A Novel Method for Topological Embedding of Time-Series Data

Sean M. Kennedy, John D. Roth, and James W. Scrofani

Abstract— **In this paper, we propose a novel method for embedding one-dimensional, periodic time-series data into higher-dimensional topological spaces to support robust recovery of signal features via topological data analysis under noisy sampling conditions. Our method can be considered an extension of the popular time delay embedding method to a larger class of linear operators. To provide evidence for the viability of this method, we analyze the simple case of sinusoidal data in three steps. First, we discuss some of the drawbacks of the time delay embedding framework in the context of periodic, sinusoidal data. Next, we show analytically that using the Hilbert transform as an alternative embedding function for sinusoidal data overcomes these drawbacks. Finally, we provide empirical evidence of the viability of the Hilbert transform as an embedding function when the parameters of the sinusoidal data vary over time.**

I. INTRODUCTION

Topological data analysis (TDA) is an emerging field of study which argues that many powerful insights about data come from the way the data points are structured with respect to each other: that is, the "shape" of the data is important [1]. The flagship tool of TDA is persistent homology (PH), which examines point clouds of data embedded into high-dimensional topological space in order to find regions of the space where n-dimensional "holes" exist [2]. As such, TDA is well-suited to analyzing data sets which are high-dimensional, i.e., each data "point" results from multiple measurements of a single sample. TDA has been used to uncover surprising insights in many different applications, and active research continues into the applicability of TDA to various classes of data problems [3][4][7].

While TDA is naturally suited to analyzing data which possesses an inherently high dimensionality, there has also been an interest in determining whether these topological techniques can also reveal insights about data taken from low-dimensional measurements, such as scalar time-series data. One immediate problem in investigating this question is that it is not obvious how scalar time-series data could be transformed into a higher-dimensional point cloud. Time delay embeddings (TDEs) provide one possible solution, and have been applied by multiple authors to real-world problems including wheeze detection in mobile devices [3], and discovering periodicity in gene expression data [4].

In this paper, we propose a novel method for performing this data transformation which extends the method of TDEs to other linear operators. The objective of this new method is to improve the robustness of TDA methods and tools (such as those provided in [3][4][5]) in the presence of degrading noise, measurement imperfection, and nonstationarity of the signal parameters.

The remainder of the paper is organized as follows: a brief overview of TDEs and the TDA framework is provided in Section II. In Section III, some drawbacks of the TDE method for sinusoidal signals are illustrated. In Section IV, we present our novel method as an extension of TDE method. In Section V, we show how to use this method to find a superior embedding function for sinusoidal signals. This embedding function is tested against noisy data simulating common communication signals with results presented in Section VI. Finally, conclusions and recommendations for future work are presented in Section VII.

II. Summary of Time Delay Embeddings and Topological Data Analysis

The method of using TDEs to transform scalar time-series data into a higher-dimensional point cloud can be summarized from [3] and [4] as follows. Given a discrete scalar time-series $x[k]$, form the multidimensional vector $X[k]$ according to

$$X[k] = (x_1[k], ..., x_m[k]), \ X[k] \in \mathbb{R}^m, \tag{1}$$

where each $x_i[k] = x[k + (i-1)j]$ $(i = 1, ..., m)$, j is a constant number of samples (the delay parameter) and m is the embedding dimension. Since the data points of $X[k]$ are also time-ordered, this point cloud can be further interpreted as samples taken from a continuous, parametric path through the m-dimensional space. As discussed in [5], points in space at which the path returns to itself create topological circles, hereafter referred to as cycles. Computing the 1-D homology of the continuous path would reveal the existence of these cycles, which in turn would inform further data analysis efforts. However, since the point cloud is a discrete sampling of this path, the homology of the point cloud may differ from the true homology of the path based on the assumed resolution at which the sampling was performed [2].

