Sinkhorn Divergence of Topological Signatures for Time Series Classification[1]

Colin Stephen – Coventry University

ICMLA 2018

[1]Slides available at https://github.com/colinstephen/icmla2018

1. **Motivation**
2. Persistence Images
3. Regularized Transport
4. Classification Pipeline
5. Results

Hénon Attractor

$$x_{n+1} = 1 - ax_n^2 + y_n \text{ and } y_{n+1} = bx_n$$

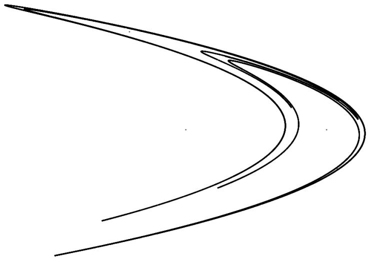


Figure 1: Hénon map for $a = 1.4, b = 0.3$

Chaotic Bifurcations

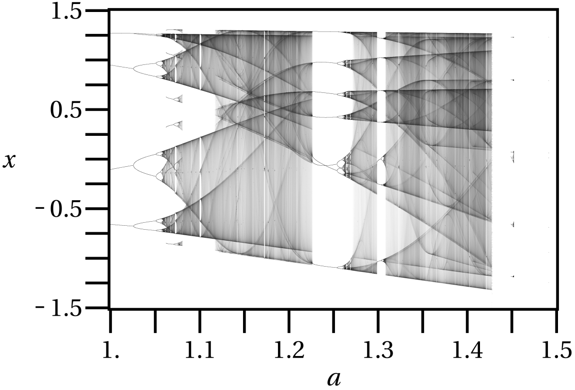


Figure 2: Possible range of x is highly sensitive to a

Distinguish Trajectory Classes

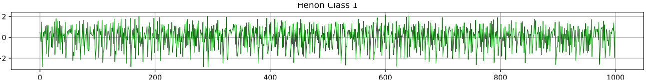


Figure 3: Sequence of x values for $a = 1.4, b = 0.3$

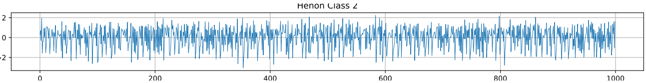
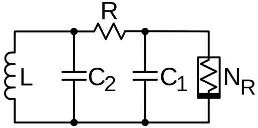
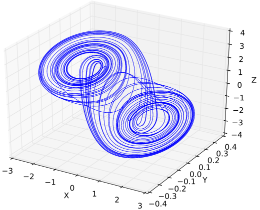


Figure 4: Sequence of x values for $a = 1.395, b = 0.3$

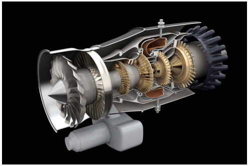
Possible Applications

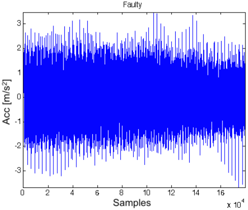
Quality control in PCB circuit fabrication:

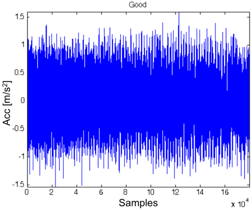




Engine component failure prediction:







Practical Challenge

Given:

- Two classes of labelled z-normalized time series measured from some chaotic system
- An unlabelled time series from one class

Find:

- A good choice of label for the unlabelled instance

Subject to:

- The underlying dynamic model is unknown
- Signal to noise ratio may be low
- Robust identification needs long time series

Many Standard Approaches

- Transformation based distances
  - dynamic time warp
  - edit distance
- Dictionary approaches
  - bag of patterns
  - SAX
- Shapelets
- Ensembles
  - COLT
  - Elastic Ensemble
- Signal decomposition approaches
  - spectral analysis
  - cepstral analysis

Topological Approaches

Q: Can topological properties distinguish time series classes?

A: Yes. Topological Data Analysis (TDA) using Takens embeddings.

