SEMARAK ILMU PUBLISHING

Journal of Advanced Research in Applied Sciences And Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Clustering Datasaurus Dozen Using Bottleneck Distance, Wasserstein Distance (WD) and Persistence Landscapes

R.U. Gobithaasan[1,2,*], Kirthana Devi Selvarajh[1], Kenjiro T. Miura[3]

[1] School of Mathematical Sciences, Universiti Sains Malaysia, 11800, Penang, Malaysia
[2] Special Interest Group of Modelling & Data Analytics, Faculty of Ocean Engineering Technology and Informatics, University Malaysia Terengganu, 21030 Kuala Nerus, Malaysia
[3] Graduate School of Engineering, Shizuoka University, Hamamatsu, 432 8018 Japan

ARTICLE INFO

Article history:
Received 8 May 2023
Received in revised form 27 August 2023
Accepted 12 November 2023
Available online 24 January 2024

Keywords:
Datasaurus Dozen; Persistent Homology; Persistence Diagram; Agglomerative Hierarchical Clustering

ABSTRACT

Topological Data Analysis (TDA) is an emerging field of study that helps to obtain insights from the topological information of datasets. Motivated by the emergence of TDA, we applied Persistent Homology (PH), one of the tools commonly used to extract topological features to cluster the Datasaurus Dozen dataset. This dataset is ideal to show PH's capability in clustering as it consists of twelve distinct point clouds (PC) that have identical mean values, standard deviation, and correlation values, yet produce dissimilar patterns. The methodology starts with normalizing Datasaurus Dozen, followed by computing H_1 Persistence Diagrams (PD) for each dataset. Two types of PD distances are computed directly: Wasserstein Distance (WD) and Bottleneck Distance (BD) and represented as proximity matrix. We also vectorized H_1 Persistence Diagrams to obtain the average of first five strips of Persistence Landscape (PL) and computed L_2 distance to represent a proximity matrix. These three distance matrices are used to generate dendrograms by using Hierarchical Agglomerative Clustering (HAC). Regardless of possessing similar descriptive statistics, PH accurately extracts the global and local geometric topological information, and clusters them accordingly. It is evident that for clustering based on global geometric information, BD is suitable and computably cheap, whereas for clustering based on local geometric information, WD and average PL vectors are suitable but may incur extra computation.

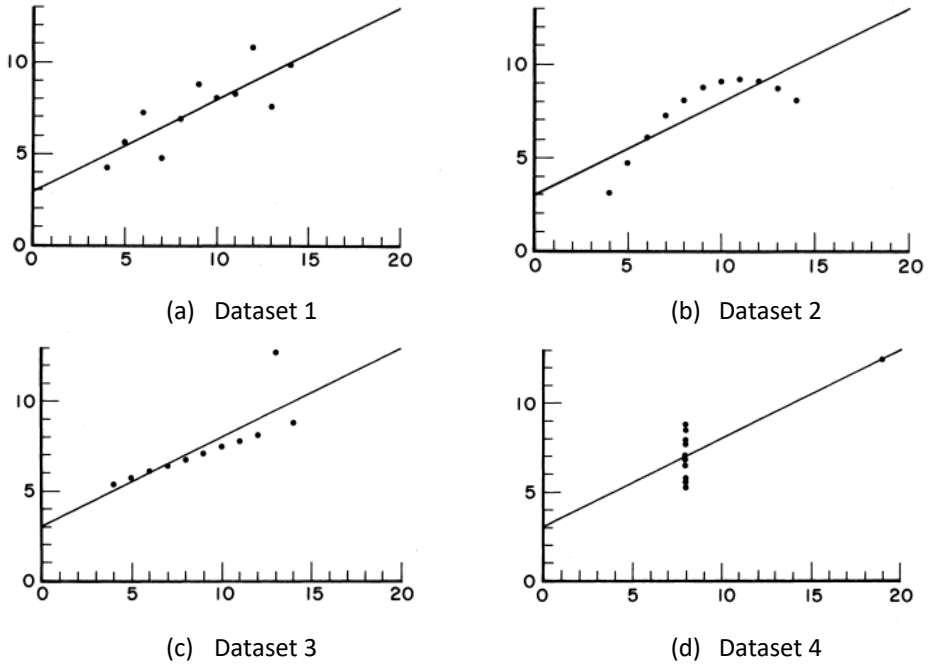
1. Introduction

In 1973, statistician Francis Anscombe created Anscombe's quartet to demonstrate the importance of data visualization before analyzing it and building a model [1]. Anscombe's quartet consists of four dataset groups which all have nearly identical statistical observations that provides the same information (variance and mean) for each x and y point. However, when these datasets are plotted, they appear very different from one another. The datasets are plotted as scatter plot with the fitted line which shows the difference between the four datasets as shown in Figure 1.

* *Corresponding author.*
E-mail address: gr@umt.edu.my

https://doi.org/10.37934/araset.38.1.1224

From Figure 1, we can deduce that (a) Dataset 1 shows that the dataset fits the linear regression model well and (b) Dataset 2 cannot fit the linear regression model but rather forms a smooth curve relation as it is a non-linear dataset. As for Figure 1(c), the datasets form in a straight line except for one observation further away from the fitted line. This shows that a single outlier is involved in the dataset. Similarly, Figure 1(d) shows the presence of a set of outliers in the dataset.



(a) Dataset 1 (b) Dataset 2

(c) Dataset 3 (d) Dataset 4

Fig. 1. Visualization of Anscombe's Quartet of four datasets in the form of scatter plot [2].

Although we will be able to visualize the dataset patterns, the failure in differentiating each dataset in terms of statistical summary leads us to seek for efficient data analysis method. In this study, we will be using Topological Data Analysis (TDA) as a method to differentiate datasets that produce similar statistical summary and show its strength by means of clustering.

