이학석사 학위논문

Bump Hunting on the Torus by Persistent Homology

지속 호몰로지를 이용한 토러스 상에서의 범프 헌팅

2023년 8월

서울대학교 대학원

통계학과

유 창 조

Bump Hunting on the Torus by Persistent Homology

지도교수 정 성 규

이 논문을 이학석사 학위논문으로 제출함
2023년 4월

서울대학교 대학원
통계학과
유 창 조

유창조의 석사 학위논문을 인준함
2023년 6월

위 원 장 이 재 용 (인)

부위원장 정 성 규 (인)

위 원 박 건 웅 (인)

Abstract

Our research aims to identify density modes within the torus space where the circular data exhibits significant concentration. We employ persistent homology, primarily utilising the von Mises kernel density estimator and mixture model. To address the uncertainty inherent in the density estimator's persistent homology, we compare four methods, including a newly proposed approach in this article. Additionally, a scale-space approach is applied. Our comprehensive discussion centers around the implementation of persistent homology on the torus space, considering both theoretical foundations and practical applications.

keywords : bootstrap, mode hunting, von Mises distribution, persistent homology, topological data analysis
Student Number : 2021-26565

Contents

1 Introduction

Multivariate angular or circular data, recognised as a notable exemplification of non-Euclidean data, have been widely employed in diverse research domains encompassing medicine, biology, and physics (Mardia and Jupp, 2000; Ley and Verdebout, 2017; Marron and Dryden, 2021). It is appropriate to consider this type of data lie on a multidimensional torus. Assuming that the density function of the underlying distribution which the data are sampled, is well-defined, our objective is to identify the modes or the bumps of the data, which indicate where the data are predominantly concentrated. This problem is often referred to as *bump hunting* (Good and Gaskins, 1980; Sommerfeld et al., 2017).

Topological Data Analysis (TDA, Carlsson (2009), Edelsbrunner and Harer (2010) and Wagner et al. (2011)) provides a useful approach to address this task, with a particular focus on *persistent homology*. Persistent homology involves calculating the homology of the upper level sets of the data's density function at different levels. Homology serves as a mathematical framework that enables effective measurement of the structural properties of a given level set. However, a comprehensive understanding of this concept requires a solid grasp of various mathematical concepts and foundational knowledge. Heuristically, homology calculations capture the essential topological features inherent in the dataset. The entirety of the information encapsulated by persistent homology is often condensed and represented in a two-dimensional diagram known as *persistence diagram* (Fasy et al., 2014).

In practice, the true density function, denoted as f, is typically unknown, necessitating the use of appropriate estimators. Considering the circular nature of the data, we use two suitable density estimators for calculating persistent homology on a torus, namely the von Mises kernel density estimator (KDE) and the von Mises mixture model (Mardia and Jupp, 2000; Taylor, 2008). Therefore, it is crucial to measure the uncertainty associated with persistent homology computed using the estimators. While previous studies including Fasy et al. (2014) and Chazal et al. (2018) have introduced bootstrap methods and concentration inequalities for the uncertainty quantification on persistence diagram, but it is only for the Euclidean case. In this paper, we extend these methods to the torus space.

Furthermore, Sommerfeld et al. (2017) introduced a *scale-space approach* for performing statistical inference on persistent homology. This approach employs multiple scale parameters instead of a single parameter thus providing a which are information on the underlying true distribution. In our work, we adopt this approach, scale-space approach, utilising the concentration parameter for the von Mises KDE and the number of components in the von Mises mixture model as scale parameters. The results of bump hunting for multi-scale density estimators are summarised and visualised via a plot called *scale-lifetime diagram*.

Significance and Contribution of This Paper This article focuses on utilising density estimators for calculating and making inference on persistent homology in the context of the torus sample space. Previous studies in this area have primarily used distance functions or conventional kernel density estimators (Fasy et al., 2014; Chazal et al., 2018). However, the use of distance functions for persistent homology is susceptible to noise and outliers, while conventional kernels are unsuitable for angular data due to difficulties in identifying boundaries inherent to circular properties. On the other hand, the von Mises distribution is well-suited for circular data, and we use the von Mises distribution to define KDE and mixture models, which are then used for persistent homology calculation.

We emphasise that the approach of using mixture model for persistent homology is firstly proposed in this work. This approach comes with some additional advantages that are not found in KDE-based or distance function-based methods for persistent homology. In particular, this approach enables plotting dendrograms of components, facilitating a direct visualisation of hierarchical structure of the topological features in the density estimate. Moreover, the locations of each mode can be identified only by using the mixture model but not by the others. Such capabilities have potential applications in machine learning tasks such as clustering.

Due to the usage of density estimators instead of the true density, the measurement of uncertainty of the persistent homology is required. In this work, we establish the theoretical applicability of bootstrap methods and introduce the use of Hoeffding's and Bernstein's inequalities for inference on the significance of bumps in the torus sample

space. In Section 4, the proposed methods are applied to SARS-CoV-2 spike glycoprotein torsion angle data in $\mathbb{T}^2$ (Walls et al., 2020).

2 Definitions and Backgrounds

2.1 Torus Space

We consider the d-dimensional torus $\mathbb{T}^d = (S^1)^d$, where $S^1 = [0, 2\pi)$ represents the unit circle in $\mathbb{R}^2$, for $d \in \mathbb{N}$. Any variable $x = (x_1, \ldots, x_d) \in \mathbb{T}^d$ possesses a modulo 2π algebraic structure, that is for some $\epsilon > 0$ and some $n \in \mathbb{Z}$, x is closer to $(x - 2\pi n) - \epsilon$ than to $x + 2\epsilon$. This characteristic, known as *boundary identification*, arises from the equivalence of the two endpoints, zero and 2π. Consequently, any inferences made on the torus should accurately account for this cyclical nature. One example of this cyclical subtraction is the definition $x \ominus y := \arg(e^{i(x-y)})$ for all $x, y \in \mathbb{T}^d$.

We consider the probability measure space $(\mathbb{T}^d, \mathcal{F}, \mathbb{P})$, and assume that a continuous probability density function $f : \mathbb{T}^d \to \mathbb{R}$ is defined corresponding to $\mathbb{P}$. Additionally, we have a set of random samples $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f$.

2.2 Persistent Homology

Given a density function $f : \mathbb{T}^d \to \mathbb{R}$, We use (zero-dimensional) *topological features* corresponding to level sets of f to capture persistent homology. In particular, given a set $U \subset \mathbb{T}^d$, a zero-dimensional feature corresponds to a connected component of U. For each t over a range $[0, \infty]$, all zero-dimensional features in the *upper level set* $U_t := \{x \in \mathbb{T}^d; f(x) \geq t\}$ are recorded. There exists a natural inclusion map from U_{t_1} to U_{t_2} when $t_1 > t_2$. As the level t decreases, new topological features are captured or cease to be detected within the corresponding upper level set. These levels associated with a particular feature are referred to as its *birth* and *death* times, respectively (Edelsbrunner and Harer, 2010).

The difference between the death and birth times of each feature is termed *persistence* (or *lifetime*). We define the *persistent homology* of f as the multiset of birth-death pairs of the topological features of f (Fasy et al., 2014). To provide a summary and visualisation, we record the death-birth pairs for each level on a two-dimensional coordinate plane, where the x-axis represents the death level and the y-axis represents the birth level. This representation is known as a *persistence diagram*, denoted as $\text{dgm}(f)$ for a given density

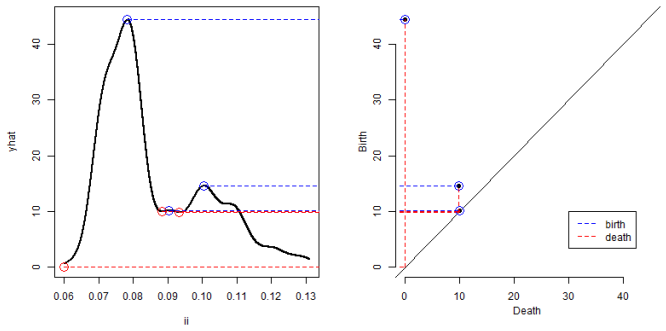

Figure 1: Illustration of persistent homology.

function f.

Given the utilisation of the density function in the calculation of persistent homology, each topological feature represents a local mode of the density function. Moreover, the birth of a feature corresponds to the peak height of the associated mode, while the death signifies the threshold at which the mode merges with another. Figure 1 provides an illustration of the concept of persistent homology. The left plot depicts a (Gaussian) kernel density estimate, while the right plot showcases the corresponding persistence diagram. The blue circles and dashed lines indicate the birth, while the reds represent the death. The diagram and plot demonstrate that this estimated density exhibits three local modes, although two of them may be spurious, either resulting from randomness or genuinely originating from the true density.

Since the true density function f is typically unknown, we rely on the estimator $\hat{f}$. Consequently, we perform statistical inference on $\mathrm{dgm}(f)$ while considering the uncertainty of $\mathrm{dgm}(\hat{f})$. This enables us to evaluate whether a local peak in $\hat{f}$ represents a genuine peak in f.

