SY09

Analyse de Données & Data Mining

TP1 - Statistique Descriptive & Analyse en Composantes Principales

Printemps 2011

Jie Chen Victor Hatinguais

1 Statistique Descriptive

1.1 Données *babies*

Dans cette première partie, nous nous intéressons à un jeu de données concernant les grossesses de 1236 femmes décrites par 23 variables et contenues dans un fichier *babies.txt*. Nous nous intéresserons plus particulièrement aux variables **bwt** : poids de naissance en ounces ; **gestation** : durée de gestation en jours ; **smoke** : permettant de distinguer les mères ayant fumé pendant leur grossesse des autres et **ed** : niveau d'éducation des mères. Ceci nous permettra d'établir les liens éventuels entre ces différentes variables.

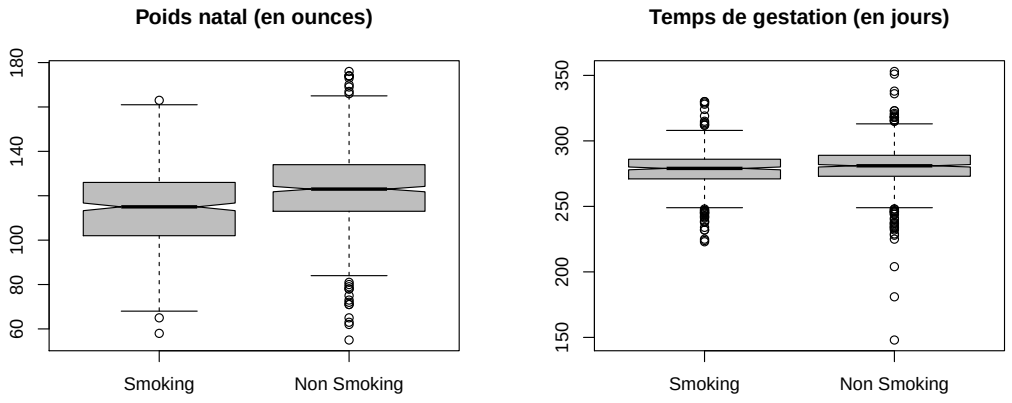
Nous chercherons plus particulièrement à déterminer si le fait de fumer durant la grossesse a une influence ou non sur le poids du bébé à la naissance, sur le temps de gestation et/ou sur le niveau d'éducation de la mère. Nous devrons également, au cours de cette étude, prendre en compte le fait que certaines données ne sont pas disponibles (NA).

Une fois les données chargées et traitées dans R, on commence par afficher un résumé numérique des variables **gestation** et **bwt** dans le cas de mères fumeuses et non fumeuses ; ceci afin de se donner une première idée des différences qui pourraient exister.

```
                      Min. 1st Qu. Median  Mean 3rd Qu. Max. NA's
bwt Smoking             58     102    115 114.1     126  163   10
bwt NonSmoking          55     113    123 123.0     134  176   10
gestation Smoking      223     271    279 278.0     286  330   14
gestation NonSmoking   148     273    281 280.2     289  353   19
```

On remarque tout de suite la présence d'une valeur étrange pour la durée de gestation : 148 jours, ce qui est possible mais semble être très court. Pour se prémunir d'une hypothétique erreur de saisie (peut-être est ce plutôt 248 jours), il conviendrait, en temps normal, de se renseigner et de prendre une décision pouvant aller jusqu'à la suppression de l'individu posant problème. Toutefois, nous considérerons ici la donnée comme correcte et la prendrons donc en compte.
On remarque également que les temps de gestation sont généralement plus courts et que les poids à la naissance des bébés sont généralement plus faible dans le cas des mères fumeuses.

Afin de mieux visualiser et évaluer ces différences, on trace les boîtes à moustaches correspondantes :



On y remarque les différents outliers, mais surtout, on constate que la différence pour le poids à la naissance est important, tandis que la différence dans les temps de gestation est très faible.

