TP - I SY09

Rémi CLAEYS - Khalida MN AZUDIN
UTC
remi.claeys@etu.utc.fr kmohdnoo@etu.utc.fr

Abstract

Dans ce rapport, nous allons expliquer nos résultats obtenus lors du TP1 de l'UV SY09 - Data Mining. Ce TP est composé de 2 parties. La première partie a pour objectif de nous familiariser avec les outils de traitement et de représentation visuelle de l'information fournis par R. La deuxième partie concerne l'analyse en composantes principales, on réalisera en premier lieu une ACP sur dataset tres reduit de 4 individus et 3 variables pour mieux comprendre le fonctionnement de l'ACP avant de l'appliquer au dataset "crabs" de la librairie MASS. Tous les résultats et graphiques fournis dans ce rapport sont obtenus grâce aux scripts R fournis en annexe et à la librairie de visualisation de données ggplot2.

1 Statistique descriptive

1.1 Le Racket du Tennis

On nous fourni un dataset concernant des paris sur des matches de tennis. Avant pré-traitement, notre dataset consiste en $N = 129271$ observations (paris) et $D = 16$ variables, dont une, qui indique l'abandon ou l'annulation du match. Un script de pré-traitement fourni par le corps enseignant de SY09 nous permet de supprimer rapidement les paris considérés obsolètes qui fausseraient notre étude.Après ce pré-traitement il nous reste donc $N = 126461$ observations et $D = 15$ variables intitulés :

- **match_book_uid**, **match_uid**, **winner** et **loser** : les identifiants du pari, du match, du gagnant, et du perdant ;
- **book** : l'identifiant du bookmaker ;
- **year** : l'année du match ;
- **odds_winner_open**, **odds_winner_close**, **odds_loser_open** et **odds_loser_close** : les cotes du joueur qui finalement gagnera le match, à l'ouverture et à la clôture des paris ; et les mêmes informations pour le joueur finalement perdant ;
- **implied_probs_winner_open**, **implied_probs_winner_close**, **implied_probs_loser_open** et enfin **implied_probs_loser_close** : les probabilités de gain du match à l'ouverture et à la fermeture des paris, déduites des cotes précédente ;
- **moved_towards_winner** : une variable indiquant si la cote a évoluée en faveur du joueur qui a finalement gagné ;

Le but de cet exercice est donc de caractériser ces matchs et les joueurs présents et de mettre en évidence les joueurs et bookmakers suspectés d'avoir participé à des paris frauduleux en truquant leurs matches. Dans la suite de cette partie nous utiliserons uniquement les données obtenus après le pré-traitement.

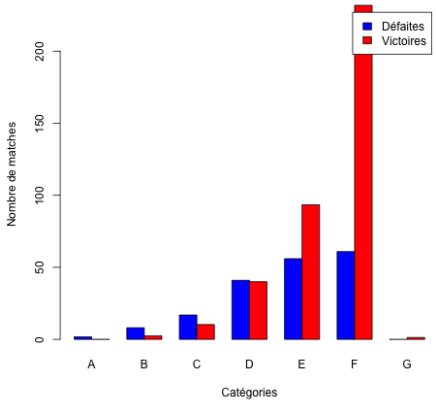
1.1.1 Analyse descriptive générale des données

Apres utilisation des fonctions R comme aggregate, length ou encore unique, on a pu déterminer que la dataset possedait **25993 matchs distincts** qui font l'objet des **126461 paris** a travers **7 bookmakers** . Nous pouvons aussi remarquer **1523 joueurs** distincts ayant joués ces matchs. Toutes ces informations concernent la période de jeu entre **2009 et 2015**.