Hence, it becomes essential to define an appropriate metric for diagrams. When comparing two diagrams, the commonly used *Bottleneck distance* denoted as $W_\infty(\mathrm{dgm}(f), \mathrm{dgm}(g))$ is employed. This distance is determined by the maximum L^∞ distance between points of $\mathrm{dgm}(f)$ and $\mathrm{dgm}(g)$, with the points optimally matched in a one-to-one manner, for non-diagonal points, or one-to-diagonal ($y = x$) if non-diagonal matches are not possible. It is noteworthy that since all persistence diagrams consist of 2-dimensional points in $\mathbb{R}^2$,

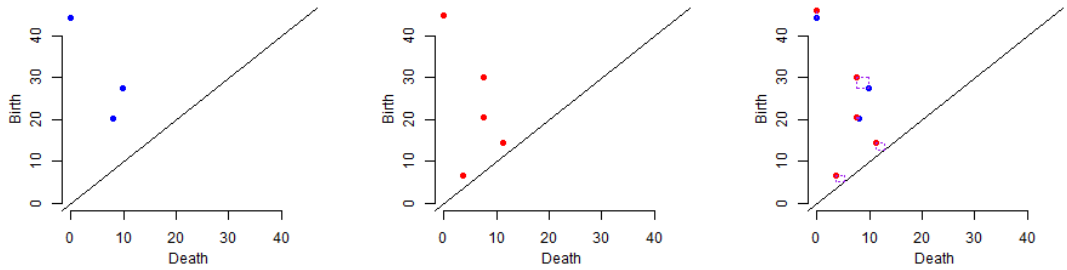

Figure 2: (Left and Middle) Two diagrams to be compared. (Right) Illustration of bottleneck matching.

the L^∞ distance corresponds to the larger value among the birth time difference and death time difference. A toy data example in Figure 2 demonstrates the definition of bottleneck matching and distance. Since the blue diagram comprises only three points, the additional two points in the red diagram are matched with the diagonal. For two density functions f and g, if $W_\infty(\mathrm{dgm}(f), \mathrm{dgm}(g)) = \delta$ holds for some $\delta > 0$, it indicates that each point in $\mathrm{dgm}(f)$ is at most δ away from a corresponding point in $\mathrm{dgm}(g)$.

2.3 Confidence Sets for Persistent Homology

To assess whether the bumps in the estimated density $\hat{f}$ are coincidental or truly representative of f, statistical inferences for the persistence diagram are conducted by constructing a confidence set. Let us fix $\alpha \in (0, 1)$ to establish a confidence level of $1-\alpha$. The confidence set for $\mathrm{dgm}(f)$ is defined as follows:

$$\mathcal{C}_n := \{\mathrm{dgm}(f); W_\infty(\mathrm{dgm}(f), \mathrm{dgm}(\hat{f})) \leq \delta_n\}$$

Here, $\delta_n \equiv \delta_n(\alpha, X_1, \ldots, X_n)$ represents the critical value dependent on the dataset and α, as used in Fasy et al. (2014) and Chazal et al. (2018). This value δ_n must satisfy the following inequality:

$$P(\mathrm{dgm}(f) \notin \mathcal{C}_n) = P(W_\infty(\mathrm{dgm}(f), \mathrm{dgm}(\hat{f})) > \delta_n) \leq \alpha. \tag{1}$$

Alternatively, we allow δ_n to represent an asymptotic critical value:

$$\limsup_{n\to\infty} P(\mathrm{dgm}(f) \notin \mathcal{C}_n) = \limsup_{n\to\infty} P(W_\infty(\mathrm{dgm}(f), \mathrm{dgm}(\hat{f})) > \delta_n) \leq \alpha. \tag{2}$$

Thus, finding the appropriate value of δ_n that tightens the inequalities (1) and (2) allows us to determine the significance of the bumps in $\hat{f}$. In other words, if p is a point in $\mathrm{dgm}(\hat{f})$, the corresponding local mode would be considered noise if the radius δ_n-L^∞-ball centered at p, $\{q \in \mathbb{R}^2; d^\infty(p,q) \leq \delta_n\}$, contains any subset of the diagonal (Fasy et al., 2014). Alternatively, a confidence band with the Euclidean width of $\sqrt{2}\delta_n$ is plotted on the diagram, and any features corresponding to points within the band are considered noise. In other words, any topological features with a persistence longer than $2\delta_n$ are deemed significant.

We note that once the value of δ_n is obtained for a given α, the remaining processes can be automated. Therefore, finding δ_n becomes one of the main objectives of this research.

Meanwhile, the *stability theorem* facilitates the estimation of δ_n in equations (1) and (2), providing computational convenience and efficiency. Recall that a smooth manofild $\mathcal{M}$ is called *triangulable* if there exists a finite simplicial complex that is homeomorphic to $\mathcal{M}$ (Cohen-Steiner et al., 2007), and a smooth function $f : \mathcal{M} \to \mathbb{R}$ is *Morse* if the Hessian matrix of f is nonsingular at every critical point (Do Carmo, 2016). Additionally, for a functions $f : \mathcal{M} \to \mathbb{R}$, $\|f\|_\infty$ denotes the L^∞ norm of the function f.

Theorem 1 (Stability theorem (Cohen-Steiner et al., 2007))**.** *Let $\mathcal{M}$ be a compact manifold that is also triangulable. If functions $f, \hat{f} : \mathcal{M} \to \mathbb{R}$ be Morse, then the persistence diagrams satisfy $W_\infty(dgm(f), dgm(\hat{f})) \leq \|f - \hat{f}\|_\infty$.*

Hence, the left-hand side of the concentration inequality (1) (and similarly for (2)) has the following upper bound:

$$P(W_\infty(\mathrm{dgm}(f), \mathrm{dgm}(\hat{f})) > \delta_n) \leq P(\|f - \hat{f}\|_\infty) > \delta_n) \leq \alpha. \tag{3}$$

It is worth noting that every compact surface is triangulable (Munkres, 2014). As $\mathbb{T}^d$ is a compact manifold, it is also triangulable. Furthermore, it is important to emphasise that in this paper, we make the assumption that the true function f is Morse, and we utilise a

Morse density estimator $\hat{f}$. Therefore our discourse shall proceed under the premise that the conditions of the stability theorem are satisfied.

3 Calculating Persistent Homology

In section 3.1, we will present two density estimation methods for calculating persistent homology, one of which is a novel approach of using mixture models, enabling the investigation of merging relationships and locations of density components by dendrogram. Moreover in section 3.2, we will discuss four approaches for uncertainty measurement of the persistence diagram. Finally, we will apply a scale-space approach to summarise multi-scale persistence diagrams and confidence sets for visualisation.

3.1 Density Estimation on the Torus

Two density estimators, the von Mises kernel density estimator (KDE) and the von Mises mixture model, for density estimation on the torus are introduced in this section.

3.1.1 Kernel Density Estimator

Previous studies on persistent homology (Fasy et al., 2014; Chazal et al., 2018) have utilised the conventional kernel density estimator $\hat{f}_h(x) := \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{\|x-X_i\|_2}{h}\right)$ with the bandwidth parameter h. However, this estimator is not suitable for the torus due to the boundary identification problem. Instead, we propose using the von Mises kernel (Mardia and Jupp, 2000), which takes the following form:

$$K_\kappa(x) := \prod_{i=1}^{d} \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos x_i},$$

where $x = (x_1, x_2, \ldots, x_d) \in \mathbb{T}^d$, and $I_\nu(\kappa)$ represents the modified Bessel function of the first kind with the order parameter ν, serving as the normalising constant for the density. It is important to note that the concentration parameter κ in the von Mises kernel plays a similar role to the bandwidth h in the conventional kernel, but it is inversely proportional to it. Therefore, as κ decreases, the shape of the kernel density estimator becomes smoother, while a larger κ leads to a more jagged shape. The von Mises kernel density estimator with concentration parameter κ is defined as

$$\hat{f}_\kappa(x) := \frac{1}{n}\sum_{i=1}^{n} K_\kappa(x \ominus X_i). \tag{4}$$

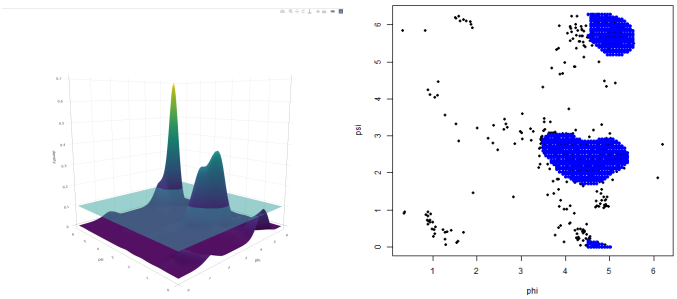

Figure 3: (Left) Illustration of kernel density estimation of 2-dimension space and (Right) Illustration of $\hat{U}_t$.

Furthermore, we define the mollified density function $f_\kappa : \mathbb{T}^d \to \mathbb{R}$ as $f_\kappa(x) = E\left(\hat{f}_\kappa(x)\right)$, which is the convolution of the true density f with the kernel. In our statistical inferences for bump hunting using the scale-space approach, we focus on inferences related to the smoothed true density f_κ, disregarding the bias between f and f_κ.

Implementations of persistent homology calculation, using a KDE, a grid approximation is typically employed. This primary approach involves the use of a cubical complex, and there are efficient implementations available for computing persistent homology using the cubical complex; See Edelsbrunner and Harer (2010) and Wagner et al. (2011). Notably, the R packages `TDA` and `rgudhi` provide various functions that encompass these efficient algorithms for computing persistent homology with a kernel density estimator. Recently, `rgudhi` begins to offer the option for boundary identification, and we use this package for computation of cubical complex for the torus. In Figure 3, we use the SARS-CoV-2 data on $\mathbb{T}^2$ to demonstrate the von Mises KDE, $\hat{f}_\kappa$ with $\kappa = 25$, and the upper level set of $\hat{f}_\kappa$ using cubical complex at level 0.1. The blue area in the right plot of Figure 3 depicts the approximated upper level set, computed using `rgudhi`. We observe that there are *two* connected components in the level set.

The birth and death of connected components given by varying levels, and the resulting persistent homology, are then summarised into the persistence diagram.

In many implementations utilising cubical complexes, the precise location of each local mode is not provided. Therefore, visual comparison between each point on the diagram and the corresponding density has been the primary means of identification of bumps.

3.1.2 Elliptical Mixture Model

In this section, we utilise a parametric model that enables us to calculate the birth, death, and location of the local modes. In particular we use a mixture model of the von Mises distributions, previously used in Hong and Jung (2022); Jung et al. (2021).

