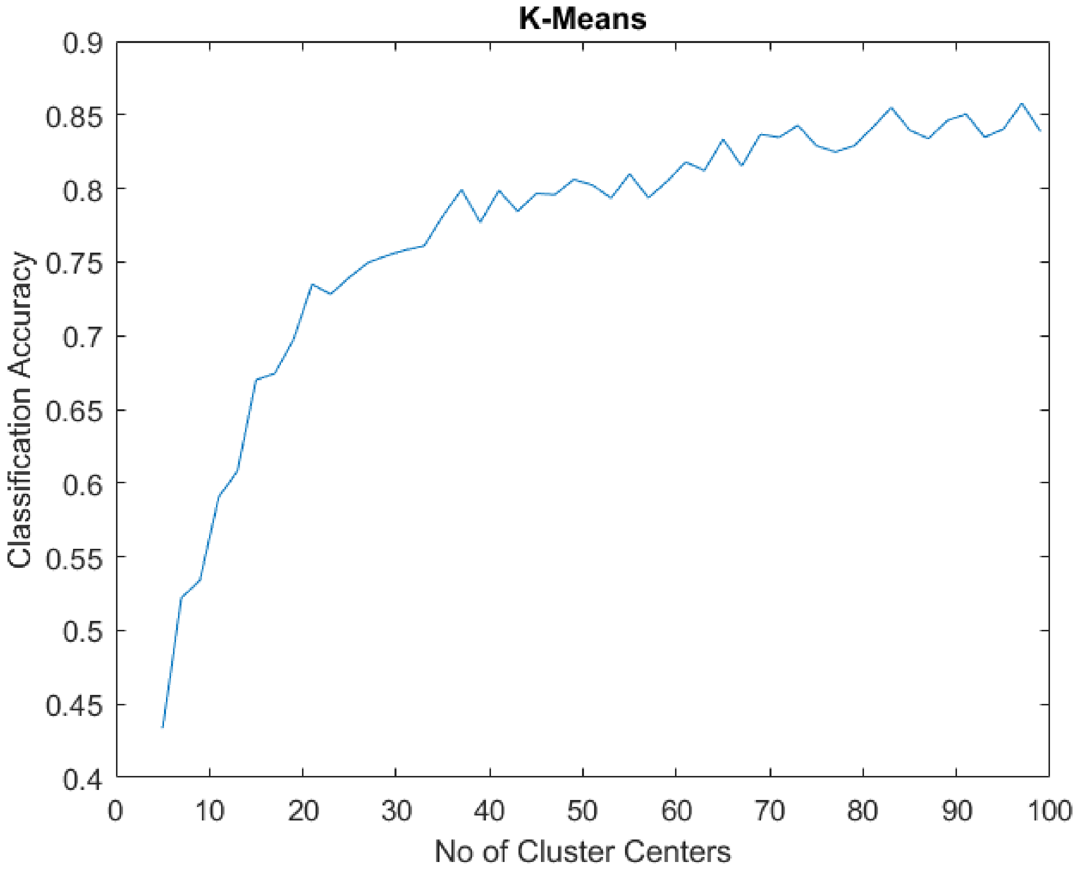
Lab 4 (K-Means Clustering)

Naman Goyal (2015CSB1021)

Experimenting with clustering and dimensionality reduction techniques

Q1) K-means clustering on the MNIST hand written digits' dataset



No of cluster centers	Accuracy
5	0. 433400
10	0. 537000
15	0. 665400

Observations

1. Classification Accuracy increases with increasing no of cluster centers.
   Explanation: As clusters get split further the accuracy is expected to go up.

2. The rate of increase of accuracy decreases with increasing cluster centers.
   Explanation: The trend is expected as rate of increase of accuracy is not linear but decay.