Topological Data Analysis (TDA) is an emerging data analysis method used in studying topological space or features. In this work, it is used to cluster the characteristics of the data. With the help of TDA, we would be able to obtain the hidden insights of the dataset feature from its topological information. TDA is also applied in many other fields such as image analysis, time series [2], medicine, sensor networks, chemistry [3] etc. Driven by TDA's performance, we employed Persistent Homology (PH) to obtain Persistence Diagrams (PD) and represented distance matrix using Bottleneck distance and Wasserstein distance respectively. The third approach is from the PDs, we vectorized its topological features using average of Persistence Landscape (PL) for each dataset, thus, by comparing these feature vectors with L_2 norm, we obtained a distance matrix. With these three types of distance matrices, we then applied cluster analysis.

Cluster Analysis or commonly known as clustering is one of an unsupervised learning method used in machine learning. With clustering, observations are separated into groups with similar characteristics and assigned into clusters. These observations are segregated based on their similarity/dissimilarity measures to help uncover the common characteristics of each cluster [4]. Cluster analysis is performed in this study to figure out the dissimilarities of the datasets using topological information. In this work, we employed Hierarchical Agglomerative Clustering (HAC), a well-known clustering algorithm. This method is used to represent clusters as well as the tree-like structure called dendrogram illustrating the flow of how observations are clustered. Four types of

clustering linkages are considered: single, complete, average and ward. After clustering the datasets, we compared the dendrograms to summarize the efficiency of PH methodology to qualitatively differentiate the dataset.

The objective of this study is twofold: (i) to extract the topological features from Persistence Diagram (PD) and obtain three types of distance matrices using Wasserstein Distance (WD), Bottleneck Distance (WD) and Persistence Landscape (PL), (ii) to compare the dendrograms obtained from BD, WD, and PL Average HAC approach. The comparison of these three approaches would be helpful in identifying the dissimilarity and suitability for clustering the dataset with similar statistical summary.

The structure of the paper is organized as follows: The following section introduces the dataset and the methods implemented which includes HAC and PH to differentiate the datasets. The subsequent section depicts findings and discussions, and the final section is the study's conclusion and future work.

2. Methodology

2.1 Datasaurus Dozen

Inspired by Anscombe's Quartet data, Matejka and Fitzmaurice [5] created additional twelve datasets which is named as Datasaurus Dozen. These datasets including Datasaurus dataset have the same statistical summary (arithmetic means, standard deviations, and correlation coefficient, all to two decimal places) as Datasaurus data but it has different appearances visually. The 12 datasets consist of horizontal, vertical, and diagonal parallel lines; fuzzier horizontal and vertical swaths; a grid and a blob of points; a big "X"; a five-pointed star, single and double circles as shown in Figure 2.

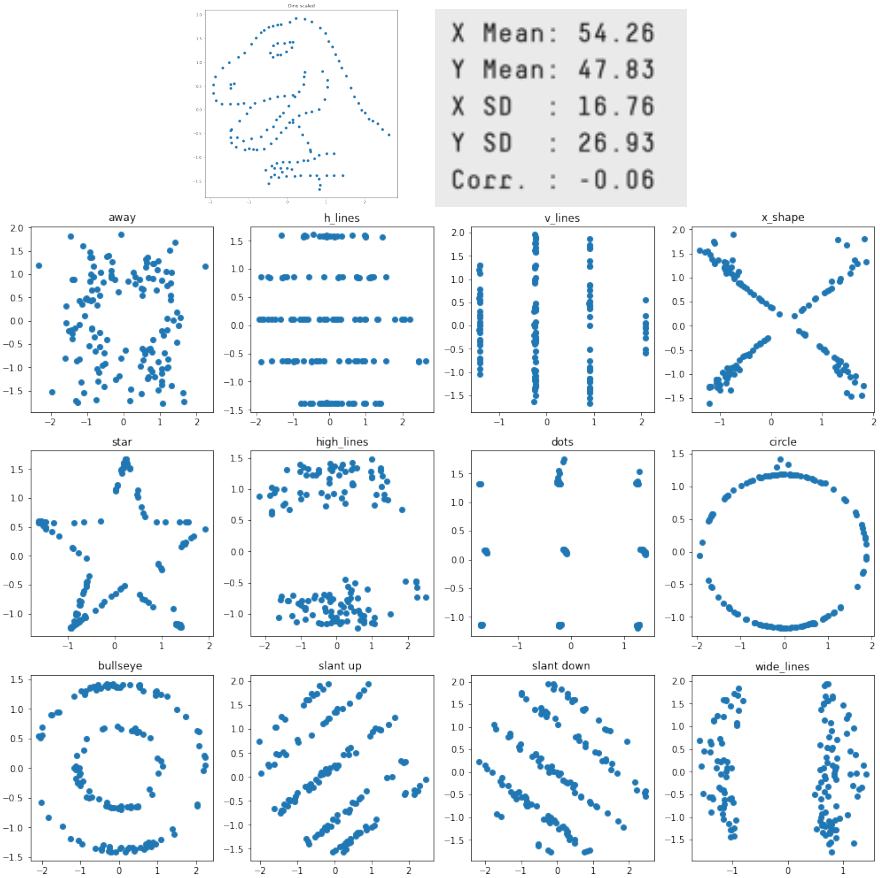


Fig. 2. The Datasaurus Dozen with different appearance but same statistical summary [5]

2.2 Persistent Homology (PH)

Topological Data Analysis (TDA) is a new method for analysing datasets which is rooted from the mathematical study of shapes and structure under deformation and stretching [6]. Most datasets are now known to be high-dimensional, incomplete, and noisy. TDA offers a comprehensive framework for evaluating such datasets in a way that is insensitive to specific metrics and allows for dimensional reduction as well as noise resistance of those features. TDA investigates and comprehends point cloud data using two methodologies: Persistent Homology (PH) and TDA Mapper. PH is a popular TDA tool that has been used successfully in a variety of fields, including financial time series analysis [7], haze clustering [8], and dynamical state analysis [9]. TDA Mapper represents the point cloud into a graph which consists of vertices and edges, hence summarizing the structure of a data. In this work, we employed the Gudhi Python package to generate PD [10] using Vietoris-Rips filtration technique with coefficient field $\mathbb{Z}_3$. $\mathbb{Z}_3$ provides the information about homology computation algebraically, in which chosen based on the dimension of the point cloud. For instance, $\mathbb{Z}_2$ has two elements (0 and 1), whereas $\mathbb{Z}_3$ has three elements (0, 1 and 2). Figure 3 shows the summaries of PH framework applied in this work.

