{G}_k)$ used in partitioned TDA to estimate the birth and death times of a loop do not directly use TDA (e.g., there is no assumption that a partition forms a loop). To detect a loop, the unbiased estimates, $(\bar{Z}^{1*}, \bar{Z}^1)$, need to be matched to a corresponding loop detected from traditional TDA, (d_j, b_j), for a loop to be detected with partitioned TDA. Algorithm 1 is designed to identify which of the loops in the $\mathcal{M}^\sigma$, $\{(d_1, b_1), \dots, (d_j, b_j), \dots (d_{\beta_1}, b_{\beta_1})\}$, are in the partitions $\mathcal{G}_{1*}$ and $\mathcal{G}_1$ by the location of the birth and death time pixel intensities. The loops which are not matched to the partitions are not considered to be part of the underlying pattern. Once (d_j, b_j) is matched with the partitions $(\mathcal{G}_{1*}, \mathcal{G}_1)$ using Algorithm 1, the birth and death times of the loop detected with traditional TDA are estimated with $(\bar{Z}_{1*}, \bar{Z}_1)$.

Algorithm 1 Localizing the birth and death times (d_j, b_j)

```
Input: $df \coloneqq (x, y, Z[x, y])$ of image $\mathcal{M}^\sigma$; partitions $\mathcal{G}_k$ for $k = \{0, 1, 1*\}$,
    birth and death times from $\mathcal{P}(\mathcal{M}^\sigma) \coloneqq \{(d_1, b_1), \dots, (d_q, b_q)\}$.
Output: $(d_j, b_j)$ matched to $(\mathcal{G}_{1*}, \mathcal{G}_1)$
Define: $df_k = \{(x, y, Z[x, y]) \in \mathcal{G}_k\}$, $k = \{0, 1, 1*\}$; $out_d = \emptyset$; $out_b = \emptyset$;
    out= $\emptyset$
for l in 1:q do
    Step 1: Find $df_k$ where $Z(x, y) = d_l$  ▷ Identify pixel location of $d_l$ in $\mathcal{G}$
    if $k = 1*$ then $out_d \leftarrow out_d \cup l$  ▷ Only keep index $l$ for $d_l \in \mathcal{G}_{1*}$
    end if
    Step 2: Find $df_k$ where $Z(x, y) = b_l$  ▷ Identify pixel location of $b_l$ in $\mathcal{G}$
    if $k = 1$ then $out_b \leftarrow out_b \cup l$  ▷ Only keep index $l$ for $b_l \in \mathcal{G}_1$
    end if
    Step 3: Calculate out$\leftarrow$$out_d \cap$ $out_b$
    if length(out)==2 then $(d_l, b_l) = (d_j, b_j)$  ▷ If $b_l \in \mathcal{G}_1$ and $d_l \in \mathcal{G}_{1*}$ loop is matched
        Stop  ▷ Match found, stop algorithm
    end if
end for
return out
```

2.5 Confidence Regions for Multiple H_1 Features

Section 2.2 introduced partitioned TDA for the setting with only one loop in $\mathcal{M}^0$, which is the setting of our motivating cell image application presented in Section 2.9. The objective of this section is to explain how the proposed method can be generalized to encompass multiple loops within a single image. While the primary emphasis is on one-spheres, it is worth noting that the methodology can be readily extended to m-spheres for higher-dimensional spaces, such as 3D images.

Assume that there are n_1 loops in $\mathcal{M}^0$ resulting in $2n_1 + 1$ partitions and that the functional value of each loop in $f(x, y)$ is μ_i and the value of the interior of each one-sphere in $f(x, y)$ is μ_{i*} for $i = \{1, \dots n_1\}$. For every loop of $\mathcal{M}^0$, the persistence diagram of the observed image represents each loop as birth death pairs: $(d_{j_1}, b_{j_1}), \dots (d_{j_{n_1}}, b_{j_{n_1}})$. The steps listed in Algorithm 1 can be extended to connect each (d_{j_i}, b_{j_i}) with $(\mathcal{G}_{i*}, \mathcal{G}_i)$ where the partitions $\mathcal{G}_k$ become $k = \{0, i, i*\}$ for $i = \{1, \dots, n_1\}$.

There are three other possible types of birth-death pairs (d_l, b_l) where $l \neq j_i$ for $i = \{1, \dots n_1\}$ detected in the image $\mathcal{M}^\sigma$ which are not loops in $\mathcal{M}^0$:

(1) loops which are in the background $(d_0, b_0 \sim \mathbf{F}_0(\mu_0, \sigma_0^2))$

$$d_0 \notin \mathcal{M}_{i*}^{\sigma} \text{ and } b_0 \notin \mathcal{M}_i^{\sigma} \forall i \neq 0 \implies \text{using Algorithm 1 } (d_0, b_0) \neq (d_{j_i}, b_{j_i}) \tag{2.11}$$

(2) loops which are only in $\mathcal{M}_i$ or only in $\mathcal{M}_{i*}$ $(d_i, b_i \sim \mathbf{F_{birth}}(\mu_i, \sigma_i^2)$ or $d_{i*}, b_{i*} \sim \mathbf{F_{death}}(\mu_{i*}, \sigma_{i*}^2))$

$$d_i, b_i \in \mathcal{M}_i^{\sigma} \implies \text{using Algorithm 1 } (d_i, b_i) \neq (d_{j_i}, b_{j_i}) \tag{2.12}$$

$$d_{i*}, b_{i*} \in \mathcal{M}_{i*}^{\sigma} \implies \text{using Algorithm 1 } (d_{i*}, b_{i*}) \neq (d_{j_i}, b_{j_i}) \tag{2.13}$$

(3) loops that traverse the background and $\mathcal{M}_i^{\sigma}$ $(b_i \sim \mathbf{F_{birth}}(\mu_i, \sigma_i^2)$ and $d_0 \sim \mathbf{F}_0(\mu_0, \sigma_0^2))$

$$d_0 \notin \mathcal{M}_{i*}^{\sigma} \implies \text{using Algorithm 1 } (d_0, b_i) \neq (d_{j_i}, b_{j_i}) \tag{2.14}$$

Since all the loops detected in the segmentation $\mathcal{M}_i^{\sigma}$ are connected to the correct (d_{j_i}, b_{j_i}), the only time a problem would arise is when $d_{j_i} = d_{j_k}$ and $b_{j_i} = b_{j_k}$ for $i \neq k$ where $i, k \in \{1, \dots, n_1\}$. In other words, if the loop $\mathcal{M}_i^{\sigma}$ and the loop in $\mathcal{M}_k^{\sigma}$ have the exact same birth and death times, the algorithm would not be able to match (d_{j_i}, b_{j_i}) and (d_{j_k}, b_{j_k}) with $(\mathcal{G}_{i*}, \mathcal{G}_i)$ and $(\mathcal{G}_{k*}, \mathcal{G}_k)$, respectively. However, this situation would happen with zero probability since all $Z^i \sim \mathbf{F_{birth}}(\mu_i, \sigma_i^2)$, $Z^k \sim \mathbf{F_{birth}}(\mu_k, \sigma_k^2)$ and $Z^{i*} \sim \mathbf{F_{death}}(\mu_{i*}, \sigma_{i*}^2)$, $Z^{k*} \sim \mathbf{F_{death}}(\mu_{k*}, \sigma_{k*}^2)$ are continuous distributions.

2.6 Segmentation of the Image

In the preceding two sections the partitions $\mathcal{G}_k$ for $k = \{0, \dots, n_p\}$ are assumed to be known; whereas in this section, the segmentation is unknown and is estimated with $\hat{\mathcal{G}}_k$ for $k = \{0, \dots, \hat{n}_p\}$. If the segmentation is incorrect the partitioned TDA estimated birth and death times in Equation (2.3) and the corresponding confidence regions in Equation (2.4) may not be accurate. Here, we propose a method to reduce the misclassification of pixels in partitions when one or more of the $\hat{\mathcal{G}}_k$'s may have some incorrect pixels assigned to it.

Recall from Equation (2.2) that if $\mathcal{G}_k$ is known $\forall k \in \{0, \dots, n_p\}$ then interior pixel intensities $Z^{i*} \sim \mathbf{F_{death}}(\mu_{i*}, \sigma^2_{i*})$ for every $Z^{i*} \in \mathcal{M}^{\sigma}_{i*}$ and pattern pixel intensities $Z^{i} \sim \mathbf{F_{birth}}(\mu_{i}, \sigma^2_{i})$ for every $Z^{i} \in \mathcal{M}^{\sigma}_{i}$, where $i \in \{1, \dots, n_1\}$, with the number of pixels in the sets defined as $|\mathcal{M}^{\sigma}_{i}| = n^{i}_{b}$, and $|\mathcal{M}^{\sigma}_{i*}| = n^{i}_{d}$.

When $\mathcal{G}_k$, is unknown $\hat{\mathcal{M}}^{\sigma}_{i}$ and $\hat{\mathcal{M}}^{\sigma}_{i*}$ are estimated using some segmentation procedure. Any segmentation procedure may be used to estimate the partitions, as long as the resulting partitions are contiguous regions. In this dissertation, we apply edge detection methods to segment the image by identifying edges, which are located at the maxima of the gradient strength obtained from a Laplacian of the Gaussian-smoothed image (Canny, 1986; Parker, 2010). For certain parameter values, the edge contours are closed creating contiguous regions and the standard deviation of the filter changes how many regions are detected. Let $\hat{e}$ be the edge set which segments the image $\mathcal{M}^{\sigma}$ into partitions $\hat{\mathcal{G}}_k$.

Assume that some part of the segmentation of a loop or its interior is incorrect so that $\hat{\mathcal{G}}_k \neq \mathcal{G}_k$ for $k = \{i, i*\}$ for some i. Then there are m_d pixel intensities, denoted by $\tilde{Z}^{i*}$, in the set $\mathcal{M}_{i*}$ which are misclassified into $\hat{\mathcal{M}}_i$ (i.e., these are the pixels that should be a part of the interior, but were assigned to the loop). Similarly, there are m_b pixel intensities, denoted by $\tilde{Z}^{i}$, in the set $\mathcal{M}_i$ which are misclassified into $\hat{\mathcal{M}}_{i*}$ (i.e., these are the pixels that should be a part of the loop, but were assigned to the interior). There are then $n_d - m_d$ pixel intensities, denoted by $\tilde{\tilde{Z}}^{i*}$, in the set $\mathcal{M}_{i*}$ which are correctly classified into $\hat{\mathcal{M}}_{i*}$ and there are $n_b - m_b$ pixel intensities, denoted by $\tilde{\tilde{Z}}^{i}$, in the set $\mathcal{M}_i$ which are correctly classified into $\hat{\mathcal{M}}_i$.

The set of pixels which comprise the interior of the loop Z^{i*} and the set of pixels which comprise the loop Z^{i} can be decomposed as follows:

$$Z^{i*} = \tilde{\tilde{Z}}^{i*} \cup \tilde{Z}^{i*} \text{ and } Z^{i} = \tilde{\tilde{Z}}^{i} \cup \tilde{Z}^{i}. \tag{2.15}$$

$\hat{\mathcal{M}}_{i*}$ denotes all the pixels which are classified as interior pixels of the loop $\hat{\mathcal{G}}_{i*}$ (i.e., $\hat{\mathcal{M}}_{i*} = \tilde{\tilde{Z}}^{i*} \cup \tilde{Z}^{i}$) and $\hat{\mathcal{M}}_i$ denotes all the pixels which are classified as loop pixels $\hat{\mathcal{G}}_{i*}$ (i.e., $\hat{\mathcal{M}}_i = \tilde{\tilde{Z}}^{i} \cup \tilde{Z}^{i*}$). Therefore $n_b - m_b + m_d$ pixels are in the birth time partition $\hat{\mathcal{G}}^{\sigma}_{i}$ and $n_d - m_d + m_b$ pixels are in the death time partition $\hat{\mathcal{G}}^{\sigma}_{i*}$.

The expected value of the (biased) estimators of the birth and death time using the incorrect partitions of the loop are:

$$E(\bar{\hat{M}}_i) = \frac{(n_b - m_b)\mu_i + m_d\mu_{i*}}{n_b - m_b + m_d} \text{ and } E(\bar{\hat{M}}_{i*}) = \frac{(n_d - m_d)\mu_{i*} + m_b\mu_i}{n_d - m_d + m_b}, \tag{2.16}$$

where $\bar{\hat{M}}_i$ and $\bar{\hat{M}}_{i*}$ are the sample means of the sets of pixels $\hat{\mathcal{M}}_i$ and $\hat{\mathcal{M}}_{i*}$, respectively.

By Assumption 3, $\mu_{i*} \leqslant \mu_i$ and assuming that the segmentation $\hat{\mathcal{G}}_i$ and $\hat{\mathcal{G}}_{i*}$ are close to the true $\mathcal{G}_i$ and $\mathcal{G}_{i*}$ (i.e., only a few pixels are misclassified), then $m_b < n_d$ and $m_d < n_b$ and any $\tilde{Z}^i \in \hat{\mathcal{M}}_{i*}$ and $\tilde{Z}^{i*} \in \hat{\mathcal{M}}_i$ are neighbors of the edge set $\hat{e}$ (i.e., $\tilde{Z}^i, \tilde{Z}^{i*} \in n_c(\hat{e})$ where c is the unit distance between two pixels.

Let q_1^i, q_1^{i*} be the first quantiles and q_3^i, q_3^{i*} be the third quantiles of $\mathbf{F_{birth}}$, $\mathbf{F_{death}}$, respectively. Assume that the noise distribution $\varepsilon(x, y) \sim \mathbf{F}(0, \sigma^2(x, y))$ is symmetric. An assumption of Algorithm 2 is that the distribution of the interior pixel intensities and the pattern pixel intensities are well-separated, as described in the following.

Assumption 4. *Assume that* $(o_i - T(\tilde{\mu}_i)) < (o_i - \tilde{T}(\mu_{i*}))$ *and* $(o_{i*} - T(\tilde{\mu}_{i*})) < (o_{i*} - T(\tilde{\mu}_i))$ *where* o_i *is an outlier in the distribution* $\boldsymbol{F_{birth}}$ *and* o_{i*} *is an outlier in the distribution* $\boldsymbol{F_{death}}$. $T(\tilde{\mu}_{i*})$ *and* $T(\tilde{\mu}_i)$ *are the truncated means of* $\boldsymbol{F_{birth}}, \boldsymbol{F_{death}}$ *with upper bound* $q_3^{i*} + 1.5(q_3^{i*} - q_1^{i*})$ *and lower bound* $q_3^i + 1.5(q_3^i - q_1^i)$, *respectively.*

Under Assumption 4, Algorithm 2 sorts the m_b and m_d misclassified pixels, $\tilde{Z}^i$ and $\tilde{Z}^{i*}$, into the edge set $\hat{e}$ and keep the outliers, $\tilde{\tilde{Z}}^i \sim \mathcal{M}_i^\sigma$ and $\tilde{\tilde{Z}}^{i*} \sim \mathcal{M}_{i*}^\sigma$ in the correct segments $\hat{\mathcal{M}}_i^\sigma, \hat{\mathcal{M}}_{i*}^\sigma$.

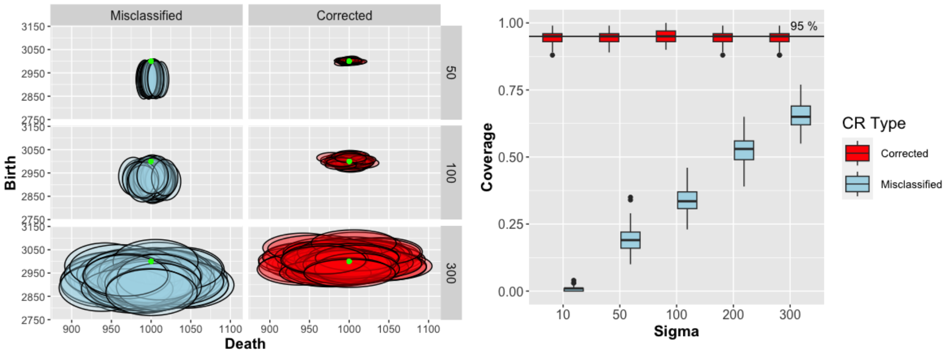
As an illustration of the performance of Algorithm 2, the following experiment was carried out and results are displayed in Figure 2.3. For three different noise settings $(\sigma = \{50, 100, 300\})$, 100 iid images with one loop, similar to Figure 1.1b with $(\mu_{1*}, \mu_1) = (1000, 3000)$, are generated and segmented incorrectly with the same edge set $\hat{e}$. In this example, six pixels are misclassified in the loop (i.e., $\tilde{Z}^{1*} \in \hat{\mathcal{G}}_1$) with the edge set $\hat{e}$. The 95% confidence regions using partitioned TDA are calculated using both this misclassified partition $\hat{e}$ and the corrected partition $\hat{e}_{\text{new}}$ generated from Algorithm 2. Lower noise levels have more biased coverage of the resulting confidence regions compared to the higher noise levels.

Algorithm 2 Remove Misclassified Pixels from Partition ($\mathcal{G}_1$,$\mathcal{G}_{1*}$)

```
Input: edge set $\hat{e}$; image $\mathcal{M}^\sigma$; partitions $\hat{\mathcal{G}}_1$ and $\hat{\mathcal{G}}_{1*}$; c=pixel side length
Output: new edge set $\hat{e}_{\text{new}}$
Define: $\hat{\mathcal{M}}_i^\sigma = \{Z^i(x,y)_l : (x,y)_l \in \hat{\mathcal{G}}_i\}$, $L_i = |\hat{\mathcal{M}}_i^\sigma|$, $P(Z^i(x,y) \leqslant q_1^i) = 0.25$, $P(Z^i(x,y) \leqslant q_3^i) = 0.75$ for $i = \{1, 1*\}$; $\text{outlier}_i = \emptyset$; $\text{outlier.idx}_i = \emptyset$; dist()=Euclidean distance; $e_1 = \emptyset$
for i in $\{1, 1*\}$ do
    for l in $1 : L_i$ do    ▷ Check if $Z^i(x,y)_l$ is an outlier and neighbors an edge in $\hat{\mathcal{G}}_i$
        if $\big((Z^i(x,y)_l > q_3^i + 1.5(q_3^i - q_1^i)) \mid (Z^i(x,y)_l < q_1^i - 1.5(q_3^i - q_1^i))\big)$ & $\Big(\exists(\tilde{x},\tilde{y}) \in \hat{e}$ s.t. $\text{dist}((x,y)_l, (\tilde{x},\tilde{y})) \leqslant \sqrt{2}c\Big)$ then $\text{outlier}_i \leftarrow \text{outlier}_i \cup Z^i(x,y)_l$, $\text{outlier.idx}_i \leftarrow \text{outlier.idx}_i \cup l$
        end if
    end for
end for
Calculate $\hat{\mu}_1 = \hat{\mathcal{M}}_1^\sigma \backslash \text{outlier}_1$ and $\hat{\mu}_{1*} = \hat{\mathcal{M}}_{1*}^\sigma \backslash \text{outlier}_{1*}$ ▷ Calculate means without outliers
for i in $\{1, 1*\}$ do
    for l in $\text{outlier.px}_i$ do
        if $|Z^i(x,y)_l - \hat{\mu}_i| \geqslant |Z^i(x,y)_l - \hat{\mu}_{i^c}|$ then ▷ $i^c$ is the complement in $\{1, 1*\}$ for i
            $e_1 \leftarrow e_1 \cup (x,y)_l$    ▷ only add $(x,y)_l$ to new edge set $e_1$ if $Z^i(x,y)_l$ is closer to $\hat{\mu}_{i^c}$
        end if
    end for
end for
$\hat{e}_{\text{new}} = \hat{e} \cup e_1$
return $\hat{e}_{\text{new}}$
```

Figure 2.3a shows all 100 estimated 95% confidence regions built using $\hat{e}$ (red) and $\hat{e}_{\text{new}}$ (blue) for the different σ values. The green dot is the true $(\mu_{1*}, \mu_1) = (1000, 3000)$ which the regions should cover 95% of the time, on average. The confidence regions for the misclassified setting are underestimating μ_1 since some Z^{1*}s pixel intensities, which are lower than those of Z^1, are included in the $\bar{Z}^1$ resulting in an estimate that is biased low. After Algorithm 2 is applied, the bias in the confidence regions appear to be corrected in terms of the birth time.

In Figure 2.3b, the coverage is calculated based on 100 iid images at each noise level ($\sigma = \{10, 50, 100, 200, 300\}$). The misclassified boxplots (red) show the coverage of the confidence regions built from $\hat{e}$, and the corrected boxplots (blue) show the coverage for confidence regions calculated with the $\hat{e}_{\text{new}}$ after running Algorithm 2. As illustrated in both plots, the algorithm significantly improves the coverage of the confidence regions. Correct segmentation is crucial for partitioned TDA, and this section emphasizes the importance of checking the segmentation.



(a) Example of confidence regions (b) Coverage of confidence regions

Figure 2.3: Confidence regions and coverage before (misclassified) and after (corrected) Algorithm 2 has been applied. The misclassified segmentation $\hat{e}$ has six pixels incorrectly classified. (a) Confidence regions for 100 images at noise level $\sigma = \{50, 100, 300\}$ are shown using $\hat{e}$ (misclassified) and $\hat{e}_{\text{new}}$ (corrected). The green dots indicate the true birth and death time location. (b) The coverage of the 95% confidence regions for $\sigma = \{10, 50, 100, 200, 300\}$ for misclassified (red) and corrected (blue) segmentations, using 100 iid images.

2.7 Alternative Method

We extend one of the methods from Fasy et al. (2014) from point-cloud data to handle an image as a way to establish a benchmark, because we are unaware of a direct basis for comparison with partitioned TDA. In this approach, a distance metric is used to derive the distribution of distances between the persistence diagrams of the smoothed data, $\mathcal{P}(\tilde{\mathcal{M}}^{\sigma})$, and the persistence diagram of the true pattern, $\mathcal{P}(\tilde{\mathcal{M}}^{0})$.

Persistence diagram stability results (Cohen-Steiner et al., 2005) are used to bound the (bottleneck) distance between the persistence diagrams by the L_{∞} distance between kernel density estimates (KDEs) of the point-cloud data and the true pattern. Asymptotic confidence regions are then built from the distribution of L_{∞} distances between $\tilde{\mathcal{M}}^{\sigma}$ and $\tilde{\mathcal{M}}^{0}$, which can be estimated using a bootstrap procedure.

This procedure is briefly outlined below and then followed by the proposed adjustments for image data. See Section 3.4 of Fasy et al. (2014) for more details.

In the context of Fasy et al. (2014), let $\mathcal{M}^{\sigma}$ be point-cloud data. One of their proposed methods for persistence diagram confidence regions considers a KDE of $\mathcal{M}^{\sigma}$, $\tilde{\mathcal{M}}^{\sigma}$, to estimate the true birth and death time, $(\tilde{\mu}_{i*}, \tilde{\mu}_{i})$, of the (true) smoothed manifold, $\tilde{\mathcal{M}}^{0}$. They define an asymptotic $(1-\alpha)100\%$ confidence region, adapted to our notation which omits the dependency on bandwidth and sample size; see Theorem 12 of Fasy et al.

(2014) for the precise statements:

$$\mathbb{P}\left(\mathcal{W}_{\infty}(\mathcal{P}(\tilde{\mathcal{M}}^{\sigma}), \mathcal{P}(\tilde{\mathcal{M}}^{0})) > c_n\right) \leqslant \mathbb{P}\left(\|\tilde{\mathcal{M}}^{\sigma} - \tilde{\mathcal{M}}^{0}\|_{\infty} > c_n\right) \leqslant \alpha + O\left(n^{-1/2}\right) \tag{2.17}$$

where c_n defines the confidence region based on the data, and the first inequality follows from the stability result of Cohen-Steiner et al. (2005). The bottleneck distance, $\mathcal{W}_{\infty}$ is defined as

$$W_{\infty}(\mathcal{P}(\tilde{\mathcal{M}}^{\sigma}), \mathcal{P}(\tilde{\mathcal{M}}^{0})) = \inf_{\eta:\mathcal{P}(\tilde{\mathcal{M}}^{\sigma}) \longrightarrow \mathcal{P}(\tilde{\mathcal{M}}^{0})} \sup_{(b,d)\in\mathcal{P}(\tilde{\mathcal{M}})} \|(b, d) - \eta(b, d)\|_{\infty} \tag{2.18}$$

where η is a bijection of the features of the diagrams, including the diagonal birth=death line (Cohen-Steiner et al., 2005; Fasy et al., 2014). Since $\tilde{\mathcal{M}}^{0}$ is unknown and there is only one realization of the data $\tilde{\mathcal{M}}^{\sigma}$, a bootstrap approach is used. In particular, the estimate of c_n is the $(1-\alpha)$-quantile of the distribution of the L_{∞} distances between the smoothed data $\tilde{\mathcal{M}}^{\sigma}$ and smoothed bootstrap realizations of the point-cloud data.

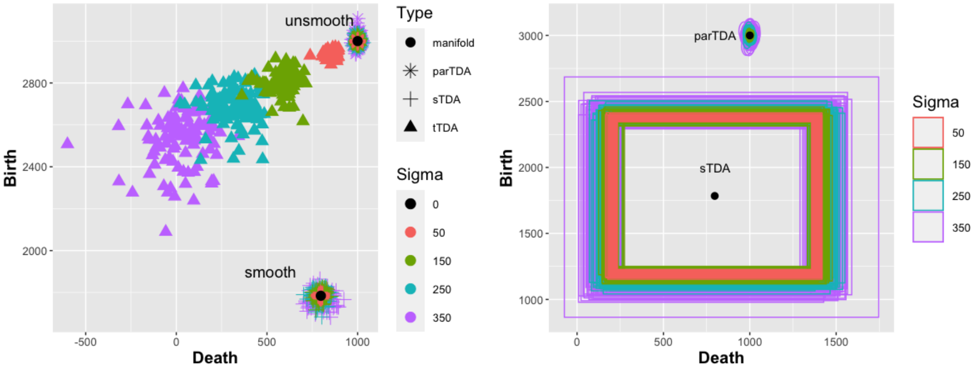
To implement this alternative approach two modifications are made: (1) Instead of a KDE on point clouds, we use local polynomial smoothing to change the raw image $\mathcal{M}^{\sigma}$ into a smoothed image $\tilde{\mathcal{M}}^{\sigma}$. In Section 2.8, we use degree two polynomials and an adaptive bandwidth of 0.3 as parameter inputs for local polynomial smoothing. These input values resulted in only one loop detected by an upper-level set filtration for the smoothed pattern, $\tilde{\mathcal{M}}^{0}$, analogous to the original image, $\mathcal{M}^{0}$. This facilitates the comparison between smooth TDA and partitioned TDA. Note that smooth TDA builds confidence regions to cover $(\tilde{\mu}_{i*}, \tilde{\mu}_i)$ (i.e., death and birth times of loops in $\tilde{\mathcal{M}}^0$) whereas partitioned TDA builds confidence regions to cover (μ_{i*}, μ_i) (i.e., death and birth times of loops in $\mathcal{M}^0$). (2) We propose a method to bootstrap an image as opposed to a point cloud. The traditional bootstrap method assumes that each observation is iid which is not a suitable assumption for an image which typically has spatial correlation. Similar to partitioned TDA, we segment the image into different strata and use the stratified bootstrap to resample the full image. Within each stratum the pixels can be viewed as being drawn from the same distribution, so pixel intensities within each stratum can be bootstrapped.

2.8 Simulation Study

In this section, we empirically evaluate the accuracy and precision of the proposed confidence regions. Accuracy is assessed by considering bias in the estimates, coverage percentage over the truth, and the identification of the number of loops in the underlying pattern, while precision is evaluated by analyzing the area of the confidence regions. A summary of all of these numerical results are displayed in Table 4.1. In our simulation study, the number of strata and the segmentation is assumed to be correct for the smooth TDA benchmark.

For the simulations, each image has one loop and follows the assumptions from Section 2.2. The birth and death times of the true pattern, $\mathcal{M}^0$, are set to $(\mu_{1*}, \mu_1) = (1000, 3000)$, which are similar intensities to those of our cell wound example (see Section 2.9). To assess the robustness of the proposed confidence regions to noise, four different noise levels are used to generate an image $\mathcal{M}^\sigma$ for $\sigma = \{50, 150, 250, 350\}$, homoscedastic Gaussian noise is used in this section. For each σ, l images are generated, denoted $\mathcal{M}_l^\sigma$ where $l = \{1, \ldots, 100\}$, and an upper-level set filtration is used to get the birth and death times for each image (i.e., the traditional TDA estimates). To test the alternative method (smooth TDA) each image is further smoothed using local polynomial smoothing, denoted $\tilde{\mathcal{M}}_l^\sigma$. Then both smooth TDA and partitioned TDA are used to get confidence regions for the underlying pattern in $\tilde{\mathcal{M}}^0$ and $\mathcal{M}^0$, respectively.

Figure 2.4 illustrates the simulation results, with examples of point estimates for the birth and death times shown in Figure 2.4a (i.e., estimated pattern) and their corresponding confidence regions are shown in Figure 2.4b (i.e., uncertainty estimate for the pattern). In both figures, each color represents a different σ value. In Figure 2.4a, the shapes are the estimated birth and death times for each method where the black dots are the true birth and death time of the smoothed $(\tilde{\mu}_{1*}, \tilde{\mu}_1)$ and unsmoothed loop (μ_{1*}, μ_1). In Figure 2.4b, the rectangles are the confidence regions using smooth TDA, denoted by sTDA, with the L_∞ distance and the ellipses are the confidence regions generated using partitioned TDA, denoted by parTDA.



(a) (Death, Birth) of loop (b) Confidence Regions

Figure 2.4: Birth and death estimates (a) and confidence regions (b) of 100 images across four noise levels, $\sigma = \{50, 150, 250, 350\}$. (a) Point estimates for (μ_{1*}, μ_1) using traditional TDA (triangle) and partitioned TDA (asterisk), estimates of $(\tilde{\mu}_{1*}, \tilde{\mu}_1)$ using smooth TDA (plus), and the true birth and death time of the manifold (circle). (b) The 95% confidence regions for (μ_{1*}, μ_1) using partitioned TDA (parTDA) and $(\tilde{\mu}_{1*}, \tilde{\mu}_1)$ using smooth TDA (smooth TDA).

Across all noise settings, point estimates from traditional TDA in Figure 2.4a are significantly biased, especially as the noise level increases. While partitioned TDA creates unbiased estimates close to μ_{1*} and μ_1 and smooth TDA creates unbiased estimates close to $\tilde{\mu}_{1*}$ and $\tilde{\mu}_1$. However, the confidence regions created using partitioned TDA are much smaller (more precise) compared to smooth TDA. Using smooth TDA, the confidence bands are large enough that a persistence of zero is within each confidence region for every loop in the data. This result suggests that no loop is distinctly identified within the underlying pattern. Whereas, partitioned TDA correctly identifies one loop for all simulated images when using Algorithm 1, and no other loops in the image are matched to the segmentation. In terms of coverage, smooth TDA covers the true birth and death times of $\tilde{\mathcal{M}}^0$ 100% of the time for a 95% confidence region. In comparison, the coverage of partitioned TDA was always approximately 95% at all noise levels.

