Topological Data Analysis of Periodic Time Series with Noise

Presented to the MA in Mathematics Program

in Partial Fulfillment of the

Requirements for Completion of the Program

Katz School of Science and Health

Yeshiva University

January 4, 2023

Brian Livian

Adviser: Dr. Yuri Katz

1 Abstract

We study the behavior of L_1 norms of persistence landscapes derived from periodic time series with various levels of noise. We experiment on synthetic time series simulated either with sin waves or auto regressive models, employing sliding window time-delayed embeddings (SWE) to construct the point clouds required for topological analysis. In both sets of experiments, we find that L_1 norms of the persistence landscapes (derived from the point clouds in the SWE) follow the expected proportionality with the variance of point cloud. Additionally, we see the expected inverse proportionality between average L_1 norm and the magnitude of covariance of the point cloud. However, in multiple instances, we see L_1 norms fluctuate from the expected proportionality between variance and inverse proportionality with magnitude of covariance (of the point clouds in the SWE). We conclude that L_1 norms are highly sensitive to random configurations of point clouds that are accidentally favorable to the emergence of homologies with a relatively high persistence.

Acknowledgement

I would like to thank my adviser, Dr. Yuri Katz, for his guidance throughout the work on this thesis. Without his guidance this work would not have been possible.

Contents

2 Introduction

The growing complexity of datasets requires new methods of analysis. Topological Data Analysis is an emerging field of applied mathematics which provides novel tools for data analysis. In recent years, TDA has been applied to a broad range of fields such as: image analysis [4], disease diagnosis [10], viral evolution [5], signal analysis [11], financial time series [9, 2, 8] and others.

TDA is a field of computational topology which involves quantifying the shape of data. The ability to quantify the shape of data provides methods to extract information one might not find through traditional clustering algorithms. Consider a point cloud, $\mathcal{P} \in \mathbb{R}^n$. This point cloud of data has an inherent shape to it which can be quantified and analyzed via TDA. We quantify the shape of data by identifying topological homologies, ie. 'holes' in 2D point clouds, 'voids' in 3D point clouds, etc. for higher dimensional homologies. In the case of 1D time series, we construct point clouds using sliding window time-delay embeddings (SWE), then continue with the TDA methodology. The progression of the TDA process is shown in figure 1.

Here, we apply TDA to 1-dimensional periodic synthetic time series. We experiment on two settings of time series: sin waves (representing a dynamic function) with added white noise, and auto regressive models (representing a stochastic time series with some beating around a mean). We apply sliding window time-delayed embeddings to the time series, implement the TDA methodology and show the behavior of L_1 norms of the persistence landscapes per point cloud. We conclude that signals of L_1 norms obtained via sliding window embeddings are highly sensitive to noise.

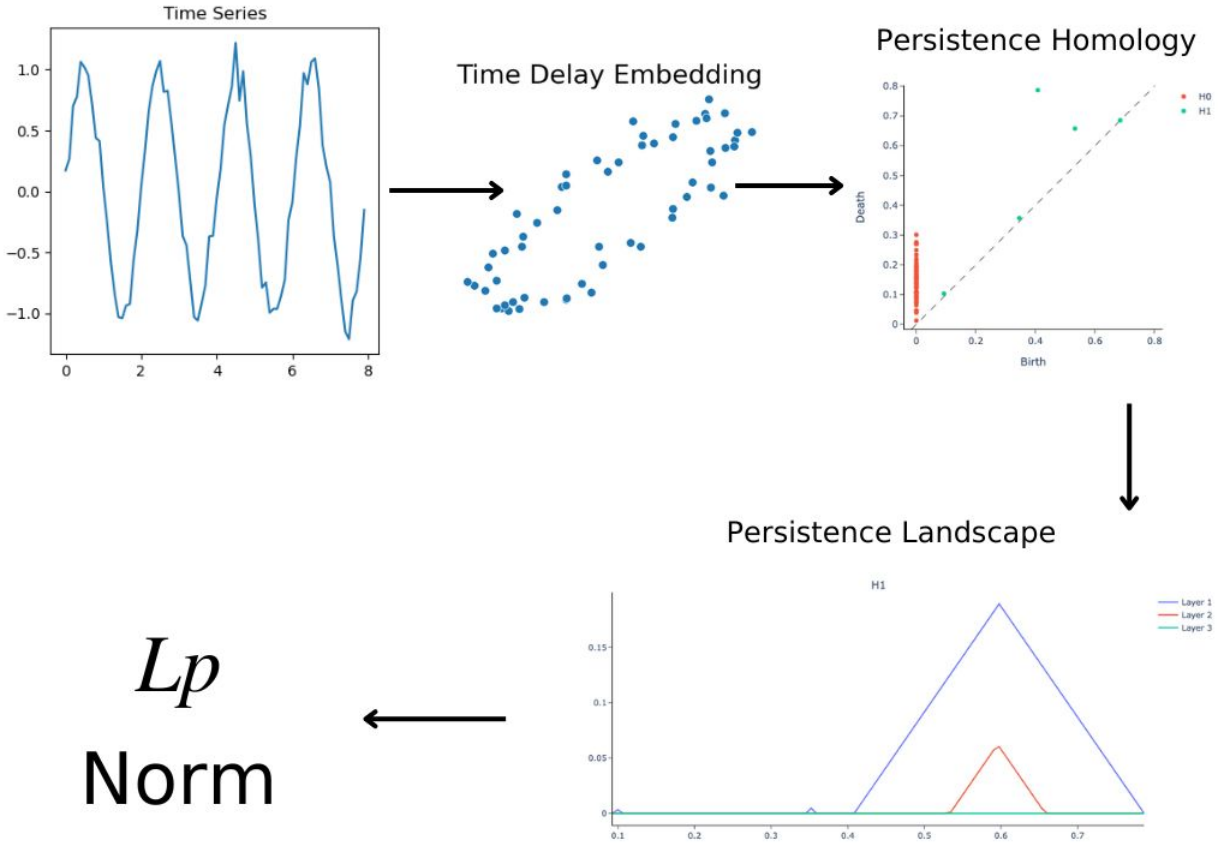

Figure 1: Flowchart Representation of TDA Process with One Step Time Delay Embedding of a Time Series

3 Methods

3.1 Sliding Window Embeddings

Consider a time series of size n, $Y = (y_1, y_2, ..., y_n)$. Let τ be the delay of the time series embedding. In our experiments, we perform time delay embeddings of the time series as following: hold Y constant and shift a time delayed Y by delay τ. Thus, the embedding becomes a multivariate time series or form: $Y^{'} = (Y_1, Y_2)$ where $Y_1 = (y_1, y_2, ..., y_{n-\tau})$ and $Y_2 = (y_{1+\tau}, y_{2+\tau}, ..., y_n)$. Note our embedded time series is of form: $Y' \in \mathbb{R}^2$, with matrix representation of 2 x (n - τ) (resulting from dropping NA values due to time series shift, which are easily seen in our results).

Being that our experiments involve sin waves, we introduce a modification to the traditional time delay embedding: phase delay embeddings. Let $Y = (y_1, y_2, ..., y_n)$ be a sin wave time series with period $P \in \mathbb{N}$. Letting $i \in \{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}, ...\}$, we write $\phi = i\pi$ as the phase delay of the time series embedding. Hence, $\tau = \frac{P}{2\pi}\phi$ as the corresponding time delay. Thus, we can write the phase delay embedding in the form of the time delay embedding.

Having our embedded time series, we next apply a sliding window of size w. Given an embedded time series, $Y^{'} = (y_1^{'}, y_2^{'}, ..., y_{n-\tau}^{'}) \in \mathbb{R}^2$, we write the sliding window embedding:

$$Y_{sw}^{'} = (\{y_{sw}^{'}\}_1, \{y_{sw}^{'}\}_2, ..., \{y_{sw}^{'}\}_{n-\tau-w+1})$$

$$= [(y_1^{'}, y_2^{'}, ..., y_w^{'}), (y_2^{'}, y_3^{'}, ..., y_{w+1}^{'}), ..., (y_{n-\tau-w+1}^{'}, y_{n-\tau-w+2}^{'}, ..., y_{n-\tau}^{'})]$$

Note, each point cloud in the sliding window embedding, $\mathcal{P}_n = \{y_{sw}^{'}\}_n \in \mathbb{R}^2$, is formed by 2 x w matrix. The TDA process (outlined in section 3.1) is then applied for each window in $Y_{sw}^{'}$, while simultaneously calculating variability and correlation. Recall that variability is defined as follows: Given a point cloud, $\mathcal{P} \in \mathbb{R}^2$, we have the covariance matrix $C_{\mathcal{P}}$. We write the variability of the point cloud as one-half the trace of the covariance matrix:

$$Var_{\mathcal{P}} = \frac{Tr(C)}{2} = \frac{C_{11} + C_{22}}{2}$$

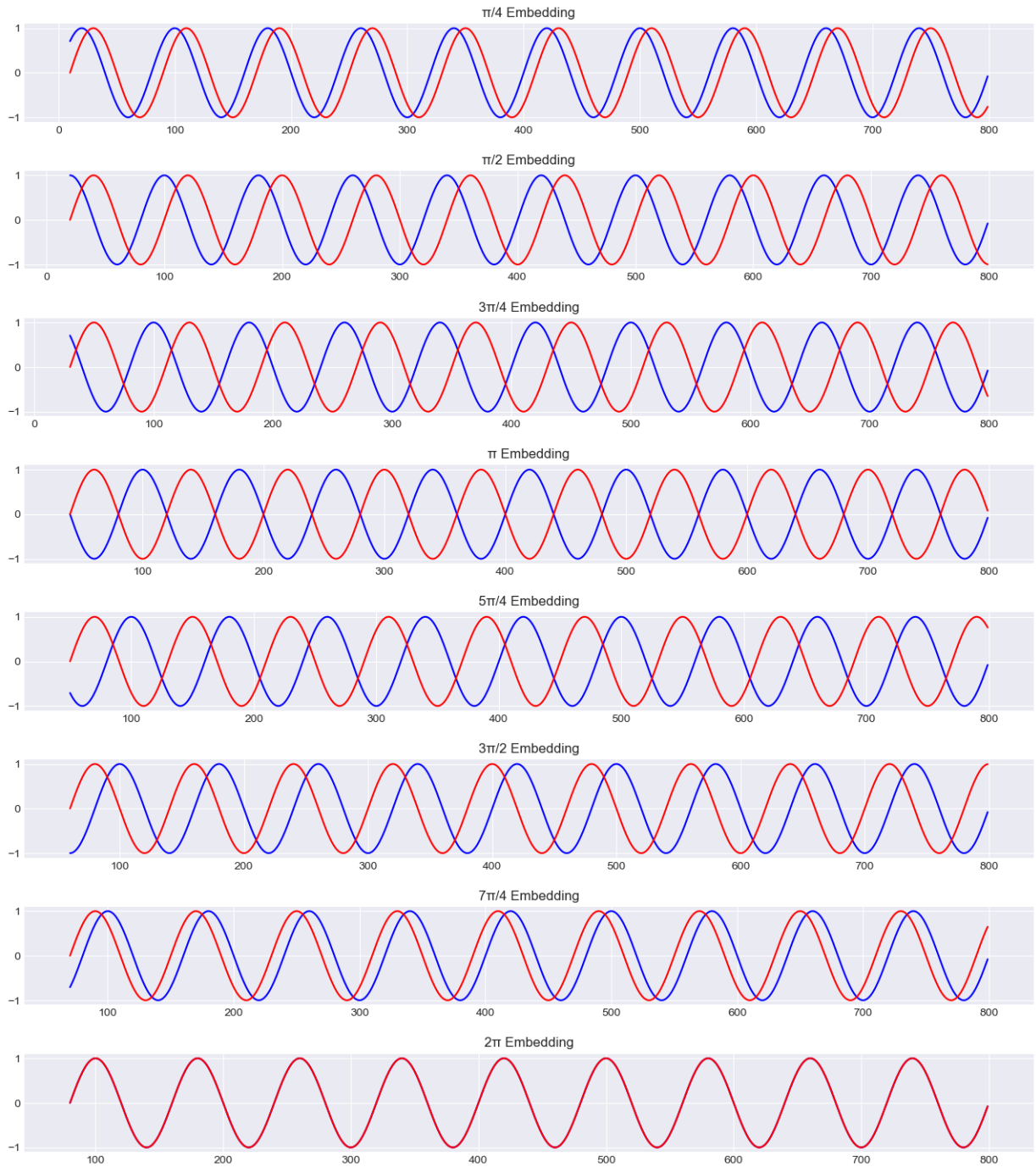

Figure 2: Phase Delay Embeddings within One Period of a Sin Wave

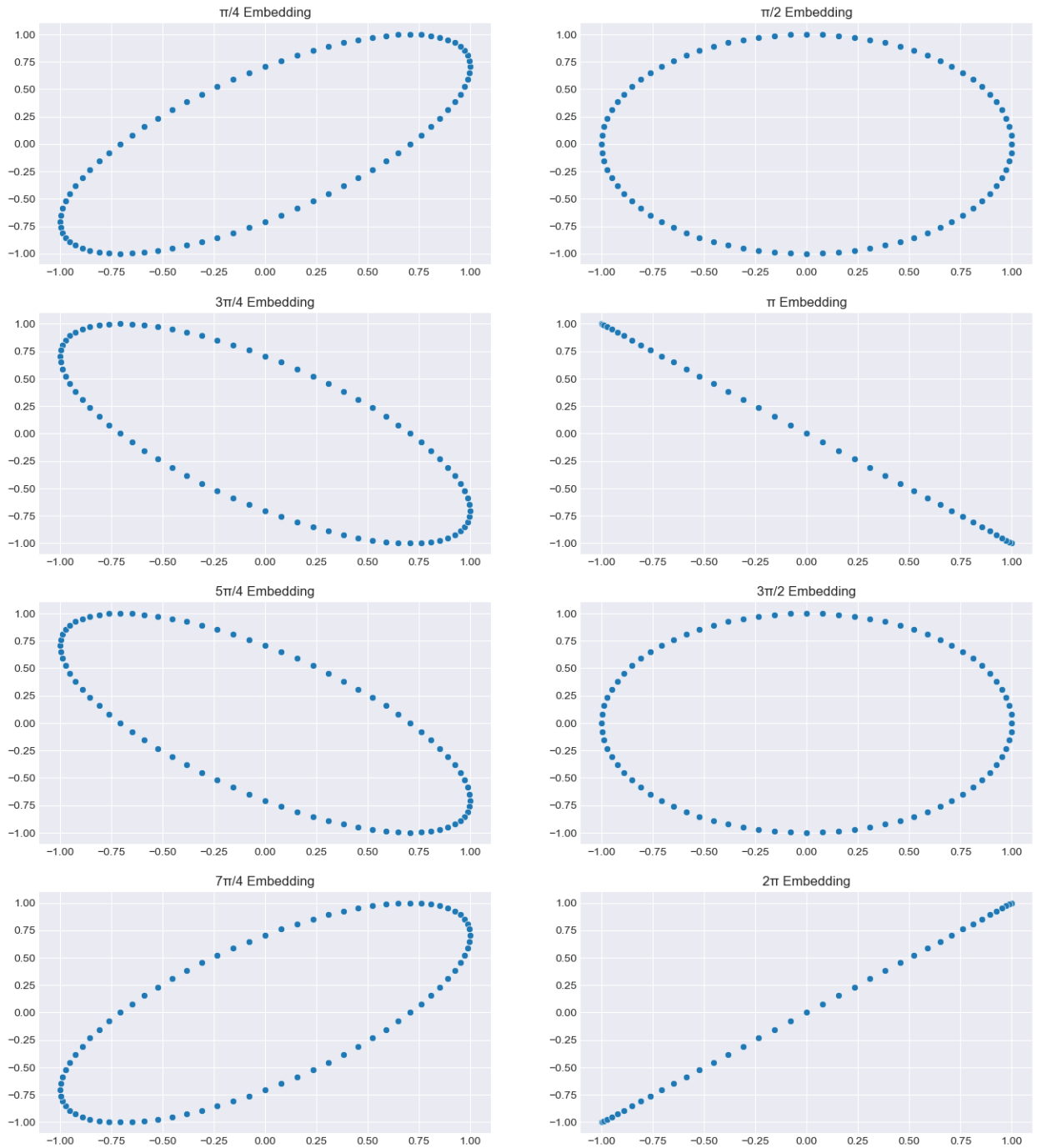

Figure 3: Scatter Plot Representation of Figure 2

3.2 Simplicial Complexes

Suppose F is a collection of sets. The nerve is defined as an abstract simplicial complex consisting of all non-empty sub collections whose sets have non empty intersection. Nerve(F) = $\{X \subset F | \cap X \neq \emptyset\}$. Furthermore, nerve theorem states that the nerve of F and the union of the sets in F have the same homotopy type [6].

3.2.1 Čech Complex

We define $\{B(z, \epsilon)\}_{z \in Z}$ as the Euclidean balls with radius ϵ around the points in point cloud Z. $[z_i, z_j]$ is considered an edge of the complex iff $B(z_i, \epsilon) \cap (z_j, \epsilon) \neq \emptyset \leftrightarrow d(z_i, z_j) < 2\epsilon$. $[z_i, z_j, z_k]$ is considered a triangle of the complex iff $B(z_i, \epsilon) \cap (z_j, \epsilon) \cap (z_k, \epsilon) \neq \emptyset$. We define the Čech Complex as $R_{\check{C}ech}(Z, \epsilon) = \{\sigma \subset Z : \cap_{z \in \sigma} B(z, \epsilon) \neq \emptyset\} \cong \text{Nerve}\{B(z_i, \epsilon) : z_i \in Z\}$ [6].

3.2.2 Vietoris-Rips Complex

As opposed to the Čech Complex, Rips Complex is defined as the following: $R_{Rips}(Z, \epsilon) = \{\sigma \subset Z : diam(\sigma) \leq 2\epsilon\}$. Similar to Čech Complex, $[z_i, z_j]$ is an edge iff $d(z_i, z_j) < 2\epsilon$. Furthermore, $[z_i, z_j, z_k]$ is a triangle of the Rips given:

$$d(z_i, z_j) < 2\epsilon$$
$$d(z_j, z_k) < 2\epsilon$$
$$d(z_i, z_k) < 2\epsilon$$

Note an advantage of the Rips Complex compared Čech Complex is that only distances between pairs of two points need to be computed [6, 7].

3.3 Simplicial Homology

Given a simplicial complex, K, we can compute the p-dimensional homology: $H_p(K)$. Thus, $H_0(K)$ corresponds to the connected components of K, $H_1(K)$ corresponds to the

'holes' in K and $H_2(K)$ corresponds to the 'cavities' of K [7]. In this paper, we focus on the 1-dimensional homologies, representing holes in 2D point clouds.

3.4 Persistence Homology

Given a simplicial complex, K, we write function $f : K \to \mathbb{R}$, with f being monotonic. Letting m be the number of simplices in K, we get n + 1 ≤ m + 1 subcomplexes, and thus can write:

$$\emptyset \subset K_0 \subset K_1 \subset ... \subset K_n = K$$

This sequence of complexes is referred to as filtration. Note that in the context of Rips filtration, the sequence is determined by the set of balls with radii $0 < \epsilon_0 < \epsilon_1 < ... < \epsilon_n$. This filtration corresponds to a series of homology groups which are connected by homeomorphisms [6]:

$$0 = H_p(K_0) \to H_p(K_1)... \to H_p(K_n) = H_p(K)$$

Let α be a homology class in $H_p(K_{\epsilon_j})$. Each α has a birth value when $\alpha \in H_p(K_{\epsilon_b})$, and death value when α merges with another class in $H_p(K_{\epsilon_d})$. Thus, we can write $Pers(\alpha) = \epsilon_d - \epsilon_b$. If α is born but never dies, we write: $Pers(\alpha) = \infty$. One can formally write the definition of persistence as the following [6]:
For:

$$H_p(K_{\epsilon i}) \xrightarrow{f_p^{i,j}} H_p(K_{\epsilon j})$$

Define:

$$H_p^{i,j} = f_p^{i,j}(H_p(K_{\epsilon_i}))$$

α is **born** in K_{ϵ_i} iff:

$$\alpha \in H_p(K_{\epsilon_i}) \text{ and } \alpha \notin H_p^{i-1,i}$$

α **dies** in K_{ϵ_i} iff:

$$f_p^{i,j-1}(\alpha) \notin H_p^{i-1,j-1} \text{ and } f_p^{i,j}(\alpha) \in H_p^{i-1,j}$$

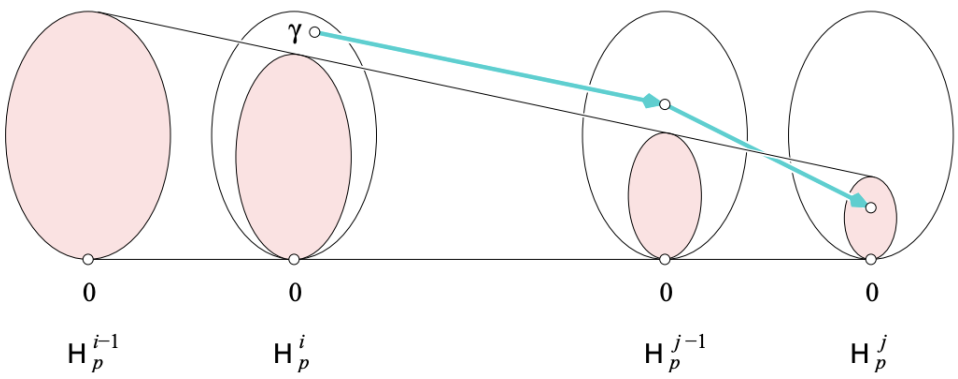

Figure 4: Illustration of Birth Death Process of α (denoted here as γ) in Filtration. Figure adapted from [6].

3.4.1 Persistence Diagram

Let α be a homology class where $\alpha \in H_p(K_{\epsilon_i})$. Let $b_\alpha = \epsilon_1$ be the point at which α is born, and $d_\alpha = \epsilon_2$ be the point at which α dies. One can assign point $z_\alpha = (b_\alpha, d_\alpha)$, known as the 'birth death' pair of α. Plotting each $z_\alpha \in \mathbb{R}^2$, we get the persistence diagram of the filtration: $\text{dgm}_p(f)$ [6]. Note that all points lie above the diagonal of the persistence diagram. In this paper, we construct persistence diagrams via Rips filtration. We compute the persistence diagram using Giotto TDA python package [14].

