

Some advances in the application of group-invariant persistent homology to topological data analysis

Patrizio Frosini

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

Homotopy Probability Theory - Saarbrücken, 4-8 September 2017

Outline



Recap of my previous talk and some remarks

Work in progress on extending the theory of GINOs

Work in progress on the metric space of GINOs

Some final remarks



Recap of my previous talk and some remarks

Work in progress on extending the theory of GINOs

Work in progress on the metric space of GINOs

Some final remarks



Assumptions in our model

We will recall the assumptions made in our previous talk:

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.

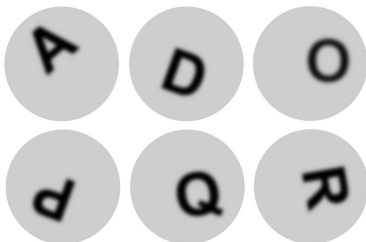


Figure: These real-valued functions share the same persistent homology.

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 $\varphi : \mathbb{R}^2 \rightarrow \mathbb{R}$ 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. (In our previous talk we have illustrated the case $\Phi = \Psi, G = H$.)

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.

Natural pseudo-distance associated with a group G



Before proceeding, let us recall the definition of natural pseudo-distance.

Definition

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



Some work in progress

In this talk we will speak about some work in progress, concerning these three lines of research:

- Change of the topologies used on X and G .
- Extension of our approach to operators taking the space Φ (where a group G acts) to a different space of functions Ψ (where another group H acts).
- Study of the metric space of GINOs both in the case $(\Phi, G) = (\Psi, H)$ and in the case $(\Phi, G) \neq (\Psi, H)$.

(Joint work with Nicola Quercioli)



Recap of my previous talk and some remarks

Work in progress on extending the theory of GINOs

Work in progress on the metric space of GINOs

Some final remarks

Some work in progress: New 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 value c with $|c| \leq \sup_{\varphi \in \Phi} \|\varphi\|_\infty$. We also consider a group $G \subseteq \text{Homeo}(X)$, 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**.

Some work in progress: New topologies on X and G



We have to justify our choice of using the initial topology on X , instead of another topology. The reason of this choice is twofold:

1. In several applications there is no information about the topology that should be used on X . In this case, it is reasonable to rely only on a topology induced by our measurements.
2. The theory is more symmetrical when the initial topology is chosen on X . Indeed, in this case the functions in Φ are used to define two “natural” pseudo-metrics on X and on G by setting
$$d_X(x_1, x_2) := \sup_{\varphi \in \Phi} |\varphi(x_1) - \varphi(x_2)|$$
 for $x_1, x_2 \in X$ and
$$D_G(g_1, g_2) := \sup_{\varphi \in \Phi} \|\varphi \circ g_1 - \varphi \circ g_2\|_\infty$$
 for $g_1, g_2 \in G$. **In plain words: Two points $x_1, x_2 \in X$ are close to each other if every function in Φ takes similar values at those points. Two homeomorphisms $g_1, g_2 \in G$ are close to each other if they act similarly on every function in Φ .**

Some work in progress: Changing (Φ, G) into (Ψ, H)



We wish to extend our theoretical approach to the case of different perception pairs. In order to do that, we consider each perception pair (Φ, G) as a **category** whose objects are the elements of the compact space Φ and whose arrows are the elements of the topological group G .

We have an arrow $g \in G$ from $\varphi_1 \in \Phi$ to $\varphi_2 \in \Phi$ if $\varphi_2 = \varphi_1 \circ g$.

Some work in progress: Changing (Φ, G) into (Ψ, H)



In our new context, each functor $F : (\Phi, G) \rightarrow (\Psi, H)$ is called a **Group Invariant Non-expansive Operator (GINO)** if:

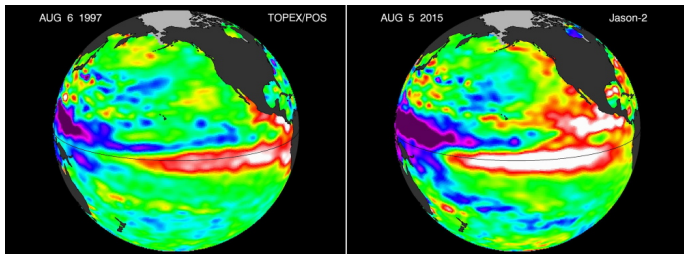
- F is group invariant: $F(\varphi \circ g) = F(\varphi) \circ F(g)$ for every $\varphi \in \Phi, g \in G$;
- F is non-expansive on Φ : $\|F(\varphi_1) - F(\varphi_2)\|_\infty \leq \|\varphi_1 - \varphi_2\|_\infty$ for every $\varphi_1, \varphi_2 \in \Phi$;
- F is non-expansive on G : $D_H(F(g_1), F(g_2))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

This definition extends the definition of GINO illustrated in my previous talk. We observe that in the previous definition we had $F(g) = g$ for every $g \in G$.

