Group-invariant persistent homology and its use for topological data analysis

Patrizio Frosini

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

OSU, 7 April 2017

Outline

Assumptions in our model

Let us recall our mathematical setting

Experiments

Assumptions in our model

We will recall the assumptions made in the previous TGDA seminar:

1. No object can be studied in a direct and absolute way. Any object is only knowable 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 these data. These operators are frequently endowed with some invariances that are relevant for the observer.
4. Only the observer is entitled to decide about data similarity.

An important remark

Classical persistent homology is not a suitable model for our purpose, because it is invariant with respect to ANY homeomorphism! In other words, it does not allow the observer to choose the invariance he/she wants. This fact justifies the introduction of G-invariant persistent homology.

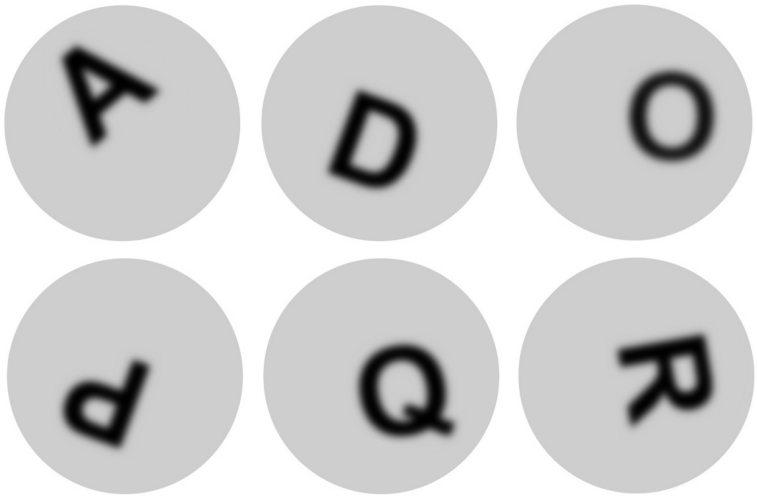


Figure: These real-valued functions share the same persistent homology.

Couldn't we maintain classical persistent homology?

One could think of using other filtering functions, possibly defined on different topological spaces. For example, we could extract boundaries of letters and consider the distance from the center of mass of each boundary. This approach presents some drawbacks:

1. It "forgets" most of the information contained in the image $\varphi : \mathbb{R}^2 \to \mathbb{R}$ that we are considering, confining itself to examine the boundary of the letter represented by φ.
2. It usually requires an extra computational cost (e.g., to extract the boundaries of the letters).
3. It can produce a different topological space for each new filtering function (e.g., this happens for letters).
4. ABOVE ALL: It is not clear how we can translate the invariance that we need into the choice of new filtering functions defined on new topological spaces.

The role of the observer in our model

In our model the observer is seen as a collection of group invariant non-expansive operators (GINOs). The observer cannot choose the data that have to be analyzed, but he/she can often choose the operators that will be applied to those functions.

Each operator transforms the data (i.e. the set Φ of the functions defined on the space X) into other data (i.e. the set Ψ of the functions defined on another space Y). This transformation usually respects some kind of invariance, expressed by suitable groups G, H of homeomorphisms.

We recall that the homeomorphisms do not concern the "objects" but the space where the measurements are made. This space is usually unique for each kind of measurement.

Assumptions in our model

Let us recall our mathematical setting

Experiments

Natural pseudo-distance associated with a group G

Let us recall the definition of natural pseudo-distance.

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.

Topologies on X and G

Let X be a set. Let Φ be a non-empty subset of the set of all bounded functions from X to $\mathbb{R}$, endowed with the norm $\|\cdot\|_\infty$. We assume that Φ is compact and contains at least the constant functions taking every finite value c with $|c| \leq \sup_{\varphi\in\Phi} \|\varphi\|_\infty$. We also consider a group $G \subseteq \mathrm{Homeo}(X)$, acting on Φ by composition on the right.

We endow X with the initial topology, i.e. the coarsest topology on X such that every function in Φ is continuous. In other words, on X we consider this pseudo-metric: $d_X(x_1, x_2) := \sup_{\varphi\in\Phi} |\varphi(x_1) - \varphi(x_2)|$.

We endow the group G with the pseudo-metric $D_G(g_1, g_2) := \sup_{\varphi\in\Phi} \|\varphi \circ g_1 - \varphi \circ g_2\|_\infty$. G is a topological group that acts continuously on Φ by composition on the right.

We will also assume that X and G are compact, and say that (Φ, G) is a perception pair.

Changing (Φ, G) into (Ψ, H)

Each perception pair (Φ, G) can be seen as a category whose objects are the elements of the compact space Φ and whose arrows are the elements of the topological group G.

Each functor $F : (\Phi, G) \to (\Psi, H)$ is called a Group Invariant Non-expansive Operator (GINO) if:

- F is group invariant: $F(\varphi \circ g) = F(\varphi) \circ F(g)$ for every $\varphi \in \Phi, g \in G$;
- F is non-expansive on Φ: $\|F(\varphi_1) - F(\varphi_2)\|_\infty \leq \|\varphi_1 - \varphi_2\|_\infty$ for every $\varphi_1, \varphi_2 \in \Phi$;
- F is non-expansive on G: $D_H(F(g_1), F(g_2)) \leq D_G(g_1, g_2)$ for every $g_1, g_2 \in G$.