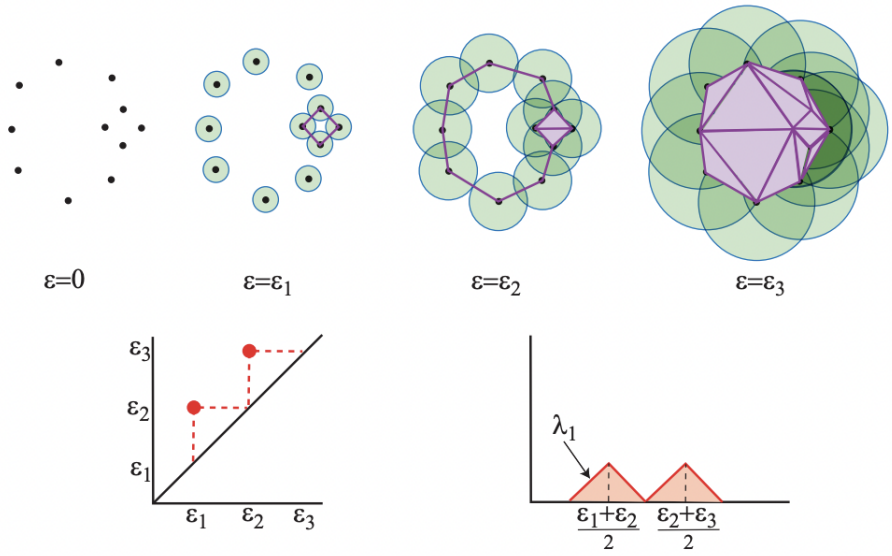

Figure 5: Illustration of Rips Filtration with Corresponding Persistence Diagram and Landscape. Figure adapted from [9]

3.5 Persistence Landscape

Let dgm_p denote a persistence diagram. Given each birth death point, $z_\alpha \in \text{dgm}_p$, we write piecewise linear function:

$$f_{(b_\alpha,d_\alpha)}(x) = \begin{cases} x - b_\alpha & \text{if } x \in (b_\alpha, \frac{b_\alpha+d_\alpha}{2}) \\ -x + b_\alpha & \text{if } x \in (\frac{b_\alpha+d_\alpha}{2}, d_\alpha) \\ 0 & \text{if } x \notin (b_\alpha, d_\alpha) \end{cases} \tag{1}$$

Thus, we can write the persistence landscape, $\lambda = (\lambda_k)_{k\in\mathbb{N}}$ as:

$$\lambda_k = k - max\{f_{(b_\alpha,d_\alpha)}(x) \mid (b_\alpha, d_\alpha) \in \text{dgm}_p\} \tag{2}$$

Where k-max is the k^{th} largest value of the function [3]. Similar to persistence diagrams, we calculate persistence landscapes using Giotto TDA python library [14].

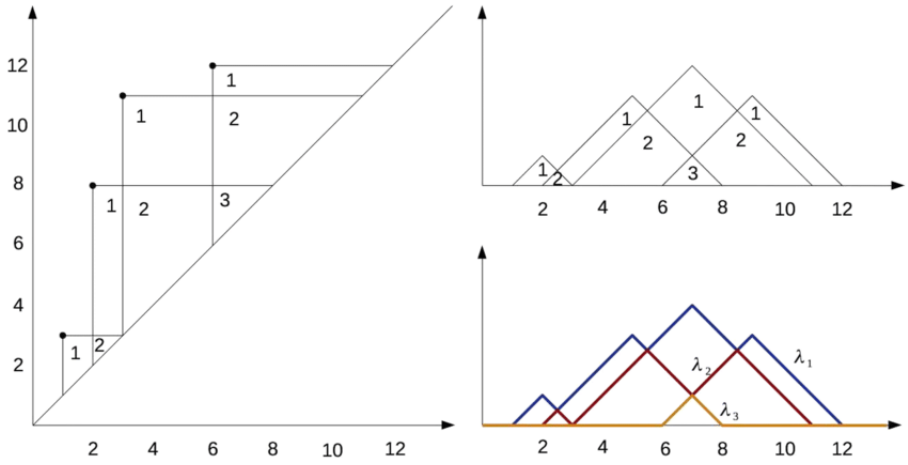

Figure 6: Illustration of the construction of persistence landscape. Figure adapted from [12]

3.6 L_p Norms

Given a persistence landscape, $\lambda = (\lambda_k)_{k \in \mathbb{N}}$, we define the L_p norm of the landscape [3]:

$$||\lambda||_p = (\sum_{k \in \mathbb{N}} ||\lambda_k||_p^p)^{(1/p)} \tag{3}$$

3.6.1 Properties of L_p Norms

Existing literature shows that persistence landscapes are elements of the Banach space, $L^p(\mathbb{N} \text{ x } \mathbb{R})$. Thus, statistical theorems and methods can be applied to the persistence landscape (such as the central limit theorem, confidence intervals, hypothesis testing, and statistical inference) [12].

A key property of L_p norms used in this paper is the relationship between L_1 norm and the covariance matrix of point clouds from the SWE. The literature shows that on average, L_1 norms of persistence landscapes of a point cloud are proportional to the variance of the point cloud. Additionally, average L_1 norms of persistence landscapes of a point cloud are

inversely proportional to the magnitude of correlation of the point cloud.

Given a point cloud of size N, $\mathcal{P} \in \mathbb{R}^2$, consider the covariance matrix $C_{\mathcal{P}}$:

$$C_{\mathcal{P}} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_2\sigma_1 & \sigma_2^2 \end{bmatrix}$$

Note $\rho \in (-1, 1)$ denotes the correlation coefficient. Following closely to [2], the number of 1D homologies in a point cloud of size N is roughly bounded by the sum:

$$n_1 = \sum_{i=4}^{N} \binom{N}{i} \leq 2^N \tag{4}$$

Where i = 4, 5, ..., N points is the size of the clusters used to calculate 1D homologies. Note that the minimum cluster size required to construct the Vietoris Rips complex is i = 4. Following proposition 1 established in [2], we write:

$$E(||\lambda||_p) \leq n_1(N) \cdot N \cdot Tr(C_{\mathcal{P}})^{\frac{p+1}{2p}} \tag{5}$$

Assume $\sigma_1 > \sigma_2$. Let θ_1 and θ_2 be the eigenvalues of $C_{\mathcal{P}}$. Note that $\theta_1 \geq \theta_2 \geq 0$ because $C_{\mathcal{P}}$ is a symmetric and positive semi-definite matrix. As established in theorem 2 of [2], we write:

$$E(||\lambda||_1) \leq n_1(N)^2 \cdot N \cdot \theta_2 \tag{6}$$

Following remark 2 from [2], we see $E(||\lambda||_1) \to 0$ when $\rho \to \pm 1$.

4 Results

4.1 L1 norms of Persistence Topological Landscapes; Time-delay Embeddings of Sin Waves

We begin the analysis of L_1 norms of time-delay embeddings using periodic time series of sin curves. In each case of sin experiments, we simulate a curve of the following form:

$$f(x) = sin(\frac{\pi x}{4}) + \epsilon$$

$$\epsilon \sim \mathcal{N}(0, \sigma)$$

We choose a sin wave with frequency $\frac{\pi}{4}$. We sample the curve using a step size of .1, leading to a period of $2\pi * \frac{4}{\pi} * 10 = 80$ data points. We sample the curve for 10 periods, causing the domain of the function to be [0,799.9]. Noise is sampled from the normal distribution, $\epsilon \sim \mathcal{N}(0, \sigma)$, with experiments for various degrees of noise: $\sigma = 0$, 0.01, 0.1, 0.3, 1, 3 (note $\sigma = 0$ being the case of no added noise).

Using a sliding window with size of 2 periods (160 points), we calculate the L_1 norms, variability and correlation for each point cloud of a particular embedding. We perform this experiment for multiple embeddings within 2 periods of the curve, specifically phase shift embeddings of: $\frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}, \pi, \frac{5\pi}{4}, \frac{3\pi}{2}, \frac{7\pi}{4}, 2\pi, \frac{9\pi}{4}, \frac{5\pi}{4}, \frac{11\pi}{4}, 3\pi, \frac{13\pi}{4}, \frac{7\pi}{4}, \frac{15\pi}{4}, 4\pi$.

4.1.1 No Noise

Our first experiment involves sin curves without added noise, representing a time series with purely dynamic cyclicality. We use this as a basis of understanding for L_1 norms when comparing to time series which do have noise. Figure 7 shows L_1 norms (normalized to maximum), variability (normalized to maximum) and correlation for multiple phase shift embeddings in 1 period of the sin wave. Unsurprisingly in the purely dynamic case, for each embedding, values of normalized L_1 norms, normalized variability and correlation stay

constant. In the cases of $n\pi$ embeddings (ie. when the time series and its lag are in phase and completely out of phase), L_1 norms are 0. This makes sense intuitively, and can easily be seen in figure 3, where π and 2π embeddings are precisely linear. Hence, there is no topological information to be extracted (ie. holes in the point clouds). Furthermore, this confirms existing research on the relationship between correlation and norms [2]. Increase correlation (magnitude) of time series leads to smaller norms, with the case of $\rho = \pm 1$ leading to norms of 0. Hence, in cases where embeddings are in phase/ completely out of phase, correlation of the point clouds are ± 1, there are no holes in the point cloud, and norms are 0.

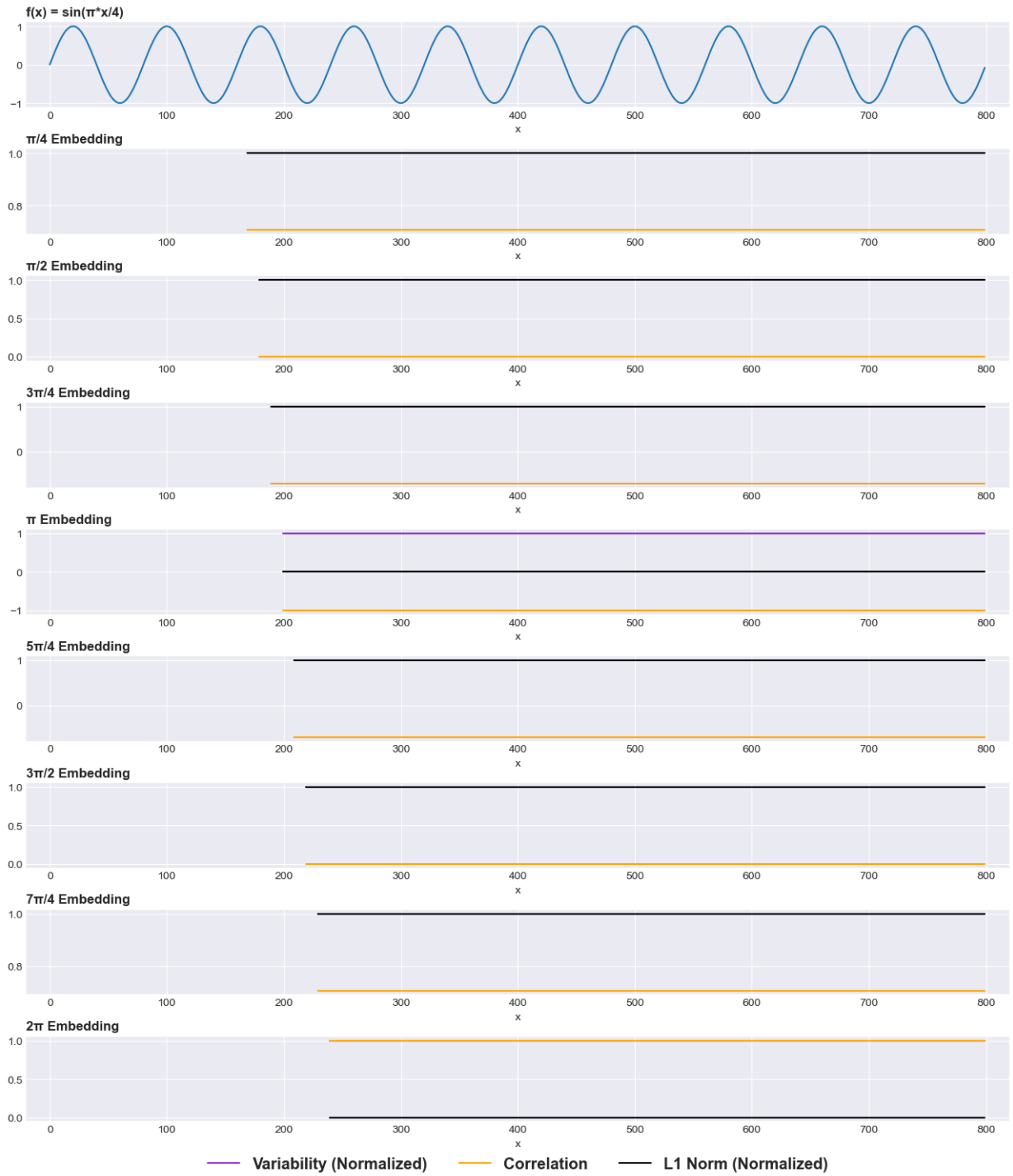

Figure 7: L_1 Norms of Sin Wave Embeddings with No Noise

4.1.2 Role of Added Noise

We then move to experiments involving sin waves with noise added. Similar to the case with no noise, we analyze the norms of multiple phase shift embeddings in the various cases of added noise explained in section 4.1.

