Robust Statistical Methods for Topological Data Analysis of Time Series

by

Sixtus Dakurah

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

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2025

Date of final oral examination: 05/01/2025

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

Jessi Cisewski-Kehe, Associate Professor, Statistics
Reid C. Van Lehn, Associate Professor, Chemical and Biological Engineering
Christopher J. Geoga, Assistant Professor, Statistics
Cécile Ané, Professor, Statistics
Chunming Zhang, Professor, Statistics

To my beloved brother, Clive Dakurah. You were the best, and your belief in me made this possible. Your memory continues to guide and inspire everything I do. This is for you.

ACKNOWLEDGMENTS

I am deeply grateful to my advisor, Jessi Cisewski-Kehe, for her invaluable guidance, mentorship, and unwavering support throughout my doctoral journey. Working with Jessi has been one of the most formative and rewarding aspects of my academic life. Her thoughtful advice and steady encouragement have shaped both the research presented in this dissertation and my ability to communicate its story with clarity and purpose.

I would like to acknowledge my committee members, Reid Van Lehn, Christopher Geoga, Cécile Ané, and Chunming Zhang, for their valuable feedback and support. Whether through discussions directly related to this dissertation or in other supportive ways, each of you has contributed meaningfully to this journey. I am especially thankful to Reid Van Lehn for his guidance during the collaboration that led to Chapter 5, and to the broader research group he is part of for welcoming me into their work. In particular, I acknowledge Victor Zavela, Rose Cersonsky, Matthew Gabbie, Fang Liu, and all the students and postdocs in the group for their contributions. Special thanks go to Lisa Je for running the experiments that generated the ionic liquid data used in parts of this dissertation.

I am grateful to Vijay Anand for introducing me to topological data analysis and, perhaps just as importantly, for all the free coffee. Thanks also to Moo Chung for mentoring me early on, and to Zijian Chen, Tahmineh Azizi, and Soumya Das, for always lending me an ear and helping me

in various ways during this journey. I appreciate Anna Liu's mentorship during my master's, and Krista Gile's help with my PhD applications. Maryclare Griffin, thank you for steering me toward Madison and for your continued guidance. I am also grateful to Collins Abaitey, Vincent Dedu, and Solomon Amoani, who supported me at critical moments as I navigated life after my bachelor's. To Henrietta, Bright, and Barbara, your enduring friendship and support made the COVID years not only bearable but meaningful, and you all have contributed in significant ways to making this journey successful.

I thank the statistics department for fostering a warm and stimulating environment. Special thanks to Nicolás Trillos for always inviting me to all the Thanksgivings, and Kris Sankaran for always supporting my recommendation requests. I also appreciate the outstanding administrative staff: John Schuppel, Dan Barnish, Nancy Brinkerhoff, and Jodie Nowakowski.

This research was supported in part by funding from the National Science Foundation and the Wisconsin Alumni Research Foundation. I am thankful to the U.S. taxpayers for making this work possible.

Finally, I thank my parents, Francis and Christina, for instilling in me a love for learning, and my siblings - Mildred, Prudence, Adeline, Edna, and Innocent, and the little ones - Rusticus, Rosaline, Nora, and Pristine - your support and companionship helped me get here. To my partner, Gifty, thank you for walking this journey with me. You've made the difficult moments easier and the joyful ones even better.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Topological data analysis seeks to uncover and characterize different topological features including connected components, loops, voids, in data. These topological features are characterized, in part, by how long they persist across different scales, and these multiscale features are summarized on a persistence diagram. One important problem is how features of topological spaces from sampled data can be used to study the underlying data-generating space. Unfortunately, perturbations due to irregular sampling, noise, outliers, and domain-specific complexity can result in many additional features that do not reflect true topological structures.

This dissertation presents methodological innovations designed to enhance the robustness of topological data analysis and enable improved statistical inference on topological features. First, a new data embedding method for constructing point cloud from irregularly-spaced time series data is introduced and shown to preserve the original state space topology in the presence of noise and varying levels of irregularity in the spacing of the time series. Second, a robust statistical inference framework is developed to assess the statistical properties of topological features, specifically the maximal persistence (longest-lived) features. This framework provides a precise quantification of statistically significant topological features without systematically reducing the strength of topological signals, a shortcoming in many existing robust inference techniques. Next, the embedding

method is applied to classify irregularly sampled radial velocity time series for exoplanet detection, where stellar activity and noise complicate the analysis. Reformulating the task as a classification problem, the embedded representation achieves strong discriminative performance even under high missingness and noise. This demonstrates the method's effectiveness in recovering dynamical information from incomplete observations, with practical relevance to astronomy and other domains involving irregular time series. Finally, we investigate the nanostructure variations in ionic liquids from molecular dynamics simulations by coupling topological and statistical techniques. Specifically, by treating a sequence of experimental ionic liquid data spaces as time series, topological methods are employed to extract interpretable nanoscale structural information and detect transition in ionic organizations. This demonstrates how robust and stable topological methods can offer insights into complex real-world systems. These methodological innovations demonstrate both substantial improvement in robustness over existing methods when handling irregular data, enhanced statistical inference for persistent features under perturbations, and broad applicability across various scientific domains.

1 INTRODUCTION

1.1 Background

Topological data analysis (TDA) has emerged as a powerful framework for extracting qualitative insights from complex datasets. Central to TDA is persistent homology, a mathematical tool that identifies and tracks topological features such as connected components, loops, voids, and their higher-dimensional analogs across multiple scales (Edelsbrunner et al., 2000; Edelsbrunner and Harer, 2022). By varying a scale parameter, such as a distance threshold in a point cloud, persistent homology produces a compact summary, often in the form of a persistence diagram, that encodes the "shape" of the data. The power of persistent homology lies in its ability to extract meaningful insights about the shape and structure of data without imposing restrictive assumptions.

Broadly, TDA serves two complementary goals. One is to use topological features to study the data generating space underlying sampled observations (Carlsson et al., 2008; Perea and Harer, 2015; Xu et al., 2019; Pike et al., 2020). The other is to extract or provide representation of topological features for use in downstream data analysis tasks (Turner et al., 2014; Cang and Wei, 2017; Berry et al., 2020). This dissertation began as an effort to explore a more robust representation for studying the data-generating space of sampled observations, specifically, time series data. Time series data is one of the most prevalent forms of structured data, and there is

considerable interest in analyzing its underlying geometric and topological properties (Brown and Knudson, 2009; Emrani et al., 2014; Perea and Harer, 2015; Tralie and Perea, 2018). A common approach in TDA is to represent time series in a multi-dimensional space using time-delay embedding, which reconstructs the data-generating space (state-space) of the time series, facilitating topological characterization of the space (Takens, 2006; Perea and Harer, 2015). This multi-dimensional transformation using time-delay embedding only works for time series observations that are uniformly-spaced in time, limiting its applicability.

In Chapter 2, we propose a subsequence method for constructing this multi-dimensional representation of irregularly-spaced time-series data that preserves certain properties of the reconstructed state space. We show that the proposed method preserves the topological features of the original underlying state space of the time series while reducing spurious shape features. Chapter 3 then develops a robust statistical inference method: "Maximal TDA" (MaxTDA) for topological features. We demonstrate that MaxTDA enhances the statistical significance of topological features by mitigating the reduction in persistence, an artifact of existing robust methods. Chapter 4 presents an application of the proposed subsequence embedding method to the classification of radial velocity time series for exoplanet detection. This chapter demonstrates that subsequence embeddings preserve essential dynamical structure and outperform imputation-based approaches, even under substantial noise and missingness. Finally, Chap-

ter 5 demonstrates an application of the combination of topological and statistical methods to ionic liquid data from molecular dynamics simulations. In this application, topological features are extracted from the coordinate representation of ionic liquids. These topological features are used to construct statistical models to study the nanostructure variation and detect transition in ionic organizations across experimental conditions. Together, these contributions demonstrate substantial improvements in robustness over existing methods for handling irregular data, more reliable statistical inference for persistent features under perturbations, and broad applicability across diverse scientific domains.

The next section provides background on TDA, introduces one of its most widely used tools, persistent homology, and reviews the foundational concepts behind it.

1.2 Homology of Simplicial Complexes

Homology is an area of mathematics that looks for holes in a topological space, and persistent homology looks for holes in data. These holes are formalized through concepts from algebraic topology and are represented by homology groups of varying dimensions (Hatcher et al., 2002; Edelsbrunner and Harer, 2022). Specifically, the zero-dimensional homology group (H_0) contains connected components (clusters), the one-dimensional homology group (H_1) contains loops, the two-dimensional homology group

(H_2) contains voids like the interior of a balloon, and more generally, the k-dimensional homology group (H_k) represents k-dimensional holes. In this work, we mainly represent topological spaces with simplicial complexes. A k-simplex $C = [\mathbf{v}_0, \cdots, \mathbf{v}_k]$ is a k-dimensional polytope of $k+1$ affinely independent points $\mathbf{v}_0, \cdots, \mathbf{v}_k$. A simplicial complex $\mathbb{C}$ is a finite set of simplices such that for any simplices $C^1, C^2 \in \mathbb{C}$, $C^1 \cap C^2$ is a face of both simplices, or the empty set; and a face of any simplex $C \in \mathbb{C}$ is also a simplex in $\mathbb{C}$. (A face of a simplex is the convex hull of any non-empty subset of points that define the simplex.) The homology is computed from these simplicial complexes built along a sequence of filtration values.

1.3 Persistent Homology on Point Clouds

The underlying topological space is often only indirectly observed through noisy point cloud data sampled from it. A common approach to constructing simplicial complexes in TDA for point clouds is the Vietoris-Rips (VR) complex (Vietoris, 1927; Edelsbrunner and Harer, 2022). A VR complex is constructed over a finite set of points $\mathbf{S} = \{\mathbf{v}_0, \mathbf{v}_1, \cdots, \mathbf{v}_n\}$ using a distance parameter δ. For any subset of k points $\{\mathbf{v}_{i_1}, \cdots, \mathbf{v}_{i_k}\}$, a $(k-1)$-dimensional simplex is formed when the pairwise Euclidean distance between all points is at most δ. A collection of all such simplices forms the VR complex denoted as $VR(\mathbf{S}, \delta)$. The composition of the simplicial complex progresses hierarchically with the distance parameter δ. This leads to the concept

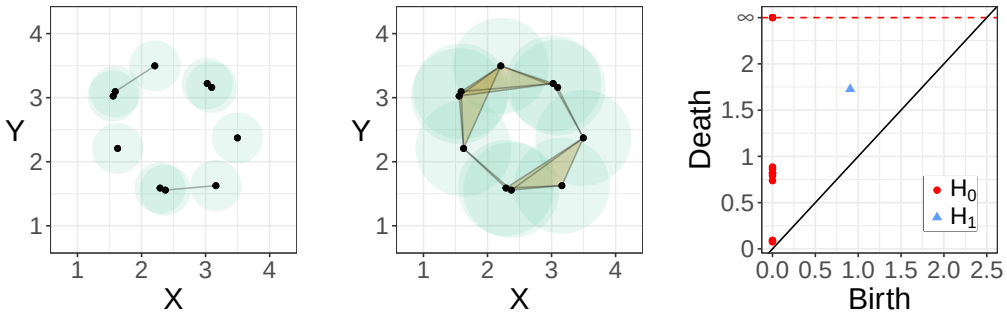
of filtration, which defines an inclusion relation between the simplicial complexes for a set of δ values. More formally, for an ordered sequence of δ values: $0 < \delta_1 < \delta_2 < \cdots < \delta_q < \infty$, the VR complexes admit a nested structure as

$$VR(\mathbf{S}, 0) \subset VR(\mathbf{S}, \delta_1) \subset \cdots \subset VR(\mathbf{S}, \delta_q) \subset VR(S, \infty). \tag{1.1}$$

The inclusion relation between the VR complexes induces a map between the k-dimensional homology groups as

$$H_k(VR(\mathbf{S}, 0)) \to H_k(VR(\mathbf{S}, \delta_1)) \to \cdots \to H_k(VR(\mathbf{S}, \infty)). \tag{1.2}$$

The notion of persistent homology is developed through these homology maps by tracking the changes in the features (i.e., homology group generators) of these nested homology groups. The birth time and death time of features along this sequence encodes topological changes in the groups. For a homology group H_k, we denote the birth time and death time of the j-th feature by b_j and d_j, respectively. The persistence of the feature is given by $d_j - b_j$, and longer persistence often is considered to be topological signal while shorter persistence often represents topological noise (Fasy et al., 2014). If we let k_j to be the homology group dimension of the j-th feature, and J the index set of the features of the homology



(a) Point set with $\delta = 0.8$ (b) Point set with $\delta = 1.5$ (c) Persistence diagram

Figure 1.1: VR filtration and persistence diagram. The zero-simplices (black points, a-b) sampled randomly around a circle. Balls (cyan) of diameter $\delta = 0.8$ and $\delta = 1.5$ are drawn around the points in (a) and (b), respectively, resulting in one-simplices (black segments) and two-simplices (orange triangles). The persistence diagram (c) has H_0 (red points) and H_1 (blue triangles) features.

groups, then the set

$$\mathrm{Dgm}(\mathbf{S}) = \{(b_j, d_j, k_j) : \forall j \in J\} \cup \Delta, \tag{1.3}$$

where Δ represent a set of points where the birth time is equal to the death time, characterizes the persistence of the features, and is used to construct a graphical summary referred to as a *persistence diagram*. Figure 1.1 illustrates the main concepts in this section, where black points (zero-simplices) in Figure 1.1a and 1.1b denotes the data with cyan balls of diameter 0.8 and 1.5, respectively. Figure 1.1c shows the corresponding persistence diagram.

A set of persistence diagrams $\{\mathrm{Dgm}(\mathbf{S})\}$ can be endowed with a distance measure, such as the bottleneck distance. The bottleneck distance gives

the minimal L_∞ distance between bijections of any two diagrams. Let $\mathbf{S}_1$ and $\mathbf{S}_2$ be two finite compact subsets of $\mathbb{R}^d$, with $\mathrm{Dgm}(\mathbf{S}_1)$ and $\mathrm{Dgm}(\mathbf{S}_2)$ as their corresponding VR filtration persistence diagrams. The bottleneck distance, d_B, between the two persistence diagrams is defined as:

$$d_B\left(\mathrm{Dgm}(\mathbf{S}_1), \mathrm{Dgm}(\mathbf{S}_2)\right) = \inf_{\gamma} \sup_{\mu \in \mathrm{Dgm}(\mathbf{S}_1)} \|\mu - \gamma(\mu)\|_\infty, \tag{1.4}$$

where the infimum is taken over all bijections $\gamma : \mathrm{Dgm}(\mathbf{S}_1) \to \mathrm{Dgm}(\mathbf{S}_2)$. Let $\mathbf{S}_1$ and $\mathbf{S}_2$ be endowed with the Euclidean metric, then their Hausdorff distance, d_H, is given by

$$d_H(\mathbf{S}_1, \mathbf{S}_2) = \max\left\{ \sup_{\mathbf{v}_1 \in \mathbf{S}_1} d_{\mathbf{S}_2}(\mathbf{v}_1), \sup_{\mathbf{v}_2 \in \mathbf{S}_2} d_{\mathbf{S}_1}(\mathbf{v}_2) \right\}, \tag{1.5}$$

where $d_{\mathbf{S}_1}(\mathbf{v}_2) = \inf_{\mathbf{v}_1 \in \mathbf{S}_1} \|\mathbf{v}_1 - \mathbf{v}_2\|$. A fundamental result on persistence diagrams is that they are stable summaries in many settings (i.e., a small change in a point cloud results in a small change in the corresponding persistence diagram) (Chazal and Michel, 2021). This stability relation can be stated as:

$$d_B\left(\mathrm{Dgm}(\mathbf{S}_1), \mathrm{Dgm}(\mathbf{S}_2)\right) \leqslant 2 d_H(\mathbf{S}_1, \mathbf{S}_2). \tag{1.6}$$

Similar results can be obtained for functions, which is discussed in the next section.

1.4 Persistent Homology on Functions

Let ϕ be any real-valued function, where $\phi : \mathcal{X} \to \mathbb{R}$ for any compact set $\mathcal{X}$. We define the *lower-level sets* of ϕ as $\{\mathbf{x} : \phi(\mathbf{x}) \leqslant \lambda\}$ and the *upper-level sets* of ϕ as $\{\mathbf{x} : \phi(\mathbf{x}) \geqslant \lambda\}$. In more specific settings, we let the function ϕ be defined on the metric space $(\mathcal{X}, d_{\mathcal{X}})$. Define the reach$(\mathbb{A})$ as the largest radius r, such that each point in $\cup_{\mathbf{x}\in\mathbb{A}} B(\mathbf{x}, r)$ has a unique projection unto $\mathcal{X}$, where $B(\mathbf{x}, r)$ is a ball with radius r centered on $\mathbf{x}$. The reach is also referred to as the "condition number," and it quantifies the smoothness of the underlying manifold (Federer, 1959; Niyogi et al., 2008). Denote by $\mathcal{K}(\mathcal{X}, \kappa)$ the class of all manifolds such that for $\mathbb{A} \in \mathcal{K}(\mathcal{X}, \kappa)$, reach$(\mathbb{A}) \geqslant \kappa$, where κ is a fixed positive constant. Let the lower bound $\underline{b}(\mathcal{K}(\mathcal{X}, \kappa))$ and the upper bound $\overline{b}(\mathcal{K}(\mathcal{X}, \kappa))$ be positive constants depending on the geometry of the class $\mathcal{K}(\mathcal{X}, \kappa)$ but not on any specific manifold in $\mathcal{K}(\mathcal{X}, \kappa)$.

Assumption 1.1. *The following assumptions are made for the density function* f *and the distribution* $\mathbb{P}$*:* (i) *the support* $\mathbb{X}$ *of the distribution* $\mathbb{P}$ *is bounded, and* (ii) f *is tame and satisfies the following:* $0 < \underline{b}(\mathcal{K}(\mathcal{X}, \kappa)) \leqslant \inf_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) \leqslant \sup_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) \leqslant \overline{b}(\mathcal{K}(\mathcal{X}, \kappa)) < \infty$*. The tameness of* f *implies it has a finite number of critical values, ensuring the topological complexity of its level sets remains systematically bounded* (*Edelsbrunner and Harer, 2022*).

Functions defined on the vertices of the simplicial complex $\mathbb{C}$ provides another means to characterize the topology of the underlying data generating space. Let $\phi : \mathcal{X} \to \mathbb{R}$, and assume ϕ is extended to the simplices of

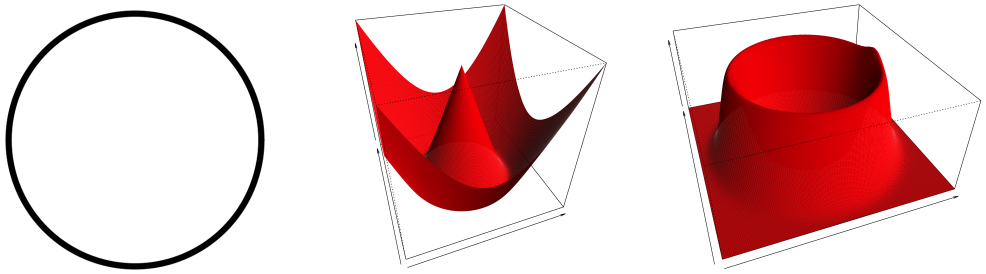
$\mathbb{C}$ such that $\phi(C) = \max_{0 \leqslant i \leqslant k} \phi(\mathbf{v}_i)$ for any simplex $C = (\mathbf{v}_0, \cdots, \mathbf{v}_k) \in \mathbb{C}$. The sequence of complexes $\mathbb{C}_\delta = \{C \in \mathbb{C} : \phi(C) \leqslant \delta\}$ creates a nested structure: $\mathbb{C}_{\delta_1} \subseteq \mathbb{C}_{\delta_2}$, $\delta_1 < \delta_2$, and defines a *lower-level set* filtration on ϕ. An *upper-level set* filtration can be defined analogously by considering the case where $\phi(C) \geqslant \delta$. We denote the resulting persistence diagram by $\text{Dgm}(\phi)$, such that topological feature $(b, d) \in \text{Dgm}(\phi)$ persists in the space $H_k(\phi^{-1}(-\infty, \delta))$, for $b \leqslant \delta < d$. Similar to the VR filtration, we can endow this space of persistence diagrams with the bottleneck distance as defined in Equation (1.4), and these persistence diagrams can also be shown to be stable summaries (i.e., small perturbations in the function space results in small changes in the persistence diagrams) (Cohen-Steiner et al., 2005; Chazal et al., 2016). This results in the following bound on the bottleneck distance for two functions ϕ and ψ under the assumption of tameness (Assumption 1.1):

$$d_B\left(\text{Dgm}(\phi), \text{Dgm}(\psi)\right) \leqslant \|\phi - \psi\|_\infty, \tag{1.7}$$

where $\|\phi - \psi\|_\infty = \sup_{x \in \mathcal{X}} |\phi(\mathbf{x}) - \psi(\mathbf{x})|$.

Two such functions ϕ that are relevant to this work are the kernel density function f_σ and the DTM function $d_{P,m}$. The empirical kernel density function $\widehat{f}_\sigma(\mathbf{x})$ with bandwidth σ is defined as:

$$\widehat{f}_\sigma(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} K_\sigma(\|\mathbf{x} - \mathbf{x}_i\|_2), \tag{1.8}$$



(a) Circle (b) DTM function (c) KDE function

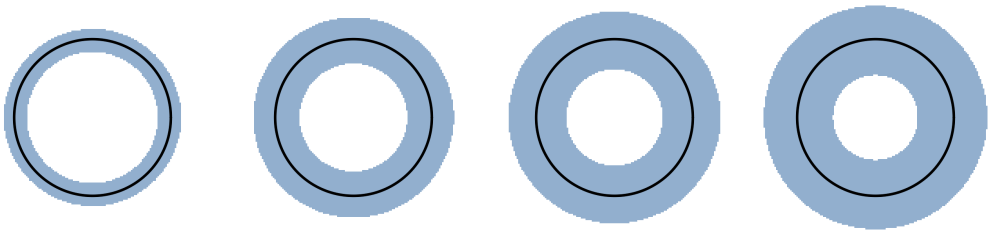
Figure 1.2: Illustration of the DTM and KDE function. (a) a 1D curve, (b) and (c) are the DTM function and the KDE function of this curve respectively.

where $K_\sigma(||\mathbf{x}||_2) = \sigma^{-d} K(||\mathbf{x}||_2/\sigma)$, and K is a d-dimensional kernel that is non-negative and integrates to one. Figure 1.2c shows an example KDE function on a 1D curve. While this kernel density function captures the shape and distribution of mass in the space $\mathcal{X}$, the DTM function provides a robust means to characterize this shape by approximating its distance function. The empirical DTM function $\widehat{d}^2_{P,m}(\mathbf{x})$ is defined as (Chazal et al., 2011):

$$\widehat{d}^2_{P,m}(\mathbf{x}) = \frac{1}{k} \sum_{\mathbf{x}_0 \in N_K(\mathbf{x})} ||\mathbf{x}_0 - \mathbf{x}||_2, \tag{1.9}$$

where $0 < m < 1$ is the resolution, and $N_K(\mathbf{x})$ is the set of k-nearest neighbors (k-NNs) to $\mathbf{x}$. The DTM filtration is a robust approximation of the VR filtration. Figure 1.2b shows an example DTM function on a 1D curve, and Figure 1.3 typical sublevel sets of the DTM function.

In practice, the estimation of filtration functions relies on the empirical probability measure $\mathbb{P}_n$, which assigns a probability mass of $1/n$ to each



(a) $\delta = 0.23$ (b) $\delta = 0.35$ (c) $\delta = 0.41$ (d) $\delta = 0.48$

Figure 1.3: Sublevel set filtration of the DTM function in Figure 1.2b, shown at four increasing thresholds.

data point $\mathbf{x}$. As a result, the empirical function ϕ_n, whether it represents the empirical KDE or the empirical DTM function, exhibits sensitivity to noise and sample density variations. This sensitivity directly affects the persistence of the resulting features.

2 A SUBSEQUENCE APPROACH TO TOPOLOGICAL DATA ANALYSIS OF IRREGULARLY-SPACED TIME SERIES DATA

The content of this chapter is published in Dakurah and Cisewski-Kehe (2024).

Abstract

A time-delay embedding (TDE), grounded in the framework of Takens's Theorem, provides a mechanism to reconstruct and analyze the state-space representation of time-series data. Recently, topological data analysis (TDA) methods have been applied to study this time series representation mainly through the lens of persistent homology. Current literature on the fusion of TDE and TDA are adept at analyzing uniformly-spaced time series observations. This work introduces a novel subsequence embedding method for *irregularly*-spaced time-series data. We show that this method preserves the original state space topology while reducing spurious homological features. Theoretical stability results and convergence properties of the proposed method in the presence of noise and varying levels of irregularity in the spacing of the time series are established. Numerical studies and an application to real data illustrates the performance of the proposed method.

2.1 Introduction

A time series measurement $x(t) \in \mathbb{R}$ at time t can be considered as the outcome of a data-generating space of some dynamical system (i.e., mathematical models that describes the evolution of variables over time) with state vector $\mathbf{s}(t) \in \mathbb{R}^N$. Constructing a meaningful approximation of this underlying data-generating space when only the scalar time series is observed can uncover latent patterns and structures not readily apparent in the raw time series. Time-delay embeddings (TDEs) are often employed for this state space reconstruction. The TDE method transforms the time-series data from the time-domain to an estimate of the state space, which can reveal properties of the system such as periodicity and other structures not apparent in the time domain. The principle underlying TDEs is Takens's Theorem, which asserts that even if the actual dynamics (i.e., the system's behavior over time) are not known, a single time series can be treated as a one-dimensional projection of the path traced by the system's state vector in a multi-dimensional space. An approximation to the actual dynamics can be constructed from this projection (Takens, 2006). Takens proved that assuming uniformly-spaced and noise-free measurements of unlimited length, there exists a diffeomorphism (i.e., a smooth and invertible function) between the true high-dimensional dynamical system and its TDE-based reconstruction. This theorem forms the foundation for much of the discussions on reconstructing the multi-dimensional state of

a system from a single time series (Ali et al., 2007).

More recently, there is renewed interest in coupling the TDE method with tools from topological data analysis (TDA) to study the dynamics of time-series data using various geometric and topological features of the TDE reconstruction of the underlying state space, such as clusters, loops, voids, and their higher dimensional analogs (El-Yaagoubi et al., 2023; Gholizadeh and Zadrozny, 2018; Seversky et al., 2016). TDA is a computational method for studying the shape of data, which can be applied to characterize the topological features of these reconstructed state spaces. The characterization is often carried out using *persistent homology*, a tool of TDA, which uses a multi-scale approach to quantify certain topological features (Edelsbrunner et al., 2000; Edelsbrunner and Harer, 2022). TDA and TDE have been successfully applied to quantify periodicity in time-series data (Perea and Harer, 2015), analyze human speech (Brown and Knudson, 2009), detect motion patterns in video (Tralie and Perea, 2018), and in wheeze detection (Emrani et al., 2014).

In the applications noted above, the observed time series is uniformly-spaced. However, time series is often not uniformly-spaced due to measurement lapses (Stark et al., 1997), process errors (Casdagli et al., 1991), or inherent features of the data generating process (Stark et al., 1997; Lekscha and Donner, 2018), etc. The standard Takens's theorem does not handle irregularly-spaced time series, but several options exist in the literature to address issues related to irregularly-spaced time series observations to

make it amenable to TDE. Broadly, these can be classified into *imputation* or *exclusion* methods. Imputation methods involve predicting the missing observations, and then the analysis is carried out assuming a uniformly-spaced time series has been observed (Harvey and Pierse, 1984; Casdagli et al., 1991; Lekscha and Donner, 2018). Exclusion methods initially ignore the presence of missing values and assume a uniformly-spaced set. The TDE maps are then constructed and any embedding vector with a missing value is excluded (Boker et al., 2018; Johnson and Munch, 2022). If the imputation model is misspecified, it can produce structures in the TDE that do not reflect true properties of the data, and the exclusion method can significantly alter the shape of the TDE space (Huke and Broomhead, 2007; Boker et al., 2018). Since TDA can provide quantification of qualitative properties of the reconstructed state space, the drawbacks of the imputation and exclusion methods may distort topological features constructed from the TDE spaces.

In this chapter, we propose a subsequence method for constructing a TDE of irregularly-spaced time-series data that preserves certain properties of the reconstructed state space. The level of irregularity of the time-series data is controlled by the *regularity score* (defined in Section 2.3). We show that the proposed method preserves the topological features of the original underlying state space of the time series while reducing spurious shape features. Theoretically, we prove stability and convergence results of the proposed subsequence method in the presence of noise and

for varying levels of irregularity in the observed time series. Further, we demonstrate the competitiveness of the proposed subsequence method through simulation studies and an application to real data.

2.1.1 A Note on Terminology

For this work, the term *uniformly-spaced* time series is used to describe time series that have equally-spaced time intervals between successive observations, and is considered the "true" time series for purposes of evaluating the proposed method. The term *irregularly-spaced* time series refers to observations with unequally-spaced time interval between successive observations. To characterize how the two forms of time series are related, it is assumed throughout this work that the irregularly-spaced time series is a subset of the uniformly-spaced time series. More formally, let $\mathbf{x} = [x(t_1), \cdots, x(t_n)]^\top$ be a uniformly spaced time series vector, such that $t_{i+2} - t_{i+1} = t_{i+1} - t_i, \forall i$; an irregularly-spaced times is any subset of $\mathbf{x}$ with observations at one or more time points randomly missing. Hence, the irregularly-spaced time series always have fewer time measurements than the corresponding uniformly-spaced time series.

The concept of a TDE is also referred to in the literature as a *delayed-coordinate embedding*, a *sliding-window embedding*, or simply a *Takens embedding*. For this work, only the term TDE is used. For any irregularly-spaced time series, it is assumed there is a "true" underlying uniformly-spaced time series. The use of "TDE" exclusively refers to an embedding con-

structed from this true underlying uniformly-spaced time series. After applying the proposed subsequence method, the resulting embedding is referred as the *subsequence embedding* (SSE). In instances where an exposition applies to both the TDE and SSE, the term *embedding map* is used as a collective reference to the two concepts.

For a given time interval, it is assumed that missing or unobserved values occur with a given probability. That is, for a given time point in a time interval, a measurement is not observed at that point with some probability. Such probabilistic mechanism governing the observations of time series values is not uncommon in the literature (e.g., Dunsmuir and Robinson 1981). This probability can be fixed for all time points or it can vary for each time point. This characterization is referred to as the *missingness structure* of the time series in context.

2.2 Basics of Time-Delay Embeddings

For this work, the discussion on TDEs is restricted to univariate time series. Assume this univariate time series is generated by a system with a state vector $\mathbf{s}(t)$ on a manifold which is a subset of some N-dimensional space $\mathbb{R}^N$. The state vector $\mathbf{s}(t)$ is not directly observable, however some measurement of it, denoted $x(t) = h(\mathbf{s}(t))$, is observed through the measurement function $h(\cdot)$. The measurement function $h(\cdot)$ can be thought of as rule that transforms the high-dimensional state vector into the observed

univariate time series $x(t)$. For instance, in astronomy, $\mathbf{s}(t)$ might include variables such as positions, velocities, and brightnesses of various celestial bodies, such as exoplanets, stars, or galaxies. However, the measurement function is specifically designed to extract a single scalar value from this vector. The specific form of $h(\cdot)$ is influenced by many considerations, for example, the limitation of observational tools. For a star, the measurement function could be designed to extract a key observable from the state vector, such as its brightness. Thus the measurement $h(\mathbf{s}(t))$ reflects the observed brightness of the star at any given time t.

The scalar value $x(t)$ is the observed time series measurement. Define the function $F : \mathbb{R}^N \to \mathbb{R}^{M+1}$ as the embedding map with the form:

$$F(\mathbf{s}(t)) = [x(t), x(t+\tau), \cdots, x(t+M\tau)]\,. \tag{2.1}$$

We emphasize that F *is a function on the state vector* $\mathbf{s}(t) \in \mathbb{R}^N$ *and not the scalar value* $h(\mathbf{s}(t))$. While some authors denote the embedding map as F_h to highlight the measurement function h, we do not adopt this notation for clarity. Figure 2.1 illustrates an example of how this function and the resulting vector are constructed. If the measurement function $h(\cdot)$ is noise-free, and the embedding dimension $M+1$ is chosen to be more than twice the dimension of the attractor (i.e., the N-dimensional region toward which the system evolves) of the system's state space, Takens's theorem guarantees that the embedding map $F(\mathbf{s}(t))$ has a one-to-one

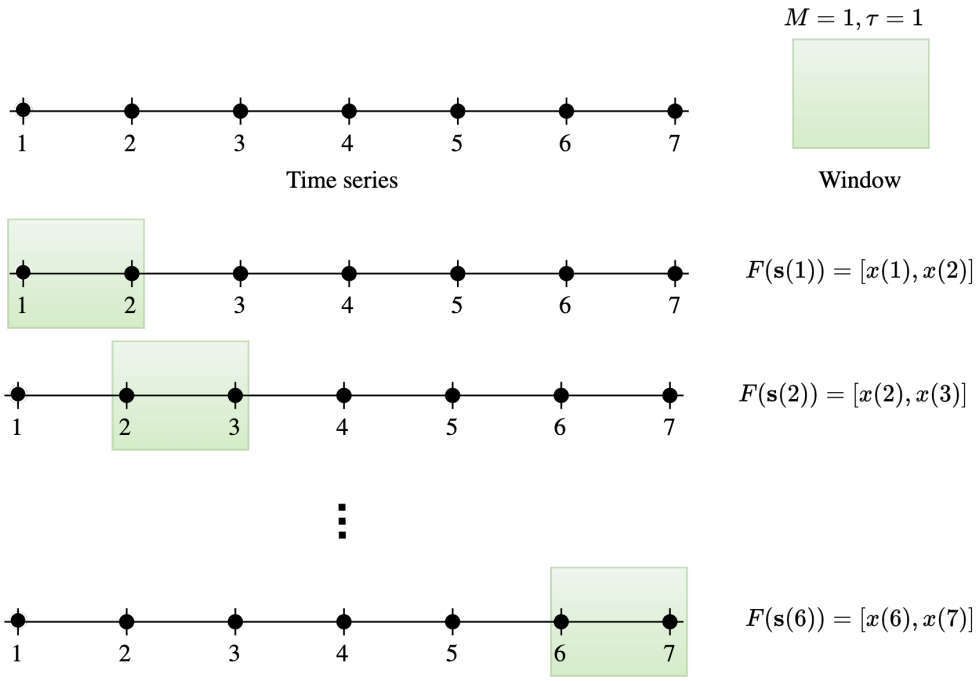


Figure 2.1: Illustration of the construction of the sliding window vector $F(\mathbf{s}(t))$. Top row: An example time series with seven time points, using embedding parameters $M = 1, \tau = 1$, along with the corresponding embedding window. The next three timelines demonstrate the sequential sliding of the embedding window to construct the embedding vectors $F(\mathbf{s}(t))$. Collectively, these vectors form the reconstructed space.

correspondence between the original state space of the system (from which the time series is derived) and the reconstructed state space formed by $F(\mathbf{s}(t))$ (Takens, 2006). This ensures that the dynamics of the system can be studied in the reconstructed space as if it were being studied in the original space. Figure 2.2 demonstrates this reconstruction process by mapping the scalar time series to the TDE matrix F to reconstruct the state space.

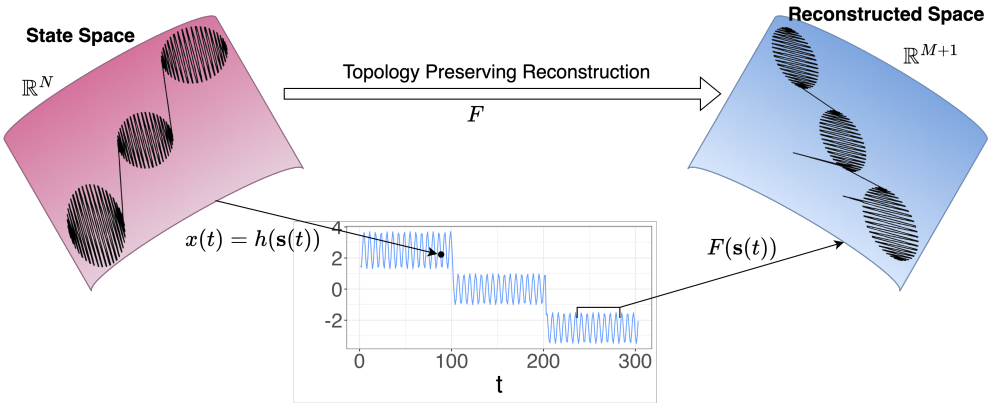


Figure 2.2: Illustration of the embedding process. Top-left: the state space, typically not observed. Middle-bottom: the time series obtained via the measurement function $h(\cdot)$. Top-right: the reconstructed space from the TDE matrix F, which preserves the topology of the original state space.

The choice of embedding dimension $M + 1$ and step size τ is a subject of considerable research in the literature (e.g., Cao 1997; Kim et al. 1999). In this work, the embedding dimension is chosen manually. However, the method of false nearest neighbors is one common method for determining this dimension, which identifies points in a low-dimensional space that appear to be near each other but are not actually neighbors when the data are viewed in a high-dimensional space. By systematically increasing the embedding dimension, and evaluating the percentage of false nearest neighbors, the dimension can be set where this percentage drops significantly, indicating a suitable dimension. More details about this and other procedures for determining M and τ can be found in Cao (1997) or Kim et al. (1999). A large value of M is often preferred as it enables the embedding to capture more details inherent in the time series. If M is

too large, there may be an insufficient number of points in the embedding space. Furthermore, if $M\tau$ is too small due to a small τ, relatively fewer points fall in each embedding window. This results in points repeatedly appearing in windows, which can lead to redundant information. If $M\tau$ is too large due to a large value of τ, the reconstructed state space can be distorted because relevant periodic behavior of the time series may not be captured (Casdagli et al., 1991). Hence the choice of τ and M is such that $M\tau$ is not too large or too small, but is application-dependent and requires empirical testing. For the purposes of this work, we assume that an appropriate embedding window $M\tau$ can be determined through a combination of the previously discussed parameter selection methods, based on empirical testing for each specific application.

2.3 Subsequence Method

The TDE construction in the previous section assumes the observed time series is uniformly-spaced, but a time series is often irregularly-spaced in real data. We propose a method to extract uniformly-spaced subsequences from the observed irregularly-spaced time series and prove its topology-preserving properties, along with consistency and convergence results.

