Topological Analysis for high-dimensional data

Louis Xiang, Fernando Schwartz, Kwai Wong
The Chinese University of Hong Kong; University of Tennessee, Knoxville; National Institute for Computational Sciences, Knoxville, TN

Overview

In this study we will focus on computing the topological invariant of high dimensional data set. By this kind of topological analysis, we are able to indicate some qualitative result about the high-dimensional data set. We will use the ICU medical data set as our object to show how the method describe the shape of the data set.

High-density Dataset

First, we reduce the dimension of the data set by selecting the most relevant 8 factors of the patients and do interpolation to fill in the missing data. Noting that direct application of simplicial complex approximation to the original 300,000 data points will lead to a wrong detection because of the outliers distributed far away from main region. To obtain a high-density subset, we use the ***density function***

$\rho_K(x) = |x - x_K|$ where x_K is the K-th nearest point of x.

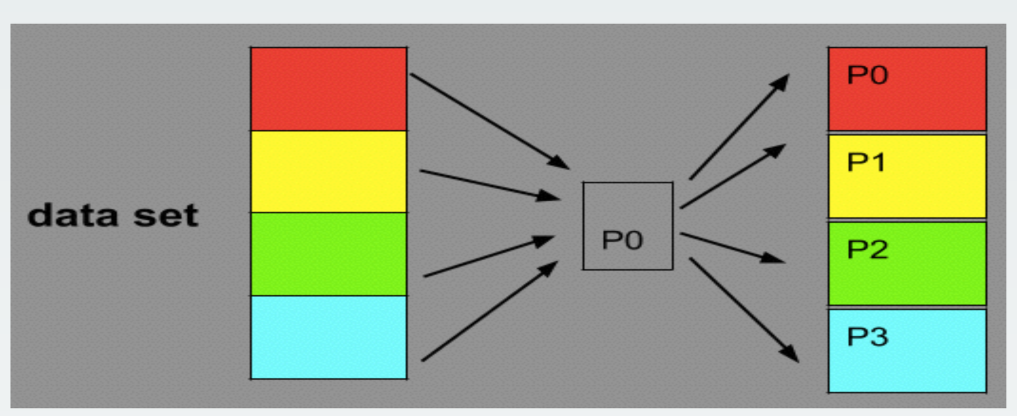
The crucial step is to find the distance matrix where

$d_{ij} = d(x_i, x_j), x_i$ and x_j is the i, j rows in the matrix.

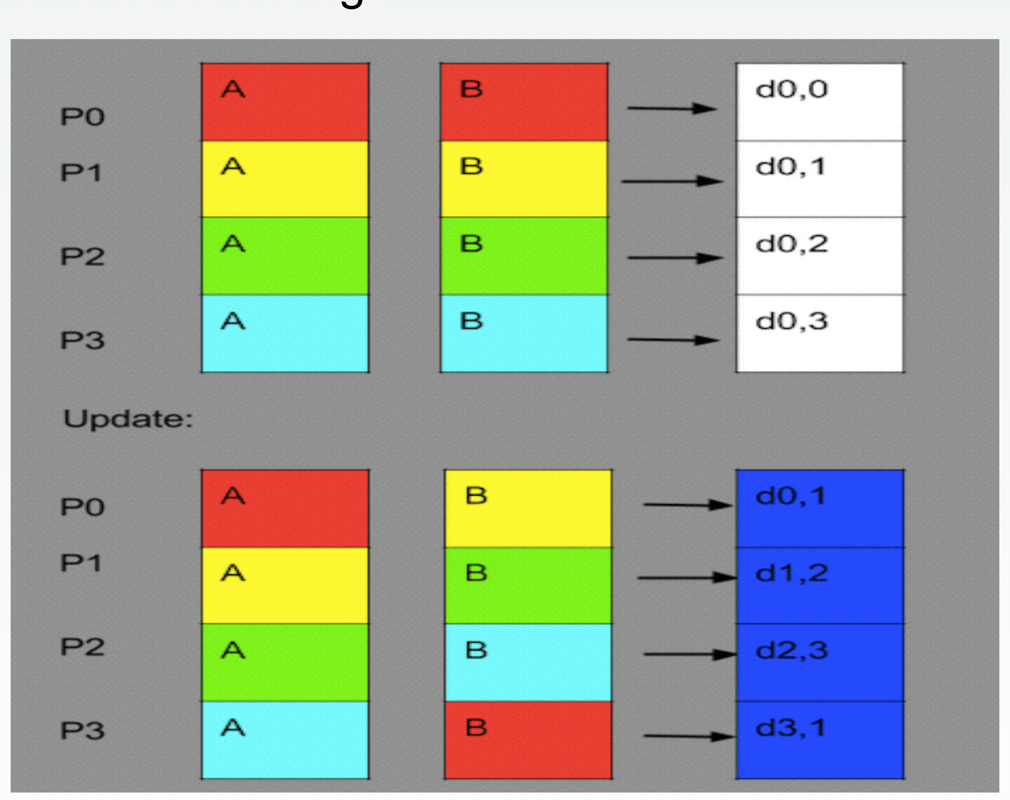
Since the original data is of big magnitude, We may use ***Darter*** in NICS as our supercomputer to do parallel computing.

