

Testing Report on Clustering Algorithms

Lieyu Shi

May 22, 2018

1 Introduction

In this report, we are testing effectiveness and validity of clustering algorithms from [[github implementation](#)], which is my C++ implementation for all the prevailing clustering methods in flow visualization. Respectively, it is constituted by, **k-means**, **k-medoids**, **BIRCH**, **agglomerative hierarchical clustering** (AHC), **DBSCAN**, **OPTICS**, **spectral clustering** (SC) and **affinity propagation** (AP).

Besides, some similar clustering algorithm testing results can be seen at [[scikit-learn](#)].

2 Testbed Data Sets

We are using 8 point cloud data sets which are either regular or irregular shapes to testify the effectiveness of the clustering algorithms. The intrinsic cluster numbers are a priori known from natural intuition. The data sets are illustrated in Fig. 1.

3 Spectral Clustering

Spectral clustering (SC) has two versions available, i.e., one with k-means as post-processing [1] and the other with eigenvector rotation minimization as post-processing [2]. The SC-kmeans suffers stability problem which might be able to detect meaningful shapes due to drawback of k-means, while SC-eigenrotation suffers from being unable to find reasonably optimal number of clusters as indicated in [[self-tuning SC github](#)]. K-means++ [3] can be applied to enhance careful seeding for initial samples of k-means of SC spectral clustering, which solves stability and local minimum problem for SC with k-means, but SC with eigenrotation however, the final clustering results is very sensitive to scaling factor setting which is hardly able to be designed.

The SC with k-means is illustrated in Fig. 2 and we can observe that SC with k-means can not work well in Fig. 2(b), 2(c) or 2(d) while working well for all other data sets.

The SC clustering with eigenvector rotation minimization is illustrated in Fig. 3, and we observe that SC with eigenrotation also fails to detect natural clusters for Fig. 3(b), 3(c) and 3(d).

4 K-means

We implemented C++ as initial sampling method and tested the clustering effect illustrated in Fig. 4. Obviously that k-means could only detect convex and spherical shape of clusters.

