Metric Structures on Datasets: Stability and Classification of Algorithms

Facundo Mémoli

Department of Mathematics, Stanford University, California, USA, and
Department of Computer Science, The University of Adelaide, Australia
memoli@math.stanford.edu

Abstract. Several methods in data and shape analysis can be regarded as transformations between metric spaces. Examples are hierarchical clustering methods, the higher order constructions of computational persistent topology, and several computational techniques that operate within the context of data/shape matching under invariances.

Metric geometry, and in particular different variants of the Gromov-Hausdorff distance provide a point of view which is applicable in different scenarios. The underlying idea is to regard datasets as metric spaces, or metric measure spaces (a.k.a. mm-spaces, which are metric spaces enriched with probability measures), and then, crucially, at the same time regard the collection of all datasets as a metric space in itself. Variations of this point of view give rise to different taxonomies that include several methods for extracting information from datasets.

Imposing metric structures on the collection of all datasets could be regarded as a "soft" construction. The classification of algorithms, or the axiomatic characterization of them, could be achieved by imposing the more "rigid" category structures on the collection of all finite metric spaces and demanding functoriality of the algorithms. In this case, one would hope to single out all the algorithms that satisfy certain natural conditions, which would clarify the landscape of available methods. We describe how using this formalism leads to an axiomatic description of many clustering algorithms, both flat and hierarchical.

Keywords: metric geometry, categories and functors, metric spaces, Gromov-Hausdorff distance, Gromov-Wasserstein distance.

1 Introduction

Nowadays in the scientific community we are being asked to analyze and probe large volumes of data with the hope that we may learn something about the underlying phenomena producing these data. Questions such as "what is the shape of data" are routinely formulated and partial answers to these usually reveal interesting science.

An important goal of exploratory data analysis is to enable researchers to obtain insights about the organization of datasets. Several algorithms have been developed with the goal of discovering structure in data, and examples of the different tasks these algorithms tackle are:

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 1–33, 2011.

- Visualization, parametrization of high dimensional data
- Registration/matching of datasets: how different are two given datasets? what is a good correspondence between sub-parts of the datasets?
- What are the features present in the data? e.g. clustering, and number of holes in the data.
- How to agglomerate/merge (partial) datasets?

Some of the standard concerns about the results produced by algorithms that attempt to solve these tasks are: the dependence on a particular choice of coordinates, the invariance to certain uninteresting deformations, the stability/sensitivity to small perturbations, etc.

1.1 Visualization of Datasets

The *projection pursuit* method (see [42]) determines the linear projection on two or three dimensional space which optimizes a certain criterion. It is frequently very successful, and when it succeeds it produces a set in $\mathbb{R}^2$ or $\mathbb{R}^3$ which readily visualizable. Other methods (Isomap [85], locally linear embedding [74], multidimensional scaling [23]) attempt to find non-linear maps to Euclidean space which preserve the distance functions on the data set to as high a degree as possible. They also produce useful two and three dimensional versions of data sets when they succeed.

Other interesting methods are the *grand tour* of Asimov [2], the *parallel coordinates* of Inselberg [44], and the *principal curves* of Hastie and Stuetzle [38].

The Mapper algorithm [80] produces representations of data in a manner akin to the Reeb graph [71] and is based on the idea of *partial clustering* and can be considered as a hybrid method which combines the ability to parametrize and visualize data, with the the ability to extract features, see Figure 1. This algorithm has been used for shape matching tasks as well for studies of breast cancer [65] and RNA [6]. The mapper algorithm is also closely related to the *cluster tree* of Stuetzle [82].

1.2 Matching and Dissimilarity between Datasets

Measuring the *dissimilarity* between two objects is a task that is often performed in data and shape analysis, and summaries or features from each of the objects are typically compared to quantify this dissimilarity.

One important instance when computing the dissimilarity between is useful is the comparison of the three dimensional shape of proteins following the underlying scientific assumption that physically similar proteins have similar functional properties [52].

The notion of zero-dissimilarity between data-sets can be dependent on the application domain. For example, in object recognition, rigid motions specifically, and more generally isometries, are often uninteresting and not important. The

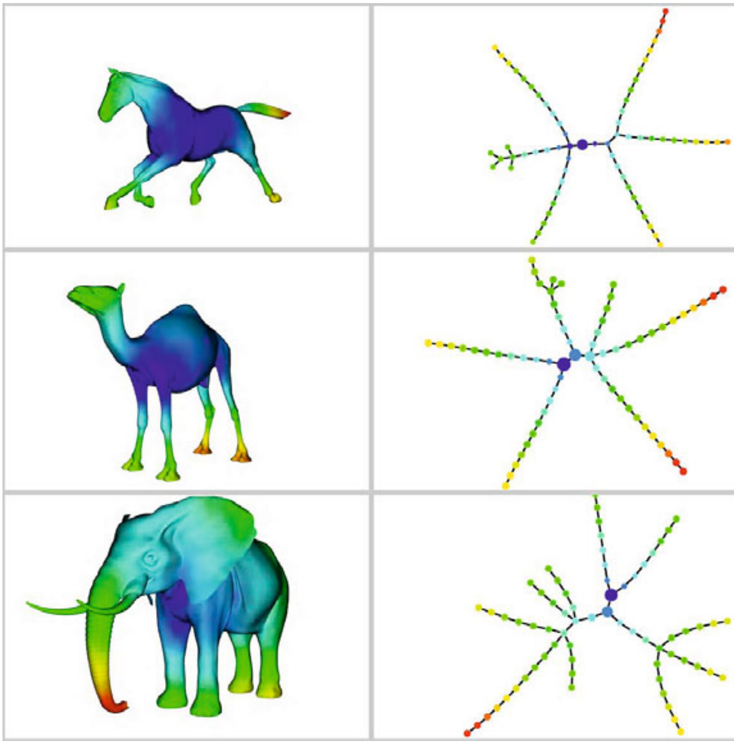


Fig. 1. A simplification of 3d models using the mapper algorithm [80]

same point applies to multidimensional data analysis, where particular choices of the coordinate system should not affect the result of algorithms. Therefore, the summaries/features extracted from the data must be insensitive to these unimportant changes.

There exists a plethora of practical methods for object comparison and matching, and most of them are based on comparing features. Given this rich and disparate collection of available methods, it seems that in order to obtain a deep understanding of the object matching problem and find possible avenues of improvement, it is of great importance to discover and establish relationships/connections between these methods. Theoretical understanding of these methods and their relationships will lead to expressing conditions of validity of each approach or family of approaches. This can no doubt help in

(a) guiding the choice of which method to use in a given practical application,
(b) deciding what parameters (if any) should be used for the particular method chosen, and
(c) clearly determining what are the *theoretical guarantees* of a particular method for the task at hand.

1.3 Features

Often, data-sets can be difficult to comprehend. One example of this is the case of high dimensional point clouds because our ability to visualize them is rather limited. To deal with this situation, one must attempt to extract *summaries* from the complicated data-set in order to capture robust global properties that signal important qualitative features present, but not apparent, in the data.

The term *feature* typically applies to the result of applying a certain simplification to a given dataset with the hope of retaining some useful information about the original data. The aim is that after this simplification it would become easier to quantify and/or visualize certain aspects of the dataset. Think for example of:

- computing the number of clusters in a given dataset, according to a given algorithm (e.g. linkage based methods, spectral clustering, k-means, etc);
- obtaining a dendrogram: the result of applying a hierarchical clustering algorithm to the data;
- computing the average distance to the barycenter of the dataset (assumed to be embedded in Euclidean space);
- computing the average distance between all pairs of points in the dataset;
- computing a histogram of all the interpoint distances between pairs of points in the dataset;
- computing persistent topology invariants of some filtration obtained from the dataset [33,17,81].

In the area of shape analysis a few examples are: the *size theory* of Frosini and collaborators [30,29,88,25,24,31]; the Reeb graph approach of Hilaga et al [39]; the *spin images* of Johnsson [49], the *shape distributions* of [68]; the *canonical forms* of [28]; the Hamza-Krim approach [36]; the spectral approaches of [72,76]; the *integral invariants* of [58,69,21]; the *shape contexts* of [3].

The theoretical question of proving that a given family of features is indeed able to signal proximity or similarity of objects in a reasonable way has hardly been addressed. In particular, the degree to which two objects with similar features are forced to be similar is in general does not seem to be well understood.

Conversely, one should ask the more basic question of whether the similarity between two objects forces their features to be similar.

Stability of features. Thus, a problem of interest is studying the extent to which a given feature is stable under perturbations of the dataset. In order to be able to say something precise in this respect we introduce some mathematical language.

To fix concepts we imagine that we have a collection $\mathcal{D}$ of all possible datasets, and a collection $\mathcal{F}$ of all possible features. A *feature map* will be any map $f : \mathcal{D} \to \mathcal{F}$. Assume further that $d_\mathcal{D}$ and $d_\mathcal{F}$ are metrics or distance functions on $\mathcal{F}$ and $\mathcal{D}$, respectively. One says that f is *quantitatively stable* whenever one can find a non-decreasing function $\Psi : [0,\infty) \to [0,\infty)$ with $\Psi(0) = 0$ such that for all $X, Y \in \mathcal{D}$ it holds that

$$d_{\mathcal{F}}(f(X), f(Y)) \leq \Psi\big(d_{\mathcal{D}}(X, Y)\big).$$

Note that this is stronger that the usual notion of *continuity* of maps, namely that $f(X_n) \to f(X)$ as $n \uparrow \infty$ whenever $(X_n)_n \subset \mathcal{D}$ is a sequence of datasets converging to X.

In subsequent sections of the paper we will describe instances of suitable metric spaces $(\mathcal{D}, d_{\mathcal{D}})$ and study the stability of different features.

2 Some Considerations

2.1 Importance of Stability and Classification of Algorithms

We claim that it would be desirable to elucidate the stability properties of the main methods used in data analysis. The underlying situation is that the output of data analysis algorithms are used in order to draw conclusions about the phenomenon producing the data, hence it is of extreme importance to make sure that these conclusions would not be grossly affected if the dataset were “noisy” or “slightly perturbed”. In order to make sense of this question one needs to ascribe mathematical meaning to “data”, “perturbations”, “algorithms”, etc.

In a similar vein, it would be clearly highly desirable to know what are the theoretical properties enjoyed by the main algorithms used in data analysis (such as clustering methods, for example). From a theoretical standpoint, it would be very nice to be able to derive algorithms from a list of desirable or required properties or axioms. In this respect, the works of Janowitz [47], Kleinberg [51], and von Luxburg [90] are very prominent.

2.2 Stability and Matching: A Duality

Assuming that datasets X and Y in $\mathcal{D}$ are given, a natural way of comparing them is to compute the $d_{\mathcal{D}}$ distance between them (whatever that distance is). Often times, however, features computed out of datasets constitute simpler structures than the datasets themselves, and as such, they are more readily amenable to direct comparisons.

So, for a family of indices A consider here the stable family $\{f_\alpha, \alpha \in A\}$ of feature maps $f_\alpha : \mathcal{D} \to \mathcal{F}$, where $\alpha \in A$ and $\mathcal{F}$ is some *feature space* which is metrized by the distance function $d_{\mathcal{F}}$. In line with the observation above, spaces of features tend to have simpler structure than the space of datasets, and in consequence the computation of $d_{\mathcal{F}}$ usually appears to be simpler. This suggests that in order to distinguish between two datasets X and Y one computes

$$\eta_A(X, Y) := \sup_{\alpha \in A} d_{\mathcal{F}}\big(f_\alpha(X), f_\alpha(Y)\big)$$

as a proxy for $d_{\mathcal{D}}(X,Y)$. This would be reasonable because since each of the features f_α, $\alpha \in A$ is stable, there exist functions Ψ_α such that

$$\eta_A(X,Y) \leq \sup_{\alpha\in A} \Psi_\alpha\big(d_{\mathcal{D}}(X,Y)\big).$$

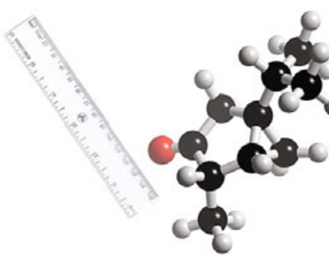
However, in order for this to be totally satisfactory it would be necessary to establish in the reverse direction! For a given subclass of datasets $\mathcal{O} \subset \mathcal{D}$, the main challenge is to find a stable family $\{f_\alpha, \alpha \in A\}$ that is rich enough so that it will discriminate all objects in $\mathcal{O}$: namely that if $X, Y \in \mathcal{O}$ and

$$f_\alpha(X) = f_\alpha(Y) \text{ for all } \alpha \in A \Longrightarrow X = Y.$$

In this respect the work of Olver [67], Boutin and Kemper [5] provide for example families of features that are able to discriminate certain datasets under rigid isometries. Other interesting and useful examples are ultrametric spaces, or in more generality trees.