The space of GINOs and $D^{\mathscr{F}}_{\mathrm{match}}$

We have seen in the previous TGDA seminar that the space of all GINOs between two persistence pairs (Φ, G), (Ψ, H) is compact, so that it can be ε-approximated by a finite set of GINOs.

This means that, in principle, the distance

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

can be computationally approximated.

We have also seen that $D^{\mathscr{F}}_{\mathrm{match}}$ is G-invariant and stable, and that it can be a proxy for the natural pseudo-distance d_G.

The reasons to use a dual approach

We recall that in general no finite subgroup H of G exists for which the pseudo-distance d_H is an arbitrarily good approximation of d_G. Therefore, differently from $D^{\mathscr{F}}_{match}$, d_G cannot be approximated by another distance of the same kind. In other words, $D^{\mathscr{F}}_{match}$ has better properties than d_G with respect to approximation.

Furthermore, the results of the experiments show that the use of some small family of simple operators may produce a pseudo-metric $D^{\mathscr{F}^*}_{match}$ that is not far from d_G and can be efficiently used for data retrieval, even if $\mathscr{F}^*$ is not a good approximation of the set of all GINOs.

These observations justify the use of $D^{\mathscr{F}}_{match}$ in place of d_G, for practical purposes.

The reasons to use a dual approach

We wish to underline the dual nature of our approach in the case $(\Phi, G) = (\Psi, H)$. When G becomes "larger and larger" the associated family $\mathscr{F}(\Phi, G)$ of all G-invariant non-expansive operators becomes "smaller and smaller", so making the computation of $D_{match}^{\mathscr{F}(\Phi,G)}$ easier and easier, contrarily to what happens for the direct computation of d_G. In other words, the approach based on $D_{match}^{\mathscr{F}(\Phi,G)}$ seems to be of use exactly when d_G is difficult to compute in a direct way.

The reasons to use a dual approach

Moreover, assuming that $\mathscr{F}^*$ is a finite subset of $\mathscr{F}$ and H is a finite subgroup of G, the duality in the definitions of $D^{\mathscr{F}}_{match}$ and d_G causes another important difference in the use of $D^{\mathscr{F}^*}_{match}$ and d_H as respective approximations. It consists in the fact that while $D^{\mathscr{F}^*}_{match}$ is a lower bound for $D^{\mathscr{F}}_{match} \leq d_G$, d_H is an upper bound for d_G:

$$D^{\mathscr{F}^*}_{match} \leq D^{\mathscr{F}}_{match} \leq d_G \leq d_H.$$

As a consequence, if we take the pseudo-metric d_G as the ground truth, the retrieval errors associated with the use of $D^{\mathscr{F}^*}_{match}$ are just false positive, while the ones associated with the use of d_H are just false negative.

The algebra of GINOs

We have seen in the previous TGDA seminar that

- The composition of GINOs is a GINO.
- The translation of a GINO is a GINO.
- The weighted average of GINOs is a GINO (provided that the sum of the weights is 1).
- The maximum of GINOs is a GINO.

We have also seen that the method we use to reduce 2D persistent Betti numbers to families of 1D persistent Betti numbers corresponds to the use of the GINO

$$F(\varphi) = \max\left\{ \frac{\min_j a_j}{a_1} \cdot (F_1(\varphi) - b_1), \ldots, \frac{\min_j a_j}{a_n} \cdot (F_n(\varphi) - b_n) \right\}$$

where $F_i(\varphi) = \varphi_i$ for $\varphi \in \Phi$, and $F(g) = F_1(g) = \ldots = F_n(g)$ for $g \in G$.

Assumptions in our model

Let us recall our mathematical setting

Experiments

Let us check what happens in practice

In the TGDA seminar we have illustrated the use of GIPHOD (`http://giphod.ii.uj.edu.pl`).

In this lecture we will present and discuss a retrieval experiment on a dataset of curves, where we set $\Phi = \Psi$ and $G = H$.

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 $\mathbf{S}^1$.

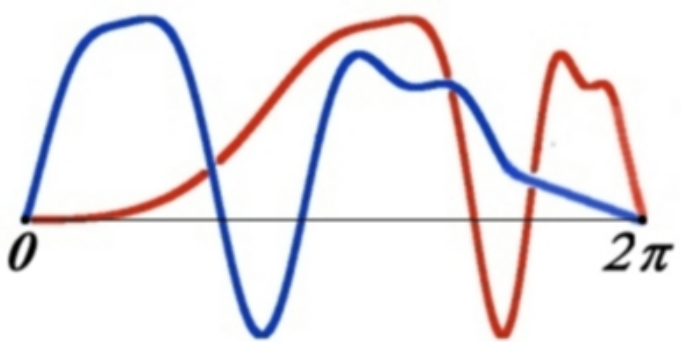
Obviously,

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

(#) 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 $[\frac{1}{2}, 2]$.

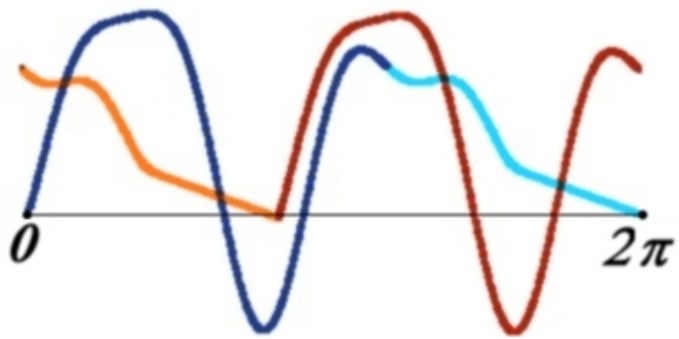
Let us check what happens in practice

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 $\mathbf{S}^1$.

