Exemples d'analyse en composantes principales

Christophe Ambroise

12/2/2020

Analyse des élèves (sans réduction)

Centrage du tableau de données

Les moyennes des cinq variables sont respectivement 9.67, 9.83, 10.22, 10.05 et 11. Le tableau centré en colonne $\boldsymbol{X}$ est obtenu en soustrayant à chaque colonne la moyenne correspondante :

```
X<- scale(X,center=TRUE,scale = FALSE)
knitr::kable(X,format="latex", caption = "Tableau avec centrage",digits = 2)
```

Matrice de variance

$$\boldsymbol{S} = \frac{1}{9}\boldsymbol{X}'\boldsymbol{X}$$

```
n<-nrow(X)
p<-ncol(X)
S<-var(X)*(n-1)/n
knitr::kable(S,format="latex", caption = "Matrice de variance",digits = 2)
```

Axes principaux d'inertie

La diagonalisation de la matrice de variance fournit les valeurs propres suivantes (rangées par ordre décroissant)

$$\lambda_1 = 28.2533, \lambda_2 = 12.0747, \lambda_3 = 8.6157, \lambda_4 = 0.0217, \lambda_5 = 0.0099.$$

Table 1: Notes de 9 <U+00E9>l<U+00E8>ves

	math	scie	fran	lati	d.m
jean	6.0	6.0	5.0	5.5	8
aline	8.0	8.0	8.0	8.0	9
annie	6.0	7.0	11.0	9.5	11
monique	14.5	14.5	15.5	15.0	8
didier	14.0	14.0	12.0	12.5	10
andre	11.0	10.0	5.5	7.0	13
pierre	5.5	7.0	14.0	11.5	10
brigitte	13.0	12.5	8.5	9.5	12
evelyne	9.0	9.5	12.5	12.0	18

Table 2: Tableau avec centrage

	math	scie	fran	lati	d.m
jean	-3.67	-3.83	-5.22	-4.56	-3
aline	-1.67	-1.83	-2.22	-2.06	-2
annie	-3.67	-2.83	0.78	-0.56	0
monique	4.83	4.67	5.28	4.94	-3
didier	4.33	4.17	1.78	2.44	-1
andre	1.33	0.17	-4.72	-3.06	2
pierre	-4.17	-2.83	3.78	1.44	-1
brigitte	3.33	2.67	-1.72	-0.56	1
evelyne	-0.67	-0.33	2.28	1.94	7

Table 3: Matrice de variance

	math	scie	fran	lati	d.m
math	11.39	9.92	2.66	4.82	0.11
scie	9.92	8.94	4.12	5.48	0.06
fran	2.66	4.12	12.06	9.29	0.39
lati	4.82	5.48	9.29	7.91	0.67
d.m	0.11	0.06	0.39	0.67	8.67

et les vecteurs propres normés ou axes principaux d'inertie suivants

$$\boldsymbol{u}_1 = \begin{pmatrix} 0.51 \\ 0.51 \\ 0.49 \\ 0.48 \\ 0.03 \end{pmatrix}, \boldsymbol{u}_2 = \begin{pmatrix} -0.57 \\ -0.37 \\ 0.65 \\ 0.32 \\ 0.11 \end{pmatrix}, \boldsymbol{u}_3 = \begin{pmatrix} -0.05 \\ -0.01 \\ 0.11 \\ 0.02 \\ -0.99 \end{pmatrix}, \boldsymbol{u}_4 = \begin{pmatrix} 0.29 \\ -0.55 \\ -0.39 \\ 0.67 \\ -0.03 \end{pmatrix}, \boldsymbol{u}_5 = \begin{pmatrix} -0.57 \\ 0.55 \\ -0.41 \\ 0.45 \\ -0.01 \end{pmatrix}.$$

Qualité de la représentation

- les inerties du nuage projeté sur les 5 axes sont égales aux valeurs propres.
- l'inertie du nuage est égale à trace(S), c'est-à-dire aussi à la somme des valeurs propres, ici 48.975.
- les pourcentages d'inertie expliquée par chaque axe sont donc de 57.69, 24.65, 17.59, 0.04 et 0.02.
- Les pourcentages d'inertie expliquée par les sous-espaces principaux sont 57.69, 82.34, 99.94, 99.98 et 100.00.
- le nuage initial est pratiquement dans un espace de dimension 3.

Composantes principales $C = \boldsymbol{X}U$

```
U<-eigen(S)$vectors ; Lambda<-eigen(S)$values ; C = X%*%U
knitr::kable(C,format="latex",
             caption = "Composantes principales",digits = 2)
```

Ces composantes principales permettent d'obtenir, par exemple, les plans de représentation l,2 et 1,3 suivants

Table 4: Composantes principales

jean	-8.70	1.70	2.55	-0.15	-0.12
aline	-3.94	0.71	1.81	-0.09	0.04
annie	-3.21	-3.46	0.30	0.17	0.02
monique	9.76	-0.22	3.34	-0.17	0.10
didier	6.37	2.17	0.96	0.07	-0.19
andre	-2.97	4.65	-2.63	-0.02	0.15
pierre	-1.05	-6.23	1.69	0.12	0.04
brigitte	1.98	4.07	-1.40	0.24	0.01
evelyne	1.77	-3.40	-6.62	-0.16	-0.06

Table 5: Contribution relative des axes aux individus

jean	0.89	0.03	0.08	0	0
aline	0.80	0.03	0.17	0	0
annie	0.46	0.53	0.00	0	0
monique	0.89	0.00	0.11	0	0
didier	0.88	0.10	0.02	0	0
andre	0.24	0.58	0.19	0	0
pierre	0.03	0.91	0.07	0	0
brigitte	0.17	0.74	0.09	0	0
evelyne	0.05	0.20	0.75	0	0

Contributions relatives des axes aux individus

```
COR<- C^2 / rowSums(X^2)
knitr::kable(COR,format="latex",
             caption = "Contribution relative des axes aux individus",
             digits = 2)
```

Contributions relatives des individus aux axes

```
CTR<- 1/n* C^2 /  matrix(eigen(S)$values,n,p,byrow = TRUE)
knitr::kable(CTR,format="latex",
             caption = "Contributions relatives des individus aux axes",
             digits = 2)
```

Analyse dans R^n

Les vecteurs $\boldsymbol{d}^\alpha$, composantes principales associées aux différentes variables, sont formés des coordonnées de toutes les variables pour un même axe $\boldsymbol{v}_\alpha$ et vérifient la relation

$$\boldsymbol{d}^\alpha = \sqrt{\lambda_\alpha}\boldsymbol{u}_\alpha.$$

On obtient

```
D<- U * matrix(sqrt(Lambda),p,p,byrow=TRUE)
knitr::kable(D,format="latex",
             caption = "Variables",digits = 2)
```

Table 6: Contributions relatives des individus aux axes

jean	0.30	0.03	0.08	0.11	0.15
aline	0.06	0.00	0.04	0.04	0.02
annie	0.04	0.11	0.00	0.15	0.00
monique	0.37	0.00	0.14	0.15	0.11
didier	0.16	0.04	0.01	0.03	0.40
andre	0.03	0.20	0.09	0.00	0.25
pierre	0.00	0.36	0.04	0.07	0.02
brigitte	0.02	0.15	0.03	0.30	0.00
evelyne	0.01	0.11	0.56	0.14	0.04

Table 7: Variables

2.73	1.97	-0.15	-0.04	0.06
2.69	1.29	-0.04	0.08	-0.05
2.62	-2.26	0.32	0.06	0.04
2.58	-1.12	0.07	-0.10	-0.05
0.16	-0.39	-2.91	0.01	0.00

