TRIED RN Exam

Carl Robinson - 19th Dec 2017

TP EX13 : Morphologie des Crabes

Objective, Data and Method

The objective of this report is to identify the existence of several distinct groups (varieties) of crabs using principal component analysis (PCA).

The dataset represents the morphology of 200 Leptograpsus crabs from Australia (Campbell & Mahon 1974). The variables are measurements of crab bodies. For use with PCA methods, the data is mean-centered and reduced.

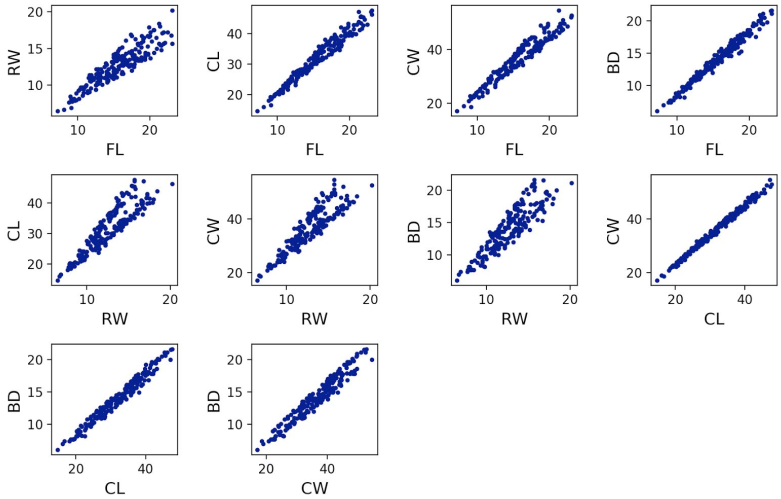
- FL - frontal lip of carapace (mm)
- RW - rear width of carapace (mm)
- CL - length along the midline of carapace (mm)
- CW - maximum width of carapace (mm)
- BD - body depth (mm)

Results and Conclusions

Correlation Coefficients and 2-by-2 Plots

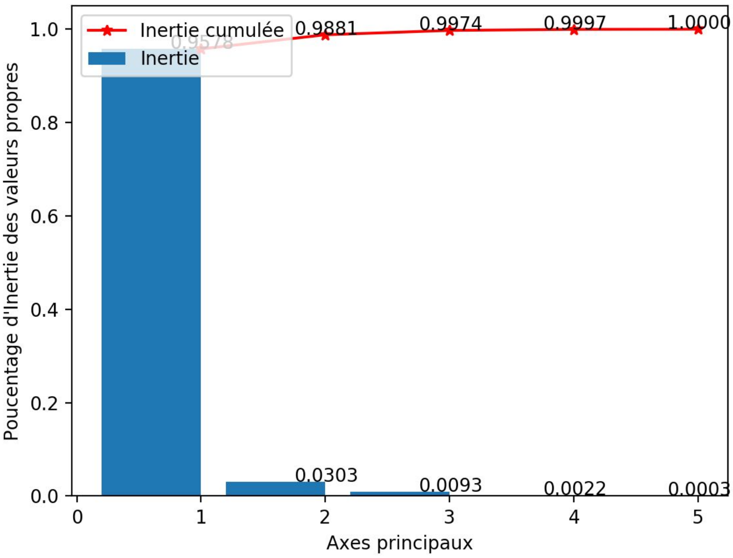
```
            FL          RW          CL          CW          BD
FL [ 1.          0.90698762  0.97884179  0.96495583  0.9876272 ]
RW [ 0.90698762  1.          0.892743    0.9004021   0.88920542]
CL [ 0.97884179  0.892743    1.          0.99502255  0.9832038 ]
CW [ 0.96495583  0.9004021   0.99502255  1.          0.96781165]
BD [ 0.9876272   0.88920542  0.9832038   0.96781165  1.        ]
```

- We can immediately see that the RW variable is poorly correlated with the other four variables, whereas the other four variables are well correlated with each other.



