Extracting Topological Features from Big Data Using Persistent Density Entropy

Jinzhong Xu\textsuperscript{1,2, a}, Xuzhi Li\textsuperscript{1} and Hongfei Wang\textsuperscript{1}

\textsuperscript{1}Key Laboratory of Space Utilization, Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing 100094, P.R. China
\textsuperscript{2}University of Chinese Academy of Sciences, Beijing 100049, P.R. China
\textsuperscript{a}Corresponding author: xujinzhong15@csu.ac.cn

Abstract. Topological data analysis is a method of extracting shape information of big data by means of algebraic topology in mathematics. Persistent homology is a very important method in topological data analysis. It constructs multi-scale simplicial complexes (also called filtration) to approximate the underlying space of the data set. By studying these simplicial complexes, the topological features of each dimension of big data are summarized. However, it does not give us the uncertainty of each simplicial complex to approximate the underlying space of the data set. This paper defines an entropy called persistent density entropy, which gives the uncertainty of each simplicial complex approximating the underlying space. The examples demonstrate that it is able to find the best simplicial complex that approximates the underlying space and can be used to detect outliers to a certain extent.

1. Introduction
Topological data analysis emerging in recent years is a mathematical method for processing high-dimensional complex big data sets. It mainly uses the algebraic topology in mathematics to extract the shape information of big data. At present, it has achieved good and successful applications in many fields: sensor networks\cite{1, 2}, bioinformatics\cite{3, 4}, neuroscience\cite{5}, medical science\cite{6-9}, shape analysis\cite{10, 11}, computer vision\cite{12}, Natural Language Processing\cite{13}, complex systems\cite{14}, and big data\cite{15, 16}, etc. The main idea of topological data analysis is to construct a suitable simplicial complex (a topological space and Vietoris-Rips complex is often used) to approximate the underlying space of big data, then obtain the shape information (or topological features) of the big data by studying these Vietoris-Rips complexes. Persistent homology is a very important technique in topological data analysis. Persistent homology was first proposed in \cite{17}. Instead of choosing the right Vietoris-Rips complex, it takes all the Vietoris-Rips complexes into account and constructs a nested Vietoris-Rips complex, also known as filtration. By studying filtration, a persistence diagram or a barcode can be obtained, which considers the feature with long duration as topological features, and features with short duration are considered topological noise. This approach does not give us the underlying space that our Vietoris-Rips complexes can better approximate the data set. And Vietoris-Rips filtrations quickly become prohibitively large as the size of the data increases, making the computation of persistence practically almost impossible. In addition, persistence diagrams of Vietoris-Rips filtration and Gromov-Hausdorff distance are very sensitive to noise and outliers. In order to solve these problems, this paper hopes to select a suitable Vietoris-Rips complex by giving the uncertainty of each complex approximation of the underlying space. Calculating only one or a few
Vietoris-Rips complexes can reduce the computational complexity. In general, entropy is a preferred measure of uncertainty. This paper calculates the entropy value for each Vietoris-Rips complex as its uncertainty about the underlying space, which is called the persistent density entropy. Finally, a suitable Vietoris-Rips complex is obtained by looking at the inflection points of the persistent density entropy diagram. The outliers of the data set appear as sudden peaks of the persistent density entropy diagram, which can find outliers to a certain extent.

2. Persistent Homology

Persistent homology summarizes the topological features of the data set by constructing a series of nested simplicial complex or a filtration.

This section gives the basic concepts needed for the paper. They include simplex (the basic module that constructs the topological space), simplicial complex (a practical topological space, such as Vietoris-Rips complex), homology group (detecting the number of holes in the topological space), and persistent homology (summarizing the important topological features of the data set).

2.1 Simplicial Complexes

The basic building blocks that make up our topological space are simplexes.

2.1.1 Simplexes

Definition 1. A $k$-simplex $\sigma$ is a $k$-dimensional polytope which is the convex hull of its $k + 1$ vertices $v_0, v_1, \ldots, v_k \in \mathbb{R}^d$. Written $\sigma = [v_0, v_1, \ldots, v_k]$. The dimension of $\sigma$ is $\text{dim} \sigma = k$.

For example, A $0$-simplex is a vertex, a $1$-simplex is an edge, a $2$-simplex is a triangle, and a $3$-simplex is a tetrahedron, and so on. See Figure 1.

Figure 1. $k$-simplexes for $k = 0, 1, 2, 3$.

