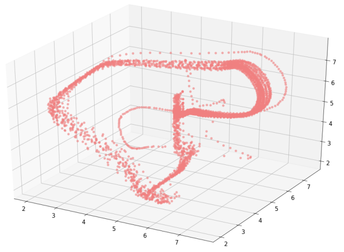
Geometric and Topological Inference for Data Analysis

Ximena Fernández

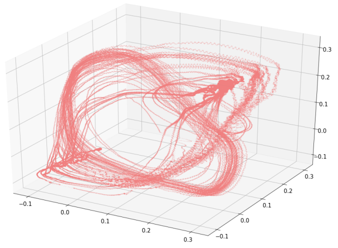
Applied Algebra and Geometry in the UK
11th meeting
15th December 2020

Liverpool-Oxford-Swansea Centre for Topological Data Analysis

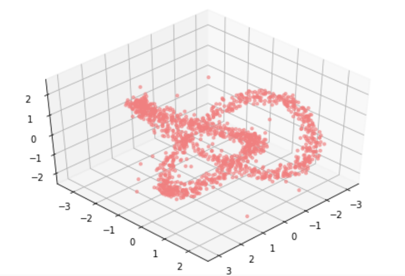
Motivation



Embedding of a ECG signal.



Embed. air sac pressure record of a canary during singing.



Trefoil knot with noise and outliers.

Geometric Inference

Geometric Inference

- Geometric inference deals with the problem of inferring information about a geometric object from a finite **sample**.

Geometric Inference

- Geometric inference deals with the problem of inferring information about a geometric object from a finite **sample**.

- Two unknown parameters are implicit in the sample:
  - the probability distribution,
  - the underlying geometry.

Geometric Inference

- Geometric inference deals with the problem of inferring information about a geometric object from a finite **sample**.

- Two unknown parameters are implicit in the sample:
  - the probability distribution,
  - the underlying geometry.

- The aim is to find estimators of:
  - the density of the distribution,
  - the dimension (of the manifold),
  - the distance (of the metric space),
  - the geometry itself,
  - the homology.

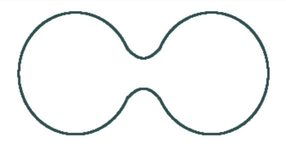
Distance learning

$(\mathcal{M}, g)$ a Riemannian manifold embedded in $\mathbb{R}^D$ with inherited geodesic distance $d_{\mathcal{M}}$.

Distance learning

$(\mathcal{M}, g)$ a Riemannian manifold embedded in $\mathbb{R}^D$ with inherited geodesic distance $d_{\mathcal{M}}$.

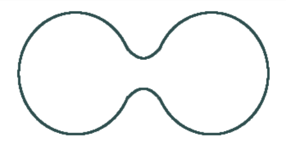
- **Good news:** *Locally*, geodesic distance can be approximated by Euclidean distance.



Distance learning

$(\mathcal{M}, g)$ a Riemannian manifold embedded in $\mathbb{R}^D$ with inherited geodesic distance $d_{\mathcal{M}}$.

- **Good news:** *Locally*, geodesic distance can be approximated by Euclidean distance.

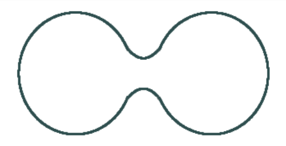


- **Bad news** *(curse of dimensionality)*: In *high dimensional* Euclidean spaces, the points essentially become uniformly distant from each other.

Distance learning

$(\mathcal{M}, g)$ a Riemannian manifold embedded in $\mathbb{R}^D$ with inherited geodesic distance $d_{\mathcal{M}}$.

- **Good news:** *Locally*, geodesic distance can be approximated by Euclidean distance.



- **Bad news** *(curse of dimensionality)*: In *high dimensional* Euclidean spaces, the points essentially become uniformly distant from each other.

- M. Bernstein, V. D. Silva, J. C. Langford, and J. B. Tenenbaum. *Graph approximations to geodesics on embedded manifolds*, 2000.

Density-based distance learning

$(\mathcal{M}, g)$ a d-dimensional Riemannian manifold embedded in $\mathbb{R}^D$ with inherited geodesic distance $d_{\mathcal{M}}$ and $f : \mathcal{M} \to \mathbb{R}_{>0}$ a density function.

Density-based distance learning

$(\mathcal{M}, g)$ a d-dimensional Riemannian manifold embedded in $\mathbb{R}^D$ with inherited geodesic distance $d_{\mathcal{M}}$ and $f : \mathcal{M} \to \mathbb{R}_{>0}$ a density function.

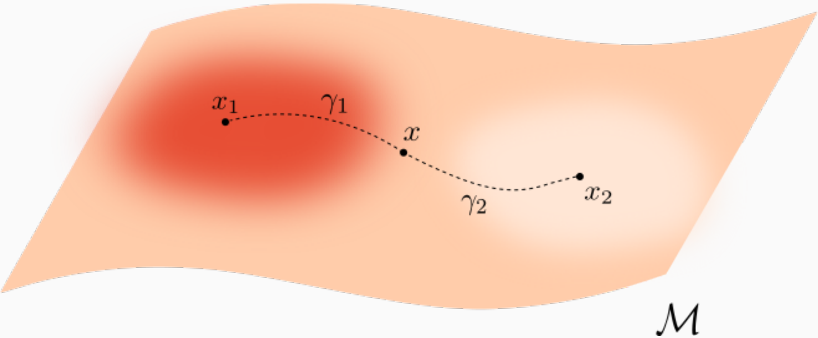
- For $p > 1$ define a new (Riemannian) metric tensor $g_p := f^{2(1-p)/d} g$.
- The induced **deformed** Riemannian distance in $\mathcal{M}$ is

$$d_{f,p}(x, y) = \inf_{\gamma} \int_I \frac{1}{f(\gamma_t)^{(p-1)/d}} ||\dot{\gamma}_t|| dt.$$

  where the infimum is taken over all piecewise smooth curves $\gamma : I \to \mathcal{M}$ with $\gamma(0) = x$, and $\gamma(1) = y$.