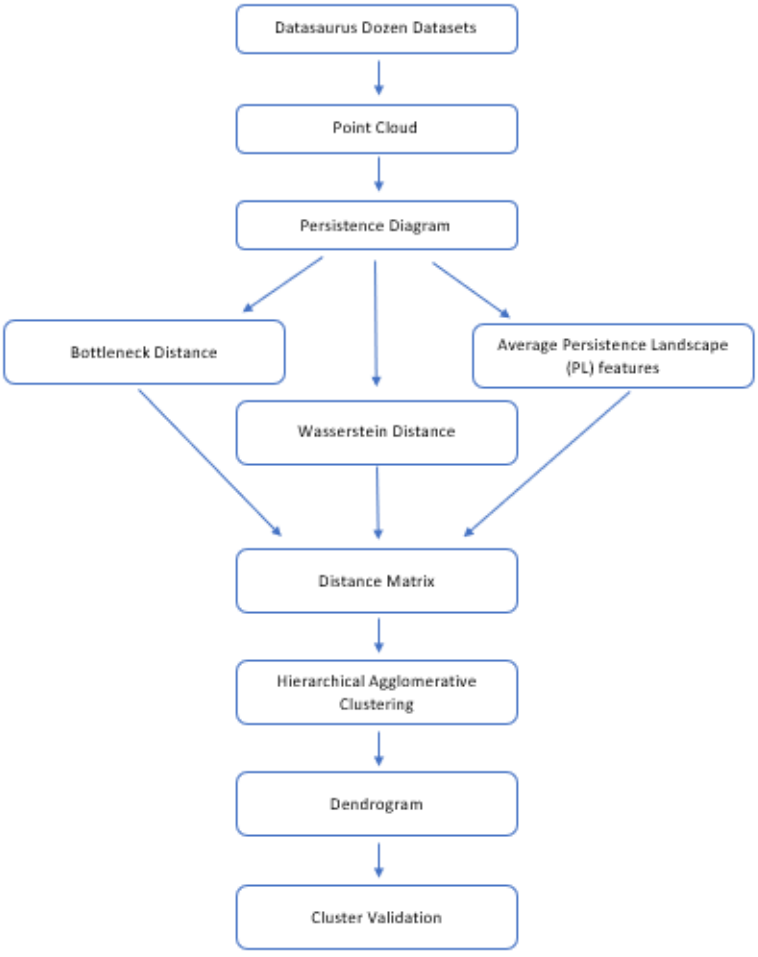


Fig. 3. General flowchart of PH clustering performed in this study

2.2.1 Persistence Diagram (PD)

In TDA, generalization of a graph concept to higher dimension is known to be as a simplicial complex, K. It is a collection of simplices $\sigma \subseteq V$, set of vertices. Filtration is a collection of simplicial complexes which is defined as

$$K_1 \subseteq K_2 \subseteq \cdots \subseteq K_N \tag{1}$$

Homology counts the number of structures such as connected components (H_0), loops (H_1), voids (H_2) and higher dimensions in a simplicial complex. In a simplicial complex, K, the *d*-dimensional simplices are denoted as $\sigma_1, \ldots, \sigma_l$. A *d*-dimensional chain is a formal sum of *d*-dimensional simplices $\alpha = \sum_{i=1}^{l} a_i \sigma_i$, where a_i is the coefficient field. The collection of all *d*-dimensional chains forms a vector space denoted as $C_d(K)$. $C_d(K)$ is the d^{th} chain group of a simplicial complex K made from all *d*-chains from simplicial complex K together with an addition operation. The boundary operator ∂_d: $C_d(K) \rightarrow C_{d-1}(K)$ is given by

$$\partial_d(\alpha) = \sum_{i=1}^{l} a_i \partial_d(\sigma_i), \tag{2}$$

where the boundary is defined as $\partial_d(\sigma) = \sum_{\tau<\sigma, \dim(\tau)=d-1} \tau$. If $\partial_d(\alpha) = 0$, a d-chain $\alpha \in C_d(K)$ is a cycle. It has a boundary when there is a (*d*+1)-chain β such that $\partial_{d+1}(\beta) = \alpha$. The group of *d*-dimensional cycles is denoted $Z_d(K)$ and the boundaries are $B_d(K)$. The *d*-dimensional homology group is defined as

$$H_d(K) = Z_d(K)/B_d(K). \tag{3}$$

An element of $H_d(K)$ is called a homology class. For instance, zero-dimensional homology is defined as $H_0(K)$ for each connected component of K [11]. In PH algorithm, we keep track of how the homology changes as filtration takes place. We start off with point cloud (PC) dataset which represents the collection of data points in any dimension. The data points in each PC represent vertices which will relate to edges continuously when a cover (usually a ball cover) for each data point intersects with other ball covers creating a simplicial complex K at various resolution of the ball radius ϵ. We call this process as the filtration of PC. There are various filtrations available to represent a PC in the form of a complex, in this work we employed Vietoris-Rips filtration which is computationally cheap as compared to other types of filtrations. For the given filtration in equation (1), we have a sequence of homology maps

$$H_d(K_1) \rightarrow H_d(K_2) \rightarrow \cdots \rightarrow H_d(K_N). \tag{4}$$

As we filter over cover with radius $\epsilon \in \mathbb{R}$, *n*-dimensional homology $H_n(K_\epsilon)$ classes are born at time b_i and die at d_i where $b_i \leq d_i$. Thus, $H_n(K_N)$ can be denoted in the form of multiset intervals $\{(b_i, d_i)\}_{i=1}^{n}$ which can be represented in the form of n^{th} persistence barcodes or a point in Persistence Diagram (PD). For the barcode, each bar corresponds to an interval $\{b_i, d_i\}$ where b_i represents birth and d_i denotes the deaths, which is the endpoint of the bar. In an n-dimensional space, a point cloud can have a collection of *n* barcodes, one for each dimension.

