Noname manuscript No.
(will be inserted by the editor)

Persistent Cohomology and Circular Coordinates

Vin de Silva[1]***, Dmitriy Morozov**[2]**, Mikael Vejdemo-Johansson**[3]**

[1] Department of Mathematics
Pomona College
e-mail: `vin.desilva@pomona.edu`

[2] Departments of Computer Science and Mathematics
Stanford University
e-mail: `dmitriy@mrzv.org`

[3] Department of Mathematics
Stanford University
e-mail: `mik@math.stanford.edu`

The date of receipt and acceptance will be inserted by the editor

Abstract Nonlinear dimensionality reduction (NLDR) algorithms such as Isomap, LLE and Laplacian Eigenmaps address the problem of representing high-dimensional nonlinear data in terms of low-dimensional coordinates which represent the intrinsic structure of the data. This paradigm incorporates the assumption that real-valued coordinates provide a rich enough class of functions to represent the data faithfully and efficiently. On the other hand, there are simple structures which challenge this assumption: the circle, for example, is one-dimensional but its faithful representation requires two real coordinates. In this work, we present a strategy for constructing circle-valued functions on a statistical data set. We develop a machinery of persistent cohomology to identify candidates for significant circle-structures in the data, and we use harmonic smoothing and integration to obtain the circle-valued coordinate functions themselves. We suggest that this enriched class of coordinate functions permits a precise NLDR analysis of a broader range of realistic data sets.

Key words dimensionality reduction, computational topology, persistent homology, persistent cohomology

* Partially supported by DARPA, through grants HR0011-07-1-0002 and HR0011-05-1-0007.

** Partially supported by the Office of Naval Research, through grant N00014-08-1-0931.

1 Introduction

Nonlinear dimensionality reduction (NLDR) algorithms address the following problem: given a high-dimensional collection of data points $X \subset \mathbb{R}^N$, find a low-dimensional embedding $\phi : X \to \mathbb{R}^n$ (for some $n \ll N$) which faithfully preserves the 'intrinsic' structure of the data. For instance, if the data have been obtained by sampling from some unknown manifold $M \subset \mathbb{R}^N$ — perhaps the parameter space of some physical system — then ϕ might correspond to an n-dimensional coordinate system on M. If M is completely and non-redundantly parametrized by these n coordinates, then the NLDR is regarded as having succeeded completely.

Principal components analysis, or linear regression, is the simplest form of dimensionality reduction; the embedding function ϕ is taken to be a linear projection. This is closely related to (and sometimes identifed with) classical multidimensional scaling [2].

When there are no satisfactory linear projections, it becomes necessary to use NLDR. Prominent algorithms for NLDR include Locally Linear Embedding [16], Isomap [18], Laplacian Eigenmaps [1], Hessian Eigenmaps [5], and many more.

These techniques share an implicit assumption that the unknown manifold M is well-described by a finite set of coordinate functions $\phi_1, \phi_2, \ldots, \phi_n : M \to \mathbb{R}$. Explicitly, some of the correctness theorems in these studies depend on the hypothesis that M has the topological structure of a convex domain in some $\mathbb{R}^n$. This hypothesis guarantees that good coordinates exist, and shifts the burden of proof onto showing that the algorithm recovers these coordinates.

In this paper we ask what happens when this assumption fails. The simplest space which challenges the assumption is the circle, which is one-dimensional but requires two real coordinates for a faithful embedding. Other simple examples include the annulus, the torus, the figure eight, the 2-sphere, the last three of which present topological obstructions to being embedded in the Euclidean space of their natural dimension. We propose that an appropriate response to the problem is to enlarge the class of coordinate functions to include circle-valued coordinates $\theta : M \to S^1$. In a physical setting, circular coordinates occur naturally as angular and phase variables. Spaces like the annulus and the torus are well described by a combination of real and circular coordinates. (The 2-sphere is not so lucky, and must await its day.)

The goal of this paper is to describe a natural procedure for constructing circular coordinates on a nonlinear data set using techniques from classical algebraic topology and its 21st-century grandchild, persistent topology. We direct the reader to [10] as a general reference for algebraic topology, and to [6] for a survey of the theory of persistence. We also recommend [19] for a more technical description of persistent homology.

1.1 Related work

There have been other attempts to address the problem of finding good coordinate representations of simple non-Euclidean data spaces. One approach [15] is to use modified versions of multidimensional scaling specifically devised to find the best embedding of a data set into the cylinder, the sphere and so on. The target space has to be chosen in advance. Another class of approaches [11, 4] involves cutting the data manifold along arcs and curves until it has trivial topology. The resulting configuration can then be embedded in Euclidean space in the usual way. In our approach, the number of circular coordinates is not fixed in advance, but is determined experimentally after a persistent homology calculation. Moreover, there is no cutting involved; the coordinate functions respect the original topology of the data.

1.2 Overview

The principle behind our algorithm is the following equation from homotopy theory, valid for topological spaces X with the homotopy type of a cell complex (which covers everything we normally encounter):

$$[X, S^1] = \mathrm{H}^1(X; \mathbb{Z}) \tag{1}$$

The left-hand side denotes the set of equivalence classes of continuous maps from X to the circle S^1; two maps are equivalent if they are homotopic (meaning that one map can be deformed continuously into the other); the right-hand side denotes the 1-dimensional cohomology of X, taken with integer coefficients. In other language: S^1 is the classifying space for H^1, or equivalently S^1 is the Eilenberg–MacLane space $K(\mathbb{Z}, 1)$. See section 4.3 of [10].

If X is a contractible space (such as a convex subset of $\mathbb{R}^n$), then $\mathrm{H}^1(X; \mathbb{Z}) = 0$ and equation (1) tells us not to bother looking for circular functions: any such function is homotopic to a constant function, and can therefore be lifted to a real-valued function. On the other hand, if X has nontrivial topology then there may well exist a nonzero cohomology class $[\alpha] \in \mathrm{H}^1(X; \mathbb{Z})$; we can then build a continuous function $X \to S^1$ which in some sense reveals $[\alpha]$.

Our strategy divides into the following steps.

1. Represent the given discrete data set as a simplicial complex or filtered simplicial complex.
2. Use persistent cohomology to identify a ‘significant’ cohomology class in the data. For technical reasons, we carry this out with coefficients in the field $\mathbb{F}_p$ of integers modulo p, for some prime p. This gives us $[\alpha_p] \in \mathrm{H}^1(X; \mathbb{F}_p)$.
3. Lift $[\alpha_p]$ to a cohomology class with integer coefficients: $[\alpha] \in \mathrm{H}^1(X; \mathbb{Z})$.

4. Smoothing: replace the integer cocycle α by a harmonic cocycle in the same cohomology class: $\bar{\alpha} \in \mathrm{C}^1(X;\mathbb{R})$.
5. Integrate the harmonic cocycle $\bar{\alpha}$ to a circle-valued function $\theta : X \to S^1$.

The paper is organized as follows. In Section 2.1, we derive what we need of equation (1). Steps (1–5) of the algorithm are addressed in Sections 2.2–2.6, respectively. The correctness of the algorithm for persistent cocycles is addressed in an appendix, Section 2.A.

In Section 3 we report some experimental results.

2 Algorithm Details

2.1 Cohomology and circular functions

Let X be a finite simplicial complex. Let X^0, X^1, X^2 denote the sets of vertices, edges and triangles of X, respectively. We suppose that the vertices are totally ordered (in an arbitrary way). If $a < b$ then the edge between vertices a, b is always written ab and not ba. Similarly, if $a < b < c$ then the triangle with vertices a, b, c is always written abc.

Cohomology can be defined as follows. Let $\mathbb{A}$ be a commutative ring (for example $\mathbb{A} = \mathbb{Z}, \mathbb{F}_p, \mathbb{R}$). We define 0-cochains, 1-cochains, and 2-cochains as follows:

$$\begin{aligned} \mathrm{C}^0 &= \mathrm{C}^0(X;\mathbb{A}) = \big\{\text{functions } f : X^0 \to \mathbb{A}\big\} \\ \mathrm{C}^1 &= \mathrm{C}^1(X;\mathbb{A}) = \big\{\text{functions } \alpha : X^1 \to \mathbb{A}\big\} \\ \mathrm{C}^2 &= \mathrm{C}^2(X;\mathbb{A}) = \big\{\text{functions } A : X^2 \to \mathbb{A}\big\} \end{aligned}$$

These are modules over $\mathbb{A}$. We now define coboundary maps $d_0 : \mathrm{C}^0 \to \mathrm{C}^1$ and $d_1 : \mathrm{C}^1 \to \mathrm{C}^2$.

$$\begin{aligned} (d_0 f)(ab) &= f(b) - f(a) \\ (d_1 \alpha)(abc) &= \alpha(bc) - \alpha(ac) + \alpha(ab) \end{aligned}$$

Let $\alpha \in \mathrm{C}^1$. If $d_1\alpha = 0$ we say that α is a *cocycle*. If $d_0 f = \alpha$ admits a solution $f \in \mathrm{C}^0$ we say that α is a *coboundary*. The solution f, if it exists, can be thought of as the discrete integral of α. It is unique up to adding constants on each connected component of X.

