Journal of the Egyptian Mathematical Society 25 (2017) 406–413

Contents lists available at ScienceDirect

Journal of the Egyptian Mathematical Society

journal homepage: www.elsevier.com/locate/joems

Original Article

On the Topological Data Analysis extensions and comparisons

H.N. Alaa*, S.A. Mohamed

Department of Mathematics, Faculty of Science Aswan University, Egypt

ARTICLE INFO

Article history:
Received 1 April 2017
Revised 16 June 2017
Accepted 8 July 2017
Available online 24 July 2017

Keywords:
55N35
55U05
62H99
Topology – homology-persistence landscape
Topological Data Analysis
Testing hypotheses

ABSTRACT

Topological Data Analysis is an emerging field at the intersection of algebraic topology and statistical inference aimed at describing the shapes objects represented as point cloud data in the multidimensional space. Since the range of applications of shape analysis is enormous, new tests have given birth to the field of TDA. In this habilitation study three TDA-oriented tests are discussed. A new test based on metric functions is proposed. A small simulation study among the preceding tests has been employed via Monte Carlo simulation. All the mentioned tests in the vignette are activated by real world data within educational field.

1. Introduction

In a wide variety of disciplines, it is of great practical importance to measure, sketch and compare the shapes between different objects. Dryden and Mardia [1] defined the shapes of certain objects as all the geometric information that remains when location, scale and rotational effects are filtered out. If the size information is also of interest, then the scale will be omitted from the definition. Here the size of the information will be taken into consideration. In other words, we can claim that two objects have the same shape if by the translation, shifting or rotation operations the two objects will coincided, see [2]. The fundamental field concerning with studying the geometric properties of the objects is topology. Indeed, topology has been present in mathematics for quite a long time without anticipating applications to real-world applications until the beginning of this century. As, Carlsson in [3] proposed his survey article which produced another new area of research known as computational topology that enables the researchers to extract the quantitative and qualitative information that describe the point cloud data's shapes.

Computational topology is a set of algorithmic methods developed to understand topological invariants such as loops and holes in high-dimensional data sets. The specialized approach that employs the statistical tools to compute and analyze the topological features is called TDA. Generally speaking, TDA refers to a collection of methods and tools that enable the researchers for finding and studying the topological invariants structure in data. The input of these procedures typically takes the form of a point cloud data which is usually represented as a large finite dataset sampled from a geometrical object in a n-dimensional metric space, possibly with some noise. The output is a collection of data summaries and diagrams that are used to estimate the statistical features of the data. Lesnick [4] divided TDA tools into two parts: the first one is the descriptors TDA which are the procedures that aim at describing, summarizing, discovering, and visualizing point cloud data. However, the second is TDA inference which uses the probability theory to investigate or test the statistical features of the sample data (e.g. mean, variance…etc.).

In the last few years, community topology has witnessed important progress in supporting complex data analysis. In consequence, TDA plays a crucial role in a variety of different fields range from industry [5] shape classification Chazal et al. in [6, 7], clustering and histology images for breast cancer analysis [8]. In addition, TDA has received recently much attention by statisticians which gives a birth to a competitor approach in the data mining. For instance, Singh et al in [9] proposed a new classification tool based on simplicial complexes figures called Mapper, Kent et al. in [10] introduced k-tree level sets which can be utilized in the classification and comparison purposes, Turner [11] defined the means and medians for the persistent homology diagrams, from [12] derived confidence band for the persistence diagram that allows us to separate topological signal from topological noise, Chazal in [13] proposed sub-sampling methods for analyzing the shape of sets and functions from point cloud data in the case of the sample is too large.

* Corresponding author.
E-mail addresses: ala2222000@yahoo.com (H.N. Alaa), statisticsMS.2010@gmail.com (S.A. Mohamed).

http://dx.doi.org/10.1016/j.joems.2017.07.001

The major motivation beyond the present study is to provide a review for the three tests based on TDA using for testing the similarities between the objects. Further, propose a new test based on metric functions can be employed for the same purpose. In addition, conducting a power comparison study between the tests based on TDA and the proposed tests a benchmarking test. This article is structured as follows: the next section will give a snapshot of TDA tools. The third section includes all the tests that can be employed for testing the closeness between the objects. The followed section is devoted for the Monte Carlo results. The final section presents the results concerned to the real life applications.

2. Topological Data Analysis

The general framework of TDA for computing topological features from point cloud data usually contains two necessary steps: constructing simplicial complexes and applying TDA techniques on the simplicial complexes frequently are the persistent homology, barcodes and the persistent landscape. The main textbook for this section is Edelsbrunner and Harer [14].

A simplicial complex S is a set consisting of a finite collection of p-simplices (simple pieces), where a 0-simplex is a vertex, a 1-simplex is an edge, a 2-simplex is a triangle, a 3-simplex is a tetrahedron, and so on. In more precise way, the simplicial complex divided the space into smaller and topologically simpler pieces, which when assembled back together carry the same aggregate topological information as the original space. These simplices should satisfy two conditions. First, for every set σ in S, every nonempty subset $\tau \subset \sigma$ also belongs in S. For instance, if tetrahedron abcd is in S, then the triangles abc, abd, acd, bcd, the edges ab, ac, ab and the vertices a, b, c, d are also in S. Second, two p-simplices are either empty or they intersect in a lower dimensional simplex. In order to obtain simplicial complex sets, Vietoris–Rips filter is advocated in this study.

