Explanatory data analysis with PCA. Decathlon data analysis

Julie Josse

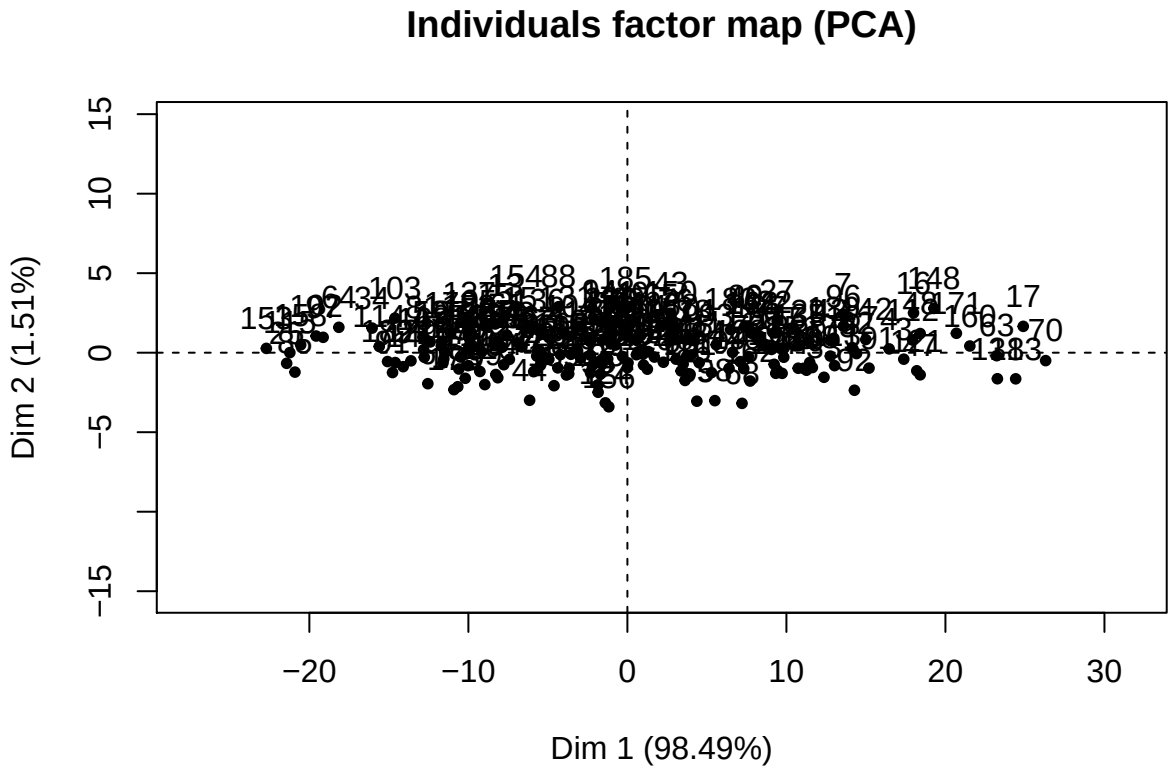
24/10/2018

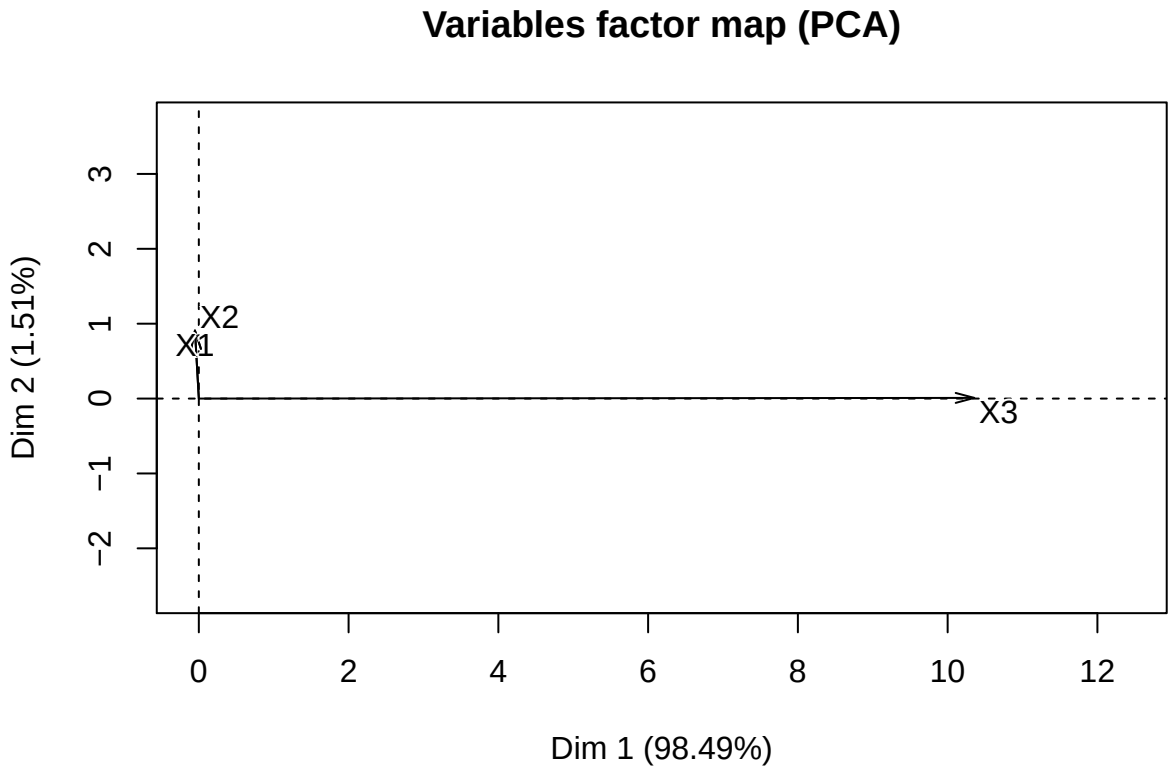