The convex hull of any nonempty subset of points $\{v_0, v_1, \ldots, v_k\}$ is called a face of the simplex $\sigma$. For example, a triangle has six faces, which are its three edges and three vertices.

2.1.2 Simplicial Complexes

Using a series of simplexes, a simplicial complex can be constructed, which is a topological space.

Definition 2. A simplicial complex $\mathcal{K}$ is a set of simplexes that satisfies the following two conditions:

1. Every face of $\sigma \in \mathcal{K}$ is also in $\mathcal{K}$.
2. If $\sigma_a, \sigma_b \in \mathcal{K}$, then $\sigma_a \cap \sigma_b$ is a face of both $\sigma_a$ and $\sigma_b$.

For example, in Figure 2, the left is a simplicial complex, while the right one is not.

Figure 2. True simplicial complex (on the left) and false simplicial complex (on the right).

The Vietoris–Rips complex is an abstract simplicial complex that is often used to characterize the topology of a point set. Because the Vietoris–Rips complex can easily extend construction to higher dimensions, they are particularly popular in topological data analysis.

Consider a set of points $X = \{x_0, x_1, \ldots, x_n\}$ with metric $d: X \times X \to \mathbb{R}$. A Vietoris-Rips complex $R_\epsilon(X)$ consists of all those simplexes whose vertices are at pairwise distance no more than $\epsilon$,

$$R_\epsilon(X) = \{ \text{conv}(\sigma) | \sigma \subseteq X, \forall x_i, x_j \in \sigma, d(x_i, x_j) \leq \epsilon \}.$$
For example, given a radius $\epsilon/2$, the Vietoris-Rips complex in Figure 3.

![Figure 3. Example of a Vietoris-Rips complex. The 7 points are 0-simplexes. Two 0-simplexes form a 1-simplex (an edge) if their distance is less than $\epsilon$. Three vertices form a 2-simplex (a triangle) if the distance between any two points is less than $\epsilon$.](image)

The nerve theorem, attributed to Borsuk[18], asserts that the homotopy type of a sufficiently nice topological space is encoded in the Čech nerve of a good cover.

**Theorem 1.** (Nerve lemma[19])

If $\mathcal{U}$ is a finite open cover of $X$ such that every nonempty intersection of sets in $\mathcal{U}$ is contractible, then $X \cong \mathcal{N}(\mathcal{U})$.

Čech complex is nerve of the union of $\epsilon$-balls

$$C_\epsilon(X) = \{ \text{conv}(\sigma) | \sigma \subseteq X, \bigcup_{x \in \sigma} B_\epsilon(x) \neq \emptyset \}.$$ According to the definition of the Vietoris-Rips complex and the Čech complex, the following relationships are established

$$C_\epsilon(X) \subset R_\epsilon(X) \subset C_{2\epsilon}(X).$$

And on big data, Vietoris-Rips complex is easier to extends to higher dimensions, so Vietoris-Rips is generally used in topology data analysis. In addition to the above two complexes, there are many other useful complexes: Alpha complex, Witness complex etc.

### 2.2 Persistent Homology

For a set of points $X$, many Vietoris-Rips complex can be constructed, which one can more accurately approximate the underlying space? Persistent Homology does not consider how to choose a specific complex, but considers everything.

2.2.1 **Homology Group**

A $k$-chain $c$ is a subset of $k$-simplexes in a simplicial complex $\mathcal{K}$. In practical computations, the coefficients is $\mathbb{Z}_2 = \{0, 1\}$. That is to say, $c = \sum_i \gamma_i \sigma_i$, where $\gamma_i \in \mathbb{Z}_2, \sigma_i$ is an arbitrary $k$-simplex.

All $k$-chains are a set, and the $+_2$ operation are defined on the set to form a group called $k$ chain group, denoted by $C_k$, where $+_2$ defined as follows:

|   | 0  | 1  |
|---|----|----|
| 0 | 0  | 1  |
| 1 | 1  | 0  |

For example, if there are two chains:

$$c_1 = [v_0, v_1] +_2 [v_1, v_2] +_2 [v_2, v_0],$$

and

$$c_2 = [v_1, v_2] +_2 [v_2, v_3] +_2 [v_3, v_1],$$

then

$$c_1 +_2 c_2 = ([v_0, v_1] +_2 [v_1, v_2] +_2 [v_2, v_0]) +_2 ([v_1, v_2] +_2 [v_2, v_3] +_2 [v_3, v_1])$$

$$= [v_0, v_1] +_2 [v_3, v_1] +_2 [v_2, v_3] +_2 [v_2, v_0]$$

$$= c_3$$

Figure 4 is its graphical representation.
The boundary $\partial_k(\sigma_k)$ of a $k$-simplex is the set of all $k - 1$-simplex faces. Such as

$$\partial_3(\sigma_3) = \partial_3([v_0, v_1, v_2]) = [v_0, v_1] + [v_1, v_2] + [v_2, v_0] = c_1.$$  