Algorithm:

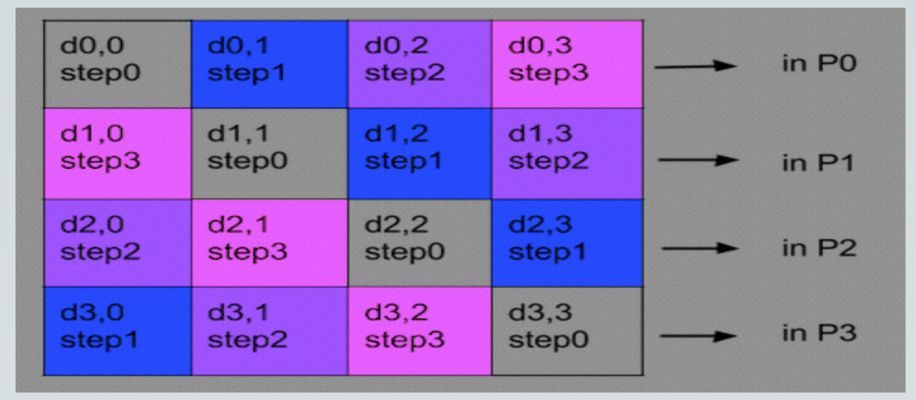
Step1: All points can form a matrix. P0 read and send each part of the matrix to other processors.



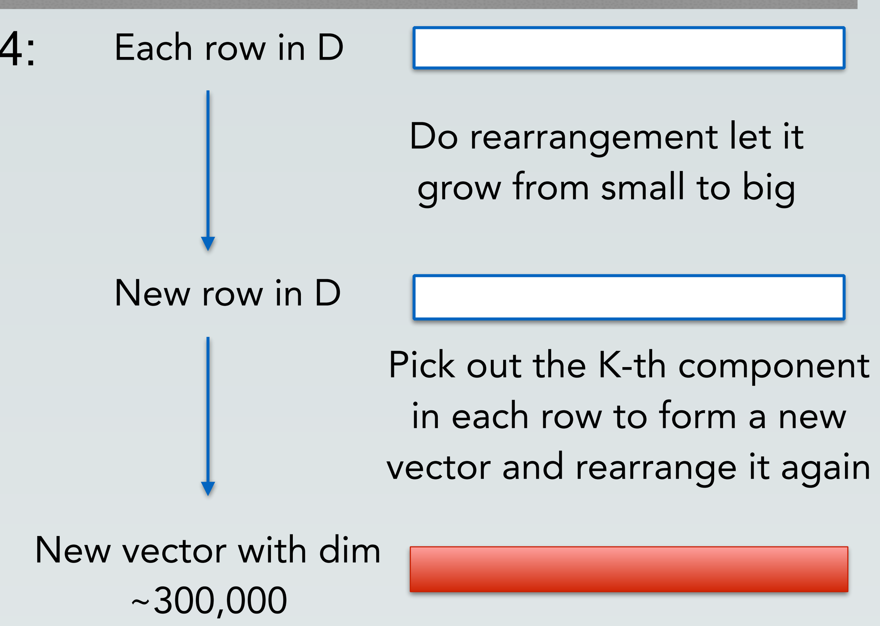
Step2:Let A and B be two collection of points. Calculate the distance matrix between A and B and then shift the B between each processors and calculate again.