Analyse dans R^n

Il est souvent préférable de représenter la projection des variables initiales normées. Il suffit de diviser chaque ligne du tableau précédent par la norme de la variables correspondante

$$\|\boldsymbol{x}^j\|^2 = \frac{1}{9}\sum_{i=1}^{9}(x_i^j)^2.$$

Les $\|\boldsymbol{x}^j\|$ correspondent en fait aux écarts-type des variables. On obtient respectivement 3.37, 2.99, 3.47, 2.81 et 2.94

```
F<- D / sqrt((1/n*colSums(X^2)))
knitr::kable(F,format="latex",
             caption = "Variables normées",digits = 2)
```

L'ACP avec FactoMineR

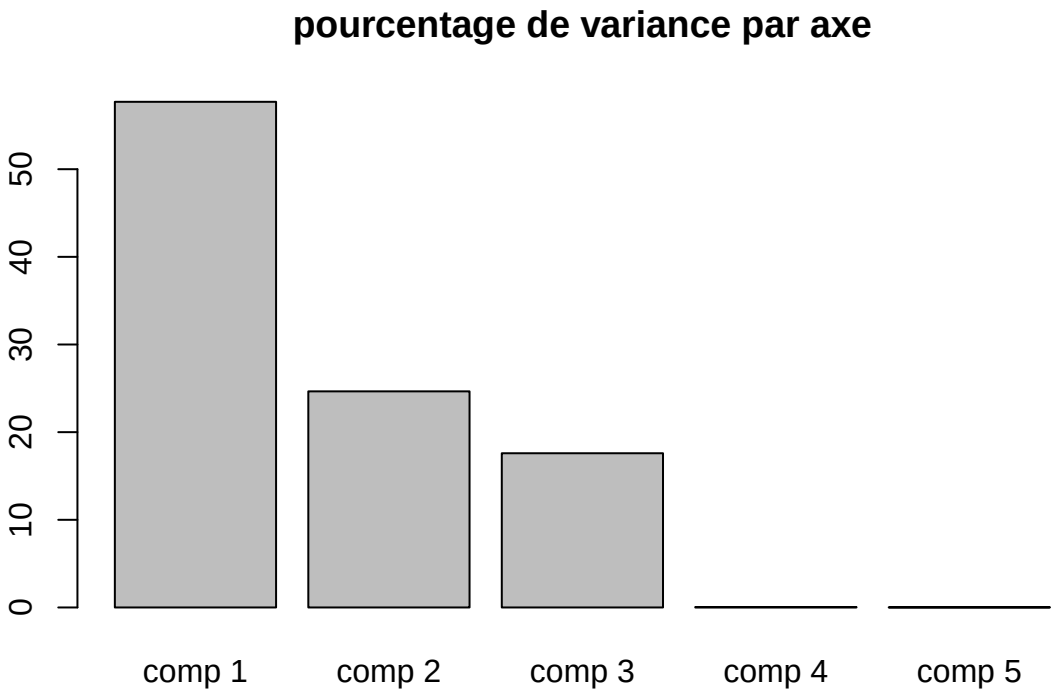
```
library(FactoMineR)
res.pca<-PCA(X, scale.unit=FALSE, ncp=5, graph=FALSE)
```

Table 8: Variables norm<U+00E9>es

0.81	0.58	-0.04	-0.01	0.02
0.90	0.43	-0.01	0.03	-0.02
0.75	-0.65	0.09	0.02	0.01
0.92	-0.40	0.02	-0.04	-0.02
0.06	-0.13	-0.99	0.00	0.00

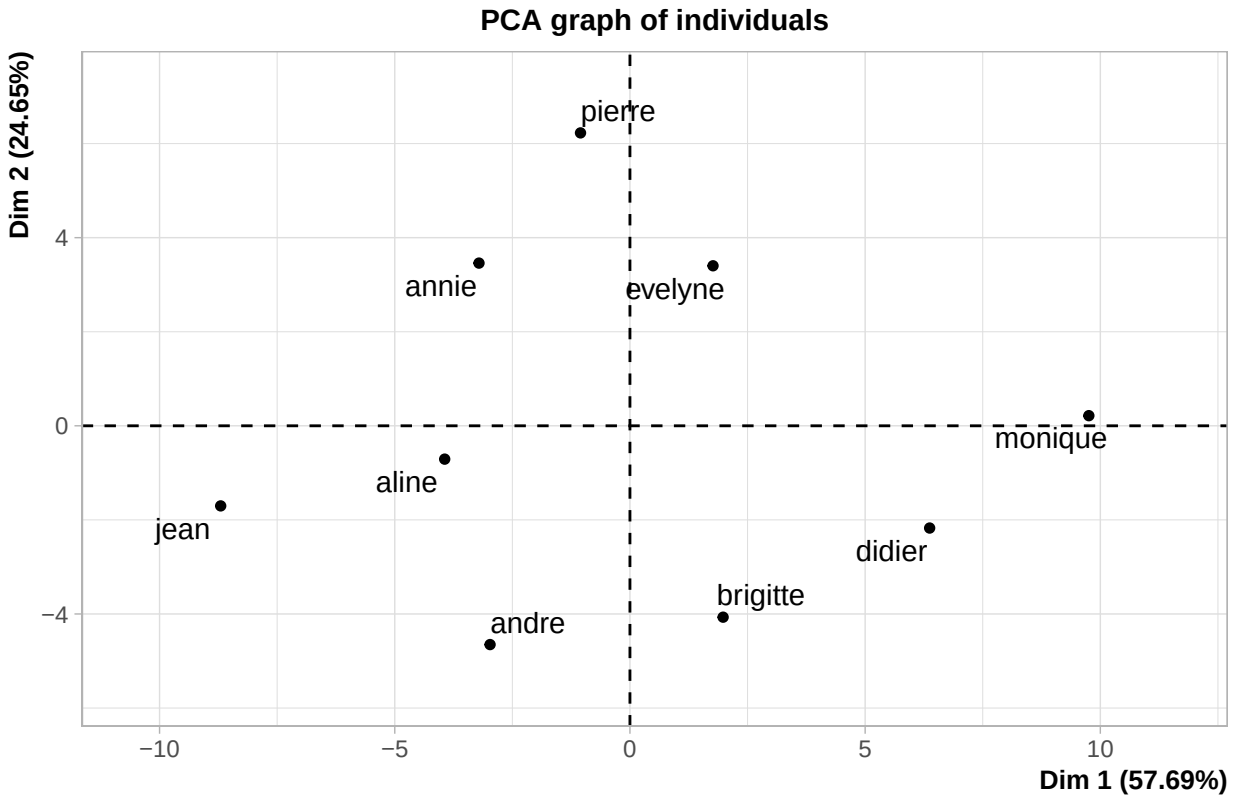
Variances expliquées

```
eigvalues<-data.frame(res.pca$eig)
barplot(eigvalues$percentage.of.variance, names.arg=row.names(eigvalues),main='pourcentage de variance p
```