The boundary operator $\partial_k$ is a group homomorphism from $C_k$ to $C_{k-1}$. And it can be proved that

$$\partial_k \circ \partial_{k+1} = 0 \quad (1)$$

For $k > 1$ and $k \in \mathbb{Z}$,

$$\partial_k \circ \partial_{k+1}[v_0, \ldots, v_{k+1}] = \sum_{i=0}^{k+1} \partial_k[v_0, \ldots, v_i^-, \ldots, v_{k+1}] = \sum_{j<i} \sum_{j>i} \partial_k[v_0, \ldots, v_j^-, \ldots, v_{k+1}] + 2 \sum_{j>i} \partial_k[v_0, \ldots, v_j^-, \ldots, v_{k+1}] = 0$$

The following chain complex can be obtained from the boundary operators $\partial_k$ and the chain group $C_k$

$$\cdots \rightarrow C_{k+1} \xrightarrow{\partial_{k+1}} C_k \xrightarrow{\partial_k} C_{k-1} \xrightarrow{\partial_{k-1}} \cdots \rightarrow C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0.$$  

If $\partial_k c_k = 0$, then $c_k$ is a $k$-cycle, which is the kernel of the boundary operator $\partial_k$

$$Z_k \triangleq \ker(\partial_k : C_k \rightarrow C_{k-1}).$$

The image of the boundary operators is the $k$-boundaries

$$B_k \triangleq \operatorname{im}(\partial_{k+1} : C_{k+1} \rightarrow C_k).$$

The following figure shows the relationship between chain complex $(C_k, \partial_k)$, cycle group $Z_k$ and boundary group $B_k$.

$$\text{Figure 5. The relationship between } C_k, Z_k \text{ and } B_k.$$

It can be seen that

$$B_k \subset Z_k \subset C_k.$$  

In Figure 4, it shows that

$$B_1 = \{c_1, 0\},$$

$$Z_1 = \{c_1, c_2, c_3, 0\},$$

And

$$B_1 +_2 c_2 = c_2 +_2 B_1 = \{c_3, c_2\} = c_3 +_2 B_1 = B_1 +_2 c_3,$$

$$B_1 +_2 c_1 = c_1 +_2 B_1 = \{0, c_1\} = 0 +_2 B_1 = B_1 +_2 0.$$  

Since each chain group $C_k$ is abelian (because $+_2$ is commutative). Then cycle group $Z_k$ and boundary group $B_k$ are abelian, and since $B_k \subset Z_k$ therefore $B_k$ is a normal subgroup of $Z_k$. Then one can define the quotient group.

**Definition 5.** The $k$-th homology group $H_k$ is the quotient group

$$H_k = Z_k/B_k.$$
The elements of $H_k$ are called homology classes. Each homology class is an equivalence class. If two elements $c_i$ and $c_j$ in the same homology class are said to be homologous, written $c_1 \sim c_2$. Let $[c]$ denote the homology class of $c \in Z_k$. The $k$-th Betti number is $\beta_k = \text{rank} (H_k)$.

Informally, the $k$-th Betti number $\beta_k$ refers to the number of $k$-dimensional holes on a topological surface. The first three Betti numbers $\beta_0, \beta_1$ and $\beta_2$ have the following definitions for 0-dimensional, 1-dimensional, and 2-dimensional simplicial complexes:

1. $\beta_0$ is the number of connected components.
2. $\beta_1$ is the number of 1-dimensional circular holes.
3. $\beta_2$ is the number of 2-dimensional voids or cavities.

For higher dimensions, $\beta_k$ represents the numbers of $k$-dimensional holes.

By definition,

$$\beta_k = \text{rank} (Z_k) - \text{rank} (B_k).$$

In Figure 4, $c_1$ and 0 are homologous, i.e., $c_1 \sim 0$.

