November 5, 2009

Homological methods in manifold learning

Peter Saveliev
Department of Mathematics, Marshall University, Huntington, WV 25755, USA
saveliev@marshall.edu

Suppose we are given a point cloud K in a euclidean space of dimension d. Suppose also that we are given a threshold r so that any two point within r from each other are to be considered "close". Then each pair of points like that is connected by an edge. Further, if three points are connected to each other, pairwise, by edges, we add a face spanned by these edges. If there are four, we add a tetrahedron, etc. From vertices to tetrahedra and beyond, we call them 0-, 1-, 2-, ..., *d-cells*. The result is a *cell complex* made of k-cells that are attached to each other along $(k-1)$-cells. (Note that there are many ways to build a cell complex from a point cloud.)

What do we want to learn about K? We want to quantify its topology by means of the so-called *Betti numbers*: B_0 is the number of connected components in K; B_1 is the number of holes or tunnels (1 for letter O or the donut; 2 for letter B and the torus); B_2 is the number of voids or cavities (1 for both the sphere and the torus), etc.

How does one compute Betti numbers? The methods come from *homology theory*. One starts by considering the collection $C_k(K)$ of all combinations of cells of the same dimension k, called *chains,*. Together they form a *chain complex* $C_*(K)$. A k-chain can be recorded as an N-vector, where N is the total number of k-cells in K. As an illustration, a component of this vector is 1 if the corresponding cell is present, -1 if it present with the opposite orientation, and 0 if it is absent. The boundary of a k-chain is the chain comprised of all $(k-1)$-faces of its cells. Then the *boundary operator* $\partial : C_k(K) \to C_{k-1}(K)$ acts on the chain complex and is represented by a matrix.

From the chain complex $C_*(K)$ the homology group is built by means of the standard tools of linear algebra. To capture the topological features one concentrates on *cycles*, i.e., chains with zero boundary, $\partial A = 0$. Further, one can verify whether two given k-cycles A and B are *homologous*: the difference between them is the boundary of a $(k+1)$-chain $T : A - B = \partial T$. In this case, A and B belong to the same *homology class* $H = [A] = [B]$. Examples of homologous A and B are: two vertices in the same component of any complex ($k = 0$); two longitudes of the torus, but not a longitude and a latitude ($k = 1$); the inner surface and the outer surface of a "thick" sphere ($k = 2$). The totality of these equivalence classes in each dimension k form the k-th *homology group* $H_k(K)$ of K, collectively $H_*(K)$. Commonly, $H_k(K)$ is simply a vector space and its dimension is equal to the corresponding Betti number B_k.

In addition to computing the Betti numbers as the global topological characteristics of the complex, homology theory can also provide local information. For every small patch P of K we compute the *relative homology* $H_*(K, K\backslash P)$ by, essentially, collapsing the complement of P to a single point. Then, if K is a manifold, its *dimension* is equal to n provided: $H_n(K, K\backslash P) \neq 0$ and $H_k(K, K\backslash P) = 0$ for $k \neq n$.

The methods for computing homology groups are well developed. In real life however the point clouds are noisy and one needs to evaluate the significance of the topological features in this uncertain environment. We measure the "robustness" of the Betti numbers as follows. Instead of using a single threshold and studying a single cell complex, we consider all thresholds and all possible cell complexes. The idea is to analyze all of them, then pool the topological features together in a single structure, and finally pick the features that lie within the user's choice of an acceptable level of noise. Homology theory provides a solution for this problem.

First we combine the homology groups of all the cell complexes into one structure. For example, we may have a sequence of complexes:

$$K^1 \hookrightarrow K^2 \hookrightarrow K^3 \hookrightarrow K^4 \hookrightarrow \ldots \hookrightarrow K^s,$$

where the arrows are inclusions: $i^{n,n+1} : K^n \hookrightarrow K^{n+1}$, and let $i^{n,m} : K^n \hookrightarrow K^m$ be defined as the compositions. This structure $\{K^n\}$ is called a *filtration*. Further, each of these inclusions generates a *homology homomorphism:* $i^{n,m}_* : H_*(K^n) \hookrightarrow H_*(K^m)$. Commonly, these are simply