3 Datasets as Metric Spaces or Metric Measure Spaces

In many applications datasets can be represented as metric spaces (see Figure 2), that is, as a pair (X, d_X) where $d_X : X \times X \to \mathbb{R}^+$ satisfies the three metric properties: (a) $d_X(x,x') = 0$ if and only $x = x'$; (b) $d_X(x,x') = d_X(x',x)$ for all $x, x' \in X$; and (c) $d_(x,x') \leq d_X(x,x'') + d_X(x'',x')$ for all $x,x',x'' \in X$. Henceforth, $\mathcal{G}$ will denote the collection of all compact metric spaces.



$$\Longrightarrow \begin{pmatrix} 0 & d_{12} & d_{13} & d_{14} & \cdots \\ d_{12} & 0 & d_{23} & d_{24} & \cdots \\ d_{13} & d_{23} & 0 & d_{34} & \cdots \\ d_{14} & d_{24} & d_{34} & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Fig. 2. Datasets as metric spaces: given the dataset, and a notion of "ruler", one induces a matrix containing the distance between all pairs of points; this distance is application dependent

We introduce some notation: for a finite metric space (X, d_X), its *separation* is the number $\mathbf{sep}\,(X) := \min_{x\neq x'} d_X(x,x')$. For any compact X, its diameter is $\mathbf{diam}\,(X) := \max_{x,x'} d_X(x,x')$.

For example in the case of Euclidean datasets, one has the following result:

Lemma 1 ([5]). *Let X and Y be finite subsets of $\mathbb{R}^k$ s.t. there exists $\phi : X \to Y$ a bijection with $\|x - x'\| = \|\phi(x) - \phi(x')\|$ for all $x, x' \in X$. Then, there exist a rigid isometry $\Phi : \mathbb{R}^d \to \mathbb{R}^d$ s.t. $Y = \Phi(X)$.*

This lemma implies that representing a Euclidean dataset (e.g. a protein, a chemical compound, etc) as a metric space by endowing it with the ambient space distance, one retains the original information up to ambient space isometries (in this case, rotations, translations, and reflections). In particular, this is not restrictive in any way, because anyhow in most conceivable cases one would not want the output of an algorithm to depend on the coordinate system in which the data is represented.

In the context of protein structure comparison, some ideas regarding the direct comparison of distance matrices can be found for example in [40].

There are other types of datasets which are not Euclidean, but also fit in the metric framework. One example is given by phylogenetic trees. Indeed, it is well known [78] that trees are exactly those metric spaces (X, d_X) that satisfy the *four point condition*: for all $x, y, z, w \in X$

$$d_X(x,y) + d_X(z,w) \leq \max\big(d_X(x,z) + d_X(y,w), d_X(x,w) + d_X(z,y)\big).$$

Another rich class of examples where the metric representation of objects arises in problems in object recognition under invariance to bending transformations, see [55,28,63,64,12,41,70,11,9,10,8].

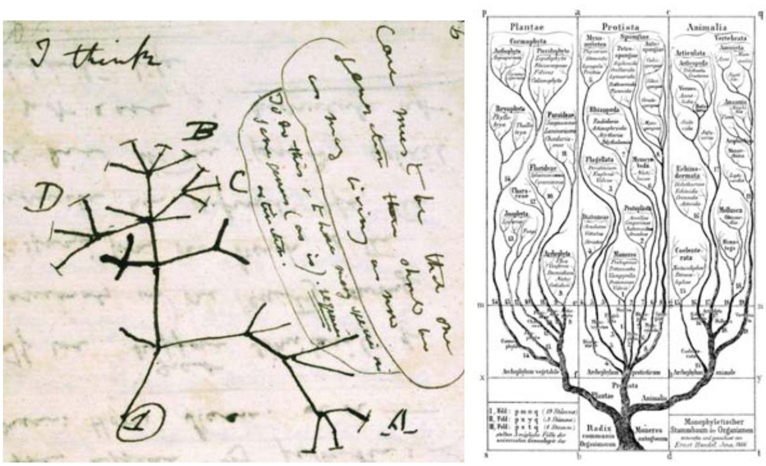


Fig. 3. Famous phylogenetic trees

mm-spaces. A *metric measure space* or *mm-space* for short, is a triple (X, d_X, μ_X) where (X, d_X) is a metric space and μ_X is a Borel probability measure on X. In the finite case, μ_X reduces to a collection of non-negative *weights*, one for each point $x \in X$, such that the sum of all weights equals 1. The interpretation is that $\mu_X(x)$ measures the "importance" of x: points with zero weight should not matter, points with lower values of the weight should be less prominent than points with larger values of the weight, etc. The representation of objects as mm-spaces can thus incorporate more information than the purely metric representation of data— when there is no application motivated choice of weights one can resort to the giving the points the *uniform distribution*, that is all points would have the same weight.

Henceforth, $\mathcal{G}_w$ will denote the collection of all compact mm-spaces.[1]

3.1 Equality of Datasets

What is the notion of equality between datasets? In the case when datasets are represented as metric spaces, we declare that $X, Y \in \mathcal{G}$ are equal whenever we cannot tell them apart by performing pairwise measurements of interpoint distances. In mathematical language, in order to check whether X and Y are equal we require that there be a surjective map $\phi : X \to Y$ which preserves distances and leaves no holes:

- $d_X(x, x') = d_Y(\phi(x), \phi(x'))$ for all $x, x' \in X$; and
- $\phi(X) = Y$.

Such maps (when X and Y are compact) are necessarily bijective, and are called *isometries*.

When datasets are represented as mm-spaces the notion of equality between them must take into account the preservation of not only the pair-wise distance information, but also that of the weights. One considers $X, Y \in \mathcal{G}_w$ to be equal, whenever there exists an isometry $\phi : X \to Y$ that *also preserves the weights*: namely that (assume that X and Y are finite for simplicity) $\mu_X(x) = \mu_Y(\phi(x))$, for all $x \in X$, see [60].

4 Metric Structures on Datasets

We now wish to produce a notion of distance between datasets that is not "too rigid" and allows substantiating a picture such as that emerging from §2.2. We will now describe the construction of distances in both $\mathcal{G}$ and $\mathcal{G}_w$.

4.1 The Case of $\mathcal{G}$

A suitable notion of distance between objects in $\mathcal{G}$ is the *Gromov-Hausdorff* distance, which can be defined as follows. We first introduce the case of finite objects and then explain the general construction.

Given objects $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ with metrics d_X and d_Y, respectively, let $R = ((r_{ij})) \in \{0,1\}^{n \times m}$ be such that

$$\sum_i r_{ij} \geq 1 \text{ for all } j \text{ and } \sum_j r_{ij} \geq 1 \text{ for all } i.$$

[1] The sub-index w is meant to suggest "weighted metric spaces".

The interpretation is that any such binary matrix R represents a notion of *friendship* between points in X and points in Y: namely, that x_i and y_j are friends if and only if $r_{ij} = 1$. Notice that the conditions above imply that every point in X has at least one friend in Y, and reciprocally, that every point in Y has at least one friend in X.

Denote by $\mathcal{R}(X,Y)$ the set of all such possible matrices, which we shall henceforth refer to as *correspondences* between X and Y.

Then, one defines the Gromov-Hausdorff distance between (X, d_X) and (Y, d_Y) as

$$d_{\mathcal{GH}}(X,Y) := \frac{1}{2} \min_{R} \max_{i,i',j,j'} \big| d_X(x_i, x_{i'}) - d_X(x_j, x_{j'}) \big| r_{ij} r_{i'j'},$$

where the minimum is taken over $R \in \mathcal{R}(X,Y)$.

The definition above has the interpretation that one is trying to match points in X to points in Y in such a way that the metrics of X and Y are optimally aligned.

The general case. In the full case of any pair of datasets X and Y (not necessarily finite) in $\mathcal{G}$, one needs to generalize the definition above. Let $\mathcal{R}(X,Y)$ denote now the collection of all subsets R of the Cartesian product $X \times Y$ with the property that the canonical coordinate projections $\pi_1 : X \times Y \to X$ and $\pi_2 : X \times Y \to Y$ are *surjective*, when restricted to R.

Then the Gromov-Hausdorff distance between compact metric spaces X and Y is defined as

$$d_{\mathcal{GH}}(X,Y) := \frac{1}{2} \inf_{R \in \mathcal{R}(X,Y)} \sup_{(x,y),(x',y') \in R} \big| d_X(x,x') - d_Y(y,y') \big|. \tag{1}$$

This definition indeed respects the notion of equality of objects that we put forward in §3.1:

Theorem 1 ([35]). *$d_{\mathcal{GH}}$ is a metric on the isometry classes of $\mathcal{G}$.*

Another expression for the GH distance. Recall the definition of the Hausdorff distance between (closed) subsets A and B of a metric space (Z, d_Z):

$$d_{\mathcal{H}}^{Z}\big(A,B\big) := \max\big(\max_{a\in A} \min_{b \in B} d_Z(a,b), \max_{b\in B} \min_{a\in A} d_Z(a,b) \big).$$

Given compact metric spaces (X, d_X) and (Y, d_Y), consider all metrics d on the disjoint union $X \sqcup Y$ s.t.

- $d(x,x') = d_X(x,x')$, all $x, x' \in X$;
- $d(y,y') = d_Y(y,y')$, all $y, y' \in Y$.

Then, according to [13, Chapter 7]

$$d_{\mathcal{GH}}(X,Y) := \inf_{d} d_{\mathcal{H}}^{(X \sqcup Y, d)}\big(X,Y\big),$$

where the infimum is taken over all the metrics d that satisfy the conditions above.

Remark 1. According to this formulation, computing the GH distance between two finite metric spaces can be regarded as a **distance matrix completion problem**. The functional is $J(d) = \max\big(\max_x \min_y d(x,y), \max_y \min_x d(x,y)\big)$ [60]. The number of constraints is roughly of order n^3 for all the **triangle inequalities**, where $n = |X| \simeq |Y|$.

Example: Euclidean datasets. Endowing objects embedded in $\mathbb{R}^d$ with the (restricted) Euclidean metric makes the Gromov-Hausdorff distance invariant under ambient rigid isometries [59]. In order to argue that similarity in the Gromov-Hausdorff sense has a meaning which is compatible and comparable with other notions of similarity that we have already come to accept as natural, it is useful to look into the case of similarity of objects under rigid motions. One of the most commonplace notions of rigid similarity is given by the Hausdorff distance under rigid isometries [43] for which one has

Theorem 2 ([59]). *Let $X, Y \subset \mathbb{R}^d$ be compact. Then*

$$d_{\mathcal{GH}}((X,\|\cdot\|),(Y,\|\cdot\|)) \leq \inf_T d_{\mathcal{H}}^{\mathbb{R}^d}(X,T(Y)) \leq c_d \cdot M^{\frac{1}{2}} \cdot \big(d_{\mathcal{GH}}((X,\|\cdot\|),(Y,\|\cdot\|))\big)^{\frac{1}{2}},$$

where $M = \max(\mathbf{diam}\,(X), \mathbf{diam}\,(Y))$ and c_d is a constant that depends only on d. The infimum over T above is taken amongst all Euclidean isometries.

Note that this theorem is a natural relaxation of the statement of Lemma 1.

4.2 The Case of $\mathcal{G}_w$

Using ideas from mass transport it is possible to define a version of the Gromov-Hausdorff distance that applies to datasets in $\mathcal{G}_w$.

Fix a metric space (Z, d_Z) and let $\mathcal{P}(Z)$ denote the collection of all the Borel probability measures. For $\alpha, \beta \in \mathcal{P}(Z)$, the **Wasserstein distance** (or order $p \geq 1$) on $\mathcal{P}(Z)$ is given by:

$$d_{\mathcal{W},p}^{(Z,d_Z)}(\alpha,\beta) := \left(\iint_{Z\times Z} \big(d_Z(z,z')\big)^p \mu(dz \times dz')\right)^{1/p},$$

where $\mu \in \mathcal{P}(Z \times Z)$ is a probability measure with marginals α and β. An excellent reference for these concepts is the book of Villani [89].

An interpretation of this definition comes from thinking that one has a pile of sand/dirt that must be moved from one location to another, where in the destination one wants build something with this material, see Figure 4. In the finite case (i.e. when all the probability measures are linear combinations of deltas), $\mu_{i,j}$ encodes information about how much of the mass initially at x_i must be moved to x_j, see Figure 4.

The **Gromov-Wasserstein distance** between mm-spaces X and Y is defined as an *optimal mass transportation* problem on $X \sqcup Y$: for $p \geq 1$

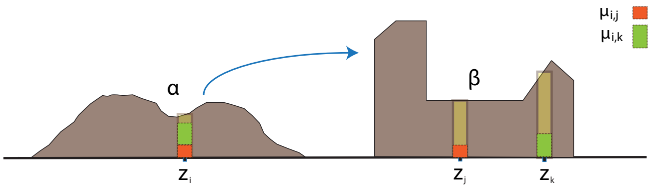


Fig. 4. An optimal mass transportation problem (in the Kantorovich formulation): the pile of sand/dirt on the left must be moved to another location on the right with the purpose of assembling a building or structure

$$d_{\mathcal{GW},p}(X,Y) := \inf_d d_{\mathcal{W},p}^{(X\sqcup Y,d)}(\mu_X,\mu_Y),$$

where as before d is a metric on $X \sqcup Y$ gluing X and Y.

The definition above is due to Sturm [83]. Notice that the underlying optimization problems that one needs to solve now are of continuous nature as opposed to the combinatorial optimization problems yielded by the GH distance. Another non-equivalent definition of the Gromov-Wasserstein distance is proposed in [60] whose discretization is more tractable.

As we will see ahead, several features become stable in the GW sense.

4.3 Stability of Hierarchical Clustering Methods

Denote by $\mathbf{P}(X)$ the set of all partitions of the finite set X.

A *dendrogram* over a finite set X is a function $\theta_X : [0,\infty) \to \mathbf{P}(X)$ with the following properties:

1. $\theta_X(0) = \{\{x_1\},\ldots,\{x_n\}\}$.
2. There exists t_0 s.t. $\theta_X(t)$ is the *single block partition* for all $t \geq t_0$.
3. If $r \leq s$ then $\theta_X(r)$ *refines* $\theta_X(s)$.
4. For all r there exists $\varepsilon > 0$ s.t. $\theta_X(r) = \theta_X(t)$ for $t \in [r, r+\varepsilon]$.

Let $\mathbf{D}(X)$ denote the collection of all possible dendrograms over a given finite set X.

Hierarchical clustering methods are maps $\mathfrak{H}$ from the collection of all finite metric spaces into the collection of all dendrograms, such that (X, d_X) is mapped into an element of $\mathbf{D}(X)$.

Standard examples of clustering methods are *single, complete and average linkage methods* [46].

A question of great interest is whether any of these clustering methods is stable to perturbations in the input metric spaces.

Linkage based agglomerative HC methods. Here we review the basic procedure of linkage based hierarchical clustering methods:

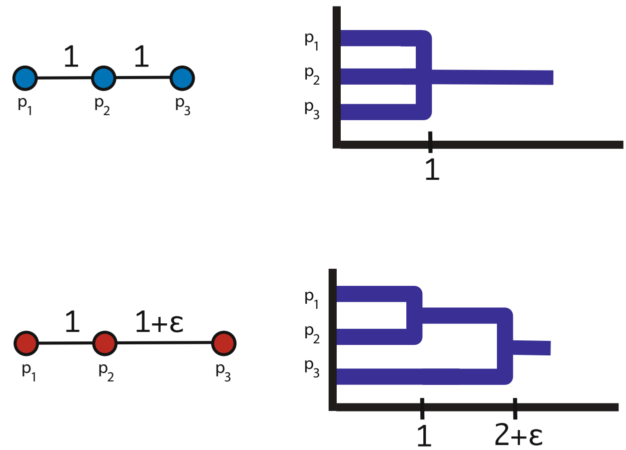


Fig. 5. Complete Linkage is not stable to small perturbations in the metric. On the left we show two metric spaces that are metrically very similar. To the right of each of them we show their CL dendrogram outputs. Regardless of $\varepsilon > 0$, the two outputs are always very dissimilar.

Assume (X, d_X) is a given finite metric space. In this example, we use the formulas for CL but the structure of the iterative procedure in this example is common to all HC methods [46, Chapter 3]. Let θ be the dendrogram to be constructed in this example.

1. Set $X_0 = X$ and $D_0 = d_X$ and set $\theta(0)$ to be the partition of X into singletons.
2. Search the matrix D_0 for the smallest non-zero value, i.e. find $\delta_0 = \mathbf{sep}\,(X_0)$, and find all pairs of points $\{(x_{i_1}, x_{j_1}), (x_{i_2}, x_{j_2}) \ldots, (x_{i_k}, x_{j_k})\}$ at distance δ_0 from eachother, i.e. $d(x_{i_\alpha}, x_{j_\alpha}) = \delta_0$ for all $\alpha = 1, 2, \ldots, k$, where one orders the indices s.t. $i_1 < i_2 < \ldots < i_k$.
3. Merge the first pair of elements in that list, (x_{i_1}, x_{j_1}), into a single group. The procedure now removes (x_{i_1}, x_{j_1}) from the initial set of points and adds a point c to represent the cluster formed by both: define $X_1 = \big(X_0 \backslash \{x_{i_1}, x_{j_1}\}\big) \cup \{c\}$. Define the dissimilarity matrix D_1 on $X_1 \times X_1$ by $D_1(a,b) = D_0(a,b)$ for all $a, b \neq c$ and $D_1(a,c) = D_1(c,a) = \max\big(D_0(x_{i_1}, a), D_0(x_{j_1}, a)\big)$ (this step is the only one that depends on the choice corresponding to CL). Finally, set
$$\theta(\delta) = \{x_{i_1}, x_{j_1}\} \cup \bigcup_{i \neq i_1, j_1} \{x_i\}.$$
4. The construction of the dendrogram θ is completed by repeating the previous steps until all points have been merged into a single cluster.

The *tie breaking* strategy used in step 3 results in the algorithm producing different non-isomorphic outputs depending on the labeling of the points. This is undesirable, but can be remedied by defining certain versions of all the linkage based HC methods that behave well under permutations [18] .

Unfortunately, even these “patched” versions of AL and CL fail to exhibit stability, see Figure 5.

It turns out, however, that single linkage does enjoy stability. Before we phrase the precise result we need to introduce the ultrametric representation of dendrograms. Furthermore, as we will see in 5, there's a sense in which SLHC is *the only HC method that can be stable.*

Dendrograms as ultrametric spaces. The representation of dendrograms as ultrametrics is well known [48,37,46].

Theorem 3 ([18]). *Given a finite set X, there is a bijection $\Psi : \mathbf{D}(X) \to \mathbf{U}(X)$ between the collection $\mathbf{D}(X)$ of all dendrograms over X and the collection $\mathbf{U}(X)$ of all ultrametrics over X such that for any dendrogram $\theta \in \mathbf{D}(X)$ the ultrametric $\Psi(\theta)$ over X generates the same hierarchical decomposition as θ, i.e.*

$$(*) \quad \text{for each } r \geq 0,\ x, x' \in B \in \theta(r) \iff \Psi(\theta)(x, x') \leq r.$$

Furthermore, this bijection is given by

$$\Psi(\theta)(x, x') = \min\{r \geq 0 |\, x, x' \text{ belong to the same block of } \theta(r)\}. \qquad (2)$$

See Figure 6.

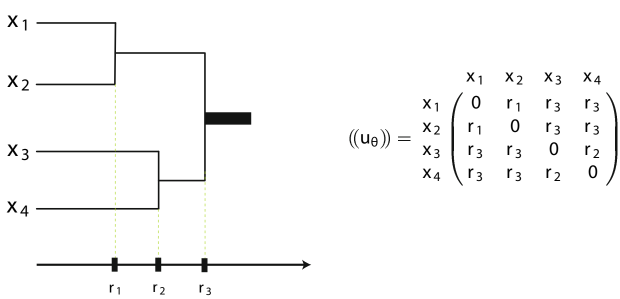


Fig. 6. A graphical representation of a dendrogram θ over $X = \{x_1, x_2, x_3, x_3\}$ and the corresponding ultrametric $u_\theta := \Psi(\theta)$. Notice for example, that according to (2), $u_\theta(x_1, x_2) = r_1$ since r_1 is the first value of the (scale) parameter for which x_1 and x_2 are merged into the same cluster. Similarly, since x_1 and x_3 are merged into the same cluster for the first time when the parameter equals r_3, then $u_\theta(x_1, x_3) = r_3$.

Let $\mathcal{U} \subset \mathcal{G}$ denote the collection of all (compact) ultrametric spaces. It follows from Theorem 3 that one can regard HC methods as maps $\mathfrak{H} : \mathcal{G} \to \mathcal{U}$. In particular [18], SLHC can be regarded as the map $\mathfrak{H}^{\mathrm{SL}}$ that assigns (X, d_X)

with (X, u_X), where u_X is the *maximal subdominant ultrametric* relative to d_X. This is given as

$$u_X(x,x') := \min\left\{\max_{i=0,\dots,k-1} d_X(x_i,x_{i+1}),\text{ s.t. } x=x_0,\dots,x_k=x'\right\}. \quad (3)$$

Stability and convergence of SLHC. In contrast with the situation for complete and average linkage HCMs, we have the following statement concerning the quantitative stability of SLHC:

Theorem 4 ([18]). *Let (X,d_X) and (Y,d_Y) be two finite metric spaces. Then,*

$$d_{\mathcal{GH}}(\mathfrak{H}^{\mathrm{SL}}(X,d_X),\mathfrak{H}^{\mathrm{SL}}(Y,d_Y)) \leq d_{\mathcal{GH}}((X,d_X),(Y,d_Y)).$$

Invoking the ultrametric representation of dendrograms and using Theorem 4, [18] proves the following convergence result, see Figure 7.

Theorem 5. *Let (Z,d_Z,μ_Z) be an mm-space and write $\mathrm{supp}\,[\mu_Z]=\bigcup_{\alpha\in A}Z^{(\alpha)}$ for a finite index set A and $\{Z^{(\alpha)}\}_{\alpha\in A}$ a collection of disjoint, compact, path-connected subsets of Z. Let (A,u_A) be the ultrametric space where u_A is the maximal subdominant ultrametric with respect to $W_A(\alpha,\alpha') := \min_{z\in Z^{(\alpha)},z'\in Z^{(\alpha')}} d_Z(z,z')$, for $\alpha,\alpha'\in A$.*

For each $n\in\mathbb{N}$, let $X_n=\{z_1,z_2,\dots,z_n\}$ be a collection of n independent random variables (defined on some probability space Ω with values in Z) with distribution μ_Z, and let d_{X_n} be the restriction of d_Z to $X_n\times X_n$. Then, $\mathfrak{H}^{\mathrm{SL}}(X_n,d_{X_n}) \xrightarrow{n} (A,u_A)$ in the Gromov-Hausdorff sense μ_Z-almost surely.

4.4 Stability of Vietoris-Rips Barcodes

Much in the same way as standard flat clustering can be understood as the zero-dimensional version of the notion of homology, hierarchical clustering can be regarded as the zero-dimensional version of persistent homology [27].

The notion of Vietoris-Rips persistent barcodes provides a precise sense in which the above statement is true. For a given finite metric space (X,d_X) and $r\geq 0$, let $R_r(X)$ denote the simplicial complex with vertex set X where $\sigma=[x_0,x_1,\dots,x_k]\in R_r(X,d_X)$ if and only if $\max_{i,j} d_X(x_i,x_j)\leq r$. This is called the Vietoris-Rips simplicial complex (with parameter r). Then, the family

$$\mathcal{R}(X,d_X) := \big\{R_r(X,d_X),\, r\geq 0\big\}$$

constitutes a *filtration*, in the sense that

$$R_r(X,d_X)\subseteq R_s(X,d_X),\ \text{whenever } s\geq r.$$

In the sequel we may abbreviate $R_r(X)$ for $R_r(X,d_X)$, and similarly for $\mathcal{R}(X)$. Now, passing to homology with field coefficients, this inclusion gives rise to a pair of vector spaces and a linear map between them:

$$\phi_r^s : H_*(R_r(X) \longrightarrow H_*(R_s(X)).$$

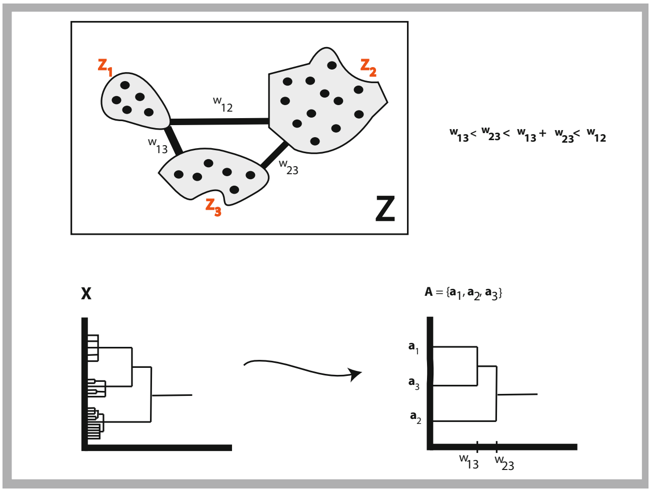


Fig. 7. Illustration of Theorem 5. *Top*: A space Z composed of 3 disjoint path connected parts, $Z^{(1)}, Z^{(2)}$ and $Z^{(3)}$. The black dots are the points in the finite sample X_n. In the figure, $w_{ij} = W_A(a_i, a_j)$, $1 \leq i \neq j \leq 3$. *Bottom Left*: The dendrogram representation of $(X_n, u_{X_n}) := \mathfrak{H}^{\mathrm{SL}}(X_n)$. *Bottom Right*: The dendrogram representation of (A, u_A). Note that $u_A(a_1, a_2) = w_{23}$, $u_A(a_1, a_3) = w_{13}$ and $u_A(a_2, a_3) = w_{23}$. As $n \to \infty$, $(X_n, u_{X_n}) \to (A, u_A)$ a.s. in the Gromov-Hausdorff sense, see text for details.