- Plotting the variables 2-by-2 clearly illustrates the extent of the correlations. All the variables are correlated, as shown by their linear relationship.
- Some variables are highly correlated (such as CL and CW together), having an almost perfect linear relationship, while others are have a looser relationship (such as RW and BD)
- The figures also reveal an interesting pattern. Almost all variables are more highly correlated at low values, but become more diffuse and less correlated at higher values.

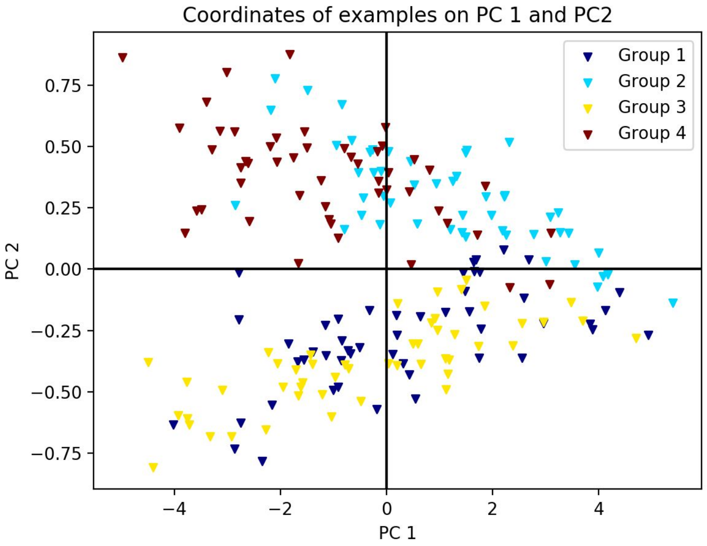
Inertia and Cumulative Inertia

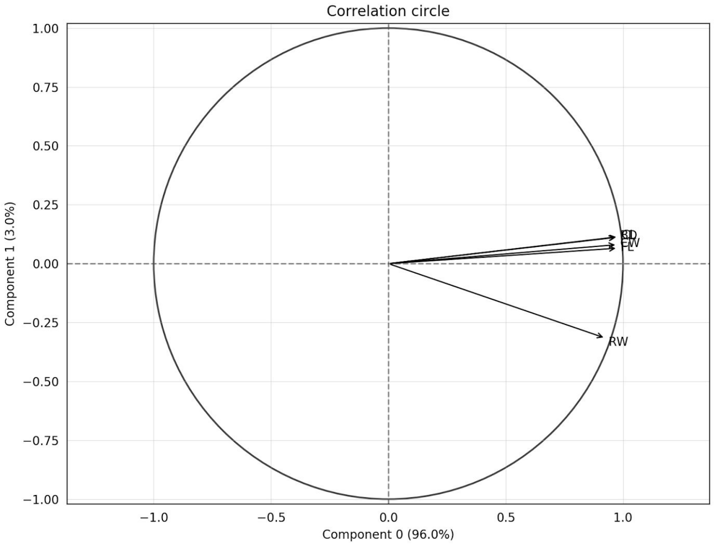


```
+-------+-----------+---------+----------------+
|       | Eigen vals | Inertia | Cumul. inertia |
+-------+-----------+---------+----------------+
| axis1 |    957.767 |   0.958 |          0.958 |
| axis2 |     30.337 |    0.03 |          0.988 |
| axis3 |      9.327 |   0.009 |          0.997 |
| axis4 |      2.227 |   0.002 |            1.0 |
| axis5 |      0.342 |     0.0 |            1.0 |
+-------+-----------+---------+----------------+
```

- Principal axis 1 (PC1) contains 95.8% of the inertia, and therefore exploits the vast majority of the variance (the information) in the data.
- Principal axis 2 (PC2) is the next-most-significant, containing only 3.0% of the inertia. The combined cumulative inertia of PC1 and PC2 equals 98.8%, which is very high, and shows that almost all information contained in the data can be represented along these two axes.
- We will therefore choose PC1 and PC2 to plot our observations against in PCA cloud of observations.

