A Stable Measure of Similarity for Time Series using Persistent Homology

Abstract

Persistent homology, the study of holes that appear in data as one thickens balls centered around its points over time, has theoretically guaranteed stability. That is, small data perturbations guarantee small changes in the lifetimes of these holes. This stability has been used to construct a measure of periodicity for a single univariate time series, denoted the sliding windows and 1-persistence scoring function, $\text{score}(f_1)$. Along with periodicity analysis of single time series, similarity between a pair of time series has also been studied. One popular measure of similarity between two time series is percent determinism (%DET), which measures the correlation between two time-series embeddings. Percent determinism relies on Euclidean distance alone, which makes it less robust to small time-series perturbations in practice. Motivated by the stability of persistent homology, coupled with the persistence-based periodicity scoring function $\text{score}(f_1)$, we introduce a novel persistent-homology based measure of time-series similarity which we denote the bi-conditional periodicity score, $\text{score}(f_1, f_2)$. We prove the stability of our measure under small time series and frequency perturbations, as well as the existence of a minimum embedding dimension for the convergence of our score. Our latter result implies that larger embedding dimensions may be necessary to reach desired levels of convergence. Since pairwise distances between points in these larger dimensions may start to concentrate, we also prove the stability of our measure under dimension reduction which guarantees that as long as the first K principal components capture a majority of the variance under orthogonal projection, the score will undergo small changes. We next introduce an algorithm for computing the bi-conditional periodicity score and deduce its computational complexity as $O(N \log N + PK^2 + P^6)$ for N the number of time series points, P the number of embedding points, and K the number of principal components. We experimentally verify the greater stability of our measure in comparison with %DET on both synthetic time series as well as real climate data. Along with the greater stability of our measure to %DET, $\text{score}(f_1, f_2)$ also requires only one parameter for its computation while %DET requires four.

Keywords: time series, persistent homology, PCA

1 Introduction

Time series analysis is a well-studied field. One time-series feature commonly studied is the extraction of its dominant frequency components. One of the earliest numerical methods introduced for estimating the frequency content of a univariate signal is the Fourier Transform [1], which quantifies the strength of a time series at a given frequency. Frequency is commonly measured in Hertz, or cycles per second. Given a continuous, real-valued time series $f(t) : T \subseteq \mathbb{R} \to \mathbb{R}$ and a frequency w, the continuous Fourier Transform of f at w is the area under the product of f with a purely sinusoidal signal at frequency w, given by

$$CFT[f](w) = \int_T f(t) \cdot e^{-i2\pi wt} dt,$$

where $e^{-i2\pi wt} = \cos(2\pi wt) - i\sin(2\pi wt)$. For discrete signals, this transform simply computes the dot product of $f(t)$ with purely sinusoidal signal. Intuitively, larger values of $CFT[f](w)$ indicate greater presence of this frequency in f. Alternative transforms have also been introduced that are more suited for studying nonstationary series (i.e. those whose frequencies change with time) or series with wavelet structure (i.e. seismic waves), such as the Short Time Fourier Transform [2] and the Wavelet Transform [3]. A newer method has also been introduced that can analyze the frequency content of a time series by measuring correlation between the sliding windows embeddings $\mathrm{SW}_{M,\tau} f(t)$ and $\mathrm{SW}_{M,\tau} e^{i2\pi wt}$ of a time series f and a reference signal $e^{i2\pi wt}$ given an embedding dimension $M \in \mathbb{N}$ and a time lag $\tau \in \mathbb{R}$. Denoted as cross-recurrence quantification analysis [4], this method summarizes all pairs of points that cross-recur (i.e. are close enough in Euclidean distance) between $\mathrm{SW}_{M,\tau} f(t)$ and $\mathrm{SW}_{M,\tau} e^{i2\pi wt}$ in a binary matrix $\mathcal{C}$. The proportion of cross-recurring states (i.e. the proportion of ones in $\mathcal{C}$) is denoted as percent recurrence (%REC), given by

$$\%\mathrm{REC}[f, e^{-i2\pi wt}] = \frac{1}{|C|} \sum_{ij} \delta_{\mathtt{tol}}(C[i,j]),$$

where $\delta_{\mathtt{tol}}(C[i,j]) = 1$ if $\| \mathrm{SW}_{M,\tau} f(t_i) - \mathrm{SW}_{M,\tau} e^{i2\pi wt_j} \| \leq \mathtt{tol}$ and $|C|$ denotes the total number of points in C. This measure essentially quantifies the correlation between the embeddings of f and $e^{-i2\pi wt}$. Further, the proportion of cross-recurring states that exist consecutively for at least $\mathtt{minDL}$ time steps (i.e. the proportion of ones in $\mathcal{C}$ that belong to a diagonal of at least $\mathtt{minDL}$ ones) denotes the percent determinism (%DET), given by

$$\%\mathrm{DET}[f, e^{-i2\pi wt}] = \frac{1}{|C|} \sum_{ij} \eta_{\mathtt{minDL}} \left(\delta_{\mathtt{tol}}(C[i,j]) \right),$$

where $\eta_{\mathtt{minDL}} \left(\delta_{\mathtt{tol}}(C[i,j]) \right) = 1$ if $\delta_{\mathtt{tol}}(C[i,j]) = 1$ and $C[i,j]$ is in a diagonal of at least $\mathtt{minDL}$ ones.

While the Fourier Transform requires less steps for frequency analysis when compared to %REC, %REC has the added capability of estimating the correlation between two time series embeddings. In practice, %REC is less robust to added noise as it relies on Euclidean distance alone. We discuss these limitations in more detail in the next section.

2 Limitations

As previously mentioned, the Fourier Transform and cross-recurrence quantification are both tools we can use to study the frequency content of a time series, however in practice %REC and %DET are less robust to added data noise. Unlike the Fourier Transform, both recurrence measures have the additional capability of quantifying time-series similarity. As we'd naturally expect real time series data to contain some inherent noise, a measure that is more robust is necessary. In Figure 1, we show an example of the instability of %REC and %DET on a synthetic sinusoidal with added Gaussian noise. When just increasing the noise in f by 5%, determinism drops by nearly 20% and recurrence drops by nearly 2%.

Persistent homology, the study of holes in different dimensions that appear in point cloud data upon the thickening of their points over time, has become a popularized method for time series analysis and prediction [5, 6]. Part of its growing popularity is due to its guaranteed stability to small amounts of added Gaussian noise to time series, which we discuss in more detail below.
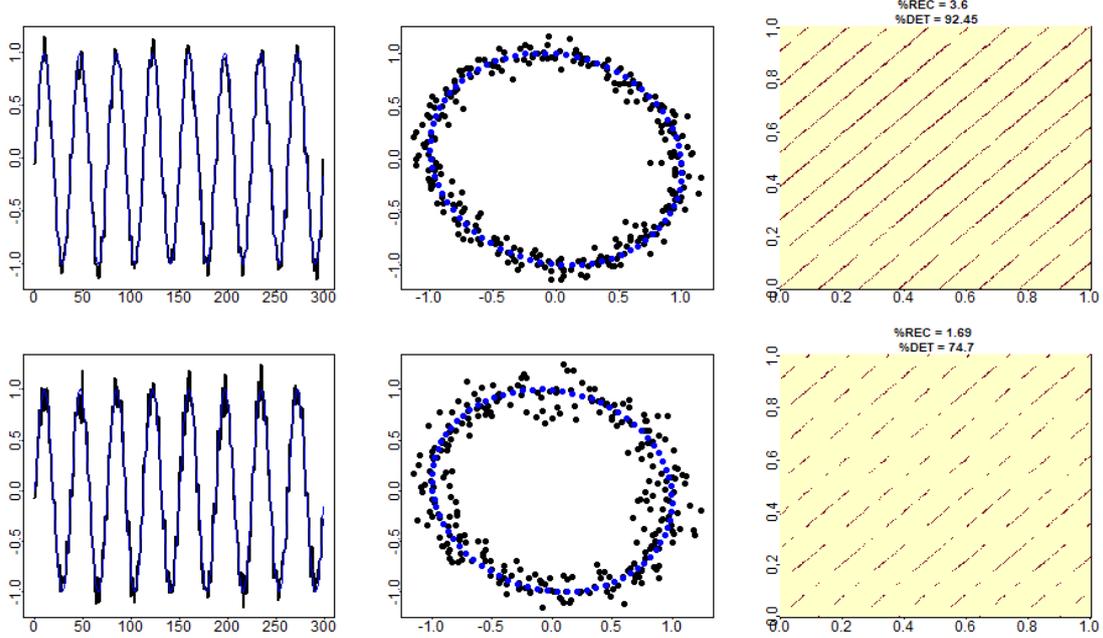
Fig. 1 The instability of percent recurrence and percent determinism when subject to added time-series noise. For each reference signal $f_2(t) = \sin(2\pi(3/300)t)$ (in blue), we define its noisy version (in black) as $f_1(t) = f_2(t) + \mathcal{N}(0, \sigma \cdot \mathrm{sd}(f_2(t)))$, where $\mathrm{sd}(f_2(t))$ denotes the standard deviation of $f_2(t)$. We vary σ beteen 0.1 (top) and $\sigma = 0.15$ (bottom). We select the embedding dimension ($M = 1$) and the time lag ($\tau = 10$) using the first minimum of the automutual information [7]. We fix the distance threshold to be 20% of the maximum pairwise distance between the embeddings of f_1 and f_2. We finally fix the minimum number of points to qualify a diagonal to `minDL = 5`.

3 Related Work

Given a set of point cloud data X in Euclidean space, the Vietoris-Rips complex (Definition 1) of X at radius ϵ, denoted $\mathcal{R}_\epsilon(X)$, is a triangulation of X in which any two of its points x_i, x_j forms an edge $[x_i x_j]$ when $||x_i - x_j|| \leq 2\epsilon$, and any three edges form a triangle, any four triangles form a tetrahedron, and so on. The nested sequences of subcomplexes formed when we increase the filtration radius ϵ over time forms the Vietoris-Rips filtration of X (Definition 2). Given a continuous and tame real-valued function f, the sublevel-set (Definition **??**) of f at level λ, denoted $\mathcal{S}_\lambda(f)$, is the set of all points in f whose value is at most λ. The nested sequence of sublevel-sets of f formed as we increase λ over time forms the sublevel-set filtration of f (Definition **??**). The birth and death radii of all holes in various dimensions that are formed during either Vietoris-Rips or sublevel-set filtration (i.e. the persistent homology of X or f) can be summarized as ordered pairs in a persistence diagram (Definition 3) or equivalently as intervals in a persistence barcode (Definition 4). One favorable result in persistent homology is its guaranteed theoretical stability to small data perturbations. That is, given two point clouds X and Y in Euclidean space, if the Hausdorff distance (Definition 5) between them is small, then the Bottleneck distance (Definition 8) between their persistence diagrams from Vietoris-Rips filtration will also be small [8, 9]. As well, given two tame and continuous functions f and g, if the infinity-norm between them is small, then the Bottleneck distance between their persistence diagrams from sublevel-set filtration will also be small [10]. These stability results are formally given below by

$$d_B(\mathrm{dgm}(\mathcal{R}(X)), \mathrm{dgm}(\mathcal{R}(Y))) \leq 2\, d_H(X, Y) \text{ and} \tag{1}$$

$$d_B(\mathrm{dgm}(\mathcal{S}(f)), \mathrm{dgm}(\mathcal{S}(g))) \leq 2||f - g||_\infty, \tag{2}$$

This guaranteed stability in part motivates the increasing usage of persistent homology for analysis of time series. In 2015, a persistence-based measure was introduced that quantifies how periodic a single univariate time series is [9]. Termed the Sliding Windows and 1-Persistence Scoring Function (denoted $\mathrm{score}(f)$), this measure was later used to quantify periodicity in gene expressions data [11]. This measure

works by first embedding a univariate time series f via and embedding dimension and a time lag to obtain a sliding windows embedding, denoted by

$$\text{SW}_{M,\tau} f(t) = (f(t), f(t+\tau), \ldots, f(t+M\tau)),$$

where $M \in \mathbb{N}$ and $\tau \in \mathbb{R}$ denote the dimension and time lag. Vietoris-Rips filtration is then performed on this embedding to obtain a persistence diagram $\text{dgm}(\text{SW}_{M,\tau} f(T))$. The normalized lifetime of the longest-surviving one dimensional hole, denoted $\text{mp}(\text{dgm}_1(\text{SW}_{M,\tau} f(t)))$ during this filtration is then used to quantify how periodic f is. The authors of [9] ultimately are able to use (1) to prove stability of $\text{score}(f)$ under sufficiently close enough Fourier series approximations of f, denoted $S_N f(t)$ (i.e. for sufficiently large enough degrees of truncation $N \in \mathbb{N}$). They are also able to prove that for sufficiently periodic time series data, the sliding windows embeddings will be roundest when the sliding window used has width proportional to the period of f. This latter finding could prove very useful when trying to determine the degree to which a frequency component appears in a signal with suspected seasonality (i.e. weather or climate data), which could in turn be useful for constructing such a persistence-based measure with guaranteed theoretical stability as a robust alternative to %DET and %REC. In the remainder of this paper, we introduce such a persistence-based measure, theoretical stability results for the same, and apply it on various synthetic series and real time series with suspected seasonality, ultimately showing that our measure consistently outperforms %DET and %REC.