$d_{f,p}$ is called p-**Fermat distance** by analogy the Fermat principle in optics.

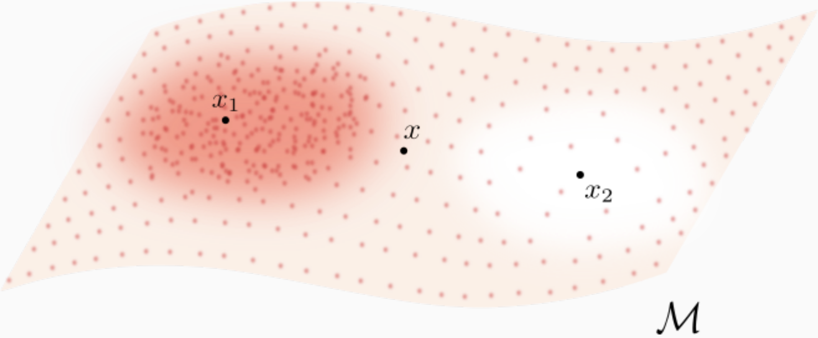
Fermat Distance



Density-based distance learning

$\mathbb{X}_n \subseteq \mathcal{M}$ a set of n sample points with common density f.

We look for a **computable estimator** of $d_{f,p}$ from the sample.

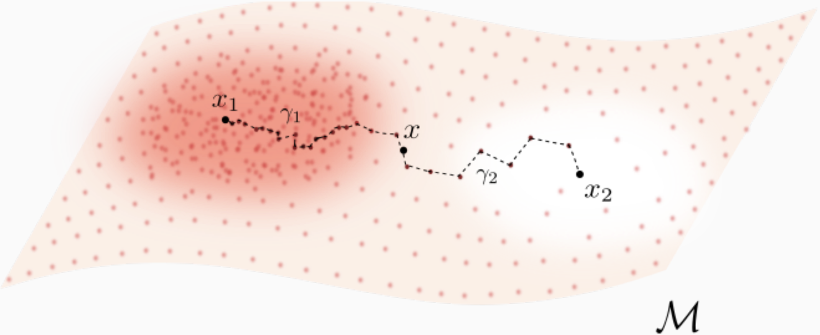


Sample Fermat distance

For $p > 1$, the **sample Fermat distance** between x, y is defined by

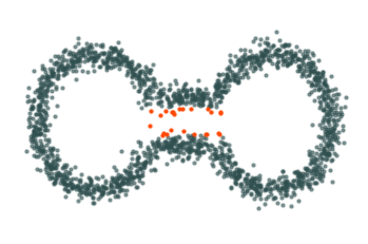
$$d_{\mathbb{X}_n,p}(x,y) = \inf_{\gamma} \sum_{i=0}^{r} |x_{i+1} - x_i|^p$$

where the infimum is taken over all paths $\gamma = (x_0, \ldots, x_{r+1})$ of finite length with $x_0 = x$, $x_{r+1} = y$ and $\{x_1, x_2, \ldots, x_r\} \subseteq \mathbb{X}_n$.



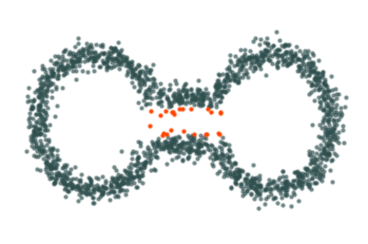
Example

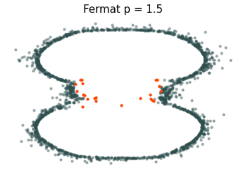
Eyeglasses curve. A sample of 2000 points with Gaussian noise.



Example

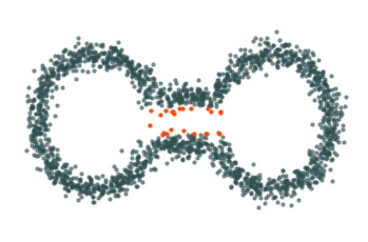
Eyeglasses curve. A sample of 2000 points with Gaussian noise.

