SY09 : Rapport TP1

Statistique descriptive, Analyses en composantes principales

CATHELAIN Valentin MARECHAL Anaig

22 avril 2016

Résumé

La classification automatique des données, tout comme l'ACP étudiée au TP précédent, va nous permettre une réduction des données et ainsi l'obtention d'une représentation simplifiée de nos données de départ pour les analyser plus aisément.
Ainsi, la classification automatique permet d'organiser nos données en clusters qui nous offrent la possibilité de distinguer les différents individus d'une population.

Première partie

Statistique descriptive

Chapitre 1

Le racket du tennis

1.1 Analyse descriptive générale des données

Notre jeu de données pré-traité concerne à présent 26532 matchs joués entre 2009 et 2015, sur lesquels 129271 positions ont été prises. 1523 joueurs ont pris part à ces match, dont 1502 ont perdu au moins un match et 899 ont gagné au moins un match.
En tout, 7 bookmakers ont pris position sur les matchs. Pour 66798 matchs, soit dans 52,82% des cas, les paris ont évolué vers le joueur gagnant.

1.2 Catégorisation des joueurs

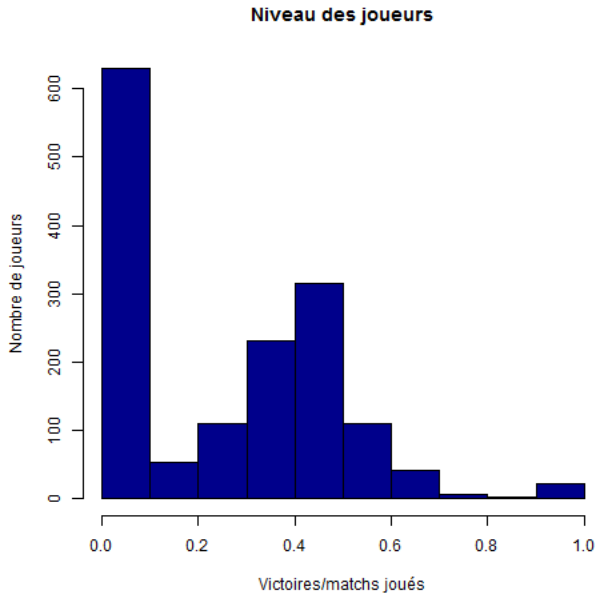

FIGURE 1.1 – Propension des joueurs à gagner

Grâce à la base de données des plus de 26000 matchs joués, nous avons pu calculer la propension de chaque joueur à gagner un match.
On remarque qu'une grande partie des joueurs a un ratio victoires/matchs joués compris entre 0 et 0.1.

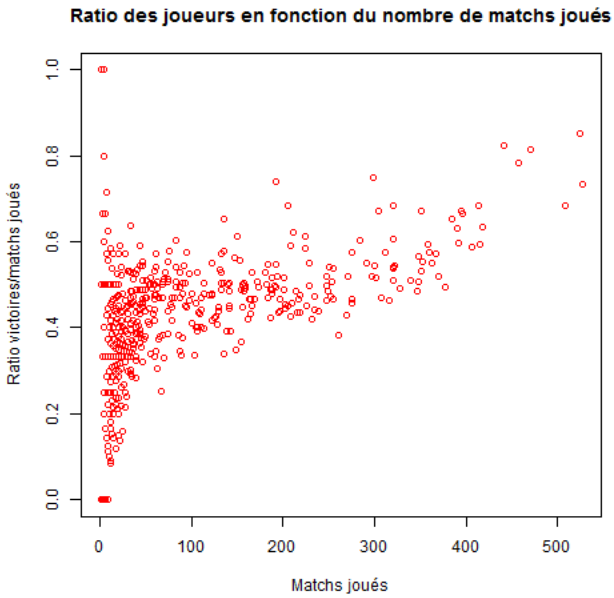

FIGURE 1.2 – Niveau des joueurs en fonction du nombre de matchs joués

De plus, si l'on affiche le ratio des joueurs en fonction du nombre de matchs qu'ils ont joués, on peut observer plusieurs choses intéressantes :

1. Les joueurs qui possèdent un ratio extrême (proche de 0 ou de 1) sont surtout ceux ayant joués que peu de matchs, l'échantillon est donc très limité.
2. La majorité des joueurs a un ratio compris entre 0.4 et 0.6 dès qu'un joueur a joué plus de de 100 matchs.
3. Cependant,après le seuil de 100 matchs, on observe que les points semblent suivre une tendance linéaire, plus les joueurs ont joués de matchs, plus leur ratio est élevé. On peut expliquer cette tendance par le fait que les joueurs qui on joué un grand nombre de matchs sont sûrement ceux étant restés longtemps dans le circuit professionnel (peut-être ces 5 années), et ils ont donc un niveau plus élevés que la majorité des joueurs.