We denote the von Mises mixture model as f_J, where J represents the number of components or local modes. The multivariate von Mises distribution has the following density form:

$$f^*(y;\mu,\Lambda,\kappa) = \{T(\kappa,\Lambda)\}^{-1}\exp\left\{\kappa^\top c(y,\mu) - \frac{1}{2}s(y,\mu)^\top \Lambda s(y,\mu)\right\} \tag{5}$$

where $y,\mu \in \mathbb{T}^d, \kappa \geq 0, (\Lambda)_{ij} = \lambda_{ij}(i \neq j, \lambda_{ii} = 0)$ is a symmetric matrix with $\lambda_{ij} \in \mathbb{R}$, $c(\theta,\mu) = (\cos(\theta_i - \mu_i))_{i=1,\ldots,d}$, $s(\theta,\mu) = (\sin(\theta_i - \mu_i))_{i=1,\ldots,d}$, and $\{T(\kappa,\Lambda)\}^{-1}$ is the normalising constant (Mardia et al., 2008, 2012). We define a von Mises mixture model (vMM):

$$f_J(y) = \sum_{j=1}^{J} \pi_j f_j^*(y;\mu_j,\Lambda_j,\kappa_j) \tag{6}$$

where each f_j^* is the multivariate von Mises denisty (5), and π_j is the mixing proportion for jth group (Mardia et al., 2012).

Locally, the density of the von Mises distribution resembles that of a Gaussian density. Exploiting the fact that any level set of the Gaussian density is an ellipse, we aim to demonstrate that the upper level set of the von Mises density can also be approximated by a union of ellipses.

We define the approximated upper level set of the jth component of von Mises mixture model as follows:

$$\hat{U}_{t,j} := \left\{y \in \mathbb{T}^d; (y \ominus \mu_j)\Sigma_j^{-1}(y \ominus \mu_j) \leq 2\log\pi_j - d\log(2\pi) - \log|\Sigma_j| - 2\log t\right\}.$$

This implies that the upper level sets of the von Mises mixture model can be obtained as an approximate representation of the union of elliptical shapes. Because one can easily determine when two ellipses intersect, we can determine not only the birth and death times of each feature but also the merging dependencies among them. These are elaborated in

Appendix.

When it comes to calculating the persistent homology using the aforementioned approximated upper level sets, various technical challenges arise. Firstly, the task of fitting a mixture model often entails the utilisation of the EM (Expectation-Maximisation) algorithm. However, it is widely acknowledged that this algorithm is associated with a high computational cost. As a viable alternative, we have embraced the max-mixture approximation and complemented it with the generalised Lloyd's algorithm, as proposed in Shin et al. (2019). This approach effectively alleviates the computational burden typically encountered with the EM algorithm. Furthermore, in order to ensure consistent results across multiple trials, we have employed the *kmeans++* initialisation technique (Arthur and Vassilvitskii, 2012). These implementations enhance the robustness and stability of the outcomes obtained from the process of fitting the mixture model. Lastly, the determination of the level at which ellipses intersect necessitates the solution of a convex optimisation problem. Through the transformation of the problem of ellipse overlapping into this particular form, we are able to derive an optimal solution. A comprehensive step-by-step procedure for this conversion can be found in Proposition 1 and Remark 1 of Gilitschenski and Hanebeck (2012).

Below, Algorithm 1 summarises the computational procedures for persistent homology using the von Mises mixture model.

This methodology provides the advantage of accurately determining the location and merging relationships of each local mode. Consequently, we can visualise the hierarchical structure using dendrograms in addition to persistence diagrams. These visualisations facilitate intuitive understanding, and have practical applications in clustering.

Figure 4 illustrates the results of calculating persistent homology using a mixture model with $J = 10$ components for the SARS-CoV-2 data. The lower left panel displays the persistence diagram obtained from the fitted mixture model. The pink band represents the confidence band, indicating that only two significant bumps, coloured in green and black, are observed in this case (see section 3.2 for the confidence band). The upper right panel shows the locations of each component, and the first and third components are determined to be significant. The upper left dendrogram depicts the hierarchical

Algorithm 1 Algorithm for persistent homology of von Mises mixture

input: data $X_1, \ldots, X_n \in \mathbb{T}^d$, number of components J
(Estimate $\theta := (\theta_1, \ldots, \theta_J)$, **where** $\theta_j = (\pi_j, \mu_j, \Sigma_j)$**)**
for do$j$ in $1 : J$
Get the initial value $\hat{\theta}^{(0)}$ using k-means++.
Approximate $\hat{f}_J(y; \hat{\theta}^{(l-1)}) \approx \max_{j=1,\ldots,J} [\hat{\pi}_j \hat{f}_j^*(y; \hat{\theta}_j^{(l-1)})]$ for $l = 1, 2, \ldots$.
Implement Generalised Lloyd's Algorithm and get $\hat{\theta}^{(l)}$ until it converges.
get $\hat{f}_J(y; \hat{\theta})$ and also acquire $\hat{g}_J(y; \hat{\theta}) := \log \hat{f}_J(y; \hat{\theta})$.
end for
(Calculate births)
for do$j$ in $1 : J$
$(\text{birth})_j = \hat{f}_J(\hat{\mu}_j; \hat{\mu}_j, \hat{\Sigma}_j)$
end for
(Calculate deaths)
for do$i$ in $1 : J$
for do$j$ in $1 : J$
define a matrix $M \in \mathbb{R}^{J \times J}$ such that $M_{i,j} = \max \{t; \hat{U}_{t,i} \text{ and } \hat{U}_{t,j} \text{ meet each other at one point } \}$
$(\text{death})_j = \max_{i; 1 \le i \le J} \{M_{i,j}; (\text{birth})_i > (\text{birth})_j\}$
we say that jth component is *merged* into ith component.
end for
end for
(Calculate persistence)
Calculate $(\text{persistence})_j = (\text{birth})_j - (\text{death})_j$.
return $\{(\text{persistence})_j)\}_{j=1,\ldots,J}, \{(\text{birth})_j\}_{j=1,\ldots,J}, \{(\text{death})_j\}_{j=1,\ldots,J}, \{M\}_{i,j=1,\ldots,J}$

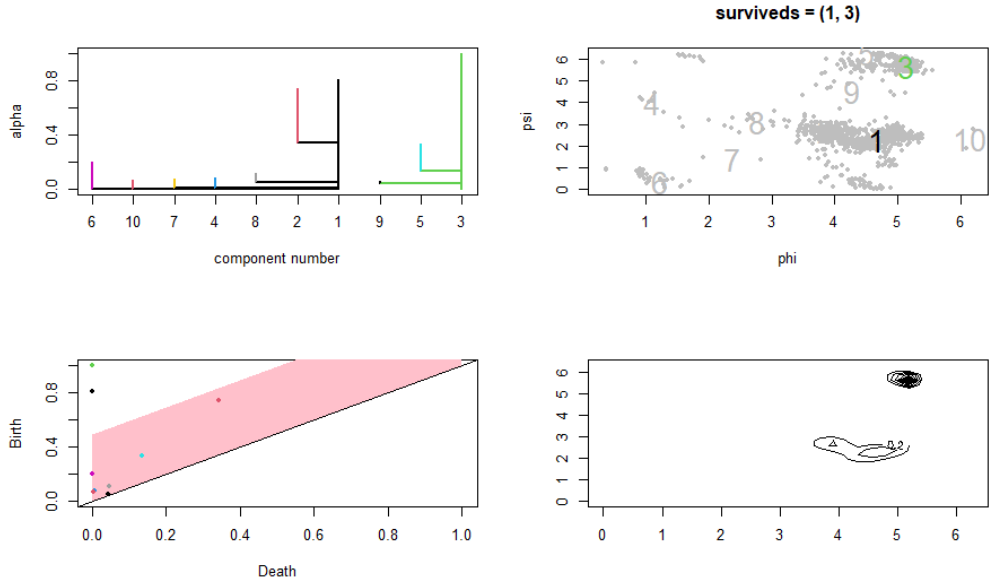

Figure 4: (Top Left) Dendrogram showing the persistence of local modes. (Top Right) Data and estimated locations of local modes. Colored ones are classified as significant, others as noise. (Bottom Left) Persistence diagram and confidence band. (Bottom Right) Contour plot of von Mises kernel density estimator.

merging relationships among the local modes in the fitted density. Each leaf corresponds to a component of the mixture model, representing a local mode. The vertical length of each branch segment indicates the lifetime* of the corresponding mode, with the upper and lower endpoints denoting the birth and death, respectively. Following the approach in Kim et al. (2017), we applied a hierarchical structure to represent the components, where horizontal lines indicate the merging of one component into another, with different colors representing different merging events. The x-axis represents the indices for the components, while the y-axis represents the empirical quantiles of the fitted density.

3.2 Uncertainty Measurement of Persistence Diagram

In this section, we present a methodology for quantifying the uncertainty associated with the estimated persistent homology. The primary objective is to determine the critical value δ_n in equations (1) and (2). To facilitate this analysis, we will employ the von Mises

*Note that the empirical quantile of the log density, denoted $\alpha \in [0, 1]$, is utilised instead of the density height, due to the practical computational challenges.

kernel density estimator and von Mises mixture model as density estimators for f.

3.2.1 Bootstrap Method

The inequality (3) given by stability theorem implies that the estimation of δ_n in (1) or (2) can be achieved by only using $\hat{f}_\kappa$ (but not involving computationally heavy construction of $\text{dgm}(\hat{f}_\kappa)$) for constructing the confidence set. One of the simplest methods for estimating δ_n is through *bootstrap*.

Recall that we have a random sample $X_1, \ldots, X_n \overset{iid}{\sim} f$ with its corresponding probability measure $\mathbb{P}$, and $\hat{f}_\kappa$ represents the von Mises KDE (4) with a given $\kappa \in (0, \infty)$. Note that $f_\kappa(x) := E(\hat{f}_\kappa(x))$ for all $x \in \mathbb{T}^d$ so that f_κ is a smoothed representation of the true density function f. In other words, we can represent that $\hat{f}_\kappa(x) = \int K_\kappa(u \ominus x)\mathbb{P}_n(dx)$ and $f_\kappa(x) = \int K_\kappa(u \ominus x)\mathbb{P}(dx)$, where $\mathbb{P}_n$ is the empirical measure corresponding to the data $X_1, \ldots, X_n$. Initially, we generate bootstrap samples $X_1^*, \ldots, X_n^*$, and compute the corresponding estimated $\hat{f}_\kappa^*$ along with its L^∞ norm, denoted as $\|\hat{f}_\kappa^* - \hat{f}_\kappa\|_\infty$. We repeat this process a sufficiently large number of times, denoted as B, to obtain the empirical bootstrap distribution of $\|\hat{f}_\kappa^* - \hat{f}_\kappa\|_\infty$. Finally, we define $\hat{\delta}_n$ as the $1-\alpha$ upper quantile of this empirical bootstrap distribution.