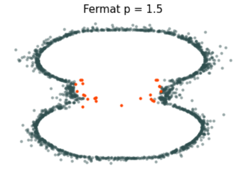


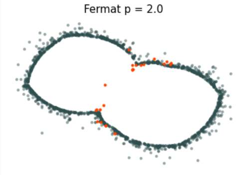


Example

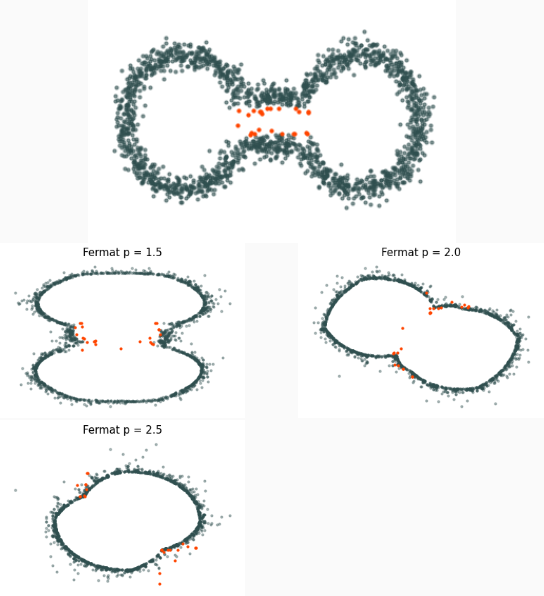
Eyeglasses curve. A sample of 2000 points with Gaussian noise.





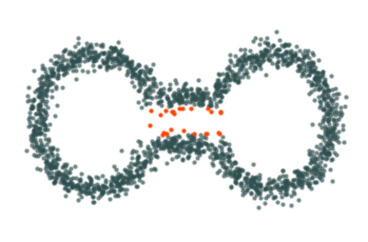


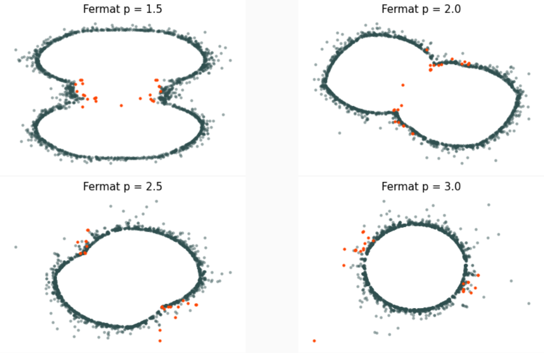
Eyeglasses curve. A sample of 2000 points with Gaussian noise.



Example

Eyeglasses curve. A sample of 2000 points with Gaussian noise.





Previous work

Sample Fermat distance was independently introduced in:

- D. Mckenzie and S. Damelin. *Power weighted shortest paths for clustering euclidean data.* Foundations of Data Science, 1(3):307, 2019.
- P. Groisman, M. Jonckheere, and F. Sapienza. *Nonhomogeneous euclidean first-passage percolation and distance learning.* arXiv:1810.09398, 2018.

Previous work

Theorem (Groisman, Jonckheere, Sapienza (2018))

Let $\mathcal{M}$ be an **isometric*** C^1 d-dimensional manifold embedded in $\mathbb{R}^D$.

Then, there exists $\mu = \mu(p, d) > 0$ such that for any $x, y \in \mathcal{M}$,

$$\lim_{n \to +\infty} \frac{n^{(p-1)/d}}{\mu} d_{\mathbb{X}_n, p}(x, y) = d_{f,p}(x, y) \text{ almost surely.}$$

*$\mathcal{M}$ is an isometric d-dimensional C^1 manifold embedded in $\mathbb{R}^D$ if there exists $S \subseteq \mathbb{R}^d$ an open connected set and $\varphi : \bar{S} \to \mathbb{R}^D$ an isometric transformation such that $\varphi(\bar{S}) = \mathcal{M}$.

Previous work

Theorem (Hwang, Damelin, Hero (2016))

Let $\mathcal{M}$ be a compact smooth d-dimensional manifold without boundary. Given $\varepsilon > 0$ and $b > 0$, there exists $\theta = \theta(\varepsilon) > 0$ such that, for all sufficiently large n,

$$\mathbb{P}\left(\sup_{x,y : d_{\mathcal{M}}(x,y) \geqslant b} \left| \frac{\frac{n^{(p-1)/d}}{\mu} L_{\mathbb{X}_n,p}(x,y)^{\dagger}}{d_{f,p}(x,y)} - 1 \right| > \varepsilon \right) \leqslant \exp(-\theta n^{1/(d+2p)})$$

In particular, for every $x, y \in \mathcal{M}$,

$$\lim_{n \to +\infty} \tfrac{n^{(p-1)/d}}{\mu} L_{\mathbb{X}_n,p}(x,y) = \mu d_{f,p}(x,y) \text{ almost surely.}$$

$^{\dagger} L_{\mathbb{X}_n}(x,y) = \inf_{\gamma} \sum_{i=0}^{r} d_{\mathcal{M}}(x_{i+1}, x_i)^p$, where the infimum is taken over all paths $\gamma = (x_0, \ldots, x_{r+1})$ with $x_0 = x$, $x_{r+1} = y$ and $\{x_1, \ldots, x_r\} \subseteq \mathbb{X}_n$.

Fermat distance learning

Theorem 1 (Borghini, F., Groisman, Mindlin, 2020)

Let $\mathcal{M}$ be a compact smooth d-dimensional manifold without boundary. Then, for every $p > 1$ and $\lambda \in \left(\frac{p-1}{pd}, \frac{1}{d}\right)$, given $\varepsilon > 0$ there exist $\theta > 0$ such that, for n large enough,

$$\mathbb{P}\left(\sup_{x,y\in\mathcal{M}} \left|\frac{n^{(p-1)/d}}{\mu} d_{\mathbb{X}_n,p}(x,y) - d_{f,p}(x,y)\right| > \varepsilon\right) \leqslant \exp\left(-\theta n^{\frac{1-\lambda d}{d+2p}}\right).$$

Manifold approximation

- **Population metric space**: $(\mathcal{M}, d_{f,p})$.
- **Sample metric space:** $\left(\mathbb{X}_n, \frac{n^{(p-1)/d}}{\mu} d_{\mathbb{X}_n,p}\right)$.

