Signatures for Geometric Shapes

Mathieu Carrière

May, 2016

1 Background

Geometric Data Analysis. Acquisition and generation of geometric data are nowadays growing very rapidly, leading to very large available geometric data sets. The classification and analysis of such data sets has thus become a problem of primary interest for practitioners. To succeed in this task, a relevant notion of proximity between geometric shapes is required, e.g. that is invariant with respect to the sampling or to rigid motions of the shapes that the data sets represent. Several scientific fields naturally consider these questions, like structural biology (analysis of protein conformations), computer graphics (registration of 3D point clouds acquired by scanners) or machine learning (classification of point clouds in arbitrary dimension).

Gromov-Hausdorff distance. Geometric data sets usually come in the form of finite metric spaces, i.e. point clouds with a given notion of metric (Euclidean norm, geodesics, diffusion distances...) between the points. Intrinsic properties of data sets depend strongly on their corresponding metrics. The canonical distance between finite metric spaces is the *Gromov-Hausdorff distance* d_{GH} [5]:

Definition 1.1. *Let A, B be compact subsets of a common metric space (X, d). The* Hausdorff distance *between A and B is:*

$$d_{\mathrm{H}}(A,B) = \max\left(\sup_{a\in A}\inf_{b\in B} d(a,b),\ \sup_{b\in B}\inf_{a\in A} d(b,a)\right).$$

Definition 1.2. *Let $(X, d_X), (Y, d_Y)$ be two compact metric spaces. The* Gromov-Hausdorff distance *between (X, d_X) and (Y, d_Y) is:*

$$d_{\mathrm{GH}}((X,d_X),(Y,d_Y)) = \inf_{Z,\gamma_X,\gamma_Y} d_{\mathrm{H}}\left(\gamma_X(X),\gamma_Y(Y)\right),$$

when γ_X, γ_Y range over all the isometric embeddings of X, Y into some same metric space (Z, d_Z).

One can see from Definition 1.2 that d_{GH} is invariant to isometries. However, even though it has been intensively studied in the recent years, its computation remains costly [1]. Much of the recent work in geometric data analysis tries to approximate d_{GH}, or at least to define another distance d between finite metric spaces that is *stable* with respect to d_{GH}, i.e. $d(X,Y) \leq C d_{\mathrm{GH}}(X,Y)$, for some constant $C > 0$, for any two finite metric spaces X and Y.

Persistence Diagrams. *Topological Data Analysis* (TDA) [6] is an area of data analysis that uses topological quantities to compare and distinguish data sets. Its main theoretical foundation is *persistence theory* [13], whose objects of study are the so-called *persistence diagrams* (PD).

Homology theory lies at the core of the PD definition. The topology of a space X is represented with a group structure, the so-called *homology groups* [25]. Basically, for any dimension k, the corresponding kth homology group of X, denoted $H_k(X)$, is the group whose basis is given by the k dimensional holes in X (connected components when $k = 0$, loops when $k = 1$, cavities when $k = 2$...). PDs can then be computed from a *filtration* of a topological space X, that is, a sequence of nested subspaces of X:

$$\emptyset = X_1 \subseteq X_2 \subseteq X_3 \subseteq \ \ldots\ \subseteq X_n = X.$$

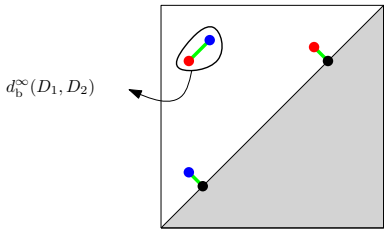


Figure 1: The bottleneck distance between the PDs with red and blue points is given by the length of the longest green segment. Black points are projections of points onto the diagonal.

A common example is the filtration given by the sublevel sets of a function $f : X \to \mathbb{R}$:

$$X_1 = f^{-1}((-\infty, \alpha_1]) \subseteq X_2 = f^{-1}((-\infty, \alpha_2]) \subseteq X_3 = f^{-1}((-\infty, \alpha_3]) \subseteq \ ... \ \subseteq X_n = f^{-1}((-\infty, \alpha_n]),$$

where $\alpha_1 \leq \alpha_2 \leq \ ... \ \leq \alpha_n$. Two other common examples are the *Čech* and *(Vietoris-)Rips filtrations*, where the time of appearance of subspaces of X in the filtrations is induced by the metric diameter of these subspaces. A filtration leads to a sequence of groups, or vector spaces, connected by linear maps, i.e. a *persistence module* [13]: $H_*(X_1) \to H_*(X_2) \to H_*(X_3) \to \ ... \ \to H_*(X_n)$. PDs simply look at generators that persist in this sequence. If a generator appears ("is born") at time $b \subseteq \{1, ..., n\}$ and disappears ("dies") at time $q \subseteq \{1, ..., n\}$, the corresponding PD includes a point with coordinates (b, d). Points in a PD always lie above the diagonal since $d \geq b$. Definitions can be extended to filtrations indexed over the real line [13].

PDs are interesting signatures for point clouds since they encode the topological structure of a point cloud at all possible scales. There is a natural metric between them, the so-called *bottleneck distance* d_{b}^{∞}:

Definition 1.3. *Let $A, B \subseteq \mathbb{R}^2$. Let $A' = A \cup \Pi(B)$ and $B' = B \cup \Pi(A)$, where Π is the projection onto the diagonal. The* bottleneck distance *between A and B is:*

$$d_{\mathrm{b}}^{\infty}(A, B) = \inf_{\gamma : A' \to B'} \max_{p \in A'} \ \mathrm{cost}(p, \gamma(p)),$$

where γ ranges over all the bijections from A' to B', and where the cost *of a pair $(p, \gamma(p))$ is set to be $\|p - \gamma(p)\|_\infty$ if p or $\gamma(p)$ is in $A \cup B$ and 0 otherwise.*

See Figure 1 for an illustration. d_{b}^{∞} can be seen as the l_∞-cost of an optimal transportation plan between the PDs, in which points are allowed to be snapped onto the diagonal. A key property justifying the use of PDs is the stability of d_{b}^{∞} w.r.t. d_{GH} [14].