Theorem 2. *For a given $\kappa \in (0, \infty)$, let $\mathcal{G}$ denote the family of all functions $K_\kappa(u \ominus x)$ mapping from $x \in \mathbb{T}^d$ to $K_\kappa(u \ominus x) \in \mathbb{R}$, indexed by all possible $u \in \mathbb{T}^d$. Additionally, let $\hat{\delta}_n$ to be the bootstrap quantile defined above. Then $\mathcal{G}$ is indeed a $\mathbb{P}$-Donsker class. Consequently, as $n \to \infty$, we have*

$$\mathbb{P}\left(\|\hat{f}_\kappa - f_\kappa\|_\infty > \hat{\delta}_n\right) = \alpha + O\left(\sqrt{\frac{1}{n}}\right). \tag{7}$$

Proof. We utilise Theorem 2.3 in Kosorok (2008) and Theorem 19.5 in van der Vaart (1998) to establish that in order to demonstrate the $\mathbb{P}$-Donsker property of $\mathcal{G}$, it suffices to prove that its bracketing integral, denoted as

$$J_{[]}(1, \mathcal{G}, L^2(\mathbb{P})) = \int_0^1 \sqrt{\log N_{[]}\left(\varepsilon, \mathcal{G}, L^2(\mathbb{P})\right)} d\varepsilon$$

is finite. Firstly, we show that $K_\kappa(\cdot \ominus x)$ is a Lipschitz function for any given $x \in \mathbb{T}^d$. Fix x in $\mathbb{T}^d$, then for every $u, v \in \mathbb{T}^d$, it holds that $|K_\kappa(u \ominus x) - K_\kappa(v \ominus x)| \leq \|K'_\kappa(w \ominus x)\| \, \|u \ominus v\|$ for some w lying between u and v. However, the lth element of $K'_\kappa(u) := \frac{\partial}{\partial u} K_\kappa(u)$ is $\frac{1}{(2\pi I_0(\kappa))^d} \kappa(-\sin u_l) \prod_{j=1}^d e^{\kappa \cos u_j}$, where u_l is the lth element of u. Consequently, irrespective of the specific values of $w, x \in \mathbb{T}^d$, the quantity $\|K'_\kappa(w \ominus \cdot)\|$ is upper-bounded by a constant denoted as M. Thus, for all $x, u, v \in \mathbb{T}^d$,

$$|K_\kappa(u \ominus x) - K_\kappa(v \ominus x)| \leq M \|u \ominus v\|. \tag{8}$$

We now demonstrate that the bracketing number of $\mathcal{G}$ is bounded above by the covering number of $\mathbb{T}^d$. Let an $\epsilon \in \left(0, \frac{1}{M}\right)$ be fixed. Since $\mathbb{T}^d$ is a compact set, we can choose N such that $\{c_1, c_2, \ldots, c_N\} \subseteq \mathbb{T}^d$ and $\cup_{i=1}^N B_{\epsilon/2}(c_i) \supseteq \mathbb{T}^d$, where $B_{\epsilon/2}(c_i)$ represents the ball of radius $\epsilon/2$ centered at c_i. For $i = 1, 2, \ldots, N$, define functions a_i, b_i from $\mathbb{T}^d$ to $\mathbb{R}$, given by $a_i(x) := K_\kappa(c_i \ominus x) - M\epsilon/2$ and $b_i(x) := K_\kappa(c_i \ominus x) + M\epsilon/2$, respectively. We will demonstrate that the set $[a_i, b_i]_{i=1,\ldots,N}$ covers $\mathcal{G}$. In other words, for each $u \in \mathbb{T}^d$, there exists an index $i \in \{1, \ldots, N\}$ such that $a_i(x) \leq K_\kappa(u \ominus x) \leq b_i(x)$ for all $x \in \mathbb{T}^d$. This can be easily established by (8) and observing that there exists an i such that c_i is within a distance of $\epsilon/2$ from u. Consequently, it holds that $N \geq N_{[]}(\epsilon M, \mathcal{G}, L^2(\mathbb{P}))$. Moreover, if we define $A := \max\{2, \frac{\epsilon}{2\pi M}\} + 1$, then the $\epsilon/2$-covering number of $\mathbb{T}^d$ can be bounded from above by the ϵ/A-covering number of $\mathbb{T}^d$, which, in turn, is bounded by $\left(\frac{2\pi}{\epsilon/A}\right)^d$. This leads us to deduce that $N_{[]}(\epsilon M, \mathcal{G}, L^2(\mathbb{P})) \leq \left(\frac{2A\pi}{\epsilon}\right)^d$. Upon substituting ϵ/M for ϵ, consequently, for all $\epsilon \in (0, 1)$, we obtain

$$N_{[]}(\epsilon, \mathcal{G}, L^2(\mathbb{P})) \leq \left(\frac{2\pi AM}{\epsilon}\right)^d.$$

Therefore,

$$\begin{aligned} J_{[]}(1, \mathcal{G}, L^2(\mathbb{P})) &\leq \int_0^1 \sqrt{d \log(\frac{2\pi AM}{\epsilon})} d\epsilon \quad = \int_\infty^{d \log 2\pi AM} \sqrt{w} 2\pi AM \left(-\frac{1}{d}\right) e^{-\frac{1}{d}w} dw \\ &\leq \int_0^\infty \sqrt{w} \frac{2\pi AM}{d} e^{-\frac{1}{d}w} dw \quad = \pi^{3/2} A \sqrt{d} \quad < \infty \end{aligned}$$

where the variable transformation $w = d \log(2\pi AM/\epsilon)$ is employed. □

Theorem 2 and the stability inequality (3) together imply that the value of $\hat{\delta}_n$ satisfies the following inequality:

$$\limsup_{n\to\infty} \mathbb{P}\left(W_\infty\left(\mathrm{dgm}(\hat{f}_\kappa), \mathrm{dgm}(f_\kappa)\right) > \hat{\delta}_n\right) \leq \alpha.$$

We emphasise that $\hat{\delta}_n$ used in Theorem 2 obviates the need for calculating the persistence diagrams and bottleneck distances. Furthermore, the computation of $\|\hat{f}_\kappa^* - \hat{f}_\kappa\|_\infty$ is conducted using the grid method, which significantly reduces the overall calculation time.

However, it is important to note that this method may not yield a tight bound due to potentially loose stability inequality (3). In this case, the resulting confidence set can be conservative, potentially leading to the misclassification of significant modes as noise. To overcome this limitation and to achieve more accurate inferences about the persistent homology, the bootstrap method utilising the Bottleneck distance can be employed, provided that certain conditions are met.

Let $X_1^*, \ldots, X_n^*$ denote the bootstrap samples, $\hat{f}_\kappa^*$ represent the kernel density estimator based on the bootstrap samples, and $\mathrm{dgm}(\hat{f}_\kappa^*)$ indicate the corresponding persistence diagram. By employing the percentile bootstrap method, we obtain the empirical bootstrap distribution of $W_\infty(\mathrm{dgm}(\hat{f}_\kappa^*), \mathrm{dgm}(\hat{f}_\kappa))$ through B iterations. We define $\hat{\delta}_n'$ as the $1-\alpha$ quantile of this empirical bootstrap distribution. The following lemma provides conditions for validity of this bootstrap procedure.

Lemma 1. *Let $\hat{f} : \mathbb{T}^d \to \mathbb{R}$ be a density estimator, $E(\hat{f})$ represent its expectation, and $\hat{\delta}_n'$ be the $1-\alpha$ quantile of the empirical bootstrap distribution mentioned above. We assume the following conditions:*

(i) *$E(\hat{f})$ is Morse.*

(ii) *The first and second derivatives of $E(\hat{f})$ are each uniformly bounded and continuous.*

(iii) *$E(\hat{f})$ has finitely many critical points.*

(iv) $\liminf_{n\to\infty} P\left(\sup_x \left\|\hat{f}^{(i)}(x) - E(\hat{f})^{(i)}(x)\right\| < \epsilon \quad \text{for } i = 0, 1, 2\right) = 1.$

Then, we have

$$\mathbb{P}\left(W_\infty\left(dgm(\hat{f}), dgm(E(\hat{f}))\right) > \hat{\delta}'_n\right) \leq \alpha + O_{\mathbb{P}}\left(\frac{\log n}{\sqrt{n}}\right). \tag{9}$$

Lemma 1 can be verified by the same argument used for showing a similar result for the standard KDE in section 6 of Chazal et al. (2018).

We note that whether the conditions (i) and (iii) of Lemma 1 satisfy critically depends on the true density f. Below we show that conditions (ii) and (iv) of Lemma 1 are satisfied by the von Mises KDE.

Theorem 3. *The von Mises kernel density estimator $\hat{f}_\kappa$ satisfies the conditions (ii) and (iv) of Lemma 1. Moreover, if the true density f is such that the conditions (i) and (iii) of Lemma 1 are satisfied, then the Bottleneck bootstrap method using the von Mises kernel density estimator is valid.*

Proof. (ii) In order to establish the condition (ii), namely $E_X(\hat{f}_\kappa(u)) = E_X\left(\frac{1}{n}\sum_{i=1}^n K_\kappa(u \ominus X)\right)$ has two uniformly bounded continuous derivatives, note that the lth element of $K'_\kappa(u) := \frac{\partial}{\partial u}K_\kappa(u)$ is $\frac{1}{(2\pi I_0(\kappa))^d}\kappa(-\sin u_l)\prod_{j=1}^d e^{\kappa\cos u_j}$, where u_l is the lth element of u. Consequently, it becomes evident that $E_X(\hat{f}\kappa'(u))$ is bounded both above and below for all values of $u \in \mathbb{T}^d$ (as seen in the proof of Theorem 2). The same holds for the second derivatives due to the presence of trigonometric terms in $K'\kappa(u)$.