1.3 Recherche des matchs suspects

Nous avons retenus deux critères qui peuvent qualifier un pari de suspect :

— Tout d'abord, une évolution significative de la probabilité de gain (+ de 0.1 en valeur absolue).

— Si la probabilité du gagnant a évolué en sa faveur au cours du match.

En pratique, ces deux éléments traduisent la situation dans laquelle un nombre de gens plus élevé que normalement aurait parié sur un joueur qui au final gagnera. En effet tout au long du match les bookmakers équilibrent les risques et revoient les probabilités en fonction des sommes pariées dans les deux camps.

1.3.1 Paris suspect et bookmakers impliqués

Au total, nous avons comptés **2657 paris suspects** .

Si l'on regarde les bookmakers impliqués, on remarque que **les principaux concernés sont les bookmakers A, B et C**.

A	B	C	D	E	F	G
609	867	704	145	137	126	69

FIGURE 1.3 – Répartition des paris suspects par bookmaker

Maintenant si l'on considère qu'un match est suspect si au moins un pari suspect concerne ce match, on obtient un total de **1752 matchs suspects.**
D'ailleurs on peut essayer de caractériser ces matchs suspects, en regardant combien de paris les concernent(ou combien de bookmaker, un bookmaker ne pouvant ouvrir qu'un pari par match).

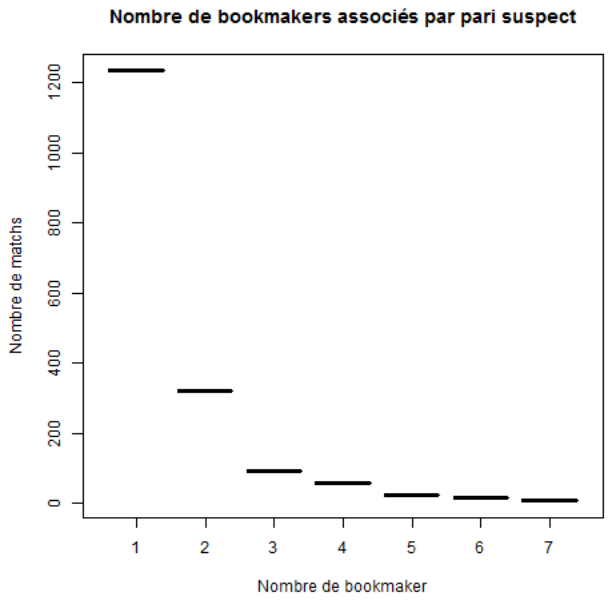

FIGURE 1.4 – Nombre de paris suspects concernant les mêmes matchs

On remarque que la grande majorité des matchs suspects ne sont liés qu'à un pari suspect, donc on peut en déduire que dans la plupart des cas les bookmakers ne se sont pas concertés.

1.3.2 Joueurs impliqués dans les pari suspects

Pour savoir quels sont les joueurs associés à des malversations, on s'intéressera uniquement aux perdants de ces matchs suspects. En effet en pratique il est bien plus facile de convaincre quelqu'un de perdre un match ....

Les joueurs associés à ces matchs suspects sont au nombre de **460**.

Mais pour qu'un joueur soit réellement suspect, on considérera qu'il doit être impliqué dans plus de 10 matchs suspects. De cette manière on obtient plus que **39 joueurs suspects**, sur les 1523 de départ.

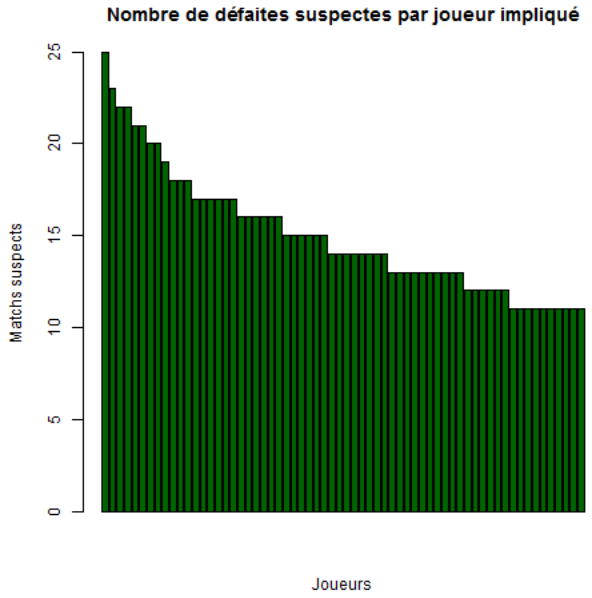

FIGURE 1.5 – Niveau des joueurs en fonction du nombre de matchs joués

Deuxième partie

Analyse en composantes principales

Chapitre 2

Données crabs

2.1 Les données

Le jeu de données "Crabs" contient les données de 200 individus, mâles ou femelles, appartenant à 2 espèces différentes. Ces crabes sont décrits par 5 variables quantitatives décrivant leur morphologie.