The TDA framework using PH overcomes the issue of discretely sampled points through the concept of persistence. When provided a point cloud of high-dimensional data, the PH algorithm iteratively computes and stores the homology of the data set by assuming increasing sampling resolutions at each step, up to a defined limit [1]. When computed in this way, each topological feature is said to be "born" at some smaller resolution and "die" at a larger resolution [1]. The persistence of a given feature is the length of the interval between birth and death. The final output of the algorithm is a description of the persistence of all homological features

The authors are with the Department of Electrical and Computer Engineering, Naval Postgraduate School, Monterey, CA 93943, USA. smkenned@nps.edu, jdroth@nps.edu, jwscrofa@nps.edu

ISBN 978-90-827970-1-5 © EURASIP 2018

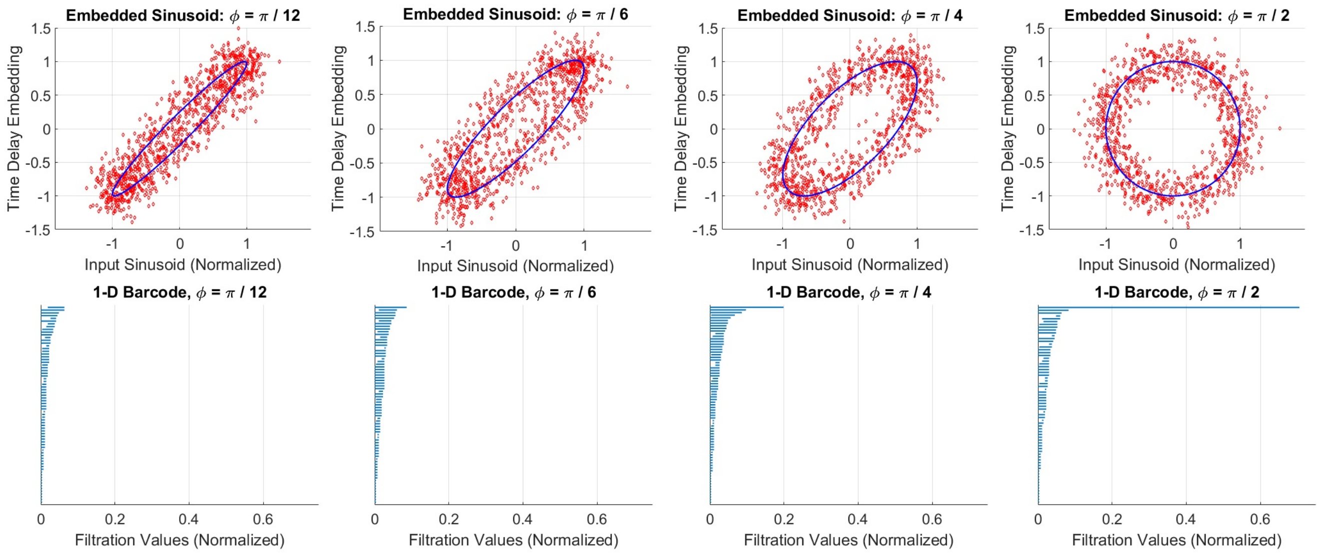

Fig. 1. **Top Row:** Lissajous ovals from a sinusoidal TDE at various phase delays. The blue oval is provided as an ideal reference for the case of dense, noiseless sampling. The red markers are provided as an example of noisy sampling (100 samples/period for 10 periods with AWGN at 12 dB SNR). **Bottom Row:** Barcodes corresponding to the 1-D persistent homology of each noisy plot in the top row, ignoring the blue ovals. The length of the longest line segment provides a measure of the size of the primary void in the associated oval. Persistent homology was computed via JavaPlex [10].

of the point cloud over all computed resolutions, and is often provided visually in the form of a barcode graph. The theory of PH argues that persistent features (i.e., those existing over a large resolution range) imply an underlying structure to the data, whereas transient features (i.e., those existing only over a small resolution range) are likely due to noisy imperfections in measurement and/or side effects of the computational procedure [1].

III. Drawbacks of Time Delay Embeddings for Sinusoidal Signals

The use of TDEs is justified mathematically by a theorem proven by Takens [6], in which Takens demonstrates that for a sufficiently large m, almost every choice of delay parameter j would allow for detection of any existing cycles, provided there are an infinite number of noiseless data points available. In real-world signals, however, data is both noisy and finite. These properties blur and discretize the path taken through the m-dimensional space, potentially obscuring meaningful cycles and/or inducing false cycles depending on the assumed resolution of the sampling. As a result, only a select few choices of j may actually provide valid results under noisy sampling. A primary difficulty with using TDEs is predicting what these valid choices of j are when the parameters of the underlying signal or system are unknown. In this section, we examine several factors which contribute to this problem.