Some work in progress: Changing (Φ, G) into (Ψ, H)

We give an example of the use of the definition of GINO between two different perception pairs (Φ, G) , (Ψ, H) .

Let us assume to be interested in the comparison of the distributions of temperatures on a sphere, taken at two different times:



Let us also imagine that only two opposite points N, S can be localized on the sphere.

Some work in progress: Changing (Φ, G) into (Ψ, H)



In this case we can set

- $X = S^2$
- $\Phi =$ set of 1-Lischitzian functions from S^2 to a fixed interval $[a, b]$
- $G =$ group of rotations of S^2 around the axis $N - S$

We can also consider the “equator” of our sphere, represented as the space S^1 .

Therefore, we can also set

- $Y =$ the equator S^1 of S^2
- $\Psi =$ set of 1-Lischitzian functions from S^1 to $[a, b]$
- $H =$ group of rotations of S^1

Some work in progress: Changing (Φ, G) into (Ψ, H)




In this case we can build a simple example of GINO from (Φ, G) to (Ψ, H) by setting

- $F(\varphi)$ equal to the function ψ that takes each point x belonging to the equator S^1 to the average of the temperatures along the meridian containing x , for every $\varphi \in \Phi$;
- $F(g)$ equal to the rotation $h \in H$ of the equator S^1 that is induced by the rotation of the sphere, for every $g \in G$.

We can easily check that F verifies the properties defining the concept of group invariant non-expansive operator.

Some work in progress: The metric space of GINOs from (Φ, G) to (Ψ, H)



We can endow the set of all GINOs from (Φ, G) to (Ψ, H) with this metric: $d_{\mathcal{F}}(F_1, F_2) := \max \left\{ \sup_{\varphi \in \Phi} \|F_1(\varphi) - F_2(\varphi)\|_{\infty}, \sup_{g \in G} D_G(F_1(g), F_2(g)) \right\}$.

Theorem

The metric space of GINOs from (Φ, G) to (Ψ, H) is compact.

Corollary

The metric space of GINOs from (Φ, G) to (Ψ, H) can be ε -approximated by a finite subset.

Some work in progress: Extending the definition of $D_{\mathcal{F}}^{\text{match}}$ to the case $(\Phi, G) \neq (\Psi, H)$



The previous corollary opens the way to the computational approximation of the following pseudo-metric, which naturally extends the one defined in our previous talk.

Let us consider a set \mathcal{F} of GINOs from (Φ, G) to (Ψ, H) .

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

$$D_{\mathcal{F}}^{\text{match}}(\varphi_1, \varphi_2) := \sup_{F \in \mathcal{F}} d_{\text{match}}(\beta_k(F(\varphi_1)), \beta_k(F(\varphi_2)))$$

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

Some work in progress: Extending the definition of $D_{\text{match}}^{\mathcal{F}}$ to the case $(\Phi, G) \neq (\Psi, H)$



Proposition

$D_{\text{match}}^{\mathcal{F}}$ is a G -invariant and stable pseudo-metric on Φ .

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

The stability of $D_{\text{match}}^{\mathcal{F}}$ means that $D_{\text{match}}^{\mathcal{F}}$ is upper-bounded by the natural pseudo-distance d_G :

$$D_{\text{match}}^{\mathcal{F}}(\varphi_1, \varphi_2) \leq d_G(\varphi_1, \varphi_2) \leq \|\varphi_1 - \varphi_2\|_{\infty}.$$



Recap of my previous talk and some remarks

Work in progress on extending the theory of GINOs

Work in progress on the metric space of GINOs

Some final remarks



The metric space of GINOs

Our approach to G-invariant TDA is based on the availability of GINOs.

How could we build new GINOs from other GINOs?

A simple method consists in using the properties of functors and producing new GINOs by composition of other GINOs:

If F_1 is a GINO from (Φ, G) to (Ψ, H) and F_2 is a GINO from (Ψ, H) to (χ, K) , then $F_2 \circ F_1$ is a GINO from (Φ, G) to (χ, K) .