Similarly, PD is an equivalent representation of barcode in which bars are now in the form of points denoted as tuples $x_i = (b_i, d_i)$ where $i = 1, 2, \ldots, p$ of points is a spread in the upper half

plane above the diagonal line. The set of all PDs can be represented with multiple distance measures where it is a rigorous metric space under the condition of local finiteness of persistence diagrams.

Two distance measures employed to find the distances between PDs are Wasserstein distance (WD) and Bottleneck distance. WD works by summing p^{th} powers of the distance to move each point and is more sensitive to noise. BD only takes the farthest distance any point needs to be moved and only sees global structure.

Given two PDs; PD_1 and PD_2, the p^{th} Wasserstein distance is defined as follows [12]:

$$WD_p(PD_1, PD_2) = \inf_{\varphi: X \to Y} \left(\sum_{a \in X} \|a - \varphi(a)\|_q^p\right)^{1/p} \tag{5}$$

According to Berwald *et al*., [12], in majority application we let the $q = \infty$. Using $q = p$ to control the geometry of the space of persistence diagrams is generally more sensible. In a simpler form, the Bottleneck distance is defined as follows [13]:

$$BD_\infty(PD_1, PD_2) = \inf_{\varphi: X \to Y} \sup_{a \in X} \|a - \varphi(a)\|_\infty. \tag{6}$$

where φ is a multi-bijective matching point between X and Y.

2.2.2 Persistence Landscape (PL)

A Persistence Landscape (PL) is created by first building a triangle that corresponds to a generalized persistence interval of the birth and death pair from PD. We denote the birth and death pair for a specific H_n in the piecewise linear function $P_{(b_i,d_i)}: \mathbb{R} \to [0, \infty]$.

$$P_{(b_i,d_i)}(t) = \begin{cases} t - b_i, \ if\ t \in (b_i, \frac{b_i+d_i}{2}), \\ -t + b_i, \ if\ t \in (\frac{b_i+d_i}{2}, d_i), \\ 0, \ if\ t \notin (b_i, d_i). \end{cases} \tag{7}$$

The birth and death pairs $\{(b_i, d_i)\}_{i=1}^n$ in PL is the order of functions $\lambda_k: \mathbb{R} \to [0, \infty]$, $k = 1, 2, 3, \ldots$ or equivalent function $\lambda_k: \mathbb{N} \times \mathbb{R} \to [0, \infty]$ where $\lambda(k, t) = \lambda_k$ given as follows:

$$\lambda_k(t) = k - \max\left\{P_{(b_i,d_i)}(t) \middle| (b_i, d_i) \in PD\right\}. \tag{8}$$

For a detailed explanation on PL, readers are referred to Bubenik and Dlotko [14].

Next, we extract the first five strips of persistence landscapes $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\}$ of H_1 and obtain the average of each strip. In general, we find the average of a strip of persistence landscapes, $\bar{\lambda}_k$ as follows:

$$\bar{\lambda}_k(t) = \frac{1}{N} \sum_{i=1}^N \lambda_k^{(i)}(t). \tag{9}$$

Let f_j represent j[th] set of Datasaurus Dozen dataset, then each dataset has the following H_1 topological feature vector:

$$f_q = (\bar{\lambda}_1, \bar{\lambda}_2, \bar{\lambda}_3, \bar{\lambda}_4, \bar{\lambda}_5), \tag{10}$$

for each dataset, thus we have $q = 1, 2, \ldots, 12$ feature vectors. To compute dissimilarities between each dataset, we used the following norm equations with $p = 2$.

$$\|f_i - f_{i+1}\|_p = \left[\sum_{i=1}^{N} |f_i - f_{i+1}|^p\right]^{1/p} \tag{11}$$

2.3 Cluster Analysis

Cluster analysis technique is used mainly to explore datasets and cluster them into different cluster based on certain aspects. There are many clustering methods namely K-Means, Mean Shift, Spectral clustering, Hierarchical clustering and DBSCAN. Throughout this study, we applied Hierarchical Clustering specifically the Hierarchical Agglomerative Cluster (HAC). HAC is applied to Datasaurus Dozen using three PH approaches to experiment the feasibility of their respective topological vectors to differentiate each dataset.

2.3.1 Hierarchical Agglomerative Clustering (HAC)

Hierarchical agglomerative clustering (HAC) is a bottom-up approach that generates dendrogram partitions by progressively merging the *n* individuals into groups [15]. In short, the observations are separated into groups with similar characteristics and assign them into clusters. Dendrograms are tree-like structures that are frequently used to depict the relationships between all the data points in a dataset [16]. For cluster observations to create a dendrogram, a measure of dissimilarity and a linkage criterion is necessary. The linkage criterion is used to calculate the distance between observation sets. A dendrogram can be produced by applying linkage techniques such as single linkage, complete linkage, group average and Ward to observations.

