Master en Biologie et Génétique en Santé

Statistiques et Analyse de Données

Section2: Analyse en Compsantes Principales (ACP)

Analyse en Composantes Principales (ACP)

Introduction

- ACP = techniques descriptive permettant d'étudier les relations qui existent entre des variables quantitatives
- Sur le plan théorique, l'Analyse en Composantes Principales est une méthode relativement complexe, dans la mesure où elle fait appel à des notions mathématiques non élémentaires : celles de matrices, de vecteur propre, de valeur propre...
- Fort heureusement, il n'est pas nécessaire de connaître ces notions pour comprendre le mécanisme d'une ACP et donc pour l'utiliser correctement.

Analyse en Composantes Principales (ACP)

2**. Définition des Composantes Principales**

Objectifs de l'ACP:

L'ACP s'intéresse à des tableaux de données rectangulaires avec des individus en lignes et des variables quantitatives en colonnes

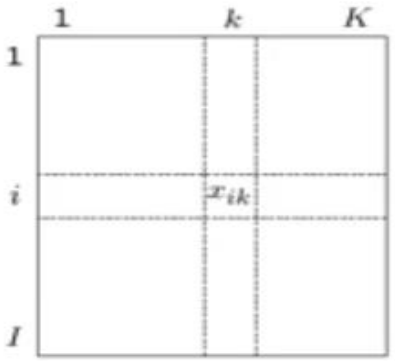


FIGURE : Tableau de données en ACP

Pour la variable k, on note :

la moyenne : $\bar{x}_k = \frac{1}{I}\sum_{i=1}^{I} x_{ik}$

l'écart-type : $s_k = \sqrt{\frac{1}{I}\sum_{i=1}^{I}(x_{ik} - \bar{x}_k)^2}$

- Un tel tableau peut contenir un grand nombre de cellule et on va essayer de résumer les données de manière à prendre plus facilement connaissance de l'information qu'elles contiennent

Analyse en Composantes Principales (ACP)

Objectifs de l'ACP:

Etude des individus

- Quand dit-on que 2 individus se ressemblent du point de vue de l'ensemble des variables ?
- Si beaucoup d'individus, peut-on faire un bilan des ressemblances ?

$\Rightarrow$ construction de groupes d'individus, partition des individus

Analyse en Composantes Principales (ACP)

Objectifs de l'ACP:

Etude des variables

- Recherche des ressemblances entre variables
- Entre variables, on parle plutôt de liaisons
- Liaisons linéaires sont simples, très fréquentes et résument de nombreuses liaisons ⇒ coefficient de corrélation

⇒ visualisation de la matrice des corrélations
⇒ recherche d'un petit nombre d'indicateurs synthétiques pour résumer beaucoup de variables (ex. d'indicateur synthétique a priori : la moyenne, mais ici on recherche des indicateurs synthétiques a posteriori, à partir des données)

Analyse en Composantes Principales (ACP)

Objectifs de l'ACP:

Lien entre les deux études

- Caractérisation des classes d'individus par les variables
  $\Rightarrow$ besoin de procédure automatique
- Individus spécifiques pour comprendre les liaisons entre variables
  $\Rightarrow$ utilisation d'individus extrêmes (en terme de variables : langage abstrait mais puissant, revenir aux individus pour voir les choses plus simplement)

Objectifs de l'ACP :

- Descriptif - exploratoire : visualisation de données par graphiques simples
- Synthèse - résumé de grands tableaux individus $\times$ variables

Analyse en Composantes Principales (ACP)

Application de la méthode à un exemple

- Exemple: la recherche a mis au point 6 variétés de riz. En vue de proposer aux paysans les variétés ayant les bonnes caractéristiques physico-culinaire, une évaluation des variétés a été faite et les résultats sont consignés dans le tableau.
- L'objectif est de synthétiser au mieux l'information contenu dans le tableau.
- Une ACP sera donc réalisée sur ce tableau