2.1.1 Analyse descriptive générale des données

Tout d'abord, nous avons procédé à une analyse descriptive des données selon le sexe et selon l'espèce des crabes, afin de voir s'il en ressort une manière d'identifier les différents individus grâce à leurs mesures.
Nous avons donc découpé le jeu de données, par sexe d'abord, par espèce ensuite. Nous avons alors fait le choix d'une représentation par boîte à moustache car cela permet d'avoir une vision assez complète sur les données et facilitant plus la comparaison qu'avec l'utilisation de nuages de points par exemple.

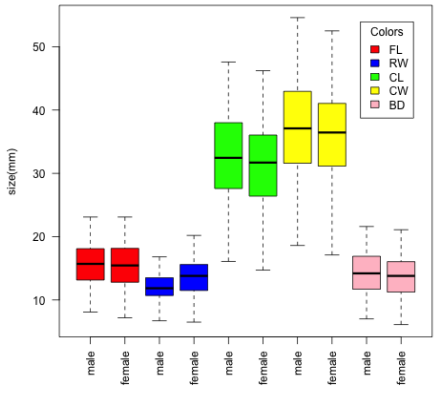

FIGURE 2.1 – Comparaison des caractéristiques morphologiques des crabes selon le sexe.

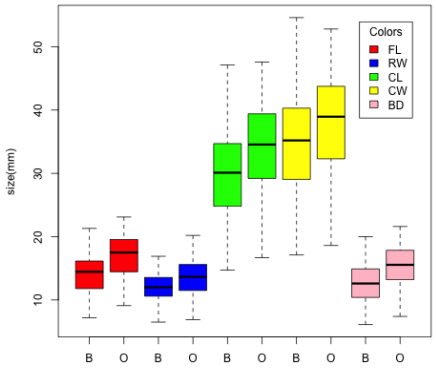

FIGURE 2.2 – Comparaison des caractéristiques morphologiques des crabes selon l'espèce.

On peut observer sur ces deux graphiques que les valeurs des variables entre les deux sexes ou entre les deux espèces sont du même ordre de grandeur d'un cas sur l'autre. De très légères différences sont visibles mais elles ne sont pas assez prononcées pour qu'on puisse identifier les individus à partir de ces différences.

La description monodimensionnelle n'apportant pas de différences significatives, nous avons décidé de représenter les différentes variables par couple grâce à un graphique matriciel.

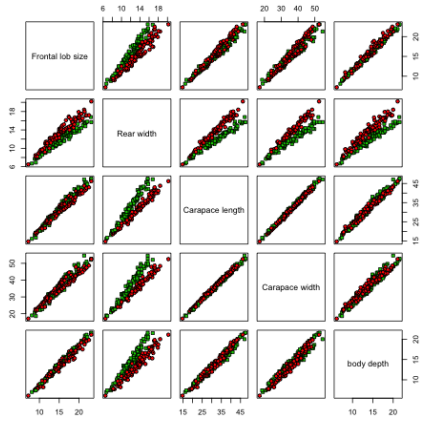

FIGURE 2.3 – Comparaison bidimensionnelle des caractéristiques morphologiques des crabes selon le sexe.

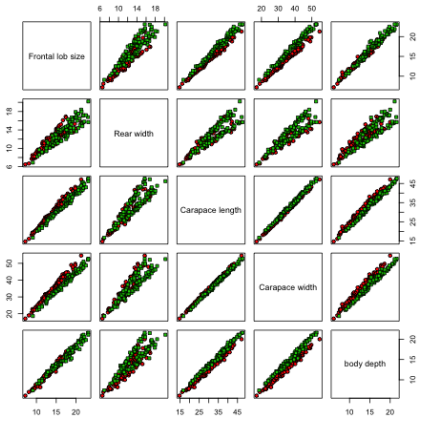

FIGURE 2.4 – Comparaison bidimensionnelle des caractéristiques morphologiques des crabes selon l'espèce.

On retrouve une liaison linéaire entre chaque variable et le coefficient de corrélation apparaît être très fort. On observe une légère distinction pour la comparaison entre mâle et femelle au niveau de la variable RW, pour lequel les graphiques affichent des coefficients directeurs légèrement différent selon le sexe de l'individu, quelque soit la variable couplée.

