Problem Set III Gustavo Grinsteins CSE 326

Gaussian Mixture Model (30%)

1. (12%) Plot the complete (in red) and incomplete (in blue) log likelihoods for 6 runs of GMM with 5 clusters. As a verification of implementation, you should see that the incomplete log likelihood is always increasing. Moreover, the complete should lower-bound the incomplete (i.e., the blue line should always be above the red line).

   (**Gustavo's Answer:**)

   For the plot of the complete and incomplete log likelihoods for 6 runs of GMM with 5 clusters refer to **Figure 1**.

   (2%) Is there a significant difference between the incomplete log likelihoods (ILL) and complete log likelihood (CLL)?

   (**Gustavo's Answer:**)

   The greatest difference in magnitude is of 214.2 and the minimum is of 93.7 for the log likelihoods for the graph plotted in **figure 1**. The average magnitude for the curves are around $3.18 * 10^4$, therefore these differences in values are relatively small. As the iterations increase, both curves seem to approach the same value making this difference even smaller.

   (2%) Do the CLLs seem to converge roughly to the same value across all runs?

   Yes, both the CLL and ILL curve seem to be approaching a value around $3.21 * 10^4$ for the log likelihood (obtained from running 50 iterations at k = 5).

2. (12%) Plot the GMM results with different number of clusters from 1 to 16. This includes: (1) all means estimated by GMM; (2) shows the incomplete log likelihood as a function of number of clusters.

   (**Gustavo's Answer:**)

   (1) Please refer to **figures 2-8**.

   (2) Please refer to **figure 9**.

   Discuss what seems to be to be a good number of clusters for this data according to your experimental results ?

   (**Gustavo's Answer:**)

   In **figure 5** for k = 8 we can see that the numbers can be seen clearly. At k = 10 we can see the numbers but they are more blurred. So a good number of clusters would be from 8 to 10.

3. (2%) How the means found by GMM for different values of K are different from K-means algorithm? How big does K have to be before you essentially see a mean for each true digit?

   (**Gustavo's Answer:**)

   The means found by gmm are more blurred than knn because of soft labeling. The clusters are defined by percentages rather than set yes or no values. Yet, the numbers for most of the clusters can be distinguished as actual digits. I believe you need a minimum of k = 10 in order to have 10 clusters that would ideally represent each label. Some digits are over-represented due to the fact that some numbers are written in different ways which would generate separate clusters for the same number.

Principal component analysis

1. (15%) Complete the program PCA provided in pca.m and test your implementation on MNIST dataset. Results listed below are required to be included in your report:

   A plot of all eigenvalues.

   (**Gustavo's Answer:**)

   Please refer to **figure 2** for the plot

   How many numbers of principal components would you choose to achieve at least 90% of the data variance? Include a plot of (number of principal components) vs. (accumulated data variance).

   (**Gustavo's Answer:**)

   Please refer to **figure 3** for the plot. Note that about 70 principal components are needed in order to account for 90% of the variance

   Plots of the first 10 eigenvectors (same dimension as digit images)

   (**Gustavo's Answer:**)

   Please refer to **figure 4** for the plot.

2. (5%) Test the Gaussian mixture model implementation for 5 classes again by using the low dimensional data constructed by PCA. Report plots of means and mixing coefficients π.

   (**Gustavo's Answer:**)

   Please refer to **figures 13-14** for the plots.

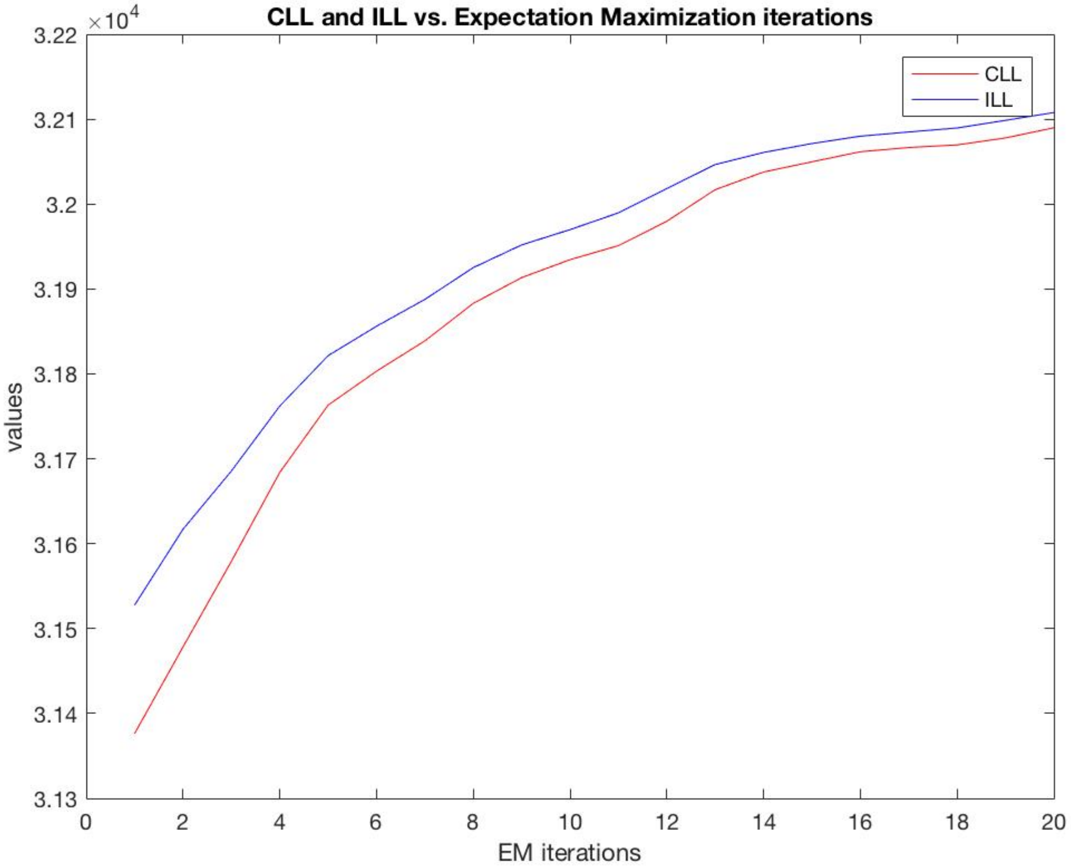

Figure 1: Plot of the complete and incomplete log likelihoods for 6 runs of GMM with 5 clusters refer

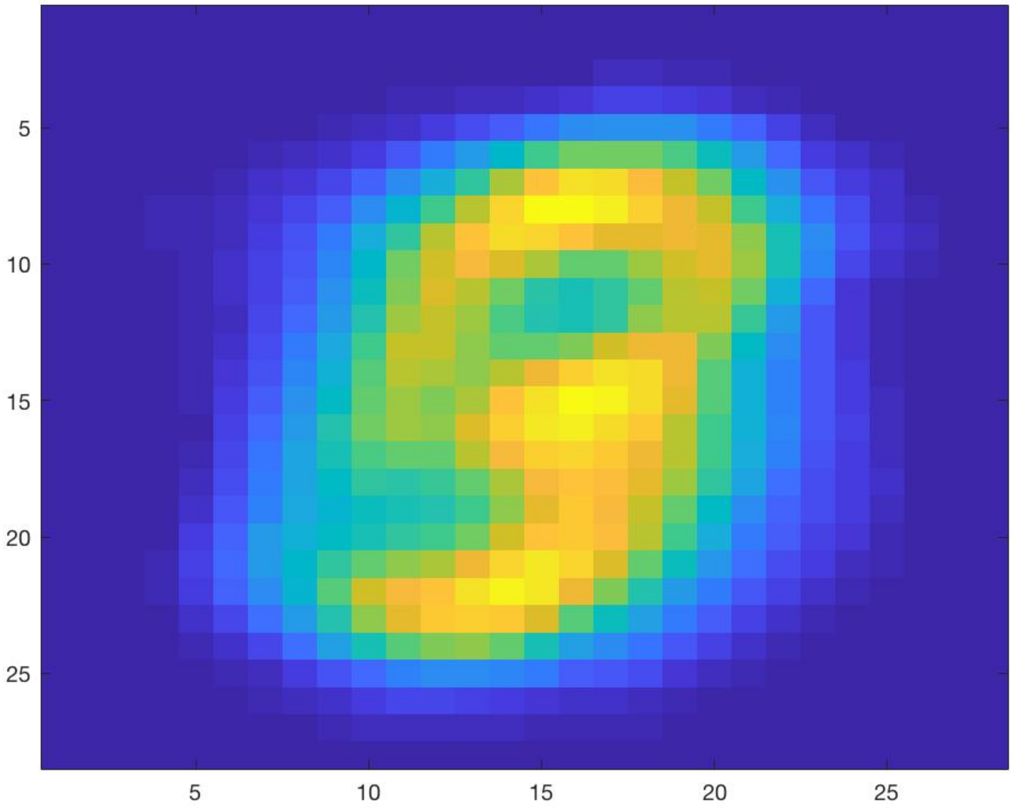

Figure 2: Plot of mean estimated by gmm at $k = 1$ and EM iteration $= 20$

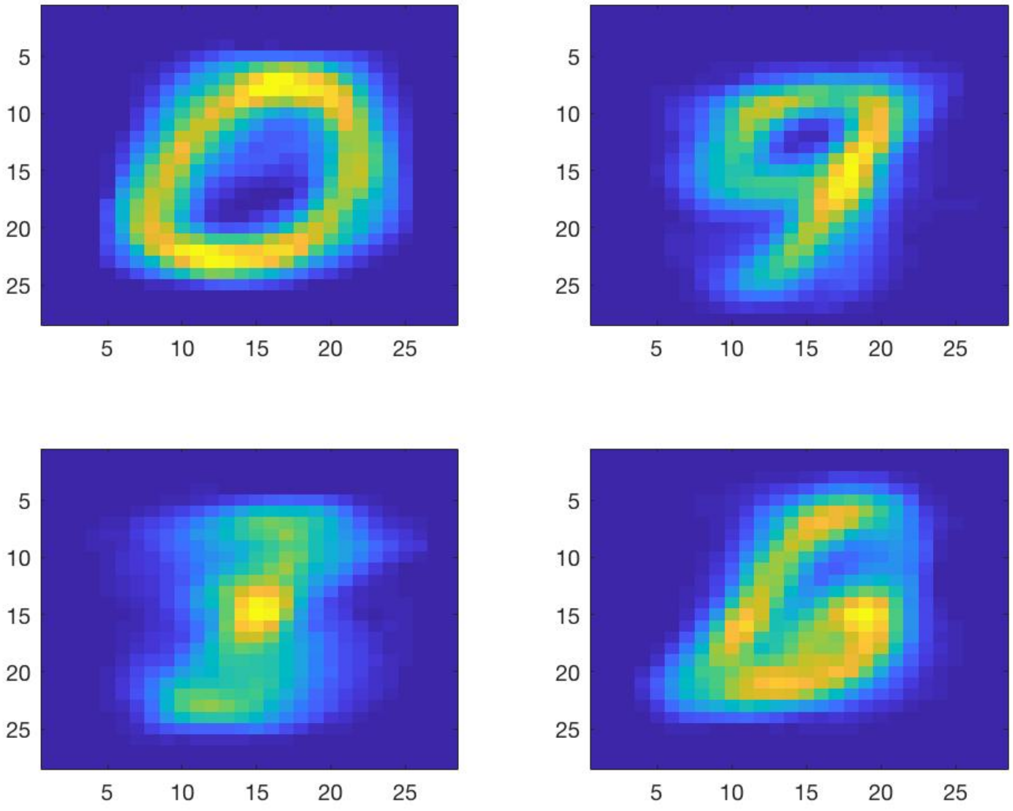

Figure 3: Plot of mean estimated by gmm at $k = 4$ EM iteration $= 20$

