Rapport SY09 -TP1

Statistique descriptive, Analyse en composantes principales

Shuhan LIN
Grégory MAYEMBA

1 Statistique Descriptive

Données babies

Le but de cet exercice est d'effectuer des études sur des données des babies, qui contiennent 1236 individuels et 23 variables. Dans notre problème, nous étudions 8 variables : le poids à la naissance, la durée de gestation, le nombre de grossesses précédentes, la taille de la mère et le poids de la mère, l'âge de la mère, si la mère fume ou non et le niveau d'éducation de la mère.

Question 01

A. Résumé numérique

Nous avons observé le résumé numérique du poids de nouveau né par rapport au tabagisme des mères.

Min	1st.Qu.	Médian	Moyenne	3rd.Qu	Max
55.0	113.0	123.0	123.0	134.0	176.0

Résultat des valeurs du poids pour bébés nés d'une mère non fumeuse.

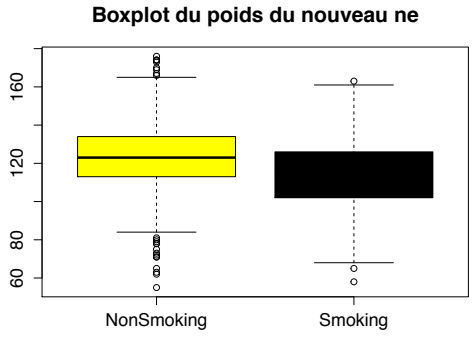
Min	1st.Qu.	Médian	Moyenne	3rd.Qu	Max
58.0	102.0	115.0	114.1	126.0	163.0

Résultat des valeurs du poids pour bébés nés d'une mère fumeuse

D'après ce résumé, nous remarquons que les valeurs du poids du nouveau né d'une mère non fumeuse est généralement supérieures que celles d'une mère fumeuse.

B. Boxplot

Ensuite, nous désignons un Boxplot pour voir plus en détail.



Ce Boxplot confirme parfaitement notre conclusion. Nous trouvons que la majorité des bébés nés d'une mère non fumeuse ont du poids entre 110 et 130, et celles de mère fumeuse ont du poids entre 100 et 120

Question 02

A. Résumé numérique

D'abord, nous avons également fait le résumé numérique du temps de gestation par rapport au tabagisme des mères.

Min	1st.Qu.	Médian	Moyenne	3rd.Qu	Max
148.0	273.0	281.0	282.2	289.0	353.0

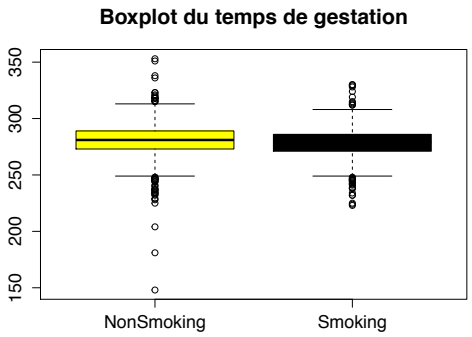
Résultat des valeurs du temps de gestation pour bébés nés d'une mère non fumeuse.

Min	1st.Qu.	Median	Moyenne	3rd.Qu	Max
223.0	271.0	279.0	278.0	286.0	330.0

Résultat des valeurs du temps de gestation pour bébés nés d'une mère fumeuse.

Nous comparons la médian et la moyenne de deux groupe. Il paraît qu'il n'y a une grande différence. Donc, d'après ce résumé, nous ne pouvons pas être sûr que le tabagisme à une influence évident sur du temps de gestation.

B. Boxplot



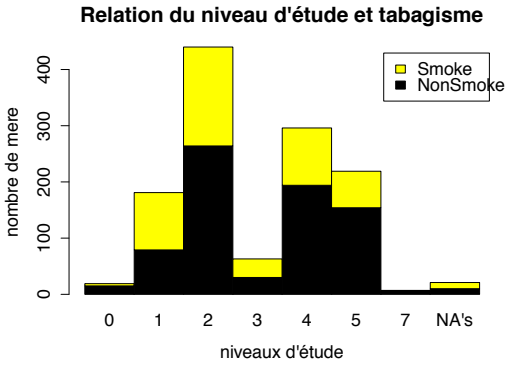
Ce graphique montre également que les deux groupes n'ont pas de différences claire. Par conséquent, nous ne pouvons pas dire si le tabagisme influence du temps de gestation.