2.1.2 Corrélation entre les variables

Le tableau de corrélation obtenu grâce à la commande

```
cor()
```

sur crabsquant confirme ce que laissait apparaître les graphiques 2.3 et 2.4 :

TABLE 2.1 – Matrice de corrélation du jeu de données crabsquant

	FL	**RW**	**CL**	**CW**	**BD**
FL	1.0000000				
RW	0.9069876	1.0000000			
CL	0.9788418	0.8927430	1.0000000		
CW	0.9649558	0.9004021	0.9950225	1.0000000	
BD	0.9876272	0.8892054	0.9832038	0.9678117	1.0000000

Ainsi, la corrélation entre les différentes variables est *très élevée*, la valeur la plus faible étant 0.89. Cette valeur peut s'expliquer par le fait qu'un crabe plus grand par exemple aura forcément des mesures morphologiques plus grandes pour toutes les variables : chaque crabe, individuellement, a, en toute logique, des membres proportionnels.
Pour s'affranchir de ce phénomène, il serait judicieux d'analyser les variables simultanément, comme une variable, voire deux, les "englobant" et représentant un individu. Cela permettrait en quelque sorte de résumer les caractéristiques morphologiques de chaque crabe, et ainsi les comparer sur un plan ou deux seulement, s'affranchissant ainsi de la corrélation qui n'apporte pas d'information pour résoudre notre problématique, tout en simplifiant notre comparaison. *L'Analyse en Composante Principale* (ACP), est une *méthode factorielle de réduction de dimension* qui permet l'analyse statistique de données regroupant un grand nombre de variables. Cette méthode semble donc tout à fait adaptée à notre cas.

2.2 Exercice théorique

On nous donne pour cet exercice un tableau de trois variables sur quatre individus. Appelons cette matrice M.

$$M = \begin{pmatrix} 3 & 4 & 3 \\ 1 & 4 & 3 \\ 2 & 3 & 6 \\ 4 & 1 & 2 \end{pmatrix}$$

2.2.1 Les axes factoriels

Tout d'abord, on commence par centrer la matrice en colonne en soustrayant à chaque colonne sa moyenne. On obtient :

$$X = \begin{pmatrix} 0.5 & 1 & -0.5 \\ -1.5 & 1 & -0.5 \\ -0.5 & 0 & 2.5 \\ 1.5 & -2 & -1.5 \end{pmatrix}$$

On procède ensuite au calcul de la matrice de variance grâce à la formule

$$V = \frac{1}{n} * M^T M$$

On obtient

$$V = \begin{pmatrix} 1.25 & & \\ -1 & 1.5 & \\ -0.75 & 0.5 & 2.25 \end{pmatrix}$$

A partir de cette matrice, on calcule les valeurs propres ainsi que les axes factoriels.

Table 2.2 – Valeurs propres

λ
3.1989
1.4685
0.3326

Table 2.3 – Axes factoriels

u_1	u_2	u_3
0.52	0.34	0.78
-0.51	-0.61	0.61
-0.69	-0.72	0.15

L'inertie du nuage est égale à la somme des valeurs propre et est donc ici 5. Les pourcentages d'inertie expliquée par chaque axe sont donc de 63.977844, 29.369722 et 6.652434.

2.2.2 Les composantes principales

Selon la formule $C = XU$, on obtient la matrice des composantes principales suivante :

Table 2.4 – Composantes principales

1	**2**	**3**
0.09	-0.80	0.92
-0.95	-1.48	-0.64
-1.97	1.62	-0.02
2.83	0.66	-0.26

Nous en avons déduit cette représentation des quatre individus dans le premier plan factoriel :

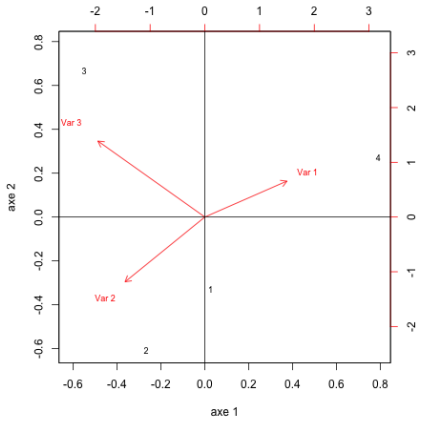

FIGURE 2.5 – Représentation des individus dans le premier plan factoriel

2.2.3 Représentation des variables dans le premier plan factoriel

Pour représenter les trois variables dans le premier plan factoriel, nous allons tracer un cercle de corrélation. Il est défini par deux composantes principales, ici nous allons choisir la première et la deuxième, et il présente les variables en fonction de leurs coefficients de corrélation avec les composantes principales.
La première étape est donc de calculer la corrélation de chaque variable avec les axes.

