Report

Topics: Sparsity (PCA and Compressive Sensing)

Assigned: Wednesday May 23

Due: Sunday June 10 by midnight

Part I: PCA

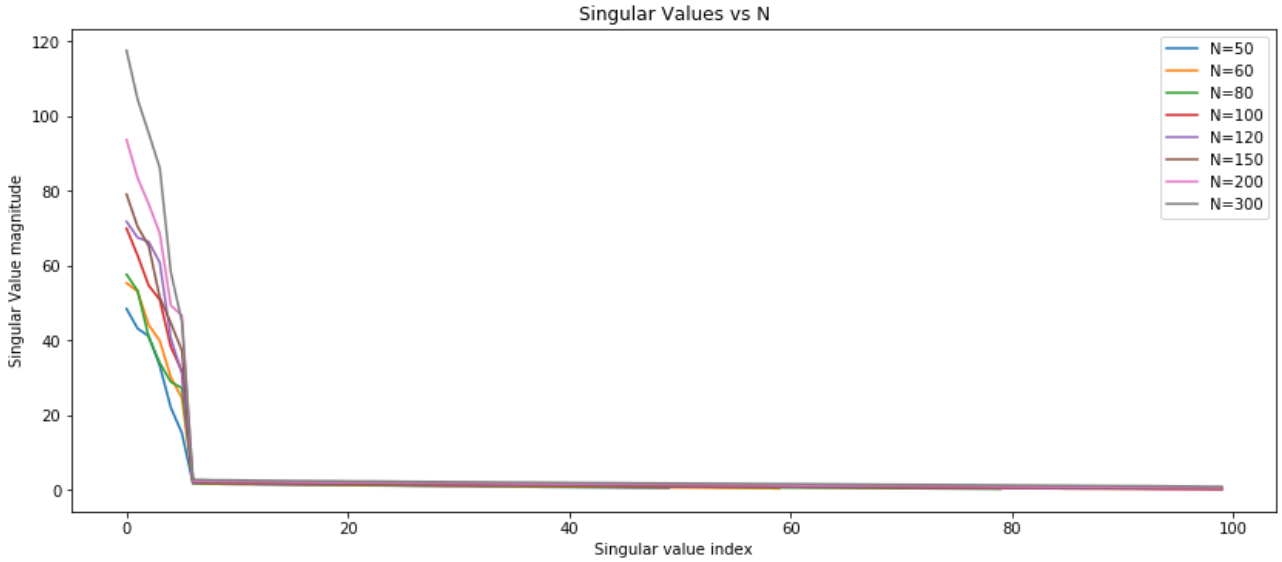
1. (a) d0 = 6 are the dominant singular values. After 6 the values are extremely low and effectively do not contribute a lot to the ddataset.

When we run a comparison for the standard deviation values. We see that when considering the 6 dominant values there is a minor error in

Reduced this Matrix from = (2000 × 100) to = (2000, 6)

