Time-optimal persistent homology representatives for univariate time series

Antonio Leitao, Nina Otter

Time-optimal persistent homology representatives for univariate time series

António Leitão[1, 2] and Nina Otter[1,3]

[1] DataShape, Inria-Saclay, France
[2]Scuola Normale Superiore di Pisa, Italy
[3] Laboratoire de Mathématiques d'Orsay, Université Paris-Saclay, Orsay, France

Abstract

Persistent homology (PH) is one of the main methods used in Topological Data Analysis. An active area of research in the field is the study of appropriate notions of PH representatives, which allow to interpret the meaning of the information provided by PH, making it an important problem in the application of PH, and in the study of its interpretability. Computing optimal PH representatives is a problem that is known to be NP-hard, and one is therefore interested in developing context-specific optimality notions that are computable in practice. Here we introduce time-optimal PH representatives for time-varying data, allowing one to extract representatives that are close in time in an appropriate sense. We illustrate our methods on quasi-periodic synthetic time series, as well as time series arising from climate models, and we show that our methods provide optimal PH representatives that are better suited for these types of problems than existing optimality notions, such as length-optimal PH representatives.

1 Introduction

Topological Data Analysis (TDA) is a field that uses insights from topology — the mathematical area that studies abstract shapes — to develop representations of data that are computable and robust in an appropriate sense [5]. Persistent homology (PH) is, arguably, one of the most successful methods used in Topological Data Analysis, and it is being increasingly applied to a variety of data analysis problems. We refer the reader to the DONUT database [14] for a vast collection or real-world applications of PH. In persistent homology, one takes as a point of departure a data set, such as a point cloud, a time series, a network, or a digital image, and associates to it a 1-parameter family of topological spaces, in which for each parameter value r one may think of the corresponding space as being an approximation, truncation or thickening of the original data set; for instance, if the parameter captures distance between points, the space at parameter value r identifies all points at distance smaller or equal than r. The output of persistent homology is then a summary, called "persistence barcode" or "persistence diagram", of the number of topological features such as connected components, holes or tunnels, voids, present in a data set, as well as how long each feature spans ("persists") across the possible parameter values.

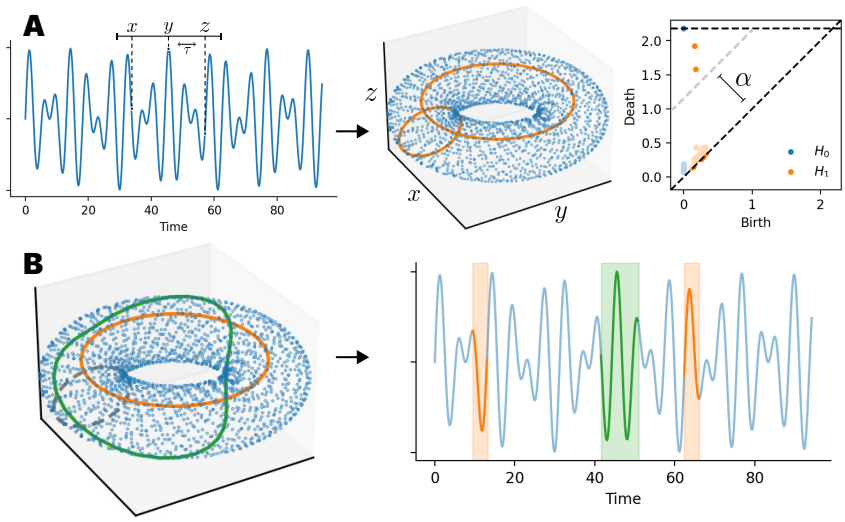


Figure 1: **Pipeline for computing time-optimal PH cycle representatives for univariate time series**. **A)** One starts by embedding a univariate time series into Euclidean space. For quasi-periodic times series, one can obtain an embedding together with a lower bound (α) on persistence that allows to distinguish between components, holes, voids, that one may consider as "significant", and features that may be due to noise. **B)** Given the significant PH features, we attempt to find representatives that are optimal with respect to their cohesion in time. Given representatives of the same PH feature (e.g., green and orange cycles on the left), the ones that correspond to a continuous trajectory (green) in the original signal are preferred over the ones that are discontinuous (orange).

Often, in application, one is interested in interpreting the meaning of such summaries, by determining which data points correspond to each individual summary. More precisely, given a persistence barcode, one asks for a choice of persistent vector basis, called "PH representatives", that is meaningful for the application at hand. Ideally, one would want to compute representatives that are optimal in an appropriate sense. This problem has been shown to be NP hard [7], and a considerable amount of work is being devoted to developing algorithms to find approximations of representatives that satisfy some minimality condition of interest in a specific application context.

In the present work we introduce algorithms to compute PH representatives for time-dependent data. To be best of our knowledge, this is the first time that such representatives have been studied. In particular, we propose two different notions: (i) vertex time-optimal PH representatives and (ii) simplex time-optimal PH representatives. We illustrate these notions on synthetic quasi-periodic time series and on time series resulting from delayed oscillator models of the El Ninõ Southern Oscillation (ENSO). We illustrate the pipeline for the computation of time-optimal representatives for univariate time series in Figure 1.

2 Related work

To the best of our knowledge, our work is the first attempt at studying PH representatives for time-varying data that explicitly take the time information into account. Our work inscribes itself in a line of research that tries to develop notions of optimal PH representatives suitable for applications, such as the length-optimal or volume-optimal PH representatives [11, 25]. A benchmarking of several existing approaches has been performed in [22]. A speed-up exploiting persistent cohomology computations has been proposed in [33].

The development of methods to study time-varying data is a very active area of research in TDA. Vineyards provide 1-parameter families of persistence barcodes for time-varying data [10]; generalisations of these to multi-parameter families have recently been proposed [18,19]. An algebraic framework for the study of time-varying persistence modules has been introduced in [31]. Univariate time series have been studied in TDA, among others, in [3, 27, 28, 30]. Several topological methods have been developed to study bifurcation diagrams of dynamical systems, including [16].

3 Background

We first give a brief overview of basic notions from persistent homology in Section 3.1; we then discuss univariate time series in Section 3.2, and we give an overview on existing optimization approaches for PH cycle representatives in Section 3.3.

3.1 Homology and persistent homology representatives

To simplify exposition, here we introduce homology and persistent homology for coefficients over the field with two elements $\mathbb{F}_2$. We refer the reader to Appendix A for a discussion of what changes for arbitrary coefficient fields.

3.1.1 Homology

Let K be a simplicial complex, and denote by $S_p(K)$ its set of p-simplices. We denote by $C_p(K)$ the vector space generated by the $\mathbb{F}_2$-linear combinations of p-simplices.

The elements of $C_p(K)$ are called ***p*-chains**. We consider the boundary operator

$$\begin{aligned} \partial_p : C_p(K) &\longrightarrow C_{p-1}(K) \\ \sigma &\mapsto \sum_{\tau \subset \sigma, \tau \in S_{p-1}(K)} \tau \end{aligned}$$

and call the elements of the kernel of ∂_p ***p*-cycles**, and the elements of the image of ∂_{p+1} ***p*-boundaries**. One can show that $\partial_{p+1} \circ \partial_p = 0$ for all p. Intuitively, this is due to the fact that the boundary of a boundary is empty. The quotient vector space $H_p(K) = \frac{\ker(\partial_p)}{\mathrm{im}(\partial_{p+1})}$ is called ***p*th simplicial homology** of K (with coefficients in $\mathbb{F}_2$).

Definition 3.1. For a given simplicial homology vector space $H_p(K) = \frac{\ker(\partial_p)}{\mathrm{im}(\partial_{p+1})}$ and $h \in H_p(K)$, we call $c \in \ker(\partial_p)$ a **cycle representative** of h if $[c] := c + \mathrm{im}(\partial_{p+1} = h$. Two cycles $c, c' \in \ker(\partial_p)$ are **homologous** if $[c] = [c']$. A **cycle basis** for $H_p(K)$ is a set of p-cycles $c_1, \ldots, c_m$ such that $[c_i] \neq [c_j]$ for $i \neq j$ and $[c_1], \ldots, [c_m]$ are a basis for $H_p(K)$.

3.1.2 Persistent homology

We now consider a finite sequence of nested simplicial complexes:

$$\varnothing \subseteq K_{t_0} \subseteq K_{t_1} \subseteq \cdots \subseteq K_{t_{n-1}} \subseteq K_{t_n} =: K\,.$$

We call $\{K_{t_i}\}_{i=0}^{n}$ a **filtration** of K. We obtain injective linear maps $\iota_{i,j}\colon C_p(K_{t_i}) \hookrightarrow C_p(K_{t_j})$ for all $0 \leq i < j \leq n$, induced by the inclusions of simplicial complexes. We can thus identify each $C_p(K_{t_i})$ with a vector subspace of $C_p(K_{t_j})$ for any $i < j$.

Definition 3.2. Given $c \in C_p(K)$, we define its **birth** to be

$$\mathrm{birth}(c) = \min\{i \mid c \in C_p(K_i)\}$$

and its **death** to be

$$\mathrm{death}(c) = \min\left\{i \mid c \in \mathrm{im}\left(\partial_{p+1}\big|_{(C_{p+1}(K_i))}\right)\right\}$$

where we use the convention $\min \emptyset = \infty$.

Similarly as for chain vector spaces, the inclusions of simplicial complexes induce (not necessarily injective) linear maps between homology vector spaces $\phi_{i,j}\colon H_p(K_{t_i}) \to H_p(K_{t_j})$ for all $0 \leq i \leq j \leq n$. One gives the following definition:

Definition 3.3. The ***p*th persistent homology** $H_p(K)$ of a filtered simplicial complex $\{K_{t_i}\}_{i=0}^{n}$ with $K_{t_n} = K$ is the tuple $(\{H_p(K_{t_i})\}_{i=0}^{n}, \{\phi_{i,j}\}_{i<j})$.

Definition 3.4. A **persistent homology cycle basis** of $H_p(K)$ is a set of p-cycles $c_1, \ldots, c_m \in C_p(K)$ such that $\mathrm{birth}(c_j) \neq \mathrm{death}(c_j)$ for all $j = 1, \ldots, m$ and for each filtration value t we have that the collection of cycles c_j with $\mathrm{birth}(c_j) \leq t \leq \mathrm{death}(c_j)$ form a cycle basis for $H_p(K_t)$. We say that a p-cycle $c \in C_p(K)$ is a **persistent homology cycle representative** for $H_p(K)$ if it is an element of a persistent homology cycle basis of $H_p(K)$.