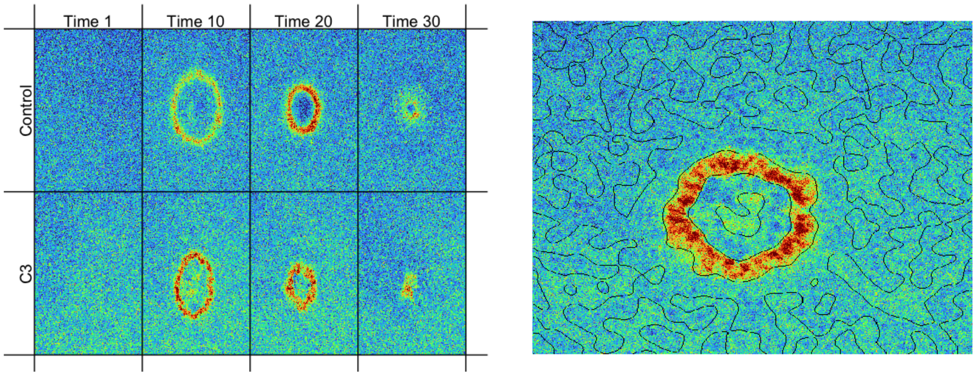
Method	Noise Level	Average confidence region area (SE)	Average coverage (SE)
sTDA	50	1390574 (5553.5)	100 (0)
	150	1460183 (9529.5)	100 (0)
	250	1577603 (14139.8)	100 (0)
	350	1746974 (28184.4)	100 (0)
parTDA	50	122.9 (0.683)	94.7 (0.2)
	150	1099.3 (7.732)	95.3 (0.2)
	250	3057.2 (18.359)	94.6 (0.2)
	350	5980.2(30.602)	94.9 (0.3)

Table 2.1: Simulations Results of a noisy loop for smooth TDA (**sTDA**) (rows 1-4) and partitioned TDA (**parTDA**) (rows 5-8). The average confidence region area and standard errors (SE) are displayed for each noise level, based on 100 iid images in each setting. The fourth column is the percent coverage of the 95% confidence regions, and corresponding SEs.

2.9 Cell Biology Application

Pattern formation is a common and critically important feature of living systems. It is a natural process that occurs across biological scales ranging from ecosystems (Pringle and Tarnita, 2017; Barbier et al., 2022), to developing tissues (Madamanchi et al., 2021; Herron et al., 2022), to individual cells (Bement et al., 2022, 2024). Further, abnormal cell or tissue pattern formation is a feature of various pathological conditions, including cancers (Paine and Lewis, 2017; A., 2020). Consequently, approaches for objectively detecting and quantifying patterns and their quality are of interest for both basic biology and medicine. In this dissertation, pattern is assessed from the perspective of TDA through estimation of the birth and death times of rings with partitioned TDA. A higher persistence (birth-death) is indicative of stronger topological signal, and can be interpreted as a stronger pattern in this context.

The proposed partitioned TDA is applied to images of two individual cells sustaining wounds at distinct time points as illustrated in Figure 2.5. One of the cells was injected with a toxin (C3 exotransferase) that inhibits healing. The other cell is only wounded with no injection and serves as a control. The image for the C3 cell is denoted as $\mathcal{M}_t^{C3}$ and the image for the Control cell denoted as $\mathcal{M}_t^{Con}$ for times $t = \{t_1, \dots, t_{30}\}$. Time $t_1 = 0$ seconds is when the cell is wounded with sequential images separated by 8 seconds. Examples of the cell images at different time points are shown in Figure 2.5a. Each of the images at every time point, $\mathcal{M}_t^{con}$ and $\mathcal{M}_t^{C3}$, was partitioned using the segmentation scheme from Section 2.6 with e_t^{Con} and e_t^{C3} representing the edge sets at time t. An example of a segmentation at t_{15} for $\mathcal{M}_t^{C3}$ is shown in Figure 2.5b.



(a) $\mathcal{M}_t^{\text{Con}}$ and $\mathcal{M}_t^{\text{C3}}$ (b) Segmentation $e_{t_{15}}^{\text{C3}}$

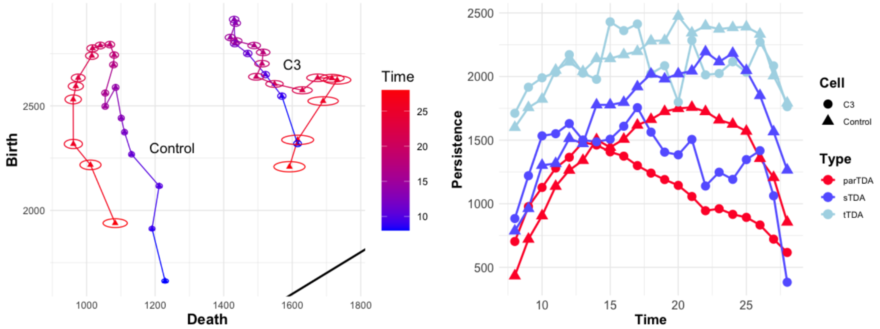
Figure 2.5: (a) The top row displays the images for the control cell $\mathcal{M}_t^{\text{Con}}$ and the bottom row is the images for the C3 cell $\mathcal{M}_t^{\text{C3}}$ for $t = \{t_1, t_{10}, t_{20}, t_{30}\}$. The columns represent different time points, t_1, t_{10}, t_{20}, and t_{30}. (b) Image of $\mathcal{M}_{t_{15}}^{\text{C3}}$ segmented by $e_{t_{15}}^{\text{C3}}$ where the black lines are the edges.

The analysis is conducted independently at each time point. For each t the number of rings in an image are detected using Algorithm 1 and a confidence region is created around the birth and death times using Equation (2.4). In this higher resolution image, Algorithm 1 has to be modified because multiple pixels in the image are equal to b_j. To address this, we smoothed the image, calculated the birth and death times, and used the smoothed birth time $\tilde{b}_j$ to help locate the pixel associated with b_j.

Using the partitioned TDA method, no rings were detected until times t_8 and t_7 for $\mathcal{M}_t^{\text{C3}}$ and $\mathcal{M}_t^{\text{Con}}$, respectively, whereas the traditional TDA method detected rings in images for $t \leqslant t_6$. When using partitioned TDA no ring was contained in e_t^{Con} and e_t^{C3} for $t < t_7$, so Algorithm 1 has no partitions to match with the rings detected in traditional TDA. From times

t_8 to t_{28} one ring is matched from e_t^{Con} to $\mathcal{M}_t^{\text{Con}}$ and from e_t^{C3} to $\mathcal{M}_t^{\text{C3}}$ using partitioned TDA and thus these are the times focused on in this section.

Two different visualizations of persistence across time for both cells are displayed in Figure 2.6. In Figure 2.6a, the partitioned TDA birth and death estimates are shown on a persistence diagram along with the confidence regions. The estimated birth and death times are connected by time, where time is indicated by different colors. Figure 2.6b, is another way to visualize persistence (y-axis) over time (x-axis). When using partitioned TDA, the estimated persistence is $\bar{Z}_t^1 - \bar{Z}_t^{1*}$, at each time t. The confidence set moves from a bivariate normal ellipse to a normal confidence interval centered at $\bar{Z}_t^1 - \bar{Z}_t^{1*}$ with approximate variance $(\hat{\sigma}_1^2)_t + (\hat{\sigma}_{1*}^2)_t$. The red lines are the estimated persistence and confidence intervals from partitioned TDA for both C3 (points) and Control (triangle) cells; the error bars are too small too see since sample size is large due to the high-resolution images. The dark blue lines use smooth TDA and the light blue lines use traditional TDA to estimate persistence across time; no confidence intervals were created for these methods. In general, smooth TDA and traditional TDA display more variability in the estimated persistences across time than partitioned TDA, and the C3 and Control cell persistences for traditional TDA are not well separated. The overall trends in smooth TDA and partitioned TDA are similar, though the partitioned TDA persistences appear to be more stable across time.



(a) (Death, Birth) estimates (b) Estimated Persistence

Figure 2.6: Estimated persistences of the C3 and Control cell images from $t = \{t_8, \ldots, t_{28}\}$. (a) The partitioned TDA birth and death times are shown on the persistence diagram along with confidence regions for both the C3 cell (right) and the Control cell (left). The black line is the diagonal line birth=death. (b) Persistence is plotted over time for the C3 cell (solid line with points) and the Control cell (dashed line with triangles) using partitioned TDA (red) denoted as parTDA, smooth TDA (purple) denoted as sTDA, and traditional TDA (light blue) denoted as tTDA.

From t_8 to t_{14}, the most rapid growth in the persistence (or strength of pattern) are observed. Originally, the C3 cell images have more pattern in terms of the ring having a higher persistence than the Control cell images. However, at t_{14} the wound ring in the Control cell continues to increase in its persistence while the wound ring in C3 cell begins to decline. In later time periods, the rings have shrunk in size, but not necessarily in intensity. The smaller size of the rings in the images result in larger confidence regions since the sample sizes of the sample means (i.e., the number of pixels in the pattern) has decreased. After t_{29} the segmentation, $e_{t_{29}}^{C3}$, does not have any rings in the partitions; the edge set in the background is almost completely connected as one edge. Two distinct edges are needed to separate the section of an image into $\mathcal{M}_i^\sigma$ and $\mathcal{M}_{i*}^\sigma$ to find a ring in the segmentation. Therefore, no ring on $\mathcal{P}(\mathcal{M}_{29}^{C3})$ is matched to any regions in $e_{t_{29}}^{C3}$ as per Algorithm 1.

During times $t_{29} - t_{30}$, the segmentation of the Control cell images continues to detect a ring where the wound is (i.e., two distinct edges separate $\mathcal{M}_1^{Con}$ and $\mathcal{M}_{1*}^{Con}$ which are matched to the ring detected in $\mathcal{P}(\mathcal{M}^{Con})$ for times t_{29}, t_{30}); however, in order to directly compare the Control cell with the C3 cell only times $t_8 - t_{28}$ are included in Figure 2.6.

2.10 Conclusion

This Chapter includes three primary developments in TDA methodology. First, partitioned TDA is proposed to estimate the birth and death times of topological features found in an image, which reduces the bias in the traditional TDA estimates (traditional TDA). Second, partitioned TDA provides a process to quantify the uncertainty associated with these new birth and death time estimates in the form of a confidence region on a persistence diagram for an image. And finally, a persistence diagram confidence region method of Fasy et al. (2014) was extended from point-cloud data to a single image as an alternative method (smooth TDA), which facilitated the creation of a new method to bootstrap an image. In general, partitioned TDA is applicable to any image to determine the underlying pattern (in terms of holes) of that image and to quantify the uncertainty in that pattern. In the next chapter, we explore methods to perform hypothesis testing on persistence diagrams of images to be used in conjunction with partitioned TDA and other applications.

3 HYPOTHESIS TEST FOR A PERSISTENCE DIAGRAM

3.1 Introduction

In Section 2.1, partitioned TDA was introduced as a method to make confidence regions for topological features on persistence diagrams generated from an upper-level set filtration on a single image. These confidence regions give unbiased estimates of the birth and death times of loops along with quantifying the uncertainty of the estimates. However, partitioned TDA is not able to successfully differentiate topological signal (where a topological feature has a persistence that is well-separated from the birth=death line) from noise for image data compared to the confidence region method proposed in Fasy et al. (2014) for persistence diagrams of point clouds. In this chapter, we address this limitation by developing two hypothesis testing frameworks to identify statistically significant topological features within images.

There are several ways to formulate a hypothesis testing framework to distinguish topological signal from topological noise on a persistence diagram generated from images. Many hypothesis tests on persistence diagrams, outside of Fasy et al. (2014), use some version of a randomization test, spatial point processes, or density estimation to do inference. Most of the randomization techniques are applied to multiple groups or compare the topological features at one point in time to another point in time (Abdallah et al., 2023; Robinson and Turner, 2017; Dawson et al., 2023; Chung et al., 2009). We introduce two hypothesis tests: (1) tests whether the maximum persistent H_m feature on a persistence diagram exhibits statistically significant differences from the maximum persistent features generated under the assumption of no underlying pattern in the image, and (2) tests if a specific H_m feature is part of the underlying pattern by comparing the mean pixel values in the birth time partition with those in the rest of the image. Both of these tests are carried out with variations of a permutation test applied to images using TDA-based test statistics.

The first hypothesis test, referred to as the *Maximum Persistence Test*, focuses on the testing if the maximum persistent H_m feature is statistically significant which is convenient for testing if the cell wound is topological signal. This test is applied in Chapter 4 to quantify the pattern in a cell wound over time in a video, as opposed to a single image. The second hypothesis test, referred to as the *Partitioned Test*, can be directly integrated into the partitioned TDA method. The confidence regions created using partitioned TDA should only include loops that have been found to be statistically significant using the Partitioned Test.

In the next section, we describe the general setup for both hypothesis tests. Following this, we define the null hypothesis and the corresponding null distribution, and provide implementation details for both the Maximum Persistence Test and the Partitioned Test. Finally, we illustrate the effectiveness of these tests using examples from both simulated and real data, with a focus on the wounded cell application.

3.2 Setup

Recall that an image in Section 2.1 is defined as $\mathcal{M}^{\sigma} = \{f(x,y) + \varepsilon(x,y) : (x,y) \in \mathcal{G}\}$ where $f(x,y)$ is the expected value of the pixel intensity, $\varepsilon(x,y)$ is the noise, and this noisy function is discretized onto a 2D grid $\mathcal{G}_{d_1 \times d_2}$, where each (x,y) coordinate represents the grid columns $x = \{1, 2, \ldots, d_1\}$ and grid rows $y = \{1, 2, \ldots, d_2\}$. To generalize this notation to higher dimensions, let $\mathcal{A}^{\sigma}$ be an array of dimension $d_1 \times \ldots \times d_M$, defined as

$$\mathcal{A}^{\sigma} = \{f(x_1, x_2, \ldots, x_M) + \varepsilon(x_1, x_2, \ldots, x_M) : (x_1, x_2, \ldots, x_M) \in \mathcal{G}\}, \quad (3.1)$$

where $\mathcal{G}$ is an M-dimensional grid, with each tuple $(x_1, x_2, \ldots, x_M)$ representing a coordinate on that grid. The function $f(x_1, x_2, \ldots, x_M)$ and the noise $\varepsilon(x_1, x_2, \ldots, x_M)$ are also defined in M dimensions, the noise assumed to follow some symmetric distribution $\mathbf{F}(0, \sigma^2(x_1, x_2, \ldots, x_M))$. The noise-free array, from which we aim to identify the topological features, is defined as:

$$\mathcal{A}^{0} = \{f(x_1, x_2, \ldots, x_M) : (x_1, x_2, \ldots, x_M) \in \mathcal{G}\}. \quad (3.2)$$

When an upper-level set filtration is applied to $\mathcal{A}^{\sigma}$, different dimensional features, from H_0 to H_{M-1}, may be detected on a persistence diagram. Some of these features may be generated from $\varepsilon(x_1, x_2, \ldots, x_M)$ (noise) and some may be a part of $f(x_1, x_2, \ldots, x_M)$ (signal). The goal of the hypothesis tests are to discriminate the real topological features from those produced by noise.

3.3 Maximum Persistence Test

To test if any of the topological features detected in the array $\mathcal{A}^{\sigma}$ are statistically significant, we apply a hypothesis testing framework where the null hypothesis assumes there is no structure in the image for the Maximum Persistence Test. Specifically, assume that $f(x_1, x_2, \ldots, x_M)$ has a real feature $(H_m)_i$ of a particular homology dimension m where $m = \{0, \ldots, M-1\}$, $i = \{1, \ldots, n_m\}$, and n_m is the total number of real (H_m) features. Then the persistence of $(H_m)_i$ should be higher than most of the persistences of the H_m features under the null hypothesis, $(H_m)_{\text{null}}$, where there is no structure present in the image. We focus on the maximum persistent feature, since a common interpretation is that a higher persistence of a feature indicates a higher likelihood of that feature being signal.

We assume the following two conditions for the Maximum Persistence Test when the null hypothesis is true.

Assumption 5. *For the null hypothesis of the Maximum Persistence Test, assume that the array $\mathcal{A}^0$ has no topological structure such that $\mathcal{A}^0 = \{\mu_0 :$*

$\forall(x_1, x_2, \ldots, x_M) \in \mathcal{G}\}$, *for a fixed* μ_0. *This implies that there are no true* H_m *features,* $m = 0, \ldots, M-1$.

