Intrinsic persistent homology via density-based metric learning

Ximena Fernández
joint work with E. Borghini, P. Groisman and G. Mindlin

IMSI Workshop on Topological Data Analysis
28th April 2021

EPSRC Centre for Topological Data Analysis

Prifysgol Abertawe
Swansea University

Image: Keenan Crane

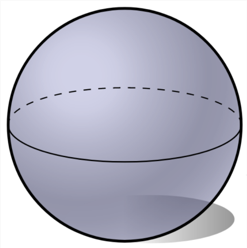
The problem

Metric learning in Riemannian manifolds

$(\mathcal{M}, g)$ a *d*-dimensional Riemannian manifold with associated Riemannian distance

$$d_{\mathcal{M}}(x, y) = \inf_{\gamma} \int_I \sqrt{g(\dot{\gamma}_t, \dot{\gamma}_t)} dt,$$

over all $\gamma : I \to \mathcal{M}$ with $\gamma(0) = x$, and $\gamma(1) = y$.

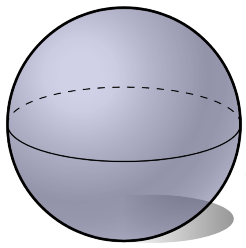


Metric learning in Riemannian manifolds

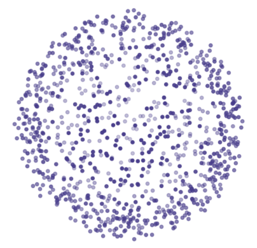
$(\mathcal{M}, g)$ a d-dimensional Riemannian manifold with associated Riemannian distance

$$d_{\mathcal{M}}(x, y) = \inf_{\gamma} \int_I \sqrt{g(\dot{\gamma}_t, \dot{\gamma}_t)} dt,$$

over all $\gamma : I \to \mathcal{M}$ with $\gamma(0) = x$, and $\gamma(1) = y$.



$\mathbb{X}_n = \{x_1, x_2, \dots, x_n\}$ a finite sample of $\mathcal{M}$.

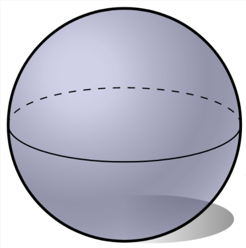


Metric learning in Riemannian manifolds

$(\mathcal{M}, g)$ a *d*-dimensional Riemannian manifold with associated Riemannian distance

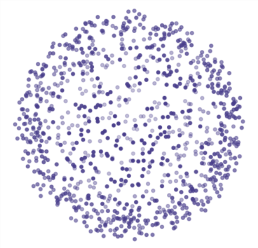
$$d_{\mathcal{M}}(x, y) = \inf_{\gamma} \int_I \sqrt{g(\dot{\gamma}_t, \dot{\gamma}_t)}dt,$$

over all $\gamma : I \to \mathcal{M}$ with $\gamma(0) = x$, and $\gamma(1) = y$.



$\mathbb{X}_n = \{x_1, x_2, \ldots, x_n\}$ a finite sample of $\mathcal{M}$.

How to infer the Riemannian distance from the sample?



Inherited Riemannian metric

If $\mathcal{M}$ is embedded in $\mathbb{R}^D$ and $g(x, y) = \langle x, y \rangle$ is the **inner product in $\mathbb{R}^D$**, the associated Riemannian distance is

$$d_{\mathcal{M}}(x, y) = \inf_{\gamma} \int_I ||\dot{\gamma}_t|| dt$$

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

Inherited Riemannian metric

If $\mathcal{M}$ is embedded in $\mathbb{R}^D$ and $g(x,y) = \langle x, y \rangle$ is the **inner product in $\mathbb{R}^D$**, the associated Riemannian distance is