Similarly to section 4.1, we simulate a sin curve of the following form: $f(x) = sin(\frac{\pi x}{4}) + \epsilon$; where in this case of 'small' noise, $\epsilon \sim \mathcal{N}(0, \sigma = 0.1)$. Figure 8 shows the scatter plots for each embedding within one period of the sin wave. Clearly, the dominant homology (hole) in each of these embeddings are weaker than in the case of no noise. Furthermore, homologies are clearly formed within the noise itself, most notably in the cases of $n\pi$ embeddings. As opposed to the case of no noise, the $n\pi$ embedding point clouds are not linear, meaning all persistent homologies are due to a weak noise. We see the effect of this in figure 9, where normalized L_1 norms visually have higher variance and are more chaotic in the $n\pi$ embeddings, despite variability and correlations staying relatively constant. Contrasted to $\frac{(n+1)\pi}{2}$ embeddings, where L_1 norms visually have less variance.

In the 2π embedding of figure 9, we see a sharp drop in L_1 norm at point x = 430. We see this drop despite variability and correlation staying relatively constant in that same region: $430 \leq x \leq 455$. We analyze this drop closely in figures 10, 11 and 12, detailing the peak point prior to the drop and the trough point following. We then analyze the point clouds at these points, and its associated persistence diagrams and persistence landscapes. The visual trend of each point cloud is extremely similar, in addition to their summary statistics. However, the point cloud at 'peak' visually has larger and more well defined homologies, leading to stronger landscapes and thus higher L_1 norm. We conclude that the dramatic fluctuation in L_1 norm at point x = 430 is due to the disappearance of mini holes in a point cloud. Particularly, white noise can lead to large fluctuations in L_1 norm signals, an incident TDA practitioners must consider when using norms as a statistical measure.

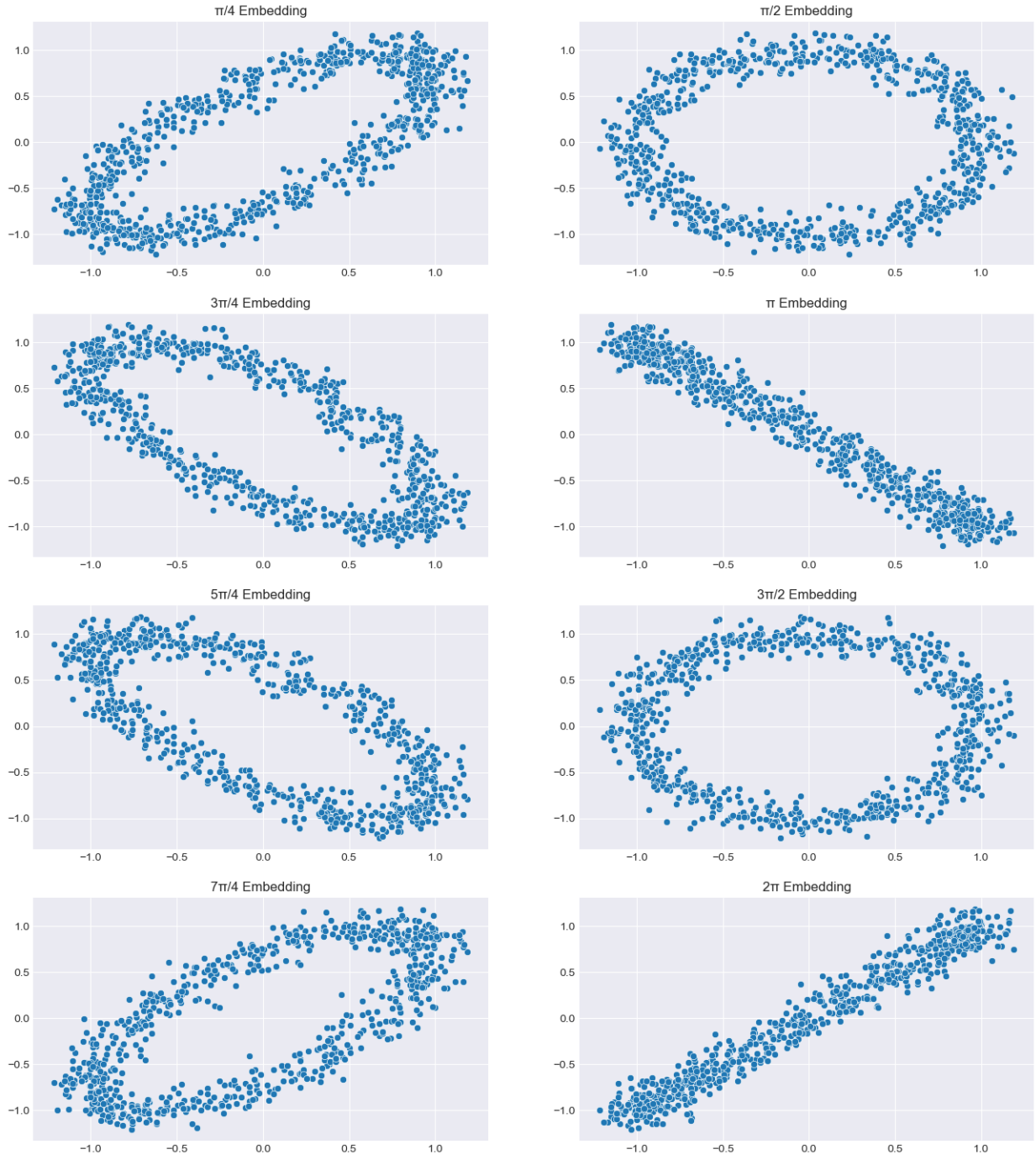

Figure 8: Phase Delay Embeddings Scatter Plots for One Period of a Sin Wave with Added Noise: $\epsilon \sim \mathcal{N}(0, \sigma = 0.1)$

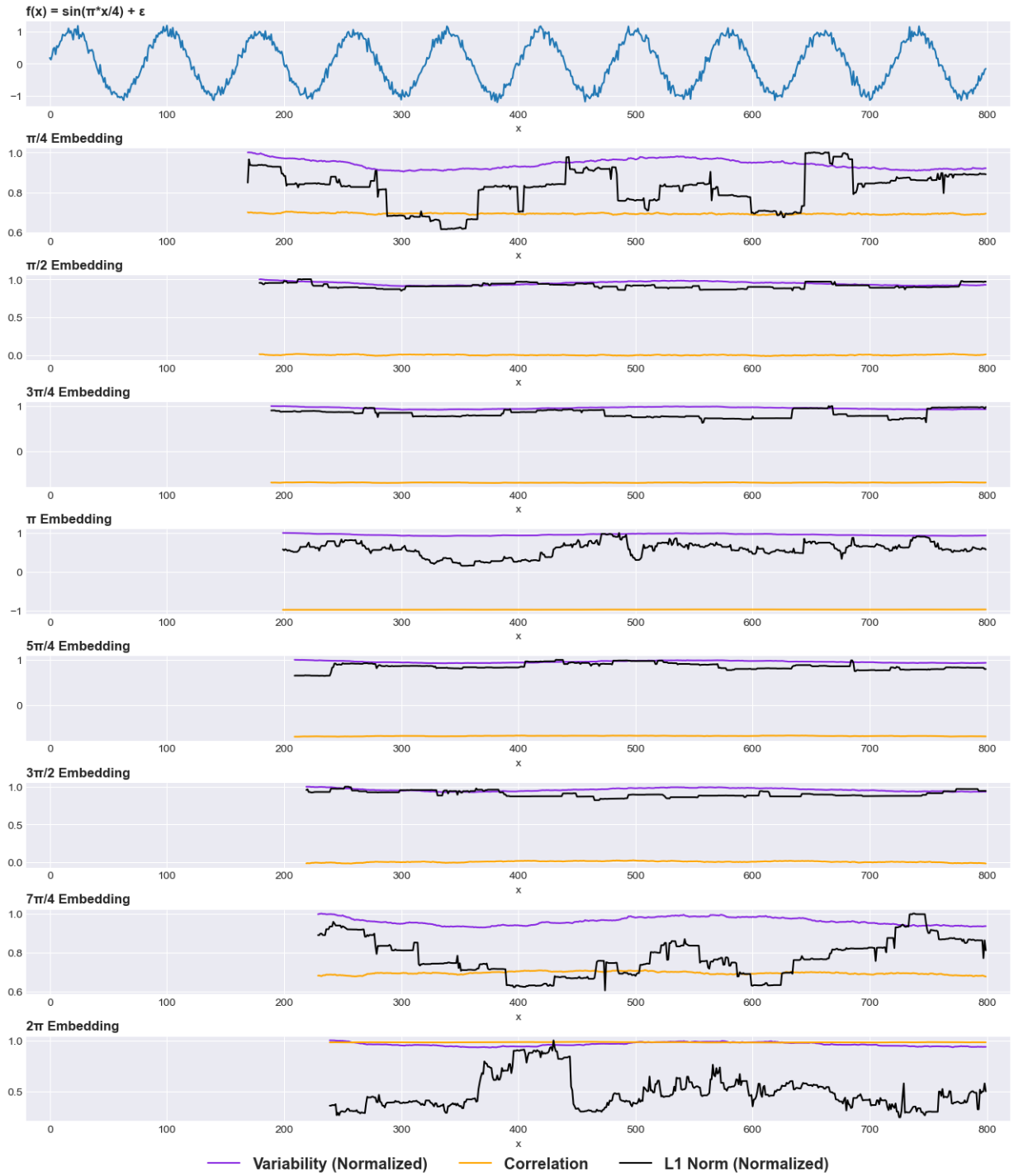

Figure 9: L_1 Norms of Sin Wave Embeddings with Added Noise: $\epsilon \sim \mathcal{N}(0, \sigma = 0.1)$