Definition 3.5. Let $c_1, \ldots, c_m \in C_p(K)$ be a persistent homology cycle basis of $H_p(K)$. We call the multiset

$$PD_p(K) := \left\{\big(\mathrm{birth}(c_j), \mathrm{death}(c_j)\big) \mid j = 1, \ldots, m\right\}$$

the **persistence diagram of $H_p(K)$, or the *p*th persistence diagram of K.** We call the number $\mathrm{death}(c_j) - \mathrm{birth}(c_j)$ the **persistence** of c_j.

It is a fundamental result in persistent homology that a persistent homology cycle basis exists for any persistent homology tuple satisfying appropriate finiteness conditions, and that the persistence diagram does not depend on the choice of PH cycle basis [32].

Example 3.6. Consider the filtration of simplicial complexes $K_1 \subset K_2 := K$ illustrated in Figure 2(ii). We can think of this filtration as the triangulation of a bent cylinder, with its two sides on the bottom, which we then scan from bottom to top, see Figure 2(i). We can compute the 1st simplicial homology of K_2, and we obtain that $H_1(K_2) \cong \mathbb{F}_2$. Some possible choices of 1-cycle representatives include for instance $< ab > + < bc > + < ac >$, $< a'b' > + < b'c' > + < a'b' >$ or $< ab' > + < b'b > + < bc > + < ca >$, where we denote by $< x_0x_1 >$ the vector corresponding to the 1-simplex with vertices x_0, x_1. We depict these three different choices in Figure 2(iii). In particular, we note that cycle representatives are in general not unique.

Example 3.7. We now use the same filtered complex from Example 3.6 to illustrates PH cycle representatives, as well as some subtleties in their interpretation. The persistence diagram of $H_1(K)$ contains two points not on the diagonal: the point $(1, 2)$ and the point $(1, \infty)$. A possible choice of PH cycle basis of $H_1(K)$ is given by $< a'b' > + < b'c' > + < a'c' >$ and $< a'b' > + < b'c' > + < a'c' > + < ab > + < bc > + < ac >$. We illustrate the simplices corresponding to the two cycle representatives in Figure 2 in orange and in cyan, respectively. We note that the second PH cycle representative is an example of representative with disjoint "support"; namely, if we consider the subcomplex corresponding to the 1-simplices with non-zero coefficients, then it is disconnected. This defies the usual interpretation of PH 1-cycle representatives as correponding to single "tunnels" or "loops" in the data. Interestingly, such PH cycle representatives do not seem to appear often in practice, see [22, Section 6.6.1].

3.2 Univariate time series

Definition 3.8. A **univariate time series** is a function $f \colon \mathbb{R} \to \mathbb{R}$.

There are mainly two approaches for associating filtrations of simplicial complexes to time series in TDA: (i) one computes sublevel-set filtrations of the time series or a transform thereof (such as a Discrete Fourier transform); or (ii) one embeds the time series in Euclidean space and associates filtrations of simplicial complexes to the resulting point cloud, such as, e.g., Vietoris-Rips complexes. See the review [29] for details about these two approaches. Here we follow the second approach, and we provide details about the computation of the embedding parameters in Appendix C.

We note that while in the current work we focus on univariate time series, our methods can be applied to a broader class of time-varying data, see the discussion in Section 6.

3.3 Optimizing (persistent) homology cycle representatives

In practice, for applications, we are interested in finding cycle representatives that are informative for the problem at hand. In particular, one often seeks to find cycles that minimise some criteria (e.g., the number of simplices contained in the cycle). Roughly, the existing approaches to computing optimal PH p-cycle representatives can be divided into those that minimise a loss function defined on p-chains (also called "edge"-loss methods, in analogy with the $p = 1$ case) and those that instead minimize a loss function defined on $p+1$-chains (also called "triangle"-loss methods). We do not consider triangle-loss methods in the current work, and we point the reader to [25] for details. In the follow we review the basic set up of edge-loss methods.

3.3.1 Edge-loss methods

Given an initial cycle $c_0 \in C_p(K)$ the problem for homology cycle representatives focuses on finding a homologous cycle that minimises a given loss function $\ell \colon C_p(K) \to \mathbb{R}$. Since adding any boundary $w \in \mathrm{im}(\partial_{p+1})$ to c_0 results in a homologous cycle, we have the following problem formulation:

$$\begin{aligned} \min \quad & \ell(c) \\ \text{subject to} \quad & c = c_0 + \partial_{p+1}(w) \\ & w \in C_{p+1}(K) \end{aligned}$$

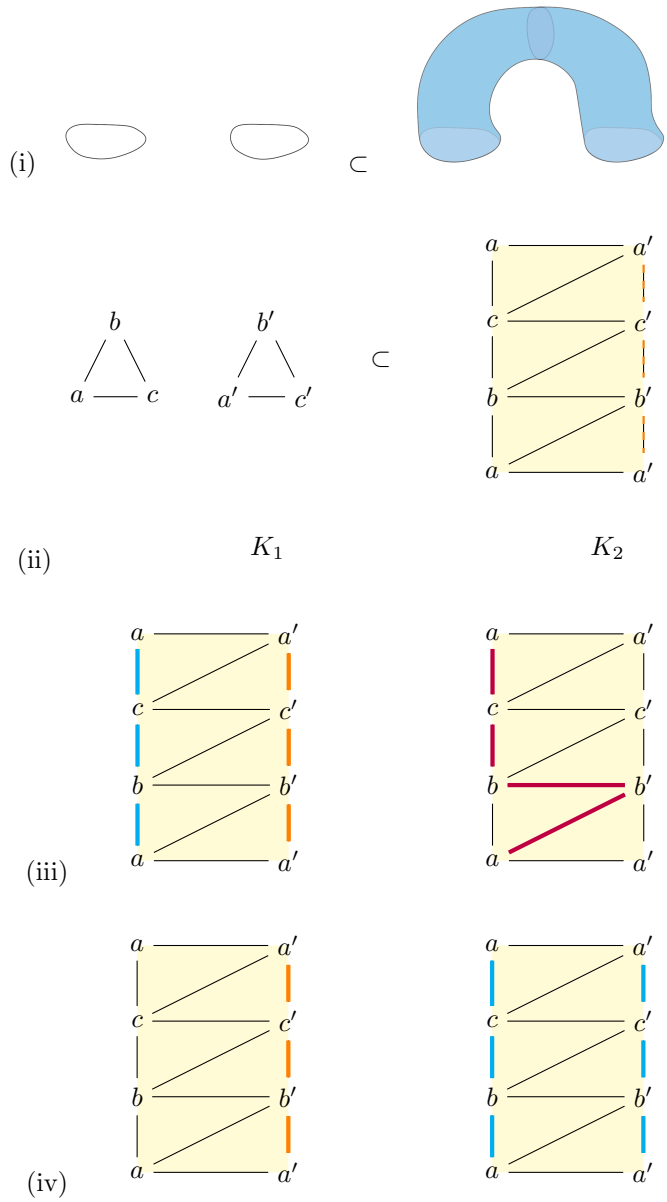


Figure 2: (i) A nested sequence of subspaces of $\mathbb{R}^3$. (ii) A filtered simplicial complex that is obtained as a triangulation of the spaces in (i). (iii) Three Possible choices of 1-cycle representatives for $H_1(K_2)$. (iv) PH 1-cycle representative basis for $H_1(K)$: we depict the two PH cycle representatives in orange and cyan. This is an example of a filtered simplicial complex for which one of the PH 1-cycle representatives consists of a linear combination of two disjoint closed curves.

In practice, obtaining the initial cycle is not a problem since most programs output a cycle basis.

We can modify the previous problem to find a cycle that has the same persistence

as the homology class it represents. Given a point (b, d) in $PD_p(K)$ we look for the solution to the following problem:

$$\begin{aligned} \min \quad & \ell(c) \\ \text{subject to} \quad & \text{birth}(c) = b \\ & \text{death}(c) = d \\ & c \in \ker\left(\partial_p\big|_{C_p(K_b)}\right) \end{aligned}$$

This problem was first studied in [8].

4 Time-optimal representatives

We start by discussing an example that will guide us in finding a suitable notion of time-optimal PH cycle representative. Consider the simplicial complex associated to the point cloud illustrated in Figure 3(a), which we can think of as being obtained by a time-delay embedding of a noisy sine curve.

Among the different possible choices of 1-cycle representatives, we give four examples in Figure 3(b), highlighted in green. Which of these choices should we consider as optimal with respect to time? For the problem we are interested in studying in this paper, namely, studying univariate time series through time-delay embeddings, we are interested in obtaining cycle representatives that correspond to time-series values that are not too far from each other in time, since this allows us to interpret the topological feature (i.e., a non-trivial PH class), through a selection of data points in the original time series that correspond to that feature. For instance, for the example application that we study in this paper — delayed oscillator models of the El Ninõ Southern Oscillation —, this allows one to ask the question of whether the topological features that one recovers have any physical meaning at all.

In particular, for the noisy sine curve from Figure 3(a), this means that we would want to choose representative 1-cycles whose vertices correspond to time-series values that are contained in an interval of length 2π, the period of the sine curve. Thus, we make the following observations:

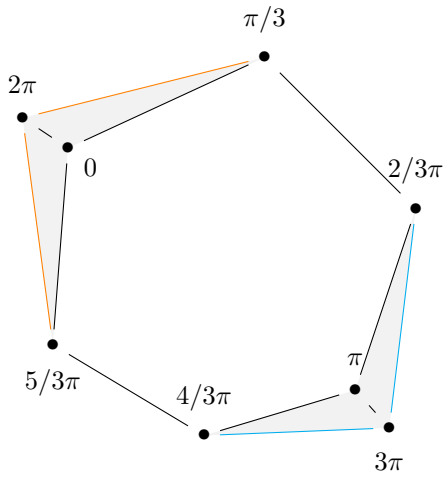
- Any reasonable notion of time-optimal cycle representative should not choose a 1-cycle that includes the cyan edges from Figure 3(a).
- On the other hand, it is likely that we will have to make a choice between the edges incident to vertices labelled by 0 and $\pi/3$, $5/3\pi$ and those incident to vertices labelled by 2π and $\pi/3$, $5/3\pi$.

Thus, based on these observations, we would wish to consider the two cycles on the left of Figure 3b as time-optimal, but not the two ones on the right.