linear operators represented by matrices. Then we have a sequence of homology groups connected by homomorphisms:

$$H_*(K^1) \to H_*(K^2) \to \ldots \to H_*(K^s) \longrightarrow 0.$$

The *homology group of filtration* $\{K^n\}$ captures all homology classes of all the complexes in a compact way:

$$H_*(\{K^n\}) = \ker i_*^{1,2} \oplus \ker i_*^{2,3} \oplus \ker i_*^{3,4} \oplus \ldots \oplus \ker i_*^{s,s+1}.$$

Indeed, from the homology group of each complex we take only the elements that are about to die (go to 0). Since each dies only once, there is no double-counting. Since the sequence ends with 0, we know that everyone will die eventually. Thus every homology class appears once and only once.

Next, the significance of a homology class in the sequence is measured by how long it lives before it ends up at 0. This number p is called the *persistence* of the homology class x: $i_*^{n,n+p}(x) = 0$ and $i_*^{n,n+p-1}(x) \neq 0$. Given a positive integer p, the *p-noise group* $N_*^p(\{K^n\})$ is comprised of the homology classes with the persistence less than p :

$$N_*^p(\{K^n\}) = \ker i_*^{1,1+p} \oplus \ker i_*^{2,2+p} \oplus \ker i_*^{3,3+p} \oplus \ldots \oplus \ker i_*^{s,s+p},$$

and the *p-persistent homology group* is

$$H_*^p(\{K^n\}) = H_*(\{K^n\})/N_*^p(\{K^n\}).$$

In other words: if the difference between two homology classes is deemed noise, they are equivalent.

The homology group of the filtration can be computed as:

$$H_*(\{K^n\}) = H_*(C_*(\{K^n\})),$$

where $C_*(\{K^n\}) = \bigoplus_n C_*(K^n)$ is the graded module of chains over the ring of polynomials $k[t]$ with $t \cdot x = (0,\ i_*^{1,2}(x_1),\ i_*^{2,3}(x_1), ...,\ i_*^{s-1,s}(x_1))$, where $x = (x_1, x_2, ..., x_s)$ and $x_n \in C_*(K^n)$. The method has been implemented as a Java-based computer program called jPlex with computational cost of $O(N^3)$, where N is the number of cells in K.

Alternatively, the homology group is computed via the *mapping cone of filtration* $\{K^n\}$, which is the following chain complex

$$Cone(\{K^n\}) = C_{*+s}(K^1) \oplus C_{*+s-1}(K^2) \oplus ... \oplus C_{*+1}(K^{s-1}) \oplus C_*(K^s).$$

The mapping cone captures this difference between the chain complexes of the elements of the filtration: everything in $C_*(K^1)$ is removed unless it also appears in $C_*(K^2)$ under i_*. Then the homology group of filtration $\{K^n\}$ is computed as:

$$H_*(\{K^n\}) = H_*(C_*(K^1) \oplus Cone(\{K^n\})).$$

A two-parameter "filtration" $\{K^{n,m}\}$ can be built from a point cloud: $n = n(r)$, where r is the radius of the ball, and $m = m(d)$, where d is the density of the cloud. Homology theory allows us to handle this situation as well. Suppose we have a partially ordered set I and a collection of complexes indexed by I, $\{K^n\}_{n \in I}$, with the inclusions $i^{n,m} : K^n \to K^m$ for all $n \leq m$. Then the homology of this "poset filtration" is

$$H_*(\{K^n\}) = \bigoplus_n \bigcap_{m \geq n} \ker i_*^{n,m}.$$

The noise groups and the persistence groups are defined as before.

REFERENCES

[1] G. Bredon, Topology and Geometry, Springer Verlag, 1993.

[2] G. Carlson, Topology and data, Bulletin of the Amer. Math. Soc., Vol. 46, No.2, pp. 255-308.

[3] T. Kaczynski, K. Mischaikow, and M. Mrozek, Computational Homology, Appl. Math. Sci. Vol. 157, Springer Verlag, NY 2004.