Analysis of high dimensional data via Topology

Louis Xiang

Oak Ridge National Laboratory

Oak Ridge, Tennessee

Contents

Abstract iii

1 Overview 1

2 Data Set 1

3 Simplicial Complex 5

4 Computation of homology 7

5 Analysis 10

6 Conclusion 10

7 Literature Cited 10

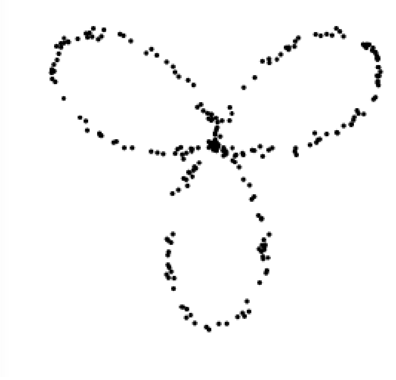
8 Acknowledgements 10

ABSTRACT

In this study we will focus on computing the topological invariant of a high dimensional data set. By this kind of topological analysis, we can know the qualitative result about the data set. Generally speaking, it can be an aid to visualisation of high dimensional data. We will use the medical data set as an example to show how the method describe the shape of the data set.

1 Overview

Given a point-cloud dataset from a space X, it is always reasonable to build a simplicial complex S to realize the topological structure of X. For example, the figure below has three loop-shaped petals.



How to indicate this kind of topological information automatically and correctly? In this paper, we will concentrate on building the simplicial complex and try to compute the $Betti$ numbers of this data set.

The Section 1 is taken up with how to deal with the data set to get 8-dimensional suitable point-cloud. Section 2 motivates and describes some background materials about simplicial complex and how to build it. Section 3 is about the $Betti$ number and how to compute the $homology$.

2 Data Set

2.1 The patches of data space

The medical data that we use records 42 kinds of factors of all the patients. But the problem is that the hospital did not give all these kinds of measurement at each time. Even some of the factors have been measured for only a few times. Below is a series of steps performed to obtain the data matrix from the set of reasonable data:

- Create a matrix with 43 columns, where the first column represent the time and the other

42 columns are used to record the factors.

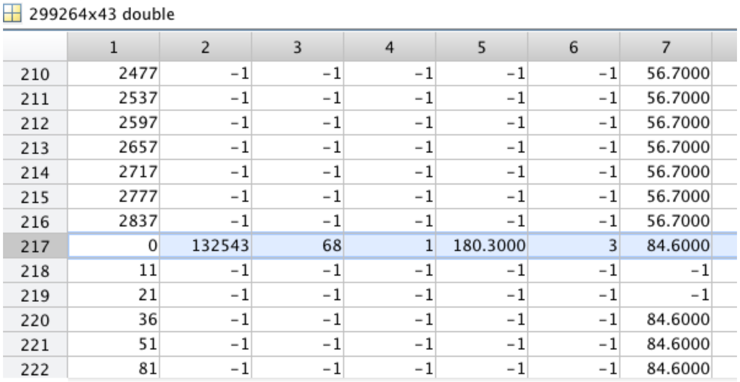
- Whenever a measurement is taken, then add a row in the matrix with missing entries filled by -1.

- Select the 8 columns from the matrix with less number of $-1s$ (i.e. the factor which has been recorded at the most of the time).

Remark Some columns can be combined together, for instance, the sys and $NIsys$, which is just the sys measured by different ways. The method is just taking average of the entries if these two entries are not -1 at the same time, if one of these two is -1, then use the other entry to replace it.

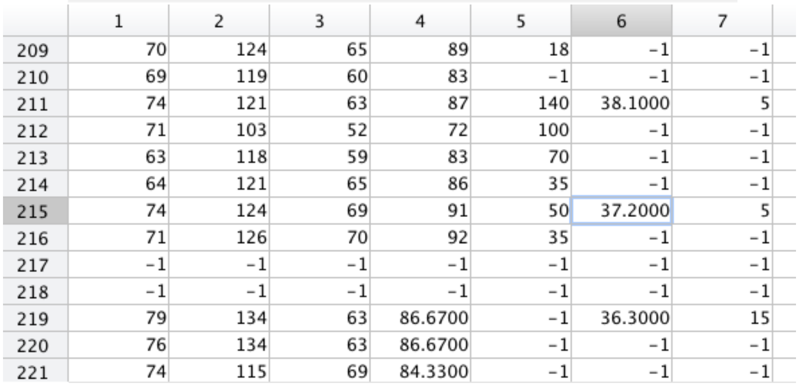
2.2 Interpolation of missing data

Dealing with the missing data is always a hot topic in nowadays research. In this topic we just use the linear interpolation to solve the problem of missing data. For example, like the matrix below,