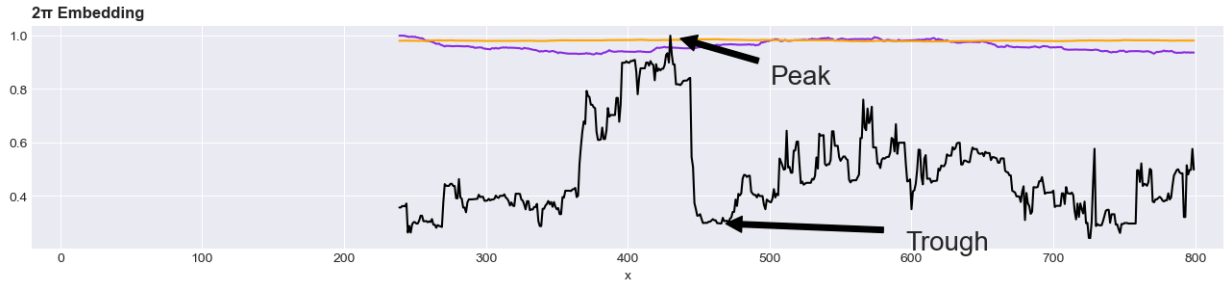

Figure 10: Detailed View of 2π Embedding in Figure 9

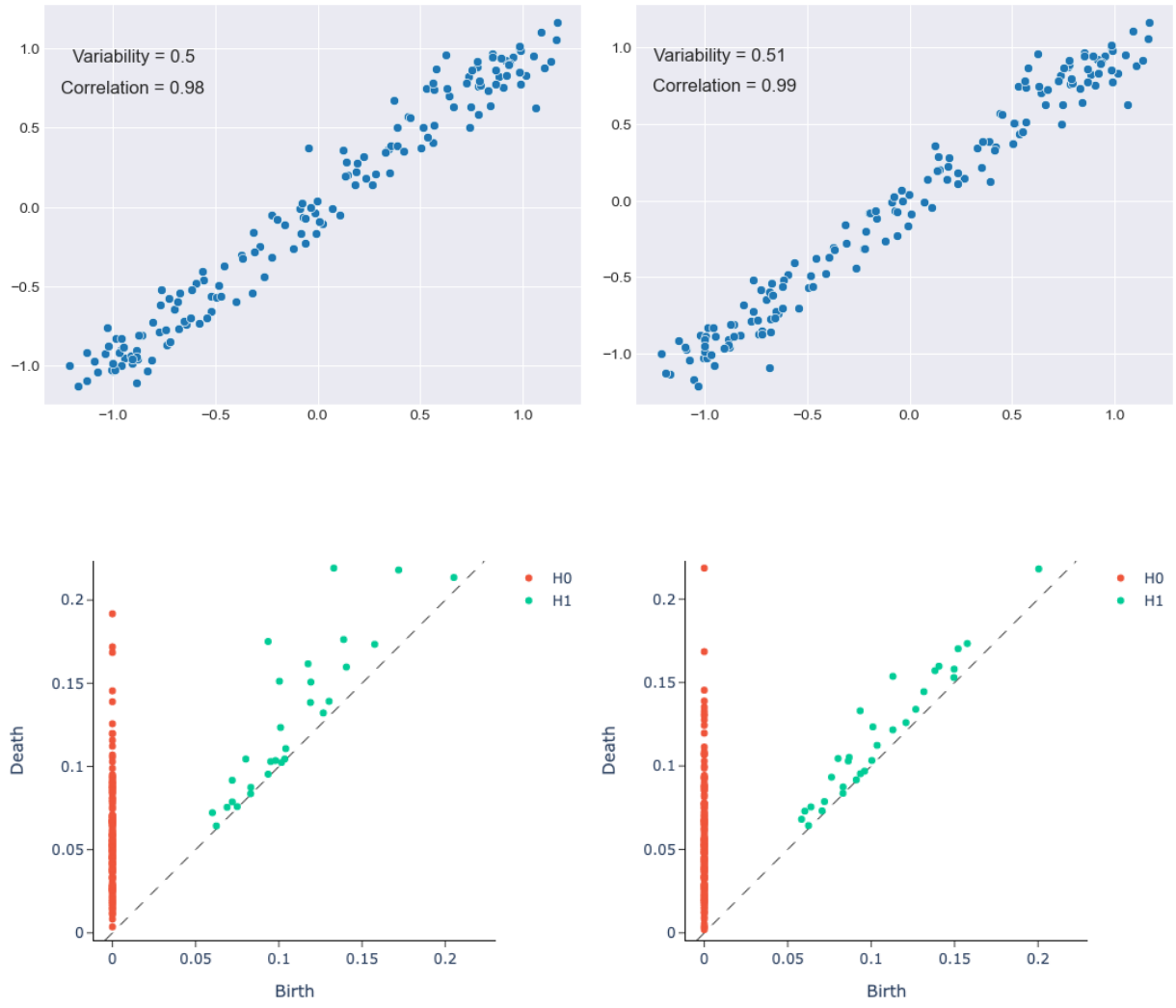

Figure 11: Left-right: peak, trough. Top-down: point cloud, persistence diagram

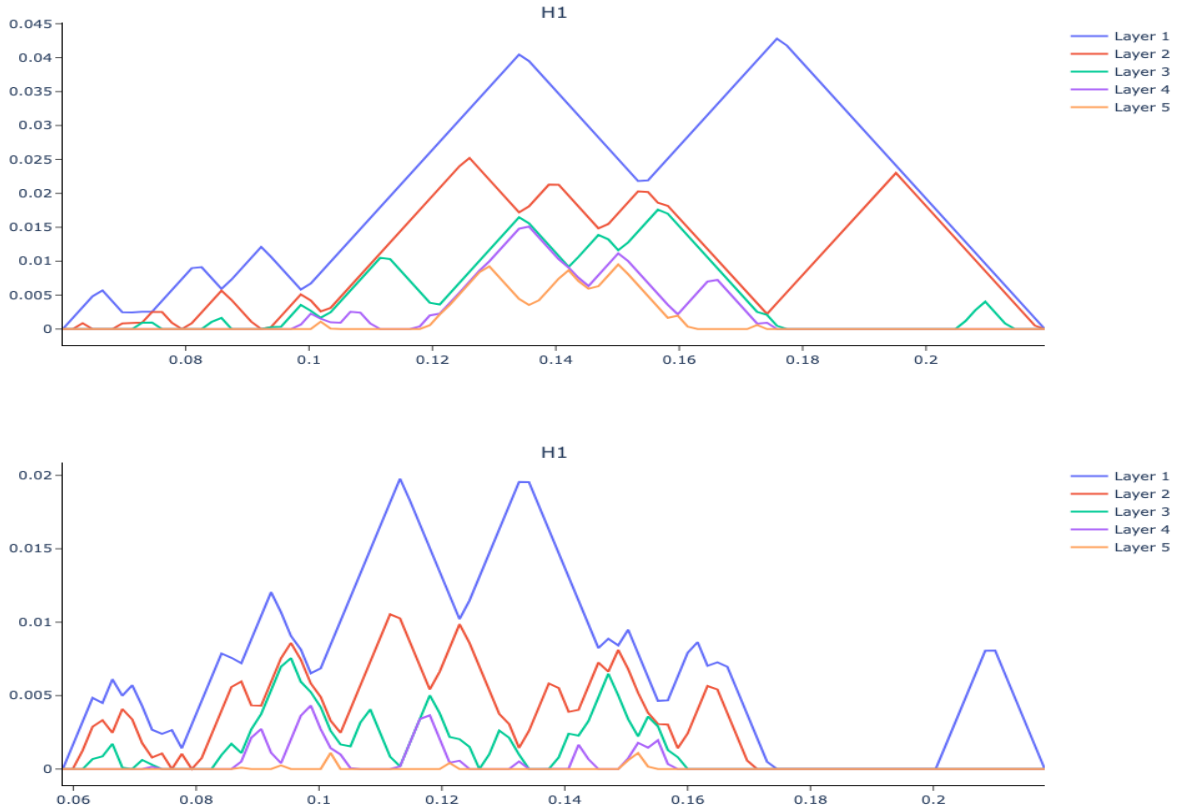

Figure 12: Top-down: persistence landscapes of peak, trough (of figure 10)

4.1.3 Average L_1 Norms of Sin Wave Embeddings

Despite norms being normalized throughout the various embedding analyses shown in figure 9, the magnitude of L_1 norm changes throughout each embedding. Hence, we analyze the average L_1 norm throughout each embedding for various levels of noise. Figure 10 shows the cyclicality of average norms taken among multiple embeddings across various levels of noise, again demonstrating the proportionality of correlation to average L_1 norm. Embeddings $\frac{(n+1)\pi}{2}$ have the highest magnitude of norms, corresponding to it having correlation $\rho = 0$. Meanwhile, embeddings of $n\pi$ have average norm magnitude of 0, as previously shown by its correlation being $\rho = \pm 1$. Adding noise reduces the cyclical signal of average norm, yet the cyclicality is still visible.

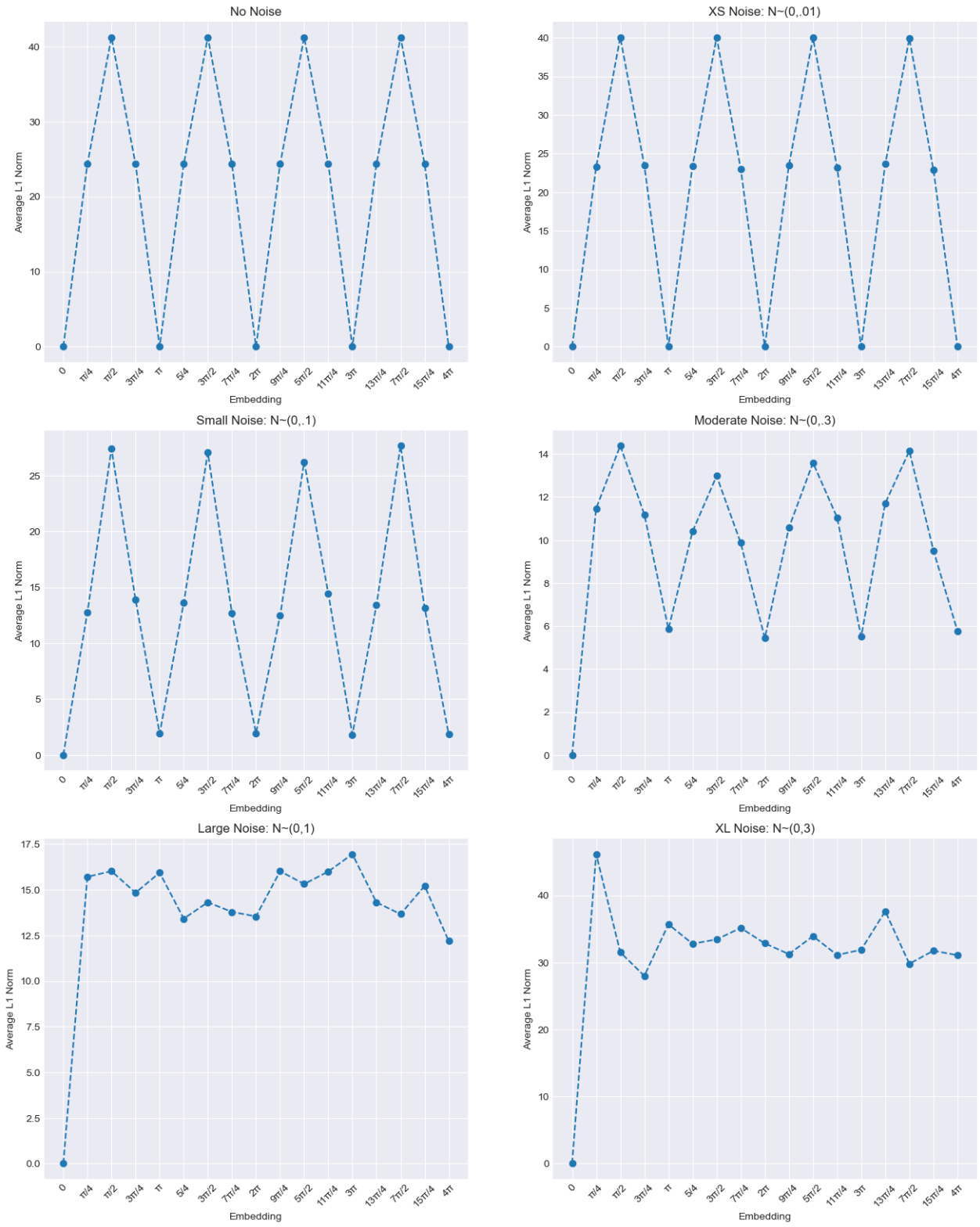

Figure 13: Average L_1 Norms for Sine Wave Embeddings across Various Degrees of Noise