Question 03

	0	1	2	3	4	5	7
fumeuse	4	102	176	33	102	65	1
non fumeuse	1	79	264	30	194	154	6

D'abord, c'est un tableau de niveau d'étude des mère fumeuse ou non fumeuse.

D'après ce tableau, nous constatons des différences claires entre différents niveaux d'étude. Parmi ces valeurs, les groupes de niveau 2,4 et 5 ont des différences plus importantes. D'ailleurs, il n'existe pas des mères qui sont dans niveau 6.

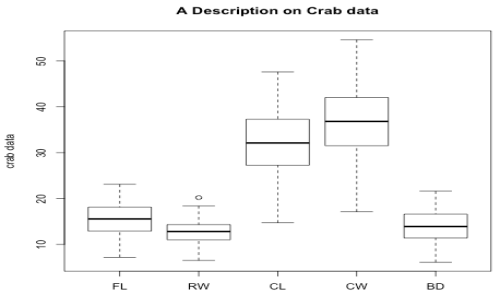


Ce graphe nous montre plus clairement que deux groupes différencient plus au niveau d'étude 2 ,4 ,5.

Données Crabes

Question01

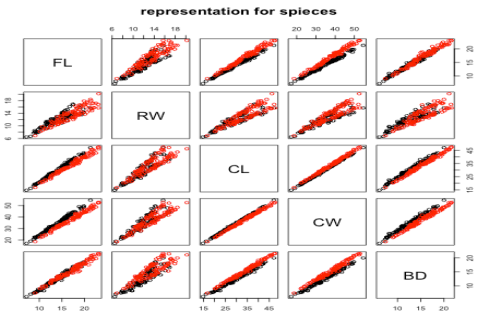
D'abord, nous avons effectué un graphe en boîte pour l'analyse descriptive des données.

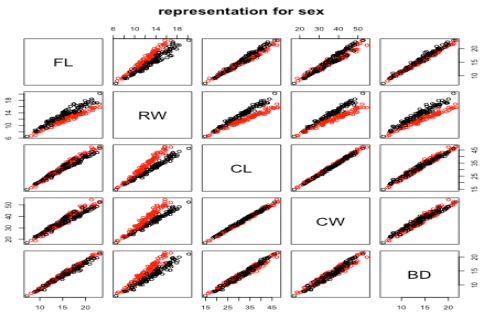


Nous constatons qu'au niveau des distributions des valeurs, les valeurs de FL, RW et BD sont tous beaucoup moins que celles de CL et CW.

Ensuite, nous avons étudié le graphe de dispersion

pour chaque deux variable au niveau des espèces et des sexes.





D'après ces deux graphes, nous remarquons que toutes les variables sont très corrélées et il est un peu difficile de classifier des espèces et sexes.

Question02

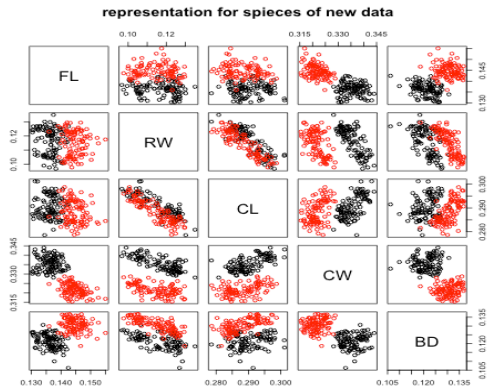
Pour laisser les données moins corrélées, nous devons éliminer les effets de taille des nos données.

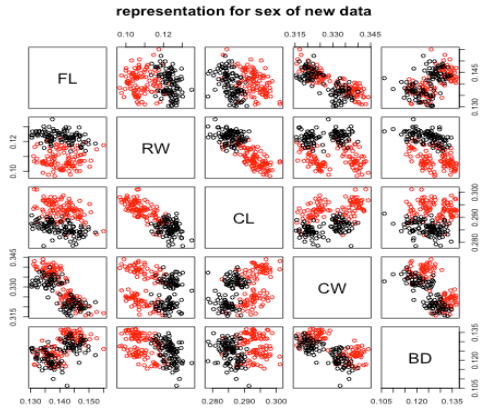
Nous divisons chaque variable par la somme de toutes les 5 variables. Et puis, nous avons calculé la matrice de corrélation.

```
          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
```

Comme cela, les données sont mois corrélé.