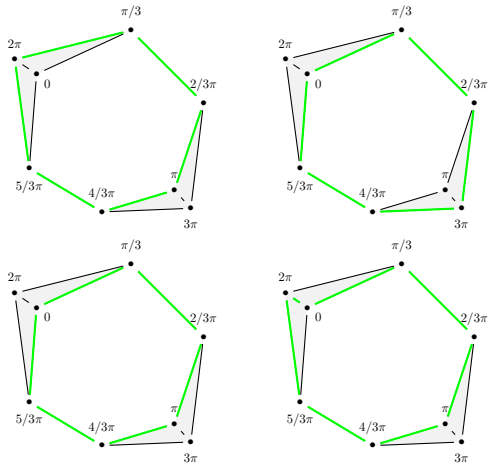
We can thus formulate our problem as follows. We let $f\colon \mathbb{R}\times\mathbb{R} \to \mathbb{R}_{\geq 0}$ be a norm. To each edge $e = \{s, s'\}$ we assign the weight $w(e) := f(T(s), T(s'))$, where for a point s in the point cloud, we denote by $T(s)$ its time label. Then for each possible 1-cycle representative, we compute its cost by considering how far apart its edges are in time. More precisely, to compute the cost of each edge, we compute the difference between its weight and the weights of the edges adjacent to it, and we retain the maximum of these differences as the cost of the edge. The cost of the 1-cycle is then the sum over the costs of its edges. A time-optimal cycle is one that minimises this cost function.

We can thus write our optimisation problem as follows:

$$c_{\text{opt}} = \underset{c \text{ a } 1-cycle}{\operatorname{argmin}} \left\{ \sum_{e \in c} \max_{\substack{e' \in c \\ \text{adjacent to } e}} \left| w(e) - w(e') \right| \right\}.$$



(a) A simplicial complex with time-labelled vertices, associated to a point cloud arising from a noisy sine curve.



(b) Different representative 1-cycles (in green).

Figure 3

While in our example we have considered 1-cycles, one can in a similar way define an optimization problem for higher-dimensional cycles. In what follows our discussion will not be restricted to any particular cycle dimension.

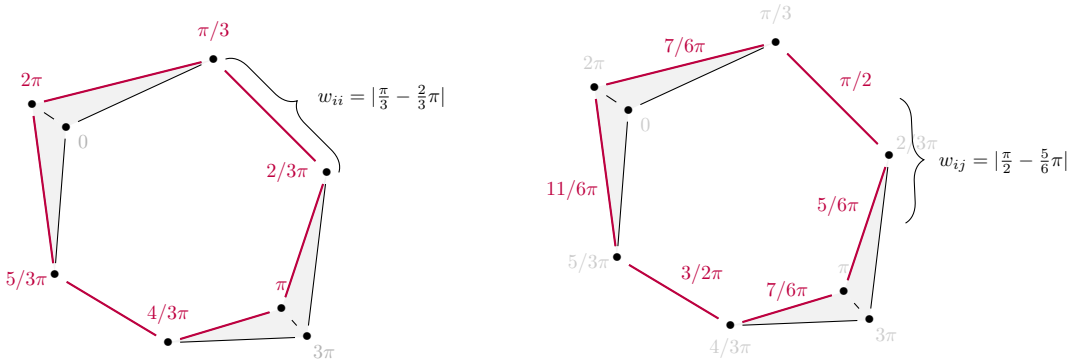
4.1 Loss function

We now look at defining an appropriate loss function for the minimization problem. Let $c \in C_p(K)$ be a chain. We wish to build a non-negative matrix W that appropriately weighs the chain c such that the optimal solution is a chain with minimal dispersion in time. We first make the notion of "time-closeness" or "minimal time dispersion" precise:

Definition 4.1. Given a p-chain, its **time dispersion** is the difference between the minimum and maximum time labels of any of the vertices contained in the p-simplices in the chain.

We note that while for other types of applications, e.g., temporal networks (see also discussion in Section 6), one might be interested in considering another notions of time dispersion, the notion of time dispersion we consider here is motivated by the study of time series, for which we want to obtain data points in the time series that are not too far apart, or dispersed.

We next choose an order on the simplices in the chain c, and label the simplices $\sigma_1, \ldots, \sigma_n$ according to this order. Each entry w_{ii} of the matrix describes the cost of



(a) Vertex-based optimal cycle and weighting

(b) Simplex-based optimal cycle and weighting

Figure 4: Illustrations of the difference between considering simplex weighting and vertex weighting for the example from Figure 3b. In both situations the highlighted cycle represents the optimal solution. **a)** In vertex-based weighting, the cost w_{ii} of selecting 1-simplex σ_i is given by the difference of the time labels of its vertices. **b)** For simplex-based weighting, we first assign a time label to each simplex, as the mean of the time label of its vertices. The cost of a chain is then the sum of the differences w_{ij} of the time labels of adjacent simplices σ_i, σ_j.

selecting the p-simplex σ_i in the chain. Similarly, the entry w_{ij} represents the cost of having both σ_i and σ_j as part of the chain. Our goal is then to minimize the norm $\|Wc\|$. In the following we consider two different ways of measuring time-optimality:

1. "vertex-based": We find chains with vertices with time labels that are close to each other.
2. "simplex-based": We first situate each simplex in time, and find a chain of closely situated simplices.

4.1.1 Vertex-based time optimality

In this scenario the cost of selecting a p-simplex depends only on the time labels of its vertices. Consider a 1-simplex $\sigma = \{v_i, v_j\}$. If we are attempting to find a 1-cycle that minimizes the dispersion in time the intuitive cost of selecting simplex σ would be the difference in the time label of the vertices $|T(v_i) - T(v_j)|$. The optimal solution is a 1-cycle composed of 1-simplices with vertices that are as close in time as possible.

Given a p-simplex $\sigma_i = \{v_{i_0}, v_{i_1}, \ldots, v_{i_{p+1}}\}$ we let

$$T_i^{\max} = \max\left\{T(v_{i_0}), T(v_{i_1}), \ldots, T(v_{i_{p+1}})\right\}$$

be the maximum time label of the vertices of σ_i. Similarly, we let $T_i^{\min}$ be the minimum time label of the vertices of σ_i. Then we define the weight matrix W as the diagonal matrix with diagonal entries given by:

$$w_{ii} = T_i^{\max} - T_i^{\min}.$$

We thus have that the cost of selecting a chain containing σ_i is given by the maximum difference of the time labels of its vertices.

4.1.2 Simplex-based time optimality

In a simplex-based approach we first assign a time label $T(\sigma)$ to each $p-$simplex. This can be seen as positioning the simplex σ in time. The solution of the optimization problem is a cycle composed of simplices where adjacent simplices are closely placed in time.

Definition 4.2. Two p-simplices σ_i, σ_j are called **adjacent** if their intersection is non-empty and is of cardinality p. That is, if there exists a $p-1$-simplex τ such that $\sigma_i \cap \sigma_j = \tau$.

Thus, we consider the weight matrix with entries defined as follows:

$$w_{ij} = \begin{cases} |T(\sigma_i) - T(\sigma_j)| & \text{if } \sigma_i, \sigma_j \text{ are adjacent} \\ 0 & \text{otherwise} \end{cases}$$

In our experiments we consider the time label of a p-simplex $\sigma = \{v_{i_0}, v_{i_1}, \ldots, v_{i_{p+1}}\}$ as the mean time label of its vertices:

$$T(\sigma) = \frac{1}{(p+1)} \sum_{k=0}^{p+1} T(v_{i_k}) \,.$$

In Figure 4 we illustrate the two notions of vertex- and simplex-optimal 1-cycle representatives obtained for the example in Figure 3b. For practical applications, we need to further modify our optimization problem, as we explain next in Section 4.2.

4.2 Approximate PH cycle representatives

When searching for a representative cycle of a PH class one forces the solution to have the same birth and death values (b, d) as the class it represents. In other words, a solution must contain the birth simplex and cannot contain any simplex that appears after the birth value. As we find in addressing our problem, this constraint limits the possible solution set of the optimization problem in a way that is too restrictive to give any meaningful cycles.

We thus relax this constraint and allow the representative cycle to not necessarily have the same persistence of the class it represents. More precisely, given a minimum persistence value ε we instead search for representatives with a birth time of $(d - \varepsilon)$, that is, with a persistence of ε.

One could then ask what a good choice for such a minimum persistence value ε might be. In our set up we follow the computational methodology from [13], which provides lower bounds on persistence that distinguish between topological noise (points in the persistence diagram with persistence smaller than the bound) and topological signal (points in the persistence diagram with persistence greater or equal than the bound). Thus, we take use this bound for several of the examples considered (in the noisy sine example, see Figure 6 and the double sine, see Figure 7). For computational reasons, for the time series arising from the ENSO models we take instead 90% of the persistence of the PH class (which thus gives a bound smaller than the significance bound from [13]), namely, we take $\varepsilon = 0.9(d - b)$ (see Figures 10,11 and 12).