Confusion Matrix K-means on original dataset (Row: Actual, Col: Prediction)

Cluster Size 5

Prediction / Actual	0	1	2	3	4	5	6	7	8	9
0	425	3	0	40	0	0	27	5	0	0
1	0	495	0	3	0	0	0	2	0	0
2	3	92	0	56	0	0	338	11	0	0
3	3	59	0	408	0	0	7	23	0	0
4	0	41	0	0	0	0	41	418	0	0
5	9	163	0	247	0	0	14	67	0	0
6	7	70	0	10	0	0	410	3	0	0
7	4	64	0	0	0	0	3	429	0	0
8	2	167	0	272	0	0	22	37	0	0
9	2	55	0	11	0	0	7	425	0	0

Cluster Size 10

Prediction / Actual	0	1	2	3	4	5	6	7	8	9
0	393	5	11	28	5	0	20	0	38	0
1	0	496	1	2	0	0	0	0	1	0
2	3	57	393	18	7	0	2	4	16	0
3	1	52	11	263	14	0	2	4	153	0
4	0	51	10	0	292	0	7	140	0	0
5	6	157	2	146	40	0	11	6	132	0
6	6	69	77	3	3	0	335	0	7	0
7	1	68	3	0	158	0	0	270	0	0
8	0	78	8	135	13	0	1	22	243	0
9	2	42	5	8	233	0	2	207	1	0

Cluster Size 15

Prediction / Actual	0	1	2	3	4	5	6	7	8	9
0	383	0	0	49	3	26	24	2	12	1
1	0	496	0	0	1	1	0	0	2	0
2	3	74	331	28	12	12	17	6	16	1
3	0	35	6	340	3	12	1	7	94	2
4	0	31	4	0	166	8	11	90	0	190
5	4	12	0	159	12	195	6	0	97	15
6	6	27	1	3	5	10	446	0	2	0
7	0	53	2	0	23	3	0	406	0	13
8	0	41	2	52	9	14	2	2	366	12
9	2	24	1	5	125	2	2	134	7	198

Observations

1. At cluster center equal to 5,
   2 is getting misclassified as 6 (merged)
   4 is getting misclassified as 7 (merged)
   5 is getting misclassified as 3 (merged)
   8 is getting misclassified as 3 (merged)
   9 is getting misclassified as 7 (merged)

2. At cluster center numbers equal to 10,
   5 is getting misclassified as 1,3,8 (merged)
   9 is getting misclassified as 7 (merged)

3. At cluster center numbers equal to 10,
   5 is getting misclassified as 1,3,8 (unmerged)
   4 is getting misclassified as 9 (unmerged)

Explanation: The no of misclassifications and merging of clusters reduces on increasing the cluster centers.

As cluster centers are less and we are labelling each cluster with most frequently occurring digit; the cluster centers are expected to merge as less of cluster centers.

Classwise Precision and Recall

Cluster Size 5

Class	Precision	Recall
0	0.93407	0.85
1	0.40943	0.99
2	0	0
3	0.38968	0.816
4	0	0
5	0	0
6	0.47181	0.82
7	0.30211	0.858
8	0	0
9	0	0

Cluster Size 10

Class	Precision	Recall
0	0.95388	0.786
1	0.4614	0.992
2	0.75432	0.786
3	0.43615	0.526
4	0.3817	0.584
5	0	0
6	0.88158	0.67
7	0.41348	0.54
8	0.41117	0.486
9	0	0

Cluster Size 15

Class	Precision	Recall
0	0.96231	0.766
1	0.62547	0.992
2	0.95389	0.662
3	0.53459	0.68
4	0.4624	0.332
5	0.68905	0.39
6	0.87623	0.892
7	0.62751	0.812
8	0.61409	0.732
9	0.45833	0.396

```
Precision = Positive/ (No of Predicted positive)
Recall = Positive/ (No of Actual positive)
```

Observations

1. For merged cluster centers the classwise precision and recall is 0.
2. Precision and Recall increases on increasing the cluster centers.

Explanation: As clusters get split further the accuracy is expected to go up.

