MAST90138 Week 5 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 formal of these data.

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

   Solution:

   ```
   data(iris)
   ```

2. Do a PC analysis of these data using only the numerical variables, this time using the prcomp command. Using the output of this function, store the eigenvectors of the covariance matrix S in a matrix and the eigenvalues in a vector. Also store the Y_{ik}'s in a matrix Y, again using the output of prcomp.

   Solution:

   ```
   PCX=prcomp(iris[,1:4],retx=T)
   vec=PCX$rotation
   lambda=PCX$sdev^2
   Y=PCX$x
   ```

3. Draw a screeplot for these data and recall the ψ_j's (cumulative proportion of variance explained by each component) from last week. How many components does this suggest you should keep?

   Solution:

   ```
   screeplot(PCX)
   cumsum(lambda)/sum(lambda)
   ```

   The first two PCs together explain 98% of the variability of the data and the screeplot confirms as quick and sharp decrease of the λ_k's. This suggests that the first two PCs capture a large fraction of the variability of the original data and that just with these two we may be able to uncover interesting features about the data.

4. What is the weight of each original variable in the linear combination use to create PC1 and PC2? Which variables are the most correlated with each PC (describe PC by PC and support your answer by some calculations)?

   Solution:

   ```
   vec
                        PC1         PC2         PC3        PC4
   Sepal.Length  0.36138659 -0.65658877  0.58202985  0.3154872
   Sepal.Width  -0.08452251 -0.73016143 -0.59791083 -0.3197231
   Petal.Length  0.85667061  0.17337266 -0.07623608 -0.4798390
   Petal.Width   0.35828920  0.07548102 -0.54583143  0.7536574
   ```

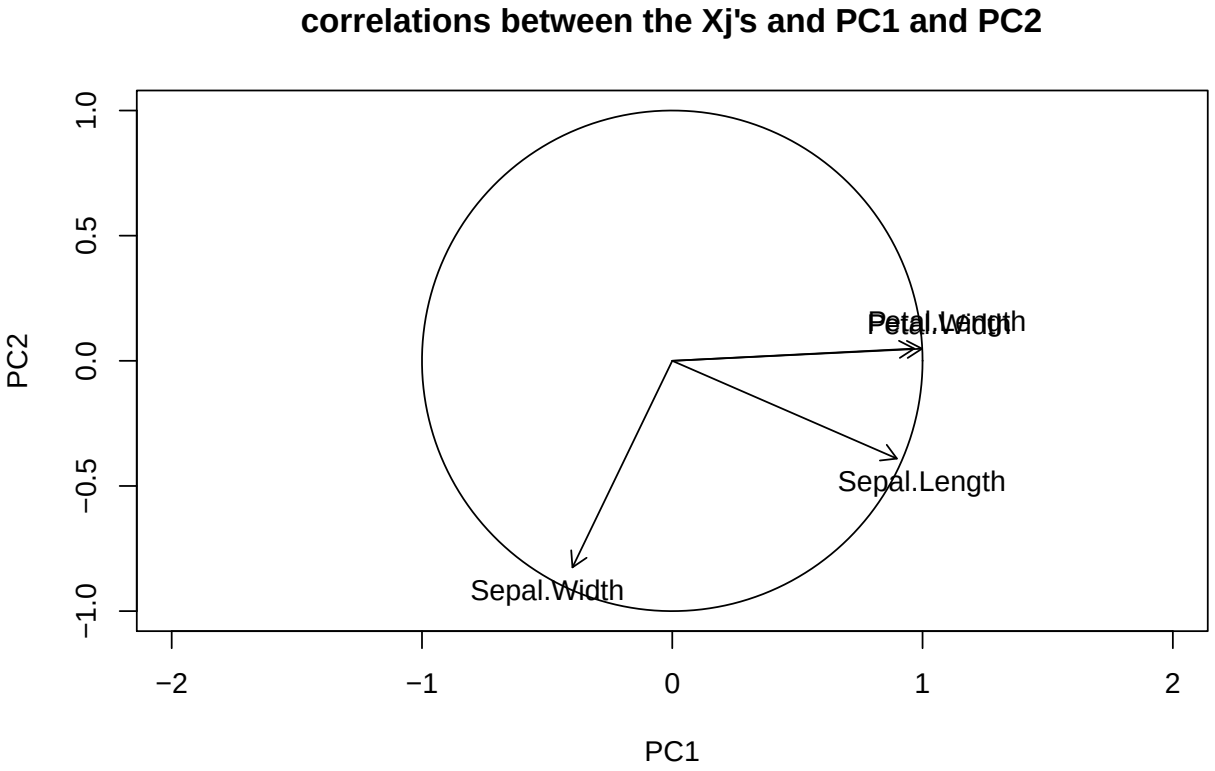
PC1 puts weight 0.3613866, -0.08452251, 0.85667061, 0.35828920 on, respectively, the sepal length, the sepal width, the petal length and the petal width. PC2 puts weights -0.6565888, -0.73016143, 0.17337266, 0.07548102 on, respectively, the sepal length, the sepal width, the petal length and the petal width. PC3 puts weights 0.5820299, -0.59791083, -0.07623608, -0.54583143 on, respectively, the sepal length, the sepal width, the petal length and the petal width. PC4 -puts weights 0.3154872, -0.31972310, -0.47983899, 0.75365743 on, respectively, the sepal length, the sepal width, the petal length and the petal width.