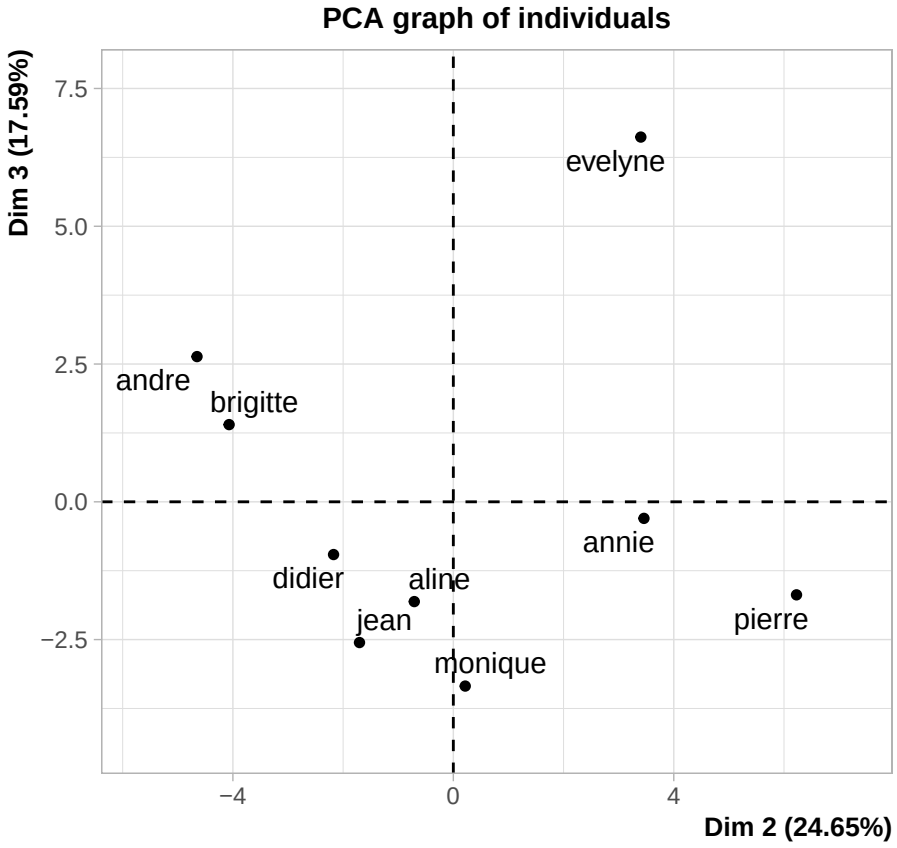


Représentation des individus

```
plot(res.pca,choix="ind",axes=1:2)
```

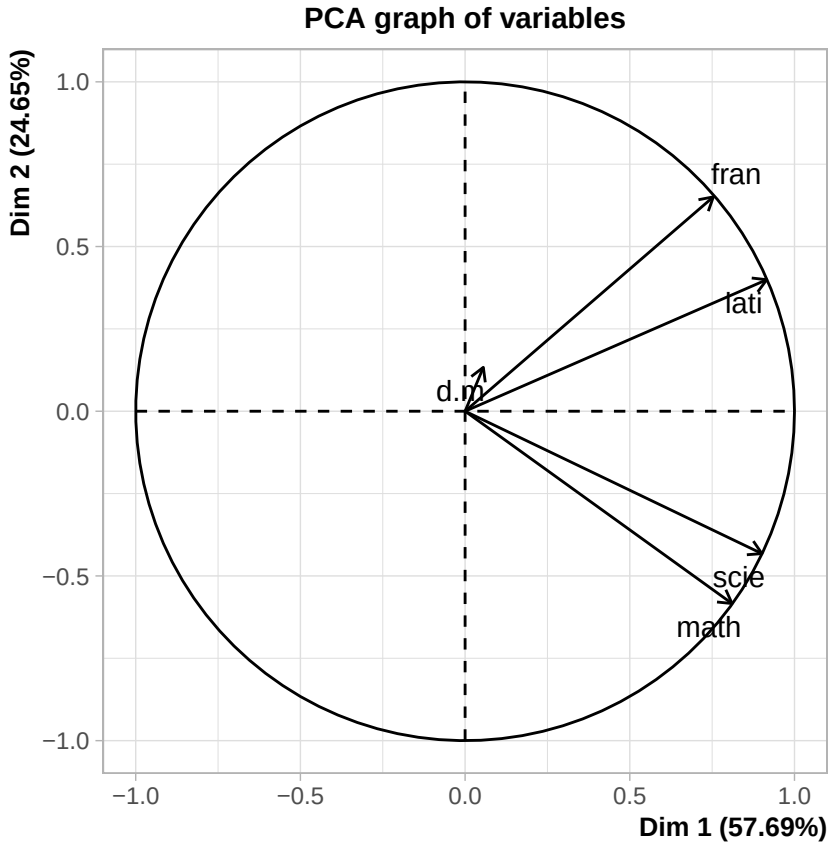



```
plot(res.pca,choix="ind",axes=2:3)
```



Représentation des variables

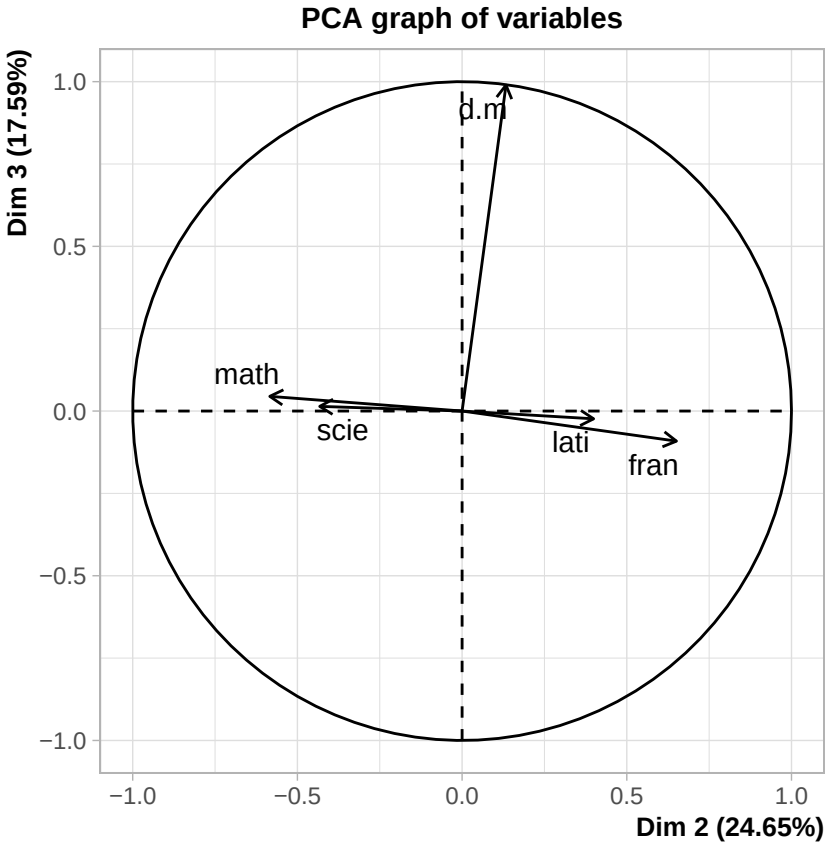
```
plot(res.pca,choix="varcor",axes=1:2)
```




```
plot(res.pca,choix="varcor",axes=2:3)
```

Table 9: Contribution relative des axes aux individus

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
jean	0.89	0.03	0.08	0	0
aline	0.80	0.03	0.17	0	0
annie	0.46	0.53	0.00	0	0
monique	0.89	0.00	0.11	0	0
didier	0.88	0.10	0.02	0	0
andre	0.24	0.58	0.19	0	0
pierre	0.03	0.91	0.07	0	0
brigitte	0.17	0.74	0.09	0	0
evelyne	0.05	0.20	0.75	0	0



