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 a high dimensional data set. By this kind of topological analysis, we are able to indicate the qualitative result about the high-dimensional 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.

Methods

First, we try to reduce the dimension of the data set by selecting the most relevant 8 factors of the patients. And interpolation is needed to fill in the missing data of the patients. But It must be pointed out that direct application of simplicial complex approximation to the original data points will unfortunately lead to wrong detection since there are points distributed far away from the high-density regions. 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.

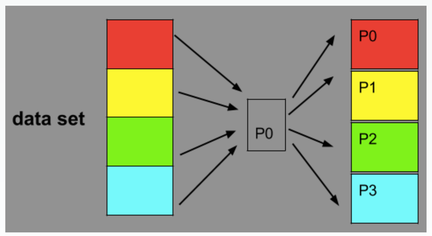
The crucial step is to find the distance matrix where

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

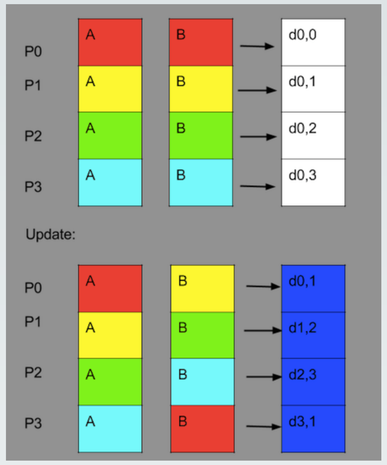
After that, we are going to form a set X(K,p). Since the original data is of big magnitude, it is suggested to use parallel computing on super computers.

Procedure

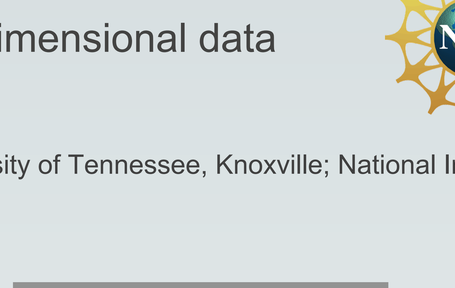
Step1: P0 read and send each part to other processors.



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 we can get the distance matrix and then do the rearrangement and take out the k-th column in each processor.



Do the rearrangement again on k-th column and record the points which is on the top p% in the rearrangement.Then these points form X(K,p), a subset of the original data.

Methods cont.

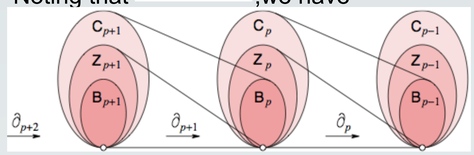
After obtaining the X(k,p), we will select the landmark points to build the simplicial complex.

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 and then the final process is to calculate the bettie number Noting that $B_k \subset Z_k \subset C_k$,we have



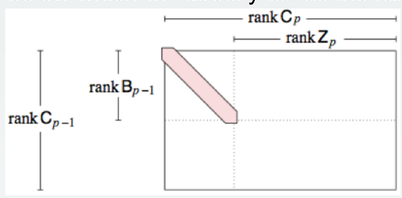
The bettie number $\beta_k = \text{rank} H_k$, where

1. $rank\ H_k = rank\ Z_k - rank B_k = null\ \partial_k - rank\ \partial_{k+1}$.
2. $rank(A) + Null(A) = N$, where N is the number of rows of the transition matrix.

Our goal now is to calculate the rank of matrix. Below is the algorithm:

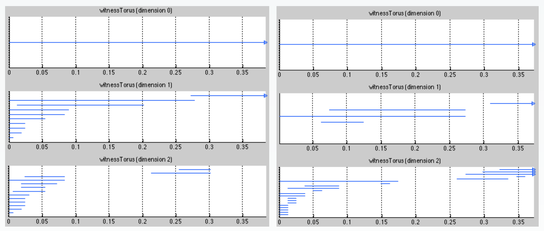
```
void REDUCE(x)
  if there exist k ≥ x, l ≥ x with N_p[k,l] = 1 then
    exchange rows x and k;  exchange columns x and l;
    for i = x + 1 to n_{p-1} do
      if N_p[i,x] = 1 then add row x to row i endif
    endfor;
    for j = x + 1 to n_p do
      if N_p[x,j] = 1 then add column x to column j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

Finally will get the Smithnormal matrix, like below which will directly tell us the rank:

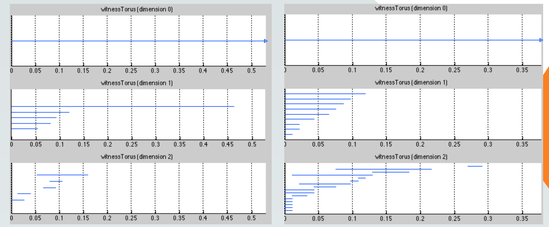


Results and Analysis

There is a very powerful software Javaplex in Matlab which can help us compute the homology. Javaplex is mainly developed by the Computational Topology workgroup at Stanford University.



X(300,20) X(300,30)



X(800,20) X(800,30)

These four graphs tell the bettie numbers of different sets X(K,p). The first box represents the bettie 0 which is the number of the connected components and bettie 1 and 2 represent the number of 1-dim and 2-dim holes in the graph. The blue line indicates the existence of the hole with the change parameter. Even though there are some short lines which is the noise, we can still concentrate on the lasting lines. It seems that there is one connected component and 2 1-dim holes and a 2-dim hole which performs like a Torus(right bottom). We will try more X(K,p) later and provide more convincing evidence.

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 present 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.