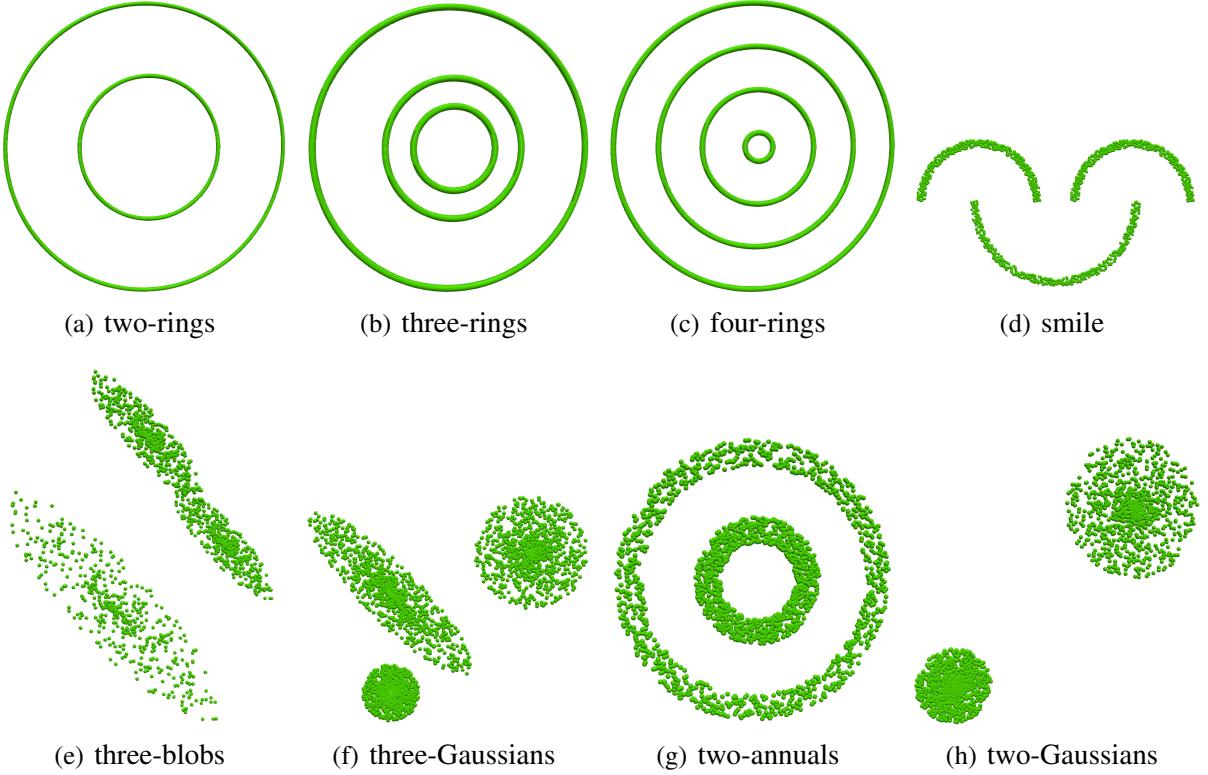


Figure 1: 8 groups of point cloud data sets which has explicitly predefined number of clusters.

5 K-medoids

Similar to k-means in Sec. 4, k-medoids also only detects convex (not strictly spherical) shape of clusters as illustrated in Fig. 5. There is only slight difference between the results of k-means and k-medoids, while computational overhead is highly increased due to medoid computing (could be median from samples or by iteration as in [geometric median]), and the results show that with former method).

6 AHC

As is well-known, AHC has three linkages, i.e., single, complete and average. The results of single, complete and average linkage are respectively shown in Fig. 6, 7 and 8. Only single-linkage AHC could authentically detect all complicated shapes of clusters, while complete-linkage and average-linkage could not, this is due to that single-linkage AHC is beneficial for detecting chaining effects and shapes.

7 BIRCH

8 DBSCAN

Some of clustering results is illustrated in Fig. 9. With careful selections of parameters, DBSCAN is able to detect naturally meaningful shapes of clusters.