Building GINOs via 1-Lipschitzian functions



We can also produce new GINOs by means of a 1-Lipschitzian function applied to other GINOs:

Proposition

Assume that two perception categories (Φ, G) , (Ψ, H) are given. Let \mathcal{L} be a 1-Lipschitzian map from \mathbb{R}^n to \mathbb{R} , where \mathbb{R}^n is endowed with the norm $\|(x_1, \dots, x_n)\|_\infty := \max_{1 \leq i \leq n} |x_i|$. Assume also that F_1, \dots, F_n are GINOs from (Φ, G) to (Ψ, H) that coincide on the homeomorphisms in G . Let us define $\mathcal{L}^*(F_1, \dots, F_n)$ by setting $\mathcal{L}^*(F_1, \dots, F_n)(\varphi)(x) := \mathcal{L}(F_1(\varphi)(x), \dots, F_n(\varphi)(x))$. We also set $\mathcal{L}^*(F_1, \dots, F_n)(g) = F_1(g) = \dots = F_n(g)$ for every $g \in G$. If $\mathcal{L}^*(F_1, \dots, F_n)(\Phi) \subseteq \Psi$, then $\mathcal{L}^*(F_1, \dots, F_n)$ is a GINO from (Φ, G) to (Ψ, H) .

From this proposition the following three results follow.



Building new GINOs via translations, weighted averages and the maximum operator

Proposition (Translation)

Let F be a GINO from (Φ, G) to (Ψ, H) . Let us consider the operator F_b that is defined as $F_b(\varphi) = \varphi - b$ on Φ and as $F_b(g) = F(g)$ on G . If $F_b(\Phi) \subseteq \Psi$, then F_b is a GINO from (Φ, G) to (Ψ, H) for every $b \in \mathbb{R}$.

Proposition (Maximum)

Assume F_1, \dots, F_n are GINOs from (Φ, G) to (Ψ, H) , and that they coincide on the homeomorphisms in G . Then the operator F that is defined as $F(\varphi) = \max_i F_i(\varphi)$ on Φ and as $F(g) = F_1(g) = \dots = F_n(g)$ on G is a GINO from (Φ, G) to (Ψ, H) , provided that $F(\Phi) \subseteq \Psi$.



Building new GINOs via translations, weighted averages and the maximum operator

Proposition (Weighted average)

Assume that F_1, \dots, F_n are GINOs from (Φ, G) to (Ψ, H) , that they coincide on the homeomorphisms in G , and that $(a_1, \dots, a_n) \in \mathbb{R}^n$ with $\sum_{i=1}^n |a_i| \leq 1$. Then the operator F that is defined as $F(\varphi) = \sum_{i=1}^n a_i F_i(\varphi)$ on Φ and as $F(g) = F_1(g) = \dots = F_n(g)$ on G is a GINO from (Φ, G) to (Ψ, H) , provided that $F(\Phi) \subseteq \Psi$.

AN IMPORTANT CONSEQUENCE OF THIS LAST PROPOSITION:
The topological space of all G -invariant non-expansive operators from (Φ, G) to (Ψ, H) that coincide on G with a given homomorphism $\bar{F} : G \rightarrow H$ is not only COMPACT, but also CONVEX.

An interesting GINO in kD persistent homology



Previous propositions imply the following statement.

Proposition

Assume F_1, \dots, F_n are GINOs from (Φ, G) to (Ψ, H) , and that they coincide on the homeomorphisms in G . Assume also that $(a_1, \dots, a_n), (b_1, \dots, b_n) \in \mathbb{R}^n$, with $a_1, \dots, a_n > 0$, $\sum_{i=1}^n a_i = 1$ and $\sum_{i=1}^n b_i = 0$. Then the operator F that is defined as

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

on Φ and as $F(g) = F_1(g) = \dots = F_n(g)$ on G is a GINO from (Φ, G) to (Ψ, H) , provided that $F(\Phi) \subseteq \Psi$.

This result can be easily generalized from the case $\Phi \subseteq C^0(X, \mathbb{R}^m)$.

An interesting GINO in kD persistent homology



Let us now take $G = H$ and $n = m$ in the previous proposition. By considering the projection operators $F_i(\varphi) := \varphi_i$ for every $\varphi = (\varphi_1, \dots, \varphi_n) \in \Phi \subseteq C^0(X, \mathbb{R}^n)$ and setting $F(g) = g$ for every $g \in G$, we obtain the operator

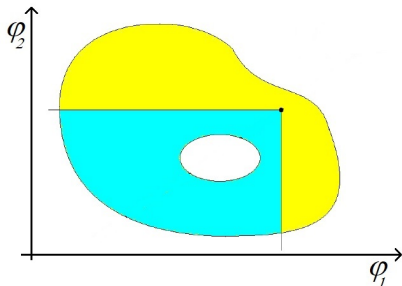
$$F(\varphi) = \max \left\{ \frac{\min_j a_j}{a_1} \cdot (\varphi_1 - b_1), \dots, \frac{\min_j a_j}{a_n} \cdot (\varphi_n - b_n) \right\}.$$

THIS OPERATOR IS IMPORTANT IN kD PERSISTENT HOMOLOGY.