In more detail, if $0 = \alpha_0 < \alpha_1 < \ldots < \alpha_m = \mathbf{diam}\,(X)$ are the distinct values assumed by d_X, then one obtains the *persistent vector space*:

$$H_*(R_{\alpha_0}(X)) \xrightarrow{\phi_0^1} H_*(R_{\alpha_1}(X)) \xrightarrow{\phi_1^2} H_*(R_{\alpha_2}(X)) \xrightarrow{\phi_2^3} \cdots$$
$$\cdots \xrightarrow{\phi_{m-2}^{m-1}} H_*(R_{\alpha_{m-1}}(X)) \xrightarrow{\phi_{m-1}^{m}} H_*(R_{\alpha_m}(X)).$$

It is well known [91] that there is a classification of such objects in terms of a finite multisets of points in the extended plane $\overline{\mathbb{R}}^2$, called the *persistence diagram* of $\mathcal{R}(X)$, and denoted $\mathrm{D}_*\mathcal{R}(X)$ which is contained in the union of the extended diagonal $\Delta = \{(x, x) : x \in \overline{\mathbb{R}}\}$ and of the grid $\{\alpha_0, \cdots, \alpha_m\} \times \{\alpha_0, \cdots, \alpha_m, \alpha_\infty = +\infty\}$. The multiplicity of the points of Δ is set to $+\infty$, while the multiplicities of the (α_i, α_j), $0 \leq i < j \leq +\infty$, are defined in terms of the ranks of the linear transformations $\phi_i^j = \phi_{j-1}^j \circ \cdots \circ \phi_i^{i+1}$ [20].

The *bottleneck distance* $d_{\mathrm{B}}^\infty(A, B)$ between two multisets in $(\overline{\mathbb{R}}^2, l^\infty)$ is the quantity $\min_\gamma \max_{p \in A} \|p - \gamma(p)\|_\infty$, where γ ranges over all bijections from A to B. Then, one obtains the following generalization of Theorem 4.

Theorem 6 ([20]). *Let (X, d_X) and (Y, d_Y) be any two finite metric spaces. Then for all $k \geq 0$,*

$$\frac{1}{2} d_{\mathrm{B}}^\infty\big(\mathrm{D}_k\mathcal{R}(X), \mathrm{D}_k\mathcal{R}(Y)\big) \leq d_{\mathcal{GH}}(X, Y).$$

This type of results are of great importance for applications of the Vietoris-Rips barcodes to data analysis.

4.5 Object Matching: More Details

Some features of mm-spaces. We define a few simple isomorphism invariants, or features, of mm-spaces, many of which will be used in §4.5 to establish lower bounds for the metrics we will impose on $\mathcal{G}_w$. All the features we discuss below have are routinely used in the data analysis and object matching communities.

Definition 1 (p-diameters). *Given a mm-space (X, d_X, μ_X) and $p \in [1, \infty]$ we define its p-**diameter** as*

$$\mathbf{diam}_p(X) := \left(\int_X \int_X \big(d_X(x,x')\big)^p \mu_X(dx)\mu_X(dx') \right)^{1/p}$$

for $1 \leq p < \infty$.

Definition 2. *Given $p \in [1, \infty]$ and an mm-space (X, d_X, μ_X) we define the p-**eccentricity function** of X as*

$$s_{X,p} : X \to \mathbb{R}^+ \quad \textit{given by} \quad x \mapsto \left(\int_X d_X(x,x')^p \mu(dx') \right)^{1/p}$$

for $1 \leq p < \infty$.

Hamza and Krim proposed using eccentricity functions (with $p = 2$) for describing objects in [36]. Ideas similar to those proposed in [36] have been revisited recently in [45]. See also Hilaga et al. [39]. Eccentricities are also routinely used as part of topological data analysis algorithms such as mapper [80].

Definition 3 (Distribution of distances). *To an mm-space (X, d_X, μ_X) we associate its **distribution of distances**:*

$$f_X : [0, \mathbf{diam}(X)] \to [0,1] \quad \textit{given by} \quad t \mapsto \mu_X \otimes \mu_X\big(\{(x,x') | d_X(x,x') \leq t\}\big).$$

See Figure 8 and [5,68].

Definition 4 (Local distribution of distances). *To a mm-space (X, d_X, μ_X) we associate its* local distribution of distances *defined by:*

$$h_X : X \times [0, \mathbf{diam}(X)] \to [0,1] \quad \textit{given by} \quad (x,t) \mapsto \mu_X\left(\overline{B_X(x,t)}\right).$$

See Figure 9. The earliest use of an invariant of this type known to the author is in the work of German researchers [4,50,1]. The so called **shape context** [3,79,75,14] invariant is closely related to h_X.

More similar to h_X is the invariant proposed by Manay et al. in [58] in the context of planar objects. This type of invariant has also been used for three dimensional objects [21,32]. More recently, in the context of planar curves, similar constructions have been analyzed in [7]. See also, [34].

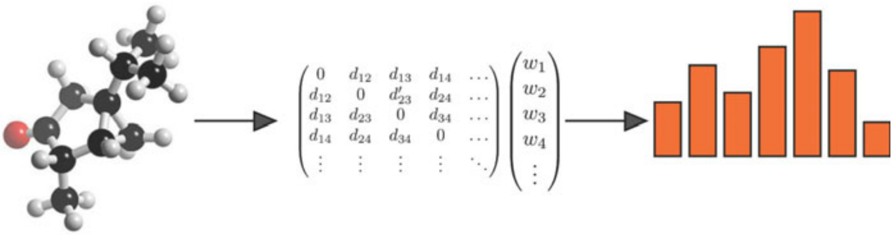


Fig. 8. Distribution of distances: from a dataset to the mm-space representation and from it to the distribution of distances

Remark 2 (**Local distribution of distances as a proxy for scalar curvature).** There is an interesting observation that in the class Riem $\subset \mathcal{G}_w$ of closed Riemannian manifolds local distributions of distance are intimately related to curvatures. Let M be an n-dimensional closed Riemannian manifold which we regard as an mm-space by endowing it with the geodesic metric and with probability measure given by the normalized volume measure. Using the well known expansion [77] of the Riemannian volume of a ball of radius t centered at $x \in M$ one finds:

$$h_M(x,t) = \frac{\omega_n(t)}{\mathbf{Vol}\,(M)}\left(1 - \frac{S_M(x)}{6(n+2)}t^2 + O(t^4)\right),$$

where $S_M(x)$ is the *scalar curvature* of M at x, $\omega_n(t)$ is the volume of a ball of radius t in $\mathbb{R}^n$ and $O(t^4)$ is a term whose decay to 0 as $t \downarrow 0$ is faster than t^4.

One may then argue that local shape distributions play a role of generalized notions of curvature.

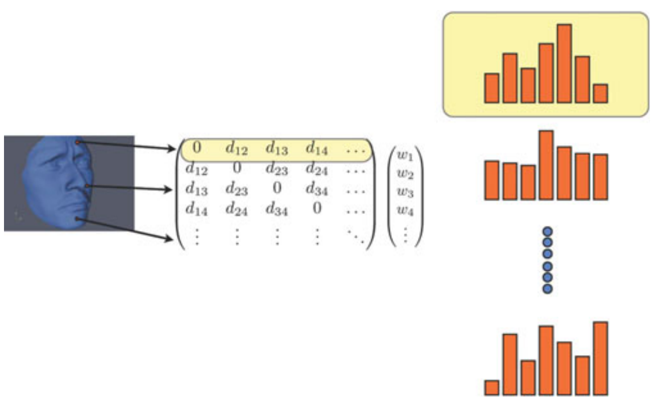


Fig. 9. Local distribution of distances: from a dataset to the mm-space representation and from it the local distribution of distances. To each point on the object one assigns the distribution of distance from this point to all other points on the object.

Precise bounds

Definition 5. *For $X, Y \in \mathcal{G}_w$ define*

$$\mathbf{FLB}(X,Y) := \frac{1}{2} \inf_{\mu \in \mathcal{M}(\mu_X, \mu_Y)} \left(\int_{X \times Y} |s_{X,1}(x) - s_{Y,1}(y)| \, \mu(dx \times dy) \right);$$
$$\mathbf{SLB}(X,Y) := \int_0^\infty |f_X(t) - f_Y(t)| \, dt;$$
$$\mathbf{TLB}(X,Y) := \frac{1}{2} \min_{\mu \in \mathcal{M}(\mu_X, \mu_Y)} \int_{X \times Y} \left(\int_0^\infty \big| h_X(x,t) - h_Y(y,t) \big| \, dt \right) \mu(dx \times dy).$$

For finite X and Y, computing the (exact) value of each of the quantities in the definition reduces to solving linear programming problems [60].

We now can state the following theorem asserting the stability of the features discussed in this section:

Theorem 7 ([60]). *For all $X, Y \in \mathcal{G}_w$, and all $p \geq 1$*

$$d_{\mathcal{GW},p}(X,Y) \geq \begin{cases} \mathbf{TLB}(X,Y) \geq \mathbf{FLB}(X,Y) \geq \frac{1}{2} |\mathbf{diam}_1(X) - \mathbf{diam}_1(Y)|. \\ \mathbf{SLB}(X,Y). \end{cases}$$

Bounds of this type, besides establishing the quantitative stability of the different intervening features, have the added usefulness that in practice they may be utilized in *layered* comparison of objects: those bounds involving simpler invariants are frequently easier to compute, whereas those involving more powerful features most often require more effort. Furthermore, hierarchical bounds of this nature that interconnect different approaches proposed in the literature allow for a better understanding of the landscape of different existing techniques, see Figure 10.

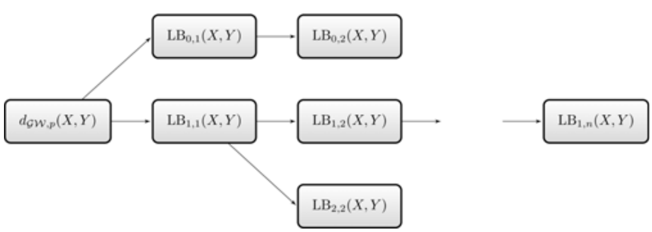


Fig. 10. Having a hierarchy (arrows should be read as $\geq$ symbols) of lower bounds such the one suggested in the figure can help in matching tasks: the strategy that suggests itself is to start the comparison using the weaker bounds and gradually increase the complexity

Also, different families of lower bounds for the GH distance have recently been found [61]; these incorporate features similar to those of [56,86].

Spectral versions of the GH and GW distances. It is possible to obtain a hierarchy of lower bounds similar to the ones above but in the context of *spectral methods* [62], see Figure 12. The motivation comes from the so called Varadhan's Lemma: if X is a compact Riemannian manifold without boundary, and k_X denotes the *heat kernel* of X, then one has

Lemma 2 ([66]). *For any compact Riemannian manifold without boundary X,*

$$\lim_{t\downarrow 0}\big(-4t\ln k_X(t,x,x')\big)=d_X^2(x,x'),$$

for all $x,x'\in X$. Here $d_X(x,x')$ is the geodesic distance between x and x' on X.

The spectral representation of objects (see Figure 12), and in particular shapes is interesting because it readily encodes a notion of *scale*. This scale parameter (the t parameter in the heat kernel) permits reasoning about similarity of shapes at different levels of "blurring" or "smoothing", see Figure 11. A (still not thoroughly satisfactory) interpretation of t as a scale parameter arises from the following observations:

- For $t\downarrow 0^+$, $k_X(t,x,x)\simeq(4\pi t)^{-d/2}\big(1+\frac{1}{6}S_X(x)+\ldots\big)$, where d is the dimension of X. Recall that S_X is the scalar curvature— therefore for small enough t, one sees local information about X.
- For $t\to\infty$, $k_X(t,x,x')\to\frac{1}{\mathbf{Vol}(X)}$. Hence, for large t all points "look the same".
- Pick $n\in\mathbb{N}$ and $\varepsilon>0$ and let $L_g=(\mathbb{R},g,\lambda)$ for $g(x)=1+\varepsilon\cos(2\pi xn)$, then the *homogenized metric* is $\overline{g}=1$. Then, by results due to Tsuchida and Davies [87,26] one has that

$$\sup_{x,x'\in\mathbb{R}}\big|k_g(t,x,x')-k_{\overline{g}}(t,x,x')\big|\leq\frac{C}{t}\text{ as }t\uparrow\infty.$$

Since for Riemannian manifolds X and Y, by Varadhan's lemma, the heat kernels k_X and k_Y determine the geodesic metrics d_X and d_Y, respectively, this suggests defining **spectral versions** of the GH and GW distances. For each $p\geq 1$, one defines [62]

$$d^{\mathrm{spec}}_{\mathcal{GW},p}(X,Y):=\frac{1}{2}\inf_{\mu}\sup_{t>0}\mathbf{F}_p\big(k_X(t,\cdot,\cdot),k_Y(t,\cdot,\cdot),\mu\big),$$

where $\mathbf{F}_p$ is a certain functional that depends on both heat kernels and the measure coupling μ (see [62]).[2] The interpretation is that one takes the supremum over all t as way of choosing *the most discriminative scale.*

One has:

Theorem 8 ([62]). $d^{\mathrm{spec}}_{\mathcal{GW},p}$ *defines a metric on the collection of (isometry classes of) Riemannian manifolds.*

[2] Here μ is a measure coupling between the *normalized* volume measures of X and Y.