$c_2$ and $c_3$ are homologous, i.e., $c_2 \sim c_3$.

The 1-the homology group is

$$H_1 = \frac{Z_1}{B_1} = \frac{\langle c_1, c_2, c_3, 0 \rangle}{\langle c_1, 0 \rangle} = \{ [c_1], [c_2] \},$$

which is isomorphic to $\mathbb{Z}_2$.

The 1-the Betti number

$$\beta_1 = \text{rank} (H_1) = \text{rank} (Z_1) - \text{rank} (B_1) = 2 - 1 = 1 = \text{rank} (\mathbb{Z}_2).$$

This shows that there is a one-dimensional hole here. This is shown in Figure 4.

2.2.2 Filtration

Through the above preparation, it can calculate the Betti numbers $\beta_k$ of the homology group $H_k$ given a parameter $\epsilon$ on the dataset $X$. In order to investigate the properties of homology groups over all parameters, the authors H. Edelsbrunner etc. introduced filtration in [17]. The filtration is a nested sequence of increasing subset. Through the filtration it can dynamically get the time when the hole in each dimension appears (birth time) and when it disappears (dead time).

Suppose there is a collection of dataset $X$ with $\# X = N$. It can be noticed that as radius $\epsilon$ grows so do the Vietoris-Rips complex $R_\epsilon (X)$, this give an inclusion of Vietoris-Rips complex as the radius grow from small to large. If there is a finite number of parameters $\epsilon_0 < \epsilon_1 < \cdots < \epsilon_{n-1} < \epsilon_n$, Then, the nested Vietoris-Rips complex $R_{\epsilon_i} (X)$ is as follows

$$R_{\epsilon_0} (X) \subset R_{\epsilon_1} (X) \subset \cdots \subset R_{\epsilon_{n-1}} (X) \subset R_{\epsilon_n} (X).$$

For example, in Figure 6, there are fifty sample points $X_{50}$ with Gaussian noise with a mean $\mu = 0$ and a variance $\sigma^2 = 0.01$ from a circle with a radius $r = 2$ and a center at $(0, 0)$.

![Figure 6. A circle with a radius of 2 and a center of (0,0) in the left column and 50 samples with Gaussian noise (\(\mu = 0, \sigma^2 = 0.01\)) from it on the right.](image)

Nested Vietoris-Rips complex (a filtration) can be obtained by choosing different parameters (or diameter) $\epsilon = 0.5, 1.5, 2.5, 3.5$. See Figure 7 below.
2.2.3 Persistence Diagram and Barcodes

When $\epsilon_0 \leq \epsilon_i \leq \epsilon_j \leq \epsilon_n$, the inclusion $R_{\epsilon_i}(X) \hookrightarrow R_{\epsilon_j}(X)$ induces a homomorphism

$$f_{k}^{i,j}: H_{k}(R_{\epsilon_i}(X)) \rightarrow H_{k}(R_{\epsilon_j}(X))$$

on the homology groups for each dimension $k$. The $k$-th persistent homology groups are the images of these homomorphisms, and the $k$-th persistent Betti numbers $\beta_{k}^{i,j}$ are the ranks of those groups.

Using GUDHI module in Python, the persistence diagram and barcodes for the data points $X_{50}$ as shown in Figures 8, 9.

In Figure 8, there is a green point indicating that a one-dimensional hole appears in the filtration, from about 0.8 birth to 3.4 death. The farther the point is from the diagonal, the more significant the topological feature is. A red point that is furthest from the diagonal indicates that a zero-dimensional hole appears in the filtration, that is, one of the connected components in the filtration has the longest duration. There are many red points in the vicinity of the diagonal. These points went through a short period of time from birth to death and are considered topological noise. At the top of the image is an infinite horizontal line $\infty$ representing the final Vietoris-Rips complex $R_{\epsilon_n}(X_{50})$.
3. Persistent Density Entropy

The persistent homology method constructs a series of Vietoris-Rips complex to approximate the underlying space \( \hat{X} \) of the dataset \( X \), i.e. filtered Vietoris-Rips complex or filtration. However, which Vietoris-Rips complex \( R_\varepsilon(X) \) is the best approximation of the underlying space \( \hat{X} \), or which one in the filtration has the least uncertainty about the underlying space \( \hat{X} \).