Principal Component Analysis

Lecture questions

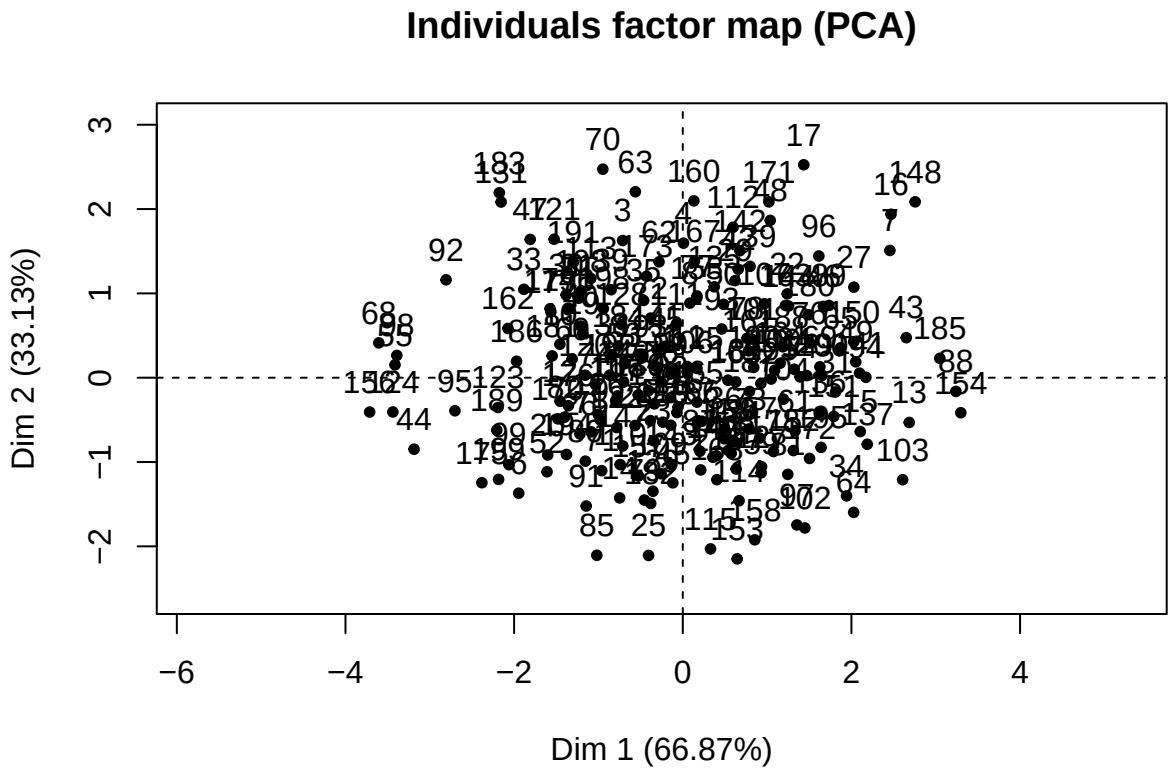
1) When do we need to scale the variables? Simulate and comment:

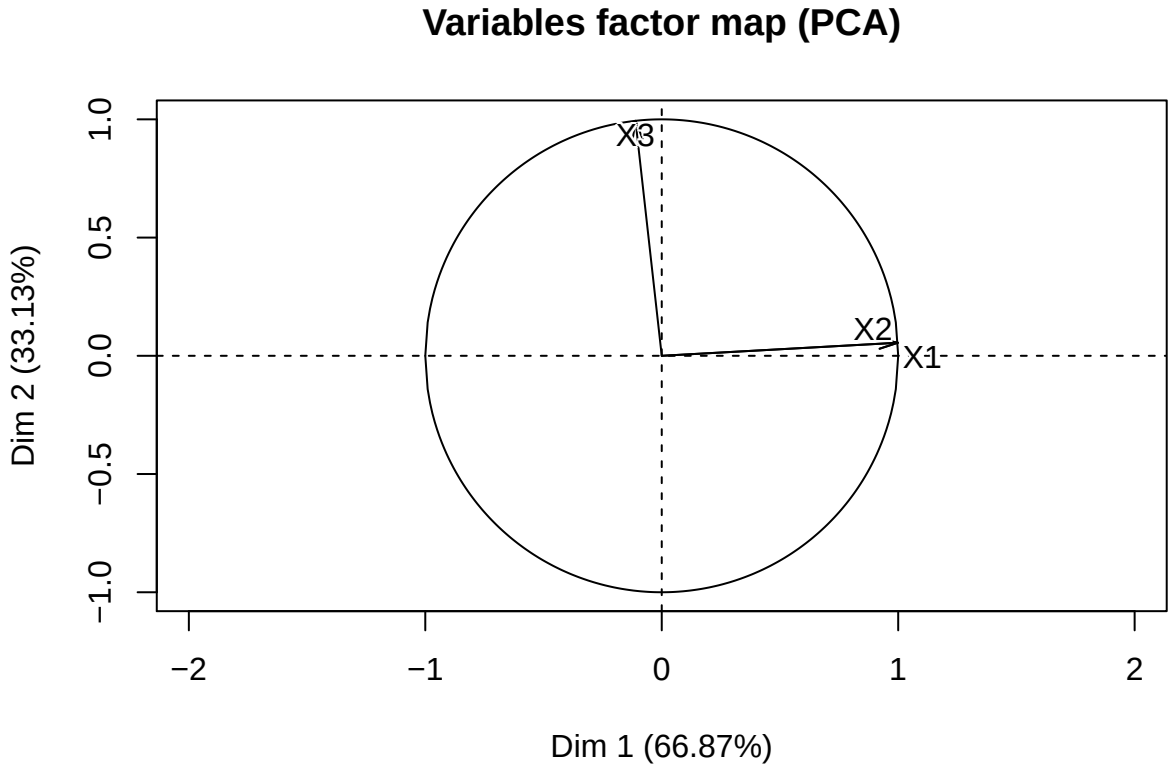
```
library(mvtnorm)
Z <- rmvnorm(n = 200, rep(0, 3), sigma = diag(3))
X1 <- Z[, 1]
X2 <- X1 + 0.001*Z[, 2]
X3 <- 10*Z[, 3]
don <- cbind.data.frame(X1, X2, X3)

library(FactoMineR)
res.pca <- PCA(don, scale = F)
```






```
res.pcascaled <- PCA(don, scale = T)
```





We must center and scale the data when the variables have different units, such as height (cm) and weight (kg). This prevents the main directions from being governed by one or more variables with a higher variance than the other variables. When the variables have the same unit, the reduction should be considered on a case-by-case basis. Sometimes we may not want to force variances to be equal (for example, with respect to grades, we may want to discriminate more against students in relation to a subject whose grades vary more).

2) TRUE or FALSE. If you perform a standardized PCA on a huge number of variables: the percentage of variability of the two first dimensions is small.

There is no general rule for this. If we generate independent Gaussian variables, we observe that the percentage of variability decreases with the dimension (see slides). Instead, consider a data set where we copy the same variable into each column. The percentage variability of the 2 dim will then be 100% regardless of the number of columns.

3) A scaled PCA has been performed on 4 data sets. Link correlation matrices and PCA results.

Correlation matrix A				
1.00	-0.46	0.54	-0.10	-0.14
-0.46	1.00	-0.55	-0.14	-0.10
0.54	-0.55	1.00	0.04	0.04
-0.10	-0.14	0.04	1.00	0.94
-0.14	-0.09	0.04	0.94	1.00

Correlation matrix B				
1.00	0.95	0.95	0.94	0.94
0.95	1.00	0.94	0.95	0.95
0.95	0.94	1.00	0.93	0.92
0.94	0.95	0.93	1.00	0.94
0.94	0.95	0.92	0.94	1.00

Correlation matrix C				
1.00	0.59	0.53	0.56	0.56
0.59	1.00	0.55	0.59	0.72
0.53	0.55	1.00	0.47	0.50
0.56	0.59	0.47	1.00	0.51
0.56	0.72	0.50	0.51	1.00

Correlation matrix D				
1.00	-0.01	0.04	-0.06	0.07
-0.01	1.00	-0.07	0.09	0.01
0.04	-0.07	1.00	0.23	0.29
-0.06	0.09	0.23	1.00	0.17
0.07	0.01	0.29	0.17	1.00

PCA 1			
	Eigenvalue	%inertia	% inertia
1	4.77	95.39	95.39
2	0.08	1.63	97.02
3	0.06	1.21	98.23
4	0.05	0.90	99.13
5	0.04	0.87	100

PCA 2			
	Eigenvalue	% inertia	% inertia
1	1.47	29.31	29.31
2	1.08	21.64	50.95
3	1.00	20.01	70.96
4	0.77	15.49	86.45
5	0.68	13.55	100

PCA 3			
	Eigenvalue	%inertia	% inertia
1	3.24	64.86	64.86
2	0.54	10.84	75.70
3	0.52	10.40	86.10
4	0.43	8.50	94.60
5	0.27	5.40	100

PCA 4			
	Eigenvalue	% inertia	% inertia
1	2.05	40.95	40.95
2	1.964	39.27	80.22
3	0.498	9.96	90.18
4	0.431	8.61	98.79
5	0.06	1.21	100

A - 4; B - 1; C - 3; D - 2.