PCA Cloud of Observations (PC1 and PC2) and Correlation Circle

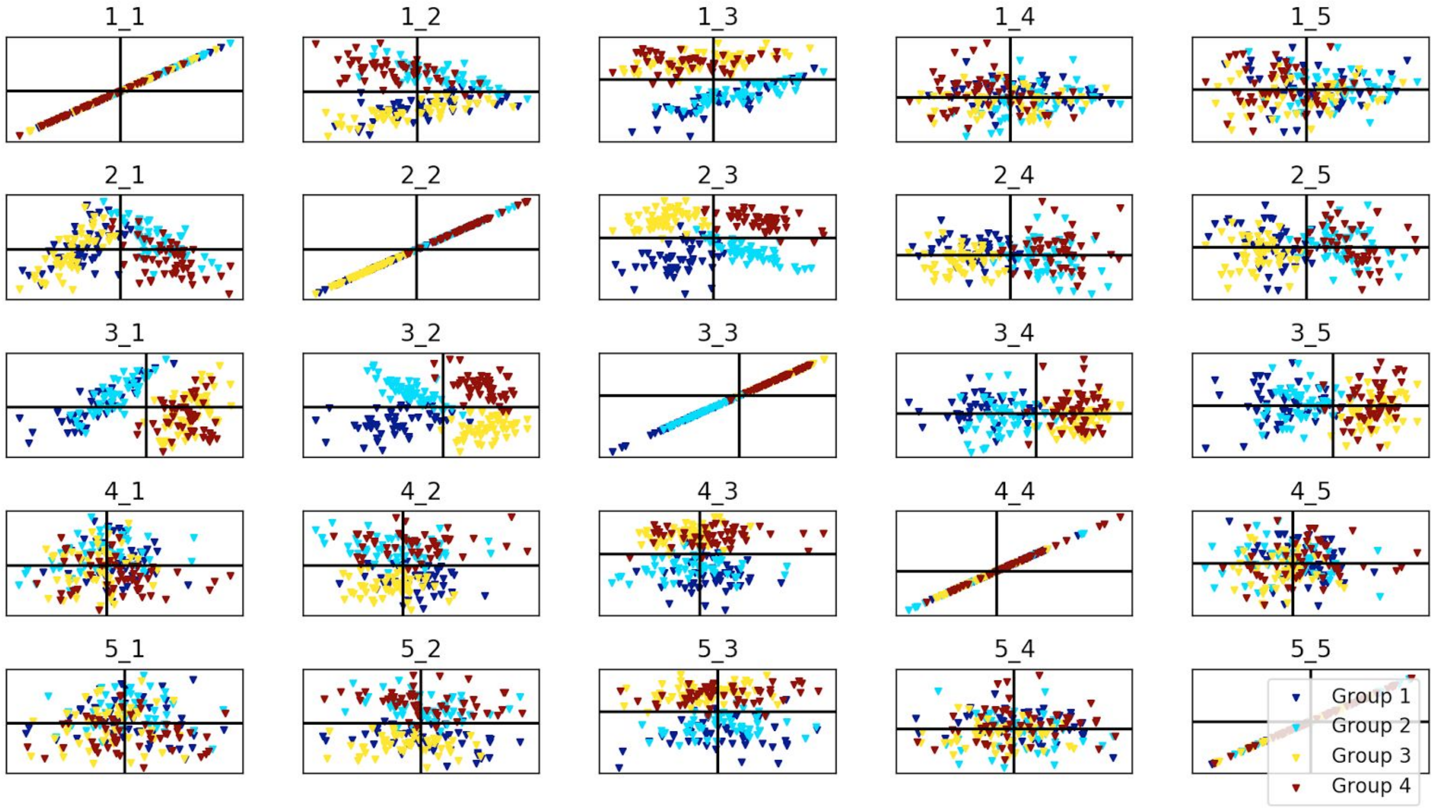




- From the clustering of the points only, and ignoring the colouring that we added using a priori knowledge, we can only see two distinct groups appear; one below the PC2=0 axis, and the other above it.
- Without use of the colouring, it is not possible to distinguish between groups 1 and 3, and between groups 2 and 4, as these clusters overlap considerably when projected onto PC1 and PC2.
- The correlation circle shows why this is the case. Four of the five variables are highly correlated with respect to PC1 and PC2, and so much of the distinguishing information that these variables contain is not represented in the cloud of points.
- It is only the RW variable that differentiates the data, having a greater PC2 component than the others. This explains why the points in the cloud are separated into two clusters above and below the PC2=0 axis.

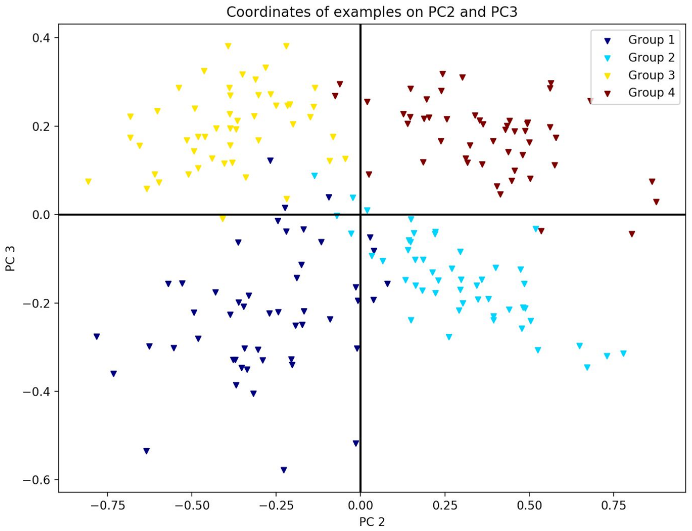
Please note that the direction of the arrows in the correlation circle is reversed in the horizontal axis, due to the coding method. Also note that the component numbers on the correlation circle axes are 0 and 1, but correspond to PC1 and PC2 respectively.

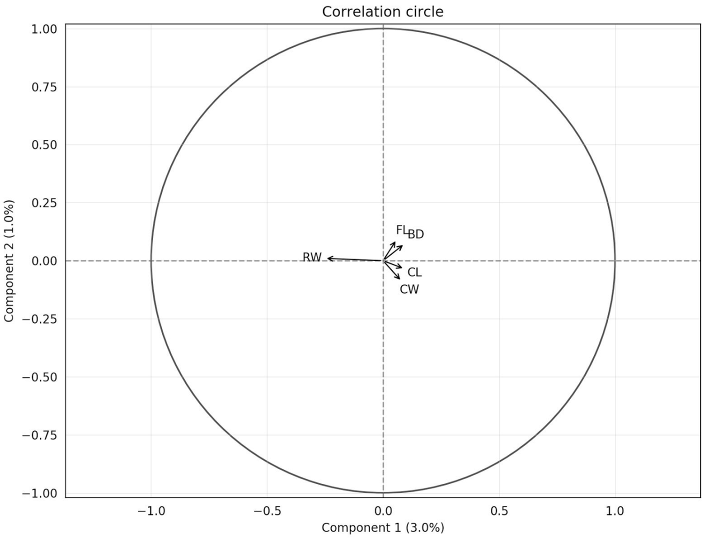
PCA Cloud of Observations (All combinations of principal axes)



- Here we plot the observations against all possible combinations of principal axes, to try and identify a useful combination of axes. We can see that for most, the four groups are not well separated.
- However we do discover that the combination of **PC2 and PC3** clearly separates the four group clusters.

PCA Cloud of Observations (PC2 and PC3) and Correlation Circle





- When projecting onto PC2 and PC3, four distinct clusters appear, allowing us to discover that there are in fact four varieties of crab.
- The arrows on the correlation circle all points in different directions (many are at 90 degrees to each other), indicating a very low or zero correlation between variables when projected onto these two axes.
- This is an interesting result, as the combined inertia of PC2 and PC3 is very low (0.03 + 0.009 = 0.039), as shown by the very short lines on the correlation circle, yet this small amount of information is in fact the most useful to us. We have thus achieved our objective of identifying the 4 varieties of crab using PCA.