4 Contributions and Organization

In this paper, we use the stability of persistent homology to construct a measure of time-series similarity with theoretical stability. We are further able to show its greater experimental stability in comparison with the commonly used measure of percent determinism on both synthetic and real data. Our novel measure has the additional advantage of only requiring one parameter for its computation when compared to the four required of percent determinism. Given a single time series f_1 and a reference sinusoidal f_2 with frequency w_2 Hertz, we first define a conditional periodicity score, $\text{score}(f_1|f_2)$, which can be used to quantify the extent to which the frequency w_2 appears in f_1. We compare this measure to spectral magnitudes produced via Fourier analysis on synthetic (Figure 2) and real (Figure 5) data. We then extend this notion in order to measure the similarity between a pair of series f_1 and f_2, where f_2 need not be a pure sinusoidal. Given an embedding dimension $M \in \mathbb{N}$, we define a conditional sliding window embedding of f_1 given f_2, which uses the estimated period of f_2 via Fourier analysis to determine the sliding window size when embedding f_1. Our measure is then computed by normalizing lifetime of the longest-surviving 1-dimensional hole from Vietoris-Rips filtration on this conditional embedding, which produces the conditional periodicity score $\text{score}(f_1|f_2)$. We then repeat this process to compute $\text{score}(f_2|f_1)$. Taking the average of both scores produces our novel similarity measure, the bi-conditional periodicity score $\text{score}(f_1, f_2)$. We are able to prove under which conditions $\text{score}(f_1|f_2)$ reduces to $\text{score}(f_1)$ (Proposition 12), as well as the stability of $\text{score}(f_1|f_2)$ to small time-series and frequency perturbations (Theorems 13 and 14). We then leverage these stability results to prove the stability of our bi-conditional periodicity score (Theorem 15). We are then able to prove the existence of a minimum embedding dimension which controls how close the conditional periodicity score is to its underlying value (Theorem 16). The latter result suggests that sufficiently large dimensions may be required to acheive desired convergence of $\text{score}(f_1|f_2)$, which motivates us to prove the stability of both $\text{score}(f_1|f_2)$ and $\text{score}(f_1, f_2)$ under dimension reduction (Theorem 17 and Corollary 18). We then construct an algorithm for computing the conditional periodicity score (Algorithm 1) and deduce its computational complexity. Following this, we use our algorithm to show that our bi-conditional periodicity score is experimentally more stable under small time-series perturbations on both synthetic and real climate data when compared to percent determinism (Figures 3 and 7).

The remainder of this paper is organized as follows. We begin with definitions to know regarding persistent homology, distance metrics for proving stability results, and our novel similarity measure (Section 5). We then provide our theoretical stability results in Section 6, followed by our experimental results in Section 7. Lastly, we provide a discussion of our results and future directions we'd like to take this project in Section 8

5 Definitions

5.1 Persistent Homology

Here we introduce standard definitions of topological features used in this paper. More information can be found in [12, 13].

Definition 1 (Vietoris-Rips Complex) *Given a point cloud* $X = \{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^d$ *and a filtration radius* $\epsilon > 0$, *the Vietoris-Rips complex of* X *at* ϵ *is given by the collection of simplices below*

$$\mathcal{R}_\epsilon(X) = \{\sigma = [x_i \ldots x_j] : i, j \in 1 \ldots n; ||x_s - x_k|| \le 2\epsilon, s, k \in 1 \ldots n\}.$$

Definition 2 (Vietoris-Rips Filtration) *Given a collection of increasing filtration radii* $0 = \epsilon_1 < \epsilon_2 < \cdots < \epsilon_r <$ *... applied to a point cloud* X, *the collection of nested subcomplexes*

$$X \subset \mathcal{R}_{\epsilon_2}(X) \subset \cdots \subset \mathcal{R}_{\epsilon_r}(X) \subset \ldots$$

denotes the Vietoris-Rips filtration of X, *given by* $\mathcal{R}(X)$.

Definition 3 (Persistence Diagram) *Given Vietoris-Rips filtration on a point cloud* X, *the collection of birth and death radii for each p-dimensional hole that appears can be summarized as ordered pairs in a persistence diagram given below by*

$$\mathrm{dgm}(\mathcal{R}(X)) = \{(b_i, d_i) : i = 1, \ldots, m_p, p = 0, 1, 2, \ldots\}.$$

Unless otherwise stated, we simplify this notation to $\mathrm{dgm}(X)$, *denoting* $\mathrm{dgm}_p(X)$ *as the collection of all ordered pairs for one dimension p from* $\mathrm{dgm}(X)$.

Definition 4 (Persistence Barcode) *Given Vietoris-Rips filtration on a point cloud* X, *the collection of birth and death radii for each p-dimensional hole that appears can be summarized as intervals in a persistence barcode given below by*

$$\mathrm{bcd}(\mathcal{R}(X)) = \{[b_i, d_i] : i = 1, \ldots, m_p, p = 0, 1, 2, \ldots\}.$$

Unless otherwise stated, we simplify this notation to $\mathrm{bcd}(X)$, *denoting* $\mathrm{bcd}_p(X)$ *as the collection of all intervals for one dimension p from* $\mathrm{bcd}(X)$.

5.2 Distance Metrics

Here we introduce standard definitions of distances used in this paper. See, for instance, the book by Burago, Bugaro, and Ivanov [14] for details.

Definition 5 (Hausdorff Distance) *Given two sets of points* X *and* Y *in a common metric space, the Hausdorff distance between them is given by*

$$\mathrm{d_H}(X, Y) = \inf\{\epsilon > 0 : X \subseteq Y^\epsilon, Y \subseteq X^\epsilon\}, \quad where \ X^\epsilon = \bigcup_{x \in X} B_\epsilon(x) \ and \ Y^\epsilon = \bigcup_{y \in Y} B_\epsilon(y)$$

denote the union of all ϵ*-balls centered at each point in either set.*

Definition 6 (Hausdorff Definition of Gromov-Hausdorff Distance) *Given two sets of points* X *and* Y, *the Gromov-Hausdorff distance between them is given by*

$$\mathrm{d_{GH}}(X, Y) = \inf\{\mathrm{d_H}(f(X), g(Y)) : f : X \to S, g : Y \to S\},$$

where f *and* g *are isometric embeddings of* X *and* Y *into a common metric space* S. *If* X *and* Y *lay in a shared metric space* S, *then* $\mathrm{d_{GH}}(X, Y) \le \mathrm{d_H}(X, Y)$ *[15]*.

Definition 7 (Distortion Definition of Gromov-Hausdorff Distance) *An alternative definition of the Gromov-Hausdorff distance between two sets of points* X *and* Y *is given by*

$$\mathrm{d_{GH}}(X, Y) = \frac{1}{2} \inf\{\mathrm{dis}(R) : R : X \to Y \in \mathcal{R}(X, Y)\},$$

where R is a relation between X and Y whose distortion is defined by

$$\mathrm{dis}(R) = \sup\{|d_X(x, x') - d_Y(y, y')| \; : \; (x, y), (x', y') \in R\},$$

where d_X and d_Y are the corresponding metrics for X and Y, respectively.

Definition 8 (Bottleneck Distance) *Given two finite sets of points X and Y, let $\mathrm{dgm}(X)$ and $\mathrm{dgm}(Y)$ denote the persistence diagrams of a chosen dimension obtained from the Vietoris-Rips (VR) filtration on X and Y, respectively. Then the Bottleneck distance between $\mathrm{dgm}(X)$ and $\mathrm{dgm}(Y)$ is given by*

$$\mathrm{d_B}(\mathrm{dgm}(X), \mathrm{dgm}(Y)) = \inf_{\phi} \sup_{x} \|x - \phi(x)\|_{\infty},$$

where $\phi : \mathrm{dgm}(X) \to \mathrm{dgm}(Y)$ denotes a bijection between $\mathrm{dgm}(X)$ and $\mathrm{dgm}(Y)$, including points along the diagonal in either diagram when they both do not share the same cardinality.

5.3 The Conditional Periodicity Scoring Function

Here we denote the definitions for our new scoring function.

Definition 9 (Conditional Sliding Windows Embedding) *Given a continuous univariate time series $f_1 : T \subseteq \mathbb{R} \to \mathbb{R}$ and a reference series f_2 with a frequency of $w_2 = \arg\max_w \left| \int_T f_2(t)e^{i2\pi wt}dt \right|$ Hertz (Hz) (i.e. cycles per second), define the time lag $\tau_{w_2} = \frac{1}{w_2(M+1)}$ for embedding dimension $M \in \mathbb{N}$. Then the conditional sliding windows embedding of f_1 given f_2 is given by*

$$\mathrm{SW}_{M, \tau_{w_2}} \, f_{1|2}(t) = (f_1(t), f_1(t + \tau_{w_2}), \ldots, f_1(t + M\tau_{w_2})).$$

Definition 10 (Conditional Periodicity Score) *Denote the lifetime of the longest-surviving 1-dimensional hole from Vietoris-Rips filtration on the conditional sliding windows embedding of a continuous time series f_1 given a reference series f_2 with frequency $w_2 = \arg\max_w \left| \int_T f_2(t)e^{i2\pi wt}dt \right|$ Hz by $\mathrm{mp}(\mathrm{dgm}_1(\mathrm{SW}_{M,\tau_{w_2}} \, f_{1|2}(T)))$. Then the conditional periodicity score of f_1 given f_2 is given by the normalization of this lifetime below:*

$$\mathrm{score}(f_1|f_2) = \frac{\mathrm{mp}(\mathrm{dgm}_1(\mathrm{SW}_{M,\tau_{w_2}} \, f_{1|2}(T)))}{\sqrt{3}}.$$

Definition 11 (Bi-Conditional Periodicity Score) *Given a continuous univariate time series $f_1 : T_1 \subseteq \mathbb{R} \to \mathbb{R}$ and a reference series $f_2 : T_2 \subseteq \mathbb{R} \to \mathbb{R}$, denote the frequency of each signal as $w_1 = \arg\max_w \left| \int_{T_1} f_1(t)e^{i2\pi wt}dt \right|$ and $w_2 = \arg\max_w \left| \int_{T_2} f_2(t)e^{i2\pi wt}dt \right|$. Given time lags $\tau_{w_1} = \frac{1}{w_1(M+1)}$ and $\tau_{w_2} = \frac{1}{w_2(M+1)}$, the bi-conditional periodicity score of f_1 and f_2 is given by*

$$\mathrm{score}(f_1, f_2) = \frac{\mathrm{score}(f_1|f_2) + \mathrm{score}(f_2|f_1)}{2}.$$

6 Theoretical Results

6.1 Reduction to the Periodicity Score

Here we discuss a result that clarifies when our conditional periodicity score $\mathrm{score}(f_1|f_2)$ simplifies to the periodicity score of f_1 as defined in [9, 11]. Specifically, as τ_{w_2} approaches $\tau_{w_1} = \frac{1}{w_1(M+1)}$, then $\mathrm{SW}_{M,\tau_{w_2}} \, f_{1|2}(t)$ approaches $\mathrm{SW}_{M,\tau_{w_1}} \, f_1(t)$ and hence $\mathrm{score}(f_1|f_2)$ approaches $\mathrm{score}(f_1)$. We formalize this notion in Proposition 12 below.