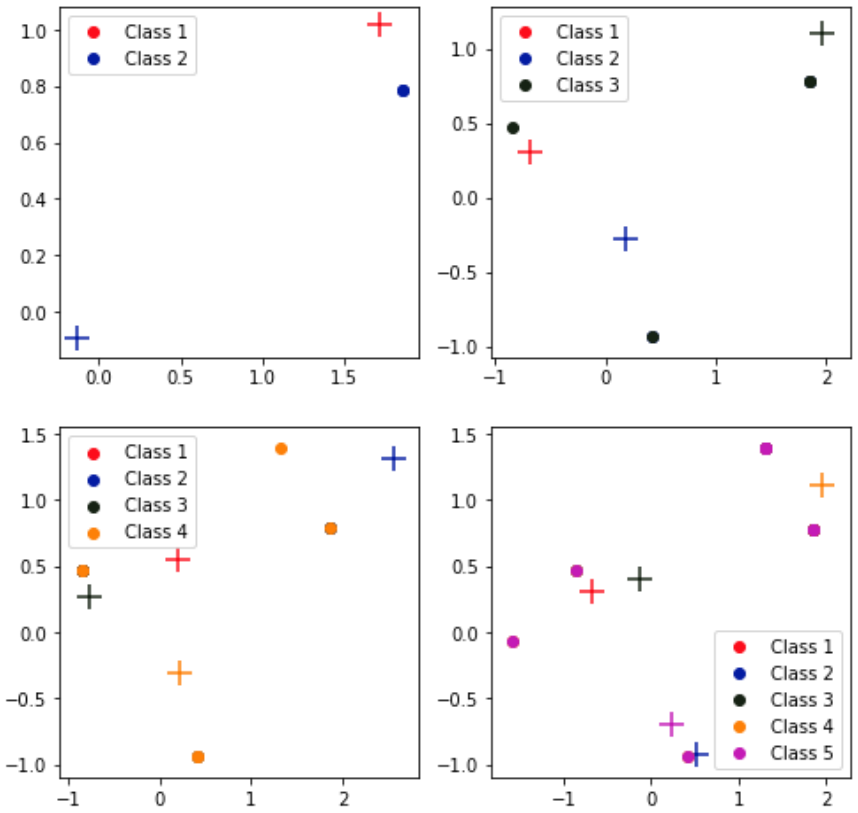
- SD-inp: 1.2395057807174528
- SD-out: 1.235722754594632
- SD-Err: 0.09679696927444671

6 dominant singular values from Matrix (2000 × 100) : [282.76080712 266.71486558 242.47285716 229.01120717 161.32588682 133.6358575 5.38798619]



1. (b) Performed PCA and implemented KMeans for the dominant d0 components. We can see the probabilities of various components based on the cluster center plot.

K= 2
comp= 1 -> Prob=0.00 -> Prob=1.00
comp= 2 -> Prob=0.80 -> Prob=0.20
comp= 3 -> Prob=0.00 -> Prob=1.00
K= 3
comp= 1 -> Prob=0.00 -> Prob=1.00 -> Prob=0.00
comp= 2 -> Prob=0.30 -> Prob=0.00 -> Prob=0.70
comp= 3 -> Prob=0.49 -> Prob=0.51 -> Prob=0.00
K=4
comp= 1 -> Prob=0.00 -> Prob=0.00 -> Prob=0.00 -> Prob=1.00
comp= 2 -> Prob=0.51 -> Prob=0.45 -> Prob=0.04 -> Prob=0.00
comp= 3 -> Prob=0.00 -> Prob=0.00 -> Prob=0.56 -> Prob=0.44
K=5
comp= 1 -> Prob=0.00 -> Prob=0.50 -> Prob=0.13 -> Prob=0.00 -> Prob=0.37
comp= 2 -> Prob=0.30 -> Prob=0.00 -> Prob=0.00 -> Prob=0.70 -> Prob=0.00
comp= 3 -> Prob=0.46 -> Prob=0.00 -> Prob=0.54 -> Prob=0.00 -> Prob=0.00



Observing the Cluster Centers

2. Insights into how the cluster centers found by K-means relate to the d0-dimensional projections of the vectors {uj} in the model

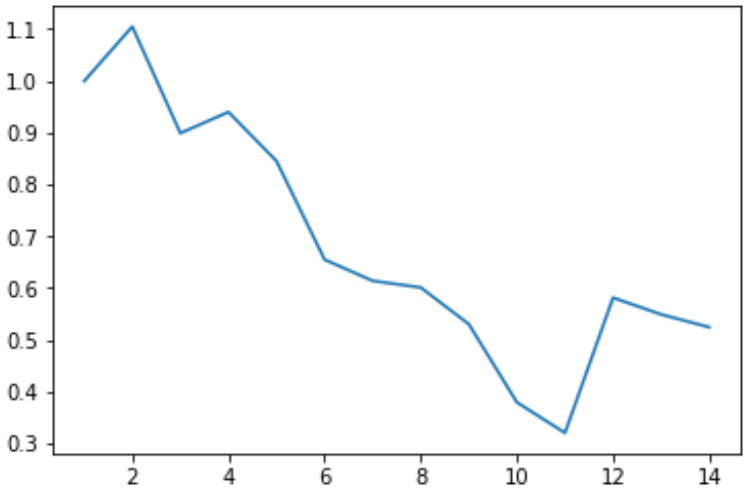
The mean of data points from component 1 is the projection of u1 into m = 6 dimension, and the mean data points from component u2 is the projection of 2 × u4 to the m = 6 dimension, and for the data points from component three, it is the projection of (2)u6 to the m = 6 dimension. Thus, for every cluster, the centers found by K-means is the weighted average of the means of each component in lower dimension, based on how many points of that component are in the cluster.