Theorem 1.4. *Let X, Y be totally bounded metric spaces, and let D_X, D_Y be the PDs of their Rips filtrations. Then:*

$$d_{\mathrm{b}}^{\infty}(D_X, D_Y) \leq 2 d_{\mathrm{GH}}(X, Y)$$

This theorem is a consequence of the general stability theorem for PDs [18]:

Theorem 1.5. *Let X be a topological space, let $f, g : X \to \mathbb{R}$ be tame functions and let $D(f), D(g)$ be the PDs of their sublevel sets filtrations. Then:*

$$d_{\mathrm{b}}^{\infty}(D(f), D(g)) \leq \|f - g\|_\infty$$

Contributions. In this PhD thesis, we study the use of PDs in machine learning and TDA.

In Section 2, we present the work [11], where we discuss the use of PDs in machine learning and present applications in 3D shape analysis. In particular, we defined a new local-to-global PD-based signature for metric spaces, that is inspired from the global PD signature defined in [12], and we addressed an issue that

has been prohibiting the use of PDs in machine learning, namely: the non conditionally negative semi-definiteness of d_b^∞, which was making the derivation of positive semi-definite kernels with d_b^∞ impossible. We defined a new type of kernel between PDs that is provably stable w.r.t. d_b^∞ and we used this kernel in two learning tasks in shape analysis, namely: 3D shape segmentation and 3D shape matching.

In Section 3, we present the work [9]. We combined PDs with the so-called *Reeb graphs* [27] as well as a well-known algorithm of TDA that approximates them, the so-called *Mapper* algorithm [30], which allows to visualize the structure of a scalar field (i.e. a point cloud and a scalar function defined on its points) in the form of a graph. Again, using PDs as signatures for the Mapper graphs, we were able to provide a theoretical framework for the analysis of the stability and structure of Mapper, which was lacking before.

2 Kernels for PDs

2.1 Kernel Methods

Kernels. The use of kernel methods allows the user to apply classical learning algorithms, like SVM, more generally when the objects he considers live in a space X that is not the traditionnal Euclidean space $\mathbb{R}^d$. The idea is simply to send these objects into a Hilbert space H, where computations are well-defined and easier. If the user does not know H in advance, a useful theorem of Moore and Aronszajn [3] states that for any positive semi-definite function $K : X \times X \to \mathbb{R}$ (a "kernel") there always exists an essentially unique Hilbert space H_K and a corresponding mapping $\Psi_K : X \to H_K$ such that $K(x_1, x_2) = \Psi_K(x_1)^T \Psi_K(x_2)$. Moreover, there is an associated metric

$$d_K(x_1, x_2) = \|\Psi_K(x_1) - \Psi_K(x_2)\| = \sqrt{K(x_1, x_1) + K(x_2, x_2) - 2K(x_1, x_2)}$$

Most common kernels are defined when X has a metric. They are:

- the linear one: $K_L(x_1, x_2) = x_1^T x_2$, $d_{K_L}(x_1, x_2) = \|x_1 - x_2\|$, when X is Hilbert,
- the polynomial one: $K_P(x_1, x_2) = (a x_1^T x_2 + c)^d$, when X is Hilbert,
- the Gaussian one: $K_\sigma(x_1, x_2) = \mathrm{e}^{-\frac{d(x_1,x_2)^2}{2\sigma^2}}$, $d_{K_\sigma}(x_1, x_2) = \sqrt{2\left(1 - \mathrm{e}^{-\frac{d(x_1,x_2)^2}{2\sigma^2}}\right)}$,

The case of PDs. A classical result of Berg et al. [4] states that K_σ is a valid kernel, i.e. positive semi-definite, if and only if d is *conditionally negative semi-definite*, i.e.

$$\sum_{i,j \in \{1,\dots,n\}} a_i a_j d(x_i, x_j) \leq 0, \forall x_1, ..., x_n, \forall a_1, ..., a_n \text{ s.t. } \sum_{i=1}^n a_i = 0$$

Unfortunately, the bottleneck distance d_b^∞ for PDs is not conditionally negative semi-definite. Actually, even the other class of distances between PDs, the Wasserstein distances, are not. Thus, the use of d_b^∞ for kernel methods is not permitted.

Our first contribution was to provide an intermediate mapping Φ from PDs to the Euclidean space $\mathbb{R}^d$, so that kernels on $\mathbb{R}^d$ can be directly applied. Indeed, for any kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$,

$$K' : (D_1, D_2) \mapsto K(\Phi(D_1), \Phi(D_2))$$

is a valid kernel. It is worth noting that K' does not use the ambient metric d_b^∞ for PDs, avoiding possibly costly matching computations. In our case, Φ is obtained by computing all pairwise distances in the PD and by truncating their distribution at a fixed threshold d. We proved in [10] that Φ is stable:

Theorem 2.1. *Let D_1, D_2 be two PDs. Then:*

$$\|\Phi(D_1) - \Phi(D_2)\| \leq C d_b^\infty(D_1, D_2),$$

for some $C > 0$ that depends on the norm used.

In particular, $C = 2$ for $\|\cdot\|_\infty$ and $C = \sqrt{\frac{N(N-1)}{2}}$, where $N = \max\{|D_1|, |D_2|\}$, for $\|\cdot\|_2$.
Theorem 2.1 leads to stability corollaries for the common kernel distances:

$$d_{K_L}(D_1, D_2) \leq C d_{\mathrm{b}}^{\infty}(D_1, D_2)$$

$$d^2_{K_\sigma}(D_1, D_2) \leq 2\left(1 - \mathrm{e}^{-\frac{C d_{\mathrm{b}}^{\infty}(D1,D2)^2}{2\sigma^2}}\right)$$

It is to be compared with other recent kernels for PDs [28, 22, 29, 16, 17, 31].