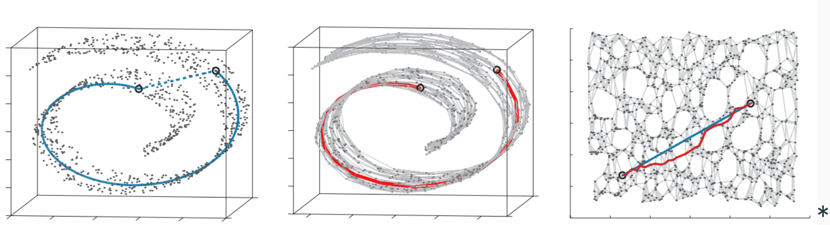
$$d_{\mathcal{M}}(x,y) = \inf_{\gamma} \int_I ||\dot{\gamma}_t|| dt$$

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

Given $\varepsilon > 0$, consider the **ε-graph** $G_\varepsilon(\mathbb{X}_n)$ and define the **estimator***

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

over all $\gamma = (x, x_1, \ldots, , x_r, y)$ with $(x_i, x_{i+1}) \in E(G_\varepsilon)$ for all $1 \leqslant i \leqslant r$.



* Bernstein, de Silva, Langford, Tenenbaum (2000).

Theorem (Bernstein, de Silva, Langford, Tenenbaum, 2000)

Let $\mathcal{M}$ be a closed d-dimensional Riemannian manifold embedded in $\mathbb{R}^D$ with inherited Riemannian distance $d_{\mathcal{M}}$. Let $\mathbb{X}_n$ be a finite sample of $\mathcal{M}$.

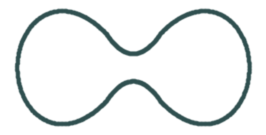
Assume $\varepsilon_n \to 0$ and $n\varepsilon_n^d \to \infty$. Then,

$$\lim_{n\to\infty} \sup_{x,y\in\mathcal{M}} |d_{\mathbb{X}_n,\varepsilon_n}(x,y) - d_{\mathcal{M}}(x,y)| = 0$$

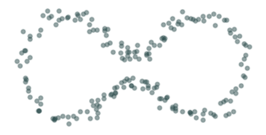
in probability, with almost sure convergence provided $n\varepsilon_n^d/\log n \to \infty$.

Dependence on ε (and density)

Manifold

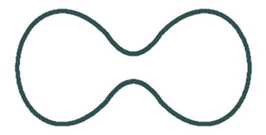


Noisy sample

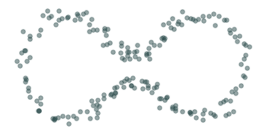


Dependence on ε (and density)

Manifold

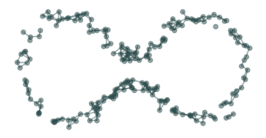


Noisy sample

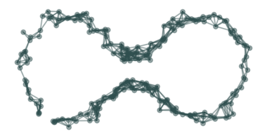


ε- graph

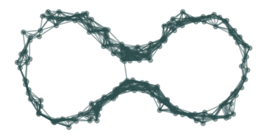
$\varepsilon = 0.2$



$\varepsilon = 0.3$



$\varepsilon = 0.4$

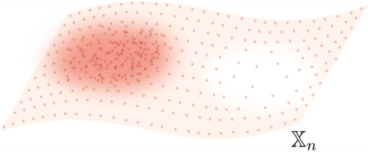


The manifold (and density) assumption

$\mathbb{X}_n = \{x_1, x_2, \ldots, x_n\}$ a finite set of points in $\mathbb{R}^D$.

We assume that:

- $\mathbb{X}_n$ lies in a d-dimensional Riemannian manifold $\mathcal{M}$,
- $\mathbb{X}_n$ is drawn according to a density f.

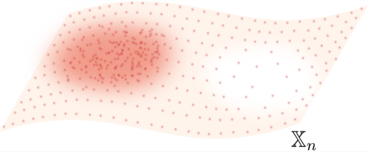


The manifold (and density) assumption

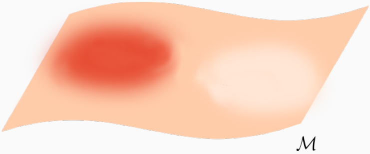
$\mathbb{X}_n = \{x_1, x_2, \ldots, x_n\}$ a finite set of points in $\mathbb{R}^D$.