Homology is a tool from algebraic topology that measures the features of a topological space such as an annulus, sphere, torus, or more complicated surface. In particular, homology can distinguish these spaces from one another by quantifying their connected components, loops, voids, and so forth. One interesting feather associated with the homology group is the Betti numbers, as they provide meaningful information about the complex. Roughly speaking, the pth Betti number β_p is the number of pth dimensional independent holes in the homology groups, so that β_0 is the number of connectedcomponents, β_1 is the number of loops, β_2 is the number of enclosed voids and so on. Persistent homology is the primary algebraic topology tool was developed by Edelsbrunner et al. [29] used in the TDA methods in order to track long persist features. It provides a way to measure the lifespan of a topological feature, which is the persistence of the feature, whereas short-lived features may be ignored as noise.

A convenient way to visualize persistent homology is through a graphical representation called a barcode which can summarize the information encoded in the persistence diagram in a different vision. There is a distinct barcode for each homology space from which we infer the Betti number. In other words, the length of every line in the Barcodes diagrams refers to the distance between the time of death j and the time of born i, the number of the lines associated to dimension zero equals to β_0, while the number of the lines associated to dimension one equals to β_1 and so on.

Another graphical way that can summarize the information contained in the persistent homology diagram is the persistent Landscape proposed by Bubenik [15]. Persistent Landscape can be considered as a rotated version of barcode plot. The main advantage of the Persistent Landscapes is it allows us to calculate and summarize the data with the standard statistics indicators e.g. means, median, variance...etc, as opposite to either persistence diagram or barcode plot. To define the landscape, construct a triangle whose base corresponds to a persistence intervals and the top vertex by tenting each persistence point using the following function:

$$\Lambda_s(\varepsilon) = \begin{cases} \varepsilon - i & \varepsilon \in \left[i, \dfrac{i+j}{2}\right] \\ j - \varepsilon & \varepsilon \in \left(\dfrac{i+j}{2}, j\right] \\ 0 & otherwise \end{cases}$$

where ε is the filtered simplicial complex time and s takes 1 to n, n is the number of the points in the persistent diagram. It should be noted that $\Lambda_s(\varepsilon)$ obtained separately to each p-dimension. Formally, $\lambda_s(\varepsilon)$ is the sth largest value of $\Lambda_s(\varepsilon)$ taken into consideration the homology dimension. When $s = 1$, of course, $\lambda_s(\varepsilon)$ can be interpreted as the maximal possible distance of an interval centered aboutε. Fig. 1 applied all the TDA's tools, mentioned above, to a sample drawn from tours.

3. Statistical shape analysis

Shape analysis is an active subject of academic research in the both of mathematical and applied sciences. It has extensive applications in many fields as it is great practical importance to carry out hypotheses tests that distinguish between objects under uncertainty. A plenty of tests have been suggested in the literature (see [2]). However, three different tests will be focused in this context. Assume that you have K-objects and that we would like to test the null hypothesis that all the objects are similar and have the same shape versus the alternative hypothesis that states that at least one object differs than the others. This can be achieved by the following tests which are so called k-sample tests.

3.1. Statistical inference using persistent homology

Gamble in [2] produced a new test which can be dependable for testing the similarity between two persistent homology diagrams using Wasserstein distance. Robinson and Turner in [16] generalized the test of Gamble in the multivariate case; as if it is required to test between two sets of persistent homology. In the present paper, it will generalize from [2], test into K samples. The test statistic that can be utilized to test between K persistent homology diagrams Pin the light of Gamble and Heo may be expressed as:

$$T_R = \frac{1}{\binom{k}{2}} \sum_{i=2}^{k} \sum_{j=1}^{i-1} W\left(P_i, P_j\right)$$

where $W(P_i, P_j)$ is the Wasserstein distance between $\mathrm{P_i}$ and $\mathrm{P_j}$. Obviously, T_R can be considered as the average of all pair wise Wasserstein distances. Robinson in [2] recommended using the Hungarian algorithm to compute the Wasserstein distance.

Given $\mathrm{p}_{1,1}, \mathrm{p}_{2,1} \dots \mathrm{p}_{\mathrm{n}_1,1}$ and $\mathrm{p}_{1,2}, \mathrm{p}_{2,2} \dots \mathrm{p}_{\mathrm{n}_2,2}$ are the points corresponding to P_1 and P_2 respectively.The Hungarian algorithm required, first, that the two persistent homology have to be the same size, this is done via adding n_2 points to the first sample and n_1 points to the second sample, which yields we have $\mathrm{n}_1 + \mathrm{n}_2$ points for the both persistent homology. The added points are copy of a diagonal that are the perpendicular distances. Then, constructing the cost matrix where its entries are the squared Euclidean distances. Next, match every row with the optimum column.[1] Finally, the Wasserstein distance is the sum up for the optimum distances,

[1] The optimum column means that the column that has least distance.

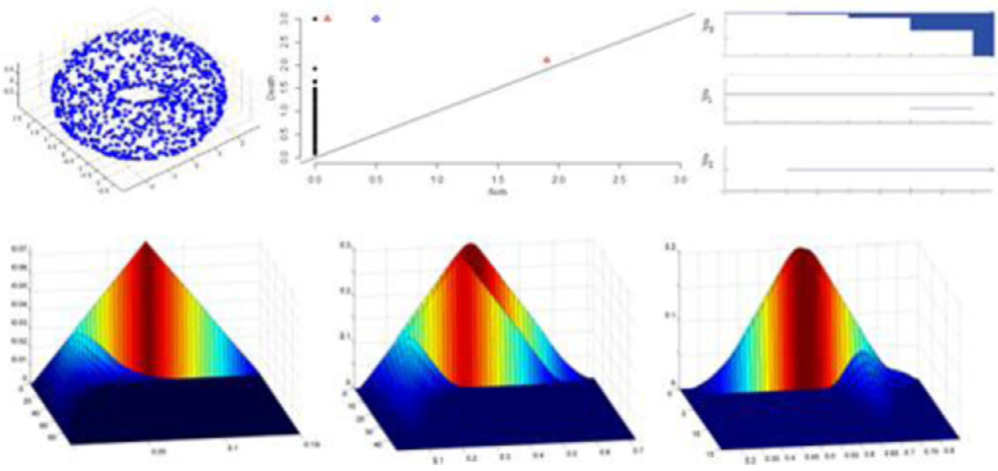

Fig. 1. Top left is tours sample data. Top middle the persistent homology. Top right the bare code. The bottom is the persistent landscape at dimension 0, 1 and 2.