A. Compression of a Topological Circle to a Line

A poor choice of the delay parameter j can compress the embedding of a periodic signal to a single line in the topological space [3], which would have trivial homology. To demonstrate this, consider the case of a discrete periodic signal whose samples are taken from an analog sinusoidal function of the form $x(t) = A\cos(2\pi ft)$, with A and f both nonzero. Constructing a TDE of this discrete signal is equivalent to sampling $y(t) = x(t + \Delta t)$, where $y(t)$ is a time-delayed version of the original analog signal. Simplifying $y(t)$ gives

$$y(t) = A\cos(2\pi f(t + \Delta t)),$$

$$y(t) = A\cos(2\pi ft + 2\pi f\Delta t),$$

$$y(t) = A\cos(2\pi ft + \phi);\ \ \phi = 2\pi f\Delta t. \tag{2}$$

As shown, $y(t)$ is simply a phase-shifted version of $x(t)$, with the amount of phase shift determined jointly by the frequency of the original signal and the chosen time delay. Plotting the analog signals $x(t)$ and $y(t)$ as a parametric equation in $\mathbb{R}^2$ produces a special case of a Lissajous curve: an oval whose eccentricity, ϵ, is determined exclusively by the phase shift [9]. For phase shifts of $0 \pm n\pi$ ($n \in \mathbb{Z}$) radians, the oval is compressed to a single line ($\epsilon = 1$). For phase shifts of $\frac{\pi}{2} \pm n\pi$ ($n \in \mathbb{Z}$), the oval becomes a circle ($\epsilon = 0$), and encloses the maximum possible area among all possible phase values. The effect of sampling $x(t)$ and $y(t)$ is a discretization of the analog path, possibly including noise. This effect is shown in the top row of Fig. 1 for multiple phase delays in both the ideal and noisy sampling cases.

B. Effects of Noisy Sampling

The 1-D barcode corresponding to each case of phase delay is given in the second row of Fig. 1. Compared with the theoretical ideal of a single line persisting throughout the entire interval, it is easy to see that an increasing ϵ corresponds to decreasing performance in detecting the cycle. Additionally, one can observe that as $\epsilon \to 1$, the embedded data points (red markers in Fig. 1) cluster toward the vertices of the semi-major axis of the ellipse. This reduces the

ISBN 978-90-827970-1-5 © EURASIP 2018

effectiveness of density clustering techniques to de-noise the data, as the density around the cycle is no longer uniform. Note, however, that this effect is not present when $\epsilon = 0$.

C. Effects of Nonstationarity in the Sinusoid Parameters

The examples shown in Fig. 1 assumed a sinusoid of constant amplitude, frequency, and phase throughout the measurement interval. However, many real signals of interest (e.g., amplitude/frequency/phase-shift keying) routinely alter these parameters to encode information. Using too long of a delay parameter such that data from disparate intervals are used to create the parametric path is unlikely to have relevant physical meaning since the intervals are generally uncorrelated [8]. Additionally, the frequency even within a given interval could vary due to, for example, sampling time jitter in the measurement equipment or the Doppler effect (in the case of mobile measurement equipment). Available space does not permit in-depth analysis of these effects here, but it should be noted that such nonstationarity in the sampled sinusoidal signal generally results in greater difficulty in selecting a proper delay parameter and/or decreased cycle detectability in the PH analysis [3].

IV. Proposed Embedding Framework

A. Reformulation of Time Delay Embeddings

Each $x_i[k]$ element in $X[k]$ from (1) is a time-delayed version of the original input data $x[k]$. This relationship can be expressed as

$$x_i[k] = x[k + (i-1)j] = x[k] * \delta[k + (i-1)j], \qquad (3)$$

where i, j, and k have the same meaning as in (1), the '$*$' operator denotes convolution, and $\delta[k]$ is the unit impulse function (i.e., Kronecker delta function $\delta_{0,k}$). Assuming $x[k] = x(k \cdot T_s)$, where T_s is the sampling period, then $x_i[k] = x_i(k \cdot T_s)$, where

$$x_i(t) = x(t) * \delta(t + (i-1)j \cdot T_s), \qquad (4)$$

and $\delta(t)$ is the Dirac delta function. As a result, each $x_i[k]$ in a TDE can be considered either as the sampled output of a continuous linear system whose input is the original analog signal, or as the output of a discrete linear system whose input is the original time-series data. In each case, the impulse response of the linear system is the appropriate time-delayed delta function.

