Arnau Mas

Contents

Chapter 1

The problem

1.1 The problem

The data that has been used for the analysis comes from a study in the field of radiomics. The aim of radiomics is to establish a relationship between a patient's response to treatment and what are known as *radiomic features* of a lesion. These are parameters which give a systematic description of a feature gathered from scans. The hypothesis is that there is information that can be extracted from this more systematic study of medical scans, as opposed to the more traditional qualitative observation by a medical professional.

On a practical level, the radiomic features are extracted from a 3d reconstruction of the lesion, built from a series of 2d sectional scans. It is then the work of a medical professional to manually delimit the contours of the actual lesion, which is known as the *region of interest*. This process is called *segmentation*. Once this is done, the actual work of computing the radiomic features takes place. For the data at hand, these had been extracted with the PyRadiomics Python package [2]. These features include shape parameters of the triangular mesh determined by the segmented region of interest, as well as statistical descriptors of the distribution of pixel intensity in the region of interest.

Chapter 2

The theory of homology

2.1 The theory of homology

2.1.1 The algebraic side: chain complexes and homology groups

The basic object in the theory of homology is a *chain complex*, which consists of a family of abelian groups $\{C_n\}_{n\in\mathbb{N}}$ together with a family of morphisms $\partial_n \colon C_n \to C_{n-1}$ such that $\partial_n \circ \partial_{n+1}$. We can then define the space of *n-cycles* as $\ker \partial_n$ and the space of *n-boundaries* as $\operatorname{im} \partial_{n+1}$. The *$n$-th homology group* of the complex is then

$$H_n := \ker \partial_n / \operatorname{im} \partial_{n+1}.$$

2.1.2 Simplices and simplicial complexes

For the purposes of data analysis, the simpler theory of simplicial homology. As we will see, it can be framed in purely combinatorial terms which is very useful when it comes to computation.

The basic geometric building block is the simplex. This is the generalisation of a triangle in two dimensions and a tetrahedron in three dimensions.

Definition 2.1 (Simplex)**.** A simplex of dimension n, or simply an n-simplex, generated by $n+1$ points of $\mathbb{R}^d$ which do not lie in an affine subspace of dimension n[1], is their convex hull, i.e. the smallest convex set which contains them. △

The n-simplex generated by the points the standard basis of $\mathbb{R}^{n+1}$, $e_0, \ldots, e_n$[2], is called

[1] which means the ambient dimension d must be at least n

[2] Because an n-simplex is generated by $n+1$-points, it will be convenient to number things starting at 0, as opposed to 1 as is more common in the rest of mathematics

the *standard n-simplex* and written Δ^n. It is easy to show that

$$\Delta^n = \left\{ (t_0, \dots, t_n) \in \mathbb{R}^{n+1} \mid \sum_{k=0}^{n} t_k = 1, \forall k \leq n \colon t_k \geq 0 \right\}.$$

The standard simplices provide a model for any other simplex. Indeed, if $\phi \colon \mathbb{R}^{n+1} \to \mathbb{R}^d$ is the linear map defined by $\phi(e_k) = p_k$ for $0 \leq k \leq n$ then $\phi(\Delta^n)$ is precisely the simplex generated by $p_0, \dots, p_n$. $(t_0, \dots, t_n)$ are called the *barycentric coordinates* of the point $\phi(t_0, \dots, t_n)$.

Orientation. For the purposes of homology, it is also important to keep track of the *orientation* of a simplex. We will use the idea of general simplices being the image of standard simplices to model this situation

Definition 2.2 (Ordered simplex)**.** An *ordered n-simplex* generated by $p_0, \dots, p_n \in \mathbb{R}^d$ is a map $\sigma \colon \Delta^n \to \mathbb{R}^d$ such that σ is the restriction to Δ^n of the linear map $\phi \colon \mathbb{R}^{n+1} \to \mathbb{R}^d$ given by $\phi(e_k) = p_k$, provided $p_0, \dots, p_k$ do indeed generate a simplex. △

Observe that giving an ordering to the vertices of a simplex completely determines an ordered simplex, thus we introduce the notation $[p_0, \dots, p_n]$ for an ordered simplex as defined above.

