Delivery 2

Topological Data Analysis

Daniel Salgado Rojo

Mathematics for Big Data. Course 2017/2018.

In this homework assignment we use the TDA package for R to analyse some real and hard-coded datasets. In particular we use persistence diagrams to analyse the dimensionality of the wholes appearing in the datasets.

1 Persistence diagrams for a real dataset

In this first section we use the famous Iris dataset that contains observations of three classes of iris plants with 50 instances for each one. The four features are the sepal and the petal lengths and widths.

In order to do our analysis with a three dimensional data set, for easy visualization, we are going to consider only three features: the sepal and petal lengths and the petal width.

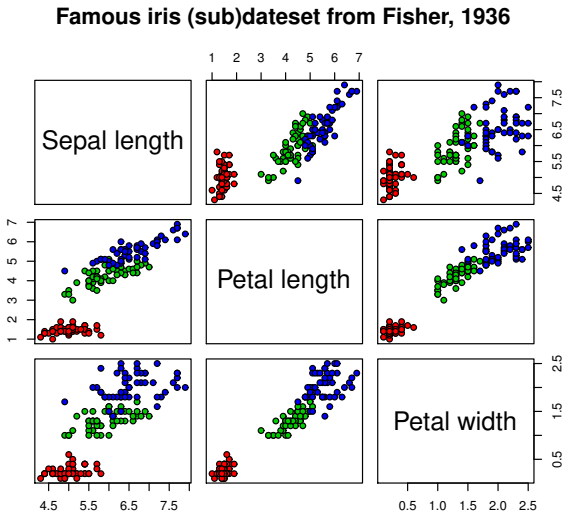


Figure 1

Figure 1 shows the iris data set for all six possible combinations of pairs of features, with a different colour for each iris plant class. Figure 2 shows a 3 dimensional representation of the considered dataset.

From the two figures we can see that one of the iris plants' classes can be linearly separated from the other two (we could find a plane such that the red class is in one side of the plane and the other two classes are in the other side).

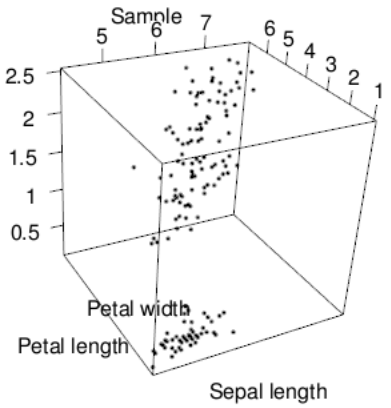


Figure 2

We now have prior knowledge about the dataset, that as we are going to see, we could have deduced from the persistence diagrams used for any pair of the three features we have selected at the beginning.

For instance, in Figure 3 we consider the persistence diagrams (the one on the left using the kernel density estimator and the Vietoris-Rips diagram on the right) for the petal and sepal lengths.

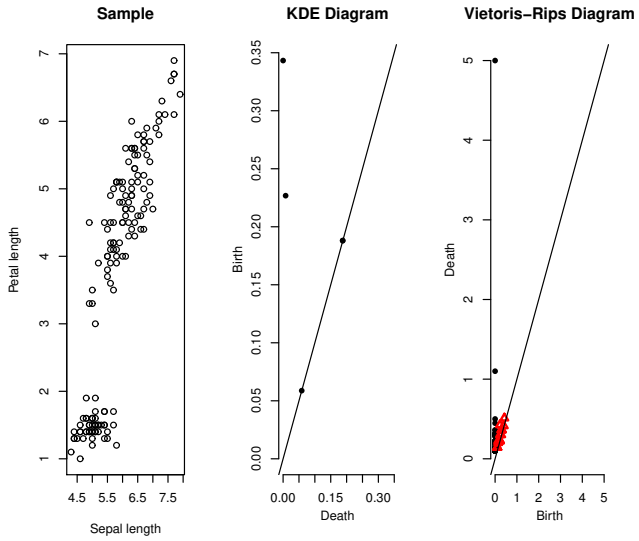


Figure 3

We can observe that there are two significantly persistent black dots that are far way from the diagonal. These two dots may indicate that in this subdataset of two features there are two wholes of dimension 0. In fact, this is what can be observed from the left-most graphic from Figure 3, and is in agreement with the fact that one of the classes in the dataset of the three iris plants dataset was linearly separable from the others (we can distinguish two points clouds clearly).

Similarly, in Figure 4 for petal width and sepal length, and in Figure 5 for the petal width and length, we can arrive to the same conclusion, since there can be distinguished two black dots with high persistence (thus, two 0-dimensional wholes).

