Analysis of high dimensional data via Topology

Louis Xiang

Overview

- Modifying the Data set
- Building the Simplicial Complex
- Computing the homology

Data set

299264x43 double

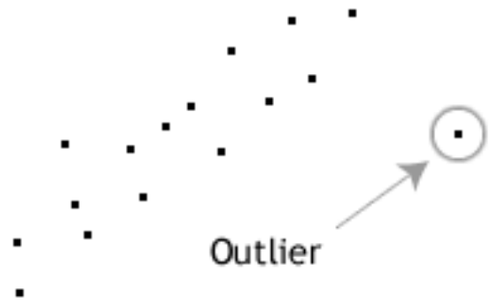
	1	2	3	4	5	6	7
210	2477	-1	-1	-1	-1	-1	56.7000
211	2537	-1	-1	-1	-1	-1	56.7000
212	2597	-1	-1	-1	-1	-1	56.7000
213	2657	-1	-1	-1	-1	-1	56.7000
214	2717	-1	-1	-1	-1	-1	56.7000
215	2777	-1	-1	-1	-1	-1	56.7000
216	2837	-1	-1	-1	-1	-1	56.7000
217	0	132543	68	1	180.3000	3	84.6000
218	11	-1	-1	-1	-1	-1	-1
219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

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	

	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

Clean out the outliers



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 $X(K,p)$.

Building the simplicial complex

1. Choosing the landmarks : maxmin method

2. Building the witness 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 $\sigma = [ab] \in W_\infty(D)$ 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 $\sigma = [a_0, a_1, ..., a_k]$ 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.

Computation of homotopy

1. Definition of chain, cycle and boundary

`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.

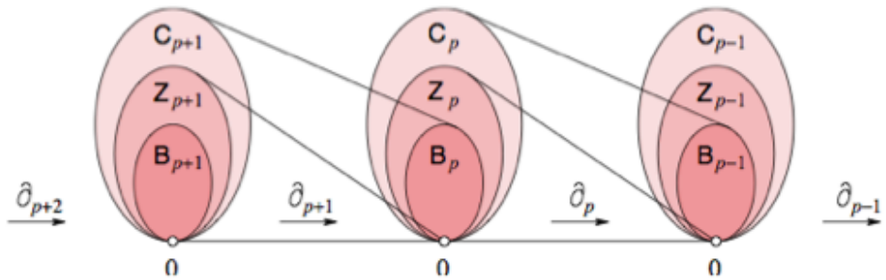


Figure IV.1: The chain complex consisting of a linear sequence of chain, cycle, and boundary groups connected by homomorphisms.

3. Homotopy groups

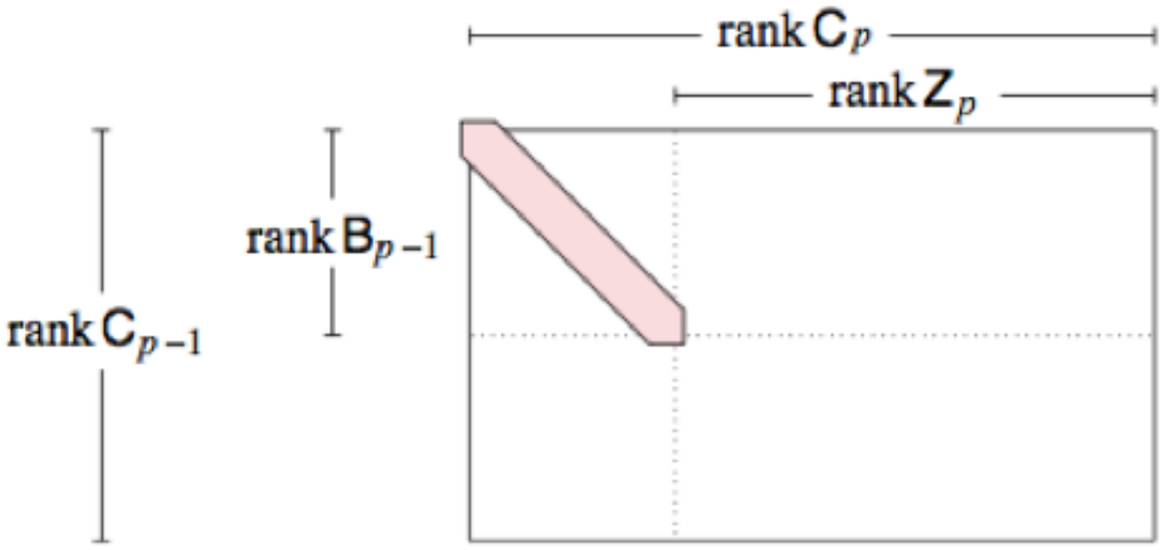
The k-th *homotopy group* is the k-th cycle group divided by the k-th boundary group, i.e. $H_k = Z_k/B_k$. The *Bettinumber* of K β_k is the rank of the k-th homology group, $\beta_k = \text{rank} H_k$ where $H_k = Z_k/B_k$.

Computation of homotopy

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 $transition\ matrix\ A$.



Computation of homotopy

Algorithm:

```
void REDUCE(x)
  if there exist k ≥ x, l ≥ 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.
```

Javaplex examples