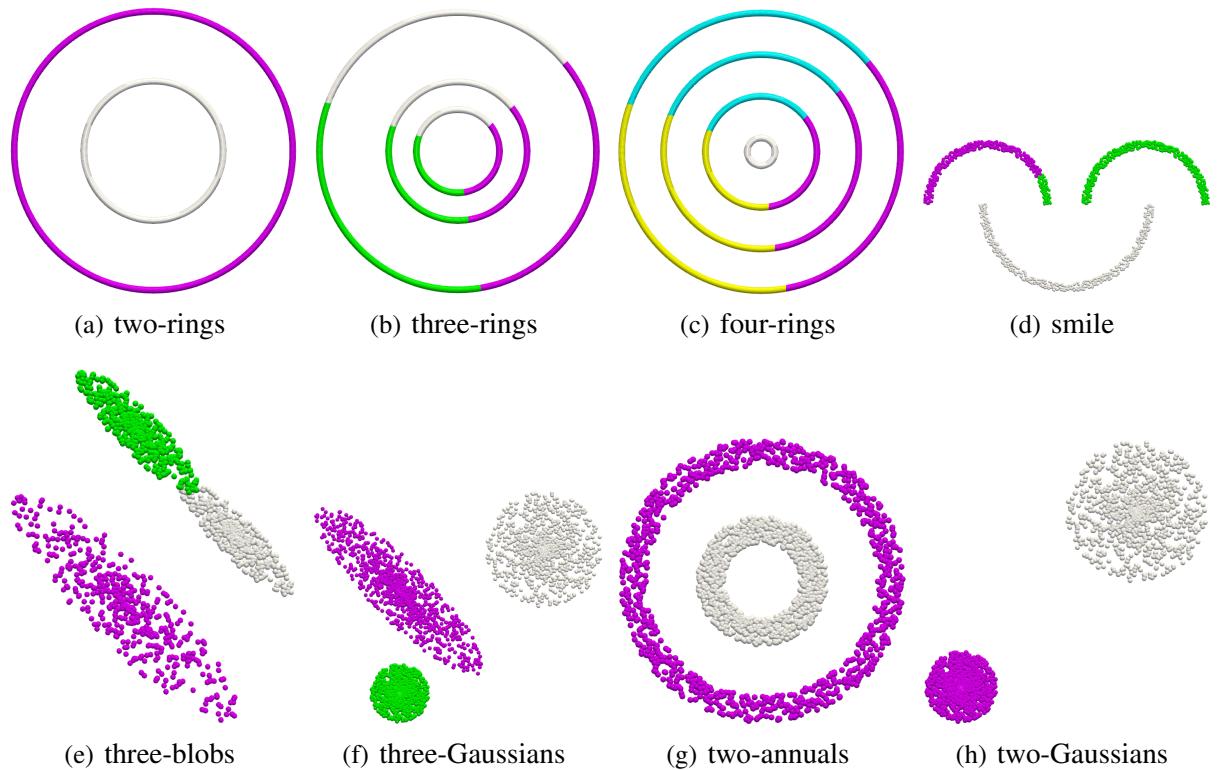


Figure 2: SC with k-means clustering results on point cloud data sets.

9 OPTICS

10 AP

References

- [1] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, pages 849–856. MIT Press, 2001.
- [2] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, NIPS’04, pages 1601–1608, Cambridge, MA, USA, 2004. MIT Press.
- [3] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA ’07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