In this case, μ_0 denotes the mean of the background partition where there are no features.

Assumption 6. *For simplicity, assume that the noise is homoskedastic in* $\mathbf{F}(0, \sigma^2(x_1, x_2, \ldots, x_M))$ *such that* $\mathcal{A}^\sigma = \{\mu_0 + \sigma_0 : \forall(x_1, x_2, \ldots, x_M) \in \mathcal{G}\}$.

Given these assumptions, any topological features detected in $\mathcal{A}^\sigma$ have pixels sampled from the following distribution:

$$Z(x_1, \ldots, x_M) \sim \mathbf{F}(\mu_0, \sigma_0^2). \tag{3.3}$$

For a given array, an upper-level set filtration on $\mathcal{A}^\sigma$ and its corresponding persistence diagram $\mathcal{P}(\mathcal{A}^\sigma)$ have β_m H_m features for each dimension $m = \{0, \ldots, M-1\}$. Any $(H_m)_i$ feature detected in the array has a birth and death time, $\{(d_1^{(m)}, b_1^{(m)}), \ldots, (d_{\beta_m}^{(m)}, b_{\beta_m}^{(m)})\}$, where $d_i^{(m)}$ represents the death time and $b_i^{(m)}$ represents the birth time for H_m feature $i = 1, \ldots, \beta_m$. All of these birth and death times are considered topological noise assuming H_{null} is true so that:

$$d_i^{(m)} \sim \mathbf{F}(\mu_0, \sigma_0^2), \quad b_i^{(m)} \sim \mathbf{F}(\mu_0, \sigma_0^2), \quad i = 1, \ldots, \beta_m.$$

For a given feature $(H_m)_i$, the null hypothesis can be stated such that the birth and death times of $(H_m)_i$ have the same distributions so that:

$$H_{\text{null}} : d_i^{(m)} \sim \mathbf{F}(\mu_0, \sigma_0^2), \quad b_i^{(m)} \sim \mathbf{F}(\mu_0, \sigma_0^2). \tag{3.4}$$

For the Maximum Persistence Test, we first focus on the maximum persistent H_m feature as this feature is the most likely to be statistically significant in terms of its persistence. The maximum persistent H_m is defined as follows:

$$\rho_{\text{max}}^{(m)} = \max_{(d_i^{(m)}, b_i^{(m)}) \in H_m(\mathcal{A}^\sigma)} \{b_i^{(m)} - d_i^{(m)} | i = 1, \ldots, \beta_m\}, \tag{3.5}$$

where $b_{\text{max}}^{(m)}$ is the birth time and $d_{\text{max}}^{(m)}$ is the death time of the most persistent H_m feature.

A one-sample permutation test is used to empirically generate the null distribution from Equation (3.4) for the Maximum Persistence Test; we can use a one-sample test since we are assuming all the pixels in the image are from the same distribution in Equation (3.3). Let B denote the total number of permutations such that there are $j = \{1, \ldots, B\}$ permuted arrays $(\mathcal{A}_1^\sigma)^*, \ldots, (\mathcal{A}_B^\sigma)^*$ from which to generate the null distribution. Let there be n total pixels with make up $\mathcal{A}^\sigma$; each pixel intensity $(Z_j)^*(x_1, \ldots, x_M)$ which make up the j^{th} permuted array $(\mathcal{A}_j^\sigma)^*$ are permuted as follows:

$$(Z_j)^*(x_1, x_2, \dots, x_M) = Z(x_{i_1}, x_{i_2}, \dots, x_{i_M})$$

where $i_1, i_2, \dots, i_M$ are permuted indexes from $i = \{1, \dots, n\}$.

For each new permutation $(\mathcal{A}_j^\sigma)^*$, we generate a persistence diagram, $\mathcal{P}((\mathcal{A}_j^\sigma)^*)$, and get the persistence $\rho^*_{j,\max}$ of the maximum persistent H_m feature:

$$\rho^*_{j,\max} = \max_{(d_i^{(m)}, b_i^{(m)}) \in H_m((\mathcal{A}_j^\sigma)^*)} \{b_i^{(m)} - d_i^{(m)} | i = 1, \dots, \beta^*_{m,b}\}. \tag{3.6}$$

The set $\boldsymbol{\rho^*}_{\mathbf{max}} = \{\rho^*_{1,\max}, \dots, \rho^*_{B,\max}\}$ are used to estimate the null distribution. The total number of H_m features detected in an upper-level set filtration for each new permutation $(\mathcal{A}_j^\sigma)^*$ is denoted by $\beta^*_{m,j}$. The birth and death times for the maximum persistent H_m feature in permuted array j are denoted as $((d_{\max}^{(m)})^*, (b_{\max}^{(m)})^*)$.

The observed test statistic for the Maximum Persistence Test is the persistence of the maximum persistent H_m feature described in Equation (3.5) for the observed data $\mathcal{A}^\sigma$ defined as:

$$\rho_{\max}^{\text{obs}} = b_{\max}^{(m)} - d_{\max}^{(m)}. \tag{3.7}$$

The permutation p-value can be then be estimated as follows:

$$\text{p-value}_{\text{max}} = \frac{\sum_{j=1}^{B} I(\rho^*_{j,\text{max}} \geqslant \rho^{\text{obs}}_{\text{max}})}{B}. \tag{3.8}$$

If the observed maximum persistence of the H_m features, $\rho^{\text{obs}}_{\text{max}}$, is in the upper α percentile of the null distribution, $\boldsymbol{\rho^*}_{\mathbf{max}} = \{\rho^*_{1,\text{max}}, \dots, \rho^*_{B,\text{max}}\}$, then the null hypothesis is rejected. The statistically significant H_m feature is considered topological signal (i.e., it is assumed to be a topological feature in $\mathcal{A}^0$).

Recall that the birth and death times of loops detected in an upper-level set filtration on an array $\mathcal{A}^\sigma$ are based on the intensities of the pixels, not on geometric information such as area of the feature. One way to indirectly include geometric information in this filtration process is to smooth the array $\tilde{\mathcal{A}}^\sigma$ so that geometrically smaller features may be eliminated by the smoothing.

An example of the pipeline for this test can be seen in Figure 3.1 where the first column shows the smoothed images that we want to perform the Maximum Persistence Test on: $\tilde{\mathcal{M}}^{C3}_{t_1}$ (row one) and $\tilde{\mathcal{M}}^{C3}_{t_{12}}$ (row two). The next column shows an example of a smoothed permutation of that image. And the last column shows a persistence diagram with the maximum persistent loop of all the permuted images (null distribution) and the maximum persistent loop from the data.

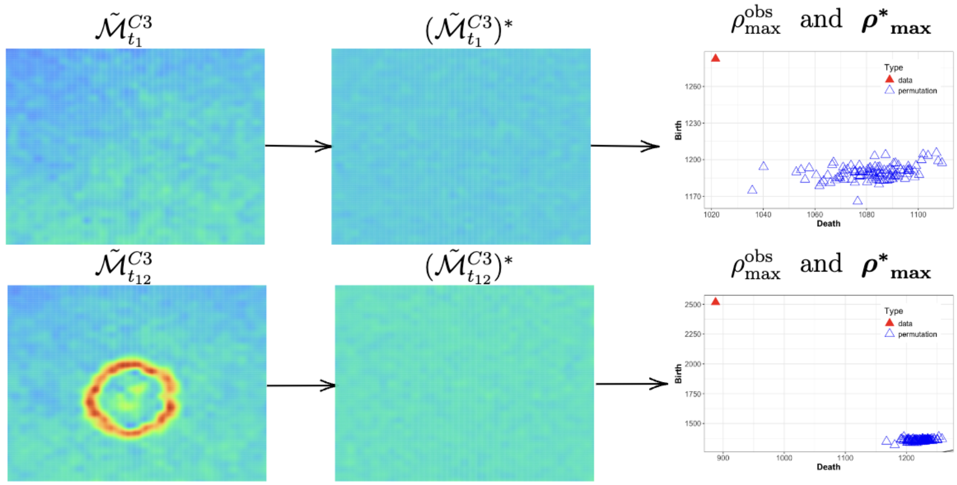


Figure 3.1: Pipeline for the Maximum Persistence Test for $\tilde{\mathcal{M}}_{t_1}^{c3}$ and $\tilde{\mathcal{M}}_{t_{12}}^{c3}$. The first column contains the smoothed images at time t_1 (row 1) and t_{12} (row 2). The second column has permuted examples of the smoothed images where the raw image $\mathcal{M}_t^{c3}$ is permuted and then smoothed to become $(\tilde{\mathcal{M}}_t^{c3})^*$. The last column has all the maximum persistent H_1 features for the permutations $\boldsymbol{\rho}_{\mathbf{max}}^*$ (blue triangles) and the maximum persistent H_1 feature of the data $\rho_{\max}^{\text{obs}}$ (red triangle).

3.4 Partitioned Test

When conducting the Maximum Persistence Test the null hypothesis assumes that there is no structure in the array, so for a test considering if an H_m feature is statistically significant the null hypothesis assumes that there are no features in any dimension $0, \dots, M-1$. This assumption can be relaxed for the Partitioned Test. We consider a specific H_m feature and test if the mean of the distribution of pixel intensities in the partition that defines it is statistically different from the mean of the distribution of pixel intensities of the rest of the array. The H_m feature to be tested is conditioned on its detection on the persistence diagram and matched to a partition using Algorithm 1 in Section 2.4. For simplicity, we have the same assumptions as in Section 2.1, where if there is structure in the underlying pattern $\mathcal{A}^0$, it is just one H_m feature and that the functional value μ_1 of the hole is the highest mean partition in the image.

The pixel intensities of the partition describing the H_m feature $\mathcal{A}_1^\sigma$ are defined as:

$$\mathcal{A}_1^\sigma = \{\mu_m + \sigma_m^2 | \forall (x_1, x_2, \ldots, x_M) \in \mathcal{G}_1\}, \tag{3.9}$$

where $\mathcal{G}_1$ is the part of the M-dimensional grid where the H_m feature is located. Pixels intensities within the partition $\mathcal{G}_1$ are sampled from the following distribution

$$Z^1 \sim \mathbf{F_1}(\mu_1, \sigma_1^2). \tag{3.10}$$

where μ_1 is the functional value of the H_m feature in $\mathcal{A}^0$ and σ_1^2 is the variance of the sampling distribution of the pixel intensities. We assume that all the pixel intensities within a partition have the same sampling distribution.

The rest of the array is partitioned as background, denoted $(\mathcal{A}_1^\sigma)^C$ (i.e., the complement of the set $\mathcal{A}_1^\sigma$), defined as follows:

$$(\mathcal{A}_1^\sigma)^C = \{f(x_1, x_2, \ldots, x_M) + \varepsilon(x_1, x_2, \ldots, x_M) | \forall (x_1, x_2, \ldots, x_M) \in (\mathcal{G}_1)^C\}, \tag{3.11}$$

where $(\mathcal{G}_1)^C$ contains all the $(x_1, x_2, \ldots, x_M)$ locations in the complement of $\mathcal{G}_1$. The pixels from this distribution have the following distribution:

$$(Z^1)^C(x_1, x_2, \ldots, x_M) \sim \mathbf{F_{1C}}(f(x_1, x_2, \ldots, x_M), \sigma^2(x_1, x_2, \ldots, x_M)), \tag{3.12}$$

The function $f(x_1, x_2, \ldots, x_M) = \mu_{1C}$ is the mean pixel intensity of the background partition where there may be superfluous background features.

Assumption 7. *If there is a feature with pixel intensities distributed as* $\mathbf{F_1}(\mu_1, \sigma_1^2)$, *then the distribution for the pixels rest of the array* $\mathbf{F_{1C}}(f(x_1, x_2, \ldots, x_M), \sigma^2(x_1, x_2, \ldots, x_M))$ *has the condition that* $f(x_1, x_2, \ldots, x_M) \neq \mu_1$.

The null hypothesis for the Partitioned Test is

$$H_{\text{null}} : \mu_{1C} = \mu_1. \tag{3.13}$$

The alternative hypothesis is that the two partitions are centered at different functional values such that $\mu_1 \neq \mu_{1C}$. Since we are doing an upper-level set filtration, for simplicity we assume that if there is a H_m feature that is separate from the background that $\mu_1 > \mu_{1C}$. The focus of this test is on the means since in partitioned TDA, the means of each partition are the estimated birth and death times of a feature.

To test this null hypothesis, we start with a segmentation on the array $\mathcal{A}^{\sigma}$ where the two partitions are denoted as $\mathcal{G}_1$ and $(\mathcal{G}_1)^C$ and there is a H_m feature on the persistence diagram $\mathcal{P}(\mathcal{A}^{\sigma})$ that is within the segmentation $\mathcal{G}_1$. Denote the two samples as $Z^1 = \{Z^1_1, \dots, Z^1_{n_1}\}$ and $(Z^1)^C = \{(Z^1_1)^C, \dots, (Z^1_{n_{1C}})^C\}$ where n_1 (n_{1C}) is the sample size for all the pixels in partition $\mathcal{G}_1$ $((\mathcal{G}_1)^C)$. A two-sample permutation test is used to compute the p-value for the Partitioned Test where there are $j = \{1, \dots, B\}$ permuted samples, denoted as $(Z^1_j)^*$ and $(Z^1_j)^{*C}$.

Let $I = \{1, \dots, n_1, n_1 + 1, \dots, n_1 + n_{1C}\}$ denote the indices for all the pixels which make up the array and let $Z = Z^1 \cup (Z^1)^C$ be all the pixels in the array. Then each permutation j is defined below as:

$$(Z_j)^*(x_1, x_2, \dots, x_M) = Z(x_{i_1}, x_{i_2}, \dots, x_{i_M}) \tag{3.14}$$

where $i_q \in I$ where $q = \{1, \dots, M\}$. The set of all the permuted pixels is denoted as $(Z_j)^* = \{(Z_j)^*(x_1, x_2, \dots, x_M)_1, \dots, (Z_j)^*(x_1, x_2, \dots, x_M)_{n_1+n_{1C}}\}$. Then the permuted pixels which form the partition for the H_m feature are defined as follows:

$$(Z^1_j)^*(x_1, x_2, \dots, x_M) = \{(Z_j)^*(x_1, x_2, \dots, x_M)_{k_1}, \dots (Z_j)^*(x_1, x_2, \dots, x_M)_{k_{n_1}}\} \tag{3.15}$$

where $k_1, \ldots, k_{n_1}$ are indexes of the pixels in $(Z_j)^*$. Then permuted pixels which form the background partition are define as:

$$(Z_j^1)^{*C}(x_1, x_2, \ldots, x_M) = \{(Z_j)^*(x_1, x_2, \ldots, x_M)_{l_1}, \ldots (Z_j)^*(x_1, x_2, \ldots, x_M)_{l_{n_{1C}}}\} \tag{3.16}$$

where $l_1, \ldots, l_{n_{1C}}$ are indexes of the pixels in $(Z_j)^*$ and $\{l_1, \ldots, l_{n_{1C}}\} \cap \{k_1, \ldots, k_{n_1}\} = \emptyset$ and $\{l_1, \ldots, l_{n_{1C}}\} \cup \{k_1, \ldots, k_{n_1}\} = I$. The null distribution $\boldsymbol{\rho}^* = \{\rho_1^*, \ldots, \rho_B^*\}$ is generated by:

$$\rho_j^* = |(\bar{Z}_j^1)^* - (\bar{Z}^1{}_j)^{*C}|, \tag{3.17}$$

where $(\bar{Z}_j^1)^* = \sum_i^{n_1}(Z_j^1)_i^*/n_1$ is the mean of the $(Z_j^1)^*$ defined in Equation (3.15) and $(\bar{Z}^1{}_j)^{*C} = \sum_i^{n_{1C}}(Z_j^1)_i^{1C}/n_{1C}$ is the mean of the $(Z_j^1)^{*C}$ defined in Equation (3.16). The observed test statistic is the mean difference for the partitions in the data defined as:

$$\rho^{obs} = |\bar{Z}^1 - (Z\bar{^1})^C|. \tag{3.18}$$

The p-value can be then be calculated as follows:

$$\text{p-value } = \frac{\sum_{j=1}^{B} I(\rho_j^* \geqslant \rho^{obs})}{B}. \tag{3.19}$$

In Figure 3.2, there is a visual pipeline of the test. The first column shows the raw images that we want to perform the Partitioned Test on: $\mathcal{M}_{t_1}^{C3}$ (row one) and $\mathcal{M}_{t_{12}}^{C3}$ (row two). The next column shows the edge partitions for each image as described in Section 2.6 where the red dots are the pixel locations which form the edges. For the C3 cell at time t_1, no loop in the partition is matched to a loop on the persistence diagram. So, we cannot run the hypothesis test and would stop. However, in the second row at time t_{12}, the pixel locations which make up the loop are highlighted in purple and the background partition is in black. This loop is in our segmentation and detected using partitioned TDA. The histograms for each partition, Z^1 and $(Z^1)^C$, is shown in the third column which is then used to do the Partitioned Test.