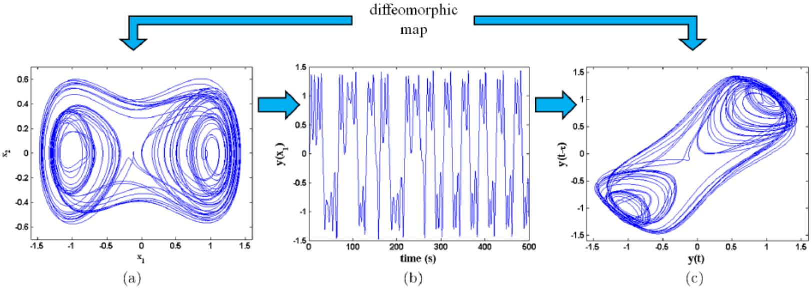


Figure 5: Takens Theorem: delay embedding of time series is diffeomorphic copy of attractor (for the right embedding!)

Challenges for TDA on Time Series

1. Takens embedding requires dimension and delay estimation.
2. The embedding moves data from 1D to nD
   - has a large complexity cost for TDA methods.
   - requires subsampling and other statistical approaches
3. Computing metric distances on topological feature spaces has high time complexity anyway (double jeopardy).

Aim of the paper: construct a TDA pipeline that...

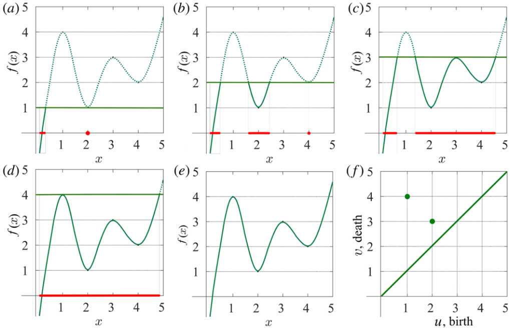
- does not require embeddings
- uses a metric on topological features that is fast to compute
- classifies with *competitive accuracy*

1. Motivation
2. **Persistence Images**
3. Regularized Transport
4. Classification Pipeline
5. Results

Persistence Diagram of a Time Series

Look at inclusions of sublevel sets $f^{-1}(-\infty, a]$ for $a \in \mathbb{R}$

Apply a *precedence condition* for merging sets (lives vs dies)



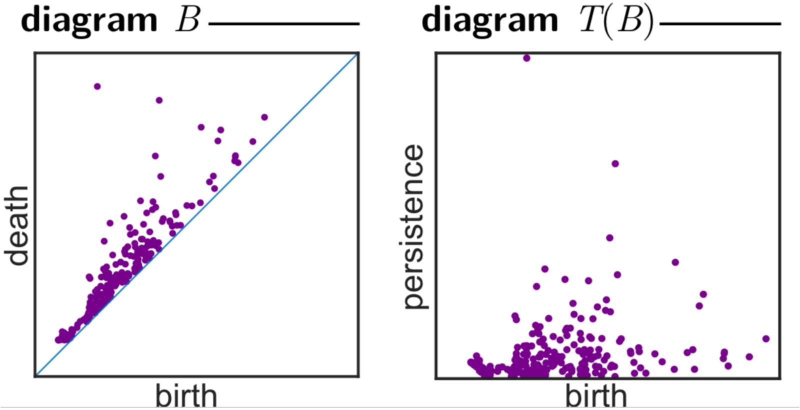
This gives a simple topological descriptor called the *persistence diagram* (bottom right)

Key result: (2007) a metric on the space of PDs (Wasserstein distance) is L_p-stable on large space of functions

Realistic Persistence

The number of persistence points is generally large.

Also we always have $b \leq d$ so can translate $T : (b, d) \mapsto (b, d - b)$



The vertical axis $d - b$ is the *persistence* of the feature

Persistence Surfaces

Practical concern:

- Large numbers of points in $T(B)$
- Suggests using KDEs instead

Constraint:

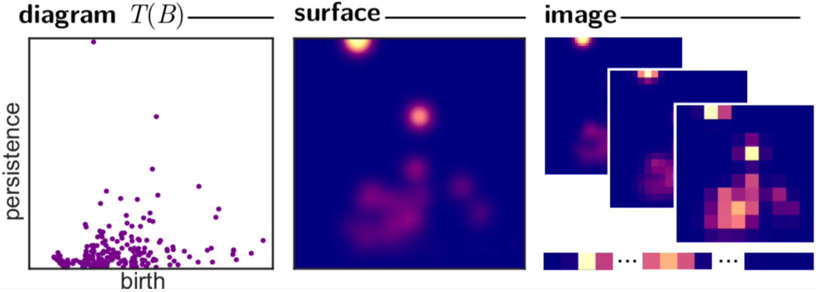
- Points near the diagonal are seen as 'topological noise'
- Suggests applying a weight function $f(b, p) \in \mathbb{R}$ that decays to zero on axis $b = 0$

$$\rho_B(z) := \frac{1}{2\pi\sigma^2} \sum_{x \in T(B)} f(x) e^{-\frac{\|z-x\|^2}{2\sigma^2}}$$

This is the persistence surface of B

Persistence Images

- Discrete approximations of surfaces can be compared more quickly
- So divide an area of $\mathbb{R}^2_+$ in to a regular grid
- Integrate ρ_B over each grid cell
  - Fast in practice using convolutions



NB: scale of Gaussian and scale of grid are independent

1. Motivation
2. Persistence Images
3. **Regularized Optimal Transport**
4. Classification Pipeline
5. Results

The Optimal Transport Problem

Transport 'probability mass' between two distributions θ, r

- corresponds to specifying a joint probability Γ

Subject to: minimal total cost of joint probabilities assigned

- So find $\min_{\Gamma} \langle \Gamma, D \rangle = \sum_{i,j} \Gamma_{i,j} D_{i,j}$

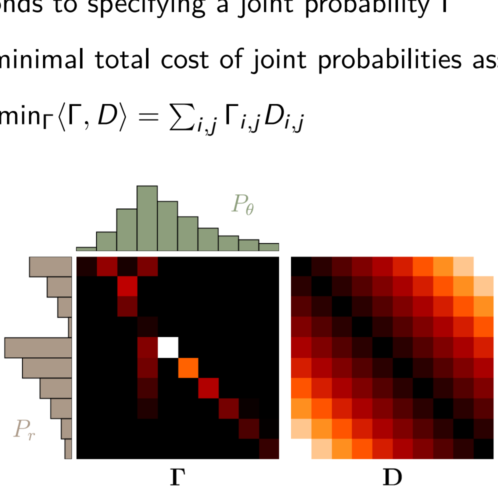


Figure 6: Marginal and joint probabilities (left) and cost matrix (right)

Regularized Optimal Transport

- Standard OT problem is $O(n^3 \log n)$ for 1D histograms
- In 2D ($n \times n$ histograms) it is $O(n^6 \log n)$
- Not feasible computationally

Key result: (method 2013, complexity 2017)

Adding a regularization term to the optimization reduces 1D problem to $O(n \log n)$.

- Define regularized optimal transport distance:

$$\mathrm{ROT}_D^{\lambda}(\theta, r) = \langle \Gamma_\lambda^*, D \rangle$$

- Subject to:

$$\Gamma_\lambda^* = \mathrm{argmin}_\Gamma(\langle \Gamma, D \rangle - \lambda H(\Gamma))$$

- For some error function H over joint probabilities

Entropic Regularization

- Choosing error penalty

$$H(\Gamma) = -\sum_{i,j} \Gamma_{i,j} \log \Gamma_{i,j}$$

  finds an unbiased – maximum ignorance – choice of Γ^*_λ.
- The entropy regularized OT distance $\mathrm{ROT}^\lambda_D(\theta, r)$ is called the Sinkhorn Divergence between the distributions.

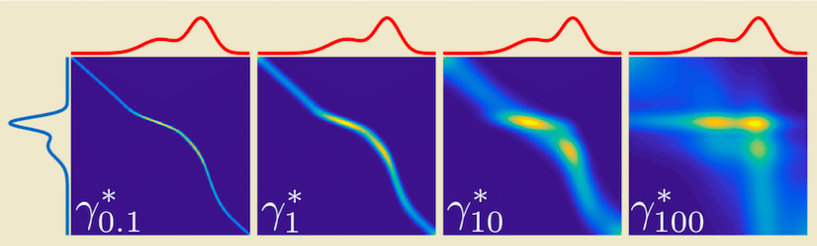


Figure 7: Regularised OT for $\lambda \in \{0.1, 1, 10, 100\}$ (γ in figure is Γ above).