A large number of spectral features are **quantitatively stable** under $d_{\mathcal{GH}}^{\text{spec}}$ [62]. Examples are the spectrum of the Laplace-Beltrami operator [73], features computed from the diffusion distance, [53,22], and the heat kernel signature [84].

A more precise framework for the geometric scales of subsets of $\mathbb{R}^d$ is worked out in [54].

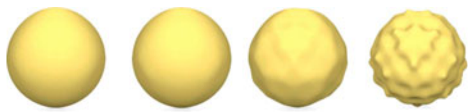


Fig. 11. A bumpy sphere at different levels of smoothing

5 Classification of Algorithms

In the next section, we will give a brief description of the theory of categories and functors, an excellent reference for these ideas is [57].

5.1 Brief Overview of Categories and Functors

Categories are mathematical constructs that encode the nature of certain objects of interest *together with a set of admissible maps between them.*

Definition 6. *A **category** $\underline{C}$ consists of:*

- *A collection of **objects** $\text{ob}(\underline{C})$ (e.g. sets, groups, vector spaces, etc.)*
- *For each pair of objects $X, Y \in \text{ob}(\underline{C})$, a set $\text{Mor}_{\underline{C}}(X,Y)$, the **morphisms** from X to Y (e.g. maps of sets from X to Y, homomorphisms of groups from X to Y, linear transformations from X to Y, etc. respectively)*
- *Composition operations: $\circ : \text{Mor}_{\underline{C}}(X,Y) \times \text{Mor}_{\underline{C}}(Y,Z) \to \text{Mor}_{\underline{C}}(X,Z)$, corresponding to **composition** of set maps, group homomorphisms, linear transformations, etc.*
- *For each object $X \in \underline{C}$, a distinguished element $id_X \in \text{Mor}_{\underline{C}}(X,X)$, called the **identity** morphism.*

The composition is assumed to be associative in the obvious sense, and for any $f \in \text{Mor}_{\underline{C}}(X,Y)$, it is assumed that $id_Y \circ f = f$ and $f \circ id_X = f$.

Definition 7 ($\underline{\mathcal{C}}$, a category of outputs of standard clustering schemes). *Let Y be a finite set, $P_Y \in \mathcal{P}(Y)$, and $f : X \to Y$ be a set map. We define $f^*(P_Y)$ to be the partition of X whose blocks are the sets $f^{-1}(B)$ where B ranges over the blocks of P_Y. We construct the category $\underline{\mathcal{C}}$ of outputs of standard clustering algorithms with $\text{ob}(\underline{\mathcal{C}})$ equal to all possible pairs (X, P_X) where X is a finite set and P_X is a partition of X: $P_X \in \mathcal{P}(X)$. For objects (X, P_X) and (Y, P_Y) one sets $\text{Mor}_{\underline{\mathcal{C}}}\big((X,P_X),(Y,P_Y)\big)$ to be the set of all maps $f : X \to Y$ with the property that P_X is a refinement of $f^*(P_Y)$.*

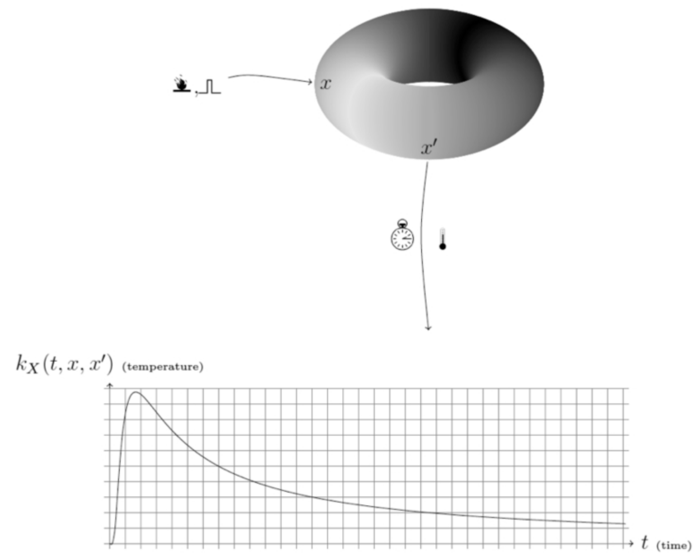


Fig. 12. A physics based way of characterizing/measuring a shape. For each pair of points x and x' on the shape X, one heats a tiny area around point x to a very high temperature in a very short interval of time around $t = 0$. Then, one measures the temperature at point x' for all later times and plots the resulting graph of the heat kernel $k_X(t, x, x')$ as a function of t. The knowledge of these graphs for all $x, x' \in X$ and $t > 0$ translates into knowledge of the heat kernel of X (the plot in the figure corresponds to $x \neq x'$). In contrast, one can think that a geometer's way of characterizing the shape would be via the use of a geodesic ruler that can be used for measuring distances between all pairs of points on X, see Figure 2. According to Varadhan's Lemma, both approaches are equivalent in the sense that they both capture the same information about X.

Example 1. Let X be any finite set, $Y = \{a, b\}$ a set with two elements, and P_X a partition of X. Assume first that $P_Y = \{\{a\}, \{b\}\}$ and let $f : X \to Y$ be any map. Then, in order for f to be a morphism in $\mathrm{Mor}_{\underline{\mathcal{C}}}\big((X, P_X), (Y, P_Y)\big)$ it is necessary that x and x' be in different blocks of P_X whenever $f(x) \neq f(x')$. Assume now that $P_Y = \{a, b\}$ and $g : Y \to X$. Then, the condition that $g \in \mathrm{Mor}_{\underline{\mathcal{C}}}\big((Y, P_Y), (X, P_X)\big)$ requires that $g(a)$ and $g(b)$ be in the same block of P_X.

We will also construct a category of *persistent sets*, which will constitute the output of hierarchical clustering functors.

Definition 8 ($\underline{\mathcal{P}}$, a category of outputs of hierarchical clustering schemes). *Let $(X, \theta_X), (Y, \theta_Y)$ be persistent sets. A map of sets $f : X \to Y$ is said to be* persistence preserving *if for each $r \in \mathbb{R}$, we have that $\theta_X(r)$ is a refinement of $f^*(\theta_Y(r))$. We define a category $\underline{\mathcal{P}}$ whose objects are persistent sets, and where $\mathrm{Mor}_{\underline{\mathcal{P}}}((X, \theta_X), (Y, \theta_Y))$ consists of the set maps from X to Y which are persistence preserving.*

Three categories of finite metric spaces. We will describe three categories $\underline{\mathcal{M}}^{iso}$, $\underline{\mathcal{M}}^{inj}$, and $\underline{\mathcal{M}}^{gen}$, whose collections of objects will all consist of the collection of finite metric spaces $\mathcal{M}$. For (X, d_X) and (Y, d_Y) in $\mathcal{M}$, a map $f : X \to Y$ is said to be **distance non increasing** if for all $x, x' \in X$, we have $d_Y(f(x), f(x')) \leq d_X(x, x')$. It is easy to check that composition of distance non-increasing maps are also distance non-increasing, and it is also clear that id_X is always distance non-increasing. We therefore have the category $\underline{\mathcal{M}}^{gen}$, whose objects are finite metric spaces, and so that for any objects X and Y, $\mathrm{Mor}_{\underline{\mathcal{M}}^{gen}}(X, Y)$ is the set of distance non-increasing maps from X to Y. It is clear that compositions of injective maps are injective, and that all identity maps are injective, so we have the new category $\underline{\mathcal{M}}^{inj}$, in which $\mathrm{Mor}_{\underline{\mathcal{M}}^{inj}}(X, Y)$ consists of the **injective distance non-increasing maps**. Finally, if (X, d_X) and (Y, d_Y) are finite metric spaces, $f : X \to Y$ is an **isometry** if f is bijective and $d_Y(f(x), f(x')) = d_X(x, x')$ for all x and x'. It is clear that as above, one can form a category $\underline{\mathcal{M}}^{iso}$ whose objects are finite metric spaces and whose morphisms are the isometries. Furthermore, one has inclusions

$$\underline{\mathcal{M}}^{iso} \subseteq \underline{\mathcal{M}}^{inj} \subseteq \underline{\mathcal{M}}^{gen} \tag{4}$$

of subcategories (defined as in [57]). Note that although the inclusions are bijections on object sets, they are proper inclusions on morphism sets.

Remark 3. The category $\underline{\mathcal{M}}^{gen}$ is special in that for any pair of finite metric spaces X and Y, $\mathrm{Mor}_{\underline{\mathcal{M}}^{gen}}(X, Y) \neq \emptyset$. Indeed, pick $y_0 \in Y$ and define $\phi : X \to Y$ by $x \mapsto y_0$ for all $x \in X$. Clearly, $\phi \in \mathrm{Mor}_{\underline{\mathcal{M}}^{gen}}(X, Y)$. This is not the case for $\underline{\mathcal{M}}^{inj}$ since in order for $\mathrm{Mor}_{\underline{\mathcal{M}}^{inj}}(X, Y) \neq \emptyset$ to hold it is necessary (but not sufficient in general) that $|Y| \geq |X|$.

Functors and functoriality. Next we introduce the key concept in our discussion, that of a *functor*. We give the formal definition first, and several examples will appear as different constructions that we use in the paper.

Definition 9 (Functor). *Let $\underline{C}$ and $\underline{D}$ be categories. Then a* functor *from $\underline{C}$ to $\underline{D}$ consists of:*

- *A map of sets $F : \mathrm{ob}(\underline{C}) \to \mathrm{ob}(\underline{D})$.*
- *For every pair of objects $X, Y \in \underline{C}$ a map of sets $\Phi(X, Y) : \mathrm{Mor}_{\underline{C}}(X, Y) \to \mathrm{Mor}_{\underline{D}}(FX, FY)$ so that*
  1. *$\Phi(X, X)(id_X) = id_{F(X)}$ for all $X \in \mathrm{ob}(\underline{C})$, and*
  2. *$\Phi(X, Z)(g \circ f) = \Phi(Y, Z)(g) \circ \Phi(X, Y)(f)$ for all $f \in \mathrm{Mor}_{\underline{C}}(X, Y)$ and $g \in \mathrm{Mor}_{\underline{C}}(Y, Z)$.*

*Given a category $\underline{C}$, an **endofunctor** on $\underline{C}$ is any functor $F : \underline{C} \to \underline{C}$.*

Remark 4. In the interest of clarity, we will always refer to the pair (F, Φ) with a single letter F. See diagram (6) below for an example.

Example 2 (Scaling functor). For any $\lambda > 0$ we define an endofunctor $\sigma_\lambda : \underline{\mathcal{M}}^{gen} \to \underline{\mathcal{M}}^{gen}$ on objects by $\sigma_\lambda(X, d_X) = (X, \lambda \cdot d_X)$ and on morphisms by $\sigma_\lambda(f) = f$. One easily verifies that if f satisfies the conditions for being a morphism in $\underline{\mathcal{M}}^{gen}$ from (X, d_X) to (Y, d_Y), then it readily satisfies the conditions of being a morphism from $(X, \lambda \cdot d_X)$ to $(Y, \lambda \cdot d_Y)$. Clearly, σ_λ can also be regarded as an endofunctor in $\underline{\mathcal{M}}^{iso}$ and $\underline{\mathcal{M}}^{inj}$.

Similarly, we define a functor $s_\lambda : \underline{\mathcal{P}} \to \underline{\mathcal{P}}$ by setting $s_\lambda(X, \theta_X) = (X, \theta_X^\lambda)$, where $\theta_X^\lambda(r) = \theta_X(\frac{r}{\lambda})$.

5.2 Clustering Algorithms as Functors

The notion of categories, functors and functoriality provide useful framework for studying algorithms. One first defines a class of input objects $\mathcal{I}$ and also a class of output objects $\mathcal{O}$. Moreover, one associates to each of these classes a class of natural maps, the morphisms, between objects, making them into *categories* $\underline{\mathcal{I}}$ and $\underline{\mathcal{O}}$. For the problem of HC for example, the input class is the set of finite metric spaces and the output class is that of dendrograms. An algorithm is to be regarded as a functor between a category of input objects and a category of output objects.