2.3.1 Subsequence construction

Let $\mathbf{x} = (x(t) : t \in \mathcal{T})$ be a time series of length n where $\mathcal{T} = \{t_1, \cdots, t_n\} \subset \mathbb{N}$. Further assume that this time series is not uniformly spaced, that is, $t_{i+1} - t_i \neq t_{i+2} - t_{i+1}$, for at least one $t_i \in \mathcal{T}$ such that $t_i < t_{i+1}$. In this work, a subsequence of the set $\mathbf{x}$ is defined as any subset that omits elements of $\mathbf{x}$ without changing the order of the remaining elements. This definition does not guarantee that $t_{i+1} - t_i = t_{i+2} - t_{i+1}, \forall t_i \in \mathcal{T}$, which is a condition we want to achieve with the proposed subsequence construction. Let $\mathbf{x}_{p,r} \subseteq \mathbf{x}$ be a subset of the original time series with time indexes $\mathcal{T}_{p,r} \subseteq \mathcal{T}$ with the condition that $t_{p,i+1} - t_{p,i} = r, \forall t_{p,i} \in \mathcal{T}_{p,r}$. The set $\mathbf{x}_{p,r}$ is the p-th subsequence of *regularity* r, and it is a uniformly-spaced subsequence. For any non-uniformly spaced time series, we can build a collection of such subsequence for various values of r. The goal is to first obtain the longest subsequence for a small r. As the subsequence length reduces for a given r, the regularity value r can increase to obtain more uniformly-spaced subsequences. An algorithm for computing this collection of subsequences is displayed in Algorithm 1, which is adapted from an algorithm that finds the longest arithmetic progression in a sequence developed in Erickson (1999). In the statement of the algorithm, the following notation is used: (i) the union symbol $\cup$ denotes the addition of a set (element) to a set (vector), (ii) the number of elements in a set or a vector A is denoted by $|A|$, and (iii) the notation $A \backslash B$ represents subset of elements in set A obtained by excluding all elements from set B. Algorithm 1 returns all

Algorithm 1 Uniform subsequence construction

Require: Regularity score r, minimum sequence length m, time points $\mathcal{T} = \{t_1, \cdots, t_n\}$ (exclude time points from subsequences generated with different r).

Ensure: $r \leqslant t_n - t_1, \quad m \leqslant n$.

Initialize: $\mathcal{T}_p \leftarrow \{\dots\}$, temporary time index, $\mathbf{T}_{reg} \leftarrow \{\}$, uniformly-spaced subsequences.

1: **while** number of elements in $\mathcal{T}$ is greater than m **do**
2: **for** $i = 1 : (|\mathcal{T}| - 1)$ **do**
3: $\mathcal{T}_{sub} \leftarrow \mathcal{T}[i]$ ▷ Initialize a subsequence.
4: **for** $j = (i + 1) : |\mathcal{T}|$ **do**
5: **if** $\mathcal{T}[j] - \mathcal{T}_{sub}[j - i] = r$ **then** ▷ Check the regularity condition.
6: $\mathcal{T}_{sub} \leftarrow \mathcal{T}_{sub} \cup \mathcal{T}[j]$; **if** $|\mathcal{T}_{sub}| > |\mathcal{T}_p|$ **then** $\mathcal{T}_p \leftarrow \mathcal{T}_{sub}$
7: **else** *break* ▷ Initialize with the next point in the sequence.
8: **if** $|\mathcal{T}_p| \geqslant m$ and $\mathcal{T}_p$ is not identical to any other subsequence in $\mathbf{T}_{reg}$ **then**
9: $\mathbf{T}_{reg} \leftarrow \mathbf{T}_{reg} \cup \mathcal{T}_p$; $\mathcal{T} \leftarrow \mathcal{T} \backslash \mathcal{T}_p$ ▷ Remove the subset from the sequence.
10: **else** *break* ▷ No uniformly-spaced subsequence of the required length exist.
11: **return** $\mathbf{T}_{reg}$ ▷ Set of all regularly spaced subsequences each of regularity r.

possible uniformly-spaced time points from the time index set $\mathcal{T}$ with regularity score r. Note that for uniformly-spaced $\mathcal{T}$, it returns the full sequence. The uniformly-spaced observations can now be obtained by simply matching these observations to the time points in each subsequence. Not all the subsequences returned by Algorithm 1 are required in the SSE (see Remark 2.1). Each time point can be used at most once among all the subsequences (see Algorithm 1, line 9). Moreover, there is no restriction

preventing two subsequences from having the same length.

2.3.1.1 Subsequence embedding method

Takens's theorem guiding the construction of the TDE in Section 2.2 involves a single measurement function $h(\cdot)$, which generates each time series measurement (Takens, 2006). A generalization considers each coordinate in the embedding maps as a measurement function (see Remark 2.9 in Sauer et al. (1991) and Theorem 2 in Deyle and Sugihara (2011)). Such generalizations allow for the extension of Takens's theorem to multiple measurement functions involving multiple time series. This motivates the proposed SSE method where each subsequence is viewed as distinct time series.

To construct the proposed SSE, a single distinct measurement function is defined on each subsequence. Let $h_p(\cdot)$ be the measurement function associated with the p-th subsequence. Then the p-th embedding mapping has the form:

$$F_p(\mathbf{s}(t_{p,i})) = [x_{p,r}(t_{p,i}), x_{p,r}(t_{p,i} + \tau_p), \cdots, x_{p,r}(t_{p,i} + M\tau_p)], \qquad (2.2)$$

where $x_{p,r}(t_{p,i}) = h_p(\mathbf{s}(t_{p,i}))$. The delay step τ_p is fixed for each subsequence map. The map is also constructed under the assumption that the length of each subsequence $n_p > \max(M + 1, M * \tau_p)$. This ensures that there are sufficient observations within each subsequence to construct

a point in the embedding space. The embedding matrix from the p-th subsequence has the form:

$$\mathbf{F}_p = \begin{bmatrix} F(\mathbf{s}(t_{p,1}))^\top & F(\mathbf{s}(t_{p,2}))^\top & \cdots & F(\mathbf{s}(t_{p,n_p-M}))^\top \end{bmatrix}^\top . \tag{2.3}$$

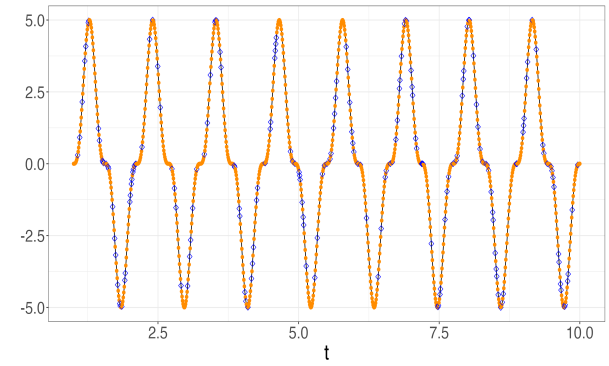
Observe that each $\mathbf{F}_p$ is a matrix of dimension $(n_p - M\tau_p) \times (M+1)$. The row dimension of $n_p - M\tau_p$ follows from the fact that, for a subsequence of length n_p, the number of points in the embedding space of dimension $M+1$ is $n_p - M\tau_p$ for step-size τ_p. The full embedding matrix for the irregularly-spaced time series, denoted by $\mathbf{F}$ is then given by:

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}_1 \\ \vdots \\ \mathbf{F}_P \end{bmatrix}_{\mathcal{N} \times (M+1)} . \tag{2.4}$$

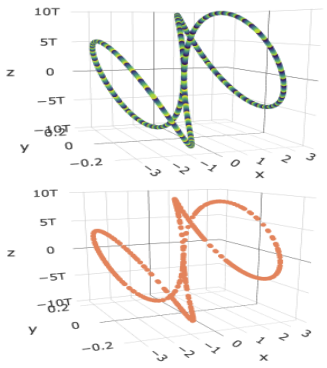
Here, P is the total number of uniformly-spaced subsequences, and since each p-th subsequence embedding matrix $\mathbf{F}_p$ has $n_p - M\tau_p$ points in the $(M+1)$-dimensional space, the row dimension of $\mathbf{F}$ is given by $\mathcal{N} = \sum_{p=1}^{P}(n_p - M\tau_p)$. Note that when the original time series is uniformly-spaced, the SSE method is identical to the TDE method. To see this, observe that the longest subsequence in the uniformly-spaced time series is the original sequence.

To illustrate the SSE framework, Figure 2.3a shows measurements at 1000 uniformly-spaced time points (orange points and blue diamonds

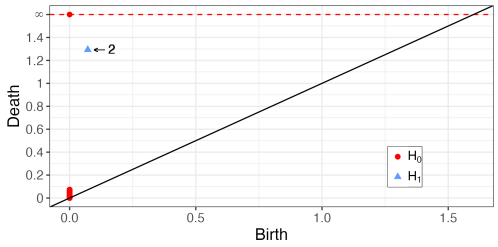
combined) of which about 20% are designated as missing values (blue diamonds), which creates an irregularly-spaced time series (orange points). Both the uniformly-spaced and irregularly-space time series were embedded into $\mathbb{R}^4$ using the TDE and SSE methods, respectively. Figure 2.3b-top gives the TDE of the uniformly-spaced 1000 measurements and contains two identical elliptical shapes. Figure 2.3b-bottom shows the proposed SSE of the irregularly-spaced time series and also contains two similar elliptical shapes, however, there is visible non-uniform spacing of the points compared to the TDE space. This is primarily due to the SSE using a subset of the original time series (i.e., it constructs a uniform subsample from the irregularly-spaced time series based on Algorithm 1); the SSE space may be considered as a sparse representation of the TDE space. The persistence diagram for the TDE is shown in Figure 2.3c. Since Figure 2.3b-top has two identical elliptical shapes, the H_1 features have overlapping birth and death time, hence the appearance of a single blue triangle. Figure 2.3d shows the persistence diagram for the SSE, and correctly identifies the two loops but the birth and death time are non-overlapping due to the non-identical spacing of the points in the two elliptical shapes. In general, the SSE converges to the TDE in terms of the topological similarity of the reconstructed spaces and in the closeness of the persistence diagrams as the time sampling becomes more uniform. A more formal theoretical justification of this assertion, and other technical considerations are discussed in the next section.



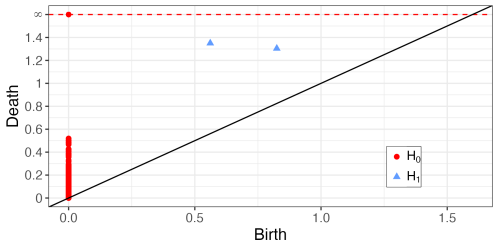
(a) Time series measurements at 1000 time points



(b) Reconstructed spaces



(c) TDE persistence diagram



(d) SSE persistence diagram

Figure 2.3: SSE method illustration. (a) One thousand time series measurements (blue and orange points). About 20% were designated as missing (hollow blue diamonds) to obtain irregularly-spaced observations (orange points). The TDE of the full time series ((b)-top) and the SSE of the irregularly-spaced time series ((b)-bottom); both time series were embedded in $\mathbb{R}^4$ and their first three principal components are plotted. The persistence diagram of the TDE (c) and SSE (d).

Remark 2.1. *The choice of the number of subsequences* P *and the number of subsequences of different regularity scores* r *depend on the context and goals of the analysis. To reconstruct an* $(M+1)$*-dimensional state space, subsequences must satisfy* $n_p > \max(M+1, M*\tau_p)$*. A set of subsequences with the same regularity score can lead to better reconstruction accuracy as it captures the dominant patterns of the underlying data-generating space of the time series more coherently.*

Combining subsequences with different regularity scores can improve topological approximations as it captures a wider range of structures of the underlying data-generating space, but it can introduce points in the embedding that may be geometric outliers leading to less accurate reconstructions; kernel smoothing may help to mitigate these issues. Thus, there is a trade-off between a better topology approximation and improved reconstruction accuracy. If subsequences with the same regularity score capture most of the time series, combining sequences with different regularity scores may offer limited benefits. While the simulations and real data analysis in this work utilize subsequences with the same regularity score, the methodology and theoretical results apply to subsequences with the same or different regularity scores.

2.4 Stability and Convergence Results

The reconstructed state space using the proposed SSE approximates the state space based on a uniformly-sampled time series (i.e., the TDE space). Persistence diagrams are used to quantify the stability of the estimate by measuring its closeness to the TDE space. In what follows, these stability results are established for the proposed SSE method and a denoising procedure to reduce the noise present in the observed time series.

2.4.1 Stability of Denoising Procedure

Time-series data are typically observed with noise. The level of noise in the reconstructed space influences the presence and the persistence of the topological features. A Fourier denoising procedure is proposed to filter out noise in the observed time series. Furthermore, stability results demonstrating how this denoising procedure preserves the underlying topological features within the persistence homology framework are established. The proposed denoising procedure, coupled with the stability guarantee, is crucial to the proposed SSE pipeline as noisy data could make it practically impossible to determine the optimal embedding window $M\tau$. Hence, a process for reducing this noise is essential, and it is important to guarantee that the denoising procedure does not alter the topological characteristics of the underlying manifold from which the time series were observed.

Let $\mathbf{x} = \begin{bmatrix} x(t_1), & x(t_2), & \cdots & , x(t_n) \end{bmatrix}^\top$ be an observed time series vector. The first step in the denoising procedure is to transform this observed signal to the frequency domain. The discrete Fourier transform (DFT) of $x(t_k)$, denoted as $\tilde{x}(t_k)$ is given by:

$$\tilde{x}(t_k) = \sum_{r=1}^{n} x(t_r) e^{-j2\pi w_r f_k} = \sum_{r=1}^{n} x(t_r)\phi_{kr}, \quad 1 \leqslant k \leqslant n, \tag{2.5}$$

where j is the imaginary unit ($j^2 = -1$), $\phi_{kr} = e^{-j2\pi w_r f_k}$, $0 \leqslant w_r \leqslant 1$ are sample points, $0 \leqslant f_k \leqslant n$ are frequencies, and $\tilde{x}(t_k)$ is the k-th sample of

the power spectrum at f_k.

To filter out noise, the power spectral density $\tilde{x}(t_k)$ is computed for each t_k. Then a threshold is chosen, and any $\tilde{x}(t_k)$ with power spectral density less than the threshold is set to zero. In selecting the threshold, the goal is to choose a value that does not smooth out the peaks in the true signal. The derivation that follows assumes the selected threshold preserves the peaks in the true signal. To simplify notations, the thresholded observations are also denoted as $\tilde{x}(t_k)$. The thresholded $\tilde{x}(t_k)$ are transformed back to the time domain to get the noise-reduced signal, which typically involves multiplying $\tilde{x}(t_k)$ by the inverse of a Fourier transform matrix.

Let $\tilde{\mathbf{x}}$ be the DFT of the time series vector with corresponding Fourier basis $\boldsymbol{\phi}_k$, such that for $k = 1, \ldots, n$:

$$\tilde{\mathbf{x}} = \begin{bmatrix} \tilde{x}(t_1), & \tilde{x}(t_2), & \cdots, & \tilde{x}(t_n) \end{bmatrix}^\top, \boldsymbol{\phi}_k = \begin{bmatrix} \phi_{k1}, & \phi_{k2}, & \cdots, & \phi_{kn} \end{bmatrix}^\top. \tag{2.6}$$

The forward transform in Equation (2.5) can be vectorized: $\tilde{\mathbf{x}} = \boldsymbol{\Phi}\mathbf{x}$, where $\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\phi}_1 & \boldsymbol{\phi}_2 & \cdots & \boldsymbol{\phi}_n \end{bmatrix}^\top$. The backward transform can then be determined by inverting the matrix $\boldsymbol{\Phi}$. However, due to the non-uniformity in the spacing of the time series $\mathbf{x}$, the columns of $\boldsymbol{\Phi}$ are not orthogonal, and it is not directly invertible, so the pseudo-inverse is used instead. The backward transform then has the form: $\mathbf{x} = \frac{1}{n}(\boldsymbol{\Phi}^H\boldsymbol{\Phi})^\dagger\boldsymbol{\Phi}^H\tilde{\mathbf{x}}$, where $\mathbf{A}^H$ and $\mathbf{A}^\dagger$ denote the complex conjugate transpose and the Moore-Penrose inverse of the matrix $\mathbf{A}$, respectively. The matrix $(\boldsymbol{\Phi}^H\boldsymbol{\Phi})^\dagger\boldsymbol{\Phi}^H$ projects the

frequency vector $\tilde{\mathbf{x}}$ onto the column space of $\mathbf{\Phi}$. Let $\Pi_{\mathbf{\Phi}}(\mathbf{x})$ denote this projection operation. The modulus of off-diagonal elements of $\mathbf{\Phi}^{H}\mathbf{\Phi}$ is bounded as: $\left|\sum_{k=1}^{n} e^{j2\pi(w_{l_1}-w_{l_2})f_k}\right| \leqslant \sum_{k=1}^{n}\left|e^{j2\pi(w_{l_1}-w_{l_2})f_k}\right| = n$, where the equality follows from the definition of the complex modulus $|z| = \sqrt{z\bar{z}}$, with $\bar{z}$ as the conjugate of the complex value z. The matrix $\mathbf{\Phi}\mathbf{\Phi}^{H}$ has the structure

$$(\mathbf{\Phi}\mathbf{\Phi}^{H})_{l_1,l_2} = \sum_{k=1}^{n} e^{(-1)^{\delta} j2\pi(f_{l_2}-f_{l_1})w_k}, \quad \delta = \mathbb{1}(l_1 \leqslant l_2), \tag{2.7}$$

for indicator function $\mathbb{1}(l_1 \leqslant l_2)$, and $(\mathbf{\Phi}\mathbf{\Phi})^{H}_{l_1,l_2}$ denotes the value in the l_1-th row and l_2-th column of $\mathbf{\Phi}\mathbf{\Phi}^{H}$. Then $\mathbf{\Phi}\mathbf{\Phi}^{H}$ is Toeplitz when the frequency components are uniformly-spaced such that $f_k = k$, and $\mathbf{\Phi}\mathbf{\Phi}^{H}$ is fully specified by its first row elements (HuoLiu and YuanTang, 1998).

Using these results, we establish a fixed sample size, non-asymptotic bound in the following proposition, which asserts that the DFT preserves topological features and is stable with respect to the bottleneck distance. In particular, the bottleneck distance between the persistence diagrams of the embeddings of the noise-free and smoothed (i.e., noise-reduced) time series is bounded above by the embeddings of the observed noisy and noise-free time series.

Proposition 2.1. *Given* $\mathbf{x}^* \in \mathbb{R}^n$ *as a possibly irregularly-spaced scalar time series with additive noise of the form* $\mathbf{x}^* = \mathbf{x} + \varepsilon$, *where* $\mathbf{x}$ *is a noise-free scalar*

time series, and ε is a zero-mean noise term, then let $\mathbf{x}'$ be the time series vector after applying the proposed Fourier denoising to $\mathbf{x}^*$, and $\mathbf{F}$, $\mathbf{F}^*$, and $\mathbf{F}'$ be the embedding matrices associated with $\mathbf{x}$, $\mathbf{x}^*$, and $\mathbf{x}'$, respectively. Also, let $Dgm(\mathbf{F})$ and $Dgm(\mathbf{F}')$ denote the persistence diagrams associated with the Vietoris-Rips complex constructed from $\mathbf{F}$ and $\mathbf{F}'$, respectively. Then the bottleneck distance between these two persistence diagrams is bounded as

$$d_B(Dgm(\mathbf{F}), Dgm(\mathbf{F}')) \leqslant 2\frac{(2n-1)}{c}\left(\sup_{i,p} \|F^*(\mathbf{s}(t_{p,i})) - F(\mathbf{s}(t_{p,i}))\|_2\right), \tag{2.8}$$

where $0 < c \leqslant 1, \quad 1 \leqslant i \leqslant n_p - M\tau_p, \quad 1 \leqslant p \leqslant P.$

Proof. It suffices to bound the Hausdorff distance between $\mathbf{F}$ and $\mathbf{F}'$, then using the stability theory in persistence homology (see Equation (1.4)), the bound on their persistence diagrams with respect to the Bottleneck distance can be established. The proof proceeds as follows.

There exists a subset $\mathbf{x}_i^* \subset \mathbf{x}^*$ that exactly equals $F^*(\mathbf{s}(t_{p,i}))$ (the i-th row of the p-th subsequence of the embedding matrix $\mathbf{F}^*$). This fact stems from the construction of $\mathbf{F}^*$, whose rows are uniformly-spaced samples of $\mathbf{x}^*$. The same guarantee holds for the pairs $(\mathbf{F}, \mathbf{x})$, and $(\mathbf{F}', \mathbf{x}')$. The distance between the projection $\Pi_{\mathbf{\Phi}}(\mathbf{x}_i^*)$ and $F(\mathbf{s}(t_{p,i}))$ is given by

$$\|\Pi_{\mathbf{\Phi}}(\mathbf{x}_i^*) - F(\mathbf{s}(t_{p,i}))\|_2 = \|\Pi_{\mathbf{\Phi}}\left(\mathbf{x}_i^* - F(\mathbf{s}(t_{p,i}))\right)\|_2, \tag{2.9}$$

where $\|\cdot\|_2$ denotes the l^2-norm. Equation (2.9) is under the assumption that the choice of frequency threshold does not smooth out the peaks in the true signal $\mathbf{x}$. Observe that each $F'(\mathbf{s}(t_{p,i}))$ is isometric to $\Pi_{\mathbf{\Phi}}(\mathbf{x}_i^*)$, where $\mathbf{x}_i^*$ is a subset of length $M+1$ of the original noisy scalar time series. Then the Gromov-Hausdorff distance between $\mathbf{F}'$ and $\mathbf{F}$ can be expressed as

$$d_{GH}(\mathbf{F}', \mathbf{F}) = d_{GH}\left(\widehat{\Pi}_{\mathbf{\Phi}}(\mathbf{x}^*), \mathbf{F}\right), \tag{2.10}$$

where $\widehat{\Pi}_{\mathbf{\Phi}}(\mathbf{x}^*)$ denotes embedding of the vector $\Pi_{\mathbf{\Phi}}(\mathbf{x}^*)$. Using the same isometric property, the Hausdorff distance between $\mathbf{F}'$ and $\mathbf{F}$ can be expressed in terms of Equation (2.9). This follows from the fact that

$$\begin{aligned}\|\Pi_{\mathbf{\Phi}}\left(\mathbf{x}_i^* - F(\mathbf{s}(t_{p,i}))\right)\|_2 &= \|\Pi_{\mathbf{\Phi}}\left(F^*(\mathbf{s}(t_{p,i})) - F(\mathbf{s}(t_{p,i}))\right)\|_2 \\ &\leqslant \|F^*(\mathbf{s}(t_{p,i})) - F(\mathbf{s}(t_{p,i}))\|_2 \|(\mathbf{\Phi}^H\mathbf{\Phi})^\dagger \mathbf{\Phi}^H\|_2.\end{aligned} \tag{2.11}$$

The matrix $\mathbf{\Phi}^H\mathbf{\Phi}$ has the form:

$$\mathbf{\Phi}^H\mathbf{\Phi} = \begin{bmatrix} n & \sum_{k=1}^n e^{j2\pi(w_2-w_1)f_k} & \dots & \sum_{k=1}^n e^{j2\pi(w_n-w_1)f_k} \\ \sum_{k=1}^n e^{-j2\pi(w_2-w_1)f_k} & n & \dots & \sum_{k=1}^n e^{j2\pi(w_n-w_2)f_k} \\ \vdots & \vdots & \dots & \vdots \\ \sum_{k=1}^n e^{-j2\pi(w_n-w_1)f_k} & \sum_{k=1}^n e^{-j2\pi(w_n-w_2)f_k} & \dots & n \end{bmatrix}. \tag{2.12}$$

The modulus of off-diagonal elements of $\mathbf{\Phi}^H\mathbf{\Phi}$ is bounded as

$$\left|\sum_{k=1}^{n} e^{j2\pi(w_{l_1}-w_{l_2})f_k}\right| \leqslant \sum_{k=1}^{n} \left|e^{j2\pi(w_{l_1}-w_{l_2})f_k}\right| = n. \tag{2.13}$$

Observe that $\|(\mathbf{\Phi}^H\mathbf{\Phi})^\dagger\mathbf{\Phi}^H\|_2 \leqslant \|(\mathbf{\Phi}^H\mathbf{\Phi})^\dagger\|_2\|\mathbf{\Phi}^H\|_2$. First we bound $\|\mathbf{\Phi}^H\|_2$, by directly using the definition:

$$\|\mathbf{\Phi}^H\|_2 = \sqrt{\lambda_{max}(\mathbf{\Phi}\mathbf{\Phi}^H)}, \tag{2.14}$$

where $\lambda_{max}(\mathbf{\Phi}\mathbf{\Phi}^H)$ is the maximum eigenvalue of $\mathbf{\Phi}\mathbf{\Phi}^H$. Since $\mathbf{\Phi}\mathbf{\Phi}^H$ is Toeplitz, a bound on the maximum eigenvalue can be established as follows (Hertz, 1992). Let $\boldsymbol{\psi} = [\psi_1, \psi_2, \cdots, \psi_n]^\top$ be a vector such that $\psi_1 = 1$ and

$$\psi_k = 2 * \cos\left(\frac{\pi}{\lfloor (n-1)/(k-1)\rfloor + 2}\right), \quad k = 2, \cdots, n. \tag{2.15}$$

Also, let $\zeta = \left[|(\mathbf{\Phi}\mathbf{\Phi}^H)_{1,1}|, |(\mathbf{\Phi}\mathbf{\Phi}^H)_{1,2}|, \cdots, |(\mathbf{\Phi}\mathbf{\Phi}^H)_{1,n}|\right]^\top$, that is, the modulus of the terms in first row of $\mathbf{\Phi}\mathbf{\Phi}^H$. Let λ_k be the k-th eigenvalue of $\mathbf{\Phi}\mathbf{\Phi}^H$. Then it follows that (Hertz, 1992):

$$\max_{1\leqslant k\leqslant n}(\lambda_k) \leqslant \zeta^\top\boldsymbol{\psi}. \tag{2.16}$$

Further observe that $\max_{1\leqslant k\leqslant n}(\psi_k) = 2$, hence, together with the bound

on the values of ζ from Equation (2.13) it follows that

$$\zeta^T\psi = n + \sum_{k=2}^{n} |(\Phi\Phi^H)_{1,k}|\psi_k \;\leqslant \left(n + n\sum_{k=2}^{n} 2\right) = 2n^2 - n. \qquad (2.17)$$

Hence the norm $\|\Phi^H\|_2 \leqslant 2n^2 - n$. It now remains to bound the quantity $\|(\Phi^H\Phi)^\dagger\|_2$. By computing the singular value decomposition of $\|(\Phi^H\Phi)^\dagger\|_2$, it follows directly that

$$\|(\Phi^H\Phi)^\dagger\|_2 \leqslant \frac{1}{\sigma^2_{min}(\Phi)}, \qquad (2.18)$$

where $\sigma^2_{min}(\Phi) > 0$ is the smallest non-zero singular value of Φ. Using the fact that

$$\sum_{k=1}^{n} \lambda_k(\Phi^H\Phi) = \mathrm{Tr}(\Phi^H\Phi) = n^2, \qquad (2.19)$$

where $\mathrm{Tr}(\Phi^H\Phi)$ is the matrix trace, it follows that the smallest non-zero eigenvalue is bounded as $0 < \lambda_{min}(\Phi^H\Phi) \leqslant n$ which implies that $\sigma^2_{min}(\Phi^H\Phi) \leqslant n$. For any such that $\sigma^2_{min}(\Phi^H\Phi)$, we can always find a $0 < c \leqslant 1$ such that $\sigma^2_{min}(\Phi^H\Phi) > cn$. Hence the bound in Equation (2.18) can be extended to

$$\|(\Phi^H\Phi)^\dagger\|_2 \leqslant \frac{1}{\sigma^2_{min}(\Phi)} \leqslant \frac{1}{cn}. \qquad (2.20)$$

Now using the bound in Equations (2.17) and (2.20), the bound in Equa-

tion (2.11) has the form

$$\|\Pi_{\mathbf{\Phi}}\left(\mathbf{x}_i^* - F(\mathbf{s}(t_{p,i}))\right)\|_2 \leqslant \frac{2n-1}{c}\|F^*(\mathbf{s}(t_{p,i})) - F(\mathbf{s}(t_{p,i}))\|_2 \tag{2.21}$$

The bound on the Hausdorff distance between $\mathbf{F}'$ and $\mathbf{F}$ is then expressed as

$$\begin{aligned} d_H(\widehat{\Pi}_{\mathbf{\Phi}}(\mathbf{x}^*), \mathbf{F}) &\leqslant \sup_{i,p}\|\Pi_{\mathbf{\Phi}}\left(\mathbf{x}_i^* - F(\mathbf{s}(t_{p,i}))\right)\|_2 \\ &\leqslant \frac{2n-1}{c}\sup_{i,p}\|F^*(\mathbf{s}(t_{p,i})) - F(\mathbf{s}(t_{p,i}))\|_2, \end{aligned} \tag{2.22}$$

where $1 \leqslant i \leqslant n_p - M\tau_p, 1 \leqslant p \leqslant P$. From the equality in Equation (2.10), it holds that

$$d_B(\mathrm{Dgm}(\mathbf{F}), \mathrm{Dgm}(\mathbf{F}')) \leqslant 2d_{\mathbf{GH}}\left(\mathbf{F}, \mathbf{F}'\right) = 2d_{\mathbf{GH}}\left(\widehat{\Pi}_{\mathbf{\Phi}}(\mathbf{x}^*), \mathbf{F}\right). \tag{2.23}$$

The Gromov-Hausdorff distance is further bounded above by the Hausdorff distance. From Equations (2.22), the bound on the bottleneck distance between $\mathrm{Dgm}(\mathbf{F})$ and $\mathrm{Dgm}(\mathbf{F}')$ is established as

$$d_B(\mathrm{Dgm}(\mathbf{F}), \mathrm{Dgm}(\mathbf{F}')) \leqslant \frac{4n-2}{c}\left(\sup_{i,p}\|F^*(\mathbf{s}(t_{p,i})) - F(\mathbf{s}(t_{p,i}))\|_2\right), \tag{2.24}$$

where $1 \leqslant i \leqslant n_p - M\tau_p, \quad 1 \leqslant p \leqslant P$. If the observed time series is 'noise-free' such that $\mathbf{x}^* = \mathbf{x}$, then $\sup_{i,p}\|F(\mathbf{s}(t_{p,i})) - F^*(\mathbf{s}(t_{p,i}))\|_2 = 0, \forall i, p$, and $d_B(\mathrm{Dgm}(\mathbf{F}), \mathrm{Dgm}(\mathbf{F}')) = 0$. □

Remark 2.2. *When the samples are uniformly-spaced in both the time and frequency domain, the matrix* $\mathbf{\Phi}$ *is Hermitian with orthogonal columns, hence the factor* $(2n-1)/c$ *is not required for the bound to hold. The constant* c *depends on the* ℓ_2*-norm of* $\mathbf{\Phi}^H\mathbf{\Phi}$*, and the factor* $(2n-1)/c$ *makes the bound conservative. However, if the denoising is done well, the bottleneck distance stays significantly below this bound. Numerical experiments in Section 2.5.1 suggest this is the case for the settings considered.*

A point to emphasize is that the dependence of the bound on sample size n *arises mainly from irregular spacing in the time series. When samples are not uniformly spaced, complexities introduced by this irregularity lead to a bound that scales with* n*. This proportionality is not indicative of a flaw in the denoising method but rather a reflection of the additional challenges posed by irregularly-spaced time sampling. The uniform bound in Equation* (2.8) *is not an asymptotic bound, so convergence with increasing sample is not expected in this context. Theorem 2.5 discusses some convergence results in the presence of irregular sampling.*

2.4.2 Stability of the Subsequence Embedding Method

The objective of this section is to show that the SSE method provides a stable approximation, in the topological sense, to the TDE (based on uniformly-spaced time series data). Recall that the SSE method is designed for cases where data are irregularly spaced, but the SSE reduces to the standard TDE for uniformly-spaced data. Hence, we present results in this section that show that the SSE construction remains close to the TDE

construction, and that small perturbation in the SSE space results in small perturbations in its topology. To simplify the notation, the embedding matrices are represented as sets where the elements of the set are the row vectors of the corresponding TDE or SSE. Also, for an embedding from a single uniformly-spaced time series, the step-size is assumed to be τ. When constructing from a set of P subsequences, a step-size of τ_p is assumed for the p-th subsequence, where $1 \leqslant p \leqslant P$.

Because the SSE can have fewer elements than the TDE, the following lemma addresses how to expand the SSE without affecting its topology by repeating already existing points in the SSE, so that distances can be computed between the TDE and the expanded SSE.

Lemma 2.3 (Topology-preserving transform)**.** *Let* $\mathbf{F}^1$ *be an embedding matrix from a uniformly-spaced time sequence of length* n *with the form:*

$$\mathbf{F}^1 = \left\{F^1(\mathbf{s}(t_1)), F^1(\mathbf{s}(t_2)), \cdots, F^1(\mathbf{s}(t_{n-M\tau}))\right\} \subset \mathbb{R}^{M+1}. \tag{2.25}$$

Also, let $\mathbf{F}^2$ *be an embedding matrix from a set of* P *subsequences with the form:*

$$\mathbf{F}^2 = \left\{F^2(\mathbf{s}(t_{1,1})), \cdots, F^2(\mathbf{s}(t_{1,n_1-M\tau_1})), \cdots, F^2(\mathbf{s}(t_{P,n_P-M\tau_P}))\right\} \subset \mathbb{R}^{M+1}, \tag{2.26}$$

where $\sum_{p=1}^{P}(n_p - M\tau_p) \leqslant n - M\tau$*. Consider the set extension* $\widehat{\mathbf{F}}^2 = \left\{\mathbf{F}^2, \mathbf{F}^2_k\right\}$*, where* $\mathbf{F}^2_k$ *is a subset of* k *elements from* $\mathbf{F}^2$*. Then the persistence diagrams as-*

sociated with $\mathbf{F}^2$ *and* $\widehat{\mathbf{F}}^2$ *are identical, that is,* $Dgm(\mathbf{F}^2) \equiv Dgm(\widehat{\mathbf{F}}^2)$*, and the three embedded spaces are related through the bottleneck distance as follows:* $d_B(Dgm(\mathbf{F}^1), Dgm(\mathbf{F}^2)) = d_B(Dgm(\mathbf{F}^1), Dgm(\widehat{\mathbf{F}}^2))$.

Proof. By construction, $\mathbf{F}^2 \subset \widehat{\mathbf{F}}^2$, thus for any $F^2(s(t_{p,i_1})) \in \mathbf{F}^2$, $\exists \widehat{F}^2(s(t_{p,i_2})) \in \widehat{\mathbf{F}}^2$ such that $F^2(s(t_{p,i_1})) = \widehat{F}^2(s(t_{p,i_2}))$. Further observe that $|\mathbf{F}^2|_u = |\hat{\mathbf{F}}^2|_u$, where $|.|_u$ is a measure of the cardinality of unique observations. Hence it follows that $\mathrm{Dgm}(\mathbf{F}^2) \equiv \mathrm{Dgm}(\widehat{\mathbf{F}}^2)$, and the conclusion is a direct consequence of this equivalence. □

Lemma 2.3 asserts that duplicating points from an embedding matrix does not change the SSE's persistence diagram. This is due to the fact that the duplicated points do not introduce new data points locations in the embedding. This is used to establish a bound on the SSE as an approximation to the TDE in the following proposition. In Lemma 2.3, when $k = (n - M\tau) - \sum_{p=1}^{P}(n_p - M\tau_p)$, the row dimension of $\widehat{\mathbf{F}}^2$ is the same as that of $\mathbf{F}^1$. In such instances, when the interest is in a row-wise comparison of $\widehat{\mathbf{F}}^2$ and $\mathbf{F}^1$ the subsequence indexing in the time variable for any $\widehat{F}^2(\mathbf{s}(t_{p,i})) \in \widehat{\mathbf{F}}^2$ is ignored for notational convenience, and a row is simply written as $\widehat{F}^2(\mathbf{s}(t_i))$, where $1 \leqslant i \leqslant n - M\tau$.

Proposition 2.2. *Let* $\mathbf{x}^1$ *be a uniformly-spaced time series vector of length* n *with TDE matrix* $\mathbf{F}^1$*. Let* $\mathbf{x}^2 \subset \mathbf{x}^1$ *be a time series vector where some of the elements are*

missing or unobserved. Denote the SSE matrix constructed from $\mathbf{x}^2$ *as* $\mathbf{F}^2$*. Define the extension* $\widehat{\mathbf{F}}^2 = \left\{\mathbf{F}^2, \mathbf{F}^2_k\right\}$*, where* $\mathbf{F}^2_k$ *is a subset of* $k = (n - M\tau) - \sum_{p=1}^{P}(n_p - M\tau_p)$ *elements from* $\mathbf{F}^2$*. Then the bottleneck distance between* $\mathbf{F}^1$ *and* $\mathbf{F}^2$ *is bounded as:* $d_B(Dgm(\mathbf{F}^1), Dgm(\mathbf{F}^2)) \leqslant 2 \sup_{1 \leqslant i \leqslant n - M\tau} \|F^1(\mathbf{s}(t_i)) - \widehat{F}^2(\mathbf{s}(t_i))\|_2$.

The choice of the k subsets of embedding vectors $\mathbf{F}^2_k$ in Proposition 2.2 is arbitrary as any subset satisfies the bound. However, since they are chosen to match the subset $\{\mathbf{F}'(\mathbf{s}_{t_l}) : \sum_{p=1}^{P}(n_p - M\tau_p) + 1 \leqslant l \leqslant n - M\tau\}$ of $\mathbf{F}^1$, the bound can be improved. The minimum bound can be attained by choosing a subset in $\mathbf{F}^2$ that has the smallest Euclidean distance to the subset $\{\mathbf{F}'(\mathbf{s}_{t_l}) : \sum_{p=1}^{P}(n_p - M\tau_p) + 1 \leqslant l \leqslant n - M\tau\}$. This is summarized as a corollary below.