(iv) Our proof for $\hat{f}_\kappa$ satisfying condition (iv) closely follows the proofs of Theorems 3.1.6 and 3.1.7 of Rao (1983), in which the domain of the density estimator was $\mathbb{R}^d$.

Let $M := \sup_{u\in\mathbb{T}^d}\|K'_\kappa(u)\|$. Given the earlier observation that each element of $K'_\kappa(u)$ is uniformly bounded, it follows that M can be identified as an absolute constant of finite value. Fix $\epsilon > 0$. Since $\mathbb{T}^d$ is compact, we can choose $\tilde{N}$ such that $\{c_1, \ldots, c_{\tilde{N}}\} \subseteq \mathbb{T}^d$ and $\cup_{j=1}^{\tilde{N}} B_{\epsilon/4M}(c_j) \supseteq \mathbb{T}^d$. Now,

$$
\begin{aligned}
P(\sup_{u\in\mathbb{T}^d}|\hat{f}_\kappa(u)-f_\kappa(u)|\geq\epsilon) &\leq \sum_{j=1}^{\tilde{N}} P\left(\sup_{u\in B_{\epsilon/4M}(c_j)}|\hat{f}_\kappa(u)-f_\kappa(u)|\geq\epsilon\right)\\
&\leq \tilde{N}\max_{j=1\ldots,\tilde{N}} P\left(\sup_{u\in B_{\epsilon/4M}(c_j)}|\hat{f}_\kappa(u)-f_\kappa(u)|\geq\epsilon\right)\\
&\leq \tilde{N}\max_{j=1\ldots,\tilde{N}}\Bigg\{P\left(\sup_{u\in B_{\epsilon/4M}(c_j)}|\hat{f}_\kappa(u)-\hat{f}_\kappa(c_j)|\geq\epsilon/3\right)+\\
&\qquad P\left(\sup_{u\in B_{\epsilon/4M}(c_j)}|f_\kappa(u)-f_\kappa(c_j)|\geq\epsilon/3\right)+\\
&\qquad P\left(|\hat{f}_\kappa(c_j)-f_\kappa(c_j)|\geq\epsilon/3\right)\Bigg\}\\
&=: \tilde{N}\max_{j=1\ldots,\tilde{N}}\left\{\mathcal{P}_j^1+\mathcal{P}_j^2+\mathcal{P}_j^3\right\}
\end{aligned}
$$

To bound the three probabilities in the last part of above inequality, we start with the first one, $\mathcal{P}_j^1$. It is evident that $\sup_{u\in B_{\epsilon/4M}(c_j)}|\hat{f}_\kappa(u)-\hat{f}_\kappa(c_j)|\leq\sup M\|u-c_j\|\leq M\cdot\epsilon/4M=\epsilon/4<\epsilon/3$ for all $j=1,\ldots,\tilde{N}$ almost surely. Likewise, we have $\sup_{u\in B_{\epsilon/4M}(c_j)}|f_\kappa(u)-f_\kappa(c_j)|=\sup|E\hat{f}_\kappa(u)-E\hat{f}_\kappa(c_j)|\leq E\sup|\hat{f}_\kappa(u)-\hat{f}_\kappa(c_j)|\leq\epsilon/4<\epsilon/3$ almost surely. Therefore, $\mathcal{P}_j^1=\mathcal{P}_j^2=0$ for all $j=1,\ldots,\tilde{N}$.

Lastly, the representation of $\hat{f}_\kappa$ is given by $\hat{f}_\kappa(x)=\sum_{i=1}^n Y_i$, where

$$
Y_i=\frac{1}{n}\frac{1}{(2\pi I_0(\kappa))^d}\prod_{j=1}^d e^{\kappa\cos(x_j\ominus X_{ij})}
$$

for $i=1,2,\ldots,n$. Note each Y_i's are independent random variables, and are bounded between $(n(2\pi I_0(\kappa))^d)^{-1}e^{-d\kappa}$ and $(n(2\pi I_0(\kappa))^d)^{-1}e^{d\kappa}$ almost surely. To further bound $\mathcal{P}_j^3$, we employ Hoeffding's inequality (Theorem 2.2.6 in Vershynin (2018)).

$$
\mathcal{P}_j^3=P\left(|\hat{f}_\kappa(c_j)-f_\kappa(c_j)|\geq\epsilon/3\right)\leq 2\exp\left[-\frac{2n\epsilon^2}{C}\right]
$$

where C is an absolute constant which does not depend on any j. Therefore, considering all the above,

$$
P(\sup_{u\in\mathbb{T}^d}|\hat{f}_\kappa(u)-f_\kappa(u)|\geq\epsilon)\leq 2\tilde{N}\exp\left[-\frac{2n\epsilon^2}{C}\right]
$$

Hence $\limsup_{n\to\infty} P(\sup_{u\in\mathbb{T}^d} |\hat{f}_\kappa(u) - f_\kappa(u)| \geq \epsilon) = 0$.

The first and second derivatives of the von Mises kernel density estimator share a similar structure, as they are comprised solely of constants, trigonometric functions, and exponentials of trigonometric functions. Consequently, these derivatives are also bounded both from below and above. Thus, we can employ the same arguments and procedures outlined above to establish the conclusion of this proof. □

In Theorem 3, we require the mollified density $f_\kappa = E(\hat{f}_\kappa)$ to satisfy the conditions (i) and (iii) of Lemma 1, which is less stringent than for f to satisfy these conditions. If f satisfies these conditions, then for sufficiently large κ, the difference between f_κ and f becomes almost negligible (Taylor, 2008), in which case f_κ also satisfies the conditions.

When the mixture model in Section 3.1.2 is used to compute persistent homology, we also use the two bootstrap procedures discussed earlier (9) and (7). However, it has been challenging to guarantee the success of bootstrap procedures for the estimators from the von Mises mixture model, as similarly done in Theorems 2 and 3 for the von Mises KDE.

3.2.2 Finite Sample Method

While the bootstrap method is straightforward to implement, it suffers from a notable drawback of being computationally intensive, especially when it involves computing the bottleneck distance. In this section, we propose to use concentration inequalities along with linear density approximations of density functions to approximate the critical value δ_n. This approach effectively mitigates the computational burden such as those encountered in the bootstrap method.

Theorem 4. *Consider N as the number of grid points used for evaluating the von Mises kernel density estimator (4), wherein N takes the form $N := m^d$ for some integer value $m \in \mathbb{Z}$. For each $l = 1, \ldots, N$ let $g_l = (g_{l1}, \ldots, g_{ld})$ be the middle point of the grid of $\mathbb{T}^d$, let $\hat{f}_\kappa$ for a given κ be a von Mises kernel density estimate computed from $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f$. we have n random samples. Also, let $\hat{f}_\kappa^\dagger$ and $f_\kappa^\dagger$ be each piecewise linear approximation of the von Mises kde and its expectation, respectively. Then, for any $\delta > 0$, the following holds.*

(1) (Hoeffding's inequality)

$$P(W_\infty(dgm(\hat{f}_\kappa^\dagger), dgm(f_\kappa^\dagger)) > \delta) \leq 2N \exp\left[-\frac{2n\delta^2(2\pi I_0(\kappa)^{2d})}{(e^{d\kappa} - e^{-d\kappa})^2}\right] \tag{10}$$

In particular, for given $\alpha \in (0,1)$*, by letting* $\delta = \hat{\delta}_n$ *such that*

$$\hat{\delta}_n = \left[\log\left(\frac{2N}{\alpha}\right)\frac{(e^{d\kappa} - e^{-d\kappa})^2}{2n(2\pi I_0(\kappa))^{2d}}\right]^{1/2}, \tag{11}$$

the inequality (1) holds.

(2) (Bernstein's inequality)

$$P(W_\infty(dgm(\hat{f}_\kappa^\dagger), dgm(f_\kappa^\dagger)) > \delta) \leq 2\sum_{l=1}^{N} \exp\left[-\frac{\delta^2/2}{n\sigma_l^2 + M\delta/3}\right] \tag{12}$$

where $M := 2(n(2\pi I_0(\kappa))^d)^{-1}e^{d\kappa}$ *and* $\sigma_l^2 = n\,Var(\frac{1}{n}\frac{1}{(2\pi I_0(\kappa))^d}\prod_{j=1}^{d} e^{\kappa\cos(g_{lj}\ominus X_{1j})})$ *for* $l = 1, \ldots, N$.

Proof of Theorem 4. The argument in this proof follows the proof of Lemma 9 in Fasy et al. (2014), in which a result similar to part 1 of Theorem 4 for usual Euclidean data is given.

We use the fact that $\hat{f}_\kappa(g_l)$ is represented by $\hat{f}_\kappa(g_l) = \sum_{i=1}^{n} Y_{il}$, where

$$Y_{il} = \frac{1}{n}\frac{1}{(2\pi I_0(\kappa))^d}\prod_{j=1}^{d} e^{\kappa\cos(g_{lj}\ominus X_{ij})}$$

for $i = 1, 2, \ldots, n$ and $l = 1, \ldots, N$. Note that each Y_i's are independent random variables, and are bounded between $(n(2\pi I_0(\kappa))^d)^{-1}e^{-d\kappa}$ and $(n(2\pi I_0(\kappa))^d)^{-1}e^{d\kappa}$ almost surely.