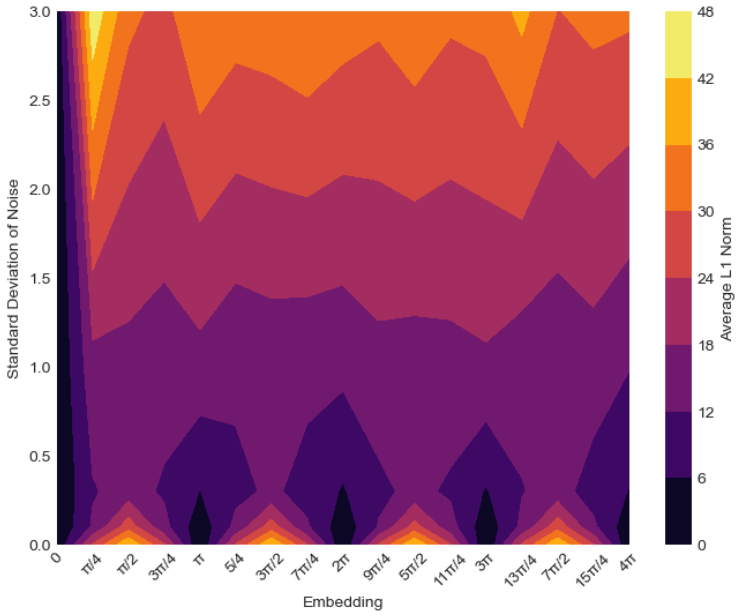

Figure 14: Heatmap Representation of Figure 13

4.2 L1 norms of Persistence Topological Landscapes; Time-delay Embeddings of Auto Regressive Time Series

Moving the analysis from sin simulations, we now experiment using auto regressive time series. As opposed to sin waves, auto regressive models represent periodic time series with a stochastic nature. Specifically, AR(2) models have been shown to have dominant frequencies, and thus are periodic [13]. In this section, we simulate a synthetic AR(2) model representing sunspot data from 1700 - current day, as performed in [13]. We obtain historical sunspot data from [1]. We use R function "arima.sim" to generate the synthetic AR(2) time series with 1000 data points:

$$X_t = 1.381X_{t-1}\text{-}0.6807X_{t-2} + \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$

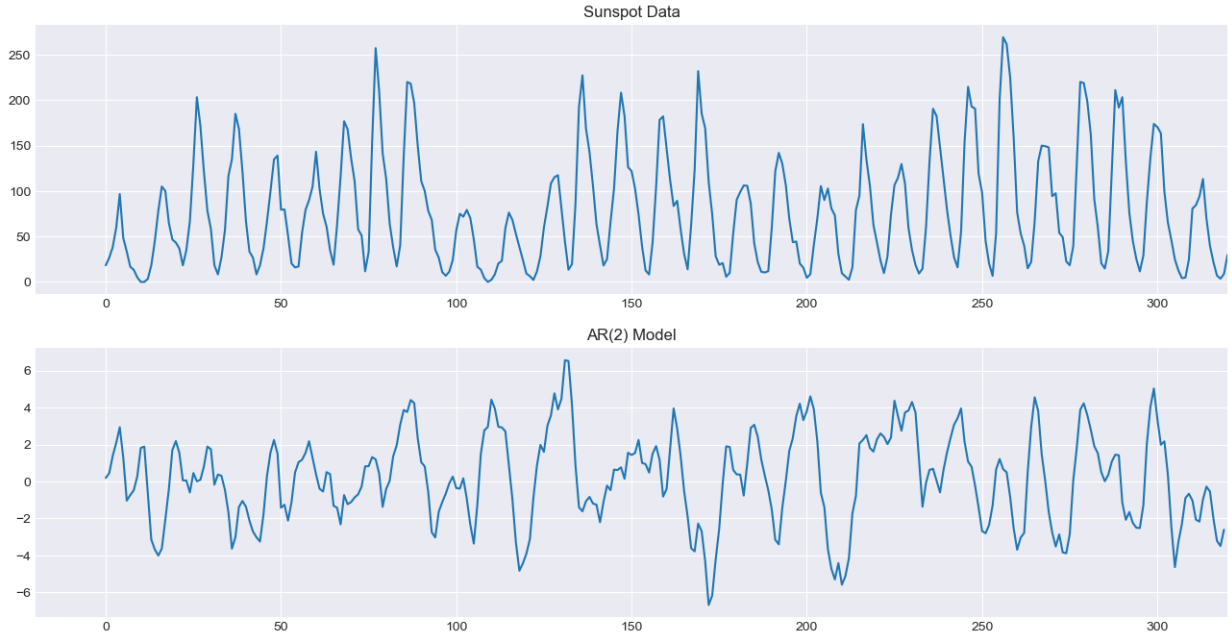

Figure 15: Sunspot and Synthetic AR(2) Time Series (First 320 Points)

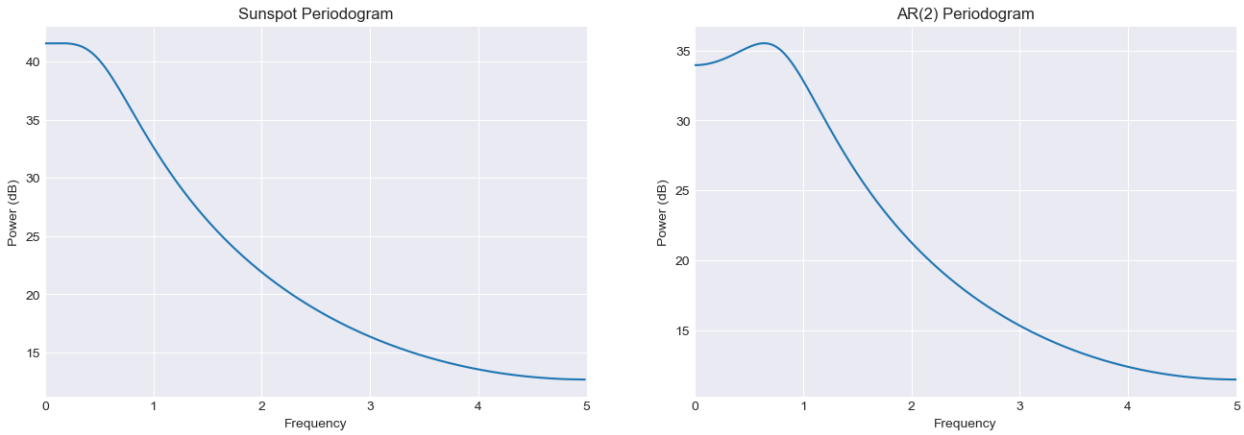

Figure 16: Periodograms of Figure 15

Analyzing the periodogram of our AR model, and following closely to [13], we see our simulated data having a dominant frequency at 0.57, leading to a period of $\mathrm{P} = 2\pi/.57 \approx 11$. Hence, we run the analysis on L_1 norms for embeddings [1-11] of our simulated AR(2) time series.

Using a window size of 50 data points, we calculate L_1 norms (normalized to maximum), variability (normalized to maximum), and correlation for each point cloud of the particular embedding and plot in figure 17. The relationship between L_1 norm and variability is visibly seen for each embedding of figure 17. We see this in figure 17, where variability and L_1 norms have a visible correlation. However, there are a few instances where L_1 norms fluctuate from the proportionality of variability. Most notably in embedding 1, in the region $470 \leq x \leq 500$, where L_1 norm spikes at point x = 485 despite variability and correlation staying relatively constant. Additionally, we see an anomaly in the reverse case of this at point x = 738 of embedding 1, where variability drops significantly while L_1 norms stay relatively constant.

Analyzing these anomalies, we produce the point clouds, persistence diagrams and landscapes for the two cases listed above. The first anomaly is analyzed in figures 18, 19 and 20, where we label the points trough 1, peak spike, and trough 2. Inspecting the 3 point clouds associated with these points, we see their trends to be extremely similar visually. Furthermore, their summary statistics are extremely similar as well. However, the point cloud at 'peak' clearly has more well defined homologies than the troughs. The effect of this is seen in the respective persistence diagrams and landscapes, where the peak has stronger signal than troughs. We conclude that L_1 norms of time delay embeddings are highly sensitive to noise, similar to the example shown in section 4.1.2. We see that fluctuations of L_1 norms can be caused by homologies due to noise, again reiterating that signals in L_1 norms can be caused by random homologies.