Pour la différence de poids, on l'évalue par la différence des moyennes :

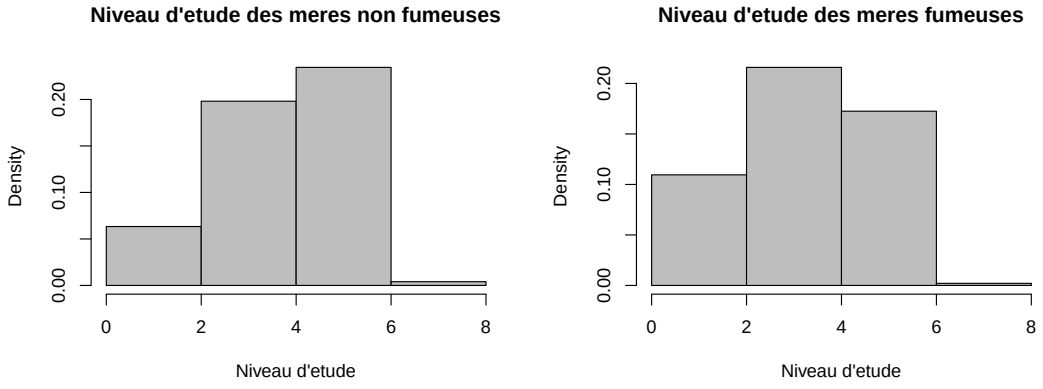
```
[1] 8.937666
```

On constate donc qu'en moyenne, un bébé né de mère non fumeuse pèse 9 ounces (environ 300 grammes) de plus qu'un bébé né de mère fumeuse. On peut également réaliser un test t de Student pour évaluer un intervalle de confiance. Par exemple, pour un risque de première espèce de $\alpha = 5\%$, on obtient une différence de poids comprise entre 6,9 et 11 ounces.

Concernant le temps de gestation, la différence semble moindre et la réalisation d'un test t de Student nous donne une p-value de 0,017.
Ceci signifie que l'on rejette l'hypothèse H_0 d'égalité des espérances pour un risque de première espèce $\alpha = 5\%$ mais qu'on l'accepte pour $\alpha = 1\%$. Autrement dit, on a peu de chance de se tromper si l'on déclare que le temps de gestation est plus long chez les femmes ne fumant pas. Néanmoins, il reste un risque d'un peu moins de 2% de chance que l'on se trompe et qu'en réalité, le fait de fumer durant la grossesse n'influe pas directement sur le temps de gestation.

L'extrait d'article présent en annexe dans le sujet mentionne que les bébés américains pèsent généralement moins que les bébés norvégiens et que l'on a longtemps pensé que la forte mortalité infantile aux Etats-Unis était dû à ce problème. Mais on s'est aperçu qu'en réalité, pour une même durée de gestation, un bébé a autant de chance de mourir qu'il soit léger ou lourd. Ce n'est donc finalement pas le poids à la naissance qui influe sur la mortalité infantile mais bien le temps de gestation. Or ce temps de gestation influant également sur le poids à la naissance (corrélation linéaire de 0,41 avec nos données), cela explique que l'on ait pu croire que le poids à la naissance influe sur la mortalité infantile mais c'est bien le temps de gestation qui influe sur ces deux variables.

De la même façon, étudions le niveau d'étude des femmes selon qu'elles fument ou non (gardons à l'esprit que nous traitons ici un cas spécifique où les femmes ont eu une grossesse et que l'on a considéré comme fumeuses uniquement les femmes ayant continué à fumer pendant leur grossesse).



On constate qu'une plus grande part de la population des non fumeuses a suivi des études supérieures, tandis que l'inverse se produit pour la population des fumeuses : une plus grande part n'a fait que peu d'études.
On pourrait alors conclure que le fait de fumer influe sur le niveau d'éducation. Mais comme on l'a expliqué précédemment, ces deux variables peuvent en réalité dépendre toutes deux d'une troisième variable ; par exemple, le milieu social. Et ceci semble être, en effet, plus réaliste.

Pour conclure, on peut affirmer que le fait de fumer pendant la grossesse diminue le poids des bébés à

