Intrinsic persistent homology via density-based metric learning

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

38th Workshop in Geometric Topology
16th June 2021

*EPSRC Centre for Topological Data Analysis

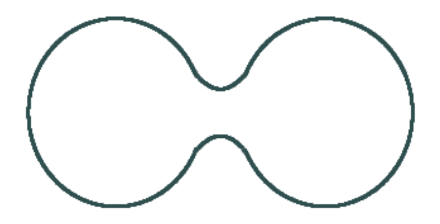
Prifysgol Abertawe
Swansea University

Image: Keenan Crane

The problem

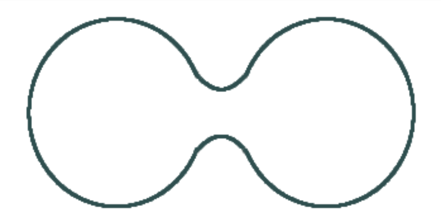
Homology inference

$(\mathcal{M}, g)$ a *d-dimensional Riemannian manifold* embedded in $\mathbb{R}^D$.



Homology inference

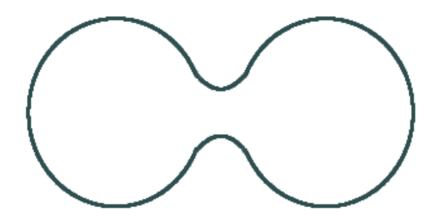
$(\mathcal{M}, g)$ a *d-dimensional Riemannian manifold* embedded in $\mathbb{R}^D$.



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

Homology inference

$(\mathcal{M}, g)$ a d-dimensional Riemannian manifold embedded in $\mathbb{R}^D$.

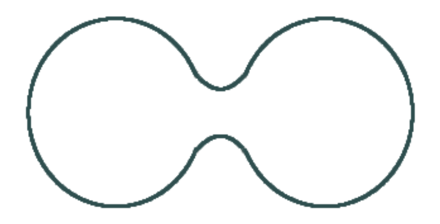


Q: How to infer the homology of $\mathcal{M}$ from the sample $\mathbb{X}_n$?

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

Homology inference

$(\mathcal{M}, g)$ a *d-dimensional Riemannian manifold* embedded in $\mathbb{R}^D$.

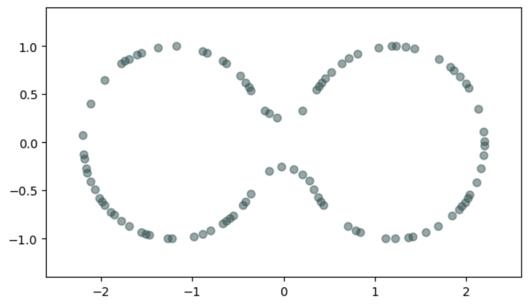


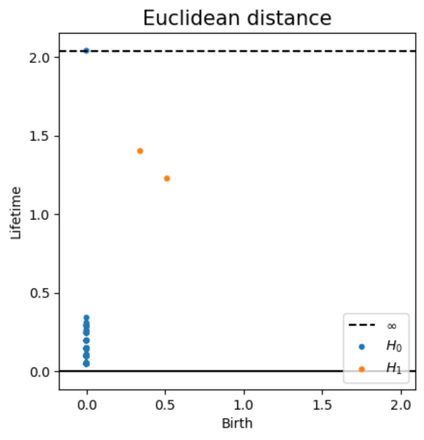
Q: How to infer the homology of $\mathcal{M}$ from the sample $\mathbb{X}_n$?

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

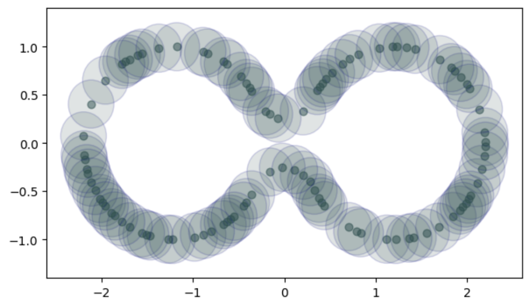
A: Compute persistent homology of $\mathbb{X}_n$.

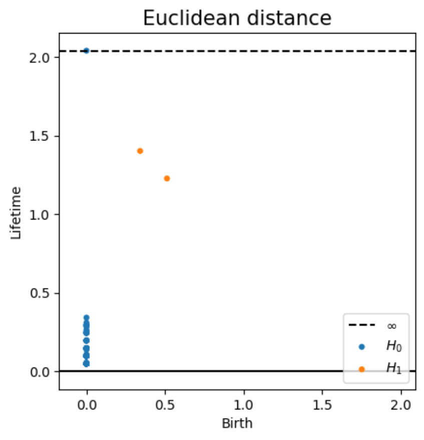
Ambient persistent homology





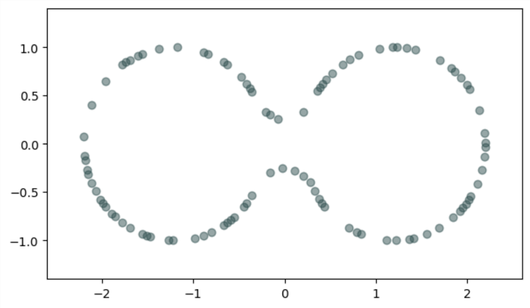
Ambient persistent homology

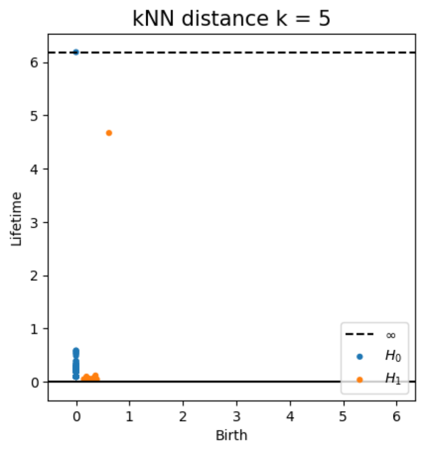




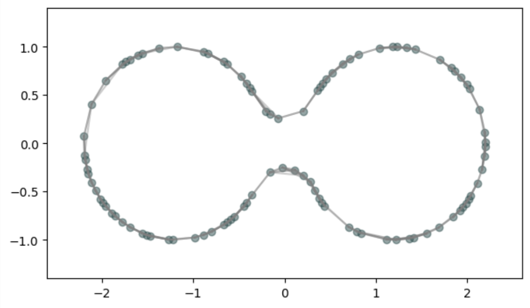
- $\mathrm{Rips}_\epsilon(\mathcal{M}, d_E) \simeq \mathcal{M}$ for $\epsilon < 2\mathrm{rch}(\mathcal{M})$