Contribution relative des axes aux individus

```
knitr::kable(res.pca$ind$cos2,format="latex",
             caption = "Contribution relative des axes aux individus",
             digits = 2)
```

Contribution relative des axes aux individus

Table 10: Contribution des individus aux axes

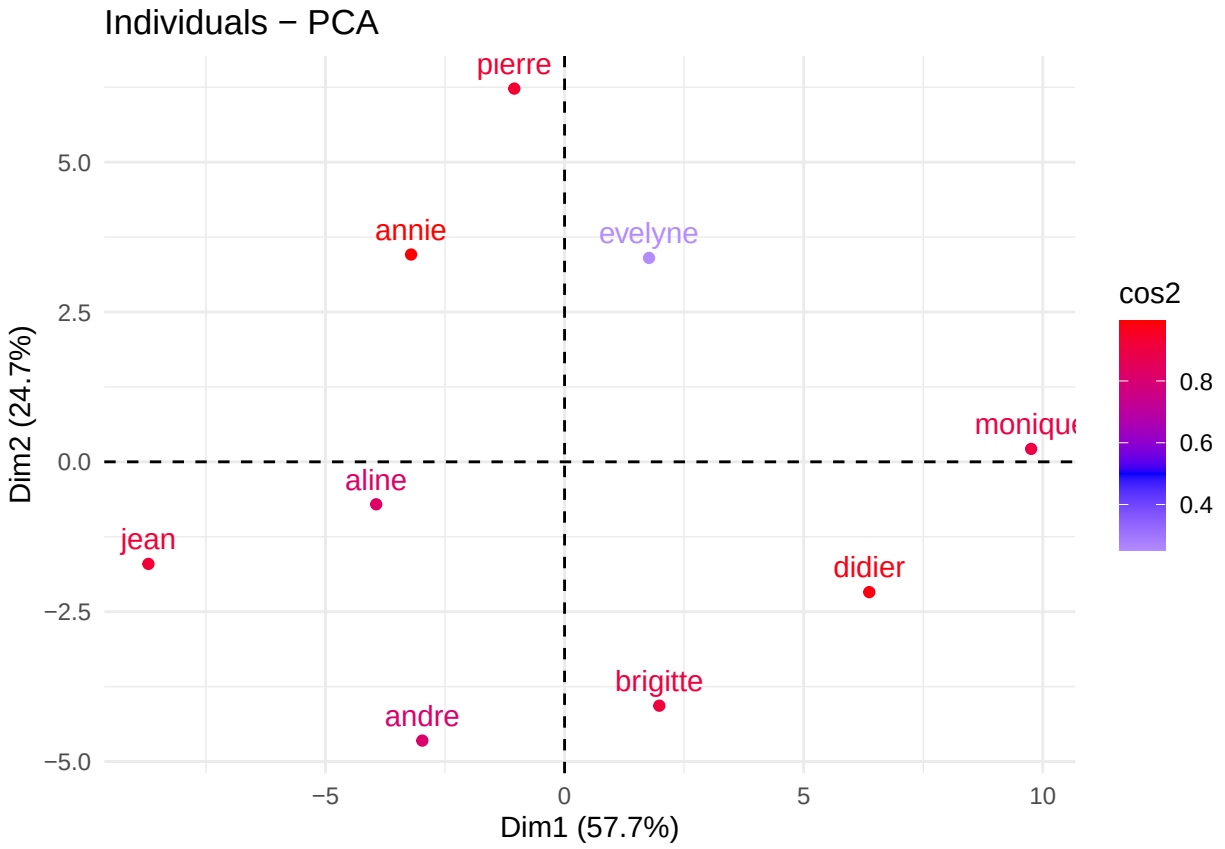
	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
math	26.47	32.14	0.26	8.34	32.78
scie	25.70	13.84	0.02	30.59	29.85
fran	24.24	42.30	1.17	15.50	16.79
lati	23.49	10.45	0.05	45.45	20.56
d.m	0.09	1.27	98.50	0.12	0.02

```
library(factoextra)
```

```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
fviz_pca_ind(res.pca, col.ind="cos2") + scale_color_gradient2(low="white", mid="blue",
high="red", midpoint=0.50) + theme_minimal()
```

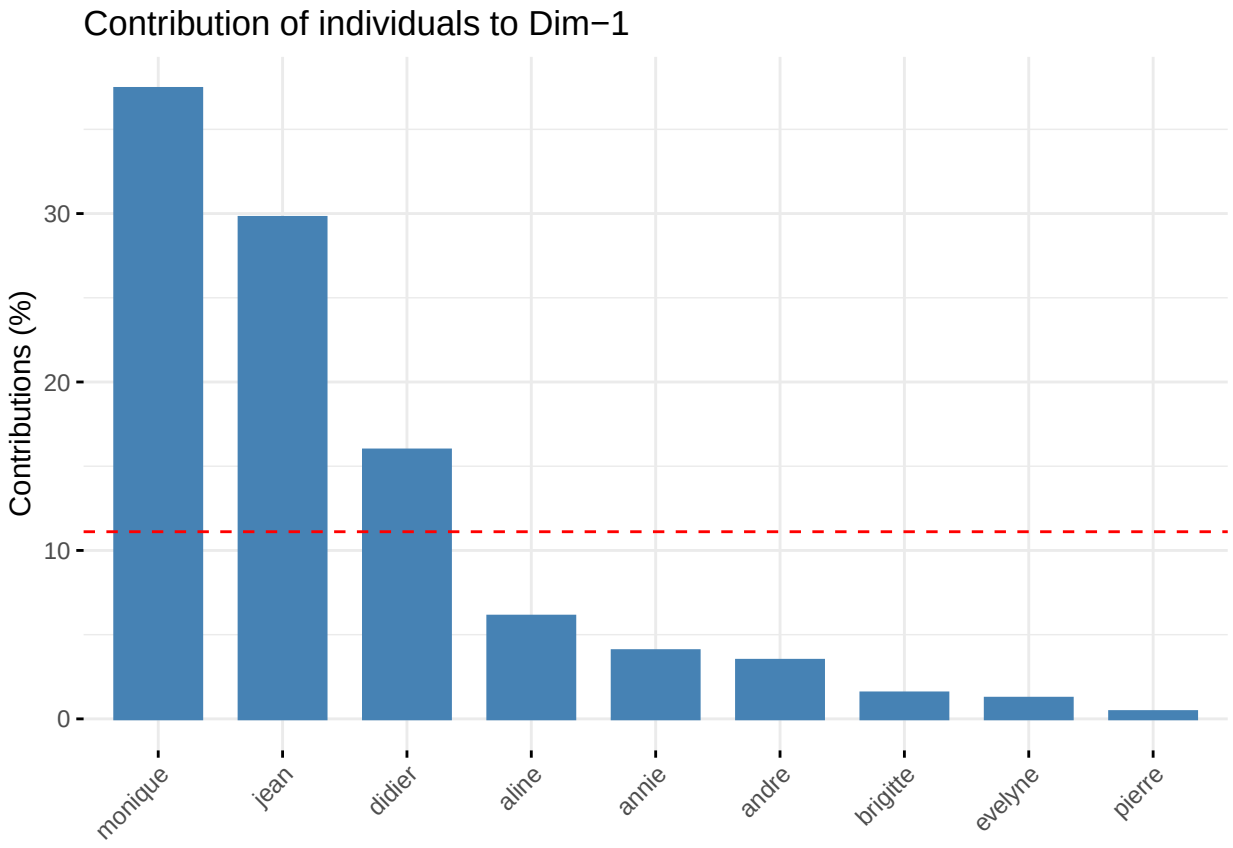


Contribution des individus aux axes

```
knitr::kable(res.pca$var$contrib,format="latex",
             caption = "Contribution des individus aux axes",
             digits = 2)
```

