1 INTRODUCTION

The purpose of this study is to perform soft clustering using EM algorithm on “Semeion Handwritten Digits” dataset and find out the value of “Number of principal components” which minimizes AIC. Subsequent part of the study includes visualizing the clusters and also perform accuracy assessment.

2 DATA

The Semeion Handwritten Digits dataset consists of 1593 handwritten digits from around 80 persons. The digits were scanned and stretched in a rectangular box 16x16 in a gray scale of 256 values. Then each pixel of each image was scaled into a boolean (1/0) value using a fixed threshold. It thus has 1593 rows and 256 columns.

3 EXPECTATION-MAXIMIZATION ALGORITHM

Step 1: Initialization:

Choose initial values for the parameters μ_k , Σ_k, and π_k using kmeans in R with several random starts and setting γ_{ik} = 1 if observation i is assigned to cluster k and zero otherwise. Also choose initial parameters using M-step.

Step 2: E-Step

Compute the class membership distribution conditional on the current parameters and the data, given in (2).

$$p(z_{ik} = 1|\mathbf{x}_i) = \gamma_{ik} = \frac{\pi_k p_{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k}(\mathbf{x}_i)}{\sum_{k'=1}^{K} \pi_{k'} p_{\boldsymbol{\mu}_{k'},\boldsymbol{\Sigma}_{k'}}(\mathbf{x}_i)}, \tag{2}$$

where

$$p_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}.$$

Step 3: M-Step

Given the current class membership distribution, compute the parameter estimates for μ_k and π_k, given in (3), and the rank-q plus noise estimate for Σ_k, given in (4).

$$\widehat{\boldsymbol{\mu}}_k = \frac{1}{N_k}\sum_{i=1}^{n} \gamma_{ik}\mathbf{x}_i \quad \text{and} \quad \widehat{\pi}_k = \frac{N_k}{n}, \tag{3}$$

where $N_k = \sum_{i=1}^{n} \gamma_{ik}$.

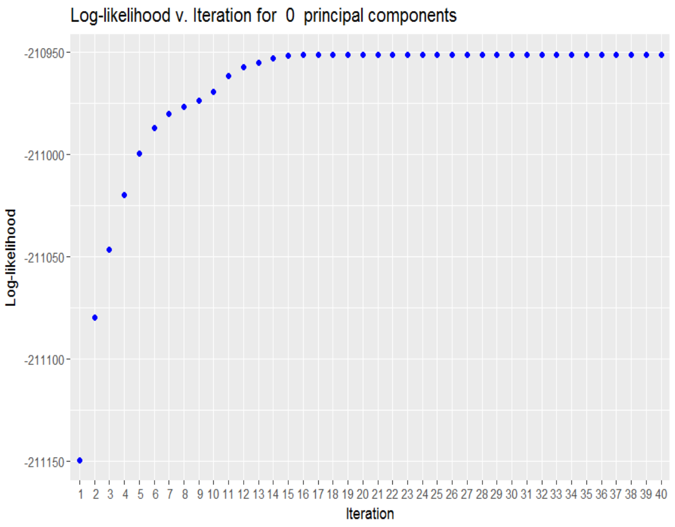
$$\widehat{\boldsymbol{\Sigma}}_k = \widehat{\mathbf{W}}_q \widehat{\mathbf{W}}_q' + \widehat{\sigma}^2 \mathbf{I}_d, \tag{4}$$

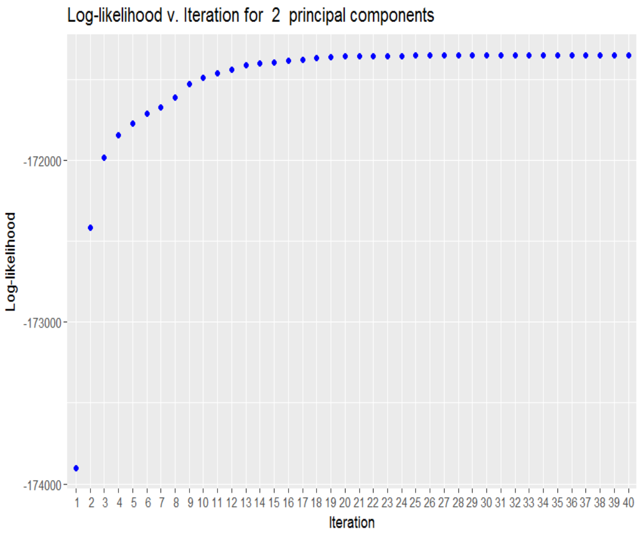
Step 4:

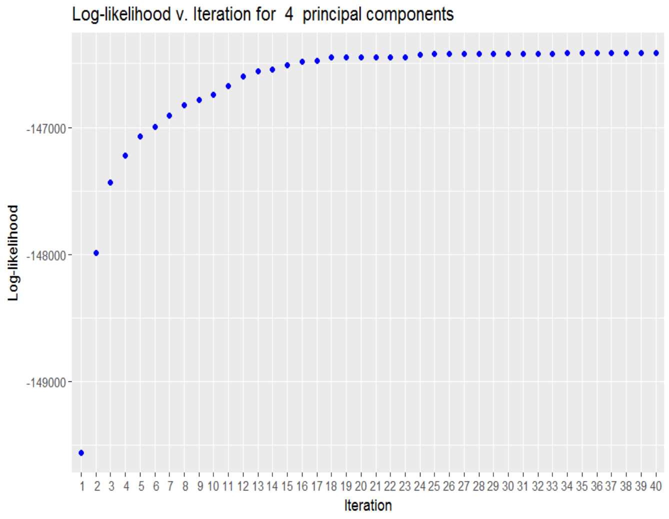
Repeat steps 2 and 3 until convergence of data log-likelihood, for this study we have used 40 iterations.

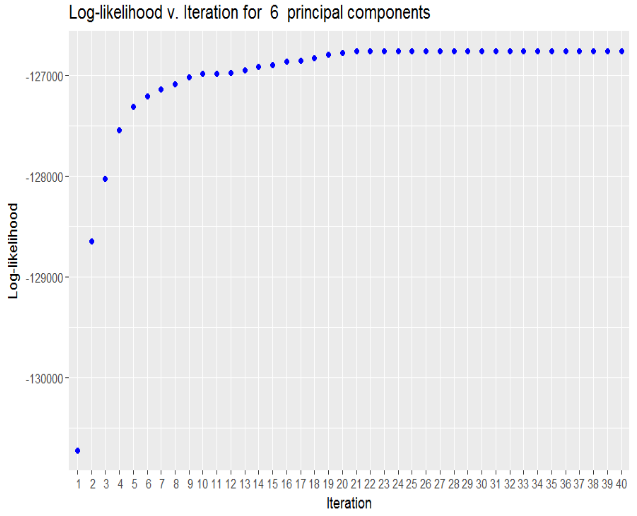
4 Convergence of Log-likelihood

Below are the plots of log-likelihood vs iteration for principal component equal to 0,2,4,6.









Following are the observations:

1. Maximum value of log-likelihood is observed for number of principal components equal to 6 i.e. q = 6.
2. Log-likelihood values become stable after 15-25 iterations for different values of q.