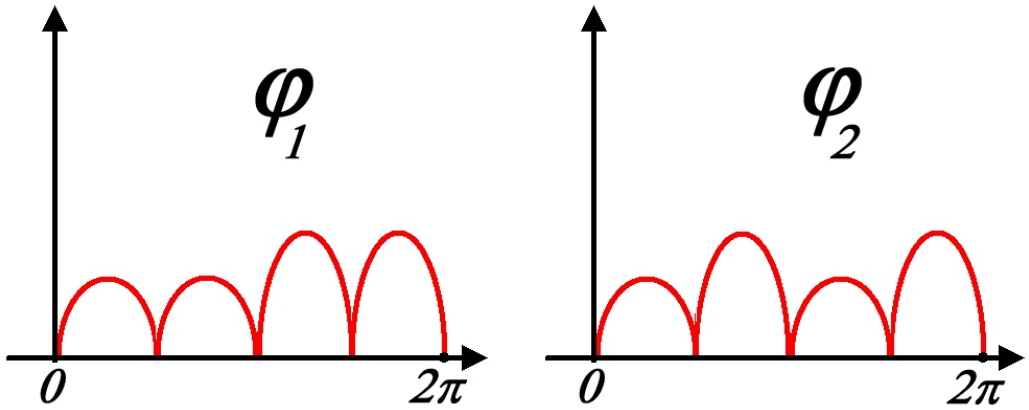
This choice produces the $\mathrm{Homeo}(\mathbf{S}^1)$-invariant pseudo-distance

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

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:



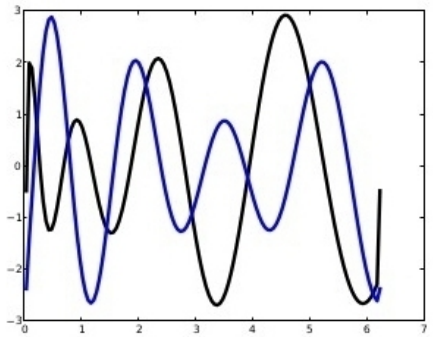
The results of an experiment: the group Homeo($\mathbf{S}^1$)

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

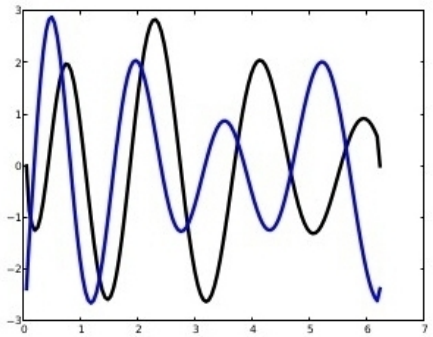
$D^{\mathcal{F}^*}_{\text{match}}(\varphi_1, \varphi_2)$