Part II: Random Projections & Compressed Sensing

3. Generated the dataset with the following variable shapes & values

```
N = 2000,
M = 20,
Y_xp  : (2000, 20),
Labels: (2000, 3),
Phi   : (20, 100),
B     : (100, 7)
```

4. Find a sparse reconstruction of s based on y using Lasso. The minimum M is 11



MSE vs M values

5. normalized MSE over many draws, with reconstruction performance.

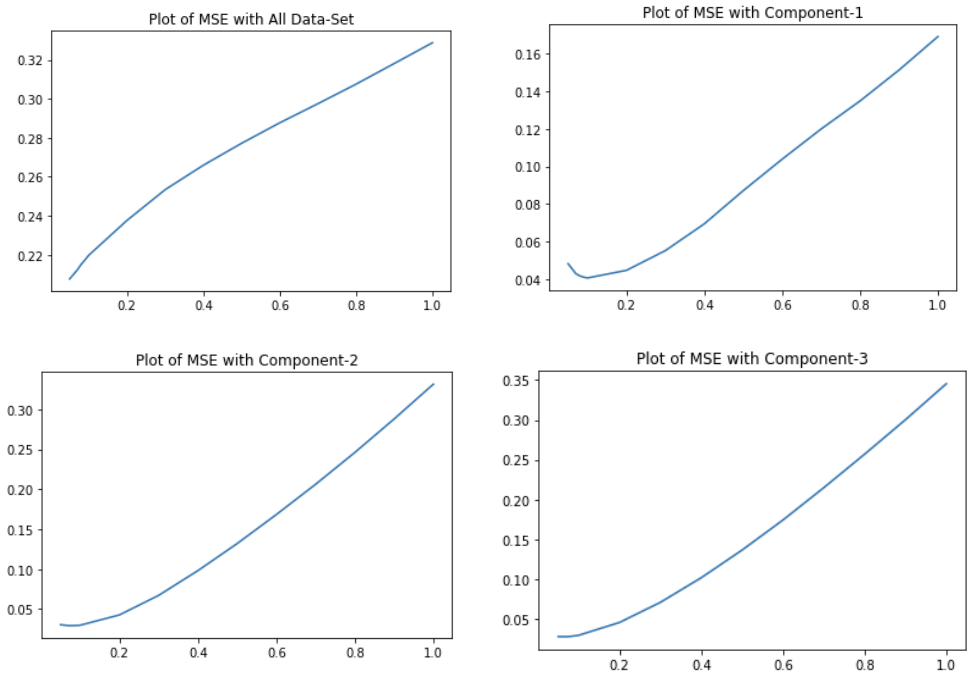


Fig: We can see the mean square error increases as the lambda value increase from 0 to 1

There was another case in which it did not perform any good irrespective of the lambda value and the plot of that is available in the code logs.

