Homology, Homotopy and Applications, vol. 9(2), 2007, pp.337–362

A STATISTICAL APPROACH TO PERSISTENT HOMOLOGY

PETER BUBENIK AND PETER T. KIM

(*communicated by Gunnar Carlsson*)

Abstract

Assume that a finite set of points is randomly sampled from a subspace of a metric space. Recent advances in computational topology have provided several approaches to recovering the geometric and topological properties of the underlying space. In this paper we take a statistical approach to this problem. We assume that the data is randomly sampled from an unknown probability distribution. We define two filtered complexes with which we can calculate the persistent homology of a probability distribution. Using statistical estimators for samples from certain families of distributions, we show that we can recover the persistent homology of the underlying distribution.

1. Introduction

There is growing interest in characterizing topological features of data sets. Given a finite set, sometimes called *point cloud data (PCD)*, that is randomly sampled from a subspace X of some metric space, one hopes to recover geometric and topological properties of X. Using random samples, P. Niyogi, S. Smale and S. Weinberger [**NSW05**] show how to recover the homology of certain submanifolds. In [**CCSL06**] the homotopy type of certain compact subsets is recovered.

A finer descriptor, developed by H. Edelsbrunner, D. Letscher, A. Zomorodian and G. Carlsson, is that of *persistent homology* [**ELZ02, ZC05**]. While it is not a homotopy invariant, it is stable under small changes [**CSEH05**]. Using the PCD and the metric, one can construct a filtered simplicial complex which approximates the unknown space X [**dSC04, CZCG04**]. This leads naturally to a spectral sequence. What is unusual is that the homology of the start of the spectral sequence is uninteresting, and so is what it converges to. Nevertheless, the intermediate homology, called *persistent homology* is of interest. It can be described using *barcodes*, which are analogues of the Betti numbers.

The aim of this paper is to take a statistical approach to these ideas. We assume that the data is sampled from a manifold with respect to a probability distribution. Given such a distribution, we construct two filtered chain complexes: the *Morse*

This research was partially funded by the Swiss National Science Foundation grant 200020-105383.
This research was partially funded by NSERC grant OGP46204.
Received July 25, 2006, revised August 1, 2007; published on September 24, 2007.
2000 Mathematics Subject Classification: 55N99, 62H11.
Key words and phrases: persistent homology, point cloud data, directional statistics, parametric statistics, expected persistent homology.

complex, and the *Čech complex*. For most of the distributions we consider, these complexes are related by Alexander duality. Using persistent homology, one can calculate the corresponding Betti barcodes, which provide a topological description of the distribution. In the case of the Čech complex we define a Betti-0 function. We apply these methods to several parametric families of distributions: the von Mises, von Mises-Fisher, Watson and Bingham distributions on S^{p-1} and the matrix von Mises distribution on $SO(3)$.

Given a sample, it is assumed that the underlying distribution is unknown, but that it is one of a parametrized family. We use statistical techniques to estimate the parameter. These are then used to estimate the barcodes. As a result, we prove that we can recover the persistent homology of the underlying distribution.

Theorem 1.1. *Let $x_1, \ldots, x_n$ be a sample from S^{p-1} according to the von Mises-Fisher distribution with fixed concentration parameter $\kappa \geqslant 0$. Given the sample, let $\hat{\kappa}$ be the maximum likelihood estimator for κ (which is given by formula* (31)*). Let β_κ and $\beta_{\hat{\kappa}}$ denote the Betti barcodes for the persistent homology of the densities associated with κ and $\hat{\kappa}$ using either the Morse or the Čech filtration. Finally let $E(\cdot)$ denote the expectation, and $\mathcal{D}$ denote the barcode metric (see Definition 3.5). Then,*

$$E(\mathcal{D}(\beta_{\hat{\kappa}}, \beta_\kappa)) \leqslant C(\kappa) n^{-1/2},$$

as $n \to \infty$, for some constant $C(\kappa)$.

We also show that the classical theory of spacings [**Pyk65**] can be used to calculate the exact expectations of the Betti barcodes for samples from the uniform distribution on S^1 together with their asymptotic behavior.

As part of the results, we show that the Morse filtrations of our distributions each correspond to a relative CW-structure for the underlying spaces. The von Mises and von Mises-Fisher distributions correspond to the decomposition $S^{p-1} \approx * \cup_* D^{p-1}$, the Watson distribution corresponds to $S^{p-1} \approx S^{p-2} \cup_{\mathrm{Id} \amalg -\mathrm{Id}} (D^{p-1} \amalg D^{p-1})$, and the Bingham distribution corresponds to

$$S^{p-1} \approx * \cup_{\mathrm{Id} \amalg -\mathrm{Id}} (D^1 \amalg D^1) \cup_{\mathrm{Id} \amalg -\mathrm{Id}} (D^2 \amalg D^2) \cup \ldots \cup_{\mathrm{Id} \amalg -\mathrm{Id}} (D^{p-1} \amalg D^{p-1}).$$

Finally, the Morse filtration on the matrix von Mises distribution on $SO(3)$ corresponds to the decomposition $\mathbb{RP}^2 \cup_f D^3$ where $f\colon S^2 \to \mathbb{RP}^2$ identifies antipodal points. Interestingly, the last decomposition is obtained by using the Hopf fibration $S^0 \to S^3 \to \mathbb{RP}^3$.

A summary of the paper goes as follows. In Section 2 we go over the background and notation used in this paper. We review both the statistical and the topological terminologies. In Section 3 we discuss filtrations and persistent homology and we develop two filtrations for densities. In Section 4 we use the theory of spacings to give exact estimates of the persistent homology of uniform samples on S^1. In Section 5 we calculate the persistent homology of some standard parametric families of densities on S^{p-1} and $SO(3)$. In Section 6 we use maximum likelihood estimators to recover the persistent homology of the underlying density.

2. Background and notation

In an attempt to make this article accessible to a broad audience, we define some of the basic statistical and topological terms we will be using.

2.1. Statistics

Given a manifold $\mathcal{M}$ with Radon measure ν, a *density* is a function $f\colon \mathcal{M} \to [0,\infty]$ such that $f d\nu$ is a *probability distribution* on $\mathcal{M}$ with $\int_{\mathcal{M}} f d\nu = 1$. A common statistical example is to take $\mathcal{M} = \mathbb{R}^p$, and $d\nu$ to be the p-dimensional Lebesgue measure. A density in this case would be a nonnegative function that integrates to unity. We can also take $\mathcal{M} = S^{p-1}$, the $(p-1)$-dimensional unit sphere, with $d\nu$ being the $(p-1)$-dimensional spherical measure. In this case a density is referred to as a *directional density*. For $\mathcal{M}$ a compact, connected, orientable Riemannian manifold, $d\nu$ would be the measure induced by the Riemannian structure.

In statistics, we think of a family of probability densities parametrized accordingly

$$\{f_\vartheta \colon \vartheta \in \Theta\}\,, \tag{1}$$

where ϑ is called a *parameter* and Θ is called the *parameter space*. The parameter space Θ can be quite general and if it is some subset of a finite-dimensional vector space, then (1) is referred to as a *parametric* family of densities; otherwise it is known as a *nonparametric* family of densities. Subsequent to this, the corresponding statistical problem will be referred to as either a parametric statistical procedure, or, a nonparametric statistical procedure, depending on whether we are dealing with a parametric, or nonparametric family of densities, respectively.

Some parametric examples are in order. Let $\mathcal{M} = \mathbb{R}^p$ and consider the normal family of location scale probability densities,

$$f_{\mu,\sigma}(x) = (2\pi\sigma^2)^{-p/2} \exp\left\{\tfrac{\|x-\mu\|^2}{2\sigma^2}\right\}, \tag{2}$$

where $\mu, x \in \mathbb{R}^p$ and $\sigma^2 \in [0,\infty)$. Letting $\vartheta = (\mu,\sigma^2)$, we note that this parametric problem has $\Theta = \mathbb{R}^p \times [0,\infty)$ as its parameter space.

If we take $\mathcal{M} = S^{p-1}$, a well-known example of a directional density, and one that will be used in this paper is given by

$$f_{\mu,\kappa}(x) = c(\kappa) \exp\left\{\kappa x^t \mu\right\}, \tag{3}$$

where $\mu, x \in S^{p-1}$, $\kappa \in [0,\infty)$, $c(\kappa)$ is the normalizing constant and superscript "t" denotes transpose. The distribution arising from $f_{\mu,\kappa}$ is called the *von Mises-Fisher distribution* where this parametric problem has $\Theta = S^{p-1} \times [0,\infty)$ as its parameter space.

Somewhat related to the above is the situation where $\mathcal{M} = SO(p)$, the space of $p \times p$ rotation matrices. Let

$$f_{\mu,\kappa}(x) = c(\kappa) \exp\left\{\kappa \operatorname{tr} x^t \mu\right\}, \tag{4}$$

where $\mu, x \in SO(p)$, $\kappa \in [0,\infty)$ and $c(\kappa)$ is the normalizing constant. The distribution arising from $f_{\mu,\kappa}$ is called the *matrix von Mises-Fisher distribution* where this parametric problem has $\Theta = SO(p) \times [0,\infty)$ as its parameter space.

A *sample* $X_1, X_2, \dots X_N$ is a sequence of independent and identically distributed random quantities on $\mathcal{M}$ drawn according to the density f_ϑ for some fixed but unknown $\vartheta \in \Theta$. The parameter of interest would be the fixed but unknown parameter ϑ, or, more generally, some transformation $\tau(\vartheta)$ thereof. Statistically, we want to find an estimator $\tilde{\tau} = \tilde{\tau}(X_1, \dots, X_N)$ of $\tau(\vartheta)$. Given some metric γ on $\tau(\Theta)$, the performance of the estimator is evaluated relative to this metric in expectation with respect to the joint probability density of the sample,

$$E_\vartheta \gamma\left(\tilde{\tau}, \tau\right) = \int_{\mathcal{M}} \cdots \int_{\mathcal{M}} \gamma\left(\tilde{\tau}, \tau\right) f_\vartheta \cdots f_\vartheta d\nu \cdots d\nu, \tag{5}$$

where the above represents an N-fold integration and $\vartheta \in \Theta$. Thus the relative merit of one estimator over another estimator can be evaluated using (5) in a statistical decision theory context; see [**Ber85**].

There are a wide variety of different distributions for a given manifold, as well as sample spaces that are different manifolds. References that discuss these topics can be found in the books by Mardia and Jupp [**MJ00**] and Chikuse [**Chi03**]. Furthermore, although nonparametric statistical procedures on compact Riemannian manifolds are available, [**Hen90, Efr00, AK05, KK00**], in this paper we will deal with parametric statistical procedures.

2.2. Topology

Let R be a commutative ring with identity. (In fact, we will only be interested in cases where R is a field, in which case R-modules are vector spaces and R-module morphisms are linear maps of vector spaces.)

Definition 2.1. A *chain complex* over R is a sequence of R-modules $\{C_i\}_{i\in\mathbb{Z}}$ and R-module morphisms $d_i \colon C_i \to C_{i-1}$ called *differentials* such that $d_i \circ d_{i+1} = 0$. This condition is often abbreviated to $d^2 = 0$. The elements of C_n are called *n-chains*. This chain complex is denoted by (C, d).

Definition 2.2. An (abstract) *simplicial complex* K is a set of finite, ordered subsets of an ordered set $\bar{K}$, such that

- the ordering of the subsets is compatible with the ordering of $\bar{K}$, and
- if $\alpha \in K$ then any nonempty subset of α is also an element of K.

The elements of K with $n+1$ elements are called n-simplices and denoted K_n.