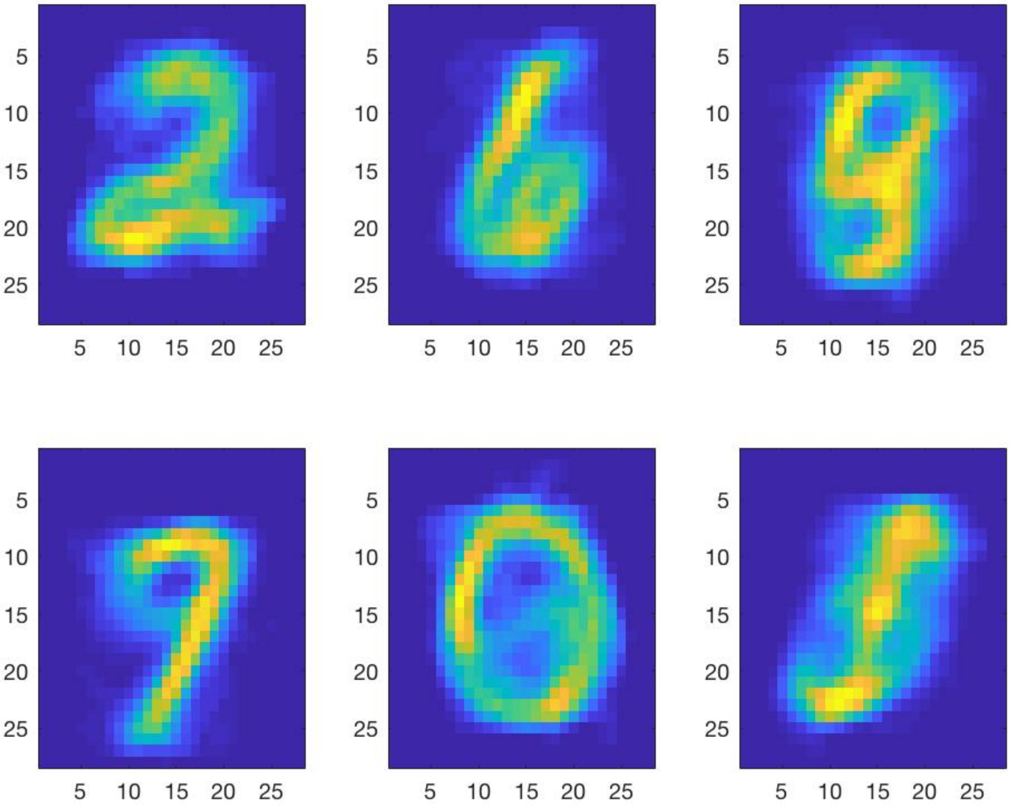

Figure 4: Plot of mean estimated by gmm at $k = 6$ EM iteration $= 20$

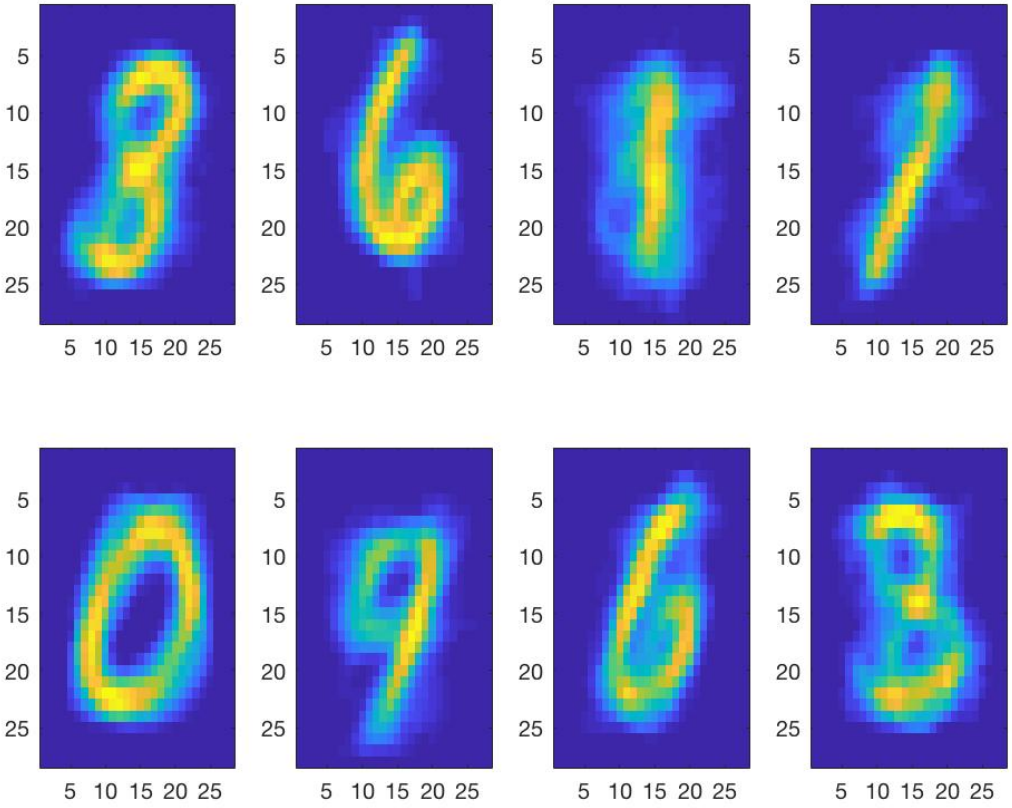

Figure 5: Plot of mean estimated by gmm at $k = 8$ EM iteration $= 20$