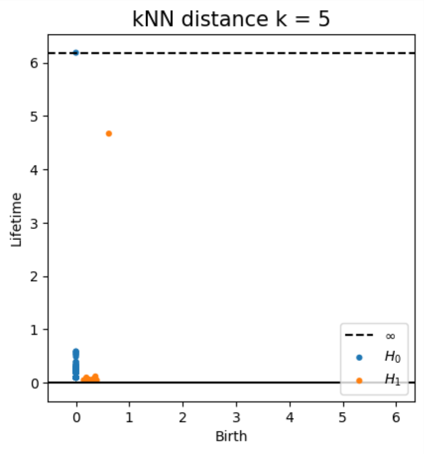
Intrinsic persistent homology





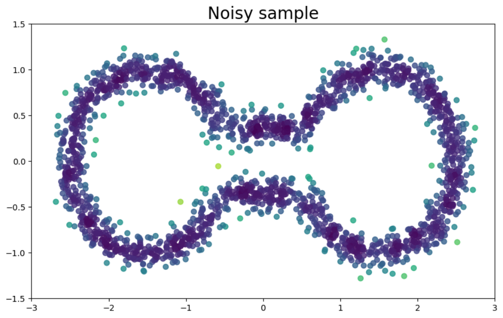
Intrinsic persistent homology

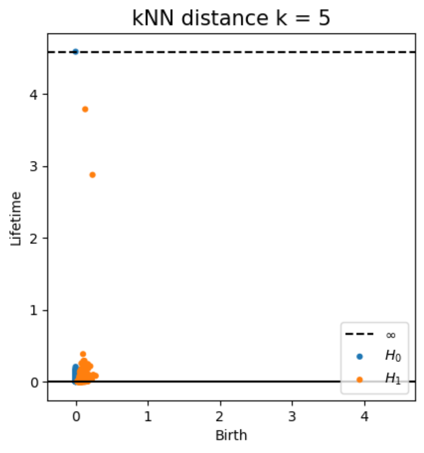




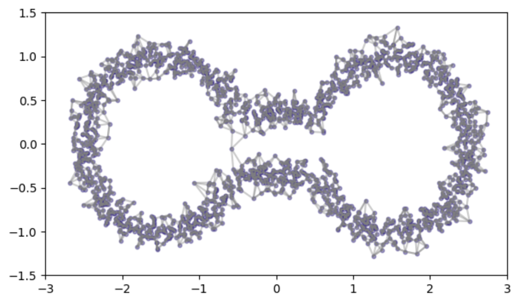
- $\mathrm{Rips}_\epsilon(\mathcal{M}, d_{\mathcal{M}}) \simeq \mathcal{M}$ for $\epsilon < \mathrm{conv}(\mathcal{M})$

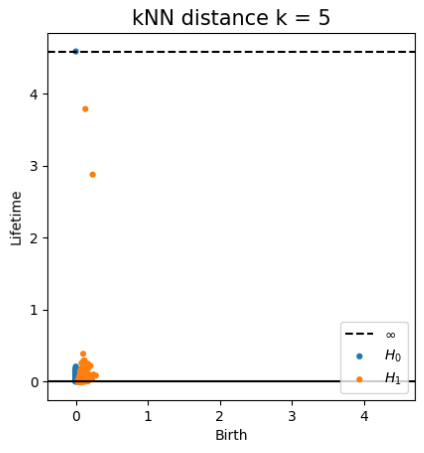
The problem of noise



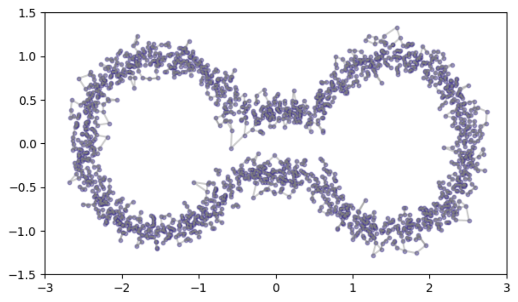


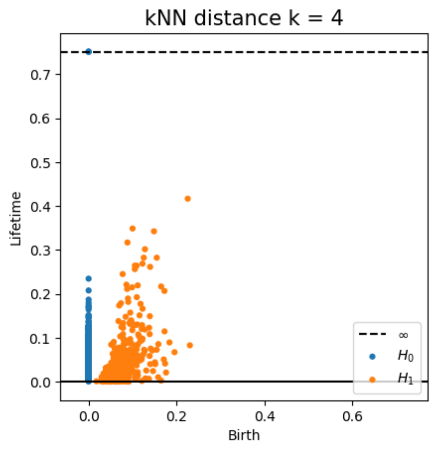
The problem of noise





The problem of noise





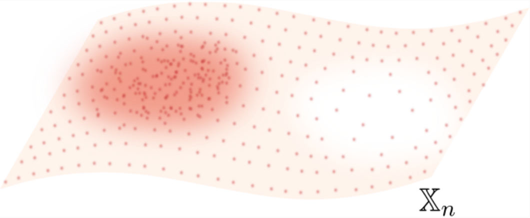
Density-based manifold learning

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 $\mathcal{M}$ embedded in $\mathbb{R}^D$,
- $**$ $\mathbb{X}_n$ is drawn according to a smooth density $f : \mathcal{M} \to \mathbb{R}_{>0}$.



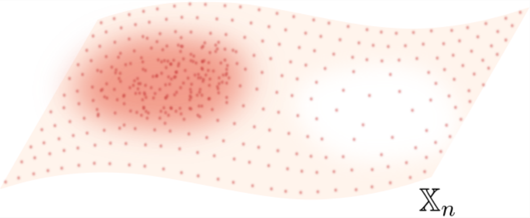
The manifold (and density) assumption

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

We assume that:

* $\mathbb{X}_n$ lies in a d-dimensional Riemannian manifold $\mathcal{M}$ embedded in $\mathbb{R}^D$,

** $\mathbb{X}_n$ is drawn according to a smooth density $f : \mathcal{M} \to \mathbb{R}_{>0}$.



Idea:

- Consider a Riemannian metric that depends on f.
- Find an estimator of the (density-based) Riemannian metric from the sample.