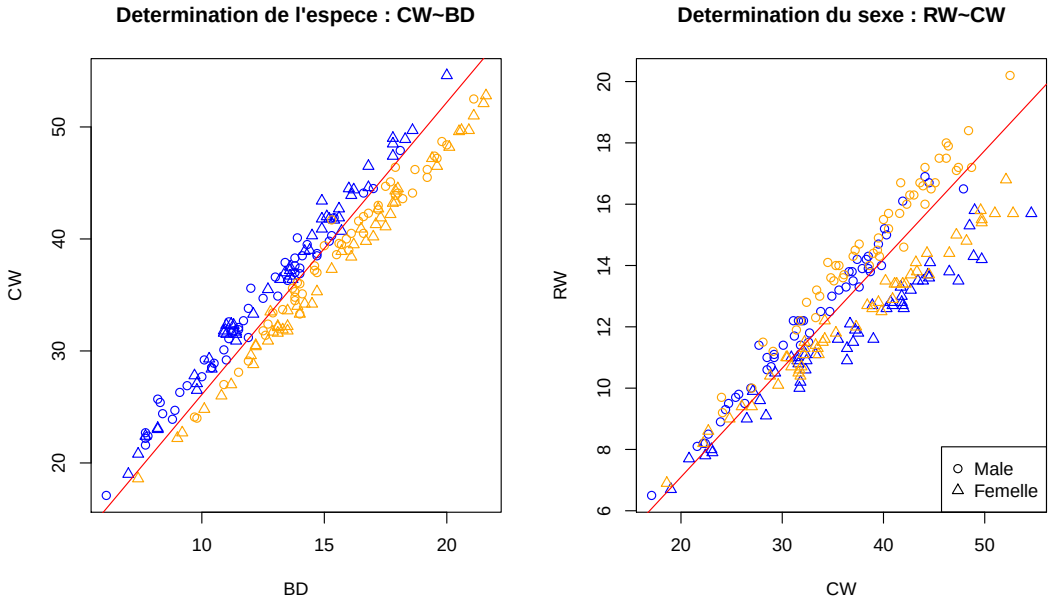
la naissance et a également de fortes chances de réduire le temps de gestation. Cependant, d'après l'article fourni en annexe, le fait de fumer pendant la grossesse n'influerait finalement pas sur le temps de gestation (ce qui reste envisageable, d'après notre étude). Or la mortalité infantile ne dépendant que du temps de gestation, on aboutit à un étrange paradoxe : dans une population de bébés de faible poids à la naissance, ceux provenant de fumeuses ont plus de chance de survivre que les autres. Ceci s'explique par le fait que dans une telle population, on aura une plus grande part de bébés provenant de fumeuses (puisque le fait de fumer diminue le poids à la naissance) dont les temps de gestation respecteront la moyenne alors que les bébés provenant de non fumeuses feront pour la plupart partie de cette population car ayant subi une gestation de plus courte durée que la moyenne. Ainsi, les bébés de faible poids provenant de femmes fumeuses auront plus de chance de survivre mais ceci n'en dit rien sur les chances de survie en général des bébés provenant de fumeuses (quelqu'en soit le poids).

1.2 Données *crabs*

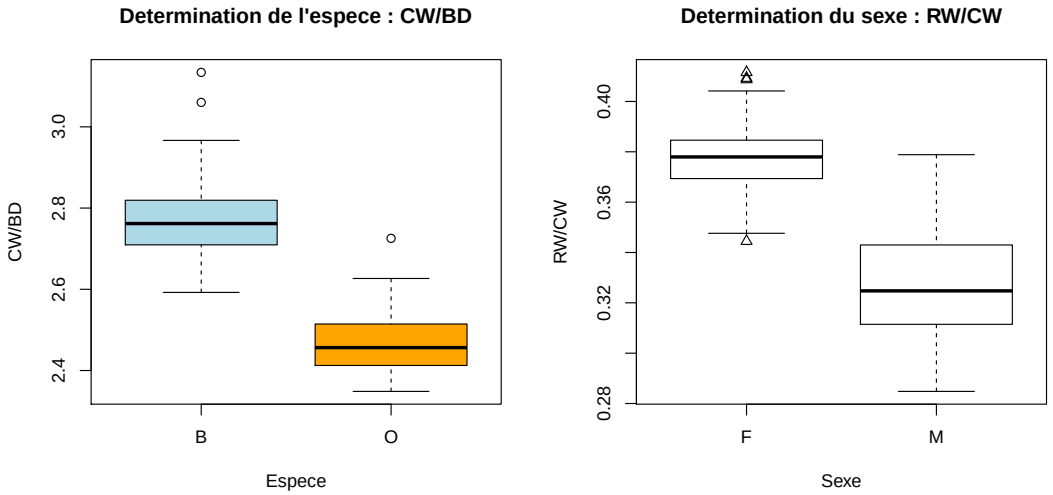
Dans cette seconde partie, nous étudierons une population constituée de 200 crabes de quatre types différents (deux espèces et deux sexes) à partir de données complètes sur cinq variables quantitatives représentant les longueurs de certaines parties de chaque crabe.