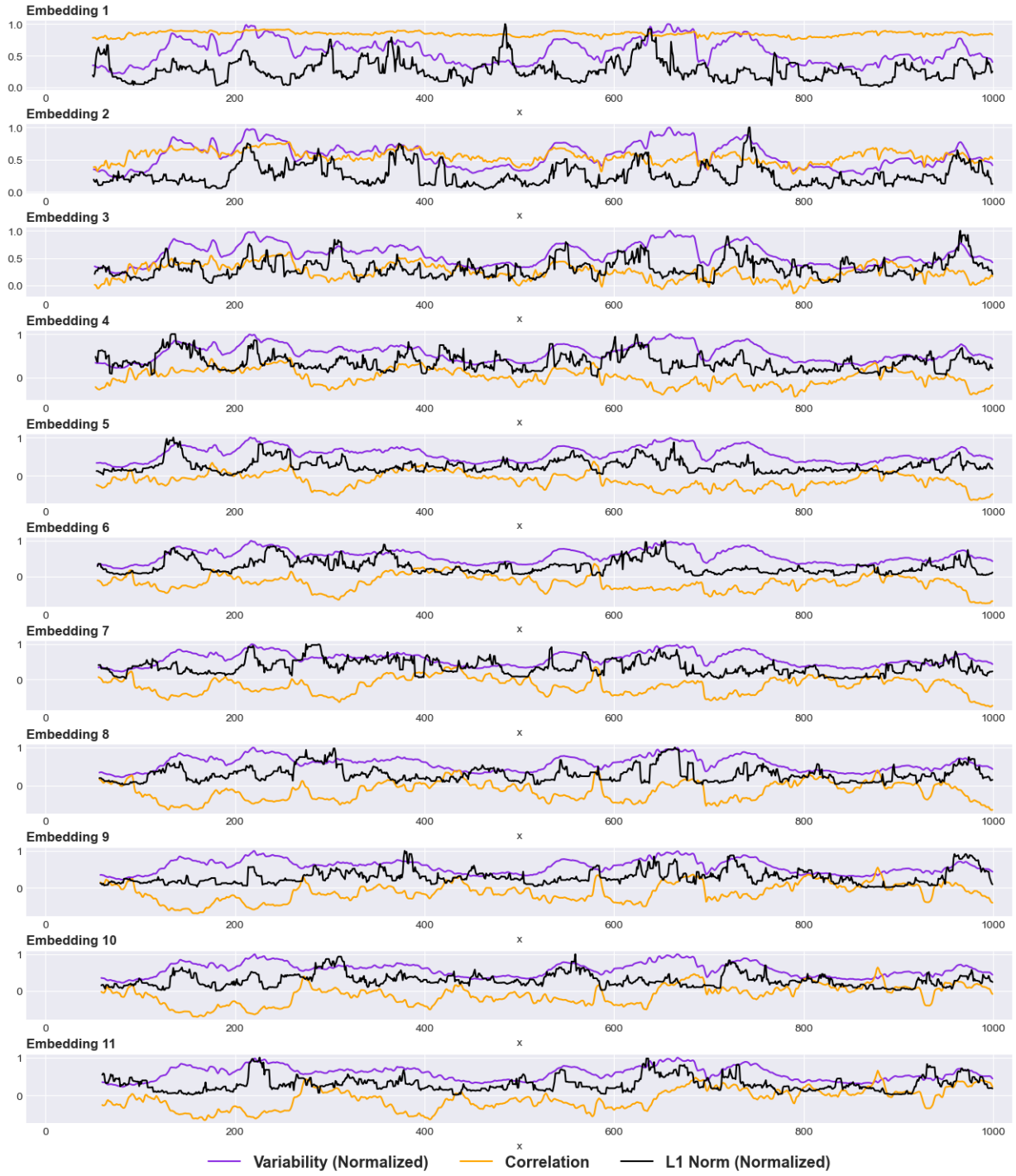

Figure 17: L_1 norms of Embeddings [1-11] of the AR(2) Synthetic Time Series

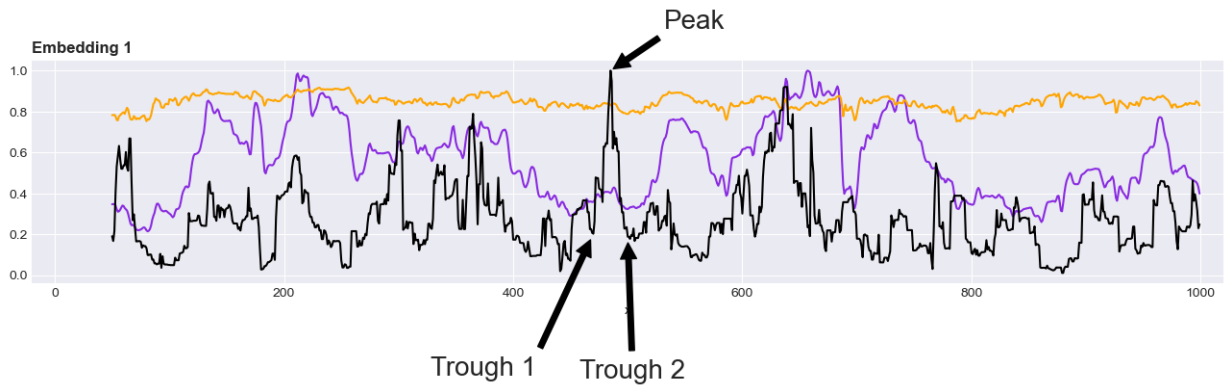

Figure 18: Detailed View of Embedding 1 in Figure 17

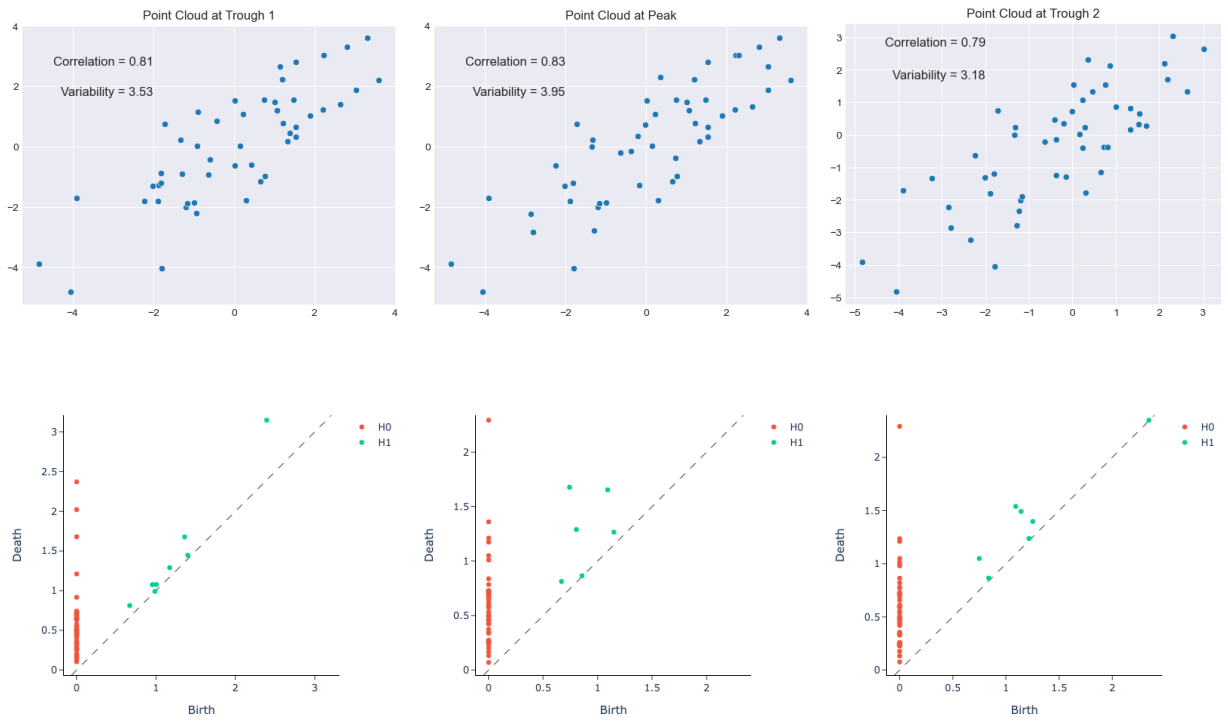

Figure 19: Left-right: trough 1, peak, trough 2. Top-down: point cloud, persistence diagram (of figure 18)

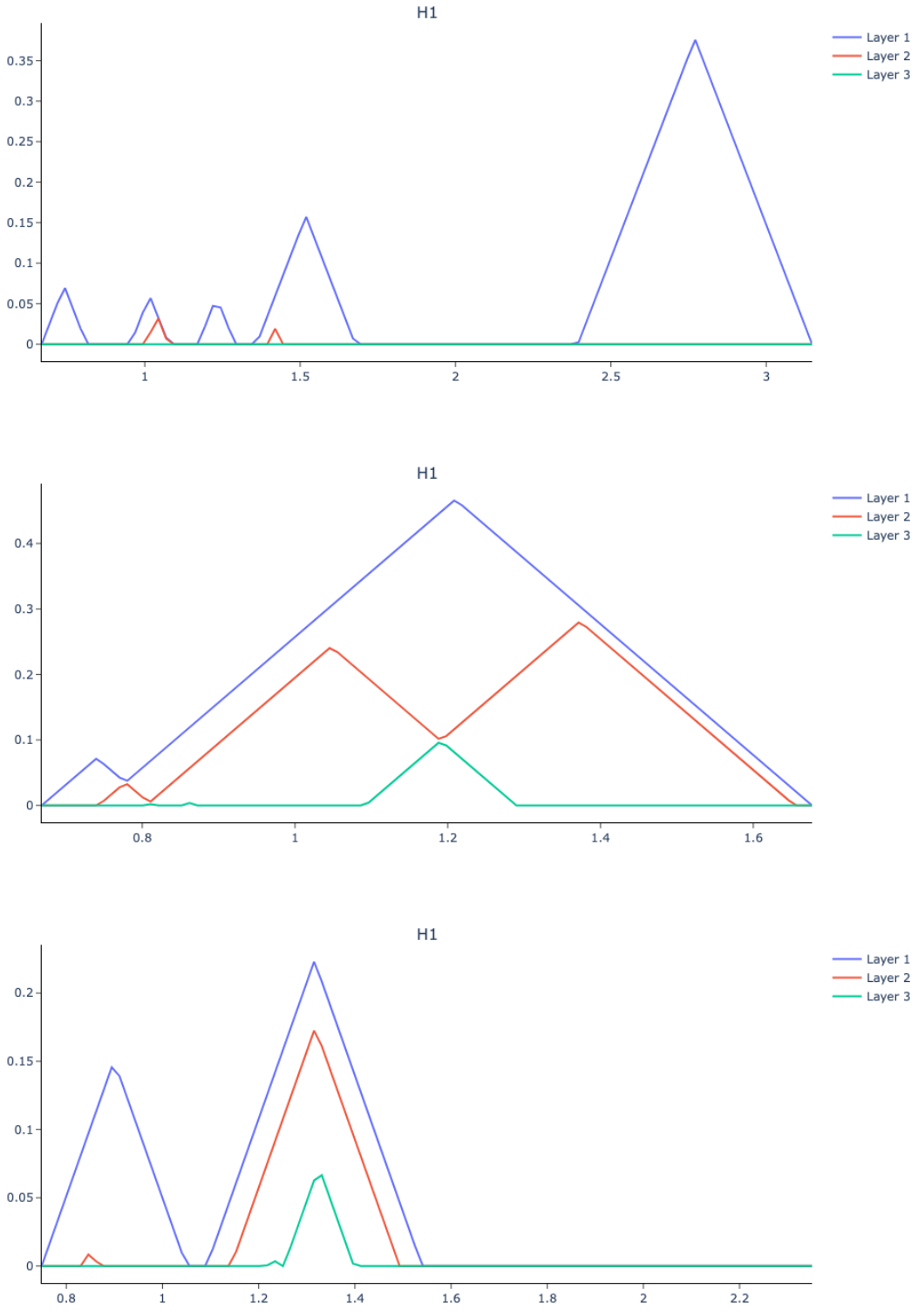

Figure 20: Top-down: persistence landscapes of trough 1, peak, trough 2 (of figure 18)

The anomaly analyzed in figures 18, 19 and 20 show a spike in L_1 norms despite variability staying relatively constant. We now look at another anomaly in the reverse case, where variability drops while L_1 norms stay relatively constant. Point x=738 shows a significant drop in variability, which we highlight in figure 21.

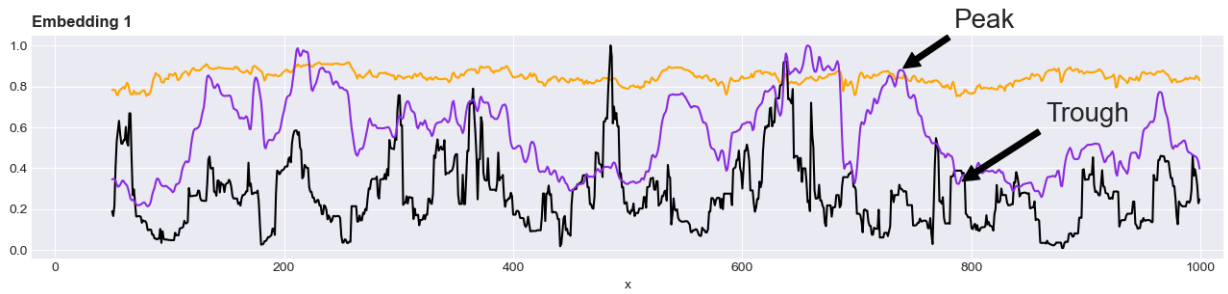

Figure 21: Detailed View of Embedding 1 in Figure 17

In figure 22 and 23, we plot the point clouds, persistence diagrams and landscapes at points x = 738 and x = 788. As opposed to the anomaly highlighted in figure 19, these points clouds do differ in their summary statistics. Of course, the peak point cloud has higher variability, in addition to its higher correlation. Furthermore, the peak's point cloud visually has a bigger homology, which can be verified in the persistence diagram. However, the trough's point cloud has many smaller homologies, leading to its L_1 norm having similar magnitude.