The value of the first eigenvalue approximately gives the number of correlated variables explained by the first dimension.

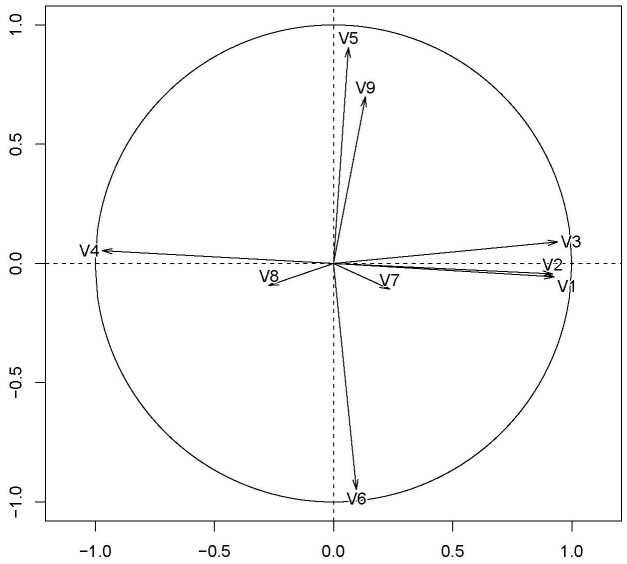
a. The matrix A seems to be a block diagonal matrix with blocks of size $3x3$ and $2x2$. The block of size $2x2$ is close to the matrix with only ones (which is of rank 1) thus one eigenvalue must be close to zero. This corresponds to PCA 4.

b. The matrix B is has all its entries near one, thus it is close to a matrix of rank one. If this matrix was of rank one, the variance could be explained with the first component only. In that case, the inertia associated with the first component would be 5 and 0 for the other components. This corresponds to PCA 1.

c. The matrix D is close to the identity. The PCA applied to the identity would give the same inertia to all components (inertia close to 1). This corresponds to PCA 2.

d. The matrix C is similar to B but the correlation are weaker. Thus the first component has a high inertia. This corresponds to PCA 3.

4) What is the percentage of variability of the first dimension? The first plane?