Manifold approximation

- **Population metric space**: $(\mathcal{M}, d_{f,p})$.
- **Sample metric space:** $\left(\mathbb{X}_n, \frac{n^{(p-1)/d}}{\mu} d_{\mathbb{X}_n,p}\right)$.

The Gromov–Hausdorff distance between $(\mathbb{X}, \rho_{\mathbb{X}}), (\mathbb{Y}, \rho_{\mathbb{Y}})$ is

$$d_{GH}\big((\mathbb{X}, \rho_{\mathbb{X}}), (\mathbb{Y}, \rho_{\mathbb{Y}})\big) := \inf\{d_H(h_1(\mathbb{X}), h_2(\mathbb{Y}))\},$$

where the infimum is over all the isometric embeddings $h_1 \colon \mathbb{X} \to \mathbb{W}$, $h_2 \colon \mathbb{Y} \to \mathbb{W}$ in a common metric space $\mathbb{W}$ and d_H stands for the Hausdorff distance.

Manifold approximation

- **Population metric space**: $(\mathcal{M}, d_{f,p})$.
- **Sample metric space:** $\left(\mathbb{X}_n, \frac{n^{(p-1)/d}}{\mu} d_{\mathbb{X}_n,p}\right)$.

Theorem 2 (Borghini, F., Groisman, Mindlin, 2020)

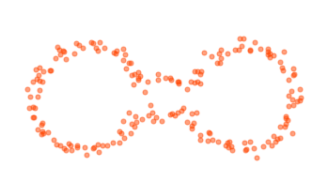
Let $\mathcal{M}$ be a compact smooth d-dimensional manifold without boundary. Then, for every $p > 1$ and $\lambda \in \left(\frac{p-1}{pd}, \frac{1}{d}\right)$, given $\varepsilon > 0$ there exist $\theta > 0$ such that, for n large enough,

$$\mathbb{P}\left(d_{GH}\left(\left(\mathcal{M}, d_{f,p}\right), \left(\mathbb{X}_n, \frac{n^{(p-1)/d}}{\mu} d_{\mathbb{X}_n,p}\right)\right) > \varepsilon\right) \leqslant \exp\left(-\theta n^{(1-\lambda d)/(d+2p)}\right)$$

Topological Inference

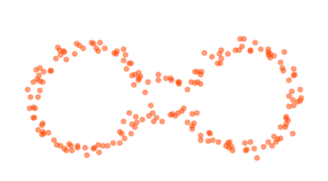
Persistent Homology

Point cloud: $(\mathbb{X}_n, \rho_n)$

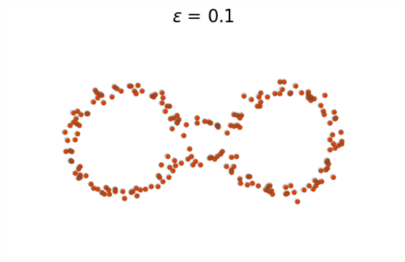


Persistent Homology

Point cloud: $(\mathbb{X}_n, \rho_n)$

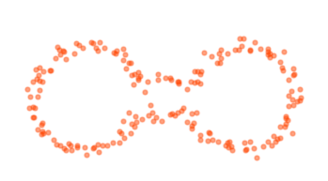


Estimator: $\bigcup_i B(x_i, \varepsilon)$

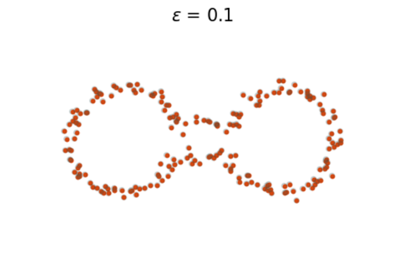


Persistent Homology

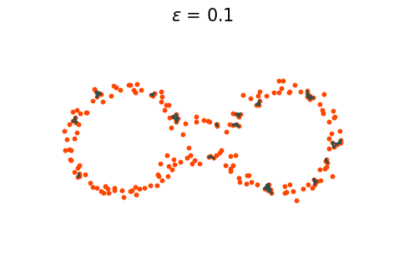
Point cloud: $(\mathbb{X}_n, \rho_n)$



Estimator: $\bigcup_i B(x_i, \varepsilon)$

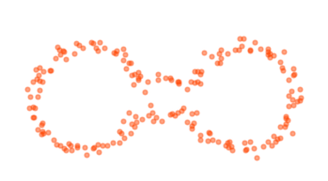


$\mathrm{Filt}_\varepsilon(\mathbb{X}_n, \rho_n)$

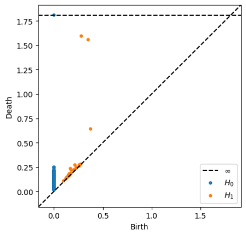


Persistent Homology

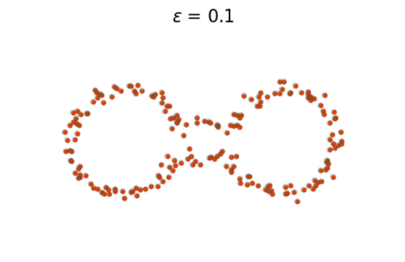
Point cloud: $(\mathbb{X}_n, \rho_n)$



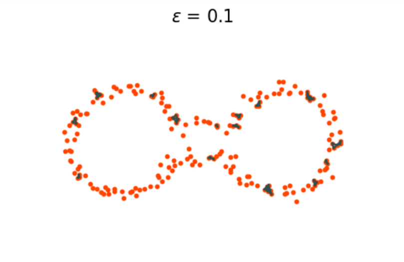
$\text{dgm}(\text{Filt}(\mathbb{X}_n, \rho_n))$



