CH 5650: Molecular Data Science
Quiz

Shashwat Patel (MM19B053)

31 March 2022

1 Objective

The molecular trajectory consists of multiple phases/structures. Apply the concept of dimensionality reduction and clustering to classify the phases. Based on your analysis, report the number of phases present in the trajectory. Draw a lower dimensional representation of the data to show all possible phases. We have been given data of different atom positions in a molecule at different time frame.

2 Methods

Two dimensionality reduction methods were used, Principal Component Analysis(PCA) and Kernel PCA was used (I tried using other non-dimensional reduction technqiues but due to requirement of large memory, these methods could not be used in my laptop). For clustering, K-means clustering has been used.

2.1 PCA

PCA is a linear dimensional-reduction technique. It transforms a large set of variables into a smaller set that still has most of the information. On performing PCA in our dataset, we got this result. We reduced our datasets features to 2 dimensions only (Figure 1)

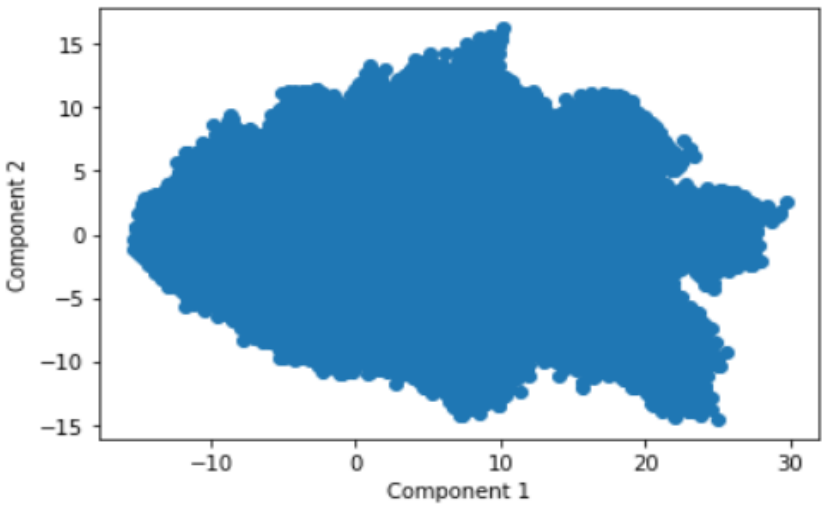

Figure 1: PCA

No distinct clusters were seen when linear dimensionality reduction was performed.

2.2 Kernel PCA

Kernel PCA is just an extension of PCA. It performs non-linear dimensionality reduction using different types of kernel. In our case we have used Radial-Basis Kernel. RBF kernels are the most generalized form of kernelization and is one of the most widely used kernels due to its similarity to the Gaussian distribution. The RBF kernel function for two points X_1 and X_2 computes the similarity or how close they are to each other.

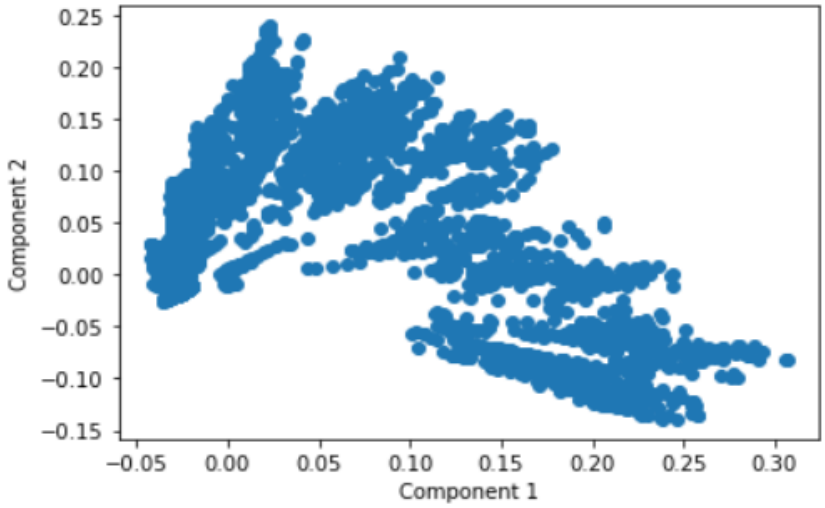

Figure 2: Kernel-PCA

Some distinct clusters can be seen in this case.

2.3 KMeans Clustering

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. The K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. The 'means' in the K-means refers to averaging of the data; that is, finding the centroid.

To determine k, we have used the elbow method. The approach consists of looking for a kink or elbow in the WCSS graph. Usually, the part of the graph before the elbow would be steeply declining, while the part after it – much smoother.

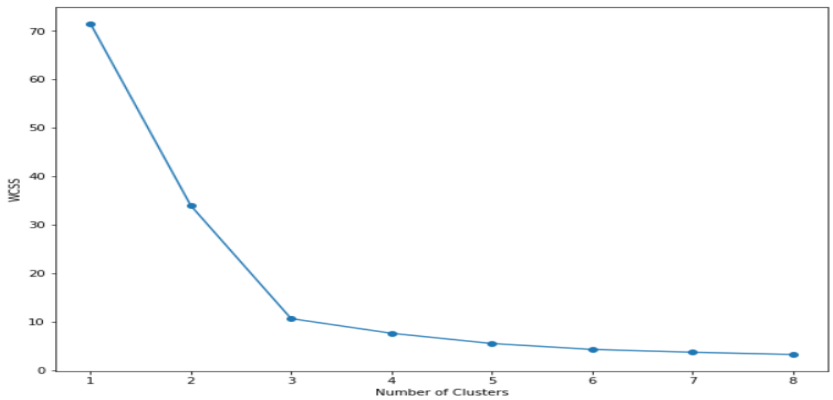

Figure 3: WCSS Graph

From figure 3, it is determined that best set of clusters is 3.

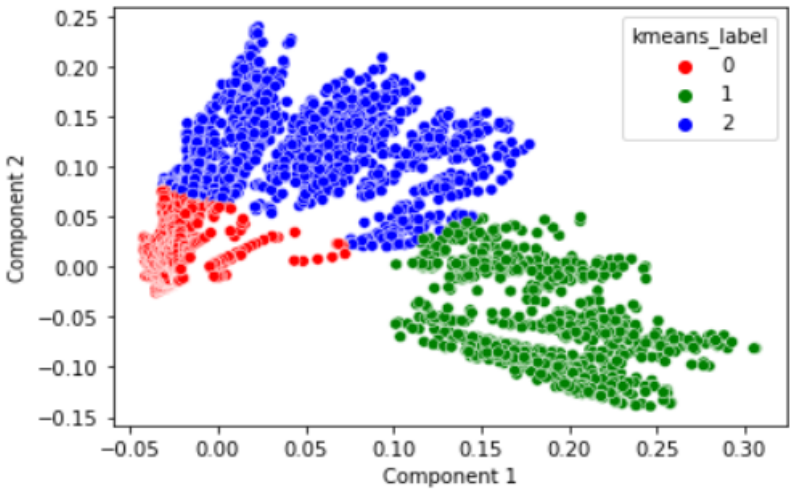

Figure 4: Clustering

This is the clustering(phases) we get. Based on the analysis, we get 3 distinct phases.