Corollary 2.4. *Let* $\mathbf{x}^1$ *be a uniformly spaced time series vector of length* n *with TDE matrix* $\mathbf{F}^1$*. Let* $\mathbf{x}^2 \subset \mathbf{x}^1$ *be a time series vector where some of the elements are unobserved, with* $\mathbf{F}^2$ *as its SSE matrix. Define the extension* $\widehat{\mathbf{F}}^2 = \left\{\mathbf{F}^2, \mathbf{F}^2_k\right\}$*, where* $\mathbf{F}^2_k$ *is a random subset of* $k = (n - M\tau) - \sum_{p=1}^{P}(n_p - M\tau_p)$ *elements from* $\mathbf{F}^2$*. For some* $F^1 \in \mathbf{F}^1$*, define the set* $\mathbf{F}^2_{k,min}$ *as follows:*

$$\mathbf{F}^2_{k,min} = \left\{F^2_{min} \in \mathbf{F}^2 : \|F^2_{min} - F^1\|_2 \leqslant \|F^2 - F^1\|_2, \forall F^2 \in \mathbf{F}^2, \text{ s.t. } F^2_{min} \neq F^2\right\}. \tag{2.27}$$

That is, $\mathbf{F}^2_{k,min}$ *is a subset of* k *embedding vectors in* $\mathbf{F}^2$ *with minimum distance to*

some points in $\mathbf{F}^1$. *Let* $\widehat{\mathbf{F}}^2_{min} = \left\{\mathbf{F}^2, \mathbf{F}^2_{k,min}\right\}$, *then it follows that*

$$\sup_{1\leqslant i\leqslant n} \|F^1(\mathbf{s}(t_i)) - \widehat{F}^2_{min}(\mathbf{s}(t_i))\|_2 \leqslant \sup_{1\leqslant i\leqslant n} \|F^1(\mathbf{s}(t_i)) - \widehat{F}^2(\mathbf{s}(t_i))\|_2, \quad (2.28)$$

where $F^1(\mathbf{s}(t_i)) \in \mathbf{F}^1$, $\widehat{F}^2(\mathbf{s}(t_i)) \in \widehat{\mathbf{F}}^2$, *and* $\widehat{F}^2_{min}(\mathbf{s}(t_i)) \in \widehat{\mathbf{F}}^2_{min}$.

An immediate consequence of Proposition 2.2 and Corollary 2.4 is that the SSE matrix approximates the TDE. In particular, for a time series $\mathbf{x} = \{x(t) : t \in \mathcal{T}\}$, and $\mathcal{T} = \{t_1, \cdots, t_n\} \subset \mathbb{N}$, the sequence of embedding matrices for each r where $1 \leqslant r \leqslant t_n - t_1$ is finite. Hence for a fixed n, the limiting persistence diagram as $r \to 1$ is close to the TDE's persistence diagram. If $r = 1$, the SSE is exactly the same as the TDE, and the persistence diagrams would be identical. This reinforces the fact that the proposed reconstruction preserves the topological structures more accurately as the level of irregularity in the observed time series decreases. A more formal treatment of these observations is presented next.

2.4.3 Topology Recovery and Convergence Results

This section presents results on the quality of the topological recovery for varying proportions of missingness and sample sizes. We assume the number of missing values increases at a slower rate than the sample size of the time series; specifically, a rate of $o(\log m)$, for sample size m. This assumption and others are formalized as follows.

Let $\mathbf{x}_1, \mathbf{x}_2, \cdots,$ be irregularly-spaced time series vectors where $|\mathbf{x}_i| < |\mathbf{x}_j|$, $i < j$. Denote by $\mathbf{F}_m$, the SSE associated with $\mathbf{x} \in \{\mathbf{x}_1, \mathbf{x}_2, \cdots,\}$, i.e., $\mathbf{F}_m = \{F(\mathbf{s}(t_1)), F(\mathbf{s}(t_2)), \cdots, F(\mathbf{s}(t_m))\}$. Note the correspondence between the subscript m and the number of points in $\mathbf{F}_m$. $\mathbf{F}_m$ depends on which time series is selected; however, indexing over this selection is not needed for the following results. Recall that $\mathbf{F}_m$ is a compact subset of $(\mathbb{R}^{M+1}, \|\cdot\|_2)$. Let the space $(\mathbb{R}^{M+1}, \|\cdot\|_2)$ be endowed with the unknown probability measure ϑ such that the $F(\mathbf{s}(\mathbf{t_k}))$'s are randomly sampled according to ϑ. Let ϑ be supported on the set $\mathbf{F}_\vartheta$, which can be considered the true underlying state space to be estimated, and let φ be the associated density function. Consider the following set of assumptions.

A1. The sample size increases such that $\mathbf{x}_i \subset \mathbf{x}_j$ whenever $i < j$.

A2. Let $\varepsilon_m(r)$ be a function of m and the regularity score r such that $\varepsilon_m(1) \to 0$ as $m \to \infty$.

A3. For any point $F_\vartheta \in \mathbf{F}_\vartheta$, $\vartheta(B(F_\vartheta, \delta)) \geqslant \min(\kappa\delta^{M+1}, 1)$, where $B(F_\vartheta, \delta)$ is a closed ball of radius $\delta > 0$ around F_ϑ, with constant $\kappa > 0$.

A4. It is possible to create joint distributions based on the marginals of $\mathbf{F}_m$ that satisfy

$$\sup_{\mathbf{F}_m} \left| \frac{\varphi\left(F(\mathbf{s}(t_1)), \cdots, F(\mathbf{s}(t_m))\right) - \varphi\left(F(\mathbf{s}(t_1))\right) \times \cdots \times \varphi\left(F(\mathbf{s}(t_m))\right)}{\varphi\left(F(\mathbf{s}(t_1))\right) \times \cdots \times \varphi\left(F(\mathbf{s}(t_m))\right)} \right| \leqslant \eta_m, \tag{2.29}$$

where η_m is such that $\sum_{m=1}^{\infty} \frac{\eta_m}{m^\beta \log(m)} < \infty$ for any $\beta > 1$.

Assumption **A4** is to address the possible lack of independence of the vectors in $\mathbf{F}_m$. Under this assumption, the dependence can be controlled and the vectors in $\mathbf{F}_m$ are regarded as the so-called η_m-almost independent samples, which allows for $\mathbf{F}_m$ to converge in Hausdorff distance to $\mathbf{F}_\vartheta$ (Aaron et al., 2017; Picado and Oliveira, 2020). This assumption, where almost-independence can be achieved for a time series in its embedding space, can be satisfied for a suitable embedding window $(M+1)\tau$ as illustrated by the following empirical example. Consider the function $h(t) = 4\cos(t)^3 \times \sin(t)^3$. We simulate a time series from this function (see Figure 2.3a for an example of the simulated time series), construct the embedding with $M = 1$ and $\tau = 3$ to obtain the point cloud $\mathbf{F}_m$. To check Assumption **A4**, we approximate η_m, denoted by η_m^*, for increasing m by evaluating the left side of Equation (2.29) via the k-nearest neighbor density estimates, where $k = 10$. For each point $F(\mathbf{s}(t_i))$, we generate $\mathcal{N} = 1000$ noise-perturbed replicates, were the noise are drawn from a normal distribution with mean 0 and standard deviation 0.01. The marginal density is estimated as:

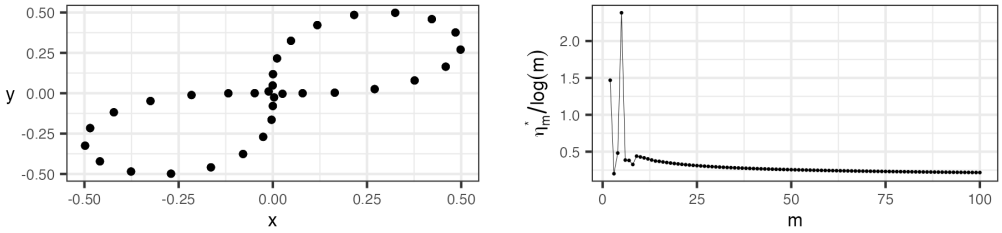
$$\varphi\left(F(\mathbf{s}(t_i))\right) = \frac{k/(\mathcal{N} \times \mathrm{vol}_{M+1})}{r_k\left[F(\mathbf{s}(t_i))\right]^{M+1}}, \tag{2.30}$$

where vol_{M+1} is the volume of a unit M-sphere, and $r_k\left[F(\mathbf{s}(t_i))\right]$ is the distance to the k-th nearest neighbor. Similarly, the joint density is also

estimated as follows:

$$\varphi\left(F(\mathbf{s}(t_1)),\cdots,F(\mathbf{s}(t_m))\right)=\frac{k/\left(\mathcal{N}\times \mathrm{vol}_{m(M+1)}\right)}{r_k\left[\varphi\left(F(\mathbf{s}(t_1)),\cdots,F(\mathbf{s}(t_m))\right)\right]^{m(M+1)}}. \tag{2.31}$$

Checking that $\eta_m^*/\log(m)$ converges to zero for large m is sufficient to validate Assumption **A4**. Figure 2.4 shows the embedded space and the



(a) A sample embedding (b) The convergence of $\eta_m^*/\log(m)$

Figure 2.4: Illustration of Assumption **A4**. (a) An example embedding constructed from the time series $h(t)=4\cos(t)^3\times\sin(t)^3$ using 100 time points; see Figure 2.3a. (b) The convergence results of $\eta_m^*/\log(m)$ for increasing m; see Equation (2.29), where η_m^* denotes the lower bound of η_m.

convergence curve. We observe that $\eta_m^*/\log(m)$ converges to zero with m, which guarantees that $\sum_{m=1}^{\infty}\frac{\eta_m^*}{m^\beta\log(m)}<\infty$.

The SSE matrix $\mathbf{F}_m$ can be regarded as an estimator of $\mathbf{F}_\vartheta$ and convergence results can be established in the context of assumptions **A1**-**A4**. These results are analogous to convergence results established on support estimation of d-dimensional sets (Cuevas and Rodríguez-Casal, 2004), its generalization to metric spaces, and on the space of persistence diagrams (Mileyko et al., 2011; Chazal et al., 2014). The following result gives the

rate of convergence of the SSE, $\mathbf{F}_m$, in estimating $\mathbf{F}_\vartheta$.

Theorem 2.5. *Let $\mathbf{x}_1, \mathbf{x}_2, \cdots,$ be a sequence of irregularly-spaced time series vectors satisfying assumption **A1**, and $\mathbf{F}_m = \{F(\mathbf{s}(t_1)), F(\mathbf{s}(t_2)), \cdots, F(\mathbf{s}(t_m))\} \subset \mathbb{R}^{M+1}$ be the SSE associated with some $\mathbf{x} \in \{\mathbf{x}_1, \mathbf{x}_2, \cdots,\}$, satisfying assumption **A2**. If the probability measure ϑ satisfies assumption **A3** and **A4**, then with probability one,*

$$\lim_{m \to \infty} \sup \left(\varepsilon_m(r)\right)^{-\frac{1}{M+1}} d_H(\mathbf{F}_m, \mathbf{F}_\vartheta) \leqslant K, \tag{2.32}$$

where K is a constant depending on κ and the embedding dimension $M+1$.

Proof. By construction of the subsequence and assumption **A2**, the Hausdorff distance between $\mathbf{F}_m$ and $\mathbf{F}_\vartheta$ has the form

$$d_H(\mathbf{F}_m, \mathbf{F}_\vartheta) = \sup_{F_\vartheta \in \mathbf{F}_\vartheta} \min_{1 \leqslant i \leqslant m} \|F(s(t_i)) - F_\vartheta\|_2. \tag{2.33}$$

Let $\mathbf{F}_0 \subset \mathbf{F}_\vartheta$ be a set of ball centers such that

$$\mathbf{F}_\vartheta \subset \bigcup_{F_0 \in \mathbf{F}_0} B(F_0, \delta), \tag{2.34}$$

that is, the minimal covering of $\mathbf{F}_\vartheta$ consisting of balls of radius δ around

$F_0 \in \mathbf{F}_0$. For any $F_\vartheta \in \mathbf{F}_\vartheta$ and $F_0 \in \mathbf{F}_0$, the following inequality holds:

$$\begin{aligned}\min_{1\leqslant i\leqslant m}\|F(s(t_i)) - F_\vartheta\|_2 &\leqslant \|F(s(t_j)) - F_\vartheta\|_2 \\ &= \|F(s(t_j)) - F_0 + F_0 - F_\vartheta\|_2 \\ &\leqslant \|F(s(t_j)) - F_0\|_2 + \|F_0 - F_\vartheta\|_2, \quad j = 1, \cdots, m.\end{aligned} \tag{2.35}$$

Observe that $\|F_0 - F_\vartheta\|_2$ is bounded by the radius δ, hence using Equation (2.34), it follows that

$$\min_{1\leqslant i\leqslant m}\|F(s(t_i)) - F_\vartheta\|_2 \leqslant \delta + \max_{F_0\in\mathbf{F}_0}\min_{1\leqslant i\leqslant m}\|F(s(t_i)) - F_0\|_2 \leqslant \varepsilon, \tag{2.36}$$

for some $\epsilon > 0$. Further, taking the supremum over all F_ϑ, the relation still holds:

$$\sup_{F_\vartheta\in\mathbf{F}_\vartheta}\min_{1\leqslant i\leqslant m}\|F(s(t_i)) - F_\vartheta\|_2 \leqslant \delta + \max_{F_0\in\mathbf{F}_0}\min_{1\leqslant i\leqslant m}\|F(s(t_i)) - F_0\|_2 \leqslant \varepsilon, \tag{2.37}$$

Then the probability that $\sup_{F_\vartheta\in\mathbf{F}_\vartheta}\min_{1\leqslant i\leqslant m}\|F(s(t_i)) - F_\vartheta\|_2$ exceeds ε is bounded as

$$\begin{aligned}\Pr\left(\sup_{F_\vartheta\in\mathbf{F}_\vartheta}\min_{1\leqslant i\leqslant m}\|F(s(t_i)) - F_\vartheta\|_2 > \varepsilon\right) &\leqslant \Pr\left(\delta + \max_{F_0\in\mathbf{F}_0}\min_{1\leqslant i\leqslant m}\|F(s(t_i)) - F_0\|_2 > \varepsilon\right) \\ &= \Pr\left(\max_{F_0\in\mathbf{F}_0}\min_{1\leqslant i\leqslant m}\|F(s(t_i)) - F_0\|_2 > \varepsilon - \delta\right).\end{aligned} \tag{2.38}$$

From Equation (2.33), $\sup_{F_\vartheta \in \mathbf{F}_\vartheta} \min_{1 \leqslant i \leqslant m} \|F(s(t_i)) - F_\vartheta\|_2 = d_H(\mathbf{F}_m, \mathbf{F}_\vartheta)$, hence a bound on the probability of $d_H(\mathbf{F}_m, \mathbf{F}_\vartheta)$ exceeding ε can be obtained as

$$\Pr\left(d_H(\mathbf{F}_m, \mathbf{F}_\vartheta) > \varepsilon\right) \leqslant \Pr\left(\max_{F_0 \in \mathbf{F}_0} \min_{1 \leqslant i \leqslant m} \|F(s(t_i)) - F_0\|_2 > \varepsilon - \delta\right). \quad (2.39)$$

Observe that $\mathbf{F}_\vartheta$ endowed with the Hausdorff metric is complete and separable (Attouch et al., 1991). Then for ε small enough, the following bound holds (Cuevas and Rodríguez-Casal, 2004):

$$\Pr\left(\max_{F_0 \in \mathbf{F}_0} \min_{1 \leqslant i \leqslant m} \|F(s(t_i)) - F_0\|_2 > \varepsilon - \delta\right) \leqslant C\left(1 - \kappa\omega(\varepsilon - \delta)^{M+1}\right)^m. \quad (2.40)$$

The constant C is the number of points in the covering of $\mathbf{F}_\vartheta$, i.e., $|\mathbf{F}_0|$, and ω is the Lebesgue measure of the unit ball in $\mathbb{R}^{M+1}$. Note that since $0 \leqslant \kappa\omega(\varepsilon - \delta)^{M+1} \leqslant 1$, it follows that $\left(1 - \kappa\omega(\varepsilon - \delta)^{M+1}\right)^m \leqslant e^{-m\kappa\omega(\varepsilon-\delta)^{M+1}}$. This allows for Equations (2.39) and (2.40) to be rewritten as

$$\Pr\left(d_H(\mathbf{F}_m, \mathbf{F}_\vartheta) > \varepsilon\right) \leqslant \Pr\left(\max_{F_0 \in \mathbf{F}_0} \min_{1 \leqslant i \leqslant m} \|F(s(t_i)) - F_0\|_2 > \varepsilon - \delta\right) \leqslant Ce^{-m\kappa\omega(\varepsilon-\delta)^{M+1}}. \quad (2.41)$$

Choose some $K > \left(\frac{2}{\kappa\omega}\right)^{\frac{1}{M+1}}$ and let $\varepsilon_m(r) = \left(r - \frac{m-l}{m}\right)$, where l is the number of missing observations in the initial time series and $m \gg l$, then

for m large enough, it follows that

$$\Pr\left((\varepsilon_m(r))^{-\frac{1}{M+1}}\, d_H(\mathbf{F}_m, \mathbf{F}_\vartheta) > K\right) \leqslant Ce^{-m\kappa\omega\left((\varepsilon_m(r)\frac{2}{\kappa\omega})^{\frac{1}{M+1}}-\delta\right)^{M+1}}. \tag{2.42}$$

The above bound can be obtained by simply substituting $\left(\frac{2\varepsilon_m(r)}{\kappa\omega}\right)^{\frac{1}{M+1}}$ for ε in Equation (2.41). Now consider the sum

$$\sum_m e^{-m\kappa\omega\left((\varepsilon_m(r)\frac{2}{\kappa\omega})^{\frac{1}{M+1}}-\delta\right)^{M+1}}, \tag{2.43}$$

and observe that it is convergent if $\varepsilon_m(r) \geqslant \left(\frac{\delta}{K}\right)^{M+1}$. This condition can always be satisfied given the restriction $K > \left(\frac{2}{\kappa\omega}\right)^{\frac{1}{M+1}}$ and for an appropriate choice of κ and δ. Then by the Borel-Cantelli lemma (Émile Borel, 1909; Cantelli, 1917; Chung and Erdös, 1952), since

$$\sum_m \Pr\left((\varepsilon_m(r))^{-\frac{1}{M+1}}\, d_H(\mathbf{F}_m, \mathbf{F}_\vartheta) > K\right) < \infty, \tag{2.44}$$

it follows that

$$\lim_{m\to\infty} \sup\, (\varepsilon_m(r))^{-\frac{1}{M+1}}\, d_H(\mathbf{F}_m, \mathbf{F}_\vartheta) \leqslant K. \tag{2.45}$$

□

From the stability relation in Equation (1.6), the Hausdorff metric can be replaced with the Gromov-Hausdorff metric and the results still holds.

This also gives a similar convergence results on the space of persistence diagrams with respect to the bottleneck distance and is summarized as:

Corollary 2.6. *Let* $\mathbf{x}_1, \mathbf{x}_2, \cdots,$ *be a sequence of irregularly-spaced time series vectors satisfying assumption* ***A1****, and let* $\mathbf{F}_m = \{F(\mathbf{s}(t_1)), F(\mathbf{s}(t_2)), \cdots, F(\mathbf{s}(t_m))\} \subset \mathbb{R}^{M+1}$ *be the SSE associated with some* $\mathbf{x} \in \{\mathbf{x}_1, \mathbf{x}_2, \cdots,\}$*, satisfying assumption* ***A2****. If the probability measure* ϑ *satisfies assumption* ***A3*** *and* ***A4****, then with probability one,*

$$\lim_{m \to \infty} \sup \left(\varepsilon_m(r)\right)^{-\frac{1}{M+1}} d_B\left(Dgm(\mathbf{F}_m), Dgm(\mathbf{F}_\vartheta)\right) \leqslant K, \tag{2.46}$$

where K *is a constant depending on* κ *and the embedding dimension* $M + 1$*.*

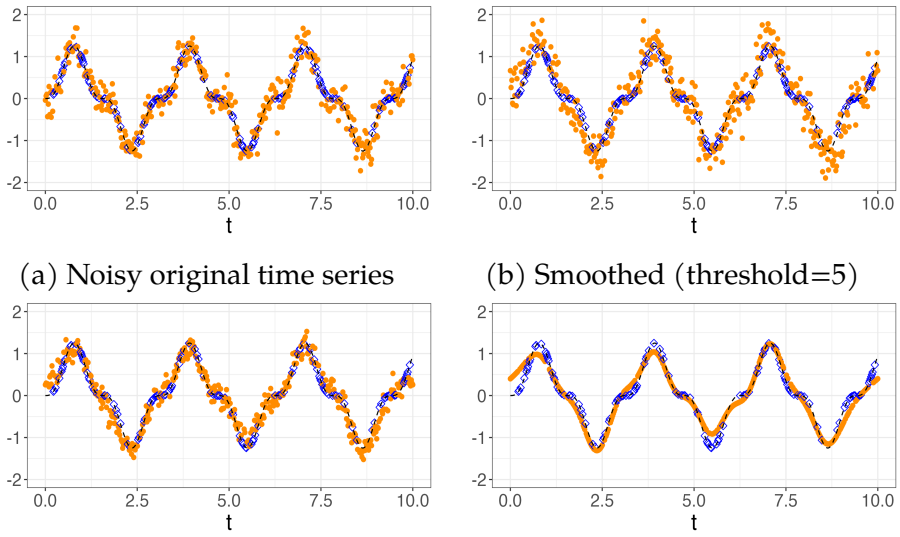
2.5 Numerical Studies

This section presents numerical studies that evaluates the performance of the proposed SSE method.

2.5.1 Evaluation of Denoising Procedure

To evaluate the performance of the denoising procedure presented in this work, and Proposition 2.1, the time series in Figure 2.3a was replicated at varying noise levels and sample sizes. The probability that any value is unobserved at a given time point is fixed at 0.25. Four noise levels

$\{0, 0.25, 0.5, 2\}$ and five samples sizes $\{50, 100, 500, 1000, 5000\}$ were used in the simulations. For each noise level and sample size combination, the denoising method outlined in Section 2.4.1 was performed and the bottleneck distance between the corresponding persistence diagrams and the theoretical upper-bound are computed. The upper bound computed does not include the multiplicative factor $\frac{2n-1}{c}$. Figure 2.5 shows a noisy time series and the outcome after denoising at various frequency thresholds. For an appropriate choice of frequency threshold, which is generally



(a) Noisy original time series (b) Smoothed (threshold=5)

(c) Smoothed (threshold=15) (d) Smoothed (threshold=25)

Figure 2.5: Illustration of the denoising method of Section 2.4.1. The time series was perturbed with noise drawn from a $N(0, 0.25)$, and the probability of a missing observation is 0.25. (a) The original 500 time series measurements. The orange points are observed values, while the blue diamonds are the missing values displayed at the true signal value without noise. The other sub-figures display the time series after denoising with a frequency threshold of 5 (b), 15 (c), and 25 (d).

application dependent, the true underlying signal can be satisfactorily recovered.

For each combination of the noise-level and sample size, the process is repeated 100 times and standard errors are obtained. The results are presented in Figure 2.6. The bottleneck distance is bounded above by the

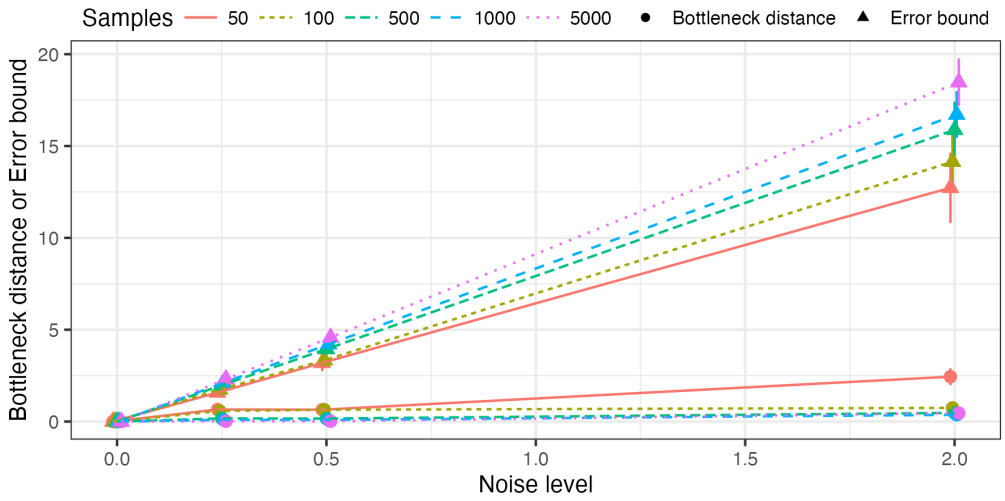


Figure 2.6: Stability results of the denoising procedure (see Proposition 2.1). The solid points represent the mean values from 100 repetitions, the vertical lines on these points indicate the error bars (which are too small to see in many cases), and the colors and line types indicate the sample size. The vertical axis represents the bottleneck distance for the circle points and the error bound (without the multiplicative factor $\frac{2n-1}{c}$) for the triangle points.

error bound for all noise levels and sample sizes as expected. At the same noise level, smaller sample sizes tend to have larger bottleneck distances. This can partly be explained by the fact that the SSE is more sparse (i.e., points in the space are more spread out since there are fewer points). The H_0 features are more likely to persist longer in such sparse settings. The

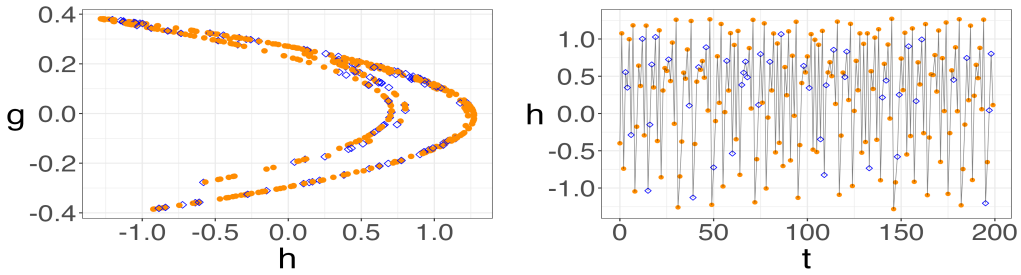
reverse is true for the error bound in Proposition 2.1, which for the same noise level, is higher for larger sample sizes. This follows from the fact that larger sample sizes increases the chance of observing highly noisy terms (and the upper bound is includes a supremum). These results demonstrate the denoising procedure's efficacy in controlling noise effects on the SSE's topological features.

Proposition 2.1 establishes a conservative bound on the bottleneck distance between persistence diagrams of a noise-free and a denoised time series using Fourier methods. The factor $(2n-1)/c$ reflects the poor-conditioning of the Fourier matrix in non-uniform domains. Empirical evidence suggests this bound could be improved in more restricted settings, which is a topic for future research.

2.5.2 Reconstruction Accuracy

The empirical study in this section was designed to assess the SSE method's effectiveness in preserving the original state space geometry using the Hénon map as an illustrative example (Hénon, 2004). The Hénon map recursively maps a point $(h_t, g_t) \in \mathbb{R}^2$ as follows: $h_{t+1} = 1 - ah_t^2 + g_t$, $g_{t+1} = bh_t$, with $a = 1.4$ and $b = 0.3$ (i.e., their classical values). The map is initialized at $(h_0, g_0) = (0, 0)$, and simulated with 500 points with observations designated as missing with a given probability. Figure 2.7 shows the 2D Hénon map and the corresponding time series for one dimensions. The measurement function (see Section 2.2) extracts observations along

the h-dimension, hence $\{h_t\}$ are used to reconstruct the space. Observations along the g-dimension could be used instead.



(a) Hénon map (b) The h dimension

Figure 2.7: The Hénon map used in assessing reconstruction accuracy. (a) The Hénon map with 500 points (blue and orange) where the blue diamonds are designated as missing. (b) The h-dimension of the Hénon map; only 200 points are displayed for visual clarity.

The *correlation dimension* is used to assess how well the geometry of the original state space is preserved in the reconstruction. Specifically, for a given $\varepsilon > 0$, it measures the probability that two random points in a space are within ε-distance of each other. To compute the correlation dimension, the correlation sum is computed using the following:

$$\text{Corr}(\varepsilon) = \lim_{m\to\infty} \frac{2}{m(m-1)} \sum_{i=1}^{m} \sum_{j=i+1}^{m} \mathbb{1}\left(\|F(\mathbf{s}(t_i)) - F(\mathbf{s}(t_j))\|_2 \leqslant \varepsilon\right), \tag{2.47}$$

for some embedding map $\mathbf{F} = \{F(\mathbf{s}(t_1)), \cdots, F(\mathbf{s}(t_m))\}$. Then the correlation dimension is estimated as: $\lim_{\varepsilon\to 0} \log(\text{Corr}(\varepsilon))/\log(\varepsilon)$. If the reconstructed space preserves relevant geometrical invariants, its correlation dimension should match that of the TDE space. Other accuracy measures

include the box-counting dimension, Hausdorff dimension, and information dimension. However, the correlation dimension is more robust to sample size, making it less noisy with fewer samples (Grassberger and Procaccia, 1983b).

The SSE method is compared to common statistical interpolation methods used to impute missing data. A range of methods were considered[1], but only the best three methods are presented, which were implemented using the R package, *imputeTS* (Moritz and Bartz-Beielstein, 2017):

(1) Kalman Smoothing (KS): This fits a structural time series model via maximum likelihood, using the linear local trend as the structural class (see referenced package for more details).

(2) Last Observation Carried Forward (LOCF): This methods replaces each missing value with the most immediate prior observed value.

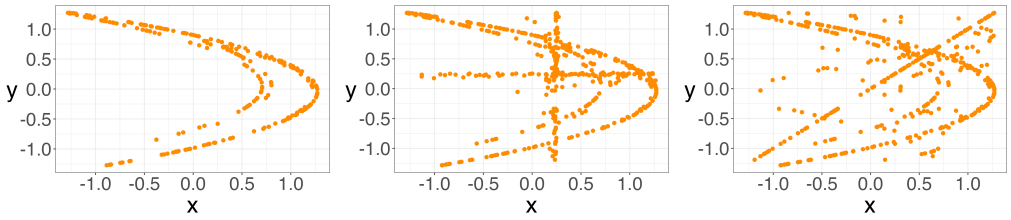
(3) Next Observation Carried Backward (NOCB): This is similar to the LOCF, but instead replaces each missing value with the most immediate *next* observed value.

The results are presented in terms of the correlation dimension, with standard errors generated by applying each method to 100 independently generated instances of the Hénon map. A noise model (with no missing

[1]The comparison methods considered are available in the the R package, *imputeTS* (Moritz and Bartz-Beielstein, 2017): *linear*, *spline*, and *Stineman* interpolation methods, *Kalman Smoothing* with a structural model and autoregressive integrated moving average model, a *moving average* method with exponential and linear weighting, *seasonal decomposition* (imputation by interpolation is done on the deseasonalized component), *seasonal split* (imputation by interpolation is done on each split), imputing with the *previous observation* (LOCF) or *next observation* (NOCB), imputing with the *mean*, *median*, *mode*, and by a *random point* in the dataset.

values) served as a baseline, with observations from a normal distribution (mean zero) and standard deviation equal to the probability of observing a missing value (for convenience).

Figure 2.8 shows example reconstructed spaces using the proposed method and two imputation methods (the NOCB result is nearly identical to the LOCF and is not shown) with 500 samples and a 0.25 missingness probability. Note that for the comparison methods, after imputation the TDE method is used to estimate the state space. Only the SSE method preserves the original geometry, while the imputation methods introduce extraneous features.



(a) SSE (b) KS imputation (c) LOCF imputation

Figure 2.8: Reconstructed state spaces of the Hénon map for: (a) proposed SSE method, (b) KS imputation, and (c) LOCF imputation.

Figure 2.9 shows the correlation dimension versus missingness probability for the SSE method and the three imputation methods. The black dashed lines indicate the established empirical estimate for the Hénon map's correlation dimension (1.22 ± 0.04) (Grassberger and Procaccia, 1983a; Sprott and Rowlands, 2001), so that a good performing method has empricial correlation dimensions within these bounds. The SSE method

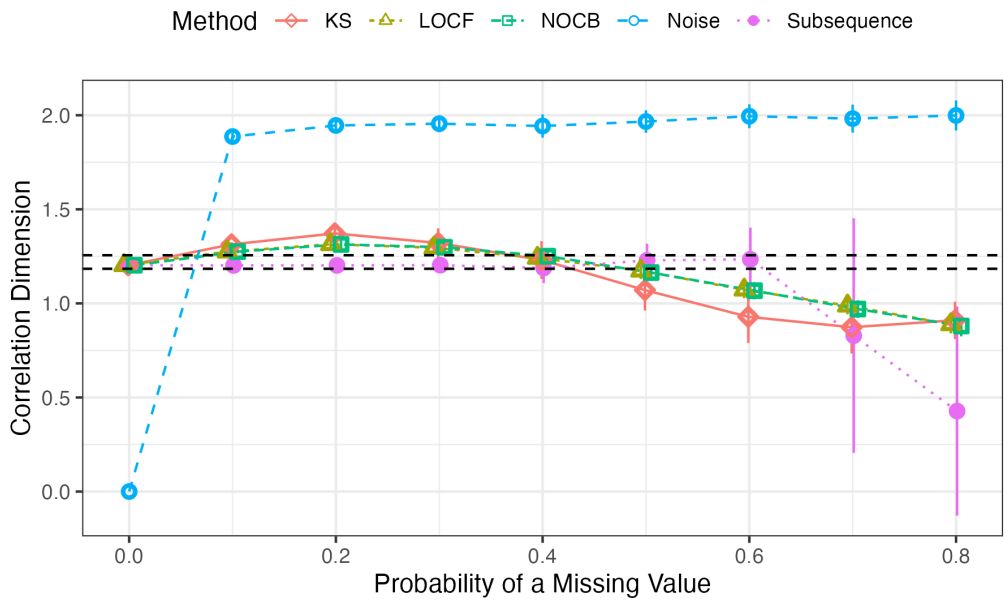


Figure 2.9: Reconstruction accuracy results of the Hénon map based on the correlation dimension. The points in different shapes are the mean correlation dimension after 100 repetitions using the proposed SSE method (solid pink points), the three imputation methods, and a baseline noise model (blue dashed), and the vertical bars represent the corresponding standard errors. The black dashed lines indicate the established empirical bounds of the Hénon map.

performs well up to a missingness probability of 0.6, staying within or close to the empirical bounds. Beyond 0.6, its comparable to the three imputation methods. However, the SSE method is more variable due to the fewer points used to compute the correlation dimension compared to the other methods (which always have 500 points).

2.5.3 Periodicity Quantification

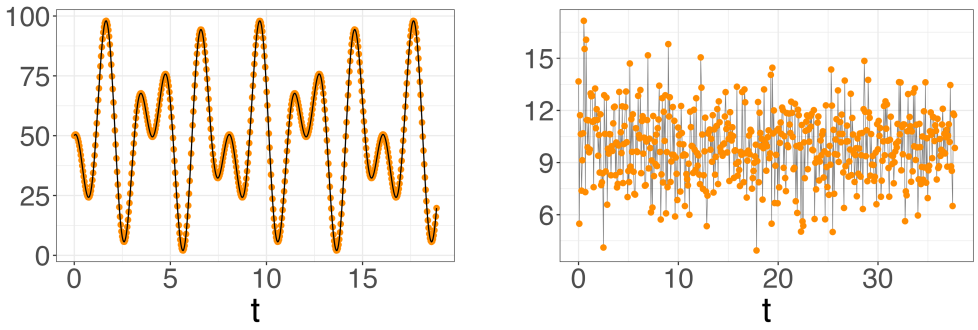
The periodicity of a time series can be quantified based on the H_1 features in the persistence diagram. This relies on the idea that periodic patterns yields elliptic curves in the reconstructed state space, and Perea and Harer (2015) use the roundness of the curves as an indicator of the periodicity in the time series. The roundness of these ellipses can be quantified by examining the maximum persistence of their associated H_1 features. For a time series vector $\mathbf{x}$ with its embedding map $\mathbf{F}$, its periodicity score $\mathrm{ps}(\mathbf{x})$ can be defined as (Perea and Harer, 2015):

$$\mathrm{ps}(\mathbf{x}) = \max_{(b,d)\in \mathrm{Dgm}(\mathbf{F})} (d-b)/\sqrt{3}, \tag{2.48}$$

where $\mathrm{Dgm}(\mathbf{F})$ is the persistence diagram, and $\max_{(b,d)\in \mathrm{Dgm}(\mathbf{F})}(d-b)$ is restricted to the H_1 features. For this calculation, the embedding map $\mathbf{F}$ is pointwise-centered and scaled. The motivation for the periodicity score is that during the VR filtration for a dataset with a large enough sample size, a unit circle (H_1 feature) dies when an inscribed equilateral triangle appears in the VR complex at filtration value $\sqrt{3}$, hence the maximum periodicity score of one is realized when either the TDE or SSE spaces has a well-sampled circle; a $\mathrm{ps}(\mathbf{x})$ closer to one indicates a stronger periodic signal in $\mathbf{x}$.

To evaluate this framework, two different set of signals were generated with sample sizes $n \in \{50, 100, 500, 1000\}$ and missingness probabilities

$\{0, 0.1, 0.2, 0.3, 0.4\}$. The first set follows $f(t) = 50 \times \cos(\pi t/4 - \lambda\pi) \times \sin(\pi t/2) + 50$ with $\lambda \in (0, 1)$ and $t \in [0, 12\pi]$, having a longest period of 8. The second set is a non-periodic signals drawn from a $N(10, 2)$. Figure 2.10 shows samples of both signals. To construct the embedding from the time



(a) Periodic signal (b) Non-periodic signal

Figure 2.10: Sample periodic (a) and non-periodic (b) signals used in the periodicity quantification simulation of Section 2.5.3. Each time series include 500 time points.

series, the time points are rescaled to integers and the step-size is set to $\tau = 1$. The periodicity score $ps(\mathbf{x})$ is then compared to those obtained from the Lomb-Scargle periodogram method for both uniformly-spaced and irregularly-spaced observations (Lomb, 1976; Scargle, 1982; Ruf, 1999), the sliding windows method (Perea and Harer, 2015), and the JTK_Cycle algorithm for uniformly-spaced samples (Hughes et al., 2010).

The results are summarized in Table 2.1 (periodic signal) and in Table 2.2 (non-periodic signal). Table 2.1 shows that all the methods rate the periodic signals as highly periodic with increasing sample size. The proposed SSE method consistently identifies a distinct H_1 across all sample

sizes and missing observations despite noisy features in the persistence diagram. The JTK_Cycle and the Lomb-Scargle method requires specifying a period search range. The proposed SSE method has the added advantage that its periodicity score has a geometric interpretation (Perea and Harer, 2015).

