Assignment 2: Due April 6 via Canvas.

- Full marks for correct answers to all 4 questions. All are worth equal marks.
- Provide all your answers in a single pdf document (exported from Word/Pages/ or similar). Start each question on a separate page.
- Paste your R code into the document at the end of each question.
- Submit your the combined pdf document on Canvas with the submission name: Assignment2-uccstudentid-name.pdf (e.g. Assignment2-11xxxxxxx-FinbarrOSullivan.pdf). Start the submission process at least 15 minutes before the Canvas submission deadline. No late assignments will be accepted.

Question 1

For this question *set.seed(uccstudentno+2000)* where *uccstudentno* = 123119660. Consider the (NXT) Z-data matrix from question 5 of Assignment 1. The red, blue and green channel images are in a Waves.Rdata Use rbinom(), to simulate a Bernouilli sequence of length (NT) with constant success probability $p = 0.7$. Put values into a matrix U with the same dimension as Z. Create a new dataset, ZNA, in which any elements of Z associated with $U = 0$ are replaced by the missing-values - code these as NA in R. Report the number and proportion of cases (rows of ZNA) that are at least 10% and 25% complete. Compute a covariance-matrix, $\hat{\Sigma}_T$ for ZNA using pairwise complete observations. Use a plot to compare the fraction of variance in the complete dataset explained using eigenvectors from the complete (Σ_T) and incomplete ($\hat{\Sigma}_T$) data covariance.

Question 2

An array of scanned digital images (28×28) extracted from hand-written postal codes on domestic letter envelopes in the US postal service is provided in the *digits* component of the Digits.Rdata dataset. Images are approximately co-registered so that the orientation, centering and horizontal and vertical scale of images are quite similar. There are 1000 samples for each digit: *digits*[,,i,d] is the i'th sample image for the d'th digit - $i = 1, 2, ...1000; d = 0, 1, ...9$.

(a) Evaluate the spectral decomposition of the total sums of squares matrix ($TSS = X'X$, where X is the 10000×784 data matrix) and use the resulting matrix of eigenvectors, Γ, to transform X to create the data matrix $Z = X\Gamma$. In a (2x2) array plot (i) the first three columns of Z against eachother highlighting the 10 points corresponding to the means of the data from each digit, and (ii) a plot of the fraction of variance explained by each eigenvector as well as the cumulative amount of variance explained as a function of the number of eigenvectors considered.

(b) Using the 1000 sample images for each digit, use *hclust()* with a Euclidean distance matrix to obtain 5 clusters and calculate the corresponding images of the cluster means, $\bar{\mu}_{dc}, c = 1, 2, 3, 4; d = 0, 1, ...9$. For digits 2 and 9 create a (3x2) array of plots showing the clustering denogram, and the 5 images of the cluster means obtained for these digits.

Question 3

(i) Consider the dataset PC generated by the first $K = 90$ columns of Γ in Question 2 - *i.e.* PC $= X\Gamma[, 1 : K]$. Carry out a linear discriminant analysis of digits based on PC-data. Provide a suitably labelled 3×3 array of boxplots, one for each of each discriminant variable, showing the distribution of the discriminant scores by digit.
(ii) Transform the K-dimensional discriminant loading vectors $\{a_1, a_2, ..., a_9\}$ into the image domain (*i.e.* $a_j \to \Gamma[, 1 : K]a_j \equiv im_j$) and use these to generate a 3×3 array of images of the loading variables.
(iii) Use (i) & (ii) to develop an interpretation of the discriminant variables.
(iv) Evaluate the average within digit covariance of the PC data and put the result into $\hat{W} = \frac{1}{10}\sum_{d=0}^{9} \hat{W}_d$.
(v) Assess, graphically and quantitatively, if each of the K diagonal elements of $\hat{W}_d$ for $d = 0, ...9$ are the same across the different digits.
(vi) In the context of linear or quadratic classification, explain why this might be an important diagnostic?

Question 4

Using lda() and qda() in the MASS library, carry out linear and quadratic multiple classification analysis of digits based on the data (PC) and also on the $N \times 9$ data-matrix (D) of discriminant scores from 3(i). Report cross-validated misclassification rate tables (well-formatted, with labels and legends) for lda and qda using both the PC and D data-matrices. Note you will have 4 tables in all. Comment on the overall misclassification rates.