Nabin Chapagain 1001551151

The problem we are solving here is a pretty common problem in the field of data science. We are aiming to get an unsupervised clustering algorithm that can provide us with the the right number of clusters of points in a n-dimensional dataset.

The algorithm that I have here does just that but if asked, also gives us information about the minimum volume bounding ellipsoid (MVBE) that binds the given cluster. The MVBE is obtained using the function *getRandMVU* and the covariance form *getMinVolPartition*. The method that was used involves computing volume of an ellipsoid in n-dimension; the ellipsoid is computed using the function *getMinVolPartition* which gives us the covariance of the ellipsoid which is required in the computation of volume of n-dimensional ellipsoid. Then we use a recursive approach where we divide the whole dataset into two sub-clusters using K-means clustering and get a MVBE that tightly bounds the two sub-clusters. We compute total volume of the sub-clusters and compare it to the volume of parent cluster. If the volume of parent cluster was smaller than total volume of sub-clusters, then it is clear we reached the limit of number of clusters and we stop there. But if the sum total of volume of sub-clusters is smaller than that of the parent cluster we use recursion and divide each of the smaller sub-clusters into further smaller sub-clusters and repeat the same approach until there is no more smaller sub-clusters or we reach the maximum number of cluster possible that is equal to number of points divided by number of dimensions plus one ; because to form a unique MVBE in two dimensions, you need at least three points, four points in three dimensions and so on.

The result as expected is the correct number of clusters of points in a dataset. Even though the code fulfills all the requirements of honors assignment, there are still a lot of room for improvement in the code. First of all, it only works with the dataset with points involving to clusters with invertible, positive definite covariance matrix. And, the use of minimum volume as the defining factor might not be always the best, something even better might be to classify the clusters on the basis of their density which will be my next step.