TABLE 2.5 – Corrélations entre les variables et les composantes

F1	**F2**	**F3**
0.8383	0.3670	0.4031
-0.7438	-0.6043	0.2855
-0.8139	0.5782	0.05675

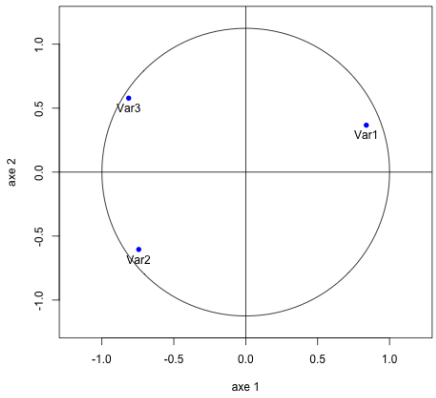

FIGURE 2.6 – Composantes principales associées aux variables.

Le fait que les points soient très proches du cercle indique une très forte corrélation entre les variables et les axes. Cependant on note ici que les variables ne se situent pas exactement sur le cercle de corrélation : la variable 3 semble être la plus proche, puis la variable 2, et enfin la variable 1.
Afin d'expliquer ce comportement, nous avons émis une hypothèse : nous avons ici associé les variables aux deux premières composantes principales. Si un des points était parfaitement situé sur le diamètre du cercle, on aurait pu en conclure que la variable était parfaitement bien représentée par ces deux composantes. Or ici nos points sont légèrement vers l'intérieur du cercle. Il faudrait donc peut-être le troisième composant pour que ces variables soient parfaitement représentées. Si notre hypothèse s'applique, cela voudrait dire que nos deux plans ne représentent pas ici exactement notre nuage initial mais s'en approchent cependant fortement, ce qui est suffisant pour faire une analyse multidimensionnelle de notre ensemble d'individus.

2.2.4 Calcul

Calculons l'expression $\sum_{\alpha=1}^{k} c_\alpha u'_\alpha$ pour les valeurs k=1, 2 et 3.

$$k=1 : \begin{pmatrix} 0.05 & -0.05 & -0.06 \\ -0.50 & 0.49 & 0.65 \\ -1.03 & 1.00 & 1.34 \\ 1.48 & -1.44 & -1.93 \end{pmatrix}, k=2 : \begin{pmatrix} -0.22 & 0.44 & -0.64 \\ -1.00 & 1.39 & -0.41 \\ -0.48 & 0.01 & 2.50 \\ 1.70 & -1.84 & -1.46 \end{pmatrix}, k=3 : \begin{pmatrix} 0,5 & 1.00 & -0.5 \\ -1.5 & 1.00 & -0.5 \\ -0.5 & 0.00 & 2.5 \\ 1.5 & -2 & -1.5 \end{pmatrix}$$

On constate que pour k=3, $\sum_{\alpha=1}^{k} c_\alpha u'_\alpha = X$, notre matrice centrée.

2.3 Utilisation des outils R

Grâce aux différentes fonctions proposées par R, nous allons tenter d'effectuer l'ACP du jeu de données de notes qui a été étudié en cours. On commence par créer la matrice.

```
 #Création de la matrice de notes
M = matrix(c(6.0, 6.0, 5.0, 5.5, 8.0, 8.0, 8.0, 8.0, 8.0, 9.0, 6.0, 7.0, 11.0, 9.5, 11.0,
14.5, 14.5, 15.5, 15.0, 8.0, 14.0, 14.0, 12.0, 12.5, 10.0, 11.0, 10.0, 5.5, 7.0, 13.0,
5.5, 7.0, 14.0, 11.5, 10.0,13.0, 12.5, 8.5, 9.5,  12.0, 9.0, 9.5, 12.5, 12.0, 18.0),
nrow = 9, byrow = T)
rownames(M) = c("jean", "aline", "annie", "monique", "didier",
"andré", "pierre", "brigitte", "evelyne")
colnames(M) = c("math", "scie", "fran", "lati","d-m")
```