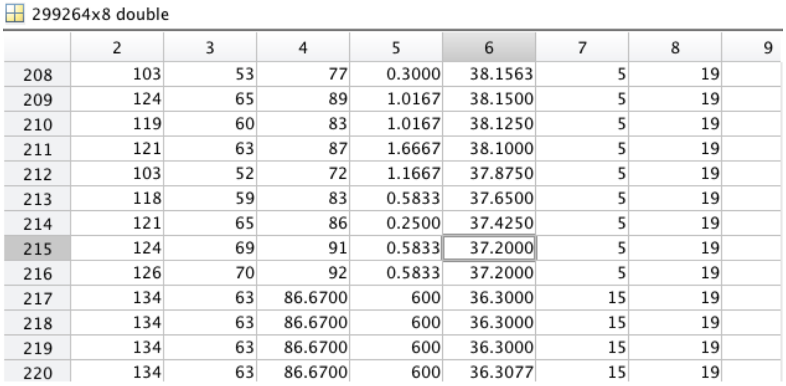
299264x43 double

	1	2	3	4	5	6	7
210	2477	-1	-1	-1	-1	-1	56.7000
211	2537	-1	-1	-1	-1	-1	56.7000
212	2597	-1	-1	-1	-1	-1	56.7000
213	2657	-1	-1	-1	-1	-1	56.7000
214	2717	-1	-1	-1	-1	-1	56.7000
215	2777	-1	-1	-1	-1	-1	56.7000
216	2837	-1	-1	-1	-1	-1	56.7000
217	0	132543	68	1	180.3000	3	84.6000
218	11	-1	-1	-1	-1	-1	-1
219	21	-1	-1	-1	-1	-1	-1
220	36	-1	-1	-1	-1	-1	84.6000
221	51	-1	-1	-1	-1	-1	84.6000
222	81	-1	-1	-1	-1	-1	84.6000



	1	2	3	4	5	6	7
209	70	124	65	89	18	-1	-1
210	69	119	60	83	-1	-1	-1
211	74	121	63	87	140	38.1000	5
212	71	103	52	72	100	-1	-1
213	63	118	59	83	70	-1	-1
214	64	121	65	86	35	-1	-1
215	74	124	69	91	50	37.2000	5
216	71	126	70	92	35	-1	-1
217	-1	-1	-1	-1	-1	-1	-1
218	-1	-1	-1	-1	-1	-1	-1
219	79	134	63	86.6700	-1	36.3000	15
220	76	134	63	86.6700	-1	-1	-1
221	74	115	69	84.3300	-1	-1	-1

The first one is the original matrix whose first column represent the time(See the time change from 2837 to 0, that means the change of patient.),the second one denoted by A is the most relevant 8 columns that we choose. By the linear interpolation, the entries between the 211 and 215 rows in the column 6 should be 37.875, 37.65 and 37.425. Also, care must be taken if we deal with the situation when there is a change-of-patient happened. Since there is a change-of-patient happening at the row 217, instead of doing linear interpolation between $A(215, 6)$ and $A(219, 6)$, we replace the -1 by $A(215, 6)$ and $A(219, 6)$ because we are not allowed to mix up the different situation between different patients. The result is as follows.



