An observer-oriented approach to topological data analysis

Part 1: From comparing subsets of $\mathbb{R}^n$ to studying metric spaces of functions

Patrizio Frosini

Department of Mathematics and ARCES, University of Bologna
patrizio.frosini@unibo.it

Second School/Conference in TDA,
Stochastic Topology and related topics
Querétaro, 7-11 December 2015

Outline

Our basic questions

Assumptions in our model

Mathematical setting and theoretical results

Experiments

Our basic questions

Assumptions in our model

Mathematical setting and theoretical results

Experiments

Our basic questions

We are interested in these questions:

- Is there a general metric model to compare data in TDA?
- What is the role of the observer in this comparison?
- How could we approximate the observer's judgement by means of a computable metric?

Our basic questions

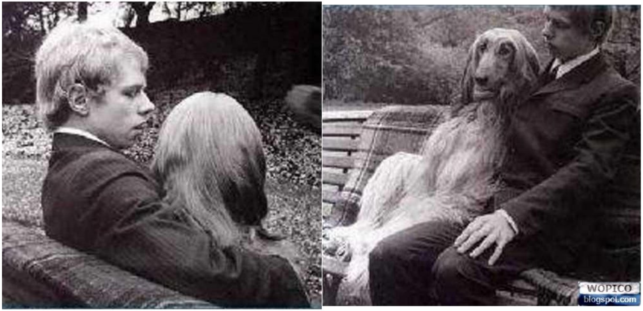
Assumptions in our model

Mathematical setting and theoretical results

Experiments

Assumptions in our model

Truth often depends on the observer's perspective:

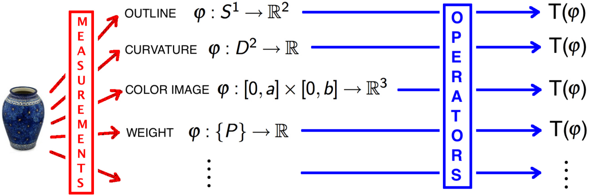


Multiple perspectives are unavoidable! (→ Sergio's talk)

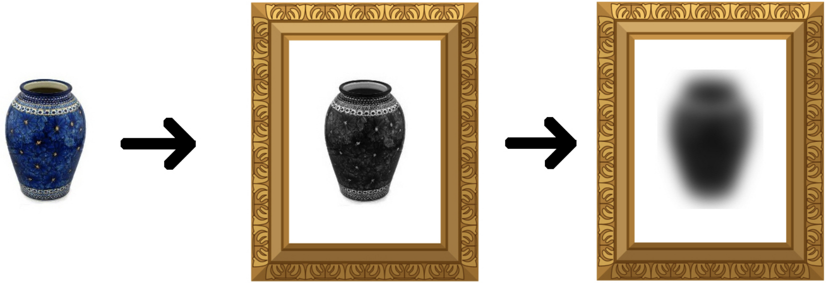
Assumptions in our model

We will make these assumptions:

1. No object can be studied directly. Any object is knowable just through acts of measurement made by an observer.
2. Any act of measurement can be represented as a function defined on a topological space.
3. The observer usually acquires measurement data by applying operators to the functions describing them.


MEASUREMENTS
OUTLINE $\varphi : S^1 \to \mathbb{R}^2$
CURVATURE $\varphi : D^2 \to \mathbb{R}$
COLOR IMAGE $\varphi : [0,a] \times [0,b] \to \mathbb{R}^3$
WEIGHT $\varphi : \{P\} \to \mathbb{R}$
$\vdots$
OPERATORS
$T(\varphi)$
$T(\varphi)$
$T(\varphi)$
$T(\varphi)$
$\vdots$

An example of operator



GRAYSCALE IMAGE

$\varphi : [0,a] \times [0,b] \longrightarrow \mathbb{R}$ $\quad$ $T(\varphi)$ = CONVOLUTION OF φ

Choice of the operators

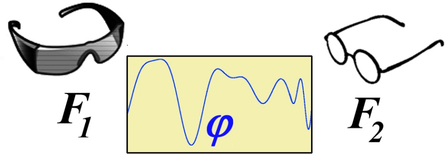
- The observer cannot usually choose the functions representing the measurement data, but can often choose the operators that will be applied to those functions.
- The choice of the operators reflects the invariances that are relevant for the observer.
- In some sense we could state that the observer can be represented as a collection of (suitable) operators, endowed with the invariance he/she has chosen.