Estimator: $\bigcup_i B(x_i, \varepsilon)$



$\text{Filt}_\varepsilon(\mathbb{X}_n, \rho_n)$



Approximation of persistence diagrams

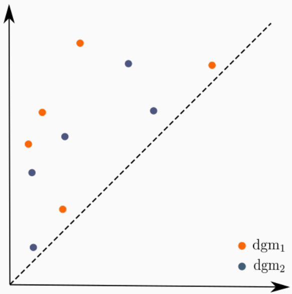
- **Population persistence diagram**: $\mathrm{dgm}(\mathrm{Filt}(\mathcal{M}, \rho))$.
- **Sample persistence diagram**: $\mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_n, \rho_n))$.

Approximation of persistence diagrams

- **Population persistence diagram**: $\mathrm{dgm}(\mathrm{Filt}(\mathcal{M}, \rho))$.
- **Sample persistence diagram**: $\mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_n, \rho_n))$.

The bottleneck distance between dgm_1 and dgm_2 is

$$d_b(\mathrm{dgm}_1, \mathrm{dgm}_2) = \inf_M \max_{(x,y)\in M} |x - y|_\infty .$$

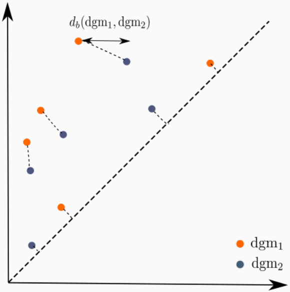


Approximation of persistence diagrams

- **Population persistence diagram**: $\mathrm{dgm}(\mathrm{Filt}(\mathcal{M}, \rho))$.
- **Sample persistence diagram**: $\mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_n, \rho_n))$.

The bottleneck distance between dgm_1 and dgm_2 is

$$d_b(\mathrm{dgm}_1, \mathrm{dgm}_2) = \inf_M \max_{(x,y)\in M} |x - y|_\infty.$$



Stability of persistence diagrams

Stability Theorem

Let X, Y be precompact metric spaces. Then,

$$d_b\big(\mathrm{dgm}(\mathrm{Filt}(X)), \mathrm{dgm}(\mathrm{Filt}(Y))\big)^{\ddagger} \leqslant 2d_{GH}(X, Y) \leqslant 2d_H(X, Y)$$

where the last inequality holds if X, Y are embedded in the same metric space.

‡Here Filt will denote either Rips or $\check{\mathrm{C}}\mathrm{ech}$ filtration.

Convergence of persistence diagrams

- **Population persistence diagram**: $\mathrm{dgm}(\mathrm{Filt}(\mathcal{M}, \rho))$.
- **Sample persistence diagram**: $\mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_n, \rho_n))$.

Theorem (Chazal, Glisse, Labruere, Michel, 2015)

Let $(\mathbb{X}, \rho)$ be a compact metric space. Let $\mathbb{X}_n$ be a sample of $\mathbb{X}$ from a measure μ with support $\mathbb{X}$ that satisfies the (a, b)-**condition**[§]. Then for every $\varepsilon > 0$

$$\mathbb{P}\left(d_b(\mathrm{dgm}(\mathrm{Filt}(\mathbb{X})), \mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_n))) > \varepsilon\right) \leqslant \min\left\{\frac{2^b}{a\varepsilon^b}\exp(-na\varepsilon^b), 1\right\}.$$

[§] For all $r > 0$ and $x \in \mathbb{X}$, $\mu(B(x, r)) \geqslant \min(1, ar^b)$.

Convergence of persistence diagrams

- **Population persistence diagram**: $\mathrm{dgm}(\mathrm{Filt}(\mathcal{M}, \rho))$.
- **Sample persistence diagram**: $\mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_n, \rho_n))$.

Theorem (Chazal, Glisse, Labruere, Michel, 2015)

Let $(\mathbb{X}, \rho)$ be a compact metric space. Let $\mathbb{X}_n$ be a sample of $\mathbb{X}$ from a measure μ with support $\mathbb{X}$ that satisfies the (a, b)-**condition**[§]. Then for every $\varepsilon > 0$

$$\mathbb{P}\left(d_b(\mathrm{dgm}(\mathrm{Filt}(\mathbb{X})), \mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_n))) > \varepsilon\right) \leqslant \min\left\{\frac{2^b}{a\varepsilon^b}\exp(-na\varepsilon^b), 1\right\}.$$

- B. T. Fasy, F. Lecci, A. Rinaldo, L. Wasserman, S. Balakrishnan, and A. Singh. *Confidence sets for persistence diagrams.* Ann. Statist., 42(6):2301–2339, 2014.

[§]For all $r > 0$ and $x \in \mathbb{X}$, $\mu(B(x, r)) \geqslant \min(1, ar^b)$.