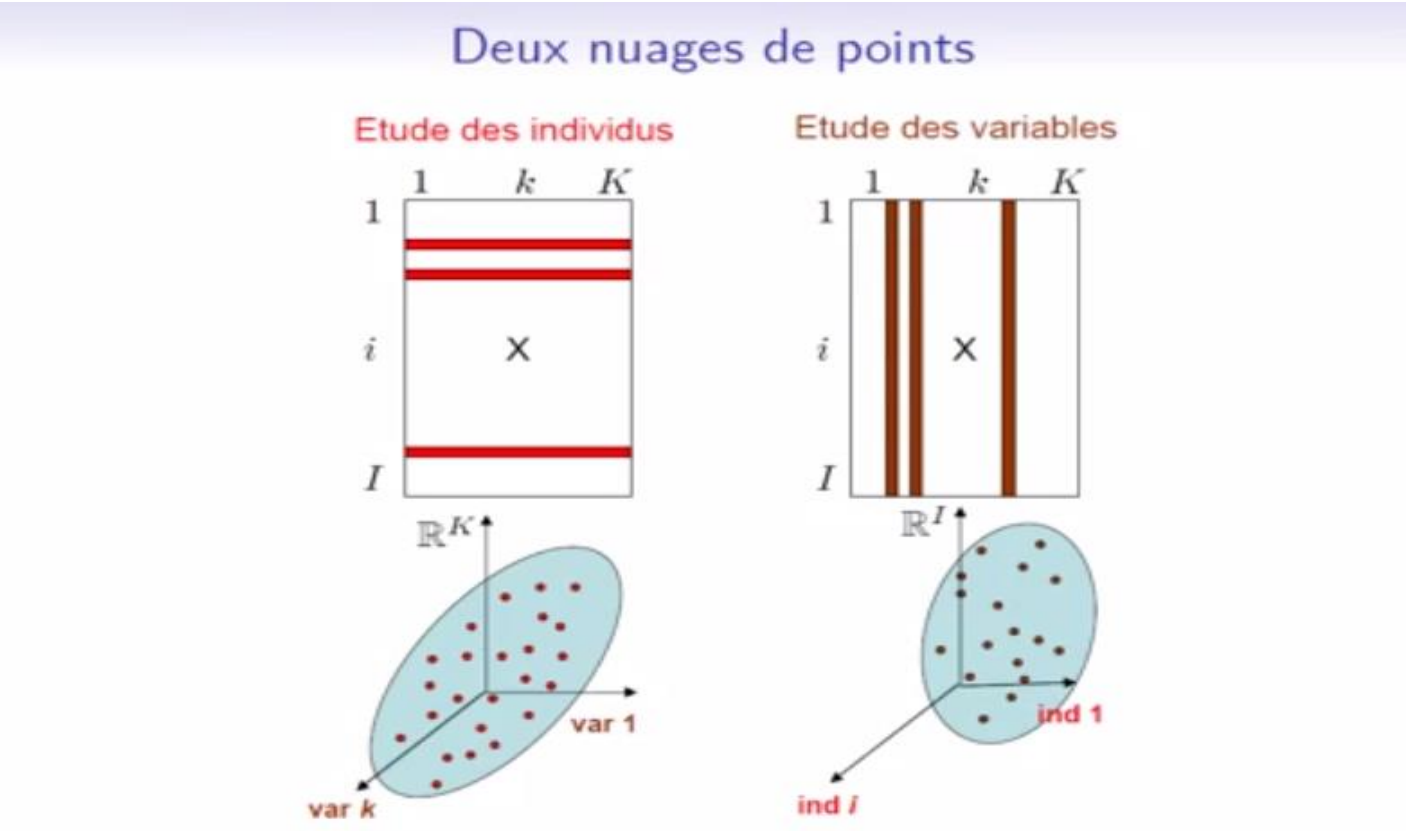
Analyse en Composantes Principales (ACP)

Tableau : Caractéristiques physico-culinaires des variétés de riz

Variétés de riz	Temps de cuisson (TC)	Elongation (EL)	Taux de brisure (TB)	Taux de riz entier (TRE)	Augmentation de volume (AV)
BOUAKE	13,33	1,35	1,83	81,21	3,34
NERICA2	23,67	1,26	1,19	77,24	3,08
WITA	23	1,31	1,67	68,45	3,53
GAMBIAKA	14	1,33	0,83	79,81	3,07
WAB	18,33	1,25	0,54	82,74	2,86
NERICA1	26,33	1,26	1,31	76,83	3,15

Analyse en Composantes Principales (ACP)

Application de la méthode à un exemple



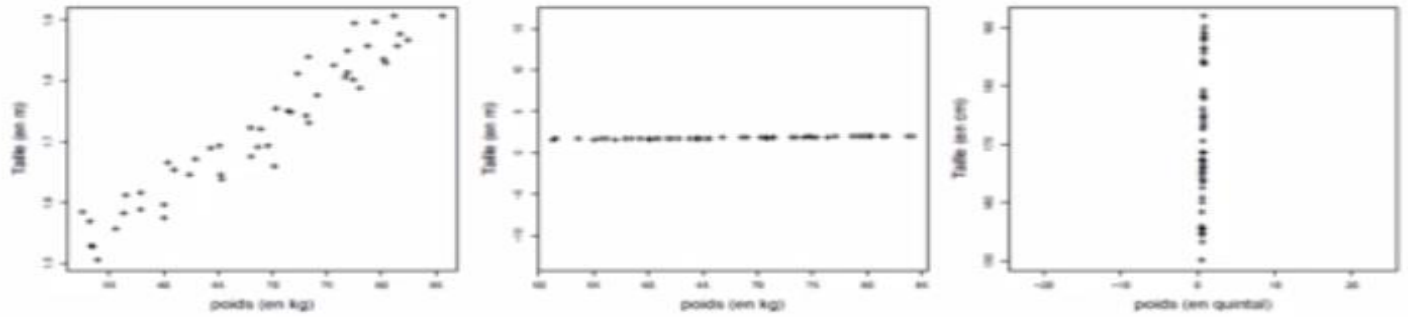
Pour étudier les individus, il faut les projeter dans un espace de dimension R^5 et les variables (ici les caractéristiques physicochimiques) dans un espace de dimension R^6). Il faut donc une méthode de réduction de dimension: une ACP

Analyse en Composantes Principales (ACP)

Application de la méthode à un exemple

Centrage – réduction des données

- Centrer les données ne modifie pas la forme du nuage
  $\Rightarrow$ toujours centrer



- Réduire les données est indispensable si les unités de mesure sont différentes d'une variable à l'autre

$$x_{ik} \hookrightarrow \frac{x_{ik} - \bar{x}_k}{s_k}$$

Analyse en Composantes Principales (ACP)

- Le principe de l'ACP est de définir des variables synthétiques, z ou **composantes principales** qui résument au mieux l'information contenue dans le tableau
- La première variable z_1 est définie de manière à contenir le maximum d'informations relatives du tableau. Il doit respecter:

1. z_1 = combinaison linéaire des variables centrées et réduites du tableau (soit x_1, x_2, x_3, x_4, x_5 ces variables ie dire TC,EL,TB,TRE,AV)

$$Z_1 = u_1x_1 + u_2x_2 + u_3x_3 + u_4x_4 + u_5x_5$$

Analyse en Composantes Principales (ACP)

2. Les coefficients u_1 , u_2, u_3, u_4 et u_5 doivent être telles que Z_1 soit normé

C'est –à dire $u^2_1 + u^2_2 + u^2_3 + u^2_4 + u^2_5 = 1$

3. En outre, ces cinq coefficients doivent être tels que la variance de soit maximale Z_1

Analyse en Composantes Principales (ACP)

- Si on souhaite prendre une plus grande part de l'information contenue dans le tableau, il faudra définir une seconde composante principale Z_2 qui doit respecter les conditions ci-après:

4. Z_2 = combinaison linéaire des variables centrées et réduites, soit:

$$Z_2 = u'_1 x_1 + u'_2 x_2 + u'_3 x_3 + u'_4 x_4 + u'_5 x_5$$

Analyse en Composantes Principales (ACP)

5. Les coefficients $u'_{1,}$ $u'_2 + u'_3$, $u'_{4,}$ $u'_{5,}$ doivent être tel que Z_2 soit normé et indépendante ie dire $u'^2_1 + u'^2_2 + u'^2_3 + u'^2_4 + u'^2_5 + = 1$ et $u_1u'_1 + u_2u'_2 + u_3u'_3 + u_4u'_4 + u_5u'_5 + = 0$

6. De plus, les 5 coefficients u'_i doivent être tels que la variance de Z_2 soit maximale

NB: Le nombre maximal de composantes principales Z pouvant être définies est égal au nombre de variables contenues dans le tableau initial.

FIGURE : Quel animal ?

- Importation et affichage des données

```
data_acp<-read.table(data_acp .txt, header=TRUE, dec=“ ,”,
  row.names=1)

attach(data_acp)

data_acp
```

```
> data_pca
             TC    EL    TB    TRE    AV
BOUAKE    13.33 1.35 1.83 81.21 3.34
NERICA2   23.67 1.26 1.19 77.24 3.08
WITA      23.00 1.31 1.67 68.45 3.53
GAMBIAKA  14.00 1.33 0.83 79.81 3.07
WAB       18.33 1.25 0.54 82.74 2.86
NERICA1   26.33 1.26 1.31 76.83 3.15
>
```

Packages nécessaires:

Package principal: FactoMineR

Packages dont il dépend: ellipse lattice cluster scatterplot3d

Installation des packages:

```
install.packages(c("FactoMineR","ellipse","lattice","cluster","scatterplot3d"))
```

ou

par menu: package-installer le(s) package(s) depuis des fichiers zip

Analyse en Composantes Principales (ACP)

library(FactoMineR)

resultat_acp<-PCA(''data_acp'', scale.unit = TRUE, ncp = 2, graph = TRUE, axes = c(1,2)) (Réalisation de l'ACP)

data_acp = tableau de données avec la colonne des identifiants.

scale.unit =TRUE si on veut centrer et réduire les données (FALSE si c'est le contraire)

ncp=nombre de composantes retenues

axes = c(1,2) permet d'afficher les graphiques des composantes 1 et 2.

Si on souhaite afficher par ex les composantes 1 et 3 on écrira axes = c(1,3)

Analyse en Composantes Principales (ACP)

Affichage des principaux résultats

resultat_acp

resultat_acp$ eig

resultat_acp$ var$ (resultat_acp$ var$cor)

resultat_acp$ ind $ (resultat_acp$ ind $cos2)

- resultat_acp = les noms des principaux tableaux

- resultat_acp$ **eig** = une matrice des valeurs propres, le pourcentage de la variance et le cumul du pourcentage de variance.

- resultat_acp$ **var** = une liste de matrices contenant les résultats des **variables** (coordonnées, corrélation entre variables et axes, cosinus carré et les contributions)

- resultat_acp$ **ind** = une liste de matrice contenant les résultats des **individus** (coordonnées, corrélation entre variables et axes, cosinus carré et les contributions)

Analyse en Composantes Principales (ACP)

Résultats de l'ACP (Valeur propres et inertie totale)

- Le Tableau1 présente les valeurs propres et proportion d'informations concentrées sur les axes avec le logiciel R

resultat_acp$eig

- Tableau 1

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	2.752079902	55.04159805	55.04160
comp 2	1.892928303	37.85856606	92.90016
comp 3	0.320597880	6.41195759	99.31212
comp 4	0.029976647	0.59953295	99.91165
comp 5	0.004417268	0.08834535	100.00000

Analyse en Composantes Principales (ACP)

Résultats de l'ACP (Valeurs propres et inertie totale)

- Les valeurs propres sont en fait les variances des valeurs des composantes principales

- L'efficacité de stockage d'une composante principale est mesurée par la proportion de sa valeur propre par rapport à la somme de de toutes les valeurs propres

Analyse en Composantes Principales (ACP)

Résultats de l'ACP (Valeur propres et inertie totale)

- On peut remarquer que la première composante explique 55,04% des informations de départ et qu'avec trois axes, on arrive à expliquer 99,31% des informations contenues dans les variables initiales, ce qui est suffisant pour garantir une précision d'interprétation du tableau

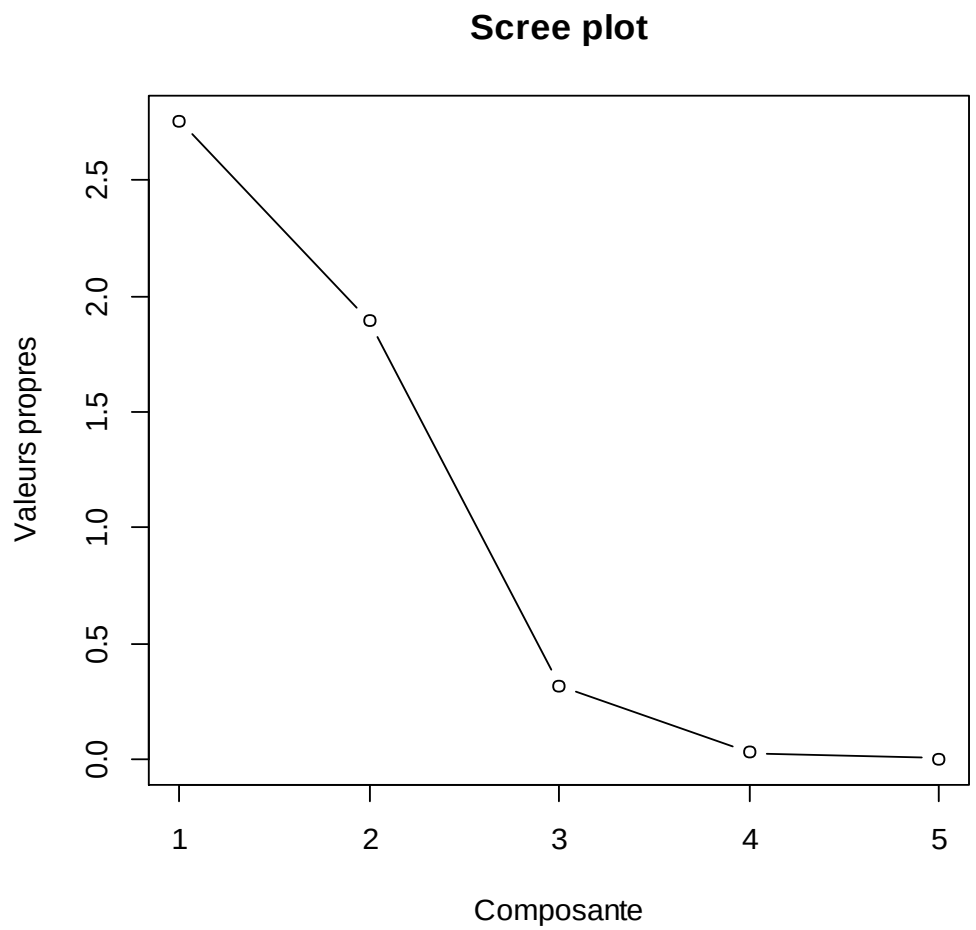
Analyse en Composantes Principales (ACP)

Choix du nombre de composantes

- De façon pratique, le nombre de composantes à retenir pour mieux résumer l'information de départ peut se faire aussi par l'observation d'une courbe expriment les valeurs propres en fonction du nombre de composantes (figure1)

Analyse en Composantes Principales (ACP)

plot(1:5, val.propres,type="b", ylab="Valeurs propres", xlab="Composante",main="Scree plot")



Analyse en Composantes Principales (ACP)

- **Résultats de l'ACP (scree plot)**
- Pour l'exemple, la courbe indique qu'il faut 3 composantes
- Mais il faut dire qu'en règle générale sauf exceptionnel, le nombre de composantes à retenir ne dépasse pas 3 pour faciliter l'interprétation

Analyse en Composantes Principales (ACP)

- **Résultats de l'ACP (Corrélation entre composantes et variables)**

- Les trois premières composantes étant retenues, il faut pouvoir connaître l'information qu'elles retiennent.

- Pour ce faire on va examiner les corrélations de ces composantes avec les 5 variables initiales (TC, EL, TB, TRE, AV) Pour l'exemple les résultats obtenus avec le logiciel R sont présentés au tableau 2.

Analyse en Composantes Principales (ACP)

- **Résultats de l'ACP (Corrélation entre composantes et variables)**
- **resultat_acp$ var$cor**

	Dim.1	Dim.2	Dim.3
TC	0.1125587	-0.97998666	0.12495135
EL	0.5892275	0.78814735	-0.12994613
TB	0.9167783	0.05885523	0.39372730
TRE	-0.7473906	0.55486013	0.36382517
AV	0.9965672	0.00667090	-0.02662911

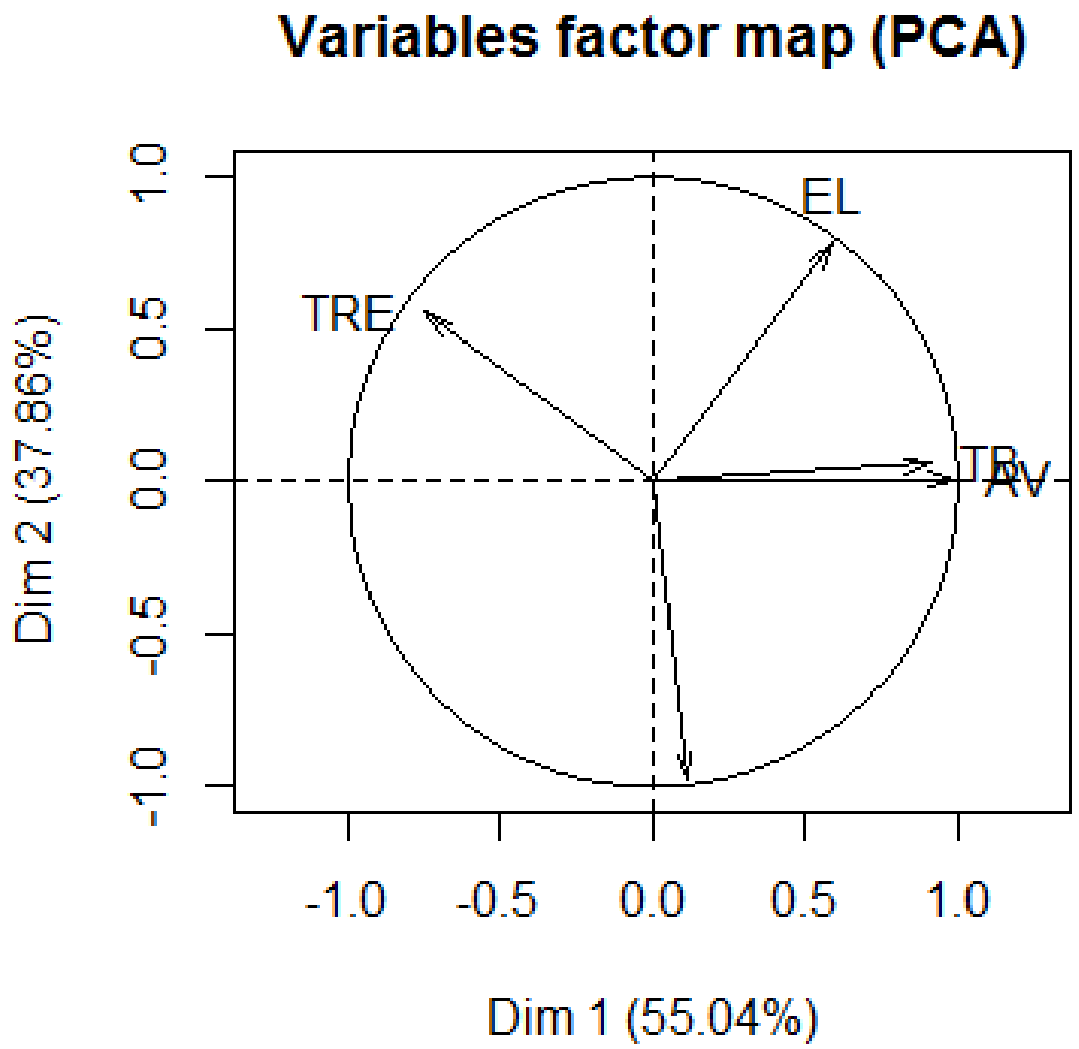
Analyse en Composantes Principales (ACP)

Résultats de l'ACP (Corrélation entre composantes et variables)

- TB, TRE, AV sont très bien représentés sur l'axe1 avec des corrélations de 0,92; -0,75; 0,99 alors que les variables TC et EL avec l'axe 2 avec des corrélations de -0, 98 et 0,79.
- Dans ce cas donc nous pouvons retenir que les deux premiers axes suffisent pour résumer l'information de départ car elle tiennent compte de toutes les variables

Analyse en Composantes Principales

Cercle de corrélation



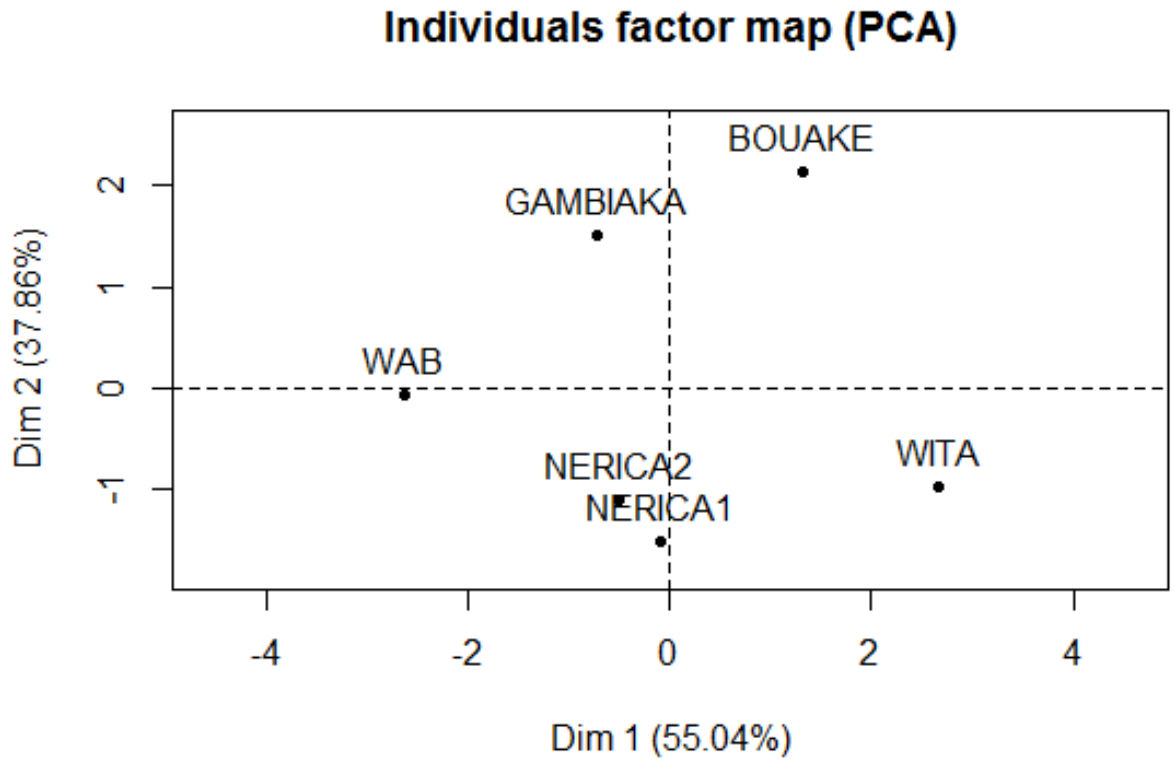
Analyse en Composantes Principales

- **Résultats de l'ACP (Corrélation entre composantes et variables)**
- On lit donc que: TB et AV sont corréles positivement avec l'axe 1 et TRE lui est négativement corrélé, donc l'axe 1 ou composante 1 est l'axe des valeurs élevées en taux de brisure, en augmentation de volume mais l'axe des valeurs faibles en taux de riz entier . En d'autre termes, les variétés situées du côté positif de cet axe sont celles qui ont des taux de brisure et une augmentation de volume élevés. Les variétés situées du côté négatif sont celles ayant des taux de riz entier élevés.

Analyse en Composantes Principales

- EL est corrélée positivement avec l'axe 2 et TC négativement avec cet axe. L'axe 2 est donc l'axe des teneurs élevées en EL et faibles en TC.
- De la même manière donc, en d'autre termes, les variétés situées du côté positif de cet axe sont celles qui ont une bonne élongation et de faible temps de cuisson. Les variétés situées du côté négatif par contre ont une élongation faible mais un temps de cuisson élevé.
- Toutes les variables étant pris en compte il serait superflu d'interpréter les variables sur le troisième axe.

Analyse en Composantes Principales
Résultat final



Des informations complémentaires peuvent se trouver sur la video youtube sur
https://www.youtube.com/watch?v=8qw0bNfK4H0