Consider now the natural action of the symmetric group, $\mathfrak{S}_n$, on $\mathbb{R}^n$ by permuting the elements of the standard basis. Thus, a reordering of an ordered n-simplex σ is of the form $\tau \circ \sigma$ for some $\tau \in \mathfrak{S}_{n+1}$. We will then say that two simplices σ_1 and σ_2 have the same *orientation* if there exists an even permutation $\tau \in \mathfrak{S}_{n+1}$ such that $\sigma_1 = \tau \circ \sigma_2$. It then follows, because the even permutations are an index 2 subgroup of $\mathfrak{S}_n$, that any n-simplex only has two possible orientations, as happens with manifolds, for instance. See fig. 2.1 for an example.

The faces of an n-simplex, σ, are the $n+1$ possible $n-1$-simplices generated by all but one of the vertices of σ.

Given a simplex $\sigma = [p_0, \dots, p_n]$, the $n+1$ possible simplices we get by removing one of the generators —i.e. $[p_1, \dots, p_n]$ and so on— are called the *faces* of σ. In barycentric coordinates, they correspond to setting one of the coordinates equal to 1.

Definition 2.3 (Simplicial complex)**.** A *simplicial complex* K is a collection of simplices of $\mathbb{R}^d$ such that

(i) if $\sigma \in K$ and $\tau \in K$ is a face of σ then $\tau \in K$,

(ii) for any $\sigma, \tau \in K$ then $\sigma \cap \tau \in K$.

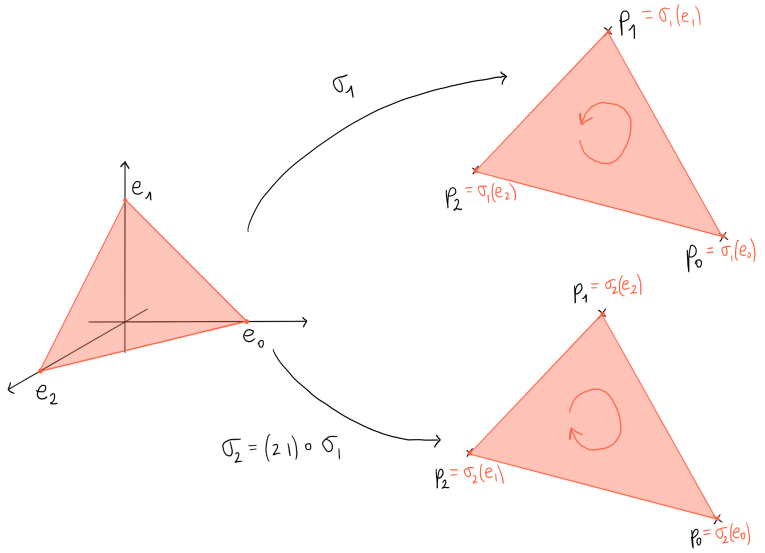


Figure 2.1: Two different orderings of a 2-simplex with the same vertices that determine different orientations.

△

The intuition is that the simplices that make up a simplicial complex are glued together only along their faces.

The simplices of K are often called its *cells*, and we will write K^n for collection of n-cells of K.

A simplicial complex K determines a topological space, called its *underlying space*, which is the subset of $\mathbb{R}^d$ determined by the union of all of the cells of K equipped with the subspace topology, which we will write as $|K|$. The theory of homology built using this class of spaces is known as *simplicial homology*. It should be noted, however, that homology can be implemented on much more general spaces by considering continous maps from the standard simplices into the space, which leads to the theory of singular homology.

2.1.3 Abstract simplicial complex

Notice that a simplex is always determined by its vertices, and vice versa, except for, of course, whenever the vertices do not generate a simplex, in the sense of definition 2.1. This requirement comes from thinking of simplices and simplicial complexes as geometric objects. But if we don't think of them in this way, and simply allow for any $n+1$ vertices to determine an n-simplex they become strictly combinatorial objects. In this case, we also have to lift the second condition in the definition of a simplicial complex, which was related to how the simplices are assembled together as geometrical objects and

therefore is senseless in this new context. This more general complexes are called *abstract simplicial complexes* and are purely combinatorial objects. From this viewpoint, the faces of a simplex are simply subsets of its vertices, thus, definition 2.3 now becomes

Definition 2.4 (Abstract simplicial complex)**.** An *abstract simplicial complex* K is any set closed under the relation of inclusion, i.e., if $\sigma \in K$ then if $\tau \subseteq K$, it must be the case that $\tau \in K$. △

For our purposes, we will assume that all of the simplices that make up a complex are finite, that there is a maximum possible dimension and that complexes are at most countably infinite, thus avoiding problems related to size.