299264x8 double

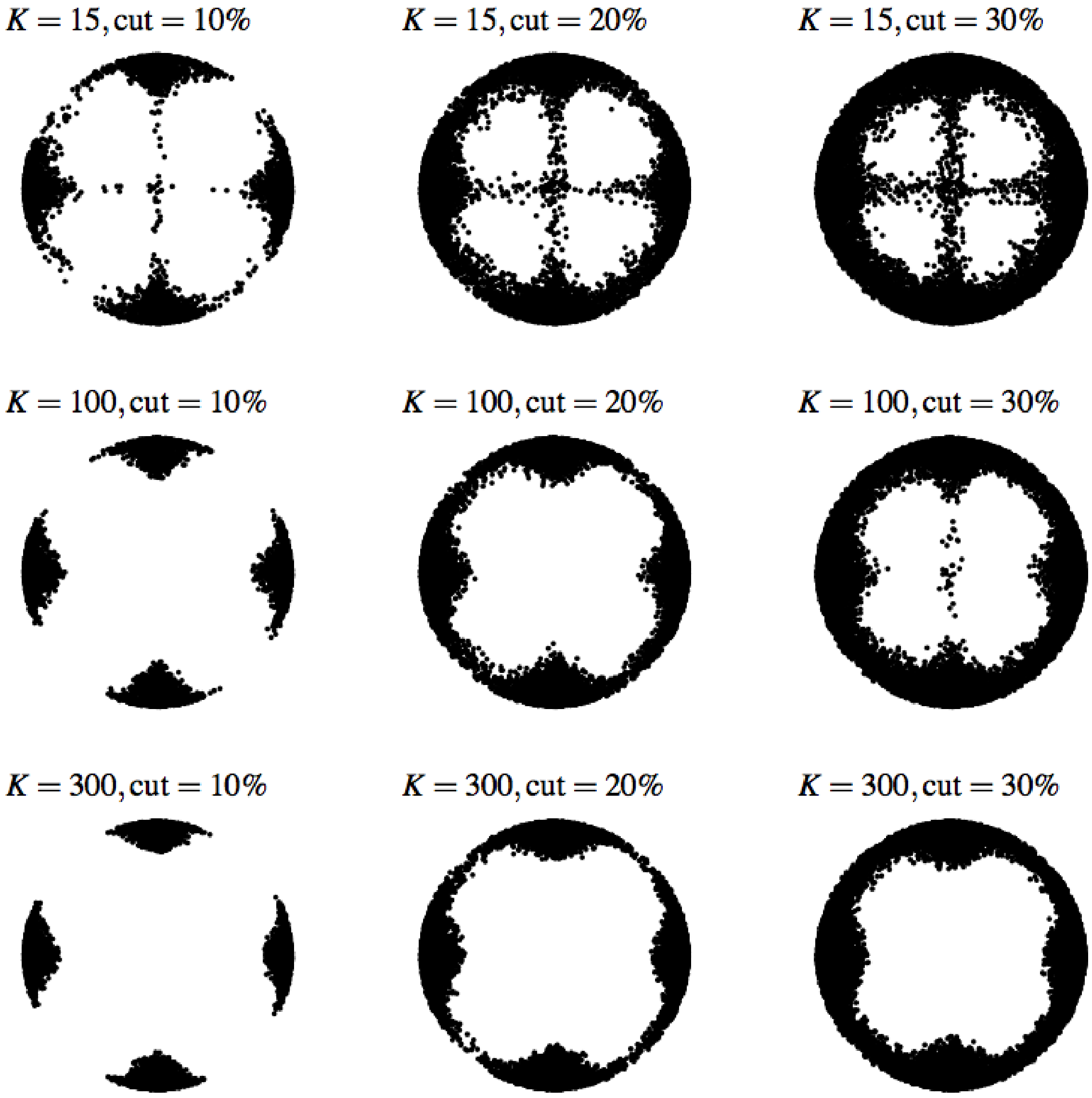
	2	3	4	5	6	7	8	9
208	103	53	77	0.3000	38.1563	5	19	
209	124	65	89	1.0167	38.1500	5	19	
210	119	60	83	1.0167	38.1250	5	19	
211	121	63	87	1.6667	38.1000	5	19	
212	103	52	72	1.1667	37.8750	5	19	
213	118	59	83	0.5833	37.6500	5	19	
214	121	65	86	0.2500	37.4250	5	19	
215	124	69	91	0.5833	37.2000	5	19	
216	126	70	92	0.5833	37.2000	5	19	
217	134	63	86.6700	600	36.3000	15	19	
218	134	63	86.6700	600	36.3000	15	19	
219	134	63	86.6700	600	36.3000	15	19	
220	134	63	86.6700	600	36.3077	15	19	

`Remark` Some of columns may have different characteristic when people measure it. For example, for the Urine, which has been accumulated in a bag. So different method should be taken when measuring this factor. In this paper we convert the original entries to urine rates (ml/min), and then do the linear interpolation.