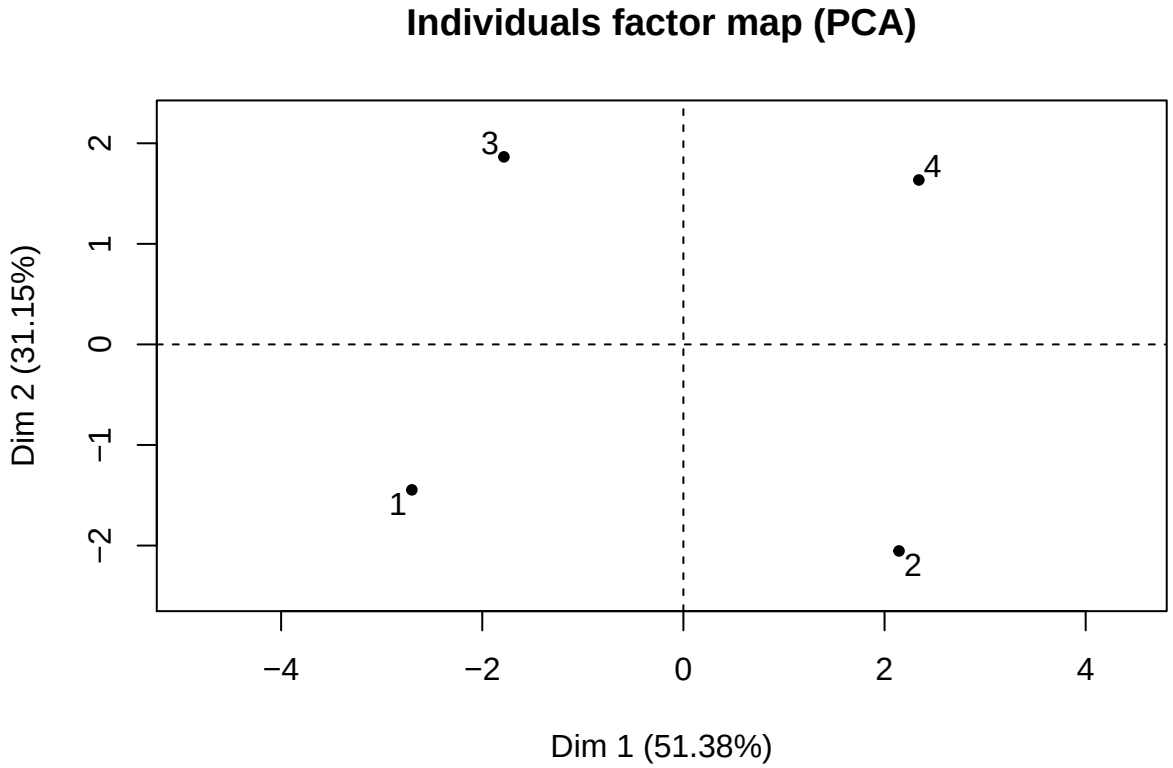


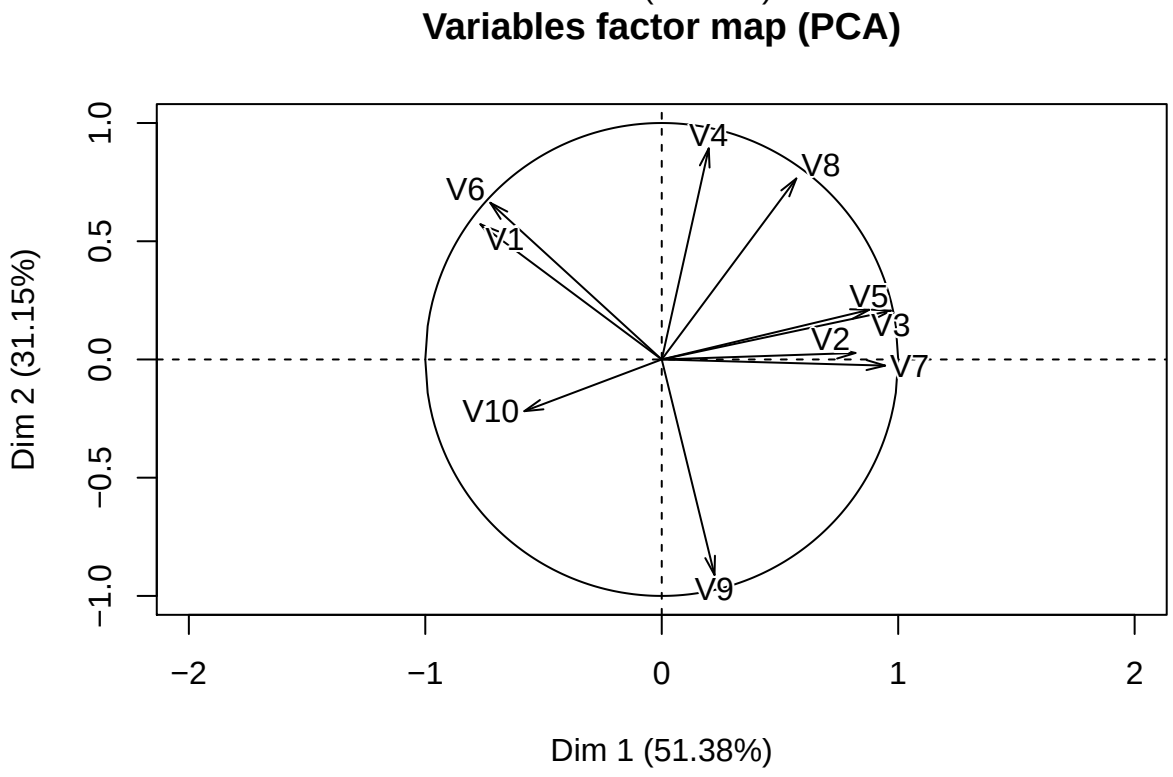
We look at the number of variables aligned with the first two dimensions and divide by the total number of variables. The first dimension explains 44% of the variability and the first plane 77%. The percentage of explained variance is equal to the sum of square norms of projection of variables along the first axis. Variables V5, V6, V9 are orthogonal to the first axis and variables V7 and V8 are not well represented by the first plan (and thus are orthogonal to this plan and in particular to the first axis). Roughly, the norm of V1, V2, V3 and V4 is one which leads to a percentage of variance for the first axis equal to 4/9. Similarly, for the first plan, we have a percentage of variance equal to 7/9 since the norm of $V5$ and $V9$ is close to one.

5) We have a data table with 4 rows and 10 independent variables. Without performing the PCA, do we have an idea of the shape of the correlation circle obtained? Simulate and comment. (You can increase the number of rows, also look at cases with 10 rows and 2000 variables, etc..)

The idea here is reminiscent of what you can see if you compute a correlation coefficient between two independent variables but on a small sample size. You can have large values even if the variables are orthogonal. You can also imagine that you have 10 variables in R^4 and they are going to be like tins in a can. So you will have the feeling to see correlations even though there is nothing.

```
nr = 4
nc = 10
mat=matrix(rnorm(nr*nc,0,1),ncol=nc)
res.pca <- PCA(mat)
```





6) Write a program to obtain the tables of the lecture slides. "Percentage of variability under the null".

```
nr=50
nc=8
iner=rep(0,1000)
for (i in 1:1000)
```

```
{
mat=matrix(rnorm(nr*nc,0,1),ncol=nc)
iner[i]=PCA(mat,graph=F)$eig[2,3]
}
quantile(iner,0.95)
```

```
##      95%
## 41.44712
```