Contribution des individus aux axes

```
fviz_contrib(res.pca, choice = "ind", axes = 1)
```



Analyse des crabes (avec et sans transformation)

Chargement des données

```
library(MASS)
data(crabs)
n=dim(crabs)[1]
```

Gardons les variables quantitatives

```
crabsquant<-crabs[,4:8]
```

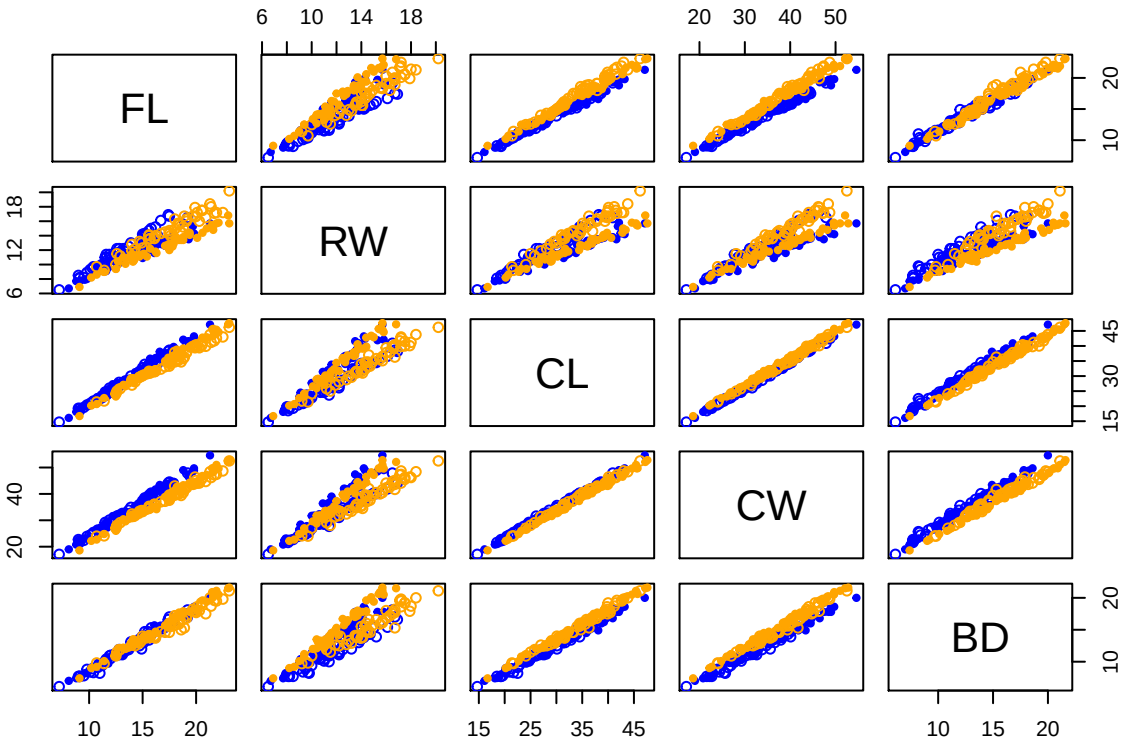
Visualisation des corrélations

```
cor(crabsquant)
```

```
##           FL        RW        CL        CW        BD
## FL 1.0000000 0.9069876 0.9788418 0.9649558 0.9876272
## RW 0.9069876 1.0000000 0.8927430 0.9004021 0.8892054
## CL 0.9788418 0.8927430 1.0000000 0.9950225 0.9832038
```

```
## CW 0.9649558 0.9004021 0.9950225 1.0000000 0.9678117
## BD 0.9876272 0.8892054 0.9832038 0.9678117 1.0000000
```

```
pairs(crabsquant,col=c("blue","orange")[crabs$sp],pch=c(21,20)[crabs$sex])
```



ACP

ACP brute

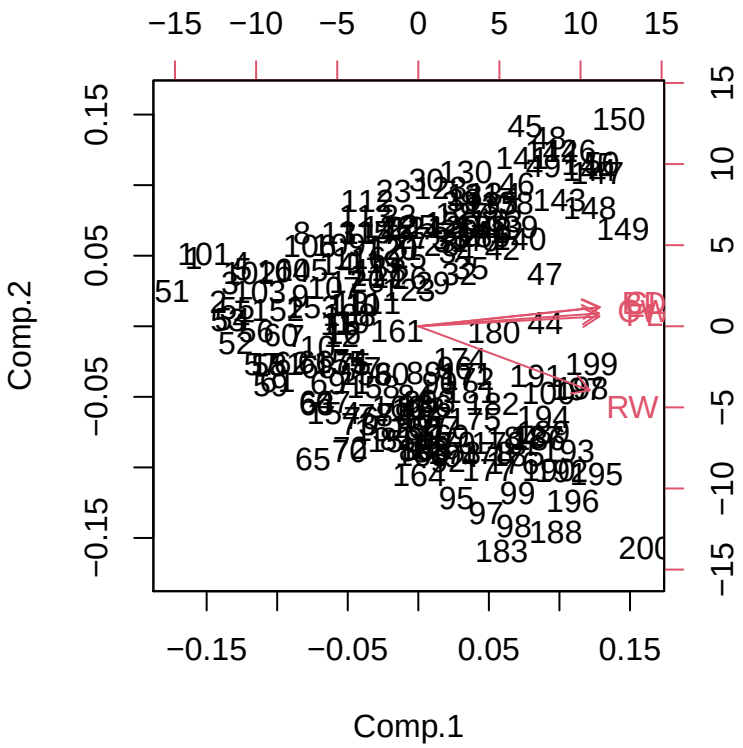
```
res<-princomp(scale(crabsquant))
```

Le premier axe capture toute la variation du nuage, et les autres n'ont plus rien à dire. On le voit bien si on regarde les valeurs propres: la première composante capture 95% de l'inertie du nuage !

```
summary(res)
```