which means that the Hungarian algorithm gives us the minimum cost value.[2]

Since the sampling distribution for T_R is unknown, a lot of statistical nonparametric can be used to estimate an empirical distribution for the test statistics, e.g. bootstrapping, jackknifing, or a permutation test. Necessary condition for implementing this methodology is that the drawn samples are representative of their populations; since our datasets are randomly generated, there is no reason to suspect the condition does not hold. According to Robinson [16], the permutation test is preferred to judge on the significance of T_R.

The permutation approach is a nonparametric tool implying that permuting (rearranging) the data by shuffling their labels in the sample, and then calculate T_R on each permutation. The collection of T_R from the permuted data constructs the null distribution. In other words, If the two compared groups are statistically indistinguishable, then random permutations applied to the observed data do not make a difference; in that case, the observed test statistic lies within the permutations values. On the other hand, if the two groups statistically different, then random permutations make a difference; in that case the observed test statistic takes an extreme value i.e. it is located in the tail of the null distribution. The stages of obtaining the P-value using the permutation test in the case $K = 3$ as follows:

```
Data: $P_1$, $P_2$ and $P_3$ with three sample sizes $n_1$, $n_2$ and $n_3$ respectively. $M$is the
  number of the permutation samples.
Results: $P$-value for $T_R$
Compute $T_R$ from the original sample data.
for $i = 1 : M$
    Randomly shuffle the group labels into disjoint sets of size
  $n_1$, $n_2$ and $n_3$;
    Compute $T_R$ for each permutation sample and save the values in $E_i$;
end.
$P$-value is the # of times that $E_i$ greater than $T_R$ divided by $M$.
```

3.2. Statistical inference using barcode sets

On another hand, Máté in [17] proposed another test for testing the similarity between certain configurations. Alternatively, they decided to depend on barcode diagrams instead of persistent points to build their test using Jaccard index as follows:

$$T^*_M = \frac{1}{\binom{k}{2}} \sum_{i=2}^{k} \sum_{j=1}^{i-1} J^*\left(B_i, B_j\right)$$

where B_i is the barcode diagram of the sample i and $J^*(B_i, B_j)$ is the Jaccard measure between B_i and B_j that is defined as the size of the intersection divided by the size of the union of the barcode diagram taking the following formula:

$$J^*\left(B_i, B_j\right) = \frac{\left|B_i \cap B_j\right|}{\left|B_i \cup B_j\right|}$$

Jaccard measure is more suitable for testing the similarity as its value lies between zero and one, as upper values means the closeness while lower values refer to the dissimilarity. In consequences, the range of T_M lies in [0,1]. A serious dilemma faced during implementing $J^*(B_i, B_j)$ to the barcode sets that it cannot straightforward to apply $J^*(B_i, B_j)$ to the barcode diagrams. As, it isn't necessary that the two barcode diagrams have the same number of the barcode sets. Further, each barcode diagram has multi-barcode sets, which is required to find a condition for a perfect matching criterion between the barcode diagrams. Therefore Máté modified $J^*(B_i, B_j)$ in the following way:

$$J\left(B_i, B_j\right) = \frac{1}{|B_i| + |B_j|}\left[\sum_t \sup_h \frac{\left|B_{it} \cap B_{jh}\right|}{\left|B_{it} \cup B_{jh}\right|} + \sum_h \sup_t \frac{\left|B_{it} \cap B_{jh}\right|}{\left|B_{it} \cup B_{jh}\right|}\right]$$

where B_{it} and B_{jh} are barcode sets within the barcode diagram of the sample i and j respectively, and $|B_i|$ is the number of sets in the barcode diagram of the sample i. Clearly, $J(B_i, B_j)$ can perfectly apply to the barcode diagrams and in the same time still ranged in the interval [0,1]. Hence, T^*_M will be become in terms of $J(B_i, B_j)$ as:

$$T_M = \frac{1}{\binom{k}{2}} \sum_{i=2}^{k} \sum_{j=1}^{i-1} J\left(B_i, B_j\right)$$

It should be noted that T_M will be operated separately for every homology dimension. The phases of calculating the T_M can be summarized in the case of $K = 3$ as follows:

[2] Wasserstein distance is computed separately for points in dimensions zero, one and two...etc.

```
**Data**: $B_1,\ B_2$ and $B_3$ with barcode sets $B_{1t}$,$B_{2t}$ and $B_{3t}$respectively.
**Results**: $T_M$ statistical test.
Compute $T_M = \frac{J(B_1,B_2)+J(B_1,B_3)+J(B_2,B_3)}{3}$
Compute $J(B_i,B_j)$ =[ $\frac{j(B_i,B_j)+j(B_j,B_i)}{|B_i|+|B_j|}$ ]
Compute$j(B_i,B_j)$
        for k=1:$|B_i|$
          for h=1:$|B_j|$
        t[k,h]= $\frac{\min(B_{ik(end)},B_{jh(end)})-\ \max(B_{ik(begin)},B_{jh(begin)})}{\max(B_{ik(end)},B_{jh(end)})-\ \min(B_{ik(begin)},B_{jh(begin)})}$%% : $end\ \wedge begin$ : the end and
begin of barcode interval
          if t[k,h]<0 set t[k,h]=0
        **end; end; end**
        $j(B_i,B_j)$=sum(max(t[k,:]))
```

Since obtaining the limiting of T_M isn't a trivial issue, permutation test is adopted to obtain the critical values as the same as done with T_R.