Notice that to specify a complex it suffices to exhibit its *maximal simplices*, those simplices which are not the face of any other simplex, or equivalently those which are maximal with respect to inclusion[3].

For the purposes of homology, we will need to define some notion of orientation for a simplex. For the case of 1-simplices this is evident, simply pick one of the two orderings of its two vertices. For 2-simplices, if we think back to the geometrical picture, we could think of the clockwise and counterclockwise orientations. Then, any two possible orderings of the vertices give the same orientation if they are related by an even permutation, thus there are two possible orientations. This is easily generalisable to arbitrary dimension, and so we will say that two orderings of the vertices of a simplex have the same orientation if they differ by an even permutation, which implies there are only two possible orientations.

2.1.4 Special complexes

In the context of topological data analysis, one often wants to give some sort of geometric structure to a set of points, and a common way of doing this is by constructing a simplicial complex out of them. There are various ways of achieving this, which we now describe. For the remainder of this section, X will denote a finite subset of $\mathbb{R}^d$, which models the raw point cloud we begin with.

Čech complex. The ϵ-Čech complex of X, $C_\epsilon(X)$, is determined by the following prescription: $\sigma \subseteq X$ is in $C_\epsilon(X)$ if and only if

$$\bigcap_{p \in \sigma} B_\epsilon(p) \neq \emptyset$$

where $B_\epsilon(p)$ is the ball of radius ϵ centered at p.

[3]This is guaranteed by requiring that the dimensions of all simplices that make up the complex be bounded.

Vietoris-Rips complex. The prescription for the ϵ-Vietoris-Rips complex, $V_\epsilon(X)$, is as follows: $\sigma \subseteq X$ defines a simplex in $V_\epsilon(X)$ if and only if for every $p, q \in \sigma$

$$B_\epsilon(p) \cap B_\epsilon(p) \neq \emptyset.$$

In other words, a set of points generate a simplex whenever each of them is at distance at most 2ϵ[4] from the rest.

Notice that the ϵ-Čech complex is always a subcomplex of the ϵ-Vietoris-Rips complex.

Clique complex. If the points of our cloud are the vertices of some graph, G, we can use this information to build the so-called clique complex, $G(X)$, by declaring that $\sigma \subseteq X$ is a simplex of $G(X)$ if and only if it is a clique, i.e. a fully-connected subgraph of G. The Vietoris-Rips complex is a special case of this, since it is the clique complex of the graph obtained by declaring to points of X adjacent when they are at distance at most 2ϵ. This is the main c

2.1.5 Simplicial chain complexes and homology groups

Once we have the geometric ideas in place, we can start constructing the algebraic objects. As we discussed, we want to define a structure that captures (at least some of the) information encoded in the fundamental groups, but which is also algebraically simpler. Given a simplicial complex K, we can consider the free abelian group of (oriented) n-cells. In the case of $n = 1$ this is, in some sense which can be made precise, related to the abelianisation of the first fundamental group.

We want the orientation to play nicely with the group structure, so the group will not be totally free as we will require that for any oriented simplex σ, $-\sigma$ is the simplex with the same vertices but opposite orientation. This group is called the n-th simplicial chain group of the complex, $C_n(K)$.

The next step towards constructing the homology groups is defining the boundary morphisms.

Definition 2.5 (Boundary morphisms)**.** We define, for $n \geq 1$, the morphisms

$$\partial_n \colon C_n(K) \longrightarrow C_{n-1}(K)$$

by giving their action on an arbitrary generator as

$$\partial_n[p_0, \ldots, p_n] \coloneqq \sum_{k=0}^{n} (-1)^k [p_0, \ldots, \hat{p}_k, \ldots, p_n].$$

[4]Some texts use a slightly different convention such that the vertices of a clique in $V_\epsilon(X)$ are at distance at most ϵ (rather than 2ϵ) from each other

△

We need to show that all of this is indeed a chain complex, which is a consequence of the following result.