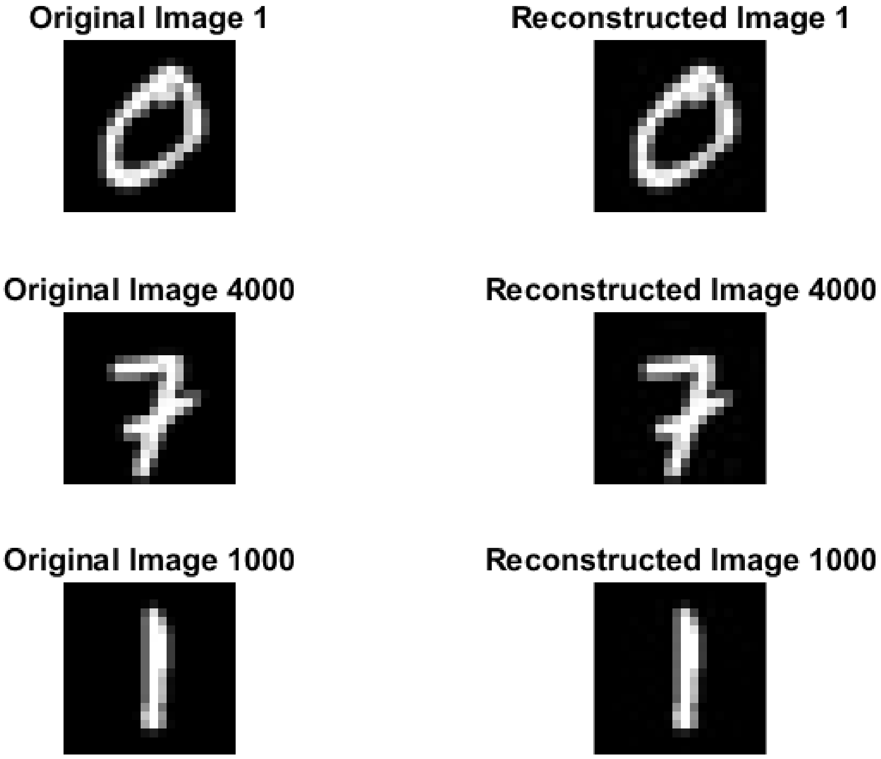
Q2) PCA to reduce the dimensionality of the digit images

The re-construction is achieved by just multiplying the transformed data with the transformation matrix U and adding the mean of the original data.

Reconstruction Error used is Mean Squared Euclidean Distance Error

It is found 191 principal components are required for should we consider for achieving a reconstruction error of 0.1