An algorithm will therefore be a procedure that assigns to each $I \in \underline{\mathcal{I}}$ an output $O_I \in \underline{\mathcal{O}}$ with the further property that it respects relations between objects in the following sense. Assume $I, I' \in \underline{\mathcal{I}}$ such that there is a "natural map" $f : I \to I'$. Then, the algorithm has to have the property that the relation between O_I and $O_{I'}$ has to be represented by a natural map for output objects.

Remark 5. Assume that $\underline{\mathcal{I}}$ is such that $\mathrm{Mor}_{\underline{\mathcal{I}}}(X, Y) = \emptyset$ for all $X, Y \in \underline{\mathcal{I}}$ with $X \neq Y$. In this case, since there are no morphisms between input objects any functor $\mathfrak{A} : \underline{\mathcal{I}} \to \underline{\mathcal{O}}$ can be specified arbitrarily on each $X \in \mathrm{ob}(\underline{\mathcal{O}})$. It is much more interesting and arguably more useful to consider categories with non-empty morphism sets.

More precisely. We view any given clustering scheme as a procedure which takes as input a finite metric space (X, d_X), and delivers as output either an object in $\underline{\mathcal{C}}$ or $\underline{\mathcal{P}}$:

- **Standard clustering**: a pair (X, P_X) where P_X is a partition of X. Such a pair is an object in the category $\underline{\mathcal{C}}$.
- **Hierarchical clustering**: a pair (X, θ_X) where θ_X is a persistent set over X. Such a pair is an object in the category $\underline{\mathcal{P}}$.

The concept of **functoriality** refers to the additional condition that the clustering procedure should map a pair of input objects into a pair of output objects in a manner which is consistent with respect to the morphisms attached to the input and output spaces. When this happens, we say that the clustering scheme is **functorial**. This notion of consistency is made precise in Definition 9 and described by diagram (6). Let $\underline{\mathcal{M}}$ stand for any of $\underline{\mathcal{M}}^{gen}$, $\underline{\mathcal{M}}^{inj}$ or $\underline{\mathcal{M}}^{iso}$.

According to Definition 9, in order to view a standard clustering scheme as a functor $\mathfrak{C} : \underline{\mathcal{M}} \to \underline{\mathcal{C}}$ we need to specify:

(1) how it maps objects of $\underline{\mathcal{M}}$ (finite metric spaces) into objects of $\underline{\mathcal{C}}$, and
(2) how a morphism $f : (X, d_X) \to (Y, d_Y)$ between two objects (X, d_X) and (Y, d_Y) in the input category $\underline{\mathcal{M}}$ induces a map in the output category $\underline{\mathcal{C}}$, see diagram (6).

$$\begin{array}{ccc} (X,d_X) & \xrightarrow{f} & (Y,d_Y) \\ \downarrow \mathfrak{C} & & \downarrow \mathfrak{C} \\ (X,P_X) & \xrightarrow{\mathfrak{C}(f)} & (Y,P_Y) \end{array} \tag{5}$$

Similarly, in order to view a hierarchical clustering scheme as a functor $\mathfrak{H} : \underline{\mathcal{M}} \to \underline{\mathcal{P}}$ we need to specify:

(1) how it maps objects of $\underline{\mathcal{M}}$ (finite metric spaces) into objects of $\underline{\mathcal{P}}$, and
(2) how a morphism $f : (X, d_X) \to (Y, d_Y)$ between two objects (X, d_X) and (Y, d_Y) in the input category $\underline{\mathcal{M}}$ induces a map in the output category $\underline{\mathcal{P}}$, see diagram (6).

$$\begin{array}{ccc} (X,d_X) & \xrightarrow{f} & (Y,d_Y) \\ \downarrow \mathfrak{H} & & \downarrow \mathfrak{H} \\ (X,\theta_X) & \xrightarrow{\mathfrak{H}(f)} & (Y,\theta_Y) \end{array} \tag{6}$$

Precise constructions will be discussed ahead.

We have 3 possible "input" categories ordered by inclusion (4). The idea is that studying functoriality over a larger category will be more stringent/demanding than requiring functoriality over a smaller one. We will consider different clustering algorithms and study whether they are functorial over our choice of the input category. The least demanding one, $\underline{\mathcal{M}}^{iso}$ basically enforces that clustering schemes are not dependent on the way points are labeled.

We will describe uniqueness results for functoriality over the most stringent category $\underline{\mathcal{M}}^{gen}$, and also explain how relaxing the conditions imposed by the morphisms in $\underline{\mathcal{M}}^{gen}$, namely, by restricting ourselves to the smaller but intermediate category $\underline{\mathcal{M}}^{inj}$, one allows more functorial clustering algorithms.

5.3 Results for Standard Clustering

Let (X, d_X) be a finite metric space. For each $r \geq 0$ we define the equivalence relation $\sim_r$ on X given by $x \sim_r x'$ if and only if there exist $x_0, x_1, \ldots, x_k \in X$ with $x = x_0$, $x' = x_k$ and $d_X(x_i, x_{i+1}) \leq r$ for all $i = 0, 1, \ldots, k-1$.

Definition 10. *For each $\delta > 0$ we define the* ***Vietoris-Rips clustering functor***

$$\mathfrak{R}_\delta : \underline{\mathcal{M}}^{gen} \to \underline{\mathcal{C}}$$

as follows. For a finite metric space (X, d_X), we set $\mathfrak{R}_\delta(X, d_X)$ to be $(X, P_X(\delta))$, where $P_X(\delta)$ is the partition of X associated to the equivalence relation $\sim_\delta$. We define how $\mathfrak{R}_\delta$ acts on maps $f : (X, d_X) \to (Y, d_Y)$: $\mathfrak{R}_\delta(f)$ is simply the set map f regarded as a morphism from $(X, P_X(\delta))$ to $(Y, P_Y(\delta))$ in $\underline{\mathcal{C}}$.

The Vietoris-Rips functor is actually just **single linkage clustering** as it is well known, see [15,18].

By restricting $\mathfrak{R}_\delta$ to the subcategories $\underline{\mathcal{M}}^{iso}$ and $\underline{\mathcal{M}}^{inj}$, we obtain functors $\mathfrak{R}_\delta^{iso} : \underline{\mathcal{M}}^{iso} \to \underline{\mathcal{C}}$ and $\mathfrak{R}_\delta^{inj} : \underline{\mathcal{M}}^{inj} \to \underline{\mathcal{C}}$. We will denote all these functors by $\mathfrak{R}_\delta$ when there is no ambiguity.

It can be seen [19] that the Vietoris-Rips functor is surjective: Among the desirable conditions singled out by Kleinberg [51], one has that of *surjectivity* (which he referred to as "richness"). Given a finite set X and $P_X \in \mathcal{P}(X)$, surjectivity calls for the existence of a metric d_X on X such that $\mathfrak{R}_\delta(X, d_X) = (X, P_X)$.

For $\underline{\mathcal{M}}$ being any one of our choices $\underline{\mathcal{M}}^{iso}$, $\underline{\mathcal{M}}^{inj}$ or $\underline{\mathcal{M}}^{gen}$, a clustering functor in this context will be denoted by $\mathfrak{C} : \underline{\mathcal{M}} \to \underline{\mathcal{C}}$. **Excisiveness** of a clustering functor refers to the property that once a finite metric space has been partitioned by the clustering procedure, it should not be further split by subsequent applications of the same algorithm.

Definition 11 (Excisive clustering functors). *We say that a clustering functor* $\mathfrak{C}$ *is* ***excisive*** *if for all* $(X, d_X) \in \mathrm{ob}(\underline{\mathcal{M}})$*, if we write* $\mathfrak{C}(X, d_X) = (X, \{X_\alpha\}_{\alpha \in A})$*, then*

$$\mathfrak{C}\left(X_\alpha, d_{X|_{X_\alpha \times X_\alpha}}\right) = (X_\alpha, \{X_\alpha\}) \text{ for all } \alpha \in A.$$

It can be seen that the Vietoris-Rips functor is excisive.

However, there exist non-excisive clustering functors in $\underline{\mathcal{M}}^{inj}$.

*Example 3 (***A non-excisive functor in*** $\underline{\mathcal{M}}^{inj}$***).** For each finite metric space X let $\eta_X := (\mathbf{sep}\,(X))^{-1}$. Consider the clustering functor $\widehat{\mathfrak{R}} : \underline{\mathcal{M}}^{inj} \to \underline{\mathcal{C}}$ defined as follows: for a finite metric space (X, d_X), we define $\widehat{\mathfrak{R}}(X, d_X)$ to be $(X, \widehat{P}_X)$, where $\widehat{P}_X$ is the partition of X associated to the equivalence relation $\sim_{\eta_X}$ on X. That $\widehat{\mathfrak{R}}$ is a functor follows from the fact that whenever $\phi \in \mathrm{Mor}_{\underline{\mathcal{M}}^{inj}}(X, Y)$ and $x \sim_{\eta_X} x'$, then $\phi(x) \sim_{\eta_Y} \phi(x')$.

Now, the functor $\widehat{\mathfrak{R}}$ is **not excisive** in general. An explicit example is the following: Consider the metric space (X, d_X) depicted in Figure 13, where the metric is given by the graph metric on the underlying graph. Note that $\mathbf{sep}\,(X) = 1/2$ and thus $\eta_X = 2$. We then find that $\widehat{\mathfrak{R}}(X, d_X) = (X, \{\{A, B, C\}, \{D, E\}\})$. Let $(Y, d_Y) = \left(\{A, B, C\}, \left(\begin{smallmatrix} 0 & 2 & 3 \\ 2 & 0 & 1 \\ 3 & 1 & 0 \end{smallmatrix}\right)\right)$. Then, $\mathbf{sep}\,(Y) = 1$ and hence $\eta_Y = 1$. Therefore,

$$\widehat{\mathfrak{R}}\left(\{A, B, C\}, \left(\begin{smallmatrix} 0 & 2 & 3 \\ 2 & 0 & 1 \\ 3 & 1 & 0 \end{smallmatrix}\right)\right) = (\{A, B, C\}, \{A, \{B, C\}\}),$$

and we see that $\{A, B, C\}$ gets further partitioned by $\widehat{\mathfrak{R}}$.

It is interesting to point out that the similar constructions of a non-excisive functor in $\underline{\mathcal{M}}^{gen}$ would not work, see [19].

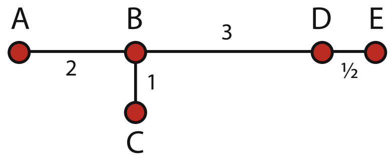


Fig. 13. Metric space used to prove that the functor $\widehat{\mathfrak{R}} : \underline{\mathcal{M}}^{inj} \to \underline{\mathcal{C}}$ is not excisive. The metric is given by the graph distance on the graph.

The case of $\underline{\mathcal{M}}^{iso}$ One can easily describe all $\underline{\mathcal{M}}^{iso}$-functorial clustering schemes. Let $\mathcal{I}$ denote the collection of all isometry classes of finite metric spaces. For each $\zeta \in \mathcal{I}$ let (X_ζ, d_{X_ζ}) denote an element of the class ζ, G_ζ the isometry group of (X_ζ, d_{X_ζ}), and Ξ_ζ the set of all fixed points of the action of G_ζ on $\mathcal{P}(X_\zeta)$.

Theorem 9 (Classification of $\underline{\mathcal{M}}^{iso}$-functorial clustering schemes, [19]). *Any $\underline{\mathcal{M}}^{iso}$-functorial clustering scheme determines a choice of $p_\zeta \in \Xi_\zeta$ for each $\zeta \in \mathcal{I}$, and conversely, a choice of p_ζ for each $\zeta \in \mathcal{I}$ determines an $\underline{\mathcal{M}}^{iso}$-functorial scheme.*

Representable Clustering Functors. In what follows, $\underline{\mathcal{M}}$ is either of $\underline{\mathcal{M}}^{inj}$ or $\underline{\mathcal{M}}^{gen}$. For each $\delta > 0$ the Vietoris-Rips functor $\mathfrak{R}_\delta : \underline{\mathcal{M}} \to \underline{\mathcal{C}}$ can be described in an alternative way. A first trivial observation is that the condition that $x, x' \in X$ satisfy $d_X(x, x') \leq \delta$ is equivalent to requiring the existence of a map $f \in \mathrm{Mor}_{\underline{\mathcal{M}}}(\Delta_2(\delta), X)$ with $\{x, x'\} \subset \mathrm{Im}(f)$. Using this, we can reformulate the condition that $x \sim_\delta x'$ by the requirement that there exist $z_0, z_1, \ldots, z_k \in X$ with $z_0 = x$, $z_k = x'$, and $f_1, f_2, \ldots, f_k \in \mathrm{Mor}_{\underline{\mathcal{M}}}(\Delta_2(\delta), X)$ with $\{x_{i-1}, x_i\} \subset \mathrm{Im}(f_i)$ $\forall i = 1, 2, \ldots, k$. Informally, this points to the interpretation that $\{\Delta_2(\delta)\}$ is the "parameter" in a "generative model" for $\mathfrak{R}_\delta$.