Proposition 12 (Reduction to Periodicity Score)

$$\lim_{\frac{T}{w_2} \to \frac{T}{w_1}^{-}} \mathrm{SW}_{M,\tau} \, f_{1|2}(t) = \mathrm{SW}_{M,\tau} \, f_1(t) \; for \; t \in [0, T].$$

Proof Let $i \in \{1, 2, \ldots, M+1\}$ for fixed $M \in \mathbb{N}$. Since f_1 is continuous and differentiable on $[0, 2\pi]$ and hence on $(0, 2\pi)$, it is continuous and differentiable on any subset $[a, b] \subset [0, 2\pi]$ and $(a, b) \subset (0, 2\pi)$. Let $t \in [0, 2\pi]$. Consider the subinterval $I_i = \left[t + \frac{2(i-1)\pi}{w_2(M+1)}, t + \frac{2(i-1)\pi}{w_1(M+1)}\right]$ of $[0, 2\pi]$. Then f_1 is also continuous and differentiable on I_i. By the Mean Value Theorem (MVT), there exists some $c_i \in I_i$ such that

$$f_1'(c_i)\left[\left(t + \frac{2(i-1)\pi}{w_1(M+1)}\right) - \left(t + \frac{2(i-1)\pi}{w_2(M+1)}\right)\right] = f_1\left(t + \frac{2(i-1)\pi}{w_1(M+1)}\right) - f_1\left(t + \frac{2(i-1)\pi}{w_2(M+1)}\right).$$

Then we get the following equality:

$$\left\|\mathrm{SW}_{M,\tau}\, f_{1|2}(t) - \mathrm{SW}_{M,\tau}\, f_1(t)\right\|_2$$

$$= \left(\sum_{i=1}^{M+1} \left|f_1\left(t + \frac{2(i-1)\pi}{w_2(M+1)}\right) - f_1\left(t + \frac{2(i-1)\pi}{w_1(M+1)}\right)\right|^2\right)^{\frac{1}{2}}$$

$$= \left(\sum_{i=1}^{M+1} |f_1'(c_i)|^2 \left|\left(t + \frac{2(i-1)\pi}{w_2(M+1)}\right) - \left(t + \frac{2(i-1)\pi}{w_1(M+1)}\right)\right|^2\right)^{\frac{1}{2}}$$

$$= \left(\sum_{i=2}^{M+1} |f_1'(c_i)|^2 \left|\frac{i-1}{M+1}\right|^2 \left|\frac{2\pi}{w_1} - \frac{2\pi}{w_2}\right|^2\right)^{\frac{1}{2}}. \qquad \square$$

A direct consequence of Proposition 12 is that $\mathrm{score}(f_1|f_2) \to \mathrm{score}(f_1)$ as $\frac{2\pi}{w_2} \to \frac{2\pi}{w_1}^-$.

6.2 Stability Under Time Series and Frequency Perturbations

We first prove that given a fixed time series f_1, small variations in the frequency of a reference sinusoidal f_2 guarantee small changes in the conditional periodicity score of f_1 given f_2, which intuitively quantifies the extent to which the frequency of f_2 is present within f_1. We prove this result formally in Theorem 13 below. We next show the opposite in Theorem 14. That is, we show that under small perturbations of the series f_1 the conditional periodicity score of f_1 given a reference sinusoidal f_2 at a fixed frequency of w_2 also undergoes small changes. We then combine these results to prove the stability of the bi-conditional periodicity score in Theorem 15.

Theorem 13 (Stability of the Conditional Periodicity Score Under Changes in Reference Frequency) *Let $f_1, f_{21}, f_{22} : \mathbb{R} \subseteq \mathbb{R} \to \mathbb{R}$ be three continuous and differentiable univariate time series with frequencies $w_1 = \arg\max_w \left|\int_T f_1(t)e^{i2\pi wt}dt\right|$, where f_{21} and f_{22} are reference sinusoidals with frequencies w_{21}, and w_{22}. Then*

$$|\mathrm{score}(f_1|f_{21}) - \mathrm{score}(f_1|f_{22})| \leq \frac{4}{\sqrt{3}}\sqrt{\frac{M}{M+1}} \left|\frac{1}{w_{21}} - \frac{1}{w_{22}}\right| \sqrt{\sum_{i=1}^{M} |f_1'(c_i)|^2}.$$

Proof Without loss of generality, suppose $w_{21} < w_{22}$. Let $M \in \mathbb{N}$. Assume $c_i \in I_i = (t + i\tau_{w_{22}}, t + i\tau_{w_{21}})$ for $i = 1, 2, \ldots, M$. Then

$$\|\mathrm{SW}_{M,\tau_{w_{21}}}\, f_{1|21}(t) - \mathrm{SW}_{M,\tau_{w_{22}}}\, f_{1|22}(t)\|^2 = \sum_{i=1}^{M} |f_1(t + i\tau_{w_{21}}) - f_1(t + i\tau_{w_{22}})|^2$$

$$= \sum_{i=1}^{M} |f_1'(c_i)|^2 |(t + i\tau_{w_{21}}) - (t + i\tau_{w_{22}})|^2 \text{ by Mean Value Theorem}$$

$$= \frac{M^2}{M+1} \left|\frac{1}{w_{21}} - \frac{1}{w_{22}}\right|^2 \sum_{I=1}^{M} |f_1'(c_i)|^2.$$

Then $\|\mathrm{SW}_{M,\tau_{w_{21}}}\, f_{1|21}(t) - \mathrm{SW}_{M,\tau_{w_{22}}}\, f_{1|22}(t)\| \leq \sqrt{\frac{M}{M+1}} \left|\frac{1}{w_{21}} - \frac{1}{w_{22}}\right| \sqrt{\sum_{I=1}^{M} |f_1'(c_i)|^2}$ ($*$). Let $\epsilon > (*)$. Then by definition,

$$d_H\left(\mathrm{SW}_{M,\tau_{w_{21}}}f_{1|21}(T),\mathrm{SW}_{M,\tau_{w_{22}}}f_{1|22}(T)\right) = \inf\{\gamma>0 : ||\mathrm{SW}_{M,\tau_{w_{21}}}f_{1|21}(t) - \mathrm{SW}_{M,\tau_{w_{22}}}f_{1|22}(T)||,$$
$$||\mathrm{SW}_{M,\tau_{w_{22}}}f_{1|22}(t) - \mathrm{SW}_{M,\tau_{w_{21}}}f_{1|21}(T)|| \le \gamma, \forall t \in T\}$$
$$\le (*).$$

Then for $T \subseteq \mathbb{R}$, by Equation 1, $d_B\left(\mathrm{dgm}_1(\mathrm{SW}_{M,\tau_{w_{21}}}f_{1|21}(T)), \mathrm{dgm}_1(\mathrm{SW}_{M,\tau_{w_{22}}}f_{1|22}(T))\right) \le 2(*)$. Let $(b_{21}^{\max}, d_{21}^{\max})$ and $(b_{22}^{\max}, d_{22}^{\max})$ be the birth and death radii of the longest-surviving 1-dimensional holes from Vietoris-Rips filtration on $\mathrm{SW}_{M,\tau_{w_{21}}}f_{1|21}(T)$ and $\mathrm{SW}_{M,\tau_{w_{22}}}f_{1|22}(T)$. Then

$$\left|\mathrm{mp}\left(\mathrm{dgm}_1\left(\mathrm{SW}_{M,\tau_{w_{21}}}f_{1|21}(T)\right)\right) - \mathrm{mp}\left(\mathrm{dgm}_1\left(\mathrm{SW}_{M,\tau_{w_{22}}}f_{1|22}(T)\right)\right)\right| = |(d_{21}^{\max} - b_{21}^{\max}) - (d_{22}^{\max} - b_{22}^{\max})|$$
$$\le |d_{21}^{\max} - d_{22}^{\max}| + |b_{22}^{\max} - b_{21}^{\max}|.$$

Let $\phi : \mathrm{dgm}_1\left(\mathrm{SW}_{M,\tau_{w_{21}}}f_{1|21}(T)\right) \to \mathrm{dgm}_1\left(\mathrm{SW}_{M,\tau_{w_{22}}}f_{1|22}(T)\right)$ be an injection given by $(b_{21}^i, d_{21}^i) \to (b_{22}^j, d_{22}^j)$. Then

$$d_B\left(\mathrm{dgm}_1\left(\mathrm{SW}_{M,\tau_{w_{21}}}f_{1|21}(T)\right), \mathrm{dgm}_1\left(\mathrm{SW}_{M,\tau_{w_{22}}}f_{1|22}(T)\right)\right) = \inf_\phi \sup_i ||(b_{21}^i, d_{21}^i) - \phi((b_{21}^i, d_{21}^i))||_\infty$$
$$= \inf_\phi \sup_i \max\left\{|b_{21}^i - b_{22}^j|, |d_{21}^i - d_{22}^j|\right\}$$
$$\ge |b_{21}^{\max} - b_{22}^{\max}|, |d_{21}^{\max} - d_{22}^{\max}|.$$

Hence

$$|\mathrm{score}(f_1|f_{21}) - \mathrm{score}(f_1|f_{22})| = \left|\mathrm{mp}\left(\mathrm{dgm}_1\left(\mathrm{SW}_{M,\tau_{w_{21}}}f_{1|21}(T)\right)\right) - \mathrm{mp}\left(\mathrm{dgm}_1\left(\mathrm{SW}_{M,\tau_{w_{22}}}f_{1|22}(T)\right)\right)\right|$$
$$\le 2\,d_B\left(\mathrm{dgm}_1\left(\mathrm{SW}_{M,\tau_{w_{21}}}f_{1|21}(T)\right), \mathrm{dgm}_1\left(\mathrm{SW}_{M,\tau_{w_{22}}}f_{1|22}(T)\right)\right)$$
$$\le 4\,d_H\left(\mathrm{SW}_{M,\tau_{w_{21}}}f_{1|21}(T), \mathrm{SW}_{M,\tau_{w_{22}}}f_{1|22}(T)\right)$$
$$\le \frac{4}{\sqrt{3}}\sqrt{\frac{M}{M+1}}\left|\frac{1}{w_{21}} - \frac{1}{w_{22}}\right|\sqrt{\sum_{i=1}^M |f_1'(c_i)|^2}.$$
$\square$

Theorem 14 (Stability of Conditional Periodicity Score Under Changes in Time Series) *Let $f_{11}, f_{12} : \mathbb{R} \to \mathbb{R}$ be two continuous and differentiable univariate time series and f_2 be a reference sinusoidal with frequency w_2 Hz. Then*

$$|\mathrm{score}(f_{11}|f_2) - \mathrm{score}(f_{12}|f_2)| \le 4||f_{11} - f_{12}||_\infty \sqrt{\frac{M}{3}}.$$

