Sparse Circular Coordinates via Principal $\mathbb{Z}$-Bundles

Jose A. Perea

Abstract We present in this paper an application of the theory of principal bundles to the problem of nonlinear dimensionality reduction in data analysis. More explicitly, we derive, from a 1-dimensional persistent cohomology computation, explicit formulas for circle-valued functions on data with nontrivial underlying topology. We show that the language of principal bundles leads to coordinates defined on an open neighborhood of the data, but computed using only a smaller subset of landmarks. It is in this sense that the coordinates are sparse. Several data examples are presented, as well as theoretical results underlying the construction.

1 Introduction

The curse of dimensionality refers to a host of phenomena inherent to the increase in the number of features describing the elements of a data set. For instance, in statistical learning, the number of training data points needs to grow roughly exponentially in the number of features, in order for learning algorithms to generalize correctly in the absence of other priors. A deeper manifestation of the curse of dimensionality is the deterioration of the concept of "nearest neighbors" in high-dimensional Euclidean space; for as the dimension increases, the distance between any two points is roughly the same [18]. One of the most popular priors in data science is the "low intrinsic dimensionality" hypothesis. It contends that while the apparent number of features describing each data point (e.g., the number of pixels in an image) might be large, the effective number of degrees of freedom (i.e., the intrinsic dimensionality) is often much lower. Indeed, images generated at random will hardly depict a cat or a natural scene.

J. A. Perea (✉)
Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, MI, USA

Department of Mathematics, Michigan State University, East Lansing, MI, USA
e-mail: joperea@msu.edu

N. A. Baas et al. (eds.), *Topological Data Analysis*, Abel Symposia 15,
https://doi.org/10.1007/978-3-030-43408-3_17

Many dimensionality reduction schemes have been proposed in the literature to leverage the "low intrinsic dimensionality" hypothesis, each making explicit or implicit use of likely characteristics of the data. For instance, Principal Component Analysis [10] and other linear dimensionality reduction methods, rely on the existence of a low-dimensional linear representation accounting for most of the variability in the data. Methods such as Locally Linear Embeddings [19] and Laplacian EingenMaps [2], on the other hand, presuppose the existence of a manifold-like object parametrizing the underlying data space. Other algorithms, like Multidimensional Scaling [11] and Isomap [22], attempt to preserve distances between data points while providing low-dimensional reconstructions.

Recently, several new methods for nonlinear dimensionality reduction have emerged from the field of computational topology [6, 17, 21]. The idea being that if the underlying space from which the data has been sampled has a particular shape, then this information can be used to generate appropriate low-dimensional representations. The circular coordinates of de Silva, Morozov, and Vejdemo-Johansson [6] pioneered the use of persistent cohomology as a way to measure the shape of a data set, and then produce circle-valued coordinates reflecting the underlying nontrivial topology. Their algorithm goes as follows. Given a finite metric space $(X, \mathbf{d})$—the data—and a scale $\alpha > 0$ so that the Rips complex

$$R_\alpha(X) := \{\sigma \subset X : \sigma \neq \emptyset \quad \text{and} \quad \mathsf{diam}(\sigma) < \alpha\}$$

has a nontrivial integer cohomology class $[\eta] \in H^1(R_\alpha(X); \mathbb{Z})$ — this is determined from the persistent cohomology of the Rips filtration $\mathcal{R}(X) = \{R_\epsilon(X)\}_{\epsilon \geq 0}$—a linear least squares optimization (of size the number of vertices by the number of edges of $R_\alpha(X)$) is solved, in order to construct a function $f_\eta : X \longrightarrow S^1 \subset \mathbb{C}$ which, roughly, puts one of the generators of $H^1(S^1; \mathbb{Z}) \cong \mathbb{Z}$ in correspondence with $[\eta] \in H^1(R_\alpha(X); \mathbb{Z})$.

1.1 Our Contribution

Two drawbacks of the perspective presented in [6] are: (1) the method requires a persistent cohomology calculation, as well as a least squares optimization, on the Rips filtration of the entire data set X. This is computationally expensive and may limit applicability to small-to-medium-sized data. (2) once the function f_η has been computed, it is only defined on the data points from X used for its construction. Here we show that these drawbacks can be addressed effectively with ideas from principal $\mathbb{Z}$-bundles. In particular, we show that it is possible to construct circular coordinates on X from the Rips filtration on a subset of landmarks $L \subset X$, Proposition 4, with similar classifying properties as in [6], Theorem 3, and that said coordinates will be defined on an open neighborhood of L containing X. We call these functions "sparse circular coordinates".

1.2 The Sparse Circular Coordinates Algorithm

Let us describe next the steps needed to construct said coordinates. The rest of the paper is devoted to the theory behind these choices:

1. Let $(X, \mathbf{d})$ be the input data set; i.e. a finite metric space. Select a set of landmarks $L = \{\ell_1 \ldots, \ell_N\} \subset X$, e.g. at random or via `maxmin` sampling, and let

$$r_L := \max_{x \in X} \min_{\ell \in L} \mathbf{d}(x, \ell)$$

   be the radius of coverage. In particular, r_L is the Hausdorff distance between L and X.
2. Choose a prime $q > 2$ at random and compute the 1-dimensional persistent cohomology $PH^1(\mathcal{R}(L); \mathbb{Z}/q)$ with coefficients in $\mathbb{Z}/q$, for the Rips filtration on the landmark set L. Let $\mathsf{dgm}(L)$ be the resulting persistence diagram.
3. If there exists $(a, b) \in \mathsf{dgm}(L)$ so that $\max\{a, r_L\} < \frac{b}{2}$, then let

$$\alpha = t \cdot \max\{a, r_L\} + (1-t)\frac{b}{2} \quad , \quad \text{for some} \quad 0 < t < 1$$

   Let $\eta' \in Z^1(R_{2\alpha}(L); \mathbb{Z}/q)$ be a cocyle representative for the persistent cohomology class corresponding to $(a, b) \in \mathsf{dgm}(L)$. If t is closer to 1, then the circular coordinates are defined on a larger domain; however, this makes step (5) below more computationally intensive.
4. Lift $\eta' : C_1(R_{2\alpha}(L); \mathbb{Z}) \longrightarrow \mathbb{Z}/q = \{0, \ldots, q-1\}$ to an integer cocycle $\eta \in Z^1(R_{2\alpha}(L); \mathbb{Z})$. That is, one for which $\eta' - (\eta \bmod q)$ is a coboundary in $C^1(R_{2\alpha}(L); \mathbb{Z}/q)$. An explicit choice (that works in practice for a prime q chosen at random) is the integer cochain:

$$\eta(\sigma) = \begin{cases} \eta'(\sigma) & \text{if } \eta'(\sigma) \leq \frac{q-1}{2} \\ \eta'(\sigma) - q & \text{if } \eta'(\sigma) > \frac{q-1}{2} \end{cases}$$

5. Choose positive weights for the vertices and edges of $R_{2\alpha}(L)$—e.g. all equal to one—and let $d^+_{2\alpha} : C^1(R_{2\alpha}(L); \mathbb{R}) \longrightarrow C^0(R_{2\alpha}(L); \mathbb{R})$ be the (weighted) Moore-Penrose pseudoinverse (solving weighted linear least squares problems) for the coboundary map

$$d_{2\alpha} : C^0(R_{2\alpha}(L); \mathbb{R}) \longrightarrow C^1(R_{2\alpha}(L); \mathbb{R})$$

   If $\iota : \mathbb{Z} \hookrightarrow \mathbb{R}$ is the inclusion, let

$$\tau = -d^+_{2\alpha}\big(\iota \circ \eta\big) \qquad \text{and} \qquad \theta = (\iota \circ \eta) + d_{2\alpha}(\tau)$$

6. Denote by $\tau_j \in \mathbb{R}$ the value of τ on the vertex $\ell_j \in L$, and by $\theta_{jk} \in \mathbb{R}$ the value of θ on the oriented edge $[\ell_j, \ell_k] \in R_{2\alpha}(L)$. If we let

$$\varphi_j(b) = \frac{|\alpha - \mathbf{d}(\ell_j, b)|_+}{\sum\limits_{k=1}^{N} |\alpha - \mathbf{d}(\ell_k, b)|_+} \qquad \text{where} \qquad |r|_+ = \max\{r, 0\}, \quad r \in \mathbb{R}$$

and $B_\alpha(\ell_k)$ denotes the open ball of radius $\alpha > 0$ centered at $\ell_k \in L$, then the sparse circular coordinates are defined by the formula:

$$\boxed{\begin{aligned} h_{\theta,\tau} : \bigcup_{k=1}^{N} B_\alpha(\ell_k) &\longrightarrow S^1 \subset \mathbb{C} \\ B_\alpha(\ell_j) \ni b &\mapsto \exp\left\{2\pi i \left(\tau_j + \sum_{k=1}^{N} \varphi_k(b)\theta_{jk}\right)\right\} \end{aligned}} \tag{1}$$

If X is a subspace of an ambient metric space $\mathbb{M}$, then the $B_\alpha(\ell_k)$'s can be taken to be ambient metric balls. This is why we call the circular coordinates *sparse*; $h_{\theta,\tau}$ is computed using only L, but its domain of definition is an open subset of $\mathbb{M}$ which, by construction, contains all of X.