Lemma 2.6. *For any $n > 1$, $\partial_{n-1} \circ \partial_n = 0$.*

Proof. It suffices to show this on any generator, which is a simple calculation:

$$\begin{aligned}
(\partial_{n-1} \circ \partial_n)[p_0, \dots, p_n] &= \partial_{n-1} \left(\sum_{k=0}^{n} (-1)^k [p_0, \dots, \hat{p}_k, \dots, p_n] \right) \\
&= \sum_{k=0}^{n} (-1)^k \partial_{n-1} [p_0, \dots, \hat{p}_k, \dots, p_n] \\
&= \sum_{k=0}^{n} (-1)^k \left(\sum_{l=0}^{k-1} (-1)^l [p_0, \dots, \hat{p}_l, \dots, \hat{p}_k, \dots, p_n] \right. \\
&\quad \left. + \sum_{l=k+1}^{n} (-1)^{l-1} [p_0, \dots, \hat{p}_k, \dots, \hat{p}_l \dots, p_n] \right) \\
&= \sum_{\substack{k=0 \\ l<k}}^{n} (-1)^k (-1)^l [p_0, \dots, \hat{p}_l, \dots, \hat{p}_k, \dots, p_n] \\
&\quad - \sum_{\substack{k=0 \\ l>k}}^{n} (-1)^k (-1)^l [p_0, \dots, \hat{p}_k, \dots, \hat{p}_l, \dots, p_n] \\
&= 0
\end{aligned}$$

because both summands in the last line are the same. □

We will drop the subscripts from the boundary morphisms whenever they can be deduced from the context.

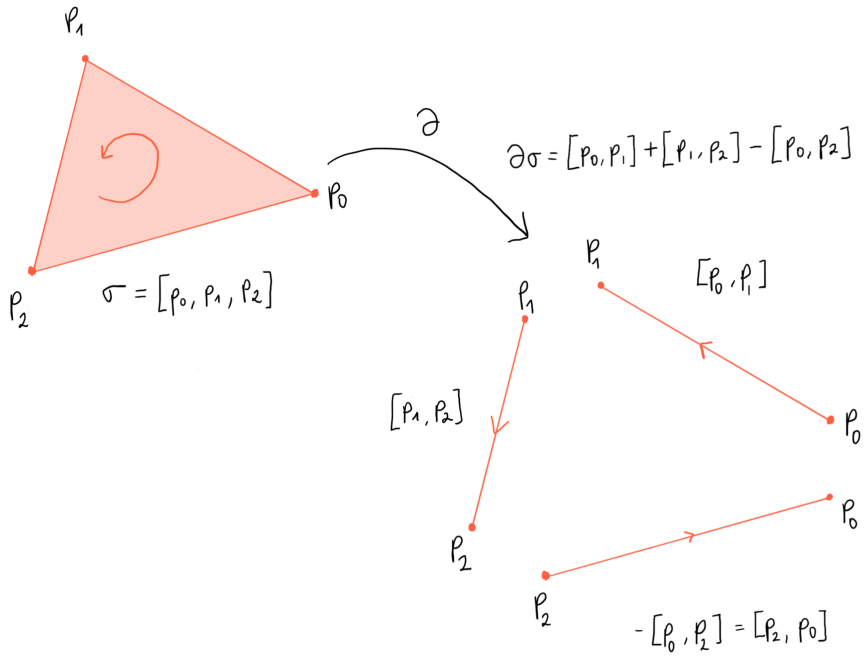
Thus, if d is the maximum dimension of the simplices of K, we have the simplicial chain complex

$$C_d(K) \xrightarrow{\partial_d} C_{d-1}(K) \longrightarrow \cdots \longrightarrow C_1(K) \xrightarrow{\partial_1} C_0(K) \longrightarrow 0$$

thus we can immediately define the simplicial homology groups by

$$H_n(K) := \ker \partial_n / \operatorname{im} \partial_{n+1}.$$

A variant of this is, instead of defining the chain groups as free abelian groups, defining the chain groups as vector spaces over a field F generated by the oriented simplices —with the condition that $-\sigma$ is σ with the opposite orientation—, which would be more sensibly



called chain spaces. Thus the homology groups would become homology spaces, which then allows for the definition of the *Betti numbers*:

$$\beta_n(X) = \dim H_n(X, F)$$

So what kind of innformation do we get from looking at the homology of a space? The 0th Betti number is exactly the number of connected components of the space. And the higher Betti numbers are, essentially, the number of higher-dimensional holes in the space. The best way to understand this is by looking at the homology of various common spaces. Of course, we have only developed the theory for simplicial complexes, but there are ways homology can be defined for (somewhat well-behaved) spaces.