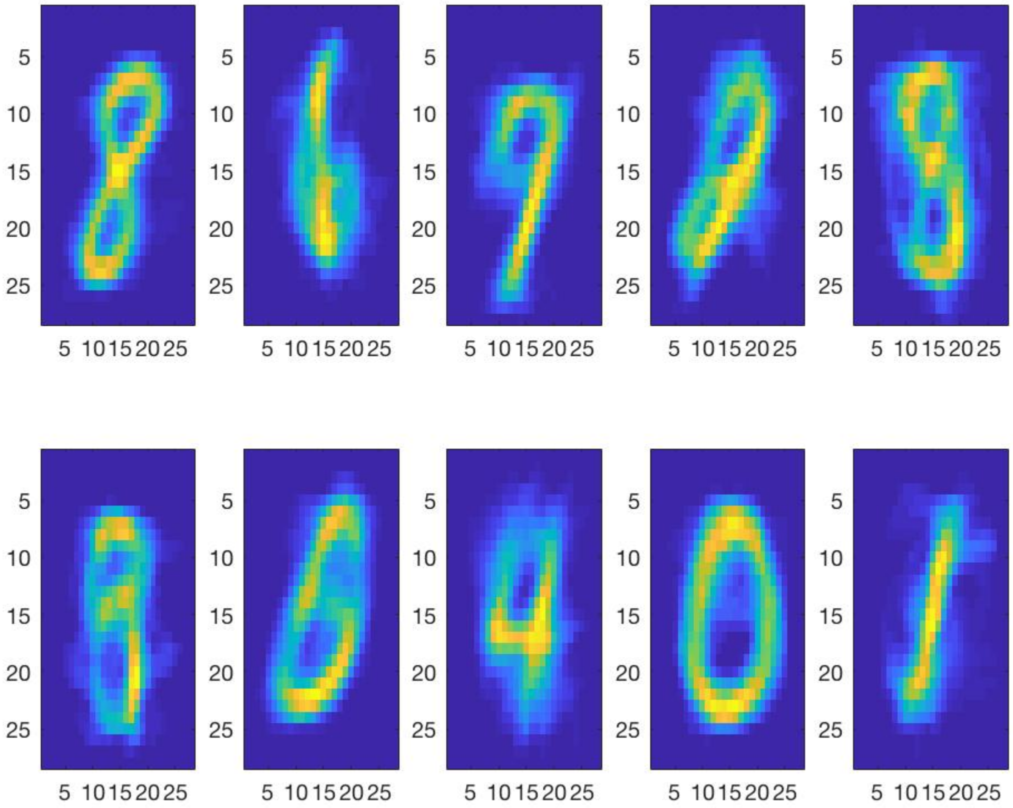

Figure 6: Plot of mean estimated by gmm at $k = 10$ EM iteration $= 20$

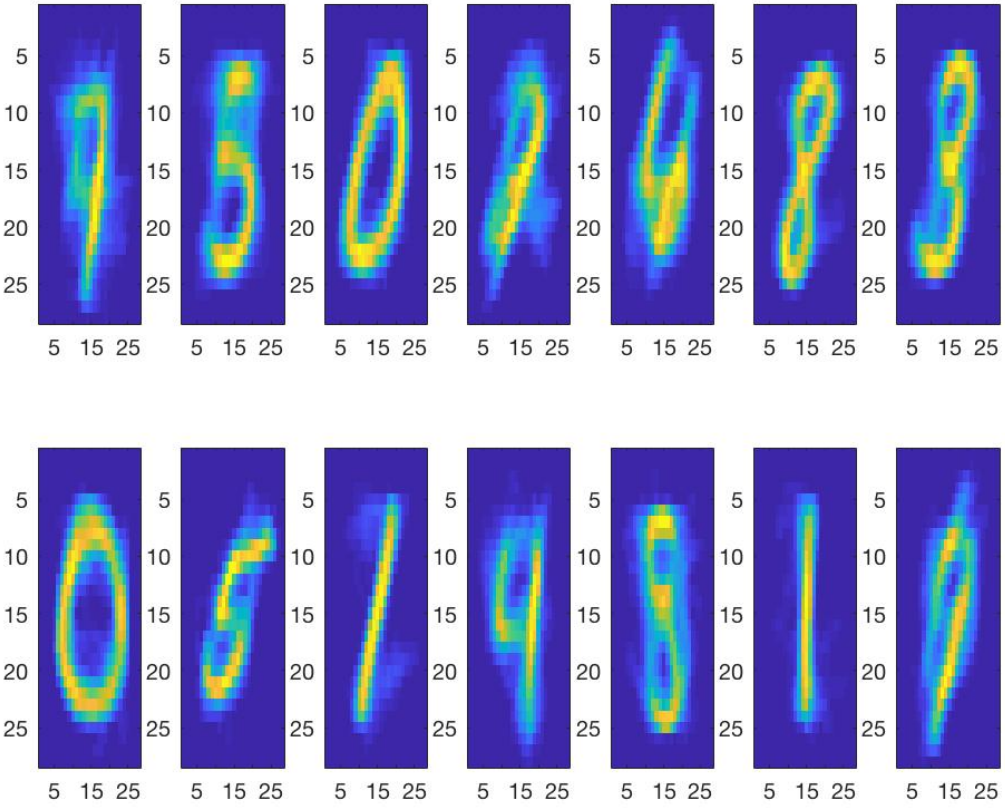

Figure 7: Plot of mean estimated by gmm at $k = 14$ EM iteration $= 20$