It is easily verified that $d_1 d_0 f = 0$ for any $f \in \mathrm{C}^0$. Thus, coboundaries are always cocycles, or equivalently $\mathrm{Im}(d_0) \subseteq \mathrm{Ker}(d_1)$. We can measure the difference between coboundaries and cocycles by defining the 1-cohomology of X to be the quotient module

$$\mathrm{H}^1(X;\mathbb{A}) = \mathrm{Ker}(d_1)/\,\mathrm{Im}(d_0).$$

We say that two cocycles α, β are *cohomologous* if $\alpha - \beta$ is a coboundary.

We now consider integer coefficients. The following proposition fulfils part of the promise of equation (1), by producing circle-valued functions from integer cocycles. It will be helpful to think of S^1 as the quotient group $\mathbb{R}/\mathbb{Z}$.

Proposition 1 *Let $\alpha \in \mathrm{C}^1(X;\mathbb{Z})$ be a cocycle. Then there exists a continuous function $\theta : X \to \mathbb{R}/\mathbb{Z}$ which maps each vertex to 0, and each edge ab around the entire circle with winding number $\alpha(ab)$.*

Proof We can define θ inductively on the vertices, edges, triangles, ... of X. The vertices and edges follow the prescription in the statement of the proposition. To extend θ to the triangles, it is necessary that the winding number of θ along the boundary of each triangle abc is zero. And indeed this is $\alpha(bc) - \alpha(ac) + \alpha(ab) = d_1\alpha(abc) = 0$. Since the higher homotopy groups of S^1 are all zero ([10], section 4.3), θ can then be extended to the higher cells of X without obstruction. □

The construction in Proposition 1 is unsatisfactory in the sense that all vertices are mapped to the same point. All variation in the circle parameter takes place in the interior of the edges (and higher cells). This is rather unsmooth. For more leeway, we consider real coefficients.

Proposition 2 *Let $\bar{\alpha} \in \mathrm{C}^1(X;\mathbb{R})$ be a cocycle. Suppose we can find $\alpha \in \mathrm{C}^1(X;\mathbb{Z})$ and $f \in \mathrm{C}^0(X;\mathbb{R})$ such that $\bar{\alpha} = \alpha + d_0 f$. Then there exists a continuous function $\theta : X \to \mathbb{R}/\mathbb{Z}$ which maps each edge ab linearly to an interval of length $\bar{\alpha}(ab)$, measured with sign.*

In other words, we can construct a circle-valued function out of any real cocycle $\bar{\alpha}$ whose cohomology class $[\bar{\alpha}]$ lies in the image of the natural homomorphism $\mathrm{H}^1(X;\mathbb{Z}) \to \mathrm{H}^1(X;\mathbb{R})$.

Proof Define θ on the vertices of X by setting $\theta(a)$ to be $f(a)$ mod $\mathbb{Z}$. For each edge ab, we have

$$\begin{aligned}\theta(b) - \theta(a) &= f(b) - f(a)\\ &= d_0 f(ab)\\ &= \bar{\alpha}(ab) - \alpha(ab)\end{aligned}$$

which is congruent to $\bar{\alpha}(ab)$ mod $\mathbb{Z}$, since $\alpha(ab)$ is an integer.

It follows that θ can be taken to map ab linearly onto an interval of signed length $\bar{\alpha}(ab)$. Since $\bar{\alpha}$ is a cocycle, θ can be extended to the triangles as before; then to the higher cells. □

Proposition 2 suggests the following tactic: from an integer cocycle α we construct a cohomologous real cocycle $\bar{\alpha} = \alpha + d_0 f$, and then define $\theta = f$ mod $\mathbb{Z}$ on the vertices of X. If we can construct $\bar{\alpha}$ so that the edge-lengths $|\bar{\alpha}(ab)|$ are small, then the behaviour of θ will be apparent from its restriction to the vertices. See Section 2.5.

2.2 Point-cloud data to simplicial complex

We now begin describing the workflow in detail. The input is a point-cloud data set: in other words, a finite set $S \subset \mathbb{R}^N$ or more generally a finite metric space. The first step is to convert S into a simplicial complex and to identify a stable-looking integer cohomology class. This will occupy the next three subsections.

The first lesson of point-cloud topology [8] is that point-clouds are best represented by 1-parameter nested families of simplicial complexes. There are several candidate constructions: the Vietoris–Rips complex $X^\epsilon = \mathrm{Rips}(S, \epsilon)$ has vertex set S and includes a k-simplex whenever all $k+1$ vertices lie pairwise within distance ϵ of each other. The witness complex $X^\epsilon = \mathrm{Witness}(L, S, \epsilon)$ uses a smaller vertex set $L \subset S$ and includes a k-simplex when the $k+1$ vertices lie close to other points of S, in a certain precise sense (see [3,9]). In both cases, $X^\epsilon \subseteq X^{\epsilon'}$ whenever $\epsilon \leq \epsilon'$. Either of these constructions will serve our purposes, but the witness complex has the computational advantage of being considerably smaller.

We determine X^ϵ only up to its 2-skeleton, since we are interested in H^1.

2.3 Persistent cohomology

Having constructed a 1-parameter family $\{X^\epsilon\}$, we apply the principle of persistence to identify cocycles that are stable across a large range for ϵ. Suppose that $\epsilon_1, \epsilon_2, \dots, \epsilon_m$ are the critical values where the complex X^ϵ gains new cells. The family can be represented as a diagram

$$X^{\epsilon_1} \longrightarrow X^{\epsilon_2} \longrightarrow \dots \longrightarrow X^{\epsilon_m}$$

of simplicial complexes and inclusion maps. For any coefficient field $\mathbb{F}$, the cohomology functor $\mathrm{H}^1(-;\mathbb{F})$ converts this diagram into a diagram of vector spaces and linear maps over $\mathbb{F}$; the arrows are reversed:

$$\mathrm{H}^1(X^{\epsilon_1};\mathbb{F}) \longleftarrow \mathrm{H}^1(X^{\epsilon_2};\mathbb{F}) \longleftarrow \dots \longleftarrow \mathrm{H}^1(X^{\epsilon_m};\mathbb{F})$$

According to the theory of persistence [7,19], such a diagram decomposes as a direct sum of 1-dimensional terms indexed by half-open intervals of the form $[\epsilon_i, \epsilon_j)$. Each such term corresponds to a cochain $\alpha \in \mathrm{C}^i(X^\epsilon)$ that satisfies the cocycle condition for $\epsilon < \epsilon_j$ and becomes a coboundary for $\epsilon < \epsilon_i$. The collection of intervals can be displayed graphically as a *persistence diagram*, by representing each interval $[\epsilon_i, \epsilon_j)$ as a point (ϵ_i, ϵ_j) in the Cartesian plane above the main diagonal. We think of long intervals as representing trustworthy (i.e. stable) topological information.

Remark. This is where we start worrying about the coefficient ring. The persistence decomposition theorem applies to diagrams of vector spaces over a field. When we work over the ring of integers $\mathbb{Z}$, however, the result is known to fail: there need not be an interval decomposition. This is unfortunate, since we require integer cocycles to construct circle maps. To finesse this problem, we pick an arbitrary prime number p (such as $p = 47$) and carry out our persistence calculations over the finite field $\mathbb{F} = \mathbb{F}_p$. The resulting $\mathbb{F}_p$ cocycle must then be converted to integer coefficients: we address this in Section 2.4.

In principle we can use the ideas in [19] to calculate the persistent cohomology intervals and then select a long interval $[\epsilon_i, \epsilon_j)$ and a specific $\delta \in [\epsilon_i, \epsilon_j)$. We then let $X = X^\delta$ and take α to be the cocycle in $\mathrm{C}^1(X; \mathbb{F})$ corresponding to the interval.

Persistent cocycle algorithm. Explicitly, persistent cocycles can be calculated in the following way. We discuss the correctness of this algorithm in Section 2.A.