We assume that:

- $\mathbb{X}_n$ lies in a d-dimensional Riemannian manifold $\mathcal{M}$,
- $\mathbb{X}_n$ is drawn according to a density f.



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

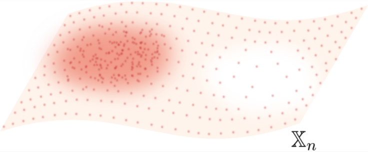


The manifold (and density) assumption

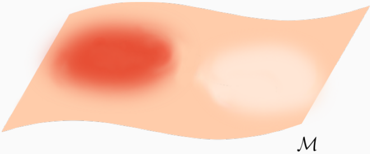
$\mathbb{X}_n = \{x_1, x_2, \ldots, x_n\}$ a finite set of points in $\mathbb{R}^D$.

We assume that:

- $\mathbb{X}_n$ lies in a d-dimensional Riemannian manifold $\mathcal{M}$,
- $\mathbb{X}_n$ is drawn according to a density f.



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



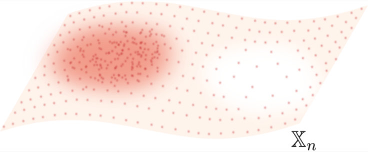
Consider a (new) Riemannian metric that depends on f.

The manifold (and density) assumption

$\mathbb{X}_n = \{x_1, x_2, \ldots, x_n\}$ a finite set of points in $\mathbb{R}^D$.

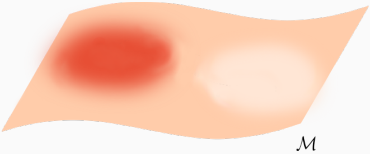
We assume that:

- $\mathbb{X}_n$ lies in a d-dimensional Riemannian manifold $\mathcal{M}$,
- $\mathbb{X}_n$ is drawn according to a density f.



Find an estimator of the (density-based) Riemannian metric from the sample.

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



Consider a (new) Riemannian metric that depends on f.

Density-based metric learning

Deformed Riemannian metric

- Let $(\mathcal{M}, g)$ be a Riemannian manifold and let $f : \mathcal{M} \to \mathbb{R}_{>0}$ be a smooth density.

*Hwang, Damelin, Hero (2016), Groisman, Jonckheere, Sapienza (2018)

Deformed Riemannian metric

- Let $(\mathcal{M}, g)$ be a Riemannian manifold and let $f : \mathcal{M} \to \mathbb{R}_{>0}$ be a smooth density.
- For $q > 0$, and consider the deformed metric tensor

$$g_q = f^{-2q} g.^*$$

*Hwang, Damelin, Hero (2016), Groisman, Jonckheere, Sapienza (2018)

Deformed Riemannian metric

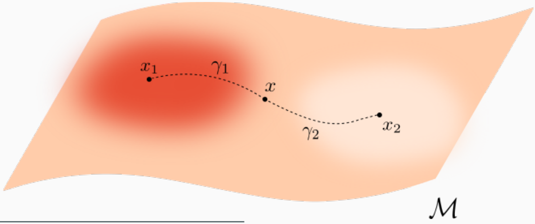
- Let $(\mathcal{M}, g)$ be a Riemannian manifold and let $f : \mathcal{M} \to \mathbb{R}_{>0}$ be a smooth density.
- For $q > 0$, and consider the deformed metric tensor

$$g_q = f^{-2q} g.^*$$

- The induced **deformed Riemannian distance** in $\mathcal{M}$ is

$$d_{f,q}(x, y) = \inf_{\gamma} \int_I \frac{1}{f(\gamma_t)^q} \sqrt{g(\dot{\gamma}_t, \dot{\gamma}_t)} dt$$

over all $\gamma : I \to \mathcal{M}$ with $\gamma(0) = x$ and $\gamma(1) = y$.