Future Work. We are now carrying out a comprehensive study comparing all recent kernels for PDs, and the possibility to make the map Φ more discriminative by looking at k-fold entries of the distance matrix.

2.2 3D Shape Analysis

Signatures for Shape Processing. Shape analysis and comparison lie at the heart of many problems in computer graphics, including shape retrieval and classification, shape labeling, shape interpolation, and deformation transfer, among many others. In recent years, a large number of approaches have been developed for these tasks, which are often based on devising new *signatures* or *descriptors*. Such descriptors facilitate comparison tasks by encoding the information about the structure of the shapes in a way that is easy to index and analyze.

The nature of such signatures may be very different : they can be global, summarizing the whole shape, or local, characterizing only a subset of the shape, such as a neighborhood of a point; they can capture different types of information (geometry, topology), they can be intrinsic or extrinsic, and also volumetric or just defined on the surface. While there is clearly no ideal signature that would be suitable for all tasks, the three key characteristics that are required for a successful descriptor are: invariance to a relevant deformation class (rigid motions, intrinsic isometries), stability to small perturbations outside of this class, and informativeness, i.e. being able to successfully distinguish points or shapes that are sufficiently different. Although the first characteristic is often easy to ensure, the other two require either extensive experimentation or non-trivial analysis, and may even be in conflict with each other. Therefore, most successful descriptor-based approaches combine multiple signatures, which themselves are often multi-dimensional, with various learning approaches [23, 21]. In this setting, another desired characteristic of a signature is to provide *complementary* information to the one present in other descriptors.

Using PDs. Because of the good properties that PDs enjoy (stability, multi-scale topological information), our second contribution was to use them to build new multiscale signatures for points on the surface of a 3D shape. Namely, given a point p, we consider the filtration given by growing geodesic balls around p. Intuitively, the corresponding PD, which we denote by D_p, includes the radii for which the homology of these balls changes. As we are dealing with connected 3D shapes, the only interesting homology is in dimension 1 so it reduces to counting the holes that appear and disappear in the geodesic balls. See Figure 2. This signature is similar to the one presented in [12], that inspired this work. However, ours is more local since it is anchored to a point.

This is in contrast to the many existing local descriptors that concentrate on the geometry of the shape around a given point and are thus insensitive to the local or global connectivity structure. Moreover, while a number of local or global descriptors have been proposed for shape analysis and comparison based on topological features, ours is the first local-to-global topological descriptor (local since each PD D_p is anchored to a specific point p, and global since the largest geodesic ball is the whole shape itself). In particular, the PDs are defined intrinsically (i.e. with respect to the distances on the surface of the shape), and can also be computed from a broad class of functions, leading to a high modularity.

We proved in [10] the stability of these PDs. Since these PDs are anchored to points, the stability result is stated with an extension of the Gromov-Hausdorff distance between *pointed metric spaces* (X, x, d_X) and (Y, y, d_Y). The only difference with the original Gromov-Hausdorff distance between (X, d_X) and (Y, d_Y) is that we replace the Hausdorff distance $d_{\mathrm{H}}(\gamma_X(X), \gamma_Y(Y))$ by $\max\{d_{\mathrm{H}}(\gamma_X(X), \gamma_Y(Y)), d_Z(\gamma_X(x), \gamma_Y(y))\}$.

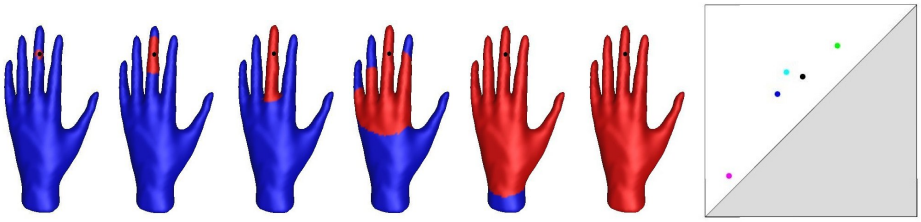
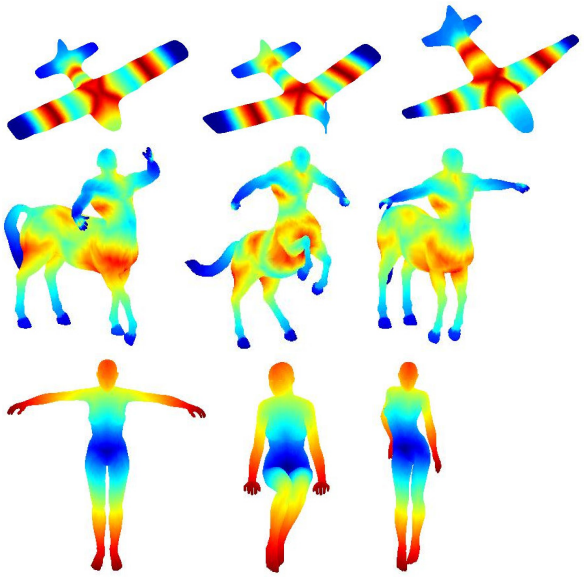


Figure 2: PD computed for a point located at the middle of the middle finger. Geodesic balls are red. There are as many points in the PDs as there are fingers.