Deformed Riemannian metric

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

Deformed Riemannian metric

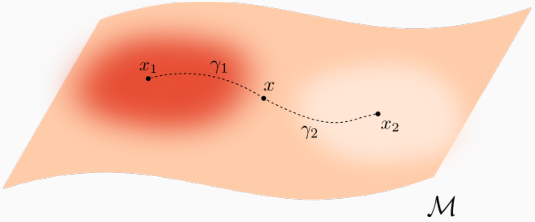
- Let $(\mathcal{M}, g)$ be a Riemannian manifold and let $f : \mathcal{M} \to \mathbb{R}_{>0}$ be a smooth density.
- For $q > 0$, and consider the deformed metric tensor $g_q = f^{-2q} g$.

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 $\gamma(1) = y$.

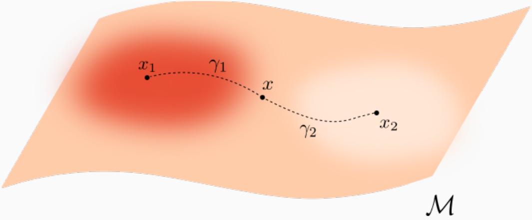


Deformed Riemannian metric

- 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 $\gamma(1) = y$.



Fermat distance

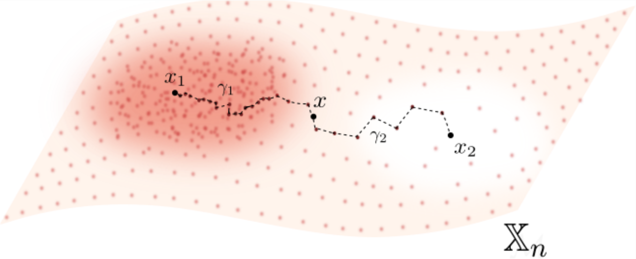
- Let $\mathbb{X}_n \subseteq \mathbb{R}^D$ a sample of points.

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

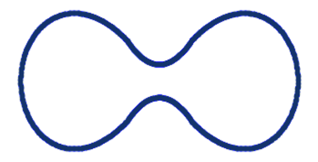
$$d_{\mathbb{X}_n,p}(x,y) = \inf_{\gamma} \sum_{i=0}^{r} |x_{i+1} - x_i|^p$$

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



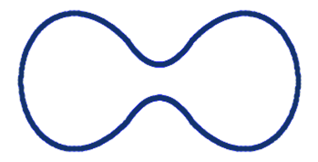
Example (Fermat distance)

Manifold

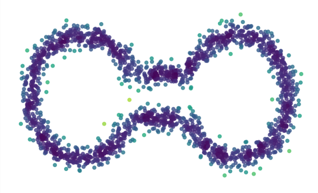


Example (Fermat distance)

Manifold

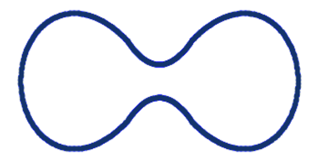


Sample

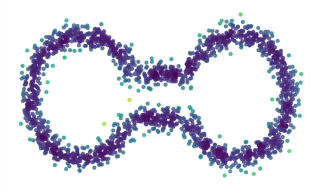


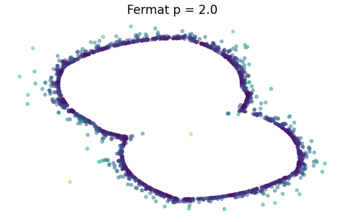
Example (Fermat distance)

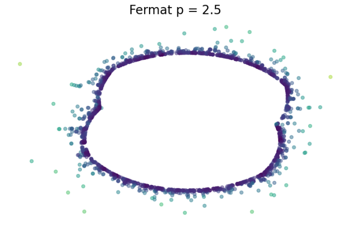
Manifold

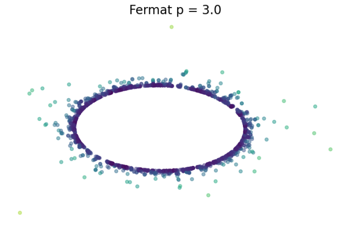


Sample









Convergence of metric spaces

For $p > 1$ and $q = (p-1)/d$,

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

Convergence of metric spaces

For $p > 1$ and $q = (p-1)/d$,

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

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

There exist a constant $C(n,p,d) > 0$ such that for every $\lambda \in \big((p-1)/pd, 1/d\big)$ and $\varepsilon > 0$ there exist $\theta > 0$ satisfying

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

for n large enough.

Convergence of persistence diagrams

For $p > 1$ and $q = (p-1)/d$,

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

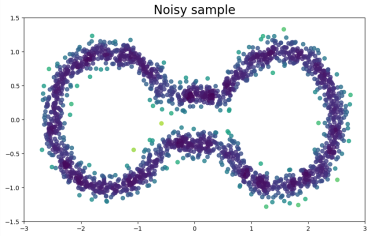
Corollary (Borghini, F., Groisman, Mindlin, 2020)

There exist a constant $C(n,p,d)$ such that for every $\lambda \in \big((p-1)/pd, 1/d\big)$ and $\varepsilon > 0$ there exist $\theta > 0$ satisfying

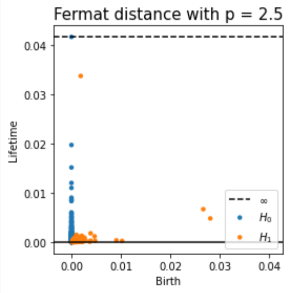
$$\mathbb{P}\Big(d_b\big(\mathrm{dgm}(\mathrm{Filt}(\mathcal{M}, d_{f,q})), \mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_n, C(n,p,d)d_{\mathbb{X}_n,p}))\big) > \varepsilon\Big) \leqslant \exp\big(-\theta n^{(1-\lambda d)/(d+2p)}\big)$$

for n large enough.

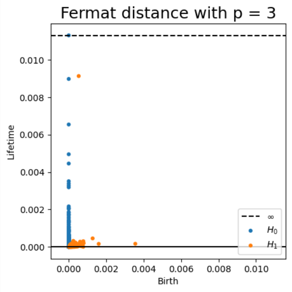
Example



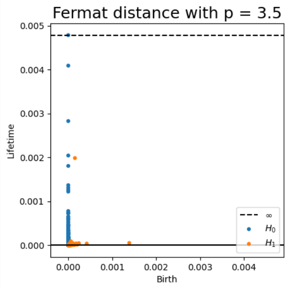
Noisy sample



Fermat distance with p = 2.5
Lifetime
Birth
∞
H_0
H_1

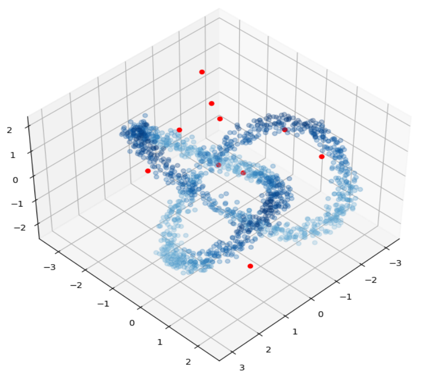


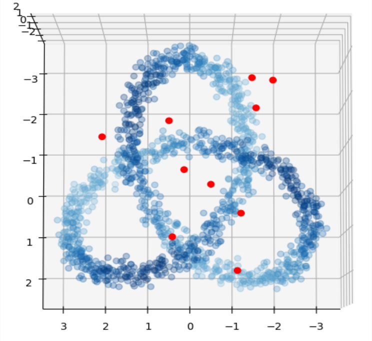
Fermat distance with p = 3
Lifetime
Birth
∞
H_0
H_1