PC1 puts the most weight on the petal length and also some weight on the sepal length and the petal width; all contribute positively to PC1. PC2 puts the most weight on the sepal length and the sepal width, which have a negative effect on PC2. PC3 put most of its weight on all but the petal length and PC4 puts most of its weight on the petal width.

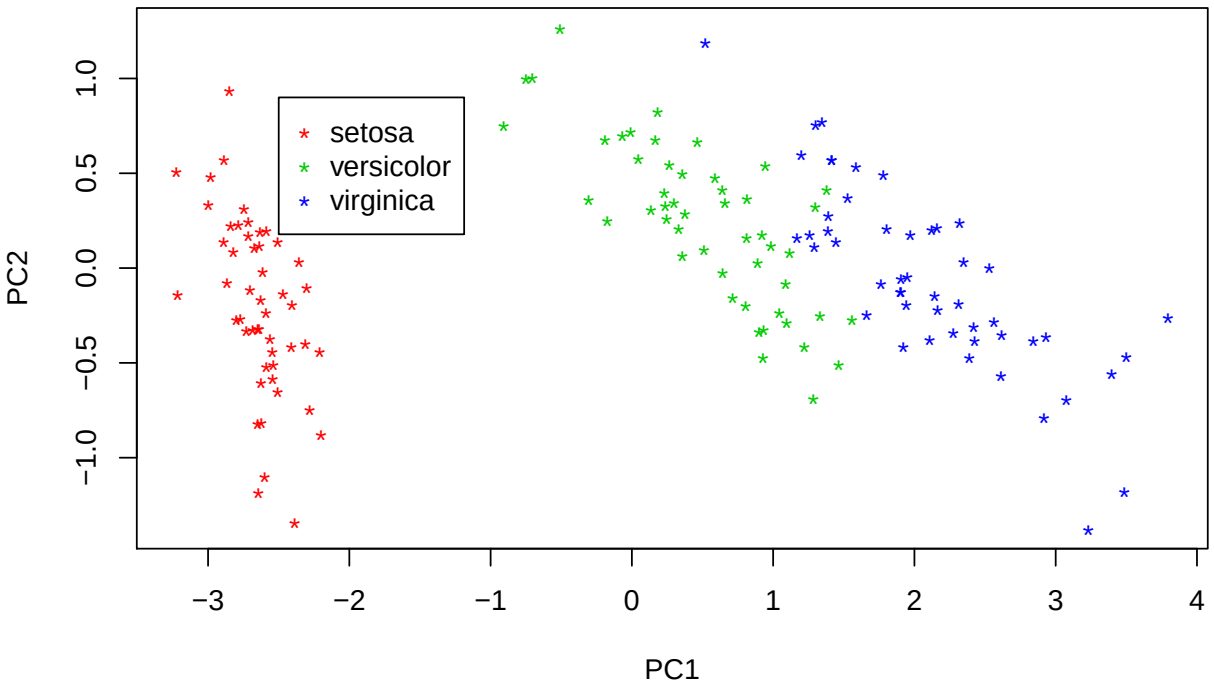
5. The correlation graph showing the correlation between each of the original variable and two PCs is given below. We also provide a table with the values of the correlations between each original variable and each PC. Use this graph and this table to provide more insight into the results of the analysis.

