Homework I

ISyE 6740
Instructor: Ben Haaland
Due: Wednesday, February 10, 2017 10:05am **Late homework will NOT be accepted.**

Name ____

Use `R` to complete programming portions. Please hand in code used to generate results. However, do not hand in only raw computer output. Conclusions and interpretation of results are more important than printouts.

You are allowed to use any resources at your disposal for the homework (books, web sites, notes, etc.). However, you are not allowed to discuss the content of the homework with anyone (no students, no faculty, no friends, no internet forums, etc.) during the week that the class is working on it.

I have read the above statement and certify that I have not communicated about the content of the exam with anyone,

signed ___________________________________ (unsigned exams will not be accepted).

Consider the Semeion Handwritten Digits dataset ("semeion.csv" on T-square). Here, we will cluster the handwritten digits data using the EM algorithm with a principle components step within each maximization as described below (**An EM Algorithm for a Principle Components Gaussian Mixture Model**). Use $K = 10$. Please hand in all `R` code (**50 points**) used to generate results.

1. **Initialization (10 points):** Use `R`'s `kmeans` function with several random starts to build a preliminary clusterin. Set $\gamma_{ik} = 1$ if observation i is assigned to cluster k and $\gamma_{ik} = 0$ otherwise. Choose initial parameter estimates as outlined in (3) and (4) below.

2. **Convergence (10 points):** After each iteration (for each q which is considered), compute the observed data log-likelihood

$$\log p(\mathbf{X}|\widehat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \widehat{\pi}_k p_{\widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}_k}(\mathbf{x}_i) \right\}$$

   Generate a plot of the observed data log-likelihood vs. iteration number (4 plots, 1 for each q).

3. **Choice of Number of Principle Components, q (10 points):** For number of principle components, $q = 0, 2, 4, 6$, compute the AIC (up to an additive constant depending on the means and class memberships) at convergence,

$$\text{AIC}(\widehat{\boldsymbol{\theta}}) = -2 \log p(\mathbf{X}|\widehat{\boldsymbol{\theta}}) + 2(\# \text{ of parameters}), \tag{1}$$

   where # of parameters can be taken as $dq + 1 - q(q-1)/2$. Note that $d = 256 = 16 \times 16$. Choose the value of $q = 0, 2, 4, 6$ which *minimizes* the AIC.

4. **Visualization of Clusters (10 points):** Make a $K = 10$ by 6 panel plot with the following entries. In column 1, plot the cluster means. In each of the columns 2 through 6, plot 1 random draw from the cluster-specific distribution (for a total of 5 random draws)

$$\mathbf{x}_{k,\text{new}} \sim \mathcal{N}(\widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}_k).$$

   Comment on the quality of clustering.

5. **Accuracy Assessment (10 points):** Compare the class labels to the clusters. For each class label, what is the mis-categorization rate (define an observation as mis-categorized if it is not in the most common categorization for the particular class)? What is the overall mis-categorization rate?

An EM Algorithm for a Principle Components Gaussian Mixture Model

1. **Initialization:** Choose initial values for the parameters $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, and π_k (subject to $\sum_{k=1}^{K} \pi_k = 1$).

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

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

3. **M step:** Given the current class membership distribution, compute the parameter estimates for $\boldsymbol{\mu}_k$ and π_k, given in (3), and the *rank-q plus noise* estimate for $\boldsymbol{\Sigma}_k$, 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 $N_k = \sum_{i=1}^{n} \gamma_{ik}$.

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

where

$$\widehat{\mathbf{W}}_q = \mathbf{V}_q\text{diag}\left\{\sqrt{\lambda_1 - \widehat{\sigma}^2}, \ldots, \sqrt{\lambda_q - \widehat{\sigma}^2}\right\}, \quad \widehat{\sigma}^2 = \frac{1}{d-q}\sum_{i=q+1}^{d} \lambda_i,$$

and $\mathbf{V}_q$ are the first q eigenvectors of the spectral decomposition $\mathbf{V}\text{diag}\{\lambda_1, \ldots, \lambda_d\}\mathbf{V}' = \frac{1}{N_k}\sum_{i=1}^{n} \gamma_{ik}(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)'$, $\lambda_1 \geq \ldots \geq \lambda_d$.

4. Repeat steps 2 and 3 until convergence of data log-likelihood.