Part - 1 Clustering:

1. Dimension of M, X, and trueLabels;
   M :(1549, 65)
   X : (1549, 64)
   trueLabels: (1549,)

2. KMeans algorithm from sckit-learn library by default uses Euclidean distance as similarity measure. Model instance has been created with 5 clusters and 50 initializations. Random state has been set so that results can be reproduced. This model selects the best output out of 50 initializations. ARI for single run is 39.78% and AMI is 0.477906
   For the average performance over 50 random initializations, I have used for loop 50 times on the same model as model1. This has given 50 best outputs. Average ARI for this model is 37.96% and average AMI is 0.477968. Using two tail hypotheses testing with 5% level significance we conclude that there is sufficient evidence to conclude that ARI = AARI and AMI = AAMI.

   From below plots, it seems there is spike between 10 to 14 iterations (outliers) and is constant for rest of the iterations.

   Fig 1: Rand index over 50 iterations

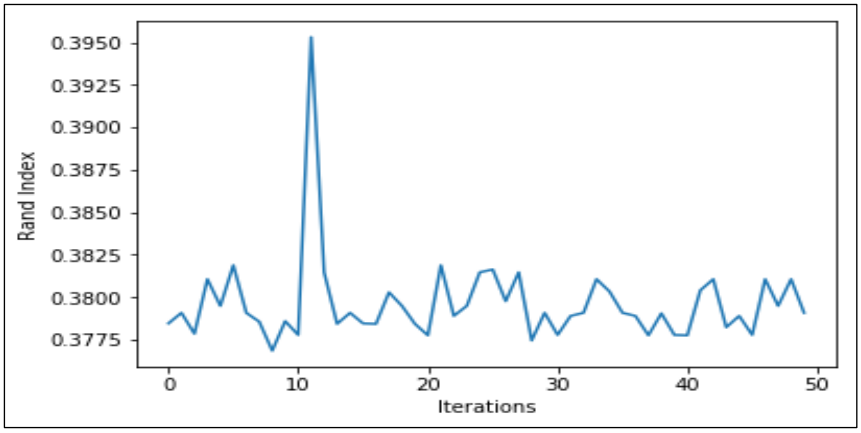
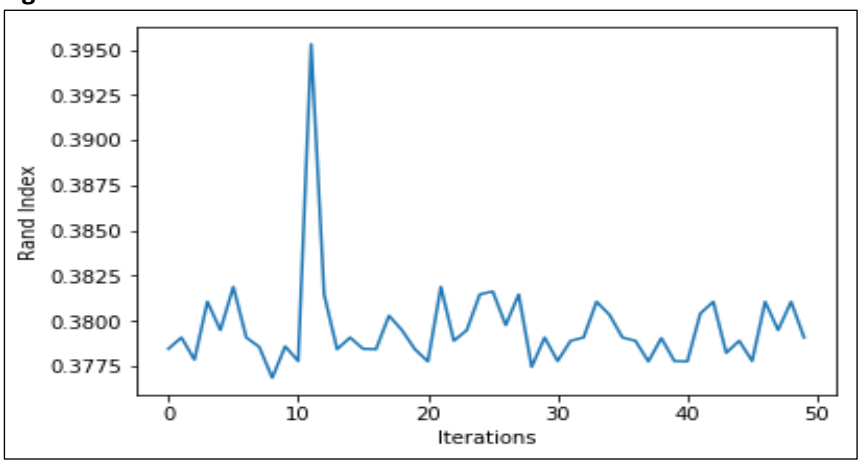


   Fig 1: Mutual Info over 50 iterations



3. KMeans algorithm for SKlearn has three initialization methods; k-means++ which selects initial cluster centres in a smart way to speed up convergence, 'random' which chooses k observations at random from data for the initial centroids, and 'ndarray' which is user defined