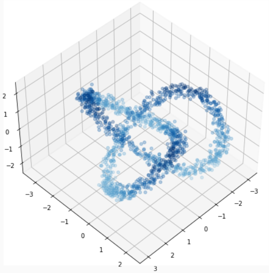
Fermat distance with p = 3.5
Lifetime
Birth
∞
H_0
H_1

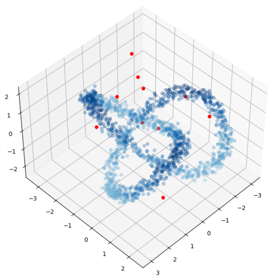
Robustness to outliers



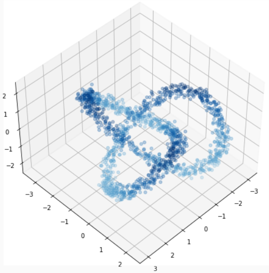


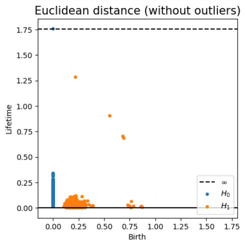
Robustness to outliers

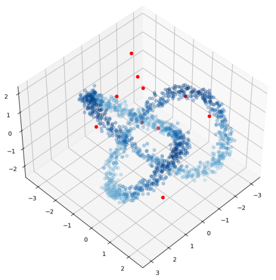




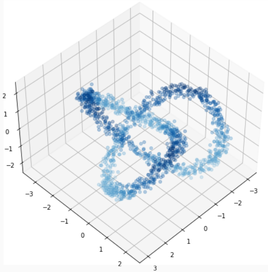
Robustness to outliers

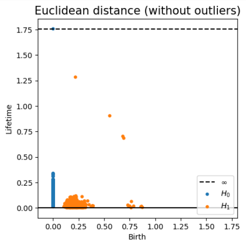


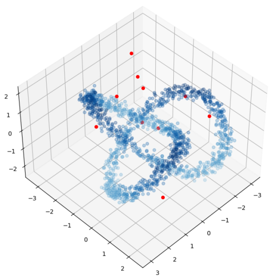


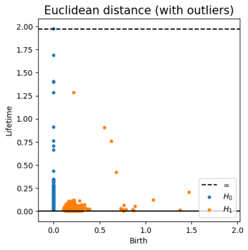


Robustness to outliers

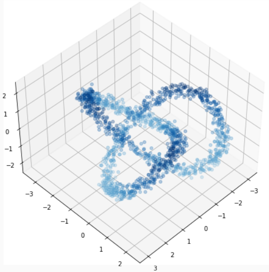


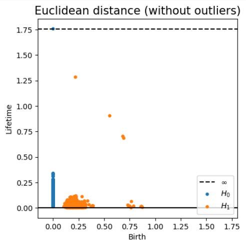


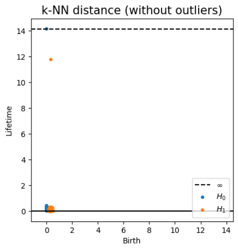


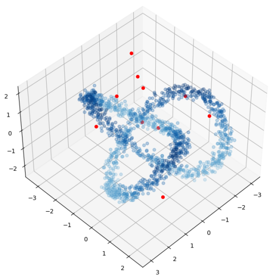


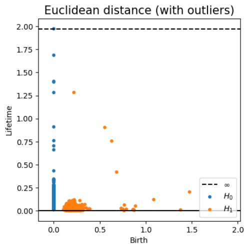
Robustness to outliers



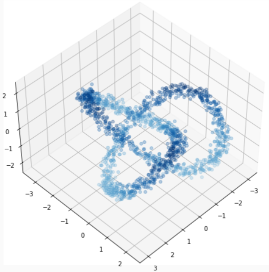


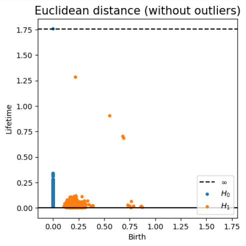


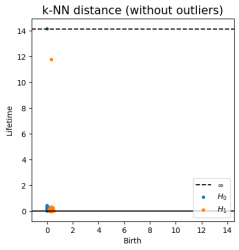


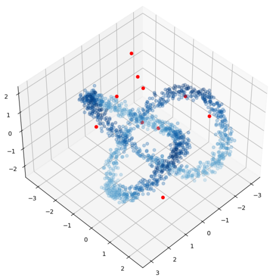


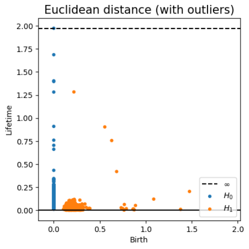
Robustness to outliers

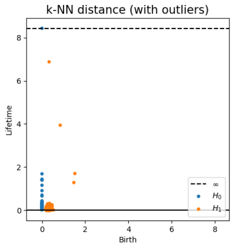




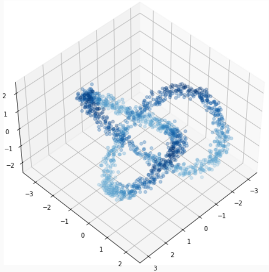


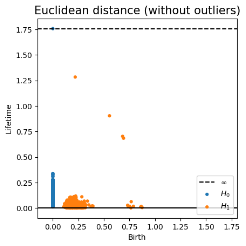


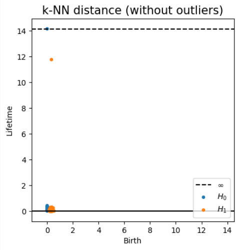


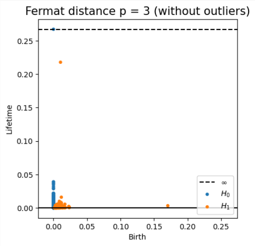


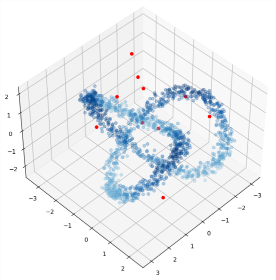
Robustness to outliers

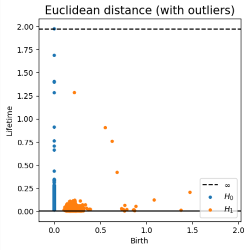


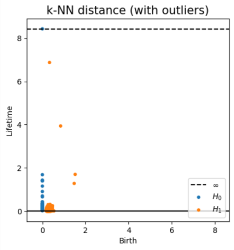




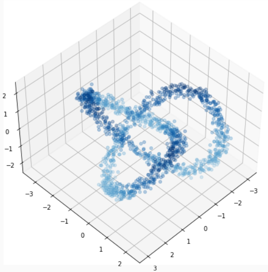


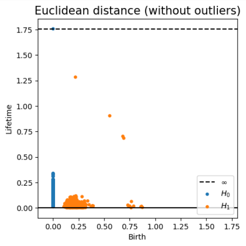


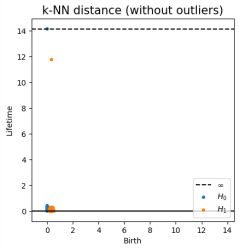


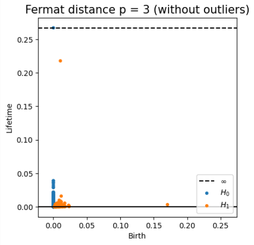


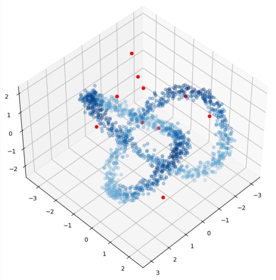
Robustness to outliers

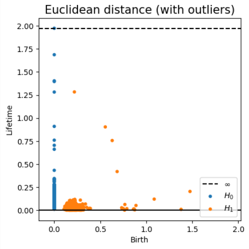


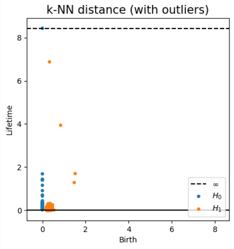


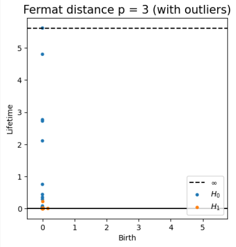




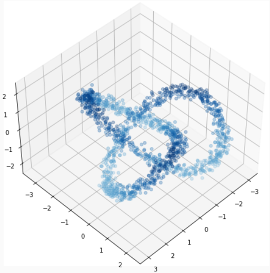


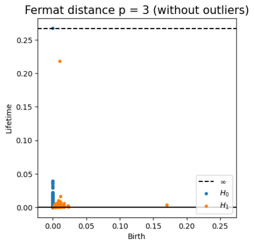


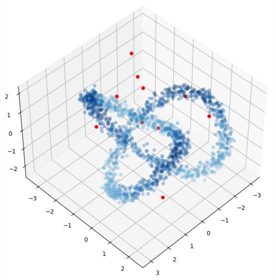


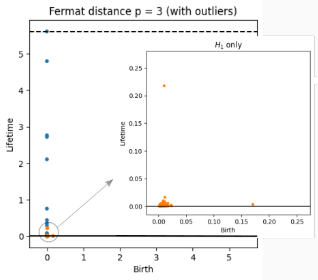


Robustness to outliers (Fermat distance)

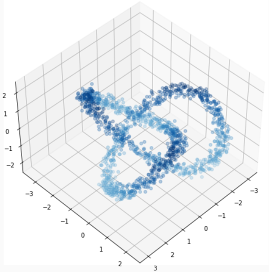


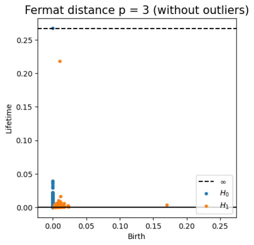


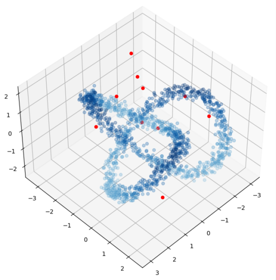


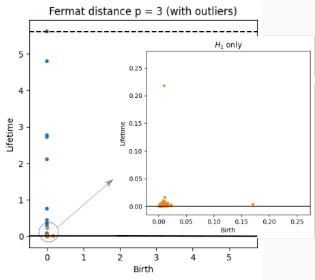


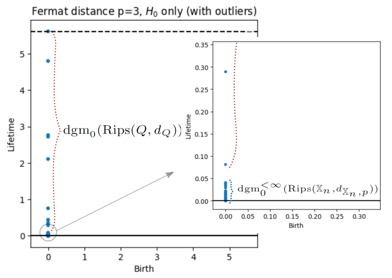
Robustness to outliers (Fermat distance)











Robustness to outliers

Proposition (Borghini, F., Groisman, Mindlin, 2021)

Let $\mathbb{X}_n$ be **sample** of $\mathcal{M}$ and let $Y \subseteq \mathbb{R}^D \setminus \mathcal{M}$ be a finite set of **outliers**.

There exists $\delta > 0$ such that for all $k > 0$ and $p > 1$,

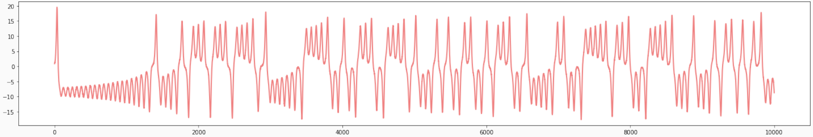
$$\mathrm{dgm}_k(\mathrm{Rips}_{<\delta^p}(\mathbb{X}_n \cup Y, d_{\mathbb{X}_n \cup Y, p})) = \mathrm{dgm}_k(\mathrm{Rips}_{<\delta^p}(\mathbb{X}_n, d_{\mathbb{X}_n, p})).$$

Here, for p large enough $\delta^p > \mathrm{diam}(\mathbb{X}_n, d_{\mathbb{X}_n, p})$.