We see here that the effect of the drop in variability on L_1 norms is dampened by homologies due to noise. In other words, we do not see the expected drop of L_1 norm (due to its expected proportionality with variability) on the period $738 \leq x \leq 788$. Therefore, we conclude that random homologies can suppress the effect of a drop in variability on L_1 norm signals.

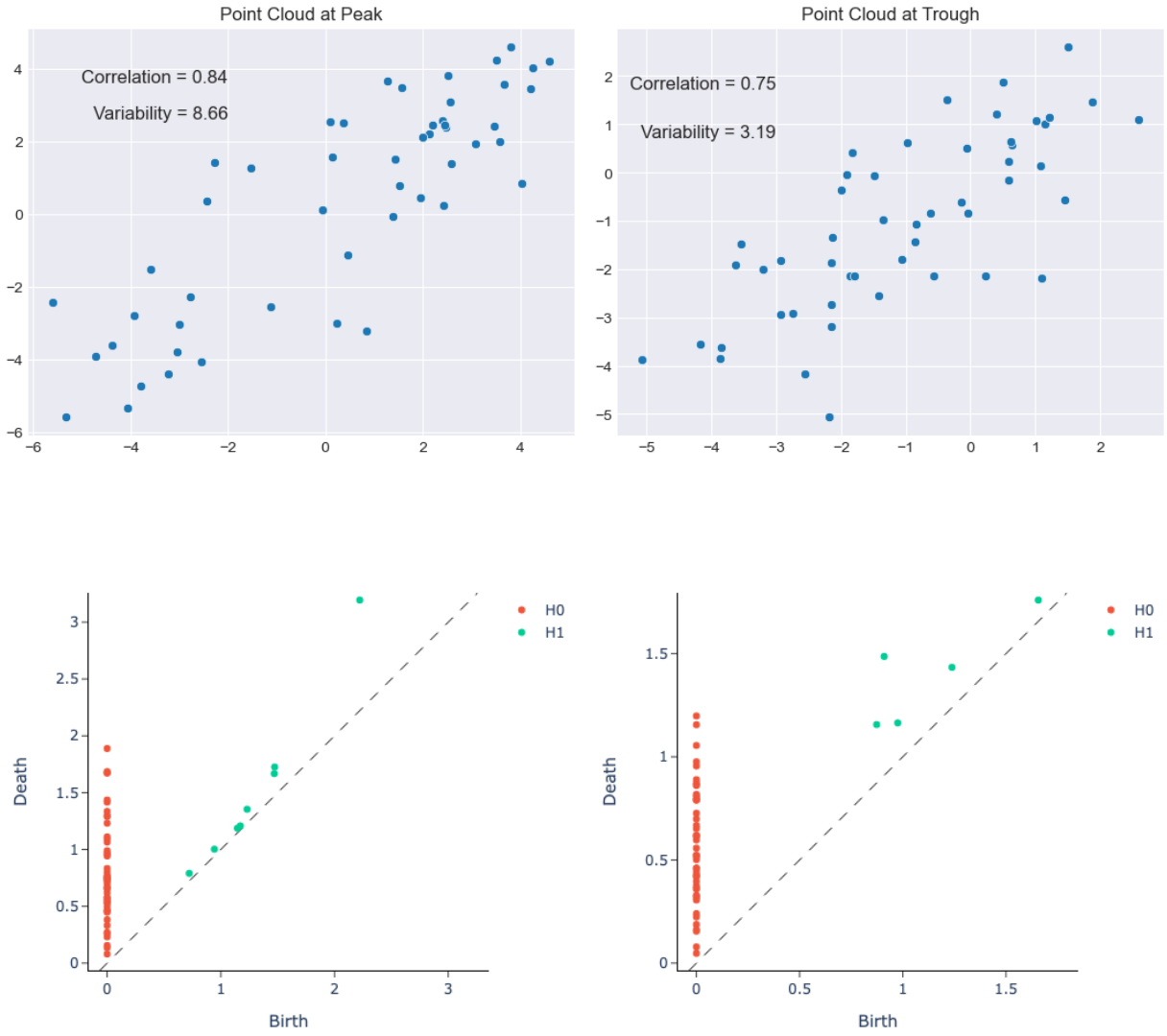

Figure 22: Left-right: peak, trough. Top-down: point cloud, persistence diagram (of figure 21)

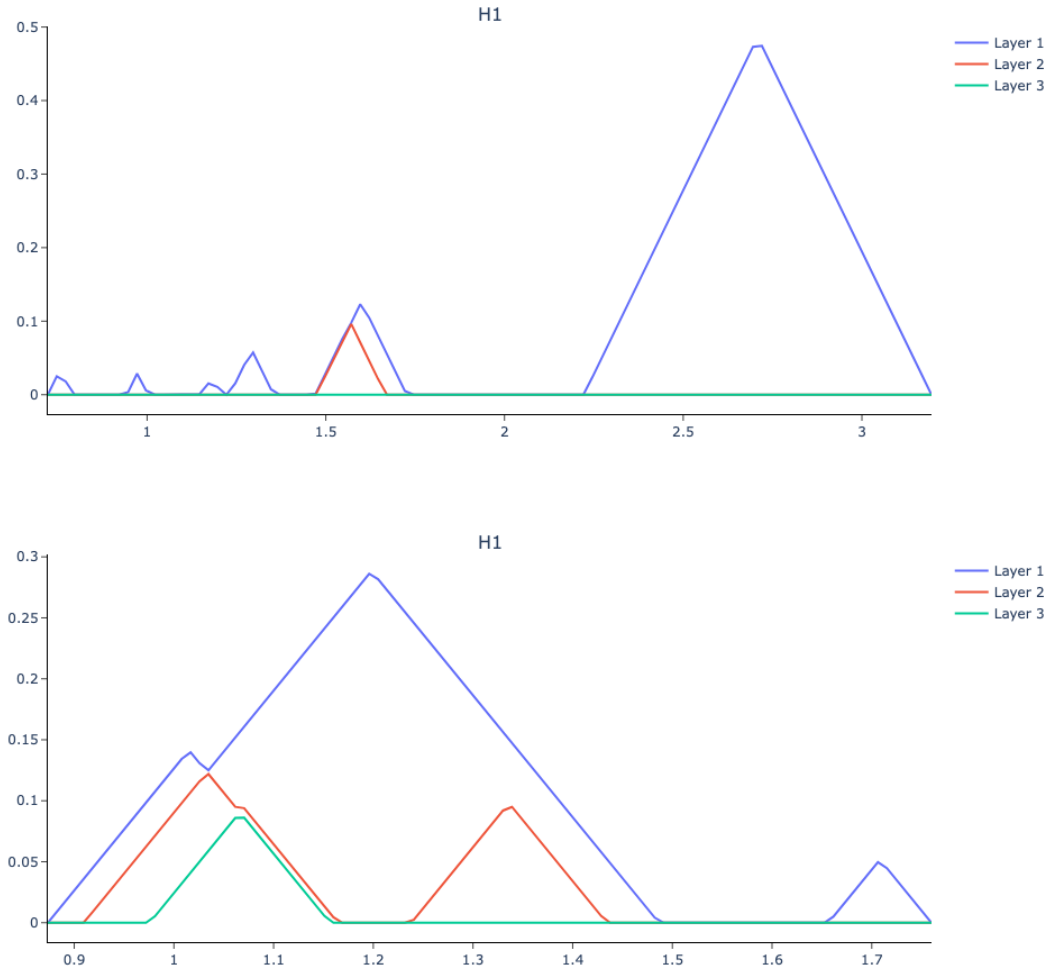

Figure 23: Top-down: persistence landscapes of peak, trough (of figure 21)

5 Conclusion

In this thesis, we studied the behavior of L_1 norms of persistence landscapes derived from periodic 1D time series using time-delayed sliding window embeddings. We start with a 1D time series, apply a sliding window time-delayed embedding, compute the persistence diagram, compute persistence landscapes and finally compute L_1 norms, variability and correlations per point cloud. Although we work with synthetic time series, our motivation for this work was studying the effects of macroeconomic periodic time series such as: unemployment time series, interest rates, etc. Future work on this topic can be extended to those applications.

We summarize by restating the main result from this paper, that L_1 norms of periodic time series using sliding window time-delayed embeddings are highly sensitive to noise. Specifically, random configurations of point clouds can lead to the emergence of homologies with relatively high persistence. This occurrence can lead to fluctuations of L_1 norms of persistence landscapes per point cloud in the SWE. We see this occur multiple times throughout our experiments. In both sets of our experiments (involving sin waves and involving AR synthetic time series), we analyze anomalies where L_1 norms fluctuate from the expected proportionality with variability of the point cloud. Future work could be extended to study this behavior not only on synthetic time series, but with analysis on real datasets.

References

[1] SILSO, World Data Center - Sunspot Number and Long-term Solar Observations, Royal Observatory of Belgium, on-line Sunspot Number catalogue, 2021.

[2] Lloyd L. Aromi, Yuri A. Katz, and Josep Vives. Topological features of multivariate distributions: Dependency on the covariance matrix. *Communications in Nonlinear Science and Numerical Simulation*, 103:105996, 2021.

[3] Peter Bubenik. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, 16(3):77–102, 2015.

[4] Gunnar Carlsson, Tigran Ishkhanov, Vin Silva, and Afra Zomorodian. On the local behavior of spaces of natural images. *International Journal of Computer Vision*, 76:1, 2008.

[5] Joseph Minhow Chan, Gunnar Carlsson, and Raul Rabadan. Topology of viral evolution. *Proceedings of the National Academy of Sciences*, 110(46):18566, 2013.

[6] Herbert Edelsbrunner and John Harer. *Computational Topology - an Introduction.*, pages 59 – 118, 169–198. American Mathematical Society, 2010.

[7] Marian Gidea. Topological data analysis of time series. Lecture notes. Yeshiva University, 2021.

[8] Marian Gidea, Daniel Goldsmith, Yuri Katz, Pablo Roldan, and Yonah Shmalo. Topological recognition of critical transitions in time series of cryptocurrencies. *Physica A: Statistical Mechanics and its Applications*, 548:123843, 2020.

[9] Marian Gidea and Yuri Katz. Topological data analysis of financial time series: Landscapes of crashes. *Physica A: Statistical Mechanics and its Applications*, 491:820–834, 2018.

[10] Monica Nicolau, Arnold J. Levine, and Gunnar Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences*, 108(17):7265, 2011.

[11] Jose Perea and John Harer. Sliding windows and persistence: An application of topological methods to signal analysis. *Foundations of Computational Mathematics*, 15:799, 2015.

[12] Mariam Pirashvili, Lee Steinberg, Francisco Belchí, Mahesan Niranjan, Jeremy Frey, and Jacek Brodzki. Improved understanding of aqueous solubility modeling through topological data analysis. *Journal of Cheminformatics*, 10:1, 2018.

[13] Suhasini Subba Rao. *A course in time series analysis*, pages 98–118. 2022.

[14] Guillaume Tauzin, Umberto Lupo, Lewis Tunstall, Julian Burella Pérez, Matteo Caorsi, Anibal Medina-Mardones, Alberto Dassatti, and Kathryn Hess. giotto-tda: A topological data analysis toolkit for machine learning and data exploration, 2020.