Hence, use Hoeffding's inequality,

$$
\begin{aligned}
P(W_\infty(\mathrm{dgm}(\hat{f}_\kappa^\dagger), \mathrm{dgm}(f_\kappa^\dagger)) > \delta) &\leq P(\|\hat{f}_\kappa^\dagger - f_\kappa^\dagger\|_\infty) > \delta) \\
&\leq P(\max_{x \in \{g_1, \ldots, g_N\}} |\hat{f}_\kappa^\dagger(x) - f_\kappa^\dagger(x)| > \delta) \\
&\leq \sum_{l=1}^{N} \mathbb{P}(|\hat{f}_\kappa^\dagger(g_l) - f_\kappa^\dagger(g_l)| > \delta) \\
&\leq 2 \sum_{l=1}^{N} \exp\left[- \frac{2\delta^2}{\sum_{i=1}^{n} \left[\frac{1}{n(2\pi I_0(\kappa))^d} (e^{d\kappa} - e^{-d\kappa}) \right]^2} \right] \\
&= 2N \exp\left[- \frac{2n\delta^2 (2\pi I_0(\kappa))^{2d}}{(e^{d\kappa} - e^{-d\kappa})^2} \right],
\end{aligned}
$$

thus verifying (10). Equation (11) is given by the unique solution of the equation

$$
2N \exp\left[- \frac{2n\delta^2 (2\pi I_0(\kappa))^{2d}}{(e^{d\kappa} - e^{-d\kappa})^2} \right] = \alpha.
$$

Secondly we use the Bernstein's inequality, stated in Theorem 2.8.4 in Vershynin (2018). Let $X_1, \ldots, X_n$ be independent, mean zero random variables, such that $|X_i| \leq K$ for some $K > 0$ for all i. Then, for every $t \geq 0$, we have

$$
\mathbb{P}\left\{ \left| \sum_{i=1}^{n} X_i \right| \geq t \right\} \leq 2 \exp\left(- \frac{t^2/2}{\sigma^2 + Kt/3} \right),
$$

where $\sigma^2 = \sum_{i=1}^{n} \mathbb{E} X_i^2 = \sum_{i=1}^{n} \mathrm{Var} X_i$ (since they are mean zero). Again we use the fact each $Y_{il} - E(Y_{il})(i = 1, \ldots, n)$ are independent, and mean zero, and its absolute value is bounded above by $M := 2(n(2\pi I_0(\kappa))^d)^{-1} e^{dk}$ almost surely. Then

$$
\begin{aligned}
P(W_\infty(\mathrm{dgm}(\hat{f}^\dagger), \mathrm{dgm}(f^\dagger)) > \delta) &\leq P(\|\hat{f}^\dagger - f^\dagger\|_\infty) > \delta) \\
&\leq P(\max_{x \in \{g_1, \ldots, g_N\}} |\hat{f}^\dagger(x) - f^\dagger(x)| > \delta) \\
&\leq \sum_{l=1}^{N} \mathbb{P}(|\hat{f}^\dagger(g_l) - f^\dagger(g_l)| > \delta) \\
&\leq 2 \sum_{l=1}^{N} \exp\left[- \frac{\delta^2/2}{\sum_{i=1}^{n} \sigma_l^2 + M\delta/3} \right]. \square
\end{aligned}
$$

In practice each σ_l^2 appeared in 12 is unknown. We propose to replace σ_l^2 by its

consistent estimator $\hat{\sigma}_l^2$ for each $l = 1, \ldots, N$, where

$$\hat{\sigma}_l^2 := \frac{n}{n-1} \sum_{i=1}^{n} \left[\frac{1}{n} \frac{1}{(2\pi I_0(\kappa))^d} \prod_{j=1}^{d} e^{\kappa \cos(g_{lj} \ominus X_{ij})} - \frac{1}{n} \sum_{i=1}^{n} \frac{1}{n} \frac{1}{(2\pi I_0(\kappa))^d} \prod_{j=1}^{d} e^{\kappa \cos(g_{lj} \ominus X_{ij})} \right]^2 .$$

To derive the cutoff value δ corresponding to a given $\alpha \in (0, 1)$, we solve the equation $2\sum_{j=1}^{N} \exp\left[-\frac{\delta^2/2}{n\sigma_j^2 + M\delta/3}\right] = \alpha$. The solution of this equation, denoted by $\hat{\delta}'_n$, must be obtained by a numerical method. We used a bisection method. This choice of $\hat{\delta}'_n$ satisfies (1) for any sample size n.

Note that the bootstrap method provides relatively tight confidence set, but instead it takes a long time to calculate. On the contrary, the Hoeffding method provides a conservative confidence set but with the shortest computational time among the four methods introduced in this article. The Bernstein method strikes a balance. This is empirically confirmed in Section 4.

3.2.3 A Scale Space Approach

The parameters κ and J play a crucial role in determining the shape of the von Mises kernel density and mixture, respectively. These parameters are now referred to as *scale parameters*.

For the concentration parameter, κ, in the von Mises KDE, a larger value indicates a more spiky shape for the density. Similarly, in the von Mises mixture model, increasing the value of J results in a higher number of peaks in the density function f_J. In general, as the scale parameter increases, the convoluted density (f_κ or f_J) exhibits reduced bias but higher variance. The challenge then becomes selecting the optimal scale parameters that best represent the true density f. Instead of identifying a single optimal parameter, we focus on observing the overall trend within a certain range of the scale parameter space. To understand how the significance of each mode changes with respect to the scale parameters, we plot the *dynamic scale-life diagram*, first introduced in Sommerfeld et al. (2017). This visual representation allows us to examine the variation of modes over different scales. This entire process is commonly referred to as a *scale-space approach*.

The movie in Movie 5 is provided for intuitive understanding. Each frame of the movie

Movie 5: Significant density bumps of SARS-CoV-2 data. (Left) 3D plots of von Mises KDE for varying values of κ. (Middle) Cumulative persistence diagram. (Right) Scale-space diagram.

corresponds to different values of κ ranged in $(0, 50)$. The left panel shows the fitted KDE with various κ's, based on the SARS-CoV-2 dataset of size $n = 972$. As κ increases, the resulting KDE becomes more jagged, which in turn leads to a gradual increase in the persistence of each mode shown in the middle and right panels.

To provide a scale space view of the persistent homology, the right panel of Movie 5 displays the scale-space diagram. The y-axis represents the value of the scale parameter, while the x-axis represents the persistence of each feature calculated using the corresponding scale parameter. The height of the confidence band, $2\hat{\delta}_n$ computed for each κ using (7) is shown as the purple curve in the right panel of Movie 5. This diagram enables us to observe how each feature evolves with respect to the scale parameter and how their significance varies.

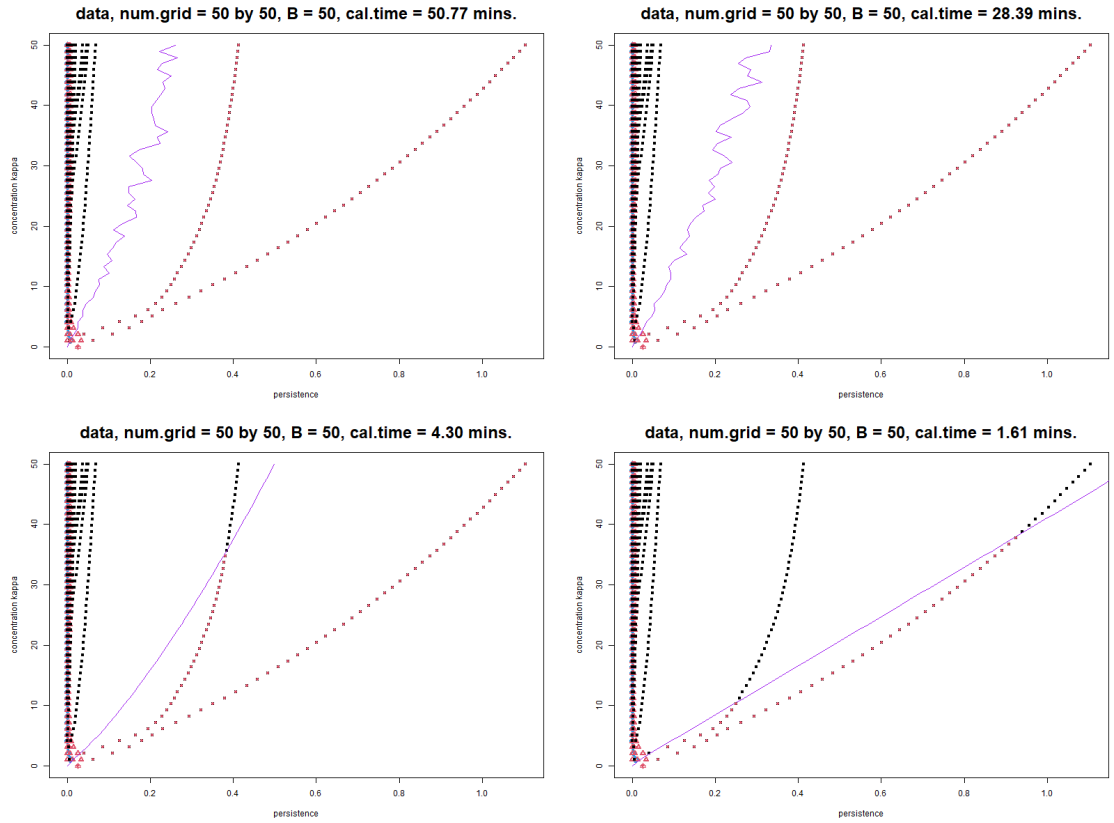

Figure 6: Scale-space diagrams by von Mises KDE for SARS-CoV-2 data. Confidence bands are given by bottleneck bootstrap (top left), functional bootstrap (top right), Bernstein's inequality (bottom left) and Hoeffding's inequaility (bottom right).

4 Experiments

We demonstrate our approaches using two primary torsion angles of the B chain in the SARS-CoV-2 spike glycoprotein, comprising a total of 972 samples in $\mathbb{T}^2$. This dataset was first appeared in Walls et al. (2020) and is accessible in `ClusTorus` R package (Hong and Jung, 2022).

Initially, we examine the variations in the von Mises KDE with the concentration parameter κ ranging from 0 to 50. The construction of confidence sets, which determine the significance of each mode, is accomplished using the four methods introduced in Section 3.2, namely functional bootstrap, Bottleneck bootstrap, Hoeffding's inequality, and Bernstein's inequality. In the implementation, we used a grid of size 50^2, and the number of bootstrap iterations is set to $B = 50$. The results obtained from each method are recorded in the scale-space diagram, shown in Figure 6.