We note that, as we have already observed in Section 3.1.2, given filtration values $r \leq r'$, we have that $K_r \subset K_{r'}$ and in particular, we can identify $C_p(K_r)$ with a

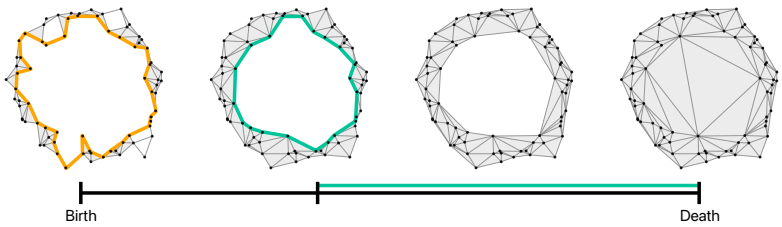


Figure 5: Example of representative 1-cycles with minimal l_1 norm for different persistence values (see Appendix B.2 for details). Left/orange: we depict a 1-cycle with full persistence. Second from left/green: we depict a 1-cycle homologous to the previous one, but with smaller persistence, and with smaller l_1 norm. We thus note that relaxing the persistence of the representative allows for a solution with smaller l_1 norm.

vector subspace of $C_p(K_{r'})$. Therefore, we have that any chain that exists in $C_p(K_r)$ also exists in $C_p(K_{r'})$. More specifically, given any optimization function f we have that:

$$r \leq r' \implies \min\{f(c) \mid c \in C_p(K_r)\} \geq \min\{f(c) \mid c \in C_p(K_{r'})\}\,.$$

In practice, if we allow the representative cycle to have a smaller persistence value than that of its class, we are guaranteed to obtain a cycle with a smaller loss value. This can very beneficial, particularly in the following situations:

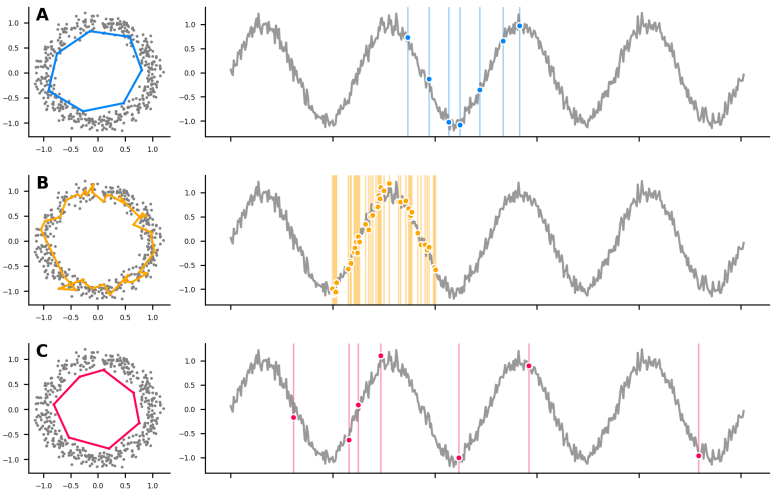


Figure 6: We illustrate different 1-cycle PH representatives for the optimal embedding of $f(t) = \sin t + \epsilon$, where ϵ is a random variable with normal distribution, in the point cloud (left), as well as in the original time series (right). **A**: Simplex-based time optimal cycle. **B**: Vertex-based time optimal cycle. **C**: Length-optimal cycle. We note that both simplex and vertex-based optimal cycles have time dispersion that is close to the period of the sine, and thus close to the desired minimum. On the other hand, the length-optimal cycle is much further spread out in time, and thus doesn't provide a meaningful cycle representative.

1. One has a lower bound on persistence, allowing to distinguish between topological "signal" and "noise". This can be given for instance, as in our case, by the computational methodology from [13] to compute embeddings of quasi-periodic time series. More generally, such a bound may be obtained as a result of a suitable statistical analysis. In all such cases, one could then use such a lower bound as a lower bound for the persistence of the cycle representative.

2. If it's likely that a birth simplex may be suboptimal for the application at hand. For instance, in an application such as the one discussed in this paper, for which one seeks time-optimal cycles, it might happen that the birth simplex connects vertices very far apart in time, and is thus a poor choice. In such a case, even small relaxations, for instance, taking a representative with 95% of the persistence of its class, can lead to much better outcomes, since the solution to the optimization problem no longer has to include the given birth simplex.

Given a homology class with barcode (b, d) and an initial representative cycle c_0, searching for an optimal homologous cycle with minimum persistence ε involves solving the problem:

$$\begin{aligned} \min \quad & \ell(c) \\ \text{subject to} \quad & c = c_0 + \partial_{p+1}(w) \\ & w \in C_{p+1}(K_{d-\varepsilon}) \end{aligned}$$

In Figure 6 we illustrate how our notions of time-optimal PH cycles compare with the existing notion of length-optimal PH cycles on a synthetic example of a noisy sine curve. We provide details about the algorithms and implementation in Appendix B.

5 Experiments

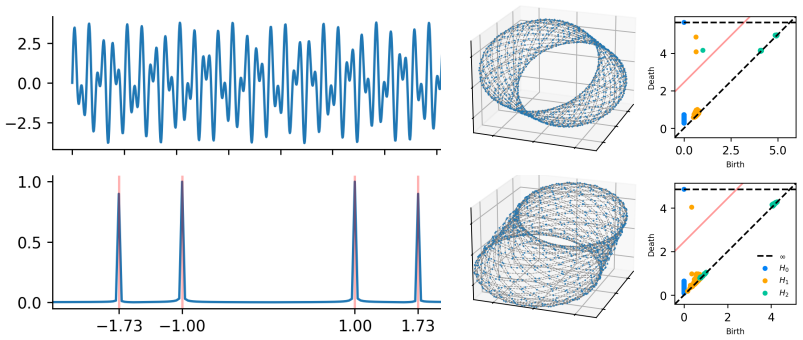
In all experiments we consider a Vietoris-Rips filtration on the embedded time-series (see Appendix C). For each PH class we obtain an initial representative $\mathbf{c}_0$ using the $R = \partial \cdot V$ decomposition of the filtration boundary matrix ∂. For each PH class with birth and death values (b, d) we restrict the domain and codomain of the boundary operator by considering the sets:

$$P = \{\sigma \in S_p(K_b) \mid \text{birth}(\sigma) \leq b\}$$

$$\hat{Q} = \{\sigma \in S_{p+1}(K_b) \mid \text{birth}(\sigma) \leq b \quad \text{and} \quad R[:, \sigma] \neq 0\},$$

which corresponds to the set of p- and $p+1$-simplices that are alive at filtration time b. This assures that the solution has a persistence value of at least $(d-b)$ (see Appendix B for more details). For a relaxed problem with minimum persistence ε we simply take $b' = d - \varepsilon$. We then solve the following linear problem [12, 22]:

$$\begin{aligned} \min \quad & \|W\mathbf{c}\|_1 = \sum_i \sum_j w_{ij}(c_j^+ + c_j^-) \\ \text{subject to } & (\mathbf{c}^+ - \mathbf{c}^-) = \mathbf{c_0} + \partial_{p+1}[P, \hat{Q}](\mathbf{w}) \\ & \mathbf{w} \in \mathbb{R}^{|\hat{Q}|} \\ & \mathbf{c} \in \mathbb{R}^{|P|} \\ & \mathbf{c}^+, \mathbf{c}^- \geq 0 \end{aligned}$$



(a) Original Signal (b) Sliding window embedding

Figure 7: Pipeline for computing the optimal embedding parameters for $f(t) = 2\sin t + 1.8sin\sqrt{3}t$. **a)** Top: graph of the function. Bottom: Fourier diagram, with peaks at 1 and $\sqrt{3}$. **b)** Two different embeddings and corresponding PDs for two different values of the delay τ. Top left: embedding for optimal value of the delay τ. Top right: corresponding PD. The continuous red diagonal line indicates the lower bound on persistence for all homology degrees (see Appendix C.4 for details). Bottom: embedding for suboptimal delay value, together with the corresponding PD. The embeddings are in $\mathbb{R}^4$ and we show the projection onto the first three principal components.

where $\partial_{p+1}[P, \hat{Q}]$ is a submatrix obtained by selecting the rows and columns of the boundary matrix ∂_{p+1}.

The experiments were run using the Gurobi solver [15] on an Apple M1 Pro with 16Gb of memory. The code to reproduce our experiments is available at our GitHub repository [1].

5.1 Synthetic quasi-periodic time series

We consider the quasi-periodic time series given by the function

$$f(t) = 2\sin t + 1.8sin\sqrt{3}t \quad 0 \leq t \leq 60\pi\,.$$

We sample 1000 points and compute the optimal embedding parameters using the methodology introduced in [26], see Figure 7 for an illustration of the pipeline, and Appendix C for details. We then proceed to compute vertex- and simplex-based time-optimal PH 1-cycle representatives for the embedding corresponding to the optimal parameters (top embedding in Figure 7(b)).

Given a a lower bound on the persistence of PH classes that we may interpret as "significant" 7(b), we use that same lower bound as a minimum persistence value for the search of approximate representatives (as opposed to representatives with same persistence as the class). The representative optimization is done with half the points (every other point sampled) due to computational constraints.

We illustrate the results of our computations in Figure 8. We observe that both simplex-based and vertex-based representatives demonstrate remarkable consistency in identifying cycle representatives that span approximately one period of the underlying signal.

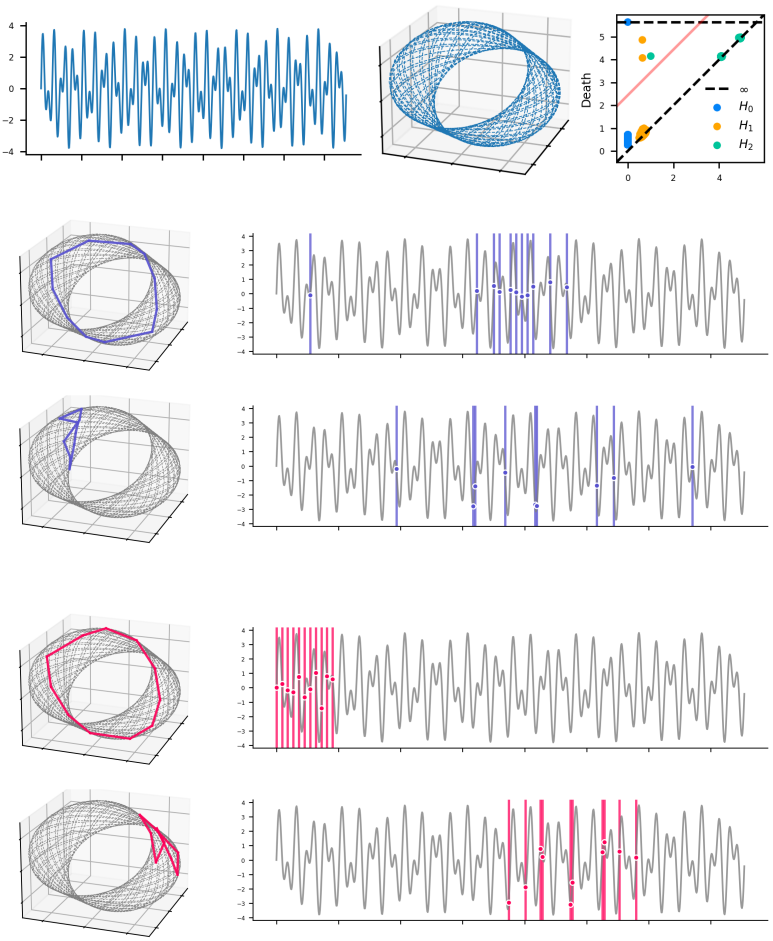


Figure 8: We illustrate the two 1-cycle PH representatives for the optimal embedding of $f(t) = 2\sin t + 1.8sin\sqrt{3}t$ in the point cloud (left), as well as in the original time series (right). Top shows the original signal (A) along with the optimal embedding (B) and the corresponding PD (C) with the lower bound. 2nd and 3rd row from top: simplex-based time-optimal cycles. 4th and 5th row from top: vertex-based time-optimal cycles.