1.3 Organization

We start in Sect. 2 with a few preliminaries on principal bundles, highlighting the main theorems needed in later parts of the paper. We assume familiarity with persistent cohomology (if not, see [16]), as well as the definition of Čech cohomology with coefficients in a presheaf (see for instance [14]). Section 3 is devoted to deriving the formulas—e.g. (1) above—which turn a 1-dimensional integer cohomology class into a circle-valued function. In Sect. 4 we describe how to make all this theory applicable to real data sets. We present several experiments in Sect. 5 with both real and synthetic data, and end in Sect. 6 with a few final remarks.

2 Preliminaries

2.1 Principal Bundles

We present here a terse introduction to principal bundles, with the main results we will need later in the paper. In particular, the connection between principal bundles and Čech chomology, which allows for explicit computations, and their classification theory via homotopy classes of maps to classifying spaces. The latter description will be used to generate our sparse circular coordinates. We refer the interested reader to [9] for a more thorough presentation.

Let B be a connected and paracompact[1] topological space with basepoint $b_0 \in B$.

Definition 1 A pair (p, E), with E a topological space and $p : E \longrightarrow B$ a continuous map, is said to be a fiber bundle over B with fiber $F = p^{-1}(b_0)$, if:

1. p is surjective
2. Every point $b \in B$ has an open neighborhood $U \subset B$ and a homeomorphism $\rho_U : U \times F \longrightarrow p^{-1}(U)$, called a local trivialization around b, so that $p \circ \rho_U(b', e) = b'$ for every $(b', e) \in U \times F$.

The spaces E and B are called, respectively, the total and base space of the bundle, and p is called the projection map.

Definition 2 Let G be an abelian topological group whose operation we write additively. A fiber bundle $p : E \longrightarrow B$ is said to be a principal G-bundle if:

1. The total space E comes equipped with a fiberwise free right G-action. That is, a continuous map

$$\cdot : E \times G \longrightarrow E$$

   satisfying the right-action axioms, with $p(e \cdot g) = p(e)$ for every pair $(e, g) \in E \times G$, and for which $e \cdot g = e$ only if g is the identity of G.
2. The induced fiberwise G-action $p^{-1}(b) \times G \longrightarrow p^{-1}(b)$ is transitive for every $b \in B$ in the base space.
3. The local trivializations $\rho_U : U \times F \longrightarrow p^{-1}(U)$ can be chosen to be G-equivariant: that is, so that $\rho_U(b, e \cdot g) = \rho_U(b, e) \cdot g$, for every $(b, e, g) \in U \times F \times G$.

Two principal G-bundles $p_j : E_j \longrightarrow B$, $j = 1, 2$, are said to be isomorphic, if there exists a G-equivariant homeomorphism $\Phi : E_1 \longrightarrow E_2$ so that $p_2 \circ \Phi = p_1$. This defines an equivalence relation on principal G-bundles over B, and the set of isomorphism classes is denoted $\mathsf{Prin}_G(B)$.

Given a principal G-bundle $p : E \longrightarrow B$ and a system of (G-equivariant) local trivializations $\left\{\rho_j : U_j \times F \longrightarrow p^{-1}(U_j)\right\}_{j \in J}$, we have that

$$\rho_k^{-1} \circ \rho_j : (U_j \cap U_k) \times F \longrightarrow (U_j \cap U_k) \times F$$

is a G-equivariant homeomorphism whenever $U_j \cap U_k \neq \emptyset$. Since the G-action on E is fiberwise free and fiberwise transitive, then $\rho_k^{-1} \circ \rho_j$ induces a well-defined continuous map

$$\rho_{jk} : U_j \cap U_k \longrightarrow G \qquad j, k \in J \tag{2}$$

[1] So that partitions of unity always exist.

defined by the equation

$$\rho_k^{-1} \circ \rho_j(b, e) = (b, e \cdot \rho_{jk}(b)) \quad , \quad \text{for all } (b, e) \in (U_j \cap U_k) \times F. \tag{3}$$

The ρ_{jk}'s are called the transition functions for the G-bundle (p, E) corresponding to the system of local trivializations $\{\rho_j\}_{j\in J}$. In fact, these transition functions define an element in the Čech cohomology of B. Indeed, for each open set $U \subset B$ let $\mathsf{Maps}(U, G)$ denote the set of continuous maps from U to G. Since G is an abelian group, then so is $\mathsf{Maps}(U, G)$, and if $V \subset U$ is another open set, then precomposing with the inclusion $V \hookrightarrow U$ yields a restriction map

$$\iota_{U,V} : \mathsf{Maps}(U, G) \longrightarrow \mathsf{Maps}(V, G)$$

This defines a sheaf $\mathscr{C}_G$ of abelian groups over B, with $\mathscr{C}_G(U) := \mathsf{Maps}(U, G)$, called the sheaf of G-valued continuous functions on B. It follows that the transition functions (2) define an element $\rho = \{\rho_{jk}\} \in \check{C}^1(\mathcal{U}; \mathscr{C}_G)$ in the Čech 1-cochains of the cover $\mathcal{U} = \{U_j\}_{j\in J}$ with coefficients in the sheaf $\mathscr{C}_G$. Moreover,

Proposition 1 *The transition functions ρ_{jk} satisfy the cocycle condition*

$$\rho_{j\ell}(b) = (\rho_{jk} + \rho_{k\ell})(b) \quad \textit{for all} \quad b \in U_j \cap U_k \cap U_\ell \tag{4}$$

In other words, $\rho = \{\rho_{jk}\} \in \check{Z}^1(\mathcal{U}; \mathscr{C}_G)$ is a Čech cocycle.

If $\{\nu_r : V_r \times F \longrightarrow p^{-1}(V_r)\}_{r\in R}$ is another system of local trivializations with induced Čech cocycle $\nu = \{\nu_{rs}\} \in \check{Z}^1(\mathcal{V}; \mathscr{C}_G)$, and

$$\mathcal{W} = \{U_j \cap V_r\}_{(j,r)\in J\times R}$$

then one can check that the difference $\rho - \nu$ is a coboundary in $\check{C}^1(\mathcal{W}; \mathscr{C}_G)$. Since $\mathcal{W}$ is a refinement for both $\mathcal{V}$ and $\mathcal{U}$, it follows that the G-bundle $p : E \longrightarrow B$ yields a well-defined element $p_E \in \check{H}^1(B; \mathscr{C}_G)$. Moreover, after passing to isomorphism classes of principal G-bundles we get that

Lemma 1 *The function*

$$\begin{array}{rcl} \mathsf{Prin}_G(B) & \longrightarrow & \check{H}^1(B; \mathscr{C}_G) \\ [(p, E)] & \mapsto & p_E \end{array}$$

is well-defined and injective.

This is in fact a bijection. To check surjectivity, fix an open cover $\mathcal{U} = \{U_j\}_{j\in J}$ for B, and a Čech cocyle

$$\eta = \{\eta_{jk}\} \in \check{Z}^1(\mathcal{U}; \mathscr{C}_G)$$

Then one can construct a principal G-bundle over B with total space

$$E_\eta = \left(\bigcup_{j \in J} U_j \times \{j\} \times G \right) \Big/ (b, j, g) \sim \big(b, k, g + \eta_{jk}(b)\big) \quad , \quad b \in U_j \cap U_k \tag{5}$$

and projection

$$p_\eta : E_\eta \longrightarrow B$$

taking the class of $(b, j, g) \in U_j \times \{j\} \times G$ in the quotient E_η, to the point $b \in B$. Notice that if $\eta_j : U_j \times G \longrightarrow E_\eta$ sends (b, g) to the class of (b, j, g) in E_η, then $\{\eta_j\}$ defines a system of local trivializations for (p_η, E_η), and that $\eta = \{\eta_{jk}\}$ is the associated system of transition functions. Therefore,

Theorem 1 *The function*

$$\begin{array}{ccc} \check{H}^1(B; \mathscr{C}_G) & \longrightarrow & \mathsf{Prin}_G(B) \\ [\eta] & \mapsto & [E_\eta] \end{array}$$

is a natural bijection.

In addition to this characterization of principal G-bundles over B as Čech cohomology classes, there is another interpretation in terms of classifying maps. We will combine these two views in order to produce coordinates for data in the next sections.