B. Extension to Other Impulse Responses

We propose extending the impulse responses of (3) and (4) to more complex functions of the input vector. Thus, each $x_i[k]$ in the multidimensional vector $X[k]$ is constructed as

$$x_i[k] = x[k] * h_i[k], \qquad (5)$$

for some set of discrete impulse responses, $h_i[k]$, or in an equivalent continuous case as

$$x_i[k] = x_i(k \cdot T_s), \qquad (6)$$

$$x_i(t) = x(t) * h_i(t) = \int_{-\infty}^{\infty} x(\tau) h_i(t-\tau) d\tau, \qquad (7)$$

for some set of continuous impulse responses, $h_i(t)$.

Analogous to the concept of a matched filter in classical signal processing, this may allow one to find and implement an optimal embedding function to maximize the recoverability of various topological features. We note that the definition of optimal will generally be context-specific for the data being analyzed, and in practice may involve a trade-off between different parameters to ensure robustness.

V. Optimizing Persistence Intervals for Single Frequency Sinusoids

To provide a simple, constructive example of the applicability of this method to signal analysis problems, we analyze the case where it is known *a priori* that the input signal of interest is a stationary sinusoid with no DC offset. Furthermore, we assume that we have no knowledge of the parameters of the sinusoid. These assumptions allow the method to be contrasted to the TDE discussion in Section III. We begin by recalling that a sinusoid can be considered as a projection of 2-D circular motion onto a single axis, and that the objective is to recover this circular motion topologically. Therefore, the minimum required embedding dimension is two. Since minimizing the number of additional dimensions reduces computational burden and memory requirements in a persistence calculation, this minimum embedding dimension is also the desired maximum. Thus, we seek one embedding function which, when combined with the original signal, forms a topological circle in $\mathbb{R}^2$.

As discussed in Section III, the TDE which produces the longest persistence interval for noisy sinusoidal data occurs when the associated phase delay is $\frac{\pi}{2} \pm n\pi$ $(n \in \mathbb{Z})$, which corresponds to a geometric circle in the plane. We can convert this observation into a constraint on the embedding function by requiring the parametric path produced between the input and output vectors to always trace such a geometric circle, regardless of the parameters of the input signal. Considering the input signal as $x(t)$, the output signal as $y(t)$, and the embedding function as $h(t)$, the problem of finding the necessary embedding function can be written as follows:

Problem: Given $x(t) = A\cos(2\pi f t)$ $(f > 0)$, find a function, $h(t)$, that satisfies

$$x^2(t) + y^2(t) = A^2, \ \forall t, \qquad (8)$$

where

$$y(t) = x(t) * h(t) = \int_{-\infty}^{\infty} x(\tau) h(t-\tau) d\tau. \qquad (9)$$

This follows immediately from combining (7) with the equation of a circle in the Cartesian plane. We now show analytically that the Hilbert transform (HT) provides one such solution. First, we note that the definition of the HT can be written as a convolution of $x(t)$ with a tempered distribution

$$\mathrm{Hilb}(x(t)) = x(t) * h(t) \equiv \frac{1}{\pi}\ p.v. \int_{-\infty}^{\infty} \frac{x(\tau)}{(t-\tau)} d\tau, \qquad (10)$$