Definition 2.3. Given a simplicial complex K, the *chain complex* on K, denoted $(C_*(K), d)$ is defined as follows. Let $C_n(K)$ be the free R-module with basis K_n. We define the differential on K_n and extend it to $C_n(K)$ by linearity. For $[v_0, \dots, v_n] \in K_n$ define

$$d[v_0, \dots, v_n] = \sum_i (-1)^i [v_0, \dots, \hat{v}_i, \dots, v_n],$$

where $\hat{v}_i$ denotes that the element v_i is omitted from the sequence.

For $n \geqslant 0$, the *standard n-simplex* is the n-dimensional polytope in $\mathbb{R}^{n+1}$, denoted Δ^n, whose vertices are given by the standard basis vectors $e_0, \dots, e_n$. It is just the

convex hull of the standard basis vectors; that is

$$\Delta^n = \left\{ x = \sum_{i=0}^{n} a_i e_i \;\middle|\; \forall i \; a_i \geqslant 0 \text{ and } \sum_{i=0}^{n} a_i = 1 \right\}. \tag{6}$$

There are inclusion maps

$$\delta_i \colon \Delta^n \to \Delta^{n+1} \tag{7}$$

(called the i-th face inclusion), which are given by

$$\delta_i(x_0, \ldots x_n) = (x_0, \ldots, x_{i-1}, 0, x_i, \ldots, x_n)$$

or $0 \leqslant i \leqslant n+1$.

Definition 2.4. Let X be a topological space. For $n \geqslant 0$, let $C_n(X)$ be the free R-module generated by the set of continuous maps $\{\phi \colon \Delta^n \to X\}$. For $n < 0$, let $C_n(X) = 0$. For $\phi \colon \Delta^n \to X$, let

$$d(\phi) = \sum_{i=0}^{n} (-1)^i \; \phi \circ \delta_i \; \in C_{n-1}(X). \tag{8}$$

Extend this by linearity to an R-module morphism $d \colon C_n(X) \to C_{n-1}(X)$. One can check that $d^2 = 0$, so this defines a differential and $C_*(X) = (\{C_n(X)\}_{n \in \mathbb{Z}}, d)$ is a chain complex, called the *singular chain complex*.

Definition 2.5. Given a chain complex (C, d), let Z_k be the submodule given by $\{x \in C_k \mid dx = 0\}$ called the *k-cycles*, and let B_k be the submodule given by $\{x \in C_k \mid \exists y \in C_{k+1} \text{ such that } dy = x\}$, called the *$k$-boundaries*. Since $d^2 = 0$, $d(dy) = 0$ and thus $B_k \subset Z_k$. The *k-th homology* of (C, d), denoted $H_k(C, d)$, is given by the R-module Z_k / B_k. The homologies $\{H_k(C, d)\}_{k \in \mathbb{Z}}$ form a chain complex with differential 0 denoted $H_*(C, d)$ and called the homology of (C, d). If R is a principal ideal domain (for example, if R is a field) and $H_k(C, d)$ is finitely generated, then $H_k(C, d)$ is the direct sum of a free group and a finite number of finite cyclic groups. The *k-th Betti number* $\beta_k(C, d)$ is the rank of the free group. If R is a field, then $\beta_k(C, d)$ equals the dimension of the vector space $H_k(C, d)$. If X is a topological space, then $H_*(X)$ denotes the homology of the singular chain complex on X.

Definition 2.6. Two spaces X and Y are said to be homotopy equivalent (written $X \approx Y$) if there are maps $f \colon X \to Y$ and $g \colon Y \to X$ such that $g \circ f$ is homotopic to the identity map on X and $f \circ g$ is homotopic to the identity map on Y.

Remark 2.7. If $X \approx Y$, then $H_*(X) \cong H_*(Y)$. So if X is a *contractible space* (that is, a space which is homotopy equivalent to a point), then $H_0(X) \cong R$ and $H_k(X) = 0$ for $k \geqslant 1$.

3. Filtrations and persistent homology

From now on, we will assume that the ground ring is a field $\mathbb{F}$.

3.1. Persistent homology

In Definition 2.5 we showed how to calculate the homology of a chain complex. Given some additional information on the chain complex, we will calculate homology in a more sophisticated way. Namely, we will show how to calculate the *persistent homology* of a *filtered chain complex*. This will detect homology classes which persist through a range of values in the filtration.

Let $\overline{\mathbb{R}}$ denote the totally ordered set of extended real numbers $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$. Then an increasing $\overline{\mathbb{R}}$-*filtration* on a chain complex (C, d) is a sequence of chain complexes $\{\mathcal{F}_r(C, d)\}_{r \in \overline{\mathbb{R}}}$ such that $\mathcal{F}_r(C, d)$ is a subchain module of (C, d) and $\mathcal{F}_r(C, d) \subset \mathcal{F}_{r'}(C, d)$ whenever $r \leqslant r' \in \overline{\mathbb{R}}$. A chain complex, together with a $\overline{\mathbb{R}}$-filtration, is called a $\overline{\mathbb{R}}$-*filtered chain complex*.

For a filtered chain complex, the inclusions $\mathcal{F}_j(C, d) \to \mathcal{F}_{j+l}(C, d)$ induce maps

$$H_k(F_j(C, d)) \to H_k(F_{j+l}(C, d)).$$

The image of this map is call the l-persistent k-th homology of $\mathcal{F}_j(C, d)$.

Let $Z_k^i = Z_k(\mathcal{F}_i(C, d))$ and let $B_k^i = B_k(\mathcal{F}_i(C, d))$. Assume $\alpha \in Z_k^i$. Then α represents a homology class $[\alpha]$ in $H_*(\mathcal{F}_i(C, d))$. Furthermore since $Z_k^i \subset Z_k^{i'}$ for all $i' \geqslant i$, α also represents a homology class in $H_*(\mathcal{F}_{i'}(C, d))$, which we again denote $[\alpha]$. One possibility is that $[\alpha] \neq 0$ in $H_k(\mathcal{F}_i(C, d))$ but $[\alpha] = 0$ in $H_k(\mathcal{F}_{i'}(C, d))$ for some $i' > i$.

Assume (C, d) is a chain complex with an $\overline{\mathbb{R}}$-filtration $\mathcal{F}_r((C, d))$ such that

$$\bigcup_{r \in \overline{\mathbb{R}}} \mathcal{F}_r(C, d) = (C, d) \text{ and } \bigcap_{r \in \overline{\mathbb{R}}} \mathcal{F}_r(C, d) = 0. \tag{9}$$

Equivalently, $\mathcal{F}_\infty(C, d) = (C, d)$ and $\mathcal{F}_{-\infty}(C, d) = 0$.

Lemma 3.1. *Let (C, d) be a filtered chain complex satisfying* (9)*. For any n-chain $\alpha \in (C, d)$, there is some smallest $r \in \overline{\mathbb{R}}$ such that $\alpha \notin \mathcal{F}_{r'}(C, d)$ for all $r' < r$ and $\alpha \in \mathcal{F}_{r''}(C, d)$ for all $r'' > r$.*

Proof. This follows from the definition of an $\overline{\mathbb{R}}$-filtration, the assumption (9), and the linear ordering of $\overline{\mathbb{R}}$. □

Lemma 3.2. *For any n-cycle $\alpha \in Z_n$, the set of all $r \in \overline{\mathbb{R}}$ such that*

$$0 \neq [\alpha] \in H_n(\mathcal{F}_r(C, d)$$

is either empty or is an interval.

Proof. Let $\alpha \in Z_n$, and let r_1 be the corresponding value given by Lemma 3.1.

If there is some $\beta \in C_{n+1}$ such that $d\beta = \alpha$, then again let r_2 be the corresponding value given by Lemma 3.1. Since $\beta \in \mathcal{F}_j(C, d)$ implies that $d\beta \in \mathcal{F}_j(C, d)$, it follows that $r_2 \geqslant r_1$. Thus α represents a nonzero homology class in $\mathcal{F}_r(C, d)$ exactly when r is in the (possibly empty) interval beginning at r_1 and ending at r_2. This interval contains r_1 if and only if $\alpha \in \mathcal{F}_{r_1}(C, d)$, and it does not contain r_2 if and only if $\beta \in \mathcal{F}_{r_2}(C, d)$.

If α is not a k-boundary then α represents a nonzero homology class in $\mathcal{F}_r(C, d)$ exactly when r is in the interval $\{x \mid x \geqslant r_1\}$ or $\{x \mid x > r_1\}$. beginning at r_1. Again this interval contains r_1 if and only if $\alpha \in \mathcal{F}_{r_1}(C, d)$. □

Definition 3.3. For $\alpha \in Z_k$ define the *persistence k-homology interval* represented by α to be the interval given by Lemma 3.2. Denote it by I_α.

Definition 3.4. Define a *Betti-k barcode* to be a set of intervals[1] $\{J_\alpha\}_{\alpha \in S \subset Z_k}$ such that

- J_α is a subinterval of I_α, and
- for all $r \in \overline{\mathbb{R}}$, $\{[\alpha] \mid \alpha \in S,\ r \in J_\alpha\}$ is an $\mathbb{F}$-basis for $H_k(\mathcal{F}_r(C,d))$.

We will sometimes use β_k to denote a Betti-k barcode.

The set of barcodes has a metric [**CZCG04**] defined as follows.

Definition 3.5. Given an interval J, let $\ell(J)$ denote its length. Given two intervals J and J', the *symmetric difference*, $\Delta(J, J')$, between them is the one-dimensional measure of $J \cup J' - J \cap J'$. Given two barcodes $\{J_\alpha\}_{\alpha \in S}$ and $\{J'_{\alpha'}\}_{\alpha' \in S'}$, a *partial matching*, M, between the two sets is a subset of $S \times S'$ where each α and α' appears at most once. Define

$$\mathcal{D}(\{J_\alpha\}_{\alpha \in S}, \{J'_{\alpha'}\}_{\alpha' \in S'}) = \min_M \left(\sum_{(\alpha,\alpha') \in M} \Delta(J_\alpha, J'_{\alpha'}) + \sum_{\alpha \notin M_1} \ell(J_\alpha) + \sum_{\alpha' \notin M_2} \ell(J'_{\alpha'}) \right),$$

where the minimum is taken over all partial matchings, and M_i is the projection of M to S_i. This defines a quasi-metric (since its value may be infinite). If desired, it can be converted into a metric.

3.2. Persistent homology from point cloud data

Let $(\mathcal{M}, \rho)$ be a manifold with a metric ρ. Let $X = \{x_1, x_2, \ldots, x_n\} \subset \mathcal{M}$. X is called *point cloud data*. One would like to be able to obtain information on $\mathcal{M}$ from X. If X contains sufficiently many uniformly distributed points one may be able to construct a complex from X that in some sense reconstructs $\mathcal{M}$.

One such construction is the following $\overline{\mathbb{R}}$-filtered simplicial complex called the Čech complex. Recall that we are working over a ground field $\mathbb{F}$. Let $\mathcal{C}_*(X)$ be the largest simplicial complex on the ordered vertex set X. That is, $\mathcal{C}_0(X) = X$ and for $k \geqslant 1$, $\mathcal{C}_k(X)$ consists of the ordered subsets of X with $k+1$ elements. Now filter this simplicial complex (along $\overline{\mathbb{R}}$) as follows. Given $r < 0$, define $\mathcal{F}_r^{\check{C}}(\mathcal{C}_n(X)) = 0$ for all n. Let $B_r(x)$ denote the ball of radius r centered at x. For $r \geqslant 0$ and $k \geqslant 1$, define $\mathcal{F}_r^{\check{C}}(\mathcal{C}_k(X))$ to be the $\mathbb{F}$-vector space whose basis is the k-simplices $[x_{i_0}, \ldots, x_{i_k}]$ such that $\cap_{j=0}^k B_r(x_{i_j}) \neq 0$. We remark that there are fast algorithms for computing $\mathcal{F}_r^{\check{C}}(\mathcal{C}_k(X))$.[2] $\mathcal{F}_r^{\check{C}}(\mathcal{C}_*(X))$ is called the r-Čech complex. It is the *nerve* of the collection of balls $\{B_r(x_i)\}_{i=1}^n$, and its geometric realization is homotopy equivalent to the union of these balls.