Table 2.1: Results for the periodic signal summarized as p-values for JTK_Cycle and Lomb-Scargle with estimated period in parentheses, and as periodicity scores for Sliding Windows (SW) and SSE methods.

n	π	M	SW	JTK_Cycle	Lomb-Scargle	SSE Method
50	0.00	2	0.74	0.00 (7.69)	0.00 (8.02)	0.74
	0.10	2	—	—	0.00 (8.03)	0.74
	0.20	2	—	—	0.00 (8.03)	0.70
	0.30	2	—	—	0.00 (8.04)	0.67
	0.40	2	—	—	0.00 (8.04)	0.44
100	0.00	6	0.53	0.00 (8.00)	0.00 (8.02)	0.53
	0.10	6	—	—	0.00 (8.02)	0.53
	0.20	6	—	—	0.00 (8.02)	0.53
	0.30	4	—	—	0.00 (8.20)	0.49
	0.40	3	—	—	0.00 (8.02)	0.43
500	0.00	8	0.93	0.00 (2.64)	0.00 (8.02)	0.93
	0.10	8	—	—	0.00 (8.02)	0.85
	0.20	6	—	—	0.00 (8.02)	0.71
	0.30	2	—	—	0.00 (8.02)	0.63
	0.40	2	—	—	0.00 (8.02)	0.60
1000	0.00	26	0.90	0.00 (1.28)	0.00 (8.02)	0.90
	0.10	15	—	—	0.00 (8.02)	0.74
	0.20	12	—	—	0.00 (8.02)	0.69
	0.30	6	—	—	0.00 (8.02)	0.44
	0.40	4	—	—	0.00 (8.01)	0.43

For the non-periodic signal, all the methods performed reasonably well across all samples and missingness mechanisms. The performance of the SSE method in the non-periodic setting is not surprising. This is because as more observations are missing, the sampled time points appear

Table 2.2: Results for the non-periodic signal are given as p-values for JTK_Cycle and Lomb-Scargle with estimated period in parentheses, and as periodicity scores for Sliding Windows (SW) and SSE methods.

n	π	M	SW	JTK_Cycle	Lomb-Scargle	SSE Method
50	0.00	3	0.30	1.00 (13.29)	0.15 (2.05)	0.30
	0.10	3	—	—	0.16 (2.04)	0.32
	0.20	3	—	—	0.28 (2.04)	0.23
	0.30	3	—	—	0.17 (2.04)	0.25
	0.40	3	—	—	0.41 (32.87)	0.14
100	0.00	3	0.26	1.00 (12.88)	0.10 (1.00)	0.26
	0.10	3	—	—	0.08 (1.00)	0.22
	0.20	3	—	—	0.27 (1.00)	0.25
	0.30	3	—	—	0.55 (16.39)	0.32
	0.40	3	—	—	0.33 (16.39)	0.29
500	0.00	11	0.14	0.28 (0.45)	0.11 (0.15)	0.14
	0.10	9	—	—	0.23 (0.80)	0.16
	0.20	0	—	—	0.13 (0.20)	0.10
	0.30	7	—	—	0.29 (0.79)	0.18
	0.40	3	—	—	0.19 (0.45)	0.28
1000	0.00	3	0.13	1.00 (1.28)	0.87 (0.25)	0.13
	0.10	3	—	—	0.50 (0.40)	0.14
	0.20	3	—	—	0.55 (0.26)	0.15
	0.30	3	—	—	0.22 (0.26)	0.16
	0.40	3	—	—	0.66 (0.26)	0.20

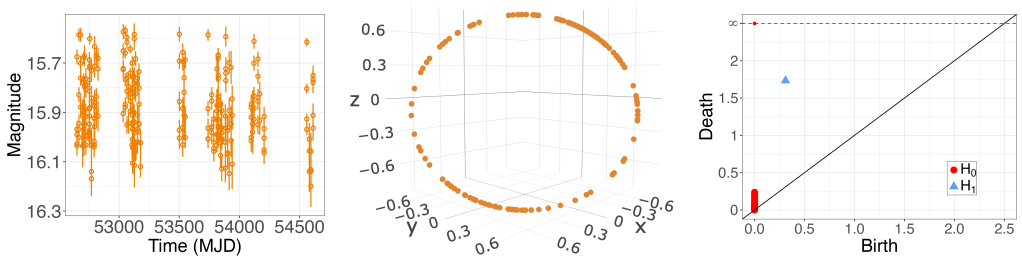
random, and the resulting time series looks more like random noise than signal.

2.5.4 Application to Real Data

Irregularly-spaced times series data are common in astronomy such as those discussed in VanderPlas (2018) and in exoplanet detection methods (Zhao et al., 2020, 2022). In this section, we examine an asteroid dataset from the Lincoln Near-Earth Asteroid Research (LINEAR) survey, which tracks near-Earth asteroids. The data include 280 magnitude measure-

ments (brightness) of LINEAR object ID 11375941 over five and a half years. Magnitude measurements are unitless, and lower values indicate brighter objects. Further details on the data and preprocessing are in Sesar et al. (2011), Palaversa et al. (2013), and VanderPlas (2018).

Figure 2.11a shows the observed magnitude over time, revealing no obvious periodic pattern due to irregular sampling. The TDE method is unsuitable for such data, but the proposed SSE method can construct a geometric representation. Using $M = 2$, rescaling the time points to integers, and taking $\tau = 1$, the SSE in Figure 2.11b reveals a circular geometric object, indicating high periodicity. The H_1 feature on the persistence diagram (Figure 2.11c) is at the point $(b, d) = (0.31, 1.74)$. The periodicity



(a) Time series (b) SSE (c) Persistence diagram

Figure 2.11: LINEAR object ID 11375941. (a) The time series of the measured magnitudes (orange circles) with error bars (vertical bars). (b) The SSE of the time series. (c) The persistence diagram for the SSE with a single highly persistent H_1 feature as expected.

score obtained using Equation (2.48) is 0.82, indicating high periodicity in the observed magnitude of LINEAR object ID 11375941. Using the Lomb-Scargle method, the optimal period was found to be 2.58 with a p-value

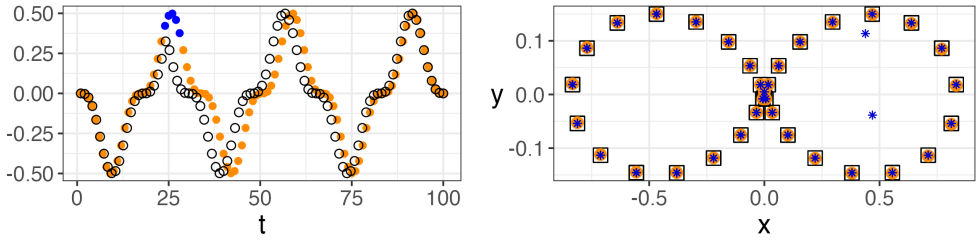
of 0.00. These results confirm the SSE method's periodicity findings and highlight its utility in quantifying and visualizing periodicity.

2.6 Discussion and Concluding Remarks

The fusion of TDE with TDA holds significant promise for discerning system dynamics and quantifying properties like periodicity in uniformly-spaced time series. This work introduces a novel subsequence embedding method for irregularly-spaced time-series data. Irregular spacing can obscure patterns and introduce noise (e.g., Figure 2.11a). While data imputation can create uniformly-spaced series, it often fails to accurately represent the TDE space (e.g., Figure 2.8). The proposed SSE method addresses these challenges, preserving the topology of the reconstructed state space and mitigating spurious homological features introduced by irregular spacing.

One may wonder if there are only a few missing values in a time series, can the missingness simply be ignored? With a investigation, we find that ignoring even a small number of missing values can change the topology of the embedding (as measured with persistence diagrams). The following discusses the results of a brief empirical study on this topic; further analysis is the topic of future investigation. The proposed SSE method seems to be robust to missingness at critical points of the time series, for example at the peaks, or at zero-crossings. To illustrate this, 100

time series observations, as shown in Figure 2.12a were used, which are subsets of the time series in Figure 2.3a. To induce and test the proposed SSE sensitivity to missing values at the peak, five observations, indicated as blue points in Figure 2.12a, were designated as missing values. We

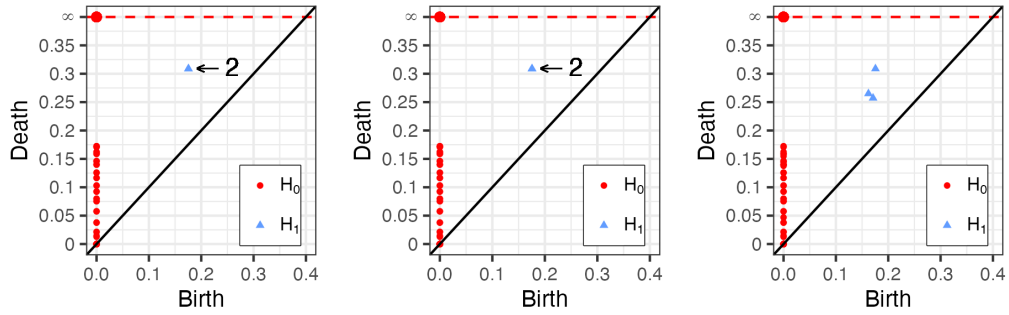


(a) The 100 time series observations (b) The SSE and TDE of the time series

Figure 2.12: The time series and its embedding. (a) The orange points are irregularly-spaced with the blue points denoting missing values. The black hollow circles are the shifted time series observations. (b) The squares denote the TDE of the full time series with no missing values, the orange points denote the SSE of the irregularly-spaced time series, while the blue asterisks. denote the TDE of the shifted time series.

compare the accuracy of the reconstruction from the proposed SSE to an approach that simply ignores the gap and shifts the time series to produce a uniform sequence. The hollow circles in Figure 2.12a denotes this shifted time series. We compare the persistence diagram of the TDE embedding when there are no missing values, the TDE embedding of the shifted time series, and the SSE of the irregularly-spaced time series. Figure 2.13 shows these persistence diagrams, where the persistence diagram of the SSE embedding (Figure 2.13a) is equivalent to the full uniform time series TDE persistence diagram (Figure 2.13b). However, the TDE from the

shifted time series persistence diagram (Figure 2.13c) differs significantly from the full uniformly-spaced time TDE persistence diagram. Similar



(a) Uniformly-spaced (b) Irregularly-spaced (c) Shifted data

Figure 2.13: The persistence diagrams of the TDE of the full uniformly-spaced data (a), the SSE of the irregularly-spaced data (b), and the TDE of the shifted data (c). The "← 2" is used to indicate that there are two H_1 features.

results were observed when valleys, zero-crossings, or any combination of peaks, valleys, and zero crossings were omitted. These findings support the robustness of the proposed SSE method when missing values are observed at critical points of the time series. It also highlights the superior performance of the SSE method compared to the TDE constructed under the assumption that the missingness can be ignored.

Section 2.5.3 demonstrates how TDEs and SSEs can be used to quantify periodicity of a time series. However, we note the need for statistical inference on periodicity scores to determine if the most persistent H_1 feature is due to a real periodic signal or chance. Existing methods for signifi-

cance testing of homology generators, such as those using kernel density estimators, allow constructing confidence sets for homology generators (Fasy et al., 2014; Xu et al., 2019). Extending this to homology generators based on direct filtration on the point-cloud space requires bounding the bottleneck-distance with the Hausdorff distance. Initial investigations produced wide confidence sets, indicating the need for a more tailored method. Finally, Algorithm 1 constructs subsequences with a fixed regularity score r. Extending this to $r \pm \epsilon$ for small ϵ would increase the length of each constructed subsequence, and the number of points in the reconstructed space, while potentially introducing outliers or perturbations in the data space. Chapter 3 introduces a robust statistical inference procedure that allows for the construction of might tight confidence intervals with improved significance detection.

3 MAXTDA: ROBUST STATISTICAL INFERENCE FOR MAXIMAL PERSISTENCE IN TOPOLOGICAL DATA ANALYSIS

The content of this chapter is published in Dakurah and Cisewski-Kehe (2025).

Abstract

Persistent homology is an area within topological data analysis (TDA) that can uncover different dimensional holes (connected components, loops, voids, etc.) in data. The holes are characterized, in part, by how long they persist across different scales. Noisy data can result in many additional holes that are not true topological signal. Various robust TDA techniques have been proposed to reduce the number of noisy holes, however, these robust methods have a tendency to also reduce the topological signal. This work introduces Maximal TDA (MaxTDA), a statistical framework addressing a limitation in TDA wherein robust inference techniques systematically underestimate the persistence of significant homological features. MaxTDA combines kernel density estimation with level-set thresholding via rejection sampling to generate consistent estimators for the maximal persistence features that minimizes bias while maintaining robustness to noise and outliers. We establish the consistency of the sampling procedure and the stability of the maximal persistence estimator. The framework

also enables statistical inference on topological features through rejection bands, constructed from quantiles that bound the estimator's deviation probability. MaxTDA is particularly valuable in applications where precise quantification of statistically significant topological features is essential for revealing underlying structural properties in complex datasets. Numerical simulations across varied datasets, including an example from exoplanet astronomy, highlight the effectiveness of MaxTDA in recovering true topological signals.

3.1 Introduction

In Chapter 2, we introduced a topologically robust method for transforming time series data into a multi-dimensional representation for topological data analysis. We alluded to the fact that assessing the statistical significance of persistence homology features requires bounding the bottleneck-distance with the Hausdorff distance, where our investigations produced wide confidence sets, indicating the need for a more tailored method. Similarly, extending Algorithm 1 to $r \pm \epsilon$ for small ϵ has the potential to introduce outliers or perturbations in the data space. In general, identifying statistically significant features, particularly, the most persistent, or maximal persistent ones is challenging because persistence diagrams lack a canonical vector space structure, meaning operations like addition, averaging, and other conventional statistical techniques are not naturally

defined. This difficulty is further compounded by noisy data. Methods such as kernel smoothing, developed within robust topological analysis (Fasy et al., 2018; Anai et al., 2020), are employed to mitigate noise but also often reduce the lifetimes (i.e., persistences) of the maximal persistent features. The systematic underestimation of the lifetimes of these features is an artifact of the smoothing mechanisms typically employed in these robust methods. To enable statistical inference for maximal persistent features, it is helpful to address these limitations. This inference challenge arises from the need to quantify uncertainty in the presence of perturbations, such as noise, outliers, or density variation in a random sample $\mathbb{X}_n = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ drawn from a probability distribution $\mathbb{P}$ with compact support $\mathbb{X}$ in a space $\mathcal{X} \subset \mathbb{R}^d$. Robust topological tools aim to recover the topology of $\mathbb{X}$ by defining a smoothing function $\phi : \mathcal{X} \to \mathbb{R}$. This function, commonly a kernel density estimate (KDE), kernel distance, or distance-to-a-measure (DTM) function, is parameterized to suppress noise or reweight outliers (Chazal et al., 2011; Fasy et al., 2014, 2018; Anai et al., 2020). A preferred outcome would maintain high persistence for true features while reducing noise features to negligible persistence levels.

The motivation for this work is to develop an inference method that builds on these robust methods, while mitigating the reduction in the persistence of the features, in order to enhance a feature's statistical significance. The proposed framework, "Maximal TDA" (MaxTDA), mitigates this reduction by first estimating a KDE over the sample as an intermediate

representation of the data sampling distribution. Then an upper-level set is defined for a carefully selected density threshold, and rejection sampling is used to draw samples from the thresholded KDE for subsequent statistical inference on the maximal persistent features. This process retains the robustness of the initial smoothing while producing a denser, more consistent sampling surface. Subsequent inference then involves further smoothing or directly computing a persistence diagram directly over this dense sample. This methodology is motivated by two key observations. First, the kernel smoothing enhances robustness against outliers and noise (Bobrowski et al., 2017; Fasy et al., 2018; Anai et al., 2020). Second, the thresholded KDE corrects for density variation in the sampling by providing for a denser and statistically consistent sampling surface (Tsybakov, 1997; Singh et al., 2009), a characteristic that is crucial for maintaining the persistence of the features. This is illustrated in Figure 3.1, where the aim is to recover and maintain the persistence of key features such as the two loops (the red and blue circles) indicated by dense clusters.

The proposed MaxTDA approach presents a robust, consistent, and less biased estimator of the most persistent features in certain homology groups, which are groups that identify different dimensional holes in data. While KDEs have been used for robust persistent homology, we show that the resulting homology estimates do not preserve the strength of the true features. In Theorem 3.2, we show that the proposed sampling technique is consistent, and in Lemma 3.1, we prove the stability of the resulting

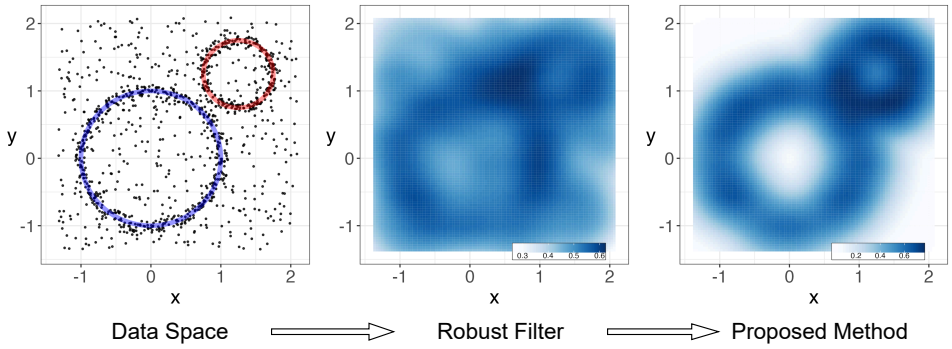


Figure 3.1: Illustration of the MaxTDA framework. For a data space (left), robust TDA methods applies a robust filter(e.g., KDE) to the data (middle). MaxTDA extends this by sampling from a thresholded KDE (right), enhancing robustness to noise and creating a denser sampling surface.

maximal persistence estimator. From a statistical perspective, we establish that MaxTDA produces estimates with reduced bias and enhanced statistical significance. The remainder of this paper is structured as follows: Section 3.2 discusses theoretical results including consistency and bias analyses, and the statistical significance of the maximal persistence estimator. Section 3.3 and 3.4 demonstrate the effectiveness of the MaxTDA through numerical simulations, including one motivated by a statistical challenge in exoplanet astronomy. Section 2.6 closes with implications and potential extensions of our work.

3.2 Maximal TDA Method

This section present the MaxTDA method and corresponding theoretical results. In particular, we show the construction of the smooth subsamples

for inference on the maximal persistence features, prove its homology preserving properties as well as some bias reduction and consistency results.

3.2.1 Overview of methodology

Before presenting the technical details of our method, we first provide an overview of our approach to estimating and performing statistical inference on the maximal persistent features. The key challenge we address is how to reliably estimate the most persistent features from noisy point cloud data while minimizing the persistence reduction (see Section 3.2.2). Traditional approaches often face a trade-off between noise reduction and feature preservation. Our method seeks to resolve this through the construction of "smooth subsamples" using a combination of kernel density and level-set estimation via rejection sampling, where we only accept proposed points where the estimated density exceeds a threshold λ. The details of this construction are discussed in Section 3.2.3, and the validity of such a construction is established in Theorem 3.2. This thresholding naturally filters out likely noise points while preserving the strength of genuine features, as true features tend to manifest in regions of high density.

The remainder of this section develops statistical inference methods for working with features constructed from the smooth subsamples of the thresholded KDE. We analyze the bias reduction properties of the

maximal persistence estimator and develop methods for assessing its statistical significance, providing both theoretical guarantees and practical tools for identifying statistically significant features in noisy data with varying density distributions. Extensions of these inference methods to functions of the maximal persistence are also discussed.

3.2.2 Stability of Maximal Persistence

The support $\mathbb{X}$ is not directly observed but is studied through the point cloud $\mathbb{X}_n$. Let $\mathrm{Dgm}(\mathbb{X}_n)$ be the VR-based filtration persistence diagram on $\mathbb{X}_n$, and $\mathrm{Dgm}(\mathbb{X})$ the true underlying persistence diagram on the support $\mathbb{X}$. For ϕ defined on $\mathcal{X}$, and its empirical estimate ϕ_n, let $\mathrm{Dgm}(\phi(\mathbb{X}))$, and $\mathrm{Dgm}(\phi_n(\mathbb{X}))$ denote their persistence diagrams from upper-level set filtrations of ϕ and ϕ_n, respectively. When an exposition applies to either a VR filtration or the upper-level set filtration of the KDE or DTM, the persistence diagram is generically denoted as $\mathrm{Dgm}(\cdot)$ or simply Dgm. Define the maximum persistence of the features on the persistence as $\mathrm{mp}\,[\mathrm{Dgm}(\cdot)]$. The following lemma presents the stability result for the maximal persistence estimator.

Lemma 3.1 (Maximal Persistence Stability)**.** *Let Δ be the persistence diagram with points only along the diagonal. Let ϕ_n be an empirical KDE or DTM function defined on the sample $\mathbb{X}_n$. Then the following results hold:*

(i) The maximum persistence can be expressed in terms of the bottleneck distance:

$$mp\left[Dgm(\cdot)\right] = 2d_B\left(Dgm(\cdot), \Delta\right). \tag{3.1}$$

(ii) Let $\widehat{\nabla}$ *be defined as* $\widehat{\nabla} = \left|mp[\widehat{Dgm}] - mp[Dgm]\right|$*, then it holds that:*

$$\widehat{\nabla} \leqslant 2d_B\left(\widehat{Dgm}, Dgm\right), \tag{3.2}$$

where $\widehat{Dgm}$ *denotes the empirical persistence diagram estimate of Dgm.*

Proof. The maximum persistence $\mathrm{mp}[\mathrm{Dgm}(\cdot)]$ is defined as: $\mathrm{mp}[\mathrm{Dgm}(\cdot)] = \max_{(\mathrm{b},\mathrm{d})\in\mathrm{Dgm}} |\mathrm{d}-\mathrm{b}|$. Similarly, the bottleneck distance $\mathrm{d_B}(\mathrm{mp}[\mathrm{Dgm}(\cdot)], \Delta)$ is defined as:

$$\mathrm{d_B}(\mathrm{mp}[\mathrm{Dgm}(\cdot)], \Delta) = \inf_{\gamma} \sup_{(\mathrm{b},\mathrm{d})\in\mathrm{Dgm}} \|(\mathrm{b},\mathrm{d}) - \gamma((\mathrm{b},\mathrm{d}))\|_\infty, \tag{3.3}$$

where $\gamma : \mathrm{Dgm} \to \Delta$ defines a bijection between Dgm and Δ. Note that since Δ is the diagonal, the optimal bijection γ is the orthogonal projection of points in $\mathrm{Dgm}(\cdot)$ to Δ, hence $\gamma((\mathrm{b},\mathrm{d})) = \left(\frac{\mathrm{b}+\mathrm{d}}{2}, \frac{\mathrm{b}+2}{2}\right)$. It then follows that:

$$\mathrm{d_B}(\mathrm{mp}[\mathrm{Dgm}(\cdot)], \Delta) = \sup_{(\mathrm{b},\mathrm{d})\in\mathrm{Dgm}(\cdot)} \left\|(\mathrm{b},\mathrm{d}) - \left(\frac{\mathrm{b}+\mathrm{d}}{2}, \frac{\mathrm{b}+\mathrm{d}}{2}\right)\right\|_\infty = \frac{|\mathrm{d}'-\mathrm{b}'|}{2}, \tag{3.4}$$

where $(\mathrm{b}', \mathrm{d}')$ are the birth-death pair with the maximal persistence. The

bound for $\widehat{\nabla}$ can be derived based on the expression:

$$\widehat{\nabla} = 2\left|d_B(\text{mp}[\widehat{\text{Dgm}}], \Delta) - d_B(\text{mp}[\text{Dgm}], \Delta)\right| \leqslant 2d_B(\text{mp}[\widehat{\text{Dgm}}], \text{mp}[\text{Dgm}]), \tag{3.5}$$

where the last inequality follows from the reverse triangle inequality for metrics. □

The main object of interest in this work is the maximal persistence mp [Dgm$(\cdot)$]. We now describe the framework to consistently estimate it while reducing the associated bias inherent in estimating these maximal values.

3.2.3 Smooth sampling surface

The methodology for constructing the smooth sampling surface that maximizes the persistence of features is described next. This approach can be used to either maximize the persistence of a single feature or multiple features, depending on the application. The goal here is to obtain samples that better approximate the true underlying topology, and these samples are subsequently used for inference on the maximal persistence features. Our approach uses kernel density estimation to create a smooth sampling surface, enabling the generation of samples that preserve the underlying topological structure. Specifically, given the observed data $\mathbb{X}_n$ drawn according to the distribution $\mathbb{P}$ with density f, we approximate this

density with the KDE estimate $\widehat{f}_\sigma$. This provides a smooth surface that captures the structure of the manifold while reducing the noise. A dense sample $\mathbb{X}_n^*$ is drawn from the smooth surface using rejection sampling (Devroye, 1986). The $\mathbb{X}_n^*$ serves to preserve the persistence of the homology features relative to $\mathbb{X}_n$. Rejection sampling requires a target distribution and a proposal distribution, where samples are drawn from a proposal distribution because of difficulties sampling from the target distribution. For the purpose of this work, the proposal distribution $\mathbb{Q}$ is a function of the volume enclosing the topological space $\mathcal{X}$. The target distribution is the KDE $\widehat{f}_\sigma$. The objective then is to draw samples $\mathbf{x}^*$ according to $\mathbb{Q}$ and accept them based on the target density $\widehat{f}_\sigma$. In particular, for some $\Gamma \geqslant \sup_{\mathbf{x}^* \in \mathcal{X}} \widehat{f}_\sigma(\mathbf{x}^*)$, the sample $\mathbf{x}^*$ is accepted with probability $\widehat{f}_\sigma(\mathbf{x}^*)/\Gamma$. Algorithm 2 outlines the sampling scheme described here. The resulting

Algorithm 2 Smooth Subsampling

Require: Observed data $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$, density threshold λ, number of generated points B.
Step 1: Fit the KDE $\widehat{f}_\sigma$ to the data sample $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$.
Step 2: For k in the range $1, \cdots, B$, **do Step 3** to **Step 4**.
Step 3: Compute $\mathbf{x}^*$ as follows:
Repeat:
(i) Draw a sample $\mathbf{x}^*$ from the proposal distribution $\mathbb{Q}$.
(ii) Compute the density associated with the sample $\mathbf{x}^*$, i.e., evaluate $\widehat{f}_\sigma(\mathbf{x}^*)$.
(iii) Sample a point $u \sim U(0, \Gamma)$.
Until: $u \leqslant \widehat{f}_\sigma(\mathbf{x}^*)$, and $\widehat{f}_\sigma(\mathbf{x}^*) \geqslant \lambda$.
Step 4: Set $\mathbf{x}_k^* = \mathbf{x}^*$.
Output: Return the samples $\mathbb{X}_B^* = \{\mathbf{x}_1^*, \cdots, \mathbf{x}_B^*\}$.

sample is used to construct a distribution of maximal persistence values by generating random persistence diagrams from the transformed data space. A threshold λ is then selected to maximize the persistence of the targeted prominent features.

We now show that the samples $\mathbb{X}_n^*$ obtained via Algorithm 2 preserves the homology of $\mathbb{X}$. This largely follows from the theory of level-set estimation, especially the work of Cuevas and Fraiman (1997). An interesting observation made in Bobrowski et al. (2017) is that recovering the homology of $\mathbb{X}$ does not rely on the consistency of the KDE $\widehat{f}_\sigma$. Also, to avoid making assumptions on the shape of the space $\mathbb{X}$, the Hausdorff metric is used to measure the closeness of the approximation.

Theorem 3.2 (Convergence of Smooth Subsamples)**.** *Let $\mathbb{P}$ be compactly supported on the set $\mathbb{X}$, having bounded density f and $f > \lambda$ for some positive constant λ. Assume the kernel function K_σ is a decreasing function of $\mathbf{x}$ such that as $||\mathbf{x}|| \to \infty$, we have $||\mathbf{x}||^{d+1} K_\sigma(||\mathbf{x}||) \to 0$. Further assume that K_σ is a bounded density such that for some r_1, r_2,*

$$K_\sigma(||\mathbf{x}||) \geqslant r_1 \mathbb{1}(\mathbf{x} \in B(0, r_2)). \tag{3.6}$$

Let β_n be of order $o(n/\log n)^{1/d}$, then $\beta_n d_H(\mathbb{X}_n^*, \mathbb{X}) \to 0$ a.s., where $\beta_n \sigma$ goes to zero with n large.

In our analysis, we used a Gaussian kernel, which is not compactly supported. Hence we make the additional assumption that the bandwidth

$\sigma \to 0$ as $n \to \infty$ in such a way that $\beta_n^{d+1}\sigma \to 0$, then the proof follows directly from Theorem 3 in Cuevas and Fraiman (1997). By constructing a smooth subsample, we intrinsically reduce the magnitude of any topological error in subsequent persistent homology computations on these smooth subsamples. For example, in the initial sample $\mathbb{X}_n$, the randomness in the sample could introduce points that results in additional features. The kernel smoothing initially reduces the presence of such outlying points, and an appropriate choice of λ (which depends on the specific application) enhances the originally significant homological features. *In summary, unless the randomness introduces features that dominate the most persistent real feature, MaxTDA guarantees a smooth recovery of this original dominant feature*. If randomness in the sample introduces more persistent features than the most persistent real feature, it is not generally feasible to recover the real feature (Fasy et al., 2014; Bobrowski et al., 2017). We demonstrate this concept in our numerical studies in Section 3.3.1 and 3.3.2. The next section discusses how to select the optimal smoothing bandwidth and the level-set threshold parameters.

3.2.3.1 Parameter selection

The choice of KDE bandwidth σ and the level-set threshold λ are essential to $\mathbb{X}_n^*$ recovering the topology of $\mathbb{X}$. These values are not known in practice, hence we provide a data-dependent estimation process for selecting these parameters. For a given homology dimension, let $\ell_i(\lambda, \sigma)$ be the lifetime

(i.e., persistence) of the i-th feature on the persistence diagram $\widehat{\mathrm{Dgm}}$ associated with $\mathbf{X}_n^*$. Consider the ordered lifetimes $\ell_1(\lambda,\sigma) \geqslant \ell_2(\lambda,\sigma) \geqslant \cdots \geqslant \ell_T(\lambda,\sigma)$, where T is the number of features of interest. The cumulative persistence of the top T features is given by: $\mathrm{CP}_T(\lambda,\sigma) = \sum_{i=1}^{T} \ell_i(\lambda,\sigma)$. The goal is to choose the parameter λ and σ that maximizes $\mathrm{CP}_T(\lambda,\sigma)$. The parameters (λ^*,σ^*) are determined by solving the optimization problem: $(\lambda^*,\sigma^*) = \arg\max_{(\lambda,\sigma)\in\Omega} \mathrm{CP}_T(\lambda,\sigma)$, where Ω denotes the feasible parameter space for λ and σ. Note that this process can be augmented to emphasize certain features by assigning weights $\{\omega_1,\cdots,\omega_T\}$ to the lifetimes. The number of features T can be chosen based on the expected topology of $\mathbb{X}$, or by adaptively by analyzing the decay of the ordered lifetimes $\ell_i(\lambda,\sigma)$. A sharp drop in $\ell_i(\lambda,\sigma)$ beyond a certain index indicates a natural cutoff for T. For the bandwidth σ, we found that the average k-NN distance (for k between 1 and 5) between points in $\mathbb{X}_n$ provides a good parameter search space.

3.2.4 Bias reduction

Existing methods for estimating a persistence diagram in the presence of noise or sampling variability can identify the maximal persistent features. This often involves smoothing out low persistence features, which consequently reduces the lifetime of the most persistent H_1 features. This results in a bias in estimating the lifetime of the maximal persistent features. In this section, we discuss this phenomenon and provide results that shows

the proposed MaxTDA method helps reduce this bias for an appropriate choice of thresholding parameter λ and for a range of bandwidths σ.

3.2.4.1 Source of bias in maximal persistence

Robust persistent homology methods, such as smoothing, subsampling, filtering, or thresholding, implicitly bias the persistence estimates by reducing the lifetimes of the features. The following example illustrates this bias. Let $\mathcal{P}$ be a class of probability distributions satisfying Assumption 1.1. Further assume that there exists positive constants c and c' such that for data $\mathbf{x} \in \mathbb{X}$ and $d' < d$:

$$\mathrm{vol}_d\left(B(\mathbf{x}, r) \cap \mathbb{X}\right) \geqslant c\left(1 - \frac{r^2}{4\kappa^2}\right)^{d'/2} r^{d'} \geqslant c' r^{d'}, \tag{3.7}$$

where $\mathrm{vol}_d(\cdot)$ denotes the volume of a d-dimensional ball, and κ is as defined in Assumption 1.1. This is the usual regularity assumption that removes certain pathological manifolds, such as those with sharp peaks or cusps. In practical terms, for every point $\mathbf{x} \in \mathbb{X}$, if you take a ball of radius r around $\mathbf{x}$, the portion of the ball lying in $\mathbb{X}$ has a volume that scales with $r^{d'}$, that is, $\mathbb{X}$ is "thick enough" in every small neighborhood such that there are no parts that are infinitesimally thin or sharply peaked. Let $\mathbb{X}_n$ be drawn according to a distribution $\mathbb{P} \in \mathcal{P}$ which is supported on $\mathbb{X}$. Let $\widehat{\mathrm{Dgm}}$ and Dgm be the persistence diagrams associated with $\mathbb{X}_n$ and

$\mathbb{X}$, respectively. Then the following inequality holds:

$$\mathrm{mp}[\widehat{\mathrm{Dgm}}] \leqslant \mathrm{d_B}(\widehat{\mathrm{Dgm}}, \mathrm{Dgm}) + \mathrm{mp}[\mathrm{Dgm}], \tag{3.8}$$

which follows from applying the triangle inequality to $\mathrm{mp}[\widehat{\mathrm{Dgm}}] = \mathrm{d_B}(\widehat{\mathrm{Dgm}}, \nabla)$. This implies the bias: $\mathrm{E}(\mathrm{mp}[\widehat{\mathrm{Dgm}}]) - \mathrm{mp}[\mathrm{Dgm}]$ is directly upper bounded by the expected bottleneck distance between $\widehat{\mathrm{Dgm}}$ based on $\mathbb{X}_n$ and Dgm based on $\mathbb{X}$. Therefore, a "good" representation of $\mathbb{X}$ can lead to a lower bias in estimating $\mathrm{mp}[\widehat{\mathrm{Dgm}}]$. Next, we discuss how the proposed framework provides a good representation of $\mathbb{X}$ with a thresholded KDE.

3.2.4.2 Role of sampling and thresholding

Consider the setup where two samples, $\mathbb{X}^*_{n,0}$ and $\mathbb{X}^*_{n,\lambda}$, are drawn using Algorithm 2 with threshold values of 0 and λ, respectively. The choice of λ and n influence the bias in estimating the maximal persistence. We consider the case of the VR filtration, but the analysis also applies to filtrations of $\phi(\cdot)$. From Equation (3.8), the maximal persistence associated with $\mathbb{X}^*_{n,\lambda}$ is given as follows:

$$\zeta(n,\lambda)\mathrm{mp}[\widehat{\mathrm{Dgm}}(\mathbb{X}^*_{n,\lambda})] = \mathrm{d_B}(\widehat{\mathrm{Dgm}}(\mathbb{X}^*_{n,\lambda}), \mathrm{Dgm}(\mathbb{X})) + \mathrm{mp}[\mathrm{Dgm}(\mathbb{X})], \tag{3.9}$$

where $\zeta(n,\lambda)$ is an unspecified sequence depending on n and λ, and goes to 1 for n large. In the limit as $\lambda \to 0$, we have by construction

that $\zeta(n, \lambda) = \zeta(n, 0)$. Hence the bias of the smoothed and unsmoothed estimators are the same as $n \to \infty$, and $\lambda \to 0$:

$$\lim_{n\to\infty,\lambda\to 0} \mathsf{E}\left(d_B(\widehat{\mathrm{Dgm}}(\mathbb{X}^*_{n,\lambda}), \mathrm{Dgm}(\mathbb{X}))\right) = \lim_{n\to\infty} \mathsf{E}\left(d_B(\widehat{\mathrm{Dgm}}(\mathbb{X}^*_{n,0}), \mathrm{Dgm}(\mathbb{X})\right). \tag{3.10}$$

Under finite sampling, the benefits of the thresholding lie in the difference in the rates of convergence of both $\mathbb{X}^*_{n,\lambda}$ and $\mathbb{X}^*_{n,0}$ to $\mathbb{X}$. For example, consider $\beta_n \to \infty$ from Theorem 3.2, Cuevas and Fraiman (1997) show any rate of order $(n/\log n)^{1/d} = O(\beta_n)$ cannot be achieved by $\mathbb{X}^*_{n,0}$ That is, consider a convergent rate that is faster than β_n for $\mathbb{X}^*_{n,\lambda}$ to $\mathbb{X}$, say $\beta^*_n \geqslant (n/\log n)^{1/d}$, then β^*_n cannot be achieved when estimating $\mathbb{X}$ with $\mathbb{X}^*_{n,0}$ (Cuevas and Fraiman, 1997). Hence, for an appropriate choice of λ, we conjecture that:

$$\mathsf{E}\left(d_B(\widehat{\mathrm{Dgm}}(\mathbb{X}^*_{n,\lambda}), \mathrm{Dgm}(\mathbb{X}))\right) \leqslant \mathsf{E}\left(d_B(\widehat{\mathrm{Dgm}}(\mathbb{X}^*_{n,0}), \mathrm{Dgm}(\mathbb{X}))\right). \tag{3.11}$$

While this inequality has been observed empirically (see Figure 3.4a), a formal theoretical proof remains an open challenge. The primary difficulty in establishing such a result lies in deriving an explicit form for $\zeta(n, \lambda)$, which would require strong assumptions on the geometric properties of the support $\mathbb{X}$ to obtain a closed-form expression.

3.2.5 Statistical significance of the maximal persistence

The statistical significance of the maximal persistence estimator $\mathrm{mp}[\widehat{\mathrm{Dgm}}]$ is determined through a lower bound for $\mathrm{mp}[\widehat{\mathrm{Dgm}}]$. This is equivalent to bounding the difference $\widehat{\nabla} = \left|\mathrm{mp}[\widehat{\mathrm{Dgm}}] - \mathrm{mp}[\mathrm{Dgm}]\right|$. A method for constructing confidence sets for persistence diagrams by bootstrapping the bottleneck distance was proposed in Fasy et al. (2018). The construction of the lower bound for $\widehat{\nabla}$ follows the same framework. We first state the following consistency result for $\widehat{\nabla}$ based on the upper-level set filtration of the density function, and similar consistency results holds for other functions such as the DTM and other kernel distances.