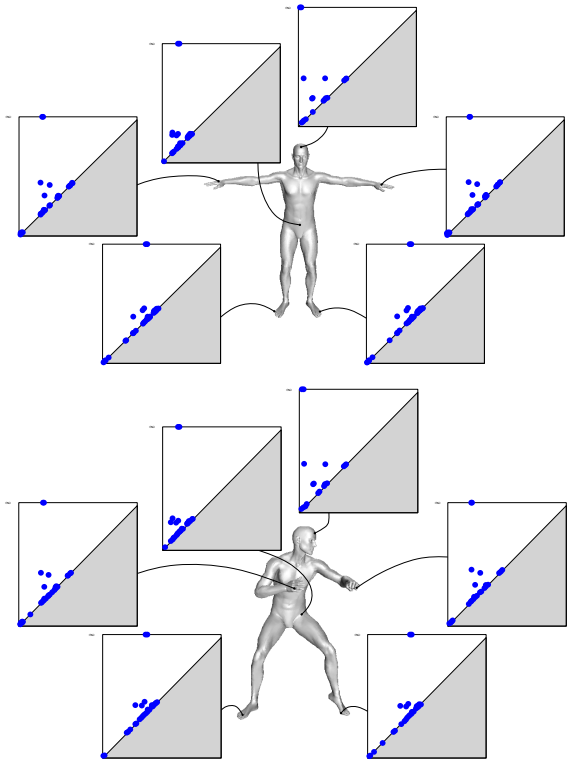
Theorem 2.2. *Let (X, d_X) and (Y, d_Y) be Riemannian manifolds equipped with their geodesic metrics. Let ρ_X (resp. ρ_Y) be the convexity radius of X (resp. Y). Let $x \in X$, $y \in Y$. If $d_{\mathrm{GH}}((X, x, d_X), (Y, y, d_Y)) \leq \frac{1}{10} \min\{\rho_X, \rho_Y\}$, then:*

$$d_{\mathrm{b}}^{\infty}(D_x, D_y) \leq 20 d_{\mathrm{GH}}((X, x, d_X), (Y, y, d_Y)).$$

The difficulty in the proof is that these PDs are computed for points on different spaces, so that classical results on stability for PDs cannot be directly applied. One can visualize stability in Figure 3. Figure 3a shows the values taken by an arbitrary coordinate of $\mathrm{im}(\Phi)$ on isometric shapes while Figure 3b displays PDs of corresponding points in two isometric shapes.



(a) Coordinate of $\mathrm{im}(\Phi)$.



(b) PDs of corresponding points.

Figure 3: Stability visualizations on isometric shapes.

Machine learning for shape processing. Then, we used the map Φ to add the PDs to the existing signatures so that each point of every shape is represented by an Euclidean vector given by the concatenation of all signatures. Using kernel-SVM methods, we performed supervised segmentation and matching tasks on 3D shapes, following the methods from [23] and [21], that show the usefulness of this signature. See Figure 4 for a sample of the results.

Figure 4: Sample of segmentation results. The first row was used for training in each class.

Our work is, to our knowledge, one of the first bridges between topological persistence theory and large-scale machine learning.

3 Reeb graphs, Mapper and PDs

Many data sets nowadays come in the form of point clouds with function values attached to the points. Such data may come either from direct measurements (e.g. think of a sensor field measuring some physical quantity like temperature or humidity), or as a byproduct of some data analysis pipeline (e.g. think of a word function in the quantization phase of the bag-of-words model). There is a need for summarizing such data and for uncovering their inherent structure, to enhance further processing steps and to ease interpretation.

As a simple alternative to the Reeb space, the Mapper has been the object of much interest by practitioners in the data sciences. It has played a key role in several success stories, such as the identification of a new subgroup of breast cancers [26], or the elaboration of a new classification of player positions in the NBA [2], due to its ability to deal with very general functions and datasets. Meanwhile, it has become the flagship component in the software suite developed by Ayasdi, a data analytics company founded in the late 2000's whose interest is to promote the use of topological methods in the data sciences.

3.1 Reeb graph and Mapper

Reeb graph. One way of characterizing the structure of a scalar field $f : X \to \mathbb{R}$ is to look at the evolution of the topology of its *level sets* – i.e. sets of the form $f^{-1}(\{\alpha\})$, for α ranging over $\mathbb{R}$. This information is summarized in a mathematical object called the *Reeb graph* of the pair (X, f), denoted by $\mathrm{R}_f(X)$ and defined as the quotient space obtained by identifying the points of X that lie in the same connected component of the same level set of f [27]. See Figure 5 for an illustration.

The Reeb graph is known to be a graph (technically, a multi-graph) when X is a smooth manifold and f is a Morse function, or more generally when f is of *Morse type* [7]. Moreover, since the map f is constant over equivalence classes, there is a well-defined quotient map $\tilde{f}$ on $\mathrm{R}_f(X)$. The connection between the topology of the Reeb graph and the one of its originating pair has been the object of much study in the past and is now well understood. It has gained increasing interest in the recent years since it has been shown that the *extended PD* [19] of the quotient map describes the structure of the Reeb graph.

Extended PDs. Given a topological space X and a tame function $h : X \to \mathbb{R}$, the *extended PD* $D(h)$ is defined as the PD computed over an extended filtration, where, in addition to the computation of the

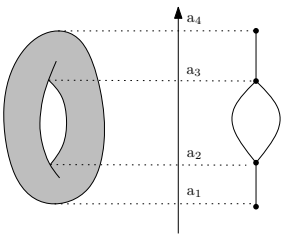


Figure 5: Reeb graph computed on the torus with the height function.

homology groups of the sublevel sets $X_i = h^{-1}((-\infty, \alpha_i])$, we also compute the *relative homology groups* [25] of the total space modulo the superlevel sets $X^i = h^{-1}([\alpha_i, +\infty))$:

$$H_*(X_1) \to H_*(X_2) \to H_*(X_3) \to \ ... \ \to H_*(X_n) = H_*(X)$$
$$\to H_*(X) = H_*(X, X^n) \to H_*(X, X^{n-1}) \to \ ... \ \to H_*(X, X^2) \to H_*(X, X^1)$$

Points in extended PDs are classified into three types: the *ordinary* ones that were born and died in the sublevel sets, the *relative* ones that were born and died in the superlevel sets, and the *extended* ones that were born in the sublevel sets and died in the superlevel sets. The set of ordinary points in dimension k in $D(h)$ is denoted by $\mathrm{Ord}_k(h)$, the set of extended points in dimension k in $D(h)$ is denoted by $\mathrm{Ext}_k(h)$ and the set of relative points in dimension k in $D(h)$ is denoted by $\mathrm{Rel}_k(h)$. We also distinguish between the sets of extended points lying under and above the diagonal, denoted be $\mathrm{Ext}_k^-(h)$ and $\mathrm{Ext}_k^+(h)$ respectively. Note that points in an extended PD do not have to lie above the diagonal anymore.