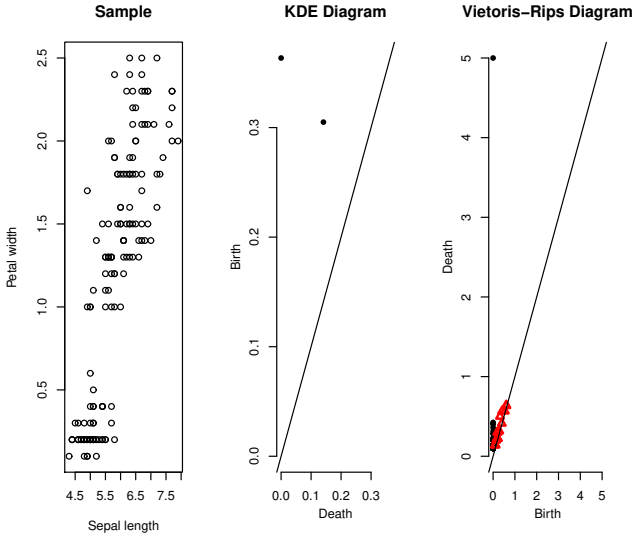


Figure 4

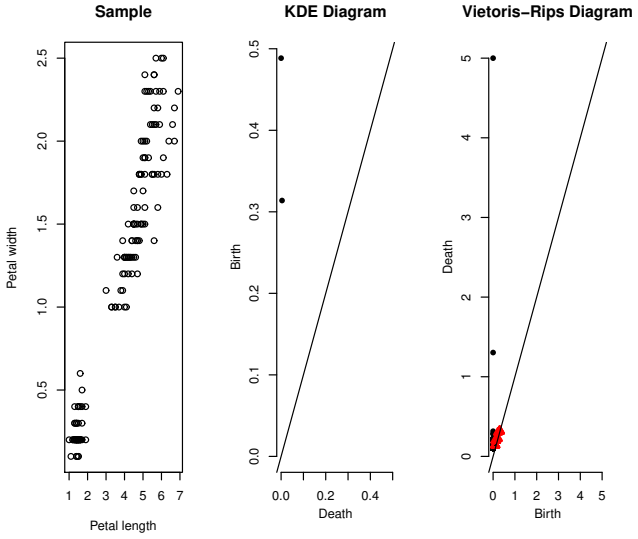


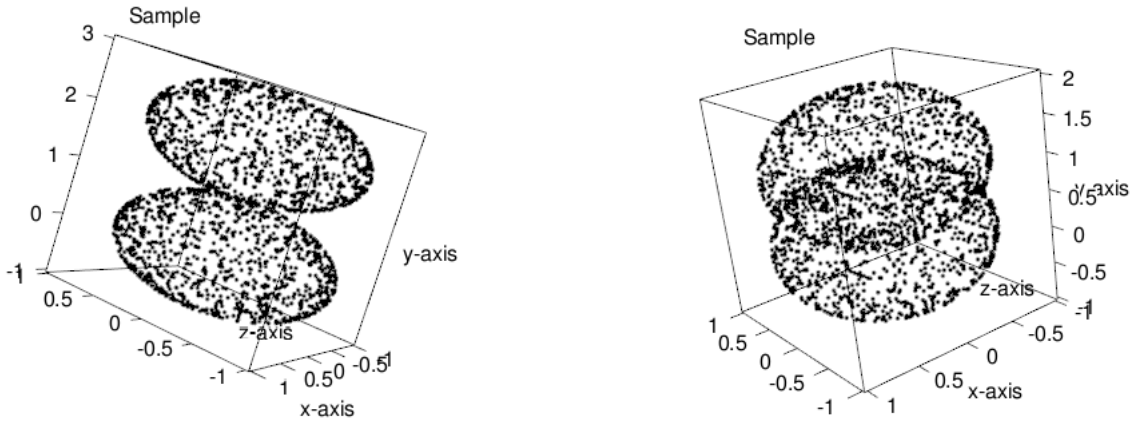
Figure 5

Finally we note that the KDE diagram gives more clear (less noisy) results than the Vietoris-Rip diagram, and moreover, it is computed faster. Thus, in the next section we are going to show just KDE diagrams.

2 Persistence diagrams for artificial datasets

In this section we analyse the persistence diagrams for two artificial datasets that we have manually created.

The first one (Figure 6a) consists on three dimensional points defining two 2-dimensional spheres tangent to each other (the usual 2d sphere centred at the origin and another displaced 2 units in the y direction). The second one (Figure 6b) are again two 2-dimensional spheres but now partially overlapped (the usual unit sphere and an sphere displaced 1 units, i.e. 1 radius distance, from the first and in the y direction).



(a) (b)

Figure 6

Before analysing the persistence diagrams, we can comment a priori which would be the expected results given the nature of our generated datasets.

For the first case where we have two tangent spheres, we expect to detect two 2-dimensional wholes, one coming from each of the spheres, and no more wholes.

For the second case where we have two overlapped spheres, we expect to detect three 2-dimensional wholes: one "shared" by the two spheres and another one for each of the spheres.