1. Motivation
2. Persistence Images
3. Regularized Transport
4. **Classification Pipeline**[2]
5. Results

[2]Python code for classifiers at: https://github.com/colinstephen/icmla2018

Training: Learn Persistence Images of Classes

Given collection of labelled time series, for each one:

- Find its persistence diagram (PD)

For each class:

- Overlay its PDs
- Compute the class persistence surface (parameters are σ, f)
- Discretize to a persistence image (parameter is $d \times d$)

Here: 100x100 persistence images for Henon classes 1 and 2, and their difference.

Prediction: compute Sinkhorn Divergence

Given an unlabelled time series:

- Find its persistence diagram
- Compute its persistence image I
  - use same values for σ, $d \times d$, and f

Fix an L_p cost matrix for some p and compute:

- Sinkhorn divergence between I and class images
- Closest one wins

In practice

- all parameters including p set in training via cross validation
- training is $O(d^2 \log d)$
- prediction is $O(n \log n)$

1. Motivation
2. Persistence Images
3. Regularized Transport
4. Classification Pipeline
5. **Results**[3]

[3] Python code for trajectory data and benchmark classifiers also at: https://github.com/colinstephen/icmla2018

Experiments

Data in paper:

- Synthetic time series from Lorenz, Hénon, and Logistic systems
- Initial conditions uniformly distributed over intervals
- Model parameters uniformly distributed over intervals too
- Two classes generated per experiment
- Approx $1,000,000$ time series classified in total

Benchmarks

Pipeline outperforms well known fequency decomposition approach:

- Euclidean distance between cepstral coefficients
- Variation on the discrete Fourier transform

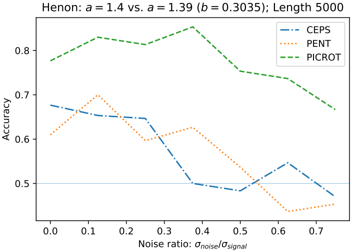
Outperforms the only TDA approach that avoids embeddings:

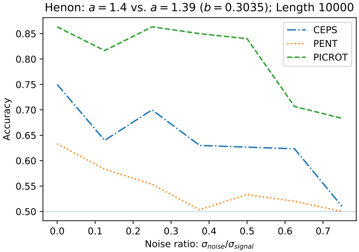
- A decision rule for class membership based on ROC curve for 'persistent entropy' of the individual time series
- Paper actually implements an improved version of this using nearest neighbours

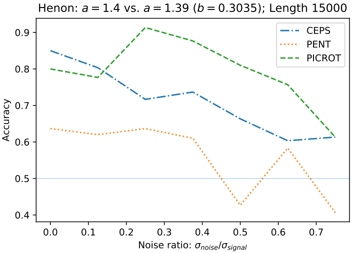
Also tested against DTW and Random Forests: these were not competitive.

Accuracy Profiles: Hénon time series

Accuracy vs noise for three lengths: 5000, 10000, 15000







Summary

If you wish to classify chaotic trajectories you can:

1. Represent topology as persistence images to give:
    - A class-based KDE of the topology
2. Quantify proximity using Sinkhorn divergence:
    - A fast metric on spaces of distributions

Result is fast estimation of class membership that is:

- Robust to noise – topological stability result
- Effective for long series – KDE over a grid; Sinkhorn algorithm
- Accurate relative to common approaches

Thank You.

References 1

[1] Adams, H., et. al. *Persistence images: A stable vector representation of persistent homology*. Journal of Machine Learning Research **18**, 1 (2017), 218–252.

[2] Altschuler, J., et. al. *Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration.* In Advances in neural information processing systems (2017), pp. 1964–1974.

[3] Cohen-Steiner, D., et. al. *Lipschitz Functions Have L_p-Stable Persistence*. Foundations of Computational Mathematics **10** (2010), 127–139.

References 2

[4] Cuturi, M. *Sinkhorn distances: Lightspeed computation of optimal transport*. In Advances in neural information processing systems (2013), pp. 2292–2300.

[5] Randall, R. B. *A history of cepstrum analysis and its application to mechanical problems.* Mechanical Systems and Signal Processing 97 (2017), 3–19.

[6] Rucco, M., et. al. *A new topological entropy-based approach for measuring similarities among piecewise linear functions*. Signal Processing **134** (2017), pp.130–138.