It is well known that entropy is a good measure to define uncertainty in information theory and statistics. The entropy defined below similar to information entropy for each Vietoris-Rips complex \( R_\varepsilon(X) \) to measure the uncertainty of the Vietoris-Rips complex approximation of the underlying space \( \hat{X} \).

3.1. Definition

Suppose there is a dataset \( X \) with metric \( d \) containing \( N \) different points, i.e. \( X = \{x_1, x_2, \ldots, x_N\} \). For a certain parameter \( \varepsilon \), the Vietoris-Rips complex is \( R_\varepsilon(X) \). According to the definition of the Vietoris-Rips complex, two points \( x_i, x_j \) are in the same simplicial complex \( \sigma \) when \( d(x_i, x_j) \leq \varepsilon \), that is to say, \( x_i \) falls within the spherical neighborhood \( B_\varepsilon(x_j) \) of \( x_j \) and \( x_j \) falls within \( B_\varepsilon(x_i) \). Let \( \text{card}_\varepsilon(x_i) \) be the cardinality of \( B_\varepsilon(x_i) \) for the point \( x_i \), that is, the number of points included in \( B_\varepsilon(x_i) \). Formulated as follows

\[
\text{card}_\varepsilon(x_i) = \#B_\varepsilon(x_i) = \#\{x_j \in X | d(x_i, x_j) \leq \varepsilon \}.
\]

Let \( \rho_\varepsilon(x_i) \) be the density of point \( x_i \)

\[
\rho_\varepsilon(x_i) = \frac{\text{card}_\varepsilon(x_i)}{\sum_{j=1}^{N} \text{card}_\varepsilon(x_j)}.
\]

Define the density entropy of the Vietoris-Rips complex \( R_\varepsilon(X) \) as \( h_\varepsilon(\varepsilon) \)

\[
h_\varepsilon(\varepsilon) = -\sum_{i=1}^{N} \rho_\varepsilon(x_i) \log(\rho_\varepsilon(x_i)). \tag{2}
\]

Where \( \varepsilon \geq 0 \).

As the parameter \( \varepsilon \) grows, a series of density entropy are defined. Call it the persistent density entropy or the entropy of the filtration of the dataset \( X \).

3.2 Basic property

The persistent density entropy has the following properties, by definition,

1. \( h_\varepsilon(X) > 0 \).
2. \( h_\varepsilon(X) \leq \log(N) \).
3. When \( \varepsilon = 0 \) or \( \varepsilon = \text{diameter}(X) \), \( h_\varepsilon(X) = \log(N) \).
Proof of the above properties.
For property 1.
Because for any point $x_i$,
\[
\text{card}_\varepsilon(x_i) = \#B_\varepsilon(x_i) \geq 1,
\]
so
\[
0 < \rho_\varepsilon(x_i) < 1.
\]
By (2),
\[
h_\varepsilon(X) > 0.
\]
For property 2.
As can be seen,
\[
\sum_{i=1}^N \rho_\varepsilon(x_i) = 1.
\]
The original problem is equivalent to solving the maximum value of the Lagrange function as follows,
\[
L(\rho) = -\sum_{i=1}^N \rho_i \log(\rho_i) + \lambda (\sum_{i=1}^N \rho_i - 1).
\]
The derivative of the function $L(\rho)$ satisfies the property that
\[
\begin{cases}
L_{\rho_1} = -\log(\rho_1) + \lambda = 0 \\
\vdots \\
L_{\rho_N} = -\log(\rho_N) + \lambda = 0
\end{cases}
\]
So
\[
\rho_1 = \cdots = \rho_N = 1/N.
\]
Hence
\[
h_\varepsilon(X) \leq \log(N).
\]
For property 3.
Since $\rho_\varepsilon(x_i) = 1/N$ as $\varepsilon = 0$. Hence
\[
h_\varepsilon(X) = \log(N).
\]
When $\varepsilon = \text{diameter}(X)$, there is $\#B_\varepsilon(x_i) = N$ for all $x_i$.
Hence $\rho_\varepsilon(x_i) = 1/N$, and then
\[
h_\varepsilon(X) = \log(N).
\]

4. Examples
This section gives some examples to verify the effectiveness of our method.

Example 1. The dataset $X_{50}$ with Gaussian noise with a mean $\mu = 0$ and a variance $\sigma^2 = 0.01$ from a circle with a radius $r = 2$ and a center at $(0, 0)$. The metric is the Euclidean distance $d(x_i, x_j) = \|x_i - x_j\|_2$. See Figure 6.