In Figure 7 we can see the dataset projected in the x-y plane on the left and the corresponding persistence diagram on the right. Black dots would denote 0-dimensional wholes, red triangles would denote 1-dimensional wholes and finally blue rhomboids would denote 2-dimensional wholes. The two significantly persistent dots are two blue rhomboids that are very far away from the diagonal. As expected, this says that there are two 2-dimensional wholes in the dataset for the two tangent spheres.

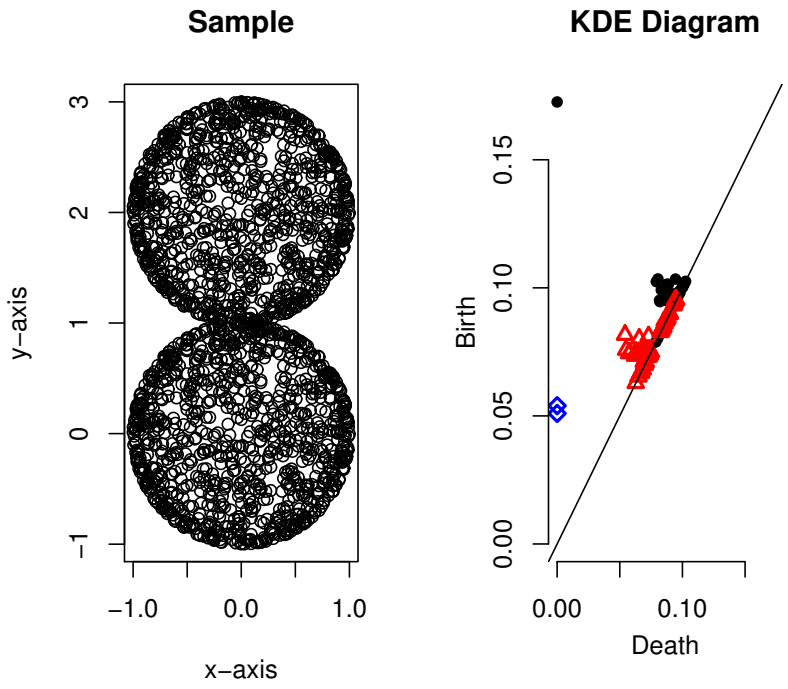


Figure 7

Similarly, in the persistence diagram for the overlapped spheres dataset, shown in Figure 8, we can distinguish three blue rhomboids that are significantly persistent and far from the diagonal, which would say there are three two-dimensional wholes in the dataset. If we compare the persistence diagram in this case with the persistence diagram for the two tangent spheres, we see that now there are some noisy points that are quite confusing (for example one of the red triangles is quite far from the diagonal, but not as much as the blue rhomboids are).

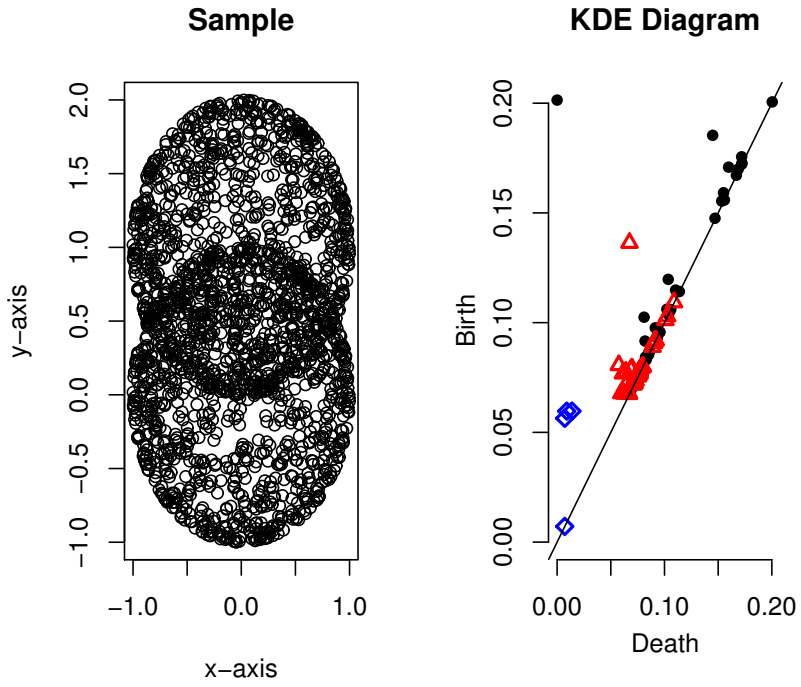


Figure 8

In conclusion, we have seen how persistence diagrams could help us to detect some features of the datasets such as are the "n-dimensional" wholes.

References

[1] Iris dataset. Famous database; from Fisher, 1936. A possible source: `https://archive.ics.uci.edu/ml/datasets/Iris`.

[2] TDA package for R. `https://cran.r-project.org/web/packages/TDA/vignettes/article.pdf`.