Theorem 3.3 (Consistency)**.** *Let ϕ be a density function defined on $\mathcal{X}$, and let ϕ_n be its empirical estimate according to Equation* (1.8) *based on the sample $\mathbb{X}_n^*$ from Algorithm 2. Let $\{c_1, \cdots, c_k\}$ and $\{c_1^n, \cdots, c_k^n\}$ be the critical points of ϕ and ϕ_n, respectively. Assume that the critical points of ϕ and ϕ_n are close enough such that the maximal difference at these critical points is bounded as: $\max_i |\phi_n(c_i^n) - \phi(c_i)| \leqslant \frac{1}{2}\min_{i\neq j}|\phi(c_i) - \phi(c_j)| - \|\phi_n - \phi\|_\infty$, and $2\|\phi_n - \phi\|_\infty \leqslant \frac{1}{2}\min_{i\neq j}|\phi(c_i) - \phi(c_j)|$. Then $mp[\widehat{Dgm}(\phi_n)]$ is a consistent estimator of mp $[Dgm(\phi)]$:*

$$\widehat{\nabla} = \left|mp[\widehat{Dgm}(\phi_n)] - mp[Dgm(\phi)]\right| \xrightarrow{P} 0, \quad as\ n \to \infty. \tag{3.12}$$

Proof. The proof follows from the regular consistency results on kernel density estimation and the critical distances lemma in Devroye and Lugosi

(2001) and Fasy et al. (2018). By the bottleneck stability theory, we have that $d_B(\mathrm{Dgm}(\phi_n), \mathrm{Dgm}(\phi)) \leqslant \|\phi_n - \phi\|_\infty$. Note that for the upper-level sets filtration of these functions, the homology only changes at the critical points. Assume that $(\phi(c_i), \phi(c_j)) \in \mathrm{Dgm}(\phi)$ and $(\phi(c_i^n), \phi(c_j^n)) \in \mathrm{Dgm}(\phi_n)$. Let $\gamma : \mathrm{Dgm}(\phi_n) \to \mathrm{Dgm}(\phi)$ be the optimal bottleneck matching between the two diagrams. Under the assumption that $\max_i |\phi_n(c_i^n) - \phi(c_i)| \leqslant \frac{1}{2}\min_{i\neq j} |\phi(c_i) - \phi(c_j)| - \|\phi_n - \phi\|_\infty$, which implies $\min_{i\neq j} |\phi(c_i) - \phi(c_j)| - \max_i |\phi_n(c_i^n) - \phi(c_i)| \geqslant \max_i |\phi_n(c_i^n) - \phi(c_i)| + 2\|\phi_n - \phi\|_\infty$, it follows that $\gamma(\phi(c_i^n), \phi(c_j^n)) = (\phi(c_i), \phi(c_j))$. By Lemma 3.1, we have that $\widehat{\nabla} \leqslant \max_i |\phi_n(c_i^n) - \phi(c_i)|$. Define $\phi_n = \widehat{f}_{\sigma_2}$, and let $\tilde{f}_{\sigma_1}$ be the KDE on $\mathbb{X}_n$ and $L_\lambda = \{\mathbf{x} : \tilde{f}_{\sigma_1}(\mathbf{x}) > \lambda\}$. Observe that $\widehat{f}_{\sigma_2} = \left(f_{\mathbb{X}_n^*} * K_{\sigma_2}\right)$ where $(\cdot * \cdot)$ denotes the convolution operation, and $f_{\mathbb{X}_n^*}(\mathbf{x}) \propto \frac{\tilde{f}_{\sigma_1}(\mathbf{x})\mathbb{1}(\mathbf{x}\in L_\lambda)}{\int_{L_\lambda} \tilde{f}_{\sigma_1}(\mathbf{y})\,d\mathbf{y}}$. The conclusion follows from the regular consistency assumption on the bandwidths $\sigma_1, \sigma_2 \to 0$ and sample size $n \to \infty$ of the KDE (Devroye and Lugosi, 2001). □

Next, we describe the framework for assessing the statistical significance of the maximal persistence features via a Monte-Carlo procedure.

3.2.6 Construction of confidence sets

It is common to consider homology features with longer persistence as topological signal (Fasy et al., 2014). Thus $H_{k>0}$ features with longer life spans can be interpreted as being more statistically significant than those

with shorter life spans. For example, in time series analysis, one method for determining periodicity examines the persistence of loops in a TDE space. A perfectly circular loop suggests an underlying periodic signal, and Perea et al. (2015) proposes estimating its period using a function of mp[Dgm], though a method for quantifying the statistical significance of this estimate was not established. The proposed MaxTDA framework addresses this gap by providing tools to determine the statistical significance of such a periodicity measure.

Methods for estimating the statistical significance of the features through confidence sets were discussed in Fasy et al. (2014). However, these methods bound the bottleneck distance with the Hausdorff distance or distances of functions defined on the data space, which shifts the randomness in the construction to the original data space. These bounds are not tight in many cases (e.g., Fasy et al. 2018; Glenn et al. 2024). A method that restricts the randomness to the persistence diagrams, by directly bootstrapping the bottleneck distance was introduced in Fasy et al. (2018). We first describe the process for constructing confidence sets for the features with maximal persistence on the persistence diagram, which in our case amount to lower bounds for the maximal persistence.

Given significance level $\alpha \in (0,1)$, the goal is to find t_α such that: $\Pr(d_B(\widehat{Dgm}, Dgm) > t_\alpha) \leqslant \alpha$ as $n \to \infty$. The confidence set on a persistence diagram can be constructed by considering points in $\widehat{Dgm}$ whose distance to the diagonal exceeds t_α, $\left\{(b,d) \in \widehat{Dgm} : |d-b| > 2t_\alpha\right\}$. This

construction extends to the maximal persistence estimator through the relation:

$$\Pr(\widehat{\nabla} > 2t_\alpha) \leqslant \Pr(d_B(\widehat{Dgm}, Dgm) > t_\alpha) \leqslant \alpha. \tag{3.13}$$

There are two ways to visualize this confidence set on $\widehat{Dgm} \subset \mathbb{R}^2$. The first is to draw d_B-balls with side length of $2t_\alpha$ centered on each point in $\widehat{Dgm}$. Then using the closeness to the diagonal, a point is considered to be be a topological noise if its d_B-ball intersects with the diagonal line. The second and equivalent option is to add a diagonal band (rejection band) of width $\sqrt{2}t_\alpha$ to $\widehat{Dgm}$, and points in $\widehat{Dgm}$ that falls within this band are elements of $\left\{(b, d) \in \widehat{Dgm} : |d - b| \leqslant 2t_\alpha\right\}$, a rejection set, and are deemed to not be statistically significant at the α significance level. Note that the rejection band is constructed individually for each homology dimension. The t_α can be estimated via a Monte-Carlo procedure described in the next section.

3.2.7 Monte-Carlo estimation procedure

In this section, a Monte-Carlo procedure is proposed to estimate the t_α. Draw the sample $\mathbb{X}_n^{*(b)}$ using Algorithm 2 as follows: first, take a bootstrap sample $\mathbb{X}_n^{(b)}$ from the original sample $\mathbb{X}_n$. Using Algorithm 2, generate $\mathbb{X}_n^{*(b)}$ with $\mathbb{X}_n^{(b)}$ as the underlying observed sample. Let $\phi_n^{(b)}$ be the function associated with the sample $\mathbb{X}_n^{*(b)}$, and ϕ_n to be the function associated with the quantity $\mathbb{X}_n^*$, obtained by applying Algorithm 2 to $\mathbb{X}_n$. Compute the empirical quantity $\hat{t}^{(b)} = d_B(\widehat{Dgm}(\phi_n^{(b)}), \widehat{Dgm}(\phi_n))$. This process is

repeated $b = 1, \cdots, N$ times. Let $\widehat{\Theta}_n$ be the empirical distribution function of this set of observations: $\{\hat{t}^{(b)} : b = 1, ..., N\}$. Let $\hat{t}_\alpha$ be the $1 - \alpha$ quantile of $\widehat{\Theta}_n$. Also let Θ_n be the distribution function of the quantity $d_B(\widehat{Dgm}(\phi_n), Dgm(\phi))$. Then the following result holds.

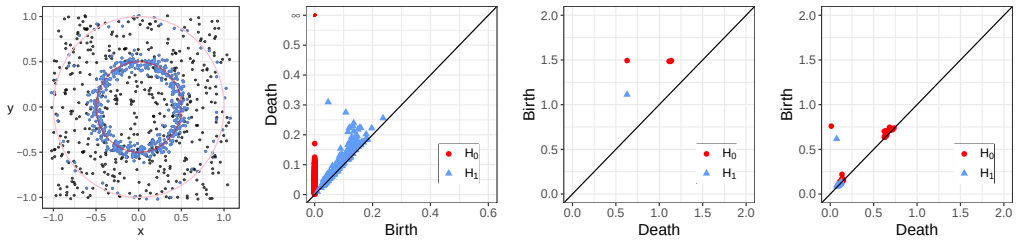
Lemma 3.4. *The* $1 - \alpha$ *quantile* $\hat{t}_\alpha$ *is a consistent estimator of* t_α*, that is,* $\sup_t |\widehat{\Theta}_n(t) - \Theta_n(t)| \xrightarrow{p} 0$.

This result follows directly from Theorem 19 and Corollary 20 in Fasy et al. (2018). This process is used to determine the statistical significance of the maximally persistent $H_{k>0}$ features.

Remark 3.5. *The inference procedure developed in this work can be extended to additive or multiplicative transformations of* $\widehat{\nabla}$*. For example, the statistical significance of the normalized periodicity score* $mp[Dgm]/\sqrt{3}$ *can be derived through the distribution of* $\widehat{\nabla}/\sqrt{3}$*, which amounts to estimating the empirical quantile function* $\sqrt{3}\hat{t}_\alpha$*. These results can also be adapted for minimal persistence.*

3.3 Experimental Validations

This section presents numerical studies that demonstrate the performance of MaxTDA. First we show the quality of the topological recovery achieved by the proposed method in terms of the number of features recovered and the persistence of these features.



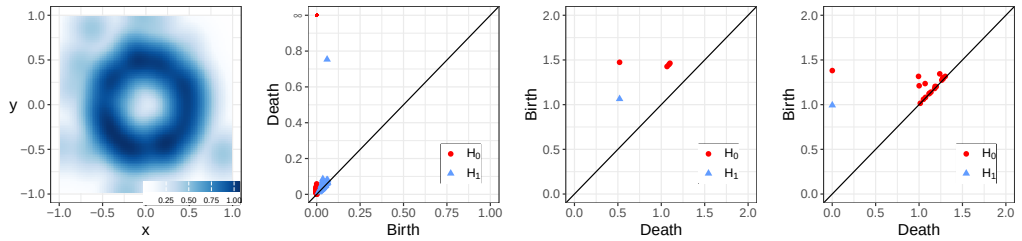
(a) Point cloud $\mathbb{X}_n$ (b) VR diagram (c) DTM diagram (d) KDE diagram

Figure 3.2: An illustration of the VR (b), DTM (c), and KDE (d) filtration on the point cloud $\mathbb{X}_n$ (a) (the blue points are signal and the black points are noise). All three methods identified one dominant H_1 feature in terms of persistence.

3.3.1 Quality of topological recovery

The first numerical experiment aims to recover a densely sampled circle while treating a sparse circle as noise. The data consist of 50 samples around a unit circle, 500 samples around a radius-0.5 circle (both perturbed by $N(0, \sqrt{0.05})$), and 450 uniform samples in $[-1, 1]^2$. These three samples give $\mathbb{X}_n$ with $n = 1000$. The VR, DTM, and KDE persistence diagrams were computed, with DTM parameter $m = 0.9$ chosen over a grid of points in the interval $(0, 1)$, and the KDE bandwidth set at 0.1. A complete comparison across various bandwidths is given in Section 3.3.1.1. The point cloud $\mathbb{X}_n$ and persistence diagrams are displayed in Figure 3.2. While the VR diagram is noisy, the DTM and KDE diagrams suppress low-persistence features, though at the cost of reduced persistence.

To demonstrate the topological recovery, $\mathbb{X}_n^*$ was constructed using Algorithm 2 with a threshold λ selected from the range $[0.1, 1]$ and a KDE



(a) Sample space (b) VR diagram (c) DTM diagram (d) KDE diagram

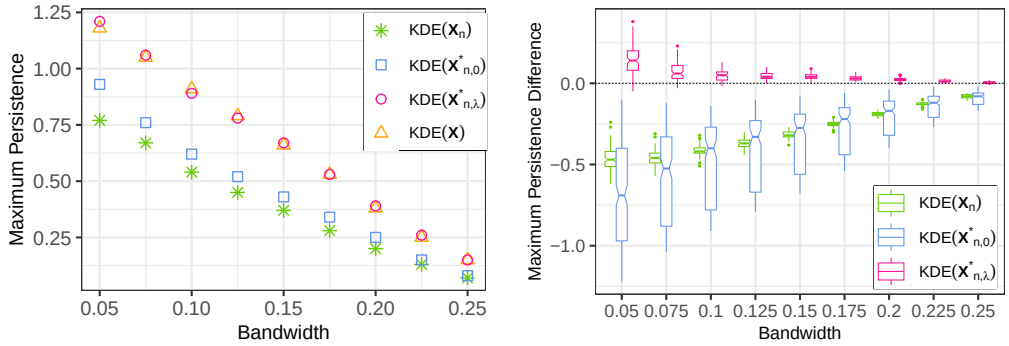
Figure 3.3: An illustration of the VR (b), DTM (c), and KDE (d) filtration on the point cloud $\mathbb{X}_n^*$ from Algorithm 2. All three methods identified one dominant and enhanced H_1 feature.

bandwidth set to the average k-NN distance ($k \in [1, 10]$) of points in $\mathbb{X}_n$. These parameters were chosen using the procedure in Section 3.2.3.1 to maximize the most persistent H_1 feature for each filtration scheme. Specifically, the optimal (λ, k) are $(0.7, 10)$, $(0.4, 2)$, and $(0.6, 8)$ for VR, DTM, and KDE filtration, respectively. The KDE sample space is shown in Figure 3.3a, along with the VR (3.3b), DTM (3.3c), and KDE (3.3d) persistence diagrams computed for $\mathbb{X}_n^*$. All three methods revealed one dominant H_1 feature. The diagrams from $\mathbb{X}_n^*$ contains fewer low-persistence H_1 features than those from $\mathbb{X}_n$.

3.3.1.1 Reducing persistence loss in prominent features

Next, we demonstrate that Algorithm 2 is robust to variations in the KDE bandwidth (and in the DTM smoothing parameter), that is, once appropriate parameters are selected for Algorithm 2, these choices remain effective across different values of σ or m. Let $\mathbb{X}_n$ be the noisy sample in Figure 3.2a,

and $\mathbb{X}^*_{n,\lambda}$ and $\mathbb{X}^*_{n,0}$ be the thresholded and non-thresholded versions, respectively. Let $\mathbb{X}$ be the noise-free data, depicted as the blue points in Figure 3.2a. Figure 3.4a shows that for a single sample, it is possible to appropriately choose the parameters of Algorithm 2 (in this case, $\lambda = 0.6$ and the KDE bandwidth is the average 8-NN distance) such that the maximal persistence associated with $\mathbb{X}^*_{n,\lambda}$ closely approximates the ground truth ($\mathbb{X}$) maximal persistence.



(a) Maximum persistence for a single sample.

(b) Distribution of the maximum persistence.

Figure 3.4: MaxTDA estimation results. (a) For an appropriately chosen threshold, the maximal persistence associated with the MaxTDA $\mathbb{X}^*_{n,\lambda}$ (red circles) closely approximates the ground truth ($\mathbb{X}$) maximal persistence (orange triangles). (b) The distribution of the difference in maximal persistence between the three data samples and the ground truth across 100 independent trials, demonstrating that $\mathbb{X}^*_{n,\lambda}$ (red) maximal persistence is less biased.

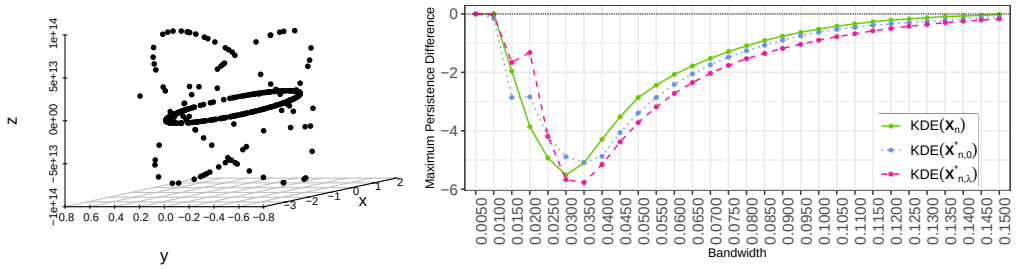
The process is repeated 100 independent times to assess the variability of the construction. The results are presented in Figure 3.4b as boxplots of the differences between the true and estimated maximal persistence, which

indicate that the distributions of the MaxTDA $\mathbb{X}^*_{n,\lambda}$ maximal persistence values are closer to the true values than those from other data spaces.

3.3.2 Data with varying sampling distributions

This section demonstrates how MaxTDA applies to data from topological spaces with similar geometries but different sampling distributions, a scenario that arises, for example, in signal processing when signals at different frequencies are embedded in the same space. Figure 3.5a shows a 3D point cloud $\mathbb{X}_n$ with four ellipses of varying density; the goal is to recover the denser ellipse as the ground truth by isolating a single maximally persistent H_1 feature. Using the parameter selection procedure in Section 3.2.3.1, a density threshold of $\lambda = 12.22$ was obtained, and a KDE bandwidth was determined as the average 1-NN distance of points in $\mathbb{X}_n$. $\mathbb{X}^*_{n,\lambda}$ and $\mathbb{X}^*_{n,0}$ were constructed using Algorithm 2. Figure 3.5b shows that for bandwidths up to 0.025, the maximal persistence of $\text{KDE}(\mathbb{X}^*_{n,\lambda})$ exceeds that of $\text{KDE}(\mathbb{X}^*_{n,0})$ and $\text{KDE}(\mathbb{X}_n)$. These bandwidths correspond to when the dense ellipse's persistence is at its maximum and increasing, whereas for larger bandwidths the persistence decreases due to over-smoothing, indicating undesirable bandwidths.

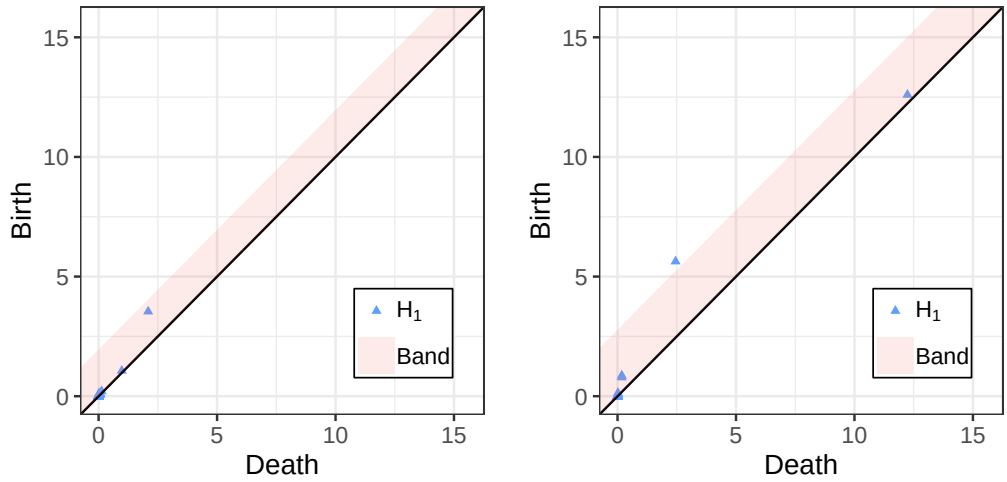
The optimal bandwidths that maximize the maximal persistence were determined to be 0.02 for $\text{KDE}(\mathbb{X}^*_{n,\lambda})$ (denoted hereafter as $\text{KDE}(\mathbb{X}^*_n)$) and 0.015 for $\text{KDE}(\mathbb{X}_n)$, and these values were used to construct the final persistence diagrams, where the most persistent H_1 features in $\mathbb{X}_n$ and $\mathbb{X}^*_n$ had



(a) Data $\mathbb{X}_n$ (b) Differences in maximal persistence

Figure 3.5: Performance of MaxTDA in estimating the maximal persistence using the sample $\mathbb{X}^*_{n,\lambda}$ compared to the original data $\mathbb{X}_n$ and the non-thresholded sample $\mathbb{X}^*_{n,0}$. (a) Data $\mathbb{X}_n$ with $n = 333$; (b) the difference in the maximal persistence from that of the dense ellipse by KDE bandwidth.

persistences of 1.45 and 3.18, respectively. To construct rejection bands, let $\widehat{\mathrm{Dgm}}(\mathrm{KDE}(\mathbb{X}_n))$ and $\widehat{\mathrm{Dgm}}(\mathrm{KDE}(\mathbb{X}^*_n))$ denote the respective persistence diagrams of the upper-level KDE filtrations of $\mathbb{X}_n$ and $\mathbb{X}^*_n$. We bootstrapped $\mathbb{X}_n$ 1000 times and, for each bootstrap sample $\mathbb{X}^{(b)}_n$, estimated a KDE with $\sigma = 0.015$ and computed the bottleneck distance $\hat{t}^{(b)}_{0.015}$ between the H_1 features of $\widehat{\mathrm{Dgm}}(\mathrm{KDE}(\mathbb{X}_n))$ and $\widehat{\mathrm{Dgm}}(\mathrm{KDE}(\mathbb{X}^{(b)}_n))$. We also computed the sample $\mathbb{X}^{*(b)}_n$ using Algorithm 2 at $\lambda = 12.22$, with $\mathbb{X}^{(b)}_n$ as the underlying input data, estimated a KDE with $\sigma = 0.02$ for $\mathbb{X}^{*(b)}_n$, and computed the bottleneck distance $\hat{t}^{(b)}_{0.02}$ between the H_1 features of $\widehat{\mathrm{Dgm}}(\mathrm{KDE}(\mathbb{X}^*_n))$ and $\widehat{\mathrm{Dgm}}(\mathrm{KDE}(\mathbb{X}^{*(b)}_n))$. The 0.95 quantile of $\{\hat{t}^{(b)}_{0.015}\}$ was 1.3115 for diagrams from $\mathbb{X}^{(b)}_n$, and that of $\{\hat{t}^{(b)}_{0.02}\}$ was 1.4195 for diagrams from $\mathbb{X}^{*(b)}_n$, which were used to construct the rejection bands. Figure 3.6a shows the persistence diagram and 95% rejection band for the ordinary KDE, where no H_1 feature is statistically significant, while Figure 3.6b shows the persistence



(a) Persistence diagram for $\mathbb{X}_n^{(b)}$ (b) Persistence diagram for $\mathbb{X}_n^{*(b)}$

Figure 3.6: Illustration of the statistical significance of the H_1 features based on 1000 bootstrap samples from $\mathbb{X}_n^{(b)}$ (a) and $\mathbb{X}_n^{*(b)}$ (b). The displayed bands (light pink) indicated the 95% rejection region for the H_1 features (blue triangles). Note that the H_0 features have been omitted.

diagram and rejection band for the KDE of the smooth subsamples, in which one statistically significant H_1 feature corresponding to the denser elliptical sample is observed. This is partly due to its enhanced persistence, further highlighting the performance of the proposed MaxTDA method.

3.4 Exoplanet Data Application

This section explores how MaxTDA enhances periodic time series analysis by linking the persistence of H_1 features to signal periodicity. Enhancing the lifetime of H_1 features can strengthen a periodicity analysis. We begin by describing a method for constructing a time series representation.

3.4.1 Time-delay embedding

Time-delay embeddings (TDEs) provide a framework for transforming time series into a multi-dimensional representation (Takens, 2006). For time series $\{x(t) : 0 \leqslant t \leqslant n\}$, an embedding matrix is constructed where each row is given by: $\mathbf{v}(t) = [x(t), x(t+\tau), \ldots, x(t+M\tau)]$, with time delay τ and $M+1$ delayed coordinates. Takens' Theorem guarantees that, under suitable conditions, this embedding preserves the shape of the underlying state space if the embedding dimension is sufficiently large (Takens, 2006). One method for determining τ is the average mutual information (AMI) (Fraser and Swinney, 1986). The AMI is computed by partitioning the range of the time series into bins: $\mathcal{I}(\tau) = \sum_{i,j} p_{i,j}(\tau) \log\left(\frac{p_{i,j}(\tau)}{p_i p_j}\right)$, where p_i is the the probability the time series has a value in the i-th bin, and p_j is the probability that $x(t+\tau)$ is in bin j, and $p_{i,j}(\tau)$ denotes the probability that $x(t)$ and $x(t+\tau)$ are in the i-th and j-th bin, respectively. The smallest value of τ where $\mathcal{I}(\tau)$ reaches a local minimum is chosen as the optimal time delay step. This corresponds to the lag at which the redundancy of information between $x(t)$ and $x(t+\tau)$ is minimized, ensuring that points in the reconstructed embedding space are sufficiently independent. Once τ is determined, the embedding dimension $M+1$ is selected using Cao's method (Cao, 1997), which evaluates how the structure of the reconstructed space changes as the embedding dimension increases. It identifies the dimension at which the reconstructed space stabilizes. TDEs remain valid under smooth linear transformations, such as principal component

analysis (PCA), motivating our subsequent use of PCA for dimensionality reduction (Sauer et al., 1991).

3.4.2 Exoplanet time series data

Exoplanets are planets that orbit stars other than our sun. One method for detecting exoplanets is the radial velocity (RV) method, which measures the forward and backward motion of a possible host star over time. This method was used to discover the first exoplanet orbiting a sun-like star (Mayor and Queloz, 1995). With this RV approach, a certain periodic signature in a star's RV over time suggests the presence of an orbiting exoplanet. The red line in Figure 3.7 displays a simulated exoplanet RV signal on a circular orbit. Detecting low-mass exoplanets, such as Earth-like planets, remains challenging as their smaller signals can be obscured by stellar activity like star spots (Huélamo et al., 2008; Dumusque, 2016; Davis et al., 2017). The green line in Figure 3.7 shows how a simulated star spot using the Spot Oscillation And Planet (SOAP) 2.0 code (Dumusque et al., 2014) can induce a periodic RV signal that resembles an exoplanet.

While statistical techniques have been developed to detect exoplanets in the presence of stellar variability (e.g., Rajpaul et al. 2015; Dumusque 2018; Holzer et al. 2021a,b; Jones et al. 2022), they do not fully mitigate the challenges (Zhao et al., 2022). This study demonstrates how MaxTDA can help identify and mitigate stellar variability in RV time series analysis; a complete analysis using real exoplanet data is the topic of future research.

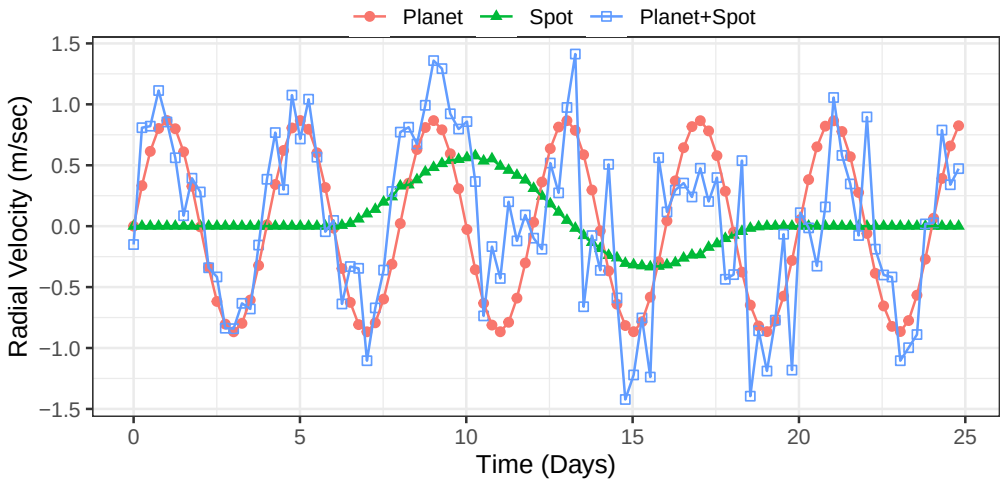


Figure 3.7: Exoplanet time-series data. Simulated RV data of an exoplanet (red circles), a 0.05% spot (green triangles), and the Planet+Spot combined (blue squares).

Our focus is on enhancing feature persistence in combined signals (e.g., Planet+Spot) and assessing the statistical significance of periodic behavior. Using simulated data (Figure 3.7), we analyze a planet, a star spot, and their combined signal (P+S) RV time series. The spot-induced signal matches the star's 25.05-day rotation, while the planet orbits with a 4-day period and 0.87 m/sec semi-amplitude. A 0.05% star spot at 30° latitude induces a 0.58 m/sec apparent RV signal. $N(0,1)$ noise was added to ensure the most persistent H_1 feature in the combined RV signal is close to the spot's H_1 feature before MaxTDA is applied.

TDE matrices were constructed for each time series, with AMI and Cao's used to select $(\tau = 4, M = 15)$ for the planet and $(\tau = 12, M = 7)$ for the spot. Instead of estimating new parameters for the combined

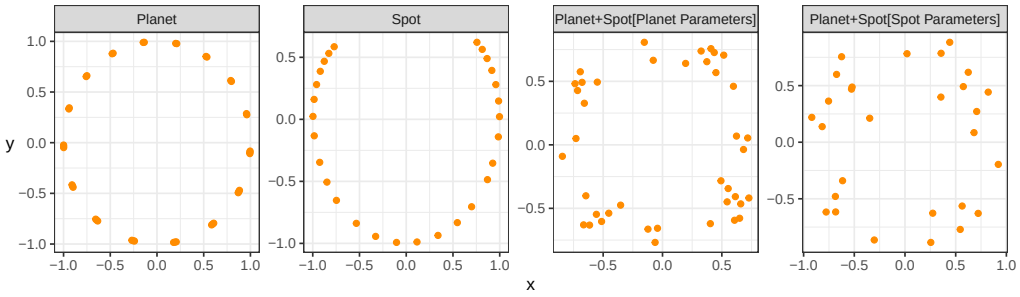


Figure 3.8: The embedded time series from Figure 3.7. The Planet and P+S[Planet Parameters] used $\tau = 4, M = 15$, while the Spot and P+S[Spot Parameters] used $\tau = 12, M = 6$.

signal, we applied the individual embeddings separately, allowing direct comparison of structural and temporal properties. This approach helps assess whether the time series geometry suggests a planet's presence. Each embedding matrix was centered, normalized, and reduced via PCA to two components for analysis (Figure 3.8).

3.4.3 Quantifying periodicity

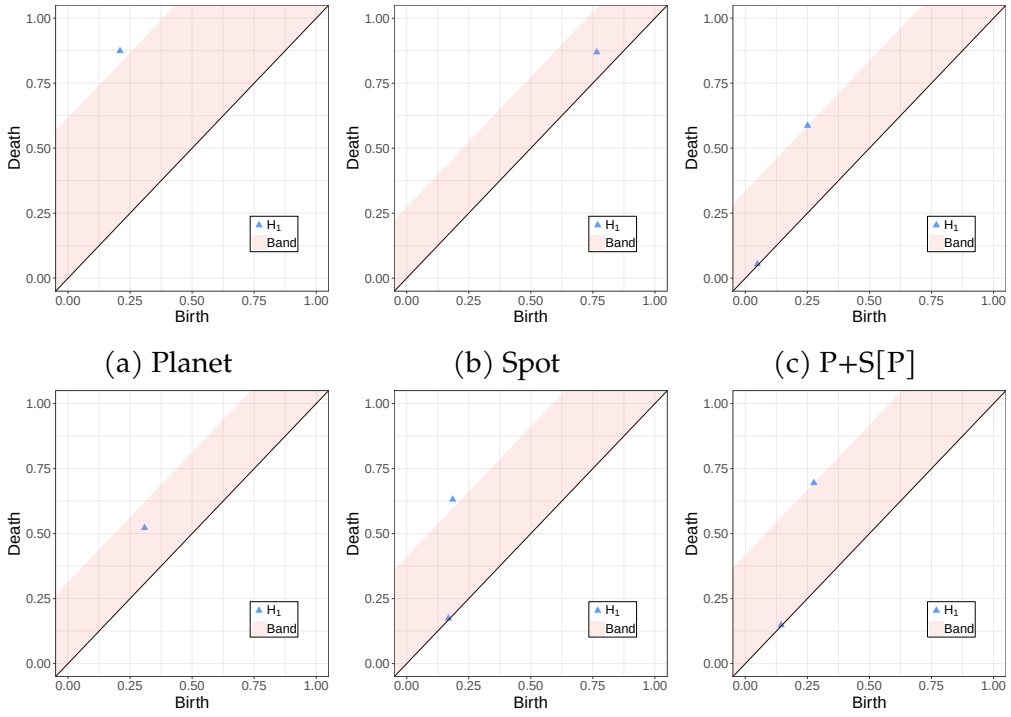
The periodicity of a time series can be assessed using the H_1 features of its TDE, where periodic patterns form elliptical shapes in thstate space (Perea et al., 2015). The roundness of these ellipses, quantified by the maximum persistence of H_1 features, serves as a periodicity score: $\max_{(b,d)\in\widehat{Dgm}_1} |d-b|$. For example, a time series that produces a well-sampled circular loop in its TDE will have high persistence and, therefore, a high periodicity score.

Algorithm 2 was applied to the P+S[Planet Parameters] and P+S[Spot

Parameters] TDEs to reduce noise, with the optimal KDE bandwidth set as the average 1-NN distance. To construct rejection bands, a DTM filtration with $m = 0.01$ was used for the Planet, the P+S[Planet Parameters], the P+S[Spot Parameters], the Smooth P+S[Planet Parameters], the Smooth P+S[Spot Parameters] embeddings, and $m = 0.05$ for the Spot embedding, which were selected to maximize the H_1 features. Figure 3.9 display the persistence diagrams. The Planet signal has the highest periodicity score (0.6647), followed by smoothed P+S[Planet Parameters] (0.4531), both statistically significant at the 5% level. The lack of significance in other embeddings is attributed to noise, data distribution variation, and the gap in the Spot's embedding. In summary, MaxTDA enhances the H_1 feature persistence in the Planet+Spot embedding. This approach is particularly useful for analyzing time series signals with missing observations (Dakurah and Cisewski-Kehe, 2024), embeddings with varying sampling density, or noisy time series where distinguishing or removing noise is impractical or undesirable in the time domain.

3.5 Discussion and Conclusion

This work introduces the MaxTDA methodology that combines kernel smoothing and level-set estimation via rejection sampling to facilitate robust statistical inference for the maximal persistence features in a topological space. Thresholding the KDEs at a suitable level creates a smooth



(a) Planet (b) Spot (c) P+S[P]

(d) P+S[S] (e) Smoothed P+S[P] (f) Smoothed P+S[S]

Figure 3.9: Persistence diagrams (only H_1 features) for the Planet (a), Spot (b), and the combined Planet+Spot embeddings and their smoothed versions (c-f) with 95% rejection bands..

and dense sampling surface. Rejection sampling is then used to obtain samples that result in improved robustness of estimated homology features with limited reduction in the lifetimes for the maximally persistent feature(s). The maximal persistence estimator is shown to be consistent, and achieves a reduction in bias relative to existing robust TDA methods. The statistical significance of the maximal persistence estimator was assessed via the construction of confidence sets. Several numerical experiments were conducted to illustrate the effectiveness of MaxTDA in uncovering,

validating, and drawing meaningful statistical inference for the maximal persistence features of datasets.

There are several important directions for future work and potential improvements. The proposed rejection sampling technique, while effective in low-dimensional settings and geometries that are relatively well-behaved, may face difficulties in ensuring that the sampled points adequately cover the features of interest in complex and high-dimensional data spaces. Dimension reduction methods, such as PCA (e.g., Section 3.4.1) or manifold learning techniques, could be applied as a preprocessing step to improve the effectiveness of the sampling scheme. Alternatively, more adaptive or data-driven sampling strategies could be explored, for example, using importance sampling or Markov chain Monte Carlo approaches that target the most relevant regions of the data space. These adjustments may lead to improved coverage of salient features in higher dimensions, and better stability and efficiency in empirical implementations.

In the exoplanet application in Section 3.4.2, a method is proposed to study the contributions of the planetary signal to the combined signal that includes stellar variability due to a spot. While this illustration highlights the scientific challenge of detecting low-mass exoplanets in the presence of stellar activity, real RV data can include multiple planets, multiple time-evolving spots, highly irregular time sampling, instrumental effects, and other complexities. The proposed approach should be considered a preliminary proof of concept requiring further validation across diverse

signal scenarios, and serves as an interesting area of future research on topological signal decomposition.

4 SUBSEQUENCE EMBEDDING FOR ROBUST CLASSIFICATION OF RADIAL VELOCITY TIME SERIES WITH MISSING DATA

Abstract

Radial velocity (RV) measurements are a foundational tool for detecting exoplanets, which are planets that orbit stars other than our sun. The RV method infers the presence of an exoplanet by measuring the periodic forward and backward motion of a potential host star over time. However, these measurements are typically perturbed by stellar activity, photon noise, and structured missingness inherent in ground based observations, making it difficult to distinguish planetary signals from stellar variability. The conventional approach to handling missingness is to use imputation techniques, which may distort the underlying dynamics, especially under structured missingness or large temporal gaps. This work proposes Subsequence Embedding (SSE) as an alternative and more robust methodology that maintains the geometric integrity of time series state space without interpolation across observational gaps. We reformulate exoplanet detection as a classification problem, employing SSE to construct a multi-dimensional representation of the irregularly-sampled RV data, from which feature vectors are extracted via a fast convolutional kernel transform. Using SOAP 2.0 to generate RV datasets with varying levels

of stellar activity, noise, and missingness, we demonstrate that SSE-based classification outperforms imputation-based approaches. Our method maintains high discriminative power ($\text{AUC} > 0.79$) even with 50% missing data and high measurement noise, while imputation methods degrade to near-random performance ($\text{AUC} \approx 0.55$) under similar conditions. This performance advantage increases with data sparsity, highlighting SSE's ability to capture essential dynamical information from incomplete observations. Our approach offers a data-driven, model-agnostic framework for analyzing irregular astronomical time series without introducing artifacts from imputation procedures, with potential applications to other fields dealing with non-uniform temporal data.

4.1 Introduction

The radial velocity (RV) method as briefly introduced in Section 3.4 of Chapter 3, has been a cornerstone of exoplanet discovery for decades. By measuring the Doppler shift of stellar spectral lines, astronomers infer the tiny wobble induced in a star by an orbiting planet, and remains one of the most direct ways to estimate exoplanet masses (Mayor and Queloz, 1995; Hara and Ford, 2023). However, RV measurements are contaminated by stellar variability: stars are not static, they exhibit oscillations, granulation, and rotating surface features such as starspots and faculae. These stellar activity signals can induce RV variations on the order of meters per second,

comparable in amplitude and sometimes in periodicity to signals from low-mass planets (Huélamo et al., 2008; Dumusque, 2016; Davis et al., 2017). As a result, distinguishing a genuine planetary Doppler signal from spot-induced RV noise is a major challenge. Compounding the difficulty, RV observations are typically irregularly sampled. Ground-based telescopes can only observe non-solar stars at night and are limited by weather, scheduling, and seasonal visibility of targets. This leads to structured missingness in the time series: large gaps (seasonal) and uneven spacing between observations. (For examples, see the left column of Figure 3 of Zhao et al. (2022), which displays RVs for four stars using the EXPRES spectrograph; there are between 22 and 58 nights of observations across over a year window in 2019 and 2020 for these stars.) Effective handling of such data with noise is crucial to ensure accurate analysis of planetary activities and detection of planetary signals.