*Hwang, Damelin, Hero (2016), Groisman, Jonckheere, Sapienza (2018)

Deformed Riemannian metric

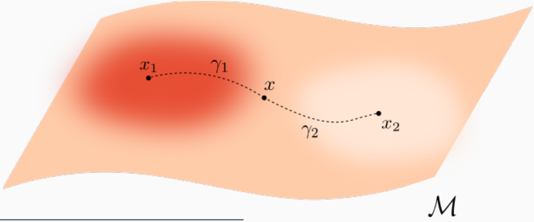
- Let $(\mathcal{M}, g)$ be a Riemannian manifold and let $f : \mathcal{M} \to \mathbb{R}_{>0}$ be a smooth density.
- For $q > 0$, and consider the deformed metric tensor

$$g_q = f^{-2q} g.^*$$

- The induced **deformed Riemannian distance** in $\mathcal{M}$ is

$$d_{f,q}(x,y) = \inf_{\gamma} \int_I \frac{1}{f(\gamma_t)^q} \sqrt{g(\dot{\gamma}_t, \dot{\gamma}_t)} dt$$

over all $\gamma : I \to \mathcal{M}$ with $\gamma(0) = x$ and $\gamma(1) = y$.



*Hwang, Damelin, Hero (2016), Groisman, Jonckheere, Sapienza (2018)

Fermat distance

- Let $\mathbb{X}_n \subseteq \mathbb{R}^D$ a sample of points.

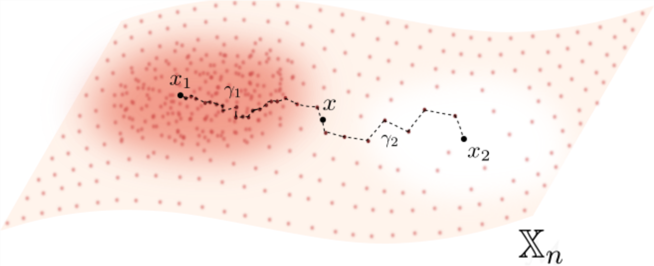
† Groisman, Jonckheere, Sapienza (2018), Mckenzie and Damelin (2019).

Fermat distance

- Let $\mathbb{X}_n \subseteq \mathbb{R}^D$ a sample of points.
- For $p > 1$, the (sample) **Fermat distance**[†] between $x, y \in \mathbb{R}^D$ is defined by

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

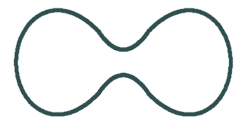
over all paths $\gamma = (x_0, \ldots, x_{r+1})$ of finite length with $x_0 = x$, $x_{r+1} = y$ and $\{x_1, x_2, \ldots, x_r\} \subseteq \mathbb{X}_n$.



[†] Groisman, Jonckheere, Sapienza (2018), Mckenzie and Damelin (2019).