The homology of $\mathbb{R}^n$ is

2.2 The idea of persistence

So far we have seen various ways of extracting information related to the shape of our data using homology. All of this ways, however, carry with them an ammount of arbitrary choice. For example, both the Vietoris-Rips and Čech complexes depend on a scale parameter ϵ, and the appropriate choice of ϵ is generally not clear from the data. The idea of persistent homology is to sidestep this problem altogether by considering the homology of the data at every scale, to determine which are the features really reflect geometric aspects and which are byproducts of background noise. The first will be,

roughly speaking, those which are present ar a large range of scales, i.e. those which are *persistent*. We now formalise these ideas.

2.2.1 Filtrations

Definition 2.7 (Filtration)**.** A *filtration* is a family of simplicial complexes $\{K_i\}_{i=0}^n$ such that K_i is a subcomplex of K_{i+1},

$$K_0 \subseteq K_1 \subseteq \cdots \subseteq K_n$$

where we will mostly assume $K_0 = \emptyset$. △

The inclusions $\iota_i^j \colon K_i \hookrightarrow K_j$, because of the functoriality of homology, give rise to maps between the homology groups at different steps of the filtration,

$$H_p(\iota_i^j) \colon H_p(K_i) \longrightarrow H_p(K_j)$$

which we will write as f_i^j for short. We say a certain class is *born at* i if it is not in the image of any f_k^i for any $k \leq i$. And we say it *dies at* j if it is in the kernel of f_i^j but not of f_i^k for $k < j$. The difference $j - i$ is called its *persistence* or *lifetime*.

This is a reflection at the level of homology of how the shape of the complex changes at every step of the filtration.

Chapter 3

The algorithm and its implementation

3.1 Justification of the algorithm

The main idea of the algorithm used to compute the persistent homology is taken from [1]. Recall that the homology groups are defined as a quotient of the subgroup of cycles by the subgroup of boundaries. But we can just as well consider the quotient of the *whole* chain group. More specifically, consider a filtration $\{K^i\}_{i=0}^N$ and its chain complex, $C_*(K^i)$. We then have the corresponding boundary morphisms,

$$\partial_p^i \colon C_p(K^i) \to C_{p-1}(K^i)$$

where the superscript i refers to the step of the filtration. refers to the step of the filtration. The group of p-cycles is $\ker \partial_p^i$ and the group of p-boundaries is $\operatorname{im} \partial_{p+1}^i$. Since $\partial_p \circ \partial_{p+1} = 0$ the boundaries are a subgroup of the cycles, which means the homology groups are well defined. But the boundaries are a subgroup of $C_p(K^i)$, which means we can define what we will call *prehomology groups* by

$$P_p(K^i) = C_p(K^i) / \operatorname{im} \partial_{p+1}^i$$

which is the result of identifying all homologous chains, not just the cycles.

Now, because $\operatorname{im} \partial_{p+1}^i \subseteq \ker \partial_p^i$, the boundary maps factor through the prehomology groups, which means the following is well-defined

$$\begin{aligned} \bar{\partial}_p^i \colon P_p(K^i) &\longrightarrow P_p(K^j) \\ [c] &\longmapsto [\partial_p^i c]. \end{aligned}$$

In particular,

$$\ker \bar{\partial}_p^i = H_p(K^i).$$

3.2 Mutual k-Nearest Neighbours Graph

As stated before, the original way of determining clusters in the point cloud was by finding the maximal cliques of the Mutual k-Nearest Neighbours graph, MkNN graph for short. We now describe this construction. Given a point cloud, C, its Nearest Neighbour directed graph is constructed according to the following prescription: there is an edge from p to q if and only if $d(p,q) = \min_{r \in C} d(p,r)$, so q is the closest point to p, its *nearest neighbour*. The reason the graph is directed is because the relation of being a nearest neighbour is not symmetric, as evidenced by the figure

This generalises to the k-Nearest Neighbours graph, in which each point is connected to its k-th nearest neighbours, that is, the k-th closest points to it. The resulting graph is still directed. We obtain the undirected *Mutual* k-Nearest Neighbours graph by connecting two points if and only if there are edges between them in both directions, i.e. if they are *mutual* k-nearest neighbours.