Mean	Max
1.648	4.831

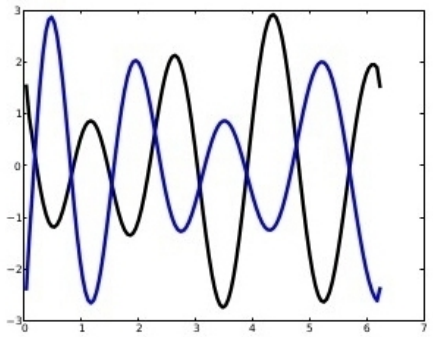
Standard deviation
0.934



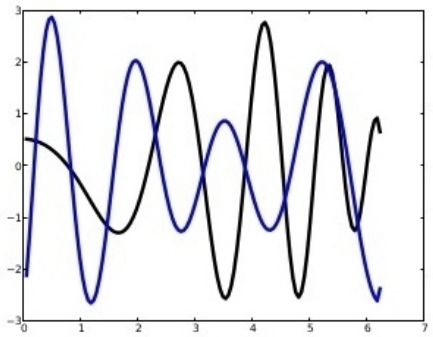
(a) φ_{516}, dist: 0.0465393



(b) φ_{381}, dist: 0.0541687



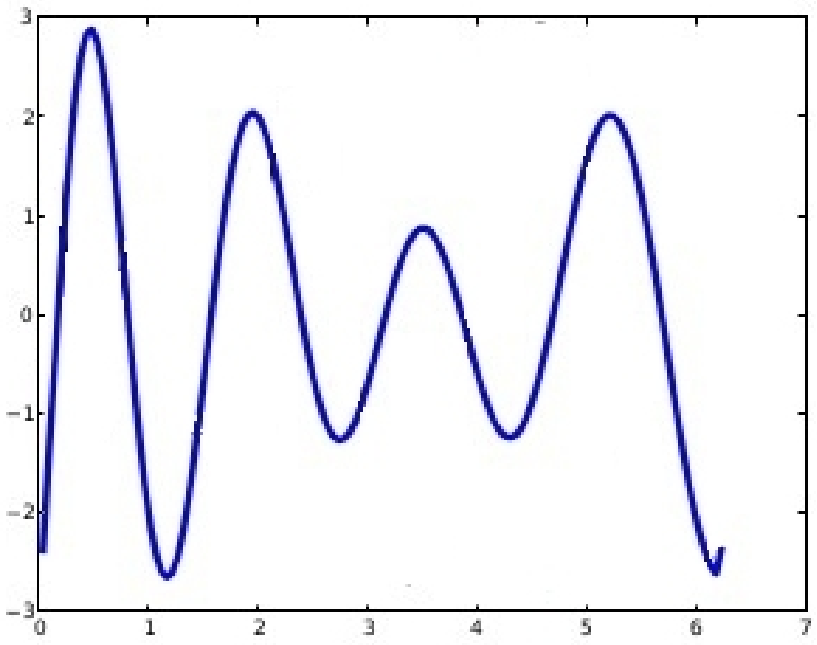
(c) φ_{7776}, dist: 0.0984192



(d) φ_{6214}, dist: 0.10376

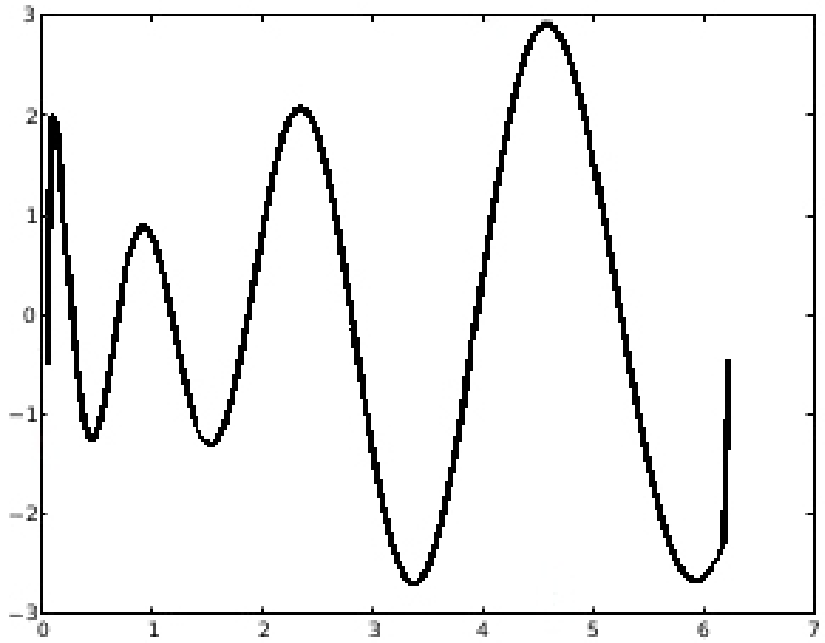
The results of an experiment: the group Homeo($\mathbf{S}^1$)

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



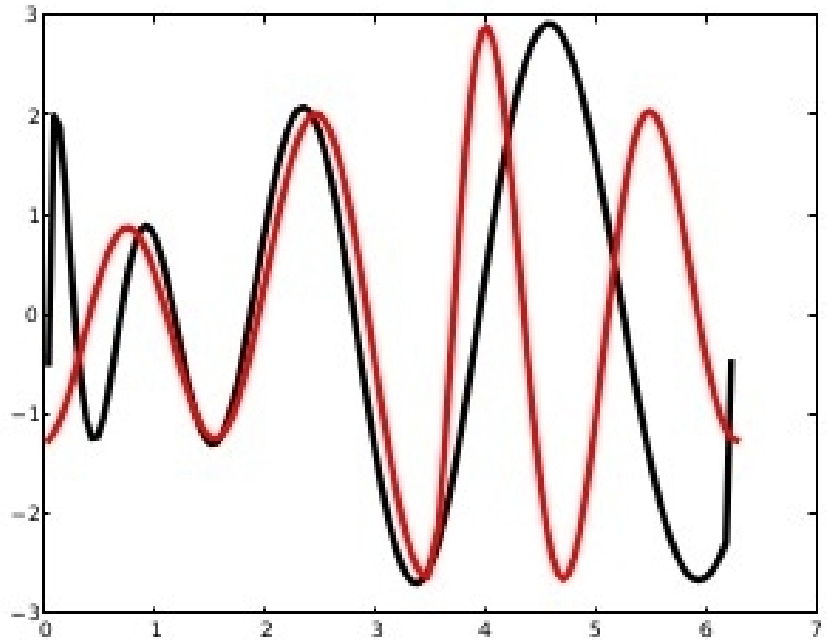
The results of an experiment: the group Homeo($\mathbf{S}^1$)

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



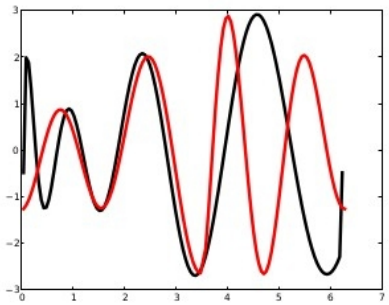
The results of an experiment: the group Homeo($\mathbf{S}^1$)

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.

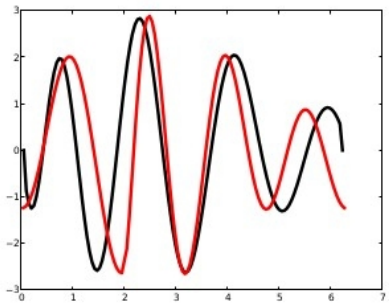


The results of an experiment: the group Homeo($\mathbf{S}^1$)

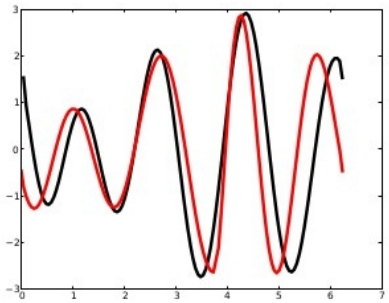
Here is the query function after aligning it to the first four retrieved functions by means of a shift (in **red**).
The first four retrieved functions are represented in **black**.