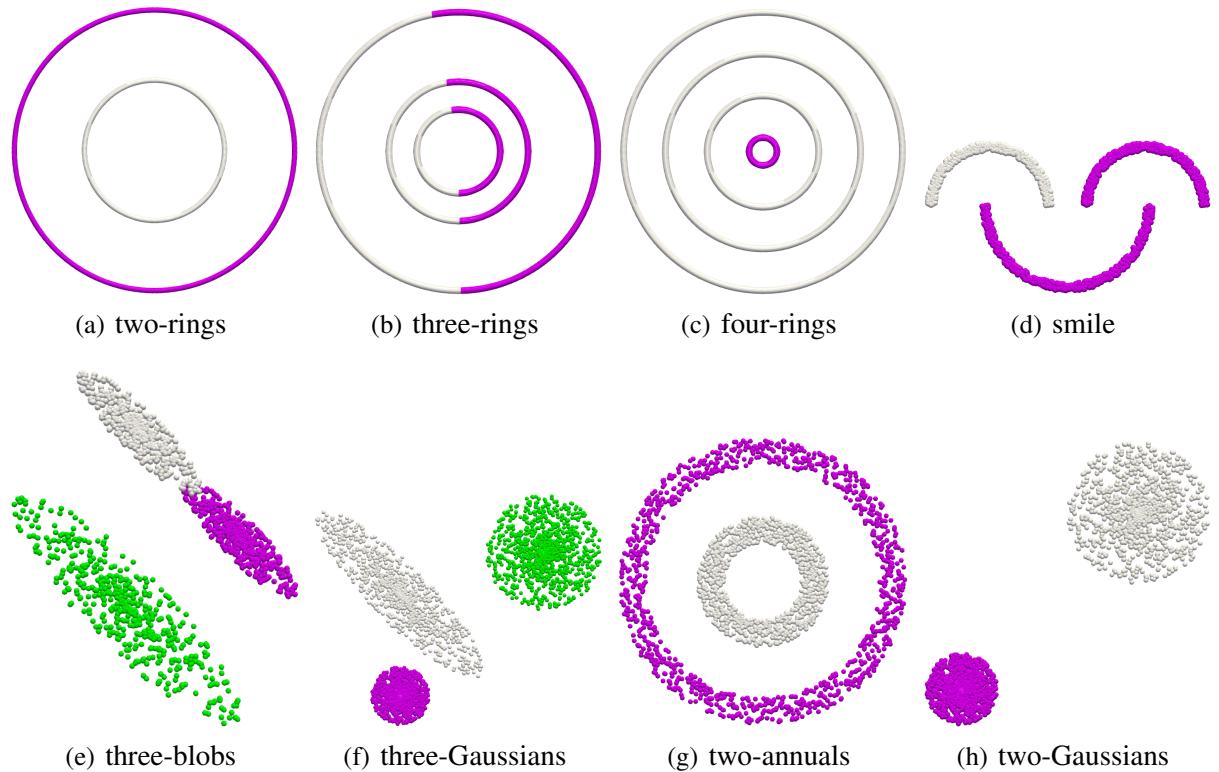


Figure 3: SC clustering with eigenvector rotation minimization results on point cloud data sets.

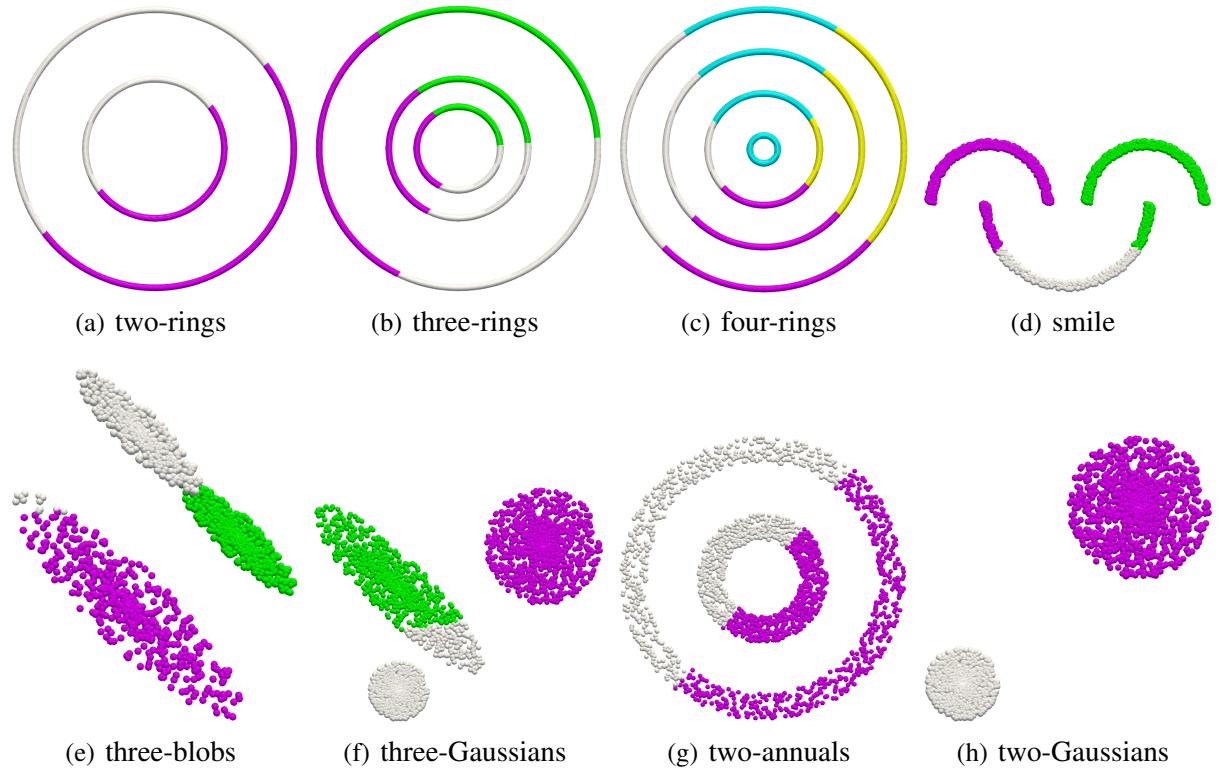


Figure 4: K-means clustering results on point cloud data sets. Notice that only two spherical clusters in Fig. 4(h) can be authentically detected.

