Assignment 1: Due March 3 via Canvas.

- Full marks for correct answers to all 5 questions. All are worth equal marks.
- Provide all your answers in a single pdf document (exported from Word/Pages). Start each question on a separate page.
- Paste your R code into the document at the end of each question.
- Submit your final pdf document Canvas submission name the file Assignment1-uccstudentid-name.pdf (e.g. Assignment1-11xxxxxxx-FinbarrOSullivan.pdf). Start the submission process at least 15 minutes before the Canvas submission deadline. No late assignments will be accepted.

Question 1

Let M be a real symmetric matrix of dimension with column/row dimension p.

(a) Formally show that the eigenvalues of M must be real.

(b) Let the eigenvalues and corresponding orthonormal eigenvectors of M be given by $\{\lambda_1, \lambda_2, ..., \lambda_p\}$ and $\{\gamma_1, \gamma_2, ..., \gamma_p\}$, respectively. By comparing the representation of M and M', show that inverse of the matrix of eigenvectors is its transpose. Verify in R using a matrix of dimension $p = 111$ with random normal entries.

(c) Suppose M is also positive definite ($x'Mx \geq 0$ for any choice of x). Use the spectral decomposition of M to show that the maximum/minimum of $y'My$ (over all y for which $y'y = 1$) is the maximum/minimum eigenvalue and that the corresponding optimal value of y is the associated eigenvector.

Question 2

Compute, directly using the data matrix [not the var function], the covariance and correlation matrix of the USJudgeRatings dataset in R.

(a) Evaluate the Eigen-decomposition of both matrices. Normalise both sets of eigenvalues by dividing by the maximum and use matplot to show both sets of normalised values in the same plot (properly label etc).

(b) Create a set of similar plots for pairs of eigenvectors corresponding to the largest, next largest, ... , min eigenvalues. Put the eigenvalue plot and the first 8 pairs of eigenvector plots on a single page as a 3x3 array, appropriately labelling.

(c) Consider the correlation matrix spectral decomposition: R = G D G'. Using the first 4 columns of G as a design matrix, model the standardised deviations of each judges data vector from the overall mean ($y_i = (x_i - \mu)/sd$) using a linear model with these explanatory variables alone (no intercept). Put the linear model coefficients into a vector b_i, $i = 1, 2, \ldots$.. Use plots to compare b_i to $G'y_i$.

(d) Compare the linear model residual sum of squares (summed over all judges) to the sum of the last 8 eigenvalues of the correlation matrix. Explain the reason for the correspondence.

Question 3

The vector μ (*aka* mu) and matrices Σ (*aka* Sigma), A and vector V are specified in the HWK2.Rdata file. (use the function load to read it)

(a) Use R to simulate a random sample, $\{X_i, i = 1, 2, ..., N\}$, of size $N = 900$ from a $P = 60$ dimensional multivariate normal $N_P(\mu, \Sigma)$. What is the theoretical mean and covariance of $Y_1 = AX_1$? Justify your answer using a formal definition of the multivariate normal.

Compute the sample mean ($\hat{\mu}$) and covariance ($\hat{\Sigma}$) of the sample values $\{Y_i = AX_i, i = 1, 2, ...N\}$ and create two appropriately labeled plots (one for mean and another for covariance) comparing estimates (y-axis) versus true/theoretical (x-axis) values.

(b) Evaluate the Mahalanobis distances:

$$Z_i = (X_i - \mu)'\Sigma^{-1}(X_i - \mu), i = 1, 2, ..., N$$

and leave-out-one values

$$\hat{Z}_i = (X_i - \hat{\mu}_{(-i)})'\hat{\Sigma}^{-1}_{(-i)}(X_i - \hat{\mu}_{(-i)}), i = 1, 2, ..., N$$

Using the theoretical density function of Z_1, create two plots with sample histograms (on a density scale) of the Z and $\hat{Z}$ datasets; superimpose lines on the histograms showing the theoretical density function of Z_1.

(c) Using a 0.05 level of significance, formally test the hypothesis that the vector V specified in HWK2.Rdata is a realization from $N_P(\mu, \Sigma)$. Provide formulas for your test-statistic and the calculation you are using to compute the *p-value* for your test.

(d) Assess the hypothesis that X_i is consistent with a multivariate normal $N_P(\hat{\mu}_{-i}, \hat{\Sigma}^{-1}_{(-i)})$ using the value $\hat{Z}_i$ as a test statistic. Indicate your *p-value* calculation. Provide a histogram of *p-values*, $\{\hat{p}_i, i = 1, 2, ..., N\}$ for these tests. Do results show conformity to a uniform? Assess this more formally using the ks.test() function in R. Interpret the result.

Question 4

The file wave-clip.mp4 contains a small 1-minute movie of a shallow-water ocean scene.

(a) Install the libraries "av" and "magick" in R-studio.

(b) Use av to create jpg files of individual movie frames

```
o = av_video_images(mp4file,destdir=junkd) ; setwd(junkd)
jpgfilenmes=system("ls ",intern=TRUE)
```

Using magick the k'th frame can be read and its data extracted using

```
o=image_read(paste0(junkd,"/",jpgfilenmes[k]))
  data=as.numeric(image_data(o))
```

Modify the above to create an .Rdata set with components corresponding to the red, green and blue channels of the movie. These should each be 3-dimensional arrays (Nx,Ny, T) with $T = 1790$

(c) Evaluate the temporal covariance of the red channel, Σ_T, and compute the associated spectral decomposition. Plot the first J eigenvalues of Σ_T. Interpret the result.

Question 5

Consider the .Rdata from question 4 - the red, blue and green channels of the 1-minute movie taken of a shallow-water ocean scene. Each channel is an array of dimension $N_x \times N_y \times T$. Let $N = N_x \cdot N_y$.

(a) Compute the 3-dimensional covariance of the data colour-space, Σ_C, and report the proportion of variance explained by each principal component. Interpret the loading vectors associated with each component.

Normalize the first loading vector to have maximum and *maximum absolute value* of unity and evaluate the corresponding projection of the data. Make a 2x2 array of labeled plots, with images of the intensity of the 3 colour channels at time $t_M = T/2$ and the first principal component projection of the color-space. Evaluate the full sequence of images corresponding to the first colour-space principal component and after adjusting each pixel time-series for a linear trend in time, create a $(N \times T)$ data-matrix Z.

(b) Evaluate the temporal covariance of the Z-data matrix, Σ_T, and compute the associated spectral decomposition. Describe the relation between the cumulative sum of the eigenvalues and the residual sums explained by a linear model involving eigenvectors of the covariance. By direct computation verify this relation for the sum of the first 5 eigenvalues.

Using the FFT evaluate the variance explained by Fourier components and order these by the amount of variance explained. Provide a plot comparing the fraction of variance explained as a function of the number of components by both Fourier and principal components.

How many Fourier and PCA components are needed to explain 95% of the variance in the data?

(c) Plot the first 6 principal component scores in a 3x3 array of images. Also provide a plot showing the time-series for each of the first 20 PC loading vectors.