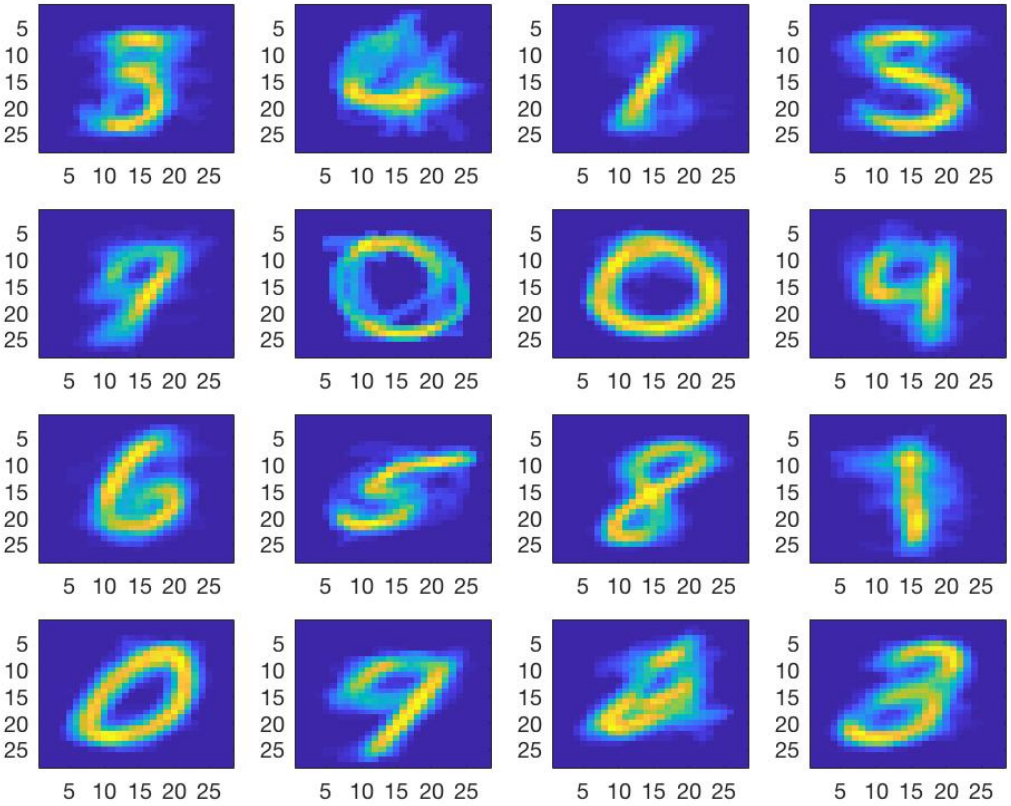

Figure 8: P

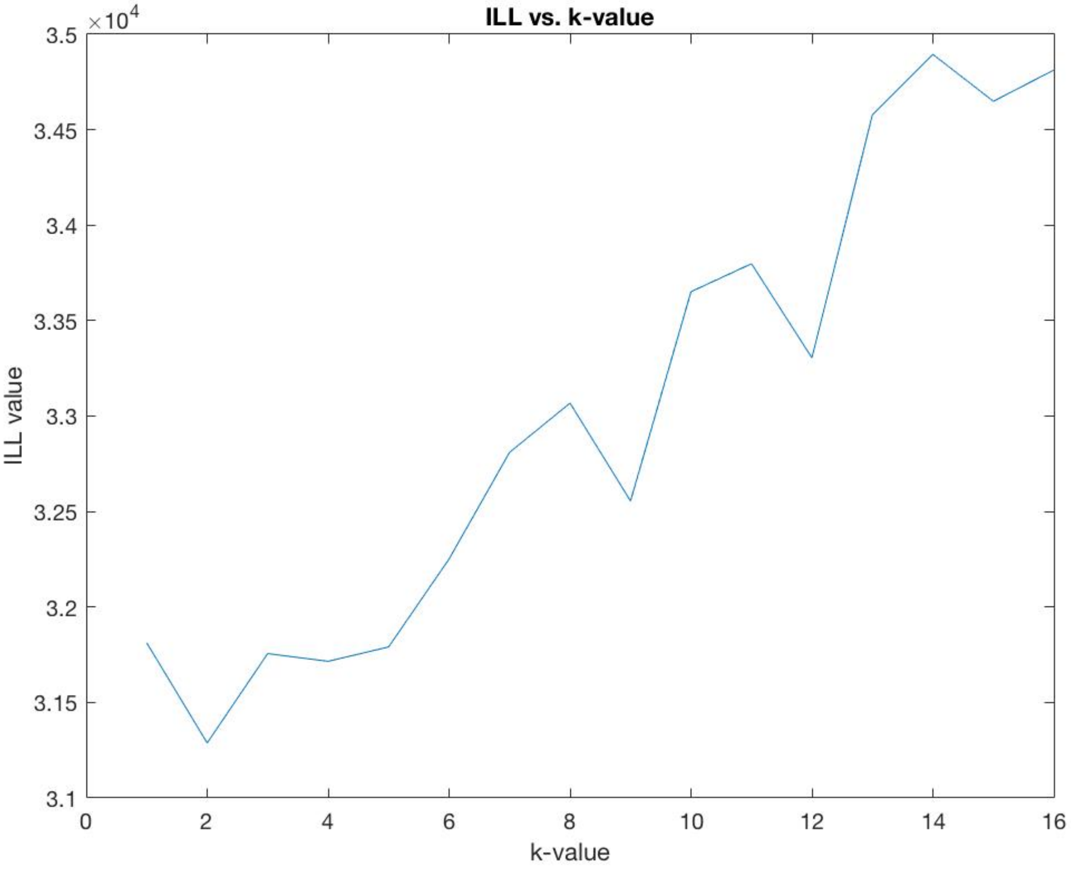

Figure 9: incomplete log likelihood as a function of number of clusters

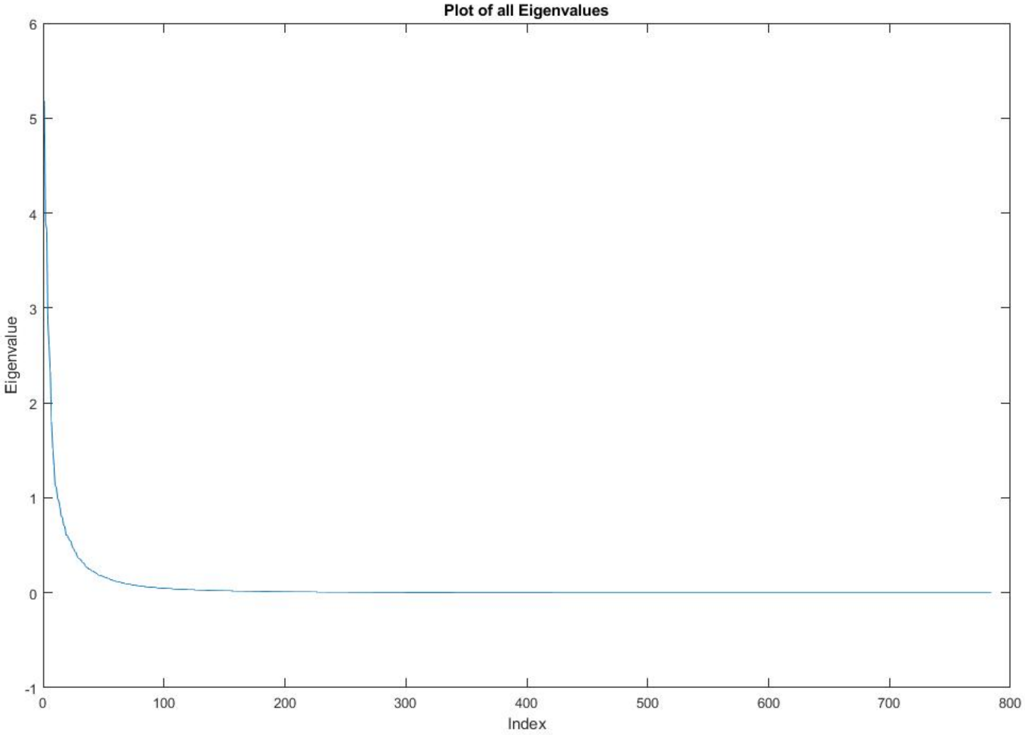

Figure 10: Plot of all Eigenvalues