Here the code to get everything:

```
simu <- function(n, p){
  res <- lapply(1:10000, function(i) rmvnorm(n, rep(0,p)))
  res.pca <- lapply(res, function(x) PCA(x, graph=FALSE))
  inertia <- sapply(1:10000, function(j) res.pca[[j]]$eig[2,3])
  return(quantile(unlist(inertia), 0.95))
}
ind.idx <- c(4:20, c(25,30,35,40,50,75,100,150,200))
var.idx <- c(4:20, c(25,30,35,40,50,75,100,150,200))

ind.idx <- c(10, 20)
var.idx <-c(5,30)

table <- sapply(ind.idx, function(n) sapply(var.idx, function(p) simu(n,p)))
library(knitr)
kable(table, caption="95 % quantile inertia on the two first dimensions of 10000 PCA on data with indepe
```

Table 1: 95 % quantile inertia on the two first dimensions of 10000 PCA on data with independent variables

95%	77.38358	66.14080
95%	43.04813	29.14502

```
PercentageNull <- function(nr, nc){
nbsim = 100
  iner=rep(0,nbsim)
for (i in 1:nbsim)
{
mat=matrix(rnorm(nr*nc,0,1),ncol=nc)
iner[i]=PCA(mat,graph=F)$eig[2,3]
}
return(quantile(iner,0.95))
}
```

PCA on decathlon data

```
# Install the package
#install.packages("FactoMineR", dependencies = TRUE)
# Load the package
library(FactoMineR)
data(decathlon)
head(decathlon)
```

This dataset contains the results of decathlon events during two athletic meetings which took place one month apart in 2004: the Olympic Games in Athens (23 and 24 August), and the Decastar 2004 (25 and 26 September). For both competitions, the following information is available for each athlete: performance for each of the 10 events, total number of points (for each event, an athlete earns points based on performance; here the sum of points scored) and final ranking. The events took place in the following order: 100 meters, long jump, shot put, high jump, 400 meters (first day) and 110 meter hurdles, discus, pole vault, javelin, 1500 meters (second day). Nine athletes participated to both competitions. We would like to obtain a typology of the performance profiles.

1) You should inspect the data set with the following commands:

```
summary(decathlon)
dim(decathlon)
#View(decathlon)
#?decathlon
rownames(decathlon)
```

2) Explain the interest of centering and scaling the data. Could you spotlight outstanding athletes ? Which inequality are you using?

When the data are standardized, it is possible to compare two variables with different units and to say sentences such as "Paul is more remarkable by his height than John is by is weight". When looking at the standardized data, we can look for the values greater or smaller than 2 for instance. We are refering to Bienaymé-Tchebychev which states that 25% of the observations will be at 2 standard deviation from their means. If we consider Gaussian data, we can refine this inequality and consider know that 4.5 % of the observations are greater than 2 in absolute value. Sebrle value for Javeline is 2.528251350 meaning that he is far above average.

The aim of conducting PCA on this dataset is to determine profiles for similar performances: are there any athletes who are better at endurance events or those requiring short bursts of energy, etc? And are some of the events similar? If an athlete performs well in one event, will he necessarily perform well in another?

3) Explain your choices for the active and illustrative variables/individuals and perform the PCA on this data set.

```
res.pca <- PCA(decathlon, quanti.sup = c(11,12) , quali.sup = 13)
summary(res.pca, nbelements = 5, ncp = 4)
barplot(res.pca$eig[,1], main = "Eigenvalues",
names.arg = paste("Dim",1:nrow(res.pca$eig), sep=""))
plot(res.pca, choix = "ind", axes = c(3, 4))
plot(res.pca, choix = "var", axes = c(3, 4))
```

To obtain a typology of the athletes based on their performances for the 10 decathlon events, such as "two athletes are close as they have similar performance profiles", the distances between two athletes are defined on the basis of their performances in the 10 events. Thus, only the performance variables are considered active; the other variables (number of points, rank, and competition) are supplementary. Here, the athletes are all considered as active individuals.

4) Comment the percentage of variability explained by the two first dimensions. What would you like (a small percentage, a high percentage) and why?

The first two dimensions summarize 50% of the total inertia, i.e. 50% of the total variability of the cloud of individuals (or variables) is represented by the first 2d plane. The importance of this percentage cannot be assessed without taking into account the number of active individuals and active variables. It may be interesting to compare this percentage with that of the 0.95 -quantile distribution of the percentages obtained by simulating data of equivalent size on the basis of independent normal distribution. According to the table provided in the slides, this quantile obtained for 40 individuals and 10 variables is about 38%: although the percentage of 50% seems relatively low, it expresses a significant structure in the data. However, the total

variability cannot be summarized by the first two dimensions only. It may also be interesting to interpret components 3 and 4 for which the inertia is greater than 1 (this value is used as a reference because it represents, in the case of standardized variables, the contribution of a single variable). It can also be said that a high percentage means that the 10 events are redundant and that we do not need 10 sport events to rank athletes.