Pour commencer, on effectue une analyse descriptive des données en regroupant les individus par type (sexe et espèce). Cependant, bien que les boîtes à moustache nous révèlent que les crabes peuvent avoir des spécificités selon le sexe ou l'espèce, il est totalement impossible de déterminer le sexe et/ou l'espèce de l'un de ces crabes à partir d'une seule de ces caractéristiques. Nous ne ferons donc pas figurer ici les différents résultats obtenus (cf. code). En revanche, nous allons tenter d'identifier le sexe et l'espèce d'un crabe à partir de plusieurs mesures.

A partir du graphique matriciel, on constate une importante corrélation linéaire entre les différentes variables dont nous reparlerons par la suite. On arrive cependant à isoler certains graphiques qui semblent pouvoir donner accès à la détermination du genre ou de l'espèce :



En effet, en réalisant le quotient d'un axe par l'autre, on obtient deux nouvelles variables qui permettent de discriminer facilement l'espèce et le sexe d'un individu :



Cependant, on remarque que l'on ne peut toujours pas isoler les quatre types d'individus sur un même graphique. Cela sera rendu possible dans la seconde partie du TP grâce à l'analyse en composantes principales.

Mais d'abord, revenons à l'importante corrélation linéaire constatée entre chaque variable sur le graphique matriciel. Le logiciel R nous permet de la chiffrer facilement grâce à la fonction **cor**. Elle est toujours supérieure à 0,88, ce qui est énorme. Cette importante corrélation linéaire entre les variables s'explique en réalité simplement par leur nature : ce sont les dimensions de différentes parties du corps des crabes. Or il est évident que quelquesoit le type du crabe, la proportionnalité impose une corrélation linéaire entre ces différentes longueurs ; il est en effet peu courant de rencontrer un petit crabe (jeune) avec une pince immense (d'adulte par exemple). Pour pallier cette situation problématique, on peut, par exemple, diviser chaque variable d'un individu par la somme totale des différentes variables du même individu. On obtient alors des longueurs relatives par rapport à la taille générale du crabe et l'on peut ainsi faire ressortir les spécificités de chaque espèce et de chaque sexe.

Après traitement, on obtient les corrélations suivantes qui correspondent à des données plus facilement exploitables pour déterminer le type d'un individu puisque bien moins liées :

```
           FL         RW          CL         CW          BD
FL  1.0000000 -0.1183474 -0.24026563 -0.7825549  0.51974244
RW -0.1183474  1.0000000 -0.82661717 -0.2020264 -0.46394103
CL -0.2402656 -0.8266172  1.00000000  0.4186543  0.09676383
CW -0.7825549 -0.2020264  0.41865431  1.0000000 -0.64982922
BD  0.5197424 -0.4639410  0.09676383 -0.6498292  1.00000000
```

2 Analyse en Composantes Principales

2.1 Exercice théorique

Dans cette partie, on part de données sur quatre individus et trois variables formant la matrice $X = \begin{pmatrix} 3 & 4 & 3 \\ 1 & 4 & 3 \\ 2 & 3 & 6 \\ 2 & 1 & 4 \end{pmatrix}$ et on cherche à réaliser une analyse en composantes principales.

Pour commencer, on centre X en colonne en soustrayant sa moyenne à chaque colonne (on peut utiliser la fonction R **scale**) : $X = \begin{pmatrix} 1 & 1 & -1 \\ -1 & 1 & -1 \\ 0 & 0 & 2 \\ 0 & -2 & 0 \end{pmatrix}$.

Puis, on calcule la matrice de variance-covariance : $S = \frac{1}{4} * {}^tX * X = \begin{pmatrix} 0{,}5 & 0 & 0 \\ 0 & 1{,}5 & -1{,}5 \\ 0 & -0{,}5 & 1{,}5 \end{pmatrix}$