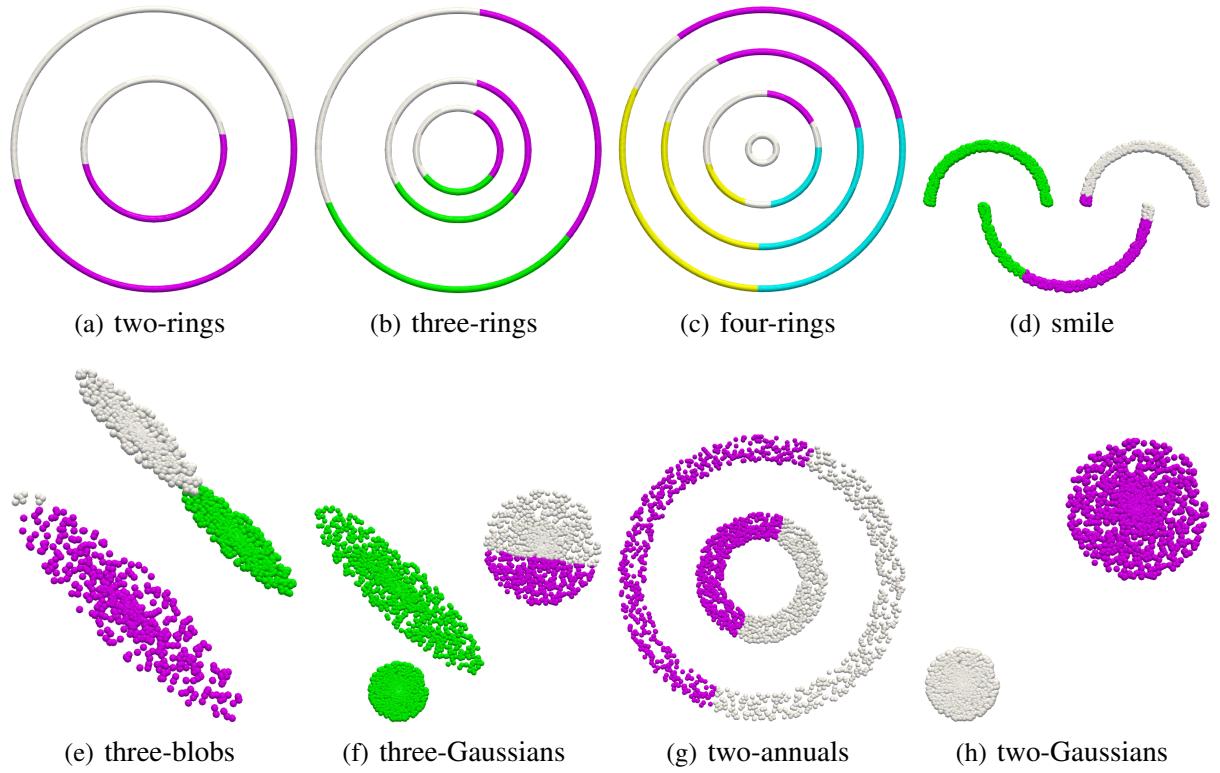


Figure 5: K-medoids clustering results on point cloud data sets. Note that only Fig. 5(h) can be authentically detected.