3.3. Statistical inference using persistent landscape

Bubenik in [18] decided to use the average of $\lambda_s(\varepsilon)$ in order to derive anew statistical test[3] that can be reliable to investigate the difference between two given shapes in the high dimensional case. He has proved, using central limit theorem, that $\lambda_s(\varepsilon)$ is asymptotically normal with mean η_s^* andvarianceVar$(\lambda_s(\varepsilon))$, where $\eta_s^* = \frac{\sum_{\varepsilon=1}^{\mathrm{T}} \lambda_s(\varepsilon)}{\mathrm{T}}$. Since the assumption of equal variances is typically violated, thus he stated that in order to test the significance differences between two given shapes, it suffices to compute Welch's Test between their persistent landscapes at different values of s and different dimensions holes. Likewise, whether it is required to test the significance differences between K shapes, one can easily operate Welch's Test between their persistent landscapes which takes following equation:

$$T_{Bs} = \frac{A_s}{1+B_s}$$

where $A_s = \sum_{j=1}^{K} W_s^j (\eta_s^j - \bar{\eta}_s)^2 / k-1,\ B_s = \frac{2(k-2)}{k^2-1} \sum_{j=1}^{K} (1-W_s^j)^2 / T-1, \quad \bar{\eta}_s = \sum_{j=1}^{K} \eta_s^j / k, \quad \eta_s^j = \sum_{j=1}^{K} W_s^j \eta_s^{*j} / \sum_{j=1}^{K} W_s^j, \quad W_s^j = T/Var(\lambda_s^j(\varepsilon))$ and η_s^{*j} is the mean of the persistent landscape corresponding to the sample j. According to [19] assuming the normality and the independency conditions, $\mathrm{T_{Bs}}$ has under H_o approximately F-distribution with degrees of freedoms $\left[k-1, \left[\frac{3}{k^2-1} \sum_{j=1}^{K} \left(1 - \frac{W_s^j}{\sum_{j=1}^{K} W_s^j}\right)^2 / T-1\right]^{-1}\right]$.

3.4. Statistical inference using metric spaces

Instead of depending on TDA tools, one can resort directly to the sample's points as an indicator of the sample's shape through computing the magnitudes within the observations' coordinates. In other words, we would like to measure how far among the points cloud data internal each sample space. Once we have recorded these numbers via any suitable metric function, we can compare them among to get a feel for how similar they are. The main advantageous of distance-based estimators that is invariant to rotation and translation operations. Let $\mathrm{X}_1, \mathrm{X}_2 \ldots \mathrm{X}_\mathrm{k}$ be K-samples with $\mathrm{X_i} =\ [\mathrm{x}_1, \mathrm{x}_2 \ldots \mathrm{x}_{\mathrm{n}_i}]$and it is required to test the similarity, this may be achieved by the following proposed test:

$$T_P = \frac{1}{\binom{k}{2}} \sum_{i=2}^{k} \sum_{j=1}^{i-1} R\big(X_i, X_j\big)$$

where

$R(X_i, X_j) = \max(\sum D(X_i), \sum D(X_j)) / \min(\sum D(X_i), \sum D(X_j))$ and $D(X_i)$ is a symmetric $n_i \times n_i$ matrix represents all pair wise distances within the sample$\mathrm{X_i}$with zeros along the diagonal. Actually the distances can be obtained using unaccountable metric functions. Yet, we will consider the following metrics for points $x_i = (x_{i1}, x_{i2})$ and $x_j = (x_{j1}, x_{j2})$ as:

1) Euclidean Distance: $d(x_i, x_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2}$
2) Taxicab Distance: $d(x_i, x_j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}|$
3) Supremum Distance: $d(x_i, x_j) = \max(|x_{i1} - x_{j1}|, |x_{i2} - x_{j2}|)$

For more details about these metric functions see [20]. One can easily deduce that the ratio of the maximum to the smallest entry in the $R(X_i,\ X_j)$ can be used as a statistic for degree of shape difference. If the configurations are very similar, then the $R(X_i,\ X_j)$ will be close to one, larger values of $R(X_i,\ X_j)$ indicate a greater degree of dissimilarity. Hence T_P can be considered as the average of the serrations for all possible pair wise of $R(X_i,\ X_j)$. Indeed, one may think that the proposed test can be considered as an extension to the test of Lele and Richtsmeier (1991) in the K-sample with slightly different. Yet, there are important differences between the two tests, as our test didn't assume the equality of the covariance matrices for which didn't rely on the variances of the distances, further our test can be computed, without any changing, at any data's dimension levels as opposite to the another test. For more details see [21,22]. Since we haven't any knowledge about the underlying distribution of T_P to get the critical values, depending on the re-sampling methods, one can gain information about T_P. Bootstrap procedure was adopted to estimate the critical values which briefly described below in the case of $K = 3$.

```
**Data**: $X_1,\ X_2$ and $X_3$ with three sample sizes $n_1,\ n_2$ and $n_3$ respectively and $M$
   is the number of the repeated samples.
**Results**: $P$-value for $T_P$
Compute: $P$-value for $T_P$ from the original sample data
Obtain the pooled sample as stack $X_1,\ X_2$ and $X_3$into a new variable $X$.
 **for** $i = 1 : M$
        Obtain a simple random sample with replacement of size $\sum_{i=1}^{3} n_i$
        Consider the first $n_1$as $X_1$, the second $n_2$ as $X_2$ and the remaining as $X_3$.
        Compute $T_P$ for a random sample and save the values in $E_i$;
**end**.
 $P$-value is the # of times that $E_i$ greater than $T_P$ divided by $M$.
```