In this talk we will confine ourselves to examine the case of operators that act on a space Φ of continuous functions and take Φ to itself.

From comparing sets in $\mathbb{R}^n$ to comparing functions

Instead of directly focusing on the objects we are interested in, we focus on the filtering functions describing the measurements we make on them, and on the “glasses” that we use “to observe” the filtering functions. In our approach, these “glasses” are G-operators which act on the filtering functions.
These operators represent the observer’s perspective.

In some sense, the family of operators defines the observer.



Our basic questions

Assumptions in our model

Mathematical setting and theoretical results

Experiments

Natural pseudo-distance associated with a group G

First of all we need a definition allowing us to formalize the comparison of data in our model.

Definition

Let X be a compact space. Let G be a subgroup of the group $\mathrm{Homeo}(X)$ of all homeomorphisms $f: X \to X$. The pseudo-distance $d_G : C^0(X,\mathbb{R}) \times C^0(X,\mathbb{R}) \to \mathbb{R}$ defined by setting

$$d_G(\varphi,\psi) = \inf_{g\in G} \max_{x\in X} |\varphi(x) - \psi(g(x))|$$

is called the natural pseudo-distance associated with the group G.

In plain words, the definition of d_G is based on the attempt of finding the best correspondence between the functions φ, ψ by means of homeomorphisms in G.

G-invariant non-expansive operators

The natural pseudo-distance d_G represents our ground truth.

Unfortunately, d_G is difficult to compute. This is also a consequence of the fact that we can easily find subgroups G of $\mathrm{Homeo}(X)$ that cannot be approximated with arbitrary precision by smaller finite subgroups of G (i.e. $G =$ group of rigid motions of $X = \mathbb{R}^3$).

Nevertheless, in this talk we will show that d_G can be approximated with arbitrary precision by means of a **DUAL** approach based on persistent homology and G-invariant non-expansive operators.

This research is based on an ongoing joint research project with Grzegorz Jabłoński (Jagiellonian University - Poland)

G-invariant non-expansive operators

Let us consider the following objects:

- A triangulable space X with nontrivial homology in degree k.
- A set Φ of continuous functions from X to $\mathbb{R}$, that contains the set of all constant functions.
- A topological subgroup G of $\mathrm{Homeo}(X)$ that acts on Φ by composition on the right.
- A subset $\mathscr{F}$ of the set $\mathscr{F}^{\mathrm{all}}(\Phi, G)$ of all non-expansive G-operators from Φ to Φ.

The operator space $\mathscr{F}^{\mathrm{all}}(\Phi, G)$

In plain words, $F \in \mathscr{F}^{\mathrm{all}}(\Phi, G)$ means that

1. $F : \Phi \longrightarrow \Phi$
2. $F(\varphi \circ g) = F(\varphi) \circ g$. ($F$ is a G-operator)
3. $\|F(\varphi_1) - F(\varphi_2)\|_\infty \leq \|\varphi_1 - \varphi_2\|_\infty$. ($F$ is non-expansive)

The operator F is not required to be linear.

Some simple examples of F, taking Φ equal to the set of all continuous functions $\varphi : \mathbf{S}^1 \to \mathbb{R}$ and G equal to the group of all rotations of $\mathbf{S}^1$:

- $F(\varphi) :=$ the constant function $\psi : \mathbf{S}^1 \longrightarrow \mathbb{R}$ taking the value $\max \varphi$;
- $F(\varphi)$ defined by setting $F(\varphi)(x) := \max\left\{\varphi\left(x - \frac{\pi}{8}\right), \varphi\left(x + \frac{\pi}{8}\right)\right\}$;
- $F(\varphi)$ defined by setting $F(\varphi)(x) := \frac{1}{2}\left(\varphi\left(x - \frac{\pi}{8}\right) + \varphi\left(x + \frac{\pi}{8}\right)\right)$.