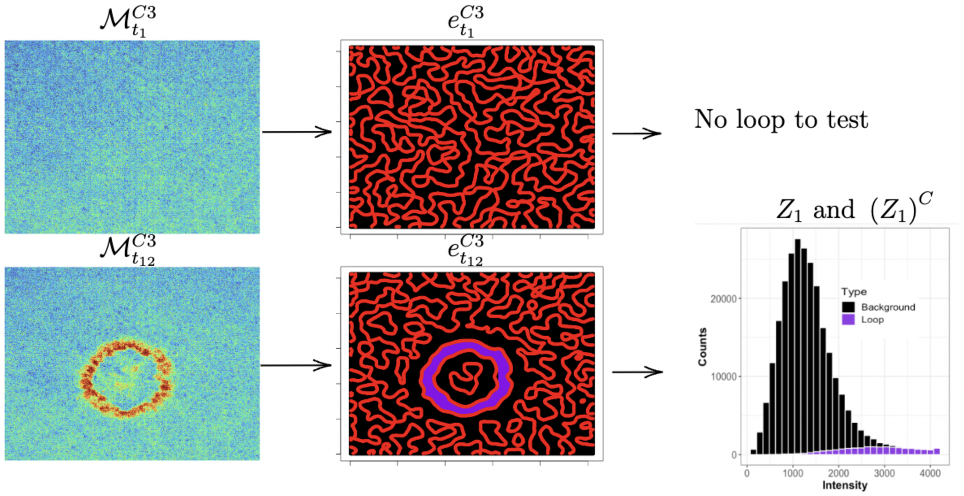


Figure 3.2: Pipeline of the Partitioned Test for $\mathcal{M}_{t_1}^{c3}$ and $\mathcal{M}_{t_{12}}^{c3}$. The first column has the raw images at times t_1 (row 1) and t_{12} (row 2). The second column is segmentation of those images where the red lines highlight the edge set, the black regions are the background partition $(\mathcal{G}_1)^C$, and the purple region is the loop partition $\mathcal{G}_1$ (row 2 only). The last column is the distribution of pixel intensities if a loop in a partition is detected.

In Sections 3.5 and 3.6, Type I error (α) and Type II error (β) are empirically evaluated for the permutation tests proposed in Section 3.3 and Section 3.4. We define empirical methods to estimate α and the power $(1-\beta)$ of the Maximum Persistence Test and the Partitioned Test, separately.

3.5 Empirical Evaluation Maximum Persistence Test

Type I Error

To evaluate the Type I error rate for the Maximum Persistence Test, we generated 1000 images with an array dimension of $M = 2$. We began by examining images with no underlying structure (see an example in Figure 3.3a), where $f(x, y) = \mu_0$ for all $(x, y) \in \mathcal{G}$. Different settings were considered with homoskedastic Gaussian noise added to the images at three different noise levels $\sigma = \{10, 100, 500\}$. When using the Maximum Persistence Test a smoothing parameter must be selected. Therefore, to see how power changes for the Maximum Persistence Test with different choices of smoothing parameters, we picked three different levels span=$\{0.05, 0.1, 0.2\}$; these levels were chosen such that there is still an H_1 feature present. The Maximum Persistence Test was performed on each image for each noise/smoothing level to detect a statistically significant loop (i.e., H_1 feature). Then the estimated Type I error rate was calculated based on how many p-values out of the 1000 for each setting were below $\alpha = 0.05$. Since there is no underlying structure, or loop in the simulated images, we expect around 5% of the p-values to be under 0.05.

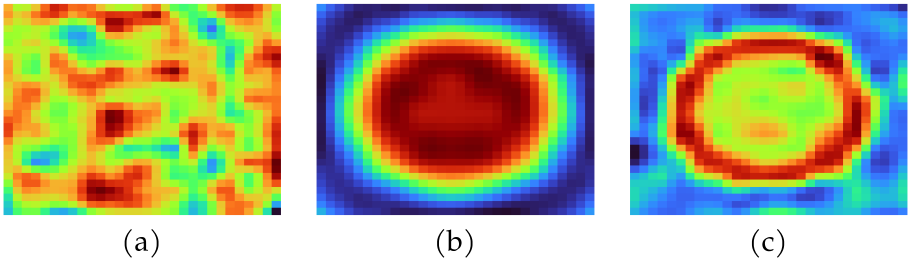


Figure 3.3: Examples of the images used to test if the underlying structure contains a loop (H_1 feature) using the Maximum Persistence Test. (a) A smoothed image with no H_0 or H_1 features to assess Type I error. (b) A smoothed image with one H_0 but no H_1 features to assess Type I error. (c) A smoothed image with an H_1 feature to estimate power.

The estimated Type I error results are presented in Table 3.1 for the different noise and smoothing settings. In general, the Type I error rates are close to 0.05 across all noise and smoothing levels, indicating the Maximum Persistence Test maintains the correct Type I error for these examples.

	span=0.05	**span=0.1**	**span=0.2**
σ=10	0.05	0.05	0.05
σ=100	0.05	0.05	0.07
σ=500	0.05	0.04	0.04

Table 3.1: Type I Error Rate for Maximum Persistence Test to identify statistically significant loops in image with no structure (i.e., when the null hypothesis is true) where $\alpha = 0.05$. Each cell is the empirical percentage of p-values below $\alpha = 0.05$ using 1000 iid images under the different smoothing (span) and noise setting..

Next, we tested 1000 images with an H_0 feature (see an example in Figure 3.3b) but no loops to assess the Type I error in this situation. These images do not fit the null hypothesis of no structure, but we do not want to find statistically significant loops in this scenario. Table 3.2 displays the results when the analysis is repeated for 1000 images containing an H_0 feature and no H_1 features.

	span=0.05	**span=0.1**	**span=0.2**
σ=10	0	0	0
σ=100	0	0	0
σ=500	0	0	0

Table 3.2: Type I Error Rate for Maximum Persistence Test to identify statistically significant loops in image an H_0 feature but no loops where $\alpha = 0.05$.

None of the p-values for all 1000 images in each category detect any significant loops. These images do not fit the null hypothesis of no structure. Therefore, rejecting the null hypothesis of no loops does not directly correspond to a significance level of α. However, the absence of detected loops is not unexpected, as our test statistic focuses solely on H_1 features as opposed to H_0 features.

Statistical Power

Next, we evaluate the statistical power of detecting a statistically significant H_m feature when there is an actual H_m feature in the underlying pattern in an image (see an example in Figure 3.3c). We generate 1000 images for the different noise and smoothing settings when there is one H_1 feature in the underlying pattern. The power test results for the Maximum Persistence Test can be seen in Table 3.3. In general, the test is powerful and robust to the choice of smoothing parameter. In higher noise settings the test is less powerful, especially for lower span values. We only consider images with one loop in the underlying pattern to match our wounded cell application.

	span=0.05	**span=0.1**	**span=0.2**
σ=10	1	1	1
σ=100	1	0.96	1
σ=500	0.026	0.40	0.97

Table 3.3: Power for Maximum Persistence Test to identify statistically significant loops in image with one H_1 feature where $\alpha = 0.05$.

3.6 Empirical Evaluations Partitioned Test

The empirical evaluation of the Type I error rate and power for the Partitioned Test follows a different approach compared to the Maximum Persistence Test. This test is used only on topological features which are both in a partition separate from the rest of the image but also detected on the persistence diagram as a H_m feature using partitioned TDA. When an image is generated from the null distribution the pixels within the image are random.

To test the Type I error, we sampled n pixels from a Gaussian distribution with mean close to the background partition 1000. Then these pixels were randomly split into a partition of size similar to the cell wound with n_1 pixels, and a partition similar in size to the background with n_{1C} pixels.

The permutation test is calculated on these two partitions and repeated 1000 times. The mean pixel intensity in the partition of the H_m feature is increased, which increases the difference between the background and feature mean pixel intensities to evaluate the statistical power. This procedure is repeated on three different noise levels $\sigma = \{10, 100, 500\}$ in order to see how power is robust to noise. Furthermore, the power analysis is run when the partition is correct and then rerun on an incorrect partition. The Partitioned Test does not test if the partition is correct; however if most of the partition contains the H_m feature then the empirical studies suggest the test can still detect the feature.

The results from the simulation are shown in Figure 3.4 where the difference between the mean pixel intensities in the two partitions are on the x-axis (i.e., a mean difference of zero is assessing Type I error and a mean difference greater than zero is estimating power), the line type is for noise level, and the color of the line is for the partition type. As expected, the lower noise levels, larger mean difference (signal), and a correct partition results in more powerful tests. The Type I error rate is around $\alpha = 0.05$ and overall this test is very powerful at higher signal-to-noise levels.

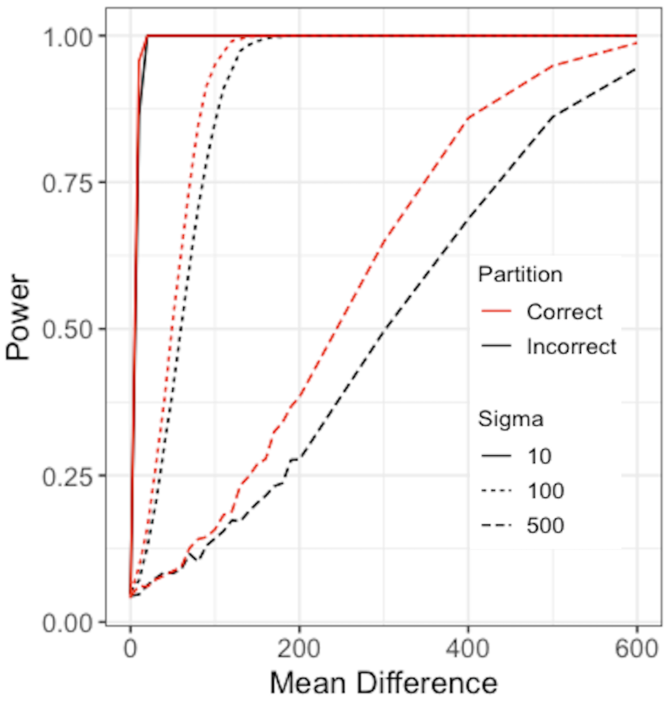


Figure 3.4: Power (y-axis) of the Partitioned Test for levels of difference in the mean of partition of the H_m feature and mean of the partition of the background (x-axis). Power was also tested for different noise levels (linetypes) and different levels of accuracy in the partitions (colors) with the correct partition in red and an incorrect partition in black.

3.7 Hypothesis Tests Applied to Cell Images

Both the Partitioned Test and the Maximum Persistence Test are applied to the cell wound images. There are two time series of images; one for a cell injected with a bacteria called C3 and one wounded but not injected with anything (Control). At each point in time, the tests are applied independent of other images in the time series. The results are shown in Table 4.1 where row one is the Maximum Persistence Test and row two is the Partitioned Test, and the columns are the average p-values (for statistically significant loops) for each time point. The second column has the time points when a statistically significant loop is detected.

The Maximum Persistence Test finds a statistically significant loop for all the time points for both the C3 cell and the Control cell. So, even though there may not be a wound yet in the image, there are still loops which are a part of the cell itself. For example in Figure 3.1, the smoothed image for the C3 cell at time t_1 has several persistent smaller holes which do not appear in the permuted version of the image. In comparison, the Partitioned Test only finds a statistically significant loop at the time points that the partition detects a loop using partitioned TDA. For the C3 cell this would be time t_8 to time t_{28} and for the Control cell this would be time t_7 to t_{30}. Once the Partitioned Test is run we can build confidence regions (see Chapter 2), for the time points which find a statistically significant loop.

Test Type	**Control**	**C3**
Maximum Persistence Test	$t_1 - t_{30}$	$t_1 - t_{30}$
Partitioned Test	$t_7 - t_{30}$	$t_8 - t_{28}$

Table 3.4: Cell image hypothesis test results. The times when a statistically significant H_1 feature is detected using the Maximum Persistence Test (row 1) or Partitioned Test (row 2) are displayed for the Control cell (column 2) and C3 cell (column 3) at an $\alpha = 0.05$ level of significance.

3.8 Conclusion

A persistence diagram is a complex summary statistic, making statistical inference challenging. This complexity has led to the development of significance testing methods for persistence diagrams, offering a more rigorous framework for identifying patterns in images or arrays beyond just confidence regions. Each test outlined in this chapter has distinct advantages and is suited for different scenarios. For instance, the Partitioned Test, which evaluates specific loop partitions in unsmoothed images, can be used in conjunction with the partitioned TDA method. In contrast, the Maximum Persistence Test examines maximum persistent topological features in smoothed images and is employed in Chapter 4 for analyzing temporal patterns across a set of images.

4 INFERENCE FOR TIME SERIES ANALYSIS USING TDA

4.1 Time Series Analysis in TDA

Partitioned TDA estimates the pattern within an image by constructing confidence regions for the birth and death times of detected loops. In the cell biology application discussed in Section 2.9, these loops represent different stages of a cell wound over time. However, partitioned TDA does not inherently link patterns across different time points t_i and t_j, relying instead on prior or scientific knowledge. To address this limitation, we propose a method called the *Maximum Void method*, which uses higher-dimensional homology groups to track the evolving pattern of the wounded cell over time. Though the focus is on an evolving cell wound in an image across time, the framework can apply more generally to other applications.

In recent years, numerous studies have integrated time into TDA analysis, focusing on tasks such as classifying one-dimensional time series (Umeda, 2017; Seversky et al., 2016), quantifying periodic or quasiperiodic behavior (Perea and Harer, 2015; Tralie and Berger, 2018), and analyzing dynamical systems (Tymochko et al., 2020). The methodology in Fasy et al. (2014) was extended to perform inference on one-dimensional time series in Myers et al. (2022). A common technique used in conjunction with TDA and time series data is a sliding window or time delay embedding, which captures the periodicity or dynamics within time series (Perea and Harer, 2015; Tymochko et al., 2020). These embeddings often focus on identifying dynamical properties of topological features in point clouds over time, rather than performing inference on the underlying patterns in the data (Tralie and Berger, 2018; Perea and Harer, 2015; Tymochko et al., 2020). Many of the studies applying TDA to time series data involve classification and prediction tasks due to the geometric and structural insights TDA can provide from complex image data sequences (Topaz et al., 2015; Mata et al., 2015). Additionally, TDA analysis of images or videos often requires converting grayscale images into binary images or point clouds, which can result in the loss of important information.

There are established topological summary statistics for time series data that could be used to estimate the temporal pattern of the cell wound, such as vineyards and CROCKER plots (Cohen-Steiner et al., 2006; Topaz et al., 2015). A CROCKER plot visualizes the evolution of Betti numbers (the cardinality of each homology group H_m) within simplicial complexes over the plane of time and scale parameters. While it lacks stability results, it has proven effective in modeling and classifying dynamic systems (Topaz et al., 2015). As discussed in Section 1.1, persistence diagrams are better representations of the cell wound compared to the Betti numbers. Vineyards, extend persistence diagrams into a time-varying persistence diagram by adding a third dimension of time to the diagrams (Cohen-Steiner et al., 2006).