The figure below shows the persistent density entropy diagram of the parameter or the diameter from 0 to 4.

In Figure 10, the red wave curve represents the persistent density entropy of $X_{50}$ and the blue line is a fitted curve. It can be seen that the minimum value of the curve appears near $\varepsilon = 0.2$. However, the persistent density entropy fluctuates sharply near this point, indicating that the influence of topological noise is too large and it is not suitable to select this point to construct the Vietoris-Rips complex approximation of the underlying space $X_{50}$. 
When $\epsilon = 1$, this point is an inflection point of the curve. At the left side of the point, the persistent density entropy increases rapidly, and the persistent density entropy growth on the right side of 1 is significantly slowed down. Referring to Figure 9, it can be seen that this point is the key point for the disappearance of topological noise. After $\epsilon = 1$, the persistent density entropy increases relatively steadily. However, the persistent density entropy still fluctuates significantly in the interval between $\epsilon = 1$ and $\epsilon = 2$.

After $\epsilon = 2$, the persistent density entropy increases steadily. Hence, choosing $\epsilon = 2$ to construct the Vietoris-Rips complex $R_2(X_{50})$ to approximate the underlying space $\hat{X}_{50}$ is more appropriate.

Figure 11 below is the Vietoris-Rips complex $R_2(X_{50})$. The Vietoris-Rips complex $R_2(X_{50})$ and $\hat{X}_{50}$ in the left of Figure 6 are homotopy equivalent.

Figure 11. The Vietoris-Rips complex $R_2(X_{50})$.

And

$\beta_0(R_2(X_{50})) = 1,$

$\beta_1(R_2(X_{50})) = 1.$

**Example 2.** The dataset $X_{50}$ with an outlier $x_{51} = (0, 10)$. For convenience, denoted by $X_{51}$.

First, draw the persistent density entropy diagram as shown in Figure 12 below.

Comparing Figure 10 and Figure 12, it is easy to see that Figure 12 has more inflection points than Figure 7. Hence, three Vietoris-Rips complexes can be extracted from persistent density entropy diagram. The increase persistent density entropy indicates that the uncertainty of the Vietoris-Rips complex approximation of the underlying space is larger.
Second, there are three steady growth points $\epsilon = 2, 6, 9$ after three inflection points $\epsilon = 1, 4, 8$, and the corresponding three Vietoris-Rips complexes $R_2(X_{S1}), R_6(X_{S1})$ and $R_9(X_{S1})$ are as follows

Third, calculate the Betti numbers of their Vietoris-Rips complex separately. See Table 1.

| Betti Numbers | $\epsilon = 2$ | $\epsilon = 6$ | $\epsilon = 9$ |
|---------------|---------------|---------------|---------------|
| $\beta_0$     | $\beta_0(R_2(X_{S1})) = 2$ | $\beta_0(R_6(X_{S1})) = 2$ | $\beta_0(R_9(X_{S1})) = 1$ |
| $\beta_1$     | $\beta_1(R_2(X_{S1})) = 1$ | $\beta_1(R_6(X_{S1})) = 0$ | $\beta_1(R_9(X_{S1})) = 0$ |
| $\beta_i, (i \geq 2)$ | $\beta_i(R_2(X_{S1})) = 0$ | $\beta_i(R_6(X_{S1})) = 0$ | $\beta_i(R_9(X_{S1})) = 0$ |

For Vietoris-Rips complex $R_2(X_{S1}), R_6(X_{S1})$ and $R_9(X_{S1})$, from Figure 12, it can see that the Vietoris-Rips complex $R_2(X_{S1})$ has the lowest density entropy, so it has less uncertainty to approximate the underlying space $X_{S1}$. Finally, choose $R_2(X_{S1})$ as an approximation of the underlying space.

Comparing the three Vietoris-Rips complexes $R_2(X_{S1}), R_6(X_{S1})$ and $R_9(X_{S1})$, it can be known that the point $(0, 10)$ should be an outlier because it causes the entropy to suddenly increase at $\epsilon = 8$.