ISBN 978-90-827970-1-5 © EURASIP 2018

where $p.v.$ denotes *principle value*. Next, we record the well-known Fourier transform of this expression to obtain the frequency response of the embedding function, $H(f)$, from

$$\mathcal{F}[\text{Hilb}(x(t))] = H(f) \cdot X(f) = (-i \text{ sgn}(f)) \cdot X(f), \quad (11)$$

where $X(f)$ is the Fourier transform of $x(t)$, and $\text{sgn}(f)$ denotes the signum function. Next, we expand the integral equation for $y(t)$ by using the commutative property of convolution, substituting the cosine expression for $x(t)$, and applying Euler's formula to obtain

$$y(t) = \int_{-\infty}^{\infty} A\cos(2\pi f(t-\tau))h(\tau)d\tau, \quad (12)$$

$$y(t) = \int_{-\infty}^{\infty} \frac{A}{2}\left[e^{2\pi i f(t-\tau)} + e^{-2\pi i f(t-\tau)}\right] h(\tau)d\tau, \quad (13)$$

$$\begin{aligned} y(t) =& \frac{A}{2}e^{2\pi i f t}\int_{-\infty}^{\infty} h(\tau)e^{-2\pi i f \tau}d\tau \\ &+ \frac{A}{2}e^{-2\pi i f t}\int_{-\infty}^{\infty} h(\tau)e^{-2\pi i(-f)\tau}d\tau. \quad (14)\end{aligned}$$

Note that the two integrals are both of the form of a Fourier transform, and so this equation can be rewritten as

$$y(t) = \frac{A}{2}e^{2\pi i f t}H(f) + \frac{A}{2}e^{-2\pi i f t}H(-f). \quad (15)$$

Since $H(f)$ is an odd function, $H(-f) = -H(f) = i\text{ sgn}(f)$. Substituting in for $H(f)$ and $H(-f)$ appropriately into the equation above and simplifying yields

$$y(t) = A\left[\frac{e^{2\pi i f t} - e^{-2\pi i f t}}{2i}\right]\text{sgn}(f), \quad (16)$$

$$y(t) = A\sin(2\pi t f)\text{ sgn}(f). \quad (17)$$

Finally, we note that $\text{sgn}^2(f) = 1$ for $(f \neq 0)$, and so

$$\begin{aligned} x^2(t) + y^2(t) =& A^2\cos^2(2\pi f t) \\ &+ A^2\sin^2(2\pi f t)\text{ sgn}^2(f), \quad (18)\\ x^2(t) + y^2(t) =& A^2, \ \forall t, \quad (19)\end{aligned}$$

which matches the problem constraint, (8). Therefore, the HT is an embedding function which unfolds a sinusoid of amplitude A into a geometric circle of radius A in $\mathbb{R}^2$, regardless of the frequency of the sinusoid.

VI. Experimental Results

A. Stationary Sinuosoid

To validate robustness in the presence of noisy sampling, a sinusoid with constant amplitude, frequency, and phase was generated in MATLAB. Identical to the TDE case of Fig. 1, the sinusoid was sampled at a rate of 100 samples per period, for 10 periods, in the presence of additive white Gaussian noise (AWGN) at a measured signal-to-noise ratio (SNR) of 12 dB. The plot of the 2-D embedding and corresponding 1-D persistence barcode is provided in Fig. 2. As can be seen, the result is very similar to the Fig. 1 plot corresponding to a phase shift of $\frac{\pi}{2}$, as expected.

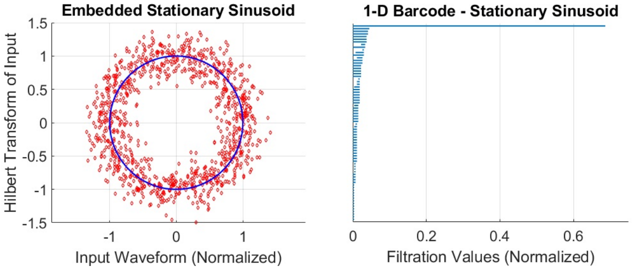

Fig. 2. **Left:** Phase plot of stationary sinusoid embedded via the HT. **Right:** Barcode corresponding to the 1-D persistent homology of left plot.