Pour pouvoir effectuer l'ACP, il est tout d'abord indispensable de centrer notre matrice. C'est ensuite à partir de notre matrice centrée en colonne qu'on va pouvoir calculer la matrice de variance, puis les axes principaux d'inertie et enfin les composantes principales.

```
> #Centrage de la matrice M en colonne
> (X <- scale(M, center=T, scale=F))
               math       scie       fran       lati d-m
jean     -3.6666667 -3.8333333 -5.2222222 -4.5555556  -3
aline    -1.6666667 -1.8333333 -2.2222222 -2.0555556  -2
annie    -3.6666667 -2.8333333  0.7777778 -0.5555556   0
monique   4.8333333  4.6666667  5.2777778  4.9444444  -3
didier    4.3333333  4.1666667  1.7777778  2.4444444  -1
andré     1.3333333  0.1666667 -4.7222222 -3.0555556   2
pierre   -4.1666667 -2.8333333  3.7777778  1.4444444  -1
brigitte  3.3333333  2.6666667 -1.7222222 -0.5555556   1
evelyne  -0.6666667 -0.3333333  2.2777778  1.9444444   7
attr(,"scaled:center")
     math      scie      fran      lati       d-m
 9.666667  9.833333 10.222222 10.055556 11.000000


> #Création de la matrice de variance
> (V <- (1/9)*(t(X)%*%X))
           math       scie       fran      lati        d-m
math 11.3888889 9.91666667  2.6574074 4.8240741 0.11111111
scie  9.9166667 8.94444444  4.1203704 5.4814815 0.05555556
fran  2.6574074 4.12037037 12.0617284 9.2932099 0.38888889
lati  4.8240741 5.48148148  9.2932099 7.9135802 0.66666667
d-m   0.1111111 0.05555556  0.3888889 0.6666667 8.66666667


> #Axes principaux d'inertie
> U <- eigen(V)

> #Valeurs propres
> (ValeursPropres <- U$values)
[1] 28.253249801 12.074723274  8.615733579  0.021732182  0.009869805
```

```
> #Vecteurs propres
> (VecteursPropres <- U$vectors)
            [,1]       [,2]        [,3]        [,4]        [,5]
[1,] -0.51453535  0.5669492 -0.05132308 -0.28874852  0.57254891
[2,] -0.50698853  0.3719958 -0.01445296  0.55305647 -0.54635285
[3,] -0.49235486 -0.6503536  0.10806565  0.39373536  0.40978192
[4,] -0.48462835 -0.3232385  0.02254331 -0.67419539 -0.45343643
[5,] -0.03062778 -0.1128933 -0.99245689  0.03443659  0.01266839
```

```
> #Composantes principales
> (C <- X%*%U$vectors)
              [,1]       [,2]       [,3]        [,4]        [,5]
jean      8.700907  1.7027046  2.5539182 -0.14945398 -0.11731596
aline     3.938596  0.7085441  1.8104644 -0.09068389  0.04349922
annie     3.209392 -3.4590552  0.3006617  0.17254286  0.01928215
monique  -9.755741 -0.2157421  3.3436726 -0.17347137  0.10041455
didier   -6.371422  2.1733326  0.9570588  0.07066256 -0.18799232
andré     2.974017  4.6509322 -2.6349457 -0.02321315  0.14809545
pierre    1.050967 -6.2271742  1.6880636  0.11529582  0.04281219
brigitte -1.980533  4.0685562 -1.4007122  0.24321198  0.01039742
evelyne  -1.766183 -3.4020982 -6.6181814 -0.16489082 -0.05919270
```

Ainsi, grâce à la fonction *princomp*, on peut obtenir les plans de représentations.

```
RP <- princomp(X)
biplot(RP, ylab = "axe 2", xlab = "axe 1", var.axes = FALSE, main="ACP :Exemple des notes"
)
abline(h=0,v=0)
```

On obtient :

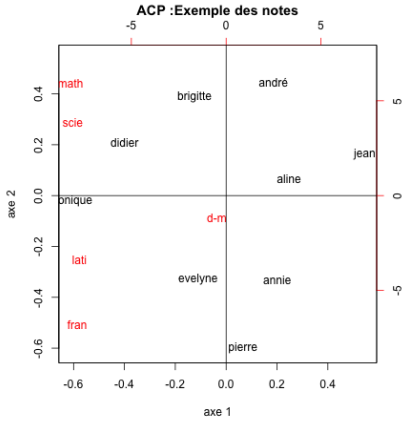

FIGURE 2.7 – Plan de représentation.

On peut également obtenir les cercles de corrélation qui associent les composantes principales normées aux variables :

```
#Cercle de corrélation
cor <- cor(M, C)

 #1er plan
plot(cor, xlab = "axe 1", ylab = "axe 2", xlim = c(-1.2, 1.2), ylim = c(-1.2, 1.2), col = "blue",
pch=c(16), main="ACP : Exemple de notes dans le premier plan factoriel")
text(cor, pos = 1, labels = c("math", "scie", "fran", "lati","d-m"))
abline(h=0,v=0)
symbols(0, 0, circle = 1, inches = F, add = T)

#2ème plan
cor2 <- cbind(cor[,1], cor[,3])
plot(cor2, xlab = "axe 1", ylab = "axe 2", xlim = c(-1.2, 1.2), ylim = c(-1.2, 1.2),
col = "blue", pch=c(16), main="ACP : Exemple de notes dans le deuxième plan factoriel")
text(cor2, pos = 1, labels = c("math", "scie", "fran", "lati","d-m"))
abline(h=0,v=0)
symbols(0, 0, circle = 1, inches = F, add = T)
```