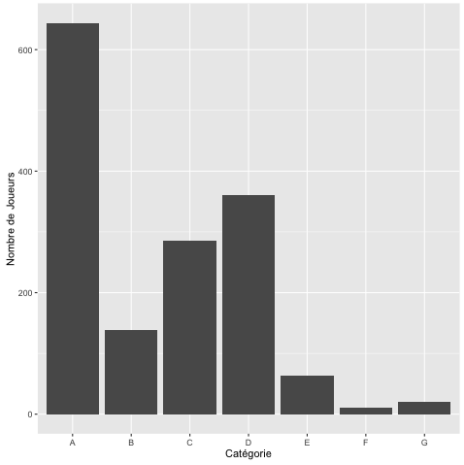
1.1.2 Catégorisation des joueurs

On nous demande par la suite de catégoriser ces joueurs en fonction de leur propension à gagner des matches, et de représenter l'information du nombre de matches joués (gagnés, perdus) en fonction du niveau du joueur . Nous avons donc procédé à la création d'un nouveau dataset avec tous les joueurs et leur nombre de victoires et défaites associées grâce à la fonction aggregate. Pour calculer cette propension a gagner nous avons simplement calculé le ratio nombre de victoires sur nombre de matches joués pour chaque joueur. À partir de cette fréquence, nous avons créé 7 catégories de joueurs représentant 7 intervalles de fréquence, visibles dans la Fig.1(a) . Le detail sur les frequences caracterisants les categories est visible ci-dessous :

- **Catégorie A** :]0.0 , 14,3] comprenant 644 joueurs;
- **Catégorie B** :]14.3 , 28.6] comprenant 138 joueurs;
- **Catégorie C** :]28.6 , 42.9] comprenant 286 joueurs;
- **Catégorie D** :]42.9 , 57.1] comprenant 361 joueurs;
- **Catégorie E** :]57.1 , 71,4] comprenant 63 joueurs;
- **Catégorie F** :]71,4 , 85.7] comprenant 10 joueurs;
- **Catégorie G** :]85,7 , 100] comprenant 21 joueurs;



(a) Distribution des matches gagnés et perdus en moyenne pour chaque catégorie



(b) Distribution du nombre de joueurs pour chaque catégorie

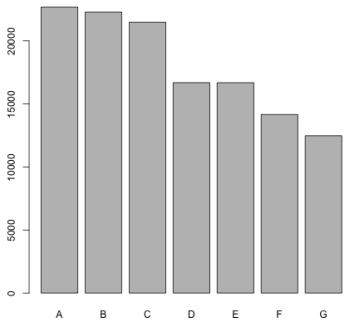
Figure 1:

On pourra s'interroger sur la pertinence de cette catégorisation. En effet ici on nous demande de représenter l'information du nombre de matches joués, gagnés comme perdus en fonction du niveau du joueur or,on a calculé ce même niveau grâce au nombre de matches gagnés. Néanmoins cette discrétisation en 7 intervalles nous a permis de remarquer des joueurs considérés comme appartenant à la meilleure catégorie alors qu'en fait ils n'ont joué seulement quelques matches mais avec 100% de victoires; cela ressort très fortement dans le graphique, pour la catégorie G. Mis à part ce cas marginal, l'information qui ressort de ce graphique est que les meilleurs joueurs sont aussi ceux qui jouent le plus de matches.

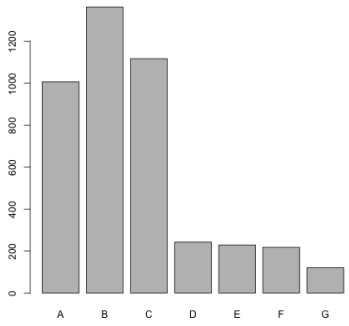
1.1.3 Étude des matches et joueurs suspects

Dans ce point nous étudions tout particulièrement les matches suspects. On considère qu'un match est suspect si la différence en valeur absolue entre la probabilité d'un gagnant(ou du perdant) au début d'un match et à sa fin est plus grande que 0.1.

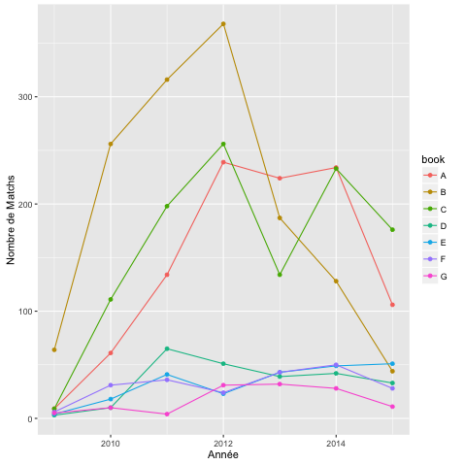
En tenant compte de ce seuil, on remarque la présence de **2798 matchs suspects**. Si on met ce chiffre en rapport avec le nombre total de matches, les matches suspects représentent 10,76% des matches, un nombre relativement important.
En premier lieu on peut caractériser ces matches par bookmaker grâce à une comparaison entre la Fig.2(a) et la Fig.2(b). On voit bien que les bookmakers A, B et C ont un taux de participation relativement plus important dans les matches suspects que les autres bookmakers. On peut aussi remarquer que cette participation dans cette activité a tout d'abord fortement augmenté avant de diminuer ces dernières années grâce à la Figure 2(c).



(a) Nombre moyen de paris par bookmaker



(b) Nombre moyen de paris suspects par bookmaker



(c) Nombre de paris suspects par bookmaker et par année

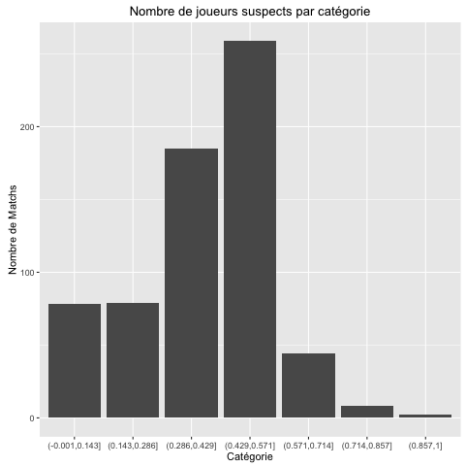
Figure 2:

À partir de ces matchs suspects, on constate que 655 joueurs, gagnants et perdants de ces matchs sont donc impliqués dans ces matchs. On constate qu'ils appartiennent majoritairement aux catégories de niveau "moyen" Fig.3(a).

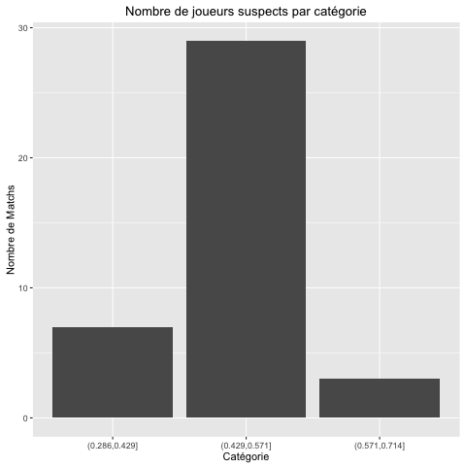
Considérer à la fois les gagnants et les perdants comme suspects nous semblent une bien trop grande approximation. Ainsi, pour mieux cibler les suspects, on va considérer seulement les

perdants ayant perdu plus de 10 matchs suspects. En effet il est plus facile d'influencer le résultat d'un match en le perdant volontairement qu'en le gagnant contre toute attente. Un autre critère important pour mieux cibler ces joueurs est de prendre seulement ceux dont la probabilité de gagner a reculé. Pour se faire on va utiliser l'attribut "moved_towards_winner" qui correspond à l'augmentation de la probabilité du gagnant entre le début et la fin du match, et donc de fait le recul de la probabilité du perdant. Ce critère nous semble pertinent dans la mesure où un joueur perdant dont la probabilité de gagner a evolué positivement au cours du match a tout simplement été meilleur que d'ordinaire mais ne présente pas un comportement frauduleux(Si c'est le cas, il remplit bien mal son rôle en abaissant la cote du gagnant et en se rapprochant de la victoire qu'il veut éviter).

Avec ces critères, on trouve un total de **39 joueurs suspects**. On peut de plus voir l'appartenance de ces joueurs à nos catégories définies auparavant et comparer la distribution avec la Fig.3(b). La catégorie majoritairement concernée et celle du milieu, qui représente les joueurs ayant quasiment autant de défaites que de victoires. Cela a du sens car avec leur propension à gagner des matches autour de 50%, ces joueurs sont quasiment indetectables dans le sens où on ne les suspectera pas forcément de perdre volontairement le match.



(a) Historigramme du nombre de joueurs impliqués dans les matchs suspects par catégorie



(b) Historigramme des perdants très suspects par catégorie

Figure 3:

1.2 Données crabs

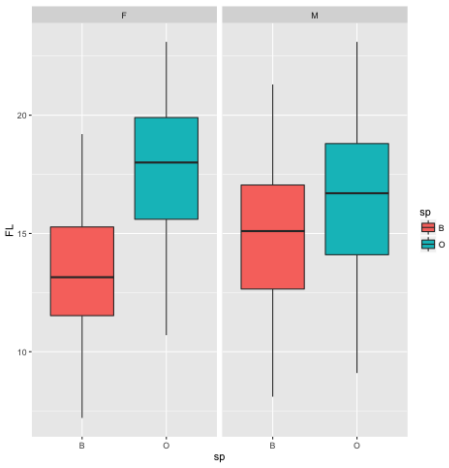
Dans cette partie, on fera une analyse descriptive des données CRABS de la librairie MASS, puis on étudiera la corrélation des différentes variables qui decrivent ces individus.

1.2.1 Analyse descriptive des données

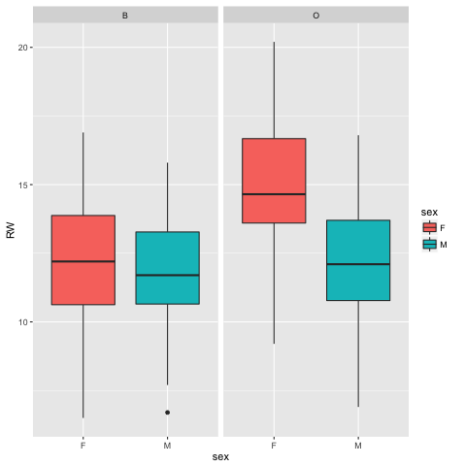
Ce dataset est composé $N = 200$ individus et $D = 8$ variables, 2 qualitatives, 5 quantitatives et 1 variable index qui ne nous servira pas pendant cette étude. Les 2 variables qualitatives, uniformément répartis, concernent le sexe et l'espèce du crabe(2 espèces), les 5 variables quantitatives concernent les caractéristiques morphologiques du crabe comme la taille du lobe frontal ou la largeur de la carapace, intitulées CW, FL, RW, CL, BD.

On va maintenant essayer de voir si il existe des différences de caractéristiques morphologiques selon l'espèce ou encore le sexe. Pour se faire, on va utiliser des boxplots. D'après les Fig.4(a) et Fig.4(b) on remarque qu'on ne peut distinguer le sexe ou l'espèce avec les variable FL et RW. Ce comportement est généralisable à toutes les variables des crabes (voir les autres boxplots en annexes). En effet, dans le rectangle du boxplot se concentre 50% des individus, et on trouve 25% des individus restants de chaque coté de ce rectangle. On remarque que tous ces boxplots se

chevauchent. Ainsi on ne peut estimer avec certitude l'espèce ou le sexe d'un individu selon ces variables.



(a) Boxplot des crabes en fonction de la variable FL, selon le sexe et l'espèce



(b) Boxplot des crabes en fonction de la variable RW, selon le sexe et l'espèce

Figure 4:

1.2.2 Étude de la corrélation des variables quantitatives

Étudions maintenant la corrélation des variables quantitatives entre elles. Nous pouvons visualiser efficacement cette information grâce à un scatterplot. La librairie R ggplot2 nous permet de faire un scatterplot customisé visible en Fig.5(a) . On peut observer toutes les combinaisons possibles des variables quantitatives en fonction des unes des autres, avec leurs coefficients de correlation avec une couleur différente pour chaque espèce .Enfin ce graphique nous montre les densités sur la diagonale de la même variable pour chacune des espèces, ce qui nous permet de davantage distinguer si on peut discriminer l'espèce avec l'aide de cette variable.

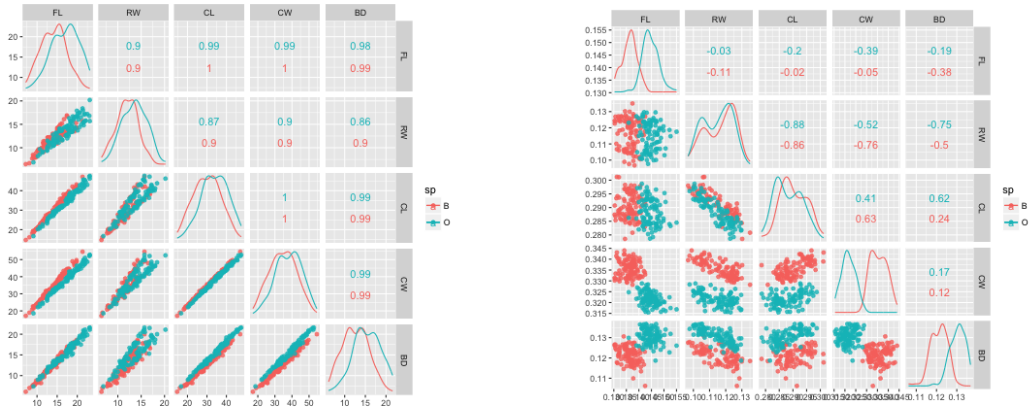
On observe bien dans la cas présent que les individus se chevauchent et donc que ces variables sont toutes très corrélées. Cela est confirmé par les densités ayant quasiment la même distribution et les coefficients de corrélation linéaire très proches de 1. Cela est dû au fait de la proportionnalité morphologique des crabes. En effet les variables quantitatives correspondent aux mesures de certaines parties d'un crabe, ainsi si une partie du crabe est grande, les autres suivent. Dans le cas contraire on obtiendrait un crabe disproportionné.

