PCA: case study

Abdelwahid Benslimane
wahid.benslimane@gmail.com

1. Introduction

This document will explain how to perform a PCA (principal component analysis) using R and to interpret the findings.

The dataset on which this case study is based is a basic dataset well known by data scientists. The version we will use is available on GitHub: https://gist.github.com/yppmark/d907dc265a84cac76ba7

It contains 40 individuals and compiles various information about mammals, notably about their sleep cycle, through 10 numerical variables:

- Body Weight (kg)
- Brain Weight (g)
- Slow wave sleep (hrs/day)
- Paradoxical sleep (hrs/day)
- Total sleep (hrs/day)
- Maximum life span (years)
- Gestation time (days)
- Predation index
- Sleep exposure index
- Overall danger index

Although the variables Predation index, Sleep exposure index and Overall danger index are numerical, they contain qualitative information simply encoded by a discrete value ranging from 1 to 5.

The page on GitHub provides these details:

predation index (1-5)
1 = minimum (least likely to be preyed upon)
5 = maximum (most likely to be preyed upon)

sleep exposure index (1-5)
1 = least exposed (e.g. animal sleeps in a
well-protected den)
5 = most exposed

overall danger index (1-5)
(based on the above two indices and other information)
1 = least danger (from other animals)
5 = most danger (from other animals)

This type of encoding called ordinal encoding can only be applied to qualitative variables whose different values have an order relationship. For example, the size of a t-shirt ranges from XS to XXXL, and these different values follow an order relationship as XS < S < M etc., they could therefore be encoded in the form of discrete values that will keep the same order relationship (XS = 1, S = 2, M = 3 etc.).

The choice was made to include these variables in the PCA, but if the data is then to be used to train a decision model to predict the value of these variables, and the PCA is a step in the process, it is not appropriate to include them in the PCA.

2. Loading of the required libraries and data

The 3 following libraries are needed to perform the analysis: MASS, factoextra and ggplot2. The following commands show how to load them and how to load the data from the .csv file:

```
> library(MASS)
> library(factoextra)
> library(ggplot2)
>
> mydata <- read.csv2(file = "<path_to_the_file>\\SleepInMammals.csv", header = TRUE, sep = ',', dec = '.')
```

You need to replace <path_to_the_file> by the actual path to the file. If you are on Windows, the separator "\" needs to be doubled: "\\".

Although there is no missing value in the version of the dataset downloaded from GitHub, to drop the incomplete rows when necessary you can use the na.omit() function like below. Also, the command below drops the first column which only contains the name of the animal species as this information is not relevant for the PCA and cannot be processed anyway.

```
> data <- na.omit(mydata[,-c(1)])
>
```

In general, we look to impute missing values using various techniques rather than deleting incomplete data.

3. Normalized PCA + summary

With the commands below, you will launch the normalized PCA (scale = TRUE in the prcomp function) and display some information about the principal components

```
> my_pca <- prcomp(data, scale = TRUE)
>
> summary(my_pca)
Importance of components:
                          PC1    PC2    PC3     PC4    PC5     PC6     PC7     PC8     PC9      PC10
Standard deviation     2.2792 1.5489 1.0182 0.74177 0.6115 0.50886 0.37983 0.15914 0.13067 2.249e-16
Proportion of Variance 0.5194 0.2399 0.1037 0.05502 0.0374 0.02589 0.01443 0.00253 0.00171 0.000e+00
Cumulative Proportion  0.5194 0.7593 0.8630 0.91804 0.9554 0.98133 0.99576 0.99829 1.00000 1.000e+00
>
```

You can see that with only the 3 first principal components, we cumulate 86.3 % of the total inertia, and with only the 2 first components 75.93% of the total inertia is already explained.