Voici le résultat alors obtenu :

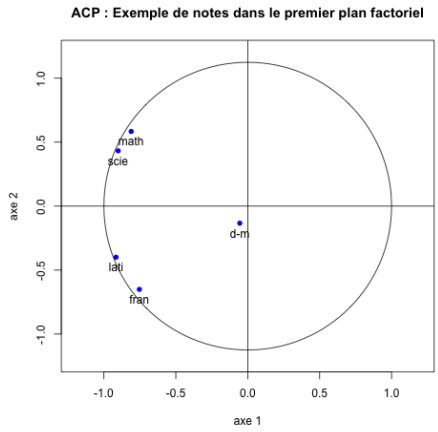

FIGURE 2.8 – Cercle de corrélation dans le premier plan.

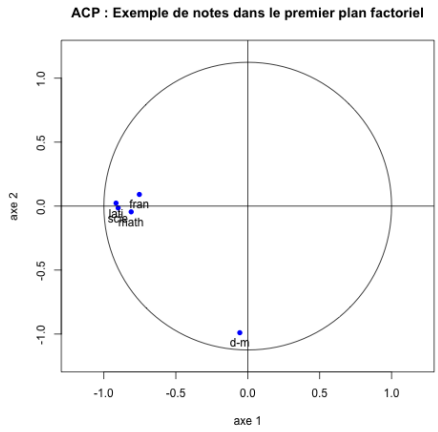

FIGURE 2.9 – Cercle de corrélation dans le deuxième plan.

La fonction *princomp* permet de réaliser une ACP. La fonction va elle même rechercher les vecteurs propres de la matrice de covariance. Ainsi, grâce à un *summary* sur le résultat obtenu grâce à *princomp*, on obtient de nombreuses informations sur les variances des différents composants : la valeur de la variance, son pourcentage et son pourcentage cumulé.

```
> summary(RP)
Importance of components:
                          Comp.1    Comp.2    Comp.3       Comp.4       Comp.5
Standard deviation     5.3153786 3.4748703 2.9352570 0.1474183919 0.0993468909
Proportion of Variance 0.5768876 0.2465472 0.1759199 0.0004437375 0.0002015261
Cumulative Proportion  0.5768876 0.8234348 0.9993547 0.9997984739 1.0000000000
```

Grâce à l'argument *$scores* sur *princomp* par exemple, on peut obtenir la liste des composantes principales.
La fonction *plot* appliquée sur *princomp* permet d'afficher la variance des composantes principales.

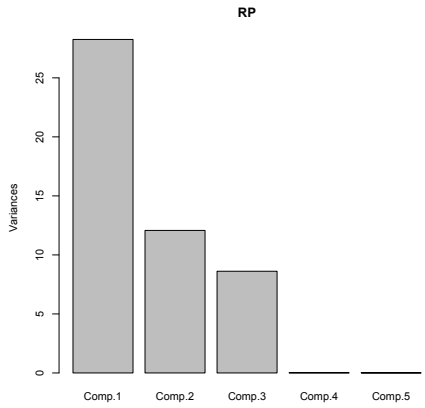

FIGURE 2.10 – Fonction plot : variance des composantes principales.

La fonction *biplot* quant à elle permet de visualiser la projection des données, individus et variables, dans le plan constitué de deux axes factoriels (les deux principaux par défaut).

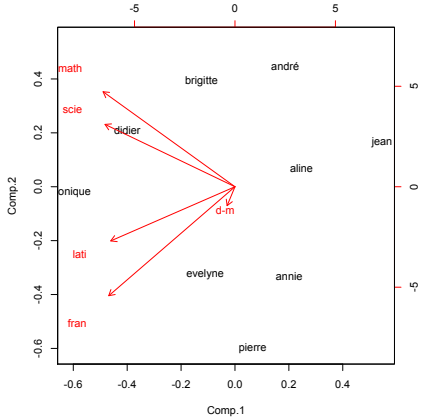

FIGURE 2.11 – Fonction biplot : variance des composantes principales.

La fonction *biplot.printcomp* prend en paramètre un objet de la classe *princomp* et des arguments comme *choices* qui permet de définir la taille des vecteurs, ou encore *scale*, valeur comprise entre 0 et 1 qui permet de définir l'échelle.

2.4 Traitement des données Crabs

2.4.1 Sans pré-traitement

On teste tout d'abord l'ACP sur *crabsquant* sans traitement préalable. On obtient :

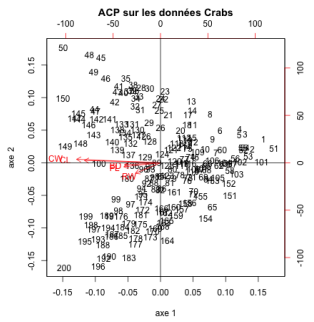

FIGURE 2.12 – Plan de représentation des données Crabs.