The pseudo-metric $D^{\mathscr{F}}_{\mathrm{match}}$

For every $\varphi_1, \varphi_2 \in \Phi$ we set

$$D^{\mathscr{F}}_{\mathrm{match}}(\varphi_1, \varphi_2) := \sup_{F \in \mathscr{F}} d_{match}(\rho_k(F(\varphi_1)), \rho_k(F(\varphi_2)))$$

where $\rho_k(\psi)$ denotes the persistent Betti number function (i.e. the rank invariant) of ψ in degree k, while d_{match} denotes the usual bottleneck distance that is used to compare the persistence diagrams associated with $\rho_k(F(\varphi_1))$ and $\rho_k(F(\varphi_2))$.

Proposition

$D^{\mathscr{F}}_{match}$ is a G-invariant and stable pseudo-metric on Φ.

The G-invariance of $D^{\mathscr{F}}_{match}$ means that for every $\varphi_1, \varphi_2 \in \Phi$ and every $g \in G$ the equality $D^{\mathscr{F}}_{\mathrm{match}}(\varphi_1, \varphi_2 \circ g) = D^{\mathscr{F}}_{\mathrm{match}}(\varphi_1, \varphi_2)$ holds.

An equivalence result

We observe that the pseudo-distance $D^{\mathscr{F}}_{\mathrm{match}}$ and the natural pseudo-distance d_G are defined in quite different ways.

In particular, the definition of $D^{\mathscr{F}}_{\mathrm{match}}$ is based on persistent homology, while the natural pseudo-distance d_G is based on the group of homeomorphisms G.

In spite of this, the following statement holds:

Theorem

If $\mathscr{F} = \mathscr{F}^{\mathrm{all}}(\Phi, G)$, then the pseudo-distance $D^{\mathscr{F}}_{\mathrm{match}}$ coincides with the natural pseudo-distance d_G on Φ.

Our main idea

The previous theorem suggests to study $D^{\mathscr{F}}_{\mathrm{match}}$ instead of d_G.

To this end, let us choose a finite subset $\mathscr{F}^*$ of $\mathscr{F}$, and consider the pseudo-metric

$$D^{\mathscr{F}^*}_{\mathrm{match}}(\varphi_1,\varphi_2) := \max_{F\in\mathscr{F}^*} d_{match}(\rho_k(F(\varphi_1)),\rho_k(F(\varphi_2)))$$

for every $\varphi_1,\varphi_2 \in \Phi$.

Obviously, $D^{\mathscr{F}^*}_{\mathrm{match}} \leq D^{\mathscr{F}}_{\mathrm{match}}$.

Furthermore, if $\mathscr{F}^*$ is dense enough in $\mathscr{F}$, then the new pseudo-distance $D^{\mathscr{F}^*}_{\mathrm{match}}$ is close to $D^{\mathscr{F}}_{\mathrm{match}}$.

In order to make this point clear, we need the next theoretical result.

Compactness of $\mathscr{F}^{\mathrm{all}}(\Phi, G)$

The following result holds:

Theorem

If Φ is a compact metric space with respect to the sup-norm, then $\mathscr{F}^{\mathrm{all}}(\Phi, G)$ is a compact metric space with respect to the distance d defined by setting

$$d(F_1, F_2) := \max_{\varphi \in \Phi} \|F_1(\varphi) - F_2(\varphi)\|_\infty$$

for every $F_1, F_2 \in \mathscr{F}$.

Approximation of $\mathscr{F}^{\mathrm{all}}(\Phi, G)$

This statement follows:

Corollary

Assume that the metric space Φ *is compact with respect to the sup-norm. Let* $\mathscr{F}$ *be a subset of* $\mathscr{F}^{\mathrm{all}}(\Phi, G)$*. For every* $\varepsilon > 0$*, a finite subset* $\mathscr{F}^*$ *of* $\mathscr{F}$ *exists, such that*

$$\left| D_{match}^{\mathscr{F}^*}(\varphi_1, \varphi_2) - D_{match}^{\mathscr{F}}(\varphi_1, \varphi_2) \right| \leq \varepsilon$$

for every $\varphi_1, \varphi_2 \in \Phi$*.*

This corollary implies that the pseudo-distance $D_{match}^{\mathscr{F}}$ can be approximated computationally, at least in the compact case.