Indeed, to each topological group G one can associate a space EG that is both weakly contractible, i.e. all its homotopy groups are trivial, and which comes equipped with a free right G-action

$$EG \times G \longrightarrow EG$$

The quotient $BG := EG/G$ is a topological space (endowed with the quotient topology), called the classifying space of G, and the quotient map

$$\jmath : EG \longrightarrow BG = EG/G$$

defines a principal G-bundle over BG, called the universal bundle. It is important to note that there are several constructions of EG, and thus of BG, but they all have the same homotopy type. One model for EG is the Milnor construction [13]

$$\mathcal{E}G := G * G * G * \cdots \tag{6}$$

with G acting diagonally by right multiplication on each term of the infinite join.

The next Theorem explains the universality of $\jmath : EG \longrightarrow BG$. Given a continuous map $f : B \longrightarrow BG$, the pullback f^*EG is the principal G-bundle over B with total space $\{(b, e) \in B \times EG : f(b) = \jmath(e)\}$, and projection map $(b, e) \mapsto b$. Moreover,

Theorem 2 *Let $[B, BG]$ denote the set of homotopy class of maps from B to the classifying space BG. Then, the function*

$$\begin{array}{ccc} [B, BG] & \longrightarrow & \mathsf{Prin}_G(B) \\ [f] & \mapsto & [f^*EG] \end{array}$$

is a bijection.

Proof See [9, Chapter 4: Theorems 12.2 and 12.4]. □

Theorem 2 implies that given a principal G-bundle $p : E \longrightarrow B$, there exists a continuous map $f : B \longrightarrow BG$ so that f^*EG is isomorphic to (p, E), and that the choice of f is unique up to homotopy. Any such choice is called a classifying map for $p : E \longrightarrow B$.

3 From Integer Simplicial Cohomology to Circular Coordinates

For an arbitrary topological group G, the Milnor construction (6) produces an explicit universal G-bundle $\jmath : \mathcal{E}G \longrightarrow \mathcal{B}G$, but the spaces $\mathcal{E}G$ and $\mathcal{B}G$ tend to be rather large. Indeed, they are often infinite-dimensional CW-complexes. For the case $G = \mathbb{Z}$ we have the more economical models $\mathcal{E}\mathbb{Z} \simeq \mathbb{R}$ and $\mathcal{B}\mathbb{Z} \simeq S^1 \subset \mathbb{C}$, with $\mathbb{Z}$ acting on $\mathbb{R}$ by right translation: $\mathbb{R} \times \mathbb{Z} \ni (r, m) \mapsto r + m$, and projection

$$\begin{array}{rcl} p : \mathbb{R} & \longrightarrow & S^1 \\ r & \mapsto & \exp(2\pi i r) \end{array}$$

Since $\mathbb{Z}$ is discrete, then $\mathbb{Z}$-valued continuous functions on B are in fact locally constant, and hence $\mathscr{C}_\mathbb{Z}$ is exactly the sheaf of locally constant functions with values in $\mathbb{Z}$, denoted $\underline{\mathbb{Z}}$. Combining the definition of the Čech cohomology group $\check{H}^1(B; \underline{\mathbb{Z}})$ with Theorems 1 and 2, yields a bijection

$$\varprojlim_{\mathcal{U}} H^1(\mathcal{N}(\mathcal{U}); \mathbb{Z}) \cong \left[B, S^1\right] \tag{7}$$

where the limit is taken over all locally finite covers $\mathcal{U}$ of B, ordered by refinement, and the groups are the 1-dimensional simplicial cohomology with $\mathbb{Z}$ coefficients of the associated nerve complexes $\mathcal{N}(\mathcal{U})$. The goal now is to produce an explicit family of compatible functions $H^1(\mathcal{N}(\mathcal{U});\mathbb{Z}) \longrightarrow [B, S^1]$ realizing the isomorphism from (7). This is done in Theorem 3, and an explicit formula is given by (11).

To begin, let $\{\varphi_j\}_{j\in J}$ be a partition of unity on B dominated[2] by $\mathcal{U} = \{U_j\}_{j\in J}$, fix a 1-cocycle $\eta = \{\eta_{jk}\} \in Z^1(\mathcal{N}(\mathcal{U});\mathbb{Z})$, and define for each $j \in J$ the map

$$\begin{array}{rcl} f_j : U_j \times \{j\} \times \mathbb{Z} & \longrightarrow & \mathbb{R} \\ (b, j, n) & \mapsto & n + \sum\limits_{\ell} \varphi_\ell(b)\eta_{j\ell} \end{array} \tag{8}$$

Since $\mathcal{U}$ is locally finite, then all but finitely many terms in this sum are zero. Note that $\mathbb{Z}$ acts on $U_j \times \{j\} \times \mathbb{Z}$ by right translation $\big((b, j, n), m\big) \mapsto (b, j, n+m)$, and that f_j is equivariant with respect to this action: $f_j(b, j, n+m) = f_j(b, j, n) + m$. If $b \in U_j \cap U_k$, then we have that

$$\begin{aligned} f_k(b, k, n + \eta_{jk}) &= n + \sum_{\ell\in J} \varphi_\ell(b)(\eta_{k\ell} + \eta_{jk}) \\ &= n + \sum_{\ell\in J} \varphi_\ell(b)\eta_{j\ell} \\ &= f_j(b, j, n) \end{aligned}$$

and hence the f_j's can be assembled to induce a continuous map $\widetilde{f_\eta} : E_\eta \longrightarrow \mathbb{R}$ on the quotient space defined by (5); here $\eta = \{\eta_{jk}\} \in Z^1(\mathcal{N}(\mathcal{U});\mathbb{Z})$ is regarded as a collection of constant functions $\eta_{jk} : U_j \cap U_k \longrightarrow \mathbb{Z}$. To be more explicit, $\widetilde{f_\eta}$ sends the class of (b, j, n) in E_η to $f_j(b, j, n) \in \mathbb{R}$. Since each f_j is $\mathbb{Z}$-equivariant, then so is $\widetilde{f_\eta}$, and hence it descends to a well defined map f_η at the level of base spaces

$$\begin{array}{rcl} f_\eta : \quad B & \longrightarrow & S^1 \subset \mathbb{C} \\ U_j \ni b & \mapsto & \exp\left(2\pi i \sum\limits_k \varphi_k(b)\eta_{jk}\right) \end{array} \tag{9}$$

Lemma 2 *The map f_η classifies the principal $\mathbb{Z}$-bundle $p_\eta : E_\eta \longrightarrow B$.*

Proof Let us see explicitly that the map f_η is well defined; in other words, that the value $f_\eta(b) \in S^1$ is independent of the open set containing b. Indeed, let $j, \ell \in J$ be so that $b \in U_j \cap U_\ell$. We contend that $\varphi_k(b)\eta_{jk} = \varphi_k(b)(\eta_{j\ell} + \eta_{\ell k})$ for every $k \in J$. If $b \notin U_k$, then the equality is trivial since $\varphi_k(b) = 0$; if $b \in U_k$, then $U_j \cap U_k \cap U_\ell \neq \emptyset$ and $\eta_{jk} = \eta_{j\ell} + \eta_{\ell k}$ since η is a cocycle. Therefore

$$\sum_k \varphi_k(b)\eta_{jk} = \eta_{j\ell} + \sum_k \varphi_k(b)\eta_{\ell k}$$

[2] That is, so that $\mathsf{support}(\varphi_j) \subset \mathsf{closure}(U_j)$ for all $j \in J$.

and given that $\eta_{j\ell} \in \mathbb{Z}$, then $\exp\left(2\pi i \sum\limits_k \varphi_k(b)\eta_{jk}\right) = \exp\left(2\pi i \sum\limits_k \varphi_k(b)\eta_{\ell k}\right)$.

Finally, let us check that taking the pullback $f_\eta^*\mathbb{R}$ of the universal $\mathbb{Z}$-bundle $\exp(2\pi i\ \cdot) : \mathbb{R} \longrightarrow S^1$ yields a principal $\mathbb{Z}$-bundle isomorphic to $p_\eta : E_\eta \longrightarrow B$. Indeed, since $f_\eta \circ p_\eta = \exp\left(2\pi i \widetilde{f_\eta}\right)$, then $\left(\widetilde{f_\eta}, f_\eta\right) : (p_\eta, E_\eta, B) \longrightarrow \left(\exp(2\pi i\ \cdot), \mathbb{R}, S^1\right)$ is a morphism of principal $\mathbb{Z}$-bundles, and the result follows from [9, Chapter 4: Theorem 4.2]. □

Theorem 3 *Let $\iota : \mathbb{Z} \hookrightarrow \mathbb{R}$ be the inclusion and*

$$\iota^* : H^1(\mathcal{N}(\mathcal{U}); \mathbb{Z}) \longrightarrow H^1(\mathcal{N}(\mathcal{U}); \mathbb{R}) \tag{10}$$

the induced homomorphism. Given $\eta \in Z^1(\mathcal{N}(\mathcal{U}); \mathbb{Z})$ and $\tau \in C^0(\mathcal{N}(\mathcal{U}); \mathbb{R})$, let $\theta = \iota^\#(\eta) + \delta^0\tau$. Denote by $\tau_j \in \mathbb{R}$ the value of τ on the vertex $j \in \mathcal{N}(\mathcal{U})$, and by $\theta_{jk} \in \mathbb{R}$ the value of θ on the oriented edge $[j,k] \in \mathcal{N}(\mathcal{U})$; in particular $\theta_{jk} = -\theta_{kj}$, and $\theta_{jk} = 0$ whenever $\{j,k\} \notin \mathcal{N}(\mathcal{U})$. If

$$\begin{array}{rccl} h_{\theta,\tau} : & B & \longrightarrow & S^1 \subset \mathbb{C} \\ & U_j \ni b & \mapsto & \exp\left\{2\pi i \left(\tau_j + \sum\limits_k \varphi_k(b)\theta_{jk}\right)\right\} \end{array} \tag{11}$$

then $h_{\theta,\tau}$ is a classifying map for the principal $\mathbb{Z}$-bundle $p_\eta : E_\eta \longrightarrow B$.