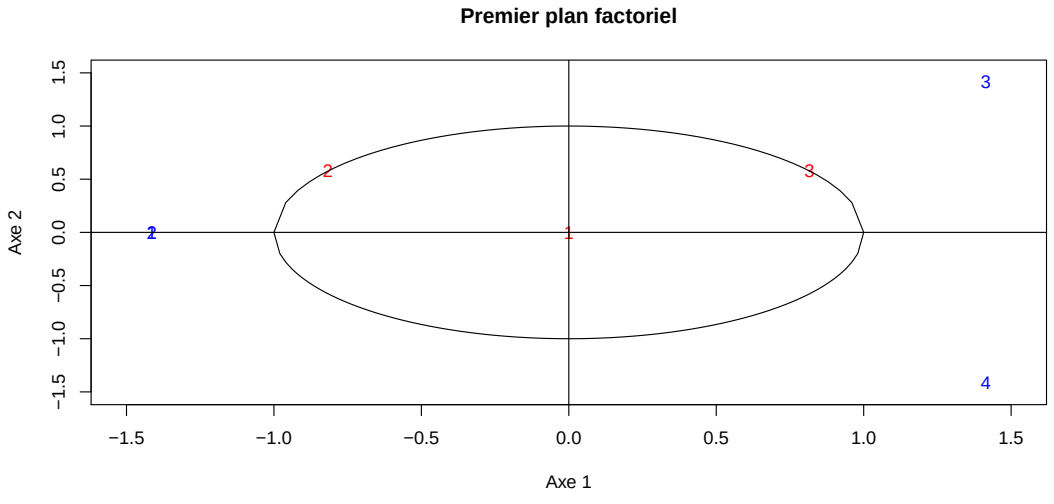
On sait alors que les axes factoriels de l'ACP correspondent aux vecteurs propres de la matrice S de variance-covariance associée à la matrice des données individus-variables centrée en colonnes. On réalise donc une décomposition spectrale à l'aide de la fonction R **eigen** et on obtient les trois vecteurs propres (axes factoriels) : $U_1 = \begin{pmatrix} 0 \\ -0,71 \\ 0,71 \end{pmatrix}; U_2 = \begin{pmatrix} 0 \\ 0,71 \\ 0,71 \end{pmatrix}; U_3 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, ainsi que leurs valeurs propres associées : $\lambda_1 = 2, \lambda_2 = 1, \lambda_3 = 0,5$.

On peut alors calculer les pourcentages d'inertie expliquée par chacun de ces axes selon la formule suivante : $E_k = \frac{\lambda_k}{\sum_{i=1}^{3} \lambda_i}$. On obtient : $E_1 = 57,14\%; E_2 = 28,57\%; E_3 = 14,29\%$.

Une fois cela fait, on peut calculer la matrice des composantes principales : $C = XMU$ où U est la matrice constituée des vecteurs propres : $U = \begin{pmatrix} 0 & -0,71 & 0,71 \\ 0 & 0,71 & 0,71 \\ 1 & 0 & 0 \end{pmatrix}$ et M = I (on utilise ici la métrique identité).

On obtient : $C = \begin{pmatrix} -1,41 & 0 & 1 \\ -1,41 & 0 & -1 \\ 1,41 & 1,41 & 0 \\ 1,41 & -1,41 & 0 \end{pmatrix}$.

On peut alors, après avoir calculé la matrice de représentation des variables (cf. code), représenter les quatre individus (en bleu) et les trois variables (en rouge) dans le premier plan factoriel :



On constate que dans ce premier plan factoriel, les variables 2 et 3 sont très bien représentées. En revanche, la variable 1 se trouve à l'origine et n'est donc pas du tout représentée. On constate notamment cela en remarquant que les individus 1 et 2 sont confondus ; en effet, ces deux-là avaient dans la matrice X de départ des valeurs identiques excepté pour la variable 1. Comme celle-ci n'est pas représentée dans cet axe factoriel, les individus sont confondus bien que pouvant être en réalité éloignés par la variable 1.

Enfin, il est aisé avec R de vérifier que la formule de reconstitution est bien valable dans notre cas :

```
> C[,1] %*% t(U[,1]) + C[,2] %*% t(U[,2]) + C[,3] %*% t(U[,3])
     [,1] [,2] [,3]
[1,]    1    1   -1
[2,]   -1    1   -1
[3,]    0    0    2
[4,]    0   -2    0
```

On retrouve bien notre matrice X centrée en colonnes.

2.2 Utilisation des outils R

L'objectif de cette partie est de se familiariser avec les fonctions du logiciel R permettant de réaliser facilement une ACP, notamment la fonction **princomp**. Pour cela, nous partirons du jeu de données *notes.txt* étudié en cours et nous tenterons de retrouver toutes les informations que peut fournir une ACP facilement à l'aide de R. Nous tâcherons également de représenter rapidement ces données à l'aide de la fonction R **biplot**.