---

[1] In Section 3.3 we will see that using the Čech filtration, the Betti-0 barcode of manifolds will have uncountably many intervals, so we will define a more appropriate descriptor, the Betti-0 function. In Section 4 it will also be useful to convert finite Betti barcodes to functions so that we can analyze limiting and asymptotic behavior.

[2] The balls of radius r centered at the points $\{x_{i_j}\}$ have nonempty intersection if and only if there is a ball of radius r containing the points $\{x_{i_j}\}$. There are fast algorithms for the smallest enclosing ball problem [**FGK03, Gae06**].

A related construction is the Rips complex. For each r, the r-Rips complex, $\mathcal{F}_r^R(\mathcal{C}_*(X))$, is the largest simplicial complex containing the 1-skeleton $\mathcal{F}_r^{\check{C}}(\mathcal{C}_1(X))$. That is, $\mathcal{F}_r^R(\mathcal{C}_*(X))$ is the $\mathbb{F}$-vector space whose basis is the set of k-simplices $[x_{i_0}, \ldots, x_{i_k}]$ such that $\rho(x_{i_j}, x_{i_\ell}) \leqslant r$ for all pairs $0 \leqslant j, \ell \leqslant k$.

Using either of these filtered chain complexes, one obtains a filtered chain complex as follows. Let $\Delta_*(\mathcal{C}_*(X))$ be the chain complex on $\mathcal{C}_*(X)$. Filter this over $\overline{\mathbb{R}}$ by letting

$$\mathcal{F}_r(\Delta_*(\mathcal{C}_*(X))) = \Delta_*(\mathcal{F}_r(\mathcal{C}_*(X))), \text{ where } \mathcal{F}_r = \mathcal{F}_r^{\check{C}} \text{ or } \mathcal{F}_r^R.$$

To simplify the notation, we write $\Delta_k(X) := \Delta_k(\mathcal{C}_*(X))$. We remark that these filtrations satisfy (9):

$$\bigcup_{r\in\overline{\mathbb{R}}} \mathcal{F}_r(\Delta_*(X)) = \Delta_*(X) \text{ and } \bigcap_{r\in\overline{\mathbb{R}}} \mathcal{F}_r(\Delta_*(X)) = 0.$$

Let α be an n-chain. By Lemma 3.1 we know that there is some $r \in \overline{\mathbb{R}}$ such that $\alpha \notin \mathcal{F}_{r'}(\Delta_n(X))$ for all $r' < r$ and $\alpha \in \mathcal{F}_{r''}(\Delta_n(X))$ for all $r'' > r$. In fact,

Lemma 3.6. *Consider an n-chain, $\alpha = \sum_{i=1}^m \alpha_i(x_{i_0}, \ldots, x_{i_n})$. For the Čech filtration let*

$$r = \max_{i=1\ldots m} \min\{r_i \mid \exists x \text{ such that } B_{r_i}(x) \ni x_{i_0}, \ldots x_{i_n}\},$$

and for the Rips filtration let

$$r = \max_{i=1\ldots m} \max_{j\neq k} \rho(x_{i_j}, x_{i_k}).$$

Then $\alpha \notin \mathcal{F}_{r'}(\Delta_n(X))$ for all $r' < r$ and $\alpha \in \mathcal{F}_{r''}(\Delta_n(X))$ for all $r'' \geqslant r$.

If α is an n-cycle then by Lemma 3.2 there is a (possibly empty) persistence n-homology interval corresponding to α. Applying Lemma 3.6 to α and, if there is some $\beta \in \Delta_{k+1}(X)$ such that $d\beta = \alpha$, applying Lemma 3.6 to β, we get the following.

Lemma 3.7. *Given an n-cycle α, the persistence n-homology interval associated to α is either empty or has the form $[r_1, r_2)$ or $[r_1, \infty]$.*

3.3. Persistent homology of densities

Let f_ϑ be a probability density on a manifold $\mathcal{M}$ for some $\vartheta \in \Theta$. We will use f_ϑ to define two increasing $\overline{\mathbb{R}}$-filtrations on $C_*(\mathcal{M})$, the singular chain complex on $\mathcal{M}$ (see Definition 2.4).

3.3.1. The Morse filtration

For $r \in \overline{\mathbb{R}}$, the *excursion sets*

$$\mathcal{M}_{\leqslant r} = \{x \in \mathcal{M} \mid f_\vartheta(x) \leqslant r\}, \tag{10}$$

(used in Morse theory [**Mil63**]) filter $\mathcal{M}$ over $\overline{\mathbb{R}}$. Hence they also provide an $\overline{\mathbb{R}}$-filtration of the singular chain complex $C_*(\mathcal{M})$,

$$\mathcal{F}_r^M(C_*(\mathcal{M})) = C_*(\mathcal{M}_{\leqslant r}),$$

which we call the *Morse filtration*. We remark that for all k,

$$H_k(\mathcal{F}_r^M C_*(\mathcal{M})) = H_k(\mathcal{M}_{\leqslant r}).$$

3.3.2. The Čech filtration

There is a dual increasing filtration to the Morse filtration which uses superlevel sets instead of sublevel sets. We modify this filtration slightly so that it mirrors the filtration on the Čech complex defined in Section 3.2, and we will call it the *Čech filtration*. We do this since the filtrations on the Čech complex and the related Rips complex are the main filtrations used in computations of persistent homology.

Notice that in the Čech complex filtration all of the points in X, even distant outliers appear when $r = 0$. So the Čech filtration starts with all of the points of M and the discrete topology, and then progressively connects the regions with decreasing density.

For $r < 0$ and all k, define $\mathcal{F}_r^{\check{C}}(C_k(\mathcal{M})) = 0$. For $r \geqslant 0$, let $\mathcal{F}_r^{\check{C}}(C_0(\mathcal{M})) = C_0(\mathcal{M})$. Assume $k \geqslant 1$. Let

$$\mathrm{Const}_k = \{\phi\colon \Delta^k \to \mathcal{M} \mid \phi \text{ is constant}\} \subset C_k(\mathcal{M}).$$

For $0 \leqslant s \leqslant \infty$, let

$$\mathcal{M}_{\geqslant s} = \{m \in \mathcal{M} \mid f_\vartheta(m) \geqslant s\}\,. \tag{11}$$

For $r \geqslant 0$, let

$$\mathcal{F}_r^{\check{C}}(C_k(\mathcal{M})) = \mathrm{Const}_k + C_k(\mathcal{M}_{\geqslant \frac{1}{r}}). \tag{12}$$

From this filtered chain complex we can calculate persistence k-homology intervals and Betti-k barcodes just as in Section 3.2.

Lemma 3.8. *For $k \geqslant 1$,*

$$H_k(\mathcal{F}_r^{\check{C}}(C_*(\mathcal{M}))) \cong H_k(\mathcal{M}_{\geqslant \frac{1}{r}}).$$

Proof. By definition, the k-cycles are $Z_k(\mathcal{F}^{\check{C}}{}_r C_*(\mathcal{M})) = \mathrm{Const}_k + Z_k C_*(\mathcal{M}_{\geqslant \frac{1}{r}})$, and the k-boundaries are $B_k(\mathcal{F}^{\check{C}}{}_r C_*(\mathcal{M})) = \mathrm{Const}_k + B_k C_*(\mathcal{M}_{\geqslant \frac{1}{r}})$. So

$$H_k(\mathcal{F}^{\check{C}}{}_r C_*(\mathcal{M}) \cong Z_k(C_*(\mathcal{M}_{\geqslant \frac{1}{r}}))/B_k(C_*(\mathcal{M}_{\geqslant \frac{1}{r}})) = H_k(\mathcal{M}_{\geqslant \frac{1}{r}}). \qquad \square$$

Let $r \geqslant 0$. Recall the notation of Section 3.1: $Z_k^r = Z_k(\mathcal{F}_r^{\check{C}}(C_*(\mathcal{M})))$ and $B_k^r = B_k(\mathcal{F}_r^{\check{C}}(C_*(\mathcal{M}))$. To start, $Z_0^r = \mathbb{F}[\mathcal{M}]$. Then

$$\mathcal{F}_r^{\check{C}}(C_1(\mathcal{M})) = \mathbb{F}[\{\phi\colon \Delta^1 \to \mathcal{M} \mid \phi \text{ is constant, or } \operatorname{im}\phi \subset \mathcal{M}_{\geqslant \frac{1}{r}}\}].$$

For two points $x, y \in M$, there is some map $\phi\colon \Delta^1 \to \mathcal{M}$ such that $\phi(0) = x$, $\phi(1) = y$ and $\operatorname{im}(\phi) \subset \mathcal{M}_{\geqslant \frac{1}{r}}$, in which case $d\phi = x - y$, if and only if x and y are in the same path component of $\mathcal{M}_{\geqslant \frac{1}{r}}$. Thus

$$H_0(\mathcal{F}_r^{\check{C}}(C_*(\mathcal{M}))) \cong \mathbb{F}[\mathcal{M}/\sim],$$

where $x \sim y$ if and only if x and y are in the same path component of $\mathcal{M}_{\geqslant \frac{1}{r}}$.

In the case where $\mathcal{M}_{\geqslant \frac{1}{r}}$ is path-connected, $H_0(\mathcal{F}_r^{\check{C}}(C_*(\mathcal{M}))) \cong \mathbb{F}[\mathcal{M}/\mathcal{M}_{\geqslant \frac{1}{r}}]$. In particular $H_0(\mathcal{F}_0^{\check{C}}(C_*(\mathcal{M}))) \cong \mathbb{F}[\mathcal{M}/\mathcal{M}_{\geqslant \infty}]$. Since f_ϑ is a probability density, $\mathcal{M}_{\geqslant \infty}$ has measure 0. Therefore almost all $m \in \mathcal{M}$ represent a distinct homology class in $\mathcal{F}_0^{\check{C}}(C_0(\mathcal{M}))$ and there are uncountably many 0-homology intervals. As a result the

Betti-0 barcode is not a good descriptor. In this section, we will describe how the 0-homology intervals can be used to describe a *Betti*-0 *function*, in the case where the density f_ϑ satisfies a continuity condition.

More generally, as long as $\mathcal{M} - \mathcal{M}_{\geqslant \frac{1}{r}}$ is uncountable and $\mathcal{M}_{\geqslant \frac{1}{r}}$ has countably many path components, then almost all homology classes in $H_0(\mathcal{F}_r^{\check{C}}(C_*(\mathcal{M})))$ have a unique representative. In this case we use this as justification to consider only those homology classes with a unique representative.

Assume that for all r, $\mathcal{M} - \mathcal{M}_{\geqslant \frac{1}{r}}$ is uncountable and $\mathcal{M}_{\geqslant \frac{1}{r}}$ has countably many path components, and that the following continuity condition holds for all $m \in \mathcal{M}$:

$$\forall \epsilon > 0,\ \exists \text{ injective } \phi\colon [0,1] \to \mathcal{M} \text{s.t.} \phi(0) = m \text{ and } f(\phi(t)) > f(m) - \epsilon. \tag{13}$$

This condition holds if f_ϑ is continuous.