Applications to signal analysis

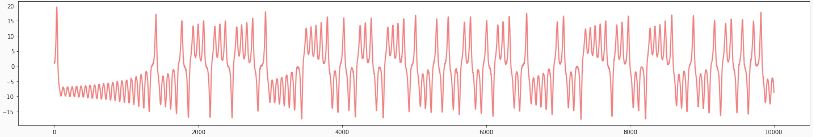
Delay embedding

- Signal $X : [t_0, t_1] \rightarrow \mathbb{R}$



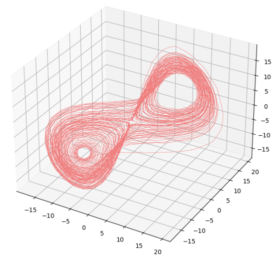
Delay embedding

- Signal $X : [t_0, t_1] \to \mathbb{R}$



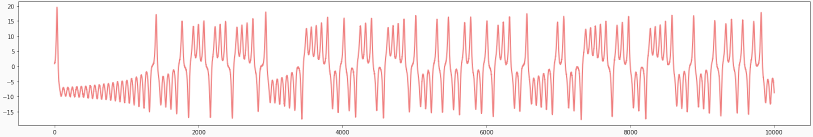
- Delay embedding

$$\mathcal{M} = \{ \big(X(t), X(t+T), X(t+2T) \dots, X(t+(D-1)T) \big) : t \in [t_0, t_1 - (D-1)T] \}$$



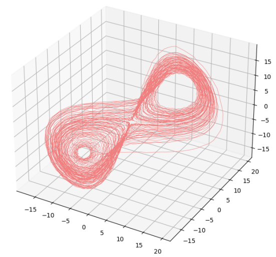
Delay embedding

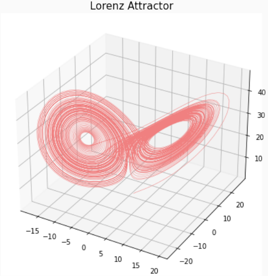
- Signal $X : [t_0, t_1] \to \mathbb{R}$



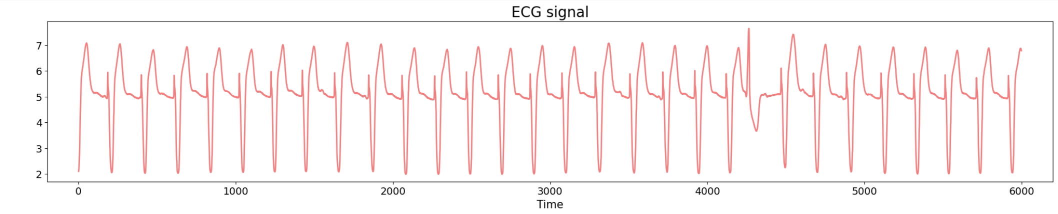
- Delay embedding

$$\mathcal{M} = \{(X(t), X(t+T), X(t+2T) \ldots, X(t+(D-1)T)) : t \in [t_0, t_1 - (D-1)T]\}$$

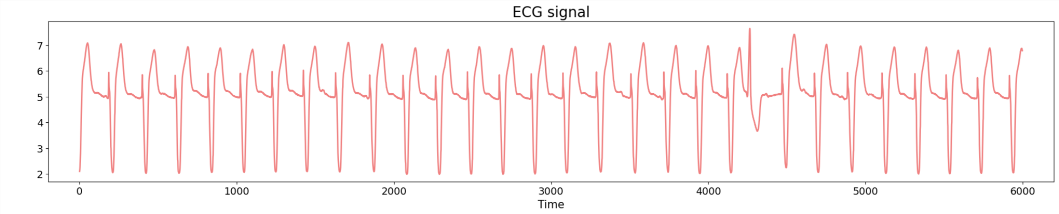


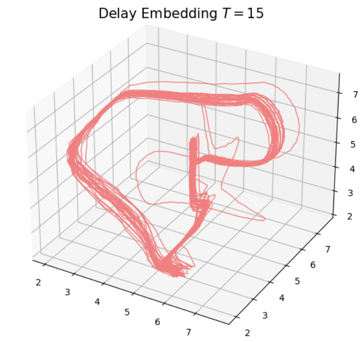


Time series: Anomaly detection

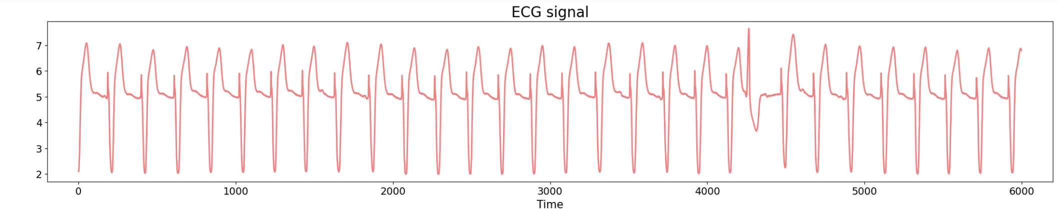


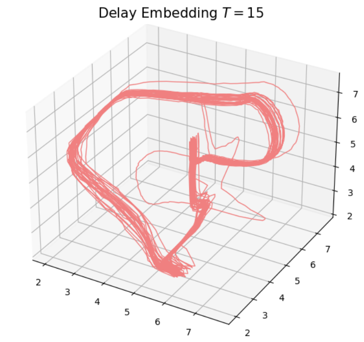
Time series: Anomaly detection



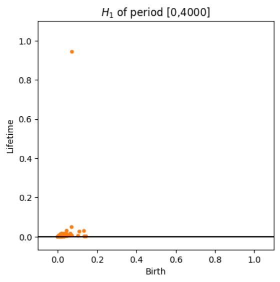


Time series: Anomaly detection

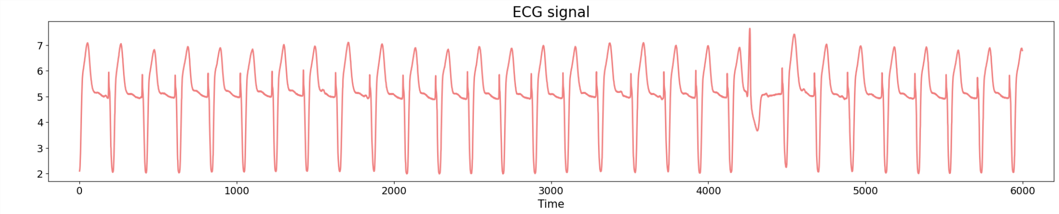


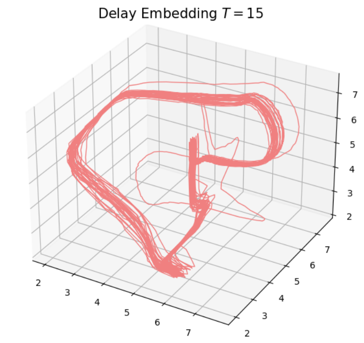


Persistence diagrams with Fermat distance for $p = 2$.