Pour commencer, voici donc comment retrouver toutes les valeurs calculées dans l'exercice précédent de façon beaucoup plus rapide avec **princomp** (cf. code pour les résultats exacts) :

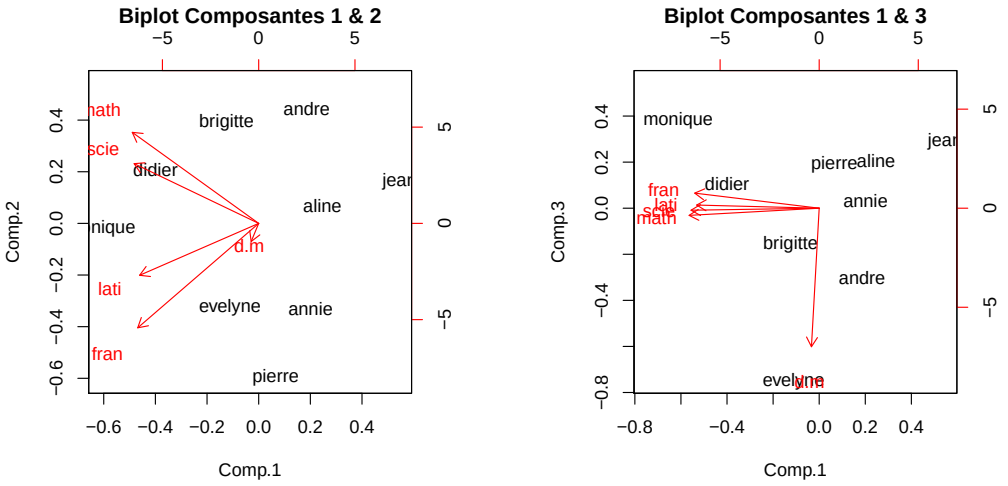
```
> N <- read.table("notes.txt")
> NP <- princomp(N)
> # Pourcentages d'inertie expliquée
> NS <- summary(NP)
> # Valeurs propres
> NvP <- NP$sdev^2
> # Vecteurs propres (axes factoriels)
> NVP <- NP$loadings
> # Composantes principales
> NCP <- NP$scores
```

Ensuite, intéressons nous aux fonctions **plot** et **biplot**.
La fonction **plot** présente peu d'intérêts : elle ne fait que représenter un histogramme montrant les valeurs propres correspondant à chaque composante principale, ce qui ne nécessite pas réellement le tracé d'un graphique pour être visualisé. Cependant, on sait que les valeurs propres sont proportionnelles aux pourcentages d'inertie expliquée. La fonction **plot** donne donc un bon aperçu de l'importance des différentes composantes.
En revanche, la fonction **biplot** présente un intérêt majeur : elle permet en une ligne de représenter un nuage de points quelconque (individus et variables) dans l'un de ses plans factoriels. Sa principale option

(en plus des options graphiques standards) est **choices** : un vecteur indiquant les deux composantes principales à utiliser.

Exemple d'utilisation :

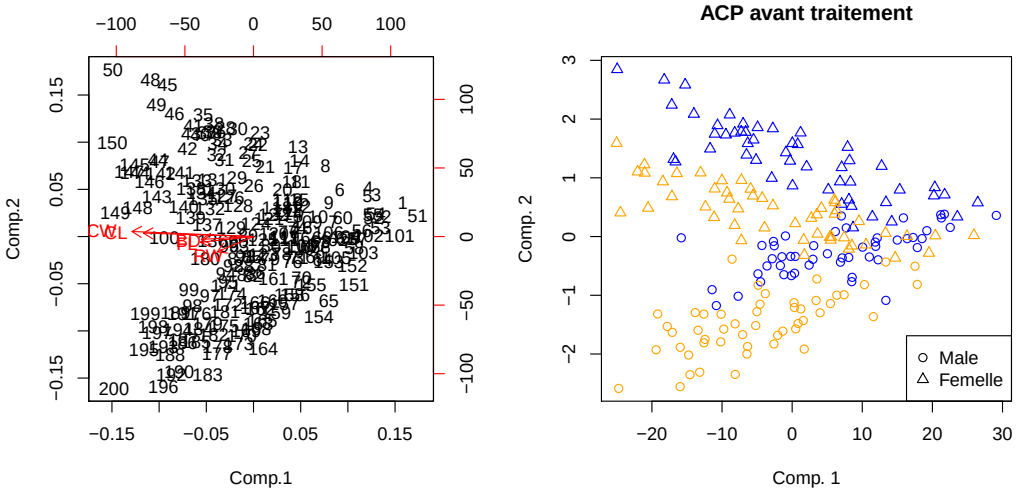
```
> par(mfrow = c(1, 2))
> biplot(princomp(N), main = "Biplot Composantes 1 & 2")
> biplot(princomp(N), choices = c(1, 3), main = "Biplot Composantes 1 & 3")
```