Proof Since f_η is a classifying map for E_η, by Lemma 2, then it is enough to check that f_η and $h_{\theta,\tau}$ are homotopic (see Theorem 2). For $b \in U_j$ we have that

$$\begin{aligned} f_\eta(b) &= \exp\left(2\pi i \sum_k \varphi_k(b)\eta_{jk}\right) \\ &= \exp\left(2\pi i \sum_k \varphi_k(b)(\theta_{jk} + \tau_j - \tau_k)\right) \\ &= \exp\left(2\pi i \left(\tau_j + \sum_k \varphi_k(b)(\theta_{jk} - \tau_k)\right)\right) \\ &= \nu_\tau(b) \cdot h_{\theta,\tau}(b) \end{aligned}$$

where $\nu_\tau(b) = \exp\left(-2\pi i \sum\limits_k \varphi_k(b)\tau_k\right)$. Since ν_τ factors through $\mathbb{R}$:

$$\begin{array}{rcccl} \nu_\tau : B & \longrightarrow & \mathbb{R} & \longrightarrow & S^1 \subset \mathbb{C} \\ b & \mapsto & \sum\limits_k \varphi_k(b)\tau_k & \mapsto & \exp\left(-2\pi i \sum\limits_k \varphi_k(b)\tau_k\right) \end{array}$$

then ν_τ is null-homotopic, hence f_η is homotopic to $h_{\theta,\tau}$, and the result follows. □

Remark 1 We note that the relation $\theta = \iota^{\#}(\eta) + \delta^0 \tau$ from Theorem 3 implies that the cochain $\tau \in C^0(\mathcal{N}(\mathcal{U}); \mathbb{R})$ encodes the degrees of freedom in choosing a cocycle representative for the class $\iota^*([\eta]) \in H^1(\mathcal{N}(\mathcal{U}); \mathbb{R})$, and thus defining the classifying map $h_{\theta,\tau} : B \longrightarrow S^1$. This choice will be addressed in the discussion about Harmonic Smoothing in Sect. 4.6.

4 Persistent Cohomology and Sparse Circular Coordinates for Data

In this section we show how the theory we have developed thus far can be applied to real data sets. In particular, we explain and justify the choices made in the construction outlined in the Introduction (Sect. 1.2). Let us begin by fixing an ambient metric space $(\mathbb{M}, \mathbf{d})$, let $L \subset \mathbb{M}$ be finite, and let

$$\begin{aligned} B_\alpha(\ell) &= \{b \in \mathbb{M} : \mathbf{d}(b, \ell) < \alpha\} \quad , \quad \alpha \geq 0,\ \ell \in L \\ \mathcal{B}_\alpha &= \{B_\alpha(\ell)\}_{\ell \in L} \\ L^{(\alpha)} &= \bigcup \mathcal{B}_\alpha \end{aligned}$$

The formulas derived in the previous section, specially (9), imply that each cocycle $\eta \in Z^1(\mathcal{N}(\mathcal{B}_\alpha); \mathbb{Z})$ yields a map $h : L^{(\alpha)} \longrightarrow S^1$. The thing to notice is that h is defined on every $b \in L^{(\alpha)}$; thus, given a large but finite set $X \subset \mathbb{M}$—the data—sampled around a continuous space $\mathbb{X} \subset \mathbb{M}$, one can select a much smaller set of landmarks $L \subset X$ and $\alpha > 0$ for which $X \subset L^{(\alpha)}$. The resulting circular coordinates $h : L^{(\alpha)} \longrightarrow S^1$ will thus be defined on all points of X, though only the landmark set is used in its construction. As we alluded to in the introduction, this is what we mean when we say that the coordinates are *sparse*.

4.1 Landmark Selection

In practice we select the landmarks $L \subset X$ either at random, or through `maxmin` sampling: Given $N \leq |X|$ and $\ell_1 \in X$ chosen arbitrarily, assume that $\ell_1, \ldots, \ell_j \in X$ have been selected, $1 \leq j < N$, and let

$$\ell_{j+1} = \underset{x \in X}{\operatorname{argmax}} \ \min \left\{ \mathbf{d}(x, \ell_1), \ldots, \mathbf{d}(x, \ell_j) \right\} \tag{12}$$

Following this inductive procedure defines a landmark set $L = \{\ell_1, \ldots, \ell_N\} \subset X$ that is in practice well-separated and well-distributed throughout the data. However, it is important to keep in mind that this process is prone to choosing outliers.

4.2 The Subordinated Partition of Unity

As for the choice of partition of unity $\{\varphi_\ell\}_{\ell \in L}$ dominated by $\mathcal{B}_\alpha$, we can use that the cover is via metric balls, and let

$$\varphi_\ell(b) = \frac{|\alpha - \mathbf{d}(\ell, b)|_+}{\sum\limits_{\ell' \in L} |\alpha - \mathbf{d}(\ell', b)|_+} \qquad \text{where} \qquad |r|_+ = \max\{r, 0\}, \quad r \in \mathbb{R} \tag{13}$$

See [17, 3.3 and Fig 6.] for other typical choices of partition of unity in the case of metric spaces, and coverings via metric balls.

4.3 The Need for Persistence

Even if the landmark set L correctly approximates the underlying topology of X, the choice of scale $\alpha > 0$ and cocycle $\eta \in Z^1(\mathcal{N}(\mathcal{B}_\alpha); \mathbb{Z})$ might reflect sampling artifacts instead of robust geometric features of the underlying space $\mathbb{X}$. This is why we need persistent cohomology. Indeed, a class $[\eta] \in H^1(\mathcal{N}(\mathcal{B}_\alpha); \mathbb{Z})$ which is not in the kernel of the homomorphism

$$H^1(\mathcal{N}(\mathcal{B}_\alpha); \mathbb{Z}) \longrightarrow H^1(\mathcal{N}(\mathcal{B}_{\alpha'}); \mathbb{Z}) \quad , \quad 0 < \alpha' < \alpha$$

induced by the inclusion $\mathcal{N}(\mathcal{B}_{\alpha'}) \subset \mathcal{N}(\mathcal{B}_\alpha)$, is less likely to correspond to spurious features as the difference $\alpha - \alpha'$ increases. Note, however, that the efficient computation of persistent cohomology classes relies on using field coefficients. We proceed, following [6] and [17], by choosing a prime $q > 2$ and a scale $\alpha > 0$ so that (1) $H^1(\mathcal{N}(\mathcal{B}_\alpha); \mathbb{Z}/q)$ contains a class with large persistence, and (2) so that the homomorphism $H^1(\mathcal{N}(\mathcal{B}_\alpha); \mathbb{Z}) \longrightarrow H^1(\mathcal{N}(\mathcal{B}_\alpha); \mathbb{Z}/q)$, induced by the quotient map $\mathbb{Z} \longrightarrow \mathbb{Z}/q$, is surjective.

4.4 Lifting Persistence to Integer Coefficients

As stated in [6], one has that:

Proposition 2 *Let K be a finite simplicial complex, and suppose that $q \in \mathbb{N}$ does not divide the order of the torsion subgroup of $H^2(K; \mathbb{Z})$. Then the homomorphism*

$$\iota_q^* : H^1(K; \mathbb{Z}) \longrightarrow H^1(K; \mathbb{Z}/q)$$

induced by the quotient map $\iota_q : \mathbb{Z} \longrightarrow \mathbb{Z}/q$, is surjective.