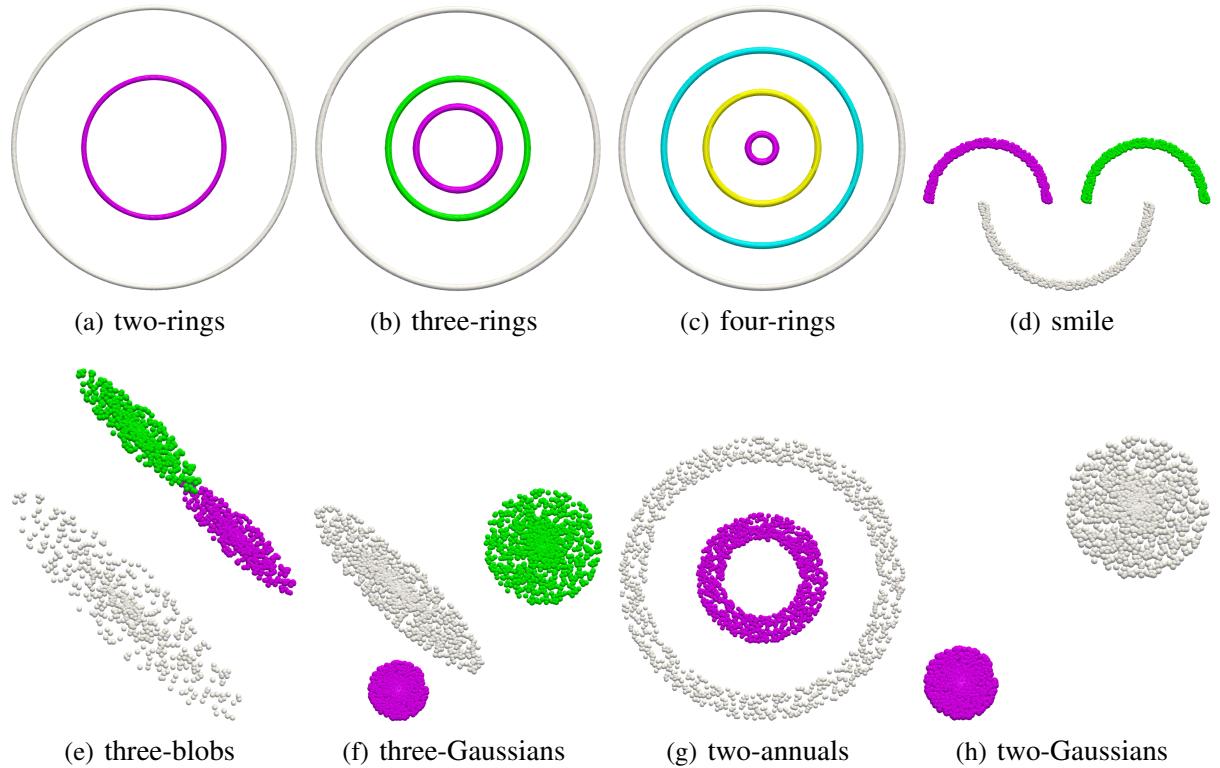


Figure 6: Single-linkage AHC clustering results on point cloud data sets. Note that only single-linkage AHC could precisely detect all shapes of clusters.