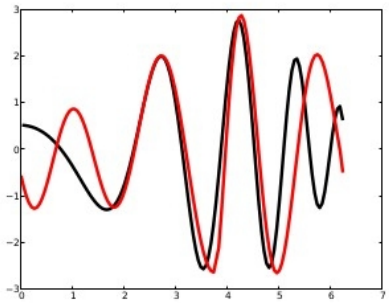
(a) φ_{516}, dist: 0.0465393



(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

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

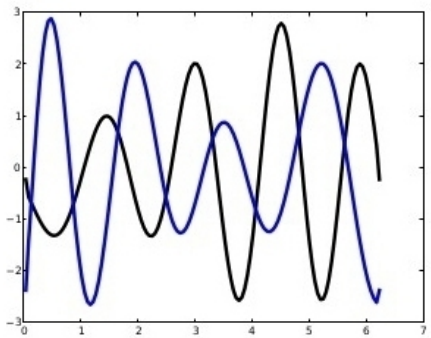
The results of an experiment: the group $R(\mathbf{S}^1)$

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

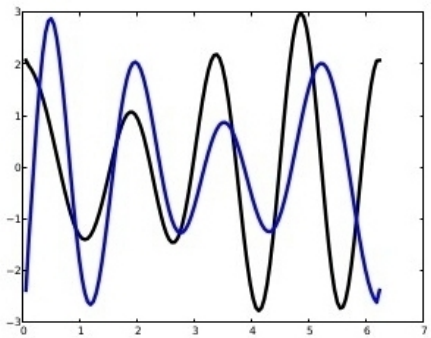
$D^{\mathcal{F}^*}_{\text{match}}(\varphi_1, \varphi_2)$

Mean	Max
1.938	4.831

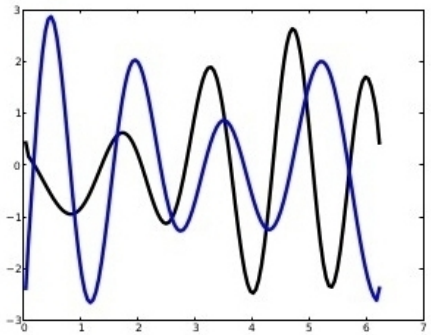
Standard deviation
0.874



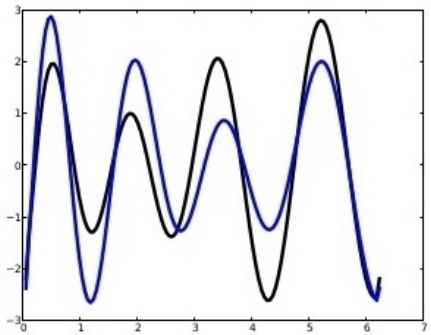
(a) φ_{5566}, dist: 0.333405



(b) φ_{8454}, dist: 0.422668



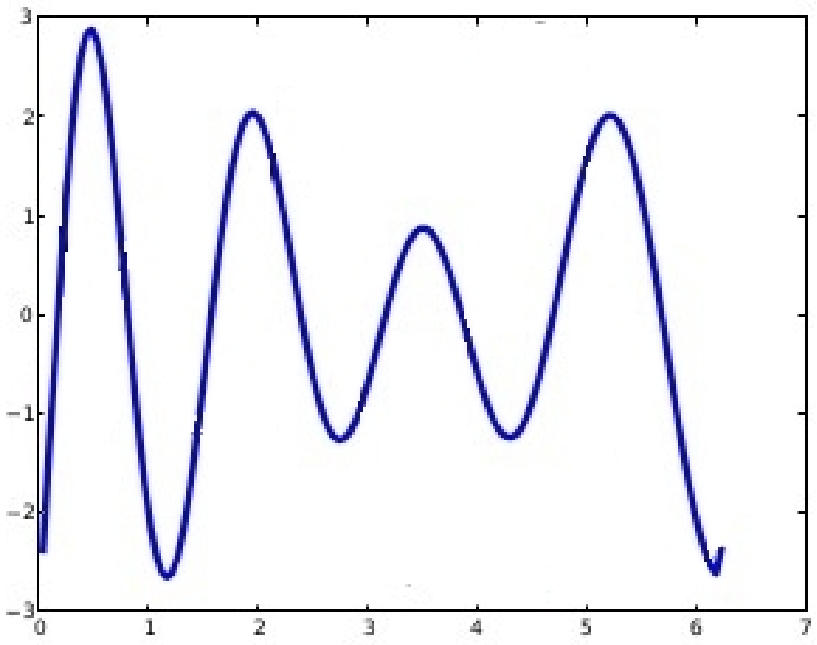
(c) φ_{8909}, dist: 0.453949



(d) φ_{4426}, dist: 0.46463

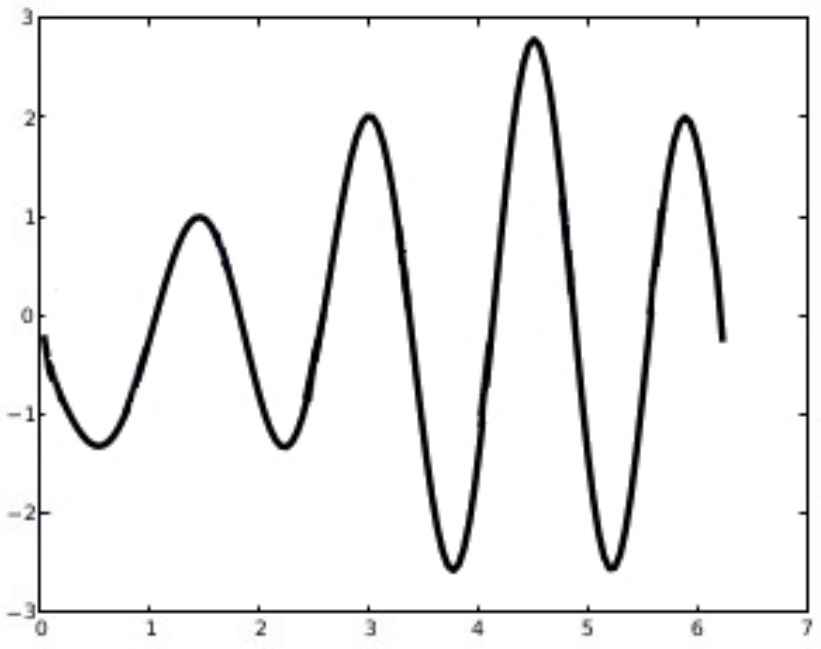
The results of an experiment: the group $R(\mathbf{S}^1)$

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



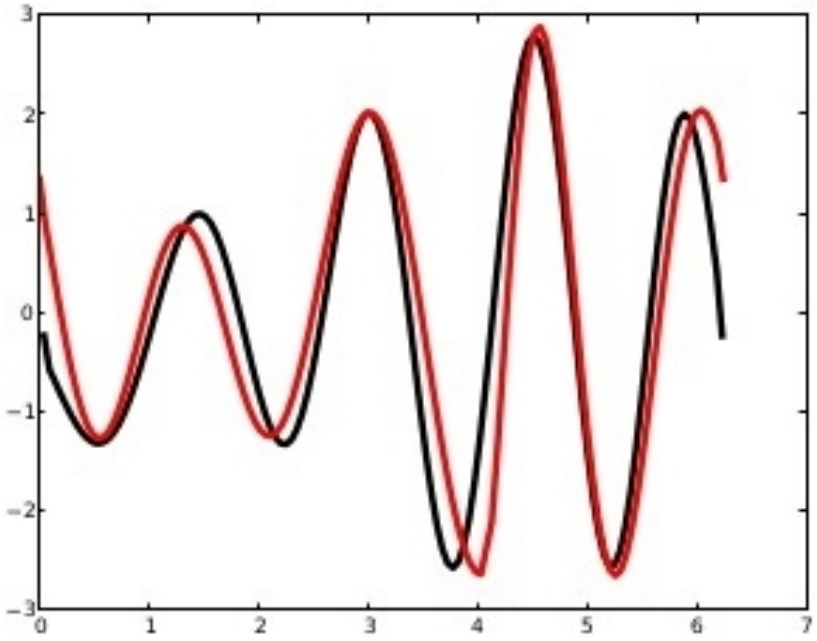
The results of an experiment: the group $R(\mathbf{S}^1)$

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



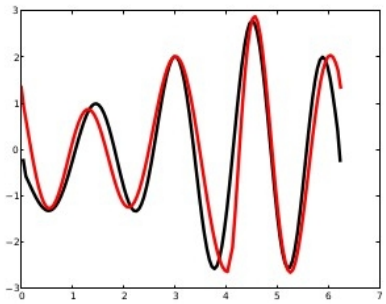
The results of an experiment: the group $R(\mathbf{S}^1)$

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.

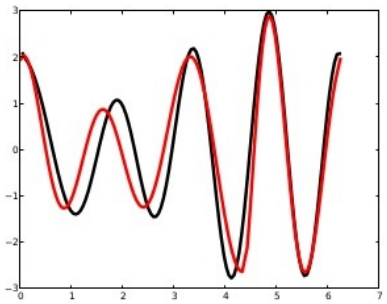


The results of an experiment: the group $R(\mathbf{S}^1)$

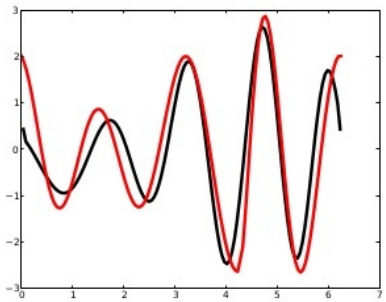
Here is the query function after aligning it to the first four retrieved functions by means of a shift (in **red**).
The first four retrieved functions are represented in **black**.



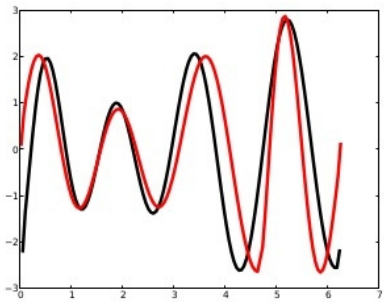
(a) φ_{5566}, dist: 0.333405



(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.

Remark: This is exactly the case where **we should not** use our dual approach! (Just compute $d_{\mathbf{I}(\mathbf{S}^1)}(\varphi_1, \varphi_2) = \|\varphi_1 - \varphi_2\|_\infty$ directly!)

