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.

![figure]

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 $-1$s (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,

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>210</td>
<td>2477</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>211</td>
<td>2537</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>212</td>
<td>2597</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>213</td>
<td>2657</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>214</td>
<td>2717</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>215</td>
<td>2777</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>216</td>
<td>2837</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>208</td>
<td>103</td>
<td>53</td>
<td>77</td>
<td>0.3000</td>
<td>38.1563</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>209</td>
<td>124</td>
<td>65</td>
<td>89</td>
<td>1.0167</td>
<td>38.1500</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>210</td>
<td>119</td>
<td>60</td>
<td>83</td>
<td>1.0167</td>
<td>38.1250</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>211</td>
<td>121</td>
<td>63</td>
<td>87</td>
<td>1.0667</td>
<td>38.1000</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>212</td>
<td>103</td>
<td>52</td>
<td>72</td>
<td>1.3667</td>
<td>37.8750</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>213</td>
<td>118</td>
<td>59</td>
<td>83</td>
<td>0.5833</td>
<td>37.6500</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>214</td>
<td>121</td>
<td>65</td>
<td>86</td>
<td>0.2500</td>
<td>37.4250</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>215</td>
<td>124</td>
<td>69</td>
<td>91</td>
<td>0.5833</td>
<td>37.2000</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>216</td>
<td>126</td>
<td>70</td>
<td>92</td>
<td>0.5833</td>
<td>37.2000</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>217</td>
<td>134</td>
<td>65</td>
<td>86.6700</td>
<td>600</td>
<td>36.3000</td>
<td>15</td>
<td>19</td>
</tr>
<tr>
<td>218</td>
<td>134</td>
<td>63</td>
<td>86.6700</td>
<td>600</td>
<td>36.2000</td>
<td>15</td>
<td>19</td>
</tr>
<tr>
<td>219</td>
<td>134</td>
<td>63</td>
<td>86.6700</td>
<td>600</td>
<td>36.3000</td>
<td>15</td>
<td>19</td>
</tr>
<tr>
<td>220</td>
<td>134</td>
<td>63</td>
<td>86.6700</td>
<td>600</td>
<td>36.3077</td>
<td>15</td>
<td>19</td>
</tr>
</tbody>
</table>

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 $R^5$. 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 regard each row as a point belongs to $R^8$.

- 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 $\{\{A\}, \{B\}, \ldots, \{D\}, \{A, B\}, \ldots, \{C, D\}, \{A, B, C\}, \ldots, \{B, C, D\}, \{A, B, C, D\}\}$ 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, $\sigma = \text{conv}\{u_0, u_1, \ldots, u_k\}$.

For example, vertex for 0-simplex, edge for 1-simplex, triangle for 2-simplex, and tetrahedron for 3-simplex, see the figure below.
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 \( l_1 \in \mathbb{Z} \) randomly. For each \( i \geq 2 \), if \( l_1, l_2, \ldots, l_{i-1} \) have been chosen, let \( l_i \in \mathbb{Z} \setminus l_1, l_2, \ldots, l_{i-1} \) be the data point which maximises the function \( f(x) = \min_{1 < j < i} D(x, j) \) 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 \( \{1, 2, \ldots, n\} \), 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 \( \sigma = [a_0, a_1, \ldots, a_k] \) belong to \( W_\infty(D) \). Then \( \sigma \in W_\infty(D) \) iff there exists a data point \( 1 \leq i \leq N \) such that \( D(a_0, i), D(a_1, i), \ldots, D(a_p, i) \) 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 $C_p=\{\sum a_i v_i | a_i \in \mathbb{Z}_2, v_i$ are $p$-simplices in $K$ and $\mathbb{Z}_2$ is the modulo 2}\}. The boundary of a $k$-simplex $\sigma$ 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, $\partial_k(c)$, is defined to be $\sum_{\sigma \in c} \partial_k(\sigma)$. Therefore we can write $\partial_p : C_p \rightarrow C_{p-1}$ 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, 

$$\cdots \xrightarrow{\partial_{p+2}} C_{p+1} \xrightarrow{\partial_{p+1}} C_p \xrightarrow{\partial_p} C_{p-1} \xrightarrow{\partial_{p-1}} \cdots$$

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

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

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

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 $\partial_k$. 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 \(\text{Im } \partial_{k+1}\). The \(k\) - cycle is the \(k\)-chain in the \(\text{ker } \partial_k\) and a \(k\)-boundry is a \(k\)-chain in the \(\text{Im } \partial_{k+1}\).

**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.](image)

### 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 Betti number of \(K\) \(\beta_k\) is the rank of the \(k\)-th homology group, \(\beta_k = \text{rank}H_k\) where \(H_k = Z_k / B_k\).

There is a Betti number for each integer \(k\). If \(B_0 = i\), that means there 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 \(\text{rank}H_k\). Note that

1. \(\text{rank } H_k = \text{rank } Z_k - \text{rank } B_k = \text{null } \partial_k - \text{rank } \partial_{k+1}\).
2. \(\text{rank}(A) + \text{null}(A) = N\), \(N\) 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 $\mathbb{Z}_2$, 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.

![Smithnormalmatrix](image)

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 $\partial_k$ denoted by $N_p$ with $n_{p-1}$ rows and $n_p$ columns.

```cpp
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,” Euro-


[3] H.Edelsbrunner, D.Letscher and A.Zomorodian ”Topological Persistence and Simplifi-


8 Acknowledgements