Homework #2

Instructions: While discussion with classmates is allowed and encouraged; please try to work on the homework independently and direct your questions to me.

Instructions: Please interpret your analysis results using concise and clear language and focusing on interesting findings. Remember to include your Python codes and only necessary Python outputs.

1. Suppose that the random variables X_1 and X_2 have the covariance matrix

$$\Sigma = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}$$

   (a) Determine the population principal components Y_1 and Y_2 for the above covariance matrix.

   (b) Calculate the proportion of the total population variance explained by the first principal component.

2. **Principal Component Analysis (PCA)**: Consider the data on optical recognition of handwritten digits, available from the UCI Machine Learning Repository:

   https://archive.ics.uci.edu/dataset/80/optical+recognition+of+handwritten+digits

   The following two files are needed: (a) optdigits.names (a description of the data set), and (b) optdigits.tra (the training set).

   (a) Load the training set optdigits.tra, which has sixty-four (p = 64) inputs plus the target variable that indicates the digit 0-9.

   (b) Remove any unary column (i.e., containing only one value)

   (c) Are there any missing values?

   (d) Data preprocessing: Check to see if the data is standardized. If not, standardize the data matrix X so that all variables are given a mean of zero and a standard deviation of one. Remember to exclude the target variable.

   (e) PCA:

      i. Run the PCA algorithm on the standardized data with number of components equal to the total number of columns in the data.

      ii. Plot the Proportion of Variance Explained (PVE) of each principal component (i.e., a scree plot) and the cumulative PVE of each principal component.

      iii. Write your observations on the variance explained by the principal components.

      iv. Create a scatter plot for the first two principal components and show the target class variable (i.e., digit number) with different symbols and colors. Write your observations on the plot.