This suggests considering more general clustering functors constructed in the following manner. Let Ω be any fixed collection of finite metric spaces. Define a clustering functor

$$\mathfrak{C}^\Omega : \underline{\mathcal{M}} \to \underline{\mathcal{C}}$$

as follows: let $(X, d) \in \mathrm{ob}(\underline{\mathcal{M}})$ and write $\mathfrak{C}^\Omega(X, d) = (X, \{X_\alpha\}_{\alpha \in A})$. One declares that points x and x' belong to the same block X_α if and only if there exist

- a sequence of points $z_0, \ldots, z_k \in X$ with $z_0 = x$ and $z_k = x'$,
- a sequence of metric spaces $\omega_1, \ldots, \omega_k \in \Omega$ and
- for each $i = 1, \ldots, k$, pairs of points $(\alpha_i, \beta_i) \in \omega_i$ and morphisms $f_i \in \mathrm{Mor}_{\underline{\mathcal{M}}}(w_i, X)$ s.t. $f_i(\alpha_i) = z_{i-1}$ and $f_i(\beta_i) = z_i$.

Also, we declare that $\mathfrak{C}^\Omega(f) = f$ on morphisms f. Notice that above one can assume that $z_0, z_1, \ldots, z_k$ all belong to X_α.

Definition 12. *We say that a clustering functor $\mathfrak{C}$ is **representable** whenever there exists a collection of finite metric spaces Ω such that $\mathfrak{C} = \mathfrak{C}^\Omega$. In this case, we say that $\mathfrak{C}$ is **represented** by Ω. We say that $\mathfrak{C}$ is **finitely representable** whenever $\mathfrak{C} = \mathfrak{C}^\Omega$ for some* finite *collection of finite metric spaces Ω.*

As we saw above, the Vietoris-Rips functor $\mathfrak{R}_\delta$ is (finitely) represented by $\{\Delta_2(\delta)\}$.

Representability and excisiveness. Notice that excisiveness is an axiomatic statement whereas representability asserts existence of generative model for the clustering functor, and interestingly they are equivalent.

Theorem 10 ([19]). *Let $\underline{\mathcal{M}}$ be either of $\underline{\mathcal{M}}^{inj}$ or $\underline{\mathcal{M}}^{gen}$. Then any clustering functor on $\underline{\mathcal{M}}$ is excisive if and only if it is representable.*

A factorization theorem. For a given collection Ω of finite metric spaces let

$$\mathfrak{T}^\Omega : \underline{\mathcal{M}} \to \underline{\mathcal{M}} \tag{7}$$

be the endofunctor that assigns to each finite metric space (X, d_X) the metric space (X, d_X^Ω) with the same underlying set and metric d_X^Ω given by the maximal metric bounded above by W_X^Ω, where $W_X^\Omega : X \times X \to \mathbb{R}_+$ is given by

$$(x,x') \mapsto \inf\big\{\lambda > 0 |\, \exists w \in \Omega \,\text{and}\, \phi \in \mathrm{Mor}_{\underline{\mathcal{M}}}(\lambda \cdot \omega, X) \,\text{with}\, \{x,x'\} \subset \mathrm{Im}(\phi)\big\}, \tag{8}$$

for $x \neq x'$, and by 0 on $\mathrm{diag}(X \times X)$. Above we assume that the inf over the empty set equals $+\infty$. Note that $W_X^\Omega(x,x') < \infty$ for all $x, x' \in X$ as long as $|\omega| \leq |X|$ for some $\omega \in \Omega$. Also, $W_X^\Omega(x,x') = \infty$ for all $x \neq x'$ when $|X| < \inf\{|\omega|, \omega \in \Omega\}$.

Theorem 11 ([19]). *Let $\underline{\mathcal{M}}$ be either $\underline{\mathcal{M}}^{gen}$ or $\underline{\mathcal{M}}^{inj}$ and $\mathfrak{C}$ be any $\underline{\mathcal{M}}$-functorial finitely representable clustering functor represented by some $\Omega \subset \mathcal{M}$. Then, $\mathfrak{C} = \mathfrak{R}_1 \circ \mathfrak{T}^\Omega$.*

This theorem implies that all finitely representable clustering functors in $\underline{\mathcal{M}}^{gen}$ and $\underline{\mathcal{M}}^{inj}$ arise as the composition of the Vietoris-Rips functor with a functor that changes the metric.

A Uniqueness theorem for $\underline{\mathcal{M}}^{gen}$. In $\underline{\mathcal{M}}^{gen}$ clustering functors are very restricted, as reflected by the following theorem.

Theorem 12 ([19]). *Assume that $\mathfrak{C} : \underline{\mathcal{M}}^{gen} \to \underline{\mathcal{C}}$ is a clustering functor for which there exists $\delta_{\mathfrak{C}} > 0$ with the property that*

- *$\mathfrak{C}(\Delta_2(\delta))$ is in one piece for all $\delta \in [0, \delta_{\mathfrak{C}}]$, and*
- *$\mathfrak{C}(\Delta_2(\delta))$ is in two pieces for all $\delta > \delta_{\mathfrak{C}}$.*

Then, $\mathfrak{C}$ is the Vietoris-Rips functor with parameter $\delta_{\mathfrak{C}}$. i.e. $\mathfrak{C} = \mathfrak{R}_{\delta_{\mathfrak{C}}}$.

Recall that the Vietoris-Rips functor is excisive.

Scale invariance in $\underline{\mathcal{M}}^{gen}$ and $\underline{\mathcal{M}}^{inj}$. It is interesting to consider the effect of imposing Kleinberg's scale invariance axiom on $\underline{\mathcal{M}}^{gen}$-functorial and $\underline{\mathcal{M}}^{inj}$-functorial clustering schemes. It turns out that in $\underline{\mathcal{M}}^{gen}$ there are only two possible clustering schemes enjoying scale invariance, which turn out to be the trivial ones:

Theorem 13 ([19]). *Let $\mathfrak{C} : \underline{\mathcal{M}}^{gen} \to \underline{\mathcal{C}}$ be a clustering functor s.t. $\mathfrak{C} \circ \sigma_\lambda = \mathfrak{C}$ for all $\lambda > 0$. Then, either*

- *$\mathfrak{C}$ assigns to each finite metric space X the partition of X into singletons, or*
- *$\mathfrak{C}$ assigns to each finite metric the partition with only one block.*

By refining the proof of the previous theorem, we find that the behavior of any $\underline{\mathcal{M}}^{inj}$-functorial clustering functor is also severely restricted [19].

5.4 Results for Hierarchical Clustering

*Example 4 (**A hierarchical version of the Vietoris-Rips functor**).* We define a functor

$$\mathfrak{R} : \underline{\mathcal{M}}^{gen} \to \underline{\mathcal{P}}$$

as follows. For a finite metric space (X, d_X), we define (X, d_X) to be the persistent set $(X, \theta_X^{\mathrm{VR}})$, where $\theta_X^{\mathrm{VR}}(r)$ is the partition associated to the equivalence relation $\sim_r$. This is clearly an object in $\underline{\mathcal{P}}$. We also define how $\mathfrak{R}$ acts on maps $f : (X, d_X) \to (Y, d_Y)$: The value of $\mathfrak{R}(f)$ is simply the set map f regarded as a morphism from $(X, \theta_X^{\mathrm{VR}})$ to $(Y, \theta_Y^{\mathrm{VR}})$ in $\underline{\mathcal{P}}$. That it is a morphism in $\underline{\mathcal{P}}$ is easy to check.

Clearly, this functor implements the hierarchical version of single linkage clustering in the sense that for each $\delta \geq 0$, if one writes $\mathfrak{R}_\delta(X, d_X) = (X, P_X(\delta))$, then $P_X(\delta) = \theta_X^{\mathrm{VR}}(\delta)$.

Functoriality over $\underline{\mathcal{M}}^{gen}$: A uniqueness theorem. We have a theorem of the same flavor as the main theorem of [51], except that one obtains existence and uniqueness on $\underline{\mathcal{M}}^{gen}$ instead of impossibility in our context.

Theorem 14 ([15]). *Let $\mathfrak{H} : \underline{\mathcal{M}}^{gen} \to \underline{\mathcal{P}}$ be a hierarchical clustering functor which satisfies the following conditions.*

(I) *Let $\alpha : \underline{\mathcal{M}}^{gen} \to \underline{Sets}$ and $\beta : \underline{\mathcal{P}} \to \underline{Sets}$ be the forgetful functors $(X, d_X) \to X$ and $(X, \theta_X) \to X$, which forget the metric and persistent set respectively, and only "remember" the underlying sets X. Then we assume that $\beta \circ \mathfrak{H} = \alpha$. This means that the underlying set of the persistent set associated to a metric space is just the underlying set of the metric space.*

(II) *For $\delta \geq 0$ let $\Delta_2(\delta) = (\{p, q\}, \left(\begin{smallmatrix} 0 & \delta \\ \delta & 0 \end{smallmatrix}\right))$ denote the two point metric space with underlying set $\{p, q\}$, and where dist$(p, q) = \delta$. Then $\mathfrak{H}(\Delta_2(\delta))$ is the persistent set $(\{p, q\}, \theta_{\Delta_2(\delta)})$ whose underlying set is $\{p, q\}$ and where $\theta_{\Delta_2(\delta)}(t)$ is the partition with one element blocks when $t < \delta$ and is the partition with a single two point block when $t \geq \delta$.*

(III) *Write $\mathfrak{H}(X, d_X) = (X, \theta^{\mathfrak{H}})$, then for any $t <$ **sep**(X), the partition $\theta^{\mathfrak{H}}(t)$ is the discrete partition with one element blocks.*

Then $\mathfrak{H}$ is equal to the functor $\mathfrak{R}$.

Extensions. There are extensions of the ideas described in previous sections that induce functorial clustering algorithms that are more sensitive to density, see [16,19].

6 Discussion

Imposing metric and or category structures on collections of datasets is useful. Doing this enables organizing the landscape composed by several algorithms commonly used in data analysis. With this in mind is possible to reason about the well posedness of some of these algorithms, and furthermore, one is able to infer new algorithms for solving data and shape analysis tasks.

Acknowledgements. This work was supported by DARPA grant number HR0011-05-1-0007 and ONR grant number N00014-09-1-0783.

References

1. Ankerst, M., Kastenmüller, G., Kriegel, H.-P., Seidl, T.: 3d shape histograms for similarity search and classification in spatial databases. In: Güting, R.H., Papadias, D., Lochovsky, F.H. (eds.) SSD 1999. LNCS, vol. 1651, pp. 207–226. Springer, Heidelberg (1999)
2. Asimov, D.: The grand tour: a tool for viewing multidimensional data. SIAM J. Sci. Stat. Comput. 6, 128–143 (1985)
3. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. 24(4), 509–522 (2002)
4. Berchtold, S.: Geometry-based Search of Similar Parts. PhD thesis. University of Munich, Germany (1998)
5. Boutin, M., Kemper, G.: On reconstructing n-point configurations from the distribution of distances or areas. Adv. in Appl. Math. 32(4), 709–735 (2004)
6. Bowman, G.R., Huang, X., Yao, Y., Sun, J., Carlsson, G., Guibas, L.J., Pande, V.S.: Structural insight into rna hairpin folding intermediates. Journal of the American Chemical Society (2008)
7. Brinkman, D., Olver, P.J.: Invariant histograms. University of Minnesota. Preprint (2010)
8. Bronstein, A.M., Bronstein, M.M., Kimmel, R.: Topology-invariant similarity of nonrigid shapes. Intl. Journal of Computer Vision (IJCV) 81(3), 281–301 (2009)
9. Bronstein, A.M., Bronstein, M.M., Kimmel, R., Mahmoudi, M., Sapiro, G.: A gromov-hausdorff framework with diffusion geometry for topologically-robust nonrigid shape matching (Submitted)
10. Bronstein, A., Bronstein, M., Bruckstein, A., Kimmel, R.: Partial similarity of objects, or how to compare a centaur to a horse. International Journal of Computer Vision
11. Bronstein, A.M., Bronstein, M.M., Kimmel, R.: Efficient computation of isometry-invariant distances between surfaces. SIAM Journal on Scientific Computing 28(5), 1812–1836 (2006)
12. Bronstein, A.M., Bronstein, M.M., Kimmel, R.: Calculus of nonrigid surfaces for geometry and texture manipulation. IEEE Trans. Vis. Comput. Graph. 13(5), 902–913 (2007)