The traditional statistical approach for handling missing data is imputation (Vacek and Ashikaga, 1980; Harvey and Pierse, 1984). For example, imputation techniques like Last Observation Carried Forward (LOCF), K-Nearest Neighbors (KNN), spline and linear interpolation remain popular. However, if the imputation model is misspecified, it can produce structures that do not reflect true properties of the data. Moreover, many imputation methods rely on inter-attribute correlations to estimate values for the missing data, which are not present in univariate time series (Vacek and Ashikaga, 1980; Harvey and Pierse, 1984; Casdagli et al., 1991; Lekscha

and Donner, 2018). In astrostatistics, several new statistical methods have been developed to address the challenge of detecting low-mass exoplanets in the presence of stellar activity (e.g., Rajpaul et al. 2015; Dumusque 2018; Holzer et al. 2021a,b; Jones et al. 2022), but none of these methods fully or generally mitigate the issues (Zhao et al., 2022)[1].

In this work, we investigate a new approach to detecting the presence of a planetary signal in noisy RV time series with missing data using *Subsequence Embedding* (SSE) (Dakurah and Cisewski-Kehe, 2024). Specifically, the goal of this work is to answer the question: does an observed RV time series contains evidence of a planetary signal or only stellar activity? We approach this as a classification problem using simulated data for ground truth. The key challenge is to engineer features from the irregular, noisy time series that preserve the underlying dynamics of planet-induced dynamics, without being distorted by irregular sampling. To accomplish this, we applied the SSE method, which transforms the univariate, noisy and irregular time series data into a multi-dimensional representation. The SSE is shown to be robust with respect to irregular sampling, that is, it preserves the dynamics of the underlying state space (Dakurah and Cisewski-Kehe, 2024). The constructed embedding is treated as a multi-dimensional time series, and we apply a fast convolutional kernel transform for time-series,

[1] In astronomy, a common approach is to use a Lomb-Scargle periodogram (Lomb, 1976; Scargle, 1982) to detect significant periodic signals in irregularly sampled data, followed by a model fitting analysis to refine orbital or other physical parameters. See Section 5.2 of Dakurah and Cisewski-Kehe (2024) for an example of this approach in asteroid data.

to generate features (Dempster et al., 2020, 2021). A *Random Forest* is trained to predict the presence (or absence) of a planetary signal. We evaluate performance in terms of area under the ROC curve (AUC), comparing SSE-based versus imputation-based embeddings across multiple missingness scenarios and noise levels. The SSE method is compared to the conventional method of handling missing observations in the literature via imputation. For the imputation, we used spline interpolation to fill-in the missing values, and apply regular time-delay embedding to construct the multi-dimensional representation of the series (Takens, 2006).

While many approaches to time series classification operate directly on the raw observations or their periodograms, this work focuses on embedding-based representations that aim to recover the underlying dynamical structure of the signal. Our emphasis on embedding methods is motivated by the irregular nature of RV observations, where interpolation or periodogram-based preprocessing may introduce artifacts or obscure weak planetary signals. Although other non-embedding methods may perform well under different scenarios, our results show that the embedding space alone contains enough geometric information to discriminate between stellar activity and planet-induced variability. This finding opens the door to future extensions where the embedding can inform downstream tasks. In the broader context of RV analysis, the ability to detect the presence of a planet from sparse and noisy data using only structural properties of the time series is a valuable step toward more interpretable

and robust detection pipelines.

Our contributions in this work are threefold. First, we formulate the exoplanet detection problem in RV data as a classification task, and construct synthetic dataset of stellar RV time series with and without planetary signals in the presence of stellar activity and observational gaps. Second, we introduce the subsequence embedding approach to this problem and demonstrate its effectiveness in handling highly irregular data. We show that SSE embeddings capture essential dynamics even with substantial missingness and noise, whereas interpolation based embeddings perform poorly. Third, we integrate SSE with a modern classification pipeline, and provide a thorough comparison of classification performance under varying levels of missing data. A methodological innovation worth noting is how the SSE transforms the univariate irregular time series into a multidimensional time series that can be treated as a multivariate time series. Our results show that SSE-based classification maintains high AUC even when about half of the data is missing, significantly outperforming the imputation-based approach. The remainder of this paper is organized as follows. Section 4.2 introduces the SSE transform tailored for classification tasks. Section 4.3 describes the data generation, preprocessing, experimental setup, and results. Section 4.4 concludes with discussion and future directions.

4.2 Deterministic Transform of SSE

Let $\mathbf{x} = (x(t) : t \in \mathcal{T})$ denote a univariate time series of RV measurements of length n, sampled at nonuniform time points. That is, $\mathcal{T} = \{t_1, \cdots, t_n\} \subset \mathbb{N}$ such that $t_{i+1} - t_i \neq t_{i+2} - t_{i+1}$, for at least one $t_i \in \mathcal{T}$ and $t_i < t_{i+1}$. Our objective is to construct a representation of $\mathbf{x}$ that can be used to detect the presence or absence of a planetary signal in the time series. We consider the *subsequence embedding* (*SSE*) method, which extends classical time-delay embedding to irregularly-sampled series (Dakurah and Cisewski-Kehe, 2024). The core idea is to construct local embeddings only from observed subsequences of $\mathbf{x}$, without imputing values over large temporal gaps. This allows the embedding to preserve the structure of the observed data regardless of the sampling pattern.

Let $\mathbf{F} \in \mathbb{R}^{N \times (M+1)}$ be the SSE embedding matrix of $\mathbf{x}$, where $M + 1$ is the chosen embedding dimension and N is the number of embedding vectors extracted from the data. We treat $\mathbf{F}$ as a multivariate time series with $M + 1$ channels and variable length N. To transform $\mathbf{F}$ into a fixed-length feature vector suitable for classification, we apply an almost deterministic very fast multivariate transform for time series, as introduced in Dempster et al. (2021). This transform automatically generates features from the input time series for use in downstream tasks. We chose it for its computational efficiency, scalability to large collections of time series, and strong empirical performance across a wide range of time series classification

tasks (Dempster et al., 2020, 2021; Middlehurst et al., 2024).

The transform works by generating a large number of fixed, nearly deterministic convolutional kernels and applying them to the embeddings. Summary statistics of the resulting convolution outputs are then used as features. Let $\{\mathcal{Q}_1, \ldots, \mathcal{Q}_L\}$ denote a collection of L univariate convolutional kernels. Each kernel $\mathcal{Q}_l$ is applied independently to each of the $M + 1$ channels (columns) of $\mathbf{F}$. A kernel $\mathcal{Q}_l$ is defined by a fixed-length filter of size 9 with weights $w = \{-1, 2\}$, a dilation $\eta_l \in \mathbb{N}$, and a fixed bias threshold determined by a chosen quantile of the convolutional output. That is, for a given channel $m \in \{1, \ldots, M+1\}$, kernel $\mathcal{Q}_l$ defines a univariate convolution:

$$[\mathbf{F} * \mathcal{Q}_l]_m(t) = \sum_{i=1}^{9} w(i) \cdot \mathbf{F}_{t+i\cdot\eta_l, m}, \tag{4.1}$$

whenever $t + i \cdot \eta_l \leqslant N$. The transform computes a summary statistic from the convolutional output for each (l, m) pair, typically the *proportion of positive values*:

$$\sigma_{l,m}(\mathbf{F}) = \frac{1}{N_l} \sum_{t} \mathbb{1}\left([\mathbf{F} * \mathcal{Q}_l]_m(t) > q_l\right), \tag{4.2}$$

where q_l is a quantile-based threshold and N_l is the number of valid convolution outputs for kernel $\mathcal{Q}_l$ on channel m. The combination of these statistics across all channels and kernels produce the final feature vector:

$$\sigma(\mathbf{x}) = (\sigma_{1,1}(\mathbf{F}), \ldots, \sigma_{L,M+1}(\mathbf{F})) \in \mathbb{R}^{L\cdot(M+1)}. \tag{4.3}$$

Note that the kernel $\mathcal{Q}_l$ is composed of the triplet $\{w_l, \eta_l, q_l\}$. The filter weights w_l are binary valued $(-1 \text{ or } 2)$, and many unique sign patterns, dilations, and thresholds are enumerated to produce L distinct kernels. To ensure a fixed-length feature vector of a given dimension, the kernels are distributed across the $M+1$ channels either evenly or proportionally. The transformed features $\sigma(\mathbf{x})$ are used as input to a classifier. In this work, we generate approximately $10,000$ features, hence we set $L(M+1) = 10000$ and solve for the number of unique kernels L based on the embedding dimension. Once $M+1$ is chosen for the embedding, we enumerates L distinct definitions (w_l, η_l, q_l), apply each to all $M+1$ channels, and combine the resulting *proportion-of-positives* to form a unified 10000-dimensional feature vector. We evaluate classification performance using only these transformed features. As the choice of classifier does not significantly affect our results, we use a Random Forest classifier for both the SSE and imputation-based approaches.

4.3 Experiments and Results

This section presents details on the generation of the RV dataset, the setup of the classification task, and a discussion of the corresponding results.

4.3.1 Dataset and preprocessing

We generate a broad range of starspot-induced RV signals using the Spot Oscillation And Planet (SOAP) 2.0 code (Dumusque et al., 2014). Various key spot parameters are systematically varied to simulate different levels of stellar activity. The spot-to-photosphere temperature contrast takes values in $\{200, 300, 400, 500, 663\}$ Kelvins, while spot sizes were selected in the range $[0.0025, 0.1]$ in fractional surface coverage, and the number of spots spans 1 to 4. For each configuration, we specify stellar inputs including a rotation period of 25.05 days and spot properties (longitude, latitude, size scaling), which are automatically written into SOAP 2.0 configuration files. SOAP 2.0 then simulates the effect of these active regions on the stellar line profiles, outputting model cross-correlation functions (CCFs) that are converted into RV values. The simulation produces a dense time series for each spot configuration by sampling the full stellar rotation with a phase step of 0.01. Each resulting RV profile is stored in a single column, yielding multiple columns that differ in temperature contrast, spot size, and spot count. We convert RV values from km/s to m/s and scale the model phase to days by multiplying by 25.05 days. Figure 4.1 shows the spot-induced RV time series for a single spot across various temperature differences and spot sizes. When the spot is behind the star, the RV is zero, and when the spot rotates in view, it produces a sinuosoid-like signal as indicated by the peak and valley. A similar pattern is observed when two spots are present, as shown in Figure 4.2, where two peaks and two valleys emerge. The

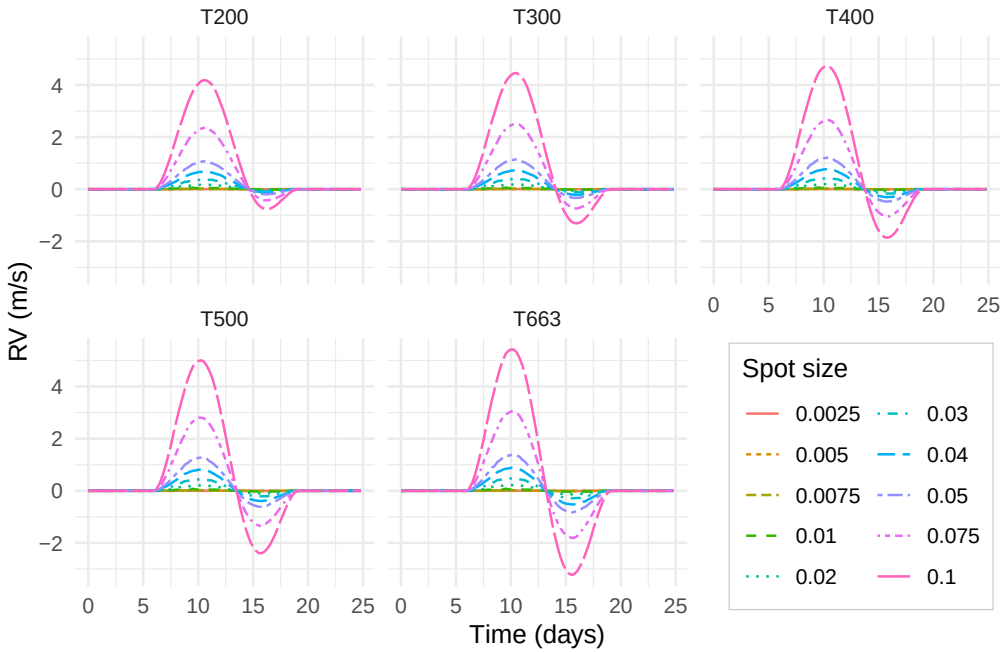


Figure 4.1: Spot-induced RV time series across various temperature differences and spot sizes when the number of spots is one. Observe that the larger the spot-size and temperature difference, the larger the RV signal.

relative size and shape of these features depend on the placement of the spots. Similar effects are seen with three or four spots.

In addition to spot-only signals, we generate a purely planetary RV time series with an orbital period 4 days and semi-amplitude 0.87 m/s. This planetary RV signal is added to the spot-induced signal to create Planet+Spot signals by superposition. Figure 4.3 shows an example of such Planet+Spot signal across various temperature differences and spot sizes when the number of spots is one. This procedure results in two main classes of synthetic RV time series: spot-only (Class 0) and Planet+Spot

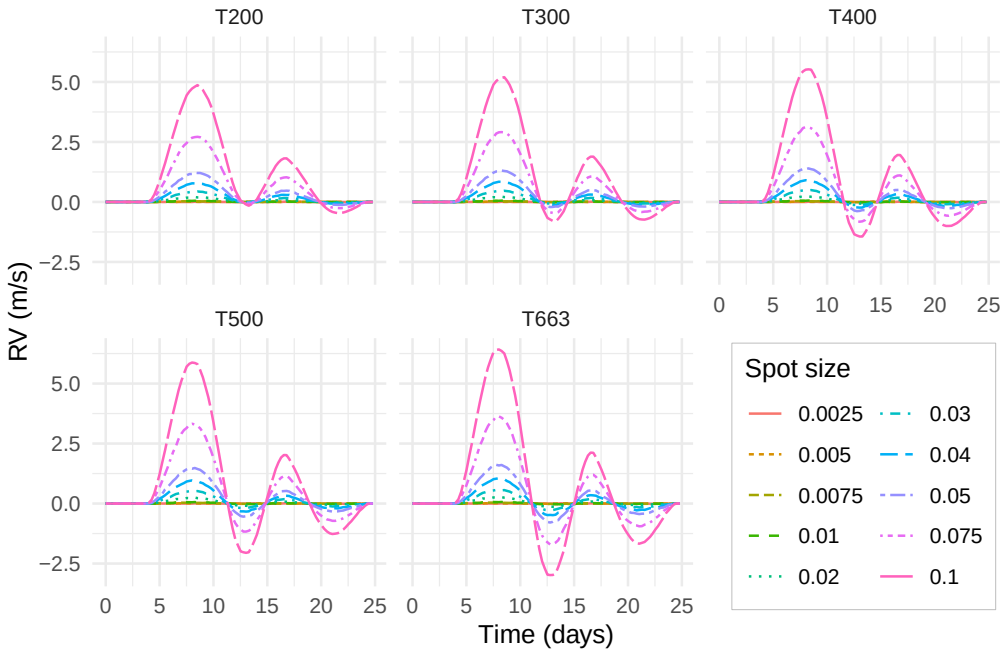


Figure 4.2: Spot-induced RV time series across various temperature differences and spot sizes when the number of spots is two. Larger the spot-size and temperature difference, the larger the RV signal.

(Class 1). Each series is labeled according to the presence or absence of the planetary signal, enabling supervised classification. In the next section, we introduce measurement noise and irregular sampling to replicate real-world observing conditions.

4.3.2 Experimental setup

The combined Planet+Spot and Spot-only signals were each perturbed with Gaussian noise, with standard deviation $\sigma \in \{0.5\,\mathrm{m/s}, 0.75\,\mathrm{m/s}, 1\,\mathrm{m/s}\}$ to mirror measurement uncertainty due to photon noise. The length of

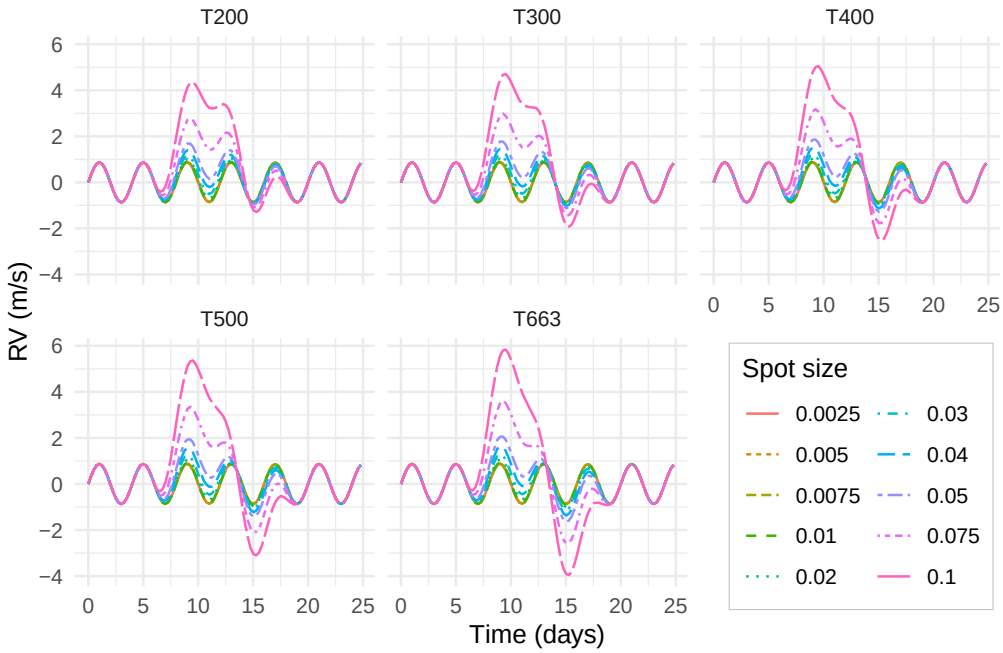


Figure 4.3: Combined Planet+Spot RV time series across various temperature differences and spot sizes when the number of spots is one.

each time series is fixed at 100. To emulate real-world observing schedules or missed observations, we impose structured missingness by randomly removing contiguous blocks of 10 time points. This process is repeated at varying levels of missingness: 10%, 20%, 30%, 40%, and 50%, allowing us to evaluate how classification performance degrades with increasing data loss. Note that these missing blocks are distributed randomly along the time series. Figure 4.4 shows an example combined Planet+Spot RV time series with random missing blocks. Consequently, each synthetic time series exhibits irregular spacing, i.e., clusters of points during observing windows and intervals of no data. Given these noisy and irregular time

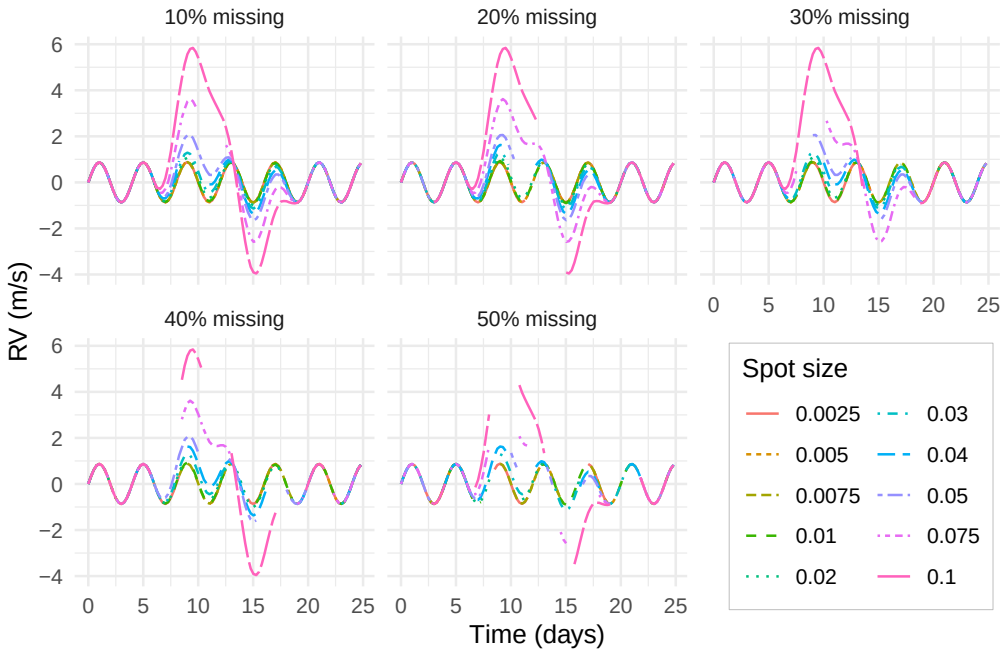


Figure 4.4: Combined Planet+Spot RV time series with random missing blocks at various proportion of missingness. The temperature differences is 663 Kelvins and the number of spots is one. The distortion caused by missing data highlights the challenge of separating planetary signals from stellar activity.

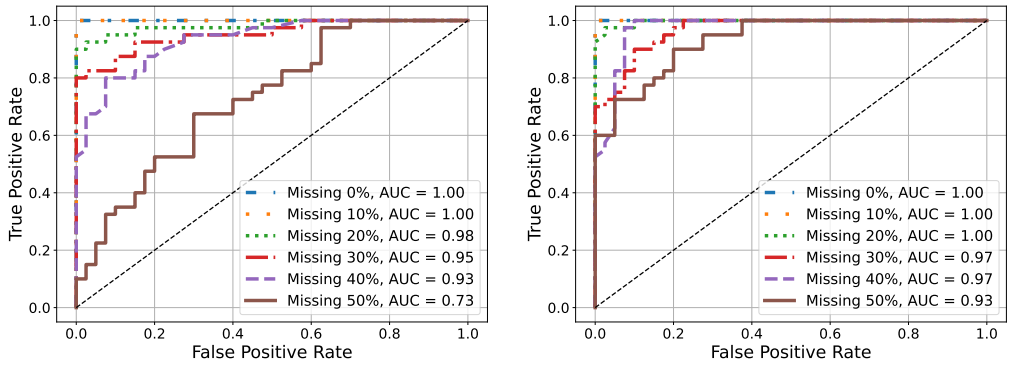
series, we prepare it for embedding in two ways. First, *Imputation-Based* uses spline interpolation to impute for the missing values and then applies a standard time-delay embedding with dimension $M = 3$ and delay $\tau = 1$. The interpolation does not recover new information but allows a uniform embedding as a baseline representation. Second, SSE is also applied to construct the embedding by first identifying observed subsequences within the irregular time series, then applying a local time-delay embedding to each subsequence, and the resulting embeddings are combined to pro-

duce the embedding matrix. The SSE matrix thus captures real dynamics without imputing over unobserved gaps. Finally, each time series, with or without a planetary signal, is represented by both an imputed embedding or by the SSE embedding. The SSE often produces fewer total rows than imputed embeddings, since it does not interpolate missing values.

The number of sample matrices is 400, with 200 in the Planet+Spot class and 200 in the Spot-only class, ensuring a balanced dataset. To fit the classifier, we divided the 400 samples into a training and testing set where 320 samples were designated as training set. The training set is then used to select the optimal parameters for the Random Forest classifier. Specifically, we search over the number of estimators $\{50, 100, 200\}$, the maximum tree depth $\{\infty, 10, 20\}$, and the minimum number of samples required to split a node $\{2, 5, 10\}$. For each level of missingness {10%, 20%, 30%, 40%, 50%}, we train and evaluate two models: (1) using the transformed samples from the TDE embedding computed on the imputed (uniformly-spaced) time series, (2) using the transformed samples from the SSE embedding without imputation. For ease of reference, we refer to these two models as the "Imputation model" and the "SSE model," respectively. We report the area under the curve (AUC) for the two models across the different missingness levels.

4.3.3 Experimental results

The experimental results, reported in terms of AUC across varying levels of missingness and noise, are presented below. Figure 4.5 shows the ROC curve and the corresponding AUC values when noise is low ($\sigma = 0.5$). Both pipelines perform well at low missingness. The AUC for the Imputa-

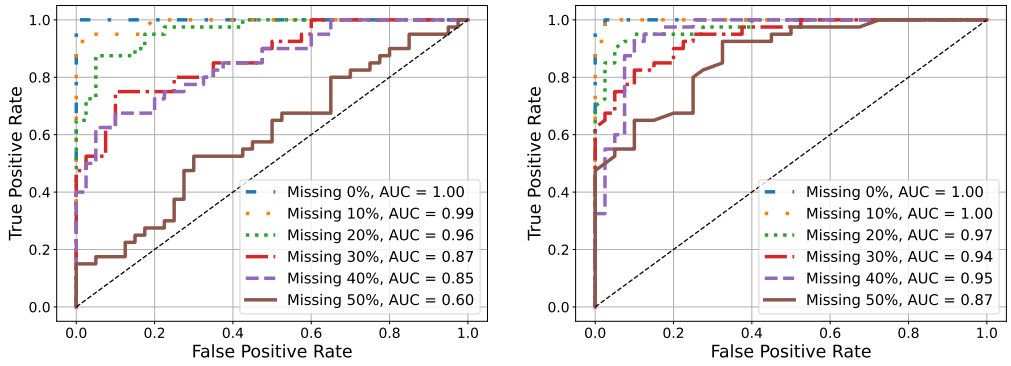


(a) Imputation model ($\sigma = 0.5$) (b) SSE model ($\sigma = 0.5$)

Figure 4.5: The ROC curves with the AUC values when $\sigma = 0.5$. For missing proportions above 10%, the SSE model consistently outperforms the Imputation model.

tion model decreases steadily from 1.0 at 0% to 0.73 at 50%. In contrast, the SSE model preserves performance more effectively, declining only from 1.00 to 0.93 across the same range. The performance gap widens as missingness increases, highlighting the advantage of avoiding interpolation when data gaps become more substantial.

Under moderate noise ($\sigma = 0.75$), performance degradation becomes more pronounced, as shown in Figure 4.6. The Imputation model shows a sharper decline, with AUC falling from 1.0 to 0.60, while the SSE model



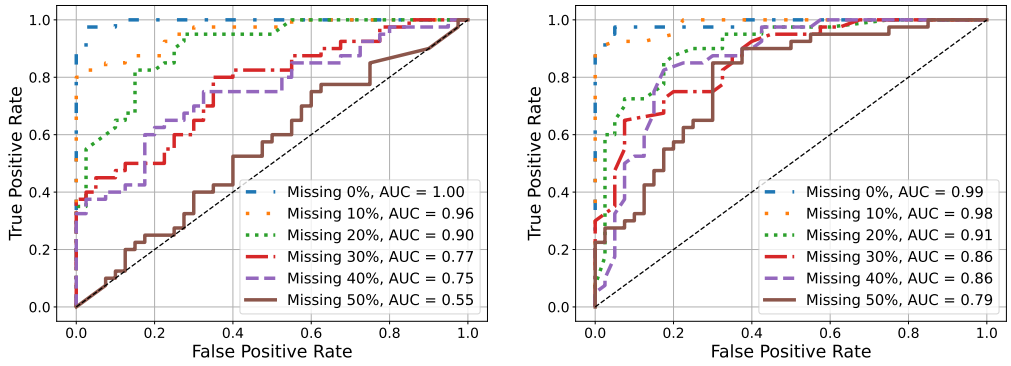
(a) Imputation model ($\sigma = 0.75$) (b) SSE model ($\sigma = 0.75$)

Figure 4.6: The ROC curves with the AUC values at $\sigma = 0.75$. The SSE-based method consistently outperforms the imputation-based method.

retains greater resilience, achieving an AUC of 0.87 even at 50% missingness. The difference in performance is most noticeable in the intermediate range of 30% to 50% missingness, where interpolation begins to distort the underlying signal structure.

Results under high noise conditions ($\sigma = 1.0$) are shown in Figure 4.7, where classification becomes significantly more challenging. The Imputation model fails to maintain discriminability, reaching an AUC of only 0.55 at 50% missingness, comparable to random guessing. The SSE-based method continues to offer better robustness, with AUC values ranging from 0.99 at 0% missingness to 0.79 at 50%. This demonstrates that SSE preserves enough signal structure to enable reliable discrimination even under severe sparsity and noise. Though both methods suffer from the high noise, SSE consistently maintains higher discriminative performance.

A further sensitivity analysis was conducted by repeatedly splitting the



(a) Imputation model ($\sigma = 1.0$) (b) SSE model ($\sigma = 1.0$)

Figure 4.7: The ROC curves with the AUC values at $\sigma = 1.0$. The imputation-based method performance degrades significantly, and at 50% missingness, it is indistinguishable form random guessing, while the SSE-based method is still significantly accurate.

dataset into training and test sets over 100 iterations, and then computing the average AUC (and its standard deviation) for each combination of missing proportion and noise level. The results, shown in Table4.1, confirm the same overall pattern: both SSE and the imputation-based method achieve near-perfect classification for low noise ($\sigma = 0.5$) and few missing data ($< 20\%$). However, as the missing proportion or noise level increases, the SSE model maintains higher average AUC and lower variance than the Imputation model. For instance, at $\sigma = 1.0$ and 50% missingness, the SSE model retains an average AUC of about 0.83 (std 0.05), whereas the imputation-based model falls to roughly 0.66 (std 0.06). This difference highlights SSE robustness in preserving essential signal structure, even under substantial irregular sampling and noise. These repeated train-test splits also indicate that the observed performance gap is not due to any

Table 4.1: The average AUC with standard deviation in brackets, across 100 repeated train-test splits, comparing SSE model and Imputation model under varying noise levels and missingness. SSE consistently achieves higher AUC and lower variance. The gray colored rows indicates combinations at which the SSE model outperforms the Imputation model.

		Proportion Missing					
σ	Model	0.00	0.10	0.20	0.30	0.40	0.50
0.50	SSE	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.99 (0.01)	0.98 (0.01)	0.95 (0.02)
	Imputation	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.98 (0.01)	0.93 (0.02)	0.84 (0.05)
0.75	SSE	1.00 (0.00)	1.00 (0.00)	0.99 (0.01)	0.97 (0.01)	0.94 (0.02)	0.90 (0.03)
	Imputation	1.00 (0.00)	1.00 (0.00)	0.98 (0.01)	0.94 (0.02)	0.85 (0.04)	0.73 (0.06)
1.00	SSE	1.00 (0.00)	0.99 (0.01)	0.97 (0.02)	0.93 (0.02)	0.89 (0.04)	0.83 (0.05)
	Imputation	1.00 (0.00)	0.98 (0.01)	0.93 (0.03)	0.86 (0.04)	0.75 (0.05)	0.66 (0.06)

particular data partition. Even at high noise and high missingness, SSE demonstrates relatively stable performance, as evidenced by its smaller standard deviations across multiple runs. By contrast, the imputation-based approach degrades more sharply, suggesting greater sensitivity to artificially filled gaps and the resulting signal distortion. Overall, these experimental results emphasize the robustness of SSE. Rather than smooth-

ing over unobserved intervals, SSE operates directly on observed data, capturing essential dynamics without introducing interpolation artifacts. While imputation methods remains a viable option for moderate missingness (e.g., at less than 10% missing values), they become unreliable when the level of missing values increases. Overall, SSE maintains consistently higher AUC across all noise and missingness levels, demonstrating its effectiveness for detecting the presence or absence of planetary signals in noisy and incomplete RV time series.

4.4 Discussion and Conclusion

The results show that subsequence embedding (SSE) improves the detectability of exoplanet signals in irregular RV time series more effectively than spline interpolation or other gap-filling strategies. By focusing only on observed measurements rather than predicting values in unobserved intervals, SSE preserves the geometry of the underlying dynamical system. This property proves especially beneficial at higher levels of data scarcity, where traditional interpolation may weaken or obscure periodic signals that are crucial for planet detection. A central reason for SSE's effectiveness lies in how it constructs time-delay embeddings: each continuous set of observations is processed independently, so large gaps are never connected. Missing portions of the time series are omitted, ensuring that embedded states represent only what has genuinely been recorded.

In contrast, interpolation-based embeddings blend observed data with extrapolated segments, which can either dampen small but meaningful oscillations or introduce new cycles that do not reflect any real feature. As our results show, this leads to systematically lower AUC, especially as data becomes sparse or noisy. By controlling how subsequences are extracted, SSE faithfully tracks the temporal and state-space structure that arises from spot-induced and planetary signals.

In astronomical contexts, this approach opens new opportunities for analyzing time series that are inherently unevenly spaced, whether due to telescope scheduling, weather, or seasonal visibility windows. The SSE framework is relatively straightforward to implement, yet it captures key properties of the original dynamical system without imposing strong assumptions about behavior in gaps. This is advantageous when classifying whether a given RV time series contains a planetary signal. Our experiments confirm that, even under substantial missingness, SSE embeddings offer higher discriminatory power compared to interpolation-based embeddings, meaning that a simple machine learning pipeline can more accurately separate planet-containing signals from purely activity-driven ones. We note that this approach is model-agnostic and relies purely on data geometry and temporal structure, making it adaptable across a wide range of time series classification tasks in other disciplines.

Our study focused on a single classification pipeline that paired SSE with Convolutional kernel feature extraction. In principle, one could re-

place this with alternative algorithms or incorporate additional feature engineering, such as topological descriptors, to further enhance classification. Further, our use of it in conjunction with a machine learning classifier is somewhat novel; previous work focused on its topological faithfulness (Perea and Harer, 2015; Dakurah and Cisewski-Kehe, 2024), but we have shown that this faithfulness translates into better machine learning performance too. In real data, spot evolution, flares, and other non-stationary effects may require adaptive choices for the delay parameters. Despite these extensions, SSE already shows notable promise, providing a robust and data-driven foundation for learning from sparsely sampled observations.

Overall, SSE offers a practical way to handle large gaps or irregular sampling in RV studies and potentially many other fields that rely on non-uniform time series. By using only the observed samples, it avoids the uncertainty of gap filling and more reliably recovers the geometric signatures of periodic or quasi-periodic processes. This property, combined with the flexibility of modern machine learning methods, can make detection pipelines both more accurate and more robust.

5 PERSISTENCE SIGNATURES IN MOLECULAR DYNAMICS SIMULATIONS OF IONIC LIQUIDS

Abstract

Ionic liquids (ILs) are room-temperature salts that often exhibit heterogeneous nanoscale organization. Understanding this internal structure is crucial because it underlies key properties such as ionic transport, viscosity, and electrochemical performance. In this work, we introduce a unified topological data analysis (TDA) framework to characterize IL nanostructures from molecular dynamics (MD) simulations. We leverage persistent homology to capture multiscale topological features of the MD-generated point clouds (ion positions), and integrate these descriptors with statistical methods. The proposed methodology encompasses persistence-based summaries, change point detection of structural transitions, and spatial point process modeling to quantify how topologically identified clusters or loops are spatially arranged. Applied to IL simulation data, this pipeline reveals interpretable descriptors of nanoscale morphology and detects structural transitions that are interpretable and relatable to physical properties of ILs. The approach is validated on two representative case studies (varying cation alkyl chain length and IL concentration), where it successfully identifies regime shifts in nanostructure. This work is a result of a collaboration with Lisa Je and Reid Van Lehn from the Department

of Chemical and Biological Engineering at the University of Wisconsin-Madison.

5.1 Introduction

In molecular systems, structure at the nanoscale refers to the non-random organization of molecules into local patterns or domains on length scale of a few nanometers. Such nanoscale structuring is physically meaningful because it arises from intermolecular forces and often governs bulk behavior. Understanding how molecules arrange themselves at nanometer scales is key to connecting microscopic interactions with macroscopic properties in molecular simulations (Wang et al., 2020; Jiang et al., 2018; Walker et al., 2018). Molecular dynamics (MD) simulations are a powerful tool for probing and modeling these complex, evolving nanoscale structures. In MD simulation, a large number of molecules are tracked in time, allowing emergent structural patterns to develop naturally from fundamental intermolecular interactions. Because MD can isolate specific interactions or molecular designs, it has been pivotal in explaining how subtle changes in molecular structure, such as alkyl chain length or functional group placement can influence the resulting nanostructure(Wang et al., 2020; Hollingsworth and Dror, 2018; Jiang et al., 2018). The MD simulations used in this work were provided by collaborators Lisa Je and Reid Van Lehn from the Department of Chemical and Biological Engineering at the

University of Wisconsin-Madison.

Extracting meaningful structural descriptors from MD data, however, poses a significant challenge. A typical MD trajectory, which is a time-ordered sequence of atomic positions generated by simulating the motion of atoms and molecules, generates a vast amount of atomic coordinates, from which one must extract meaningful patterns that reflect how the system is organized at small scales. Traditional measures like radial distribution functions and cluster analysis summarize structural features over time, but they can miss or obscure unique or *transient* structures (Smith et al., 2023; Je et al., 2022; Jiang et al., 2018). More specifically, MD analysis strategies such as tracking a few predefined order parameters, performing clustering in coordinate space, or applying linear dimensionality reduction often struggle to capture the full richness of the system's behavior as they focus on two-body correlations or require apriori definitions of an order parameter. Important collective motions or structural changes may be missed when using overly simplistic descriptors. The core methodological challenge is thus one of *structure detection at the nanoscale, feature extraction, and interpretability of the structure and transient features.* Achieving this requires the development of new analysis pipelines that are both quantitatively robust and physically interpretable.

In persistent homology, a single snapshot at fixed resolution is not considered, rather a multi-scale family of complexes is built from the point cloud which tracks the birth and death of the topological features as the

observation scale varies. For instance, in MD simulations, the 3D Cartesian coordinates of molecules form a point cloud that captures snapshots of molecular structures. In this context, the point cloud is analyzed through a VR filtration introduced in Section 1.3 of Chapter 1. Instead of focusing on specific low-dimensional descriptors derived from domain knowledge that capture essential aspects of a molecular system's behavior, TDA examines the overall shape and structure of the data by identifying patterns or structures that are consistently present throughout the simulation. This approach allows for the quantification of certain shape features in the data that are useful for characterizing the molecular system's physical properties. TDA provides an alternative and complementary framework to conventional analysis, one that is sensitive to global structural patterns and intrinsic geometry rather than just local correlations.

In this work, we present a unified methodology that integrates persistent homology with statistical techniques to characterize IL nanostructure in MD simulations. The novelty of our approach lies in combining topological quantification provided by persistent homology with statistical analyses that enhances robustness, stability and physical interpretability. First, we perform persistent homology on MD-generated point clouds of ions, obtaining persistence diagrams that serve as descriptors of structure. Summary persistence statistics, such as the mean persistence, variance of persistences, or maximum persistence, are then computed across different point clouds from different simulation conditions. By treating these sum-

mary measures over different frames as time series data (see Section 5.3.1 for details), we apply change point detection algorithms to identify statistically significant changes in the IL's structural state over time. This allows us to automatically detect the onset of a structural transition, for example, when an initially homogeneous mixture begins to segregate. In addition, we incorporate a spatial point process perspective to interpret the geometry of the identified structures. Specifically, if persistent homology indicates the presence of certain features, we model the location of these features as points in space and analyze their spatial distribution. Using tools from spatial statistics, we can determine whether the topology-derived structures themselves are randomly distributed, form a regular lattice, or exhibit higher-order clustering.