Table 1: Correlations between original variables and the principal components:

	PC1	PC2	PC3	PC4
Sepal length	0.8974018	-0.3906044	0.19656672	0.05882002
Sepal width	-0.3987485	-0.8252287	-0.38363030	-0.11324764
Petal length	0.9978739	0.0483806	-0.01207737	-0.04196487
Petal width	0.9665475	0.0487816	-0.20026170	0.15264831



Solution:



All four variables are close to the circle of radius 1, which indicates that they are strongly correlated with the first two PCs. The angle between the arrows of Petal length and width are almost parallel to the axis for PC1, and almost at a right angle with the axis of PC2, which suggests that these two variables are strongly positively correlated with PC1 and almost not correlated with PC2. Sepal length has a small angle with axis of PC1 and angle smaller than, but close to, 90 degrees with PC2, it is positively correlated with PC1 and has small negative correlation with PC2. Sepal width has small angle with the axis of PC2 but points in opposite direction, and angle with PC1 a bit larger than 90 degrees, it is strongly negatively correlated with PC2 and had small negative correlation with PC1.

Petal and width and length are strongly positively correlated, and positively correlated with sepal length. Arrow of Sepal width is at a right angle with that of sepal length, so these two variables have very small correlation. Sepal width has moderate negative correlation with petal width and length because its arrow points at angle moderately larger than 90 degrees with arrows of those variables. All this is only valid because arrows are near periphery of circle. Let's check our claims by looking at correlation matrix of original data:

```
cor(iris[,-5])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

We also know that together the first two PCs explain a large fraction of the variability of the data, so the direction of the arrows can be used in conjunction with the scatterplot of the first two PCs to learn the effect of those three original variables on the individuals. In particular, it appears that the setosa tend to be very different from the versicolor and the virginica: they

tend to have a larger sepal width than these two. The versicolor and the virginica tend to have larger values of petal width and length and of sepal length. The virignica tend to have larger values of petal width and length than the versicolor.

Going back to the original data using `pairs(iris[,1:4],col=c(2,3,4)[iris$Species])`, we can see that indeed this is the case.

6. Now compute the variances of the original variables. What does it suggest for the PC analysis done above and what can we consider doing instead (do it)?

```
diag(cov(iris[,-5]))
Sepal.Length  Sepal.Width  Petal.Length   Petal.Width
    0.6856935    0.1899794     3.1162779    0.5810063
```

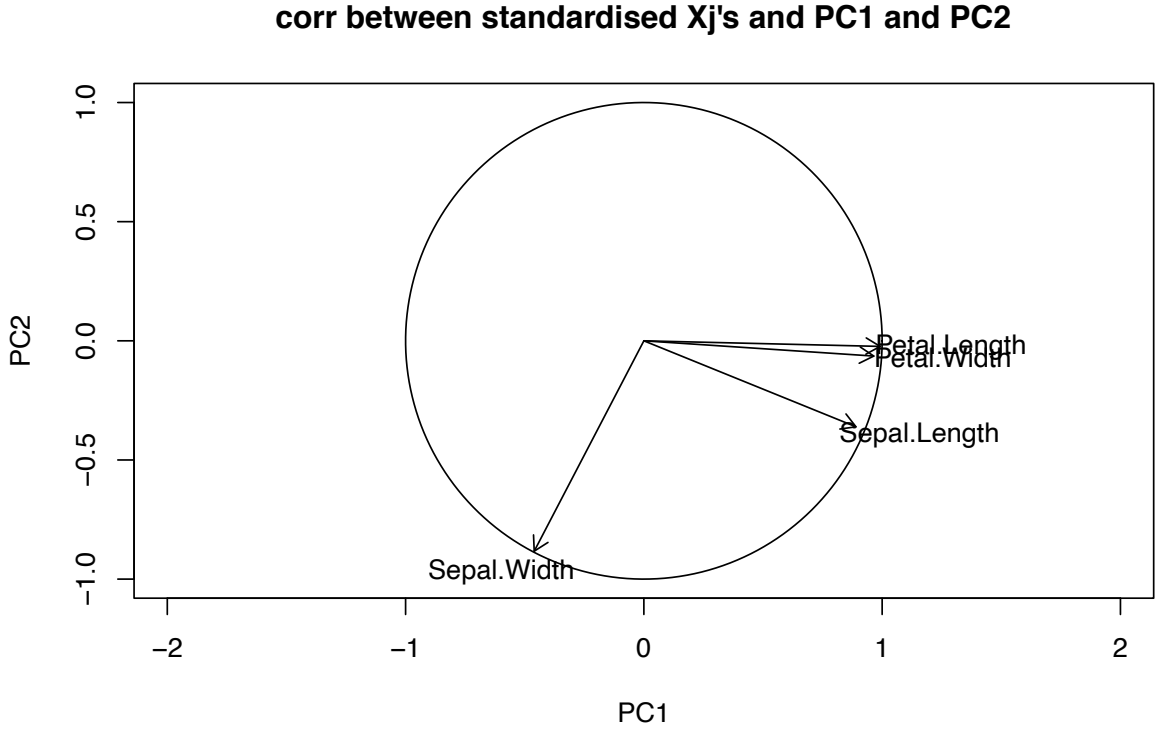
The variance of petal length is much larger than that of the other variables, which is why PC1 has focused a lot on that variable. We can consider doing PCA of the scaled variables:

```
PCX=prcomp(iris[,1:4],retx=T,scale=T) vec=PCX$rotation
lambda=PCX$sdev^2
Y=PCX$x
screeplot(PCX)
cumsum(lambda)/sum(lambda)
0.7296245 0.9581321 0.9948213 1.0000000
```

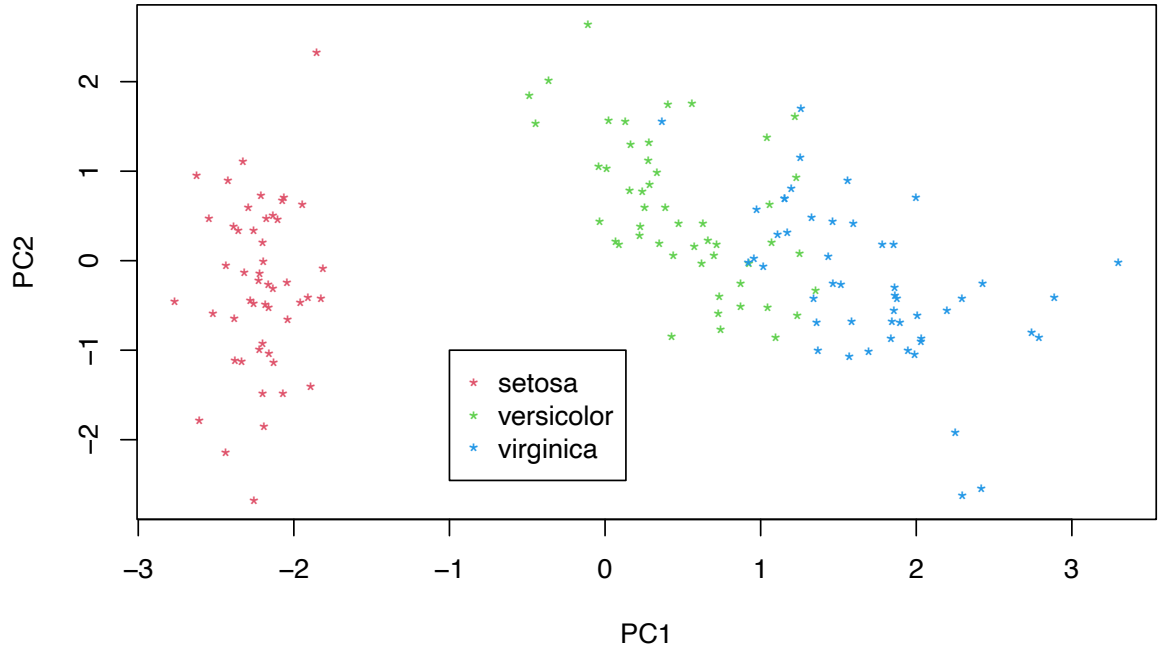
Now we find that the first two PCs explain 96% of the variability of the data. We have

```
vec
                     PC1         PC2         PC3        PC4
Sepal.Length   0.5210659 -0.37741762   0.7195664  0.2612863
Sepal.Width   -0.2693474 -0.92329566  -0.2443818 -0.1235096
Petal.Length   0.5804131 -0.02449161  -0.1421264 -0.8014492
Petal.Width    0.5648565 -0.06694199  -0.6342727  0.5235971
```

and we see that now PC1 puts more or less equal weight to sepal length, petal width and petal length. the correlation plots and coloured scatterplots now look like this:



Solution:



In this particular example, the graphs and conclusions for PCA on scaled variables are almost identical to those of non scaled variables. However, by rescaling the data, we have lost a little of our ability to distinguish between versicolor and virginca. This is because petal length is actually an important variable to distinguish between the flower types, and since the PCA on the non scaled data was putting a lot of weight on that variable, it was able to better distinguish between the flower types. This illustrates that rescaling the data does not always improve the descriptive analysis.