B. Nonstationary Sinusoid

To demonstrate the viability of the HT as an embedding function in the presence of nonstationarity, three representative waveforms modeling those commonly used in radar and wireless signaling were generated via MATLAB. These include a quadratic chirp (or frequency-swept signal), a quadrature phase-shift keyed (QPSK) signal, and a multi-level frequency-shift keyed (MFSK) signal. The time-series data for the waveforms were constructed by assuming a sampling rate of 1000 samples per second for 6 seconds, with a uniformly-random, noncompounding timing jitter within $\pm 1\%$. All three waveforms were further degraded by AWGN at a measured SNR of 12 dB. As a control signal, a fourth time-series composed entirely of AWGN is also included. The parameters of each signal are provided below:

- *Quadratic Chirp* – The instantaneous frequency of a constant-amplitude sine wave was continuously swept from an initial minimum of 1 Hz to final maximum of 495 Hz (99% of the Nyquist frequency) in a quadratic fashion.
- *QPSK* – The phase of a constant-amplitude sine wave, with a nominal frequency of 4 Hz, was set for each 1 second interval according to the following sequence: 45°, 225°, 135°, 315°, 45°, 135°.
- *MFSK* – The frequency of a constant-amplitude sine wave with zero phase was set for each 1 second interval according to the following sequence: 4 Hz, 8 Hz, 16 Hz, 12 Hz, 4 Hz, 16 Hz.
- *AWGN* – For comparison, a pure AWGN (no signal) waveform is generated at the same signal power and sampled at the same rate as the other three waveforms.

The 2-D embedding via HT of the chirp, QPSK, MFSK, and AWGN time-series are shown in the top row of Fig. 3, respectively. Clearly, the three information-bearing signals are differentiable from the pure noise scenario, and all three primary signals appear to embed as noisy circles. This is particularly remarkable since the frequency and/or phase of these signals vary in time throughout the measurement window, and the algorithm has no knowledge of these parameter transitions.

One can observe that points exist near the center of noisy circles which would hinder attempts at homological feature recovery. However, these points are not very dense in the space. Furthermore, the density of the point cloud appears

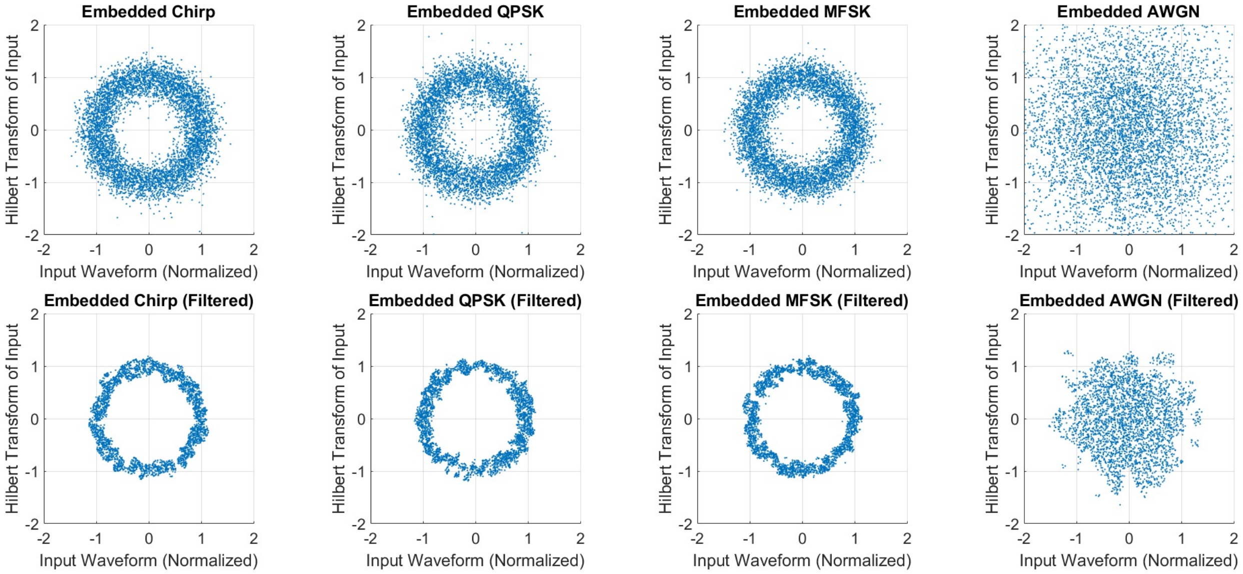

Fig. 3. **Top Row:** Plots of the chirp, QPSK, MFSK, and AWGN signals embedded with the Hilbert transform (all data points included). **Bottom Row:** Filtered versions of the top row with only the densest 50% of points included, as measured by shortest distance to the 25th nearest neighbor.