Given a pair (X, f), its corresponding Reeb graph $\mathrm{R}_f(X)$ and its quotient map $\tilde{f} : \mathrm{R}_f(X) \to \mathbb{R}$, each point of the extended PD of $\tilde{f}$ is matched with a *feature* (branch or hole) of the Reeb graph in a one-to-one manner. Furthermore, the coordinates of the point characterize the *span* of the feature, that is, the interval of $\mathbb{R}$ spanned by its image through $\tilde{f}$. The vertical distance of the point to the diagonal $\Delta = \{(x, x) \mid x \in \mathbb{R}\}$ measures the length of that interval and thereby quantifies the prominence of the feature. Thus, the extended PD plays the role of a "bag-of-features" type signature, summarizing the Reeb graph through its list of features together with their spans, and forgetting about the actual layout of those features.

Mapper. The *Mapper* was introduced by Singh, Mémoli and Carlsson [30] as a new mathematical object to summarize the topological structure of a general pair $(X, f : X \to \mathbb{R}^d)$. Its construction depends on the choice of a cover $\mathcal{I}$ of the image of f by open sets. Pulling back $\mathcal{I}$ through f gives an open cover of the domain X. This cover may have some elements that are disconnected, so it is refined into a connected cover by splitting each element into its various connected components. Then, the Mapper is defined as the nerve of the connected cover, having one vertex per element, one edge per pair of intersecting elements, and more generally, one k-simplex per non-empty $(k+1)$-fold intersection. See Figure 6.

From a philosophical point of view, the Mapper can be thought of as a *pixelized version* of the Reeb space, where the resolution is prescribed by the cover $\mathcal{I}$. From a practical point of view, its construction from point cloud data is very easy to describe and to implement, using standard graph traversals to detect connected components.

3.2 PDs as signatures for the Mapper.

In order to use PDs for the Mapper, we drew an explicit connection between the Mapper and the original space, that happens through the Reeb graph. We then derived guarantees on the structure of the Mapper and quantities to measure its stability. Specifically, given a pair (X, f), its corresponding Reeb graph $\mathrm{R}_f(X)$, its quotient map $\tilde{f} : \mathrm{R}_f(X) \to \mathbb{R}$ and its Mapper $\mathrm{M}_f(X, \mathcal{I})$:

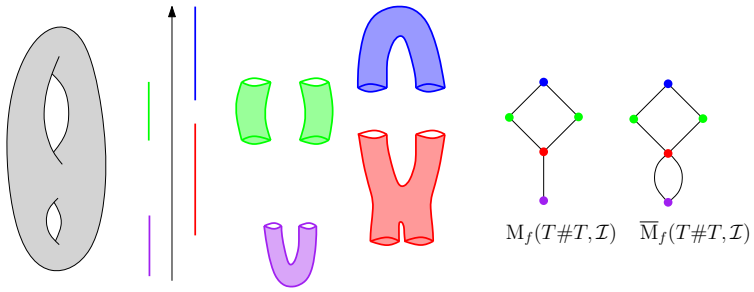


Figure 6: MultiNerve Mapper (rightmost) and Mapper computed on the double torus $T\#T$ with the height function f and a cover $\mathcal{I}$ with four intervals.

- We established the relationship between the extended PD of f and the one of $\tilde{f}$:

  Theorem 3.1. *One has the following equalities:*

$$\mathrm{Ord}_0(\tilde{f}) = \mathrm{Ord}_0(f) \text{ and } \mathrm{Ord}_k(\tilde{f}) = \emptyset, k \geq 1$$
$$\mathrm{Rel}_0(\tilde{f}) = \emptyset, \mathrm{Rel}_1(\tilde{f}) = \mathrm{Rel}_1(f), \text{ and } \mathrm{Rel}_k(\tilde{f}) = \emptyset, k \geq 2$$
$$\mathrm{Ext}_0(\tilde{f}) = \mathrm{Ext}_0(f), \mathrm{Ext}_1^+(\tilde{f}) = \emptyset, \mathrm{Ext}_1^-(\tilde{f}) = \mathrm{Ext}_1^-(f) \text{ and } \mathrm{Ext}_k(\tilde{f}) = \emptyset, k \geq 2$$

- To connect the Mapper and the Reeb graph, we defined an intermediate object, called the *MultiNerve Mapper*, denoted by $\overline{\mathrm{M}}_f(X,\mathcal{I})$, which is simply the *multinerve* of the connected pullback cover in the sense of [20]. See Figure 6 for an illustration. $\overline{\mathrm{M}}_f(X,\mathcal{I})$ and $\mathrm{M}_f(X,\mathcal{I})$ are related through the usual Nerve-vs-MultiNerve connection.

- We showed that this MultiNerve Mapper itself is a Reeb graph, for a perturbed pair (X', f'). Furthermore, we were able to track the changes that occur in the structure of the Reeb graph as we go from the initial pair (X, f) to its perturbed version (X', f'). More precisely, we can match the extended PDs of the quotient maps $\tilde{f}$ and $\tilde{f}'$ with each other, and thus draw a correspondence between the features of the MultiNerve Mapper and the ones of the Reeb graph of (X, f). This correspondence is oblivious to the actual layouts of the features in the two graphs, which in principle could differ.

- The previous connection allowed us to derive signatures for $\overline{\mathrm{M}}_f(X,\mathcal{I})$ and $\mathrm{M}_f(X,\mathcal{I})$, which take the form of extended PDs. The points in these diagrams are in one-to-one correspondence with the features (branches, holes) in the corresponding (MultiNerve) Mapper. Thus, like the extended PD of the quotient map $\tilde{f}$ for the Reeb graph, our diagrams for the (MultiNerve) Mapper serve as bag-of-features type signatures.