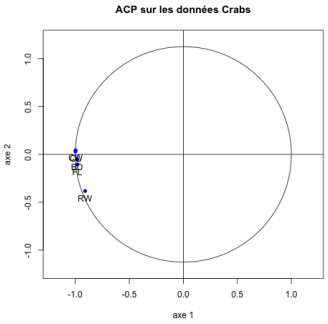

FIGURE 2.13 – Corrélation entre les variables des données Crabs et les axes.

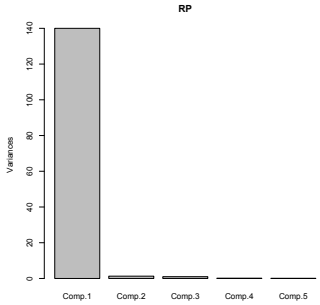

FIGURE 2.14 – Variance des composantes principales.

On observe sur les figures 2.12 et 2.13 que les variables sont très corrélées aux composantes et que la variance de la première composante est très élevée. Avec un *summary* sur emphprincomp on constate que la première composante a un pourcentage de 98%.

2.4.2 Avec pré-traitement

Afin d'améliorer la qualité de notre représentation en terme de visualisation des différents groupes, nous avons donc décidé de retirer la variable CW des données. En effet, sa valeur nous semble très corrélée à l'axe 1 au vu du graphique 2.13 et l'importance de la variable est donc moindre (La variable CL aurait été un deuxième choix). On obtient alors :

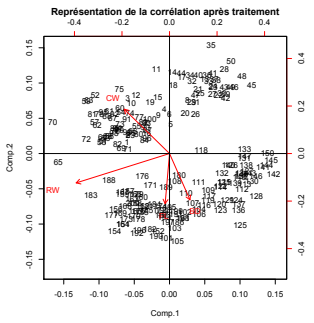

FIGURE 2.15 – Plan de représentation des données Crabs pré-traitées.

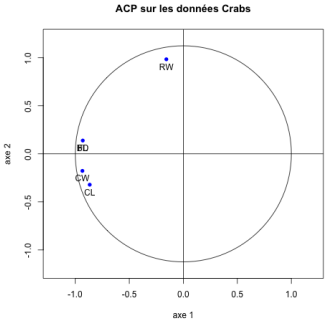

FIGURE 2.16 – Cercle des corrélations des données Crabs pré-traitées.

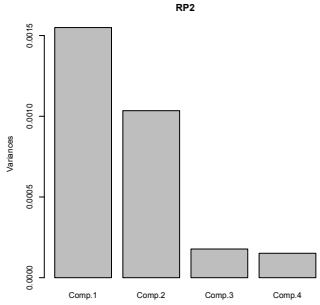

FIGURE 2.17 – Variance des composantes.

Ainsi, avec ce pré-traitement on obtient une nouvelle comparaison des caractéristiques morphologiques des crabes par sexe et par espèce, qui sont désormais *facilement différenciables*, en faisant par exemple le rapport RW/FL pour le sexe et CL/FL pour l'espèce :

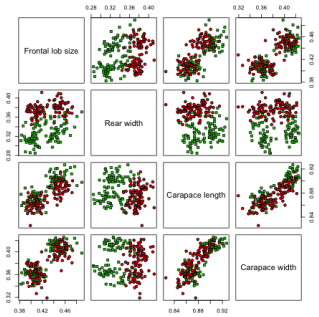

FIGURE 2.18 – Comparaison des caractéristiques morphologiques des crabes par sexe.

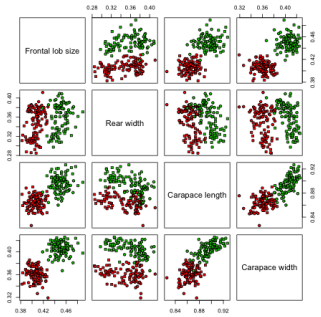

FIGURE 2.19 – Comparaison des caractéristiques morphologiques des crabes par espèce.

2.5 Conclusion

En conclusion, nous avons pu d'une première part au cours de ce rapport, utiliser de façon concrète la puissance et les multiples possibilités qui s'offrent à nous en terme de traitement statistique de données avec R. Grâce à la statistique descriptive, nous avons appris à faire ressortir les données qui nous intéressent à travers un très large jeu. Nous avons ensuite appliqué la méthode de l'ACP qui nous a considérablement aidé à obtenir des données analysables en les étudiant sous un plan qui ne s'offrait pas directement à nous.