Stable Vectorization of Multiparameter Persistent Homology using Signed Barcodes as Measures

David Loiseaux, Luis Scoccola,
Mathieu Carrière, Steve Oudot, Magnus Botnan

Python package

Multipers : https://github.com/DavidLapous/multipers

References

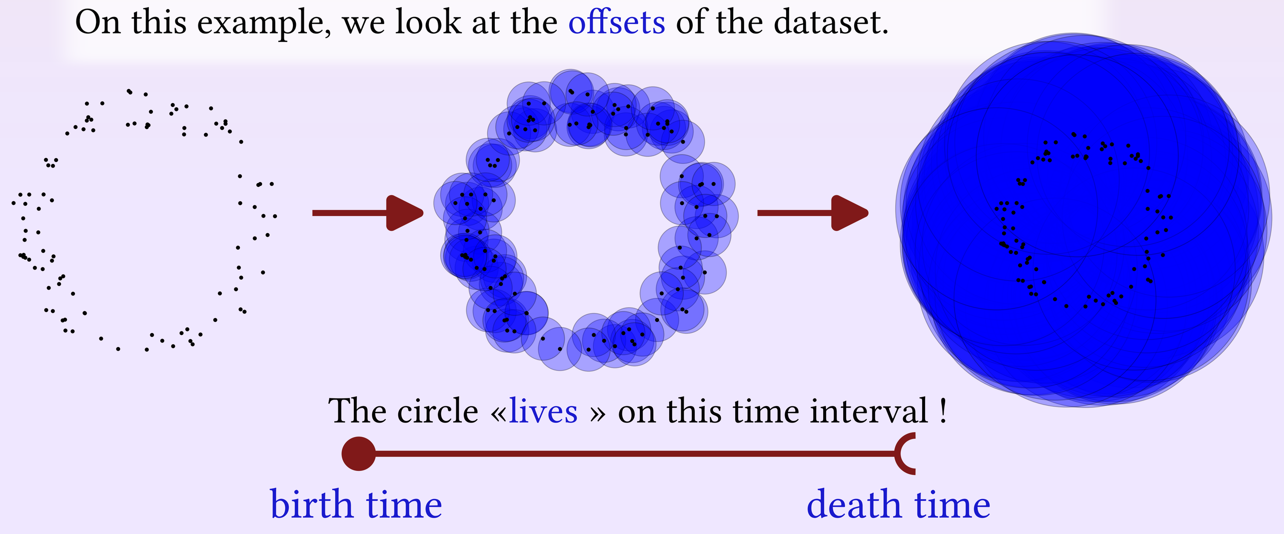
[1] Chazal F. de Silva V. Glisse M. Oudot S. The Structure and Stability of Persistence Modules
[2] M. B. Botnan and M. Lesnick. An introduction to multiparameter persistence
[3] S. Oudot and L. Scoccola. On the stability of multigraded betti numbers and hilbert functions
[4] Carrière M. Blumberg A. Multiparameter Persistence Images for Topological Machine Learning
[5] M. B. Botnan, S. Oppermann, and S. Oudot. Signed barcodes for multi-parameter persistence via rank decompositions

Motivation

The huge variety of dataset in the wild has brought many difficulties from a statistical point of view, such as the so-called « curse of the dimensionality ». Fortunately, data sets usually lie close to some hidden structure; which, if taken into account in the learning pipeline, can help mitigate this effect.
Topological Data Analysis (TDA) is a strategy that aims for a solution to this challenge, by providing compact descriptors inferring the topological features of this hidden structure, such as connectivity, loops, cavities; with nice guarantees.
However, these main descriptors, the persistent modules, still suffers from some technical limitations; for instance, in computational biology, there are, in some cases no canonical way to compute them, as the input contains too much information.
This motivates their generalization : Multiparameter Persistence Modules.
The price to pay for this generalization is their computational cost. In this work we propose a strategy to encode these structure, as point signed measures.
By leveraging on the structure similarities with persistent modules we end up with an easy to compute, and statistically robust topological encoding of datasets.