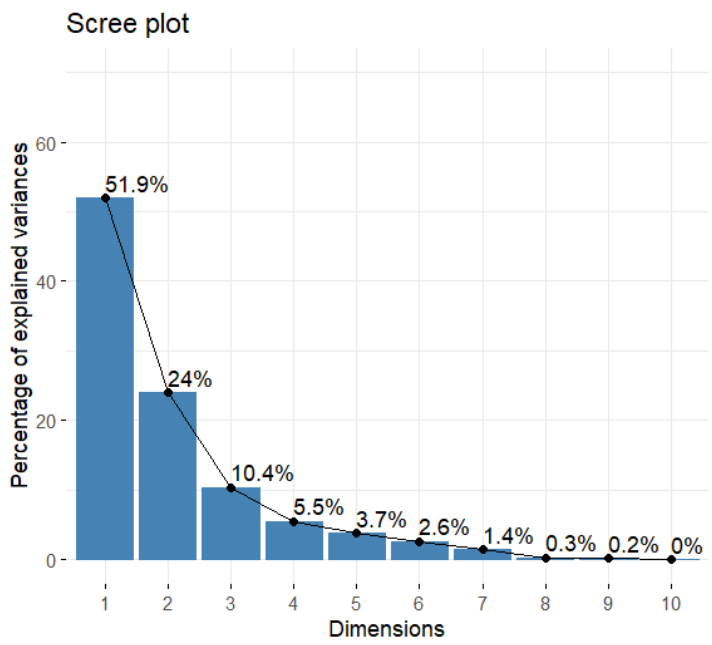
You can also note that the proportion of inertia explained by the last principal components is null.

4. choosing the ideal number of principal components

To be able to use the elbow method in order to select the ideal number of components, we can use the fviz_eig() function and plot the proportion of explained inertia against the different dimensions/principal components.

```
> fviz_eig(my_pca,
+          addlabels = TRUE,
+          ylim = c(0, 70))
> |
```

Output:



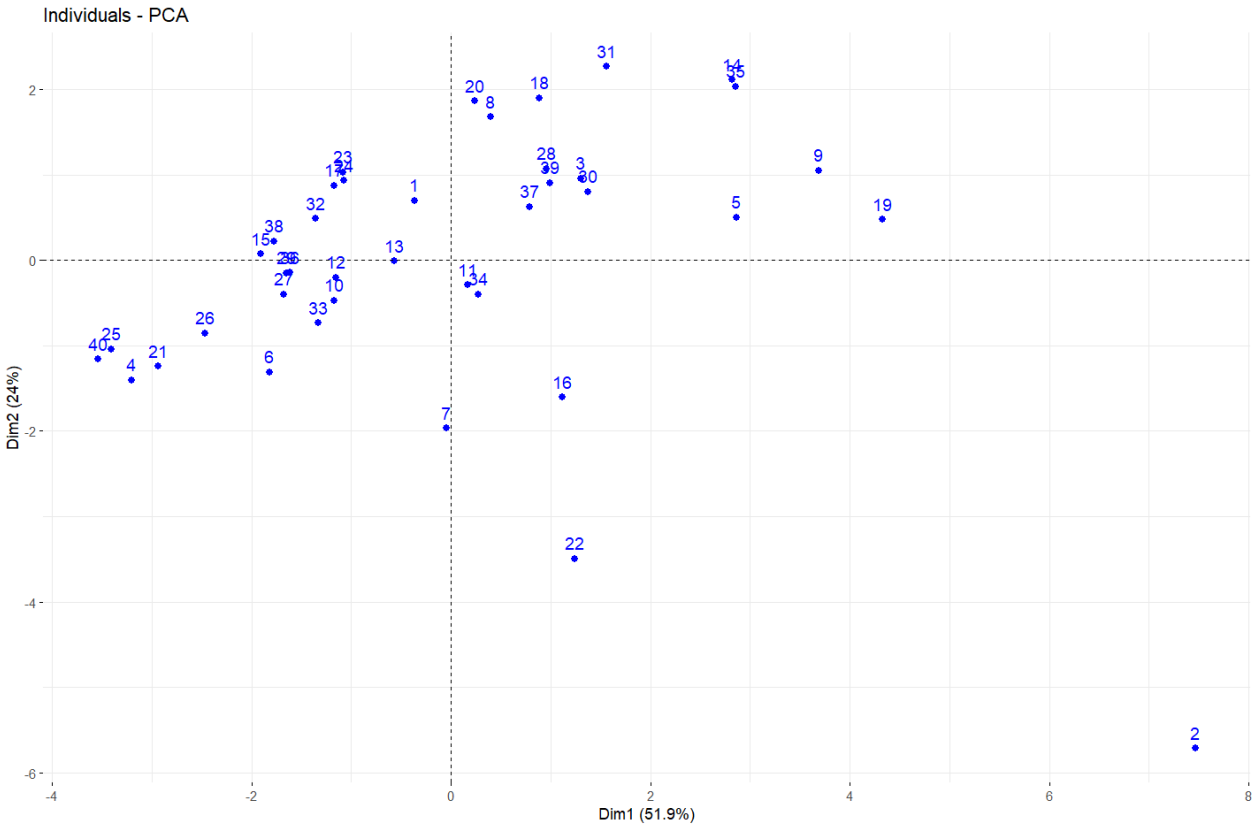
The breaking point or “elbow” seems to be at the 3rd dimension/principal component. According to the elbow method the ideal number of principal components is then 3.

5. Projection of individuals onto the first factorial plane

To project the individuals onto the first factorial plane formed by axes which correspond to the first and second principal components, you can use the fviz_pca_ind() function:

```
> fviz_pca_ind(my_pca, axes = c(1, 2), geom = c("point", "text"),
+              label = "all", invisible = "none", labelsize = 4,
+              pointsize = 2, habillage = "none",
+              col.ind = "blue")
> |
```

Output:



The labels are the numbers corresponding to the indexes of the individuals (line number) in the initial data table.

We can already make some observations that would have been difficult to make from the set of initial variables.

Sheep and goat

For example, at a glance we can see that individuals 14 (goat) and 35 (sheep) are close, therefore their values should be close on most variables compared to how widespread they are.

Goat	27.66	115	3.3	0.5	3.8	20	148	5	5	5
Sheep	55.5	175	3.2	0.6	3.8	20	151	5	5	5

The clearest difference is in the variables Body Weight (27.66 vs 55.5) and Brain Weight (115 vs 175) which have a high variability (this is easily verified by looking at the lines in the table for these variables).

Mouse and musk shrew

The same assumption can be made for individuals 23 (mouse) and 24 (musk shrew, it looks pretty much like a mouse)

Mouse	0.023	0.4	11.9	1.3	13.2	3.2	19	4	1	3
Musk shrew	0.048	0.33	10.8	2	12.8	2	30	4	1	3

The clearest (but still small compared to the range of observed values) difference is for the Gestation time (in days).

Musk shrew

Picture: https://commons.wikimedia.org/wiki/File:Crocidura_HC2.JPG

Mouse

Picture: hhttps://en.wikipedia.org/wiki/Mouse

Asian elephant and water opossum

In contrast, individuals 2 (Asian elephant) and 40 (water opossum) are very far apart and even opposite on the first factorial axis (Dim 1), so we would expect them to be very different.

We can also observe that the Asian elephant is very far from all the other individuals and at one extreme on the 2 axes, which could lead to consider it an outlier.

This is confirmed by the figures. Indeed, the Asian elephant stands out for its very strong/extreme values for the variables Body Weight, Brain Weight, Maximum life span and Gestation time. In the dataset, it is only beaten by humans being as regards Maximum lifespan.

Performing a new analysis after the outliers have been eliminated may be relevant in order to allow the other observations to be expressed more fully.

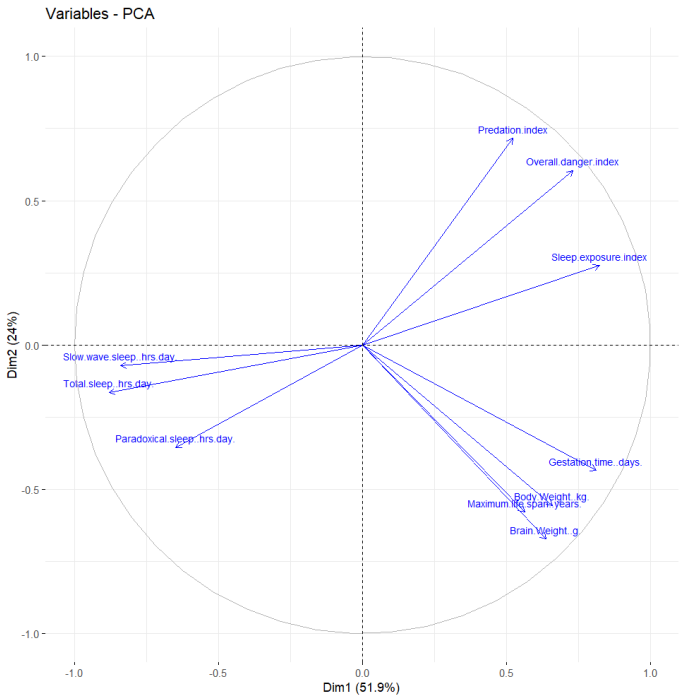
Asian elephant	2547	4603	2.1	1.8	3.9	69	624	3	5	4
Water opossum	3.5	3.9	12.8	6.6	19.4	3	14	2	1	1