Pour s'affranchir de cet effet de comparaison entre les grands et petits crabes, on doit ramener tous les crabes à une même échelle. Pour se faire, on crée une variable représentant l'addition de toutes les autres variables, pour chaque crabe et on divise les 5 variables quantitatives par cette nouvelle variable pour chacun des crabes. Ainsi on ne peut plus reconnaitre les grands des petits crabes.

On obtient un nouveau scatterplot, Fig.5(b), on peut voir les spécifités de chaque espèce avec des variables décorellées. Cela se confirme avec la valeur des coefficients de corrélation présents, bien moindre que les précédents. On distingue en effet facilement des clusters pour chaque espèce. On obtient les mêmes informations en réalisant ce scatterplot en fonction du sexe.

2 Analyse en Composantes Principales

Dans cette section nous allons nous pencher sur la technique dite de l'Analyse en Composantes Principales qui consiste à transformer des variables corrélées en nouvelles variables décorrélées les



(a) Avant la transformation

(b) Après la transformation

Figure 5: Scatterplot des variables quantitatives en fonction de l'espèce incluant le traçé des densités et les coefficients de correlation

unes des autres. On réalisera d'abord l'ACP sur un dataset réduit $N = 4$ observations et $D = 3$ variables, sur le dataset notes étudié en cours puis sur le dataset des crabes.

2.1 Exercice Théorique

On associe les mêmes pondérations à tous les individus, et on munit $\mathcal{R}^P$ de la métrique euclidienne. Trois variables mesurées sur quatre individus fournissent la matrice A.

2.1.1 Axes Factoriels et Pourcentages d'inertie

Pour calculer les axes factoriels de l'ACP du nuage defini, on commence par centrer la matrice A en la plaçant à l'origine le centre de gravité du nuage des individus en soustrayant sa moyenne a chaque colonne. On obtient la matrice A_c. On calcule la matrices des variances en utilisant la relation: $V = \frac{1}{4} * A_c^T * \mathrm{A}_c$. On obtient par la suite les vecteurs propres / axes d'inertie avec la matrice U dont chaque colonne représente un axe d'inertie et les valeurs propres $\lambda_1, \lambda_2, \lambda_3$ rangées dans l'ordre décroissant.

$$A = \begin{bmatrix} 3 & 4 & 3 \\ 1 & 4 & 3 \\ 2 & 4 & 6 \\ 4 & 1 & 2 \end{bmatrix}, \; A_c = \begin{bmatrix} 0.5 & 1 & -0.5 \\ -1.5 & 1 & -0.5 \\ -0.5 & 0 & 2.5 \\ 1.5 & -2 & -1.5 \end{bmatrix}, \; U = \begin{bmatrix} 0.524 & 0.339 & 0.781 \\ -0.509 & -0.611 & 0.606 \\ -0.683 & 0.716 & 0.147 \end{bmatrix},$$

$$\lambda_1 = 3.1988922, \lambda_2 = 1.4684861, \lambda_3 = 0.3326217$$

On calcule ainsi que le premier axe représente **64%** de l'inertie, le deuxième **29.4%** et le troisième **6.6%**. Ainsi **93.4%** de "l'information" se concentre sur les 2 premiers axes, ce qui rend pertinent le fait de représenter ces variables sur le premier plan factoriel.