- An interesting property of our signatures is to be predictable given the extended PD of the quotient map $\tilde{f}$. Indeed, it is obtained from this diagram by removing the points lying in certain *staircases* that are defined solely from the cover $\mathcal{I}$ and that encode the mutual positioning of the intervals of the cover. Given an interval $I = (a, b)$ (indifferently open, closed or half-open), let

$$Q_I^+ = \{(x, y) \in \mathbb{R}^2 \mid a \leq x \leq y \leq b\}$$

  be the half-square above the diagonal, and

$$Q_I^- = \{(x, y) \in \mathbb{R}^2 \mid a \leq y < x \leq b\}$$

  the half-square strictly below the diagonal.

  Decomposing each interval $I \in \mathcal{I}$ as $I = I_\cap^- \sqcup \tilde{I} \sqcup I_\cap^+ \in \mathcal{I}$, we can define the staircases: $Q_O^{\mathcal{I}} = \bigcup_{I\in\mathcal{I}} Q^+_{\tilde{I}\cup I_\cap^+}$, $Q_R^{\mathcal{I}} = \bigcup_{I\in\mathcal{I}} Q^-_{\tilde{I}\cup I_\cap^-}$, and $Q_{E^-}^{\mathcal{I}} = \bigcup_{I\in\mathcal{I}} Q_I^-$. Our structure theorem follows:

Theorem 3.2. *Let D_O, D_R, D_E^+ and D_E^- denote the ordinary, relative and above- and below-diagonal extended parts of the signature of $\overline{\mathrm{M}}_f(X, \mathcal{I})$. Then:*

$$\text{(i) } D_O = \mathrm{Ord}(\tilde{f}) \setminus Q_O^{\mathcal{I}} \quad \text{(iii) } D_E^- = \mathrm{Ext}^-(\tilde{f}) \setminus Q_{E^-}^{\mathcal{I}}$$
$$\text{(ii) } D_R = \mathrm{Rel}(\tilde{f}) \setminus Q_R^{\mathcal{I}} \quad \text{(iv) } D_E^+ = \mathrm{Ext}^+(\tilde{f}) \cup (\mathrm{Ext}^-(f) \cap Q_{E^-}^{\mathcal{I}})$$

Moreover, the signature of $\mathrm{M}_f(X, \mathcal{I})$ is included in the one of $\overline{\mathrm{M}}_f(X, \mathcal{I})$. It can be obtained from the one of $\overline{\mathrm{M}}_f(X, \mathcal{I})$ by removing the points that lie in a thickened staircase $Q_E^{\mathcal{I}}$. Thus, the signature for the (MultiNerve) Mapper is a subset of the one for the Reeb graph, which provides theoretical evidence to the intuitive claim that the Mapper is a pixelized version of the Reeb graph. Then, one can easily derive sufficient conditions under which the bag-of-features structure of the Reeb graph is preserved in the (MultiNerve) Mapper, and when it is not, one can easily predict which features are preserved and which ones disappear. See Figure 7.

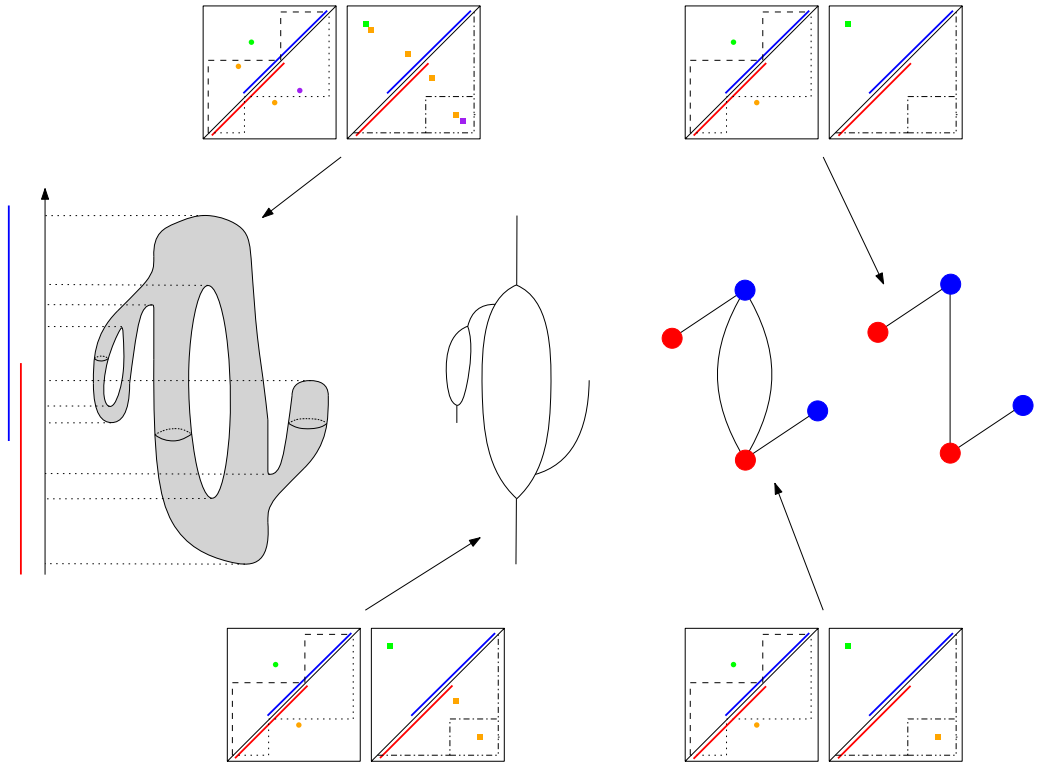


Figure 7: From left to right: a 2-manifold equipped with the height function; the corresponding Reeb graph, MultiNerve Mapper, and Mapper. For each object, we display the extended PDs of dimension 0 (green points), 1 (orange points) and 2 (purple points). Extended points are squares while ordinary and relative points are disks (above and below the diagonal respectively). The staircases are represented with dashed ($Q_O^{\mathcal{I}}$), dotted ($Q_{E^-}^{\mathcal{I}}$), dash-dotted ($Q_R^{\mathcal{I}}$), and dash-dot-dotted ($Q_E^{\mathcal{I}}$) lines. One can see how to go from the extended PD of the height function to the one of the quotient map (remove the points in dimension 2 and the points in dimension 1 above the diagonal), then to the one of the MultiNerve Mapper (remove the points inside the staircases corresponding to their type), and finally, to the one of the Mapper (remove the extended points in $Q_E^{\mathcal{I}}$).