Nous constatons les graphes comme ci-dessous :





D'après le graphe, les données sont beaucoup mieux séparées.

La conclusion est que nous devons faire attention aux effets de taille des données, surtout sur des données biologiques.

2. Analyse ACP

2.1 Exercice théorique

Le principe de cet exercice est de faire l'ACP de trois variables mesurées sur quatre individus. Le nombre de données à traiter étant faible, cet exercice nous permet de nous familiariser avec la méthode de l'ACP sous R.

Question 01 Calcul des axes factoriels

Nos données sont dans la matrice

$$X = \begin{bmatrix} 3 & 4 & 3 \\ 1 & 4 & 3 \\ 2 & 3 & 6 \\ 2 & 1 & 2 \end{bmatrix}$$

On calcul ensuite la matrice de variance

$$V = X^t D_p X = \begin{bmatrix} 4.5 & 6 & 7 \\ 6 & 10.5 & 11 \\ 7 & 11 & 14.5 \end{bmatrix}$$

La diagonalisation de la matrice V fournit les valeurs propres et les axes factoriels associés (les vecteurs propres). Ci dessous un tableau contenant les % d'inertie et % d'inertie cumulés correspondants aux valeurs propres associées.

Valeurs Propres	Λ1= 27.39	Λ2= 1.34	Λ3= 0.76
% d'inertie	92.86	4.55	2.59
% d'inertie cumulé	92.86	97.41	100

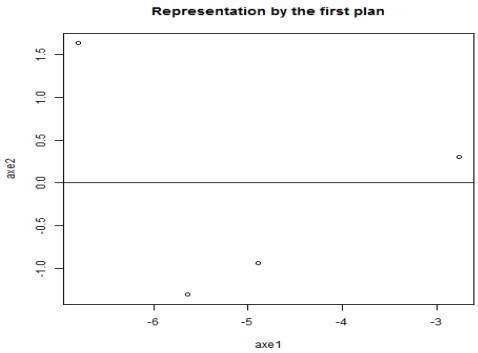
Les 2 premiers axe principal donne 97.41 % de l'information. Nous pouvons donc visualiser 97.41% de l'information sur le plan factoriel composé des deux premiers axes factoriels.

Question 02 Calcul des composantes principales

D'après la formule, on a

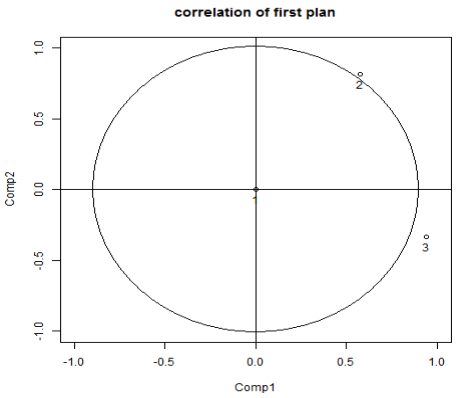
$$C = XMU = \begin{bmatrix} -5.64 & -1.31 & 0.72 \\ -4.89 & 0.94 & -1.10 \\ -6.80 & 1.64 & -0.26 \\ -2.76 & 0.30 & 1.12 \end{bmatrix}$$

Ci-dessous, la représentation des individus dans le premier plan factoriel :



Question03 Représentation des variables dans le premier plan factoriel

Pour avoir les coordonnés des variables dans le premier plan factoriel, il faut calculer les corrélations grâce à la matrice D= cor(X, C)



Les variables 2 et 3 sont représentatives des individus car elles sont proches du cercle, de plus ces deux variables sont corrélées car leurs coordonnés sur l'axe 2 sont très proches.

