Computer experiment 3

1. (Ch5 [Theodoridis 2009])
   - **a.** (i) Generate four sets, each one consisting of 100 two-dimensional vectors, from the normal distributions with mean values $[-10,\ -10]^T, [-10,\ 10]^T, [10,\ -10]^T, [10,\ 10]^T$ and covariance matrices equal to $0.2 * I$. These sets constitute the data set for a four-class two-dimensional classification problem (each set corresponds to a class).
   - **a.** (ii) Compute the $Sw, Sb,$ and Sm scatter matrices.
   - **a.** (iii) Compute the value for the criterion $J3$.
   - **b.** Repeat (a) when the mean vectors of the normal distributions that generate the data are $[-1,\ -1]^T, [-1,\ 1]^T, [1,\ -1]^T, [1,\ 1]^T$.
   - **c.** Repeat (a) when the covariance matrices of the normal distributions that generate the data are equal to $3 * I$.

2. The Fisher's discriminant ratio (FDR) is defined by:

$$FDR = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}.$$

FDR is sometimes used to quantify the separability capabilities of **individual** features.

   - **a.** (i) Generate two sets, each one consisting of 100 two-dimensional vectors, from the normal distributions with mean values $[2,\ 4]^T$ and $[2.5,\ 10]^T$ and covariance matrices equal to the 2×2 identity matrix I. Their composition forms the data set for a two class two dimensional classification problem (each set corresponds to a class).
   - **a.** (ii) Compute the value of the FDR index for both features.
   - **b.** Repeat (a) when the covariance matrices of the normal distributions that generate the data are both equal to $0.25 * I$.
   - **c.** Discuss the results.

3. 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 a certain digit from the training dataset. The dimension of each instance should be $d = 784$.
   (b) Compute the principal components of the covariance matrix of $\mathbf{X}$ as well as the corresponding variances (eigenvalues). Refer to P 27 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 N and/or the reduced dimension l. You may also compare the results when choosing different digits (e.g., "3" and "8") from MNIST in your experiments and draw your conclusion.