- The staircases also play a role in the stability of the (MultiNerve) Mapper, since they prescribe which features will (dis-)appear as the function f is perturbed. Stability is then naturally measured by a

slightly modified version of d_{b}^{∞}, denoted by $d_{\mathcal{I}}$, in which the staircases play the role of the diagonal. Our stability guarantees follow easily from the general stability theorem for extended persistence [19]:

Theorem 3.3. *Let $f, g : X \to \mathbb{R}$ be Morse-type functions on a topological space X. Then:*

$$d_{\mathcal{I}}(\overline{\mathrm{M}}_f(X,\mathcal{I}), \overline{\mathrm{M}}_g(X,\mathcal{I})) \leq \|f-g\|_{\infty}$$

Similar guarantees hold when the domain X or the cover $\mathcal{I}$ is perturbed.

- These stability guarantees can be exploited in practice to approximate the signatures of the Mapper and MultiNerve Mapper from point cloud data. The approach boils down to applying known scalar field analysis techniques [15] then pruning the obtained extended PDs using the staircases. The approach becomes more involved if one wants to further guarantee that the approximate signature does correspond to some perturbed Mapper or MultiNerve Mapper.

All proofs can be found in the full version of the article [8].

Future Work. Several questions remain open for future work. We are working on the possibility to link our approach to the cosheaf approach developed at the same time by Munch and Wang [24]. We are also deriving statistical guarantees for the convergence of the Mapper to the Reeb graph, like confidence regions and convergence rates. Finally, we will also try to use our signature in machine learning tasks, using e.g. kernels on graphs.

References

[1] P. Agarwal, K. Fox, A. Nath, A. Sidiropoulos, and Y. Wang. Computing the Gromov-Hausdorff Distance for Metric Trees. arXiv 1509.05751, 2015.

[2] M. Alagappan. From 5 to 13: Redefining the Positions in Basketball. MIT Sloan Sports Analytics Conference, 2012.

[3] N. Aronszajn. Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, 68, 1950.

[4] C. Berg, J. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*. Springer, 1984.

[5] D. Burago, Y. Burago, and S. Ivanov. *A Course in Metric Geometry*. American Mathematical Society, 2001.

[6] G. Carlsson. Topology and Data. *Bulletin American Mathematical Society*, 46:255–308, 2009.

[7] G. Carlsson, V. de Silva, and D. Morozov. Zigzag Persistent Homology and Real-valued Functions. In *Proceedings 25th Symposium Computational Geometry*, pages 247–256, 2009.

[8] M. Carrière and S. Oudot. Structure and Stability of the 1-Dimensional Mapper. arXiv 1511.05823, 2015.

[9] M. Carrière and S. Oudot. Structure and Stability of the 1-Dimensional Mapper. To appear in *Proceedings 13th Symposium Geometry Processing*, 2016.

[10] M. Carrière, S. Oudot, and M. Ovsjanikov. Local Signatures using Persistence Diagrams. HAL preprint, 2015.

[11] M. Carrière, S. Oudot, and M. Ovsjanikov. Stable Topological Signatures for Points on 3D Shapes. In *Proceedings 13th Symposium Geometry Processing*, 2015.

[12] F. Chazal, D. Cohen-Steiner, L. Guibas, F. Mémoli, and S. Oudot. Gromov-Hausdorff Stable Signatures for Shapes using Persistence. *Proceedings 7th Symposium Geometry Processing*, pages 1393–1403, 2009.

[13] F. Chazal, V. de Silva, M. Glisse, and S. Oudot. The Structure and Stability of Persistence Modules. arXiv 1207.3674, 2012.

[14] F. Chazal, V. de Silva, and S. Oudot. Persistence Stability for Geometric Complexes. *Geometriae Dedicata*, 173(1):193–214, 2013.

[15] F. Chazal, L. Guibas, S. Oudot, and P. Skraba. Analysis of Scalar Fields over Point Cloud Data. In *Proceedings 20th Symposium Discrete Algorithm*, pages 1021–1030, 2009.

[16] Y.C. Chen, D. Wang, A. Rinaldo, and L. Wasserman. Statistical Analysis of Persistence Intensity Functions. arXiv 1510.02502, 2015.

[17] S. Chepushtanova, T. Emerson, E. Hanson, M. Kirby, F. Motta, R. Neville, C. Peterson, P. Shipman, and L. Ziegelmeier. Persistence Images: An Alternative Persistent Homology Representation. arXiv 1507.06217, 2015.

[18] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of Persistence Diagrams. *Discrete Computational Geometry*, 37(1):103–120, 2007.

[19] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Extending persistence using Poincaré and Lefschetz duality. *Foundation Computational Mathematics*, 9(1):79–103, 2009.

[20] É. Colin de Verdière, G. Ginot, and X. Goaoc. Multinerves and Helly numbers of acyclic families. In *Proceedings 28th Symposium Computational Geometry*, pages 209–218, 2012.

[21] É. Corman, M. Ovsjanikov, and A. Chambolle. Supervised Descriptor Learning for Non-Rigid Shape Matching. In *Proceedings 6th Workshop on Non-Rigid Shape Analysis and Deformable Image Alignment*, 2014.

[22] B. DiFabio and M. Ferri. Comparing Persistence Diagrams through Complex Vectors. arXiv 1505.01335, 2015.