Convergence of persistence diagrams

- **Population persistence diagram**: $\mathrm{dgm}(\mathrm{Filt}(\mathcal{M}, d_{f,p}))$.
- **Sample persistence diagram**: $\mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_n, \frac{n^{(p-1)/d}}{\mu} d_{\mathbb{X}_n,p}))$.

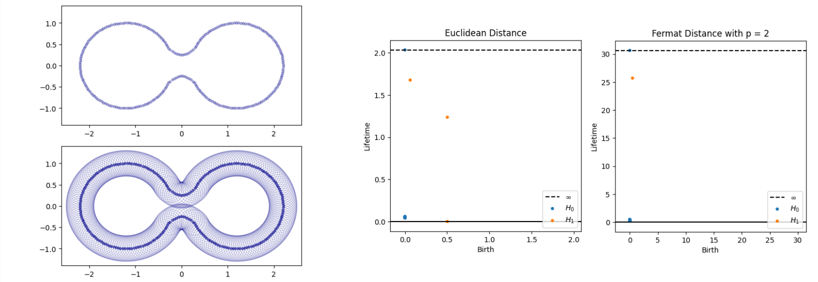
Theorem 3 (Borghini, F., Groisman, Mindlin, 2020)

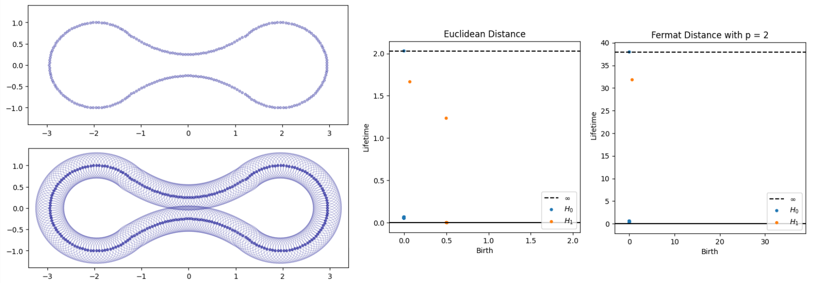
Given $\varepsilon > 0$ and $\lambda \in \left(\frac{p-1}{pd}, \frac{1}{d}\right)$ there exists a constant $\theta > 0$ such that

$$\mathbb{P}\Big(d_b\big(\mathrm{dgm}(\mathrm{Filt}(\mathcal{M}, d_{f,p})), \mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_n, \tfrac{n^{(p-1)/d}}{\mu} d_{\mathbb{X}_n,p}))\big) > \varepsilon\Big) \leqslant \exp\left(-\theta n^{(1-\lambda d)/(d+2p)}\right)$$

for n large enough.

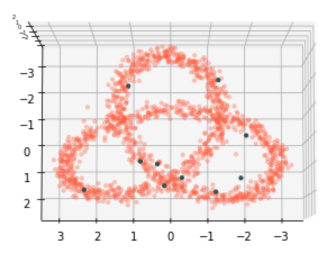
Example


Euclidean Distance
Fermat Distance with p = 2
Lifetime
Birth
∞
H_0
H_1


Euclidean Distance
Fermat Distance with p = 2
Lifetime
Birth
∞
H_0
H_1

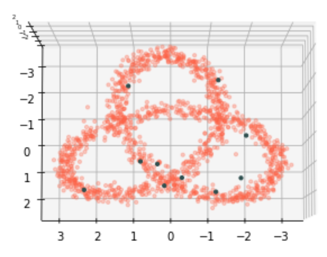
Experiment with outliers & noise

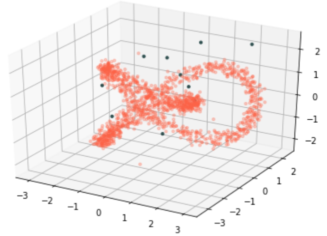
A sample of 1500 points from the **trefoil knot** with **noise and outliers**.



Experiment with outliers & noise

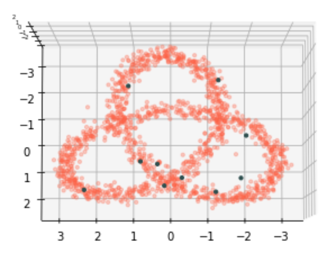
A sample of 1500 points from the **trefoil knot** with **noise and outliers**.

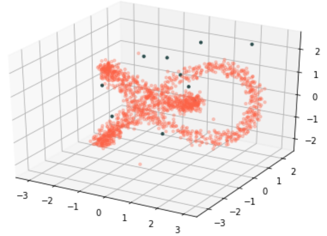


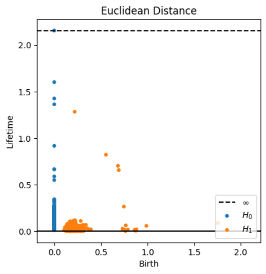


Experiment with outliers & noise

A sample of 1500 points from the **trefoil knot** with **noise and outliers**.