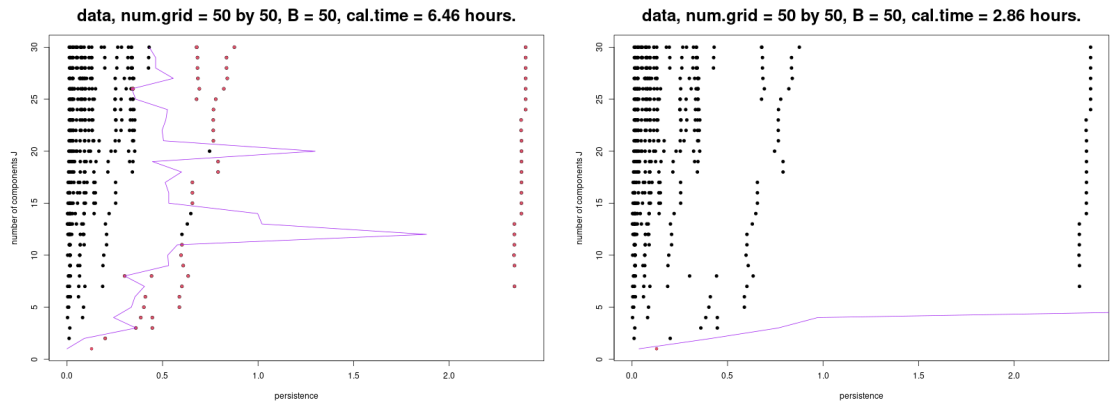

Figure 7: Scale-space diagrams by von Mises mixture models for SARS-CoV-2 data. Confidence bands are given by bottleneck bootstrap (left) and functional bootstrap (right).

In Figure 6, we observed that the bootstrap methods (shown in the top panels) provides shorter confidence bands compared to the finite sample methods (shown in the bottom panels). Moreover, the bottleneck bootstrap provides the tightest confidence band and is most preferable. On the other hand, due to repeated computations of both persistence diagrams and bottleneck distances, it takes the longest time. Overall, the confidence band given by Bernstein's inequality performs good enough and takes moderate computation times. Based on this analysis, we may conclude that SARS-CoV-2 dataset has two significant modes.

Next, we perform the same analysis using the von Mises mixture model. The scale-space diagrams corresponding to the number J of components ranging from 1 to 30 are shown in Figure 7. Note that the mixture model fits are not "continuous" with respect to the changes of J. Nevertheless, one can apply the bootstrap methods to build the confidence bands. Utilising the bottleneck bootstrap (see the left panel), we observed that two modes are identified as significant overall. These modes correspond to the first and third components shown in Figure 4. On the other hand, the functional bootstrap method shown in the right panel does not yield any significant modes. We conjecture that the hypothesis of the stability theorem, which requires $\hat{f}_J$ and its expectation to be Morse, may not hold for this case.

In addition, we repeat the same implementation using one more dataset, which is intentionally simulated data on $\mathbb{T}^2$, with the sample size $n = 1100$. One can see its

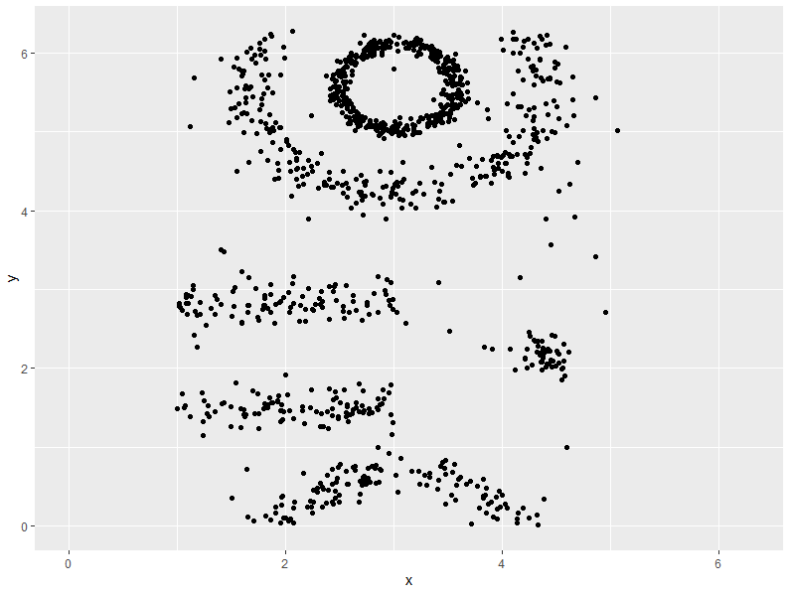

Figure 8: The scatterplot for the simulated data on $\mathbb{T}^2$. Each axis of x and y circulates between 0 and 2π.

scatterplot on Figure 8. This dataset is originally simulated on Euclidean space, which is accessible in `factoextra` R package. We preprocessed this data to be embedded well over a two-dimensional torus space circulating between 0 and 2π for each dimension.

Figure 9 and Figure 10 depict the outcomes of the simulated data experiment. The experiment employed identical grid parameters, bootstrap iteration counts, and methodologies for assessing uncertainty in persistent homology relative to the SARS-CoV-2 dataset. In Figure 9, each data point within a circular arrangement signifies its corresponding zero-dimensional topological characteristic, a connected component. Conversely, each triangular data point designates its corresponding one-dimensional topological aspect, denoting a ring-shaped structure. As we observe in Figure 8, it is reasonable and intuitive a ring-shaped topological feature should be captured to be significant. The bootstrap methodologies effectively identified the aforementioned ring-like feature as well as several other connected components of significance. In contrast, Bernstein's and Hoeffding's methods failed to achieve analogous outcomes. This trend persists in Figure 10, where the von Mises mixture model was applied to the simulated data. However, due to the inherent assumptions of the mixture model framework, the detected significant components deviated from capturing an entire ring-like feature to identifying several pronounced connected components resembling a ring-like configuration. Notably, the functional boot-

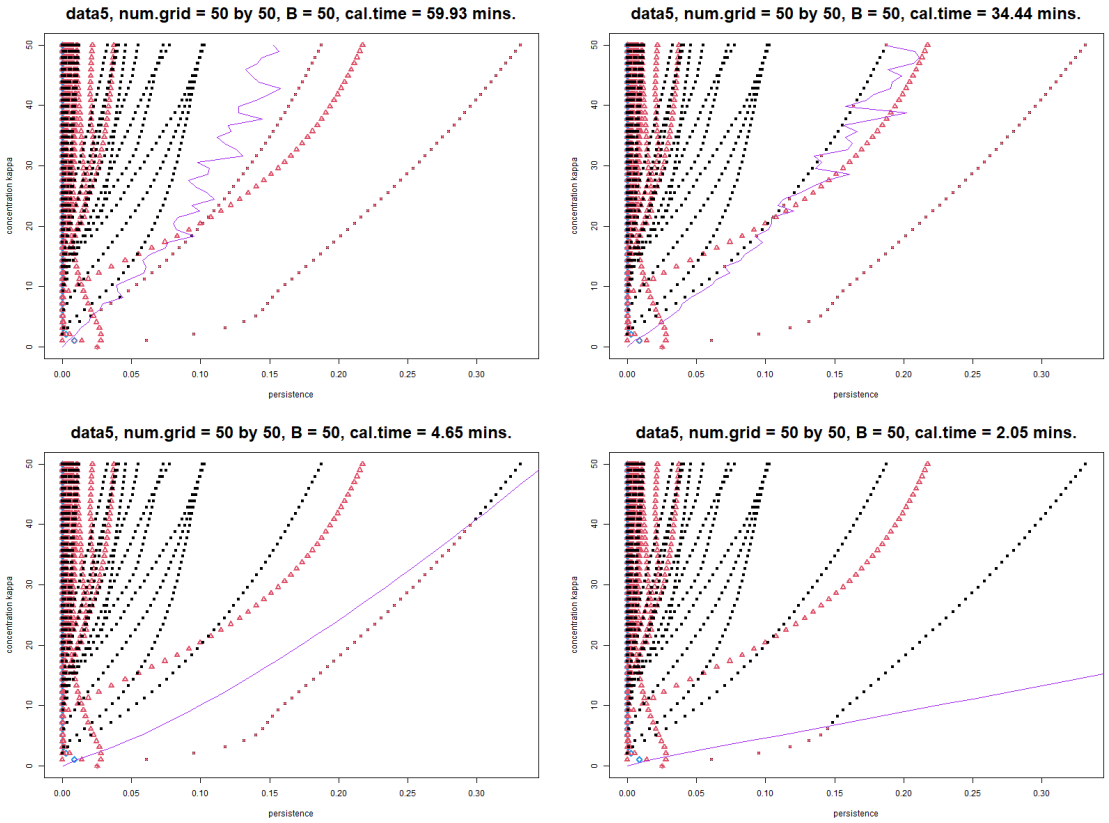

Figure 9: Scale-space diagrams by von Mises KDE for the simulated data. Confidence bands are given by bottleneck bootstrap (top left), functional bootstrap (top right), Bernstein's inequality (middle left) and Hoeffding's inequaility (middle right).

strap approach yielded no significant components apart from the scenario where $J = 1$. This outcome aligns with the observations in the SARS-CoV-2 data case, likely due to analogous underlying factors.

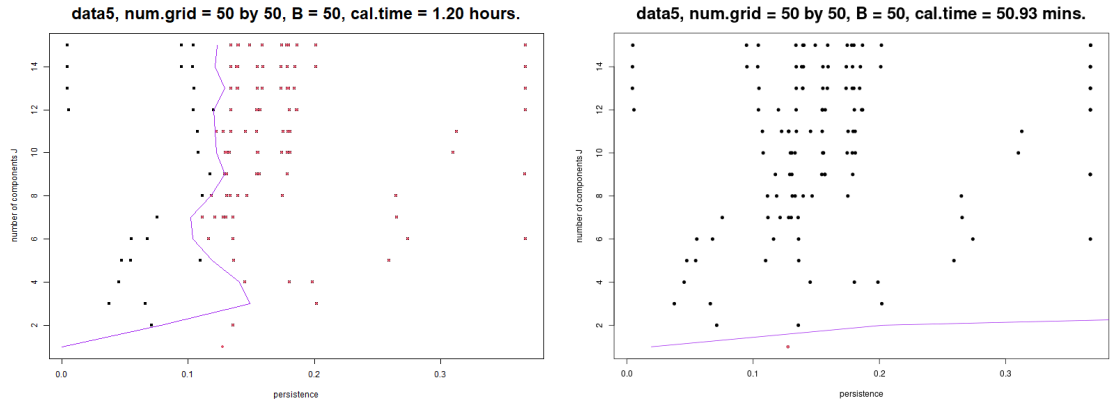

Figure 10: Scale-space diagrams by von Mises mixture models for the simulated data. Confidence bands are given by bottleneck bootstrap (left) and functional bootstrap (right).