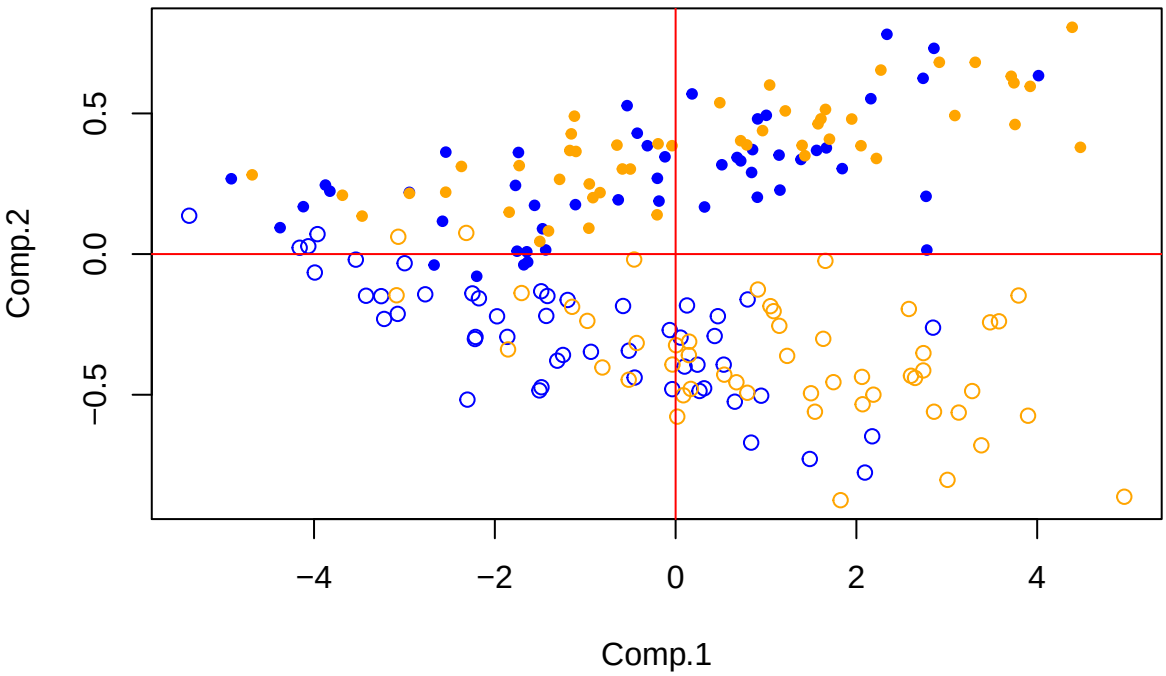
```
## Importance of components:
##                           Comp.1     Comp.2      Comp.3      Comp.4       Comp.5
## Standard deviation     2.182863 0.38849296 0.215406149 0.105260061 0.0412688656
## Proportion of Variance 0.957767 0.03033704 0.009326595 0.002227071 0.0003423355
## Cumulative Proportion  0.957767 0.98810400 0.997430593 0.999657664 1.0000000000
```

```
biplot(res)
```



L'effet taille est ici très visible. Toutes les variables sont très corrélées, et l'on ne départage pas bien du tout les crabes, ni du point de vue de l'espèce.

```
plot(res$scores[,1:2],col=c("blue","orange")[crabs$sp],pch=c(21,20)[crabs$sex])
abline(h=0,v=0,col="red")
```



Si l'on veut la suppression de l'effet taille il suffit simplement de regarder les composantes principales qui sont au dela de la première

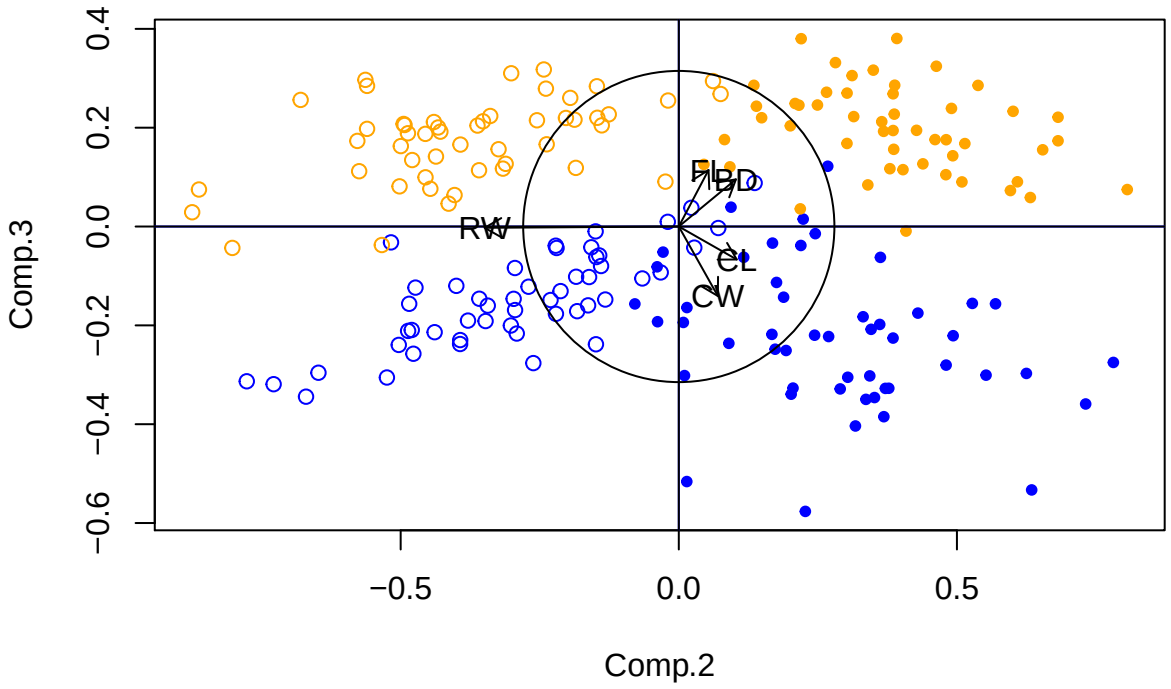
```r
# Corrélations variables-facteurs principaux
rho2 <- res$loadings[,2] * res$sdev[2]
rho3 <- res$loadings[,3] * res$sdev[3]
corr <- cbind(rho2,rho3)
print(corr,digits=2)
```

```
##      rho2    rho3
## FL  0.053  0.1143
## RW -0.349 -0.0026
## CL  0.104 -0.0667
## CW  0.070 -0.1406
## BD  0.103  0.0955
```

```r
plot(res$scores[,2:3],col=c("blue","orange")[crabs$sp],pch=c(21,20)[crabs$sex])
abline(h=0,v=0,col="blue")
#plot(c(-1,1),c(-1,1),type="none")
arrows(0,0,rho2,rho3, xlim=c(-1,1), ylim=c(-1,1), type = "n",length=0.1)
```

```
## Warning in arrows(0, 0, rho2, rho3, xlim = c(-1, 1), ylim = c(-1, 1), type =
## "n", : graphical parameter "type" is obsolete
```

```r
abline(h=0,v=0)
text(rho2,rho3, labels=names(crabsquant), cex=1)
symbols(0,0,circles=1,inches= F, add=T)
```



On voit clairement que la variable RW sépare les males et les femelles, et que les crabes bleus ont une large et longue carapace (CW et CL grands) alors que les crabes oranges ont une petite (CW et CL petits) carapace épaisse (BD grand).