Proof Let $i = 1, 2, \ldots, M$. Because for all $t \in \mathbb{R}$, $t + i\tau_{w_2}$ is also a real number, we have the following

$$||\mathrm{SW}_{M,\tau_{w_2}}f_{11|2}(t) - \mathrm{SW}_{M,\tau_{w_2}}f_{12|2}(t)||^2 = \sum_{i=1}^M |f_{11}(t + i\tau_{w_2}) - f_{12}(t + i\tau_{w_2})|^2$$
$$\le \sum_{i=1}^M \left(\max_{t\in\mathbb{R}}\left\{|f_{11}(t + i\tau_{w_2}) - f_{12}(t + i\tau_{w_2})|\right\}\right)^2$$
$$= \sum_{i=1}^M ||f_{11} - f_{12}||_\infty^2$$
$$= M||f_{11} - f_{12}||_\infty^2.$$

Then $||\mathrm{SW}_{M,\tau_{w_2}}f_{11|2}(t) - \mathrm{SW}_{M,\tau_{w_2}}f_{12|2}(t)|| \le \sqrt{M}||f_{11} - f_{12}||_\infty$. Then by similar arguments as in the proof of Theorem 13, for some $T \subseteq \mathbb{R}$ we have that

$$d_H(\mathrm{SW}_{M,\tau_{w_2}} f_{11|2}(T), \mathrm{SW}_{M,\tau_{w_2}} f_{12|2}(T)) \leq \sqrt{M}||f_{11} - f_{12}||_\infty \text{ and}$$

$$\left| \mathrm{mp}\left(\mathrm{dgm}_1(\mathrm{SW}_{M,\tau_{w_2}} f_{11|2}(T)) \right) - \mathrm{mp}\left(\mathrm{dgm}_1(\mathrm{SW}_{M,\tau_{w_2}} f_{12|2}(T)) \right) \right|$$

$$\leq 2\, d_B\left(\mathrm{dgm}_1(\mathrm{SW}_{M,\tau_{w_2}} f_{11|2}(T)), \mathrm{dgm}_1(\mathrm{SW}_{M,\tau_{w_2}} f_{12|2}(T)) \right)$$

Therefore, by Equation 1 we have that

$$|\mathrm{score}(f_{11}|f_2) - \mathrm{score}(f_{12}|f_2)| = \frac{1}{\sqrt{3}} \left| \mathrm{mp}\left(\mathrm{dgm}_1(\mathrm{SW}_{M,\tau_{w_2}} f_{11|2}(T)) \right) - \mathrm{mp}\left(\mathrm{dgm}_1(\mathrm{SW}_{M,\tau_{w_2}} f_{12|2}(T)) \right) \right|$$

$$\leq \frac{2}{\sqrt{3}}\, d_B\left(\mathrm{dgm}_1(\mathrm{SW}_{M,\tau_{w_2}} f_{11|2}(T)), \mathrm{dgm}_1(\mathrm{SW}_{M,\tau_{w_2}} f_{12|2}(T)) \right)$$

$$\leq \frac{4}{\sqrt{3}}\, d_H(\mathrm{SW}_{M,\tau_{w_2}} f_{11|2}(T), \mathrm{SW}_{M,\tau_{w_2}} f_{12|2}(T))$$

$$\leq 4||f_{11} - f_{12}||_\infty \sqrt{\frac{M}{3}}.$$

$\square$

Theorem 15 (Stability of the Bi-Conditional Periodicity Score) *Let $f_1, f_{11}, f_{12}, f_2, f_{21}, f_{22} : \mathbb{R} \to \mathbb{R}$ be three continuous and differentiable univariate time series with frequencies given by $w_k = \arg\max_w \left| \int_\mathbb{R} f_k(t) e^{i2\pi wt} dt \right|$ for $k = 1, 11, 12, 2, 21, 22$. Without loss of generality, suppose that $w_{21} < w_{22}$ and $w_{11} < w_{12}$. Then given some $c_i \in I_i = (t + i\tau_{w_{22}}, t + i\tau_{w_{21}})$ and $d_i \in J_i = (t + i\tau_{w_{12}}, t + i\tau_{w_{11}})$, we have that*

$$|\mathrm{score}(f_1, f_{21}) - \mathrm{score}(f_1, f_{22})| \leq \frac{2}{\sqrt{3}} \left(\sqrt{\frac{M}{M+1}} \left| \frac{1}{w_{21}} - \frac{1}{w_{22}} \right| \sqrt{\sum_{i=1}^M |f_1'(c_i)|^2} + ||f_{21} - f_{22}||_\infty \sqrt{M} \right) \quad and$$

$$|\mathrm{score}(f_{11}, f_2) - \mathrm{score}(f_{12}, f_2)| \leq \frac{2}{\sqrt{3}} \left(\sqrt{\frac{M}{M+1}} \left| \frac{1}{w_{11}} - \frac{1}{w_{12}} \right| \sqrt{\sum_{i=1}^M |f_2'(d_i)|^2} + \sqrt{M}||f_{11} - f_{12}||_\infty \right).$$

Proof By Theorem 13,

$$|\mathrm{score}(f_1|f_{21}) - \mathrm{score}(f_1|f_{22})| \leq \frac{4}{\sqrt{3}} \sqrt{\frac{M}{M+1}} \left| \frac{1}{w_{21}} - \frac{1}{w_{22}} \right| \sqrt{\sum_{i=1}^M |f_1'(c_i)|^2} \quad \text{and}$$

$$|\mathrm{score}(f_2|f_{11}) - \mathrm{score}(f_2|f_{12})| \leq \frac{4}{\sqrt{3}} \sqrt{\frac{M}{M+1}} \left| \frac{1}{w_{11}} - \frac{1}{w_{12}} \right| \sqrt{\sum_{i=1}^M |f_2'(d_i)|^2}.$$

By Theorem 14,

$$|\mathrm{score}(f_{21}|f_1) - \mathrm{score}(f_{22}|f_1)| \leq 4||f_{21} - f_{22}||_\infty \sqrt{\frac{M}{3}} \quad \text{and}$$

$$|\mathrm{score}(f_{11}|f_2) - \mathrm{score}(f_{12}|f_2)| \leq 4||f_{11} - f_{12}||_\infty \sqrt{\frac{M}{3}}.$$

Then we have that

$$|\mathrm{score}(f_1, f_{21}) - \mathrm{score}(f_1, f_{22})| = \frac{1}{2} |(\mathrm{score}(f_1|f_{21}) + \mathrm{score}(f_{21}|f_1)) - (\mathrm{score}(f_1|f_{22}) + \mathrm{score}(f_{22}|f_1))|$$

$$\leq \frac{1}{2} (|\mathrm{score}(f_1|f_{21}) - \mathrm{score}(f_1|f_{22})| + |\mathrm{score}(f_{21}|f_1) - \mathrm{score}(f_{22}|f_1)|)$$

$$\leq \frac{1}{2} \left(\frac{4}{\sqrt{3}} \sqrt{\frac{M}{M+1}} \left| \frac{1}{w_{21}} - \frac{1}{w_{22}} \right| \sqrt{\sum_{i=1}^M f_1'(c_i)|^2} + 4||f_{21} - f_{22}||_\infty \sqrt{\frac{M}{3}} \right).$$

A similar argument proves that

$$\left|\operatorname{score}(f_{11}, f_2) - \operatorname{score}(f_{12}, f_2)\right| \leq \frac{1}{2}\left(4\|f_{11} - f_{12}\|_\infty \sqrt{\frac{M}{3}} + \frac{4}{\sqrt{3}}\sqrt{\frac{M}{M+1}}\left|\frac{1}{w_{11}} - \frac{1}{w_{12}}\right|\sqrt{\sum_{i=1}^{M}|f_2'(d_i)|^2}\right).$$

$\square$

6.3 A Minimum Embedding Dimension

For the conditional periodicity score, the only parameter required is the embedding dimension $M \in \mathbb{N}$. We are hence motivated to identify a minimum dimension required to optimize the performance of $\operatorname{score}(f_1|f_2)$, which we formalize in Theorem 16 below.

Theorem 16 (A Minimum Embedding Dimension) *Let $\epsilon > 0$. Suppose $f_1, f_2 : \mathbb{R} \to \mathbb{R}$ are continuous and differentiable univariate time series and the frequency of f_2 is given by $w_2 = \arg\max_w \left|\int_{\mathbb{R}} f_2(t)e^{i2\pi wt}dt\right|$. Then, any embedding dimension $M_2 > M_1 \geq \mathcal{M} \in \mathbb{N}$ for $\mathcal{M} = \left\lceil\frac{1}{w_2\epsilon}\right\rceil$ guarantees that the change in respective conditional periodicity scores $\operatorname{score}_{M_2}(f_1|f_2)$ and $\operatorname{score}_{M_1}(f_1|f_2)$ are linearly bounded above by ϵ.*

Proof We first show that $\tau(M) = \frac{1}{w_2(M+1)}$ is a Cauchy sequence of M. By the definition of $\mathcal{M}$, $\mathcal{M} \geq \frac{1}{w_2\epsilon}$ and hence $\frac{1}{w_2\mathcal{M}} \leq \epsilon \implies \frac{1}{w_2}\left(\frac{1}{\mathcal{M}}\right) \leq \epsilon$. Since $M_2 > M_1 \geq \mathcal{M}$, $M_2 + 1 > M_1 + 1 \geq \mathcal{M} + 1$ and hence $\frac{1}{M_2+1} < \frac{1}{M_1+1} \leq \frac{1}{\mathcal{M}+1} < \frac{1}{\mathcal{M}}$. Thus we get

$$\begin{aligned}
|\tau(M_1) - \tau(M_2)| &= \frac{1}{w_2}\left|\frac{1}{M_1+1} - \frac{1}{M_2+1}\right| \\
&= \frac{1}{w_2}\left(\frac{1}{M_1+1} - \frac{1}{M_2+1}\right) \\
&\leq \frac{1}{w_2}\left(\frac{1}{\mathcal{M}+1} - \frac{1}{M_2+1}\right) \\
&< \frac{1}{w_2}\left(\frac{1}{\mathcal{M}}\right) \leq \epsilon.
\end{aligned} \tag{3}$$

Define the zero-padded conditional sliding windows embedding of f_1 given f_2 for the smaller embedding dimension M_1 by $\operatorname{SW}_{M_1, \tau(M_1)} f_{1|2}(t) = \left(f_1(t), \ldots, f_1(t + M_1\tau(M_1)), \underbrace{0, \ldots, 0}_{M_2 - M_1}\right)^T$. Then for $i = 1, \ldots, M_1 + 1$, by the MVT there exists some $c_i \in I_i = \left(t + (i-1)\tau(M_2), t + (i-1)\tau(M_1)\right)$ such that

$$\left|f_1'(c_i)\right|\left|(i-1)\left(\tau(M_1) - \tau(M_2)\right)\right| = \left|f_1\left(t + (i-1)\tau(M_1)\right) - f_1\left(t + (i-1)\tau(M_2)\right)\right|. \tag{4}$$

$$\tag{5}$$

Applying the results in Equations (3) and (4), we obtain the following relation.

$$\begin{aligned}
&\left\|\operatorname{SW}_{M_1, \tau(M_1)} f_{1|2}(t) - \operatorname{SW}_{M_2, \tau(M_2)} f_{1|2}(t)\right\|_2^2 \\
&= \sum_{i=1}^{M_1+1}\left|f_1\left(t + (i-1)\tau(M_1)\right) - f_1\left(t + (i-1)\tau(M_2)\right)\right|^2 + \sum_{M_1+1}^{M_2+1}\left|f_1\left(t + (i-1)\tau(M_2)\right)\right|^2 \\
&= \sum_{i=1}^{M_1+1}|f_1'(c_i)|^2|i-1|^2|\tau(M_1) - \tau(M_2)|^2 \quad + \quad \sum_{M_1+1}^{M_2+1}\left|f_1\left(t + (i-1)\tau(M_2)\right)\right|^2 \\
&\leq \epsilon^2(M_1+1)^3\sqrt{\sum_{i=1}^{M_1+1}|f_1'(c_i)|^2} \quad + \quad \sum_{M_1+1}^{M_2+1}\left|f_1\left(t + (i-1)\tau(M_2)\right)\right|^2 \\
&= \epsilon^2 \cdot g(M_1, f_1) \quad + \quad h(M_1, M_2, f_1) \\
&\leq \left(\epsilon \cdot \sqrt{g(M_1, f_1)} + \sqrt{h(M_1, M_2, f_1)}\right)^2,
\end{aligned}$$

since $g(M_1, f_1)$ and $h(M_1, M_2, f_1)$ are nonnegative constants for any given comparison, and $\epsilon > 0$.