Idea

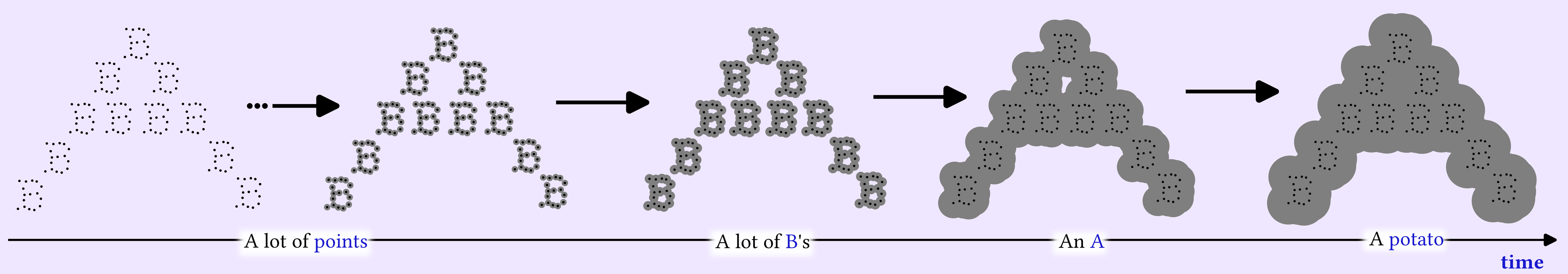
A standard goal is to recover « topological features » from a dataset.
E.g., from points sampled on a circle, we want to retrieve this circle.
On this example, we look at the offsets of the dataset.



This construction can be generalized to « filter functions »; we look at the topology of the sublevelsets $(\{x \in \mathcal{X} \mid f(x) \leq t\})_{t\in\mathbb{R}}$ for a function $f : \mathcal{X} \to \mathbb{R}$

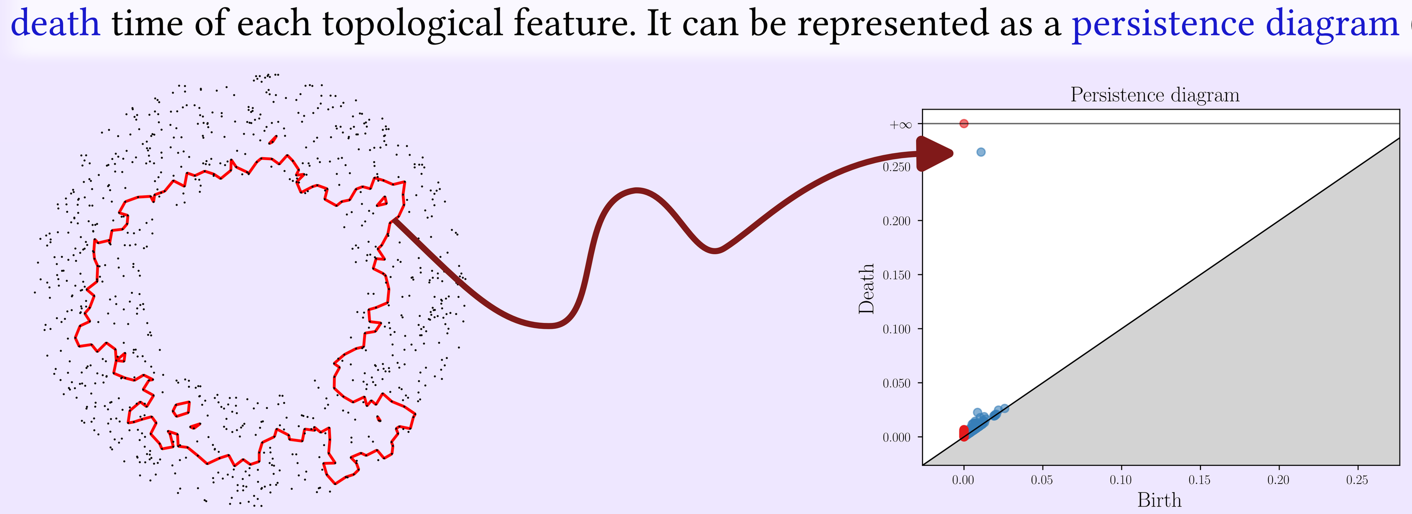
In that context, « growing balls » ≈ take the sublevelsets of the distance function to the dataset.

All scales at once !



Byproducts of TDA

The mathematical structure behind is the persistent homology, which encodes birth and death time of each topological feature. It can be represented as a persistence diagram (Dgm).