Suppose that the simplices in the filtered complex are totally ordered, and labelled $\sigma_1, \sigma_2, \dots, \sigma_m$ so that σ_j arrives at time ϵ_j, where the sequence (ϵ_j) is non-decreasing. Write $X_\ell = \sigma_1 \cup \sigma_2 \cup \cdots \cup \sigma_\ell$. A cochain $\alpha \in \mathrm{C}^*(X_\ell) = \mathrm{C}^*(X_\ell; \mathbb{F})$ can be represented as a vector $(a_1, a_2, \dots, a_\ell)$, where $a_j = \alpha(\sigma_j)$. The cochains corresponding to the standard basis vectors are denoted $\hat{\sigma}_1, \hat{\sigma}_2, \dots, \hat{\sigma}_\ell$.

We iterate over $\ell = 0, 1, \dots, m$, maintaining the following information as we go:

- a set of indices $I_\ell \subseteq \{1, 2, \dots, \ell\}$ associated with 'live' cocycles;
- a list of cocycles $(\alpha_i : i \in I_\ell)$ in $\mathrm{C}^*(X_\ell)$.

The cocycle α_i involves only σ_i and those simplices of the same dimension that appear later in the filtration sequence (thus only σ_j with $j \geq i$).

Initialize ($\ell = 0$): Set $I_0 = \emptyset$. The list of cocycles is empty.

Update (from $\ell - 1$ to ℓ): Our convention is to extend each cochain $\alpha = (a_1, a_2, \dots, a_{\ell-1})$ in $\mathrm{C}^*(X_{\ell-1})$ to a cochain $\alpha = (a_1, \dots, a_{\ell-1}, 0)$ in $\mathrm{C}^*(X_\ell)$ by appending 0. We still call it α.

Begin by computing, for each $i \in I_{\ell-1}$, the coboundaries of the cocycles α_i of $X_{\ell-1}$ within the larger complex X_ℓ. Since $d\alpha_i = 0$ in $\mathrm{C}^*(X_{\ell-1})$, it follows that the coboundary $d\alpha_i$ in $\mathrm{C}^*(X_\ell)$ must be a multiple of the newest basis vector $\hat{\sigma}_\ell = (0, \dots, 0, 1)$. Write $d\alpha_i = c_i \hat{\sigma}_\ell$.

- If all the c_i are zero, then we have one new cocycle: let $I_\ell = I_{\ell-1} \cup \{\ell\}$ and define $\alpha_\ell = \hat{\sigma}_\ell$.
- Otherwise, we lose a cocycle. Let $j \in I_{\ell-1}$ be the largest index for which $c_j \neq 0$. Delete α_j by setting $I_\ell = I_{\ell-1} \setminus \{j\}$, and restore the earlier cocycles by setting $\alpha_i \leftarrow \alpha_i - (c_i/c_j)\alpha_j$. The 'lost' cocycle is recorded for posterity: write the persistence interval $[\epsilon_j, \epsilon_\ell)$ to the output, together with its associated cocycle α_j.

Finish ($\ell = m$): Surviving cocycles are associated with semi-infinite intervals. For each $i \in I_m$, write the interval $[\epsilon_i, \infty)$ to the output, together with its associated cocycle α_i.

Remark. The reader may be more familiar with persistence diagrams in homology rather than cohomology. In fact, the universal coefficient theorem [10] implies that the two diagrams are identical. The salient point is that cohomology is the vector-space dual of homology, when working with field coefficients. That said, we cannot simply use the usual algorithm for persistent homology: we are interested in obtaining explicit cocycles, whereas the classical algorithm [19] returns cycles.

After completing the persistent cocycle calculation, up to some parameter value $\epsilon_{\max}$, we are left with a collection of finite and semi-infinite persistence intervals. For the next step, we select one such interval and a parameter value $\delta \leq \epsilon_{\max}$ contained in it. Henceforth, we fix our attention on the complex X^δ. The cocycle associated to the interval can be regarded as a cocycle on X^δ, by restriction. If we are working over the field $\mathbb{F}_p$, we denote this cocycle α_p.

In some of the experimental examples in Section 3, we consider several persistence intervals at once, and use a value of δ common to all of them. This can be done elegantly using the persistence diagram. Select a point (δ, δ) on the diagonal and draw the upper-left quadrant at that point. The chosen persistence intervals must appear in the diagram as points in that quadrant. We use this visual convention in all of our examples.

2.4 Lifting to integer coefficients

We now have a simplicial complex $X = X^\delta$ and a cocycle $\alpha_p \in \mathrm{C}^1(X; \mathbb{F}_p)$. The next step is to 'lift' α_p by constructing an integer cocycle α which reduces to α_p modulo p.

Theory. To show that this is (almost) always possible, note that the short exact sequence of coefficient rings $0 \longrightarrow \mathbb{Z} \xrightarrow{\cdot p} \mathbb{Z} \longrightarrow \mathbb{F}_p \longrightarrow 0$ gives rise to a long exact sequence, called the Bockstein sequence (see Section 3.E of [10]). Here is the relevant section of the sequence:

$$\to \mathrm{H}^1(X;\mathbb{Z}) \to \mathrm{H}^1(X;\mathbb{F}_p) \overset{\beta}{\to} \mathrm{H}^2(X;\mathbb{Z}) \overset{\cdot p}{\to} \mathrm{H}^2(X;\mathbb{Z}) \to$$

By exactness, the Bockstein homomorphism β induces an isomorphism between the cokernel of $\mathrm{H}^1(X;\mathbb{Z}) \to \mathrm{H}^1(X;\mathbb{F}_p)$ and the kernel of $\mathrm{H}^2(X;\mathbb{Z}) \overset{\cdot p}{\to} \mathrm{H}^2(X;\mathbb{Z})$, and this kernel is precisely the set of p-torsion elements of $\mathrm{H}^2(X;\mathbb{Z})$. If there is no p-torsion, then it follows immediately that the cokernel of the first map is zero. In other words $\mathrm{H}^1(X;\mathbb{Z}) \to \mathrm{H}^1(X;\mathbb{F}_p)$ is surjective; any cocycle $\alpha_p \in \mathrm{C}^1(X;\mathbb{F}_p)$ can be lifted to a cocycle $\alpha \in \mathrm{C}^1(X;\mathbb{Z})$.

If we are unluckily sabotaged by p-torsion, then we pick another prime and redo the calculation from scratch: it is enough to pick a prime that does not divide the order of the torsion subgroup of $\mathrm{H}^2(X;\mathbb{Z})$, so almost any prime will do.

Practice. We construct α by taking the coefficients of α_p in $\mathbb{F}_p$ and replacing them with integers in the correct congruence class modulo p. The default is to choose coefficients close to zero; that is, in the range

$$\{-(p-1)/2, \ldots, -1, 0, 1, \ldots, (p-1)/2\}$$

when p is an odd prime. (We do not recommend using $p = 2$; there is no way to distinguish 1 from -1.)

We then evaluate $d_1\alpha$. If $d_1\alpha = 0$ then α is a cocycle and we are done. Otherwise, it becomes necessary to do some repair work. Certainly $d_1\alpha \equiv 0$ modulo p, so we can write $d_1\alpha = p\eta$ for some $\eta \in \mathrm{C}^2(X;\mathbb{Z})$. To effect the repair, we must write η as a coboundary by solving the equation $\eta = d_1\zeta$ for $\zeta \in \mathrm{C}^1(X;\mathbb{Z})$. Given a solution, the 1-cochain $\alpha - p\zeta$ is the required lift of α_p, since $d_1(\alpha - p\zeta) = p\eta - p\eta = 0$.

When can this fail? We know that $p\eta$ is a coboundary (indeed $p\eta = d_1\alpha$), and we know that η is a cocycle (since $p(d_1\eta) = d_1(p\eta) = d_1 d_1\alpha = 0$). Thus we have a cohomology class $[\eta]$ in $\mathrm{H}^1(X;\mathbb{Z})$ such that $p[\eta] = [p\eta]$ is zero in cohomology. If $\mathrm{H}^2(X;\mathbb{Z})$ has no p-torsion, then $[\eta]$ must itself be zero, meaning that η is a coboundary and there exists a solution to $\eta = d_1\zeta$. On the other hand, if $\mathrm{H}^2(X;\mathbb{Z})$ has p-torsion then there is no such guarantee.

This is all very well. Unfortunately, the equation $\eta = d_1\zeta$ is a Diophantine linear system. At present, we can provide no particular guidance as to how to solve the system (other than by vague appeal to off-the-shelf Diophantine or integer linear programming solvers), even if we know that a solution exists. Fortunately, and mysteriously, this has not proved necessary in any of our examples. In our experiments, the heuristic of lifting to integer coefficients close to zero (that is, between $\pm(p-1)/2$) produces a cocycle every time. We wonder why.

To finish this section, we draw attention to a basic fact from classical algebraic topology.