Then $\mathrm{d_H}\left(\mathrm{SW}_{M_1,\tau(M_1)} f_{1|2}(t), \mathrm{SW}_{M_2,\tau(M_2)} f_{1|2}(t)\right) \leq \epsilon \cdot \sqrt{g(M_1, f_1)} + \sqrt{h(M_1, M_2, f_1)}$,

which implies that

$$\mathrm{d_B}\left(\mathrm{dgm}_1(\mathrm{SW}_{M_1,\tau(M_1)} f_{1|2}(t)), \mathrm{dgm}_1(\mathrm{SW}_{M_2,\tau(M_2)} f_{1|2}(t))\right) \leq \epsilon \cdot 2\sqrt{g(M_1, f_1)} + 2\sqrt{h(M_1, M_2, f_1)}.$$

By a similar argument as in the proof of Theorem 13, this means that

$$\left| \mathrm{mp}\left(\mathrm{dgm}_1(\mathrm{SW}_{M_1,\tau(M_1)} f_{1|2}(t))\right) - \mathrm{mp}\left(\mathrm{dgm}_1(\mathrm{SW}_{M_2,\tau(M_2)} f_{1|2}(t))\right)\right|$$
$$\leq \epsilon \cdot 4\sqrt{g(M_1, f_1)} + 4\sqrt{h(M_1, M_2, f_1)}$$
$$\implies \quad |\mathrm{score}_{M_1}(f_1|f_2) - \mathrm{score}_{M_2}(f_1|f_2)| \leq \frac{4}{\sqrt{3}}\left(\epsilon \cdot \sqrt{g(M_1, f_1)} + \sqrt{h(M_1, M_2, f_1)}\right).$$

Hence the difference between the scores under M_1 and M_2 is bounded above by a linear function of ϵ with positive slope. Therefore, the smaller the ϵ-precision is, the closer these scores are to each other. $\qquad\square$

6.4 Stability Under Dimension Reduction

As shown by Theorem 16, for sufficiently high enough precision of our observed score $\mathrm{score}_M(f_1|f_2)$ to the true underlying score $\mathrm{score}(f_1|f_2)$ (i.e. for sufficiently small enough $\epsilon > 0$), larger embedding dimensions are required. For sufficiently large embedding dimensions, pairwise distances between points have been shown to concentrate [16]. As well, the bottleneck when using the periodicity score in practice is the computation of the dimension 1 persistence diagram of the Vietoris-Rips filtration of the SWE. In general, for a point cloud X with N points, the computation of $\mathrm{dgm}_1(X)$ runs in $O(N^6)$ time, although faster approaches may be available in lower dimensions [17]. Both of these limitations in the high-dimensional setting inspire us to prove the stability of our scoring function under dimension reduction of the conditional sliding windows embedding. That is, given a sufficiently large enough dimension M to reach this desired precision, so long as the top principle components capture a majority of the variance in the conditional embedding upon orthogonal projection, the conditional periodicity score won't change much. We hence study the stability of our conditional and bi-conditional periodicity scores under principal component analysis (PCA), a widely used dimension reduction technique [18].

Theorem 17 (Stability of Conditional Periodicity Score Under Dimension Reduction) *Let $K \leq M+1$ for $K \in \mathbb{N}$. Suppose $f_1, f_2 : \mathbb{R} \to \mathbb{R}$ are continuous and differentiable univariate time series. For $T \subseteq \mathbb{R}$, define the orthogonal projection of $X = \mathrm{SW}_{M,\tau_{w_2}} f_{1|2}(T)$ onto its top K principal components by $\phi(X) = (\langle \mathbf{c}_1, X\rangle, \ldots, \langle \mathbf{c}_K, X\rangle)^T$ for orthonormal eigenvectors and corresponding eigenvalues $\{\mathbf{c}_k, \lambda_k\}_{k=1}^N$ produced by PCA. Suppose X contains $N \in \mathbb{N}$ points. Denote the conditional periodicity score under ϕ, $\mathrm{score}_\phi(f_1|f_2)$, as the normalized maximum 1-d persistence from VR-filtration on $\phi(X)$. Then*

$$|\mathrm{score}(f_1|f_2) - \mathrm{score}_\phi(f_1|f_2)| \leq \sqrt{\frac{8}{3}} \sqrt[4]{\sum_{i=K+1}^N \lambda_i^2}.$$

Proof Our proof is inspired by methods used in [19]. Notice that ϕ is a relation in $\mathcal{R}(X, Y)$ for $Y = \phi(X)$. Then by Definition 7 we have that $\mathrm{d_{GH}}(X, Y) \leq \frac{1}{2}\mathrm{dis}(\phi)$, where $\mathrm{dis}(\phi)^2 = \|D_X - D_Y\|_{\max}^2$ with D_X being the matrix of pairwise distances in X. We get that [19, Lemma 3.9]

$$\|D_X - D_Y\|_{\max}^2 \leq \left\|D_X^{\circ 2} - D_Y^{\circ 2}\right\|_{\max},$$

where $D_X^{\circ 2}$ denotes the matrix of squared pairwise Euclidean distances in X.

Recall that for a matrix A, $\|A\|_{\max}^2 = \left(\max_{ij}|A_{ij}|\right)^2 = \max_{ij}|A_{ij}|^2 \leq \sum_{ij}|A_{ij}|^2 = \|A\|_F^2$, where $\|\cdot\|_F$ denotes the Frobenius norm. Then we have

$$\left\|D_X^{\circ 2} - D_Y^{\circ 2}\right\|_{\max} \leq \left\|D_X^{\circ 2} - D_Y^{\circ 2}\right\|_F.$$

Define the eigendecomposition of the covariance of X as $XX^T = Q\Lambda Q^T$, where $Q = [\mathbf{c}_1, \ldots, \mathbf{c}_N]$ is the $N \times N$ matrix of orthonormal eigenvectors and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$ is the $N \times N$ diagonal matrix of eigenvalues corresponding to the orthogonal projection of X. Then the eigendecomposition of the K-dimensional

subspace containing the top K principal components of X can be defined by $YY^T = Q\Lambda|_K Q^T$, where $\Lambda|_K = \text{diag}(\lambda_1, \ldots, \lambda_K, \underbrace{0, \ldots, 0}_{N-K})$. Then, if we center the squared distances in $D_X^{\circ 2}$ and $D_Y^{\circ 2}$, we obtain the relations

$$XX^T = -\frac{1}{2}C_N D_X^{\circ 2} C_N \text{ and}$$

$$YY^T = -\frac{1}{2}C_N D_Y^{\circ 2} C_N,$$

where C_N is the $N \times N$ centering matrix with diagonal entries $1 - \frac{1}{N}$ and off-diagonal entries $-\frac{1}{N}$ [20]. Then $D_X^{\circ 2} = -2C_N(Q\Lambda Q^T)C_N$ and $D_Y^{\circ 2} = -2C_N(Q\Lambda|_K Q^T)C_N$. Since $Q^T Q = QQ^T = I$, we obtain that

$$\left\| D_X^{\circ 2} - D_Y^{\circ 2} \right\|_F = 2 \left\| C_N(Q\Lambda Q^T - Q\Lambda|_K Q^T)C_N \right\|_F$$

$$= 2 \left\| Q(\Lambda - \Lambda|_K)Q^T \right\|_F \quad \text{(since } C_N \approx I \text{ for sufficiently large } N)$$

$$= 2\sqrt{\text{tr}(Q(\Lambda - \Lambda|_K)^2 Q^T)}$$

$$= 2\sqrt{\text{tr}((\Lambda - \Lambda|_K)^2)} \quad \text{(since trace is cyclically invariant)}$$

$$= 2\sqrt{\sum_{i=K+1}^{N} \lambda_i^2}.$$

Thus we get that $\text{dis}(\phi) \leq \sqrt{2} \sqrt[4]{\sum_{i=K+1}^{N} \lambda_i^2}$ and hence $\text{d}_{\text{GH}}(X, Y) \leq \frac{\sqrt{2}}{2} \sqrt[4]{\sum_{i=K+1}^{N} \lambda_i^2}$. Then by Equation 1, we get that

$$\text{d}_B(\text{dgm}_1(X), \text{dgm}_1(Y)) \leq 2\,\text{d}_{\text{GH}}(X, Y) \leq \sqrt{2}\sqrt[4]{\sum_{i=K+1}^{N} \lambda_i^2}.$$

Hence by similar arguments as in the proof of Theorems 13 and 14, we have that

$$|\text{score}(f_1|f_2) - \text{score}_\phi(f_1|f_2)| = \frac{1}{\sqrt{3}} |\text{mp}(\text{dgm}_1(X)) - \text{mp}(\text{dgm}_1(\phi(X)))|$$

$$\leq \frac{2}{\sqrt{3}} \text{d}_B(\text{dgm}_1(X), \text{dgm}_1(\phi(X)))$$

$$\leq 2\sqrt{\frac{2}{3}} \sqrt[4]{\sum_{i=K+1}^{N} \lambda_i^2}$$

$$= \sqrt{\frac{8}{3}} \sqrt[4]{\sum_{i=K+1}^{N} \lambda_i^2}. \square$$

Corollary 18 (Stability of Bi-Conditional Periodicity Score Under Dimension Reduction) *Let $K \leq M+1$ for $K \in \mathbb{N}$. Suppose $f_1, f_2 : \mathbb{R} \to \mathbb{R}$ are continuous and differentiable univariate time seies. For $T \subseteq \mathbb{R}$, define the eigenvalues obtained from the orthogonal projection (ϕ) of $\text{SW}_{M,\tau_{w_2}} f_{1|2}(T)$ and $\text{SW}_{M,\tau_{w_2}} f_{2|1}(T)$ onto their top K principal components respectively by $\{\lambda_i\}_{i=1}^{N}$ and $\{\gamma_i\}_{i=1}^{N}$. Define the bi-conditional periodicity score of f_1 and f_2 under this orthogonal projection by $\text{score}_\phi(f_1, f_2)$. Then*

$$|\text{score}(f_1, f_2) - \text{score}_\phi(f_1, f_2)| \leq \sqrt{\frac{2}{3}}\left(\sqrt[4]{\sum_{i=K+1}^{N} \lambda_i^2} + \sqrt[4]{\sum_{i=K+1}^{N} \gamma_i^2} \right).$$

Proof By Theorem 17,

$$|\text{score}(f_1, f_2) - \text{score}_\phi(f_1, f_2)| = \frac{1}{2}\left| \big(\text{score}(f_1|f_2) + \text{score}(f_2|f_1)\big) - \big(\text{score}_\phi(f_1|f_2) + \text{score}_\phi(f_2|f_1)\big) \right|$$

$$\leq \frac{1}{2}\big(|\text{score}(f_1|f_2) - \text{score}_\phi(f_1|f_2)| + |\text{score}(f_2|f_1) - \text{score}_\phi(f_2|f_1)| \big)$$

$$\leq \frac{1}{2}\sqrt{\frac{8}{3}}\left(\sqrt[4]{\sum_{i=K+1}^{N} \lambda_i^2} + \sqrt[4]{\sum_{i=K+1}^{N} \gamma_i^2} \right)$$

$$= \sqrt{\frac{2}{3}} \left(\sqrt[4]{\sum_{i=K+1}^{N} \lambda_i^2} + \sqrt[4]{\sum_{i=K+1}^{N} \gamma_i^2} \right).$$