Nice properties

Universality

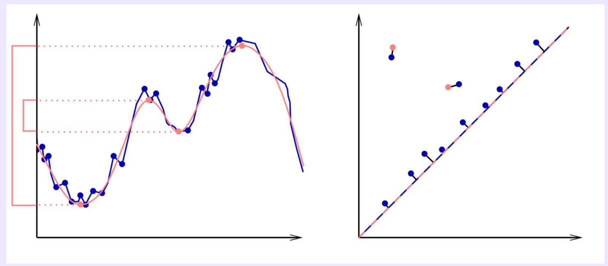
Any dataset having topological or geometrical signal can be used in this pipeline, and the output has always the same diagram structure.

Convergence

If $X^{(n)} = (X_1, \ldots, X_n) \sim \mu^{\otimes n}$ is a nice sampling of a space X_μ,
Then $\mathrm{Dgm}\left(X^{(n)}\right) \xrightarrow[n\to\infty]{d_b} \mathrm{Dgm}\left(X_\mu\right)$

Stability

For two functions $f, g : \mathcal{X} \to \mathbb{R}$,
$d_b\left(\mathrm{Dgm}(f), \mathrm{Dgm}(g)\right) \leq \|f - g\|_\infty$.

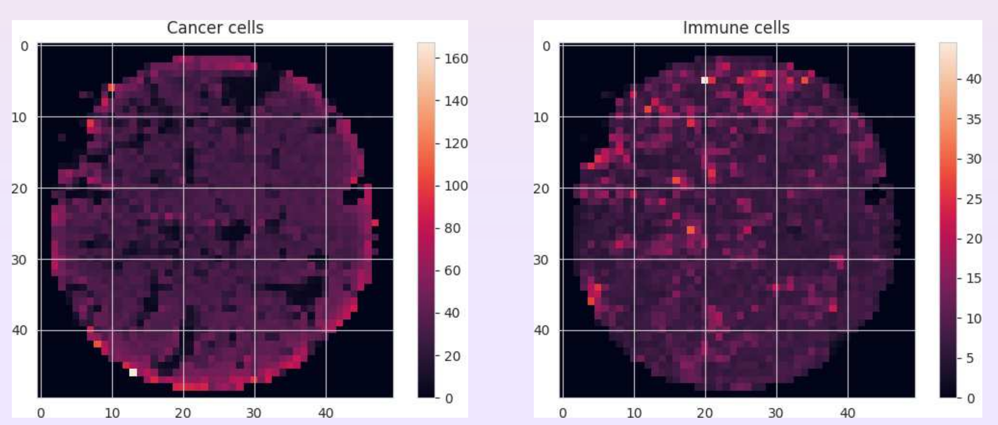


Multiparameter Persistence

Some datasets contain more than geometric information, or have an interesting sampling measure, which will not be taken into account in with PH. This motivates the construction of Multiparameter Persistent Homology (MPH) which looks at the topological persistence of a multi-filtered function $f : \mathcal{X} \to \mathbb{R}^n$.

A Multiparameter Persistent Module M or an n-parameter persistent module is a family of vector spaces $(M_x)_{x\in\mathbb{R}^n}$ with some linear maps $M(x \leq y) : M_x \to M_y$ for $x \leq y \in \mathbb{R}^n$, satisfying

$$\forall x \leq y \leq z \in \mathbb{R}^n, \quad M(y \leq z) \circ M(x \leq y) = M(x \leq z) \quad \text{and} \quad M(x \leq x) = \mathrm{id}$$



From Signed Barcodes to Signed Measures

Hilbert Decomposition Signed Measure
The dimension of any *finitely presented* n-parameter persistence module M is characterized by a *unique* signed point measure $\mu_M \in \mathcal{M}(\mathbb{R}^n)$ such that