Our basic questions

Assumptions in our model

Mathematical setting and theoretical results

Experiments

Let us check what happens in practice

A RETRIEVAL EXPERIMENT
ON A DATASET OF CURVES

Let us check what happens in practice

We have considered

1. a dataset of 10000 functions from $\mathbf{S}^1$ to $\mathbb{R}$, depending on five random parameters (#);
2. these three invariance groups:
   - ○ the group $\mathrm{Homeo}(\mathbf{S}^1)$ of all self-homeomorphisms of $\mathbf{S}^1$;
   - ○ the group $R(\mathbf{S}^1)$ of all rotations of $\mathbf{S}^1$;
   - ○ the trivial group $\mathbf{I}(\mathbf{S}^1) = \{id\}$, containing just the identity of $\mathbf{S}^1$.

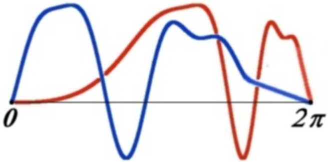
Obviously,

$$\mathrm{Homeo}(\mathbf{S}^1) \supset R(\mathbf{S}^1) \supset \mathbf{I}(\mathbf{S}^1).$$

(#) For $1 \leq i \leq 10000$ we have set $\bar{\varphi}_i(x) = r_1 \sin(3x) + r_2 \cos(3x) + r_3 \sin(4x) + r_4 \cos(4x)$, with $r_1, .., r_4$ randomly chosen in the interval $[-2, 2]$; the i-th function in our dataset is the function $\varphi_i := \bar{\varphi}_i \circ \gamma_i$, where $\gamma_i(x) := 2\pi(\frac{x}{2\pi})^{r_5}$ and r_5 is randomly chosen in the interval $[\frac{1}{2}, 2]$.

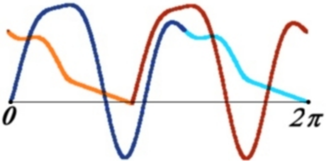
Let us check what happens in practice

The choice of $\mathrm{Homeo}(\mathbf{S}^1)$ as an invariance group implies that the following two functions are considered equivalent. Their graphs are obtained from each other by applying a horizontal stretching. Also shifts are accepted as legitimate transformations.



Let us check what happens in practice

The choice of $R(\mathbf{S}^1)$ as an invariance group implies that the following two functions are considered equivalent. Their graphs are obtained from each other by applying a rotation of $\mathbf{S}^1$. Stretching is not accepted as a legitimate transformation.



Finally, the choice of $\mathbf{I}(\mathbf{S}^1) = \{id\}$ as an invariance group means that two functions are considered equivalent if and only if they coincide everywhere.

The results of an experiment: the group $\mathrm{Homeo}(\mathbf{S}^1)$

What happens if we decide to assume

that the invariance group is the group $\mathrm{Homeo}(\mathbf{S}^1)$

of all self-homeomorphisms of $\mathbf{S}^1$?

The results of an experiment: the group $\mathrm{Homeo}(\mathbf{S}^1)$

If we choose $G = \mathrm{Homeo}(\mathbf{S}^1)$, to proceed we need to choose a finite set of non-expansive $\mathrm{Homeo}(\mathbf{S}^1)$-operators. In our experiment we have considered these three non-expansive $\mathrm{Homeo}(\mathbf{S}^1)$-operators:

- $F_0 := id$ (i.e., $F_0(\varphi) := \varphi$);
- $F_1 := -id$ (i.e., $F_0(\varphi) := -\varphi$);
- $F_2(\varphi) :=$ the constant function $\psi : \mathbf{S}^1 \to \mathbb{R}$ taking the value $\frac{1}{5} \cdot \sup\{-\varphi(x_1) + \varphi(x_2) - \frac{1}{2}\varphi(x_3) + \frac{1}{2}\varphi(x_4) - \varphi(x_5) + \varphi(x_6)\}$, $(x_1, \ldots, x_6)$ varying among all the counterclockwise 6-tuples on $\mathbf{S}^1$.

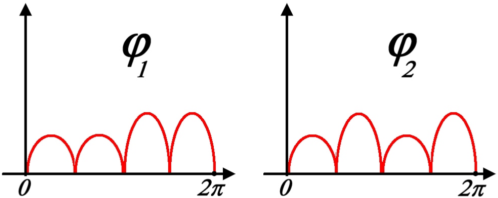
This choice produces the $\mathrm{Homeo}(\mathbf{S}^1)$-invariant pseudo-distance

$$D_{match}^{\mathscr{F}^*}(\varphi_1, \varphi_2) := \max_{0 \le i \le 2} d_{match}(\rho_k(F_i(\varphi_1)), \rho_k(F_i(\varphi_2))).$$

An important remark

It is important to use several operators. The use of just one operator still produces a pseudo-distance $D^{\mathscr{F}^*}_{match}$ that is invariant under the action of the group G, but this choice is far from guaranteeing a good approximation of the natural pseudo-distance d_G.