4. A small simulation study

In this part, the practical performance for the above tests is investigated. We compare between the proposed test T_P corresponding to different metric functions respectively to these existing tests T_R,$\mathrm{T_M}$ and $T_{B1,\,2}$, where $T_{B1,\,2}$ is testing based on the average of the first two largest landscape values. We have applied the mentioned tests to the common geometric objects that can be conducted through GEOZOO Package [23], then recorded the p-values of each test using TDA [24] and ONEWAYTESTS (see [1]) packages. When the two groups are generated from the same geometric objects, the p value in this case is denoted as the size of the test. Otherwise, the p value is considered as the power of the test.

Since it maybe intractable to do any theoretical comparisons about the performance of the previous tests, thus one has to resort to compare through Monte Carlo simulation. Monte Carlo simulation is now a much-used scientific tool for problems that are analytically intractable and for which experimentation is too time-consuming, costly, or impractical. It depends basically on generating artificial random sampling many times, 1000 times for instance, in order to estimate the statistical models and the mathematical functions. Even though, simulation also has disadvantages; it can require huge computing resources, it doesn't give exact solutions, and results are only as good as the model and inputs used.

[3] The Test I will be used here.

Table 1
Simulated size and the power of the test statistics under the study.

Sample size	Model	a	Dimension	T_R	T_M	$T_{B1,\,2}$	T_P		
							E	T	S
20	Model(1)	$a=0$	0	.02	.05	.03	.01	.02	.01
			1	.03	–	–			
		$a=\frac{6}{\sqrt{N}}$	0	.20	.30	.10	.25	.70	.80
			1	.68	–	–			
	Model(2)	$a=0$	0	.01	.03	.20	.02	.02	.01
			1	.03	–	–			
		$a=\frac{6}{\sqrt{N}}$	0	.75	.40	.90	.30	.90	.93
			1	.40	–	–			
	Model(3)	$a=0$	0	.01	.15	.24	.03	.02	.02
			1	.02	.10	.40			
		$a=\frac{6}{\sqrt{N}}$	0	.72	.32	.95	.30	.80	.90
			1	.24	–	–			
	Model(4)	$a=0$	0	.02	.03	.40	.02	.03	.01
			1	.06	–	.30			
		$a=\frac{6}{\sqrt{N}}$	0	.30	.20	.90	.20	.54	.45
			1	.10	–	1.0			
50	Model(1)	$a=0$	0	.01	.02	.01	.01	.01	.01
			1	.01	.01	–			
		$a=\frac{6}{\sqrt{N}}$	0	.30	.40	.20	.55	1.0	1.0
			1	.80	.25	–			
	Model(2)	$a=0$	0	.01	.01	.01	.01	.01	.01
			1	.01	.01	.20			
		$a=\frac{6}{\sqrt{N}}$	0	1.0	.87	.85	1.0	1.0	1.0
			1	.85	.30	1.0			
	Model(3)	$a=0$	0	.01	.20	.04	.02	.01	.01
			1	.02	.30	.02			
		$a=\frac{6}{\sqrt{N}}$	0	1.0	.70	1.0	.35	1.0	1.0
			1	.50	.30	1.0			
	Model(4)	$a=0$	0	.01	.01	.30	.01	.01	.01
			1	.01	.07	.20			
		$a=\frac{6}{\sqrt{N}}$	0	.80	.22	1.0	.35	.84	.90
			1	.25	.10	1.0			

Table 2
The empirical *P*-values for testing the similarity based on the statistical tests.

	Dimension	T_R	T_M	$T_{B1,\,2}$	T_P		
					E	T	S
GPvsMS	0	<.001	<.001	.90	<.001	<.001	<.001
	1	<.001	.06	.33			
Mathematics vs Language	0	.53	.97	.51	.75	.89	.80
	1	.93	.04	.25			

Comparisons between the statistical tests should be conducted under various situations which can be summarized as:

1 Different sample sizes: We will operate our simulation under two different sample sizes which are 20 and 50.
2 Different dimension holes: At β_0 and β_1 the size and the power of each test, except tests based on metric functions, are calculated which allows us to demonstrate at which dimension the tests can capture perfectly the topological feathers of the objects.
3 In the light of [25], it will be considered only the case K = 3, where the first and the second objects are generated from:
a. Model(1): The circle with radius equals one
b. Model(2): The torus with radius from the center equals two.
c. Model(3): The klein bottle with inner radius equals one.
d. Model(4): The standard multivariate normal with three variables.

while the third object corresponding to each model is generated through $(1-a)$ Model(i) where $i = 1 \ldots 4$, a takes the both values zero and $2\ /\sqrt{N}$ and N is the gross sample size.