2.3 Data Selection

It must be pointed out that direct application of simplicial complex approximation to the 310^6 data points will unfortunately lead to wrong detection since there are points distributed far away from the high-density regions.These kind of problem is very possible. For example, some measurements mix up the millilitre and litre which will be destined to give a large distance between other points. To obtain a high-density subset, we rely on a simple density function $\rho_K(x) = |x - x_K|$ where x_K is the K-th nearest point of x. The choice of K affects the results qualitatively which can be shown in the following example. The different choice of the objective data are projected onto the first two coordinates of $\mathbf{R}^8$. The nine panels show the 10 percent, 20 percent and 30 percent of points having the smallest values of $\rho 15$, $\rho 100$ and $\rho 300$.

The 30 cut with K = 300 appears to be centred entirely on an annulus. But If K = 15, on the other hand, there is a cross-like feature that is already present once the cut is large enough for the annulus to be fully formed.



Algorithm: The algorithm to obtain the $X(K, p)$:

- Let A be the matrix which contains the most relevant 8 columns for each measurement, then we can regards each row as a point belongs to $R8$.
- Find the distance matrix d where $d_{ij} = d(x_i, x_j), x_i$ and x_j is the i, j rows in the matrix. Then d will be a 300,000*300,000 matrix.

- Rearrange the d such that entries in each row grow from small to large.
- Take out the K-th column, and put it in order again form small to large, record the point which is in the top p-percent in the rearrangement column.Then these points form the $X(K, p)$.

`Remark` This step is of big significance for obtaining the topological characteristics of the data. Some of the data is a kind of mess, and some are because of the wrong measurement taken by the hospital (look at the 217 row and 5 column of the matrix in the third figure in the section 2.2, it seems very unreasonable) . The $X(k, p)$ can easily help us to take away these kinds of stuff. Also, it should be mentioned that these kind of point will cause a big trouble when we pick out the landmark points which we will discuss later.

3 Simplicial Complex

3.1 Background knowledge about simplicial complex

An abstract simplicial complex S is specified by the following:
An abstract simplicial complex is a finite collection of sets A such that $\alpha \in A$ and $\beta \subset \alpha$ implies $\beta \in A$.
For example, a tetrahedron with vertex A, B, C, D has simplicial complex $\{\{A\}, \{B\}, ..., \{D\}$, $\{A, B\}, .., \{C, D\}, \{A, B, C\}, ..., \{B, C, D\}, \{A, B, C, D\}\}$ which is the power set of the set $\{A, B, C, D\}$.
K-simplex: A k-$simplex$ is the convex hull of $k+1$ affinely independent points, $\sigma = conv\{u_0, u_1, ..., u_k\}$.
For example, $vertex$ for 0-simplex, $edge$ for 1-simplex, $triangle$ for 2-simplex, and $tetrahedron$ for 3-simplex, see the figure below.

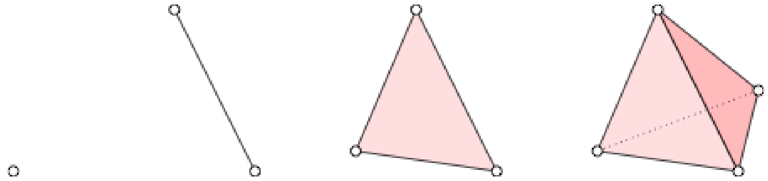


Figure III.1: From left to right: a vertex, an edge, a triangle, and a tetrahedron. We note that an edge has two vertices, a triangle has three edges, and a tetrahedron has four triangles as faces.

3.2 selection of landmarks

We recommend obtaining the landmark set by using the $maxmin$. method.Maxmin is the following inductive procedure. Initialise by selecting $l_1 \in Z$ randomly. For each $i \geq 2$, if $l_1, l_2, ..., l_{i-1}$ have been chosen, let $l_i \in Z \setminus l_1, l_2, ..., l_{i-1}$ be the data point which maximises the function $f(x) = \min_{1<=j<=i-1} D(x, j)$ where D is the metric. Continue in this way until the desired number of landmark points have been chosen.