13. Burago, D., Burago, Y., Ivanov, S.: A Course in Metric Geometry. AMS Graduate Studies in Math, vol. 33. American Mathematical Society, Providence (2001)
14. Bustos, B., Keim, D.A., Saupe, D., Schreck, T., Vranić, D.V.: Feature-based similarity search in 3d object databases. ACM Comput. Surv. 37(4), 345–387 (2005)
15. Carlsson, G., Mémoli, F.: Persistent Clustering and a Theorem of J. Kleinberg. ArXiv e-prints (August 2008)
16. Carlsson, G., Mémoli, F.: Multiparameter clustering methods. Technical report, technical report (2009)
17. Carlsson, G.: Topology and data. Bull. Amer. Math. Soc. 46, 255–308 (2009)
18. Carlsson, G., Mémoli, F.: Characterization, stability and convergence of hierarchical clustering methods. Journal of Machine Learning Research 11, 1425–1470 (2010)
19. Carlsson, G., Mémoli, F.: Classifying clustering schemes. CoRR, abs/1011.5270 (2010)
20. Chazal, F., Cohen-Steiner, D., Guibas, L., Mémoli, F., Oudot, S.: Gromov-Hausdorff stable signatures for shapes using persistence. In: Proc. of SGP (2009)
21. Clarenz, U., Rumpf, M., Telea, A.: Robust feature detection and local classification for surfaces based on moment analysis. IEEE Transactions on Visualization and Computer Graphics 10 (2004)
22. Coifman, R.R., Lafon, S.: Diffusion maps. Applied and Computational Harmonic Analysis 21(1), 5–30 (2006)
23. Cox, T.F., Cox, M.A.A.: Multidimensional scaling. Monographs on Statistics and Applied Probability, vol. 59. Chapman & Hall, London (1994) With 1 IBM-PC floppy disk (3.5 inch, HD)
24. d'Amico, M., Frosini, P., Landi, C.: Natural pseudo-distance and optimal matching between reduced size functions. Technical Report 66, DISMI, Univ. degli Studi di Modena e Reggio Emilia, Italy (2005)
25. d'Amico, M., Frosini, P., Landi, C.: Using matching distance in size theory: A survey. IJIST 16(5), 154–161 (2006)
26. Davies, E.B.: Heat kernels in one dimension. Quart. J. Math. Oxford Ser. (2) 44(175), 283–299 (1993)
27. Edelsbrunner, H., Harer, J.: Computational Topology - an Introduction. American Mathematical Society, Providence (2010)
28. Elad (Elbaz), A., Kimmel, R.: On bending invariant signatures for surfaces. IEEE Trans. Pattern Anal. Mach. Intell. 25(10), 1285–1295 (2003)
29. Frosini, P.: A distance for similarity classes of submanifolds of Euclidean space. Bull. Austral. Math. Soc. 42(3), 407–416 (1990)
30. Frosini, P.: Omotopie e invarianti metrici per sottovarieta di spazi euclidei (teoria della taglia). PhD thesis. University of Florence, Italy (1990)
31. Frosini, P., Mulazzani, M.: Size homotopy groups for computation of natural size distances. Bull. Belg. Math. Soc. Simon Stevin 6(3), 455–464 (1999)
32. Gelfand, N., Mitra, N.J., Guibas, L.J., Pottmann, H.: Robust global registration. In: SGP 2005: Proceedings of the Third Eurographics Symposium on Geometry Processing, p. 197. Eurographics Association, Aire-la-Ville (2005)
33. Ghrist, R.: Barcodes: The persistent topology of data. Bulletin-American Mathematical Society 45(1), 61 (2008)
34. Grigorescu, C., Petkov, N.: Distance sets for shape filters and shape recognition. IEEE Transactions on Image Processing 12(10), 1274–1286 (2003)
35. Gromov, M.: Metric structures for Riemannian and non-Riemannian spaces. Progress in Mathematics, vol. 152. Birkhäuser Boston Inc., Boston (1999)

36. Ben Hamza, A., Krim, H.: Geodesic object representation and recognition. In: Nyström, I., Sanniti di Baja, G., Svensson, S. (eds.) DGCI 2003. LNCS, vol. 2886, pp. 378–387. Springer, Heidelberg (2003)
37. Hartigan, J.A.: Statistical theory in clustering. J. Classification 2(1), 63–76 (1985)
38. Hastie, T., Stuetzle, W.: Principal curves. Journal of the American Statistical Association 84(406), 502–516 (1989)
39. Hilaga, M., Shinagawa, Y., Kohmura, T., Kunii, T.L.: Topology matching for fully automatic similarity estimation of 3d shapes. In: SIGGRAPH 2001: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, pp. 203–212. ACM, New York (2001)
40. Holm, L., Sander, C.: Protein structure comparison by alignment of distance matrices. Journal of Molecular Biology 233(1), 123–138 (1993)
41. Huang, Q.-X., Adams, B., Wicke, M., Guibas, L.J.: Non-rigid registration under isometric deformations. Comput. Graph. Forum 27(5), 1449–1457 (2008)
42. Huber, P.J.: Projection pursuit. The Annals of Statistics 13(2), 435–525 (1985)
43. Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J.: Comparing images using the Hausdorff distance. IEEE Transactions on Pattern Analysis and Machine Intelligence 15(9) (1993)
44. Inselberg, A.: Parallel Coordinates: Visual Multidimensional Geometry and Its Applications. Springer-Verlag New York, Inc., Secaucus (2009)
45. Ion, A., Artner, N.M., Peyre, G., Marmol, S.B.L., Kropatsch, W.G., Cohen, L.: 3d shape matching by geodesic eccentricity. In: IEEE Computer Society Conference on, Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2008, pp. 1–8 (June 2008)
46. Jain, A.K., Dubes, R.C.: Algorithms for clustering data. Prentice Hall Advanced Reference Series. Prentice Hall Inc., Englewood Cliffs (1988)
47. Janowitz, M.F.: An order theoretic model for cluster analysis. SIAM Journal on Applied Mathematics 34(1), 55–72 (1978)
48. Jardine, N., Sibson, R.: Mathematical taxonomy. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons Ltd., London (1971)
49. Johnson, A.: Spin-Images: A Representation for 3-D Surface Matching. PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA (August 1997)
50. Kastenmüller, G., Kriegel, H.P., Seidl, T.: Similarity search in 3d protein databases. In: Proc. GCB (1998)
51. Kleinberg, J.M.: An impossibility theorem for clustering. In: Becker, S., Thrun, S., Obermayer, K. (eds.) NIPS, pp. 446–453. MIT Press, Cambridge (2002)
52. Koppensteiner, W.A., Lackner, P., Wiederstein, M., Sippl, M.J.: Characterization of novel proteins based on known protein structures. Journal of Molecular Biology 296(4), 1139–1152 (2000)
53. Lafon, S.: Diffusion Maps and Geometric Harmonics. PhD thesis, Yale University (2004)
54. Le, T.M., Mémoli, F.: Local scales of embedded curves and surfaces. preprint (2010)
55. Ling, H., Jacobs, D.W.: Using the inner-distance for classification of articulated shapes. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 719–726 (2005)
56. Lu, C.E., Latecki, L.J., Adluru, N., Yang, X., Ling, H.: Shape guided contour grouping with particle filters. In: IEEE 12th International Conference on, Computer Vision 2009, pp. 2288–2295. IEEE, Los Alamitos (2009)
57. Lane, S.M.: Categories for the working mathematician, 2nd edn. Graduate Texts in Mathematics, vol. 5. Springer, New York (1998)

58. Manay, S., Cremers, D., Hong, B.W., Yezzi, A.J., Soatto, S.: Integral invariants for shape matching 28(10), 1602–1618 (2006)
59. Mémoli, F.: Gromov-Hausdorff distances in Euclidean spaces. In: IEEE Computer Society Conference on, Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2008, pp. 1–8 (June 2008)
60. Mémoli, F.: Gromov-wasserstein distances and the metric approach to object matching. In: Foundations of Computational Mathematics, pp. 1–71 (2011) 10.1007/s10208-011-9093-5
61. Mémoli, F.: Some properties of gromov-hausdorff distances. Technical report, Department of Mathematics. Stanford University (March 2011)
62. Mémoli, F.: A spectral notion of Gromov-Wasserstein distances and related methods. Applied and Computational Mathematics 30, 363–401 (2011)
63. Mémoli, F., Sapiro, G.: Comparing point clouds. In: SGP 2004: Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing, pp. 32–40. ACM, New York (2004)
64. Mémoli, F., Sapiro, G.: A theoretical and computational framework for isometry invariant recognition of point cloud data. Found. Comput. Math. 5(3), 313–347 (2005)
65. Nicolau, M., Levine, A.J., Carlsson, G.: Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. Proceedings of the National Academy of Sciences 108(17), 7265–7270 (2011)
66. Norris, J.R.: Heat kernel asymptotics and the distance function in Lipschitz Riemannian manifolds. Acta. Math. 179(1), 79–103 (1997)
67. Olver, P.J.: Joint invariant signatures. Foundations of computational mathematics 1(1), 3–68 (2001)
68. Osada, R., Funkhouser, T., Chazelle, B., Dobkin, D.: Shape distributions. ACM Trans. Graph. 21(4), 807–832 (2002)
69. Pottmann, H., Wallner, J., Huang, Q., Yang, Y.-L.: Integral invariants for robust geometry processing. Comput. Aided Geom. Design (2008) (to appear)
70. Raviv, D., Bronstein, A.M., Bronstein, M.M., Kimmel, R.: Symmetries of non-rigid shapes. In: IEEE 11th International Conference on, Computer Vision, ICCV 2007, October 14-21, pp. 1–7 (2007)
71. Reeb, G.: Sur les points singuliers d'une forme de Pfaff complètement intégrable ou d'une fonction numérique. C. R. Acad. Sci. Paris 222, 847–849 (1946)
72. Reuter, M., Wolter, F.-E., Peinecke, N.: Laplace-spectra as fingerprints for shape matching. In: SPM 2005: Proceedings of the 2005 ACM Symposium on Solid and Physical Modeling, pp. 101–106. ACM Press, New York (2005)
73. Reuter, M., Wolter, F.-E., Peinecke, N.: Laplace-Beltrami spectra as "Shape-DNA" of surfaces and solids. Computer-Aided Design 38(4), 342–366 (2006)
74. Roweis, S.T., Saul, L.K.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science 290(5500), 2323–2326 (2000)
75. Ruggeri, M., Saupe, D.: Isometry-invariant matching of point set surfaces. In: Proceedings Eurographics 2008 Workshop on 3D Object Retrieval (2008)
76. Rustamov, R.M.: Laplace-beltrami eigenfunctions for deformation invariant shape representation. In: Symposium on Geometry Processing, pp. 225–233 (2007)
77. Sakai, T.: Riemannian geometry. Translations of Mathematical Monographs, vol. 149. American Mathematical Society, Providence (1996)
78. Semple, C., Steel, M.: Phylogenetics. Oxford Lecture Series in Mathematics and its Applications, vol. 24. Oxford University Press, Oxford (2003)

79. Shi, Y., Thompson, P.M., de Zubicaray, G.I., Rose, S.E., Tu, Z., Dinov, I., Toga, A.W.: Direct mapping of hippocampal surfaces with intrinsic shape context. NeuroImage 37(3), 792–807 (2007)
80. Singh, G., Mémoli, F., Carlsson, G.: Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition, pp. 91–100. Eurographics Association, Prague (2007)
81. Singh, G., Memoli, F., Ishkhanov, T., Sapiro, G., Carlsson, G., Ringach, D.L.: Topological analysis of population activity in visual cortex. J. Vis. 8(8), 1–18 (2008)
82. Stuetzle, W.: Estimating the cluster type of a density by analyzing the minimal spanning tree of a sample. J. Classification 20(1), 25–47 (2003)
83. Sturm, K.-T.: On the geometry of metric measure spaces. I. Acta. Math. 196(1), 65–131 (2006)
84. Sun, J., Ovsjanikov, M., Guibas, L.: A concise and provably informative multi-scale signature based on heat diffusion. In: SGP (2009)
85. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science 290(5500), 2319–2323 (2000)
86. Thureson, J., Carlsson, S.: Appearance based qualitative image description for object class recognition. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3022, pp. 518–529. Springer, Heidelberg (2004)
87. Tsuchida, T.: Long-time asymptotics of heat kernels for one-dimensional elliptic operators with periodic coefficients. Proc. Lond. Math. Soc (3) 97(2), 450–476 (2008)
88. Verri, A., Uras, C., Frosini, P., Ferri, M.: On the use of size functions for shape analysis. Biological cybernetics 70(2), 99–107 (1993)
89. Villani, C.: Topics in optimal transportation. Graduate Studies in Mathematics, vol. 58. American Mathematical Society, Providence (2003)
90. von Luxburg, U., Ben-David, S.: Towards a statistical theory of clustering. presented at the pascal workshop on clustering, london. Technical report, Presented at the PASCAL Workshop on Clustering, London (2005)
91. Zomorodian, A., Carlsson, G.: Computing persistent homology. In: SCG 2004: Proceedings of the Twentieth Annual Symposium on Computational Geometry, pp. 347–356. ACM, New York (2004)