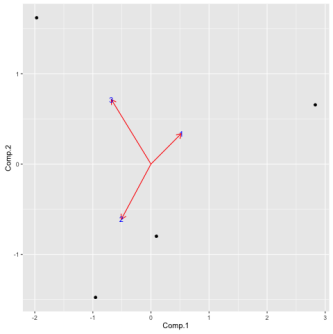
2.1.2 Calcul des Composantes Principales

Les composantes principales sont obtenues par la relation $C = A_c M U = A_c U$, visibles dans chacune des colonnes de la matrice C :

$$C = \begin{bmatrix} 0.094 & -0.800 & 0.923 \\ -0.954 & -1.48 & -0.640 \\ -1.97 & 1.62 & -0.0217 \\ 2.83 & 0.656 & -0.262 \end{bmatrix}$$

2.1.3 Représentation des variables dans le premier plan factoriel

On peut maintenant représenter les variables et les individus ensemble dans le premier plan factoriel visible dans la Fig.6(a)



(a) Représentation des variables et individus dans le premier plan factoriel

Figure 6:

On remarque que les 3 variables sont représentées assez équitablement par les composantes 1 et 2.

2.1.4 Reconstitution de la matrice

On va donc maintenant restituer la matrice de départ grâce à la relation $\sum_{\alpha=1}^{k} c_\alpha {u_\alpha}'$ pour les valeurs k = 1, 2 et 3

$$k = 1 : \begin{pmatrix} 0.0492 & -0.048 & -0.064 \\ -0.500 & 0.486 & 0.651 \\ -1.032 & 1.003 & 1.344 \\ 1.482 & -1.441 & -1.931 \end{pmatrix} \quad k = 2 : \begin{pmatrix} -0.221 & 0.440 & -0.636 \\ -1.000 & 1.388 & -0.406 \\ -0.483 & 0.013 & 2.503 \\ 1.704 & -1.841 & -1.461 \end{pmatrix} \quad k = 3 : \begin{pmatrix} 0.5 & 1 & -0.5 \\ -1.5 & 1 & -0.5 \\ -0.5 & 0 & 2.5 \\ 1.5 & -2 & -1.5 \end{pmatrix}$$

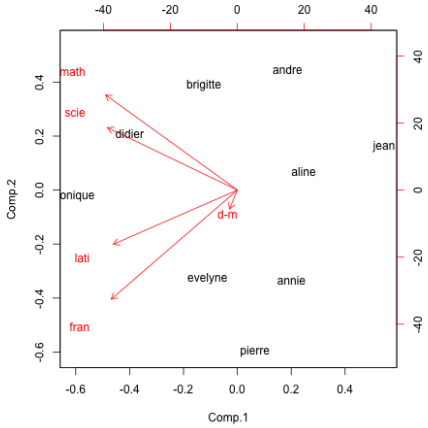
Avec $k = 3$ on retrouve logiquement la matrice A_c sur laquelle on a pratiqué cet ACP, en effet sur les trois composantes il est normal de retrouver 100% de l'information de départ

2.2 Utilisation des Outils R

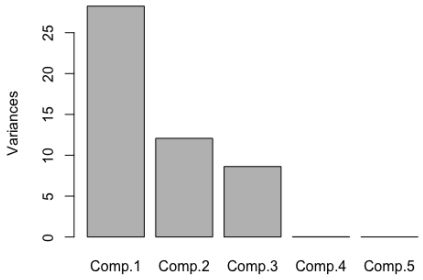
L'objectif de cet exercice est de se familiariser avec les fonctions R permettant d'effectuer une ACP, en particulier les fonctions princomp, summary, loadings, plot et biplot. On utilisera pour se faire le jeu de données notes étudiées en cours. Notes concernent $N = 9$ individus et $D = 5$ variables. La fonction princomp nous permet de réaliser l'ACP. La fonction summary nous donne la variance des composantes principales et le pourcentage de variance expliquée pour chaque composante. Les attributs loadings et scores nous donnent respectivement les axes d'inertie et les composantes principales. On remarquera aussi qu'on peut obtenir les valeurs propres fournies par l'attribut sdev.

La fonction biplot nous permet de visualiser les variables et les individus dans le premier plan factoriel, visibles dans la Fig.7(a). On peut detecter rapidement les corrélations entre les variables, par exemple entre les maths et les sciences, le français et le latin. On remarque d'ailleurs que la composante 2 est celle qui va nous permettre de discriminer les élèves bons en sciences en général dans le positif, et les moins bons, dans le négatif.

La fonction plot nous permet de voir les proportions de variances qu'expliquent les différentes composantes principales dans la Fig.7(b). On voit ici que les composantes 1,2 et 3 concentrent la majorité de l'information et pourraient décrire le dataset à elles seules.