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

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

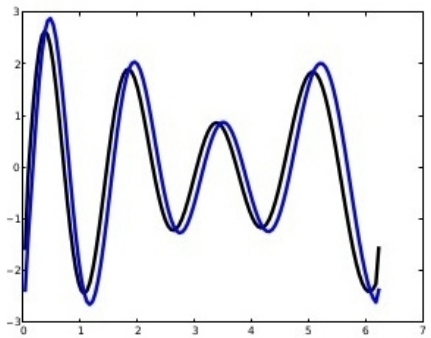
The results of an experiment: the group $\mathbf{I}(\mathbf{S}^1)$

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

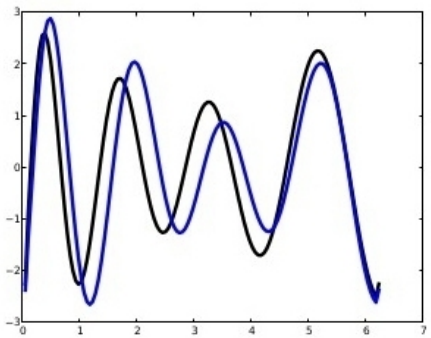
$D^{\mathcal{F}^*}_{\text{match}}(\varphi_1, \varphi_2)$

Mean	Max
2.022	4.831

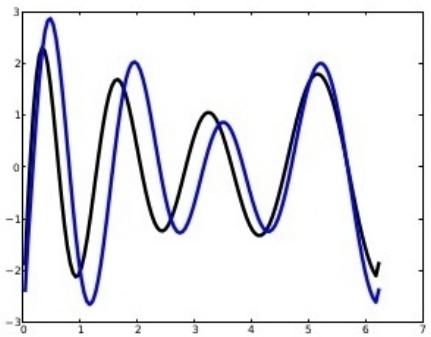
Standard deviation
0.828



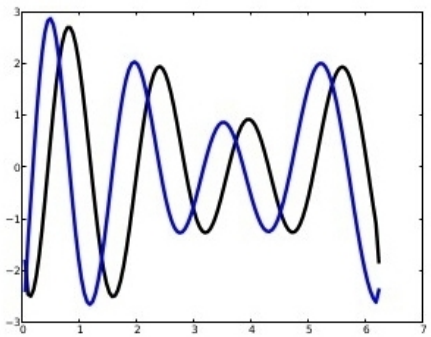
(a) φ_{7133}, dist: 0.415802



(b) φ_{7001}, dist: 0.598145



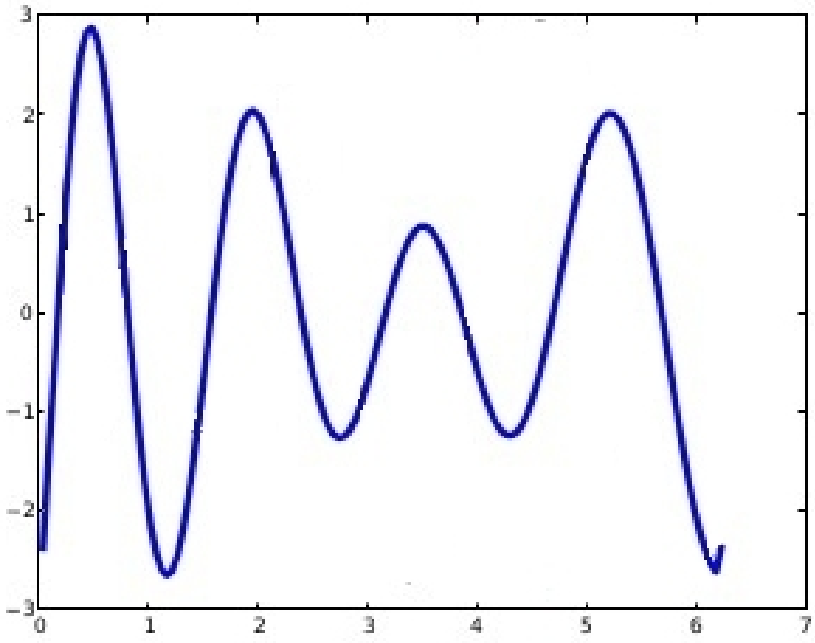
(c) φ_{389}, dist: 0.617218



(d) φ_{5723}, dist: 0.617981

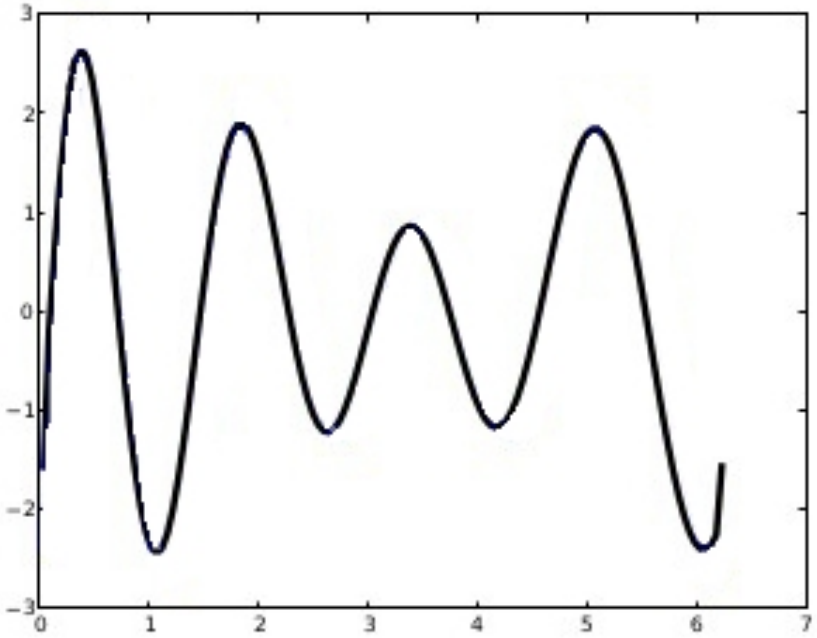
The results of an experiment: the group $\mathbf{I}(\mathbf{S}^1)$

