MAST90138 Week 4 Lab

Problems:
The iris data contain various measurements (sepal length, sepal width, petal length and petal width) of 50 flowers from each of 3 species of iris flowers. Type `help(iris)` to learn about the format of these data.

1. Load the `iris` data in R (they are already in R).

   ```
   data(iris)
   ```

2. Print on the screen the name of each column of this data set.

   ```
   colnames(iris)
   ```

3. Draw a scatterplot of each pair of numerical variable of this dataset.

   ```
   plot(iris[,1:4])
   ```

4. Draw a scatterplot of each pair of numerical variable of this dataset, this time using a different colour per species (in other words, use 3 colours, one per species). Add a title to the main graph, well above the tickmarks of the pairwise plots. Do the three species appear to be well separated in these graphs?

   ```
   plot(iris[,1:4],col=c(2,3,4)[iris$Species],main='Iris data')
   ```

   Note, using main to put the title produces the title well above the tickmarks, which is not the case if you use `title(Iris data')`.

5. Compute the eigenvectors and eigenvalues of the covariance matrix of these data (using the four numerical variables) and store them in, respectively, a matrix and a vector called `vec` and `lambda`.

   ```
   EVV=eigen(cov(iris[,1:4]))
   G=EVV$vectors
   lambda=EVV$values
   ```

6. Compute the four principal components of these data (using the four numerical variables). What is the variance of each of the four PCs? Which fraction of the sum of the variances of the four components of the original data do each of the PCs explain? What does this suggest?

   ```
   n=nrow(iris)
   X=as.matrix(iris[,1:4])
   #Compute PCs like this:
   PCX=(X-matrix(rep(1,n),nrow=n)%*%colMeans(X))%*%G
   #Can also do like this:
   PCX=scale(X,scale=F)%*%G
   fracvar=lambda/sum(lambda)
   ```

   We find that `fracvar= 0.924618723 ,0.053066483, 0.017102610 ,0.005212184` , so the first PC already explains 92.5% of the variability that all four components explain together. This suggests that the first PC alone captures a lot of the information about the data.

7. Plot the first PCs along a horizontal line, using a different colour for each species. Label by PC1 the x axis and remove entirely the second axis from the figure (keep only the vertical lines but remove all ticks etc). Does the first PC separate well all 3 species?

   `plot(PCX[,1],rep(0,n),col=iris$Species,xlab='PC1',ylab='',yaxt="n")` No, versicolor and virginica are not well separated.

8. Do a scatterplot of the first two PCs, using a different colour for each species and putting x and y labels to the axes (called them PC1 and PC2). Do the first two PCs together separate well all 3 species?

   ```
   plot(PCX[,1],PCX[,2],col=iris$Species,xlab='PC1',ylab='PC2')
   ```
   Yes, by adding a second PC we have a clearer separation of the three species.