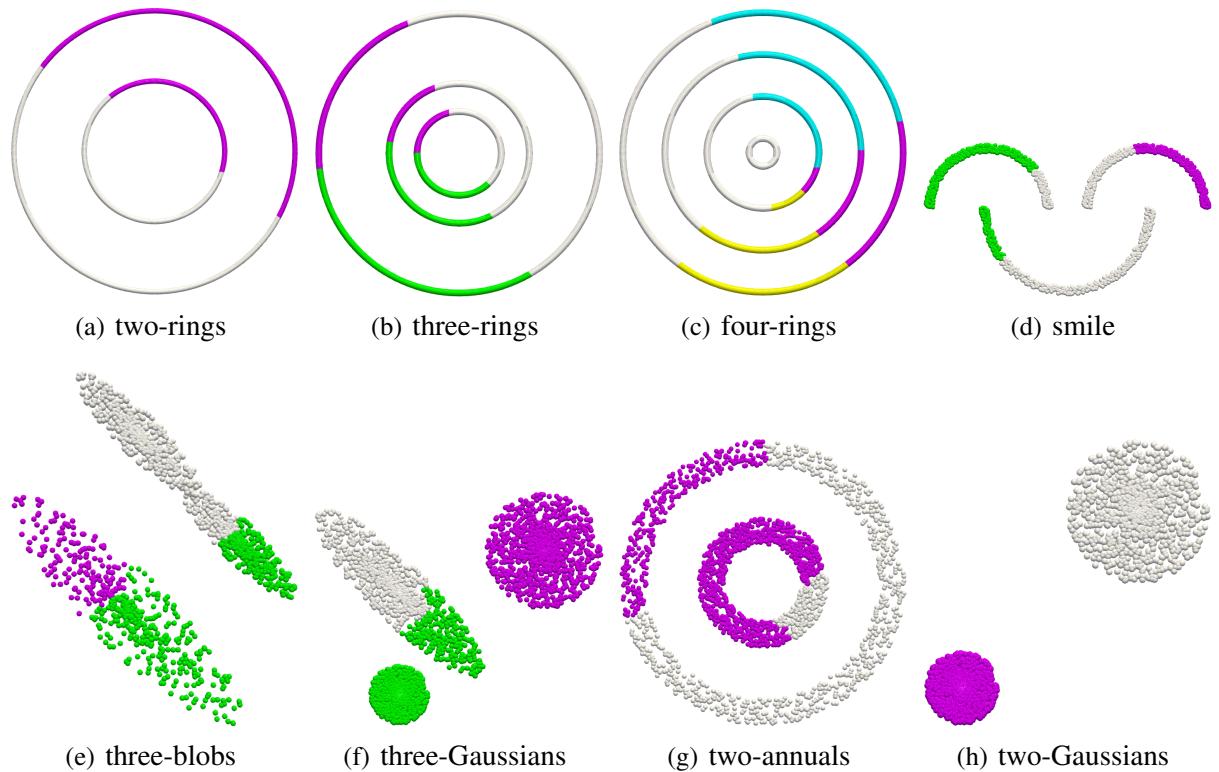


Figure 7: Complete-linkage AHC clustering results on point cloud data sets.