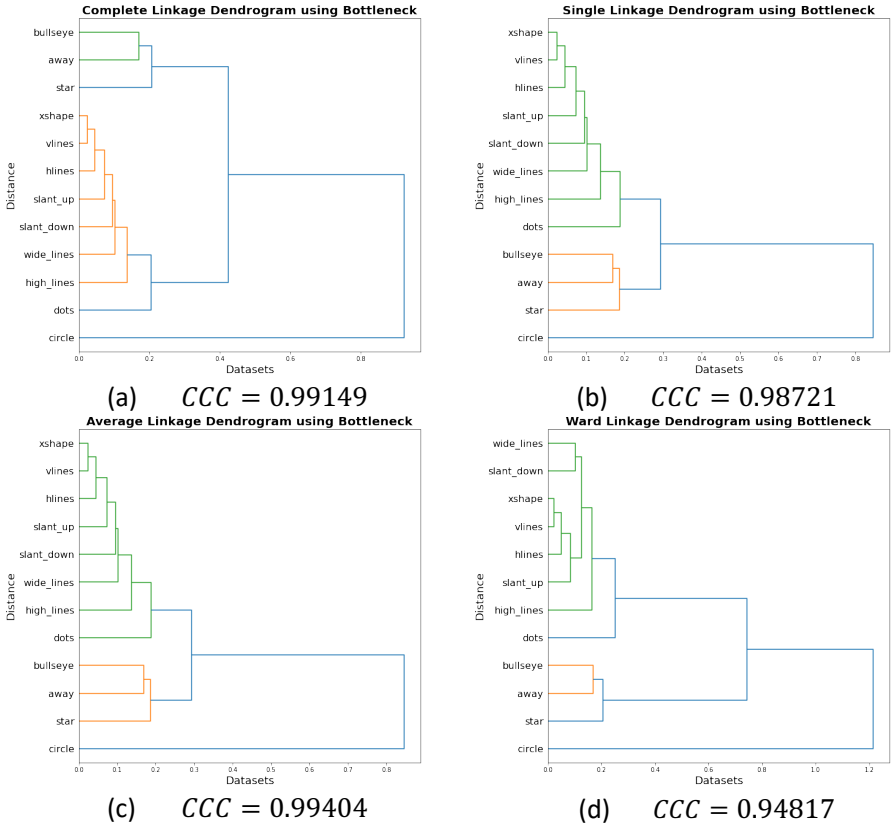
The distance matrix also called as dissimilarity matrix used in this work is Bottleneck distance (BD) of PD, Wasserstein distance matrix (WD) of PD and PL Average (PL) distance as inputs along with its linkage to construct the dendrogram. Four linkages (single, average, ward and complete) were used in this work. A single linkage is the shortest distance between a point in one cluster and a point in another. Complete linkage is the distance between clusters with the most distant observations. Average linkage is defined as the average distance between each cluster point to every point in the other cluster [17]. Ward linkage employs the variance for clusters instead of measuring the distance directly, hence it is said to suit quantitative variables.

Cluster validation is an important step in cluster analysis. It aids in determining the quality of clustering algorithm results. In this work, we employed Cophenetic Correlation Coefficient (CCC) as cluster validation. It is used to assess the dendrogram and the accuracy of the dissimilarity matrix generated based on Pearson coefficient (r_{cof}) where it is calculated between original dissimilarity and the cophenetic distance of a dendrogram. If the CCC is close to one, the dendrogram produced precisely represents the dissimilarity matrix [18].

3. Results

3.1 Clustering Based on Bottleneck Distance (BD-HAC)

In this part, we computed the distance between the PD of each dataset with one another using the Bottleneck Distance (BD). Each distance was compiled and used as a new input to find the dissimilarity matrix to generate the dendrogram. Along with the dissimilarity matrix obtained, we employed four different types of linkage to obtain four different dendrograms (Figure 4).

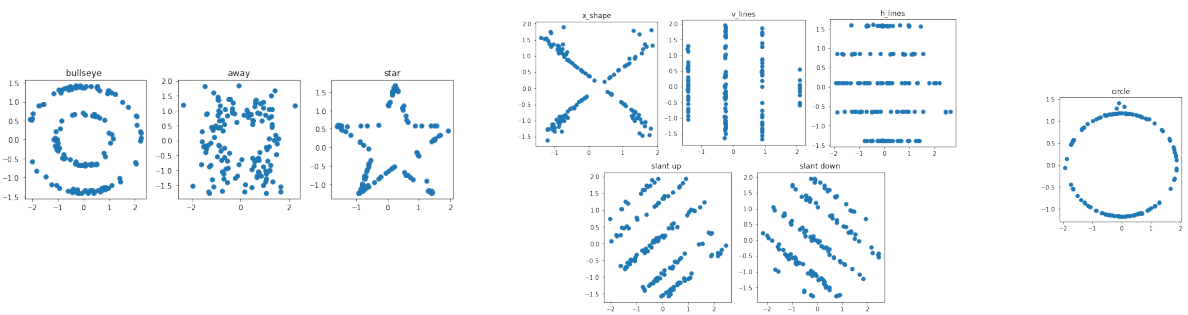


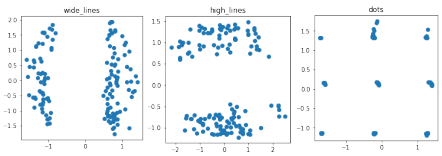
(a) $CCC = 0.99149$ (b) $CCC = 0.98721$

(c) $CCC = 0.99404$ (d) $CCC = 0.94817$

Fig. 4. Four different Bottleneck Distance dendrogram based on different linkages

From these four dendrograms (Figure 4), we chose the complete linkage dendrogram (Figure 4(a)) as an example to show the cluster formation and dissimilarity of each cluster. Complete linkage dendrogram has one of the highest CCC value and looks well clustered compared to the other 3 dendrograms. The BD distance measure used allows the clustering of the Datasaurus Dozen into three distinct clusters. The first cluster (Figure 5(a)) consists of three datasets with 1D loop-like structure. The second cluster (Figure 5(b)) has datasets with no loop whereas the last cluster (Figure 5(c)) has one stand-alone circle dataset which has one distinct loop.