Figure: Using 191 principal components - Reconstruction Error of 0.1



Original Image 1

Reconstructed Image 1

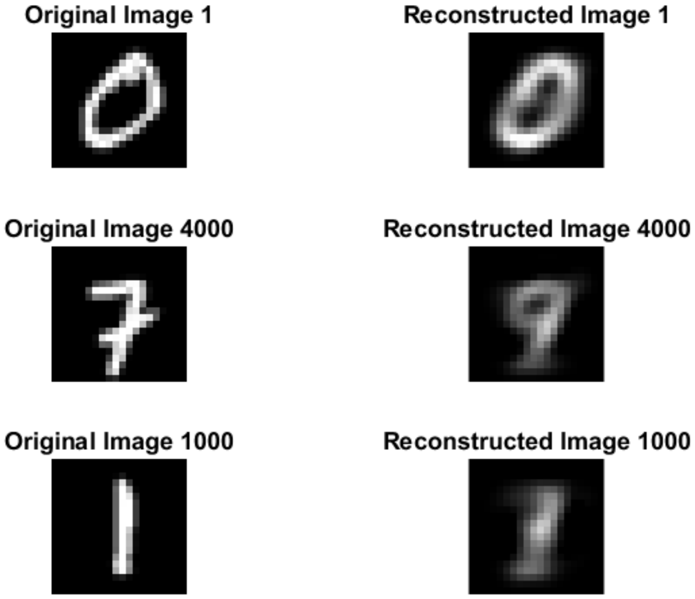
Original Image 4000

Reconstructed Image 4000

Original Image 1000

Reconstructed Image 1000

Figure 2: Using 4 principal components - Reconstruction Error of 16.1346



Original Image 1

Reconstructed Image 1

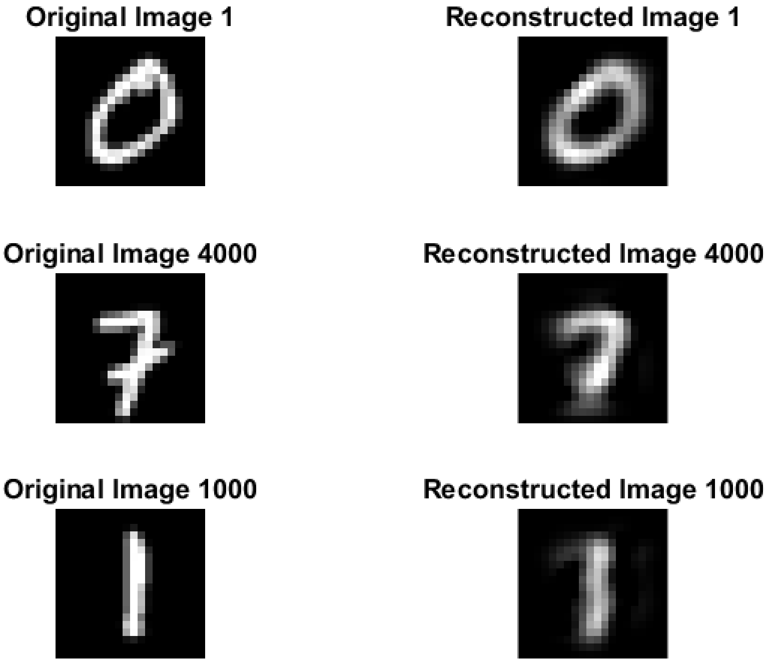
Original Image 4000

Reconstructed Image 4000

Original Image 1000

Reconstructed Image 1000

Figure 2: Using 11 principal components - Reconstruction Error of 10.2344



Original Image 1

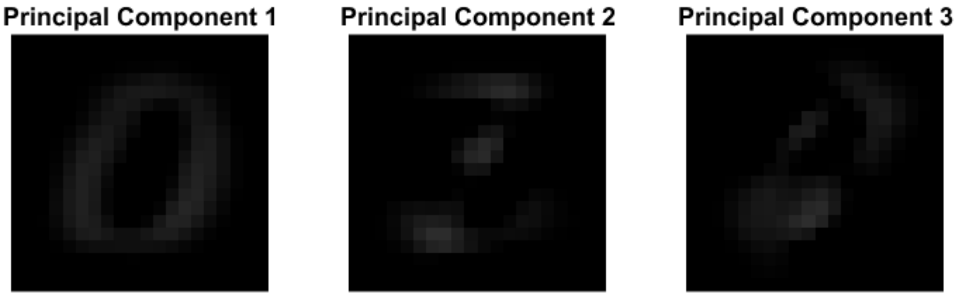
Reconstructed Image 1

Original Image 4000

Reconstructed Image 4000

Original Image 1000

Reconstructed Image 1000



The 3 Principal Components

The 1st principal component estimates roundness and is more like 0.

The 2nd principal component is more like 3.

Q3) K-Means on Projected Dataset

Confusion Matrix K-means on Projected dataset (Row: Actual, Col: Prediction)

Cluster Size 5

Prediction / Actual	0	1	2	3	4	5	6	7	8	9
0	454	4	0	0	10	0	24	8	0	0
1	1	496	0	0	2	0	0	1	0	0
2	27	118	0	0	11	0	335	9	0	0
3	291	125	0	0	46	0	6	32	0	0
4	0	30	0	0	290	0	21	159	0	0
5	208	138	0	0	56	0	14	84	0	0
6	13	69	0	0	11	0	407	0	0	0
7	0	49	0	0	196	0	0	255	0	0
8	101	174	0	0	59	0	18	148	0	0
9	8	22	0	0	244	0	4	222	0	0

Cluster Size 10

Prediction / Actual	0	1	2	3	4	5	6	7	8	9
0	397	4	4	27	5	0	20	3	40	0
1	0	497	0	1	1	0	0	0	1	0
2	4	91	329	31	14	0	13	5	13	0
3	1	50	15	282	6	0	2	15	129	0
4	0	35	3	0	203	0	9	250	0	0
5	5	142	1	139	20	0	13	43	137	0
6	7	75	5	2	43	0	359	0	9	0
7	1	51	1	1	34	0	0	412	0	0
8	0	71	3	114	21	0	3	36	252	0
9	2	24	0	9	122	0	2	340	1	0

