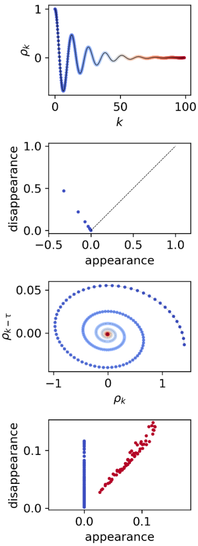
Topological Classification of Time Oscillations



Introduction

Motivation: Detecting periodicity in timeseries is applicable to many scientific fields, such as identifying brain oscillations in neural signals (Donoghue et al. 2022) or detecting gravitational waves (Bresten et al. 2019).

Traditional Methods: Periodic frequencies are typically parameterized in the frequency domain which decomposes signals into sine waves, but this may not capture the complex non-sinusoidal features of timeseries (Donoghue et al. 2022).

Approach: This project introduces methods for classifying timeseries as periodic (oscillatory) or aperiodic (non-oscillatory) using topological data analysis (TDA).

TDA Pipeline

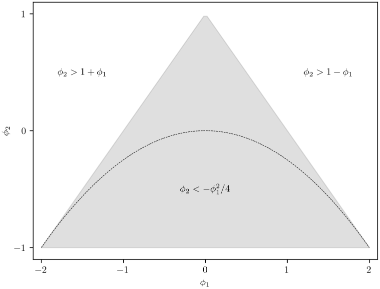
1. **Generate periodic and aperiodic timeseries** and their corresponding autocorrelation functions.
2. **Create persistence diagrams** from both the timeseries and autocorrelation functions.
3. **Convert these persistence diagrams into vectors** using measures like amplitude, persistence entropy, and number of points.
4. **Train a Logistic Regression classifier** using the extracted topological features from both time and autocorrelation domains.
5. **Evaluate and compare classification accuracies** of these methods in distinguishing periodic from non-periodic signals.

Notation, Autoregressive Timeseries

$x_t = \{x_1, x_2, ..., x_T\}$ is a discrete, length T, timeseries from a stationary ar-2 process:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + e_t, \quad e_t \overset{\text{iid}}{\sim} \mathcal{N}(0,1). \tag{1}$$

The stationarity triangle in the (ϕ_1, ϕ_2) plane defines where x_t is stationary ($\phi_2 < 1 + \phi_1$, $\phi_2 < 1 - \phi_1$, $\phi_2 > -1$), and the theoretical boundary between **periodic** and **aperiodic** processes exists at $\phi_2 = -\phi_1^2/4$.



Notation, Time Domain

Taken's embedding maps $x_t : \mathbb{Z} \to \mathbb{R}$ to $\mathbb{R}^d$ in Euclidean space:

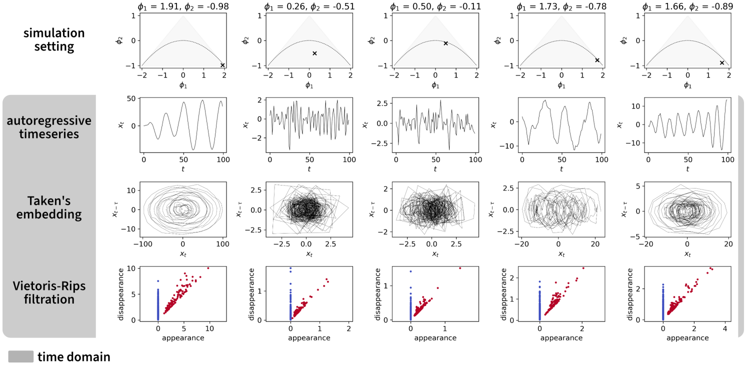
$$\text{emb}(x_t) \mapsto \{x_t, x_{t-\tau}, x_{t-2\tau}, ..., x_{t-(d-1)\tau}\}, \tag{2}$$

where τ is the time delay and d is the embedding dimension.

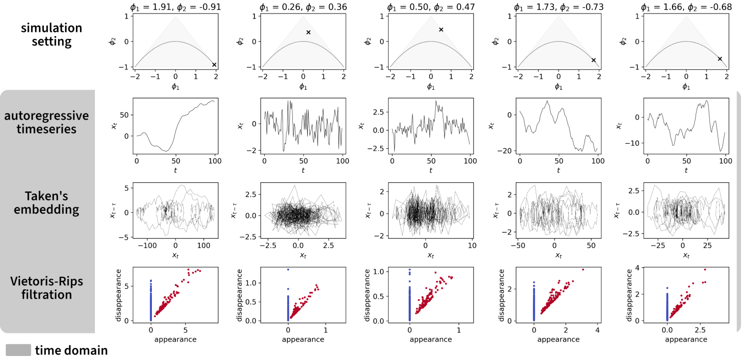
Vietoris-Rips filtration contains a subset of $\text{emb}(x_t)$ as a simplex if all pairwise distances are $\leq s$. Since $\text{emb}(x_t)$ is collection of points in $\mathbb{R}^d$ with a distance function $d(\cdot)$:

$$\text{VR}_s(\text{emb}(x_t)) = \{[v_0, ..., v_n] \mid \forall i, j; d(v_i, v_j) \leq s\} \tag{3}$$

Periodic, Time Domain



Aperiodic, Time Domain



Notation, Autocorrelation Domain

Autocorrelation function for stationary x_t depends only on time lag k:

$$\rho_k = \mathrm{corr}(x_t, x_{t-k}), \tag{4}$$

Sublevel Set filtration of the autocorrelation function, for a given threshold a is defined as:

$$\mathrm{SL}a(\rho) = k \mid \rho_k \leq a. \tag{5}$$

As a increases, the sublevel sets form a filtration:

$$\mathrm{SL}a_1(\rho) \subseteq \mathrm{SL}a_2(\rho) \subseteq \ldots \subseteq \mathrm{SL}a_n(\rho) \quad \text{for} \quad a_1 \leq a_2 \leq \ldots \leq a_n \tag{6}$$