Our approach can also be used for finding representatives for PH classes in degrees higher than 1. We illustrate time-optimal PH representatives for the single significant PH class in PD_2 in Figure 9.

We believe that such time-optimal representatives are critical for meaningful signal interpretation, not only for the reasons elucidated earlier, namely the need to be able to physically interpret the results, e.g., in a dynamical systems application, but also because, as the example in Figure 9 illustrates, visualising cycle representatives in the embedding point clouds is a challenging problem: while visualising appropriately

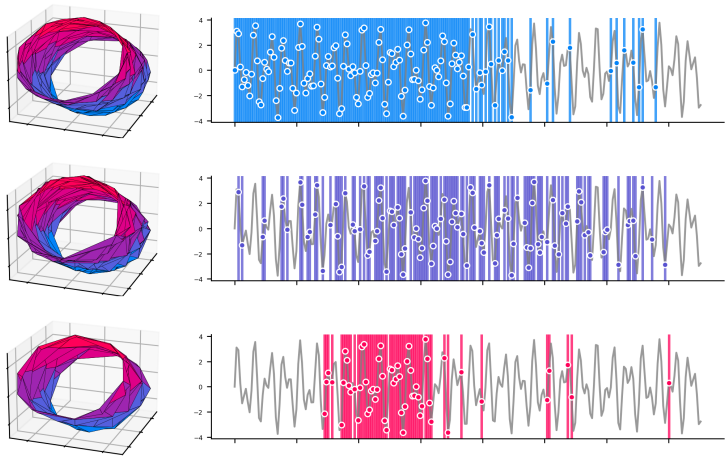


Figure 9: Time-optimal representative 2-cycles for the only significant PH class in PD_2, for the optimal embedding point cloud from Figure 7. Top: initial representative, given as output by the implementation. Middle: simplex-based optimization. Bottom: vertex-based optimization. We note that the embedding is 4-dimensional and the triangulation shown on the left column is a projection onto the first three principal coordinates.

1-cycle representatives is challenging for point clouds in more than 3-dimensions, having any reasonable visualisation in the embedding point cloud becomes an even more challenging task for cycle representatives in higher homology degrees.

5.2 Delayed oscillator models of El Ninõ Southern Oscillation

We consider quasi-periodic time series arising from a delayed oscillator model of the El Ninõ Southern Oscillation [9]. In particular, the model depends on a parameter κ encoding the strength of the ocean-atmosphere coupling, and we study three different time series for the values $\kappa \in \{1.4, 1.65, 1.9\}$. We provide details about this data set in Appendix D.

Given such a time series, we first compute its embedding parameters by using the methodology developed in [26]. We give more details about the parameter computations in Appendix C. We then compute persistent homology with the Vietoris-Rips complex of the resulting points clouds, and for each point in the persistence diagram PD_1, we compute time-optimal PH representatives. The optimization problem is performed with an evenly spaces subsample of 500 points due to computational constraints. We illustrate the results of the computations in Figures 10, 11 and 12.

To be able to obtain representatives that are easier to interpret for our application, we compute a relaxation of time-optimal PH representatives, in which we allow simplices to have later birth times. We provide more details about this relaxation technique in Section 4.2. Here we report results for the relaxed version of the PH cycle representatives that have 90% persistence of the corresponding class (as opposed to full persistence). We report the results for cycle representatives with full persistence in Appendix E.

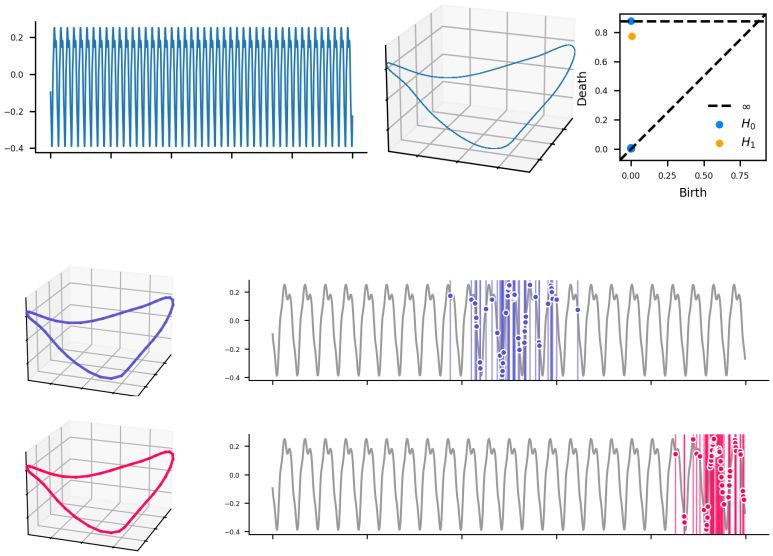


Figure 10: Computations for ENSO model for $\kappa = 1.65$. Top: original time series, optimal embedding and resulting PD. Middle: simplex-wise time representatives, with relaxed persistence. Bottom: vertex-wise time representatives, with relaxed persistence.

We believe that our approach offers a new perspective for climate science, providing a sophisticated method to characterize oscillatory behaviors that goes beyond the standard analysis of state space embeddings. More specifically we observe, in certain coupling scenarios, cycle representatives that cluster tightly within specific temporal windows yet have very dispersed spatial configurations (see bottom row of Figure 11). We aim to further investigate the insights that our methods can bring in the study of the El Ninõ Southern Oscillation in future work.

6 Conclusions and open questions

In the present work we have introduced algorithms for computing PH representatives that correspond to data points that are close in time, in an appropriate sense. We have illustrated the outputs of the different algorithms with synthetic quasi-periodic signals, and several univariate time series arising as models of the El Niño Southern Oscillation. In most examples that we consider, both methods (simplex and vertex based) manage to extract cycle representatives that are within one period of oscillation, while, overall vertex-based optimization performs better.

One key aspect in extracting representatives that are physically meaningful is the relaxation of birth times for representatives. We aim to further investigate such relaxation techniques, as well as their stability, in future work. In particular, we note that such relaxation techniques are widely applicable to many different notions of

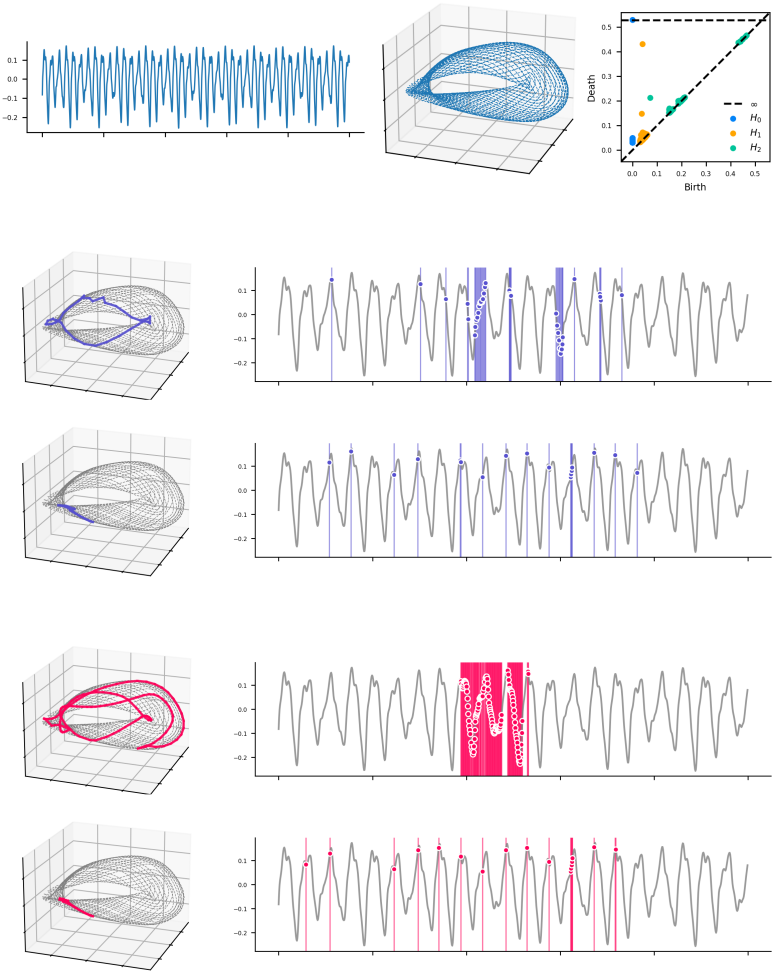


Figure 11: Computations for ENSO model for $\kappa = 1.4$. Top: original time series, optimal embedding and resulting PD. Middle: simplex-wise time representatives, with relaxed persistence. Bottom: vertex-wise time representatives, with relaxed persistence.

optimal PH cycle representatives.

Our methods can be used to compute time-optimal p-cycles, for any $p \geq 0$. In the current work we focused on $p = 1$, but we also provide an example for $p = 2$, for the synthetic quasi-periodic time series, see Figure 9. We aim to optimise our implementation in future work to be able to compute higher-degree time-optimal cycles for the ENSO model time series. Being able to compute such higher-degree PH representatives is particularly important in the study of quasi-periodic time series, which have many topological features of interest in homological degrees greater than 1. We also note that it might be of interest to explore to which extent the choice of norm in the simplex-based method plays a role in the computation of time-optimal

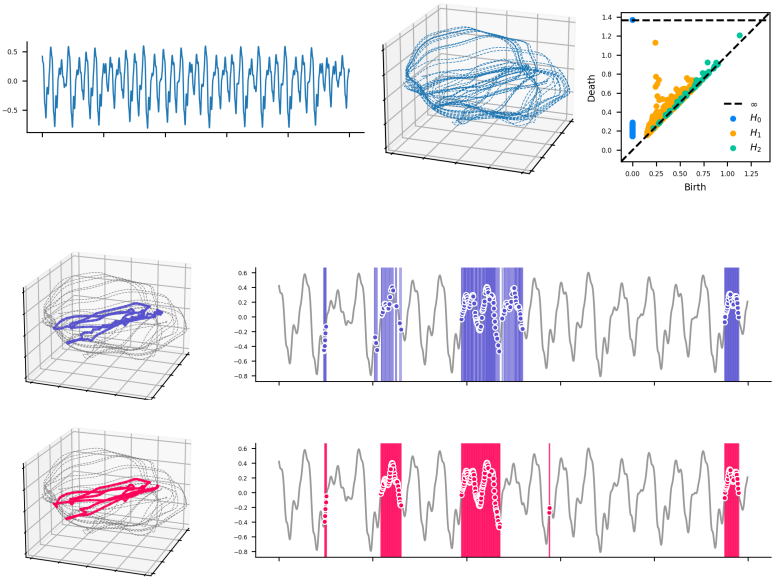


Figure 12: Computations for ENSO model for $\kappa = 1.9$. Top: original time series, optimal embedding and resulting PD. Middle: simplex-wise time representatives, with relaxed persistence. Bottom: vertex-wise time representatives, with relaxed persistence.

cycles.