5. Conclusion
Topological data analysis is well suited for big data analysis. It analyzes big data by studying the shape information and then has an overall grasp of high dimensional complex big data. Persistent homology is an important method in topological data analysis. It avoids considering the specific shape information but indirectly considers the long duration feature as the true topological feature of the big data. This kind of thinking has a good inspiration. The persistent density entropy proposed in this paper, as a topological method for analyzing big data, is not to replace the persistent homology, but as
a complement to make the constructed topological big data shape interpretable. From the above examples, it can be found that the outlier has a great influence on the persistent density entropy diagram. If a suitable distance function can be defined to measure the distance between the two persistent density entropy diagram, it can be easily judged whether the newly added point is an outlier. Or, the newly added points make a big change in the shape of the data.

References
[1] V. De Silva, R. Ghrist, Coverage in sensor networks via persistent homology. *Algebra. Geom. Topol.* 7, (1), 339-358. (2007)
[2] V. De Silva, R. Ghrist, Homological sensor networks. *Notices of the American mathematical society* 54, (1). (2007)
[3] P. M. Kasson, A. Zomorodian, S. Park, N. Singhal, L. J. Guibas, V. S. Pande, Persistent voids: a new structural metric for membrane fusion. *Bioinformatics* 23, (14), 1753-1759. (2007)
[4] C. M. Topaz, L. Ziegelmeier, T. Halverson, Topological data analysis of biological aggregation models. *PloS one* 10, (5), e0126383. (2015)
[5] G. Singh, F. Memoli, T. Ishkhanov, G. Sapiro, G. Carlsson, D. L. Ringach, Topological analysis of population activity in visual cortex. *J. Vis.* 8, (8), 11-11. (2008)
[6] M. K. Chung, P. Bubenik, P. T. Kim In *Persistence diagrams of cortical surface data*, International Conference on Information Processing in Medical Imaging, Springer: pp 386-397, (2009)
[7] Y. Dabaghian, F. Mémoli, L. Frank, G. Carlsson, A topological paradigm for hippocampal spatial map formation using persistent homology. *PLoS computational biology* 8, (8), e1002581. (2012)
[8] G. Sarikonda, J. Pettus, S. Phatak, S. Sachithanantham, J. F. Miller, J. D. Wesley, E. Cadag, J. Chae, L. Ganesan, R. Mallios, CD8 T-cell reactivity to islet antigens are unique to type 1 while CD4 T-cell reactivity exists in both type 1 and type 2 diabetes. *J. Autoimmun.* 50, 77-82. (2014)
[9] T. S. Hinks, X. Zhou, K. J. Staples, B. D. Dimitrov, A. Manta, T. Petrossian, P. Y. Lum, C. G. Smith, J. A. Ward, P. H. Howarth, Innate and adaptive T cells in asthmatic patients: relationship to severity and disease mechanisms. *J. Allergy Clin. Immunol.* 136, (2), 323-333. (2015)
[10] J. Gamble, G. Heo, Exploring uses of persistent homology for statistical analysis of landmark-based shape data. *J. Multivar. Anal.* 101, (9), 2184-2199. (2010)
[11] Z. Zhou, Y. Huang, L. Wang, T. Tan, Exploring generalized shape analysis by topological representations. *Pattern Recognit. Lett.* 87, 177-185. (2017)
[12] D. Freedman, C. Chen, Algebraic topology for computer vision. *Computer Vision*, 239-268. (2009)
[13] X. Zhu In *Persistent Homology: An Introduction and a New Text Representation for Natural Language Processing*, IJCAI, pp 1953-1959, (2013)
[14] E. Merelli, M. Rucco, P. Sloot, L. Tesei, Topological characterization of complex systems: Using persistent entropy. *Entropy* 17, (10), 6872-6892. (2015)
[15] V. Snášel, J. Nowaková, F. Xhafa, L. Barolli, Geometrical and topological approaches to Big Data. *Future Gener. Comput. Syst.* 67, 286-296. (2017)
[16] M. Offroy, L. Duponchel, Topological data analysis: A promising big data exploration tool in biology, analytical chemistry and physical chemistry. *Anal. Chim. Acta* 910, 1-11. (2016)
[17] H. Edelsbrunner, D. Letscher, A. Zomorodian In *Topological persistence and simplification*, Proceedings 41st Annual Symposium on Foundations of Computer Science, pp 454-463, 2000, (2000)
[18] K. Borsuk, On the imbedding of systems of compacta in simplicial complexes. *Fundam. Math.* 35, (1), 217-234. (1948)
[19] D. Kozlov, *Combinatorial algebraic topology*. Springer Science & Business Media: Vol. 21, (2007)