The proposed framework enables detection of nanostructural phase transitions, identification of distinct regimes, and a nuanced interpretation of local aggregation in IL systems. For example, as simulation conditions evolve, our approach can pinpoint the moment an IL switches from one nanostructural regime to another by detecting abrupt shifts in persistence summaries. This capability is particularly valuable for ILs where transitions may be gradual or not apparent in traditional methods like radial distribution function. Furthermore, the integration of spatial point process models provides interpretative context for the detected topological structures. Rather than simply stating that a certain persistence summary increases, we can interpret this as evidence that loop-like structures in the

point cloud are arranged in a connected network, similar to pathways or channels running through the system, or that charged particle groups are unevenly distributed, which suggests the formation of larger and distinct regions within the point cloud. Such insights bridge the gap between abstract topological measures and the tangible structural concepts familiar to domain experts. To demonstrate the utility of this framework, we apply it to two illustrative case studies (detailed in Section 5.4.2 and 5.4.3). In the first case study, we examine a family of ILs with varying alkyl chain length on the cation. Our analysis captures how increasing the chain length gradually intensifies nanosegregation and eventually triggers a transition to a more percolated domain structure, *all identified via changes in persistent homology signatures*. In the second case study, we investigate an IL system at different concentrations, which allows us to probe how diluting the IL affects its internal organization. The pipeline detects the emergence (and dissipation) of ionic aggregates as the concentration changes, effectively mapping out distinct structural regimes from isolated ion pairs in dilute conditions to extensive ionic networks in more concentrated conditions. *Notably, these regime boundaries and structural insights arise naturally from our unified analysis, without a priori assumptions.*

This methodology highlight how a TDA-guided approach can uncover clear, physically meaningful patterns in IL simulations. The rest of this work is structured as follows: Section 5.2 provides a motivation example as a preview to the methodology, Section 5.3 details the proposed analysis

method while Section 5.4 demonstrates the pipeline with two case studies. Finally, Section 5.5 concludes the work and discusses directions for future research.

5.2 Motivating Example

First, we demonstrate how we can apply TDA to a simple, 2D toy dataset that convey some properties of the more complex 3D simulations analyzed in subsequent sections. A group of datasets with 500 points each were randomly generated and shown in Figure 5.1. This group of data is composed of four different point clouds numbered Groups 1 - 4. Group 1 consists of uniform samples randomly drawn from the interval $[-4, 4]$, while Groups 2 - 4 are generated from an asymmetric knot and perturbed with Gaussian noise. The noise has a mean of 0 and decreasing standard deviations of 0.2, 0.1, and 0.025 for Groups 2, 3, and 4, respectively. The goal is to develop a

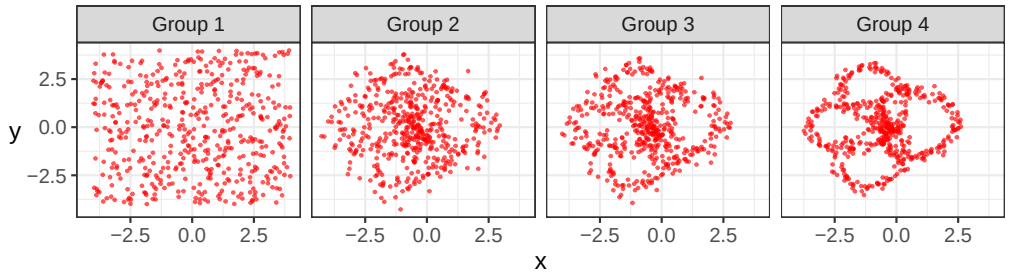


Figure 5.1: Group 1 is randomly distributed with no apparent pattern. Group 2 point appears to be clustering around the center. Group 3 and Group 4 manifest four elliptical empty shells with Group 4 being more prominent relative to Group 3.

methodology for systematically quantifying the evolving structure, in this example, the shape of the point cloud across Groups 1–4, capturing both qualitative and quantitative properties. TDA tools, such as persistence diagrams, allow for the quantification of differences between these four point clouds by measuring the systematic aggregation of points, which leads to the formation of loops.

A persistence diagram is generated from each point cloud dataset to quantify the number of components and loops the data contain. For the toy dataset displayed in Figure 5.1, the corresponding persistence diagrams are displayed in Figure 5.1. By observing the birth and death of

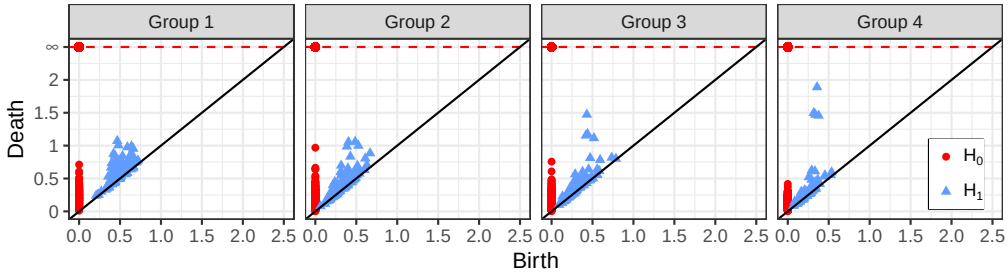
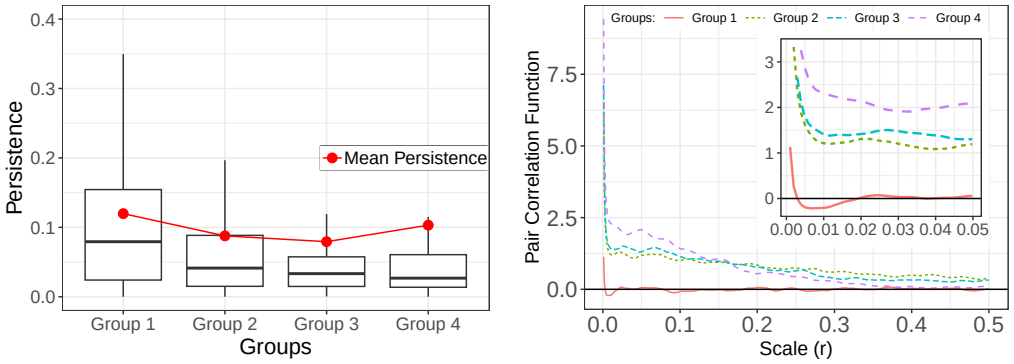


Figure 5.2: The persistence diagrams corresponding to the four groups of point clouds in Figure 5.1. Observe the H_1 features across the four groups. Groups 3 and 4 have four blue triangles that are distinctively above the rest of the blue triangles, indicative of the four elliptical empty territories in Group 3 and Group 4 of Figure 5.1.

topological features over a set of filtration values, we can observe robust and stable topological features in the point cloud dataset. For example, in the persistence diagrams for Group 3 and Group 4 in Figure 5.2, we observe four dominant blue triangles, indicated by H_1, which identifies



(a) Boxplot of persistence values. (b) The pair correlation function.

Figure 5.3: (a) A boxplot (displayed without outliers) of the persistence of the H_1 features. (b) The pair correlation function (with an inset for visual clarity) used to measure the degree of aggregation of the sequence.

the four empty elliptical loops or circles shown in Figure 5.1 for Group 3 and Group 4.

Our second TDA graphical summary are side-by-side box plots that display the distributions of the computed persistence of the point cloud over a series (shown in Figure 5.3a). For this demonstration, we only use the mean persistence, calculated for each point cloud. As the series progresses, the summary persistence exhibits significant shifts that indicate changes in the overall trend. For example, the mean persistence decreases from Group 1 to Group 3 and then increases in Group 4. An abrupt change (i.e, a shift in the trajectory of a persistence summary over the group sequence) can be captured using change point analysis tools, as discussed in Section 5.3.2, which provides a more nuanced perspective on the data's evolving structure. Group 3 is identified as a change-point in

this sequence, and serve as an anchor point for further analysis of structural change. Other summary statistics on the persistence offer additional insights into the number of connected components and holes present in the dataset.

The last technique is the pair correlation function (PCF), which is not a TDA method, but is produced from the point cloud data to measure the level of aggregation or clustering at different scales or distances between the points. Similar to the TDA summary statistics, we can qualitatively observe any clustering of points through the peak intensity and peak location of the pair correlation function. Figure 5.3b shows the pair correlation function of the data sequence in Figure 5.1. A high-intensity peak means there are a lot of nearby points whereas the location of the peak on the x-axis indicates the distance scale a which the pair-wise interaction is found.

5.3 Nanostructure in Persistence Summaries

In this section, we outline the post-processing tools and methods applied to persistence summaries to study nanostructure variations in the time-averaged trajectories of both dilute and pure IL, where pure IL refers to an ionic liquid with no or minimal added solvent, and dilute IL refers to the same ionic liquid mixed with a molecular solvent to reduce ion concentration.. The first objective is to identify inherent nanostructure variations

using persistence summaries and then examine how these variations are reflected in the original 3D point cloud in terms of the aggregation of the ions by applying spatial point processes. Finally, we quantify the statistical significance of these detected features through hypothesis testing.

5.3.1 Data representations for TDA

The dataset used in this work can be described as a collection of point clouds in $\mathbb{R}^3$, indexed by two variables p and q. We define $\mathbf{X}_{p,q} \subset \mathbb{R}^3$ to represent the point cloud at the p-th instance of the q-th category. It is assumed that each category contains n instances, and there are m categories in total. Each point in a category q is organized by occurrence yielding the sequence: $\mathbf{X}_{1,q}, \mathbf{X}_{2,q}, \cdots, \mathbf{X}_{n,q}$. For this work, different categories may be distinguished by various features such as concentration or alkyl chain length (as illustrated in Case Study I and II). A sequential arrangement of the point clouds can therefore be constructed across categories, ordered by the magnitude of these distinguishing features. This results in an overall sequence: $\mathbf{X}_{1,1}, \mathbf{X}_{2,1}, \cdots, \mathbf{X}_{n,1}, \mathbf{X}_{1,2}, \cdots, \mathbf{X}_{n,m}$. This sequence can be viewed as a pseudo-time series, where the transition from one instance to the next represents both internal dynamics within each category and a progression from simpler to more complex configurations as governed by the feature attribute (Paparoditis, 2018; Muggeo and Adelfio, 2011). In the sections that follow, when discussing point clouds more generally, the indices are omitted, and the point cloud is simply be denoted as $\mathbf{X}$. Next,

we describe the construction of topological features on these point cloud representations.

5.3.2 Change point analysis

Change point analysis (CPA) techniques seek to identify significant shifts or variations in the underlying structure of sequential data. In its basic construction, an appropriate model is proposed to represent the evolution of the sequence. The primary objective is to detect instances where a numerical quantity measured from a model exhibits a statistically significant change, where statistical significance is defined and quantified in Section 5.3.2.2. We use one CPA technique, called cumulative sum (CUSUM) of residuals applied to persistence summaries to study nanostructure variations in MD simulation trajectories.

5.3.2.1 Model of persistence summaries

The sequence of persistence summaries $y_{p,q}$ (see Section 5.3.1) can be considered as a process. Assume this process is piecewise stationary, meaning the process remains constant within distinct segments but can change abruptly between segments. The goal then is to detect the indexes where these changes occur. For the purpose of this work, and in subsequent analysis, our focus is to identify the index with the most significant shift in the sequence, that is, a single change-point location. This often involves proposing a model for the process. Consider the following proposed

model of the persistence summaries:

$$y_{p,q} = \mu_{p,q} + \varepsilon_{p,q}. \tag{5.1}$$

Here, $\mu_{p,q}$ is a fixed term, and is the expected value associated with the observation at index (p, q), and the error term $\varepsilon_{p,q}$ has mean 0 and variance σ^2. Note that normality is not assumed for the error term $\varepsilon_{p,q}$. This model is used to detect mean shifts in the MD simulation trajectory's persistence summary statistics. The process of detecting a change point reduces to testing the following null hypothesis of "no structural change": $\mu_{p,q} = \mu$ for all indexes (p, q), implying the mean observation is constant across the sequence. Under this null hypothesis, the ordinary least squares (OLS) residuals and an estimate of its variance can be obtained, respectively, as:

$$\hat{\varepsilon}_{p,q} = y_{p,q} - \frac{1}{nm} \sum_{1 \leqslant p \leqslant n, 1 \leqslant q \leqslant m} y_{p,q}, \quad \widehat{\sigma}^2 = \frac{1}{nm} \sum_{1 \leqslant p \leqslant n, 1 \leqslant q \leqslant m} \varepsilon_{p,q}^2. \tag{5.2}$$

Figure 5.4 provides a graphical illustration of the hypothesis described in this section. A popular approach to testing this hypothesis involves analyzing the cumulative sum of the residuals and rejecting the hypothesis if the fluctuations are deemed excessive. This procedure is described in the next section.

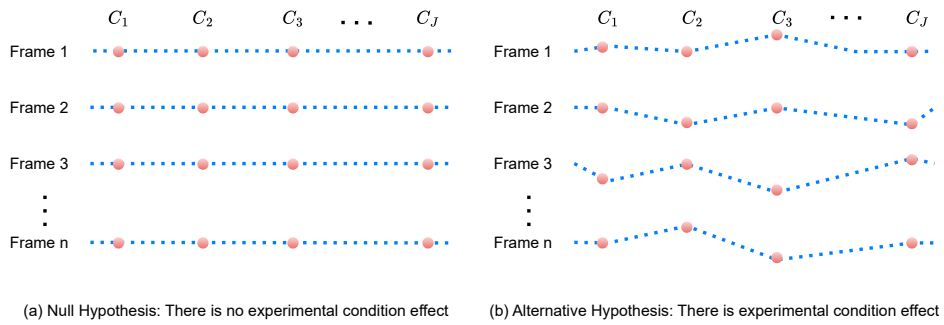


Figure 5.4: A graphical illustration of the hypothesis to be tested for the change-point analysis. The solid red points denotes persistence summaries, and the dotted blue lines indicates their trajectories. (a) The null hypothesis, indicating a persistence summary measure is the same across all experimental conditions or categories for each frame. (b) The alternative hypothesis indicating there is a difference in a persistence summary measure across the different experimental conditions or categories.

5.3.2.2 Fluctuations in residuals

Generalized fluctuation tests, a statistical framework for detecting structural changes in models, construct empirical processes that capture fluctuations in residuals. The underlying premise is that any changes in the signal ($\mathbf{y}$) are reflected in these residuals. One such empirical process proposed by Ploberger and Krämer (1992) is to compute the cumulative sum of the residuals. The CUSUM process is defined as follows:

$$\mathcal{B}^{(nm)}(c) = \frac{1}{\widehat{\sigma}\sqrt{nm}} \sum_{(p,q)\in\Omega_c} \hat{\varepsilon}_{p,q}, \quad 0 \leqslant c \leqslant 1, \tag{5.3}$$

where $\Omega_c \subset \{(p,q) : 1 \leqslant p \leqslant n, 1 \leqslant q \leqslant m\}$ contains not more than the first c proportions of the indexes. For example, suppose the set $\{p, q : 1 \leqslant$

$p \leqslant n, 1 \leqslant q \leqslant m\}$ contains a total of ten indexes, then $\Omega_{0.1}$ will contain the first pair of index values, $\Omega_{0.2}$ will contain the first two pair of index values, and so on. The assumption of equal means in the sequence is rejected if the maximum CUSUM process is sufficiently large. Therefore, a suggested test statistic is given by Ploberger and Krämer (1992):

$$\sup_{0 \leqslant c \leqslant 1} \left| \mathcal{B}^{(nm)}(c) \right|. \tag{5.4}$$

To conduct the test with this statistic, it is necessary to derive its distribution. The finite sample null distribution of this test statistics is not known. However, Sen (1982) showed that for infinitely large nm, the test statistic $\mathcal{W}^{(nm)}(c)$ converges in distribution to a standard Brownian bridge:

$$\mathcal{B}^{(nm)}(c) \xrightarrow{d} \mathcal{B}(c), \quad \mathcal{B}(c) = W(c) - cW(1). \tag{5.5}$$

Here, $W(c)$ is a real-valued continuous-time stochastic process, commonly known as the Wiener process in the literature. For more details on its characterization—such as its independent increments, which are normally distributed with variance equal to the difference in the time indices—refer to (Billingsley, 1968, p. 61-65). This limiting process (Equation (5.5)) is circular, starting at 0 when $c = 0$ and returning to 0 at $c = 1$. Consequently, there exist a point γ_0 where the maximum fluctuation in the residuals occurs. The test statistic in Equation (5.4) has the following limiting

cumulative distribution (Billingsley, 1968, p. 85, Eq. (11.39)):

$$\Theta(x) = 1 + 2\sum_{i=1}^{\infty}(-1)^i \exp(-2i^2x^2). \tag{5.6}$$

For a significance level α, for example say $\alpha = 5\%$ (the level that we use in subsequence analysis), the critical value $\Theta^{-1}(\alpha)$ obtained is 1.36. Values of Equation (5.4) exceeding this threshold indicate statistically significant fluctuations. Although this critical value produces constant linear boundaries, a more effective framework for detecting variations adopted in this work, utilizes elliptical boundaries defined by: $\Theta^{-1}(1-\alpha)\sqrt{c(1-c)}$. Equivalently, the statistical significance of the fluctuations can be determined by computing a p-value. Due to the OLS formulation and the CUSUM constructions, this change point testing procedure is called the 'OLS-CUSUM test.' The computation of the empirical fluctuation process and its boundaries as well as this hypothesis testing procedure are implemented in the R software package *strucchange* (Zeileis et al., 2002).

5.3.3 Spatial point processes applied to TDA representations

After detecting structural shifts in the point cloud sequence, we employ spatial point process techniques to analyze the point cloud configurations at the identified change boundaries, allowing us to characterize the local structural changes driving these transitions. This is motivated by the fact

that simplicial complexes are constructed by drawing spheres of a given radius around the points in the space. While the distribution of these points are reflected in the resulting persistence diagram and subsequent persistence summaries, features of this distribution can also be captured using tools from spatial point analysis, specifically, the pair correlation function (PCF). The PCF is a statistical tool used to analyze the spatial distribution of points in a point cloud. Consider a point cloud $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_{\mathcal{P}}\}$ observed within a region $\mathcal{V} \subset \mathbb{R}^3$ with volume $|\mathcal{V}|$. The intensity λ of the point process is defined as $\lambda = \mathcal{P}/|\mathcal{V}|$, representing the average number of points per unit volume. The PCF, denoted $g(r)$, quantifies the likelihood of finding a pair of points separated by a distance r relative to what would be expected under complete spatial randomness (CSR). Under CSR, the PCF is constant: $g(r) = 1$ for all r. Deviations from this baseline indicate clustering ($g(r) > 1$) or regularity ($g(r) < 1$) in the point distribution.

To estimate the PCF empirically, we consider the pairwise distances between points in $\mathbf{X}$. The empirical PCF $\widehat{g}(r)$ is computed by smoothing the observed pairwise distances using a kernel function. In this work, the Epanechnikov kernel with a fixed bandwidth $h = 0.26/\lambda^{1/3}$ (a rule-of-thumb bandwidth) is employed for smoothing. The Epanechnikov kernel is defined as:

$$K_{Epa}(x; h) = \frac{3}{4h}\left(1 - \frac{x^2}{h^2}\right)\mathbb{1}(|x| \leqslant h), \tag{5.7}$$

where $\mathbb{1}(\cdot)$ is the indicator function. The empirical PCF is then given by:

$$\widehat{g}(r) = \frac{1}{4\pi r^2 \lambda \mathcal{P}} \sum_i \sum_{j \neq i} K_{Epa}\left(\|\mathbf{x}_i - \mathbf{x}_j\| - r; h\right) e(\mathbf{x}_i, r), \tag{5.8}$$

where $e(\mathbf{x}_i, r)$ is an edge correction factor accounting for points near the boundary of $\mathcal{V}$. Specifically, $e(\mathbf{x}_i, r)$ is defined as the inverse of the fraction of the sphere centered at $\mathbf{x}_i$ with radius r that lies within $\mathcal{V}$ (Baddeley et al., 1993, 2015). This correction ensures that boundary effects do not bias the estimation of the PCF. For interpretability, the PCF is often scaled by subtracting 1, yielding a reference value of 0 under CSR. This scaled version is adopted in this work, providing a clearer baseline for identifying deviations from randomness in the spatial distribution of points. Figure 5.12 illustrates example distribution of point clouds and their corresponding PCFs. The $g(r)$ is centered at 0 by subtracting 1. The PCF of the random point cloud (solid orange line) lies close to zero, reflecting the randomness of the point distribution. For the clustered point cloud, the PCF (dotted green line) is above zero at small distances, indicating within-cluster proximity, with the location of the maximum suggesting the most frequent short inter-point distance. Conversely, the PCF of the regular point cloud (dashed blue line) initially decreases, indicating a tendency for points to be farther apart than expected randomly.

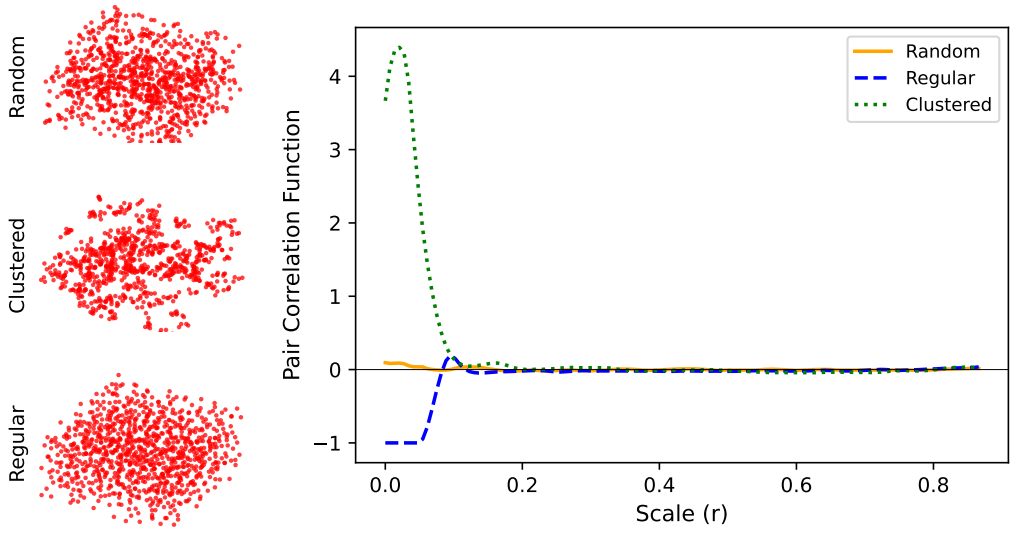


Figure 5.5: Example distributions of point clouds (1000 points each in $\mathbb{R}^3$) and their corresponding pair correlation functions. Top-left: a randomly generated point cloud following a homogeneous Poisson process, with its pair correlation function (solid orange line) near zero, indicating randomness. Middle-left: a clustered point cloud, where the pair correlation function (dotted green line) is above zero for small distances, showing within-cluster proximity. Bottom-left: points with near-uniform pairwise distances exhibit repulsion, and the pair correlation function (blue dashed line) is below zero for smaller scales.

5.4 Applications

This section demonstrates the application of the methods introduced in Section 5.3 to all-atom MD simulations of ILs. ILs are room-temperature salts with unique electrochemical properties, making them ideal candidates for battery applications. In addition, ILs have interesting spatial and structural patterns at the atomic level. We explore two case studies: 1) the effect of varying the alkyl chain length of the cation within an IL

family, and 2) the impact of varying the concentration of a specific IL in an acetonitrile solvent. For the first case study, eight distinct MD simulation trajectories are generated, while the second case study involves 13 trajectories. Each trajectory is processed by averaging the Cartesian coordinates of the IL over all time frames, providing a static representation of the system's spatial configuration for analysis. These case studies illustrate the utility of the proposed methods in capturing and quantifying nanostructural patterns in complex ionic systems.

5.4.1 Data normalization and robustness

The realizations of different ILs, observed as 3D point clouds, may differ in the number of points in this 3D space, and by extension, their densities might be different. To mitigate the impact of varying point cloud densities on computed topological features, particularly for the H_1 features, we employ a data normalization technique based on scaling the point cloud by the average first nearest neighbor (1-NN) distance of each point cloud. For a given point cloud $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_{\mathcal{P}}\}$, each data point's coordinates are scaled by dividing by the average 1-NN distance:

$$\bar{\mathbf{x}}_i = \frac{\mathbf{x}_i}{\bar{d}_{\text{1-NN}}}, \quad \bar{d}_{\text{1-NN}} = \frac{1}{\mathcal{P}} \sum_{i=1}^{\mathcal{P}} \min_{1 \leqslant j \leqslant n, i \neq j} \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbb{R}^3}, \tag{5.9}$$

where $\|\cdot\|_{\mathbb{R}^3}$ is the Euclidean distance between any two points in the space.

The scaled data that is used for all the analyses is denoted as: $\bar{\mathbf{X}} =$

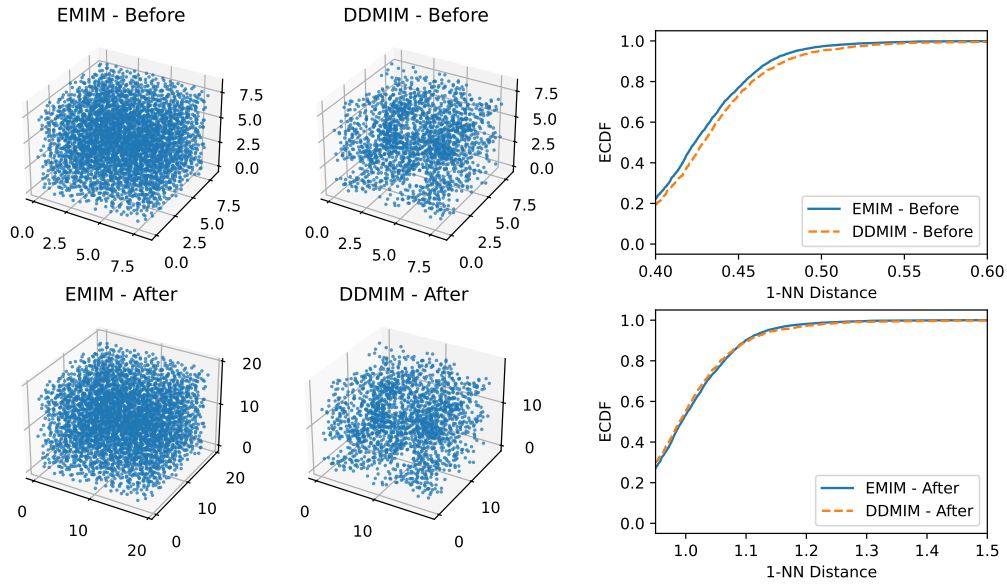


Figure 5.6: Distribution of the point cloud before and after scaling with average 1-NN distance. Top-left: unscaled point cloud data for EMIM; top-middle: unscaled point cloud data for DDMIM; top-right: empirical cumulative distribution functions (ECDFs) of the unscaled datasets. Bottom-left: scaled point cloud data for EMIM; bottom-middle: scaled point cloud data for DDMIM; bottom-right: ECDFs of the scaled datasets.

$\{\bar{\mathbf{x}}_1, \cdots, \bar{\mathbf{x}}_{\mathcal{P}}\}$. To demonstrate the effectiveness of this normalization, consider Figure 5.6 where the IL 1-ethyl-3 methylimidazolium (C_2Mim) BF_4 and 1-dodecyl-3 methylimidazolium (C_{12}Mim) BF_4 are compared. The point cloud 1-ethyl-3 methylimidazolium (C_2Mim) BF_4 have 4008 points while 1-dodecyl-3 methylimidazolium (C_{12}Mim) BF_4 have 1914 points. From Figure 5.6-right top, the comparison of their empirical cumulative distribution of the 1-NN distance is shown before the proposed scaling, indicating some difference in their distribution. However, after scaling by the proposed method the difference in their observed 1-NN distance

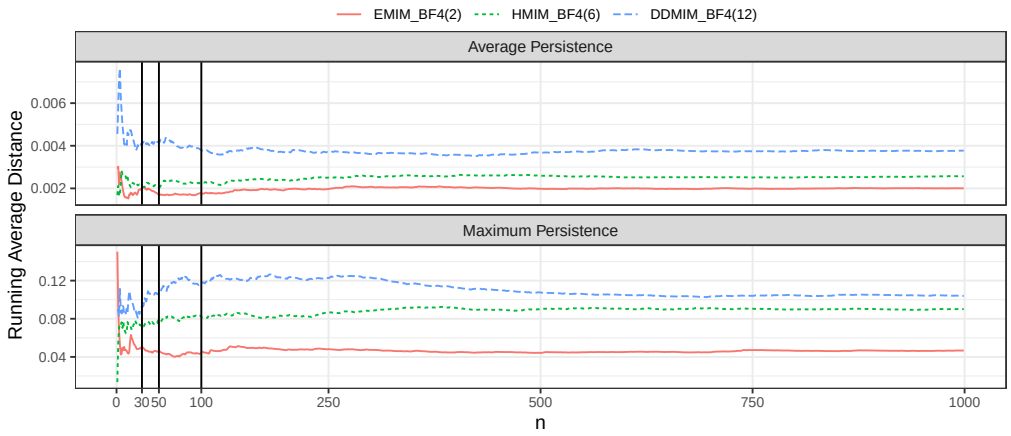


Figure 5.7: The running average distance of the average and maximum persistence summaries for three selected ILs. The n is the number of samples, where sample is defined as a frame or its point cloud representation. The differences are less wiggly after 30 samples, and more stable results can be obtained after 100 samples.

are very similar as shown in Figure 5.6-right bottom. In this particular example, the magnitude of the difference might not appear pronounced, but such disparities could significantly affect the persistence of the homology features when aggregated over a large set of observations. Since the VR filtration is constructed by forming simplices based on proximity, this normalization standardizes the scale of distances and ensures that the topological features identified are intrinsic to the time-averaged trajectories structure rather than artifacts of point density.

A final robustness check in our analysis pipeline involves determining the minimum number of samples (defined as frames or their point cloud representations) required to obtain consistent results. This is particularly relevant for the CPA, where we construct sequences as defined

in Section 5.3.1. Specifically, we aim to identify the minimum number of samples needed to reliably detect the same change point. To assess this, we computed the Running Average Distance (RAD) of the persistence summaries. The RAD at index $(p > 1, q)$ is defined as:

$$\mathrm{RAD}_{p,q} = \frac{1}{(p-1)+(q-1)\cdot n}\left(\sum_{k=1}^{p-1}\sum_{l=1}^{m}(y_{p,q}-y_{k,l})+\sum_{k=1}^{n}\sum_{l=1}^{q-1}(y_{p,q}-y_{k,l})\right).$$

Figure 5.7 illustrates the RAD results for $(C_2Mim)\ BF_4$, $(C_6Mim)\ BF_4$, and $(C_{12}Mim)\ BF_4$. Our analysis shows that the successive average differences stabilize after 30 samples, with more consistent results achieved beyond 100 samples. This stability suggests the analysis can be performed with fewer frames, reducing computational costs while maintaining accuracy.

5.4.2 Case Study I: Varying Alkyl Chain Length

This case study focuses on a well-characterized class of ILs: imidazoliums. These are defined by base cations consisting of a five-membered ring containing two nitrogen atoms. Specifically, we consider imidazolium ILs with varying alkyl chain lengths. Bulk molecular dynamics simulations were conducted for each IL using an $8 \times 8 \times 8\ \mathrm{nm}^3$ simulation box. The following cation–anion pairs were used:

i) 1-ethyl-3-methylimidazolium $(C_2Mim)BF_4 - EMIM(2)$

ii) 1-propyl-3-methylimidazolium $(C_3Mim)BF_4 - PMIM(3)$

iii) 1-butyl-3-methylimidazolium $(C_4Mim)BF_4 - BMIM(4)$

iv) 1-pentyl-3-methylimidazolium $(C_5Mim)BF_4 - PTMIM(5)$

v) 1-hexyl-3-methylimidazolium $(C_6Mim)BF_4 - HMIM(6)$

vi) 1-octyl-3-methylimidazolium $(C_8Mim)BF_4 - OMIM(8)$

vii) 1-decyl-3-methylimidazolium $(C_{10}Mim)BF_4 - DMIM(10)$

viii) 1-dodecyl-3-methylimidazolium $(C_{12}Mim)BF_4 - DDMIM(12)$

These ILs were selected for this case study because it is well established in the literature that as the alkyl chain length n in $(C_nMim)BF_4$ increases, the bulk IL transitions from a homogeneous to a heterogeneous nanostructure, typically around $n = 4$ - 6 (Wei et al., 2021). This structural transition is visually apparent in the MD simulations, as shown in Figure 5.8, where longer alkyl chains lead to increased nanoscale segregation of the nonpolar cation alkyl chains.

The data normalization described in Section 5.4.1 is then applied to each IL simulation point cloud. The proposed analysis pipeline is then applied to this normalized point cloud. First, the persistence diagrams are constructed for each point cloud and from which several persistence summary statistics, including the minimum, quartiles (25th, 50th, and 75th percentiles), average, maximum, and variance are computed. However, only the average and maximum values were utilized in the subsequent analysis. Specifically, we focus on the H_1 features, which correspond to

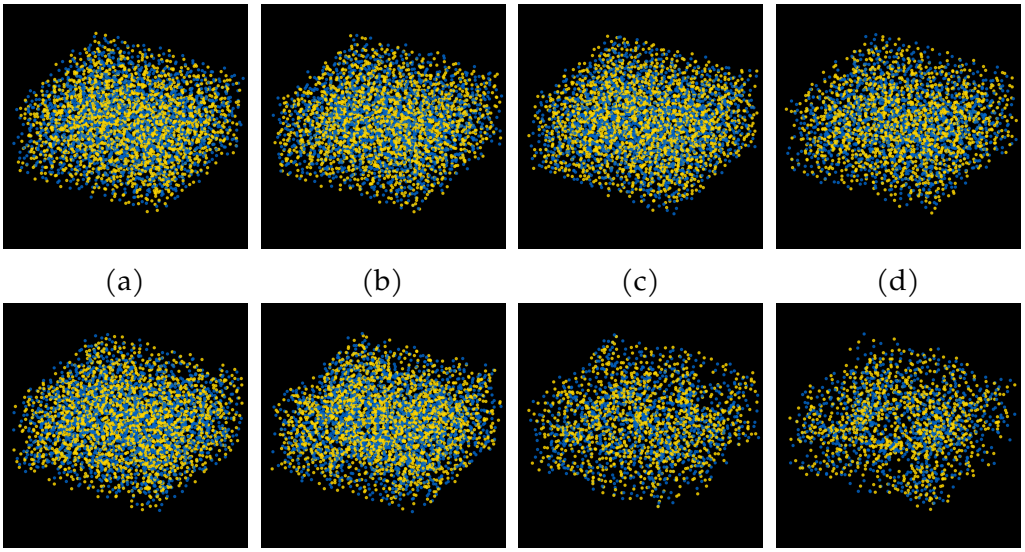


Figure 5.8: Example MD simulation point cloud data where yellow indicates cation and blue indicates anion for (a) EMIM(2) (b) PMIM(3) (c) BMIM(4) (d) PTMIM(5) (e) HMIM(6) (f) OMIM(8) (g) DMIM(10) (h) DDMIM(12).

loops formed by the arrangement of atoms in molecular dynamics IL simulations. These loops can provide insight into the structural organization and clustering behavior of ions within the liquid. The average and maximum persistences are displayed as a scatterplot in Figure 5.9. It can be seen that the average persistence is increasing with alkyl chain length up to HMIM(6), and there is a dip in the average persistence. A similar observation can also be made for the maximum persistence, with the only difference being that there is a slight jump in the maximum persistence after HMIM(6). These shifts are also apparent in the smoothed density plots in Figure 5.10. As the alkyl chain length increases, the spread of

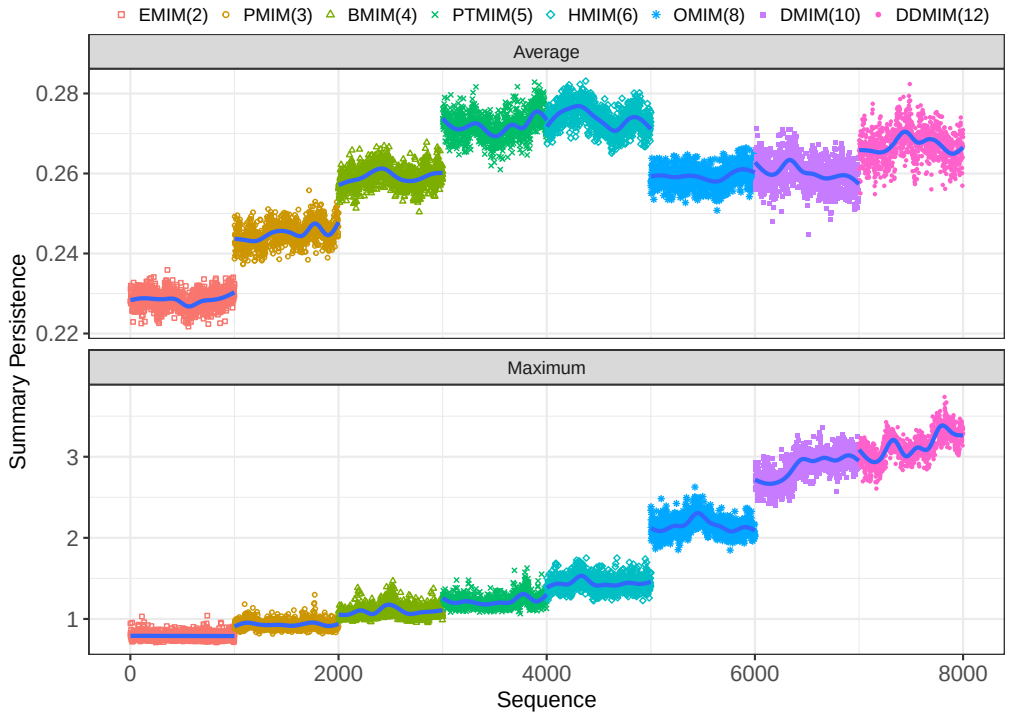
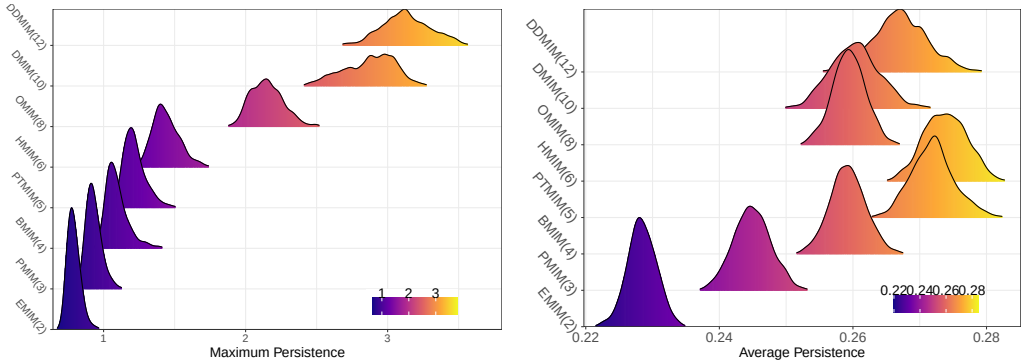


Figure 5.9: Summary of the persistence of the H_1 features. Top: The average persistence. Bottom: The maximum persistence.

the persistence summaries changes, visual difference can be observed in the groups with alkyl chain length above six and groups with alkyl chain length not exceeding six. The CPA is used to more rigorously study this observed differences between these imidazolium groups and quantify their statistical significance. The average and maximum empirical fluctuation process defined in Equation (5.3) is shown in Figure 5.11. The empirical fluctuation process (EFP) for both persistence summaries do not fall within the elliptical boundary at all sequence points, indicating statistically significant fluctuations. The point at which the two processes



(a) Density plot of the average persistence of the H_1 features (H_1).