Cluster Size 15

Prediction / Actual	0	1	2	3	4	5	6	7	8	9
0	381	1	0	38	4	37	34	1	2	2
1	0	494	0	2	0	1	0	0	2	1
2	3	80	259	32	4	9	88	9	11	5
3	1	39	9	381	0	12	3	6	37	12
4	0	23	3	1	316	6	9	11	0	131
5	6	15	1	206	18	201	12	0	10	31
6	6	30	0	6	6	9	443	0	0	0
7	0	48	0	0	23	3	3	402	1	20
8	0	44	2	144	19	11	3	4	259	14
9	2	25	0	9	171	2	2	132	1	156

No of cluster centers	Accuracy (Projected)
5	0. 380400
10	0. 546200
15	0. 658400

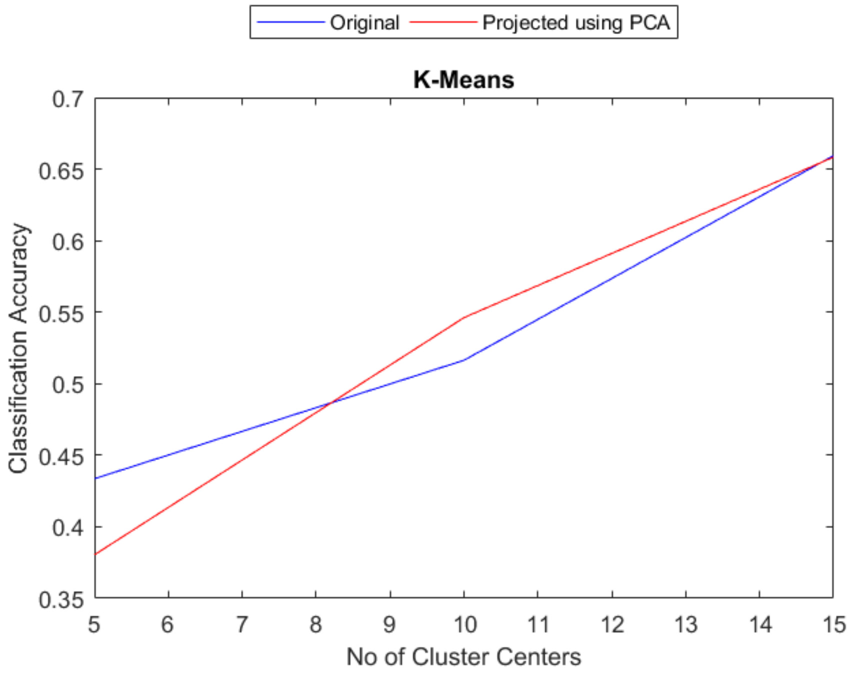
Observations

1. Classification Accuracy increases with increasing no of cluster centers on Projected dataset.
   Explanation: Same as original dataset, as clusters get split further the accuracy is expected to go up.