[23] E. Kalogerakis, A. Hertzmann, and K. Singh. Learning 3D Mesh Segmentation and Labeling. *ACM Transactions on Graphics*, 29(4), 2010.

[24] E. Munch and B. Wang. Convergence between Categorical Representations of Reeb Space and Mapper. To appear in *Proceedings 13th Symposium Geometry Processing*, 2016.

[25] J. Munkres. *Elements of Algebraic Topology*. Westview Press, 1993.

[26] M. Nicolau, A. Levine, and G. Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings National Academy Science*, 108(17):7265–7270, 2011.

[27] G. Reeb. Sur les points singuliers d'une forme de pfaff complètement intégrable ou d'une fonction numérique. *Compte Rendu Académie Science Paris*, 222:847–849, 1946.

[28] J. Reininghaus, U. Bauer, S. Huber, and R. Kwitt. A Stable Multi-scale Kernel for Topological Machine Learning. In *Proceedings 27th IEEE Conference Computer Vision and Pattern Recognition*, 2015.

[29] V. Robins and K. Turner. Principal Component Analysis of Persistent Homology Rank Functions with case studies of Spatial Point Patterns, Sphere Packing and Colloids. arXiv 1507.01454, 2015.

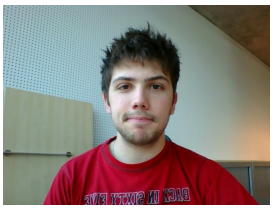
[30] G. Singh, F. Mémoli, and G. Carlsson. Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. In *Symposium Point Based Graphics*, 2007.

[31] S. Verovsek. Tropical Coordinates on the Space of Persistence Barcodes. arXiv 1604.00113, 2016.

Mathieu Carrière

PhD Student in Applied Mathematics and Data Science

65 B4 Boulevard du Maréchal Joffre
92340 Bourg-la-Reine, France
+33(0)682611323
mathieu.carriere@inria.fr
French/American
Birth date: 09/08/1991



Education

2013–2014 **Master in Mathematics, Vision and Learning**, *ENS Cachan*.
Passed with the Highest Honours. **Relevant modules:** Probabilistic Graphical Models, Geometry and Shape Space, Computer Vision and Learning (SVM, Boosting, Random Forests, Neural Networks).

2011–2014 **Engineering Degree**, *Ecole Centrale Paris*.
Relevant modules: Geometric and Topological Modeling, C++ Programming, Statistics and Data Mining, Convex Optimization.

2009–2011 **Intensive Preparation for Engineering Schools**, *Pierre de Fermat High School, Toulouse*.

2007–2009 **High School Diploma**, *Rive Gauche High School, Toulouse*.

Research

2014–present **PhD Thesis**, *DataShape team, INRIA Saclay*.
Title: Signatures for Geometric Shapes. **Supervisor:** Steve Oudot. **Topic:** Stable topological signatures for high dimensional shapes that can be used in Machine Learning classical algorithms.

04–09/2014 **Master Internship**, *DataShape team, INRIA Saclay*.
Title: Topological Signatures for 3D Shapes. **Supervisors:** Steve Oudot, Maks Ovsjanikov. **Topic:** Topological persistence for stable signatures that can be used in 3D shape segmentation.

2013–2014 **3rd Year Engineering Project**, *DataShape team, INRIA Saclay*.
Supervisor: Steve Oudot. **Topic:** Topological clustering in C++ with the ToMATo algorithm.

2012–2013 **2nd Year Engineering Project**, *MAS Laboratory, Ecole Centrale Paris*.
Supervisor: Patrick Callet. **Topic:** 3D simulation of the Royaumont church with Blender.

Articles

2015 **Stable Topological Signatures for Points on 3D Shapes**.
M. C., S. Oudot, M. Ovsjanikov. Proceedings of the 13th Symposium of Geometry Processing, 2015

2015 **Local Signatures using Persistence Diagrams**.
M. C., S. Oudot, M. Ovsjanikov. HAL preprint, 2015.

2016 **Structure and Stability of the 1-Dimensional Mapper (conference version)**.
M. C., S. Oudot. Proceedings of the 32nd Symposium of Computational Geometry, 2016

2016 **Structure and Stability of the 1-Dimensional Mapper (full version)**.
M. C., S. Oudot. arXiv preprint, 2016

Participation to Conferences and Workshops

06/2015 **Symposium of Geometry Processing**, *Graz, Austria*.

11/2015 **Journées de Géométrie Algorithmiques**, *Cargèse, Corsica*.

12/2015 **Computational and Methodological Statistics**, *London, UK*.

04/2016 **Stochastic Geometry and its Applications**, *Nantes, France*.

06/2016 **Symposium of Computational Geometry**, *Boston, USA*.

07/2016 **Applied Topology: Methods, Computation and Science**, *Torino, Italy*.

Teaching

2015–2016 **Teaching Assistant in Topological Data Analysis**, *Ecole Polytechnique*.

2015–2016 **Supervisor of 2nd year project with Unity**, *IUT d'Orsay*.

Skills

Programming	**C++, Python, Matlab, R, Linux, LaTeX, HTML**.	
Website	**http://geometrica.saclay.inria.fr/team/Mathieu.Carriere/**.	
English	**Highly proficient in spoken and written English**	*TOEFL 627/677*
Spanish	**Good command**	*B2 Level*

Interests

Running *I like running. I participated to several competitions and adventure races.*

Climbing *I have been climbing since the age of 12. I often go climbing in the mountains.*

Music *I have been playing piano (classical/jazz) for 20 years. I participated to several concerts.*

References

Dr Steve Oudot
INRIA Saclay
Palaiseau (France)
steve.oudot@inria.fr
+33 (0) 174854216

Dr Maks Ovsjanikov
LIX Laboratory
Palaiseau (France)
maks@lix.polytechnique.fr
+33 (0) 177578011

Dr Nikos Paragios
MAS Laboratory
Chatenay-Malabry (France)
nikos.paragios@ecp.fr