Taken's embedding maps $\rho_k : \mathbb{Z} \to \mathbb{R}$ to $\mathbb{R}^d$ in Euclidean space:

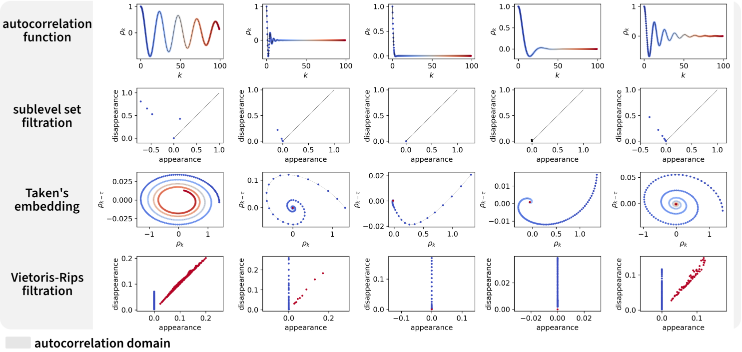
$$\mathrm{emb}(\rho_k) \mapsto \{\rho_k, \rho_{k-\tau}, \rho_{k-2\tau}, ..., \rho_{k-(d-1)\tau}\}. \tag{7}$$

where τ is the time delay and d is the embedding dimension.

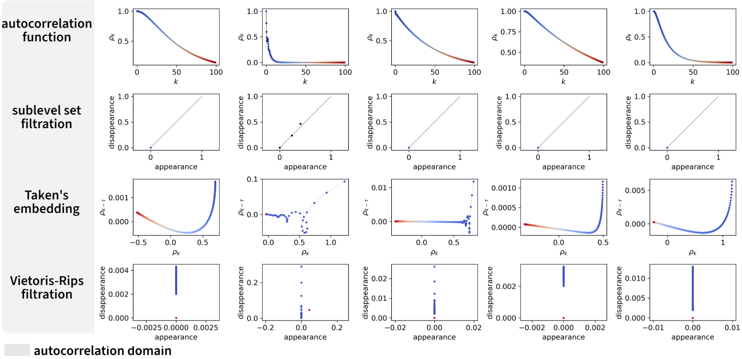
Vietoris-Rips filtration contains a subset of $\mathrm{emb}(\rho_k)$ as a simplex if all pairwise distances are $\leq s$. Since $\mathrm{emb}(\rho_k)$ is collection of points in $\mathbb{R}^d$ with a distance function $d(\cdot)$:

$$\mathrm{VR}_s(\mathrm{emb}(\rho_k)) = \{[v_0, ..., v_n] \mid \forall i, j; d(v_i, v_j) \leq s\} \tag{8}$$

Periodic, Autocorrelation Domain



Aperiodic, Autocorrelation Domain



Vectorization of Persistence Diagrams

As implemented in *giotto-tda* (Tauzin et al. 2021), these three methods were used to map **persistence diagrams** $\rightarrow$ **feature vectors** for use in downstream classifiers:

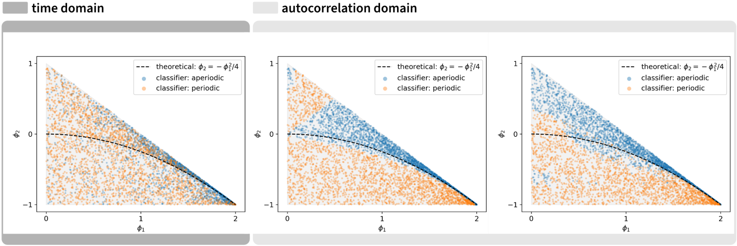
1. **Amplitude** was calculated as the L^2 distance between persistence landscapes, for each homology dimension.
2. **Persistence entropy** was calculated as the Shannon entropies of points on a persistence diagram, for each homology dimension.
3. **Number of points** was calculated the number of off-diagonal points in persistence diagrams, for each homology dimension.

Results

Simulation setting: $T = 250$ timepoints, $K = 50$ autocorrelation lags, $N = 5000$ repeats.

Classifier: $L2$ Logistic Regression, $N = 4000$ training set, $N = 1000$ validation set.

Performance Metrics: Accuracy (fraction of correct classifications), Area Under ROC Curve (true positives vs false positives).



Accuracy on train: 0.618	Accuracy on train: 0.875	Accuracy on train: 0.903
ROC AUC on train: 0.664	ROC AUC on train: 0.889	ROC AUC on train: 0.964
Accuracy on valid: 0.611	Accuracy on valid: 0.885	Accuracy on valid: 0.909
ROC AUC on valid: 0.664	ROC AUC on valid: 0.907	ROC AUC on valid: 0.971

Conclusions

Summary of Results for the effectiveness of topological methods for classifying timeseries data as periodic or aperiodic.

1. **Time domain, Taken's embedding, Vietoris-Rips filtration** performed poorly, likely due to noisy embeddings.
2. **Autocorrelation domain, Sublevel Set filtration** showed large a accuracy improvement, indicating the utility of autocorrelations in making periodicity more detectable.
3. **Autocorrelation domain, Taken's embedding, Vietoris-Rips filtration** performed best, suggesting that mapping the timeseries through multiple stages (from time to autocorrelation, then embedding) captures more robust topological features that are effective for classification.

These results emphasize the importance of preprocessing and feature extraction in topological data analysis for timeseries classification.