5 Conclusion and Discussions

We summarise the findings of our study and identify topics for future research.

Firstly, we have explored the use of von Mises KDE and mixture model. The kernel method and the confidence band corresponding to it are proven to be valuable and theoretically supported. Mixture models offer useful advantages over the KDE, but further theoretical development is required. In particular, theoretical results for bootstrap confidence band and the detection of one or higher dimensional homology can be studies further.

Secondly, the bootstrap method provides tight confidence intervals, allowing for the detection of more significant features. However, its computation time poses a major challenge. In this regard, the Bernstein method has shown moderate computation time and power, making it a viable alternative when using the KDE for persistent homology.

Thirdly, our study demonstrates that under certain theoretical conditions, with an appropriate density estimator, Topological data analysis can be applied to more general sample spaces, such as directional data on a sphere.

Appendix

The assertion regarding the approximation of the upper level set of the von Mises distribution as a union of ellipses relies on the approximating technique that involves the Gaussian density (Hong and Jung, 2022). By employing Taylor approximation for $\cos y \approx 1 - \frac{1}{2}y^T y$ and $\sin y \approx y$, (5) can be simplified as:

$$f^*(y; \mu, \Lambda, \kappa) \approx (2\pi)^{-d/2}|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}\left[\kappa^{\mathrm{T}}(2 - 2c(y, \mu)) + s(y, \mu)^{\mathrm{T}}\Lambda s(y, \mu)\right]\right\},$$

where Σ contains the elements of κ. Again, by using the large concentration of $\cos y \approx 1 - \frac{1}{2}y^T y$ and $\sin y \approx y$, we further approximate the density as:

$$f^*(y; \mu, \Lambda, \kappa) \approx (2\pi)^{-d/2}|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}\left[(y \ominus \mu)\Sigma^{-1}(y \ominus \mu)\right]\right\}.$$

Consequently, the approximated jth component of $f_J(y)$ takes the following form:

$$\begin{aligned}\left\{y \in \mathbb{T}^d; \pi_j f_j^*(y) \geq t\right\} &= \left\{y \in \mathbb{T}^d; 2\log \pi_j + 2\log f_j^*(y) \geq 2\log t\right\} \\ &\approx \left\{y \in \mathbb{T}^d; (y \ominus \mu_j)\Sigma_j^{-1}(y \ominus \mu_j) \leq 2\log \pi_j - d\log(2\pi) - \log|\Sigma_j| - 2\log t\right\}.\end{aligned}$$

References

Arthur, D. and Vassilvitskii, S. (2012), "k-means: the advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms*, pp. 396–401.

Carlsson, G. (2009), "Topology and data," *Bulletin (new series) of the American Mathematical Society*, 46, 255–308.

Chaudhuri, P. and Marron, J. S. (1999), "SiZer for Exploration of Structures in Curves," *Journal of the American Statistical Association*, 94, 807–823.

Chazal, F., de Silva, V., Glisse, M., and Oudot, S. (2016), *The structure and stability of persistence modules*, Springer.

Chazal, F., Fasy, B., Lecci, F., Michel, B., Rinaldo, A., and Wasserman, L. (2018), "Robust Topological Inference: Distance To a Measure and Kernel Distance," *Journal of machine learning research*, 18, 1–40.

Chazal, F., Fasy, Brittany, T., Lecci, F., Rinaldo, A., Singh, A., and Wasserman, L. (2015), "On the Bootstrap for Persistence Diagrams and Landscapes," *arXiv preprint arXiv:1311.0376*.

Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. (2007), "Stability of Persistence Diagrams," *Discrete & computational geometry*, 37, 103–120.

Di Marzio, M., Panzera, A., and Taylor, C. C. (2011), "Kernel density estimation on the torus," *Journal of statistical planning and inference*, 141, 2156–2173.

Do Carmo, M. P. a. (2016), *Differential geometry of curves & surfaces*, Dover Publications, Inc.

Dvoretzky, A.and Kiefer, J. and Wolfowitz, J. (1956), "Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator," *The Annals of mathematical statistics*, 27, 642–669.

Edelsbrunner, H. and Harer, J. (2010), *Computational topology*, Providence, R.I. : American Mathematical Society.

Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. (2014), "Confidence Sets for Persistence Diagrams," *The Annals of statistics*, 42, 2301–2339.

Gilitschenski, I. and Hanebeck, U. D. (2012), "A robust computational test for overlap of two arbitrary-dimensional ellipsoids in fault-detection of Kalman filters," in *15th International Conference on Information Fusion*, pp. 396–401.

Gine, E. and Zinn, J. (1990), "Bootstrapping General Empirical Measures," *The Annals of probability*, 18, 851–869.

Giné, E. and Nickl, R. (2008), "Uniform central limit theorems for kernel density estimators," *Probability theory and related fields*, 141, 333–387.

Good, I. J. and Gaskins, R. A. (1980), "Density Estimation and Bump-Hunting by the Penalized Likelihood Method Exemplified by Scattering and Meteorite Data," *Journal of the American Statistical Association*, 75, 42–56.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009), *The elements of statistical learning*, Springer.

Hatcher, A. (2002), *Algebraic topology*, Cambridge University Press.

Hong, S. and Jung, S. (2022), "ClusTorus: An R Package for Prediction and Clustering on the Torus by Conformal Prediction," *The R Journal*, 14, 186–207.

Jung, S., Park, K., and Kim, B. (2021), "Clustering on the torus by conformal prediction," *The annals of applied statistics*, 15, 1583–1603.

Kahn, D. W. (1995), *Topology*, Dover.

Kim, J., Chen, Y.-C., Balakrishnan, S., Rinaldo, A., and Wasserman, L. (2017), "Statistical Inference for Cluster Trees," *arXiv preprint arXiv:1605.06416*.

Kosorok, M. R. (2008), *Introduction to empirical processes and semiparametric inference*, New York: Springer.

Lei, J., Rinaldo, A., and Wasserman, L. (2013), "A conformal prediction approach to explore functional data," *Annals of mathematics and artificial intelligence*, 74, 29–43.

Ley, C. and Verdebout, T. (2017), *Modern directional statistics*, CRC Press.

Mardia, K. V., Hughes, G., Taylor, C. C., and Singh, H. (2008), "A multivariate von mises distribution with applications to bioinformatics," *Canadian journal of statistics*, 36, 99–109.

Mardia, K. V. and Jupp, P. E. (2000), *Directional statistics*, J. Wiley.

Mardia, K. V., Kent, J. T., Zhang, Z., Taylor, C. C., and Hamelryck, T. (2012), "Mixtures of concentrated multivariate sine distributions with applications to bioinformatics," *Journal of applied statistics*, 39, 2475–2492.

Marron, J. S. and Dryden, I. L. (2021), *Object Oriented Data Analysis*, Chapman and Hall.

Munkres, J. R. (2014), *Topology*, Pearson, 2nd ed.

Rao, P. B. L. S. (1983), *Nonparametric functional estimation*, Academic Press.

Shin, J., Alessandro, R., and Larry, W. (2019), "Predictive clustering," *arXiv preprint arXiv:1903.08125*.

Sommerfeld, M., Heo, G., Kim, P., Rush, S. T., and Marron, J. S. (2017), "Bump hunting by topological data analysis," *Stat*, 6, 462–471.

Taylor, C. C. (2008), "Automatic bandwidth selection for circular density estimation," *Computational statistics & data analysis*, 52, 3493–3500.

van der Vaart, A. W. (1998), *Asymptotic statistics*, Cambridge University Press.

Vershynin, R. (2018), *High-dimensional probability*, Cambridge University Press.

Wagner, H., Chen, C., and Vuçini, E. (2011), "Efficient Computation of Persistent Homology for Cubical Data," *Mathematics and Visualization*, 91–106.

Walls, A. C., Park, Y.-J., Tortorici, M. A., Wall, A., McGuire, A. T., and Veesler, D. (2020), "Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein," *Cell*, 183, 1735.

국문초록

본 연구는 각도 데이터 등 순환하는 자료들이 토러스 공간 위에 있다고 가정하고 그들을 통해 밀도함수의 최빈값(mode)들을 찾음으로써 자료들이 집중적으로 분포된 곳을 탐색함을 목표로 한다. 이를 위해 밀도함수를 폰 미시스(von Mises) 커널 밀도함수 추정량과 혼합모형을 이용하여 추정하고 이들을 통해 지속 호몰로지(persistent homology) 방법을 사용할 것이다. 밀도함수 대신 추정량을 사용함으로 파생된 지속 호몰로지의 불확실성을 계량하기 위해 우리는 네 가지 방법을 비교할 것인데, 그 중 셋은 선행연구에서 제시된 것이고, 하나는 우리가 이 연구를 통해 새롭게 제시하는 방법이다. 또한 기존 위상학적 자료 분석 선행연구에서 제시된 측도모수공간 방법(scale-space approach)을 적극적으로 활용하여 여러 측도모수에 의한 밀도함수 추정량 최빈값들, 지속 호몰로지와 그 유의성의 변화를 살펴볼 것이다. 이러한 연구는 이론적인 내용뿐만 아니라 현존하는 데이터들을 이용한 실험을 통해 그 유용성을 검증하는 절차도 포함하고 있다.

주요어 : 범프 헌팅, 붓스트랩, 위상학적 자료 분석, 지속 호몰로지, 폰 미시스 분포

학 번 : 2021-26565