6. Down projected the data to 6 dimensions with m =11

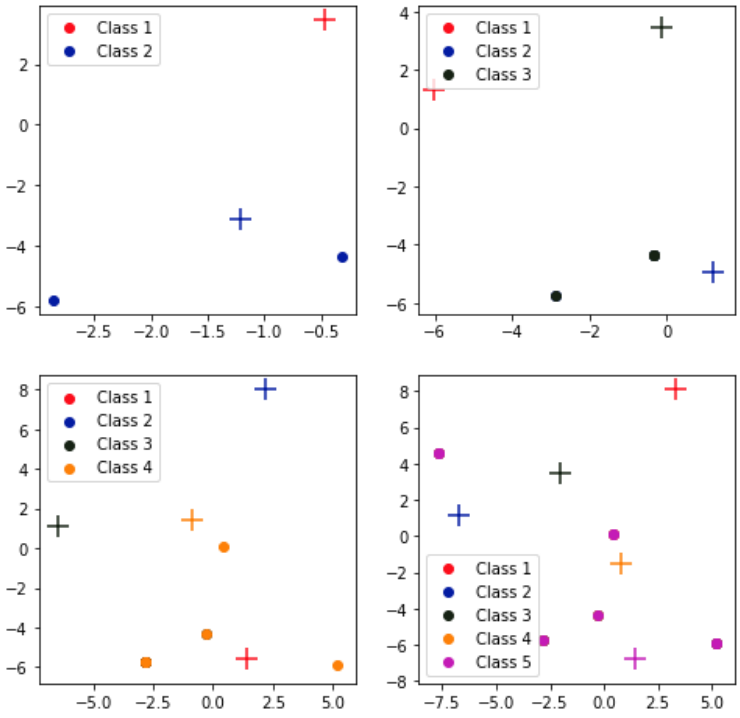
The projected distance are mentioned below in both the spaces. We can see that it did maintain some consistency across the values however not accurate.

```
projected_dist_matrix
        [[  0.    37.45  56.36  73.45  23.64  46.82]
        [ 37.45   0.    66.91  66.55  34.91  87.55]
        [ 56.36  66.91   0.   131.27  59.64 114.45]
        [ 73.45  66.55 131.27   0.    70.18  67.18]
        [ 23.64  34.91  59.64  70.18   0.    43.18]
        [ 46.82  87.55 114.45  67.18  43.18   0.  ]]
dist_matrix
        [[ 0.  2.  2.  7.  3.  4.]
        [ 2.  0.  2.  3.  3.  2.]
        [ 2.  2.  0.  7.  3.  4.]
        [ 7.  3.  7.  0. 10.  7.]
        [ 3.  3.  3. 10.  0.  7.]
        [ 4.  2.  4.  7.  7.  0.]]
```

7. K-means algorithm post-projection

Dataset Dimension: (2000, 11)

K= 2
comp= 1 -> Prob=0.69 -> Prob=0.31
comp= 2 -> Prob=0.63 -> Prob=0.37
comp= 3 -> Prob=0.27 -> Prob=0.73
K= 3
comp= 1 -> Prob=0.00 -> Prob=0.31 -> Prob=0.69
comp= 2 -> Prob=0.43 -> Prob=0.02 -> Prob=0.55
comp= 3 -> Prob=0.08 -> Prob=0.66 -> Prob=0.26
K= 4
comp= 1 -> Prob=0.23 -> Prob=0.03 -> Prob=0.00 -> Prob=0.74
comp= 2 -> Prob=0.01 -> Prob=0.30 -> Prob=0.39 -> Prob=0.30
comp= 3 -> Prob=0.57 -> Prob=0.01 -> Prob=0.06 -> Prob=0.36
K= 5
comp= 1 -> Prob=0.00 -> Prob=0.00 -> Prob=0.47 -> Prob=0.43 -> Prob=0.10
comp= 2 -> Prob=0.25 -> Prob=0.37 -> Prob=0.28 -> Prob=0.10 -> Prob=0.00
comp= 3 -> Prob=0.01 -> Prob=0.05 -> Prob=0.13 -> Prob=0.41 -> Prob=0.40



Observing the Cluster Centers

8. geometric insight. cluster centers found by K-means relate to the m=11 dimensional projections of the vectors

K-means seeks to represent all n data vectors via small number of cluster centroids so K-means can be seen as a highly-sparse PCA.

The mean of data points from component 1 is the projection of u1 into m = 11 dimension, and the mean data points from component u2 is the projection of 2 × u4 to the m = 11 dimension, and for the data points from component three, it is the projection of (2)u6 to the m = 11 dimension. Thus, for every cluster, the centers found by K-means is the weighted average of the means of each component in lower dimension, based on how many points of that component are in the cluster.