The procedure to encode this graph is as follows. First, the distance matrix, D, for the data set is computed. If we index the points of the cloud by $C = \{p_i\}_{i=1}^n$, then this matrix contains the difference between every pair of points:

$$D_{ij} = d(p_i, p_j).$$

This can be done with NumPy and is implemented as a method of the `Cloud` class, which models a point cloud. Then, again using NumPy, the indices of each column of D are sorted into a new matrix such that the i-th column of this new matrix lists the points of C by increasing order of distance to p_i. In particular, the first k are p_i's k-th nearest neighbours. From this we can then construct the adjacency matrix of the kNN graph, A, which is in general not symmetric since the graph is directed. The adjacency matrix of the MkNN graph, M, is given by

$$M = A \odot A^\top$$

where $\odot$ denotes the entrywise prodct. Indeed, we have

$$M_{ij} = A_{ij} A_{ji}^\top = A_{ij} A_{ji}$$

so that M_{ij} is nonzero if and only if both A_{ij} and A_{ji} are nonzero, i.e. if there is an edge going from p_i to p_j and one from p_j to p_i. Furthermore

$$M_{ij} = A_{ij} A_{ji} = A_{ji} A_{ij} = M_{ji}$$

which means that M is symmetric and therefore the adjacency matrix of an undirected graph, as we claimed. This adjacency matrix is used to represent the MkNN graph as a NetworkX graph.

3.3 Building the filtration

The work of constructing the filtration is packaged up in the `Filtration` class. This class is passed an instance of `Cloud`. The most straightforward way to represent a filtration is to simply store an ordered list of the simplices, in this case cliques, which appear at every step. There are more efficient data structures that can be used, see , but since the data sets analysed are of small size, this more naive approach sufficed.

The NetworkX package has methods which can compute every clique in a graph. In a loop over k from 1 to one minus the number of points in the cloud, the correspoinding MkNN graph is computed and a list of its cliques is extracted. Of these, all those which are new and have size at most one more than the ambient dimension —such that the simplices they determine have dimension at most the ambient dimension— are appended to a list. The `Simplex` wraps a clique as a list of its points as well as keeping track of the step, k, at which the clique is born. In addition it implements methods used in the actual computation of the homology groups.

This list is then sorted by birth, such that cliques born earlier appear firts, and cliques with the same birth are sorted by increasing size. This guarantees that any simplex is always preceded by its faces, which is required for later computations.

3.4 Computing the homology

The classes `Simplex` and `Chain` can be used to calculate with the chain groups. As mentioned before, a `Simplex` simply wraps a clique as well as its birth. It also has the property `faces`, which returns a list of the faces of the simplex.

The elements of the chain group are linear combinations of simplices (of the same dimension), but since we are working over $\mathbb{F}_2$, this amounts to a list of simplices, which is what `Chain` stores. Furthermore, this class implements the addition of chains which, again because the field we ar working over has characteristic 2, reduces to taking the symmetric difference of the lists of simplices of the two chains we are adding. Indeed, if we have two chains of the form $c_1 = \sum_{i=1}^{n} \epsilon_i \sigma_i$ and $c_2 = \sum_{i=1}^{n} \delta_i \sigma_i$ with $\epsilon_i, \delta_i \in \mathbb{F}_2$, then

$$c_1 + c_2 = \sum_{i=1}^{n} (\epsilon_i + \delta_i)\sigma_i$$

which means the coefficient of σ_i in $c_1 + c_2$ is $\epsilon_j + \delta_j$. And this will be 1 provided only one of ϵ_j or δ_j is equal to 1, and will be 0 whenever *both* ϵ_j and δ_j are 1 or 0. So σ_j will be present in $c_1 + c_2$ whenever it is present in c_1 or c_2, but not both.

This makes it very easy to implement the boundary morphisms. For a single simplex, wrap the list of faces inside a `Chain` object. And for a larger chain, add the boundaries of each of its constituent simplices. Again, this works because we are taking coefficients from $\mathbb{F}_2$, so that, as explained in the alternating signs that appear in the definition of the boundary all disappear.

With all of this in place we can

Bibliography

[1] Martín Campos. “Filtracions en homología persistene mediante estimadores kernel de densidad”. Universitat Autònoma de Barcelona, Facultat de Matemàtiques, June 2018.

[2] J. J. M. van Griethuysen et al. “Computational radiomics system to decode the radiographic phenotype”. In: *Cancer Research* 77.21 (2017), pp. 104–107. DOI: 10.1158/0008-5472.can-17-0339.