One can note that when $a = 0$, the p value refers to the size of the test, whilst $a > 0$, the p value refers to the power of the test. In the same manner, the alternatives go to the null as the sample size increases. Results based on Monte Carlo simulation are implemented with 1000 replicates and 100 repeated samples at 99% confidence interval under the above conditions and corresponding to the Vietoris–Rips complex which are consistent with the prior studies (see [19]). Final results are organized and reported in the Table 1. A number of conclusions is drawn from the overall results and summarized in the following points:

1) The results depicted the sample size has strong effect on the simulated size and the power of the tests, as increasing the sample size yields the tests' size tends to the correct nominal level and the tests' power increases in spite of the effect of factor a. Thus it can be recommended to use these tests at large sample sizes.
2) It is obvious that the Betti dimension can be considered as an effect or factor on the behavior of all tests based on TDA. Generally speaking, the performance of T_R and T_M is better at β_0. This phenomenon can be explained by the fact that the number of the points at β_0 is greater than the number of the points at upper dimensions yields that the final decision based β_0 dimension is more accurate compared to other dimension. Another problematic point related to the

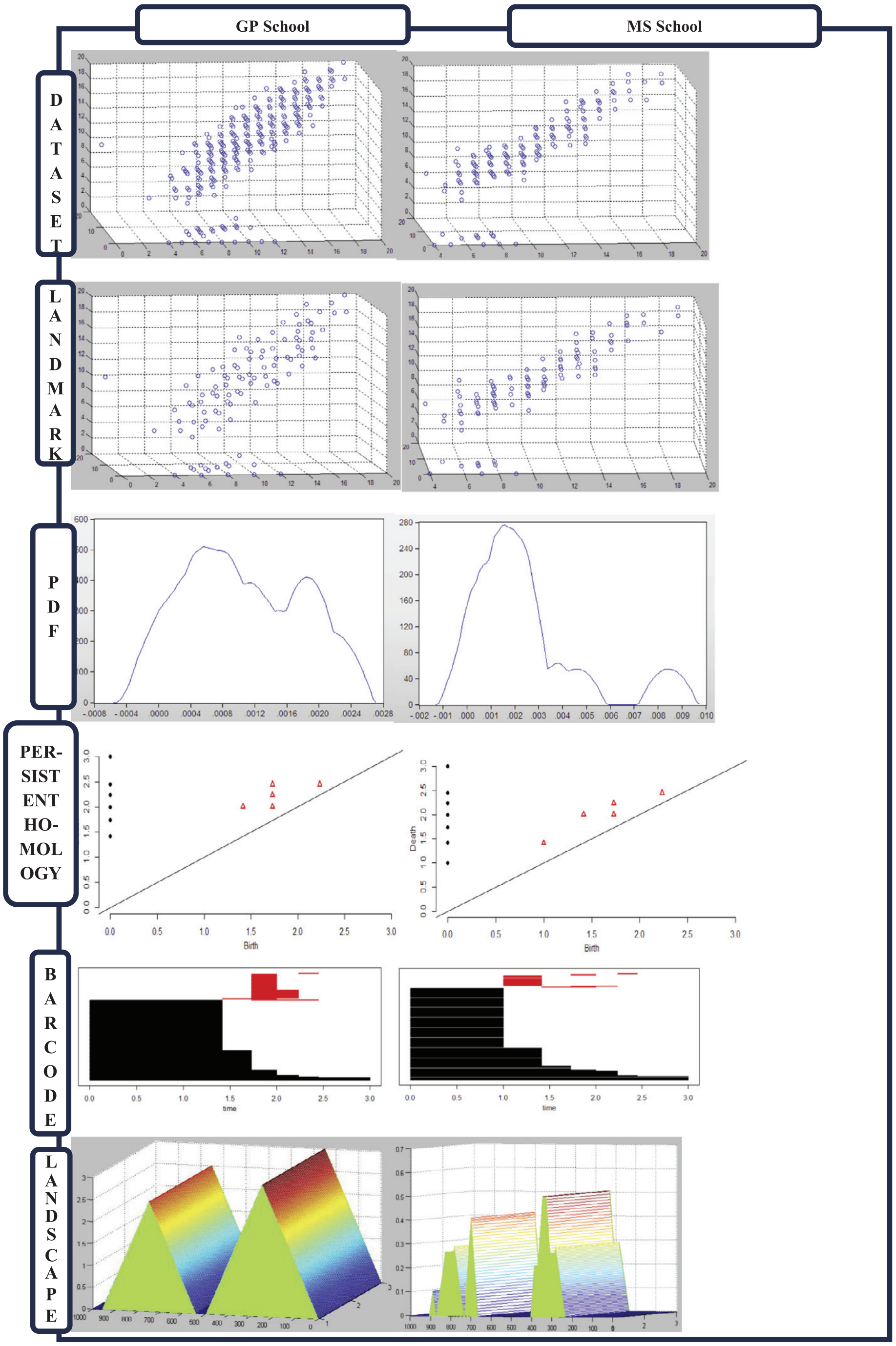

Fig. 2. The statistical features for the two portuguese schools databases respectively.

holes' upper dimensions is at low sample sizes: the holes most likely not to appear, which prevents the tests to be computed based on.

3) It is easily to notice the superiority of T_R to all the tests based on TDA in terms of either the size or the power. However T_M the least statistical poweramong TDA tests. Whereas $T_{B1,\,2}$ seems to be anticonservative for $\alpha = .01$.
4) It can be obviously seen that all the tests based on metric functions have satisfied type I error, even at small sample size. Regarding to the power, T_P based on Euclidean metric achieves the lowest level among the proposed tests.
5) It is observed that T_P based on supremum metric is the winner among all the tests in terms of the size and the power. In contrast, T_M has clearly poor power in almost simulated cases, while $T_{B1,\,2}$ has inflated type I error for which not recommended at the small samples.