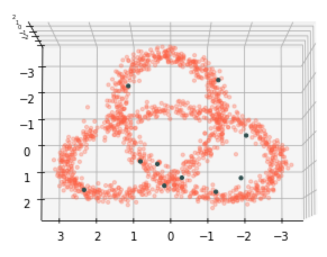


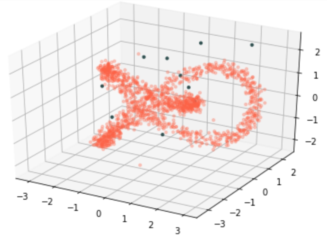


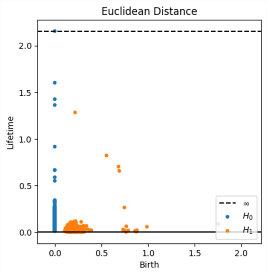


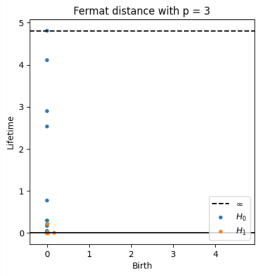
Experiment with outliers & noise

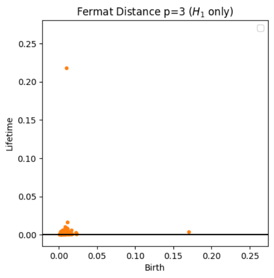
A sample of 1500 points from the **trefoil knot** with **noise and outliers**.







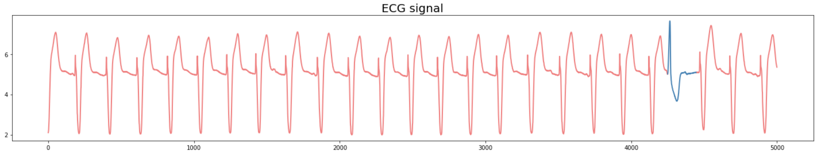




Applications

Time series: Anomaly detection

Electrocardiogram signal with abnormal heartbeat (arrhythmia).

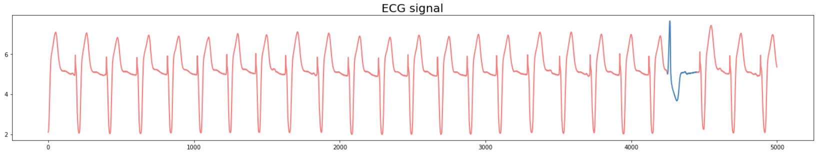


¶

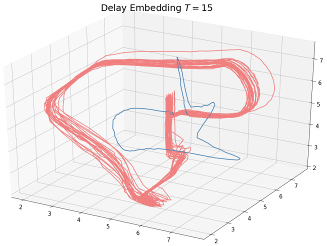
¶Data from Physionet database, MIT Laboratory for Computational Physiology.

Time series: Anomaly detection

Electrocardiogram signal with abnormal heartbeat (arrhythmia).



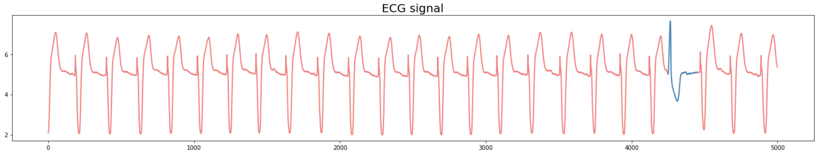
¶



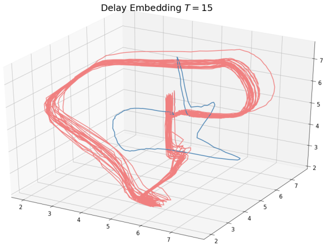
¶Data from Physionet database, MIT Laboratory for Computational Physiology.

Time series: Anomaly detection

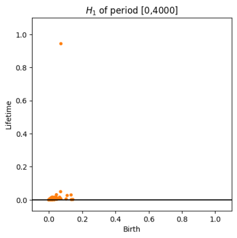
Electrocardiogram signal with abnormal heartbeat (arrhythmia).



¶



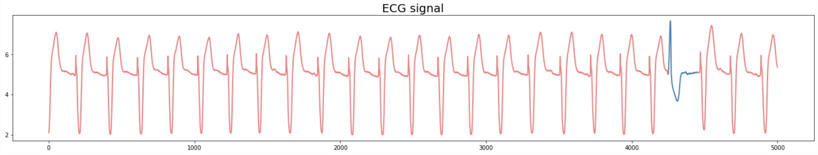
Persistence diagrams with Fermat distance for $p = 2$.



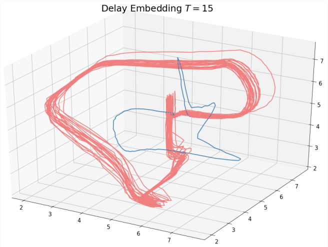
¶Data from Physionet database, MIT Laboratory for Computational Physiology.

Time series: Anomaly detection

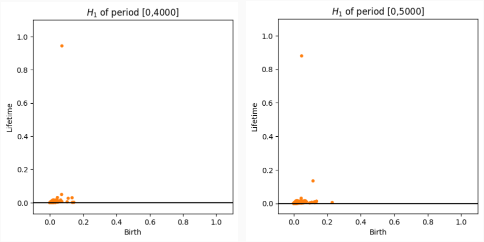
Electrocardiogram signal with abnormal heartbeat (arrhythmia).



¶



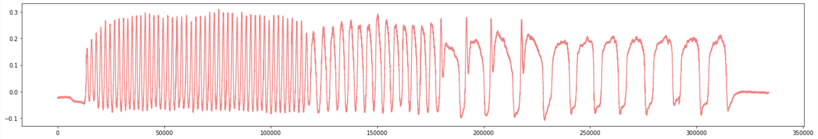
Persistence diagrams with Fermat distance for $p = 2$.