2.3 Traitement des données *crabs*

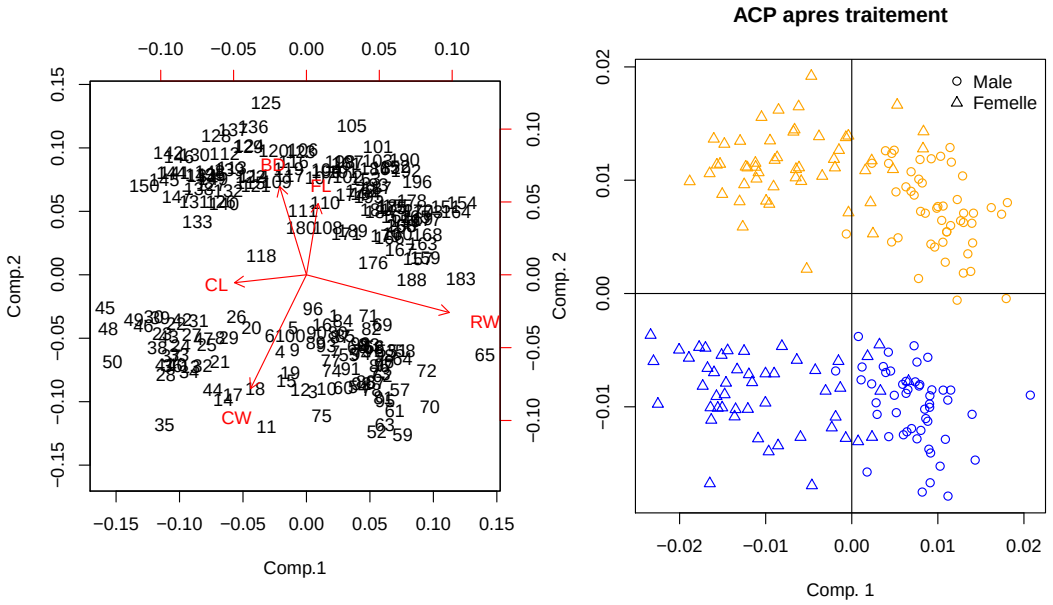
Dans cette dernière partie, on reprend les données *crabs* traitées dans la première partie mais on va utiliser, cette fois-ci, la puissance de l'analyse en composantes principales pour pouvoir distinguer visuellement les quatre différents groupes de crabes dans un même plan.

Pour commencer, on réalise une ACP directement sur les données fournies à l'aide des fonctions R **princomp** et **biplot** :



Cependant, on constate que le résultat est peu satisfaisant : il n'est pas facile d'isoler les quatre différents groupes et, de plus, on peut remarquer que la plupart des informations (variables) sont portées sur la première composante. Ce résultat médiocre s'explique par l'importante corrélation linéaire existant entre les variables (dont on a expliqué les origines dans la première partie).

Il convient de réaliser au préalable le même traitement que dans la première partie sur les données, à savoir : diviser chaque variable par la somme des variables associées au même individu.
L'ACP réalisée avec R sur ces nouvelles données donnent un résultat bien plus satisfaisant :



Finalement, à partir des données sur cinq variables, il semblerait que l'on puisse identifier de façon quasi-certaine à quel groupe appartient un crabe quelconque grâce à l'analyse en composantes principales réalisée après traitement des données pour s'affranchir d'une trop importante corrélation linéaire.

Conclusion

Pour conclure, nous avons découvert au cours de ce TP qu'une analyse descriptive des données est toujours essentielle afin de ne pas passer à côté de résultats importants mais qu'il est parfois indispensable de se servir d'outils mathématiques puissants tel que l'analyse en composantes principales afin de pouvoir visualiser certaines choses. Nous avons également constaté qu'il était important de toujours garder en tête la nature des données afin de bien comprendre le sujet et de faire subir à ces données les bons traitements lorsque nécessaire. L'importance des graphiques pour observer des hypothèses d'un seul coup d'oeil a également été démontrée au cours de ce TP.