to be approximately uniform as function of distance from the unit circle, which was only true for TDEs which induced a relative phase shift of $\frac{\pi}{2}$. This symmetry can be exploited to reduce the total number of data points while simultaneously improving cycle detection, by using the method of dense core subsets included in [10]. By retaining only the 50% densest points as measured by distance to the 25th nearest neighbor, we obtained the 2-D plots provided in the bottom row of Fig. 3. After this process, which can be loosely considered as a sort of topological filtering, the three cycles appear much more dense around the ideal path of a circle of radius 1, while the pure noise signal becomes more dense around the origin. While not shown due to space constraints, it is obvious that the 1-D barcodes corresponding to the filtered, information-bearing signals would contain a large feature that the AWGN signal would not.

VII. Conclusion

By extending the set of possible embedding functions of time-series data beyond simple delta functions, we were able to overcome many of the drawbacks of TDEs for 1-D sinusoidal data while simultaneously reducing the amount of data necessary for a persistence computation. In particular, we found a new application for the Hilbert transform in the domain of signal processing via TDA. The authors wish to emphasize that while the focus of the paper was on periodic, sinusoidal signals as a standard case-study, the results of Section VI suggest broad applicability to many classes of quasi-periodic signals. These include modulated communication signals, radar pulses, and acoustic vocalizations in both real-time and static analysis scenarios. As such, there are many potential applications, including, for example, improved wheeze detection such as discussed in [3], novel methods for estimating signal parameters such as those proposed in [5], cognitive radio spectrum sensing, radar pulse detection, and exploratory data analysis. In future work, we plan to formalize the effects of noise and filtering on the persistent homology of the periodic signal, and attempt to derive “optimal” embedding functions for more complex signals and datasets (e.g., multi-carrier modulated signals).

References

[1] G. Carlsson, “Topology and data,” *Bull. AMS*, vol. 46, pp. 255-308, 2009.

[2] R. Ghrist, “Barcodes: the persistent topology of data,” *Bull. AMS*, vol.45, pp. 61-75, 2008.

[3] S. Emrani, T. Gentimis, and H. Krim, “Persistent homology of delay embeddings,” *IEEE Trans. Signal Process.*, vol. 21, no. 4, pp. 459-463, 2014.

[4] J. Perea, A. Deckard, S. Haase, and J. Harer, “Sw1pers: sliding windows and 1-persistence scoring; discovering periodicity in gene expression time series data,” *BMC Bioinform.*, vol. 16, no. 1, p. 257, 2015.

[5] V. de Silva, P. Skraba and M. Vejdemo-Johansson, “Topological analysis of recurrent systems,” *Workshop of Algebraic Topology and Machine Learning, NIPS 2010*, 2012. Preprint available at http://sites.google.com/site/nips2012topology/contributed-talks.

[6] F. Takens, “Detecting strange attractors in turbulence,” *Dynamical Systems and Turbulence, vol. 898 of Lecture Notes in Mathematics*, pp. 366-381, 1981.

[7] F. Erden, “Period estimation of an almost periodic signal using persistent homology with application to respiratory rate measurement,” *IEEE Trans. Signal Process.*, vol. 24, no. 7, pp. 958-962, 2017.

[8] M. Casdagli, S. Eubank, J. D. Farmer, and J. Gibson, “State space reconstruction in the presence of noise,” *Physica*, vol. 51, pp. 52-98, 1991.

[9] H. Al-Khazali and M. Askari, “Geometrical and graphical representations analysis of lissajous figures in rotor dynamic system,” *IOSR J. of Engineering*, vol. 2, iss. 5, pp. 971-978, 2012.

[10] A. Tausz, M. Vejdemo-Johansson, and H. Adams, “Javaplex: a research software package for persistent (co)homology,” *Proc. 4th Int. Conf. Math. Softw.*, pp. 129-136, 2014. [Online]. Available: http://appliedtopology.github.io/javaplex/.