An alternative approach, presented in Ciocanel et al. (2021); Dawson et al. (2023), analyzes time-varying cell images similar to our dataset. These methods track the persistence of maximum or multiple loops over time and assess statistical significance by generating topological features from a null distribution, where significant features are not expected to occur. This method is more comparable to partitioned TDA but lacks the robust uncertainty quantification provided by partitioned TDA's confidence regions. The primary aim of this section is to estimate the degree of disorganization in wounded cells over time, a task that partitioned TDA does not fully address. In other words, our goal extends beyond merely analyzing patterns through the persistence of loops at each time point. The Maximum Void method integrates functional, geometric, and spatial information to connect topological features across time frames and evaluate their significance.

4.2 Maximum Void Method for Image Time Series

The Maximum Void method consists of four steps which are outlined below and applied to the C3 cell and Control cell images.

Step 1: Add temporal dimension to identify higher dimensional homology groups (H_2 features).

Step 2: Find statistically significant H_2 features which represents the cell wound throughout time.

Step 3: Use the birth time of statistically significant H_2 feature to identify lower dimensional features (H_0 and H_1) which make up the cell wound at each point in time.

Step 4: Connect the lower dimensional features (H_0 and H_1) at consecutive time points.

Step 1: Add Time Dimension

Extending the notation from Section 3.2, the time series of images $\{\mathcal{M}^{\sigma}_{t_1}, \ldots, \mathcal{M}^{\sigma}_{t_{30}}\}$ form an array $\mathcal{A}^{\sigma}$ where t_1 is the first image in the video and t_{30} is the last image. In the cell wounding application, $t_1 = 0$ seconds and a frame of the video is captured every eight seconds. There are 30 total image frames in the video meaning that the last frame is $t_{30} = 240$ seconds. Let $\mathcal{A}^0 = \{f(x, y, t) : (x, y, t) \in \mathcal{G}\}$ denote the noise free array which is described by some function $f(x, y, t)$ discretized onto a 3D grid $(x, y, t) \in \mathcal{G}$, where (x, y) is the spatial location of an image snapshot and t is the time index.

In practice, there is some zero-centered noise $\varepsilon(x, y, z)$ which follows a symmetric distribution $\mathbf{G}(0, \sigma^2(x, y, z))$ added to the function such that data, $\mathcal{A}^{\sigma}$, is defined as follows:

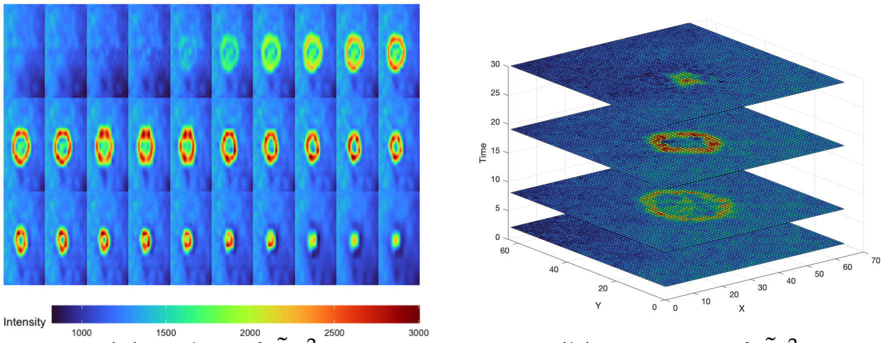
$$\mathcal{A}^{\sigma} = \{f(x, y, t) + \varepsilon(x, y, t) : (x, y, t) \in \mathcal{G}\}, \tag{4.1}$$

where $f(x, y, t)$ is the mean for the pixel intensity in the 3D space location (x, y, t). Each pixel $Z(x, y, t)$ in the array is drawn from the following distribution:

$$Z(x, y, t) \sim \mathbf{G}(f(x, y, t), \sigma^2(x, y, z)). \tag{4.2}$$

When examining the temporal evolution of the pattern $f(x, y, t)$ in the array $\mathcal{A}^\sigma$, the additional time dimension can give rise to higher-dimensional homology groups which connect the H_0 and H_1 features at each point in time. Persistent homology describes the shape and structure of a 2D space X through its connected components $(H_0(X))$ and loops $(H_1(X))$. In addition to these lower dimensional features $(H_0(X)$ and $H_1(X))$, holes such as voids or cavities, which help characterize 3D space X, belong to the two-dimensional homology group $(H_2(X))$.

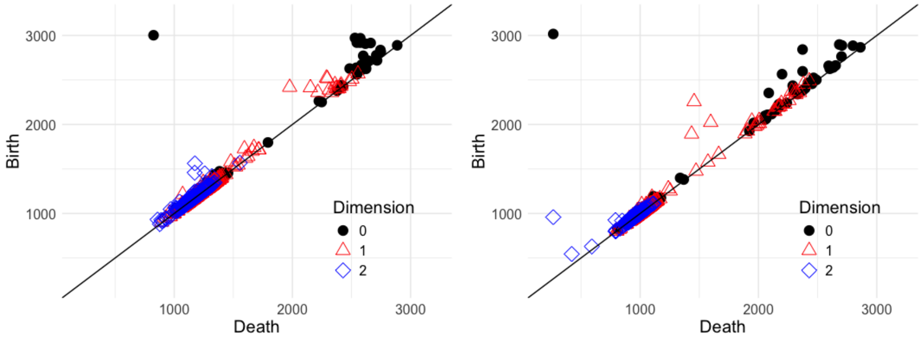
In the time series of cell images, an H_2 feature can form from a loop that emerges shortly after the wounding procedure, evolves gradually over time, and eventually vanishes as the cell undergoes a healing process. An example of the evolving loops can be seen in Figure 4.1 where the times series of smoothed images for the C3 cell $\{\tilde{\mathcal{M}}^{C3}_{t_1}, \ldots, \tilde{\mathcal{M}}^{C3}_{t_{30}}\}$ are shown in time order in Figure 4.1a. For this analysis, the image resolution has been reduced to make computations less intensive. These images are transformed into the array $\tilde{\mathcal{A}}^{C3}$ in Figure 4.1b where only four slices in time of the array are shown for visualization purposes. The H_2 feature has a clear cylindrical form throughout the stack of images due to the cell wounds from the individual time points.



(a) Video of $\tilde{\mathcal{M}}_t^{c3}$ (b) 3D array of $\tilde{\mathcal{A}}_t^{c3}$

Figure 4.1: (a) Time series of images of the C3 cell where each row is a sequence of ten consecutive images (e.g. row one is $t_1 - t_{10}$, row two is $t_{11} - t_{20}$, and row 3 is $t_{21} - t_{30}$). (b) The time series of images transformed into an array $\tilde{\mathcal{A}}^{C3}$ by adding a time dimension on the z-axis.

The initial step in the Maximum Void method for analyzing cell data involves obtaining two arrays, $\tilde{\mathcal{A}}^{C3}$ and $\tilde{\mathcal{A}}^{Con}$, and apply an upper-level set filtration to identify various H_2 features. Figure 4.2 shows the persistence diagrams for these arrays with the C3 cell in Figure 4.2a and the Control cell in Figure 4.2b. The blue diamonds are the H_2 features.



(a) $\mathcal{P}(\tilde{\mathcal{A}}_t^{c3})$ (b) $\mathcal{P}(\tilde{\mathcal{A}}_t^{Con})$

Figure 4.2: Persistence diagrams for (a) $\tilde{\mathcal{A}}^{C3}$ and (b) $\tilde{\mathcal{A}}^{Con}$ where the blue diamonds are H_2 feature, the red triangles are H_1 features, and the black dots are H_0 features. The death time is one the x-axis and the birth time is on the y-axis.

The C3 cell has several persistent H_2 features on the persistence diagram (blue diamonds located further from the birth=death line), whereas the Control cell shows only one prominently persistent H_2 feature.

Step 2: Find Statistically Significant H_2 Feature

The function $f(x, y, t)$ which forms the mean of the pixel intensity sampling distribution and describes the pattern in the arrays can be partitioned as described in the following assumption.

Assumption 8. *For each point in time* $t = \{t_1, \dots, t_{30}\}$ *an image* $\tilde{\mathcal{M}}_t^{\sigma}$ *in the array* $\tilde{\mathcal{A}}^{\sigma}$ *can be partitioned into* k *contiguous regions, where* k *could depend on time (e.g.* k_t*), and each of the* k *partitions has a constant function value defined as:*

$$\theta_k(t) = \{f(x, y, t) : (x, y) \in \mathcal{G}_k(t) \textit{ and } f(x, y) = \tilde{\mu}_k\}, \tag{4.3}$$

where $\mathcal{G}_k(t)$ *is the* (x, y) *coordinates which make up partition* k *at time* t *and* μ_k *is the value of the function for partition* k*.*

For all the times the wound exists $\mathbf{t^*}$, let $\theta_1(t) = \{f(x, y, t) : (x, y) \in \mathcal{G}_1(t), t \in \mathbf{t^*},\text{ and } f(x, y) = \mu_1\}$ be the functional value of the wound and $\mathcal{G}_1(t)$ be the partition of the wound at time t. A partition of $\tilde{\mathcal{A}}^0$, denoted by $\mathcal{G}_1$, contains the entire cell wound throughout time in the form of H_2 features where $\mathcal{G}_1$ is defined as follows:

$$\mathcal{G}_1 = \{\mathcal{G}_1(t) : \forall t \in \mathbf{t^*}\} \tag{4.4}$$

The birth time of the most persistent H_2 feature $(\hat{\theta}_1)$ in the data array which has been smoothed across time $\tilde{\mathcal{A}}^\sigma$, is used to estimate the birth time of the H_2 features which make up the wound in $\mathcal{G}_1$ for the underlying smoothed pattern $\tilde{\mathcal{A}}^0$. This birth time, $\hat{\theta}_1$, is a good estimate of the birth time of the H_2 features which make up the cell wound as long as the following assumptions hold.

Assumption 9. *For every time* $t \in \mathbf{t^*}$ *where the wound exists, the persistence of the corresponding* H_m *feature is higher than the persistence of the background* $(k \neq 1)$ *such that:* $\theta_1(t) > \theta_k(t)$ *if the wound is an* H_0 *feature and* $\theta_1(t) - \theta_{1*}(t) > \theta_k(t) - \theta_{k*}(t)$ *if the wound is an* H_1 *feature, where* θ_{1*} *and* θ_{k*} *denote the partitions in the interior of loop* 1 *and loop* k*, respectively.*

Assumption 10. *For every time* $t \in \mathbf{t}^*$*, the functional value of the partition which describes the wound* $\theta_1(t)$ *is closer to the functional value of the wound at another point in time than it is to the background partition so that:* $\theta_1(t_i) - \theta_1(t_j) < \theta_1(t_i) - \theta_k(t_i)$ *where* $k \neq 1$ *and* $i \neq j$*.*

We assume that the H_2 feature representing the wound corresponds to the maximum persistent feature in the arrays $\tilde{A}^{C3}$ and $\tilde{A}^{Con}$ aligning with Assumptions 9 and 10. Therefore, the second step in the Maximum Void method applies the Maximum Persistence Test on the entire array to identify the possible statistically significant H_2 feature, which represents the cell wound at most points in time, instead of on each image individually. Steps 1-4 of Algorithm 3 applies the Maximum Persistence Test to the C3 and the Control cell arrays $\tilde{A}^{C3}$ and $\tilde{A}^{Con}$, respectively. Using $m = 2$ for the dimension of the H_m feature in the Maximum Persistence Test, each column of Table 4.1 shows the p-value$_{max}$, the 95th percentile of the null distribution $\boldsymbol{\rho}^*_{\mathbf{max}}$, and the persistence of the H_2 feature for both the C3 cell (row 1) and the Control cell (row 2).

Cell Type	**p-value$_{max}$**	**95th percentile ρ^*_{max}**	**persistence**
C3	0	99	388
Control	0	89	693

Table 4.1: Results of the Maximum Persistence Test applied to the smoothed arrays where row is the cell type, C3 and Control, respectively. The columns are the p-value$_{max}$ from the Maximum Persistence Test, the 95th percentile of the null distribution of no structure in the array, and the persistence of the most persistent H_2 feature.

Since the most persistent H_2 features in $\tilde{A}^{C3}$ and $\tilde{A}^{Con}$ are statistically significant, the birth times of these features ($\hat{\theta}_1^{C3}$ and $\hat{\theta}_1^{Con}$) can be used to identify the part of the array where the wound is ($\mathcal{G}_1^{C3}$ and $\mathcal{G}_1^{Con}$).

Step 3: Identify H_0 and H_1 Features at Each Point in Time

The partitions of the cell wound in the arrays for the C3 cell and the Control cell are estimated as follows:

$$\begin{aligned}\hat{\mathcal{G}}_1^{C3}(\hat{\theta}_1^{C3}) &= \{(x,y,t) : f^{-1}(\theta) = (x,y,t) \text{ for } \theta \geqslant \hat{\theta}_1^{C3}\}\\ \hat{\mathcal{G}}_1^{Con}(\hat{\theta}_1^{Con}) &= \{(x,y,t) : f^{-1}(\theta) = (x,y,t) \text{ for } \theta \geqslant \hat{\theta}_1^{Con}\},\end{aligned} \tag{4.5}$$

where only (x,y,t) coordinates from pixels above the birth times of the most persistent H_2 feature are a part of the estimated partitions. These partitions can be broken into slices of the cell wound at each point in time t; let the (x,y) coordinates defining the wound at time t_i be estimated as follows:

$$\begin{aligned}\hat{\mathcal{G}}_1^{C3}(t_i,\hat{\theta}_1^{C3}) &= \{(x,y) : f^{-1}(\theta) = (x,y,t_i) \text{ for } \theta \geqslant \hat{\theta}_1^{C3}\}\\ \hat{\mathcal{G}}_1^{Con}(t_i,\hat{\theta}_1^{Con}) &= \{(x,y) : f^{-1}(\theta) = (x,y,t_i) \text{ for } \theta \geqslant \hat{\theta}_1^{Con}\}.\end{aligned} \tag{4.6}$$

These (x,y) coordinates in $\hat{\mathcal{G}}_1^{C3}(t_i,\hat{\theta}_1^{C3})$ and $\hat{\mathcal{G}}_1^{Con}(t_i,\hat{\theta}_1^{Con})$ are the zero-simplices which form the simpicial complex at each time point ($\mathcal{K}_{t_i}^{\hat{\theta}_1^{C3}}$ and $\mathcal{K}_{t_i}^{\hat{\theta}_1^{Con}}$) which can be used to calculate the homology at each point in time ($H_m(\mathcal{K}_{t_i}^{\hat{\theta}_1^{C3}})$ and $H_m(\mathcal{K}_{t_i}^{\hat{\theta}_1^{Con}})$). The partitions of the wounds in the arrays

($\hat{\mathcal{G}}_1^{C3}(\hat{\theta}_1^{C3})$ and $\hat{\mathcal{G}}_1^{Con}(\hat{\theta}_1^{C3})$) connect the homology of the wound in the C3 and Control cells at each point in time t. The homology of the wound at time t can manifest as an H_0 feature, an H_1 feature, the empty set, or multiple features of both dimensions.