The number of landmarks should be chosen by setting a lower bound on the ratio N/n. We do not have a systematic answer to what this lower bound should be, but N/n $\geq$220 seems to work quite well for data sampled from a two - dimensional surface.

3.3 Building the simplicial complex

Let D be an $n \times N$ matrix of non-negative entries, regarded as the matrix of distances between a set of n landmarks and N data points. We define the (strict) witness complex $W_\infty(D)$, with vertex set {1,2,...,n}, as follows:

- The edge σ =$[ab] \in W_\infty(D)$ iff there exists a data point $1 \leq i \leq 2N$ such that D(a,i) and D(b,i) are the smallest two entries in the i-th column of D, in some order.

- For any p: suppose all the faces of the p-simplex σ = $[a_0, a_1, ..., a_k]$ belong to $W_\infty(D)$. Then $\sigma \in W_\infty(D)$ iff there exists a data point 1$\leq$ i$\leq$ N such that $D(a_0, i), ..., D(a_p, i)$ are the smallest $p + 1$ entries in the i-th column of D.

`Remark` Keep in mind that the construction of distance matrix D can be done by the Euclidean or any other metric. For instance, you can use the intrinsic graph metric, which is defined by computing shortest paths in a suitable graph G on the set of all data points. We will not go into it too far.

4 Computation of homology

The Computation of homology includes a lot of theory about topology. And this kind of computation can be done by $Javaplex$ which is a MATLAB program mainly developed by the Computational Topology workgroup at Stanford University. We will include some brief introduction to compute the homology.

4.1 Chain, cycle and boundary

Let K be a simplicial complex.The p-$chain$ is defined to be a formal sum of p-$simplices$ in K. The p-chains together with the addition operation form the group of p-chains denoted by C_p=$\{\sum_i a_i v_i | a_i \in \mathbf{Z}_2, v_i$ are p-simplices in K and $\mathbf{Z}_2$ is the $modulo$ $2\}$. The boundary of a k-simplex σ is the set contains all its $(k-1)$-dimensional faces which is a (k-1)-chain and it is denoted by $\partial_k(\sigma)$. The boundary of the p-$chain$, ∂_k(c), is defined to be $\sum_{\sigma \in c} \partial_k(\sigma)$. Therefore we can write $\partial_p : C_p \to C_{p-1}$ as the boundary map which takes a p-$chain$ to a (p-1)-$chain$. The chain complex is the sequence of chain groups connected by boundary homomorphisms,

$$\ldots \xrightarrow{\partial_{p+2}} \mathsf{C}_{p+1} \xrightarrow{\partial_{p+1}} \mathsf{C}_p \xrightarrow{\partial_p} \mathsf{C}_{p-1} \xrightarrow{\partial_{p-1}} \ldots$$

Having the definition of the map ∂_k, we recall that the kernel of this is the collection of $k - chains$ with empty boundary and the image set of ∂_k is the collection of $(k-1) - chains$ that are boundaries of k-chains, i.e.

ker ∂_k = $\{c \in C_k | \partial_k(c) = \emptyset\}$,

Im ∂_k = $\{\partial_k(c) \in C_{(k-1)}|$ for some $c \in \partial_k\}$.

Actually,a p-$cycle$ is a p-$chain$ with empty boundary, so as the same, we denote the group of p-$cycles$ with addition operation as Z_k which is just ker ∂_k. Also, a p-$boundary$ is a p-$chain$

that is the boundary of $(p+1)$-*chain* and with addition operation it gives us the group of p-*boundaries*, denoted by B_p which is just Im ∂_{k+1}. The $k - cycle$ is the k-chain in the ker ∂_k and a k-boundry is a k-chain in the Im ∂_{k+1}.

`Fundamental lemma of topology`: For $\forall$ integer p and (p+1)-*chain* c, $\partial_p \partial_(p+1)c$ = 0. This implies that $B_k \subset Z_k \subset C_k$, then we can have a relation illustrated by the figure below.