Lemma 3.9. *Each $m \in M$ is a unique representative for $[m]$ for exactly those values of $r \in \left[0, \frac{1}{f_\vartheta(m)}\right)$ or $r \in \left[0, \frac{1}{f_\vartheta(m)}\right]$.*

Proof. Let $m \in \mathcal{M}$. Since $dm = 0$, $m \in Z_0^r$ for $r \geqslant 0$. Let $[m] \in H_*(\mathcal{F}_r^{\check{C}}(C_*(\mathcal{M})))$ denote the homology class represented by m. By definition, $m \in \mathcal{M}_{\geqslant \frac{1}{r}}$ if and only if $r \geqslant \frac{1}{f_\vartheta(m)}$. Thus m is the unique representative for $[m]$ for $r < \frac{1}{f_\vartheta(m)}$. By assumption, for any $\epsilon > 0$ there is an injective map $\phi\colon [0,1] \to \mathcal{M}$ such that $\phi(0) = m$ and $f_\vartheta(\phi(t)) > f_\vartheta(m) - \epsilon$. Then $\phi \in \mathcal{F}_r^{\check{C}}(C_1(\mathcal{M}))$ where $r = \frac{1}{f_\vartheta(m) - \epsilon}$. This implies that for any $\epsilon > 0$ there is a nonconstant continuous map $\phi\colon \Delta^1 \to \mathcal{M}$ with $\phi(0) = m$ such that $\phi \in \mathcal{F}^{\check{C}}_{\frac{1}{f_\vartheta(m)} + \epsilon}(C_1(\mathcal{M}))$. Hence m is not a unique representative for $[m]$ for $r > \frac{1}{f_\vartheta(m)}$. Therefore m is a unique representative for $[m]$ for either $r \in \left[0, \frac{1}{f_\vartheta(m)}\right)$ or $r \in \left[0, \frac{1}{f_\vartheta(m)}\right]$. □

Before we formally define the Betti-0 function, we give the following intuitive picture. We draw each of our intervals $\left[0, \frac{1}{f_\vartheta(m)}\right]$ or $\left[0, \frac{1}{f_\vartheta(m)}\right)$ vertically starting at $r = 0$ and ending at $r = f_\vartheta(m)$. Furthermore we order the intervals from left to right according to their length. In fact we draw all of the intervals between $x = 0$ and $x = 1$, where the x-axis is scaled according to the probability distribution $f_\vartheta d\nu$. The increasing curve traced by the tips of the intervals will be called the Betti-0 function.

Definition 3.10. Formally, define the *Betti*-0 *function* $\beta_0\colon (0,1] \times \Theta \to [0,\infty]$ as follows.[3] For $r \in [0,\infty]$, let

$$g_\vartheta(r) = \int_{\mathcal{M}_{\geqslant \frac{1}{r}}} f_\vartheta d\nu. \tag{14}$$

Since f_ϑ is a probability density, g_ϑ is an increasing function $g_\vartheta\colon [0,\infty] \to [0,1]$ for each fixed $\vartheta \in \Theta$. Also recall that $\mathcal{M}_{\geqslant \infty}$ has measure 0 and by definition $\mathcal{M}_{\geqslant 0} = \mathcal{M}$.

[3] While our definition of β_0 below (15) is valid for $x = 0$, we get $\beta_0(0,\vartheta) \equiv 0$. This does not provide any information, and is furthermore inappropriate in cases such as the von Mises distribution with $\kappa = 0$ (see Section 5.1 below) where $\beta_0(x,\vartheta)$ is constant and no-nzero for $x > 0$.

So $g_\vartheta(0) = 0$ and $g_\vartheta(\infty) = 1$. For $0 < x \leqslant 1$, let

$$\beta_0(x, \vartheta) = \inf_{g_\vartheta(r) \geqslant x} r. \tag{15}$$

If g_ϑ is continuous and strictly increasing,[4] then

$$\beta_0(x, \vartheta) = g_\vartheta^{-1}(x), \tag{16}$$

for $\vartheta \in \Theta$. That is, $\beta_0(x, \vartheta)$ is the unique value of r such that $\int_{M \geqslant \frac{1}{r}} f_\vartheta d\nu = x$.

3.3.3. Alexander duality

The Morse and Čech filtrations on S^{p-1} are related by Alexander duality. Let f be a density on S^{p-1}. Assume that $r \in \mathrm{im}(f)$ and that $r < \sup(f)$. Then $S^{p-1}_{f \leqslant r}$ is a proper, nonempty subset of S^{p-1}. Assume that $S^{p-1}_{f \leqslant r}$ is compact and a neighborhood retract.

Theorem 3.11 (Alexander duality for the Morse and Čech filtrations on S^{p-1})**.** *Let $\tilde{H}$ denote reduced homology, let $\mathbb{F}$ be a field, and let $s = \frac{1}{r}$.*

$$\tilde{H}_i(S^{p-1}_{f > \frac{1}{s}}; \mathbb{F}) \cong \tilde{H}^{p-2-i}(S^{p-1}_{f \leqslant r}; \mathbb{F}) \cong \tilde{H}_{p-2-1}(S^{p-1}_{f \leqslant r}; \mathbb{F}).$$

4. Expected barcodes of PCD

4.1. Betti barcodes of uniform samples on S^1

Let f be the uniform density on S^1. Let $X = \{X_1, \ldots X_n\} \subset S^1$ be a sample drawn according to f. X is called the point cloud data. In this section we consider the Betti barcodes obtained for the persistent homology of $\mathcal{F}^R_*(\Delta_*(X))$ the Rips complex on X (see Section 3.2). The metric we use on S^1 is $\frac{1}{2\pi}$ times the shortest arc length between two points (we have normalized so that the total length of S^1 is one).

Before we continue, we introduce some notation. Choose α such that $X_1 = e^{i\cos(\alpha)}$. For $k = 2, \ldots n$ choose $U_k \in [0, 1]$ such that

$$X_k = e^{2\pi i(\alpha + U_k)}.$$

We remark that each U_k is uniformly distributed on $[0, 1]$. Now reorder the $\{U_k\}$ to obtain the order statistic[5]:

$$0 < U_{n:1} < U_{n:2} < \ldots < U_{n:n-1} < 1.$$

Let $U_{n:0} = 0$ and $U_{n:n} = 1$. Reorder the $\{X_k\}$ as $\{X_{n:k}\}$ to correspond with the $\{U_{n:k}\}$. Then for $1 \leqslant k \leqslant n$ define

$$S_k = U_{n:k} - U_{n:k-1}.$$

The set $S = \{S_1, \ldots, S_n\}$ is called the set of spacings [**Pyk65**]. We remark that if $U_k = U_{n:j}$ with $1 \leqslant j \leqslant n-1$ and take the usual orientation of S^1, then the distances from X_k to its nearest backward neighbor and nearest forward neighbor are S_j and S_{j+1}, respectively. Also the distance from X_1 to its neighbors is S_n and S_1. It is well known (for example, [**Dev81**]) that

[4] In this case we can define $\beta_0(x, \vartheta)$ for $x \in [0, 1]$.

[5] Equality among any of the terms occurs with probability zero.

Lemma 4.1. *$(S_1, \ldots, S_n)$ is uniformly distributed on the standard $(n-1)$-simplex $\{(x_1, \ldots x_n) | x_i \geqslant 0, \sum_{i=1}^n x_i = 1\}$. It follows that*

$$P[S_1 > a_1; \cdots; S_n > a_n] = \begin{cases} (1 - \sum_{i=1}^n a_i)^{n-1} & \text{if } \sum_{i=1}^n a_i < 1, \\ 0 & \text{otherwise,} \end{cases}$$

and, as calculated by Whitworth in 1897,

$$P(S_{n:n} > x) = \sum_{\substack{k \geqslant 1 \\ kx < 1}} (-1)^{k+1} (1 - kx)^{n-1} \binom{n}{k}, \quad \forall x > 0. \tag{17}$$

Finally, order the spacings to obtain

$$0 < S_{n:1} < S_{n:2} < i \ldots < S_{n:n-1} < 1.$$

Now we are ready to calculate the degree-0 homology. Recall that $\beta_0(\mathcal{F}_r^R(\Delta_*(X)))$ equals the dimension of $H_0(\mathcal{F}_r^R(\Delta_*(X))$, which equals the number of path components of $\mathcal{F}_r^R(\Delta_*(X))$. Recall that $\mathcal{F}_r^R(\Delta_0(X))$ is the empty set for $r < 0$ and is the set X for $r \geqslant 0$. So at $r = 0$, there are (almost surely) exactly n distinct homology classes in $H_0(\mathcal{F}_r^R(\Delta_*(X)))$. Each homology class $[X_k]$ will no longer have a distinct representative when the distance from X_k to one of its neighbors is equal to r. That is, each time r passes one of the S_k the dimension of $H_*(\mathcal{F}_r^R(\Delta_*(X)))$ decreases by one. Therefore for $k = 0, \ldots n-2$,

$$r \in [S_{n:k}, S_{n:k+1}) \implies \beta_0(\mathcal{F}_r^R(\Delta_*(X))) = n - k.$$

When $r \geqslant S_{n:n-1}$, $\mathcal{F}_r^R(\Delta_*(X))$ is path-connected so $\beta_0(\mathcal{F}_r^R(\Delta_*(X))) = 1$. Translating this, we see that the Betti-0 barcode is the collection of homology intervals

$$[0, S_{n:k}) \text{ for } k = 1, \ldots n-1 \text{ and } [0, \infty].$$

Finally, let us consider the homology in degree 1. Let

$$\alpha = (X_{n:1}, X_{n:2}) + \ldots + (X_{n:n-1}, X_{n:n}) + (X_{n:n}, X_{n:1}).$$

This is a 1-cycle in $\Delta_*(X)$.

Lemma 4.2. *If $S_{n:n} \leqslant \frac{1}{2}$, then the Betti-1 barcode is the single (possibly empty) persistence homology interval*

$$I_\alpha = [S_{n:n}, R), \quad \text{where} \quad R \in [\tfrac{1}{3}, \tfrac{1}{2});$$

otherwise it is empty.

Remark 4.3. If the large spacing $S_{n:n}$ is greater than or equal to $\frac{1}{2}$, then all of the points $X_1, \ldots X_n$ are concentrated on a semicircle, and $\mathcal{F}_r^R(\Delta_*(X))$ does not contain any nontrivial 1-cycles. By (17), $P[S_{n:n} > \frac{1}{2}] = \frac{n}{2^{n-1}}$.

Proof. Assume that $S_{n:n} \leqslant \frac{1}{2}$. If $r \geqslant S_{n:n}$, then $\alpha \in \mathcal{F}_r^R(\Delta_1(X))$. We claim that by using the definition of the Rips filtration and the geometry of S^1, α becomes a boundary at some $R \in [\frac{1}{3}, \frac{1}{2}]$. Since half the perimeter of S^1 is $\frac{1}{2}$, when $r \geqslant \frac{1}{2}$, $(X_i, X_j) \in \mathcal{F}_r^R(\Delta_1(X))$ for all $X_i, X_j \in X$. Thus when $r \geqslant \frac{1}{2}$ then $\mathcal{F}_r^R(\Delta_*(X)) = \Delta_*(X)$ which

is the full $(n-1)$-simplex on the vertices $X_1, \dots X_n$. In particular, if $r \geqslant \frac{1}{2}$, then α is a boundary.

Since $S_{n:n} < \frac{1}{2}$, the geometric realization of α is a n-gon containing the center of S^1. Thus if there is some $\beta = \sum \beta_{ijk}(X_i, X_j, X_k) \in \mathcal{F}_r^R(\Delta_2(X))$ such that $d\beta = \alpha$ then for some $(X_i, X_j, X_k) \in \mathcal{F}_r^R(\Delta_2(X))$ the geometric realization of (X_i, X_j, X_k) contains the center of S^1. The smallest r for which this can happen is $\frac{1}{3}$. So if $r < \frac{1}{3}$ then α cannot be a boundary.