Examples of how the homology of the wound could manifest at three consecutive slices of time is shown in Figure 4.3 with $\mathcal{G}_1^1$ and $\mathcal{G}_1^2$. The grey part of the cylinder are empty regions and the slices are three consecutive time points t_1, t_2, t_3. The colors show which H_1 features are connected in time through the H_2 features. If there is more regularity in the wound (e.g. $\mathcal{G}_1^2$ similar to the Control cell), then the homology at each point in time is easy to connect to the times before and after. For instance, $\mathcal{G}_1^2$ would be defined by one red loop which is a part of the wound at times t_1, t_2, and t_3. However, if the wound is more disorganized (e.g. $\mathcal{G}_1^1$ similar to the C3 cell) and has multiple H_m features which describe it, the homology at each point in time can be difficult to connect to the time before. For instance, $\mathcal{G}_1^1$ is defined by a red loop which is a part of the wound at times t_1, t_2, and t_3, a blue loop which is a part of the wound at time t_2 but then merges with the red loop at time t_3, and a completely separate purple loop which is a part of the wound at times t_2 and t_3.

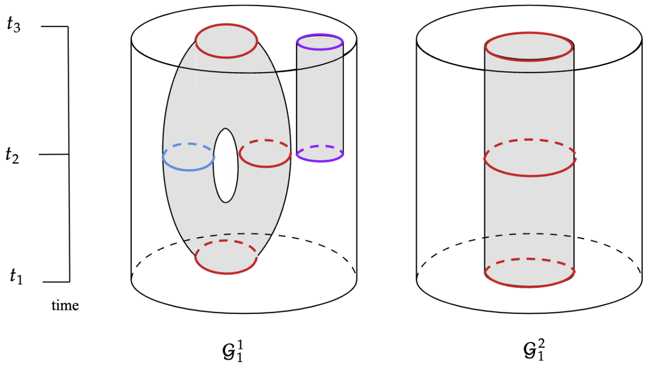


Figure 4.3: Examples of different types of wounds, or $\mathcal{G}_1$s. The loop color groups loops which are connected in time, and the slices are three time points t_1, t_2, t_3. The white region is the H_1 feature and the gray region is empty space or the part of the image which makes up the background.

The third step of the Maximum Void method identifies $\hat{\mathcal{G}}_1^{C3}(\hat{\theta}_1^{C3})$ and $\hat{\mathcal{G}}_1^{Con}(\hat{\theta}_1^{Con})$ from $\hat{\theta}_1^{C3}$ and $\hat{\theta}_1^{Con}$. Then the simplicial complexes at each point in time $t_i \in \mathbf{t^*}$ ($\mathcal{K}_{t_i}^{\hat{\theta}_1^{C3}}$ and $\mathcal{K}_{t_i}^{\hat{\theta}_1^{Con}}$) are found from $\hat{\mathcal{G}}_1^{C3}(t_i, \hat{\theta}_1^{C3})$ and $\hat{\mathcal{G}}_1^{Con}(t_i, \hat{\theta}_1^{Con})$. Step 5 of Algorithm 3 describes how to get the list of simplicial complexes at each point in time $\mathcal{K}_{t_i}^{\hat{\theta}_1^{C3}}$ and $\mathcal{K}_{t_i}^{\hat{\theta}_1^{Con}}$. Examples of $\hat{\mathcal{G}}_1^{C3}(t_i, \hat{\theta}_1^{C3})$ for times t_{11}, t_{12}, t_{13} can be seen in Figure 4.4 and the corresponding simplicial complexes $\mathcal{K}_{t_i}^{\hat{\theta}_1^{C3}}$ for times t_{11}, t_{12}, t_{13} can be seen in Figure 4.5.

Algorithm 3 MV Method for Stacked Images

```
Input: $\mathtt{df} := (x, y, t, Z[x, y, t])$ of array $\mathcal{A}^{\sigma}$; $B$ = number of permutations; $m$ = homology group dimension; $P$ = persistence diagram of $\mathcal{A}^{\sigma}$; $r^{obs} := \max_m(P)$ persistence of most persistent $H_m$ feature; $b$ = birth time of maximum persistent $H_m$ feature; $T$ total number of time points
Output: $S$ list of simplicial complexes for representing the wound at each point in time
Define: $\mathbb{Z} = \{Z[x, y, t] \mid (x, y, t, Z[x, y, t]) \in \mathtt{df}\}$; $p = \emptyset$; $r^* = \emptyset$; $L$ =nrows($\mathbb{Z}$)
for j in 1:B do
    Step 1: Define $A^* \in \mathbb{R}^3$
    for l in 1:L do                                    ▷ permute $Z[x, y, t]$
        $A^*$(x,y,t)$\leftarrow$ sample($\mathbb{Z}$, without replacement)
    end for
    Step 2: $P^* \leftarrow \mathrm{pers}(A^*)$         ▷ calculate Persistence Diagram
    Step 3:
    $r^* \leftarrow \max_m(P^*)$                        ▷ Get max persistence
end for
Step 4: $p \leftarrow \frac{\sum j=1^B \mathbb{I}(r_j^* > r^{obs})}{B}$     ▷ Get p-value for Maximum Persistence Test
Step 5:
if $p > \alpha$ then $S \leftarrow \emptyset$          ▷ If fail to reject null there is no feature
else
    Define $S = \emptyset$; $G = \emptyset$             ▷ If reject null find $(x, y, t)$ coordinates of feature
    for j in 1:T do
        $G_t = \leftarrow \{(x, y, Z[x, y, t]) \mid t = j\}$     ▷ Subset dataframe by each time
        $S_t \leftarrow \{(x, y) \mid (Z[x, y, t] \geqslant b) \cap ((x, y, Z[x, y, t]) \in G_t)\}$     ▷ Get $(x, y)$ for birth of feature
    end for
end if
return S
```

Step 4: Connect the H_1 and H_0 features Across Time

The final step of the Maximum Void method involves connecting and summarizing the homology of $\mathcal{K}_{t_i}^{\hat{\theta}_1}$ and $\mathcal{K}_{t_j}^{\hat{\theta}_1}$ for all the consecutive time points t_i and t_j where the wound exists $(\mathbf{t}^*)$ in both the C3 and Control cells. This is achieved using ZigZag persistence, a technique designed for analyzing time-varying dynamic graphs and clustering dynamic data. ZigZag persistence captures the evolution of H_1 and H_0 features over time by examining how homology changes during the filtration process where δ is time rather than pixel intensity (Carlsson and De Silva, 2010; Mata et al., 2015; Tausz and Carlsson, 2011; Carlsson et al., 2009; Tymochko et al., 2020).

ZigZag Persistence

In traditional persistent homology calculations, the simplicial complexes built using $\mathcal{M}^{-1}(\delta, \infty) = \{(x, y) \in \mathbb{R}^2 | Z(x, y) > \delta\}$ form a *filtration*, or a finite sequence of nested sub-complexes, as δ decreases from ∞ to 0. The topology of each simplicial complex is quantified through homology groups and the δ value when homological feature j is born (b_j) and dies (d_j). Throughout the filtration, each simplicial complex is nested within simplicial complexes on top of upper-level sets with smaller thresholds such that:

$$\mathcal{K}_{\delta_1} \hookrightarrow \mathcal{K}_{\delta_2} \hookrightarrow \ldots \hookrightarrow \mathcal{K}_{\delta_l=0}, \tag{4.7}$$

where $\mathcal{K}_\delta$ is the simplicial complex on the upper level set $\mathcal{M}^{-1}(\delta, \infty)$. The inclusion arrow $\hookrightarrow$ shows that $\mathcal{K}_{\delta_1} \subset \mathcal{K}_{\delta_2}$ for $\delta_1 \geqslant \delta_2$.

The inclusion maps between simplicial complexes induce linear maps between the homology groups of those complexes where m denotes the dimension. All the linear maps point in the same direction.

$$H_m(\mathcal{K}_{\delta_1}) \to H_m(\mathcal{K}_{\delta_2}) \to \ldots \to H_m(\mathcal{K}_{\delta_l=0}). \tag{4.8}$$

This graphical representation of vector spaces $H_m(\mathcal{K})$ is called a *persistence module* (Carlsson and De Silva, 2010).

The mathematical relationship shown above allows for clear interpretation, calculation, and a multi-scale view homology over a space. However, when a space is evolving throughout time, topological features can disappear and reappear making the birth and death time calculations difficult. ZigZag persistence is a technique which describes the persistent homology of a family of space without requiring nesting (Carlsson and De Silva, 2010). This technique is commonly applied to time series data where the filtration is through time, not space, or in our case pixel intensity.

ZigZag persistence is a generalization of quiver theory where the direction of each inclusion map in Equation (4.7) and the linear map in

Equation (4.8) is arbitrary instead of the same direction throughout the filtration. In most applications the inclusion direction alternates earning the name ZigZag Carlsson and De Silva (2010). An example of this structure for a filtration over time (e.g. t=time) is shown below:

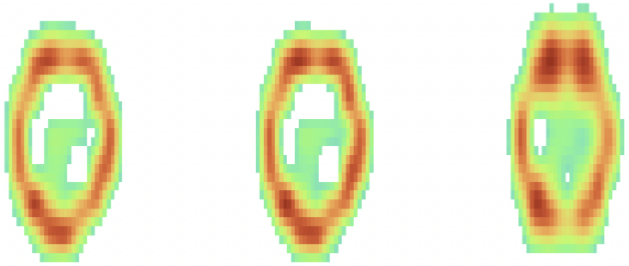
$$\begin{aligned} &\mathcal{K}_{t_1} \hookrightarrow \mathcal{K}_{t_2} \hookleftarrow \mathcal{K}_{t_3} \hookrightarrow \ldots \hookleftarrow \mathcal{K}_{t_{l-1}} \hookrightarrow \mathcal{K}_{t_l} \\ &\text{ or } \\ &\mathcal{K}_{t_1} \hookleftarrow \mathcal{K}_{t_2} \hookrightarrow \mathcal{K}_{t_3} \hookleftarrow \ldots \hookrightarrow \mathcal{K}_{t_{l-1}} \hookleftarrow \mathcal{K}_{t_l}, \end{aligned} \tag{4.9}$$

where t_l is the last point in time.

The inclusion maps induce a linear map on the mth homology group which can now change direction creating new persistence modules:

$$\begin{aligned} &H_m(\mathcal{K}_{t_1}) \rightarrow H_m(\mathcal{K}_{t_2}) \leftarrow H_m(\mathcal{K}_{t_3}) \rightarrow \ldots \leftarrow H_m(\mathcal{K}_{t_{l-1}}) \rightarrow H_m(\mathcal{K}_{t_l}) \\ &\qquad\qquad \text{or} \\ &H_m(\mathcal{K}_{t_1}) \leftarrow H_m(\mathcal{K}_{t_2}) \rightarrow H_m(\mathcal{K}_{t_3}) \leftarrow \ldots \rightarrow H_m(\mathcal{K}_{t_{l-1}}) \leftarrow H_m(\mathcal{K}_{t_l}) \end{aligned} \tag{4.10}$$

To integrate ZigZag persistence into the Maximum Void method in order to interpret the homology of the wound, Algorithm 3 reduces the time series of images $\boldsymbol{\tilde{\mathcal{M}}_t^\sigma} = \{\tilde{\mathcal{M}}_{t_1}^\sigma, \tilde{\mathcal{M}}_{t_2}^\sigma, \ldots, \tilde{\mathcal{M}}_{t_{30}}^\sigma\}$ to simplicial complexes $\boldsymbol{\mathcal{K}_t^{\hat{\theta}_1}} = \{\mathcal{K}_{t_1}^{\hat{\theta}_1}, \mathcal{K}_{t_2}^{\hat{\theta}_1}, \ldots, \mathcal{K}_{t_{30}}^{\hat{\theta}_1}\}$. These simplicial complexes are constructed solely from pixels with intensity values above $\hat{\theta}_1$ and vary in structure over time. Figure 4.4 illustrates slices of the wound for the C3 cell at time points t_{11}, t_{12}, t_{13} where each slice in time may depict the wound as a loop, connected component, or empty space, allowing for linear mappings in diverse directions. We want to induce this alternating structure to the persistence modules of $H_m(\mathcal{K}_t^{\hat{\theta}_1^{C3}})$.



$$H_m(\mathcal{K}_{t_{11}}^{\hat{\theta}_1}) \overset{?}{\longleftrightarrow} H_m(\mathcal{K}_{t_{12}}^{\hat{\theta}_1}) \overset{?}{\longleftrightarrow} H_m(\mathcal{K}_{t_{13}}^{\hat{\theta}_1})$$

Figure 4.4: Slices of the cell wound $\hat{\mathcal{G}}_1^{C3}(t_{11}, \hat{\theta}_1^{C3})$, $\hat{\mathcal{G}}_1^{C3}(t_{12}, \hat{\theta}_1^{C3})$, $\hat{\mathcal{G}}_1^{C3}(t_{13}, \hat{\theta}_1^{C3})$ where color is the intensity value of the pixels which are above the threshold $\hat{\theta}_1^{C3}$. The nesting structure of the homology of the wound throughout times t_{11}, t_{12}, t_{13} is arbitrary.

Since, there is no natural zigzag order to the maps, this relationship can be induced through unions and intersections of the simplicial complexes. This helps connect features throughout time which are disappearing, merging, separating, or staying the same.

$$H_m(\mathcal{K}_{t_1}^{\hat{\theta}_1}) \rightarrow H_m(\mathcal{K}_{t_1}^{\hat{\theta}_1} \cup \mathcal{K}_{t_2}^{\hat{\theta}_1}) \leftarrow H_m(\mathcal{K}_{t_2}^{\hat{\theta}_1}) \rightarrow \ldots \leftarrow H_m(\mathcal{K}_{t_{l-1}}^{\hat{\theta}_1} \cup \mathcal{K}_{t_l}^{\hat{\theta}_1}) \rightarrow H_m(\mathcal{K}_{t_l}^{\hat{\theta}_1})$$

or

$$H_m(\mathcal{K}_{t_1}^{\hat{\theta}_1}) \leftarrow H_m(\mathcal{K}_{t_1}^{\hat{\theta}_1} \cap \mathcal{K}_{t_2}^{\hat{\theta}_1}) \rightarrow H_m(\mathcal{K}_{t_2}^{\hat{\theta}_1}) \leftarrow \ldots \rightarrow H_m(\mathcal{K}_{t_{l-1}}^{\hat{\theta}_1} \cap \mathcal{K}_{t_l}^{\hat{\theta}_1}) \leftarrow H_m(\mathcal{K}_{t_l}^{\hat{\theta}_1}) \tag{4.11}$$

The choice between intersections and unions did not significantly alter the analysis; therefore, the remainder of the Chapter will primarily focus on unions as the set operation that connects homology over time. An example of these unions between slices of the statistically significant H_2 feature representing the wound of the C3 cell, is illustrated in Figure 4.5 for times t_{11}, t_{12}, t_{13}. The color of each loop indicates the grouping across time.