Proof This follows directly from the Bockstein long exact sequence in cohomology, corresponding to the short exact sequence $0 \longrightarrow \mathbb{Z} \xrightarrow{\times q} \mathbb{Z} \xrightarrow{\iota_q} \mathbb{Z}/q \longrightarrow 0.$ □

More generally, let $\{K_\alpha\}_{\alpha\geq 0}$ be a filtered simplicial complex with $\bigcup\limits_{\alpha\geq 0} K_\alpha$ finite. Since each complex K_α is finite, and the cohomology groups $H^2(K_\alpha;\mathbb{Z})$ change only at finitely many values of α, then there exists $Q \in \mathbb{N}$ so that the hypotheses of Proposition 2 will be satisfied for each $q \geq Q$, and all $\alpha \geq 0$. In practice we choose a prime q at random, with the intuition that for scientific data only a few primes are torsion contributors.

Let $\mathbb{Z}/q = \{0, 1, \ldots, q-1\}$ and for $\eta' \in Z^1(K_\alpha;\mathbb{Z}/q)$ let $\eta \in C^1(K_\alpha;\mathbb{Z})$ be defined on each 1-simplex $\sigma \in K_\alpha$ as:

$$\eta(\sigma) = \begin{cases} \eta'(\sigma) & \text{if } \eta'(\sigma) \leq \frac{q-1}{2} \\ \eta'(\sigma) - q & \text{if } \eta'(\sigma) > \frac{q-1}{2} \end{cases} \tag{14}$$

Thus, η takes values in $\left\{-\frac{q-1}{2}, \ldots, 0, \ldots, \frac{q-1}{2}\right\} \subset \mathbb{Z}$ and it satisfies $(\eta \bmod q) = \eta'$. For the examples we have observed, the cochain defined by (14) produces an integer cocycle. One of the reviewers of an earlier version of this paper remarked that this is not always the case; the outlined procedure tends to fail (in real world-examples) when the cohomology computation involves division by 2. As highlighted in [6, 2.4], solving a Diophantine linear system can be used to fix the problem.

4.5 Use Rips, Not Nerves

Constructing the filtered complex $\{\mathcal{N}(\mathcal{B}_\alpha)\}_{\alpha\geq 0}$ can be rather expensive for a general ambient metric space $(\mathbb{M}, \mathbf{d})$. Indeed, the inclusion of an n-simplex into the nerve complex is predicated on checking if the intersection of $n+1$ ambient metric balls is nonempty. This is nontrivial on curved spaces. On the other hand, the Rips complex

$$R_\alpha(L) = \{\sigma \subset L : \mathsf{diam}(\sigma) < \alpha\} \quad , \quad \alpha \geq 0$$

provides a straightforward alternative, since we can use that

$$R_\alpha(L) \subset \mathcal{N}(\mathcal{B}_\alpha) \subset R_{2\alpha}(L)$$

for every $\alpha \geq 0$. Here is how. Let $q > 2$ be a prime so that

$$\iota_q^* : H^1(R_\alpha(L);\mathbb{Z}) \longrightarrow H^1(R_\alpha(L);\mathbb{Z}/q)$$

is surjective for all $\alpha \geq 0$, and let

$$\jmath : H^1(R_{2\alpha}(L); \mathbb{Z}/q) \longrightarrow H^1(R_\alpha(L); \mathbb{Z}/q)$$

be the homomorphism induced by the inclusion $R_\alpha(L) \subset R_{2\alpha}(L)$. Moreover, let $\eta' \in Z^1(R_{2\alpha}(L); \mathbb{Z}/q)$ be so that $[\eta'] \notin \mathsf{Ker}(\jmath)$, and fix an integer lift

$$\eta \in Z^1(R_{2\alpha}(L); \mathbb{Z})$$

That is, one for which $\eta' - (\eta \ \mathsf{mod}\ q) \in Z^1(R_{2\alpha}(L); \mathbb{Z}/q)$ is a coboundary, e.g. (14).

The diagram below summarizes the spaces and homomorphisms used thus far:

$$\begin{tikzcd}
\left[\eta'\right] \in H^1(R_{2\alpha}(L); \mathbb{Z}/q) \arrow[d] \arrow[dd, bend right=60, "\jmath"'] & H^1(R_{2\alpha}(L); \mathbb{Z}) \ni [\eta] \arrow[l, "\iota_q^*"'] \arrow[d, "\iota_{\mathbb{Z}}^*"] \\
H^1(\mathcal{N}(\mathcal{B}_\alpha); \mathbb{Z}/q) \arrow[d] & H^1(\mathcal{N}(\mathcal{B}_\alpha); \mathbb{Z}) \ni [\widetilde{\eta}] \arrow[l, "\iota_q^*"] \\
H^1(R_\alpha(L); \mathbb{Z}/q) &
\end{tikzcd}$$

Since the diagram commutes, then $[\eta]$ is not in the kernel of $\iota_{\mathbb{Z}}^*$, and hence we obtain a nonzero element $\iota_{\mathbb{Z}}^*([\eta]) = [\widetilde{\eta}] \in H^1(\mathcal{N}(\mathcal{B}_\alpha); \mathbb{Z})$. This is the class we would use as input for Theorem 3.

4.6 Harmonic Smoothing

The final step is selecting an appropriate cocycle representative (refer to Fig. 1 to see why this matters)

$$\widetilde{\theta} \in Z^1(\mathcal{N}(\mathcal{B}_\alpha); \mathbb{R})$$

for the class $\iota^*([\widetilde{\eta}]) \in H^1(\mathcal{N}(\mathcal{B}_\alpha); \mathbb{R})$, see (10). Again, since one would hope to never compute the nerve complex, the strategy is to solve the problem in $Z^1(R_{2\alpha}(L); \mathbb{R})$ for $\iota^\#(\eta)$, and then transfer the solution using $\iota_{\mathbb{R}}^\# : C^1(R_{2\alpha}(L); \mathbb{R}) \to C^1(\mathcal{N}(\mathcal{B}_\alpha); \mathbb{R})$.

Inspecting (11) reveals that the choice of $\widetilde{\theta}$ which promotes the smallest total variation in $h_{\widetilde{\theta},\widetilde{\tau}}$, is the one for which the value of $\widetilde{\theta}$ on each 1-simplex of $\mathcal{N}(\mathcal{B}_\alpha)$ is as small as possible. Consequently, we will look for the cocycle representative

$$\theta \in Z^1(R_{2\alpha}(L); \mathbb{R})$$

of $\iota^*([\eta])$, which in average has the smallest squared value[3] on each 1-simplex of $R_{2\alpha}(L)$. That said, not all edges in the rips complex $R_\epsilon(L)$ are created equal. Some might have just entered the filtration, i.e. $\mathbf{d}(\ell_j, \ell_k) \approx \epsilon$, which would make them unstable if L is corrupted with noise, or perhaps $X \cap \left(B_{\epsilon/2}(\ell_j) \cup B_{\epsilon/2}(\ell_k)\right)$ is a rather small portion of the data, which could happen if ℓ_j and ℓ_k are outliers selected during `maxmin` sampling.

These observations can be encoded by choosing weights on vertices and edges:

$$\omega_\epsilon : L \times L \longrightarrow [0, \infty) \quad , \quad \epsilon \geq 0 \tag{15}$$

where ω_ϵ is symmetric for all $\epsilon > 0$, it satisfies

$$\omega_{\epsilon'}(\ell, \ell') \leq \omega_\epsilon(\ell, \ell') \quad , \qquad \text{for} \quad \epsilon' \leq \epsilon$$

and $\omega_\epsilon(\ell, \ell') = 0$ only when $0 < \epsilon \leq \mathbf{d}(\ell, \ell')$. Here $\omega_\epsilon(\ell, \ell)$ is the weight of ℓ, and $\omega_\epsilon(\ell, \ell')$ is the weight of the edge $\{\ell, \ell'\}$. For instance, one can take

$$\omega_\epsilon(\ell, \ell') = |\epsilon - \mathbf{d}(\ell, \ell')|_+$$

but we note that we have not yet systematically investigated the effects of this choice. See [20, Apendix D] for a different heuristic.

It follows that ω_ϵ defines inner products $\langle \cdot, \cdot \rangle_\epsilon$ on $C^0(R_\epsilon(L); \mathbb{R})$ and $C^1(R_\epsilon(L); \mathbb{R})$, by letting the indicator functions 1_σ on k-simplices ($k = 0, 1$) $\sigma \in R_\epsilon(L)$ be orthogonal, and setting

$$\langle 1_\sigma, 1_\sigma \rangle_\epsilon = \omega_\epsilon(\sigma) \tag{16}$$

Using $\langle \cdot, \cdot \rangle_\epsilon$, for $\epsilon = 2\alpha$, we let $\beta \in B^1(R_{2\alpha}(L); \mathbb{R})$ be the orthogonal projection of $\iota^\#(\eta)$ onto the space of 1-coboundaries, and define

$$\theta = \iota^\#(\eta) - \beta \tag{17}$$

A bit of linear algebra shows that,

Proposition 3 *The 1-cocycle θ defined by (17) is a minimizer for the weighted least squares problem*

$$\min_{\phi \sim \iota^\#(\eta)} \sum_\sigma \omega_{2\alpha}(\sigma) \cdot \phi(\sigma)^2 \tag{18}$$

Here the sum runs over all 1-simplices $\sigma \in R_{2\alpha}(L)$, and the minimization is over all 1-cocycles $\phi \in Z^1(R_{2\alpha}(L); \mathbb{R})$ which are cohomologous to $\iota^\#(\eta)$.