Thus α becomes a boundary when $r = R$ for some $R \in [\frac{1}{3}, \frac{1}{2}]$. If $S_{n:n} \geqslant \frac{1}{3}$ it is possible that $R = S_{n:n}$, and α is not a nontrivial boundary in any $\mathcal{F}_r^R(\Delta_*(X)$. $\square$

Remark 4.4. If $S_{n:n} < \frac{1}{3}$, then the Betti-1 barcode is a single nonempty persistence homology interval. Using (17), $P[S_{n:n} \geqslant \frac{1}{3}] < n\left(\frac{2}{3}\right)^{n-1}$.

4.2. Expected values of the Betti barcodes

Let $U_1, \dots U_{n-1}$ be a sample from the uniform distribution on $[0,1]$. Let $0 < U_{n:1} < U_{n:2} < \dots < U_{n:n-1} < 1$ be the corresponding order statistic.[6] Define $U_{n:0} = 0$ and $U_{n:n} = 1$. For $k = 1, \dots n$, let $S_k = U_{n:k} - U_{n:k-1}$. Recall (Lemma 4.1) that the set of spacings $S = \{S_1, \dots S_n\}$ is uniformly distributed on the standard $(n-1)$-simplex.

Let $0 < S_{n:1} < \dots < S_{n:n} < 1$ be the order statistic for the spacings. Then one can show [**SW86**, 21.1.15] that

Proposition 4.5. *For $1 \leqslant i \leqslant n$, the expected value of the spacings is given by*

$$ES_{n:i} = \frac{1}{n}\sum_{j=1}^{i} \frac{1}{n+1-j} = \frac{1}{n}\sum_{j=n+1-i}^{n} \frac{1}{j}.$$

So the expected Betti-0 barcode is the collection of intervals

$$\left\{\left[0, \frac{1}{n}\sum_{j=1}^{i} \frac{1}{n+1-j}\right)\right\}_{i\in\{1,\dots,n-1\}} \cup \{[0,\infty]\},$$

and the expected Betti-1 barcode is

$$\left\{\left[\frac{1}{n}\sum_{j=1}^{n} \frac{1}{n+1-j}, \infty\right]\right\}.$$

To obtain the Betti-0 function from the Betti-0 barcode let

$${}_n\tilde{\beta}_0(x,0) = ES_{n:\lceil (n-1)x\rceil}.$$

The Betti-0 function is a normalized version of this ${}_n\beta_0(x,0) = c_n\, {}_n\tilde{\beta}_0(x,0)$ so that $\int_0^1 {}_n\beta_0(x,0)dx = 1$. (In fact, $c_n = \frac{n-1}{1-ES_{n:n}}$, which for large values of n is approximately equal to n.) Thus,

$${}_n\beta_0(x,0) = \frac{c_n}{n}\sum_{j=1}^{\lceil (n-1)x\rceil} \frac{1}{n+1-j} = \frac{c_n}{n}\sum_{j=n+1-\lceil (n-1)x\rceil}^{n} \frac{1}{j}.$$

[6] We use n here to match the notation of Section 4.1 where $\{U_1, \dots, U_{n-1}\}$ is derived from $\{X_1, \dots, X_n\} \in S^1$.

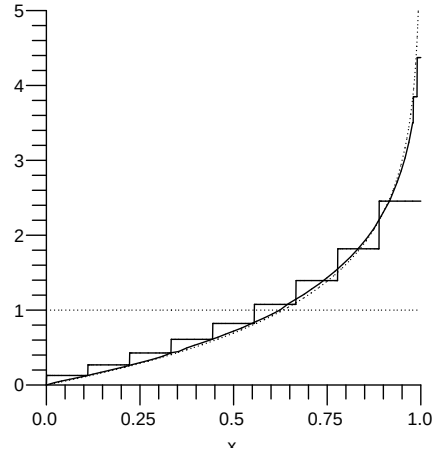

Figure 1: Graphs of the expected Betti 0-functions for $n = 10$ and $n = 100$ together with graphs of $f(x) = 1$ and $f(x) = -\ln(1-x)$.

Proposition 4.6. *For* $0 < x < 1$, *as* $n \to \infty$,

$$ {}_n\beta_0(x, 0) \to -\ln(1-x). $$

Proof. By the definition of c_n, $\lim_{n\to\infty} \frac{c_n}{n} = 1$. The result then follows from the observation that

$$ \frac{1}{n} + \int_k^n \frac{1}{x} dx < \sum_{j=k}^{n} \frac{1}{j} < \frac{1}{k} + \int_k^n \frac{1}{x} dx $$

and the fact that

$$ \lim_{n\to\infty} \ln\left(\frac{n}{n+1-\lceil (n-1)x \rceil}\right) = -\ln(1-x). \qquad \square $$

In Figure 1, we display graphs of the expected Betti-0 functions $y = {}_{10}\beta_0(x, 0)$ and $y = {}_{100}\beta_0(x, 0)$ and the limiting function $y = -\ln(1-x)$. For comparison, we also graph $y = 1$, the limiting function one would obtain if the spacings became relatively equal in the limit.

5. Barcodes of certain parametric densities

5.1. The von Mises distribution

Let $\mathcal{M} = S^1 = \{e^{i\theta} \mid \theta \in [-\pi, \pi)\} \subset \mathbb{R}^2$. We will use this parametrization to identify $\theta \in [-\pi, \pi)$ with an element of S^1. Consider the von Mises density on S^1 with respect to the uniform measure,

$$ f_{\mu,\kappa}(\theta) = \tfrac{1}{I_0(\kappa)} e^{\kappa \cos(\theta - \mu)}, \quad \theta \in [-\pi, \pi), $$

where $\mu \in [-\pi, \pi)$, $\kappa \in [0, \infty)$ and $I_0(x)$ is the modified Bessel function of the first kind and order 0, where the general ν-th order Bessel function of the first kind is

$$I_\nu(\kappa) = \frac{(\kappa/2)^\nu}{\Gamma(\nu+\frac{1}{2})\Gamma(\frac{1}{2})} \int_{-1}^{1} e^{\kappa t}(1-t^2)^{\nu-\frac{1}{2}} dt, \tag{18}$$

and $\Gamma(\cdot)$ denotes the gamma function.

Our homologies will be independent of μ, so assume that $\mu = 0$ and in this case the parameter $\vartheta = \kappa$.

We will filter the chain complex on S^1 using both the Čech and Morse filtrations. Recall that by (11) and (10), $S^1_{\geqslant \frac{1}{r}} = \{\theta \in S^1 \mid f_\kappa(\theta) \geqslant \frac{1}{r}\}$ and $S^1_{\leqslant r} = \{\theta \in S^1 \mid f_\kappa(\theta) \leqslant r\}$. Choose $\alpha_{r,\kappa} \in [-\pi, \pi)$ such that

$$f_\kappa(\alpha_{r,\kappa}) = r.$$

Specifically, let $\alpha_{r,\kappa} = \cos^{-1}(\frac{1}{\kappa}\ln(\frac{r}{c(\kappa)}))$. Our calculations of the persistent homology will follow from the following straightforward result.

Lemma 5.1. *For* $0 \leqslant r < \frac{1}{\max f_\kappa}$, $S^1_{\geqslant \frac{1}{r}} = \phi$, *and for* $r < \min f_\kappa$, $S^1_{\leqslant r} = \phi$.
For $\frac{1}{\max f_\kappa} \leqslant r < \frac{1}{\min f_\kappa}$,

$$S^1_{\geqslant \frac{1}{r}} = \{\theta \mid -\alpha_{\frac{1}{r},\kappa} \leqslant \theta \leqslant \alpha_{\frac{1}{r},\kappa}\}.$$

For $\min f_\kappa \leqslant r < \max f_\kappa$,

$$S^1_{\leqslant r} = \{\theta \mid \alpha_{r,\kappa} \leqslant \theta \leqslant 2\pi - \alpha_{r,\kappa}\}.$$

For $r \geqslant \frac{1}{\min f_\kappa}$, $S^1_{\geqslant \frac{1}{r}} = S^1$, *and for* $r \geqslant \max f_\kappa$, $S^1_{\leqslant r} = S^1$.

Since its analysis is simpler, we start with the Morse filtration on S^1. By Lemma 5.1, $S^1_{\leqslant r}$ is empty if $r < \min f_\kappa$, is contractible (see Remark 2.7) if $\min f_\kappa \leqslant r < \max f_\kappa$ and is equal to S^1 if $r \geqslant \max(f_\kappa)_\kappa$. It follows that the Betti-0 barcode for the Morse filtration is the single interval

$$[\min f_\kappa, \infty] = \left[\frac{1}{I_0(\kappa)e^\kappa}, \infty\right],$$

the Betti-1 barcode is the single interval

$$[\max f_\kappa, \infty] = \left[\frac{e^\kappa}{I_0(\kappa)}, \infty\right],$$

and all other Betti-k barcodes are empty.

Now consider the Čech filtration on S^1. We will derive a formula for the Betti-0 function, $\beta_0(x, \kappa)$, and calculate the Betti-k barcodes for $k > 0$.

If $\kappa = 0$ then $f_0 = 1$. So for $r < 1$, $S^1_{\geqslant \frac{1}{r}} = \emptyset$, and for $r \geqslant 1$, $S^1_{\geqslant \frac{1}{r}} = S^1$. By definition (14),

$$g_\kappa(r) = \begin{cases} 0 & \text{if } r < 1, \\ 1 & \text{if } r \geqslant 1. \end{cases}$$

So by definition (15), $\beta_0(x, 0) = 1$.

For $\kappa > 0$, let $\min(f_\kappa) = \frac{1}{I_0(\kappa)}e^{-\kappa}$ and $\max(f_\kappa) = \frac{1}{I_0(\kappa)}e^{\kappa}$. For $r < \frac{1}{\max(f_\kappa)}$, $S^1_{\geqslant \frac{1}{r}} = \emptyset$, and for $r \geqslant \frac{1}{\min(f_\kappa)}$, $S^1_{\geqslant \frac{1}{r}} = S^1$. For $\frac{1}{\max(f_\kappa)} \leqslant r < \frac{1}{\min(f_\kappa)}$, since f_κ is even and decreasing for $\theta > 0$,

$$S^1_{\geqslant \frac{1}{r}} = \{\theta \mid -\alpha_{r,\kappa} \leqslant \theta \leqslant \alpha_{r,\kappa}\},$$

where $\alpha_{r,\kappa} \in (0, \pi)$ and $f_\kappa(\alpha_{r,\kappa}) = \frac{1}{r}$.