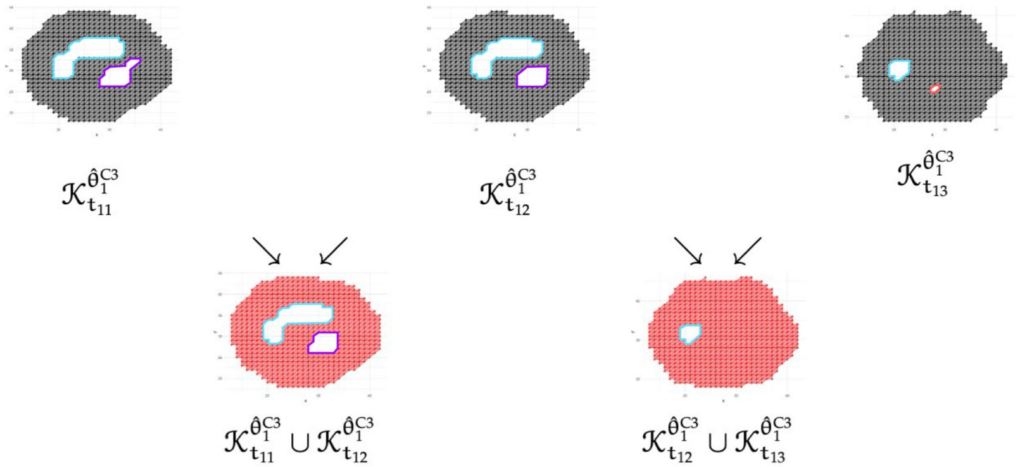
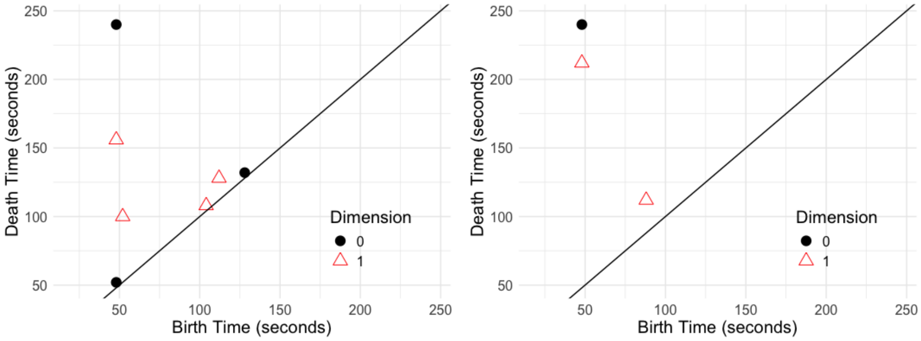

$\mathcal{K}_{t_{11}}^{\hat{\theta}_1^{C3}}$
$\mathcal{K}_{t_{12}}^{\hat{\theta}_1^{C3}}$
$\mathcal{K}_{t_{13}}^{\hat{\theta}_1^{C3}}$
$\mathcal{K}_{t_{11}}^{\hat{\theta}_1^{C3}} \cup \mathcal{K}_{t_{12}}^{\hat{\theta}_1^{C3}}$
$\mathcal{K}_{t_{12}}^{\hat{\theta}_1^{C3}} \cup \mathcal{K}_{t_{13}}^{\hat{\theta}_1^{C3}}$

Figure 4.5: The black simplicial complexes are the simplicial complexes representing the wound at time points t_{11}, t_{12}, t_{13} and the red simplicial complexes are the unions between simplicial complexes representing the wounds at two consecutive time points for the C3 cell. The loop color groups H_1 features through the time slices.

These set operations introduce new times points $\mathcal{K}_{t_{11}}^{\hat{\theta}_1^{C3}} \cup \mathcal{K}_{t_{12}}^{\hat{\theta}_1^{C3}}$ and $\mathcal{K}_{t_{12}}^{\hat{\theta}_1^{C3}} \cup \mathcal{K}_{t_{13}}^{\hat{\theta}_1^{C3}}$, denoted as $t_{11.5}$, and $t_{12.5}$, which link the simplicial complexes and help interpret the evolution of the wound over the time series. The light blue loop persists in the filtration at all time points: $t_{11}, t_{11.5}, t_{12}, t_{12.5}, t_{13}$. In contrast, the purple loop is present at time points $t_{11}, t_{11.5}, t_{12}$ and disappears (i.e. dies) at time $t_{12.5}$, as it is not part of $\mathcal{K}_{t_{12}}^{\hat{\theta}_1^{C3}} \cup \mathcal{K}_{t_{13}}^{\hat{\theta}_1^{C3}}$. The red loop appears (i.e. is born) at time t_{13}, appearing for the first time in $\mathcal{K}_{t_{13}}^{\hat{\theta}_1^{C3}}$. The union at time $t_{12.5}$ demonstrates that the red loop is not the same at the purple loop since these H_1 features are in different locations on the simplicial complex representing the wound.

The Maximum Void method differs from traditional approaches to applying TDA to time series, specifically time series of images, in two significant ways: (i) it connects time through the data space not through topological summary statistics and (ii) it uses a statistical test with $\hat{\theta}_1$ to threshold the image, thereby generating a simplicial complex at each time point for ZigZag persistence analysis. Most often ZigZag persistence is applied to networks or simplicial complexes built on point clouds. In many TDA applications on images, thresholding is typically based on criteria not grounded in a statistical test.

In Figure 4.6, the ZigZag diagrams are shown where the Birth (x-axis) and Death (y-axis) times are displayed in seconds (e.g. $t_{12} = 96$ seconds). The C3 cell is shown in Figure 4.6a and the Control cell is shown in Figure 4.6b. In general, the Control cell is more organized and consistent with only two loops which make up the wound. The C3 cell has much more disorganization in the wound with multiple loops and connected components making up the wound at different points in time.



(a) ZigZag $\mathcal{P}(\tilde{\mathcal{A}}^{c3})$ (b) ZigZag $\mathcal{P}(\tilde{\mathcal{A}}^{\text{Con}})$

Figure 4.6: ZigZag diagrams for the array of the C3 cell (a) and the array of the Control cell (b) where the birth time (in seconds) is on the x-axis and the death time (in seconds) is on the y-axis. The red triangles are the H_1 features and the black dots are the H_0 features.

4.3 Conclusion

For the wounded cell example, the Maximum Void method captures the pattern in an sequence of images by tracking changes in the H_1 and H_0 features that constitute the wound across time. In Section 2.9, partitioned TDA found that the C3 cell exhibits higher persistence in the ring structure of the wound compared to the Control cell at earlier time points. However, this relationship changed at later time points, where the ring of the wound in the Control cell became more persistent. This new method shows that the C3 cell has more topological features representing the wound throughout the time series compared to the Control cell, as illustrated in Figure 4.6a. The Maximum Void method provides additional information for quantifying cell wound patterns and approaches time series analysis from a different perspective than traditional TDA methods.

5 DISCUSSION AND CONCLUSION

This dissertation presents novel methods for inference on both single images and sequences of images evolving over time through the use of TDA. In Chapter 2, we introduced the partitioned TDA approach, which improves confidence regions for topological summary statistics by estimating the mean and variance of partitions associated with the birth and death times of homology group generators. This method, compared to traditional TDA and smooth TDA, provides more accurate coverage, smaller confidence regions, and unbiased estimates of the birth and death times of homology group generators.

The Maximum Persistence Test and Partitioned Test, introduced in Chapter 3, are designed to identify statistically significant topological features. These hypothesis tests allow for the separation of topological signal from noise on the persistence diagrams which the confidence regions in partitioned TDA do not provide. To further improve partitioned TDA which does not connect patterns across different time points Chapter 4 introduces the Maximum Void method. This method tracks evolving patterns over time using higher-dimensional homology groups, the Maximum Persistence Test, and ZigZag persistence.

All of the methods in this dissertation were applied to cell biology data to differentiate between the wounded C3 and Control cells based on persistence and the amount of topological features of the wound over time. These techniques offer useful insights for estimating and quantifying uncertainty in cellular patterns. Future research could extend partitioned TDA to point-cloud data or images where the distribution of pixels within the partition of the loop changes across the partition. With the extension of partitioned TDA to point cloud data, our method can be more directly compared to Fasy et al. (2014). To address difference in sampling distributions on the partition of the feature, a local partitioned TDA approach could use neighboring pixel intensities to estimate the means and variances of the birth and death times of the feature. Additionally, the hypothesis tests could be extended to situations involving the task of finding multiple statistical significant topological features. And lastly, the Maximum Void Method could also be generalized for time series without H_2 features by using pairwise arrays instead of a complete array across all time points. Furthermore, improvements could be made to reduce computational complexity and integrate uncertainty estimates into ZigZag diagrams.

In summary, this dissertation addresses gaps in the TDA literature regarding inference on a single image. While much of the TDA research has traditionally focused on tasks outside of estimation, this dissertation introduces novel insights into addressing biases and quantifying uncertainty in birth and death times of topological features. For example, we shift from estimating the distribution of distances between persistence diagrams to using summary statistics derived from pixel distributions to estimate the true persistence diagram. Additionally, our hypothesis tests and method to connect of shapes in images temporally were performed directly within the images, rather than solely on persistence diagrams. These techniques demonstrate compelling, new approaches to performing inference on persistence diagrams.

DISCARD THIS PAGE

COLOPHON

As with all academic work, "I stand on the shoulders of giants." As such this work could not have been made possible without William C. Benton creating this template.

REFERENCES

A., Uthamacumaran. 2020. Cancer: A turbulence problem. *Neoplasia* 22(12):759–769.

Abdallah, Hassan, Adam Regalski, Mohammad Behzad Kang, Maria Berishaj, Nkechi Nnadi, Asadur Chowdury, Vaibhav A Diwadkar, and Andrew Salch. 2023. Statistical inference for persistent homology applied to simulated fmri time series data. *Foundations of Data Science* 5(1):1–25.

Barbier, Içvara, Hadiastri Kusumawardhani, and Yolanda Schaerli. 2022. Engineering synthetic spatial patterns in microbial populations and communities. *Current Opinion in Microbiology* 67.

Bement, William M, Andrew B Goryachev, Ann L Miller, and George von Dassow. 2024. Patterning of the cell cortex by rho GTPases. *Nat Rev Mol Cell Biol* 25:290–308.

Bement, William M, Ann L Miller, and George von Dassow. 2022. Rho GTPase activity zones and transient contractile arrays. *Bioessays* 28(10): 983–93.

Boissonnat, Jean-Daniel, Frédéric Chazal, and Mariette Yvinec. 2018. *Geometric and topological inference*, vol. 57. Cambridge University Press.

Bukkuri, Anuraag, Noemi Andor, and Isabel Darcy. 2021. Applications of topological data analysis in oncology. *Frontiers in Artificial Intelligence* 4:659037.

Burkel, Brian M, Helene A Benink, Emily M Vaughan, George von Dassow, and William M Bement. 2012. A rho GTPase signal treadmill backs a contractile array. *Developmental cell* 23(2):384–396.

Canny, John F. 1986. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-8:679–698.

Carlsson, Gunnar, and Vin De Silva. 2010. Zigzag persistence. *Foundations of computational mathematics* 10:367–405.

Carlsson, Gunnar, Vin De Silva, and Dmitriy Morozov. 2009. Zigzag persistent homology and real-valued functions. In *Proceedings of the twenty-fifth annual symposium on computational geometry*, 247–256.

Chazal, Frédéric, and Bertrand Michel. 2021. An introduction to topological data analysis: Fundamental and practical aspects for data scientists. *Frontiers in Artificial Intelligence* 4.

Chung, Moo K., Peter Bubenik, and Peter T. Kim. 2009. Persistence diagrams of cortical surface data. In *Information processing in medical imaging, 21st international conference, ipmi 2009, williamsburg, va, usa, july 5-10, 2009. proceedings*, vol. 5636 of *Lecture Notes in Computer Science*, 386–397. Springer.

Ciocanel, Maria-Veronica, Riley Juenemann, Adriana T. Dawes, and Scott A. McKinley. 2021. Topological data analysis approaches to uncovering the timing of ring structure onset in filamentous networks. *Bulletin of Mathematical Biology* 83(3).

Cohen-Steiner, David, Herbert Edelsbrunner, and John Harer. 2005. Stability of persistence diagrams. vol. 37, 263–271.

Cohen-Steiner, David, Herbert Edelsbrunner, and Dmitriy Morozov. 2006. Vines and vineyards by updating persistence in linear time. In *Proceedings of the twenty-second annual symposium on computational geometry*, 119–126.

Dawson, Madeleine, Carson Dudley, Sasamon Omoma, Hwai-Ray Tung, and Maria-Veronica Ciocanel. 2023. Characterizing emerging features in cell dynamics using topological data analysis methods. *Mathematical Biosciences and Engineering* 20(2):3023–3046.

Edelsbrunner, Herbert, and John Harer. 2010. *Computational topology - an introduction.* American Mathematical Society.

Fasy, Brittany Terese, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, and Aarti Singh. 2014. Confidence sets for persistence diagrams. *The Annals of Statistics* 42(6).

Gholizadeh, Shafie, and Wlodek Zadrozny. 2018. A short survey of topological data analysis in time series and systems analysis. *arXiv preprint arXiv:1809.10745*.

Gupta, Saumya, Yikai Zhang, Xiaoling Hu, Prateek Prasanna, and Chao Chen. 2023. Topology-aware uncertainty for image segmentation. `2306.05671`.

Haglund, Kaisa, Ioannis P. Nezis, and Harald Stenmark. 2019. Structure and functions of stable intercellular bridges formed by incomplete cytokinesis during development. *Communicative & Integrative Biology* 4(1): 1–9.

Herron, John Cody, Shiqiong Hu, Bei Liu, Takashi Watanabe, Klaus M. Hahn, and Timothy C. Elston. 2022. Spatial models of pattern formation during phagocytosis. *PLOS Computational Biology* 18.

Madamanchi, Aasakiran, Mary C Mullins, and David M Umulis. 2021. Diversity and robustness of bone morphogenetic protein pattern formation. *Development* 148(7).

Mandato, Craig A., and William M. Bement. 2001. Contraction and polymerization cooperate to assemble and close actomyosin rings around xenopus oocyte wounds. *The Journal of Cell Biology* 154:785 – 798.

Mata, Gadea, Miguel Morales, Ana Romero, and Julio Rubio. 2015. Zigzag persistent homology for processing neuronal images. *Pattern Recognition Letters* 62:55–60.

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

Myers, Audun D., Firas A. Khasawneh, and Brittany T. Fasy. 2022. Anapt: Additive noise analysis for persistence thresholding. *Foundations of Data Science* 4(2):243–269.

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

Paine, I.S., and M.T. Lewis. 2017. The terminal end bud: the little engine that could. *J Mammary Gland Biol Neoplasia* 22:93–108.

Parker, James R. 2010. *Algorithms for image processing and computer vision*. New York: John Wiley & Sons, Inc.

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

Pollard, Thomas D., and Ben O'Shaughnessy. 2019. Molecular mechanism of cytokinesis. *Annual Review of Biochemistry* 88(1):661–689.

Pringle, Robert M., and Corina E. Tarnita. 2017. Spatial self-organization of ecosystems: Integrating multiple mechanisms of regular-pattern formation. *Annual Review of Entomology* 62(1):359–377.

Rabadan, Raul, and Andrew J. Blumberg. 2019. *Topological data analysis for genomics and evolution: Topology in biology*. Cambridge University Press.

Ravishanker, N, and R Chen. 2019. Topological data analysis (tda) for time series. arxiv 2019. *arXiv preprint arXiv:1909.10604*.

Robinson, Andrew P., and Katharine Turner. 2017. Hypothesis testing for topological data analysis. *Journal of Applied and Computational Topology* 1: 241–261.

Seversky, Lee M, Shelby Davis, and Matthew Berger. 2016. On time-series topological data analysis: New data and opportunities. In *Proceedings of the ieee conference on computer vision and pattern recognition workshops*, 59–67.

Singh, Yashbir, Colleen M. Farrelly, Quincy A. Hathaway, Tim Leiner, Jaidip Jagtap, unnar E. Carlsson, and Bradley J. Erickson. 2023. Topological data analysis in medical imaging: current state of the art. *Insights Imaging* 14(58).

Skaf, Yara, and Reinhard Laubenbacher. 2022. Topological data analysis in biomedicine: A review. *Journal of Biomedical Informatics* 130:104082.

Tausz, Andrew, and Gunnar Carlsson. 2011. Applications of zigzag persistence to topological data analysis. *arXiv preprint arXiv:1108.3545*.

Topaz, Chad M, Lori Ziegelmeier, and Tom Halverson. 2015. Topological data analysis of biological aggregation models. *PloS one* 10(5):e0126383.

Tralie, Christopher J, and Matthew Berger. 2018. Topological eulerian synthesis of slow motion periodic videos. In *2018 25th ieee international conference on image processing (icip)*, 3573–3577. IEEE.

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

Tymochko, Sarah, Elizabeth Munch, and Firas A Khasawneh. 2020. Using zigzag persistent homology to detect hopf bifurcations in dynamical systems. *Algorithms* 13(11):278.

Umeda, Yuhei. 2017. Time series classification via topological data analysis. *Information and Media Technologies* 12:228–239.

Wang, Jinyu, Kun Meng, and Fenghai Duan. 2023. Hypothesis testing for medical imaging analysis via the smooth euler characteristic transform.