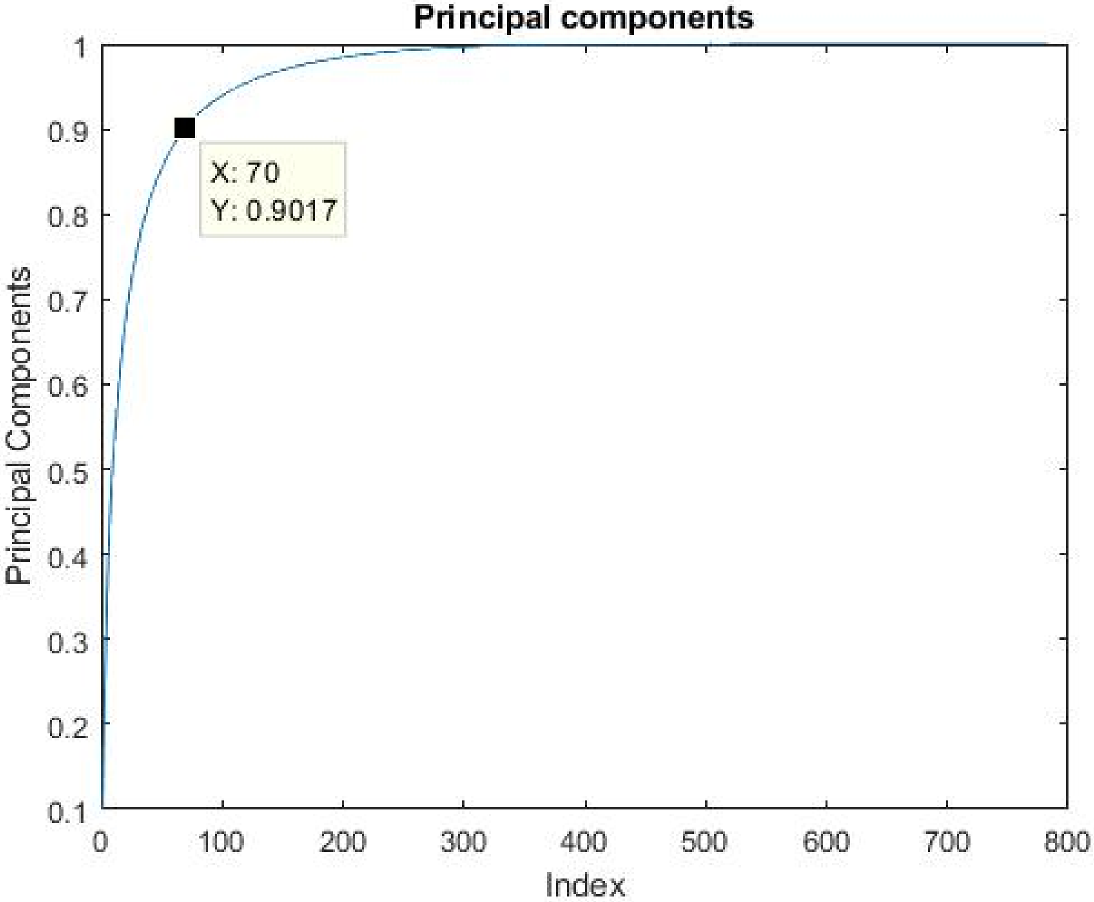

Figure 11: As seen from this graph, about only the 70 principal components with highest corresponding eigen values are needed in order to account for 90% of the variance