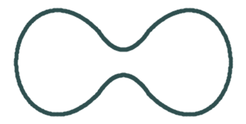
Example

Manifold

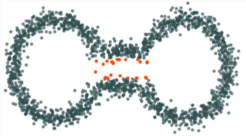


Example

Manifold

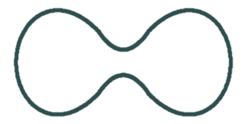


Sample

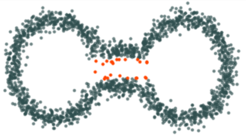


Example

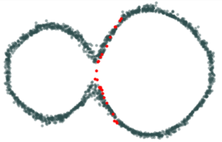
Manifold



Sample

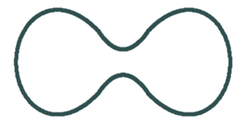


Isomap

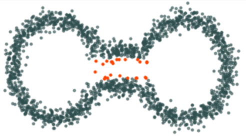


Example

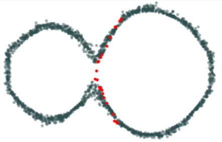
Manifold



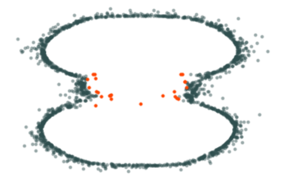
Sample



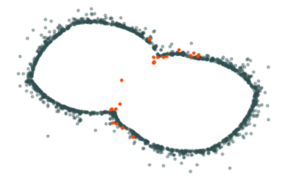
Isomap



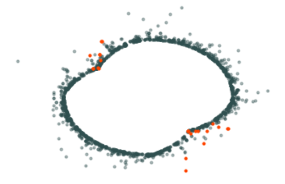
Fermat p = 1.5



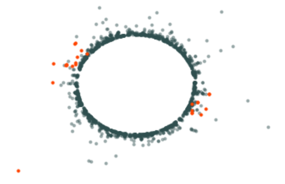
Fermat p = 2.0



Fermat p = 2.5



Fermat p = 3.0



Previous work

Theorem (Groisman, Jonckheere, Sapienza (2018))

Let $(\mathcal{M}, g)$ be an **isometric**[‡] C^1 d-dimensional Riemannian manifold embedded in $\mathbb{R}^D$ with inherited metric tensor. Let $\mathbb{X}_n \subseteq \mathcal{M}$ be a set of n independent sample points with common smooth density $f : \mathcal{M} \to \mathbb{R}_{>0}$.

Given $p > 1$, there exists $\mu = \mu(p, d) > 0$ such that for any $x, y \in \mathcal{M}$,

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

with $q = (p-1)/d$.

[‡] $\mathcal{M}$ 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 $\varphi : \bar{S} \to \mathbb{R}^D$ such that $\varphi(\bar{S}) = \mathcal{M}$ and $\varphi : \bar{S} \to \mathcal{M}$ is a Riemannian isometry.

Previous work

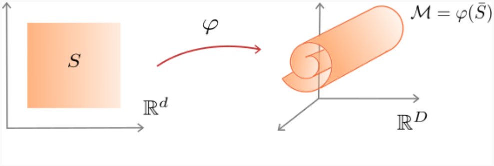
Theorem (Groisman, Jonckheere, Sapienza (2018))

Let $(\mathcal{M}, g)$ be an **isometric**‡ C^1 d-dimensional Riemannian manifold embedded in $\mathbb{R}^D$ with inherited metric tensor. Let $\mathbb{X}_n \subseteq \mathcal{M}$ be a set of n independent sample points with common smooth density $f : \mathcal{M} \to \mathbb{R}_{>0}$.

Given $p > 1$, there exists $\mu = \mu(p, d) > 0$ such that for any $x, y \in \mathcal{M}$,

$$\lim_{n \to +\infty} \tfrac{n^q}{\mu} d_{\mathbb{X}_n, p}(x, y) = d_{f,q}(x, y) \text{ almost surely}$$

with $q = (p - 1)/d$.



‡ $\mathcal{M}$ 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 $\varphi : \bar{S} \to \mathbb{R}^D$ such that $\varphi(\bar{S}) = \mathcal{M}$ and $\varphi : \bar{S} \to \mathcal{M}$ is a Riemannian isometry.

Previous work

Theorem (Hwang, Damelin, Hero (2016))

Let $(\mathcal{M}, g)$ be a closed smooth d-dimensional manifold with associated Riemannian distance $d_{\mathcal{M}}$. Let $\mathbb{X}_n \subseteq \mathcal{M}$ be a set of n independent sample points with common smooth density $f : \mathcal{M} \to \mathbb{R}_{>0}$.

---

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

Previous work

Theorem (Hwang, Damelin, Hero (2016))

Let $(\mathcal{M}, g)$ be a closed smooth d-dimensional manifold with associated Riemannian distance $d_{\mathcal{M}}$. Let $\mathbb{X}_n \subseteq \mathcal{M}$ be a set of n independent sample points with common smooth density $f : \mathcal{M} \to \mathbb{R}_{>0}$.

Given $p > 1$ (and $q = (p-1)/d$), there exists $\mu = \mu(p, d) > 0$ such that for all $\varepsilon > 0$ and $b > 0$

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

for some $\theta = \theta(\varepsilon) > 0$ and sufficiently large n.