ACP après transformation

Transformation des données

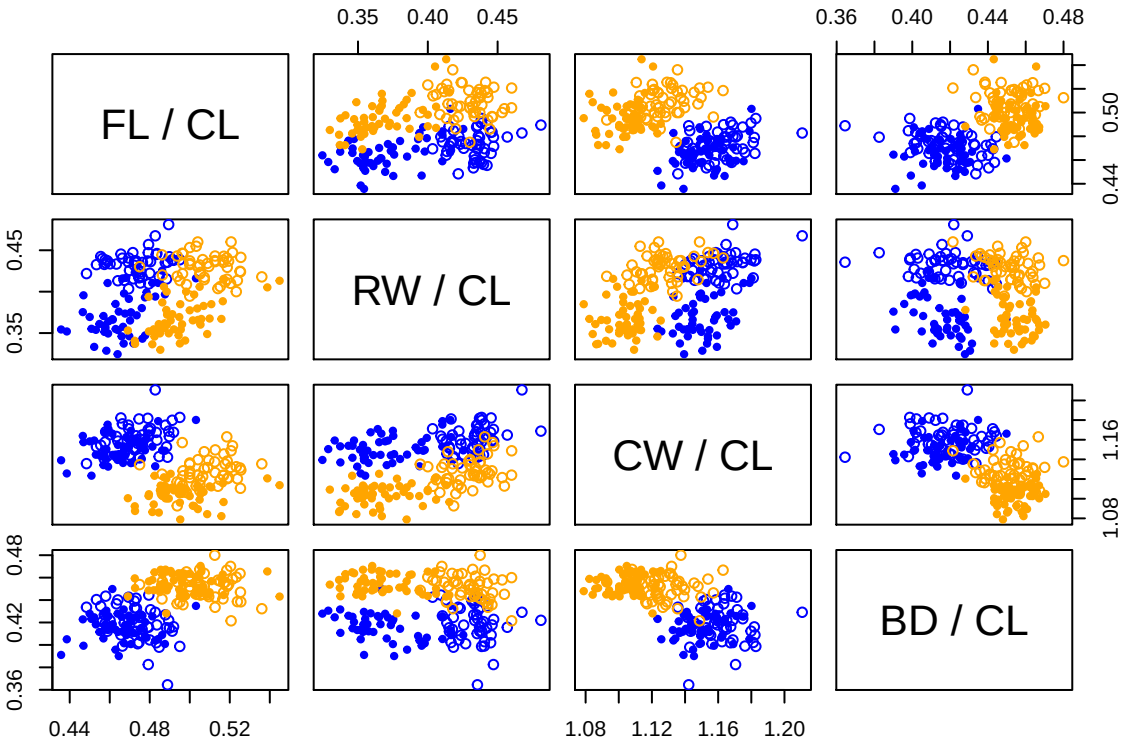
Une simple petite transformation peut aussi enlever l'effet taille, si toutes les variables sont renormalisées par la 3ème (variable la plus corrélée avec les autres).

```
crabsquant2<-(crabsquant/crabsquant[,3])[,-3]
```

```
j=0
for(i in c(1,2,4,5))
{
 j=j+1
 names(crabsquant2)[j]<-c(paste(names(crabsquant)[i],"/",names(crabsquant[3])))
}
```

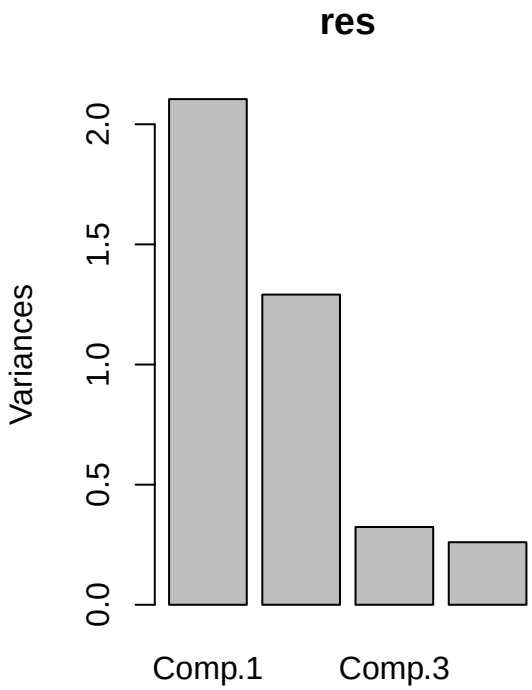
Visualisons à nouveau les corrélations entre variables:

```
pairs(crabsquant2,col=c("blue","orange")[crabs$sp],pch=c(21,20)[crabs$sex])
```



Effectuons une nouvelle fois l'ACP:

```
res<-princomp(scale(crabsquant2))
par(mfrow=c(1,2))
plot(res)
```



Analyse de la sortie

```
str(res)
```

```
## List of 7
##  $ sdev    : Named num [1:4] 1.451 1.136 0.569 0.51
##   ..- attr(*, "names")= chr [1:4] "Comp.1" "Comp.2" "Comp.3" "Comp.4"
##  $ loadings: 'loadings' num [1:4, 1:4] 0.51 -0.134 -0.589 0.613 0.485 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:4] "FL / CL" "RW / CL" "CW / CL" "BD / CL"
##   .. ..$ : chr [1:4] "Comp.1" "Comp.2" "Comp.3" "Comp.4"
##  $ center  : Named num [1:4] -9.48e-16 -5.07e-16 -8.39e-17 1.13e-15
##   ..- attr(*, "names")= chr [1:4] "FL / CL" "RW / CL" "CW / CL" "BD / CL"
##  $ scale   : Named num [1:4] 1 1 1 1
##   ..- attr(*, "names")= chr [1:4] "FL / CL" "RW / CL" "CW / CL" "BD / CL"
##  $ n.obs   : int 200
##  $ scores  : num [1:200, 1:4] -0.665 -1.129 -1.907 -1.242 -0.864 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:200] "1" "2" "3" "4" ...
##   .. ..$ : chr [1:4] "Comp.1" "Comp.2" "Comp.3" "Comp.4"
##  $ call    : language princomp(x = scale(crabsquant2))
##  - attr(*, "class")= chr "princomp"
```

On trouve les éléments suivants:

- [sdev] Ecarts types des composantes principales, soit la racine carrée des valeurs propres.
- [loadings] Matrice des vecteurs propres : axes principaux
- [center] Moyennes utilisées pour le centrage des données.
- [scale] Ecarts-types utilisés pour la réduction des données.
- [n.obs] Nombre d'observations
- [scores] Composantes principales

- [call] Rappel de l'appel fait à 'princomp".