As an example in the case $G = \mathrm{Homeo}(\mathbf{S}^1)$, if we use just the identity operator (i.e., we just apply classical persistent homology), we cannot distinguish these two functions $\varphi_1, \varphi_2 : \mathbf{S}^1 \to \mathbb{R}$, despite the fact that they are different for d_G:



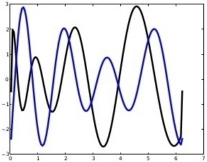
The results of an experiment: the group Homeo($\mathbf{S}^1$)

Here is a query (in **blue**), and the first four retrieved functions (in **black**):

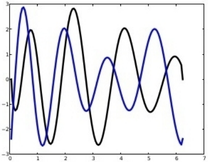
$D^{\mathcal{F}^*}_{\text{match}}(\varphi_1, \varphi_2)$

Mean	Max
1.648	4.831

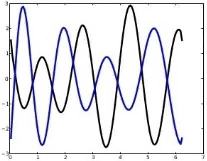
Standard deviation
0.934



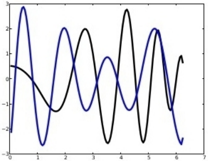
(a) φ_{516}, dist: 0.0465393



(b) φ_{381}, dist: 0.0541687



(c) φ_{7776}, dist: 0.0984192

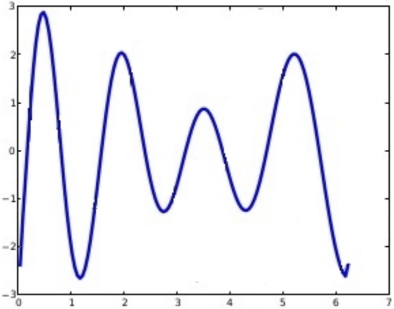


(d) φ_{6214}, dist: 0.10376

The results of an experiment: the group $\mathrm{Homeo}(\mathbf{S}^1)$

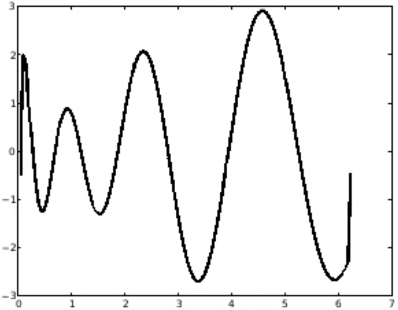
Let's have a closer look at the query and at the first retrieved function:

Here is the query:



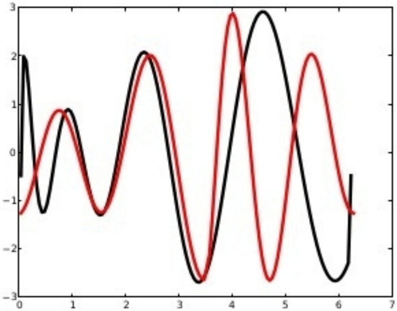
The results of an experiment: the group $\mathrm{Homeo}(\mathbf{S}^1)$

Here is the first retrieved function with respect to $D^{\mathscr{F}^*}_{match}$:



The results of an experiment: the group $\mathrm{Homeo}(\mathbf{S}^1)$

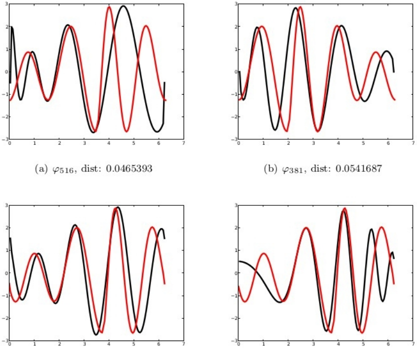
Here is the query function after aligning it to the first retrieved function by means of a shift (in **red**).
The first retrieved function is represented in **black**.
The figure shows that the retrieved function is approximately equivalent to the query function, by applying a shift and a stretching.