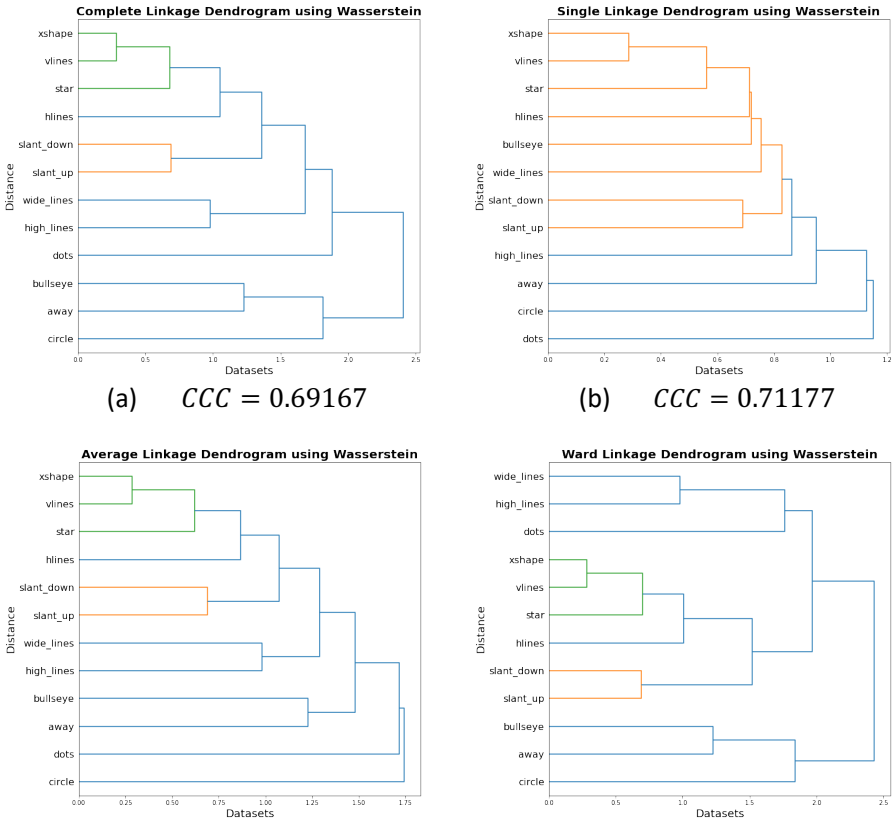


(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

Fig. 5. Three distinct clusters based on Complete Linkage dendrogram of BD of PD input

3.2 Clustering Based on Wasserstein Distance (WD-HAC)

Similar to BD-HAC, we computed the distance between the PD of each dataset using the Wasserstein Distance (WD). Each distance was compiled into a dissimilarity matrix to generate the dendrogram. Next, we employed four different types of linkages to obtain four different dendrograms as shown in Figure 6.

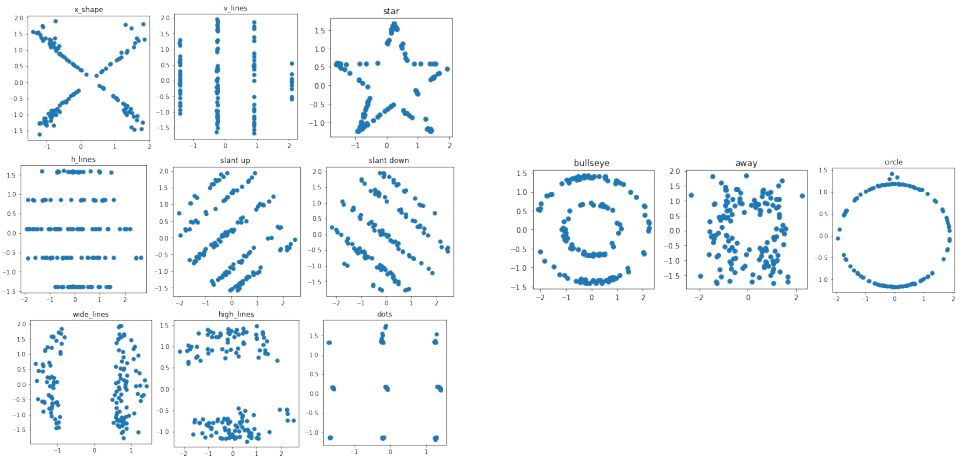


(a) $CCC = 0.69167$ (b) $CCC = 0.71177$

(c) $CCC = 0.75949$ (d) $CCC = 0.67754$

Fig. 6. Four different Wasserstein Distance dendrogram based on different linkages

We chose the complete linkage dendrogram (Figure 6(a)) as an example to show the cluster formation and dissimilarity of each cluster. Although both Complete linkage and Ward dendrogram has the lowest CCC value and looks well clustered, complete linkage has the higher CCC compared to Ward. We have two distinct clusters; the first cluster (Figure 7(a)) consists of all datasets without loop except for star shape which has an edgy shape whereas the second cluster (Figure 7(b)) consists of datasets with both smooth curvy shape loop.

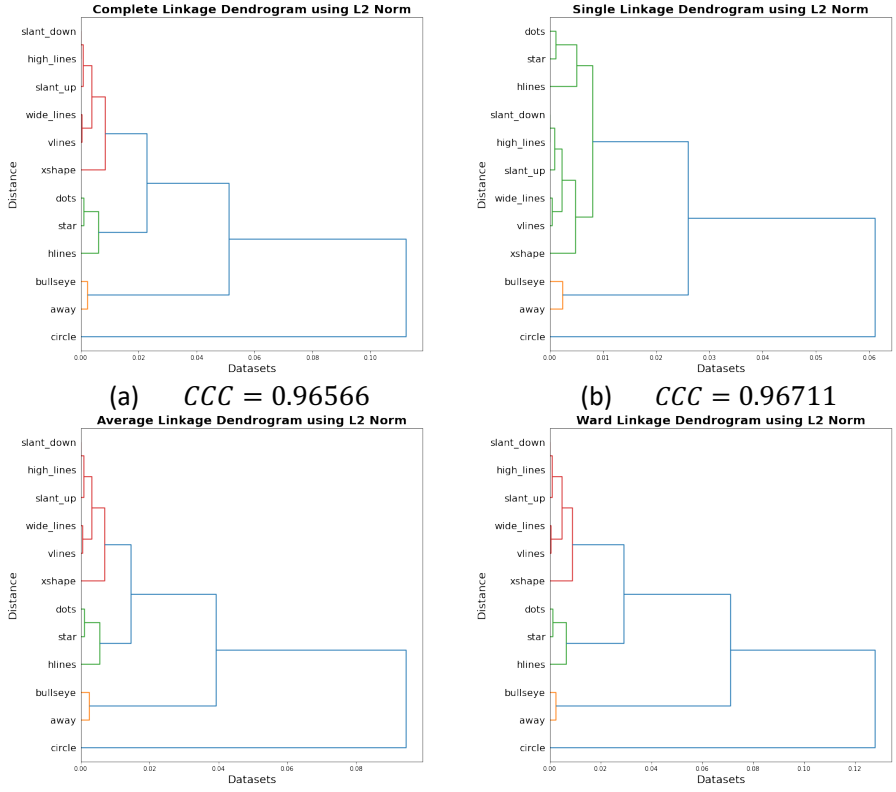


(a) Cluster 1 (b) Cluster 2

Fig. 7. Two distinct clusters based on Complete Linkage dendrogram of BD of PD input

3.3 Clustering Based on PL Average (PL-HAC)

In this section, we computed the average of PL of each dataset and used this topological vector as an input to find the dissimilarity matrix to generate the dendrogram with linkages as stated above. All the dendrogram (Figure 8) looks similar with same group of clusters.