6. Correlation circle and analysis

To draw the correlation circle (in the first factorial plane), you need to use the fviz_pca_var() function

```
> fviz_pca_var(my_pca, col.var = "blue", labelsize = 3)
>
```

Output:



We can make several observations. Firstly, we can see that all the variables are all fairly close to the circle/well represented, which means that the variance of all the variables is well explained by the first 2 factorial axes/principal components.

There are obviously 3 distinct groups of variables that are more closely correlated with each other:

- group 1: Predation index, Overall danger index and Sleep exposure index
- group 2: Gestation time, Body weight, Maximum life span and Brain weight (in particular, Maximum life span and Brain weight are very strongly correlated)
- group 3: Paradoxical sleep, Total sleep and Slow wave sleep

According to the plot, we can assume that:

- individuals located in the top right-hand quarter, provided they are far enough from the centre, are characterised by rather high values for the variables in group 1
- individuals located in the bottom right-hand quarter, provided they are far enough from the centre, are characterised by rather high values for the variables in group 2
- individuals located on the left, around the first factorial axis, provided they are far enough from the centre, are characterised by rather high values for the variables in group 3.

The goat and the sheep are in the top right-hand quarter and are indeed characterized by high values for variables of the group 1 (and low values for the variables of the group 3).

The Asian elephant is in the bottom right-hand quarter and is indeed characterized by very high values for the variables of the group 2 (and low values for the variables of the group 3) as we have already seen.

The water opossum is on the left and is indeed characterized by high values for variables of the group 3 (and low values for variables of the groups 1 and 2).

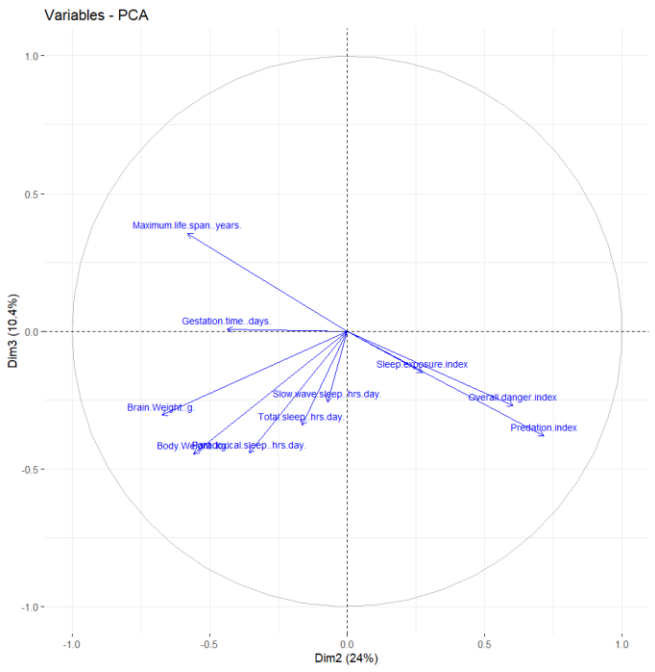
Nota Bene

The analysis was based on observations made on the first factorial plane formed by the first 2 factorial axes, but it is entirely possible, although of little interest, to display the projections of the individuals and the correlation circle on a factorial plane formed by other factorial axes (for example the 1st and 2nd factorial axes, or the 2nd and 3rd factorial axes, or the 3rd and 4th factorial axes, etc.).

A correlation circle on a plane not containing the first factorial axis should have most of the variables more distant from the circle, therefore these variables would be less well represented by the plane, and the analysis would be less reliable and relevant.

Let's do a test with the factorial plane formed by the 2nd and 3rd factorial axes.

```
> fviz_pca_var(my_pca, col.var = "blue", axes = c(2,3), labelsize = 3)
>
```



The variables are more distant overall from the circle, which is particularly obvious for the variables Sleep exposure, Slow wave sleep, Total sleep and Gestation time.