Proposition 3 *Let X be a finite simplicial complex. Then* $\mathrm{H}^1(X;\mathbb{Z})$ *is torsion free, and* $\mathrm{H}^2(X;\mathbb{Z})$ *has the same torsion as* $\mathrm{H}_1(X;\mathbb{Z})$.

Proof More generally, $\mathrm{H}^{k+1}(X;\mathbb{Z})$ and $\mathrm{H}_k(X;\mathbb{Z})$ have isomorphic torsion subgroups. This is a consequence of the universal coefficient theorems for homology and cohomology: see [10, Corollary 3.3]. For the first statement, note that $\mathrm{H}_0(X;\mathbb{Z})$ is the free abelian group generated by the connected components of X. It is therefore torsion-free, hence so is $\mathrm{H}^1(X;\mathbb{Z})$. □

REMARK. We expect that p-torsion is extremely rare in 'real' data sets, since it is symptomatic of rather subtle topological phenomena. For instance, the simplest examples which exhibit 2-torsion are the nonorientable closed surfaces (such as the projective plane and the Klein bottle). For a 'randomly' chosen prime p, one would be very surprised to find p-torsion arising from a statistical data set. We do not know how to quantify this.

At any rate, the arguments in this section show us that we can recognize torsion trouble when it occurs, by observing the failure of $d_1\alpha = 0$ for the chosen lift α. We then have the choice of changing primes or setting up an appropriate integer linear programming problem.

2.5 Harmonic smoothing

Given an integer cocycle $\alpha \in \mathrm{C}^1(X;\mathbb{Z})$, or indeed a real cocycle $\alpha \in \mathrm{C}^1(X;\mathbb{R})$, we wish to find the 'smoothest' real cocycle $\bar{\alpha} \in \mathrm{C}^1(X;\mathbb{R})$ cohomologous to α. It turns out that what we want is the harmonic cocycle representing the cohomology class $[\alpha]$.

We define smoothness. Each of the spaces $\mathrm{C}^i(X;\mathbb{R})$ comes with a natural Euclidean metric:

$$\begin{aligned} \|f\|^2 &= \sum_{a \in X^0} |f(a)|^2, \\ \|\alpha\|^2 &= \sum_{ab \in X^1} |\alpha(ab)|^2, \\ \|A\|^2 &= \sum_{abc \in X^2} |A(abc)|^2. \end{aligned}$$

A circle-valued function θ is 'smooth' if its total variation across the edges of X is small. The terms $|\bar{\alpha}(ab)|^2$ capture the variation across individual edges; therefore what we must minimize is $\|\bar{\alpha}\|^2$.

Proposition 4 *Let $\alpha \in \mathrm{C}^1(X;\mathbb{R})$. There is a unique solution $\bar{\alpha}$ to the least-squares minimization problem*

$$\operatorname*{argmin}_{\bar{\alpha}} \left\{ \|\bar{\alpha}\|^2 \mid \exists f \in \mathrm{C}^0(X;\mathbb{R}),\ \bar{\alpha} = \alpha + d_0 f \right\}. \tag{2}$$

Moreover, $\bar{\alpha}$ is characterized by the equation $d_0^* \bar{\alpha} = 0$, where d_0^* is the adjoint of d_0 with respect to the inner products on $\mathrm{C}^0, \mathrm{C}^1$.

Proof Note that if $d_0^* \bar{\alpha} = 0$ then for any $f \in \mathrm{C}^0$ we have

$$\begin{aligned} \|\bar{\alpha} + d_0 f\|^2 &= \|\bar{\alpha}\|^2 + 2\langle \bar{\alpha}, d_0 f\rangle + \|d_0 f\|^2 \\ &= \|\bar{\alpha}\|^2 + 2\langle d_0^* \bar{\alpha}, f\rangle + \|d_0 f\|^2 \\ &= \|\bar{\alpha}\|^2 + \|d_0 f\|^2 \end{aligned}$$

which implies that such an $\bar{\alpha}$ must be the unique minimizer. For existence, note that

$$d_0^* \alpha + d_0^* d_0 f = 0$$

certainly has a solution f if $\mathrm{Im}(d_0^*) = \mathrm{Im}(d_0^* d_0)$. But this is a standard fact in finite-dimensional linear algebra: $\mathrm{Im}(A^{\mathrm{T}}) = \mathrm{Im}(A^{\mathrm{T}} A)$ for any real matrix A; this follows from the singular value decomposition, for instance. □

It is customary to construct the Laplacian $\Delta = d_1^* d_1 + d_0 d_0^*$. The twin equations $d_1 \bar{\alpha} = 0$ and $d_0^* \bar{\alpha} = 0$ immediately imply (and conversely, can be deduced from) the single equation $\Delta\bar{\alpha} = 0$; in other words $\bar{\alpha}$ is *harmonic*.

Remark. The space of harmonic 1-forms $\mathcal{H}^1 = \mathrm{Ker}(\Delta)$ is naturally isomorphic to both the cohomology $\mathrm{H}^1(X;\mathbb{R})$ and the homology $\mathrm{H}_1(X;\mathbb{R})$ with

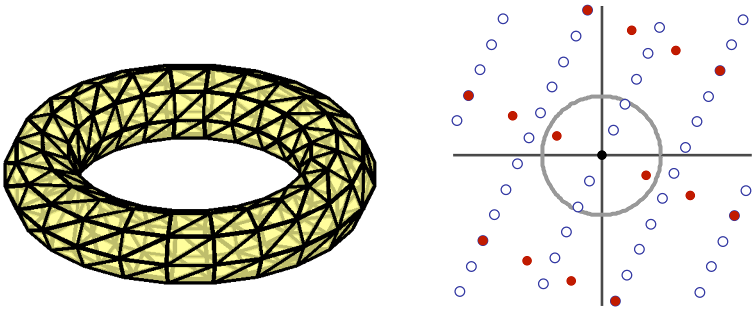


Fig. 1 A torus, and the integer cohomology ● and homology ○ lattices of its harmonic space $\mathcal{H}^1$. The two lattices are dual with respect to the inner product whose unit circle is shown. We seek points in the cohomology lattice.

real coefficients. These are related to the integer cohomology and homology groups via natural maps:

$$\mathrm{H}^1(X;\mathbb{Z}) \to \mathrm{H}^1(X;\mathbb{R}) = \mathcal{H}^1(X) = \mathrm{H}_1(X;\mathbb{R}) \leftarrow \mathrm{H}_1(X;\mathbb{Z})$$

For our purposes (following Propositions 1 and 2) we seek points in the image of the map $\mathrm{H}^1(X;\mathbb{Z}) \to \mathcal{H}^1(X)$. The set of these points is a full-rank discrete lattice of the real vector space $\mathcal{H}^1(X)$. The Diophantine nature of our calculations arises from the fact that we are trying to work in a lattice.

Remark. Dual to the integer cohomology lattice is the integer homology lattice, which is the image of the map $\mathrm{H}_1(X;\mathbb{Z}) \to \mathcal{H}^1(X)$. The two lattices are generally different. This is why we must compute persistent cocycles rather than cycles. See Figure 1.

2.6 Integration

The least-squares problem in equation (2) can be solved using a standard algorithm such as LSQR [14]. By Proposition 2 we can use the solution parameter f to define the circular coordinate θ on the vertices of X: simply let θ be the reduction of f modulo $\mathbb{Z}$. This works because the original cocycle α has integer coefficients.

Remark. More generally, if $\bar{\alpha}$ is an arbitrary real cocycle such that

$$[\bar{\alpha}] \in \mathrm{Im}(\mathrm{H}^1(X;\mathbb{Z}) \to \mathrm{H}^1(X;\mathbb{R})),$$

it is a straightforward matter to integrate $\bar{\alpha}$ to a circle-valued function θ on the vertex set X^0. Suppose that X is connected (if not, each connected component can be treated separately) and pick a starting vertex x_0 and assign $\theta(x_0) = 0$. One can use Dijkstra's algorithm to find shortest paths to each remaining vertex from x_0. When a new vertex b enters the structure via an edge ab, we assign $\theta(b) = \theta(a) + \bar{\alpha}(ab)$ (or $\theta(a) - \bar{\alpha}(ba)$ if the edge is correctly identified as ba). If a vertex a is connected to x_0 by multiple paths then the different possible values of $\theta(a)$ differ by an integer; this is where we use the hypothesis that $\bar{\alpha}$ is cohomologous to an integer cocycle.

2.7 Summary

The procedure described above seeks a 1-cocycle $\bar{\alpha}$ with real coefficients which is:

- harmonic (for smoothness)
- in the integer cohomology lattice (for integrability to $S^1 = \mathbb{R}/\mathbb{Z}$)
- persistent (for geometric significance)

The circular coordinate θ is obtained by integrating $\bar{\alpha}$, either by brute force or as a side-effect of the smoothing step.