Let $x \in [0, 1]$ and assume that $\beta_0(x, \kappa) = r$. Since $\kappa \geqslant 0$, $g_\kappa(r) = \int_{S^1_{\geqslant \frac{1}{r}}} f_\kappa(\theta)d\theta$ is continuous and strictly increasing. So,

$$x = \int_{S^1_{\geqslant \frac{1}{r}}} f_\kappa(\theta)d\theta.$$

Define $\alpha_{r,\kappa} \in [0, \pi]$ by the condition that $f_\kappa(\alpha_{r,\kappa}) = \frac{1}{r}$. So

$$r = \frac{1}{f_\kappa(\alpha_{r,\kappa})}. \tag{19}$$

For $\psi \in [0, \pi]$, let

$$F_\kappa(\psi) = \int_0^\psi f_\kappa(\theta)d\theta.$$

Then

$$x = \int_{S^1_{\geqslant \frac{1}{r}}} f_\kappa d\nu = \int_{-\alpha_{r,\kappa}}^{\alpha_{r,\kappa}} f_\kappa(\theta)d\theta = 2F_\kappa(\alpha_{r,\kappa}). \tag{20}$$

Since F_κ is strictly increasing, it is invertible. So $\alpha_{r,\kappa} = F_\kappa^{-1}(\frac{x}{2})$, and thus

$$\beta_0(x, \kappa) = r = \frac{1}{f_\kappa(F_\kappa^{-1}(\frac{x}{2}))}. \tag{21}$$

Since f_κ and F_κ are smooth, by the inverse function theorem, so is F_κ^{-1}. So

$$\beta_0(x, \kappa) = (F_\kappa^{-1})'\left(\frac{x}{2}\right).$$

We remark that as $\kappa \to 0$, $\beta_0(x, \kappa) \to 1 = \beta_0(x, 0)$. We can also describe the graph of $r = \beta_0(x, \kappa)$ parametrically by combining (19) and (20) (see Figure 2):

$$h_\kappa(t) = \left(2F_\kappa(t), \frac{1}{f_\kappa(t)}\right), t \in [0, \pi]. \tag{22}$$

For $k \geqslant 1$, recall that

$$\mathcal{F}_r^{\check{C}}(C_k(S^1)) = \mathrm{Const}_k + C_k(S^1_{\geqslant \frac{1}{r}}).$$

Also recall that for $r < \frac{1}{\max(f_\kappa)}$, $S^1_{\geqslant \frac{1}{r}} = \emptyset$, for $\frac{1}{\max(f_\kappa)} \leqslant r < \frac{1}{\min(f_\kappa)}$, $S^1_{\geqslant \frac{1}{r}}$ is the arc from $-\alpha_{r,\kappa}$ to $\alpha_{r,\kappa}$ where $f_\kappa(\alpha_{r,\kappa}) = \frac{1}{r}$ and for $r \geqslant \frac{1}{\min(f_\kappa)}$, $S^1_{\geqslant \frac{1}{r}} = S^1$. It follows that

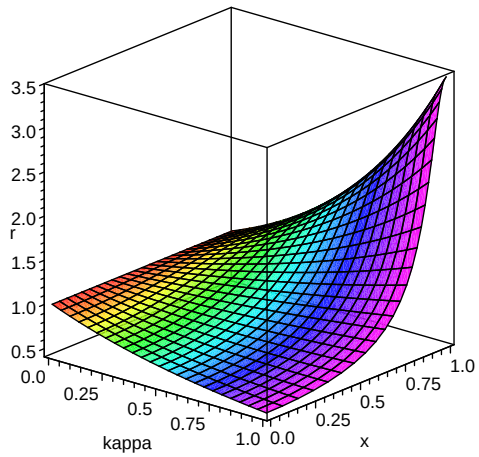

Figure 2: Graph of the Betta 0-function of the von Mises density for a range of concentration parameters.

for $k \geqslant 1$,

$$H_k(\mathcal{F}_r^{\check{C}}(C_*(S^1))) = \begin{cases} \mathbb{F} & \text{for } k = 1 \text{ and } r \geqslant \frac{1}{\min(f_\kappa)}, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore the Betti-1 barcode has the single interval

$$\left[\frac{1}{\min(f_\kappa)}, \infty\right] = [I_0(\kappa)e^\kappa, \infty] \tag{23}$$

and for $k > 1$ the Betti-k barcode is empty.

5.2. The von Mises-Fisher distribution

Now consider $\mathcal{M} = S^{p-1}$, $p \geqslant 3$ and the unimodal von Mises-Fisher density given by

$$f_{\mu,\kappa}(x) = c(\kappa) \exp\left\{\kappa x^t \mu\right\}, \quad x \in S^{p-1}$$

where $\kappa \in [0, \infty)$, $\mu \in S^{p-1}$, and

$$c(\kappa) = \left(\frac{\kappa}{2}\right)^{\frac{p}{2}-1} \frac{1}{\Gamma(\frac{p}{2}) I_{\frac{p}{2}-1}(\kappa)} \tag{24}$$

is the normalizing constant with respect to the uniform measure. This is also known as the Langevin distribution. Note that the minimum and maximum of f also do not depend on μ: $\min(f_\kappa) = c(\kappa)e^{-\kappa}$ and $\max(f_\kappa) = c(\kappa)e^\kappa$. In fact, by symmetry the homologies will not depend on μ. Hence once again take $\vartheta = \kappa$.

Consider the Morse filtration (defined in Section 3.3.1) on S^{p-1}. If $r < \min(f_\kappa)$, then $S^{p-1}_{\leqslant r} = \phi$ and if $r \geqslant \max(f_\kappa)$ then $S^{p-1}_{\leqslant r} = S^{p-1}$. For $\min(f_\kappa)) \leqslant r < \max(f_\kappa)$

$$S^{p-1}_{\leqslant r} = \{x \in S^{p-1} | x^t \mu \leqslant a_{r,\kappa}\},$$

where $a_{r,\kappa} = \frac{1}{\kappa}\ln\left(\frac{r}{c(\kappa)}\right) \in [-1,1]$. So $S^{p-1}_{\leqslant r}$ is the closure of S^{p-1} minus a right circular cone with vertex 0 and centered at μ. In particular, $S^{p-1}_{\leqslant r}$ is contractible (see Remark 2.7) so $H_0(\mathcal{F}^M_r(C_*(S^{p-1}))) = \mathbb{F}$ and for $k \geqslant 1$, $H_k(\mathcal{F}^M_r(C_*(S^{p-1}))) = 0$.

Thus the Betti-0 barcode is the single interval $[\min(f_\kappa), \infty)$, the Betti-$(p-1)$ barcode is the single interval $[\max(f_\kappa), \infty)$ and all other barcodes are empty.

Consider the Čech filtration (defined in Section 3.3.2) on S^{p-1}. For $\frac{1}{\max(f_\kappa)} \leqslant r < \frac{1}{\min(f_\kappa)}$,

$$S^{p-1}_{\geqslant \frac{1}{r}} = \{x \in S^{p-1} \mid x^t\mu \geqslant a_{\frac{1}{r},\kappa}\}.$$

So $S^{p-1}_{\geqslant \frac{1}{r}}$ is the intersection of S^{p-1} and a right circular cone with vertex 0 and centered at μ. In particular for $\frac{1}{\max(f_\kappa)} \leqslant r < \frac{1}{\min(f_\kappa)}$, $S^{p-1}_{\geqslant \frac{1}{r}}$ is contractible, so for $k \geqslant 1$, $H_k(S^{p-1}_{\geqslant \frac{1}{r}}) = 0$.

Assume $\kappa = 0$. Then $f_0 = c(0)$, and

$$S^{p-1}_{\geqslant \frac{1}{r}} = \begin{cases} \phi & \text{if } r < \frac{1}{c(0)}, \\ S^{p-1} & \text{if } r \geqslant \frac{1}{c(0)}. \end{cases}$$

Thus

$$g_\kappa(r) = \begin{cases} 0 & \text{if } r < \frac{1}{c(0)}, \\ 1 & \text{if } r \geqslant \frac{1}{c(0)}. \end{cases}$$

Therefore $\beta_0(x,0) := \inf_{g_\kappa(r) \geqslant x} r = \frac{1}{c(0)}$.

Assume $\kappa > 0$. Then for $k = 0$,

$$x = g_\kappa(r) = \int_{S^{p-1}_{\geqslant \frac{1}{r}}} f_\kappa = c(\kappa)\frac{s_{p-2}}{s_{p-1}} \int_0^{\arccos\left(-\frac{\ln(rc(\kappa))}{\kappa}\right)} e^{\kappa\cos\theta}\sin^{p-2}\theta\, d\theta, \tag{25}$$

where $s_{p-1} = \frac{2\pi^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2}\right)}$. When $\kappa > 0$, $g_\kappa(r)$ is continuous and strictly increasing. Hence

$$\beta_0(x,\kappa) = g_\kappa^{-1}(x) \tag{26}$$

for $x \in [0,1]$ and $\kappa > 0$. As we did for the von Mises distribution (22), we can describe the graph of $r = \beta_0(x,\kappa)$ more explicitly using a parametric equation:

$$h_\kappa(t) = \left(c(\kappa)\frac{s_{p-2}}{s_{p-1}}\int_0^t e^{\kappa\cos\theta}\sin^{p-2}\theta\, d\theta, \frac{e^{-\kappa\cos t}}{c(\kappa)}\right), \quad t \in [0,\pi]. \tag{27}$$

For $k \geqslant 1$, by Lemma 3.8,

$$H_k(\mathcal{F}^{\check{C}}_r(C_*(S^{p-1}))) = H_k(S^{p-1}_{\geqslant \frac{1}{r}}) = \begin{cases} \mathbb{F} & \text{if } k = p-1 \text{ and } r \geqslant \frac{1}{\min(f_\kappa)}, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore for $k \geqslant 1$ the Betti-k barcode has the single interval:

$$\left[\frac{1}{\min(f_\kappa)}, \infty\right] = \left[\frac{e^\kappa}{c(\kappa)}, \infty\right] \tag{28}$$

for $k = p-1$ and is empty otherwise.

5.3. The Watson distribution

Let $\mathcal{M} = S^{p-1}$ and consider the following bimodal distribution

$$f_{\mu,\kappa}(x) = d(\kappa)\exp\{\kappa(x^t\mu)^2\}, \tag{29}$$

where $\kappa \geqslant 0$ and $x, \mu \in S^{p-1}$, called the *Watson distribution*. We remark that this density is rotationally symmetric, where μ is the axis of rotation. The minimum and maximum densities are given by

$$\min f = d(\kappa), \quad \max f = d(\kappa)e^{\kappa}.$$

The maximum is achieved at $x = \pm\mu$ and the minimum is achieved at all x such that $x^t\mu = 0$.

Using the Morse filtration we get the following Betti barcodes. For $p = 2$, we remark that for $r < \min f$, $S^1_{\leqslant r} = \phi$. For $r = \min f$, $S^1_{\leqslant r}$ consists of two points. As r increases, these points become two arcs of increasing size, which connect when $r = \max f$. So the Betti-0 barcode consists of the homology intervals $[\min f, \infty]$ and $[\min f, \max f)$, and the Betti-1 barcode has the single interval $[\max f, \infty]$. All other Betti barcodes are empty.

For $p > 2$, we observe similar behavior. When $r < \min f$, $S^{p-1}_{\leqslant r} = \phi$. For $r = \min f$, $S^{p-1}_{\leqslant r}$ is the equator which is homeomorphic to S^{p-2}. As r increases, the equator expands until it reaches the poles when $r = \max f$. So the Betti-0, Betti-$(p-2)$ and Betti-$(p-1)$ barcodes each consist of a single homology interval: $[\min f, \infty]$, $[\min f, \max f)$, and $[\max f, \infty]$, respectively. All other Betti barcodes are empty.

Using the Čech filtration, $S^{p-1}_{\geqslant \frac{1}{r}}$ is either empty, or consists of two contractible components, or is all of S^{p-1}. So the Betti-$(p-1)$ barcode is the single homology interval $[\frac{1}{\min f}, \infty]$ and the Betti-k barcodes for all other $k \geqslant 1$ are empty. The Betti-0 function is given by $\beta_0(x, \kappa) = g_\kappa^{-1}(x)$, where

$$g_\kappa(r) = \int_{S^{p-1}_{\geqslant \frac{1}{r}}} f_\kappa = 2\frac{s_{p-2}}{s_{p-1}}\int_0^{\alpha_\kappa(r)} d(\kappa)e^{\kappa\cos^2(\theta)}\sin^{p-2}(\theta)d\theta,$$

with $\alpha_\kappa(r) = \cos^{-1}\left(\sqrt{-\frac{1}{\kappa}\ln(d(\kappa)r)}\right)$ and $s_{p-1} = \frac{2\pi^{p/2}}{\Gamma(p/2)}$. As with the von Mises (22) and von Mises-Fisher distributions (27), the Betti-0 function can also be described parametrically.