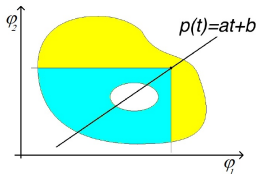
What is kD persistent homology?



kD persistent homology is the natural generalization of persistent homology to functions taking values in \mathbb{R}^k instead of \mathbb{R} . In plain words, in place of the sublevel sets associated with real numbers we consider sublevel sets associated with real vectors. This approach leads us to define k-dimensional persistent Betti number functions.



What is kD persistent homology?



The collection of the 1D-filtrations associated with the lines $p(t) = (a, 1 - a)t + (b, -b)$ such that $a, b \in \mathbb{R}$ with $0 < a < 1$ is equivalent to the 2D-filtration associated with the filtering function $p \mapsto (x(p), y(p))$.

[A. Cerri, B. Di Fabio, M. Ferri, P. Frosini, C. Landi, *Betti numbers in multidimensional persistent homology are stable functions*, *Mathematical Methods in the Applied Sciences*, **36**(2013),1543-1557.]

An interesting GINO in kD persistent homology



In other words, each multidimensional persistent Betti number function is equivalent to a family of 1-dimensional persistent Betti number functions. This reduction is formally done by means of the operator that takes each \mathbb{R}^n -valued function $\varphi = (\varphi_1, \dots, \varphi_n)$ to the \mathbb{R} -valued functions

$$\varphi_{(a,b)} := \max \left\{ \frac{\min_j a_j}{a_1} \cdot (\varphi_1 - b_1), \dots, \frac{\min_j a_j}{a_n} \cdot (\varphi_n - b_n) \right\}$$

with $a_1, \dots, a_n > 0$, $\sum_{j=1}^n a_j = 1$ and $\sum_{j=1}^n b_j = 0$.

This is exactly the operator that we considered previously. Therefore, the study of multidimensional persistent homology naturally leads to study the topological space of GINOs and its properties.



Recap of my previous talk and some remarks

Work in progress on extending the theory of GINOs

Work in progress on the metric space of GINOs

Some final remarks

Let us recap some good properties in our theory



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

This means that, in principle, the distance

$$D_{\text{match}}^{\mathcal{F}}(\varphi_1, \varphi_2) := \sup_{F \in \mathcal{F}} d_{\text{match}}(\beta_k(F(\varphi_1)), \beta_k(F(\varphi_2)))$$

can be computationally approximated.

We have also seen that $D_{\text{match}}^{\mathcal{F}}$ is **G -invariant and stable**, and that it can be **a proxy for the natural pseudo-distance d_G** .

The space of all GINOs that take (Φ, G) to (Ψ, H) and coincide on G with a given homomorphism $\bar{F} : G \rightarrow H$ is **convex**.

Some final remarks about the use of duality



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 $D_{match}^{\mathcal{F}}$, d_G cannot be approximated by another distance of the same kind. In other words, $D_{match}^{\mathcal{F}}$ 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 $D_{match}^{\mathcal{F}^*}$ that is not far from d_G and can be efficiently used for data retrieval, even if \mathcal{F}^* is not a good approximation of the set of all GINOs.

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

Some final remarks about the use of duality



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 $\mathcal{F}^{all}(\Phi, G)$ of all G -invariant non-expansive operators becomes “smaller and smaller”, so making the computation of $D_{match}^{\mathcal{F}^{all}(\Phi, G)}$ easier and easier, contrarily to what happens for the direct computation of d_G . In other words, the approach based on $D_{match}^{\mathcal{F}^{all}(\Phi, G)}$ seems to be of use exactly when d_G is difficult to compute in a direct way.

Some final remarks about the use of duality



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

$$D_{match}^{\mathcal{F}^*} \leq D_{match}^{\mathcal{F}} \leq d_G \leq d_H.$$

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



Open questions

After defining an observer as a collection of GINOs, our purpose consists in looking for methods to approximate the observer by a finite (and possible small) set of simple GINOs.

This leads us to the following open questions:

- How can we build a good library of GINOs?
- How can we find a method to choose a finite set \mathcal{F}^* of GINOs that allows for both a good approximation of the natural pseudo-distance d_G and a fast computation?
- **How can we provide a suitable probability theory for group invariant non-expansive operators?**

Further research is needed.



Conclusions

- 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. The operators are allowed to change both the space of functions and the invariance group.
- The functions describing the data can be compared by means of the natural pseudo-distance associated with any subgroup G of $\text{Homeo}(X)$.
- 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.
- **The topological space of GINOs deserves further research.**



Thanks for your attention!