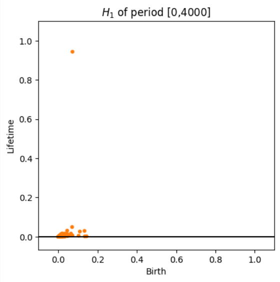


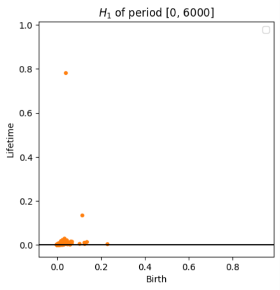
Time series: Anomaly detection



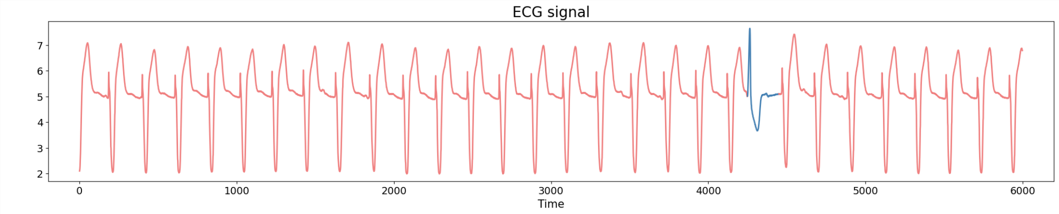


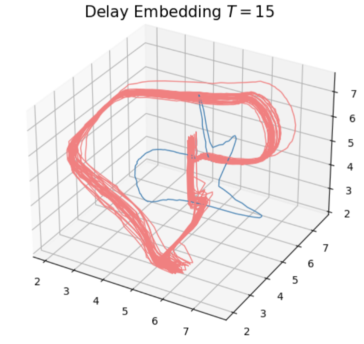
Persistence diagrams with Fermat distance for $p = 2$.



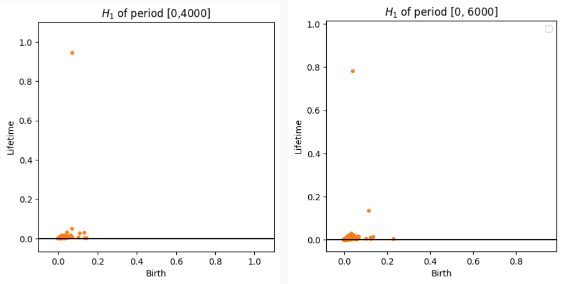


Time series: Anomaly detection

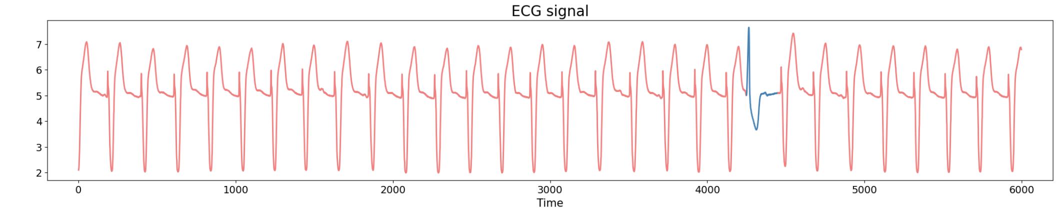


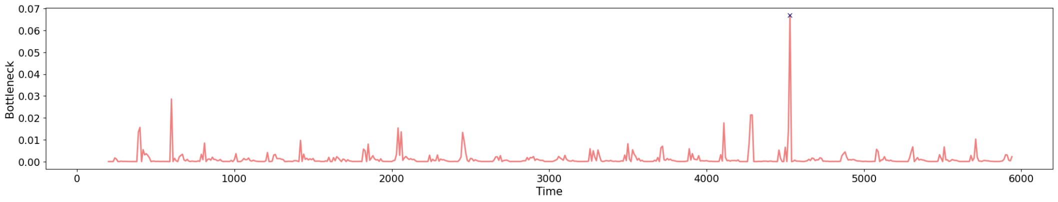


Persistence diagrams with Fermat distance for $p = 2$.