Finally, we note that our methods can be applied to other types of time-varying data, including time-varying networks and point clouds. There are several notions of time-varying networks; we refer the reader to [20] for an overview. If one considers time-varying networks in which edges may be deleted or added over time, then one does in general not obtain a nested sequence of simplicial complexes, but rather a zigzag of simplicial complexes, for which one can compute PH with the zigzag algorithm [4, 21, 24]. In particular, we note that for the study of temporal networks, one might be interested in a notion of time dispersion different from the one considered here for time series, and is closer in spirit to the simplex-based approach; namely, one might be interested in defining the time dispersion of a p-chain as the difference between minimum and maximum time labels of any of the 1-simplices (i.e., edges in the network) contained in the chain. Zig-zag algorithms have also been used to study time-varying point clouds, for instance in the analysis of flocking behaviour [21]. We will study applications of our methods to such types of data sets in future work.

Acknowledgments

We thank Hannah Christensen for sharing with us the code to compute the ENSO models time series.

References

[1] GitHub repository for Time-optimal PH representatives. URL: https://github.com/antonio-leitao/optimal-cycles.

[2] Ulrich Bauer. Ripser: efficient computation of vietoris–rips persistence barcodes. *Journal of Applied and Computational Topology*, 5(3):391–423, 2021.

[3] Kenneth A. Brown and Kevin P. Knudson. Nonlinear statistics of human speech data. *International Journal of Bifurcation and Chaos*, 19(07):2307–2319, 2009. arXiv:https://doi.org/10.1142/S0218127409024086, doi:10.1142/S0218127409024086.

[4] Gunnar Carlsson, Vin de Silva, and Dmitriy Morozov. Zigzag persistent homology and real-valued functions. In *Proceedings of the Twenty-Fifth Annual Symposium on Computational Geometry*, SCG '09, page 247–256, New York, NY, USA, 2009. Association for Computing Machinery. doi:10.1145/1542362.1542408.

[5] Gunnar E. Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46:255–308, 2009. URL: https://api.semanticscholar.org/CorpusID:1472609.

[6] Nicholas J. Cavanna, Mahmoodreza Jahanseir, and Donald R. Sheehy. A geometric perspective on sparse filtrations, 2015. URL: https://arxiv.org/abs/1506.03797, arXiv:1506.03797.

[7] C. Chen and D. Freedman. Hardness results for homology localization. *Discrete & Computational Geometry*, 45:425–448, 2011. URL: https://doi.org/10.1007/s00454-010-9322-8.

[8] Chao Chen and Daniel Freedman. Quantifying Homology Classes. In Susanne Albers and Pascal Weil, editors, *25th International Symposium on Theoretical Aspects of Computer Science*, volume 1 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 169–180, Dagstuhl, Germany, 2008. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. URL: https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.STACS.2008.1343, doi:10.4230/LIPIcs.STACS.2008.1343.

[9] Hannah M. Christensen, Judith Berner, Danielle R. B. Coleman, and Tim N. Palmer. Stochastic parameterization and el niño–southern oscillation. *Journal of Climate*, 30(1):17 – 38, 2017. URL: https://journals.ametsoc.org/view/journals/clim/30/1/jcli-d-16-0122.1.xml, doi:10.1175/JCLI-D-16-0122.1.

[10] David Cohen-Steiner, Herbert Edelsbrunner, and Dmitriy Morozov. Vines and vineyards by updating persistence in linear time. In *Proceedings of the Twenty-Second Annual Symposium on Computational Geometry*, SCG '06, page 119–126, New York, NY, USA, 2006. Association for Computing Machinery. doi:10.1145/1137856.1137877.

[11] Tamal K. Dey, Tao Hou, and Sayan Mandal. Persistent 1-cycles: Definition, computation, and its application. In Rebeca Marfil, Mariletty Calderón, Fernando Díaz del Río, Pedro Real, and Antonio Bandera, editors, *Computational*

Topology in Image Context, pages 123–136, Cham, 2019. Springer International Publishing.

[12] Emerson G Escolar and Yasuaki Hiraoka. Optimal cycles for persistent homology via linear programming. In *Optimization in the Real World: Toward Solving Real-World Optimization Problems*, pages 79–96. Springer, 2016.

[13] H. Gakhar and J.A. Perea. Sliding window persistence of quasiperiodic functions. *Journal of Applied and Computational Topology*, 8:55–92, 2024. URL: https://doi.org/10.1007/s41468-023-00136-7.

[14] Barbara Giunti, Jānis Lazovskis, and Bastian Rieck. DONUT: Database of Original & Non-Theoretical Uses of Topology, 2022. https://donut.topology.rocks.

[15] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2024. URL: https://www.gurobi.com.

[16] İsmail Güzel, Elizabeth Munch, and Firas A. Khasawneh. Detecting bifurcations in dynamical systems with crocker plots. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 32(9):093111, 09 2022. arXiv:https://pubs.aip.org/aip/cha/article-pdf/doi/10.1063/5.0102421/19806050/093111_1_online.pdf, doi:10.1063/5.0102421.

[17] A. Hatcher. *Algebraic Topology*. Algebraic Topology. Cambridge University Press, 2002. URL: https://books.google.fr/books?id=BjKs86kosqgC.

[18] Abigail Hickok. Computing Persistence Diagram Bundles. *arXiv e-prints*, page arXiv:2210.06424, October 2022. arXiv:2210.06424, doi:10.48550/arXiv.2210.06424.

[19] Abigail Hickok. Persistence Diagram Bundles: A Multidimensional Generalization of Vineyards. *arXiv e-prints*, page arXiv:2210.05124, October 2022. arXiv:2210.05124, doi:10.48550/arXiv.2210.05124.

[20] Petter Holme and Jari Saramäki. Temporal networks. *Physics Reports*, 519(3):97–125, 2012. Temporal Networks. doi:10.1016/j.physrep.2012.03.001.

[21] Woojin Kim, Facundo Mémoli, and Zane Smith. Analysis of dynamic graphs and dynamic metric spaces via zigzag persistence. In Nils A. Baas, Gunnar E. Carlsson, Gereon Quick, Markus Szymik, and Marius Thaule, editors, *Topological Data Analysis*, pages 371–389, Cham, 2020. Springer International Publishing.

[22] Lu Li, Connor Thompson, Gregory Henselman-Petrusek, Chad Giusti, and Lori Ziegelmeier. Minimal cycle representatives in persistent homology using linear programming: An empirical study with user's guide. *Frontiers in Artificial Intelligence*, 4, 2021. doi:10.3389/frai.2021.681117.

[23] M Munnich, Mark A Cane, and Stephen E Zebiak. A study of self-excited oscillations of the tropical ocean-atmosphere system. *J. Atmos. Sci*, 48:1238–1248, 1991.

[24] Audun Myers, David Muñoz, Firas A. Khasawneh, and Elizabeth Munch. Temporal network analysis using zigzag persistence. *EPJ Data Science*, 12, 2023.

[25] Ippei Obayashi. Volume-optimal cycle: Tightest representative cycle of a generator in persistent homology. *SIAM Journal on Applied Algebra and Geometry*, 2(4):508–534, 2018. doi:10.1137/17M1159439.

[26] Jose A Perea. Topological time series analysis. *Not Am Math Soc*, 66(5):686–694, 2019.

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

[28] Jose A. Perea and John Harer. Sliding windows and persistence: An application of topological methods to signal analysis. *Foundations of Computational Mathematics*, 15(3):799–838, June 2015. doi:10.1007/s10208-014-9206-z.

[29] Nalini Ravishanker and Renjie Chen. An introduction to persistent homology for time series. *WIREs Computational Statistics*, 13(3):e1548, 2021. doi:10.1002/wics.1548.

[30] Wojciech Reise, Bertrand Michel, and Frédéric Chazal. Topological signatures of periodic-like signals. *arXiv e-prints*, page arXiv:2306.13453, June 2023. arXiv:2306.13453, doi:10.48550/arXiv.2306.13453.

[31] Katharine Turner. Representing vineyard modules, 2023. URL: https://arxiv.org/abs/2307.06020, arXiv:2307.06020.

[32] A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete & Computational Geometry*, 33:249–274, 2005. URL: https://doi.org/10.1007/s00454-004-1146-y.

[33] Matija Čufar and Žiga Virk. Fast computation of persistent homology representatives with involuted persistent homology, 2023. URL: https://www.aimsciences.org/article/id/64252c216565307f274c6b30, doi:10.3934/fods.2023006.

A (Persistent) homology with coefficients over arbitrary fields

The main ingredient in defining simplicial homology and persistent homology over arbitrary coefficient fields is given by a notion of *orientation* of simplices.

An **orientation** of a simplex is the choice of a total order on its vertices. Two orientations of a simplex are defined to be equivalent if they differ by an even permutation.

In practice, one usually chooses a total order on the set of vertices of a simplicial complexes, and then gives to each simplex the orientation induced by the order on its vertices. We illustrate an example of a simplex with different orientations in Figure 13.

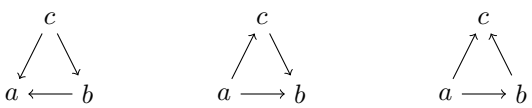


Figure 13: A simplex with different orientations, where given an order $<$ on the vertices, we denote $i < j$ by drawing an arrow $j \to i$. Left: orientation given by lexicographical order $a < b < c$. Middle: orientation given by $b < c < a$. Right: Orientation given by $c < b < a$. This orientation is equivalent to the one given by the lexicographical order.