In order to compute persistent cocycles we are forced to work over a field, so we choose $\mathbb{F}_p$ and then attempt to lift the results to $\mathbb{Z}$. This step may fail if $\mathrm{H}^2(X;\mathbb{Z})$ (or equivalently $\mathrm{H}_1(X;\mathbb{Z})$) has nontrivial p-torsion. Even when the lifting problem has a solution, we might have to solve a Diophantine linear system to find it.

2.A Correctness of the cocycle algorithm.

The persistent cocycle algorithm is a stripped-down version of a more complete calculation, which we describe now. The output of this calculation is the following information:

- A partition $\{1, 2, \ldots, m\} = I \cup P \cup Q$, (where I, P, Q are disjoint).
- A bijective pairing between the sets P, Q. We write $p \lhd q$ to indicate that p is paired with q.
- An 'echelon basis' $\alpha_1, \alpha_2, \ldots, \alpha_m$ for $\mathrm{C}^*(X_m)$. By 'echelon' we mean that α_j involves σ_j (with a nonzero coefficient) and subsequent cells only. In vector notation, each α_j is of the form
$$\alpha_j = (0, \ldots, 0, a^j_j, a^j_{j+1}, \ldots, a^j_m)$$
where $a^j_j \neq 0$.
- The coboundaries of the basis cochains α_j are:
$$\begin{aligned} d\alpha_i &= 0 && \text{for } i \in I, && (*_i)\\ d\alpha_p &= \alpha_q && \text{for } p \in P \text{ with } p \lhd q. && (*_p)\\ d\alpha_q &= 0 && \text{for } q \in Q, && (*_q) \end{aligned}$$

Note that the echelon form implies that the kernel of each restriction map $\mathrm{C}^*(X_m) \to \mathrm{C}^*(X_j)$ is spanned by the cochains $\alpha_{j+1}, \dots, \alpha_m$.

The key point is that the persistent cohomology of the filtered complex can be deduced from any partition, pairing, and echelon basis which satisfy the coboundary equations $(*_i)$, $(*_p)$ and $(*_q)$. Indeed, the equations imply that the space of coboundaries in $\mathrm{C}^*(X_j)$ has basis consisting of the (restrictions of the) cochains

$$\alpha_q \quad \text{for } q \in Q \text{ with } q \leq j,$$

and the space of cocycles has basis consisting of these boundary cochains together with the (restrictions of the) cochains

$$\begin{aligned} &\alpha_i && \text{for } i \in I \text{ with } i \leq j, \\ &\alpha_p && \text{for } p \in P \text{ with } p \lhd q \text{ and } p \leq j < q. \end{aligned}$$

Thus, each α_i, for $i \in I$, restricts to a nonzero cocycle over the index range $\{i, \dots, m\}$; and each α_p, for $p \in P$ with $p \lhd q$, restricts to a nonzero cocycle over the index range $\{p, \dots, q-1\}$. These give us persistence intervals $[\epsilon_i, \infty)$ and $[\epsilon_p, \epsilon_q)$ respectively.

We now describe the computation, carried out iteratively. Suppose we have determined a partition

$$\{1, \dots, \ell - 1\} = I_{\ell-1} \cup P_{\ell-1} \cup Q_{\ell-1},$$

a pairing $\lhd$, and an echelon basis $\alpha_1, \dots, \alpha_{\ell-1}$ for $\mathrm{C}^*(X_{\ell-1})$, with coboundaries as above. We now add the cell σ_ℓ.

The immediate impact is that coboundaries computed in $\mathrm{C}^*(X_\ell)$ have an extra coefficient for the new cell. Thus, for some scalars $c_1, c_2, \dots, c_{\ell-1}$ we have

$$\begin{aligned} d\alpha_i &= c_i \hat{\sigma}_\ell && \text{for } i \in I_{\ell-1}, \\ d\alpha_p &= \alpha_q + c_p \hat{\sigma}_\ell && \text{for } p \in P_{\ell-1} \text{ with } p \lhd q. \\ d\alpha_q &= c_q \hat{\sigma}_\ell && \text{for } q \in Q_{\ell-1}, \end{aligned}$$

We can begin defining a new echelon basis $\bar{\alpha}_1, \bar{\alpha}_2, \dots, \bar{\alpha}_\ell$ as follows:

$$\begin{aligned} \bar{\alpha}_p &= \alpha_p && \text{for } p \in P_{\ell-1} \\ \bar{\alpha}_q &= d\bar{\alpha}_p = \alpha_q + c_p \hat{\sigma}_\ell && \text{for } q \in Q_{\ell-1} \text{ with } p \lhd q. \end{aligned}$$

Note that the leading term of $\bar{\alpha}_q$ is unchanged from α_q, and that $d\bar{\alpha}_q = d(d\alpha_p) = 0$.

Now we must consider $\bar{\alpha}_i$ for $i \in I_{\ell-1}$, and $\bar{\alpha}_\ell$.

Case 1: each $c_i = 0$, for $i \in I_{\ell-1}$. Then we can set $\bar{\alpha}_i = \alpha_i$ for each $i \in I_{\ell-1}$, and $\bar{\alpha}_\ell = \hat{\sigma}_\ell$. We set

$$I_\ell = I_{\ell-1} \cup \{\ell\}, \quad P_\ell = P_{\ell-1}, \quad Q_\ell = Q_{\ell-1},$$

and the coboundary equations $(*_i)$, $(*_p)$, $(*_q)$ are clearly satisfied.

Case 2: some $c_i \neq 0$, for $i \in I_{\ell-1}$. Let j be the largest such index. Define

$$\begin{aligned} &\bar{\alpha}_j = \alpha_j, \\ &\bar{\alpha}_i = \alpha_i - (c_i/c_j)\alpha_j \quad \text{for } i \in I_{\ell-1} \text{ with } i \neq j, \\ &\bar{\alpha}_\ell = d\alpha_j = c_j\hat{\sigma}. \end{aligned}$$

The echelon property still holds (since j was chosen largest). If we set

$$I_\ell = I_{\ell-1} \setminus \{j\}, \quad P_\ell = P_{\ell-1} \cup \{j\}, \quad Q_\ell = Q_{\ell-1} \cup \{\ell\},$$

and extend the pairing by adding the relation $j \lhd \ell$, then it is easily seen that the coboundary equations are satisfied.

The persistent cocycle algorithm can be thought of as a 'forgetful' or 'neglectful' version of the the calculation above. We maintain only the I_ℓ and the echelon basis vectors α_i. The index sets P_ℓ and Q_ℓ, the pairing $\lhd$, and the remaning basis vectors are not necessary for this. We write each interval $[\epsilon_p, \epsilon_q)$ to output as soon as we identify a pair $p \lhd q$, but we immediately discard the pairing information from memory. At the end we collect the remaining intervals $[\epsilon_i, \infty)$.

Thus, the correctness of the cocycle algorithm follows from the correctness of the full cohomology algorithm. The correctness of the cohomology algorithm follows from the fact that the persistent cohomology can be deduced from any partition, pairing and echelon basis which satisfy the coboundary equations.

3 Experiments

3.1 Software

Early experimental trials were performed with the Java-based jPlex simplicial complex software [17]. The present results and timings are obtained with the C++ library Dionysus [12]. We used Paige and Saunders' implementation of LSQR [13] for the least-squares problems in the harmonic smoothing step.

3.2 General procedure

We tested our methods on several synthetic data sets with known topology, ranging from the humble circle itself to a genus-2 surface ('double torus'). Most of the examples were embedded in $\mathbb{R}^2$ or $\mathbb{R}^3$, with the exception of a sample from a complex projective curve (embedded in $\mathbb{C}P^2$) and a synthetic image-like data set (embedded in $\mathbb{R}^{120000}$).

In each case we selected vertices for the filtered simplicial complex: either the whole set, or a smaller well-distributed subset of 'landmarks' selected by iterative furthest-point sampling. We then built a Rips or witness complex, with maximum radius generally chosen to ensure around 10^5 simplices in the complex.

In most cases, we show the persistence diagram produced by the cocycle computation. The chosen value δ is marked on the diagonal, with its upper-left quadrant indicated in green lines. The persistent cocycles available at parameter value δ are precisely those contained in that quadrant. Each of those cocycles then produces a circular coordinate.

There are various figures associated with each example. Most important are the correlation scatter plots: each scatter plot compares two circular coordinate functions. These may be functions produced by the computation ('inferred coordinates') or known parameters. These scatter plots are drawn in the unit square, which is of course really a torus $S^1 \times S^1$.