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

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

Let $(\mathcal{M}, g)$ be a closed smooth d-dimensional Riemannian manifold embedded in $\mathbb{R}^D$ with inherited metric tensor. Let $\mathbb{X}_n \subseteq \mathcal{M}$ be a set of n independent sample points with common smooth density $f : \mathcal{M} \to \mathbb{R}_{>0}$.

Given $p > 1$ and $q = (p-1)/d$, there exists a constant $\mu = \mu(p, d)$ such that for every $\lambda \in \big((p-1)/pd, 1/d\big)$ and $\varepsilon > 0$ there exist $\theta > 0$ satisfying

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

for n large enough.

‘Metric space’ learning

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

'Metric space' learning

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

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

Given $p > 1$ and $q = (p-1)/d$, there exists a constant $\mu = \mu(p, d)$ such that for every $\lambda \in \big((p-1)/pd, 1/d\big)$ and $\varepsilon > 0$ there exist $\theta > 0$ satisfying

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

for n large enough.

Proof. Thm 1 + some additional work.

Intrinsic persistent homology

Convergence of persistence diagrams

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

Convergence of persistence diagrams

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

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

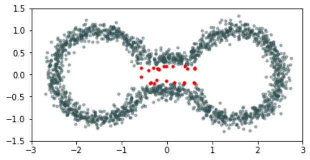
Given $p > 1$ and $q = (p-1)/d$, there exists a constant $\mu = \mu(p, d)$ such that for every $\lambda \in \big((p-1)/pd, 1/d\big)$ and $\varepsilon > 0$ there exist $\theta > 0$ satisfying

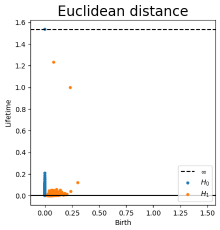
$$\mathbb{P}\Big(d_b\big(\mathrm{dgm}(\mathrm{Filt}(\mathcal{M}, d_{f,q})), \mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_n, \tfrac{n^q}{\mu} d_{\mathbb{X}_n,p}))\big) > \varepsilon\Big) \leqslant \exp\big(-\theta n^{(1-\lambda d)/(d+2p)}\big)$$

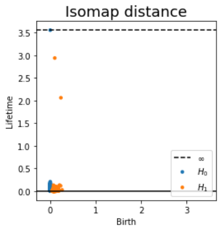
for n large enough.

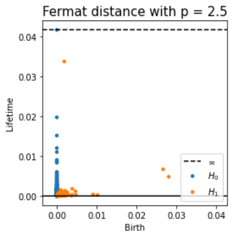
Proof. Thm 2 + Stability Thm.

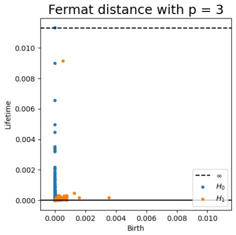
Example

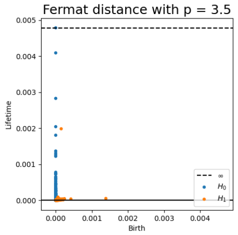



Euclidean distance
Lifetime
Birth
∞
H_0
H_1


Isomap distance
Lifetime
Birth
∞
H_0
H_1


Fermat distance with p = 2.5
Lifetime
Birth
∞
H_0
H_1

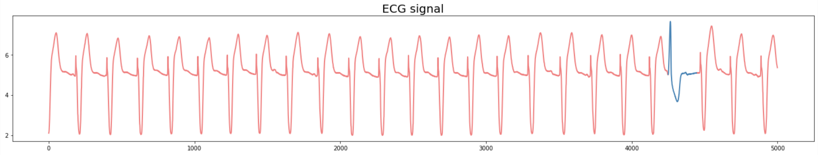

Fermat distance with p = 3
Lifetime
Birth
∞
H_0
H_1


Fermat distance with p = 3.5
Lifetime
Birth
∞
H_0
H_1

Applications

Time series: Anomaly detection

Electrocardiogram signal with abnormal heartbeat (arrhythmia).