The results of an experiment: the group Homeo($\mathbf{S}^1$)

Here is the query function after aligning it to the first four retrieved functions by means of a shift (in **red**).

The first four retrieved functions are represented in **black**.



(a) φ_{516}, dist: 0.0465393

(b) φ_{381}, dist: 0.0541687

(c) φ_{7776}, dist: 0.0984192

(d) φ_{6214}, dist: 0.10376

The results of an experiment: the group $R(\mathbf{S}^1)$

What happens if we decide to assume

that the invariance group is the group $R(\mathbf{S}^1)$

of all rotations of $\mathbf{S}^1$?

The results of an experiment: the group $R(\mathbf{S}^1)$

If we choose $G = R(\mathbf{S}^1)$, in order to proceed we need to choose a finite set of non-expansive $R(\mathbf{S}^1)$-operators. Obviously, since F_0, F_1 and F_2 are $\mathrm{Homeo}(\mathbf{S}^1)$-invariant, they are also $R(\mathbf{S}^1)$-invariant. In our experiment we have added these five non-expansive $R(\mathbf{S}^1)$-operators (which are not $\mathrm{Homeo}(\mathbf{S}^1)$-invariant) to F_0, F_1 and F_2:

- $F_3(\varphi)(x) := \max\{\varphi(x), \varphi(x+\pi)\}$
- $F_4(\varphi)(x) := \frac{1}{2}\cdot\left(\varphi(x)+\varphi(x+\frac{\pi}{4})\right)$
- $F_5(\varphi)(x) := \max\{\varphi(x), \varphi(x+\pi/10), \varphi(x+\frac{2\pi}{10}), \varphi(x+\frac{3\pi}{10})\}$
- $F_6(\varphi)(x) := \frac{1}{3}\cdot\left(\varphi(x)+\varphi(x+\frac{\pi}{3})+\varphi(x+\frac{\pi}{4})\right)$
- $F_7(\varphi)(x) := \frac{1}{3}\cdot\left(\varphi(x)+\varphi(x+\frac{\pi}{3})+\varphi(x+\frac{2\pi}{3})\right)$

This choice produces the $R(\mathbf{S}^1)$-invariant pseudo-distance

$$D_{match}^{\mathscr{F}^*}(\varphi_1,\varphi_2) := \max_{0\le i\le 7} d_{match}(\rho_k(F_i(\varphi_1)), \rho_k(F_i(\varphi_2))).$$

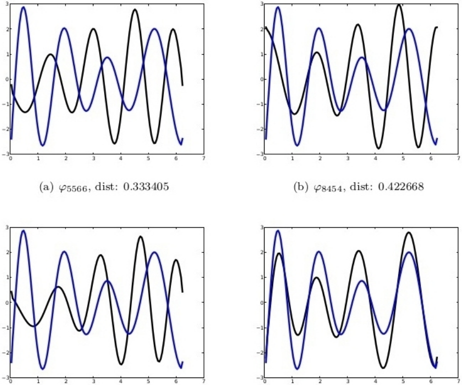
The results of an experiment: the group $R(\mathbf{S}^1)$

Here is a query (in **blue**), and the first four retrieved functions (in **black**):

$$D^{\mathcal{F}^*}_{\text{match}}(\varphi_1, \varphi_2)$$

Mean	Max
1.938	4.831

Standard deviation
0.874



(a) φ_{5566}, dist: 0.333405

(b) φ_{8454}, dist: 0.422668

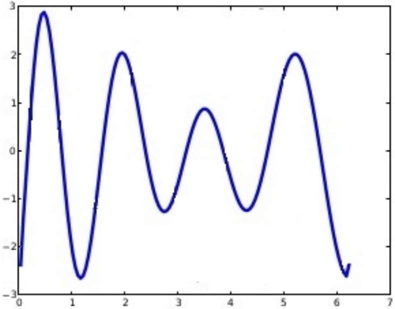
(c) φ_{8909}, dist: 0.453949

(d) φ_{4426}, dist: 0.46463

The results of an experiment: the group $R(\mathbf{S}^1)$

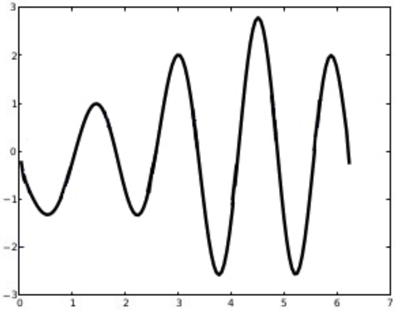
Let's have a closer look at the query and at the first retrieved function:

Here is the query:



The results of an experiment: the group $R(\mathbf{S}^1)$

Here is the first retrieved function with respect to $D^{\mathscr{F}^*}_{match}$:

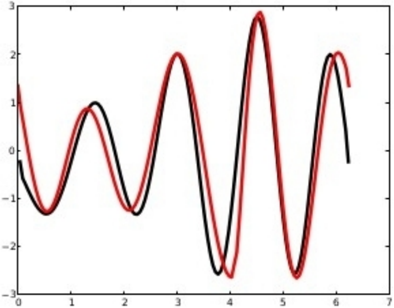


The results of an experiment: the group $R(\mathbf{S}^1)$

Here is the query function after aligning it to the first retrieved function by means of a shift (in **red**).

The first retrieved function is represented in **black**.

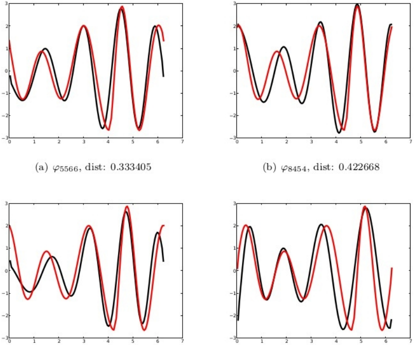
The figure shows that the retrieved function is approximately equivalent to the query function, via a shift.



The results of an experiment: the group $R(\mathbf{S}^1)$

Here is the query function after aligning it to the first four retrieved functions by means of a shift (in **red**).

The first four retrieved functions are represented in **black**.



(a) φ_{5566}, dist: 0.333405

(b) φ_{8454}, dist: 0.422668

(c) φ_{8909}, dist: 0.453949

(d) φ_{4426}, dist: 0.46463

The results of an experiment: the group $\mathbf{I}(\mathbf{S}^1)$

Finally, what happens if we decide to assume

that the invariance group is the group $\mathbf{I}(\mathbf{S}^1) = \{id\}$

containing only the identity of $\mathbf{S}^1$?

This means that the "perfect" retrieved function

should coincide with our query.

The results of an experiment: the group $\mathbf{I}(\mathbf{S}^1)$

If we choose $G = \mathbf{I}(\mathbf{S}^1) = \{id\}$, in order to proceed we need to choose a finite set of non-expansive operators (obviously, every operator is an $\mathbf{I}(\mathbf{S}^1)$-operator).

In our experiment we have considered these three non-expansive operators (which are not $R(\mathbf{S}^1)$-operators):

- $F_8(\varphi)(x) := \sin(x)\varphi(x)$
- $F_9(\varphi)(x) := \frac{\sqrt{2}}{2}\sin(x)\varphi(x) + \frac{\sqrt{2}}{2}\cos(x)\varphi(x+\frac{\pi}{2})$
- $F_{10}(\varphi)(x) := \sin(2x)\varphi(x)$

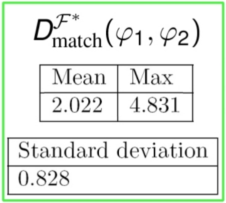
We have added F_8, F_9, F_{10} to $F_1, \ldots, F_7$.

This choice produces the pseudo-distance

$$D^{\mathscr{F}^*}_{match}(\varphi_1,\varphi_2) := \max_{0 \le i \le 10} d_{match}(\rho_k(F_i(\varphi_1)), \rho_k(F_i(\varphi_2))).$$

The results of an experiment: the group $\mathbf{I}(\mathbf{S}^1)$

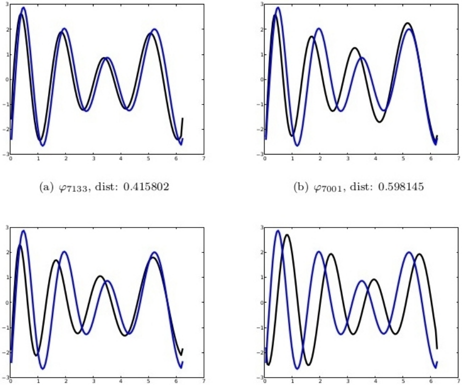
Here is a query (in **blue**), and the first four retrieved functions (in **black**):