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



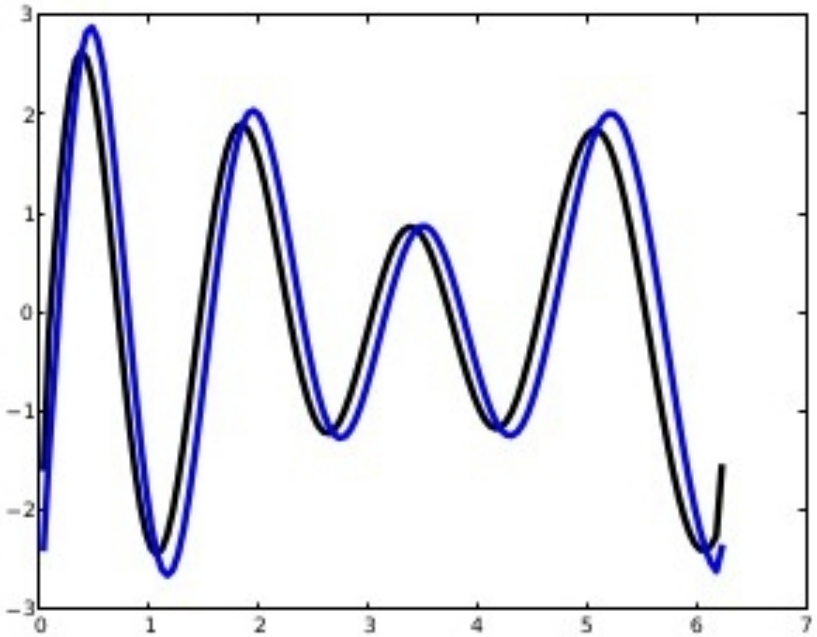
The results of an experiment: the group $\mathbf{I}(\mathbf{S}^1)$

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



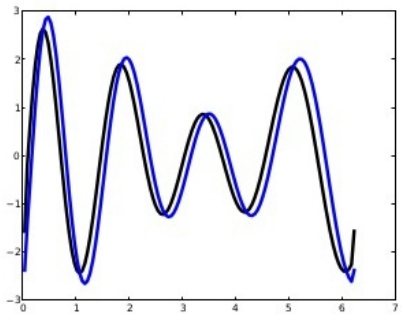
The results of an experiment: the group $\mathbf{I}(\mathbf{S}^1)$

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.

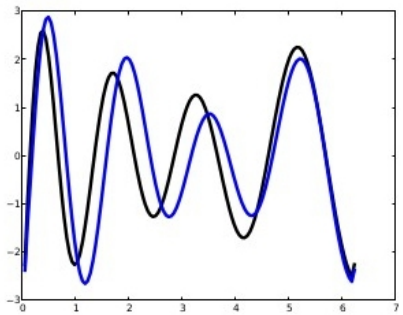


The results of an experiment: the group $\mathbf{I}(\mathbf{S}^1)$

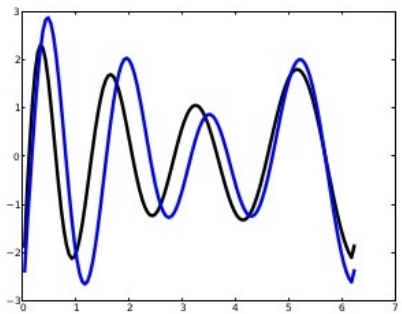
Here we show again the query function and the first four retrieved functions (in **black**).
The figure shows that the retrieved functions are approximately coinciding with the query function.

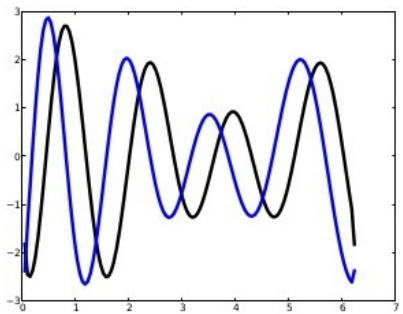


(a) φ_{7133}, dist: 0.415802



(b) φ_{7001}, dist: 0.598145





Open questions

- How can we build a good library of GINOs?
- How can we find a method to choose a finite set $\mathscr{F}^*$ of GINOs that allows for both a good approximation of the natural pseudo-distance d_G and a fast computation?

In other words: **How can we obtain an efficient and good approximation of the observer in our model?**

Further research is needed.

Conclusions

In this lecture we have supported these statements:

- Data comparison is based on acts of measurement made by an observer. Each set of acts of measurement can be represented as a function defined on a topological space X.
- The observer can be seen as a collection of GINOs, 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 GINOs. This method is stable with respect to noise.

We have also illustrated an experiment showing the practical use of our theoretical approach.


Thanks for your attention!
TDA