5) Comment:

- the correlation between the 100 m and long.jump
- the correlation between long.jump and Pole.vault
- can you describe the athlete Casarsa?
- the proximity between Sebrle and Clay
- the proximity between Schoenbeck and Barras
- The 100m and long.jump are negatively correlated: therefore, an athlete who runs 100 meters quickly will generally jump a long way. The variables 100m, 400m, and 110m hurdles are positively correlated, that is, some athletes perform well in all four events while others do not.
- Since long.jump is well represented in the first plan and Pole.vault is not, we can deduce that long.jump and Pole.vault are approximately orthogonal, meaning that the corresponding variables are roughly uncorrelated.

Overall, the variables relating to speed are negatively correlated with the first principal component while the variables shot put and long jump are positively correlated with this component. The coordinates of these active variables can be found in the object res.pca$var which also gives the representation quality of the variables (cosine squared) and their contributions to the construction of the components.

```
round(cbind(res.pca$var$coord[,1:4],res.pca$var$cos2[,1:4],
res.pca$var$contrib[,1:4]),2)
```

Bourguignon and Karpov have very different performance profiles since they are opposed according to the main axis of variability.

- Casarsa is located on the top left corner. The first dimension is highly correlated with the number of points: this indicates that he does not have a large number of points. The second dimension is correlated with the Shot.put, High.jump and Discus. This indicates that Casarsa had good results in these three sports. Remember that the second dimension is calculated orthogonally to the first. So Casarsa has good results in these three sports compared to other “bad” athletes.
- Sebrle and Clay are close to one another and both far from the center of gravity of the cloud of points. The quality of their projection is therefore good, and we can be certain that they are indeed close in the original space. This means that they have similar profiles in their results across all sports events.
- Schoenbeck and Barras are close to one another but they are also close to the center of gravity of the cloud of points. When looking at their cos2 they are not well projected, We cannot interpret their distance based on this plot only.

The coordinates of these individuals can be found in res.pca$ind. Below is a sample output provided by the PCA function. We obtain a data table with the coordinates, the cosine squared (which gives an idea of the quality of the projection of the individuals on a component), and the contributions for each individual (to know how much an individual contributes to the construction of a component).

```
round(cbind(res.pca$ind$coord[,1:4],res.pca$ind$cos2[,1:4],
res.pca$ind$contrib[,1:4]),2)
```

6) Enhance the graphical outputs with the following options:

```
plot.PCA(res.pca, choix = "ind", habillage = ncol(decathlon), cex = 0.7)
plot.PCA(res.pca, choix = "ind", habillage = ncol(decathlon), cex = 0.7,
autolab = "no")
plot(res.pca, select = "cos2 0.8",  invisible = "quali")
plot(res.pca, select = "contrib 10")
plot(res.pca, choix = "var", select = "contrib 8", unselect = 0)
plot(res.pca, choix = "var", select = c("400m", "1500m"))
```

7) In which trials those who win the decathlon perform the best? Could we say that the decathlon trials are well selected?

The supplementary variable "number of points" is almost collinear to the first principal component. Therefore, the athletes with a high number of points are particularly good in the trials correlated with the first principal component. Those who win the decathlon perform the best in 100m, 110m hurdles and long jump. This means that the ranking of the decathlon is governed by those three sports.

8) Compare and comment the performances during both events: Decastar and Olympic. Could we conclude on the differences? Perform a test or plot Confidence ellipses:

```
dimdesc(res.pca)
```

```
## $Dim.1
## $Dim.1$quanti
##    correlation    p.value
## V3   0.9691672 0.03083284
##
##
## $Dim.3
## NULL
```

```
dimdesc(res.pca, proba = 0.2)
```

```
## $Dim.1
## $Dim.1$quanti
##    correlation    p.value
## V3   0.9691672 0.03083284
## V7   0.9447628 0.05523718
## V5   0.8773726 0.12262744
## V2   0.8193234 0.18067662
##
##
## $Dim.2
## $Dim.2$quanti
##    correlation    p.value
## V4   0.8926825 0.10731752
## V9  -0.9111279 0.08887214
```

```
plotellipses(res.pca)
```

```
## NULL
```

This function is very useful when there are a great number of variables. We see here that the first component is mainly due to the variable number of points (with a correlation coefficient of 0.96), and the variable 100m (with a negative correlation). The second component is described by two quantitative variables only (discus and shot put). No category of any categorical variable characterizes components 1 and 2 with a confidence level of 95%.

However, with the confidence level of 0.8, we can say that both of the two categories Olympic Games and

Decastar have coordinates that are significantly different from 0 on the first component. As the value is positive (negative) for the Olympic Games (Decastar) we can say that individuals who participated in the Olympic Games tended to have positive coordinates (or negative, respectively) on component 1.

9) To select the number of dimensions, you should have a look at

```
?estim_ncp
```

which performs cross validation as detailed in the lecture slides.