5. Real life applications

Testing the similarities using TDA has been used in a wide variety of disciplines, as it is very helpful to tool in analyzing and exploring a large amount of datasets. During the past few decades, there is increasing recently of using TDA in various fields. In this study, empirical world data set related to student data from two public Portugal secondary schools are analyzed. This data is collected from Gabriel Pereira (GP)(772 students) and Mousinho da

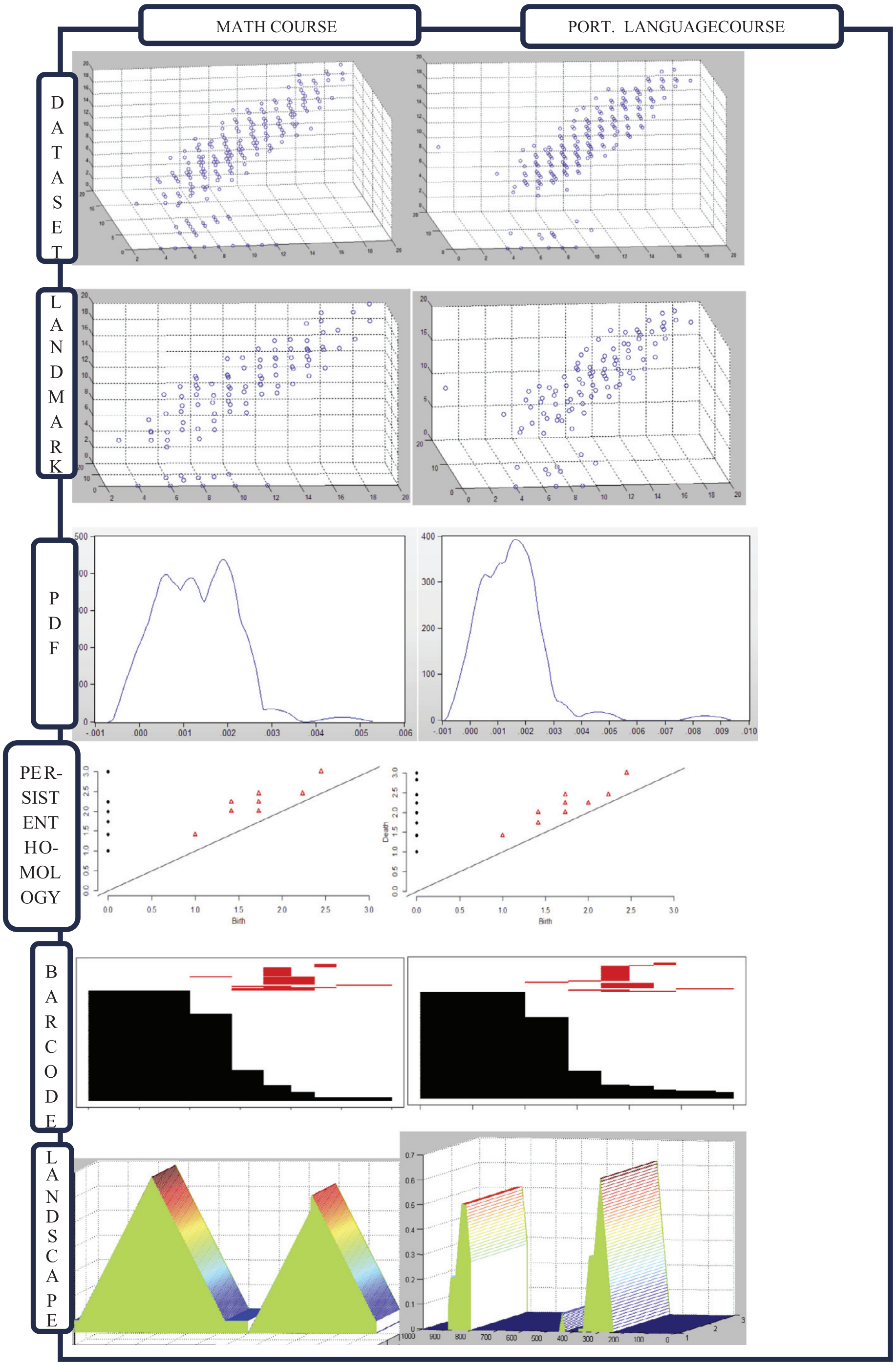

Fig. 3. The statistical features for the math and portuguese language courses databases respectively.

Silveira (MS) schools (272 students) during 2005–2006 and consists from several demographic, social and school related attributes (e.g. student's gender, alcohol consumption, grades' students). Further, in [26] divided separately these students into two categories: 1- students studied Math course (395), 2- students studied Portuguese language course(649), and used several data mining techniques for forecasting the students' final grades.

The present work tends to assess the similarities between GP and MS schools with respect to the students' grades. In more precise way, we would like to determine whether the students' marks for the two schools are similar or not. According to the argument of [27] the two schools have a similar pattern if they are applying the same common criteria, instructions and procedures to evaluate the performance of the students, for which we can guarantee that the students into the two schools get equal treatment. Likewise, assuming that the mathematical course's markers are completely independent than the Portugal course's markers, one can test the similarities between the mathematical students and the Portuguese students in terms of grades.

In our data, students are evaluated three times yearly ranged from 0 to 20. In order to make the process of the analysis faster and easier, we resort to select 100 landmarks points from each school (subject) using the Sequential Maximum Landmark Method operated by JPLEX package. Figs. 2 and 3 reveal the most important statistical feathers according to each school and course subject respectively. From a first look at Fig. 2, one can clearly deduce that the two schools are highly different with respect to the students' grades. In contradiction, it is easily to notice from the diagrams of Fig. 3 that the two courses are slightly differing with respect to the students' grades. In the light of The *P*-values appeared in Table 2, we can conclude that the GP's graders mark in a different way than MS's graders. Whilst, the mathematical markers are assigning grades that are nearly similar than those from the Portuguese markers, which means that within each school its mathematical and Portuguese markers are using the same common criteria.