5.4. The Bingham distribution

Again let $\mathcal{M} = S^{p-1}$ with the probability density

$$f_K(x) = d(K)\exp\{x^tKx\},$$

where $x \in S^{p-1} \subset \mathbb{R}^3$ and K is a symmetric $p \times p$ matrix. We remark that $f_K(x) = d(K)\exp\{\operatorname{tr} Kxx^t\}$. By a change of coordinates we can write $K = \operatorname{diag}(k_1, \ldots, k_p)$, where $k_p \geqslant \ldots \geqslant k_1$ are the eigenvalues of K. Let v_i be the eigenvector associated to k_i.

Assume that $k_p > \ldots > k_1 > 0$. Then the minimum and maximum values of f_K

are given by

$$\min f_K = d(K)e^{k_1}, \quad \max f_K = d(K)e^{k_p},$$

and are attained at $\pm v_1$ and $\pm v_p$.

The Betti-k barcodes (for $k \geqslant 1$) when $p = 2$ are the same as for the Watson distribution. When $p \geqslant 3$, the Bingham distribution differs significantly from the Watson distribution. For example, the minimum of the function is attained at only $\pm v_1$ which is certainly not homeomorphic to S^{p-2}.

Consider the Morse filtration. We can calculate the Betti-k barcodes inductively. If we consider v_p to be the north pole, then there is a homotopy from $S^{p-1} - \{v_p, -v_p\}$ to S^{p-2} which collapses the sphere with missing its poles to the equator. When $r < k_p$, by the symmetry of f_K this homotopy also gives a homotopy from $S^{p-1}_{\leqslant r}$ to $S^{p-2}_{\leqslant r}$ where the filtration on S^{p-2} is the Morse filtration associated to the Bingham distribution with $K = \operatorname{diag}(k_1, \dots k_{p-1})$.

As a result, the Betti-0 barcode is given by the homology intervals $[d(K)e^{k_1}, \infty]$ and $[d(K)e^{k_1}, d(K)e^{k_2})$. For $1 \leqslant k \leqslant p-2$, the Betti-$i$ barcode is given by the interval $[d(K)e^{k_{i+1}}, d(K)e^{k_{i+2}})$. Finally, the Betti-$(p-1)$ barcode is given by the interval $[d(K)e^{k_p}, \infty]$.

We remark that this barcode corresponds to the cellular construction of S^{p-1} that repeatedly attaches northern and southern hemispheres of increasing dimension.

For the Čech filtration we can use the same argument starting with v_1. The Betti-0 barcode is given by the homology intervals $\frac{1}{d(K)}\left[e^{-k_p}, \infty\right]$ and $\frac{1}{d(K)}\left[e^{-k_p}, e^{-k_{p-1}}\right)$. For $1 \leqslant i \leqslant p-2$, the Betti-$i$ barcode is given by the interval $\frac{1}{d(K)}\left[e^{-k_{p-i}}, e^{-k_{p-i-1}}\right)$. The Betti-$(p-1)$ barcode is given by the single interval $\frac{1}{d(K)}\left[e^{-k_1}, \infty\right]$.

We remark that the correspondence between the two sets of barcodes is a manifestation of Alexander duality.

5.5. The matrix von Mises distribution and a Hopf fibration

The Lie group of rotations of $\mathbb{R}^3$, $SO(3)$, can be given the matrix von Mises density

$$f_{A,\kappa}(X) = c(\kappa) \exp\left\{\kappa \operatorname{tr}(X^t A)\right\}, \tag{30}$$

where $A \in SO(3)$ and $\kappa > 0$ is a concentration parameter. We determine the Morse and Čech filtrations of $SO(3)$ via the Hopf fibration $S^3 \to \mathbb{RP}^3$.

The special orthogonal group $SO(3)$ is diffeomorphic to the real projective space $\mathbb{RP}^3$. The map $S^3 \to \mathbb{RP}^3$ which identifies each point on the sphere with the one-dimensional subspace on which it lies is a Hopf fibration whose fiber is $S^0 = \{-1, 1\}$. Thus, S^3 is a double-cover of $SO(3)$ (and since S^3 is simply-connected, it is the universal cover).

If we represent S^3 with the unit quaternions and $\mathbb{RP}^3$ with $SO(3)$, then the Hopf fibration above is represented by the Cayley-Klein map $\rho\colon S^3 \to SO(3)$:

$$\rho\begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix} = I + 2p_1 B + 2B^2, \text{ where } B = \begin{pmatrix} 0 & -p_4 & p_3 \\ p_4 & 0 & -p_2 \\ -p_3 & p_2 & 0 \end{pmatrix}.$$

We can use this map to relate the matrix von Mises density (30) on $SO(3)$ to the

Watson density (29) on S^3 by making the following observation. If $P = \rho(p)$ and $Q = \rho(q)$, then

$$\mathrm{tr}(P^t Q) = 4(p^t q)^2 - 1.$$

Then if $\rho(a) = A$,

$$\rho^{-1}\{X \in SO(3) \mid f_{A,\kappa}(x) = r\} = \{x \in S^3 \mid f_{a,4\kappa}(x) = kr\}, \text{ where } k = \frac{d(4\kappa)e^{\kappa}}{c(\kappa)}.$$

It follows that

$$\rho^{-1}(SO(3)_{\leqslant r}) = S^3_{\leqslant kr} \text{ and } \rho^{-1}(SO(3)_{\geqslant \frac{1}{r}}) = S^3_{\geqslant k\frac{1}{r}},$$

where the filtration on S^3 is with respect to Watson density $f_{a,4\kappa}$.

Recall (Section 5.3) that for $\frac{1}{\max f} \leqslant kr < \frac{1}{\min f}$, $S^3_{\geqslant \frac{1}{kr}}$ consists of two contractible components. The Hopf fibration $S^3 \to \mathbb{RP}^3$ and equivalently the map $\rho\colon S^3 \to SO(3)$ identify these two components. So $SO(3)_{\geqslant \frac{1}{r}}$ is contractible. Therefore, for the Čech filtration the Betti-3 barcode is the single homology interval $[\frac{1}{\min f}, \infty)$ and all other Betti-k barcodes for $k \geqslant 1$ are empty. The Betti-0 function is identical to the one for the Watson density on S^3.

For $\min f \leqslant kr < \max f$, $S^3_{\leqslant kr}$ is homotopy equivalent (via a projection onto its equator) to S^2. The Hopf fibration $S^3 \to \mathbb{RP}^3$ restricted to the equator gives the Hopf fibration and double cover $S^2 \to \mathbb{RP}^2$. The homotopy equivalences $S^3_{\leqslant kr} \approx S^2$ induces a homotopy equivalence $SO(3)_{\leqslant r} \approx \mathbb{RP}^2$. Thus for the Morse filtration, the Betti-0 and Betti-3 barcodes are the single homology intervals $[\min f, \infty)$ and $[\max f, \infty)$ and all Betti-k barcodes for $k > 3$ are empty. However, since the fundamental group and integral homology group of degree one of $\mathbb{RP}^2$ are the cyclic group of order two, the Betti-1 and Betti-2 barcodes depend on the choice of the field of coefficients $\mathbb{F}$. If $\mathbb{F}$ is a field of characteristic 0 (e.g. the rationals) then both are empty. However if $\mathbb{F}$ is the field of characteristic two ($\mathbb{Z}/2\mathbb{Z}$), then both are the single homology interval $[\min f, \max f)$.

6. Statistical estimation of the Betti barcodes

In this section we will calculate the expected persistent homology using statistics sampled from various densities.

6.1. The von-Mises and von-Mises Fisher distributions

For point cloud data $x_1, \ldots, x_n$ on S^{p-1} sampled from the von Mises-Fisher distribution (3): $f_{\mu,\kappa}(x) = c(\kappa)\exp\{\kappa x^t \mu\}$, we will give the statistical estimators for the (unknown) parameters. We will show that these can be used to obtain good estimates of the persistent homology of the underlying distribution.

Letting $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ denote the sample mean, consider the decomposition

$$\bar{x} = \|\bar{x}\| \left(\frac{\bar{x}}{\|\bar{x}\|}\right).$$

The statistical estimator for μ is $\bar{x}/\|\bar{x}\|$ while the statistical estimator for κ is solved [**MJ00**, Section 10.3.1] by inverting $A_p(\hat{\kappa}) = \|\bar{x}\|$, where $A_p(\lambda) = \frac{I_{p/2}(\lambda)}{I_{p/2-1}(\lambda)}$ and $I_\nu(\lambda)$

is the modified Bessel function of the first kind and order ν. Hence,

$$\hat{\kappa} = A_p^{-1}(\|\bar{x}\|). \tag{31}$$

A large sample asymptotic normality calculation for (31) is [**MJ00**, Section 10.3.1]

$$\sqrt{n}\,(\hat{\kappa} - \kappa) \rightsquigarrow N\left(0, A_p'(\kappa)^{-1}\right), \tag{32}$$

as $n \to \infty$, where $\rightsquigarrow$ means convergence in distribution and $N(0, \sigma^2)$ stands for a normally distributed random variable with mean 0 and variance $\sigma^2 > 0$. Using this estimate of κ we obtain estimates for the β_κ barcodes for the Morse and Čech filtrations. For the Morse filtration, we estimate the β_0 barcode and β_{p-1} barcode to be $[c(\hat{\kappa})e^{-\hat{\kappa}}, \infty]$ and $[c(\hat{\kappa})e^{\hat{\kappa}}, \infty]$, respectively. For the Čech filtration, we estimate the β_{p-1} barcode to be $[\frac{e^{\hat{\kappa}}}{c(\hat{\kappa})}, \infty]$.

Recall that the space of barcodes has a metric $\mathcal{D}$ (see Definition 3.5). Let $\beta_i^M(f)$ and $\beta_i^{\check{C}}(f)$ denote the Betti-i barcode for the density f using the Morse and Čech filtrations. Then the expectations of the distance from the estimated persistent homology to the persistent homology of the underlying density can be bounded as follows.

Theorem 6.1. *For the von Mises-Fisher distribution on S^{p-1} and $\kappa \in [\kappa_0, \kappa_1]$, where $0 < \kappa_0 \leqslant \kappa_1 < \infty$,*

$$E(\mathcal{D}(\beta_i^M(f_{\hat{\kappa}}), \beta_i^M(f_\kappa))) \leqslant C(\kappa)n^{-1/2}$$

as $n \to \infty$ for all i, and

$$E(\mathcal{D}(\beta_i^{\check{C}}(f_{\hat{\kappa}}), \beta_i^{\check{C}}(f_\kappa))) \leqslant C(\kappa)n^{-1/2}$$

as $n \to \infty$ for all $i \geqslant 1$, for some constant $C(\kappa)$.

