Computer experiment 3

1. Download the MNIST data set of handwritten digits from: https://www.kaggle.com/datasets/oddrationale/mnist-in-csv
   (a) Create a data matrix $\mathbf{X} = [\mathbf{x}_1 - \boldsymbol{\mu}_\mathbf{x}, \mathbf{x}_2 - \boldsymbol{\mu}_\mathbf{x}, \dots, \mathbf{x}_N - \boldsymbol{\mu}_\mathbf{x}] \in R^{d \times N}$ from $N = 2000$ randomly sampled instances of the digit "3" from the training dataset (mnist_train.csv). The dimension of each instance is $d = 784$.
   (b) Compute the principal components of the covariance matrix of $\mathbf{X}$ as well as the corresponding variances (eigenvalues). Please refer to P. 26 of Ch6 slides to show similar results (e.g., the PCA bases and the dimension-reduced reconstructions)
   (c) Evaluate the reconstruction error with different settings of the reduced dimension $l = 1, 10, 50, 250$ and draw your conclusion.

(EM algorithm for Gaussian mixtures)

2. Generate 150 2-D samples from one Gaussian $N(\mu_1, \Sigma_1)$, another 300 samples from $N(\mu_2, \Sigma_2)$, and 100 samples from $N(\mu_3, \Sigma_3)$, with mean vectors $\mu_1 = \begin{bmatrix}1\\1\end{bmatrix}$, $\mu_2 = \begin{bmatrix}5\\5\end{bmatrix}$, $\mu_3 = \begin{bmatrix}9\\1\end{bmatrix}$ and covariance matrices $\Sigma_1 = \begin{bmatrix}1 & 0.4\\0.4 & 1\end{bmatrix}$, $\Sigma_2 = \begin{bmatrix}1 & -0.6\\-0.6 & 1\end{bmatrix}$, $\Sigma_3 = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$, respectively.
   (a) Use EM algorithm and the generated samples to estimate the unknown parameters $\boldsymbol{\mu}_i, \Sigma_i, P_i$ $(i = 1,2,3)$. Please specify your experimental settings (e.g., initialization, stopping criterion) in the report.
   (b) Repeat the mixture density estimation by EM when the mean vectors are $\begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}2\\2\end{bmatrix}, \begin{bmatrix}3\\1\end{bmatrix}$, respectively.
   (c) Compare the results (in terms of confusion matrices or 2-D visualization) and draw your conclusions.

(K-means algorithm)

3. Use K-means algorithm on the two data sets you generate in Problem 2, for $K = 2,3,4$. Compare the results and draw your conclusions.