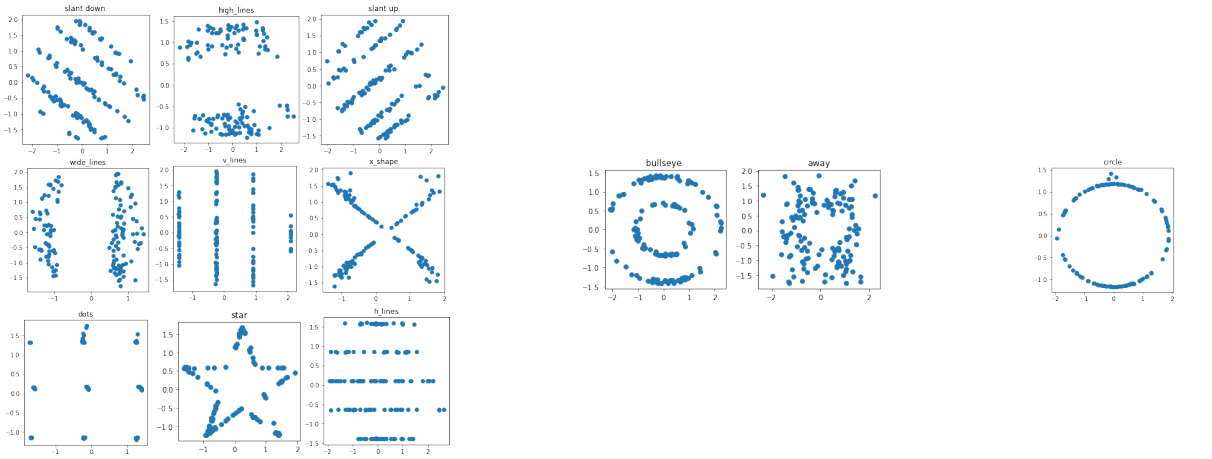
(a) $CCC = 0.96566$ (b) $CCC = 0.96711$

(c) $CCC = 0.96787$ (d) $CCC = 0.95284$

Fig. 8. Four different Average Persistence Landscape dendrogram based on different linkages

We chose complete linkage dendrogram (Figure 8(a)) as an example to show the cluster formation and dissimilarity of each cluster although all the dendrogram gives the same result. The first cluster (Figure 9(a)) shows that the cluster consists of edgy shaped datasets. The second cluster

(Figure 8(b)) consists of dataset shape that has rough 1D loop whereas the third cluster (Figure 8(c)) is a stand-alone cluster just like in BD-HAC which is a circle.



(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

Fig. 9. Three distinct clusters based on Complete Linkage dendrogram of Average PL input

3.4 Comparisons of BD-HAC, WD-HAC, and PL-HAC

Cluster comparisons serve as the foundation of clustering evaluation and temporal evaluation of clusters. Due to grouping of elements into clusters, similarity measures must be considered for other aspects of clustering such as number of clusters, size distribution of clusters, cluster overlaps and scaling relations between levels of hierarchical clustering. Therefore, we used element-centric similarity which gives additional specific insights of how two clustering are different based on the similarity calculated at the level of individual variable in a dataset. Element-centric similarity technique captures cluster-induced relationships between the observations through cluster affiliation graph, a bipartite network in which one vertex set corresponds to the original elements and the other to the clusters. In short, it is a per-observation measure, and is obtained by constructing cluster induced element graph.

The element-centric similarity score lies between 0 and 1 where 0 means clusters do not share a single element and 1 indicates the clustering is identical [19]. In this study, we compared three dendrograms obtained from WD-HAC, BD-HAC and PL-HAC approach using element-centric similarity. BD-HAC approach has two number of clusters whereas WD-HAC and PL-HAC has three number of clusters. Based on the element-centric similarity value (Table 1), the dendrogram obtained using BD is 87% similar to the dendrogram obtained using WD. It is not surprising as these two approaches employ distance between PD directly rather than extracting the topological vector as represented by average PL. It is also apparent that all three has about 82% similarities, hence indicating the stability of PDs as discussed by Otter *et al*., [6] and Bubenik *et al*., [14]. Comparing PDs obtained directly is simpler and straightforward as shown in equation (5) and (6), using either WD or BD. However, obtaining PL requires extra computation work as shown in equation (7) to equation (11). As we compare the three approaches, BD approach only considers the most apparent topological feature (furthest PD from diagonal line), whereas WD and average PL considers all topological feature; whether its far or near the diagonal line. Hence, WD and average PL clusters are based on geometric information, namely the curvature of the dataset. This finding is evident as we inspect how the star-shape dataset is clustered. For WD and average PL based clusters, the star-shaped dataset is clustered far away from those datasets having curvy edges. Hence, this is yet

another proof where the points closer to diagonal line in PD is indeed capturing local geometric information as reported by Bubenik *et al*., [20].

Table 1
Element-Centric Similarity value compared between complete linkage for each HAC approach

	BD-HAC	WD-HAC	PL-HAC
BD-HAC			
WD-HAC	87.18%		
PL-HAC	82.55%	84.03%	

4. Conclusions

With the aid of Datasaurus Dozen dataset, we have shown how clustering Persistent Homology successfully cluster datasets with the same statistical summary. Since the dataset is 2D, we can easily visualize and compare unlike those in high dimension. Overall characteristics analysis of clusters formation in three different HAC approach shows that BD-HAC clusters are based on global geometric information, whereas WD-HAC and PL-HAC are based on local geometric information. Finally, we compared all three HAC approaches by using element-centric similarity (ECS), which showed that the similarity range of all approach is within the range of 82-87%. These show that the way each dataset cluster in different approach is almost similar and stable. BD-HAC approach can be used if the data used is high-dimensional since its computationally cheaper compared to WD-HAC and PL-HAC. As for PL-HAC, it can be used for clustering based on local geometric information, although it may involve extra computation.