Question 04 Calcul pour déduire le donnée de sources ,

Nous avons fait le calcul: $k_i = \sum_{i=1}^{p} c_i * u_i'$

$$k1 = \begin{bmatrix} 2.11 & 3.36 & 4.01 \\ 1.83 & 2.91 & 3.48 \\ 2.54 & 4.05 & 4.84 \\ 1.03 & 1.65 & 1.97 \end{bmatrix}$$

$$k2 = \begin{bmatrix} 2.35 & 4.28 & 3.11 \\ 2.00 & 3.57 & 2.83 \\ 2.24 & 2.90 & 5.96 \\ 0.98 & 1.44 & 2.17 \end{bmatrix}$$

$$k3 = \begin{bmatrix} 3 & 4 & 3 \\ 1 & 4 & 3 \\ 2 & 3 & 6 \\ 2 & 1 & 2 \end{bmatrix}$$

Nous trouvons que la matrice k3 est exactement la même comme la matrice des données source.

2.2 Utilisation des outils R

Question 01 En utilisant ces fonctions, effectuer l'ACP du jeu de données notes étudiées en cours.

Dans cet exercice, nous utilisons la fonction R, *princomp* pour mise en œuvre l'ACP en formule :

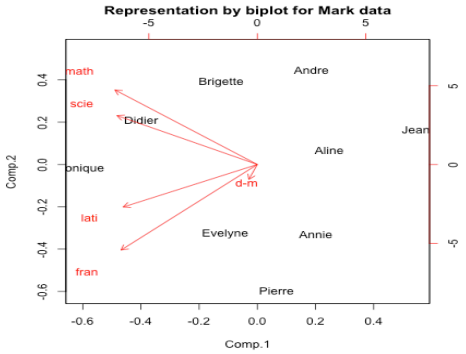
$acp = princomp(data)$

Valeurs propres: $acp\$sdev$

Vecteurs propres : $acp\$loadings$

Composantes : $acp\$scores$

Et puis, nous avons désigné le graphe de la représentation par premières deux composantes principales.



D'après ce graphe, nous avons constatons que sauf la variable '*d-m*', les autres sont assez bien représentés.

Question02 La fonctionnement de plot et biplot

D'abord, '*plot*' est une fonction générale pour designer des graphes. Par conséquent, il peut designer des graphes de représentation pour deux composantes

D'ailleurs, *'biplot'* nous permet de mettre les individus et les variables dans le même graphe.

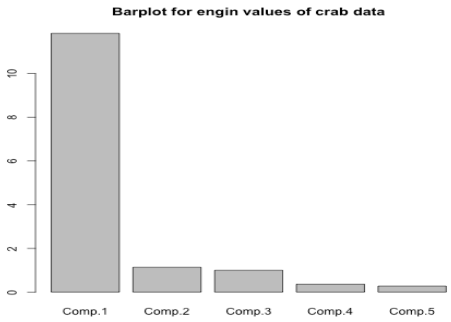
Ensuite, *'biplot.princomp'* est une fonction plus avancé et spécialisé pour l'ACP. Sauf la représentation de deux composantes, il peut également évaluer la corrélation entre chaque variable et sa composante, qui nous permet d'évaluer la qualité de représentation.

Nous avons aussi étudié les options de *'biplot.princomp'*. L'option *'choices'* se sert de définir la taille de vecteurs et *'scale'* est pour obtenir une représentation standard avec des données standard (*scaled*). En finale, l'option *'pc.plot'* se sert de désigner un plot avec des observations standardisé par la racine de N.

2.3 Traitement des données Crabs

Question01 Tester tout d'abord l'ACP sur *crabsquant* sans traitement préalable.

Nous avons d'abord effectué l'ACP sans traitement. D'abord, nous regardons les valeurs propres (Inertie d'expliquée pour chaque axe)

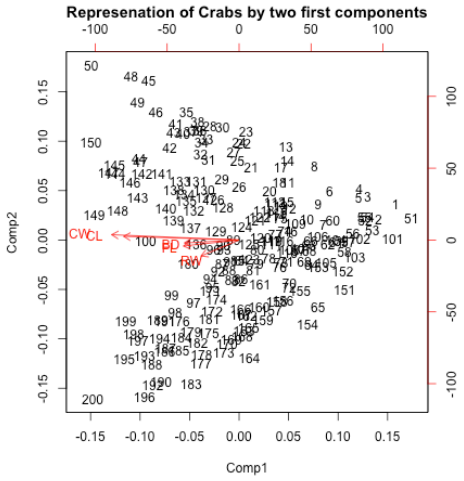


Nous remarquons que la première composante a un pourcentage très important par rapport les autre composantes.