[3] That is, we use the harmonic cocycle representative for appropriate inner products on cochains.

Similarly, and if

$$d_{2\alpha} : C^0(R_{2\alpha}(L); \mathbb{R}) \longrightarrow C^1(R_{2\alpha}(L); \mathbb{R})$$

denotes the coboundary map, then we let

$$\tau \in \mathsf{Ker}(d_{2\alpha})^\perp \subset C^0(R_{2\alpha}(L); \mathbb{R})$$

in the orthogonal complement of the kernel of $d_{2\alpha}$, be so that $d_{2\alpha}(\tau) = -\beta$. Hence τ is the 0-chain with the smallest norm mapping to $-\beta$ via $d_{2\alpha}$. Consequently, if

$$d_{2\alpha}^+ : C^1(R_{2\alpha}(L); \mathbb{R}) \longrightarrow C^0(R_{2\alpha}(L); \mathbb{R})$$

is the weighted Moore-Penrose pseudoinverse of $d_{2\alpha}$ (see [3, III.3.4]), then

$$\tau = -d_{2\alpha}^+(\iota^\#(\eta)) \qquad \text{and} \qquad \theta = \iota^\#(\eta) + d_{2\alpha}(\tau) \tag{19}$$

This is how we compute τ and θ in our implementation. Now, let

$$\widetilde{\tau} = \iota^\#_{\mathbb{R}}(\tau) \in C^0(\mathcal{N}(\mathcal{B}_\alpha); \mathbb{R})$$
$$\widetilde{\theta} = \iota^\#_{\mathbb{R}}(\theta) \in Z^1(\mathcal{N}(\mathcal{B}_\alpha); \mathbb{R})$$

If we were to be completely rigourous, then $\widetilde{\tau}$ and $\widetilde{\theta}$ would be the cochains going into (11); this would require the 1-skeleton of the nerve complex. However, as the following proposition shows, this is unnecessary:

Proposition 4 *For all $b \in B_\alpha(\ell_j)$, and every $j = 1, \ldots, N$, we have that*

$$\exp\left\{2\pi i\left(\widetilde{\tau}_j + \sum_{k=1}^{N}\varphi_k(b)\widetilde{\theta}_{jk}\right)\right\} = \exp\left\{2\pi i\left(\tau_j + \sum_{k=1}^{N}\varphi_k(b)\theta_{jk}\right)\right\}$$

That is, we can compute sparse circular coordinates using only the Rips filtration on the landmark set.

Proof Since $\mathcal{N}(\mathcal{B}_\alpha)$ and $R_{2\alpha}(L)$ have the same vertex set, namely L, then $\widetilde{\tau} = \tau$ as real-valued functions on L. Moreover, for all $k = 1, \ldots, N$ we have that

$$\varphi_k(b)\widetilde{\theta}_{jk} = \varphi_k(b)\theta_{jk}$$

for if $b \notin B_\alpha(\ell_k)$, then both sides are zero, and if $b \in B_\alpha(\ell_j) \cap B_\alpha(\ell_k)$, then the edge $\{\ell_j, \ell_k\}$ is in both $R_{2\alpha}(L)$ and $\mathcal{N}(\mathcal{B}_\alpha)$, which shows that $\widetilde{\theta}_{jk} = \theta_{jk}$. □

5 Experiments

In all experiments below, persistent cohomology is computed using a MATLAB wrapper for Ripser [1] kindly provided by Chris Tralie (http://www.ctralie.com/). The Moore-Penrose pseudoinverse was computed via MATLAB's `pinv`. In all cases we run the algorithm from the Introduction in Sect. 1.2 using the indicated persistence classes, or linear combinations thereof as made explicit in each example.

5.1 Synthetic Data

5.1.1 A Noisy Circle

We select 1000 points from a noisy circle in $\mathbb{R}^2$; the noise is Gaussian in the direction normal to the unit circle. Fifty landmarks were selected via `maxmin` sampling (5% of the data), and circular coordinates were computed for the two most persistent classes η_1 and η_2, using (19) as input to (11)—this is the harmonic cocycle column—or (9) with either η_1 or η_2 directly—the integer cocycle column. We show the results in Fig. 1 below. Computing persistent cohomology took 0.079423 s (the Rips filtration is constructed from zero to the diameter of the landmark set); in each case computing the harmonic cocycle takes about 0.037294 s. This example highlights the inadequacy of the integer cocycle and of choosing cohomology classes associated to sampling artifacts (i.e., with low persistence). From now on,

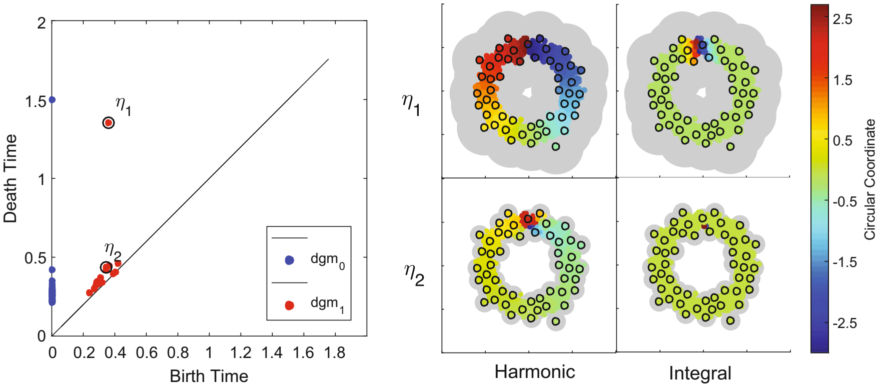

Fig. 1 A noisy circle. Left: persistence diagrams in dimension 0 (blue) and 1 (red) for the Rips filtration on the landmarks. Right: Circular coordinates from the two most persistent classes η_1 (top row) and η_2 (bottom row). The columns indicate if the harmonic or integral cocycle was used. The dark rings are the landmarks. The colors are: the domain of definition for the circular coordinate (gray), and its value on each point (dark blue, $-\pi$, through dark red, π). Please refer to an electronic version for colors

we only present circular coordinates computed with the relevant harmonic cocycle representative.

5.1.2 The 2-Dimensional Torus

For this experiment we sample 1000 points uniformly at random from the square $[0, 2\pi) \times [0, 2\pi)$, and for each selected pair (ϕ_1, ϕ_2) we generate a point $\left(e^{i\phi_1}, e^{i\phi_2}\right) \in S^1 \times S^1$ on the surface of the torus embedded in $\mathbb{C}^2$. The resulting finite set is endowed with the ambient distance from $\mathbb{C}^2$, and 100 landmarks (i.e., 10% of the data) are selected through maxmin sampling. We show the results in Fig. 2 below, for the circular coordinates computed with the two most persistent classes, η_1 and η_2, and the maps (11) associated to the harmonic cocycle representatives (19). Computing persistent cohomology for the Rips filtration on the Landmarks (from zero to the diameter of the set) takes 0.398252 s, and computing the harmonic cocycles takes 0.030832 s.

5.1.3 The Klein Bottle

We model the Klein bottle as the quotient space

$$K = S^1 \times S^1/(z, w) \sim (-z, \overline{w})$$

and endow it with the quotient metric. Just like in the case of the 2-torus, we sample 1000 points uniformly at random on (the fundamental domain $[0, \pi) \times [0, 2\pi)$ of) K, and select 100 landmarks via maxmin sampling and the quotient metric. Below in Fig. 3 we show the results of computing the persistent cohomology, with coefficients in $\mathbb{Z}/13$, of the Rips filtration on the landmark set (left), along with the circular coordinates corresponding to the most persistent class (right).