Choix des axes

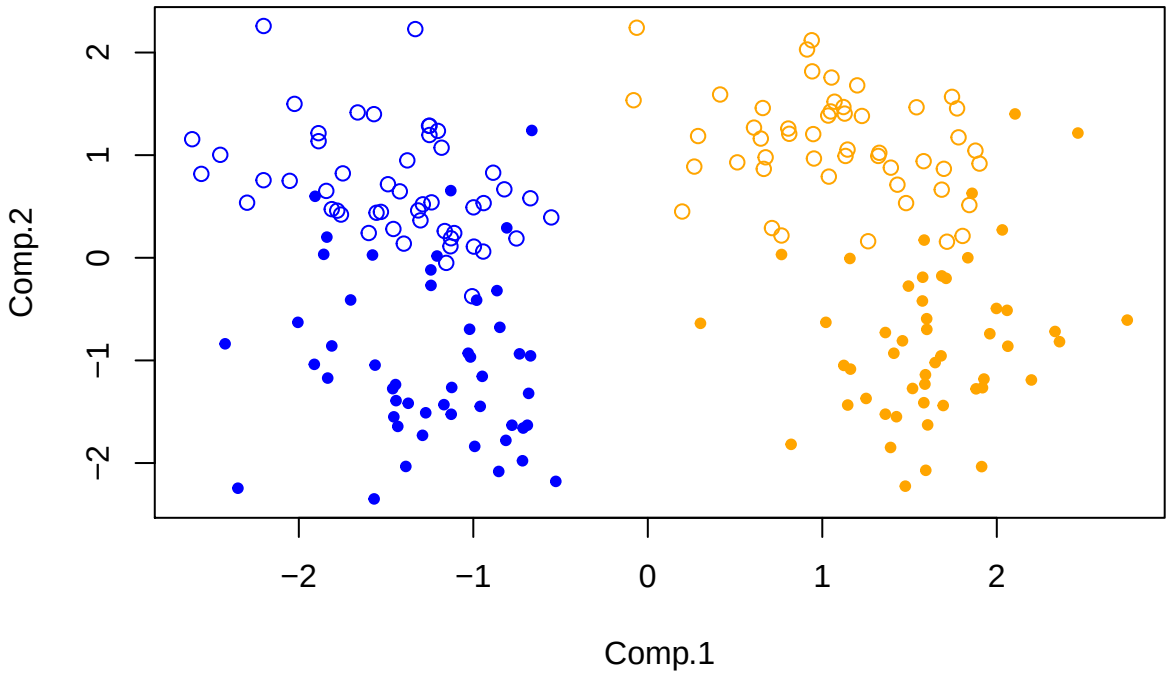
```
summary(res)
```

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3     Comp.4
## Standard deviation     1.4507641 1.1362749 0.56899999 0.51029581
## Proportion of Variance 0.5288233 0.3244022 0.08134698 0.06542759
## Cumulative Proportion  0.5288233 0.8532254 0.93457241 1.00000000
```

On conserve ici les deux premiers axes, qui portent tous les deux une variance supérieure à la moyenne (1). Au total, le plan des deux premiers facteurs représentera 85% de la variance totale.

Cette fois, ce plan parvient à très bien distinguer les espèces et les sexes. L'axe 1 sépare les 2 espèces, tandis que l'axe 2 sépare les femelles des mâles.

```
plot(res$scores[,1:2],col=c("blue","orange")[crabs$sp],pch=c(21,20)[crabs$sex])
```



Analyse des variables : construction du cercle des corrélations

```
# Corrélations variables-facteurs principaux
rho1 <- res$loadings[,1] * res$sdev[1]
rho2 <- res$loadings[,2] * res$sdev[2]
corr <- cbind(rho1,rho2)
print(corr,digits=2)
```

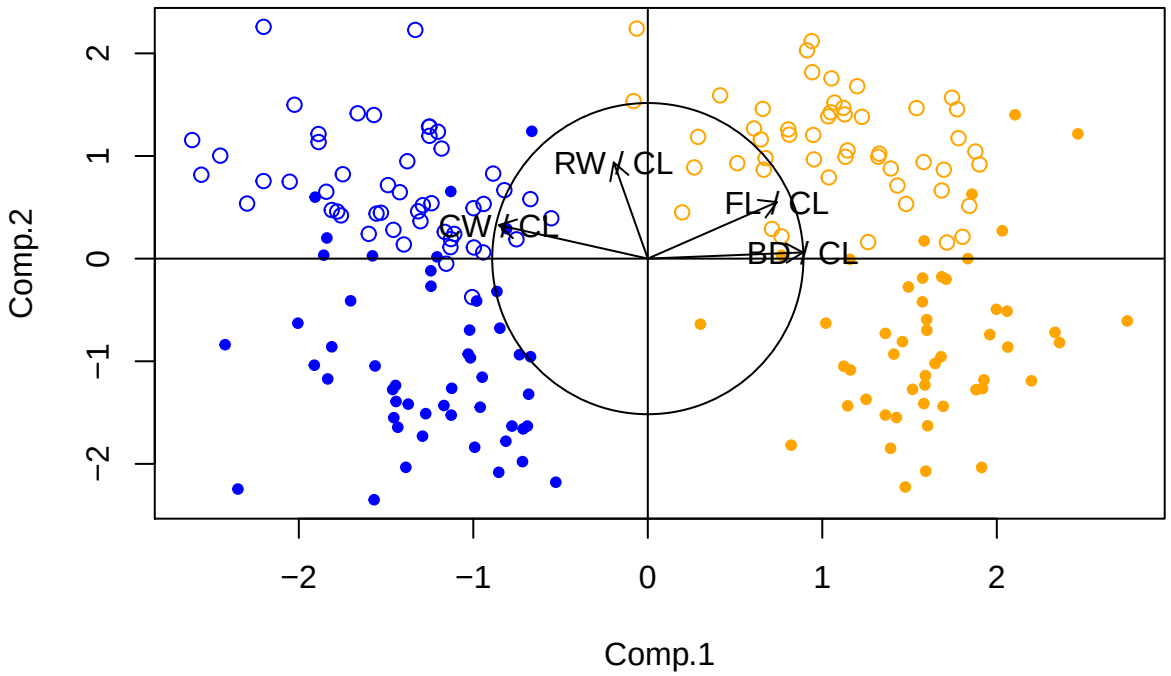
```
##          rho1  rho2
## FL / CL  0.74 0.551
## RW / CL -0.19 0.937
## CW / CL -0.85 0.326
```

```
## BD / CL  0.89 0.061
```

```
# Affichage du cercle des corrélations
plot(res$scores[,1:2],col=c("blue","orange")[crabs$sp],pch=c(21,20)[crabs$sex])
arrows(0,0,rho1,rho2, xlim=c(-1,1), ylim=c(-1,1), type = "n",length=0.1)
```

```
## Warning in arrows(0, 0, rho1, rho2, xlim = c(-1, 1), ylim = c(-1, 1), type =
## "n", : graphical parameter "type" is obsolete
```

```
abline(h=0,v=0)
text(rho1,rho2, labels=names(crabsquant2), cex=1)
symbols(0,0,circles=1,inches= F, add=T)
```



Le premier axe est positivement corrélé aux variables BD/CL et FL/CL, négativement corrélé à CW/CL. Le second axe est principalement (négativement) corrélé à RW/CL.

```
# Carrés des corrélations (cosinus carrés)
print(corr^2,digits=2)
```

```
##          rho1   rho2
## FL / CL 0.547 0.3031
## RW / CL 0.038 0.8781
## CW / CL 0.729 0.1063
## BD / CL 0.790 0.0037
```

```
# Cumul des carrés des corrélations
print(t(apply(corr^2,1,cumsum)),digits=2)
```

```
##          rho1 rho2
## FL / CL 0.547 0.85
## RW / CL 0.038 0.92
## CW / CL 0.729 0.84
## BD / CL 0.790 0.79
```

Analyse des individus

```
ctrb <- NULL
for (k in 1:2){
  ctrb <- cbind( ctrb,res$scores[,k]^2/res$sdev[k]^2/nrow(crabs))
}
o1 <-order(ctrb[,1],decreasing=T)
o2 <-order(ctrb[,2],decreasing=T)
best1 <- cbind(ctrb[o1,1],res$scores[o1,1],crabs$sp[o1],crabs$sex[o1])
best2 <- cbind(ctrb[o2,2],res$scores[o2,2],crabs$sp[o2],crabs$sex[o2])
print(best1[1:10,])
```

```
##           [,1]      [,2] [,3] [,4]
## 125 0.01793696  2.747807    2    2
## 59  0.01619717 -2.611148    1    1
## 52  0.01555194 -2.558610    1    1
## 105 0.01443306  2.464853    2    2
## 63  0.01426134 -2.450146    1    1
## 11  0.01393991 -2.422377    1    2
## 136 0.01322529  2.359470    2    2
## 35  0.01310833 -2.349013    1    2
## 137 0.01293975  2.333860    2    2
## 75  0.01252811 -2.296437    1    1
```

```
print(best2[1:10,])
```

```
##           [,1]      [,2] [,3] [,4]
## 28  0.02137378 -2.349303    1    2
## 70  0.01975185  2.258407    1    1
## 35  0.01952067 -2.245152    1    2
## 183 0.01945851  2.241575    2    1
## 65  0.01922960  2.228351    1    1
## 145 0.01917601 -2.225244    2    2
## 45  0.01838197 -2.178685    1    2
## 177 0.01740040  2.119718    2    1
## 48  0.01678720 -2.082033    1    2
## 144 0.01660556 -2.070738    2    2
```