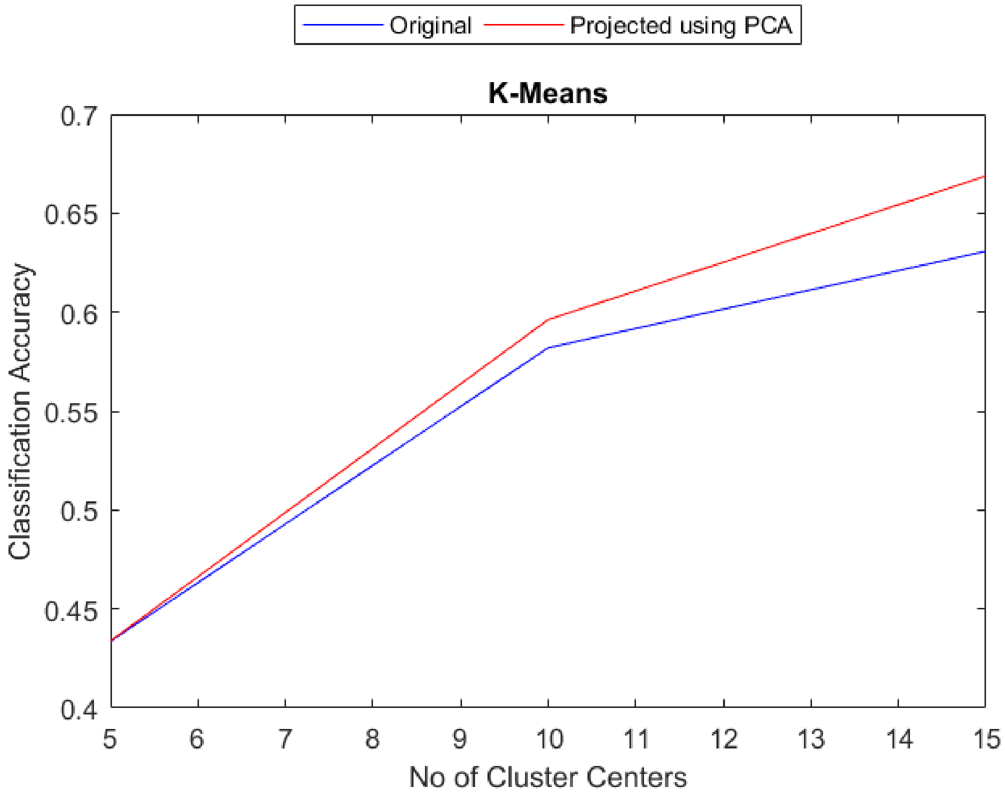
2. The rate of increase of accuracy decreases with increasing cluster centers on Projected dataset.
   Explanation: Same as original dataset, The trend is expected as rate of increase of accuracy is not linear but decay.

3. Similar to original dataset; classes are merged in the projected dataset on lower number of cluster centers.

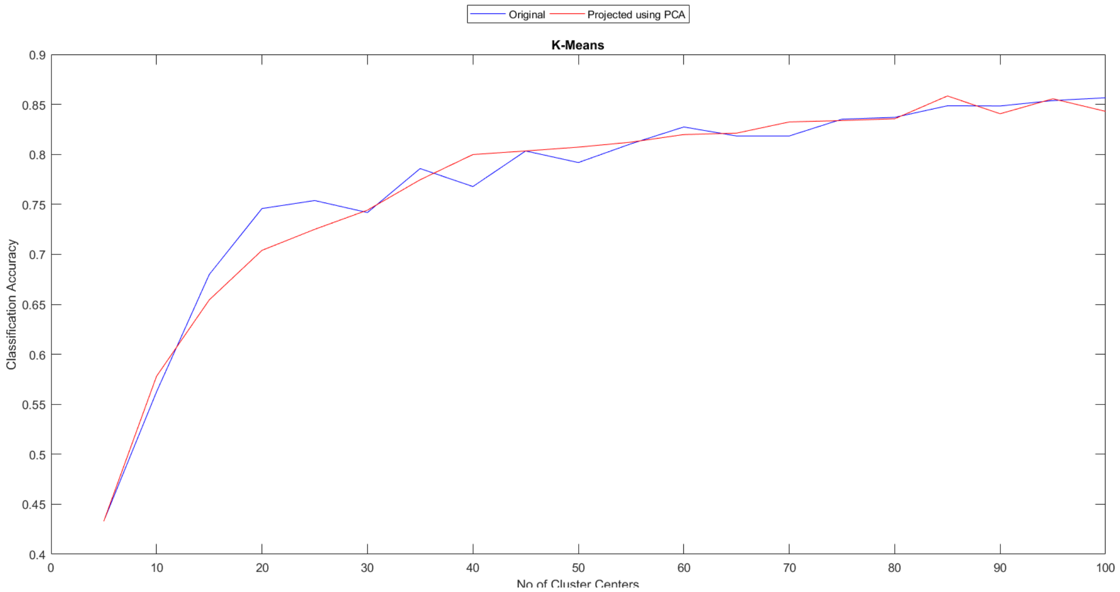
Comparing clustering in the low dimensional space and the original space



Run 1



Run 2



Observations

1. The trends of 2 graphs the original and projected dataset is nearly similar.
2. The differences across the run mainly arises because of the change of random seeds between iterations of K-Means.

Explanation:

Shows the projected data of l<u>ower dimensional able to represent the original dataset</u> to good extend and all the trends on original dataset present in projected also.