When the original data are embedded in $\mathbb{R}^2$ or $\mathbb{R}^3$, we also display the circular coordinates directly on the data set, plotting each point in color according to its coordinate value interpreted on the standard hue-circle. This works less well in grayscale reproductions, of course.

Finally, in certain cases we plot coordinate values against frequency, as a histogram. This distributional information can sometimes be useful in the absence of other information.

Remark. When the goal is to infer the topology of a data set whose structure is unknown, we do not have any 'known parameters' available to us. We can still construct correlation scatter plots between pairs of inferred coordinates, and the distributional histograms for each coordinate individually. We exhort the reader to view the following examples through the lens of the topological inference problem: what structures can be distinguished using scatter plots and histograms (and persistence diagrams) alone?

3.3 Noisy circle

We begin with the circle itself, and its tautological circle-valued coordinate.

We picked 200 points distributed along the unit circle. We added a uniform random variable from $[0.0, 0.4]$ to each coordinate. A Rips complex was constructed in 0.07 seconds with maximal radius 0.5, resulting in 23475 simplices. The computation of cohomology finished in 0.03 seconds.

Parametrizing at 0.4 yielded a single coordinate function, which very closely reproduces the tautological angle function. Parametrizing at 0.14 yielded several possible cocycles. We selected one of those with low persistence; this produced a parametrization which 'snags' around a small gap in the data.

See Figure 2. The left panel in each row shows the histogram of coordinate values; the middle panel shows the correlation scatter plot against the known angle function; the right panel displays the coordinate using color. The high-persistence ('global') coordinate correlates with the angle function with topological degree 1. Variation in that coordinate is uniformly

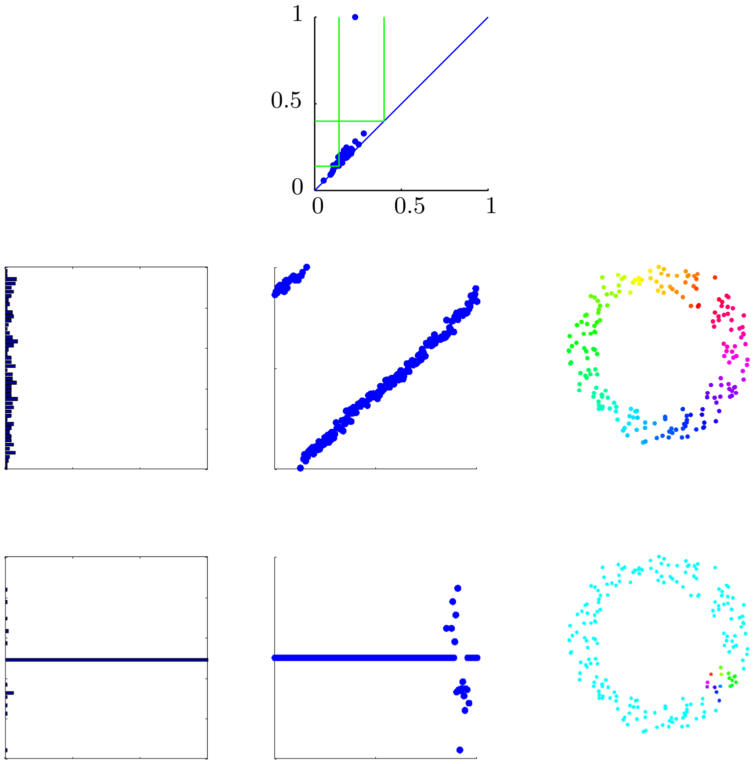


Fig. 2 Noisy circle. Persistence diagram (top). Global coordinate (middle row), local coordinate (bottom row). In the coordinate rows: histogram of coordinate values (left), correlation scatter plot against known angle function (middle), inferred coordinate in color (right).

distributed, as seen in the histogram. In contrast, the low-persistence ('local') coordinate has a spiky distribution.

3.4 Trefoil torus knot

Another example with circle topology: see Figure 3. We picked 400 points distributed along the $(2, 3)$ torus knot on a torus with radii 2.0 and 1.0. We jittered them by a uniform random variable from $[0.0, 0.2]$ added to each coordinate. We generated a Rips complex in 0.11 seconds up to radius 1.0, acquiring 36936 simplices. We computed persistent cohomology in 0.05 seconds. As expected, the inferred coordinate correlates strongly with the known parameter with topological degree 1. The histogram shows three 'bulges' corresponding to the three high-density regions of the sampled curve, which occur when the curve approaches the central axis of the torus.

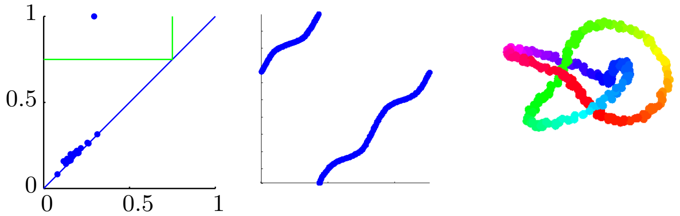


Fig. 3 Trefoil torus knot. Persistence diagram (left), correlation scatter plot of inferred coordinate against known parametrization (middle), inferred coordinate in color (right).

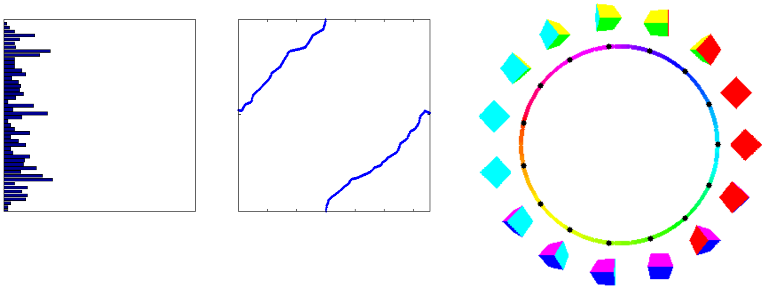


Fig. 4 Images of a rotating cube. Histogram of coordinate values (left); scatter plot against known angle function (middle); a selection of images matched to recovered circle coordinate (right).

3.5 Rotating cube

For a more elaborate data set with S^1-topology, we generated a sequence of 657 rendered images of a colorful cube rotating around one axis. Each image was regarded as a vector in the Euclidean space $\mathbb{R}^{200\cdot 200\cdot 3}$. From this data we built a witness complex with 50 landmark points and constructed a single circular coordinate. Interpolating the resulting function linearly between the landmarks gave us coordinates for all the points in the family.

See Figure 4. The frequency distribution is comparatively smooth (by which we mean that there are no large spikes in the histogram), which indicates that the coordinate does not have large static regions. The correlation plot of the inferred coordinate against the original known sequence of the cube images shows a correlation with topological degree 1. We show the progression of the animation on an evenly-spaced sample of representative points around the circle.

3.6 Pair of circles

See Figure 5 for these two examples.

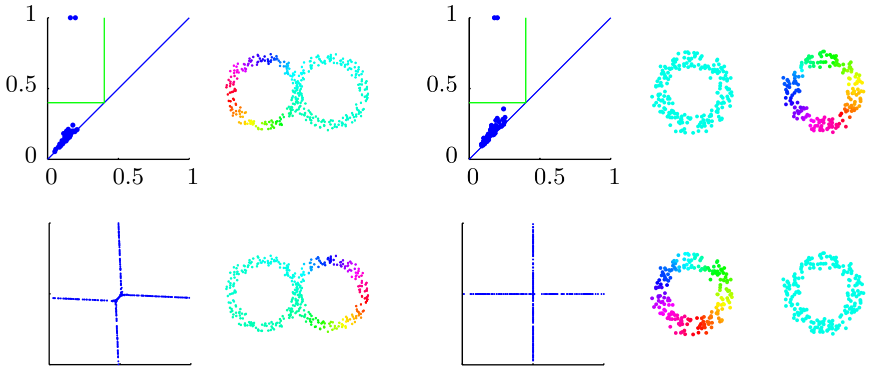


Fig. 5 Two conjoined circles (left); two disjoint circles (right). In each case we show the persistence diagram (top left), the two inferred coordinates (right column), the correlation scatter plot (bottom left).

Conjoined circles: we picked 400 points distributed along circles in the plane with radius 1 and with centres at $(\pm 1, 0)$. The points were then jittered by adding noise to each coordinate taken uniformly randomly from the interval $[0.0, 0.3]$. A Rips complex was constructed in 0.26 seconds with maximal radius 0.5, resulting in 76763 simplices. The cohomology was computed in 0.10 seconds.