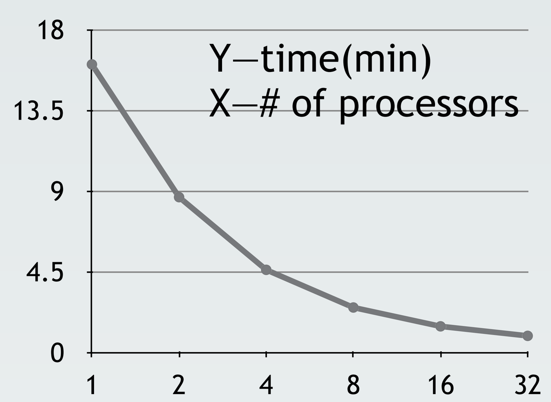
Step3:Continuing in this way until we get the whole distance matrix.

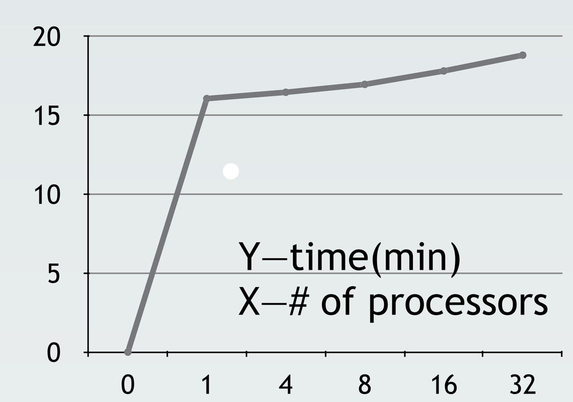


Step4:



Step5: Record the points which are on the top p % in the rearrangement of new vector. Then these points form X(K,p), a subset of the original data. Below is the scaling of this method for a relative small data set, say, 10,000 points.





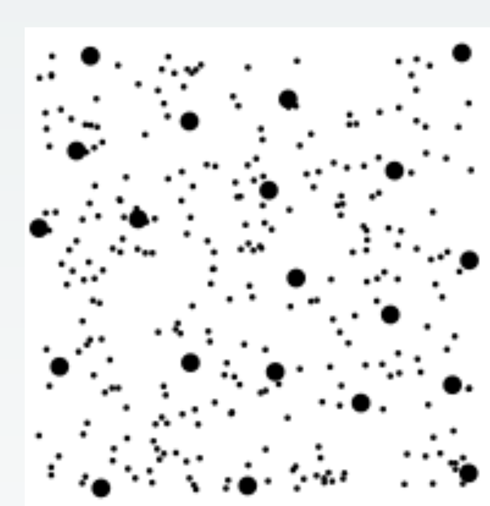
Javaplex for Betti Number

We recommend selecting the landmark points by maxmin method.

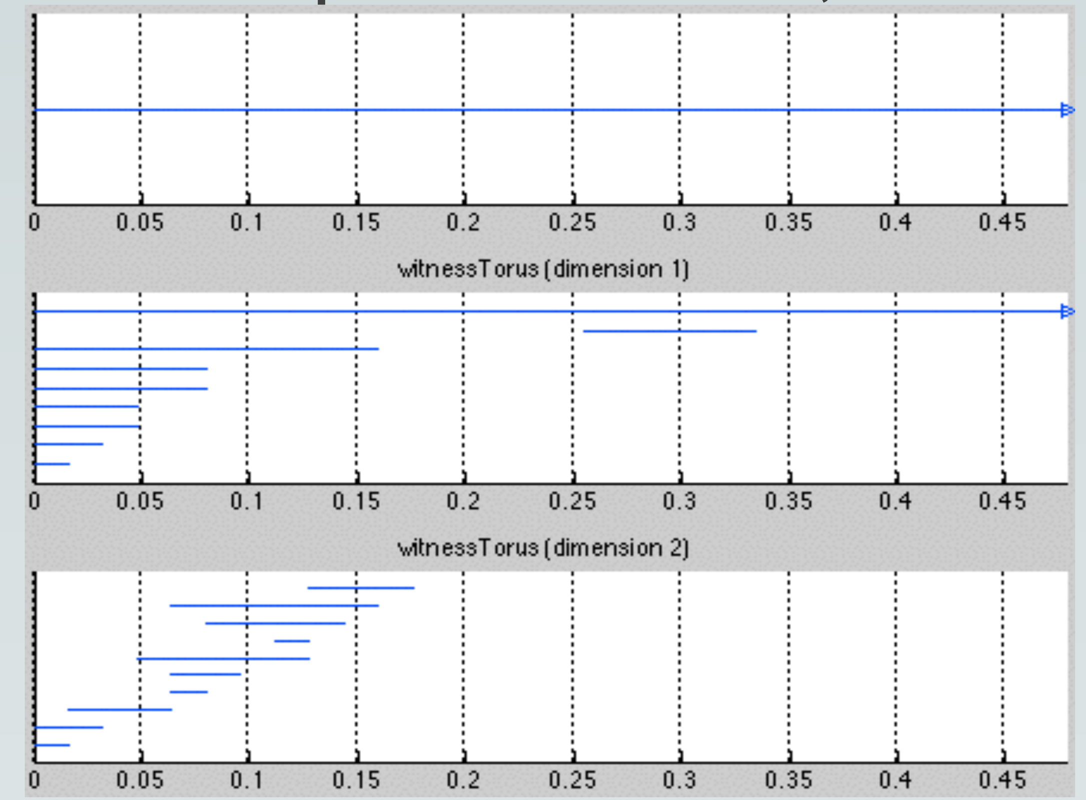
Algorithm:

- 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 normal metric.

example of landmarks from data set:



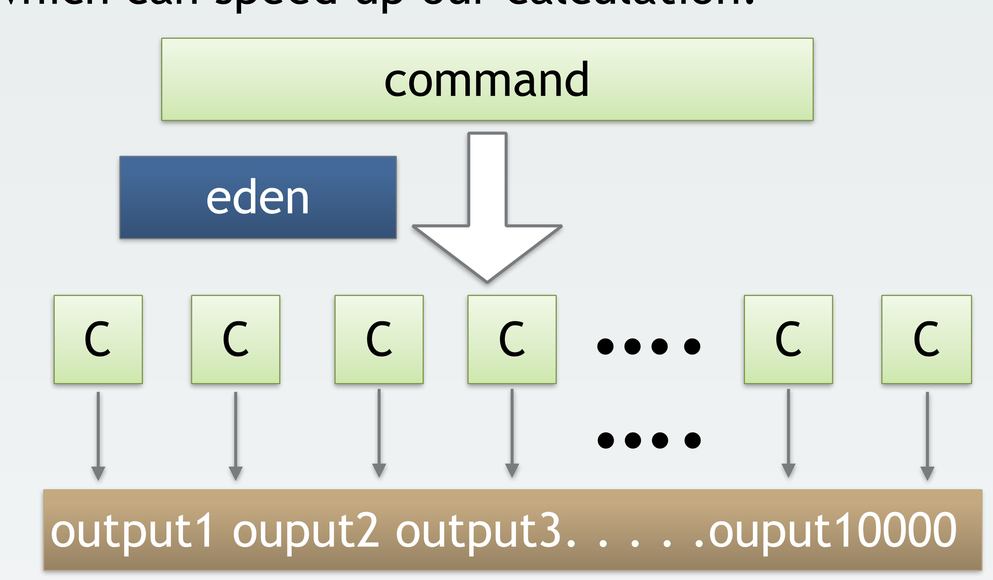
We use the landmarks to build the simplicial complex. Then we are able to compute homology. The final goal is to get the ***betti*** number which indicates how many holes of the shape in each dimension. For more details about the knowledge of computing the betti number, see reference [2]. **Javaplex** is a Java package developed in Standord. It can directly give us the barcode of holes in each dimension for input data.For the algorithm, please refer to the reference [1].Here is the example of the output.



- The first barcode represents the betti 0 which is the number of the connected components.
- Betti 1 and 2 represent the number of 1-dim and 2-dim holes in the graph represented by the 2nd and 3rd barcodes.
- The blue line indicates the existence of the hole with the change of parameter.
- Even though there are some short lines which are the noise, we can still concentrate on the lasting lines.

Numerical Simulation

Because of the undetermined characteristic of landmarks, we need to run Javaplex for 10,000 times for each K and p to give a statistical certainty of 97% of the betti number.To implement the simulation, we use eden on the Nautilus which can speed up our calculation.



Basically, we can put the command of running Javaplex for 10,000 times into the command file and create the head file, then Eden can do it for 10,000 times and gives out the output.