The main difference to defining simplicial homology over the field with 2 elements then relies on how the boundary maps on chains are defined. One defines

$$\partial_p : C_p(K) \longrightarrow C_{p-1}(K)$$
$$\sigma = (x_0, \ldots, x_p) \mapsto \sum_{i=0}^{p} (-1)^i (x_0, \ldots, \widehat{x}_i, \ldots, x_p)$$

where the symbol $\widehat{x}_i$ means that the ith vertex has been deleted.

We refer the reader to [17] for more details on simplicial homology over general coefficient fields.

B Details on algorithms and implementation

B.1 Optimal homologous cycle

Given an initial p-cycle $c_0 \in C_p(K)$, finding an homologous cycle that minimizes a given loss function ℓ is defined as:

$$\begin{aligned} \min \quad & \ell(c) \\ \text{subject to} \quad & c = c_0 + \partial_{p+1}(w) \\ & w \in C_{p+1}(K) \end{aligned}$$

The search space consists of the initial cycle plus the boundary of a higher dimensional simplex. Guaranteeing, by definition, that the solution c is homologous to the initial cycle c_0.

In a simplicial filtration, finding a representative for a homology class with birth and death (b, d) requires that the optimal representative also shares the same birth and death. This reflects in restricting the search to the simplicial complex existing at birth time (K_b). Which resolves to the following problem [8]:

$$\begin{aligned} \min \quad & \ell(c) \\ \text{subject to} \quad & c = c_0 + \partial_{p+1}(w) \\ & w \in C_{p+1}(K_b) \\ & \partial_{p+1}(w) \in C_p(K_b) \end{aligned}$$

Where the search is contained in cycles composed of simplices that have a birth *before or at* the birth of the homology class. The third constraint is redundant since if $w \in K_b$ then $\partial(w) \in K_b$, that is, the birth of a simplex is always later than or equal to the ones that make up its boundary.

B.2 Linear Programming

If the function we optimize is linear, then it is possible to solve the previous problem with linear optimization [12]. The added step is separating the coefficients vector c into a positive c^+ and a negative c^- part such that $c = c^+ - c^-$ allowing us to set both $c^+, c^- \geq 0$.

The optimization can be done both by allowing only binary coefficients ($\mathbb{F}_2$) which requires a solver capable of mixed linear programming or can be relaxed to $\mathbb{R}$. In general using the l_1 norm over $\mathbb{R}$ is sufficient to provide sparse solutions [22].

Let $\mathbf{c} = \mathbf{c}^+ - \mathbf{c}^-$ be the vector of coefficients in $\mathbb{R}$. Let W be a non-negative weight matrix, then the linear programming formulation is the following:

$$\begin{aligned} \min \quad & \|W\mathbf{c}\|_1 = \sum_i \sum_j w_{ij}(c_j^+ + c_j^-) \\ \text{subject to } & (\mathbf{c}^+ - \mathbf{c}^-) = \mathbf{c_0} + \partial_{p+1}(\mathbf{w}) \\ & \mathbf{w} \in \mathbb{R}^m \\ & \mathbf{c} \in \mathbb{R}^n \\ & \mathbf{c}^+, \mathbf{c}^- \geq 0 \end{aligned}$$

The initial representative $\mathbf{c}_0$ can be obtained by most persistent homology methods. It appears naturally from the decomposition $R = \partial_k V$ of the filtration boundary matrix. Given the birth b of the homology class we consider the sets of p and $p+1$ simplices that are alive at filtration time b:

$$P = \{\sigma \in S_p(K_b) \mid \text{birth}(\sigma) \leq b\}$$

$$Q = \{\sigma \in S_{p+1}(K_b) \mid \text{birth}(\sigma) \leq b\}$$

The solution of the previous problem with the rows and columns of the boundary operator ∂_{p+1} restricted to $\partial_{p+1}[P, Q]$ is a cycle with persistence (b, d).

A substantial speedup comes from restricting the domain of the boundary operator even further by considering the set [22]:

$$\hat{Q} = \{\sigma \mid R[:, \sigma] \neq 0\}$$

Which are the nonzero columns of the R matrix resulting from the previous decomposition $R = \partial_{p+1} V$. This significantly reduces the number of conditions without affecting the search space. The resulting problem is the following:

$$\begin{aligned} \min \quad & \|W\mathbf{c}\|_1 = \sum_i \sum_j w_{ij}(c_j^+ + c_j^-) \\ \text{subject to } & (\mathbf{c}^+ - \mathbf{c}^-) = \mathbf{c_0} + \partial_{p+1}[P, \hat{Q}](\mathbf{w}) \\ & \mathbf{w} \in \mathbb{R}^{|\hat{Q}|} \\ & \mathbf{c} \in \mathbb{R}^{|P|} \\ & \mathbf{c}^+, \mathbf{c}^- \geq 0\,. \end{aligned}$$

C Time-delay embeddings of quasi-periodic time series

In this section we provide details on a framework developed in [13], which provides a rigorous methodology to compute optimal parameters for sliding window embeddings of quasi-periodic functions. This computational methodology allows to obtain embedding parameters so that the persistent homology, computed with respect to the Vietoris-Rips complex of the obtained embedding point cloud, reflects in a robust manner the persistent homology of a hypertorus in as many dimensions as the frequencies characterising the quasi-periodic function. We emphasise that there are no contribution by the authors in this section, and all contributions are due to the authors of [13].

C.1 Quasiperiodic Functions

Definition C.1 (Quasi-periodic function)**.** Let $N \in \mathbb{N}$ and $\mathbb{T}^N = (\mathbb{R}/2\pi\mathbb{Z})^N$. A function $f\colon \mathbb{R} \to \mathbb{C}$ is **quasi-periodic** if there exists a vector $\omega = (\omega_1, \ldots, \omega_N) \in \mathbb{R}^{\mathbb{N}}_{\geq 0}$ with components ω_i linearly independent over $\mathbb{Q}$ and a function $F : \mathbb{T}^N \to \mathbb{C}$ such that:

$$f(t) = F(\omega_1 t, \ldots, \omega_N t)\,.$$

The vector ω is called **frequency vector** of f, and the function F **parent function** of f.

In other words, the function f is of the form

$$f(t) = c_1 e^{i\omega_1 t} + c_2 e^{i\omega_2 t} + \cdots + c_N e^{i\omega_N t}$$

where $\omega_1, \ldots \omega_N$ are linearly independent over $\mathbb{Q}$ (incommensurable) and $c_i > 0$.

The sliding window embedding of $f(t)$ with delay $\tau > 0$ and embedding dimension $d + 1$ is defined for any $t \in \mathbb{R}$ as:

$$SW_{d,\tau} f(t) = \begin{bmatrix} f(t) \\ f(t+\tau) \\ \vdots \\ f(t+d\tau) \end{bmatrix} \in \mathbb{C}^{d+1}.$$

The geometry of the embedding $SW_{d,\tau}f$ depends on the parameters τ and d, as well as the properties of f.

For practical computations, $f(t)$ is approximated by its truncated Fourier series. Given positive integers N and K, one considers a restricted integral square box of side length bounded by $2K$ in an N-dimensional grid with integer grid points:

$$I_K^N = \{\mathbf{k} \in \mathbb{Z}^N \mid \|\mathbf{k}\|_\infty < K\}.$$

One then defines the truncation:

$$S_K f(t) = \sum_{\mathbf{k} \in I_K^N} \hat{F}(\mathbf{k}) e^{i\langle \mathbf{k}, \omega t \rangle},$$

where the integer K can be thought of as controlling the fidelity of the truncation.

C.2 Optimal dimension

We can estimate the nonzero Fourier coefficients $\hat{F}(\mathbf{k})$ and their frequency locations $\langle \mathbf{k}, \omega \rangle$ as follows:

$$\mathrm{supp}(\hat{F}_K) := \left\{ \mathbf{k} \in \mathbb{Z}^N \;\middle|\; \hat{F}(\mathbf{k}) \neq 0 \text{ and } \|\mathbf{k}\|_\infty \leq K \right\}.$$

We take the embedding dimension d to be the cardinality of $\mathrm{supp}(\hat{F}_K)$, (the number of prominent peaks in the spectrum of f).

C.3 Optimal delay

The choice of delay τ influences the appearances of the homological features in given degrees of the hypertorus in the sliding window embedding. In particular, poor choices can obscure them. Figure 7 shows an example of how a suboptimal delay parameter (bottom) squashes the homology groups. The sliding window embedding, for dimension d and delay τ, of the truncated Fourier approximation of f is the following

$$\begin{aligned} SW_{d,\tau} S_K f(t) &= \begin{bmatrix} 1 & \dots & 1 \\ e^{i\langle \mathbf{k}_1, \omega \rangle \tau} & \dots & e^{i\langle \mathbf{k}_\alpha, \omega \rangle \tau} \\ \vdots & \vdots & \vdots \\ e^{i\langle \mathbf{k}_1, \omega \rangle \tau d} & \dots & e^{i\langle \mathbf{k}_\alpha, \omega \rangle \tau d} \end{bmatrix} \cdot \begin{bmatrix} \hat{F}(\mathbf{k}_1) e^{i\langle \mathbf{k}_1, \omega \rangle t} \\ \vdots \\ \hat{F}(\mathbf{k}_\alpha) e^{i\langle \mathbf{k}_\alpha, \omega \rangle t} \end{bmatrix} \\ &= \Omega_{K,f} \cdot x_{K,f}(t) \end{aligned}$$

Note that only the $\Omega_{K,x}$ matrix depends on the delay value τ. The optimal choice of delay is the one that best improves the conditioning number of $\Omega_{K,x}$ so that no toroidal features are squashed. This is done by minimizing a scalar function that measures the extent to which columns in $\Omega_{K,f}$ are pairwise orthogonal.

C.4 Persistence Significance Bounds

The embedding quality depends on the Fourier approximation (choice of K) and the embedding parameters (d, τ) as well as the smallest singular value $\sigma_{\min}$ of $\Omega_{K,f}$. We have that [13, Theorem 6.8] provides lower bounds on the lifespan $b - a$ of points (a, b) in a persistence diagram. Such bounds can be interpreted as giving a separation between noise (points with lifespan smaller than the bounds) and signal (points with lifespan greater or equal to the bounds).

D ENSO model time series

The El Niño–Southern Oscillation (ENSO) is a set of coupled ocean-atmosphere phenomena characterized by an irregular cycle of warming (El Niño) and cooling (La Niña) in the eastern tropical Pacific along with a corresponding variation in sea level pressure [9,23]. ENSO significantly impacts global weather patterns and is associated with heavy rain in Peru, drought in Indonesia, intensity of the Indian monsoon and the number of hurricanes in North America [9]. The parameter κ represents the coupling strength of the ocean-atmosphere coupling, different values of this parameter simulate a variety of ENSO behaviors.