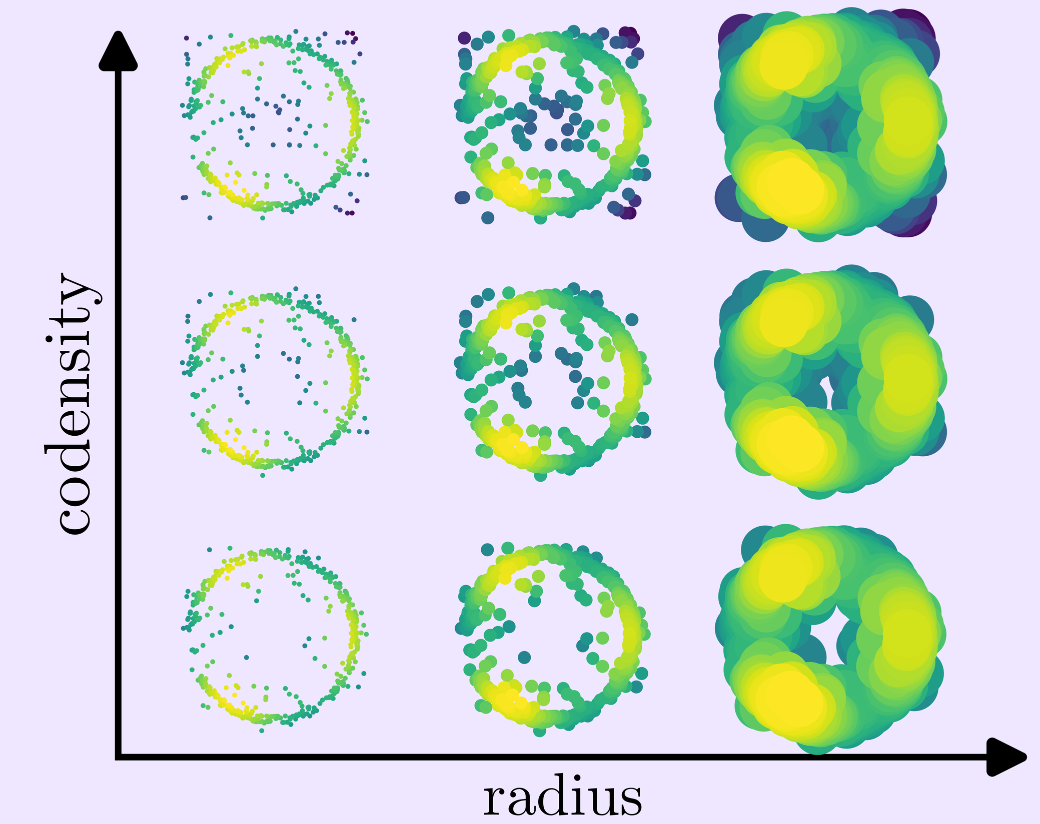
$$\forall x \in \mathbb{R}^n, \quad \dim(M_x) = \mu_M(\{y \in \mathbb{R}^n : y \leq x\})$$

Euler Decomposition Signed Measure
The Euler Characteristic of any n-parameter *finite* simplicial complex filtration F is characterized by a *unique* signed point measure $\mu_F \in \mathcal{M}(\mathbb{R}^n)$ such that

$$\forall x \in \mathbb{R}^n, \quad \chi(F_x) = \mu_F(\{y \in \mathbb{R}^n : y \leq x\})$$

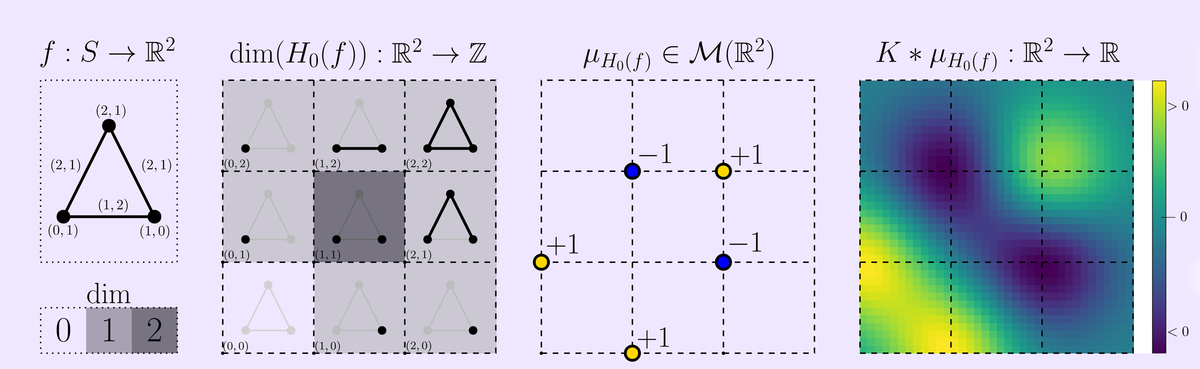
Stability theorem
Let $n \in \mathbb{N}$, let S be a finite simplicial complex, and let $f, g : S \to \mathbb{R}^n$ be monotonic.