Voici le tableau de pourcentages cumulés :

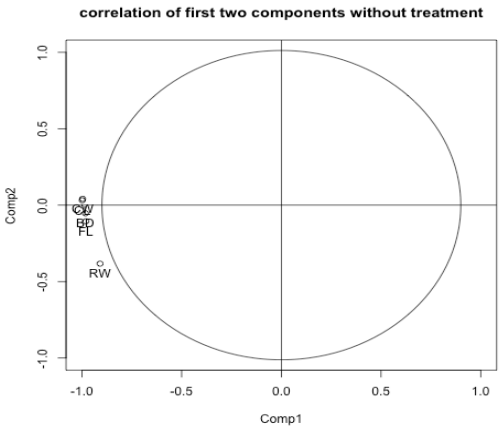
1	2	3	4	5
80.9%	88.7%	95.4%	98.0%	100%

Ensuite, nous étudions la représentation par biplot.



D'après ce graphe, les angles entre les variables sont très petits. Autrement dit, toutes les variables sont fortement corrélées.

Egalement, nous regardons le graphe de corrélation :



Nous constatons que les variables sont assez bien représentées au niveau du pourcentage de l'inertie. Par contre, d'après le graphe de corrélation, les variables sont corrélés avec le premier axe et mal corrélés avec le deuxième axe.

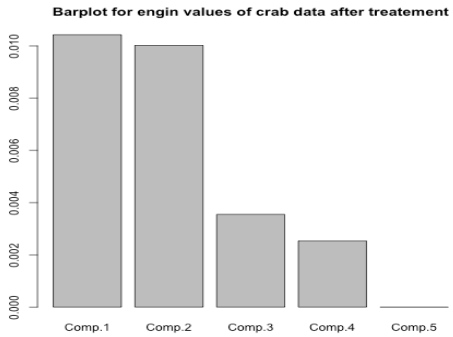
Les variables sont trop corrélées à cause de l'effet taille.

Question02 Trouver une solution pour améliorer la qualité de votre représentation

Solution01

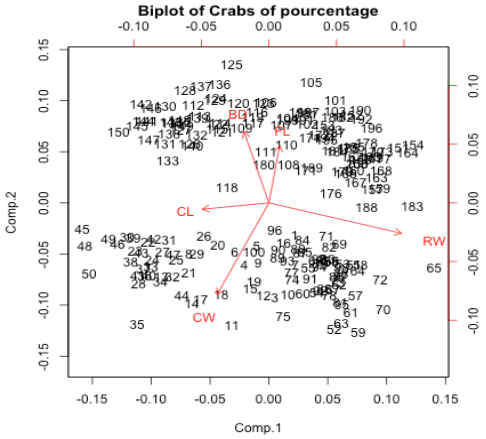
Nous avons fait la même manipulation comme l 'exercice précédant, c'est que nous divisons chaque variable par le somme de toutes des variables et obtenons des pourcentage. Comme cela, nous éliminons l'effet de taille.

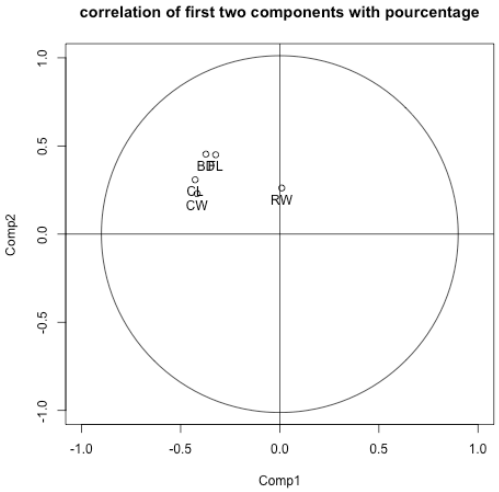
Nous avons effectué également la barplot pour des inerties expliquées



1	2	3	4	5
39.2%	77.0%	90.4%	99.9%	100%

Et puis, nous avons effectué le biplot :





Nous trouvons que les données sont beaucoup moins corrélées maintenant. Et puis, nous trouvons que la qualité de représentation est pire que celle sans traitement au niveau de pourcentage de l'inertie. En revanche, sauf RW, les