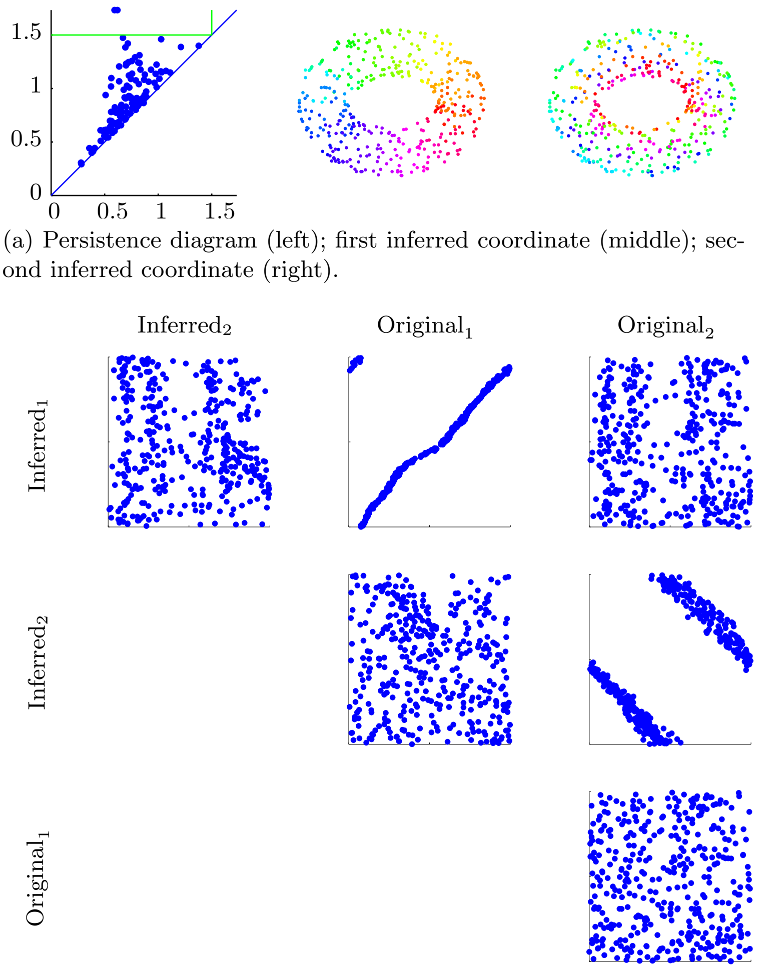
Disjoint circles: 400 points were distributed on circles of radius 1 centered around $(\pm 2, 0)$ in the plane. These points were subsequently disturbed by a uniform random variable from $[0.0, 0.5]$. We constructed a Rips complex in 0.14 seconds with maximum radius 0.5, which gave us 45809 simplices. The cohomology computation finished in 0.06 seconds.

In both cases, our method detects the two most natural circle-valued functions. The scatter plots appear very similar. In the conjoined case, there is some interference between the two circles, near their meeting point.

3.7 Torus

See Figure 6. We picked 400 points at random in the unit square, and then used a standard parametrization to map the points onto a torus with inner and outer radii 1.0 and 3.0. These were subsequently jittered by adding a uniform random variable from $[0.0, 0.2]$ to each coordinate. We constructed a Rips complex in 0.20 seconds with maximal radius $\sqrt{3}$, resulting in 61522 simplices. The corresponding cohomology was computed in 0.09 seconds.

The two inferred coordinates at the radius 1.6 in this (fairly typical) experimental run recover the original coordinates essentially perfectly: the first inferred coordinate correlates with the meridional coordinate with topological degree -1, while the second inferred coordinate correlates with the longitudinal coordinate with degree 1.



(a) Persistence diagram (left); first inferred coordinate (middle); second inferred coordinate (right).

(b) Correlation scatter plots between the two original and two inferred coordinates.

Fig. 6 Torus in $\mathbb{R}^3$.

When the original coordinates are unavailable, the important figure is the inferred-versus-inferred scatter plot. In this case the scatter plot is fairly uniformly distributed over the entire coordinate square (i.e. torus). In other words, the two coordinates are decorrelated. This is slightly truer (and more clearly apparent in the scatter plot) for the two original coordinates. Contrast these with the corresponding scatter plots for a pair of circles (conjoined or disjoint).

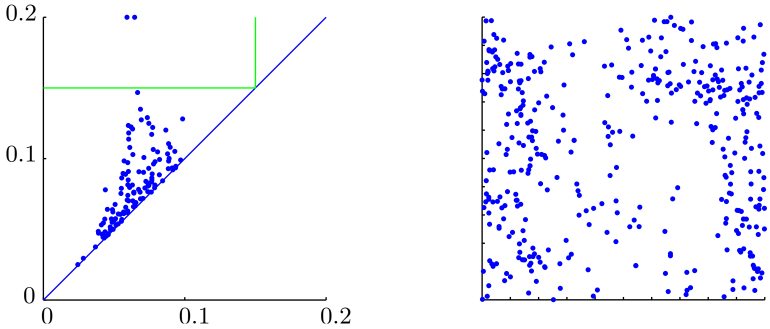


Fig. 7 Elliptic curve. Persistence diagram (left), correlation scatter plot between the two coordinates (right).

3.8 Elliptic curve

See Figure 7. For fun, we repeated the previous experiment with a torus abstractly defined as the zero set of a homogeneous cubic polynomial in three variables, interpreted as a complex projective curve. We picked 400 points at random on $S^5 \subset \mathbb{C}^3$, subject to the cubic equation

$$x^2y + y^2z + z^2x = 0.$$

To interpret these as points in $\mathbb{C}P^2$, we used the projectively invariant metric

$$d(\xi, \eta) = \cos^{-1}(|\bar{\xi} \cdot \eta|)$$

for all pairs $\xi, \eta \in S^5$. With this metric we built a Rips complex in 0.08 seconds with maximal radius 0.15. The resulting complex had 44184 simplices, and the cohomology was computed in 0.06 seconds. We found two dominant coclasses that survived beyond radius 0.15, and we computed our parametrizations at the 0.15 mark.

The resulting scatter plot quite clearly exhibits the decorrelation which is characteristic of the torus.

3.9 Double torus

See Figure 9. We constructed a torus by generating 1600 points, uniformly distributed in the unit square, and then using a standard parametrization of the torus to wrap the points onto a torus surface with inner and outer radii 1.0 and 3.0. This was done twice, translating the two tori to place centers 5.7 apart from each other. The points, from each torus, that overrun the intersection plane were dropped, resulting in a data set with 2885 points distributed on a double torus. We build a Rips complex on these points in 12.97 seconds up to radius 1.25 which yields 1,879,805 simplices. The persistent cohomology computation took 8.46 seconds, and identified the four most significant cocycles. The resulting persistence diagram is in Figure 8.

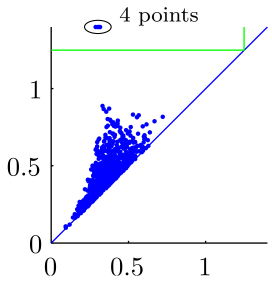


Fig. 8 Double torus: persistence diagram.

The identified cocycles and the resulting parametrizations are not especially perspicuous; we present them in Figure 9(a). On the other hand, by taking linear combinations we can find a new basis of circular coordinate functions whose correlation scatter-plot matrix is much more suggestive of the double torus: see Figure 9(b).

This particular coordinate transformation was obtained ‘by inspection’. Open question: is there a systematic way to transform a basis of circular coordinate functions so that the structure of the data is revealed as helpfully as possible?

After the update, coordinates 1 and 2 are ‘coupled’, in the sense that they are supported over the same subtorus of the double torus. The scatter plot shows that the two coordinates appear to be completely decorrelated except for a large mass concentrated at a single point. This mass corresponds to the other subtorus, on which coordinates 1 and 2 are essentially constant. A similar discussion holds for coordinates 3 and 4.