□

7 Experimental Results

7.1 An Algorithm for the Conditional Periodicity Score

We introduce a procedure for computing the conditional periodicity score of a discrete time series $f_i(T)$ given another $f_j(T)$, for $T = \{t_0, t_1, \ldots, t_{N-1}\}_{N \in \mathbb{N}}$. We then estimate the frequency of f_j using the arg maximum of the spectral magnitude produced via the discrete fast Fourier transform. We particularly compute the discrete fast Fourier spectra of f_j, which we denote below by the absolute value of the dot product of the discrete signal with a pure sinusoidal with frequency k cycles every N time steps:

$$|DFT[f_j](k)| = \frac{1}{N} \left| \sum_{n=0}^{N-1} (f_j)_n \cdot e^{-i2\pi \cdot \frac{k}{N} \cdot n} \right| = \frac{1}{N} \left| \sum_{n=0}^{N-1} (f_j)_n \cos\left(2\pi\frac{k}{N}n\right) - i \sum_{n=0}^{N-1} (f_j)_n \sin\left(2\pi\frac{k}{N}n\right) \right|$$

$$= \frac{1}{N} \left(\sum_{n=0}^{N-1} (f_j)_n \cos\left(2\pi\frac{k}{N}n\right) \right)^2 + \frac{1}{N} \left(\sum_{n=0}^{N-1} (f_j)_n \sin\left(2\pi\frac{k}{N}n\right) \right)^2.$$

We then estimate w_j as the frequency corresponding to the maximum spectral density of f_2, denoted by $w_j = \arg\max_k |DFT[f_j](k)|$. We then use this estimate to define the time lag as $\tau_{w_j} = \frac{1}{w_j(M+1)}$ given an embedding dimension $M \in \mathbb{N}$. We then include an option for computing the top $K \in N$ principal components of the conditional sliding windows embedding $\mathrm{SW}_{M, \tau_{w_j}} f_{i|j}(T)$ under an orthogonal projection $\phi : \mathbb{R}^{M+1} \to K$. We finally perform Vietoris-Rips filtration on the conditional sliding windows embedding, obtaining the conditional periodicity score.

Algorithm 1 Procedure for computing $\mathrm{score}(f_i|f_j)$

Inputs: An embedding dimension $M \in \mathbb{N}$, two discrete univariate time series $\{f_i(T), f_j(T)\}$ for $T = \{t_0, t_1, \ldots, t_{N-1}\}_{N \in \mathbb{N}}$, PCA $\in \mathbb{R}$. **Estimate** w_j via spectral analysis using the (Discrete) Fast Fourier Transform up to the Nyquist frequency:

$$w_j = \underset{k=0,\ldots,N/2-1}{\arg\max} |DFT[f_j](k)|$$

Define $\tau_{w_j} = \left\lfloor \frac{1}{w_j(M+1)} \right\rfloor$.
Compute $X = \mathrm{SW}_{M, \tau_{w_j}} f_{i|j}(T)$.
If PCA $= 1$,
    For $K = \min\{2, M+1\}$,
        Compute $\{\phi(X)\}$ for orthogonal projection ϕ onto the first K principle components of X.
        Define $Y = \phi(X)$.
Define $Y = \phi(X)$.
Center and normalize Y to obtain Z.
Compute $\mathrm{dgm}_1(Z)$ from Vietoris-Rips filtration on Z and extract $\mathrm{mp}(\mathrm{dgm}_1(Z))$ from $\mathrm{dgm}_1(Z)$.
Compute $\mathrm{score}(f_i|f_j)$ using $\mathrm{mp}(\mathrm{dgm}_1(Z))$.

Return: $\mathrm{score}(f_i|f_j)$.

Remark 19 (Computational Complexity of the Bi-Conditional Periodicity Score) *Algorithm 1 runs in* $O\left(N \log N + PK^2 + P^6\right)$ *time, where N is the number of points in the discrete univariate input signals f_i and f_j, $P = N - M\tau_{w_j}$ is the number of points in the conditional sliding windows embedding of f_i given f_j, and $K < M + 1$ is the number of principal components computed from the conditional sliding windows embedding for embedding dimension $M \in \mathbb{N}$. The discrete Fast Fourier Transform on f_j can be computed in $O(N \log N)$ time [21]. The PCA computations run in $O(PK^2)$ time [18, 22]. The bottleneck step is usually the computation of the 1D persistence diagram using the VR filtration, which runs in $O(P^6)$ time [17].*

Notice that if we don't perform dimension reduction on the conditional embedding, then computing score$(f_i|f_j)$ *takes $O(N \log N + P^6)$ time. Hence with PCA, the bi-conditional periodicity score,* score(f_i, f_j), *requires $O(N \log N + PK^2 + P^6)$ computations, and without PCA requires only $O(N \log N + P^6)$.*

7.2 Synthetic Data

7.2.1 Stability of the Conditional Periodicity Score

For our first experiment, we assess the stability of our conditional periodicity score when quantifying the extent to which varied frequencies w_2 are present in a time series f_1 with varied levels of added Gaussian noise. We further compare this stability to the Fourier Spectrum of the noisy signal f_1 as well as percent recurrence. We fix $f_1(t) = \sin(2\pi w_1 T)$ for $T = \{0, 1, \ldots, 299\}$ and frequency $w_1 = 8/300$ Hz. We further define the noisy signal as $f_1^\sigma(T) = f_1(T) + \mathcal{N}(0, \sigma * \mathrm{sd}(f_1))$, where $\mathrm{sd}(f_1)$ denotes the standard deviation of f_1. We then vary the frequency w_2 of the reference sinusoidal f_2 among the values $w_2 = \{k/300\}$ for $k = 5, 5.25, \ldots, 13$. Additionally, we vary the standard deviation of Gaussian noise among the values $\sigma = \{0.01, 0.05, 0.1, 0.15\}$ (i.e. 1%, 5%, 10%, and 15%). For computing %REC and score$(f_1|f_2)$, we fix the embedding dimension to $M = 1$. To compute %REC, we additionally fix the time lag to $\tau = 10$ by taking the first minimum of the automutual information of f_1 [7]. We finally fix the distance threshold to denote a cross-recurrence as $\mathtt{tol} = 0.1 \cdot \max(\|\mathrm{SW}_{M,\tau} f_1(T) - \mathrm{SW}_{M,\tau} f_2(T)\|)$ (i.e. 10% of the maximum distance between the embeddings of f_1 and the reference sinusoidal). We finally define our conditional periodicity score as

$$\mathrm{score}(f_1|f_2) = 1 - \frac{\mathrm{mp}\left(\mathrm{dgm}_1(\mathrm{SW}_{M,\tau_{w_2}} f_{1|2}(T)\right)}{\sqrt{3}}.$$

See our results in Figure 2 below. As expected, each measure shows a peak around the underlying frequency of f_1 for 0% noise. Additionally, our measure and the Fourier magnitude are much more robust to added Gaussian noise than %REC is. The stability of our conditional periodicity score and the spectral magnitude motivates us to use this magnitude to compute a stable measure of time series correlation, which we discuss below.
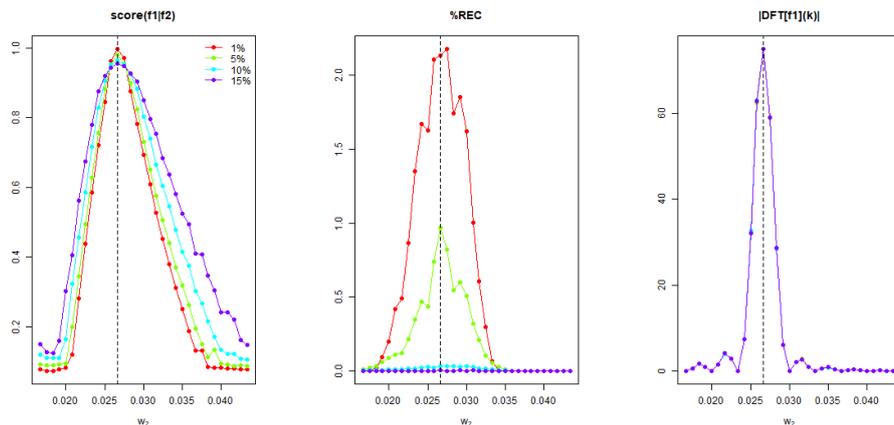

Fig. 2 Given a time series f_1 of $N = 300$ points, we compare the stability of score$(f_1|f_2)$, $|\mathrm{REC}[f_1](k)|$ for $w_2 = k/N$, and %REC$[f_1, e^{i2\pi w_2 t}]$ when subject to added gaussian noise to f_1. We include a vertical dashed line at the underlying frequency w_1.

7.3 Stability of the Bi-Conditional Periodicity Score

In this experiment, we compare the stability of $score(f_1, f_2)$ as computed via Algorithm 1 to %REC against added Gaussian noise to both f_1 and f_2. We again fix f_1 as before and vary the frequency of f_2 between the same values of w_2 as previously mentioned. We again show the average of each measure over 20 samples of each signal $f_1^\sigma(t)$ and $f_2^\sigma(t)$ at each Gaussian noise level as described above. We define the embedding dimension as $M = 1$ for both measures. For computing %REC and %DET, we fix the distance threshold to $\texttt{tol} = 0.2 \cdot \max\{|| \text{SW}_{M,\tau} f_1(T) - \text{SW}_{M,\tau} f_2(T)||\}$ and the time lag to be the first minimum of the automutual information of f_1 when $\sigma = 0$ (i.e. $\tau = 10$) [7] . To compute %DET, we additionally fix the minimum number of diagonal points to $\texttt{minDL} = 5$.

Notice in Figure 3, that our measure is much more stable to small levels of added Gaussian noise to both f_1 and f_2 than %REC. This suggests that our bi-conditional periodicity score is indeed a more stable measure of time-series correlation.
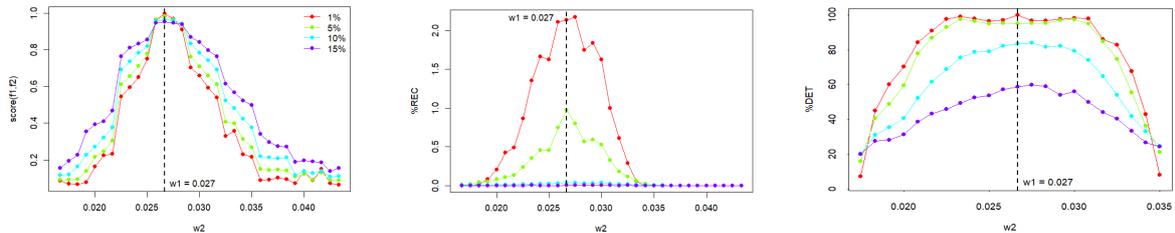

Fig. 3 Comparing the average stability of the bi-conditional periodicity score (left), percent recurrence (middle), and percent determinism (right) to added Gaussian noise to the sinusoidals $f_1(t) = \sin(2\pi w_1 t)$ and $f_2 = \sin(2\pi w_2 t)$. We fix the frequency of f_1 to $w_1 = 8/300$ and vary that of f_2. Then $score(f_1, f_2)$ is the average of $score(f_1|f_2)$ and $score(f_2|f_1)$. We denote the underlying frequency of f_1 via a vertical dashed line.