(b) Density plot of the maximum persistence of the H_1 features.

Figure 5.10: A density plot of the distribution of the average and maximum persistence summaries.

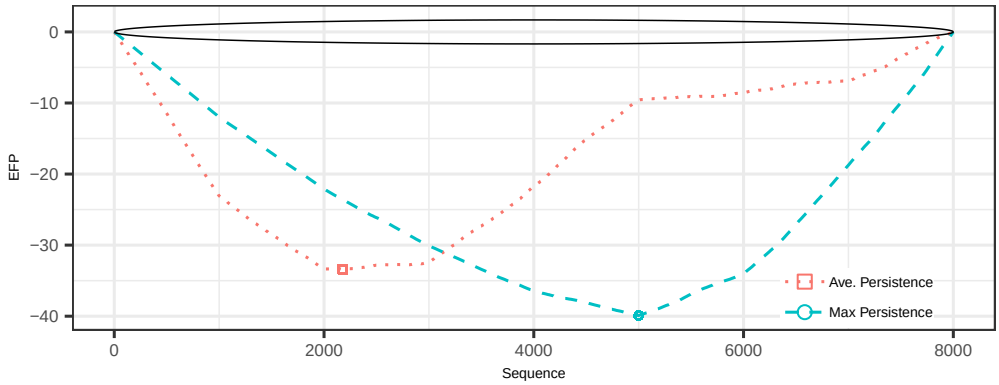


Figure 5.11: The empirical fluctuation process (EFP) identifying the location of the change in mean for the average persistence EFP (dotted red lines) and maximum persistence EFP (dasjed cyan lines). The solid black lines are the EFP boundaries. The square red point indicate the location (2179) of the change-point for the average persistence. The cyan circle indicates the change-point location (5000) for the maximum persistence.

are at their minima as well as their significance are summarized in the top-half of Table 5.1. Both change-point (CP) locations are statistically significant.

	Statistic	CP Location	Test Statistic	CP Group
C_nMim BF_4	Average	2179	33.39	BMIM(4)
	Maximum	5000	39.91	HMIM(6)
C_2Mim BF_4	Average	9967	20.54	1.650M
	Maximum	1964	25.90	0.125M

Table 5.1: The summary results of the CPA of the two data bases applied to the average and maximum processes. Both change-point (CP) locations are statistically significant for the pure IL and their CP locations are not at the boundary or do not fall within a boundary group.

To relate the CPs detected to the local variations in the original point cloud, we construct the PCFs of the various imidazolium groups. From the CPA results, we expect groups with chain-length above six to differ in local structural variations compared to those at or below chain length of six. The PCF of each MD imidazolium simulation via the single-atom representation is shown in Figure 5.12. This is similar to a traditional MD radial distribution function that looks at molecules or atoms with respect to their nearest neighbor to dictate the radius of the first solvation shell which highlights the nanostructural organization of ILs. The imidazolium groups with alkyl chain length not exceeding six exhibits different structural organization and clustering behavior compared to the groups with chain length exceeding six. In general the higher the chain length, the more pronounced the clustering pattern, and the clustering pattern is observed to be pronounced at scale between 0.05 and 0.075 for all groups. These observations are consistent with the CPA results, in that for the CPA, we observed that longer alkyl chain lengths are associated with broader

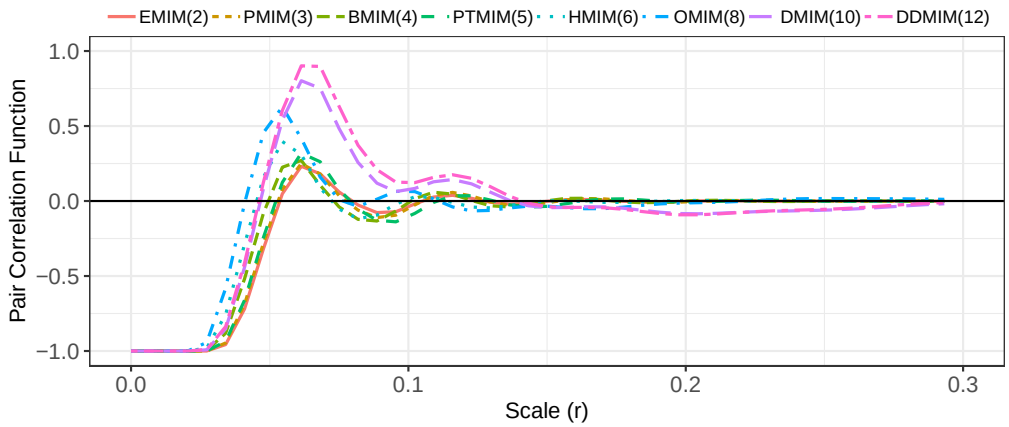


Figure 5.12: The pair correlation function averaged over 1000 frames quantifying the degree of clustering.

distribution of persistence, resulting in higher variance. This broader distribution of persistence with higher variance is indicative of patterns of aggregation. This provides a distinctive way to observe shifts in the nanostructure variations in the experimental sequence.

5.4.3 Case Study II: Varying Molar Concentration

In contrast to the first case study, which explored a range of ILs, this second case study focuses exclusively on a single compound: 1-ethyl-3-methylimidazolium tetrafluoroborate, denoted as $(C_2Mim)BF_4$. This IL has been widely studied in experimental literature due to its promising properties in energy-related applications, particularly in batteries and electrocatalysis(Liu et al., 2022). One prevailing hypothesis is that ionic nanoclustering alters the Debye screening length, potentially influenc-

ing charge transport properties (Gebbie et al., 2023). To investigate these effects from a molecular perspective and identify nanoclustering, MD simulations were carried out at thirteen different concentrations of $(C_2Mim)BF_4$ in acetonitrile solvent, ranging from highly dilute to pure IL conditions (0.025 M, 0.125 M, 0.3 M, 0.4 M, 0.5 M, 0.7 M, 0.9 M, 0.997 M, 1.0 M, 1.65 M, 3.28 M, 4.91 M, 6.5 M). The point clouds corresponding to these simulations were normalized, and persistence diagrams computed for all thirteen concentrations. Similar to Case Study II, persistence summary statistics were computed and the average and maximum persistence are shown in Figure 5.13. At low concentrations, considerable variation was observed in both the average and maximum persistence values. This variability is consistent with the sparsity of ions and the increased number of possible spatial configurations in dilute regimes. As concentration increases, the variance in persistence reduces, reflecting a more constrained ionic environment. An interesting pattern to observe is that when controlling for the variation in the persistence, the topology of the lower concentrations appears to be replicated by that of the higher concentrations, exhibiting some form of cyclical pattern in the topology. This cyclic pattern is more clearly seen in the smoothed density plots shown in Figure 5.14. This suggests that it might be possible to recover the topology of the lower concentration from that of the higher concentrations and vice versa. Further, this cyclic pattern implies that locations of significant nanostructure variations are likely to occur at the extreme groups or closer to the extreme.

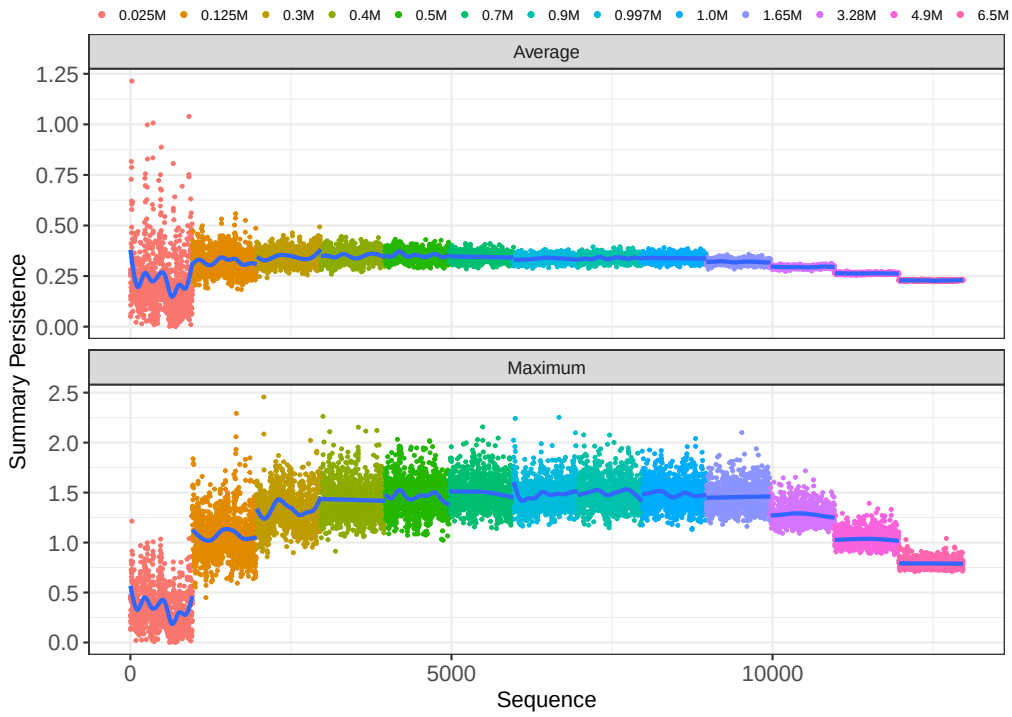
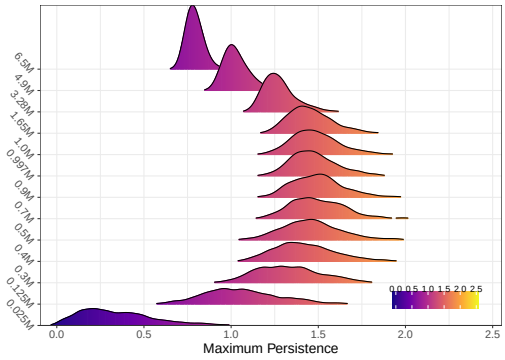
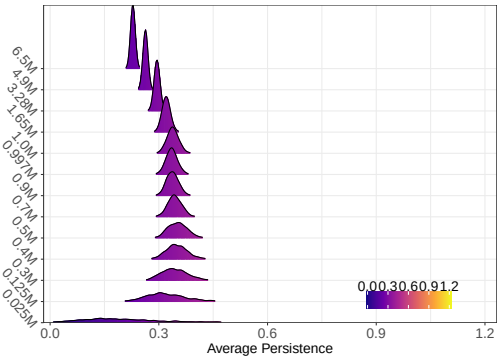


Figure 5.13: Summary of the H_1 features persistence for the C_2Mim BF_4 dataset. Top: The average persistence. Bottom: The maximum persistence.

To quantify the existence of any such locations with significant nanostructure variation, we applied the CPA. The average and maximum empirical fluctuation process is shown in Figure 5.15. The EFP for both persistence summaries attained their peaks close to the boundary of either direction. The point at which the two processes are at their peaks as well as their significance are summarized in the bottom-half of Table 5.1. Both CP locations are statistically significant for the dilute IL, but their CP locations occur close to the least and highest concentration groups. The closeness of these change-point locations to the boundaries indicates there is no



(a) Density plot of the maximum persistence of H_1.



(b) Density plot of the average persistence of H_1.

Figure 5.14: A density plot of the distribution of the average and maximum persistence summaries computed from the persistence diagrams constructed for the C_2Mim BF_4 dataset.

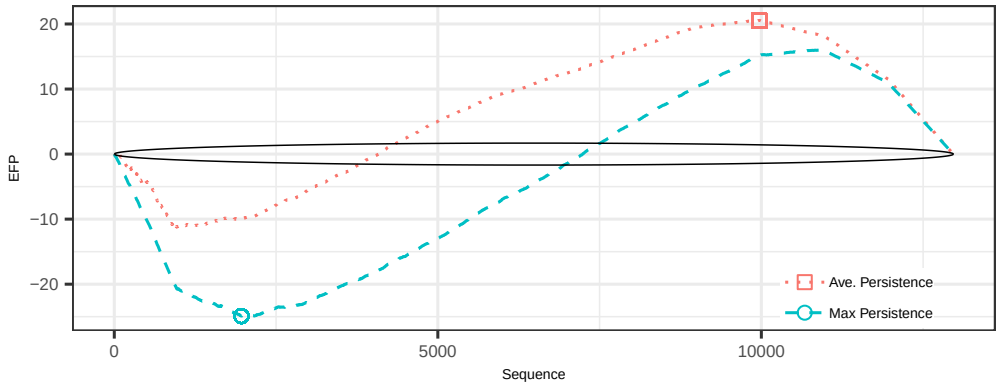


Figure 5.15: The empirical fluctuation process (EFP) identifying the location of the change in mean. The solid black ellipse is the empirical bound. The dotted red lines, and the dashed cyan lines indicates the EFP of the average persistence and maximum persistence respectively. The square red point indicate the location (9967) of the change-point for the average persistence. The cyan circle indicates the change-point location (1964) for the maximum persistence.

significant nanostructure variations that distinguishes the different concentrations. To confirm our observations, we look at the PCF computed for

the thirteen concentrations, which is shown in Figure 5.16. In general, the peaks of the PCF for all the thirteen concentrations are relatively similar, with the only difference being the scale at which they occurred. Hence

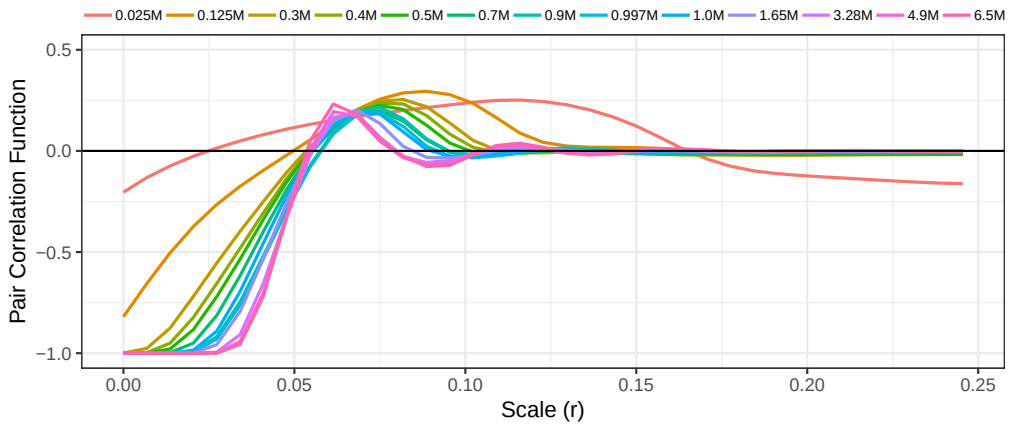


Figure 5.16: The pair correlation function averaged over 1000 frames quantifying the degree of clustering for the C_2Mim BF_4 dataset.

the proposed pipeline reveals similar nanostructure variations across the different concentrations. Specifically, clusters tend to form holes rather than fully agglomerated ion structures. The lifetimes of these holes are comparable across all thirteen concentrations, making it challenging to pinpoint a clear turning point in the non-monotonic trend for both the average and maximum persistence.

5.5 Discussion and Conclusion

This work presents a topological framework for analyzing MD simulations of ILs, with a focus on characterizing nanoscale structure and identifying

nanostructural transitions. Building on the persistent homology formalism within TDA, the proposed methodology leverages the geometric and topological structure of MD-generated point clouds to extract physically interpretable descriptors of ionic organization. Through the integration of change point detection and spatial point process modeling, the framework facilitates the detection of regime shifts and spatial aggregation patterns without relying on predefined structural assumptions.

The two case studies demonstrate how variations in molecular architecture (alkyl chain length) and composition (solute concentration) influence the emergence of topological features such as clusters and loops. In the first case, changes in persistent homology signatures revealed a gradual transition to percolated domain structures, consistent with known nanosegregation behavior (Wei et al., 2021). In the second case, the framework captured the formation and dissolution of ionic aggregates as a function of concentration, enabling the delineation of distinct structural regimes. These findings underscore the potential of TDA to complement traditional MD analysis tools by capturing global organizational patterns that may not be apparent through conventional descriptors.

Overall, the results highlight the utility of TDA as a robust and flexible framework for probing nanoscale structure in complex molecular systems. By enabling quantitative and interpretable analysis of MD simulations through a topological lens, this approach opens new avenues for understanding structure–function relationships in ILs and related mate-

rials. Future work may extend this methodology to include additional molecular components, incorporate temporal dynamics more explicitly, or explore alternative topological descriptors that capture other structural or functional aspects of interest.

6 CONCLUSION AND FUTURE DIRECTIONS

This dissertation develops robust statistical methods for persistent homology that extend the applicability of topological data analysis to complex and irregular data settings. The work contributes to three main areas: methodology for irregularly sampled time series, statistical inference for persistent features, and domain-specific application to molecular simulation data. A novel subsequence-based delay embedding is proposed for irregularly spaced time series, addressing the limitations of classical embeddings that assume uniform sampling. The method is supported by theoretical guarantees and numerical studies showing improved preservation of topological structure in the presence of noise and irregularity. To address the challenge of statistical inference on persistence features, we also introduce MaxTDA, a framework for estimating and evaluating the significance of the most persistent topological features. Standard robust approaches tend to shrink persistent features and obscure genuine signal. MaxTDA combines thresholded kernel density estimation with a sampling-based procedure to reduce bias in the presence of noise and outliers. This enables more accurate inference on maximal persistence features, allowing for improved feature extraction and hypothesis testing in applied settings. Finally, persistent homology is used to characterize nanoscale structure in ion distributions and to detect transitions between structural regimes in molecular dynamics simulations.

The methods developed in this work illustrate the importance of integrating statistical principles into the persistent homology pipeline. By addressing irregular sampling, uncertainty, and interpretability, this dissertation contributes to a more principled foundation for topological data analysis. The conclusions in Chapter 2, 3 and 4 discusses various directions for future research. These include extending the proposed embedding method to multivariate and spatially indexed time series, developing formal statistical guarantees for other persistence-based summaries, and applying robust topological methods in experimental or observational settings across scientific domains. Overall, this work highlights how robust statistical techniques can enhance the reliability and applicability of topological methods in modern data analysis.

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

Aaron, Catherine, Alejandro Cholaquidis, and Antonio Cuevas. 2017. Detection of low dimensionality and data denoising via set estimation techniques. *Electronic Journal of Statistics* 11(2):4596 – 4628.

Ali, Saad, Arslan Basharat, and Mubarak Shah. 2007. Chaotic invariants for human action recognition. In *2007 ieee 11th international conference on computer vision*, 1–8. IEEE.

Anai, Hirokazu, Frédéric Chazal, Marc Glisse, Yuichi Ike, Hiroya Inakoshi, Raphaël Tinarrage, and Yuhei Umeda. 2020. Dtm-based filtrations. In *Topological data analysis: The abel symposium 2018*. Springer.

Attouch, Hedy, Roberto Lucchetti, and Roger J-B Wets. 1991. The topology of the ρ-Hausdorff distance. *Annali di Matematica pura ed applicata* 160(1): 303–320.

Baddeley, Adrian, Ege Rubak, and Rolf Turner. 2015. *Spatial point patterns: methodology and applications with r*. CRC press.

Baddeley, AJ, RA Moyeed, CV Howard, and A Boyde. 1993. Analysis of a three-dimensional point pattern with replication. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 42(4):641–668.

Berry, Eric, Yen-Chi Chen, Jessi Cisewski-Kehe, and Brittany Terese Fasy. 2020. Functional summaries of persistence diagrams. *Journal of Applied and Computational Topology* 4(2):211–262.

Billingsley, Patrick. 1968. *Convergence of probability measures*. John Wiley & Sons.

Bobrowski, Omer, Sayan Mukherjee, and Jonathan E Taylor. 2017. Topological consistency via kernel estimation. *Bernoulli* 23(1):288 – 328.

Boker, Steven M, Stacey S Tiberio, and Robert G Moulder. 2018. Robustness of time delay embedding to sampling interval misspecification. *Continuous time modeling in the behavioral and related sciences* 239–258.

Brown, Kenneth A, and Kevin P Knudson. 2009. Nonlinear statistics of human speech data. *International Journal of Bifurcation and Chaos* 19(07): 2307–2319.

Cang, Zixuan, and Guo-Wei Wei. 2017. Topologynet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLoS computational biology* 13(7):e1005690.

Cantelli, FP. 1917. Sulla probabilità como limite della frequenza. *Atti Reale Academia Nazionale dei Lincei*.

Cao, Liangyue. 1997. Practical method for determining the minimum embedding dimension of a scalar time series. *Physica D: Nonlinear Phenomena* 110(1-2):43–50.

Carlsson, Gunnar, Tigran Ishkhanov, Vin De Silva, and Afra Zomorodian. 2008. On the local behavior of spaces of natural images. *International journal of computer vision* 76:1–12.

Casdagli, Martin, Stephen Eubank, J Doyne Farmer, and John Gibson. 1991. State space reconstruction in the presence of noise. *Physica D: Nonlinear Phenomena* 51(1-3):52–98.

Chazal, Frédéric, David Cohen-Steiner, and Quentin Mérigot. 2011. Geometric inference for probability measures. *Foundations of Computational Mathematics* 11:733–751.

Chazal, Frédéric, Vin De Silva, Marc Glisse, and Steve Oudot. 2016. *The structure and stability of persistence modules*, vol. 10. Springer.

Chazal, Frédéric, Marc Glisse, Catherine Labruère, and Bertrand Michel. 2014. Convergence rates for persistence diagram estimation in topological data analysis. In *International conference on machine learning*, 163–171. PMLR.

Chazal, Frédéric, and Bertrand Michel. 2021. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *Frontiers in artificial intelligence* 4:108.

Chung, Kai Lai, and Paul Erdös. 1952. On the application of the Borel-Cantelli lemma. *Transactions of the American Mathematical Society* 72(1): 179–186.

Cohen-Steiner, David, Herbert Edelsbrunner, and John Harer. 2005. Stability of persistence diagrams. In *Proceedings of the twenty-first annual symposium on computational geometry*, 263–271.

Cuevas, Antonio, and Ricardo Fraiman. 1997. A plug-in approach to support estimation. *The Annals of Statistics* 2300–2312.

Cuevas, Antonio, and Alberto Rodríguez-Casal. 2004. On boundary estimation. *Advances in Applied Probability* 36(2):340–354.

Dakurah, Sixtus, and Jessi Cisewski-Kehe. 2024. A subsequence approach to topological data analysis for irregularly-spaced time series. *arXiv preprint arXiv:2410.13723*.

———. 2025. Maxtda: Robust statistical inference for maximal persistence in topological data analysis. *arXiv preprint arXiv:2504.03897*.

Davis, Allen B, Jessi Cisewski, Xavier Dumusque, Debra A Fischer, and Eric B Ford. 2017. Insights on the spectral signatures of stellar activity and planets from PCA. *The Astrophysical Journal* 846(1):59.

Dempster, Angus, François Petitjean, and Geoffrey I Webb. 2020. Rocket: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery* 34(5):1454–1495.

Dempster, Angus, Daniel F Schmidt, and Geoffrey I Webb. 2021. Minirocket: A very fast (almost) deterministic transform for time series

classification. In *Proceedings of the 27th acm sigkdd conference on knowledge discovery & data mining*, 248–257.

Devroye, Luc. 1986. *Non-uniform random variate generation*. New York: Springer-Verlag.

Devroye, Luc, and Gábor Lugosi. 2001. *Combinatorial methods in density estimation*. Springer Science & Business Media.

Deyle, Ethan R, and George Sugihara. 2011. Generalized theorems for nonlinear state space reconstruction. *Plos one* 6(3):e18295.

Dumusque, X. 2016. Radial velocity fitting challenge-I. Simulating the data set including realistic stellar radial-velocity signals. *Astronomy & Astrophysics* 593:A5.

Dumusque, X, I Boisse, and NC Santos. 2014. SOAP 2.0: a tool to estimate the photometric and radial velocity variations induced by stellar spots and plages. *The Astrophysical Journal* 796(2):132.

Dumusque, Xavier. 2018. Measuring precise radial velocities on individual spectral lines-I. Validation of the method and application to mitigate stellar activity. *Astronomy & Astrophysics* 620:A47.

Dunsmuir, William, and PM Robinson. 1981. Estimation of time series models in the presence of missing data. *Journal of the American Statistical Association* 76(375):560–568.

Edelsbrunner, Herbert, and John L Harer. 2022. *Computational topology: an introduction*. American Mathematical Society.

Edelsbrunner, Herbert, David Letscher, and Afra Zomorodian. 2000. Topological persistence and simplification. In *Proceedings 41st annual symposium on foundations of computer science*, 454–463. IEEE.

El-Yaagoubi, Anass B, Moo K Chung, and Hernando Ombao. 2023. Topological data analysis for multivariate time series data. *Entropy* 25(11): 1509.

Émile Borel, M. 1909. Les probabilités dénombrables et leurs applications arithmétiques. *Rendiconti del Circolo Matematico di Palermo (1884-1940)* 27(1):247–271.

Emrani, Saba, Thanos Gentimis, and Hamid Krim. 2014. Persistent homology of delay embeddings and its application to wheeze detection. *IEEE Signal Processing Letters* 21(4):459–463.

Erickson, Jeff. 1999. Finding longest arithmetic progressions. *University of Illinois at Urbana-Champaign*.

Fasy, Brittany, Fabrizio Lecci, Larry Wasserman, et al. 2018. Robust topological inference: Distance to a measure and kernel distance. *Journal of Machine Learning Research* 18(159):1–40.

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

Federer, Herbert. 1959. Curvature measures. *Transactions of the American Mathematical Society* 93(3):418–491.

Fraser, Andrew M, and Harry L Swinney. 1986. Independent coordinates for strange attractors from mutual information. *Physical review A* 33(2): 1134.

Gebbie, Matthew A., Beichen Liu, Wenxiao Guo, Seth R. Anderson, and Samuel G. Johnstone. 2023. Linking electric double layer formation to electrocatalytic activity. *ACS Catalysis* 13(24):16222–16239. `https://doi.org/10.1021/acscatal.3c04255`.

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

Glenn, Susan, Jessi Cisewski-Kehe, Jun Zhu, and William M Bement. 2024. Confidence regions for a persistence diagram of a single image with one or more loops. *arXiv preprint arXiv:2405.01651*.

Grassberger, Peter, and Itamar Procaccia. 1983a. Characterization of strange attractors. *Physical review letters* 50(5):346.

———. 1983b. Measuring the strangeness of strange attractors. *Physica D: nonlinear phenomena* 9(1-2):189–208.

Hara, Nathan C, and Eric B Ford. 2023. Statistical methods for exoplanet detection with radial velocities. *Annual Review of Statistics and Its Application* 10:623–649.

Harvey, Andrew C, and Richard G Pierse. 1984. Estimating missing observations in economic time series. *Journal of the American statistical Association* 79(385):125–131.

Hatcher, A., Cambridge University Press, and Cornell University. Department of Mathematics. 2002. *Algebraic topology*. Algebraic Topology, Cambridge University Press.

Hénon, Michel. 2004. A two-dimensional mapping with a strange attractor. *The theory of chaotic attractors* 94–102.

Hertz, David. 1992. Simple bounds on the extreme eigenvalues of Toeplitz matrices. *IEEE transactions on information theory* 38(1):175–176.

Hollingsworth, Scott A, and Ron O Dror. 2018. Molecular dynamics simulation for all. *Neuron* 99(6):1129–1143.

Holzer, Parker H, Jessi Cisewski-Kehe, Debra Fischer, and Lily Zhao. 2021a. A Hermite-Gaussian based exoplanet radial velocity estimation method. *The Annals of Applied Statistics* 15(2):527–555.

Holzer, Parker H, Jessi Cisewski-Kehe, Lily Zhao, Eric B Ford, Christian Gilbertson, and Debra A Fischer. 2021b. A stellar activity F-statistic for exoplanet surveys (SAFE). *The Astronomical Journal* 161(6):272.

Huélamo, Nuria, P Figueira, X Bonfils, NC Santos, F Pepe, Michaël Gillon, R Azevedo, T Barman, Matilde Fernández, E Di Folco, et al. 2008. TW Hydrae: evidence of stellar spots instead of a Hot Jupiter. *Astronomy & Astrophysics* 489(2):L9–L13.

Hughes, Michael E, John B Hogenesch, and Karl Kornacker. 2010. Jtk_cycle: an efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. *Journal of biological rhythms* 25(5): 372–380.

Huke, Jeremy P, and David S Broomhead. 2007. Embedding theorems for non-uniformly sampled dynamical systems. *Nonlinearity* 20(9):2205.

HuoLiu, Qing, and Xue YuanTang. 1998. Iterative algorithm for nonuniform inverse fast Fourier transform (NU-IFFT). *Electronics Letters* 34(20): 1913–1914.

Je, Lisa, George W Huber, Reid C Van Lehn, and Victor M Zavala. 2022. On the integration of molecular dynamics, data science, and experiments for studying solvent effects on catalysis. *Current Opinion in Chemical Engineering* 36:100796.

Jiang, Haihui Joy, Rob Atkin, and Gregory G Warr. 2018. Nanostructured ionic liquids and their solutions: Recent advances and emerging challenges. *Current Opinion in Green and Sustainable Chemistry* 12:27–32.

Johnson, Bethany, and Stephan B Munch. 2022. An empirical dynamic modeling framework for missing or irregular samples. *Ecological Modelling* 468:109948.

Jones, David E, David C Stenning, Eric B Ford, Robert L Wolpert, Thomas J Loredo, Christian Gilbertson, and Xavier Dumusque. 2022. Improving exoplanet detection power: Multivariate Gaussian process models for stellar activity. *The Annals of Applied Statistics* 16(2):652–679.

Kim, H_S, R Eykholt, and JD Salas. 1999. Nonlinear dynamics, delay times, and embedding windows. *Physica D: Nonlinear Phenomena* 127(1-2): 48–60.

Lekscha, Jaqueline, and Reik V Donner. 2018. Phase space reconstruction for non-uniformly sampled noisy time series. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 28(8):085702.

Liu, Beichen, Wenxiao Guo, and Matthew A Gebbie. 2022. Tuning ionic screening to accelerate electrochemical co2 reduction in ionic liquid electrolytes. *ACS Catalysis* 12(15):9706–9716.

Lomb, Nicholas R. 1976. Least-squares frequency analysis of unequally spaced data. *Astrophysics and space science* 39:447–462.

Mayor, Michel, and Didier Queloz. 1995. A Jupiter-mass companion to a solar-type star. *Nature* 378(6555):355–359.

Middlehurst, Matthew, Patrick Schäfer, and Anthony Bagnall. 2024. Bake off redux: a review and experimental evaluation of recent time series classification algorithms. *Data Mining and Knowledge Discovery* 38(4): 1958–2031.

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

Moritz, Steffen, and Thomas Bartz-Beielstein. 2017. imputeTS: time series missing value imputation in R. *R J.* 9(1):207.

Muggeo, Vito MR, and Giada Adelfio. 2011. Efficient change point detection for genomic sequences of continuous measurements. *Bioinformatics* 27(2):161–166.

Niyogi, Partha, Stephen Smale, and Shmuel Weinberger. 2008. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry* 39:419–441.

Palaversa, Lovro, Željko Ivezić, Laurent Eyer, Domagoj Ruždjak, Davor Sudar, Mario Galin, Andrea Kroflin, Martina Mesarić, Petra Munk, Dijana Vrbanec, et al. 2013. Exploring the variable sky with LINEAR. III.

Classification of periodic light curves. *The Astronomical Journal* 146(4): 101.

Paparoditis, Efstathios. 2018. Sieve bootstrap for functional time series. *The annals of Statistics* 46(6B):3510–3538.

Perea, Jose A, Anastasia Deckard, Steve B Haase, and John Harer. 2015. Sw1pers: Sliding windows and 1-persistence scoring; discovering periodicity in gene expression time series data. *BMC bioinformatics* 16(1): 1–12.

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.

Picado, Nuno, and Paulo Eduardo Oliveira. 2020. Denoising and interior detection problems. *arXiv preprint arXiv:2010.16360*.

Pike, Jeremy A, Abdullah O Khan, Chiara Pallini, Steven G Thomas, Markus Mund, Jonas Ries, Natalie S Poulter, and Iain B Styles. 2020. Topological data analysis quantifies biological nano-structure from single molecule localization microscopy. *Bioinformatics* 36(5):1614–1621.

Ploberger, Werner, and Walter Krämer. 1992. The cusum test with ols residuals. *Econometrica: Journal of the Econometric Society* 271–285.

Rajpaul, Vinesh, Suzanne Aigrain, Michael A Osborne, Steven Reece, and S Roberts. 2015. A Gaussian process framework for modelling stel-

lar activity signals in radial velocity data. *Monthly Notices of the Royal Astronomical Society* 452(3):2269–2291.

Ruf, T. 1999. The Lomb-Scargle periodogram in biological rhythm research: analysis of incomplete and unequally spaced time-series. *Biological Rhythm Research* 30(2):178–201.

Sauer, Tim, James A Yorke, and Martin Casdagli. 1991. Embedology. *Journal of statistical Physics* 65:579–616.

Scargle, Jeffrey D. 1982. Studies in astronomical time series analysis. ii-statistical aspects of spectral analysis of unevenly spaced data. *Astrophysical Journal, Part 1, vol. 263, Dec. 15, 1982, p. 835-853.* 263:835–853.

Sen, Pranab Kumar. 1982. Invariance principles for recursive residuals. *The Annals of Statistics* 10(1):307–312.

Sesar, Branimir, J Scott Stuart, Željko Ivezić, Dylan P Morgan, Andrew C Becker, and Przemysław Woźniak. 2011. Exploring the variable sky with LINEAR. I. Photometric recalibration with the sloan digital sky survey. *The Astronomical Journal* 142(6):190.

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, Aarti, Clayton Scott, and Robert Nowak. 2009. Adaptive hausdorff estimation of density level sets. *The Annals of Statistics*.

Smith, Alexander, Spencer Runde, Alex K Chew, Atharva S Kelkar, Utkarsh Maheshwari, Reid C Van Lehn, and Victor M Zavala. 2023. Topological analysis of molecular dynamics simulations using the euler characteristic. *Journal of Chemical Theory and Computation* 19(5):1553–1567.

Sprott, Julien Clinton, and George Rowlands. 2001. Improved correlation dimension calculation. *International Journal of Bifurcation and Chaos* 11(07): 1865–1880.

Stark, Jaroslav, David S Broomhead, Michael Evan Davies, and J Huke. 1997. Takens embedding theorems for forced and stochastic systems. *Nonlinear Analysis: Theory, Methods & Applications* 30(8):5303–5314.

Takens, Floris. 2006. Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, warwick 1980: proceedings of a symposium held at the university of warwick 1979/80*, 366–381. Springer.

Tralie, Christopher J, and Jose A Perea. 2018. (Quasi) Periodicity quantification in video data, using topology. *SIAM Journal on Imaging Sciences* 11(2):1049–1077.

Tsybakov, Alexandre B. 1997. On nonparametric estimation of density level sets. *The Annals of Statistics* 25(3):948–969.

Turner, Katharine, Sayan Mukherjee, and Doug M Boyer. 2014. Persistent homology transform for modeling shapes and surfaces. *Information and Inference: A Journal of the IMA* 3(4):310–344.

Vacek, P, and T Ashikaga. 1980. An examination of the nearest neighbor rule for imputing missing values. *Proc. Statist. Computing Sect., Amer. Statist. Ass* 326–331.

VanderPlas, Jacob T. 2018. Understanding the Lomb-Scargle periodogram. *The Astrophysical Journal Supplement Series* 236(1):16.

Vietoris, Leopold. 1927. Über den höheren zusammenhang kompakter räume und eine klasse von zusammenhangstreuen abbildungen. *Mathematische Annalen* 97(1):454–472.

Walker, Theodore W, Alex K Chew, Huixiang Li, Benginur Demir, Z Conrad Zhang, George W Huber, Reid C Van Lehn, and James A Dumesic. 2018. Universal kinetic solvent effects in acid-catalyzed reactions of biomass-derived oxygenates. *Energy & Environmental Science* 11(3):617–628.

Wang, Yong-Lei, Bin Li, Sten Sarman, Francesca Mocci, Zhong-Yuan Lu, Jiayin Yuan, Aatto Laaksonen, and Michael D Fayer. 2020. Microstructural and dynamical heterogeneities in ionic liquids. *Chemical reviews* 120(13): 5798–5877.

Wei, Chunlei, Kun Jiang, Timing Fang, and Xiaomin Liu. 2021. Effects of anions and alkyl chain length of imidazolium-based ionic liquids at the au (111) surface on interfacial structure: a first-principles study. *Green Chemical Engineering* 2(4):402–411.

Xu, Xin, Jessi Cisewski-Kehe, Sheridan Beckwith Green, and Daisuke Nagai. 2019. Finding cosmic voids and filament loops using topological data analysis. *Astronomy and Computing* 27:34–52.

Zeileis, Achim, Friedrich Leisch, Kurt Hornik, and Christian Kleiber. 2002. strucchange: An r package for testing for structural change in linear regression models. *Journal of statistical software* 7:1–38.

Zhao, Lily, Debra A Fischer, Eric B Ford, Gregory W Henry, Rachael M Roettenbacher, and John M Brewer. 2020. The EXPRES stellar-signals project. I. description of data. *Research Notes of the AAS* 4(9):156.

Zhao, Lily L, Debra A Fischer, Eric B Ford, Alex Wise, Michaël Cretignier, Suzanne Aigrain, Oscar Barragan, Megan Bedell, Lars A Buchhave, João D Camacho, et al. 2022. The EXPRES stellar signals project II. state of the field in disentangling photospheric velocities. *The Astronomical Journal* 163(4):171.