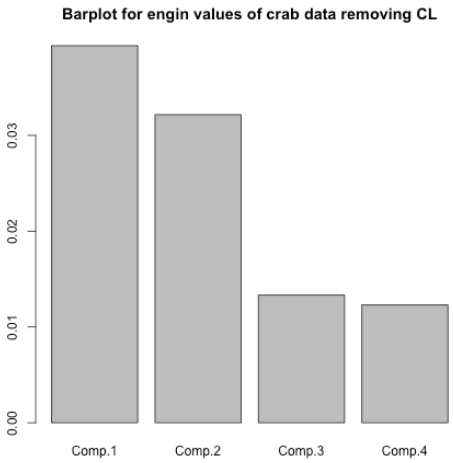
autres variables ont mieux corrélés selon le graphe de corrélation.

D'ailleurs, selon le biplot, nous pouvons constatons généralement une séparations des données. (Il y a généralement 4 groupes pour des crabs de deux espèces et deux sexes).

Solution02

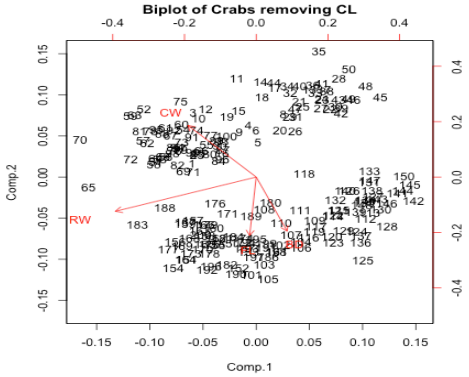
Il y a une autre méthode pour éliminer l 'effet taille. C'est que nous divisons chaque variable par le troisième variable *'CL'*. Et puis, nous enlevons cette variable. Comme cela, nous pouvons aussi éliminé l'effet taille.

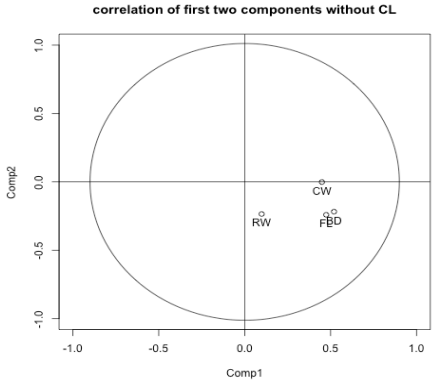
D'abord, nous avons effectué le barplot pour inerties expliquées :



1	2	3	4
40.2%	73.5%	87.2%	100%

Ensuite, nous avons désigné le biplot et graphe de corrélation:





Le résultat ressemble à première solution.
C'est que la qualité de représentation est pire que celle sans traitement au niveau du pourcentage d'inertie et elle se présente mieux au niveau du graphe de corrélation.

Et puis, nous pouvons constaté en gros une séparation des groupes d'après le graphe de biplot.

Conclusion

Après avoir fait ce projet, nous avons d'abord étudié les méthodes exploratoires élémentaires des données par le premier exercice.

Pour des données monodimensionnelles, nous utilisons des statistiques élémentaires (Moyenne, Variance, Médiane etc.), le histogramme et diagramme en boîte comme les méthodes d'analyse.

Pour des données multidimensionnelles (Dans le

premier exercice, nous intéressons plutôt aux données bidirectionnelles), nous avons des méthodes comme graphe de dispersion, graphe de corrélation etc.

Ensuite, par le deuxième exercice, nous avons étudié l 'analyse des composantes principales.
Nous avons effectué l'ACP manuellement et par le logiciel R, avec lequel nous avons appris des façons différentes à représenter des données (plot, biplot, biplot.princomp etc.)

D'ailleurs, nous avons pratiqué l'ACP sur les données Crabs, où nous avons étudié les principes à estimer la qualité de représentation pour l'ACP. Nous avons calculé le pourcentage de l'inertie expliquée et nous avons désigné le graphe de corrélation des variables.

Comme conclusion, nous avons trouvé que le pourcentage de l'inertie expliquée n'est pas le seul indicateur de la qualité de l'ACP (Comme dans les données crabs). Les corrélations des variables est aussi très important.