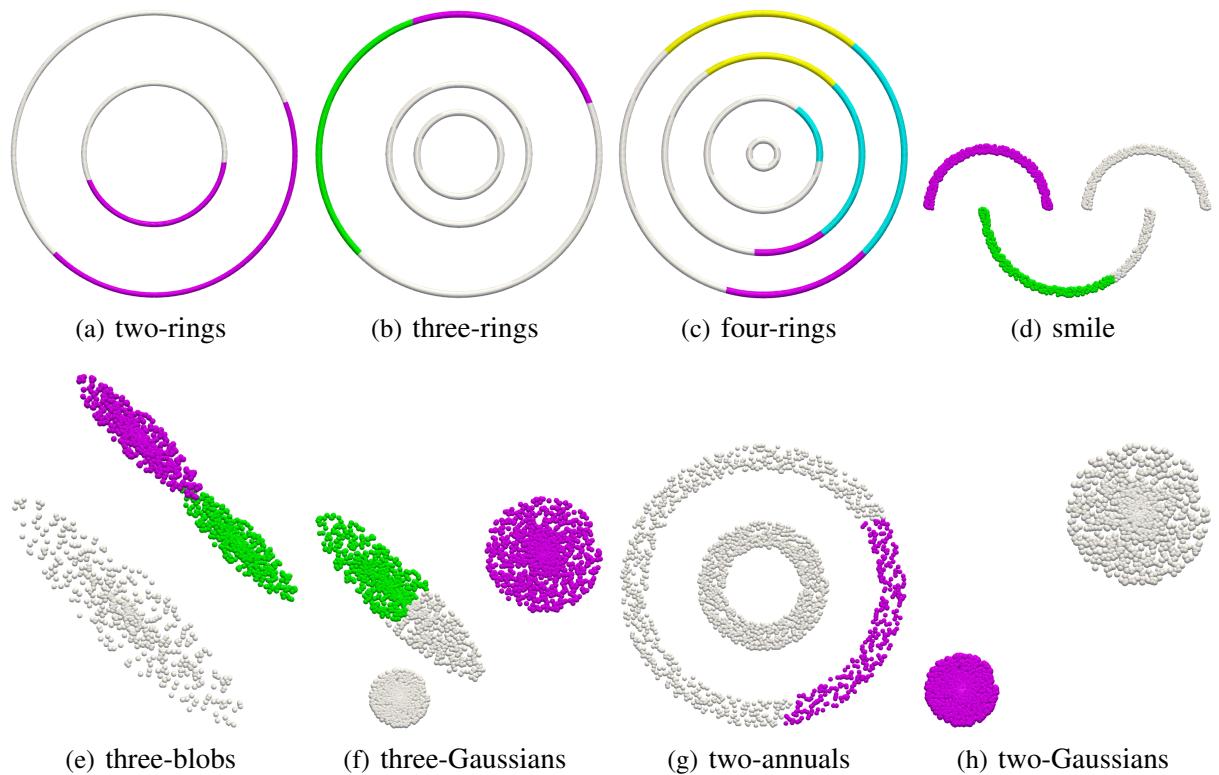


Figure 8: Average-linkage AHC clustering results on point cloud data sets.

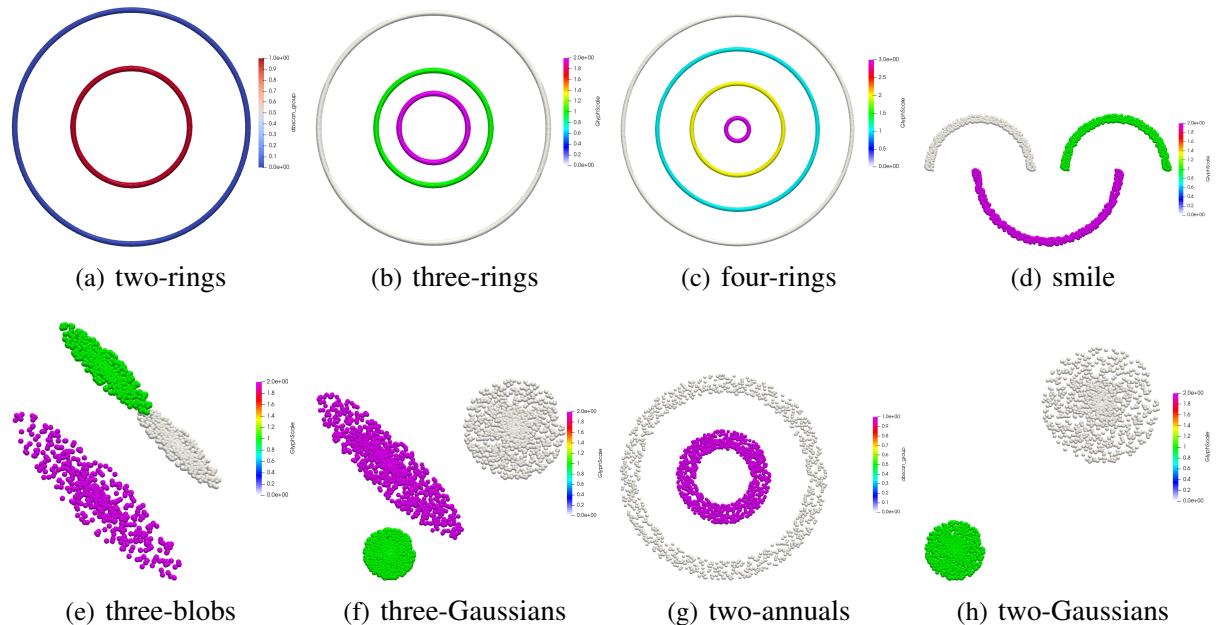


Figure 9: DBSCAN clustering results on point cloud data sets. To note that the white color denotes the NOISE. The parameter setting is based on that, one parameter minPts is for user-input, and another parameter ϵ is decided by the average minPts -th distance of every point to all other points.