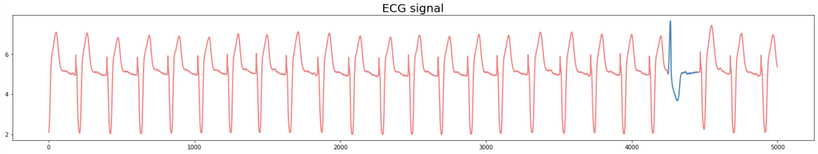


¶

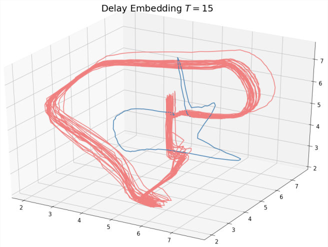
¶ Data from Physionet database, MIT Laboratory for Computational Physiology.

Time series: Anomaly detection

Electrocardiogram signal with abnormal heartbeat (arrhythmia).



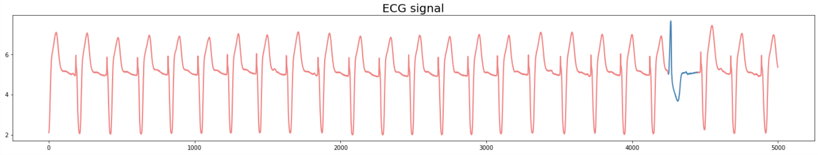
¶



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

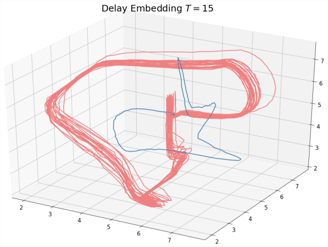
Time series: Anomaly detection

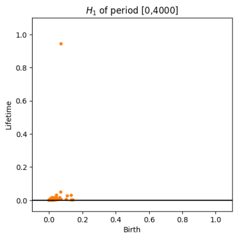
Electrocardiogram signal with abnormal heartbeat (arrhythmia).



¶

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

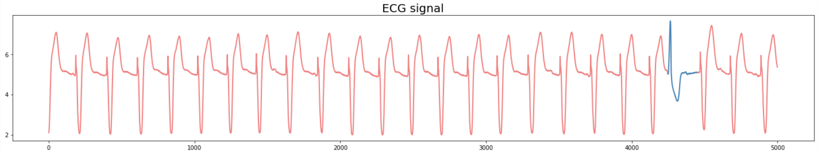




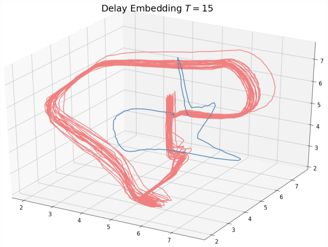
¶Data from Physionet database, MIT Laboratory for Computational Physiology.

Time series: Anomaly detection

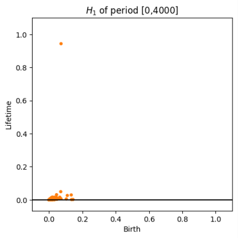
Electrocardiogram signal with abnormal heartbeat (arrhythmia).

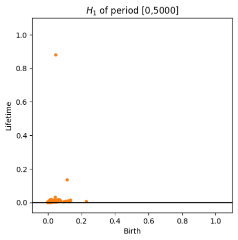


¶



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

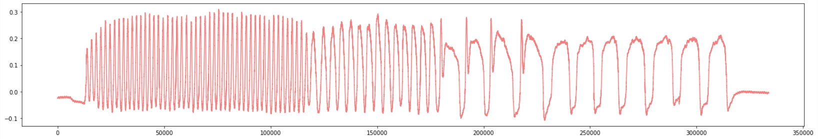




¶Data from Physionet database, MIT Laboratory for Computational Physiology.