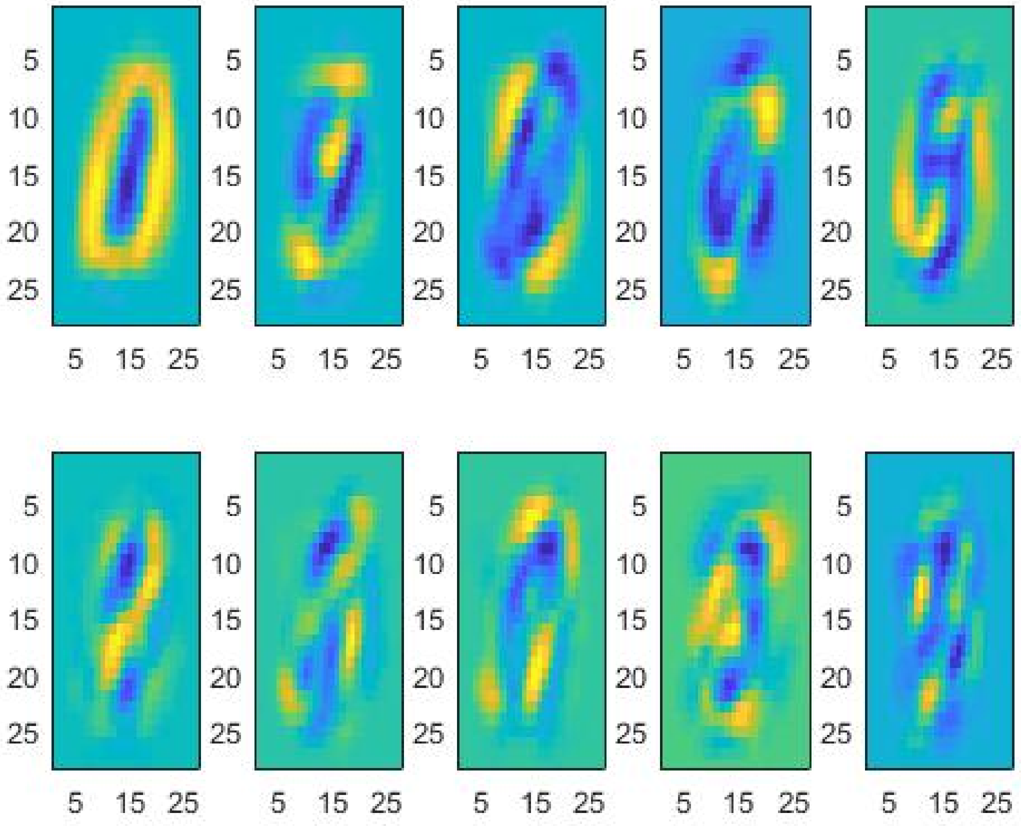

Figure 12: Plot of the first 10 eigen values

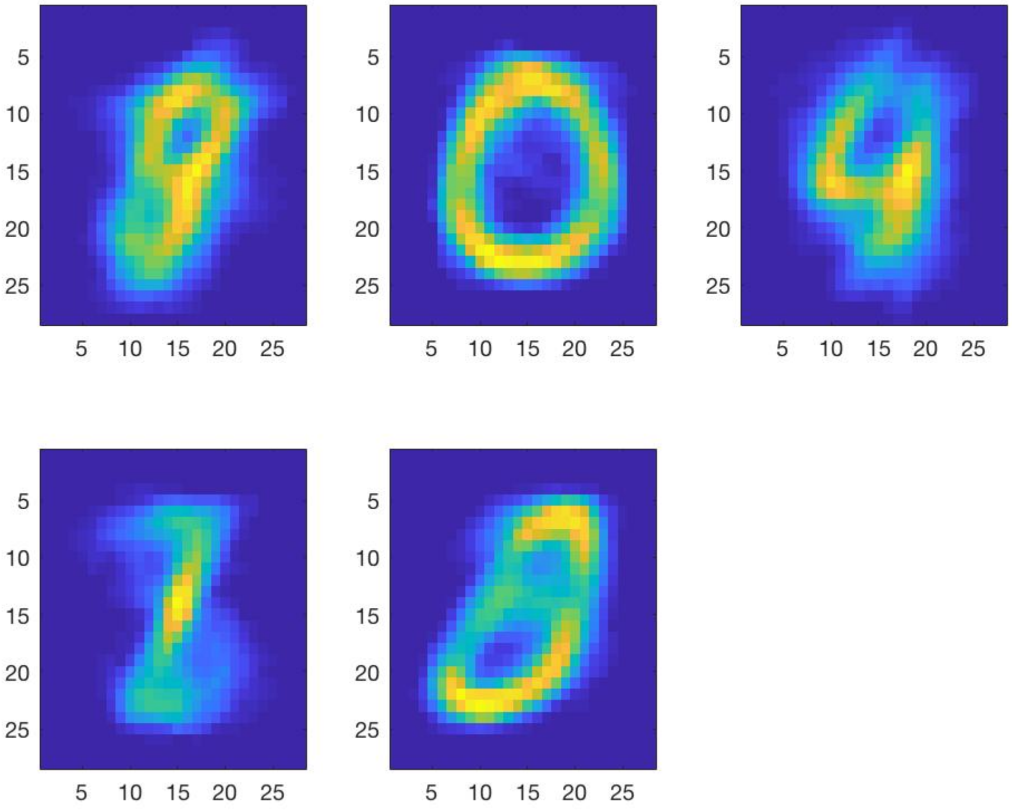

Figure 13: Report of plots of means

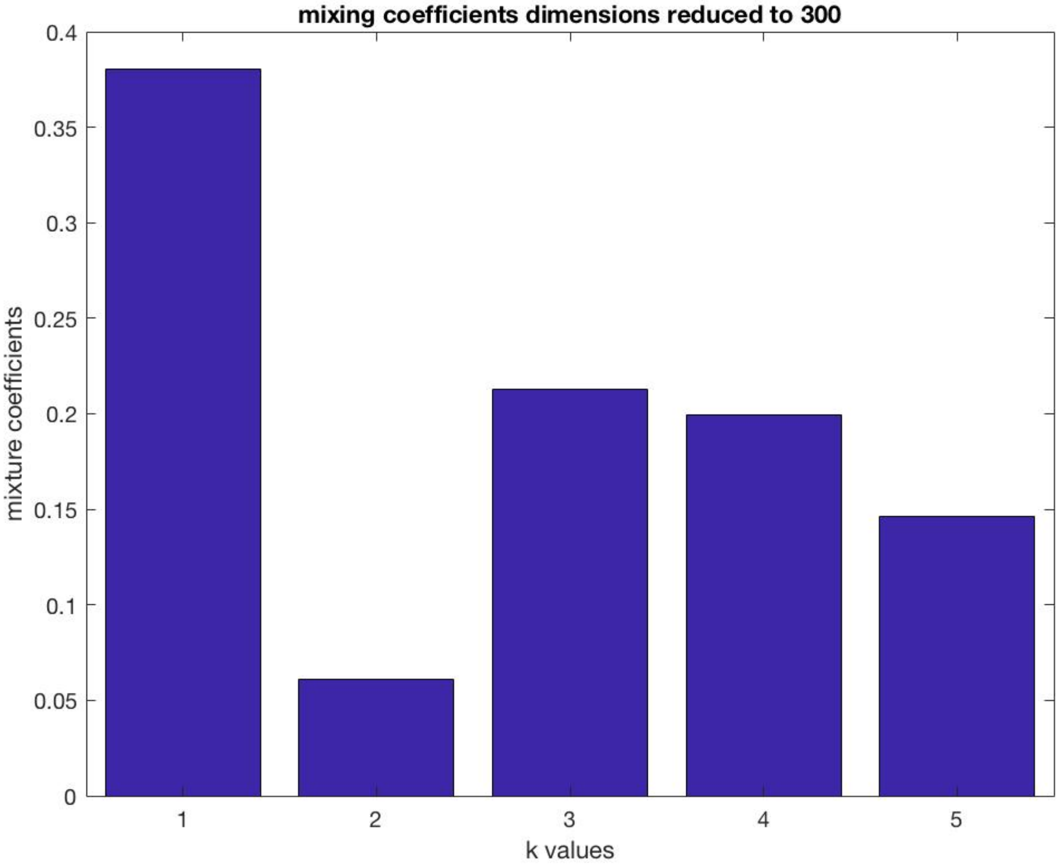

Figure 14: mixing coefficients π vs. K values