7.4 Real Data

For this experiment, we compare the similarity of climate data from three different locations near Australia. We specifically compare three different time series of average hourly cooling temperatures (in degrees Farenheit) from the National Oceanic and Atmospheric Administration [23]. The hourly cooling degree typically measures the energy demand (i.e. for air conditioning), by denoting how many degrees the average hourly temperature is above the baseline of $65°$. These hourly averages are taken over 30 years, between the years of 1981 and 2010. We compare the corresponding time-series data from three weather stations - Pago Pago airport, Yap Island airport, and Koror weather service office. In Figure 4, we show the respective locations of these stations on a map in black, blue, and green, along with their time series for the first 300 hours (12.5 weeks). We denote these time series as f_{Pago} (black), f_{Yap} (blue), and f_{Koror} (green), which are defined on the hours $T = \{0, 1, \ldots, 299\}$. To pre-process these series, we first interpolate any missing values via cubic splining, and then fit a continuous function to each discrete signal evaluated at more discretized time points $T_{\text{fine}} = \{0, 0.5, \ldots, 299\}$.

In order to embed these time series, we fix the embedding dimension to $M = 1$. The remaining parameters we define are used for computing %DET, including the time lag $\tau_{\text{DET}} = 5$, the distance threshold of $\texttt{tol} = c \cdot \max\{|| \text{SW}_{M,\tau} f_i(T_{\text{fine}}) - \text{SW}_{M,\tau} f_j(T_{\text{fine}})||\}$ for $c = 0.00105$, and the minimum number of points to qualify a diagonal as $\texttt{minDL} = 9$. We obtain the time lag by first extracting the first minimum of automutual information for each of the three time series ($\tau_{\text{Pago}} = 5, \tau_{\text{Yap}} = 5, \tau_{\text{Koror}} = 5$) and then taking their maximum [7]. We denote their resulting embeddings as $\text{SW}_{M,\tau} f_{\text{Pago}}(t)$, $\text{SW}_{M,\tau} f_{\text{Yap}}(t)$, and $\text{SW}_{M,\tau} f_{\text{Koror}}(t)$. We obtain our distance threshold, $\texttt{tol}$, as 0.105% of the maximum pairwise distance between any two embeddings and the minimum $\texttt{minDL} = 9$ by conducting a grid search of for each pair of series over the parameters $c = \{0.00104, 0.00105, 0.00106\}$ and $\texttt{minDL} = \{8, 9, 10\}$ and obtaining stable measures of %DET for all parametric combinations. When comparing f_{Pago} and f_{Yap}, f_{Pago} and f_{Koror}, and f_{Yap} and f_{Koror}, we obtain %DET values in the respective ranges $[5.11, 15.9]$, $[9.2, 18.61]$, and $[23.8, 32.56]$ with respective averages of 10.33%, 13.91%, and 28.19%. We hence chose the median parameters of $c = 0.00105$ and $\texttt{minDL} = 9$ for our computations. In Figure 4, we show all three pairs of
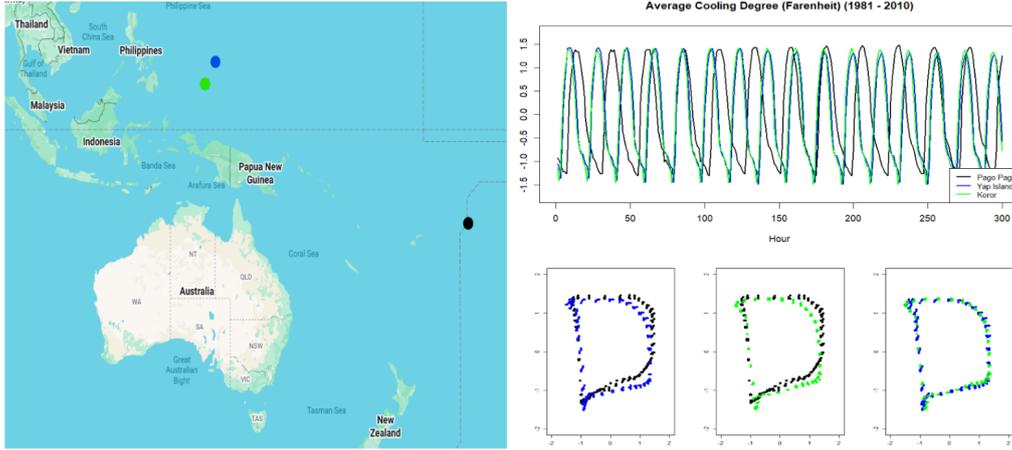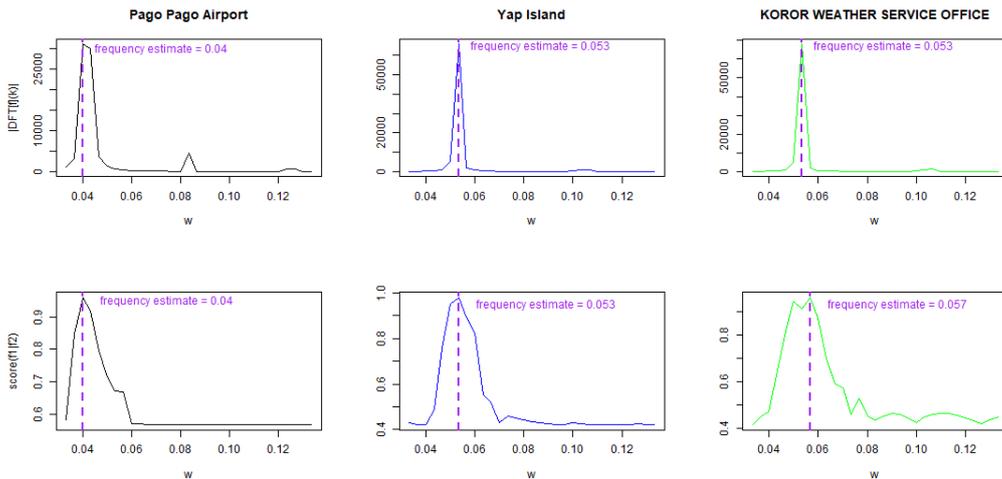
Fig. 4 Average Hourly Cooling Degrees recorded from three stations between 1981 and 2010, Pago Pago (black), Yap Island (blue), and Korer (green). We show the locations of these stations on the left, the first 300 hours of data in the top right, and the corresponding pairs of time-series embeddings which we compare via %DET in the bottom right.

time series embeddings for Pago Pago (black), Yap (blue), and Koror (green) which we compare using %DET.

In Figure 4, we can see that the average cooling degrees at Yap and Koror stations are more similar over time in comparison to Pago Pago. This makes sense intuitively, since these two stations are closer geographically, so we'd expect their oceanic environments to behave more similarly. As a result, we can also see that the embeddings of Yap and Koror data (shown in blue and green) also share more overlap than those of Yap and Pago Pago or Koror and Pago Pago.

7.4.1 Comparing Frequency Estimates of Climate Data

We first compare the frequency distributions produced from the discrete spectral magnitudes $|\text{DFT}[f_i](k)|$ for $k = \{10, 11, \ldots, 40\}$ and $\text{score}(f_i|f_w)$. We specifically plot these measures over the frequencies $w = \{k/600\}_{k=10}^{40}$ in cycles per hour, and denote the corresponding reference sinusoidal for the conditional periodicity score to f_w. We again fix the embedding dimension for $\text{score}(f_i|f_w)$ to $M = 1$. We show the resulting distributions of $|\text{DFT}[f_i](k)|$ (top) and $\text{score}(f_i|f_w)$ (bottom) in Figure 5. We show the peaks (i.e. the frequency estimates) for Pago Pago, Yap, and Koror stations via purple vertical dashed lines.

Fig. 5 The frequency distributions of three time series of average hourly cooling degrees over thirty years as computed by the Discrete Fourier spectral magnitudes (top) and our conditional periodicity score (bottom). Yap Island and Koror both appear to contain temperatures that oscillate at a rate of about 19 hours per cycle, while Pago Pago temperatures oscillate at a slower rate of 25 hours a cycle. We fix the embedding dimension to compute $\text{score}(f_1|f_2)$ to $M = 1$.

16

According to Algorithm 1, we use the estimates as produced via $|\mathrm{DFT}[f_i](k)|$, but we show the results from $\mathrm{score}(f_i|f_w)$ as well to highlight how close the estimates are for each. Using the estimates from $|\mathrm{DFT}[f_i](k)|$, we fix the time lags for the three conditional sliding windows embeddings as $\tau_{\mathrm{Pago}} = \frac{1}{0.04(M+1)}$, $\tau_{\mathrm{Yap}} = \frac{1}{0.053(M+1)}$, and $\tau_{\mathrm{Koror}} = \frac{1}{0.053(M+1)}$. For example, the conditional sliding windows embeddings of Pago given f_w is denoted $\mathrm{SW}_{M,\tau_w} f_{\mathrm{Pago}|w} = \big(f_{\mathrm{Pago}}(t), \ldots, f_{\mathrm{Pago}}(t + M\tau_w)\big)$, for $\tau_w = \frac{1}{w(M+1)}$ the time lag for the reference sinusoid f_w.

We also point out that the frequency estimates for both measures show greater similarity of Yap and Koror cooling degrees over time (about 19 hours per cycle) when compared to Pago (25 hours per cycle), further validating our intuition about their closeness geographically.

7.4.2 Comparing Similarity Estimates of Climate Data

Here we compare the correlation measures produce from percent determinism to our bi-conditional periodicity score for each combination of time series. As previously mentioned, for both measures we fix the embedding dimension to $M = 1$, and the remaining parameters for %DET we fix to $\tau_{\mathrm{DET}} = 5$, minDL = 9, and tol = $0.00105 \cdot \max\{\| \mathrm{SW}_{M,\tau} f_i(t) - \mathrm{SW}_{M,\tau} f_j(t)\|\}$. We use the frequency estimates from the discrete Fourier magnitudes in Figure 5 to define the time lags of the conditional sliding windows embeddings for each pair of time series comapred. For instance, the conditional sliding windows embedding and corresponding conditional periodicity score of Pago Pago given Yap are given below by

$$\mathrm{SW}_{M,\tau_{\mathrm{Yap}}} f_{\mathrm{Pago}-\mathrm{Yap}}(t) = \big(f_{\mathrm{Pago}}(t), \ldots, f_{\mathrm{Pago}}(t + M\tau_{\mathrm{Yap}})\big),$$

$$\mathrm{score}(f_{\mathrm{Pago}}|f_{\mathrm{Yap}}) = 1 - \frac{\mathrm{mp}\big(\mathrm{SW}_{M,\tau_{\mathrm{Yap}}} f_{\mathrm{Pago}-\mathrm{Yap}}(T_{\mathrm{fine}})\big)}{\sqrt{3}},$$