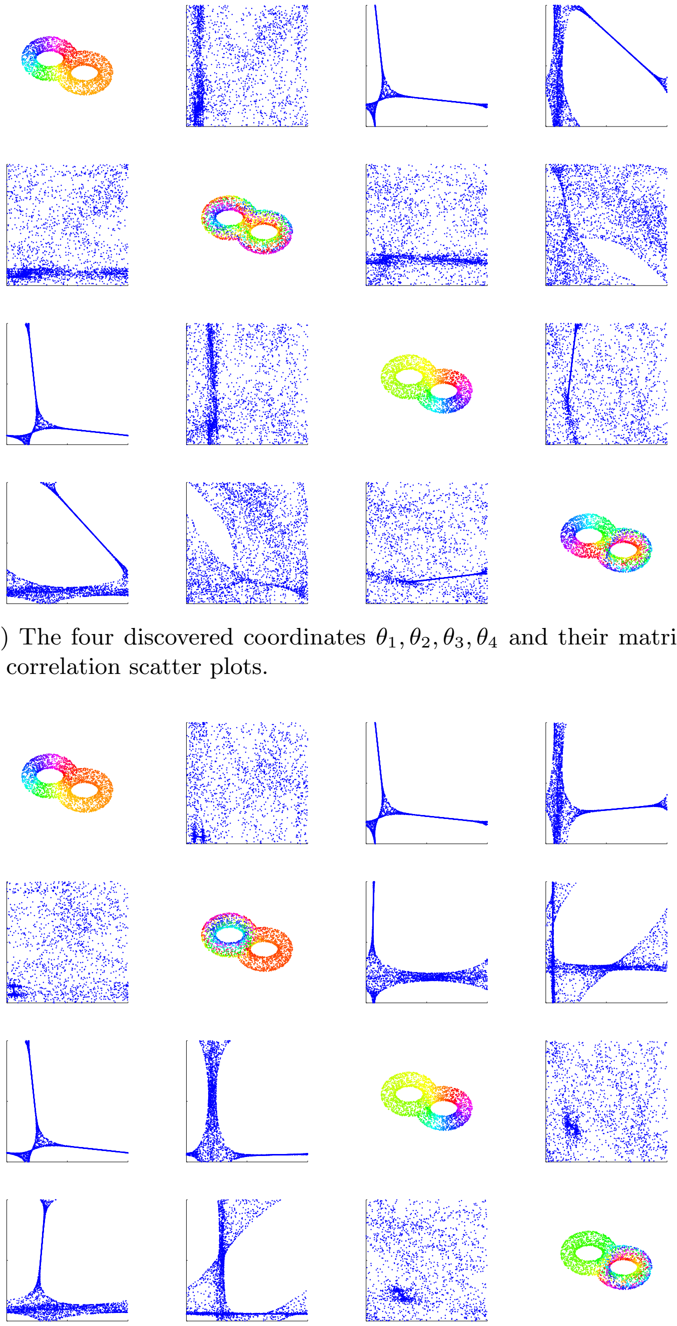
The uncoupled coordinate pairs (1,3), (1,4), (2,3), (2,4) produce scatter plots reminiscent of two conjoined circles.

4 Discussion

Although our procedure works well in these simple examples, there are various unanswered questions about the behaviour of this algorithm in general. We discuss these now.

Diophantine algebra

- When lifting from $\mathbb{F}_p$ coefficients to $\mathbb{Z}$ coefficients, why does the ‘close to zero’ heuristic work perfectly in the given examples? In fact, coefficients of cocycles produced by the persistence algorithm appear to be almost always $0, \pm 1$. What makes this happen?
- Are there efficient ways to repair an integer lift α of an $\mathbb{F}_p$-cocycle α_p, when $d_1\alpha \neq 0$? What about under special conditions, such as $d_1\alpha$ being sparse?
- Are there *a priori* geometric estimates on the largest torsion prime in $\mathrm{H}^2(X; \mathbb{Z})$? In other words, can one quantify the assertion “p-torsion is rare”?



(a) The four discovered coordinates $\theta_1, \theta_2, \theta_3, \theta_4$ and their matrix of correlation scatter plots.

(b) Taking linear combinations for a geometrically more ‘natural’ basis of circular coordinates: $\phi_1 = \theta_1$, $\phi_2 = \theta_2 + \theta_3 + \theta_4$, $\phi_3 = \theta_3$, and $\phi_4 = \theta_1 + \theta_4$. The pairs ϕ_1, ϕ_2 and ϕ_3, ϕ_4 respectively parametrize the left and right halves of the double torus.

Fig. 9 Double torus in $\mathbb{R}^3$.

– The cohomology group $\mathrm{H}^1(X;\mathbb{Z})$ is torsion-free, and hence isomorphic to some $\mathbb{Z}^n$. Are there efficient ways to compute an independent set of generators?

Generalized multidimensional scaling (mds)

– The real coordinates in classical mds have an absolute scale, which can be related to the metric structure on the input data. Circular coordinates, on the other hand, have no absolute scale. Is there a meaningful way to assign radius values to each circular coordinate, for instance to estimate the longitudinal and meridional radii of a general torus?
– The methods presented in this paper will recover topologically independent circle coordinates (since the generators of the persistence diagram are by definition linearly independent elements of H^1). Classical mds, similarly, recovers statistically independent real coordinates. Is there some way to combine the two approaches to obtain mixed families of real and circular coordinates? What is the appropriate notion of independence?

Higher dimensions

– Can we apply similar methods to obtain sphere-valued coordinates, for spheres S^n with $n \geq 2$? The simplest analogue of (1) in 2 dimensions is

$$[X, \mathbb{CP}^\infty] \cong \mathrm{H}^2(X;\mathbb{Z})$$

where $\mathbb{CP}^\infty$ can be thought of as S^2 with a sequence of attached disks $D^4, D^6, D^8, \ldots$ in even dimensions. One can therefore define S^2-valued maps up to the 3-skeleton of X, which are homotopy-unique up to the 2-skeleton. Is there a tractable smoothing procedure analogous to the harmonic smoothing used here for S^1-maps?

Our hope is that the methods presented here are simply the first steps in a larger, more ambitious theory of *topological* multidimensional scaling and structure discovery.

Acknowledgements Thanks to Jennifer Kloke for sharing her analysis of a visual image data set; this example did not make the present version of this paper. We also thank Gunnar Carlsson, for his support and encouragement as leader of the topological data analysis research group at Stanford; and Robert Ghrist, as leader of the DARPA-funded project Sensor Topology and Minimal Planning (SToMP).

References

1. M. Belkin and P. Niyogi. Laplacian Eigenmaps and spectral techniques for embedding and clustering. In T. Diettrich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 585–591. MIT Press, Cambridge, Massachussetts, 2002.

2. T. F. Cox and M. A. A. Cox. *Multidimensional Scaling.* Chapman & Hall, London, 1994.
3. V. de Silva and G. Carlsson. Topological estimation using witness complexes. In M. Alexa and S. Rusinkiewicz, editors, *Eurographics Symposium on Point-Based Graphics*, ETH, Zürich, Switzerland, 2004.
4. M. Dixon, N. Jacobs, and R. Pless. Finding minimal parameterizations of cylindrical image manifolds. In *CVPRW '06: Proc. 2006 Conference on Computer Vision and Pattern Recognition Workshop*, page 192, Washington, DC, USA, 2006. IEEE Computer Society.
5. D. L. Donoho and C. Grimes. Hessian Eigenmaps: New locally linear embedding techniques for high-dimensional data. Technical Report TR 2003-08, Department of Statistics, Stanford University, 2003.
6. H. Edelsbrunner and J. Harer. Persistent homology — a survey. In J. E. Goodman, J. Pach and R. Pollack, editors, *Surveys on Discrete and Computational Geometry: Twenty Years Later*, pages 257–282. *Contemporary Mathematics*, 453. American Mathematical Society, Rhode Island, 2008.
7. H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete and Computational Geometry*, 28:511–533, 2002.
8. R. Ghrist. Barcodes: the persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008.
9. L. J. Guibas and S. Y. Oudot. Reconstruction using witness complexes. In *Proc. 18th ACM-SIAM Symposium on Discrete Algorithms*, pages 1076–1085, 2007.
10. A. Hatcher. *Algebraic Topology.* Cambridge University Press, Cambridge, 2002.
11. J. A. Lee and M. Verleysen. Nonlinear dimensionality reduction of data manifolds with essential loops. *Neurocomputing*, 67:29–53, 2005.
12. D. Morozov. Dionysus library for computing persistent homology. http://www.mrzv.org/software/dionysus/.
13. C. C. Paige and M. A. Saunders. LSQR: Sparse equations and least squares. http://www.stanford.edu/group/SOL/software/lsqr.html.
14. C. C. Paige and M. A. Saunders. LSQR: An algorithm for sparse linear equations and sparse least squares. *ACM Transactions on Mathematical Software*, 8(1):43–71, March 1982.
15. R. Pless and I. Simon. Embedding images in non-flat spaces. In *Conference on Imaging Science Systems and Technology*, pages 182–188, 2002.
16. S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, Dec. 2000.
17. H. Sexton and M. Vejdemo-Johansson. jPlex simplicial complex library. http://comptop.stanford.edu/programs/jplex/.
18. J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, Dec. 2000.
19. A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete and Computational Geometry*, 33(2):249–274, 2005.