Rock Crabs 3

Principal components

This is the **third** in a series of questions on the exploration of some data on rock crabs.

The context for the data as well as the data itself is explained in more detail in the **background** document accompanying the assignment.

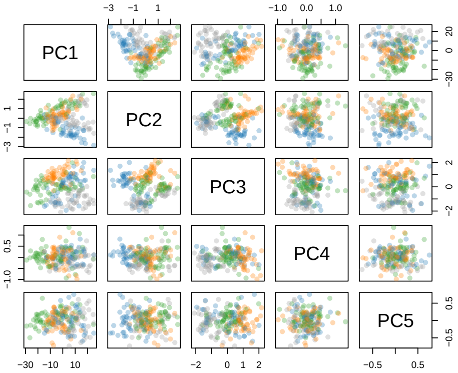
Since you will be using loon's interactive graphics to explore the data, you might also want to occasionally save some plots from your interactive analysis to include in your solution(s). Information on how to do this appears in the document **SavingLoonPlots**. Be sure to read that document (and even its "Rmd" version) before attempting to save loon plots.

17 marks

a. Principal components. Rather than work with the original variates, we might prefer to work with the principal component axes instead. The coordinates in this axis system are easily had in `R` via the function `prcomp()` using default arguments. You will need the `x` component of the value returned by `prcomp` applied on the `lepto` data.

   i. **(3 marks)** Produce a `pairs` plot of `x` with `pch=19` and alpha blending, where each point of `x` is coloured according to which of the subgroups you identified earlier. Show your code.

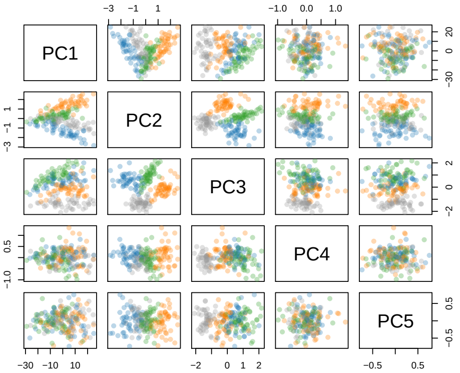
```
pairs(prcomp(lepto)$x, pch=19, col=adjustcolor(hex12tohex6(pic5["color"]), 0.3))
```



   ii. **(3 marks)** Produce another `pairs` plot of `x` with `pch=19` and alpha blending, but this time colour each point differently depending on its value of `Species` and `Sex` from the data set `crabSpecies` (also from loon.data. Match the colours to be the same as your earlier groups (i.e. match to the predominant `Species` and `Sex` in each of your original groups). Show your code.

```
colour <- matrix(1:200)
colour[crabSpecies$Species=="blue"&crabSpecies$Sex=="male"]=unique(pic5["color"])[3]
colour[crabSpecies$Species=="blue"&crabSpecies$Sex=="female"]=unique(pic5["color"])[1]
colour[crabSpecies$Species=="orange"&crabSpecies$Sex=="male"]=unique(pic5["color"])[2]
colour[crabSpecies$Species=="orange"&crabSpecies$Sex=="female"]=unique(pic5["color"])[4]

pairs(prcomp(lepto)$x, pch=19,
      col=adjustcolor(hex12tohex6(colour), 0.3))
```



iii. **(2 marks)** Which pair of principal components best separate the four groups? How well did your grouping fare in separating the same groups (on those two principal components)?

PC2 best seperate the four groups. It shows a clear clusters for four groups

b. Principal components continued.

Suppose you have the principal components from part f as the matrix `x`. You will now cluster the data interactively using a mix of the principal component coordinates and the original coordinates.

To that end, execute the following code:

```
library(loon); library(PairViz)
x <- prcomp(lepto)$x
nav <- l_navgraph(x[, 1:3])
p <- nav$plot
g <- l_glyph_add_serialaxes(p, data = lepto[, eseq(ncol(lepto))],
                            showAxes = TRUE, showArea = TRUE, label = "serialaxes")
p["glyph"] <- g
```

i. **2 marks** A popular automatic clustering method is "kmeans". This can be used on the crabs data and displayed as colours in the plot `p` just constructed.

```
km <- kmeans(lepto, centers = 4)
p["color"] <- km$cluster
```

Using the navigation graph interface, comment on what you perceive to be the quality of the clustering produced by "kmeans"? That is, how do the groups given by kmeans compare to any grouping you might perceive in the plot?

Remember that the scientists answers are contained in `crabSpecies`. An easy way to compare the kmeans result with the true values is as follows:

```
classes <- paste(crabSpecies[,1], crabSpecies[,2], sep = ":")
table(classes, km$cluster)
```

```
##
## classes          1  2  3  4
##   blue:female    3 16 19 12
##   blue:male     10 19 14  7
##   orange:female 18 21  9  2
##   orange:male   14 14 17  5
```

PC1:PC2 and PC2:PC3 are bad, PC1:PC3 is relatively better. Generally, k means fails to provide good quality of the custering. Lengh and width have great influence on groupings.

ii. **4 marks** Change all of the colours in the scatterplot to the same colour (e.g. p["color"] <- "grey50"'). Now, using the navigation graph, explore the space of the first three principal components. Based on what you see, divide the crabs into 4 different groups (using different colours).

Describe the features of the plot you used to make decisions about the groups.

Using `table()` show how well you did. How well did you do compared to kmeans?

```
pic6 <- l_getSavedStates(file = "pc6")
table(classes, pic6["color"])
```

```
##
## classes         #3333A0A02C2C #999999999999 #E3E31A1A1C1C #FFFFFFFF0000
##   blue:female               1             1             0            48
##   blue:male                 3             0            41             6
##   orange:female             4            46             0             0
##   orange:male              50             0             0             0
```

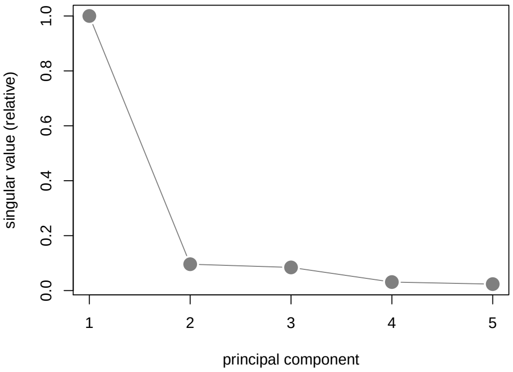
The plot I used has clear borders between clusters and Colors are given to every large cluster. The result is pretty good as only one group has dominate number for each grouping.

iii. **3 marks** The `sdev` component of the `prcomp()` output contains the singular values for the principal component analysis. Use these to produce a "scree plot". Show your plot.

According to the scree plot, how many principal components should be sufficient?

Would that number of principal components have been enough to have identified the groups? Why? Or why not?

```
d <- prcomp(lepto)$sdev
plot(d/max(d),type="b", xlab="principal component",
ylab="singular value (relative)",
col="grey50", pch=16, cex=2)
```



One principle component should be sufficient to have identified groups. However, two and three components are needed because from the previous questions, we see that two principle components works well and one does not work effectively.