$D^{\mathcal{F}^*}_{\text{match}}(\varphi_1, \varphi_2)$

Mean	Max
2.022	4.831

Standard deviation
0.828



(a) φ_{7133}, dist: 0.415802

(b) φ_{7001}, dist: 0.598145

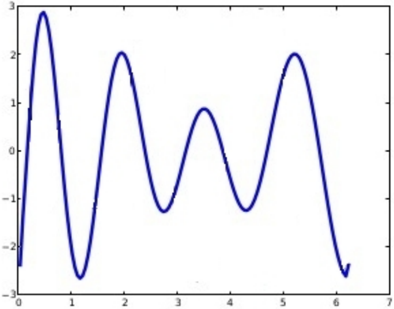
(c) φ_{389}, dist: 0.617218

(d) φ_{5723}, dist: 0.617981

The results of an experiment: the group $\mathbf{I}(\mathbf{S}^1)$

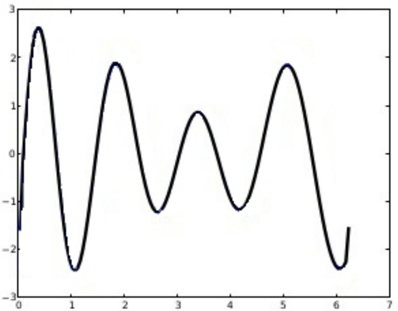
Let's have a closer look at the query and at the first retrieved function:

Here is the query:



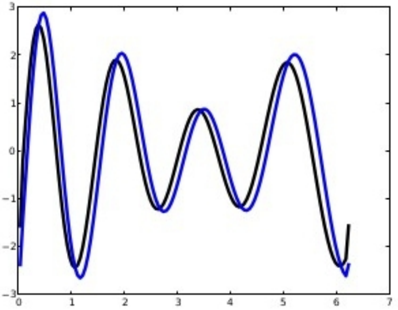
The results of an experiment: the group $\mathbf{I}(\mathbf{S}^1)$

Here is the first retrieved function with respect to $D^{\mathscr{F}}_{match}$:



The results of an experiment: the group $\mathbf{I}(\mathbf{S}^1)$

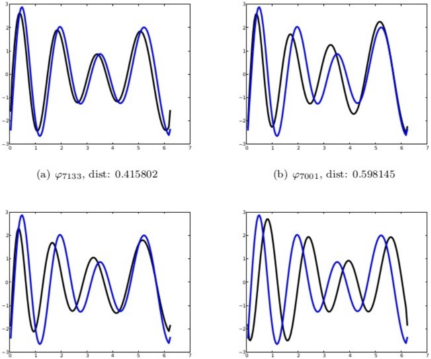
The first retrieved function is represented in **black**.
As expected, no aligning shift is necessary here.
The figure shows that the retrieved function is approximately equal to the query function.



The results of an experiment: the group $\mathbf{I}(\mathbf{S}^1)$

Here we show again the query function and the first four retrieved functions (in **black**).

The figure shows that the retrieved functions are approximately coinciding with the query function.



(a) φ_{7133}, dist: 0.415802

(b) φ_{7001}, dist: 0.598145

An open problem

We have proven that if Φ is compact, then $D^{\mathscr{F}}_{match}$ can be approximated computationally.

However, this result does not say which set of operators allows for both a good approximation of $D^{\mathscr{F}}_{match}$ and a fast computation.

Further research is needed in this direction.

Conclusions

In this talk we have supported these statements:

- Shape comparison is based on acts of measurement made by an observer. The acts of measurement can be represented as a function defined on a topological space X. The observer can be seen as a collection of G-invariant operators, applied to the functions describing the data.
- These functions can be compared by means of the natural pseudo-distance associated with any subgroup G of $\mathrm{Homeo}(X)$.
- Persistent homology can be used to approximate the natural pseudo-metric d_G. This can be done by means of a method that is based on non-expansive G-operators. This method is stable with respect to noise.

For more information about the approach described in these slides use the following link: http://arxiv.org/pdf/1312.7219v3.pdf.

THANKS
FOR YOUR
ATTENTION
UNAM
JURIQUILLA