5 Choice of Number of Principle Components

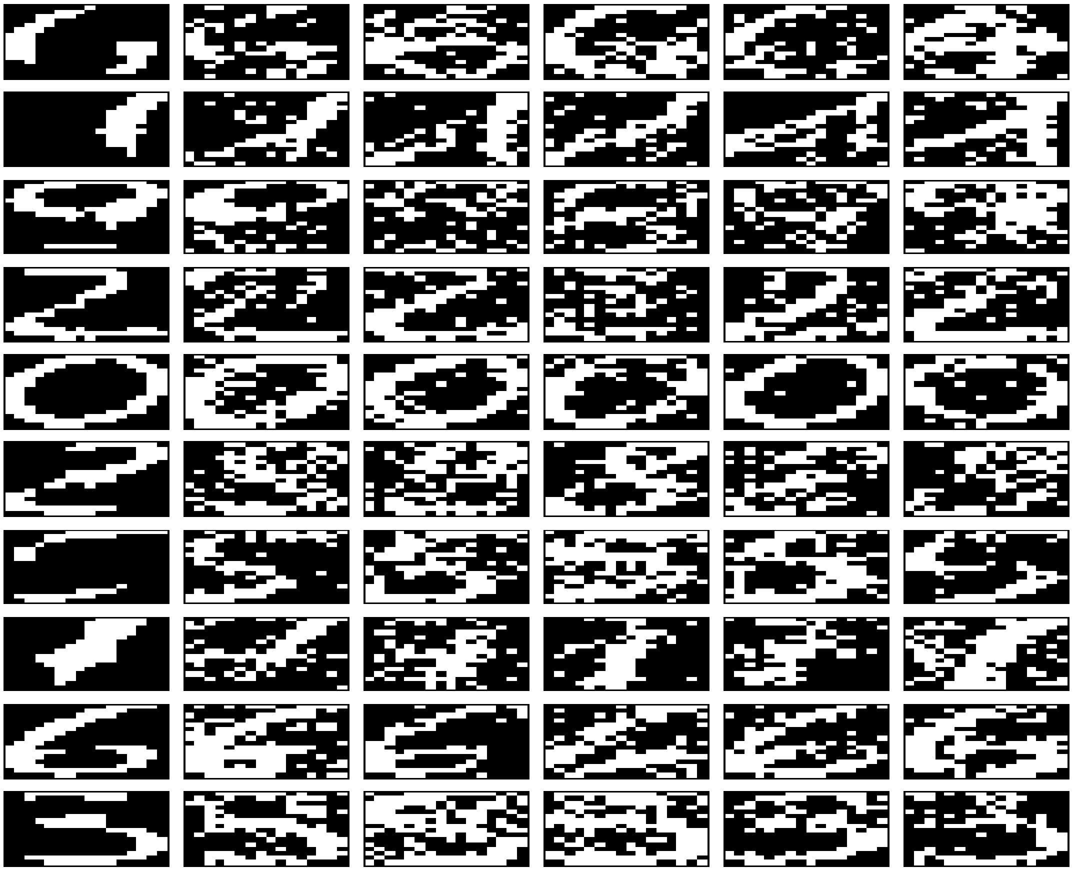
Below is the table for AIC values for different values of Principal Components:

# of principal components	AIC Value
0	421904
2	343723
4	294862
6	256561

According to the above values the least value of AIC is obtained for Number of principal components equal to 6. It can be concluded that for the given values of q the choice would be q=6. This algorithm should be run for more values of q in order to obtain the optimum number of principal components which minimizes the AIC.

6 VISUALIZATION OF CLUSTERS

Below is the visualization of clusters for q = 6. First column contains the visualized cluster means, the other 5 columns contain five random draws from the cluster-specific distribution, in order to see how well the clusters had been defined.



From above visuals it can be observed that the cluster with numeric digit 0 is most well defined. The other clusters may improve on increasing the number of principal components.

7 Accuracy Estimate

For each cluster the most frequent occurring digit is recognized and misclassification rate was calculated. The overall misclassification rate was found to be 30.8%. The results are summarized below:

Cluster	Most Common Digit	# of occurences of the most common digit	Total number of observations	Mis-classification Rate
Cluster 1	4	82	140	41.40%
Cluster 2	1	97	119	18.50%
Cluster 3	9	108	169	36.10%
Cluster 4	2	113	123	8.10%
Cluster 5	0	152	153	0.70%
Cluster 6	8	86	168	48.80%
Cluster 7	5	89	95	6.30%
Cluster 8	7	137	223	38.60%
Cluster 9	6	101	153	34.00%
Cluster 10	3	137	250	45.20%

As can be observed from the table, 0 was the most accurately categorized digit by the algorithm, followed by 2, 5, and 1 which can be seen in the cluster visualization also. The most difficult digits to categorize are 8 and 3. The misclassification rate can be reduced by increasing the number of principal components.