Analysis of output

The output is 10,000 collections of barcodes. It is impossible to analysis each output by hand. Better way is to transfer the result to matrix and write some programs in Matlab to judge the number of holes in each dimension. My basic idea is below:

```
Dimension: 0
[0.0, infinity)
Dimension: 1
[0.0, 0.025)
[0.055, infinity)
[0.0, infinity)
Dimension: 2
[0.0, 0.05)
```

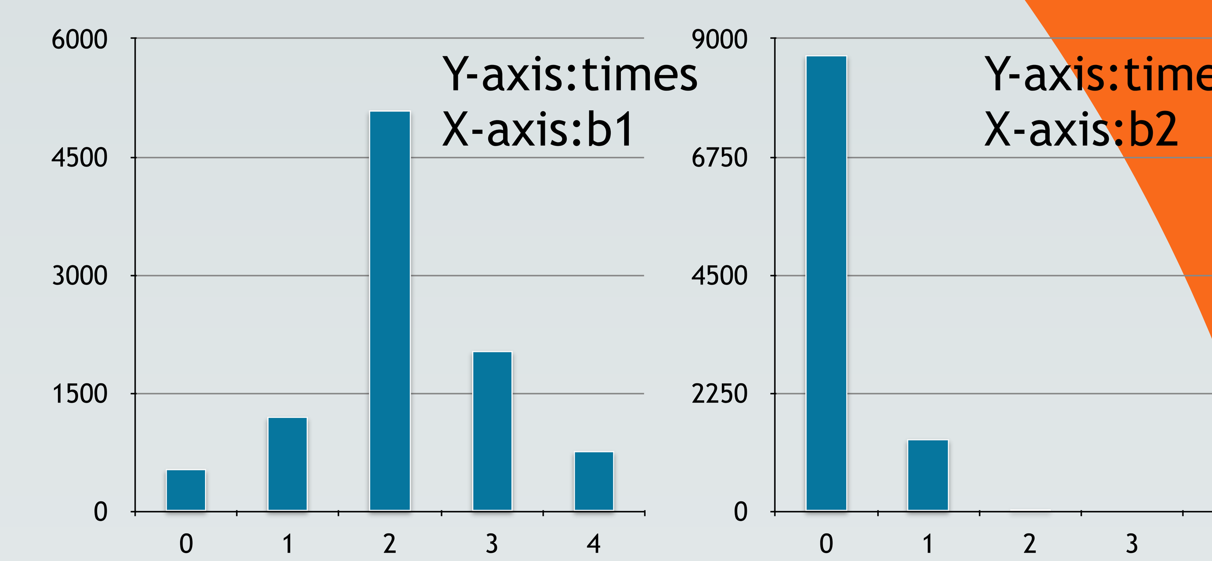
read the every interval's endpoints into a 3-dimensional array

	1	2	3
1	0	Inf	0
2	0	0.0250	1
3	0.0550	Inf	1
4	0	Inf	1
5	0	0.0500	2

we say if length of the interval exceed 0.18(depends on your self), then we improve the betti number by 1 in that dimension

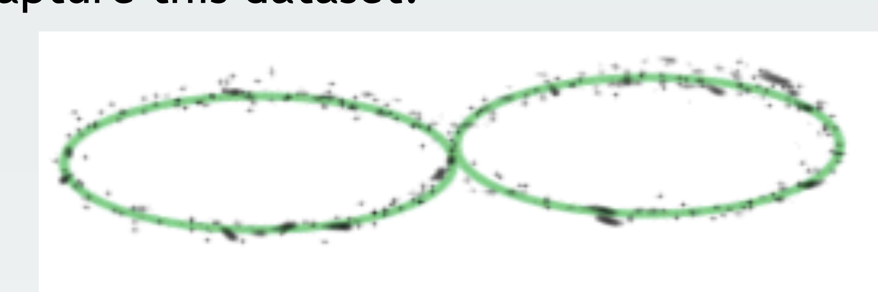
NUMBER	NUMBER	NUMBER
b0	b1	b2
1	2	0

For K=50,P=50, the histograms for betti numbers of X(K,P) are below.(b0 are always 1 since only 1 connected component is detected.)



Conclusion

- Statistically speaking, through the distribution of the histogram above,we can have 97% certainty that the b1 and b2 are 2 and 0.
- However, this is just the case for K=p=50, more experiments for different K and p are in the process. Through the experiments that we have conducted until now, the result that b1=2 and b2=0 is quite solid. For this property, the shape below may have much possibility to capture this dataset.



- Interesting thing is that bi=0 for any 3<=i<=7 in every iteration. This result tells us that there may be some relationships between some of the factors of these 8. Further analysis will be expected.

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 imges*, Springer, LLC2007.

Acknowledgements

The research was conducted under the Computational Science for REU project and is supported by the JICS, founded by the UTK and ORNL. The authors acknowledge ORNL for allowing access to high-performance computing resources.