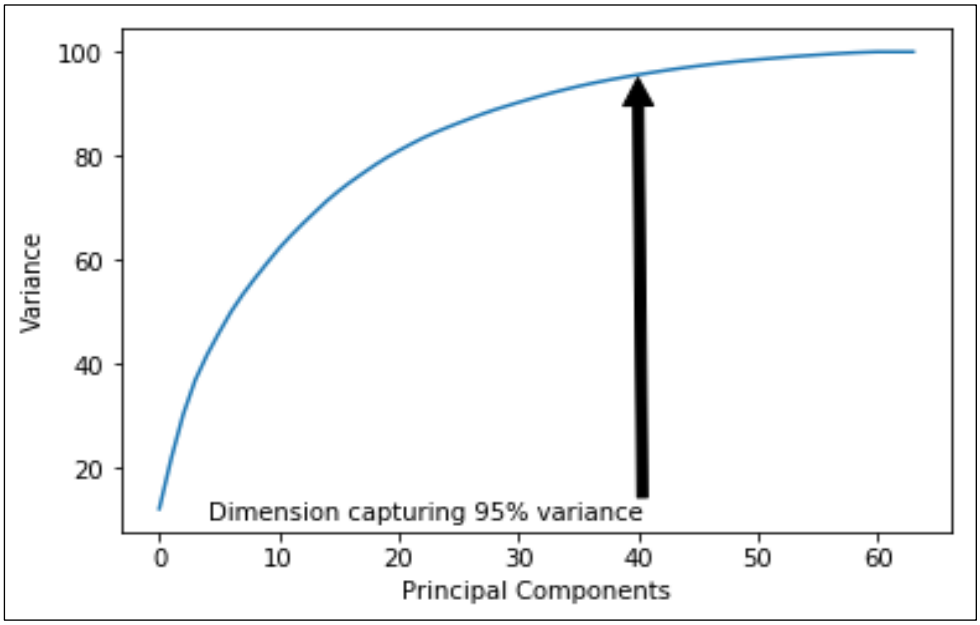
initial centres. We can set the iterations of the initialization which runs with different centroid seeds. However, the final results will be the best output out of all the iterations. Accuracy of the clusters depends on optimum number of clusters and centroids identified by the algorithm or defined by the user. If we run KMeans algorithm 20 times, ARI value and AMI value of single run should be within the range of average ARI and average AMI. Thus, we can say that ARI value of single run will be similar to the average ARI value over 20 runs of KMeans algorithm. This has been shown via hypothesis testing in question 2.

4. I have used KMeans clustering algorithm from nltk library using cosine distance similarity measure for 50 random initializations. ARI of Kmeans cluster algorithm using cosine distance similarity measure is 39.64% and AMI is 0.4709. ARI of Kmeans cluster algorithm using Euclidean distance similarity measure is 39.78% and AMI is 0.477. Accuracy of the clustering for both the similarity measure is similar.

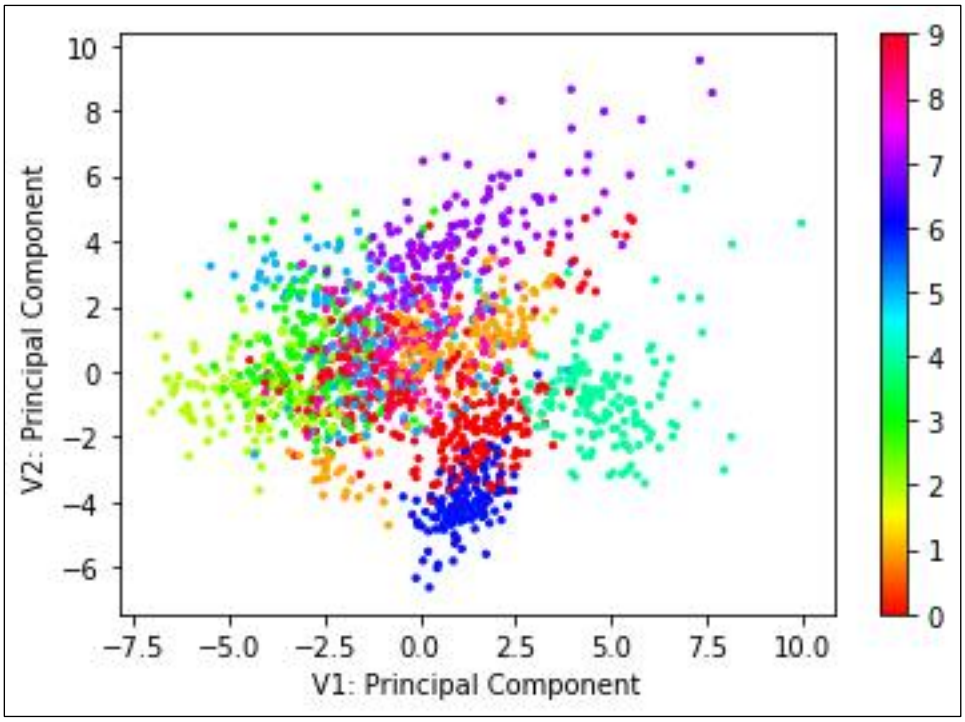
Part-2 Dimensionality Reduction using PCA/SVD:

1. Minimum dimension that captures at least 95% variance is 40 (95.2% variance).

Fig 3: Dimensions with cumulative variance



2. Scatter plot of trueLabels (0 to 9 digit) on X (V1) and Y (V2) subspace using color map instead of legend.


V2: Principal Component
V1: Principal Component
10
8
6
4
2
0
−2
−4
−6
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
9
8
7
6
5
4
3
2
1
0