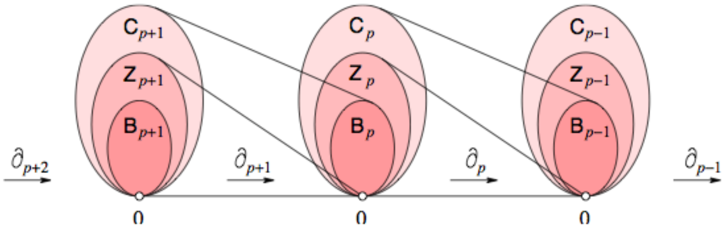


Figure IV.1: The chain complex consisting of a linear sequence of chain, cycle, and boundary groups connected by homomorphisms.

4.2 Homotopy groups

The k-th *homotopy group* is the k-th cycle group divided by the k-th boundary group, i.e. $H_k = Z_k/B_k$. The $Bettinumber$ of K β_k is the rank of the k-th homology group, β_k = rankH_k where H_k = Z_k/B_k.

There is a Betti number for each integer k. If $B_0 = i$, that means theres is i connected component in the data set, and if $B_k = j$, then there are j $k - dimensional$ holes in the data set. Now the goal is to find the rankH_k. Note that

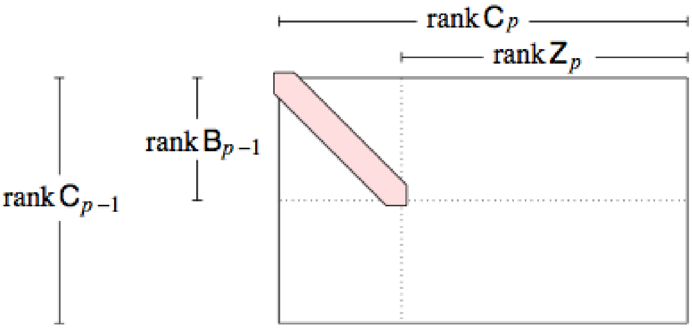
1. $rank\ H_k = rank\ Z_k - rank B_k = null\ \partial_k - rank\ \partial_{k+1}$.

2. $rank(A)$ + $Null(A)$ = N, N is the number of rows of the $transition\ matrix\ A$.

From these two important results, we could calculate the $Betti$ number and suggest the topological characteristics of the data set.

4.3 Algorithm

From linear algebra, we know that $Gaussian - Elimination$ method is always helpful for us to find a rank of the matrix. Care must be taken since all the entries in the matrix are belongs to $\mathbf{Z}_2$, i.e. 0 and 1, so $1 + 1$ should give 0 instead of 2 for modulo 2 arithmetic. The optimal matrix is the $Smithnormalmatrix$ which has an initial segment of the diagonal filled by 1 and everything else is 0, as in Figure below.



To $Smithnormalmatrix$, the basic idea is to move 1 to the upper left corner, and we can eliminate the left $1s$ in the first column and first row by simple addition. Continuing in this way, the optimal matrix is available. Let the transition matrix of ∂_k denoted by N_p with n_{p-1} rows and n_p columns.

```
void REDUCE($x$)
  if there exist $k \geq x, l \geq x$ with $N_p[k, l] = 1$ then
    exchange rows $x$ and $k$;  exchange columns $x$ and $l$;
    for $i = x + 1$ to $n_{p-1}$ do
      if $N_p[i, x] = 1$ then add row $x$ to row $i$ endif
    endfor;
    for $j = x + 1$ to $n_p$ do
      if $N_p[x, j] = 1$ then add column $x$ to column $j$ endif
    endfor;
    REDUCE($x + 1$)
  endif.
```

Hence, we have at most $n_{p-1} + n_p$ operation for each per recursive calculation. So we have $(n_{p-1} + n_p)min\{n_{p-1}, n_p\}$ operations in total.

5 Analysis

6 Conclusion

7 Literature Cited

References

[1] V. de Silva and G. Carlsson “Topological estimation using witness complexes,” Eurographics Symposium on Point-Based Graphics, 2000.

[2] H.Edelsbrunner, COMPUTATIONAL TOPOLOGY: An Introduction, 2008.

[3] H.Edelsbrunner, D.Letscher and A.Zomorodian ”Topological Persistence and Simplification,” Discrete Comput Geom, 28:511-533,2002.

[4] G. Carlsson, T.Ishkhanov, V. de Silva, A.Zomorodian, On the Local Behavior of Spaces of Natural Images, Springer, LLC2007.

8 Acknowledgements