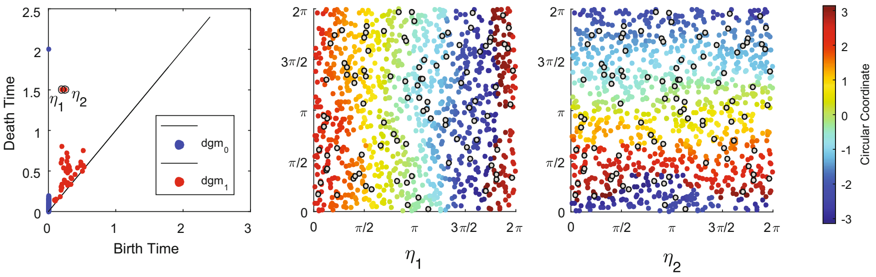

Fig. 2 The torus. Left: Persistence in dimensions 0 and 1 for the Rips filtration on the landmark set. Center and Right: the landmark set is depicted with dark rings, and the colors correspond to the circular coordinates computed with (the harmonic representatives from) the two most persistent classes η_1 (center) and η_2 (right). Please refer to an electronic version for colors

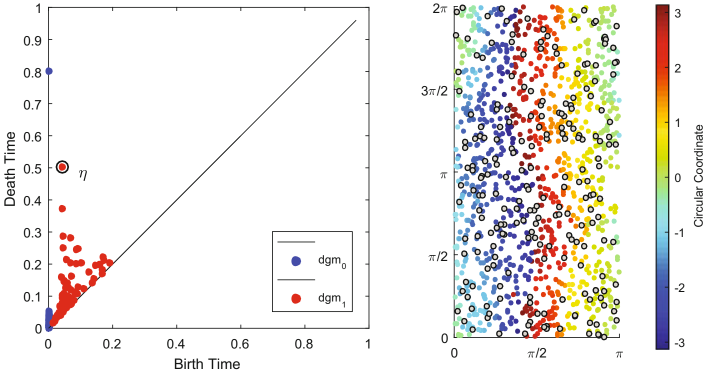

Fig. 3 Circular coordinates on the Klein bottle. Left: Persistence with coefficients in $\mathbb{Z}/13$ for the Rips filtration on the landmark set. Right: Circular coordinates computed from the harmonic representative from the class η with largest persistence. Dark rings indicate landmarks, and the colors (dark blue through dark red) are the angular values of the circular coordinate on each data point. Please refer to an electronic version for colors

5.2 *Real Data*

5.2.1 COIL 20

The Columbia University Image Library (COIL-20) is a collection of 448 × 416-pixel gray scale images from 20 objects, each of which is photographed at 72 different rotation angles [15]. The database has two versions: a processed version, where the images have been cropped to show only the rotated object, and an unprocessed version with the 72 raw images from 5 objects. We will analyze the unprocessed database, of which a few examples are shown in Fig. 4 below.

Regarding each image as a vector of pixel intensities in $\mathbb{R}^{448\times 416}$ yields a set X with 360 points; this set becomes a finite metric space when endowed with the ambient Euclidean distance. Below in Fig. 5 (left) we show the result of computing persistence (this time visualized as barcodes) for the Rips complex on the entire data set (0.293412 s). Each one of the six most persistent classes $\eta_1, \ldots, \eta_6$ yields a circle-valued map on the data $h_j : X \longrightarrow S^1$, $j = 1, \ldots, 6$. Multiplying these maps together, using the group structure from $S^1 \subset \mathbb{C}$, yields a map $h : X \longrightarrow S^1$. We do this at the level of maps, as opposed to adding up the cocycle representatives, because there is no scale α at which all these classes are alive. We also show in Fig. 5 (right) an Isomap [22] projection of the data onto $\mathbb{R}^2$, and we color each projected data point with its h value.

As we show in Fig. 6 below, a better system of coordinates for the data (i.e. one without crossings) is given by the computed circular coordinate of each data point, and the cluster (computed using single linkage) to which it belongs to.

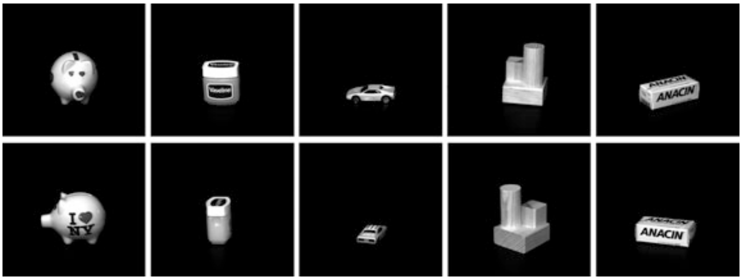

Fig. 4 Some examples from the unprocessed COIL20 image database

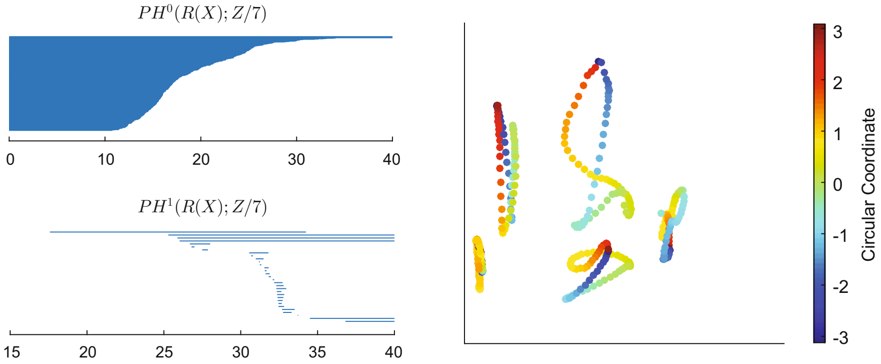

Fig. 5 COIL-20 unprocessed. Left: persistence of the Rips filtration. Right: Isomap projection colored by circular coordinate

5.2.2 The Mumford Data

This data set was first introduced in [12], with an initial exploration of its underlying topological structure done in [5], and then a more thorough investigation in [4]. The data set in question is a collection of roughly four million 3 × 3-pixel gray-scale images with high-contrast, selected from monochrome photos in a database of 4000 natural scenes [8]. The 3 × 3-pixel image patches are preprocessed, intensity-centered and contrast-normalized, and a linear change of coordinates is performed yielding a point-cloud $\mathcal{M} \subset S^7 \subset \mathbb{R}^8$. The Euclidean distance in $\mathbb{R}^8$ endows $\mathcal{M}$ with the structure of a finite metric space. Following [4], we select 50,000 points at random from $\mathcal{M}$ and then let X be the top 30% densest points as measured by the distance to their 15th nearest neighbor. This results in a data set with 15,000 points, which we analyze below.

We select 700 landmarks from X via `maxmin` sampling, i.e. 4.7% of X, and compute persistence for the associated Rips filtration. This takes about 2.2799 s and the result is shown in Fig. 7.

Each bar in the barcode yields a class η_j, which we order from largest (η_1) to smallest (η_5) persistence. Below in Fig. 8 we show the circular coordinates

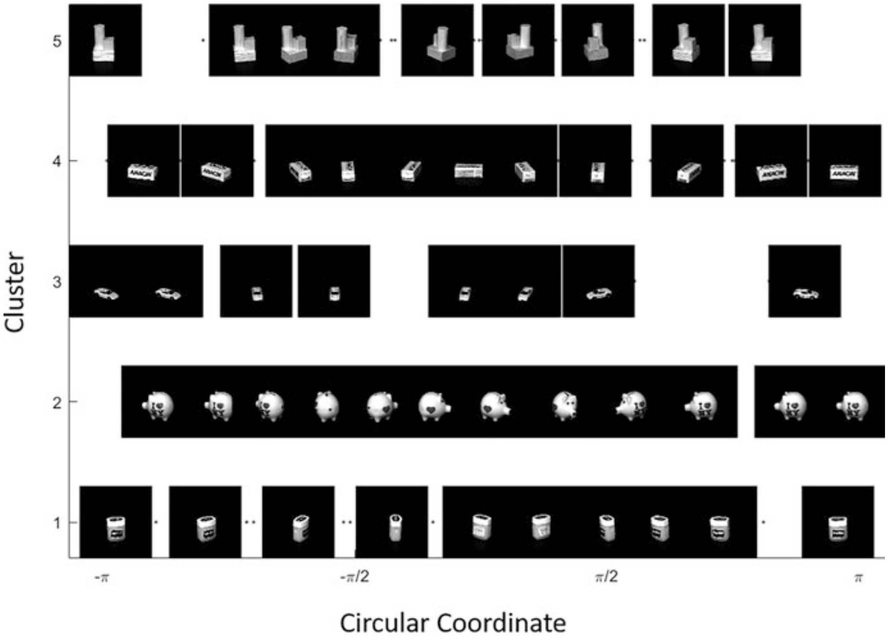

Fig. 6 COIL-20 unprocessed: clusters vs circular coordinates

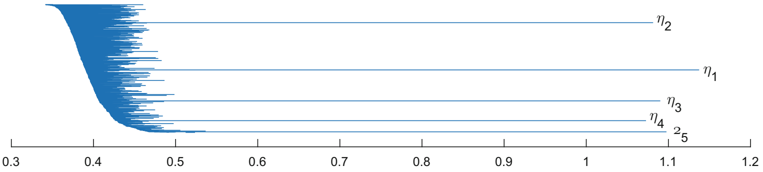

Fig. 7 Barcodes from persistence on the Rips filtration of the landmark set $L \subset X$

associated to the classes η_2, $\eta_1 + \eta_5$ and $\eta_3 + \eta_4$, respectively. Each of the three panels shows a scatter plot of $X \subset \mathbb{R}^8$ with respect to the first two coordinates, dark rings are the selected landmarks, and the colors (dark blue through dark red) are the circular coordinates corresponding to the indicated persistence classes. The computation of each cocycle representative takes about 7.1434 s, so the entire analysis is less than 25 s.