$$\forall n \in \{1,2\},\ i \in \mathbb{N}, \qquad \|\mu_{H_i(f)} - \mu_{H_i(g)}\|_{W^1} \leq n \cdot \|f - g\|_1,$$
$$\forall n \in \mathbb{N}, \qquad \|\mu_{\chi(f)} - \mu_{\chi(g)}\|_{W^1} \leq \|f - g\|_1.$$



Notations.
$\mathcal{M}(\mathbb{R}^n)$: The space of finite signed point measures on $\mathbb{R}^n$.
$\mathcal{M}_0(\mathbb{R}^n) \subseteq \mathcal{M}(\mathbb{R}^n)$: The space of measures of mass 0.

Signed Measures Representations

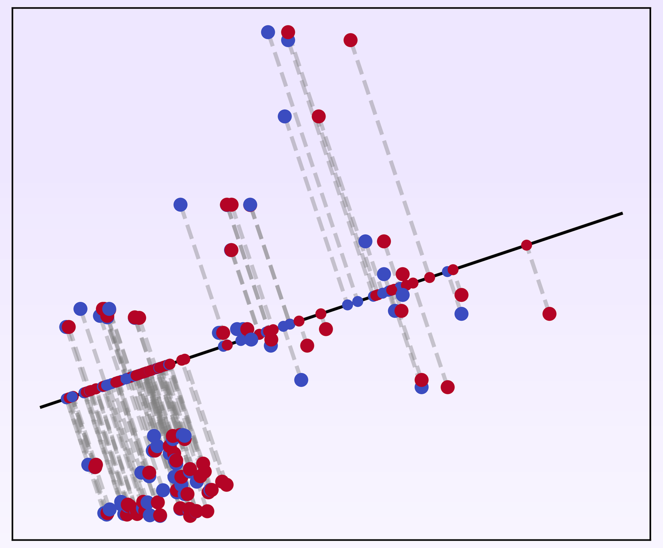


Convolution stability
If a a kernel function $K \in L^2(\mathbb{R}^n \to \mathbb{R}_{\geq 0})$, satisfies, for some positive $c > 0$

$$\forall x, y \in \mathbb{R}^n, \quad \|K(\cdot - x) - K(\cdot - y)\|_2 \leq c \cdot \|x - y\|_2$$

then, if $\mu, \nu \in \mathcal{M}(\mathbb{R}^n)$ have the same total mass,

$$\|K * \mu - K * \nu\|_2 \leq c \cdot \|\mu - \nu\|_{W^2}$$



Sliced Wasserstein distance
For any measure α on S^{n-1}, and $\mu, \nu \in \mathcal{M}_0(\mathbb{R}^n)$ point measures of total mass 0, define

$$\mathrm{SW}^\alpha(\mu,\nu) := \int \left\|\pi^\theta_* \mu - \pi^\theta_* \nu\right\|_1^K \mathrm{d}\alpha(\theta), \text{ and, } k^\alpha_{\mathrm{SW}} = \exp(-\mathrm{SW}^\alpha(\mu,\nu)).$$

where $\pi^\theta : \mathbb{R}^n \to \mathbb{R}$ is the orthogonal projection on the line of slope θ .

Sliced Wasserstein Kernel
For any $n \in \mathbb{N}$, there *exists* a Hilbert space $\mathcal{H}$ and a map $\Phi^\alpha_{\mathrm{SW}} : \mathcal{M}_0(\mathbb{R}^n) \to \mathcal{H}$,
such that for any $\mu, \nu \in \mathcal{M}_0$,

$$\|\Phi^\alpha_{\mathrm{SW}}(\mu) - {\Phi_{\mathrm{SW}}}^\alpha(\nu)\|_{\mathcal{H}} \leq 2\alpha(S^{n-1})\,\|\mu - \nu\|_2^K \quad \text{and} \quad k^\alpha_{\mathrm{SW}}(\mu,\nu) = \langle \Phi^\alpha_{\mathrm{SW}}(\mu), \Phi^\alpha_{\mathrm{SW}}(\nu)\rangle_{\mathcal{H}}$$

Poster link

Code link