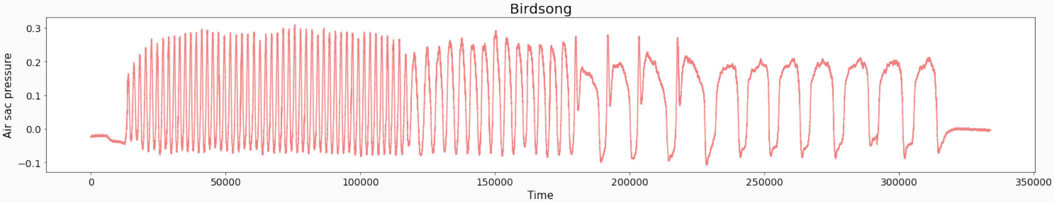


Time series: Anomaly detection

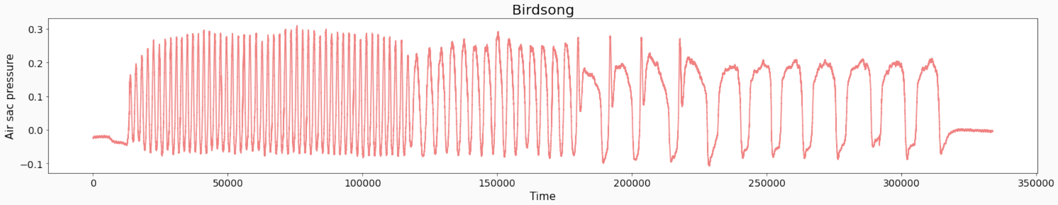


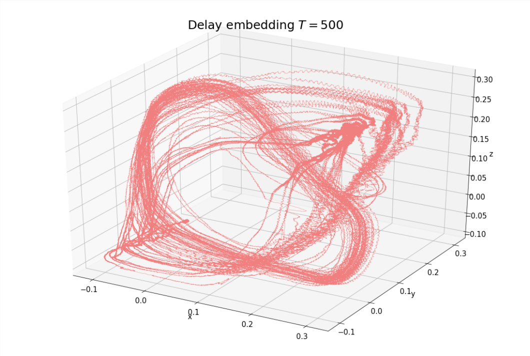


Time series: Pattern recognition

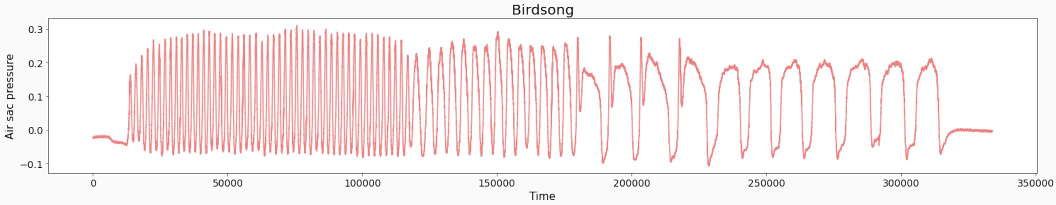


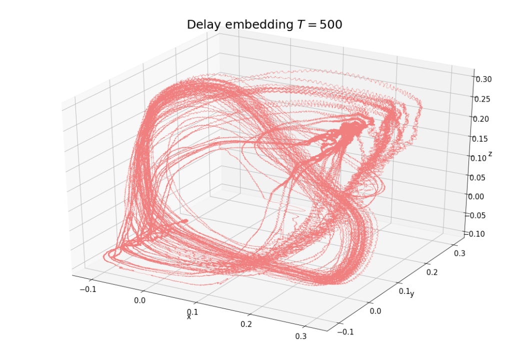
Time series: Pattern recognition

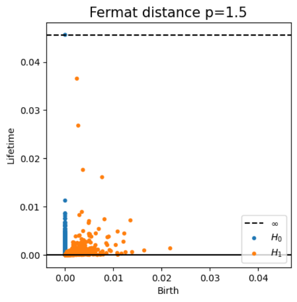




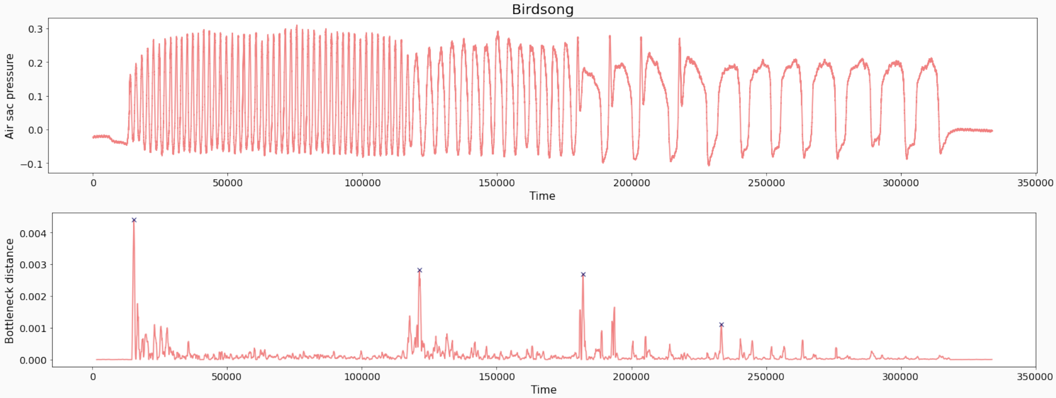
Time series: Pattern recognition



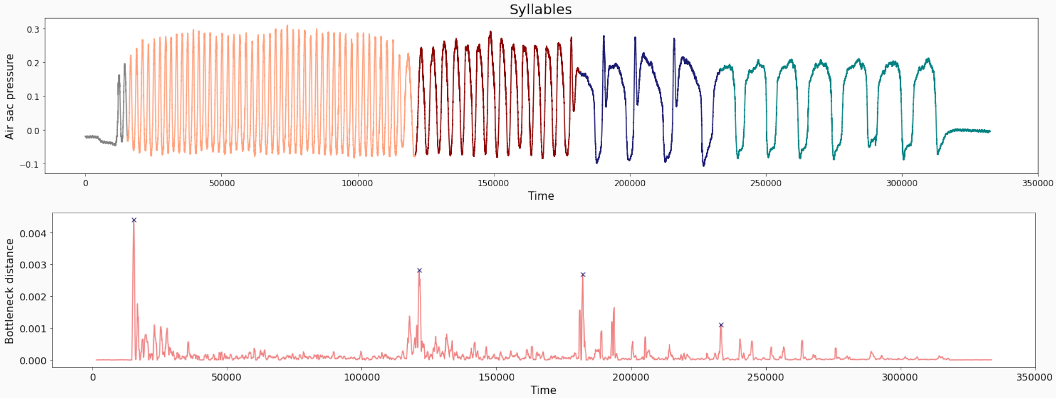




Time series: Pattern recognition

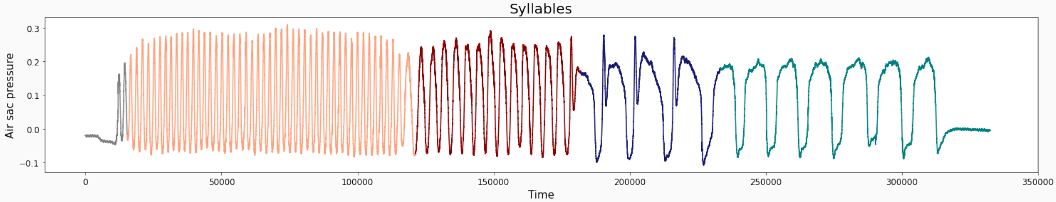


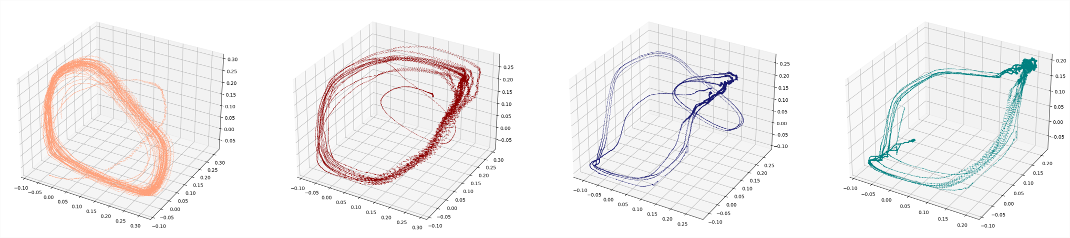
Time series: Pattern recognition



Time series: Pattern recognition

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





References

- *Preprint:* E. Borghini, X. F., P. Groisman, G. Mindlin. *Intrinsic persistent homology via density-based metric learning.* arXiv:2012.07621 (2020) [Updated version soon]
- *Code: https://github.com/ximenafernandez/intrinsicPH*
- *Python library:* `fermat`.

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

Thanks for your attention!