** More interpretation**

The representations of both the cloud of individuals and the cloud of variables are to be analysed together. In other words, differences between individuals can be explained by the variables, and relationships between variables can be illustrated by individuals. On the whole, the first component opposes performance profiles that are uniformly high" (i.e., athletes that are good in all events) such as Karpov at the Olympics to performance profiles that are (relatively!) weak in all events" such as Bourguignon at the Decastar meeting. Furthermore, the first component is mainly linked to the events using qualities relating to a burst of energy (100m, 400m, 110m hurdles and long jump). These four variables constitute a relatively homogeneous group: the correlation between any two of these performances is higher than 0.52 (see the correlation matrix). With just one exception, these variables have the highest coefficients. This group of variables draws" (i.e., contributes to the construction of) the first principal component and the overall score. It must here be emphasized that the first principal component is the combination that best sums up all the variables. In this example, the automatic summary provided by the PCA corresponds almost exactly with the official summary (the number of points). The second component opposes the variables of endurance (400m and 1500m) and power (discus, shot put). Notably, it separates the performance profiles that are considered weak“, which suggests that the best performance profiles are balanced: even among the weakest profiles, athletes can be specialised. Note that power (discus and shot put) is not correlated with speed (100m, long jump, 110m hurdles). As these two variables are not linearly related, there are powerful and fast individuals (all-round athletes with high values on the first component), powerful individuals who are not so fast (corresponding to high values on com- ponent 2) and individuals who are not powerful but fast (with low coordinates on component 2). This can be illustrated by comparing Casarsa and Lorenzo from the stan- dardised data (see the table on the next page). Casarsa performs well in the events that require power and poorly in speed related events, while the oppo- site is true for Lorenzo. It must be noted that these two athletes have a low coordinate on the first principal component and therefore do not have good overall performances. Their strengths, which are lessened by the second com- ponent, must therefore be relativized compared to their overall performance. The variable number of points seems to be entirely unrelated to this component (correlation of 0.02, see list of coordinates of the supplementary variables). The third component is mainly related to the 1500 meters and to a lesser extent, to the pole vault. It opposes these two events: athletes that do not perform well in the 1500 metres (N.B. as it is a variable relating to speed, a high value indicates rather poor performance) do however obtain good results in the pole vault (i.e., see standardized values in the centered and scaled data for Terek: 1.96 in the pole vault and 0.98 in the 1500 meters). This third component mainly highlights four individuals that are particularly weak in the 1500 meters: Clay and Karpov at Decastar (with standardised val- ues of 1.95 and 1.84, respectively), and Terek and Kokhizoglou at the Olympic Games (with standardised values of 0.98 and 3.29). These four individuals contribute up to 34.7% of the inertia of component 3. The fourth component is correlated with the variable javelin and, to a lesser extent, the variable pole vault. Three profiles are characterized by these two events: Bernard at the Decastar meeting, and Sebrle and Nool at the Olympic Games. These three athletes contribute up to 31.3% of the inertia of this component. It must be noted that the representations of the individuals and variables are only approximate representations of the data table on the one hand, and of the correlation (or variance{covariance) matrix on the other. It is therefore necessary to support the interpretation by referring back to the data by looking at the means the standard deviations by variable, the standardized data, and the correlation matrix.

All athletes who participated in both decathlons certainly focused their phys- ical preparation on their performances at the Olympic Games. Indeed, they all performed better at the Olympic Games than at the Decastar meeting. We can see that the dots representing a single athlete (for example, Sebrle) are in roughly the same direction. This means, for example, that Sebrle is good at the same events for both decathlons, but

that the dot corresponding to his performance at the Olympic Games is more extreme, so he obtained more points during the Olympics than at the Decastar meeting. This data can be interpreted in two different ways: 1. Athletes that participate in the Olympic Games perform better (on average) than those participating in the Decastar meeting. 2. During the Olympics, athletes are more motivated by the challenge, they tend to better, etc. From the point of view of component 2, however, there is no overall difference between the Olympics and Decastar. Overall, athletes' perfor- mances may have improved but their profiles have not changed. Only Zsivoczky changed from a rather powerful profile to a rather fast profile. His per- formances in shot put and javelin are worse at Decastar compared to the Olympics, with throws of 15.31m (standardised value of 1.02) and 13.48 (1:22) for shot put and 63.45m (1.08) and 55.37m (0:62) for javelin. However, he improved his performance in the 400 meters: 49.40 seconds and then 48.62 seconds and for the 110m hurdles: 14.95 seconds and then 14.17 seconds

Other application to practice

Chicken microarray data and fatty acid concentration

This is an experiment with 27 chickens with six diet conditions: normal diet (N), fasting for 16 hours (F16), fasting for 16 hours then refed for 5 hours (F16R5), fasting for 16 hours then refed for 16 hours (F16R16) fasting for 48 hours (F48), and fasting for 48 hours then refed for 24 hours (F48R24). At the end of the diet, the genes were analyzed using DNA chips, and the expression of 7407 genes retained for all the chickens. The data were then preprocessed in a standard manner for DNA chips (normalisation, eliminating the chip effect, etc.). The aim of the study is to see whether the genes are expressed differently depending on the situation of stress. More precisely, it may be interesting to see how long the chicken needs to be refed after fasting before it returns to a normal state, i.e., a state comparable to the state of a chicken with a normal diet. Might some genes be underexpressed during fasting and overexpressed during feeding? Data can be found at http://factominer.free.fr/book/chicken.csv