Time series: Periodicity

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

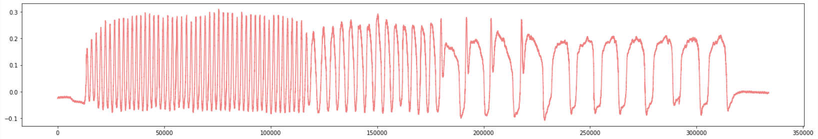


‖

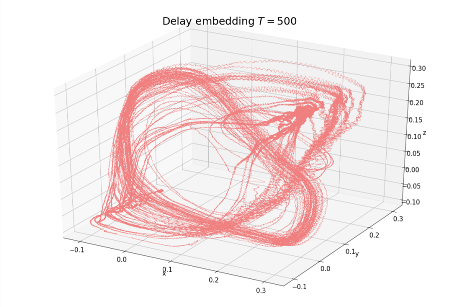
‖ Data from experimental records, Laboratory of Dynamical Systems, Physics Department, University of Buenos Aires.

Time series: Periodicity

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



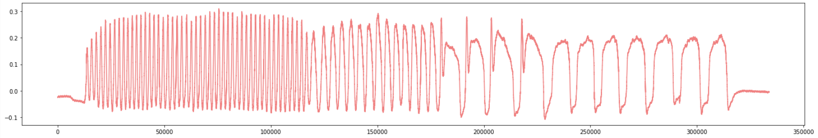
‖



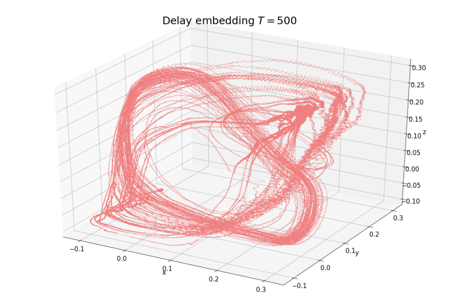
‖ Data from experimental records, Laboratory of Dynamical Systems, Physics Department, University of Buenos Aires.

Time series: Periodicity

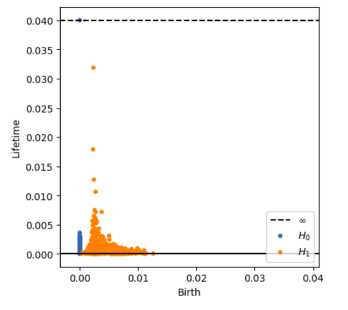
Observation of the pressure in the air sacs of a canary during singing.



‖

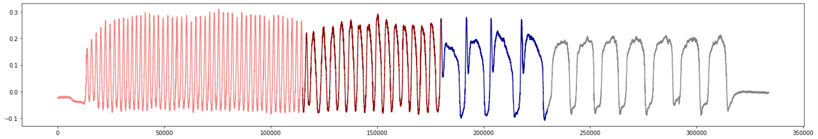


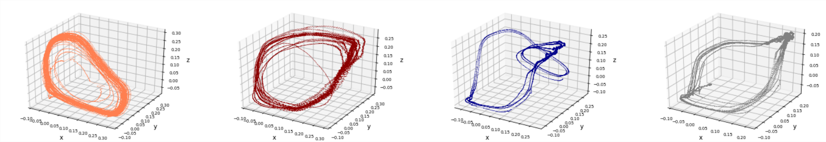
Persistence diagram Fermat distance $p = 1.5$.



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

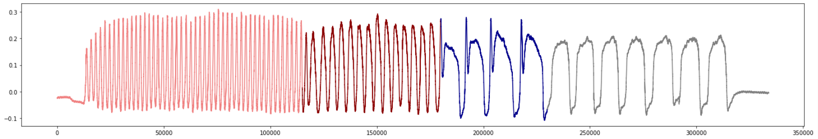
A canary song is composed by a concatenation of different syllabus patterns in the pressure in their air sacs.

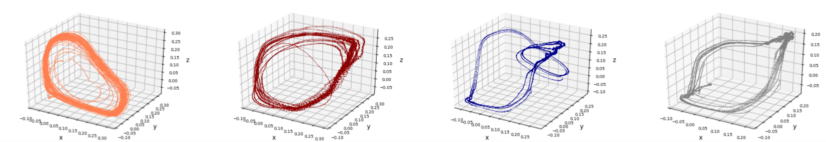




Time series: Periodicity

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

- *Preprint:* 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`.

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

Thanks for your attention!