Proof. Since the barcodes have a particularly simple form, we only need to know the barcode metric for the following case:

$$\mathcal{D}(\{[a, \infty]\}, \{[b, \infty]\}) = |a - b|.$$

Using our previous calculations of the Betti barcodes, we have:

$$\begin{aligned}
\mathcal{D}(\beta_0^M(f_{\hat{\kappa}}), \beta_0^M(f_\kappa)) &= |c(\hat{\kappa})e^{-\hat{\kappa}} - c(\kappa)e^{-\kappa}| \\
\mathcal{D}(\beta_{p-1}^M(f_{\hat{\kappa}}), \beta_{p-1}^M(f_\kappa)) &= |c(\hat{\kappa})e^{\hat{\kappa}} - c(\kappa)e^{\kappa}| \\
\mathcal{D}(\beta_{p-1}^{\check{C}}(f_{\hat{\kappa}}), \beta_{p-1}^{\check{C}}(f_\kappa)) &= |c(\hat{\kappa})^{-1}e^{\hat{\kappa}} - c(\kappa)^{-1}e^{\kappa}|.
\end{aligned}$$

We note that the normalizing constant can be re-expressed as

$$c(\kappa) = \frac{B\left(\frac{p-1}{2}, \frac{1}{2}\right)}{\int_{-1}^{1} e^{\kappa t}(1-t^2)^{\frac{p-3}{2}}\,dt},$$

where $B(\cdot, \cdot)$ is the beta function. Furthermore,

$$c'(\kappa) = -B\left(\frac{p-1}{2}, \frac{1}{2}\right) \frac{\int_{-1}^{1} e^{\kappa t} t(1-t^2)^{\frac{p-3}{2}}\,dt}{\left(\int_{-1}^{1} e^{\kappa t}(1-t^2)^{\frac{p-3}{2}}\,dt\right)^2}$$

and

$$A_p'(\kappa) = 1 - A_p(\kappa)^2 - \frac{p-1}{\kappa}A_p(\kappa).$$

For $0 \leqslant \kappa_0 \leqslant \kappa_1 < \infty$ and $\kappa \in [\kappa_0, \kappa_1]$, we observe $0 < c(\kappa), |c'(\kappa)|, A'_p(\kappa) < \infty$, and by the mean value theorem,

$$E|c(\hat{\kappa})e^{\hat{\kappa}} - c(\kappa)e^{\kappa}| = E|(c(\kappa^*) + c'(\kappa^*))e^{\kappa^*}(\hat{\kappa} - \kappa)|,$$

where κ^* is a value between $\hat{\kappa}$ and κ. Consequently,

$$\begin{aligned} E|c(\hat{\kappa})e^{\hat{\kappa}} - c(\kappa)e^{\kappa}| &\leqslant \bar{C}(\kappa)\left\{E|\hat{\kappa} - \kappa|^2\right\}^{1/2} \\ &\leqslant C(\kappa)n^{-1/2} \end{aligned}$$

where the first inequality is by the Hölder inequality, and the second is by (32).

Similarly,

$$E|c(\hat{\kappa})e^{-\hat{\kappa}} - c(\kappa)e^{-\kappa}| = E|(c'(\kappa^*) - c(\kappa^*))e^{-\kappa^*}(\hat{\kappa} - \kappa)|,$$

and

$$E\left|\frac{e^{\hat{\kappa}}}{c(\hat{\kappa})} - \frac{e^{\kappa}}{c(\kappa)}\right| = E\left|\left(\frac{c(\kappa^*) - c'(\kappa^*)}{c(\kappa^*)^2}\right)e^{\kappa^*}(\hat{\kappa} - \kappa)\right|. \qquad \square$$

Expressing the estimated β_0-function is more challenging. For the case of the sphere S^2, an exact expression can be obtained. One can calculate that $c(\kappa) = \frac{\kappa}{\sinh(\kappa)}$, and from (25),

$$g_\kappa(r) = \frac{e^\kappa}{2\sinh(\kappa)} - \frac{1}{2\kappa r},$$

from which we use (26) to obtain,

$$\beta_0(x, \kappa) = \frac{e^{2\kappa} - 1}{2\kappa[(1 - x)e^{2\kappa} + x]} \tag{33}$$

for $x \in (0, 1]$ and $\kappa > 0$. Notice that $\beta_0(x, \kappa) \to 1$ as $\kappa \to 0$ and $\beta_0(x, \kappa) \to 0$ as $\kappa \to \infty$, for all $x \in (0, 1)$. Furthermore, for (31), [**MJ00**, 9.3.9]

$$A_3(\kappa) = \coth\kappa - \tfrac{1}{\kappa}. \tag{34}$$

We have the following:

Theorem 6.2. *For the von Mises-Fisher distribution on S^2, and fixed $\kappa > 0$,*

$$E\left|\left|\beta_0(x, \hat{\kappa}) - \beta_0(x, \kappa)\right|\right|_\infty \leqslant C(\kappa)n^{-1},$$

as $n \to \infty$.

Proof. By the mean value theorem,

$$\beta_0(x, \hat{\kappa}) - \beta_0(x, \kappa) = \frac{\partial}{\partial\kappa}\beta_0(x, \tilde{\kappa})(\hat{\kappa} - \kappa), \tag{35}$$

where $\tilde{\kappa}$ is between $\hat{\kappa}$ and κ. One can calculate that

$$\frac{\partial}{\partial\kappa}\beta_0(x, \kappa) = \frac{-(1 - x)e^{4\kappa} + (1 + 2\kappa - 2x)e^{2\kappa} + x}{2\kappa^2\left[(1 - x)e^{2\kappa} + x\right]^2}.$$

Recall that the domain of $\beta_0(x, \kappa)$ is $(0, 1]$. For $x \in (0, 1]$, $\left|\frac{\partial}{\partial\kappa}\beta_0(x, \kappa)\right|$ is bounded: for

instance,

$$\left|\frac{\partial}{\partial\kappa}\beta_0(x,\kappa)\right| \leqslant \frac{e^{4\kappa}+(1+2\kappa)e^{2\kappa}+1}{2\kappa^2}. \tag{36}$$

Combining (35), (36), (32) and (34) produces the desired result. □

6.2. The Watson distribution

Recall that the Watson distribution on S^{p-1} is given by

$$f_{\mu,\kappa}(x) = d(\kappa)\exp\{\kappa(x^t\mu)^2\}, \text{ where } \mu \in S^{p-1} \text{ and } \kappa > 0. \tag{37}$$

Let us parametrize μ using the spherical angles: $\mu = \mu(\phi)$, where $\phi = (\phi_1, \ldots, \phi_{p-1})^t$. Let $X_1, \ldots X_n$ be a random sample from the Watson distribution.

If we take the sample to be fixed and the underlying parameters to be unknown, then the log-likelihood function of (37) is given by:

$$\ell(\phi,\kappa) = n\log d(\kappa) + \kappa\sum_{j=1}^{n}(X_j^t\mu(\phi))^2.$$

The maximum likelihood estimation of μ and κ comes from the estimating equation:

$$\nabla_{\phi,\kappa}\ell(\phi,\kappa) = 0, \tag{38}$$

where $\nabla_{\phi,\kappa}$ denotes the gradient. Let $\hat{\phi}$ and $\hat{\kappa}$ be the solutions to (38), which are the maximum likelihood estimators. Then the standard theory of maximum likelihood estimators [**CH74**, pp.294-296] shows that the large sample asymptotics satisfy:

$$\sqrt{n}\left[\begin{pmatrix}\hat{\phi}\\ \hat{\kappa}\end{pmatrix} - \begin{pmatrix}\phi\\ \kappa\end{pmatrix}\right] \to_d N_p(0, I(\phi,\kappa)^{-1}) \tag{39}$$

as $n \to \infty$, where "$\to_d$" means convergence in distribution, $I(\phi,\kappa)$ is the Fisher information matrix[7] and N_p stands for the p-dimensional normal distribution with given mean and covariance. It turns out that in the case of the Watson distribution,

$$I(\phi,\kappa) = \left[\begin{array}{c|c} * & 0 \\ \hline 0 & -\frac{\partial^2}{\partial\kappa^2}\log d(\kappa)\end{array}\right].$$

Consequently, from (39), we have that

$$\sqrt{n}(\hat{\kappa}-\kappa) \to_d N_1\left(0, -\left(\frac{\partial^2}{\partial\kappa^2}\log d(\kappa)\right)^{-1}\right),$$

as $n \to \infty$.

[7]The Fisher information matrix is defined to be $I(\phi,\kappa) = -E\nabla^2_{\phi,\kappa}\ell(\phi,\kappa)$, where $\nabla^2_{\phi,\kappa}$ is the $p \times p$ Hessian matrix.

References

[**AK05**] J.-F. Angers and Peter T. Kim, Multivariate Bayesian function estimation, *Ann. Statist.* **33** (2005), 2967–2999.

[**Ber85**] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, second edition, Springer Series in Statistics, Springer-Verlag, New York, 1985.

[**CCSL06**] F. Chazal, D. Cohen-Steiner, and A. Lieutier, A sampling theory for compacts in Euclidean space, in *SoCG'06: Proc. of the Twenty-Second Annual Symposium on Computational Geometry, 2006*, 319–326, ACM Press, New York, 2006.

[**CH74**] D. R. Cox and D. V. Hinkley, *Theoretical Statistics*, Chapman and Hall, London, 1974.

[**Chi03**] Y. Chikuse, *Statistics on Special Manifolds*, Lecture Notes in Statistics **174**, Springer-Verlag, New York, 2003.

[**CSEH05**] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer, Stability of persistence diagrams, in *SCG '05: Proc. of the Twenty-First Annual Symposium on Computational Geometry*, pp. 263–271, ACM Press, New York, NY, 2005.

[**CZCG04**] G. Carlsson, A. Zomorodian, A. Collins, and L. Guibas, Persistence barcodes for shapes, in *SGP '04: Proc. of the 2004 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing*, pp. 124–135, ACM Press, New York, NY, 2004.

[**Dev81**] L. Devroye, Laws of the iterated logarithm for order statistics of uniform spacings, *Ann. Probab.* **9** (1981), 860–867.

[**dSC04**] V. de Silva and G. Carlsson, Topological estimation using witness complexes, in *SPBG'04 Symposium on Point-Based Graphics 2004*, pp. 157–166, 2004.

[**Efr00**] S. Efromovich, On sharp adaptive estimation of multivariate curves, *Math. Methods Statist.* **9** (2000), 117–139.

[**ELZ02**] H. Edelsbrunner, D. Letscher, and A. Zomorodian, Topological persistence and simplification, in *Discrete and Computational Geometry and Graph Drawing* (Columbia, SC, 2001), *Discrete Comput. Geom.* **28** (2002), 511–533.

[**FGK03**] K. Fischer, B. Gaertner, and M. Kutz, Fast smallest-enclosing-ball computation in high dimensions, in *Algorithms - ESA 2003*, *Lecture Notes in Computer Science* **2832**, pp. 630–641, Springer-Verlag, New York, 2003.

[**Gae06**] B. Gaertner, Smallest enclosing ball code, `http://www.inf.ethz.ch/personal/gaertner/miniball.html`, 2006.

[**Hen90**] H. Hendriks, Nonparametric estimation of a probability density on a Riemannian manifold using Fourier expansions, *Ann. Statist.* **18** (1990), 832–849.

[KK00] P. T. Kim and J.-Y. Koo, Directional mixture models and optimal estimation of the mixing density, *Canad. J. Statist.* **28** (2000), 383–398.

[Mil63] J. Milnor, *Morse Theory*, Based on lecture notes by M. Spivak and R. Wells, *Ann. of Math. Studies*, No. 51, Princeton University Press, Princeton, NJ, 1963.

[MJ00] K. V. Mardia and P. E. Jupp, *Directional Statistics*, Wiley Series in Probability and Statistics, John Wiley & Sons Ltd., Chichester, 2000; Revised reprint of *Statistics of Directional Data* by Mardia [MR0336854 (49 #1627)].

[NSW05] P. Niyogi, S. Smale, and S. Weinberger, Finding the homology of submanifolds with high confidence from random samples, *Discrete Computat. Geom.*, to appear. DOI:10.1007/s00454-006-1250-7.

[Pyk65] R. Pyke, Spacings (with discussion), *J. Royal Statist. Soc. Ser. B* **27** (1965), 395–449.

[SW86] G. R. Shorack and J. A. Wellner, *Empircal Processes with Applications to Statistics*, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons Inc., New York, 1986.

[ZC05] A. Zomorodian and G. Carlsson, Computing persistent homology, *Discrete Comput. Geom.* **33** (2005), 249–274.

Peter Bubenik `p.bubenik@csuohio.edu`

Department of Mathematics
Cleveland State University
2121 Euclid Ave. RT 1515
Cleveland OH 44115-2214
USA

Peter T. Kim `pkim@uoguelph.ca`

Department of Mathematics and Statistics
University of Guelph
Guelph, Ontario N1G 2W1
Canada

This article is available at `http://intlpress.com/HHA/v9/n2/a12`