where $\tau_{\mathrm{Yap}} = \frac{1}{0.053(M+1)}$. The bi-conditional periodicity score of Pago and Yap is then given as the average, $\mathrm{score}(f_{\mathrm{Pago}}, f_{\mathrm{Yap}}) = \frac{\mathrm{score}(f_{\mathrm{Pago}}|f_{\mathrm{Yap}}) + \mathrm{score}(f_{\mathrm{Yap}}|f_{\mathrm{Pago}})}{2}$. We show our results in Figure 6.
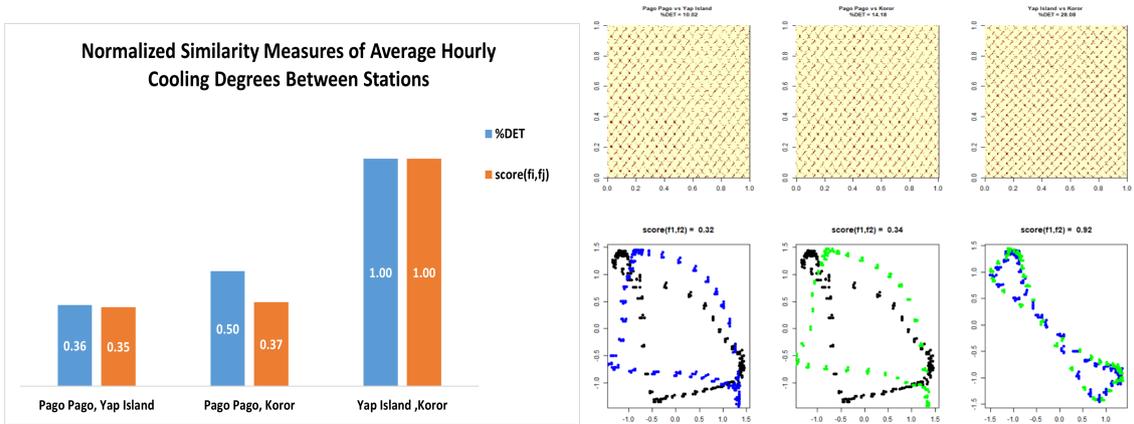
Fig. 6 The similarity estimates of our biconditional periodicity score and percent determinism. We show the normalized measures of %DET in blue and $\mathrm{score}(f_1, f_2)$ on the left for each pair of time series. On the right we show the cross-recurrence matrices for each pair (top) and the corresponding conditional sliding windows embeddings (bottom). We fix the embedding dimension to $M = 1$ for both measures and color the conditional embeddings according to the series being embedded (i.e. the green embedding in the middle column corresponds to the conditional embedding of Koror given Pago). To compute %DET, we fix the distance threshold to be 0.105% of the maximum pairwise distance between both embeddings (i.e. $c = 0.00105$) and minDL = 9. We obtain these values by retaining a stable range of values for %DET among $c = \{0.00104, 0.00105, 0.00106\}$ and minDL = $\{8, 9, 10\}$. We finally fix the time lag for %DET by first identifying the first minimum of automutual information of each series and taking the maximum to be $\tau = 5$.
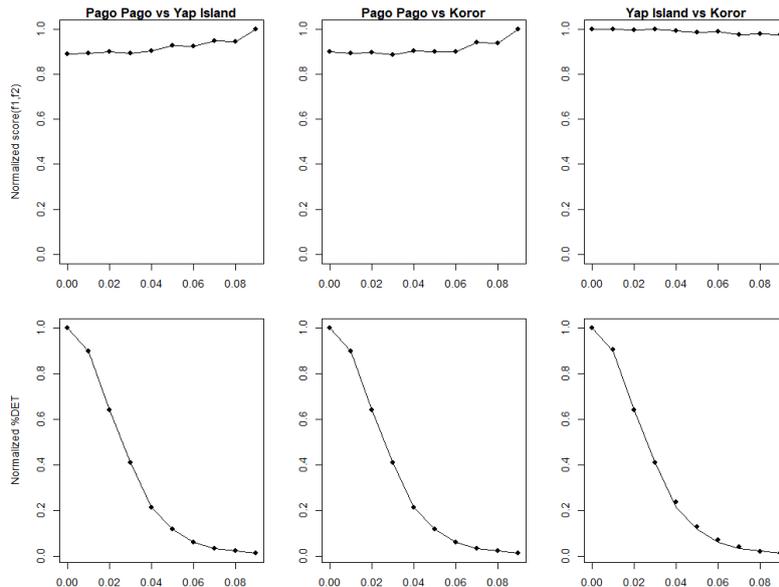
We plot the normalized bi-conditional periodicity scores in orange and the normalized %DET measures in blue. We can again see that both measures detect greater similarity in cooling degrees over time between the closer stations of Yap Island and Koror. In the right hand side of Figure 6, we show the cross-recurrence matrices and %DET values for each pair of time series being compared (top), along with the corresponding conditional sliding windows embeddings used to compute the bi-conditional periodicity scores (bottom). For example, the %DET estimated when comparing the time series for Pago Pago

and Yap Island is 10.04%, obtained from the cross-recurrence matrix on the left. The conditional sliding windows embeddings of Pago Pago given Yap and Yap given Pago Pago are shown directly below this matrix in black and blue, respectively, whose conditional periodicity scores are used to obtain the bi-conditional periodicity score of 0.32. The normalized measures of %DET and score($f_{\text{Pago}}, f_{\text{Yap}}$) are then given as 0.36 and 0.35 in the left cluster of bars in the bar plot.

We again wish to highlight that both measures follow our intuition - since Pago Pago and Yap Island are stations that are closer together geographically, we naturally expect the behavior of their average cooling degrees over time to be more similar. This greater similarity is highlighted in Figure 6, as both percent determinism and our bi-conditional periodicity score are largest when comparing these two time series.

7.4.3 Comparing Stability of Similarity Estimates on Climate Data

We finally compare the stability of the bi-conditional periodicity score to percent determinism when adding increasing levels of Gaussian noise to each time series. Here, we vary the standard deviation of Gaussian noise between the values $\sigma = \{0, 0.01, \ldots, 0.09\}$ and report the average (normalized) measure over 30 samples of time series pairs at each noise level. For example, we define the time series of Pago Pago at noise level σ as $f_{\text{Pago}}^{\sigma}(t) = f_{\text{Pago}}(t) + \mathcal{N}(0, \sigma \cdot \text{sd}(f_{\text{Pago}}))$, where $\text{sd}(f_{\text{Pago}})$ denotes the standard deviation of the original time series. We fix all parameters for both measures to those used in the previous experiment. See our results in Figure 7. As shown, for small amounts of added noise, our measure remains much more stable. Our bi-conditional periodicity score stays within 10% of its original value in all three time-series comparisons for up to 10% noise, while %DET drops to less than 50% of its original value for only 3% noise.

Fig. 7 The normalized average bi-conditional periodicity score (top) and percent determinism (bottom) for three pairs of time series containing average cooling degrees per hour against added levels of Gaussian noise.

Notice that the conditional embeddings of Pago given Yap and Pago Given Koror have more clustered points (see the bottom right images of Figure 6). Hence, adding small amounts of noise will spread them out more evenly, causing the conditional embeddings to be more rounded and the max persistence to increase (i.e. the score will increase). This is why we see a small increase in the bi-conditional score in the top row of Figure 7. Oppositely, when comparing Yap and Koror, both pairs of conditional embeddings are much skinnier, hence adding small amounts of noise to their points will fill in the hole that is already there, causing a subtle decrease in the max persistence and hence the score.

8 Discussion and Future Work

In this work, we are successfully able to introduce a novel persistence-based measure of time-series similarity with both stability and parametric advantages to a previously defined method. We are not only able to prove theoretical stability of our measure in comparison to percent determinism, but we are also able to experimentally verify this greater stability on synthetic and real time series data. While this work is novel and our measure avoids the limitations posed by percent determinism, we are still interested in continuing this work by extending our measure to higher homology dimensions. As well, we are interested in constructing a measure of similarity that uses a cross-recurrence matrix that measures correlation between the 0- (or higher-) dimensional Betti curves from filtration on a pair of time series. As 0-dimensional Betti-curves are stable to small time-series perturbations, we are further motivated to try this approach as an alternatively stable measure of time-series similarity to percent determinism.

References

[1] Bracewell, R.N.: The Fourier Transform and Its Applications, 3rd edn. McGraw-Hill, Boston (2000)

[2] Qian, S., Chen, D.: Joint time-frequency analysis. IEEE Signal Processing Magazine **16**(2), 52–67 (1999) https://doi.org/10.1109/79.752051

[3] Russell, B., Han, J.: Jean morlet and the continuous wavelet transform. CREWES Res. Rep **28**, 115 (2016)

[4] Coco, M.I., Dale, R.: Cross-recurrence quantification analysis of categorical and continuous time series: an r package. Frontiers in Psychology **5**, 510 (2014) https://doi.org/10.3389/fpsyg.2014.00510

[5] Ichinomiya, T.: Machine learning of time series data using persistent homology. Scientific Reports **15**(20508) (2025) https://doi.org/10.1038/s41598-025-06551-3

[6] Güzel, İ., Munch, E., Khasawneh, F.A.: Detecting bifurcations in dynamical systems with CROCKER plots. Chaos: An Interdisciplinary Journal of Nonlinear Science **32**(9), 093111 (2022) https://doi.org/10.1063/5.0102421

[7] Wallot, S., Leonardi, G.: Analyzing multivariate dynamics using cross-recurrence quantification analysis (crqa), diagonal-cross-recurrence profiles (dcrp), and multidimensional recurrence quantification analysis (mdrqa) - a tutorial in r. Frontiers in Psychology **9**, 2232 (2018) https://doi.org/10.3389/fpsyg.2018.02232

[8] Cohen-Steiner, D., Edelsbrunner, H., Harer, J.: Stability of persistence diagrams. Discrete & Computational Geometry **37**(1), 103–120 (2007) https://doi.org/10.1007/s00454-006-1276-5

[9] Perea, J., Harer, J.: Sliding Windows and Persistence: An Application of Topological Methods to Signal Analysis. Foundations of Computational Mathematics **15**, 799–838 (2015) https://doi.org/10.1007/s10208-014-9206-z 1307.6188

[10] Cohen-Steiner, D., Edelsbrunner, H., Harer, J.: Stability of persistence diagrams. Discrete & Computational Geometry **37**, 103–120 (2007) https://doi.org/10.1007/s00454-006-1276-5

[11] Perea, J.A., Deckard, A., Haase, S.B., Harer, J.: SW1PerS: Sliding windows and 1-persistence scoring; discovering periodicity in gene expression time series data. BMC Bioinformatics **16**(1), 257 (2015) https://doi.org/10.1186/s12859-015-0645-6

[12] Munkres, J.R.: Elements of Algebraic Topology, p. 454. Addison–Wesley Publishing Company, Menlo Park (1984)

[13] Otter, N., Porter, M.A., Tillmann, U., Grindrod, P., Harrington, H.A.: A roadmap for the computation of persistent homology. EPJ Data Science **6**(17), 1–38 (2017) https://doi.org/10.1140/epjds/s13688-017-0109-5

[14] Burago, D., Burago, Y., Ivanov, S.: A Course in Metric Geometry. Graduate Studies in Mathematics, vol. 33. American Mathematical Society, ??? (2001)

[15] Adams, H., Frick, F., Majhi, S., McBride, N.: Hausdorff vs Gromov-Hausdorff distances (2024)

[16] Aggarwal, C.C., Hinneburg, A., Keim, D.A.: On the surprising behavior of distance metrics in high dimensional space. In: Bussche, J., Vianu, V. (eds.) Database Theory — ICDT 2001, pp. 420–434. Springer, Berlin, Heidelberg (2001)

[17] Koyama, M.A., Memoli, F., Robins, V., Turner, K.: Faster computation of degree-1 persistent homology using the reduced Vietoris-Rips filtration (2024)

[18] Jolliffe, I.T.: Principal Component Analysis vol. 89. Springer, ??? (2002)

[19] May, N.H., Krishnamoorthy, B., Gambill, P.: A Normalized Bottleneck Distance on Persistence Diagrams and Homology Preservation Under Dimension Reduction. La Matematica, 1–23 (2024) https://doi.org/10.1007/s44007-024-00130-0

[20] Anselin, L.: Dimension Reduction Methods (2): Distance Preserving Methods. Available as part of GeoDa: An Introduction to Spatial Data Science (2020). https://geodacenter.github.io/workbook/7ab_mds/lab7ab.html

[21] Brigham, E.O., Morrow, R.E.: The fast Fourier transform. IEEE Spectrum **4**(12), 63–70 (1967) https://doi.org/10.1109/MSPEC.1967.5217220

[22] Yiğit, H.: Time Complexity of PCA 1. General Time Complexity of PCA. Posted on ResearchGate (2024). https://doi.org/10.13140/RG.2.2.12847.34728

[23] Arguez, A., Durre, I., Applequist, S., Squires, M., Vose, R., Yin, X., Bilotta, R.: NOAA's U.S. Climate Normals (1981-2010). https://doi.org/10.7289/V5PN93JP . Monthly temperature normals for Pago Pago airport (AQW00061705), Yap Island airport (FMW0040308), and Koror weather service station (PSW00400309). Accessed 2024-01-15