The ENSO data consists in 3 timeseries, one for each value of $\kappa = \{1.4, 1.65, 19\}$, each with 100000 points representing the modeled oscillation. Each time series is embedded in $\mathbb{R}^n$ using the optimal parameters. Below we show the different series and persistence diagrams of optimal (labeled MIN in the figures) and suboptimal embeddings. The persistent homology is calculated using Ripser [2] using a subsample of 1000 points [6].

D.1 ENSO model ($\kappa = 1.4$)

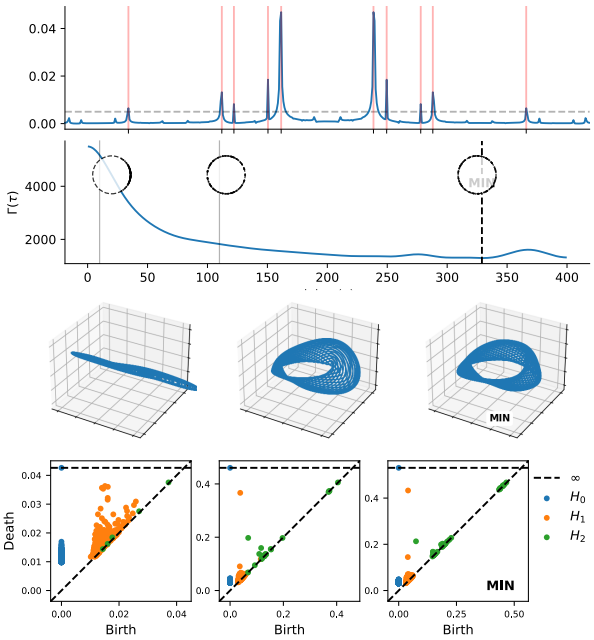


Figure 14: Optimal sliding window embedding of the ENSO model for the parameter $k = 1.4$. **Top:** Fourier decomposition of the signa, and the average orthogonality of the $\Omega_{K,f}$ matrix (Appendix C) with respect to a chosen delay value. **Middle:** PCA reduction of the original embeddings in $\mathbb{R}^{10}$ for optimal (right) and suboptimal delay values (right and middle). **Bottom:** Persistence Diagram for each embedding.

D.2 ENSO model ($\kappa = 1.65$)

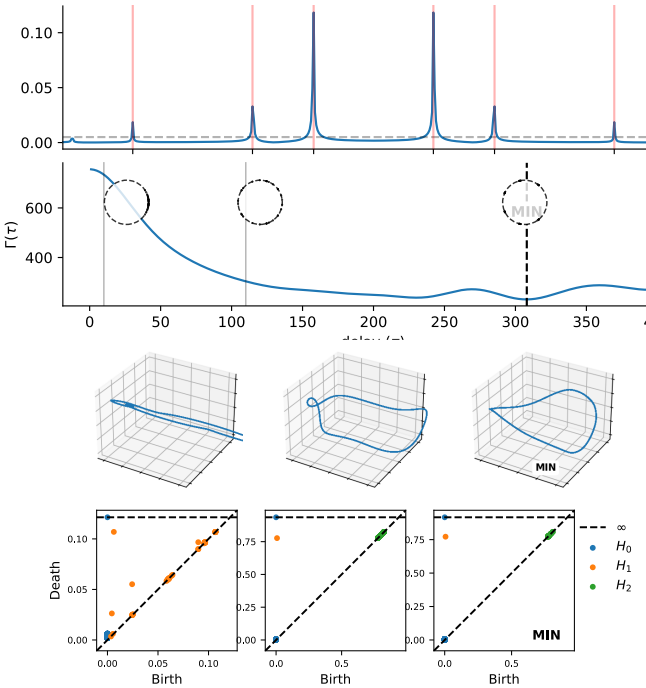


Figure 15: Optimal sliding window embedding of the ENSO model for the parameter $k = 1.65$. **Top:** Fourier decomposition of the signal, and the average orthogonality of the $\Omega_{K,f}$ matrix (Appendix C) with respect to a chosen delay value. **Middle:** PCA reduction of the original embeddings in $\mathbb{R}^6$ for optimal (right) and suboptimal delay values (right and middle). **Bottom:** Persistence Diagram for each embedding.

D.3 ENSO model ($\kappa = 1.9$)

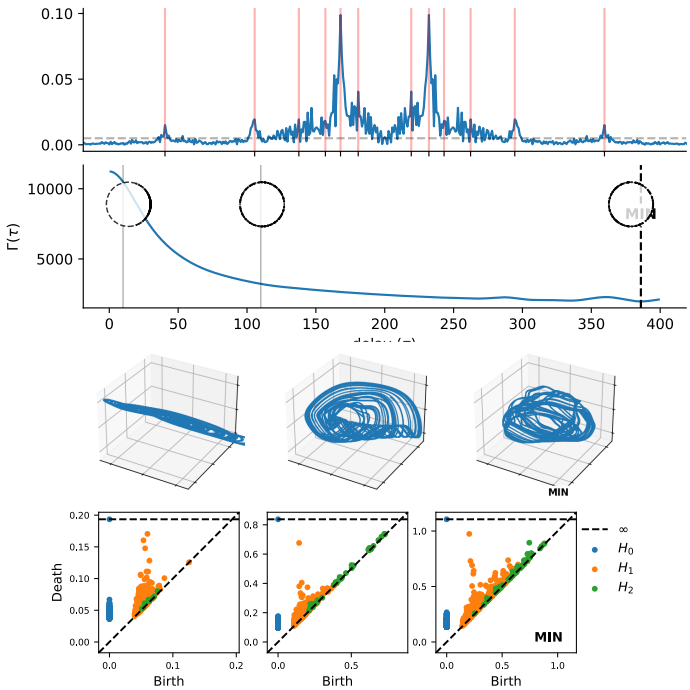


Figure 16: Optimal sliding window embedding of the ENSO model for the parameter $k = 1.9$.**Top:** Fourier decomposition of the signal, and the average orthogonality of the $\Omega_{K,f}$ matrix (Appendix C) with respect to a chosen delay value. **Middle:** PCA reduction of the original embeddings in $\mathbb{R}^{12}$ for optimal (right) and suboptimal delay values (right and middle). **Bottom:** Persistence Diagram for each embedding.

E Additional computations: full persistence time-optimal PH cycle representatives

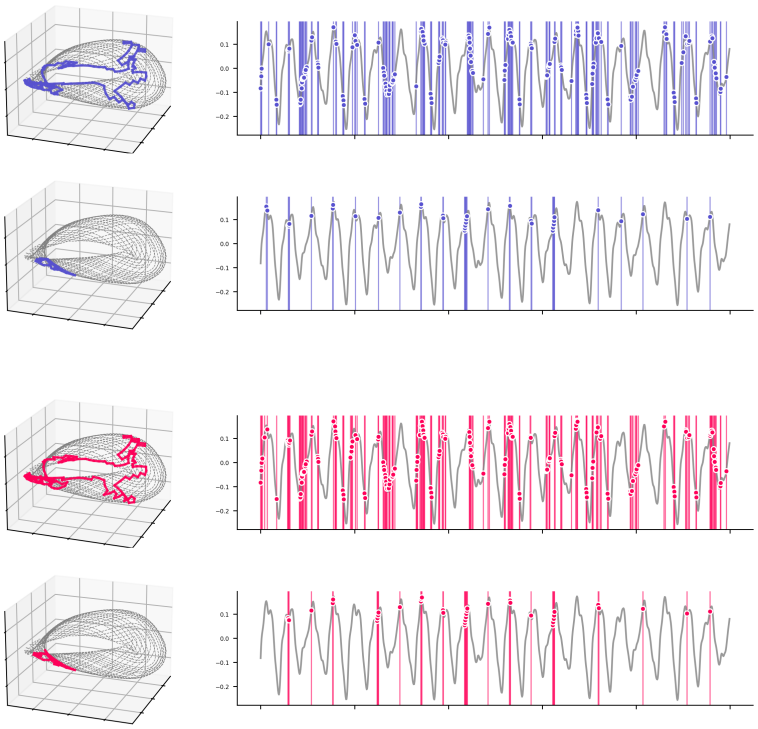


Figure 17: Time representatives for the ENSO model with $\kappa = 1.4$, simplex-based (top) and vertex-based (bottom), with full persistence

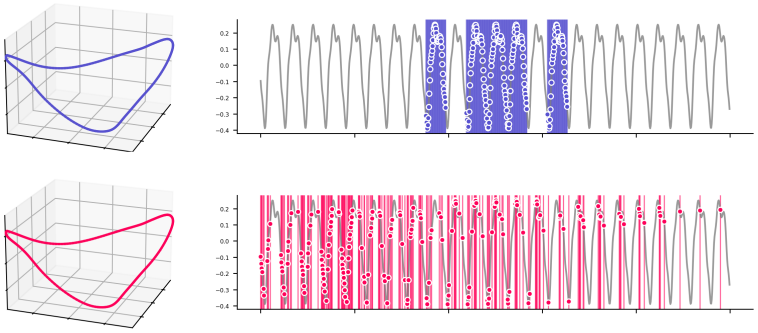


Figure 18: Time representatives for the ENSO model with $\kappa = 1.65$, simplex-based (top) and vertex-based (bottom), with full persistence

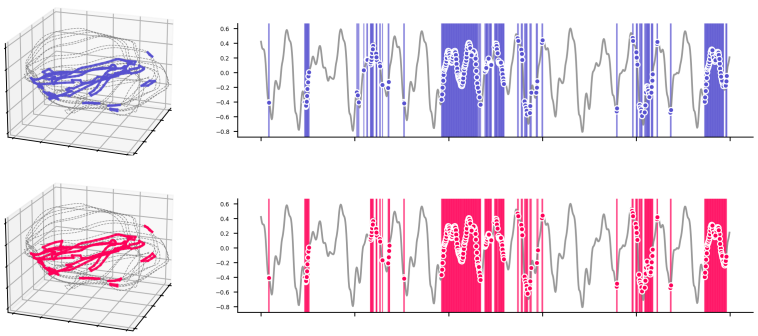


Figure 19: Time representatives for the ENSO model with $\kappa = 1.9$, simplex-based (top) and vertex-based (bottom), with full persistence.