(a) Individus et Variables projetés sur le premier plan factoriel



(b) Variance expliquée pour chaque composante

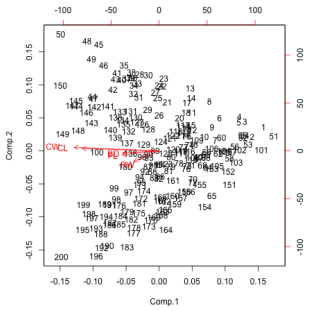
Figure 7:

2.3 Traitement des données Crabs

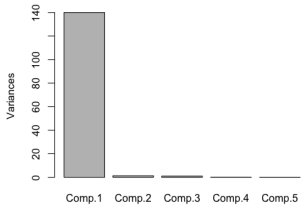
On va maintenant pouvoir tester l'ACP sur les données crabes étudiées dans la section 1.2.2 pour trouver une représentation visuelle des crabes qui permettent de distinguer visuellement différents groupes.

2.3.1 Sans décorrélation des variables

On utilise pour cela la fonction princomp puis biplot Fig.8(a) et Fig.8(b). On constate sans surprise qu'une composante principal, la composante 1, concentre l'écrasante majorité de l'information. En effet toutes les variables sont orientés selon la composante 1 dans le biplot et les composantes principales 2,3,4 et 5 sont quasiment nulles dans le plot. Cela est du au fait que l'on a pas encore décoréllé ces variables comme on a pu le faire dans la section 1.2.2. On ne peut donc pas comparer et distinguer des clusters si on peut encore faire la distinction entre les "gros" et "petits" crabes.



(a) Individus et Variables dans le premier plan factoriel avant transformation

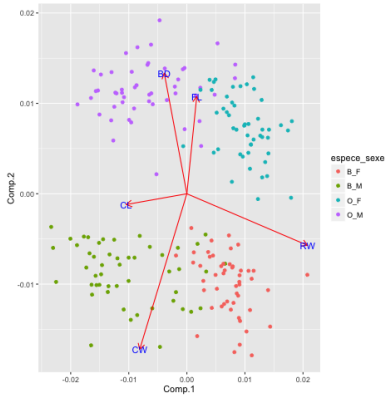


(b) Variance expliquée pour chaque composante avant transformation

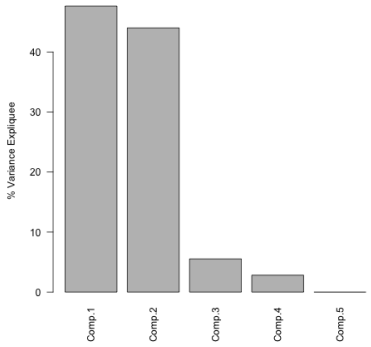
Figure 8: Résultat de l'ACP avant transformation, les 5 variables sont alors très corrélées

2.3.2 Après décorrélation des variables

Afin de décoreller les variables, on réalise la manipulation décrite en 1.2.2. Suite à cela nous avons combiner les 2 variables qualitatives, sexe et espèce en une variable qualitative qui prend 4 valeurs possibles de la forme Male/Femelle_O/B. C'est ainsi qu'on distingue 4 clusters relativement distincts dans la Fig.9(a). Plus généralement on constate que l'information du sexe est encodée dans la composante 1, et l'information de l'espèce est encodée dans la composante 2. On remarque aussi que les variables BO, FL et CW sont décrites principalement par la composante 2, et que les variables CL et RW le sont par la composante 1, cette répartition est confirmée dans la Fig.9(b) qui nous montre que 2 composantes principales concentrent le maximum de l'information



(a) Individues et Variables dans le premier plan factoriel après transformation



(b) Variance expliquée pour chaque composante après transformation

Figure 9: Résultat de l'ACP après transformation

3 Conclusion

Ce projet nous aura permis dans un premier temps d'explorer les outils de R dédiés à la statistique descriptive et le langage R en lui-même afin de dégager des informations non triviales. On a pu dans un deuxième temps comprendre et éxécuter l'ACP sur divers jeu de données. Cette puissante méthode nous permet de visualiser différemment et de manière efficace les données, en les projetant sur des nouveaux axes, non corrélés, éliminant ainsi la redondance. Sur des données très corrélées on pourra même s'attendre à des valeurs propres nulles et donc une réduction de dimension du dataset.