Acknowledgement
This research was funded by a grant from Ministry of Higher Education of Malaysia (FRGS Grant 2019/STG06/UMT/02/2). We thank the reviewers who helped to improve the presentation of this paper.

References
[1] Anscombe, Francis J. "Graphs in statistical analysis." *The american statistician* 27, no. 1 (1973): 17-21. https://doi.org/10.2307/2682899
[2] Gobithaasan, R. U., Zabidi Abu Hasan, Krithana Devi Selvarajh, Khai-Sam Wong, Shukri Mamat, Mohd Zaharifudin Muhamad Ali, Kenjiro T. Miura, and Pawe Dotko. "Clustering Selected Terengganu's Rainfall Stations Based on Persistent Homology." *Thai Journal of Mathematics* (2022): 197-211. http://thaijmath.in.cmu.ac.th/index.php/thaijmath/article/view/6120/354355147
[3] Chazal, Frédéric, and Bertrand Michel. "An introduction to topological data analysis: fundamental and practical aspects for data scientists." *Frontiers in artificial intelligence* 4 (2021): 108. https://doi.org/10.3389/frai.2021.667963
[4] Lee, Seulbi, Jaehoon Kim, Jongyeon Hwang, EunJi Lee, Kyoung-Jin Lee, Jeongkyu Oh, Jungsu Park, and Tae-Young Heo. "Clustering of time series water quality data using dynamic time warping: A case study from the Bukhan River water quality monitoring network." *Water* 12, no. 9 (2020): 2411. https://doi.org/10.3390/w12092411
[5] Matejka, Justin, and George Fitzmaurice. "Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing." In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pp. 1290-1294. 2017. https://doi.org/10.1145/3025453.3025912
[6] Otter, Nina, Mason A. Porter, Ulrike Tillmann, Peter Grindrod, and Heather A. Harrington. "A roadmap for the computation of persistent homology." *EPJ Data Science* 6 (2017): 1-38. https://doi.org/10.1140/epjds/s13688-017-0109-5
[7] Gidea, Marian, and Yuri Katz. "Topological data analysis of financial time series: Landscapes of crashes." *Physica A: Statistical Mechanics and its Applications* 491 (2018): 820-834. https://doi.org/10.1016/j.physa.2017.09.028

[8] Zulkepli, Nur Fariha Syaqina, Mohd Salmi Md Noorani, Fatimah Abdul Razak, Munira Ismail, and Mohd Almie Alias. "Hybridization of hierarchical clustering with persistent homology in assessing haze episodes between air quality monitoring stations." *Journal of environmental management* 306 (2022): 114434. https://doi.or/10.1016/j.jenvman.2022.114434

[9] Myers, Audun, Elizabeth Munch, and Firas A. Khasawneh. "Persistent homology of complex networks for dynamic state detection." *Physical Review E* 100, no. 2 (2019): 022314. https://doi.org/10.1103/physreve.100.022314

[10] *Gudhi Library*. GUDHI library. (n.d.). Retrieved January 15, 2023, from https://gudhi.inria.fr/

[11] Myers, Audun, Elizabeth Munch, and Firas A. Khasawneh. "Persistent homology of complex networks for dynamic state detection." *Physical Review E* 100, no. 2 (2019): 022314. https://doi.org/10.1103/PhysRevE.100.022314

[12] Berwald, Jesse J., Joel M. Gottlieb, and Elizabeth Munch. "Computing Wasserstein distance for persistence diagrams on a quantum computer." *arXiv preprint arXiv:1809.06433* (2018).

[13] Cao, Yueqi, Anthea Monod, Athanasios Vlontzos, Luca Schmidtke, and Bernhard Kainz. "Topological information retrieval with dilation-invariant bottleneck comparative measures." *Information and Inference: A Journal of the IMA* 12, no. 3 (2023): iaad022. https://doi.org/10.1093/imaiai/iaad022

[14] Bubenik, Peter, and Paweł Dłotko. "A persistence landscapes toolbox for topological statistics." *Journal of Symbolic Computation* 78 (2017): 91-114. https://doi.org/10.1016/j.jsc.2016.03.009

[15] Everitt, B. S.; Dunn, G. Cluster Analysis. In *Applied Multivariate Data Analysis*, 2nd ed.; John Wiley & Sons Ltd, United Kingdom, 2001, pp. 125-158. https://doi.org/10.1002/9781118887486.ch6

[16] Chen, Jin, Alan M. MacEachren, and Donna J. Peuquet. "Constructing overview+ detail dendrogram-matrix views." *IEEE transactions on visualization and computer graphics* 15, no. 6 (2009): 889-896. https://doi.org/10.1109/TVCG.2009.130

[17] Everitt, B. S., Landau, S., Leese, M., and Stahl, D. "Hierarchical Clustering". In *Cluster Analysis*, 5th ed.; Jon Wiley & Sons Ltd, United Kingdom, 2011, pp. 71-110. https://doi.org/10.1002/9780470977811

[18] Clarke, K. Robert, Paul J. Somerfield, and Raymond N. Gorley. "Clustering in non-parametric multivariate analyses." *Journal of Experimental Marine Biology and Ecology* 483 (2016): 147-155. https://doi.org/10.1016/j.jembe.2016.07.010

[19] Gates, Alexander J., Ian B. Wood, William P. Hetrick, and Yong-Yeol Ahn. "Element-centric clustering comparison unifies overlaps and hierarchy." *Scientific reports* 9, no. 1 (2019): 8574. https://doi.org/10.1038/s41598-019-44892-y

[20] Bubenik, Peter, Michael Hull, Dhruv Patel, and Benjamin Whittle. "Persistent homology detects curvature." *Inverse Problems* 36, no. 2 (2020): 025008. https://doi.org/10.1088/1361-6420/ab4ac0