These three circular coordinates allow us to map the data set X into the 3-dimensional torus $T^3 = S^1 \times S^1 \times S^1$, which we model as the 3-dimensional cube $[-\pi, \pi] \times [-\pi, \pi] \times [-\pi, \pi]$ with opposite faces identified. We show in Fig. 9 below the result of mapping the data into T^3.

As we can see from the scatter plot, these three coordinates provide a faithful realization of the data in the three circle model proposed in [4]. Below in Fig. 10 we

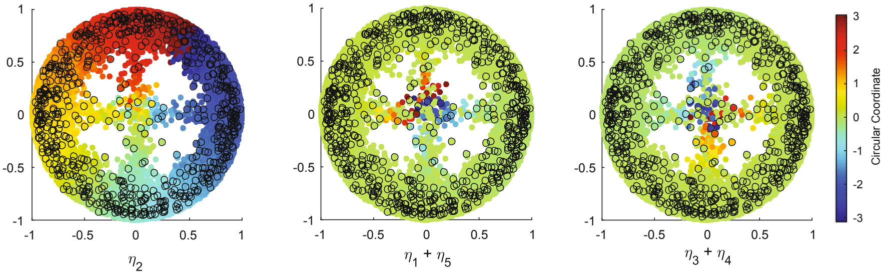

Fig. 8 Circular coordinates for the points in $X \subset \mathbb{R}^8$, plotted according to their first two coordinates, and colored by the circular coordinates associated to each one of the classes η_2 (left), $\eta_1 + \eta_5$ (center) and $\eta_3 + \eta_4$ (right)

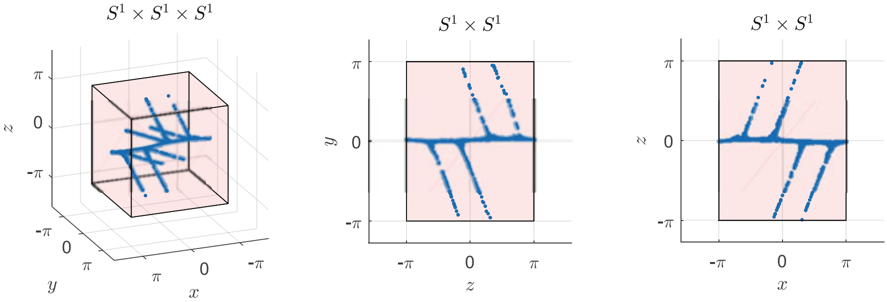

Fig. 9 Scatter plot in the 3-torus (left) for X, along with two 2-d projections (center, left). The horizontal line on the xy plane is a circle (the primary circle), and each one of the four V-shaped curves in T^3 is a hemisphere of a (secondary) circle

show some of these image patches in their T^3-coordinate to better illustrate what the actual circles are.

6 Discussion

We have presented in this paper an application of the theory of principal bundles to the problem of finding topologically and geometrically meaningful coordinates for scientific data. Specifically, we leverage the 1-dimensional persistent cohomology of the Rips filtration on a subset of the data (the landmarks), in order to produce S^1-valued coordinates on the entire data set. The coordinates are designed to capture 1-dimensional topological features of a continuous underlying space, and the theory on which the coordinates are built, indicates that they classify $\mathbb{Z}$-principal bundles on the continuum.

The use of bundle theory allows for the circular coordinates to be sparse, which is fundamental for analyzing geometric data of realistic size. We hope that these

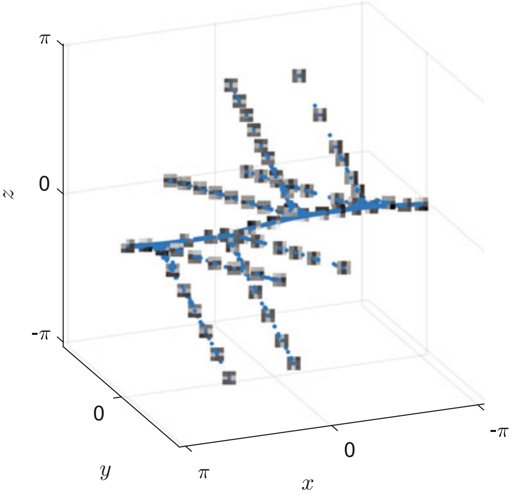

Fig. 10 Image patches from X, plotted at their location in the 3-torus, according to the computed circular coordinates

coordinates will be useful in problems such as the analysis of recurrent dynamics in time series data (as in [23, 24] or [7]), and nonlinear dimensionality reduction as indicated in the Experiments Sect. 5.

An interesting direction from this work is the question of stability and Lipschitz continuity of sparse circular coordinates. The main theoretical challenge is to determine how the edge and vertex weights on the Rips complex can be used to stabilize the harmonic cocycle representative with respect to an appropriate notion of (hopefully Hausdorff) noise on the landmark set. We hope to address this question in upcoming work.

Acknowledgements This work was partially supported by the NSF under grant DMS-1622301 and DARPA under grant HR0011-16-2-003.

References

1. Bauer., U.: Ripser: a lean C++ code for the computation of Vietoris-Rips persistence barcodes. 2017. Software: https://github.com/Ripser/ripser.
2. Belkin, M., Niyogi P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural computation, **15**(6), 1373–1396 (2003)
3. Ben-Israel, A., Greville, T. N.: *Generalized inverses: theory and applications*, volume 15. Springer Science & Business Media, (2003)
4. Carlsson, G., Ishkhanov, T., de Silva, V., Zomorodian, A.: On the local behavior of spaces of natural images. International journal of computer vision, **76**(1), 1–12, (2008)
5. de Silva V., Carlsson, G. E.: Topological estimation using witness complexes. SPBG, **4**, 157–166, (2004)
6. de Silva, V., Morozov, D., Vejdemo-Johansson, M.: Persistent cohomology and circular coordinates. Discrete & Computational Geometry, **45**(4), 737–759, (2011)

7. de Silva, V., Skraba, P., Vejdemo-Johansson, M.: Topological analysis of recurrent systems. In Workshop on Algebraic Topology and Machine Learning, NIPS, (2012)
8. v. Hateren, J. H., v. d. Schaaf, A.: Independent component filters of natural images compared with simple cells in primary visual cortex. In Proceedings: Biological Sciences, **265**(1394), 359–366, (1998)
9. Husemoller, D.: Fibre bundles, volume 5. Springer, (1966)
10. Jolliffe, I.: Principal component analysis. Wiley Online Library, (2002)
11. Kruskal, J. B.: Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika, **29**(1), 1–27, (1964)
12. Lee, A. B., Pedersen, K. S., Mumford, D.: The nonlinear statistics of high-contrast patches in natural images. International Journal of Computer Vision, **54**(1–3), 83–103, (2003)
13. Milnor, J.: Construction of universal bundles, ii. Annals of Mathematics, pages 430–436, (1956)
14. Miranda, R.: Algebraic curves and Riemann surfaces, volume 5. American Mathematical Soc., (1995)
15. Nene, S. A., Nayar, S. K., Murase, H., et al: Columbia object image library (coil-20). 1996. Data available at http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php.
16. Perea, J. A.: A brief history of persistence. Morfismos, **23**(1), 1–16, (2019)
17. Perea, J. A.: Multiscale projective coordinates via persistent cohomology of sparse filtrations. Discrete & Computational Geometry, **59**(1), 175–225, (2018)
18. Radovanović, M., Nanopoulos, A., Ivanović, M.: Hubs in space: Popular nearest neighbors in high-dimensional data. Journal of Machine Learning Research, **11**(Sep), 2487–2531, (2010)
19. Roweis, S. T., Saul, L. K.: Nonlinear dimensionality reduction by locally linear embedding. Science, **290**(5500), 2323–2326, (2000)
20. Rybakken, E., Baas, N., Dunn, B.: Decoding of neural data using cohomological learning. Neural computation **31**(1), 68–93, (2019)
21. Singh, G., Mémoli, F., Carlsson, G. E.: Topological methods for the analysis of high dimensional data sets and 3d object recognition. In SPBG, pages 91–100, (2007)
22. Tenenbaum, J. B., de Silva, V., Langford, J. C.: A global geometric framework for nonlinear dimensionality reduction. Science, **290**(5500), 2319–2323, (2000)
23. Tralie, C. J., Berger, M.: Topological Eulerian synthesis of slow motion periodic videos. In: 2018 25th IEEE International Conference on Image Processing (ICIP), pages 3573–3577, (2018)
24. Xu, B., Tralie, C. J., Antia, A., Lin, M., Perea, J. A.: Twisty Takens: A geometric characterization of good observations on dense trajectories. Journal of Applied and Computational Topology, (2019). https://doi.org/10.1007/s41468-019-00036-9