¶Data from Physionet database, MIT Laboratory for Computational Physiology.

Time series: Periodicity

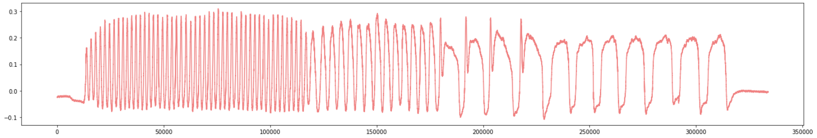
Observation of the pressure in the air sacs of a canary during singing.



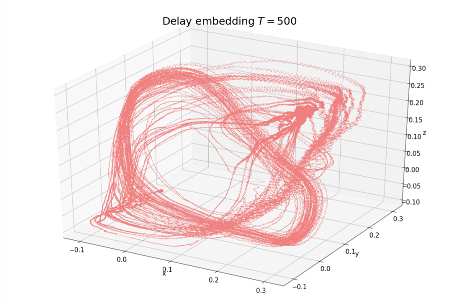
‖ Data from experimental records, Laboratory of Dynamical Systems, Physics Department, University of Buenos Aires.

Time series: Periodicity

Observation of the pressure in the air sacs of a canary during singing.



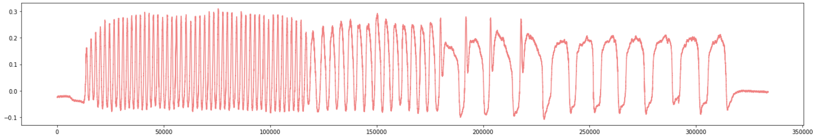
‖



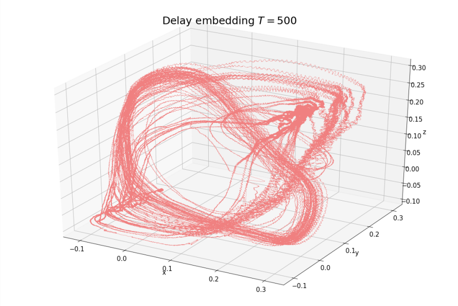
‖ Data from experimental records, Laboratory of Dynamical Systems, Physics Department, University of Buenos Aires.

Time series: Periodicity

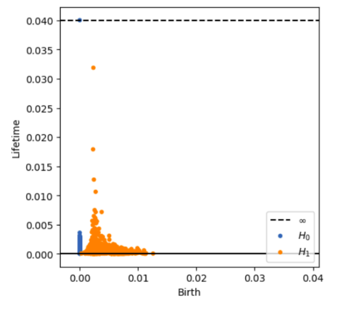
Observation of the pressure in the air sacs of a canary during singing.



‖



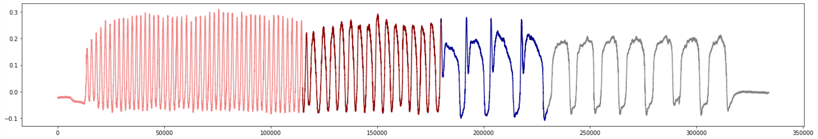
Persistence diagram with Fermat distance $p = 2$.

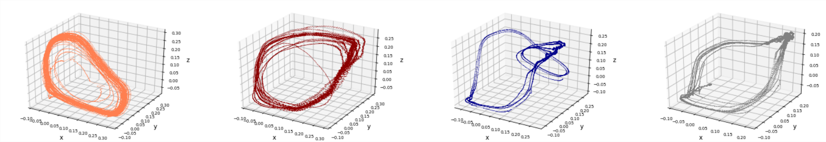


‖Data from experimental records, Laboratory of Dynamical Systems, Physics Department, University of Buenos Aires.

Time series: Periodicity

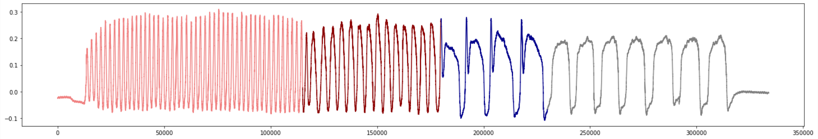
A canary song is composed by a concatenation of different syllabus patterns in the pressure in their air sacs.

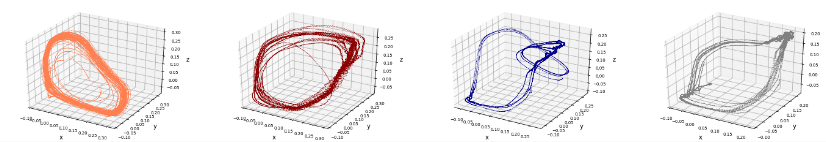




Time series: Periodicity

A canary song is composed by a concatenation of different syllabus patterns in the pressure in their air sacs.





Work in progress: Fit parameters of physical models of the underlying dynamical system using this correspondence between pressure patterns and 1-dimensional cycles.

References

- E. Borghini, X. F., P. Groisman, G. Mindlin. *Intrinsic persistent homology via density-based distance learning.* arXiv:2012.07621 (2020)
- Code: *https://github.com/ximenafernandez/intrinsicPH*
- Python library `fermat`. Author: F. Sapienza
  Documentation: *http://www.aristas.com.ar/fermat/index.html*.

email: `x.l.fernandez@swansea.ac.uk`

Thanks!