6. Conclusions

In this article, we have shown the implementations of TDA tools in testing the similarities among different configurations. A new test based on metric function is suggested using several distance function. A comparison among the tests based on persistent homology, barcode sets, persistent landscape and the test based on metric function is conducted at different patterns under two criteria: 1- The size of the test. 2- The power of the test. Our results indicated that, tests based on persistent homology and metric function have more suitable Type I error and satisfied power than the others. Generally speaking, at dimension zero tests based on TDA has satisfied properties and increasing the sample size of the point cloud data has a positive effect on the whole tests. Further, we have illustrated the strength of the preceding tests on the Wisconsin breast cancer dataset.

Considering future researches, there is a plenty of work to do. For instance, comparing between the different methods given in [28] with the above mentioned methods, developing TDA tests in order to improve its performance, studying in depth clustering based on TDA and comparing with the other statistical known methods…etc. Lastly, we would like to mention that several issues deserve future attention that we believe that as we will progress in TDA tools, more researchers will adopt the topological analysis in their work.

Acknowledgments

The authors will like to express their heartiest thanks to the TDA Team: Brittany Terese Fasy, Jisu Kimand Clement Mariafor their support, guidance, immense knowledge and their generously sharing their time. Further, special thanks to Dr. Fabrizio Lecci for his co-operating about sending his thesis.

References

[1] O. Dag, A. Dolgun, M. Konar, One way tests, 2015. http://cran.r-project.org/web/packages/onewaytests/index.html.
[2] J. Gamble, G. Heo, Exploring uses of persistent homology for statistical analysis of landmark-based shape data, J. Multivariate Anal. 101 (9) (2010) 2184–2199.
[3] G. Carlsson, Topology & data, Bull. Am. Math. Soc. 46 (2) (2009) 255–308.
[4] M. Lesnick, Multidimensional Interleavings and Applications to Topological Inference PhD thesis, Stanford University, 2012.
[5] O. Artamonov, Topological Methods for the Representation and Analysis of Exploration Data in Oil Industry PhD thesis, University of Kaiserslautern, 2010.
[6] F. Chazal, D. Cohen-Steiner, L. Guibas, F. M'emoli, S.Y. Oudot, Gromov-Hausdorff stable signatures for shapes using persistence, Comput. Graphics Forum (2009) 1393–1403.
[7] F. Chazal, L. Guibas, S.Y. Oudot, P. Skraba, Persistence-based clustering in Riemannian manifolds, J. ACM 60 (2013) 6–41.
[8] N. Singh, H. Couture, S. Marron, C. Perou, M. Niethammer, Topological descriptors of histology images, Mach. Learn. Med. Imag. (2014) 231–239.
[9] G. Singh, F. Mémoli, G.E. Carlsson, Topological methods for the analysis of high dimensional data sets and 3d object recognition, SPBG. Citeseer (2007) 91–100.
[10] B.P. Kent, Level Set Trees for Applied Statistics PhD thesis, Carnegie Mellon University, 2013.
[11] K. Turner, Means and medians of sets of persistence diagrams. (2013), arXiv:1307.8300v1.
[12] B. Fasy, F. Lecci, A. Rinaldo, Wasserman Larry, Balakrishnan Sivaraman, A. Singh, Confidence sets for persistence diagrams, Ann. Stat. 42 (6) (2014) 2301–2339.
[13] F. Chazal, B. Fasy, F. Lecci, A. Rinaldo, A. Singh, L. Wasserman, On the bootstrap for persistence diagrams and landscapes, (2014), arXiv:1311.0376v2.
[14] H. Edelsbrunner, J. Harer, Computational Topology: An Introduction, Amer Mathematical Society, 2010.
[15] P. Bubenik, Statistical topology using persistence landscapes, (2012) Available at arxiv.org/abs/1207.6437v1.
[16] A. Robinson, K. Turner, Hypothesis testing for topological data analysis, (2013), arXiv:1310.7467.
[17] G. Máté, A. Hofmann, N. Wenzel, W. Heermann, A topological similarity measure for proteins, Biochimicaet Biophysica Acta (2014) 1180–1190.
[18] P. Bubenik, Statistical topological data analysis using persistence landscapes, J. Mach. Learn. Res. 16 (2015) 77–102.
[19] V. Kovacev-Nikolic, Persistent Homology in Analysis of Point-Cloud Data Msc thesis, Alberta University, 2012.
[20] S. Balchin, E. Pillin. Comparing metrics on arbitrary spaces using topological data analysis,(2015), arXiv:1503.04619v1.
[21] S. Lele, T. Richtsmeier, Euclidean distance matrix analysis: a coordinate free approach for comparing biological shapes using landmark data, Am. J. Phys. Anthropol. 86 (1991) 415–427.
[22] C. Brombin, L. Salmaso, Permutation Tests in Shape Analysis, Springer Science+Business Media, New York, 2013.
[23] B. Scloerke, Zoo of geometric objects, 2015. http://cran.r-project.org/web/packages/geozoo/index.html.
[24] B. Fasy, J. Kim, F. Lecci, C. Maria, Introduction to the R package TDA, 2015. http://cran.r-project.org/web/packages/TDA/index.html.
[25] P. Martínez-Camblor, J. Uña-Álvarez, N. Corral, k-Sample test based on the common area of kernel density estimator, J. Stat. Plan. Inference 138 (12) (2008) 4006–4020.
[26] P. Cortez, A. Silva, Using data mining to predict secondary school student performance, in: Proceedings of 5th Future Business, Technology Conference, 2008, pp. 5–12.
[27] P. Álvarez-Esteban, E. del Barrio, A. Cuesta-Albertos, C. Matrán, Similarity of samples and trimming, Bernoulli, 18 (2012) 606–634.
[28] I. Dryden, K. Mardia, Statistical Shape Analysis, John Wiley and Sons, 1998.
[29] H. Edelsbrunner, D. Letscher, A. Zomorodian, Topological persistence and simplification, Discrete Comput. Geo. 28 (2002) 511.