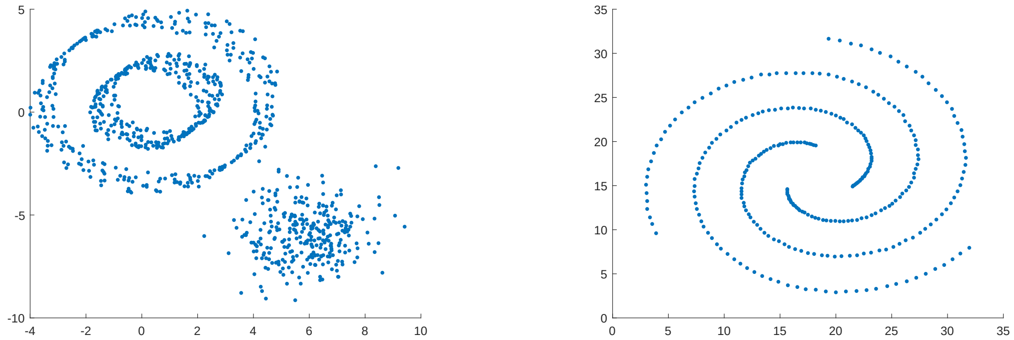
Computational Linear Algegbra For Large Scale Problems

Nenna Giulio, Ornella Elena Grassi

Spectral Clustering Homework

The aim of this homework is to implement and apply **Spectral Clustering** to two different sets of datapoints in $\mathbb{R}^2$. The two sets are shown in Figure 1



(a) Scatterplot of the data stored in `Circle.mat` (b) Scatterplot of the data stored in `Spiral.mat`

Figure 1: Scatterplot of the two Datasets

As it is clearly visible through visual inspection, both datasets contain 3 different shapes that can be classified as different clusters. In the `Circle` dataset there are two concentrical circles and a cloud of points in the bottom right while in the `Spiral` dataset there are 3 spirals. Traditional clustering algorithms, that mainly rely on euclidean distance, may fail in recognizing the presence of shapes in our data hence our need to rely on a different technique called **Spectral clustering**.

1 K-Nearest Neighborhood Graph

First, we need to define a similarity function that measures "how much our points are similar to each other". Let X_i and X_j be two points in our data, then we will use a similarity measure defined as:

$$s_{i,j} = \exp\left(-\frac{\|X_i - X_j\|^2}{2\sigma^2}\right) \tag{1.1}$$

Then, a *K-Nearest Neighborhood* similarity graph is a Graph $G = (V, E)$ where each vertex $v_1, \ldots v_n$ represents a point and two vertices v_i and v_j are connected by an undirected edge $e_{i,j}$ if the similarity between v_i and v_j is among the K-th highest similarities between v_i and other vertices in V. For such graph we can define the relative adjacency matrix as $W_{i,j} = s_{i,j}$ where each entry $W_{i,j}$ is nonzero only if there exists an edge between v_i and v_j. W has zero-values on diagonal by definition.

The following MATLAB code was use to generate the K-NN similarity graph of our data: