|  |  |
| --- | --- |
| M:\ADMIN\henallux_montgolfiere.png  Implantation IESN | Business intelligence  IG3 — C. Charlier |

# ANALYSE EN COMPOSANTES PRINCIPALES (ACP).

(ref : WikiStat : « Introduction à l’analyse en composantes principales »//  « Analyse en composantes principales »  
« Exploration des données » D.T . Larose)

<http://www.showme.com/sh/?h=P632jqK>

<http://www.showme.com/sh/?h=xGMNCtM>

## Introduction

Comment généraliser le nuage de points tracé dans le cas de deux variables et aborder la structure de corrélation présente entre plus de deux variables ? En appliquant l’analyse en composantes principales.

Mathématiquement, l’analyse en composantes principales est un simple changement de base : passer d’une représentation dans la base canonique des variables initiales à une représentation dans la base des facteurs définis par les vecteurs propres de la matrice des corrélations.

Prenons l’ensemble de données ci-dessous qui indique les cotes (entre 0 et 20) obtenues par 9 étudiants dans 4 disciplines.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Math | Physique | Franc | Anglais |
| Jean | 6 | 6 | 5 | 5,5 |
| Alain | 8 | 8 | 8 | 8 |
| Annie | 6 | 7 | 11 | 9,5 |
| Monique | 14,5 | 14,5 | 15,5 | 15 |
| Didier | 14 | 14 | 12 | 12,5 |
| André | 11 | 10 | 5,5 | 7 |
| Pierre | 5,5 | 7 | 14 | 11,5 |
| Brice | 13 | 12,5 | 8,5 | 9,5 |
| Evelyne | 9 | 9,5 | 12,5 | 12 |

## Représentation matricielle.

Soient p (=4) variables statistiques réelles observées sur n (=9) individus.   
Ainsi, sur le ième individu. Ces observations sont regroupées dans une matrice X d’ordre n.

Dans l’exemple :

* A chaque individu i est associé le vecteur contenant la ième ligne de X. C’est un élément d’un espace vectoriel noté E de dimension p. E est appelé espace des individus.
* A chaque variable est associé le vecteur contenant la jème colonne centrée. C’est un élément d’un espace vectoriel F de dimension n. F est appelé l’espace des variables.
* Lorsque les variables sont centrées et représentées par des vecteurs de F :
  + la longueur d’un vecteur représente un écart-type
  + le cosinus de l’angle entre deux vecteurs représente une corrélation.

## Décomposition spectrale de la matrice des covariances.

Exercice 1 : Etude des 4 variables séparément.

Analysez séparément ces 4 variables : graphiquement et numériquement.

Exercice 2 : Etude des variables 2 à 2

Analysez les liaisons entre les variables 2 à 2.

Pour l’étude simultanée des 4 variables, les choses se compliquent car chaque individu est représenté dans un espace de dimension 4. Les objectifs de l’ACP sont

* de revenir à un espace de dimension réduite q < p(par exemple 2) en déformant le moins possible la réalité. Nous voulons donc obtenir le résumé le plus pertinent des données initiales;
* de représenter les individus de manière « optimale » en minimisant les déformations du nuage des points, dans un sous-espace Eq de dimension q < p ;
* de représenter graphiquement les variables dans un sous-espace Fq en explicitant « au mieux » les liaisons initiales entre ces variables ;
* de faciliter l’utilisation de techniques comme la régression linéaire en fournissant des variables orthogonales.

Remarquons que, dans notre exemple, les données sont toutes du même ordre de grandeur. Si les variables ne sont pas homogènes, elles doivent être préalablement réduites.

Etape 1 :

Réaliser une standardisation des données.

Etape 2 :

Calculer la matrice symétrique des variances-covariances notée où mesure le degré avec lequel les deux variables varient ensemble. Une covariance positive indique que quand une variable s’accroît, l’autre tend à s’accroitre. Une covariance négative indique que quand une des variables s’accroit, l’autre a tendance à décroitre

Les composantes principales représentent un nouveau système de coordonnées, trouvé en faisant tourner le système original le long des directions de variabilité maximum. Remarque : si sont indépendantes alors leur covariance est nulle.

NB : la covariance n’est pas redimensionnée donc changer les unités de mesure change la covariance, ce qui n’est pas le cas avec le coefficient de corrélation linéaire : .

Etape 3 :

Matrice des corrélations : .

Lorsque les données sont standardisées, la matrice des variances-covariances et la matrice des corrélations sont les mêmes.

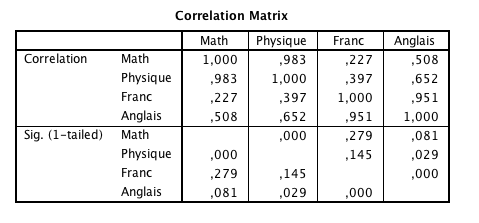
Etape 4 :

La i-ième composante principale de la matrice des données standardisées Z=[Z1,Z2,…,Zn] est donnée par   
 où est le i-ième vecteur propre de la matrice des corrélations.   
Les composantes principales sont les combinaisons linéaires des variables standardisées de Z telles que les variances des soient les plus élevées possible et que les ne soient pas corrélées.

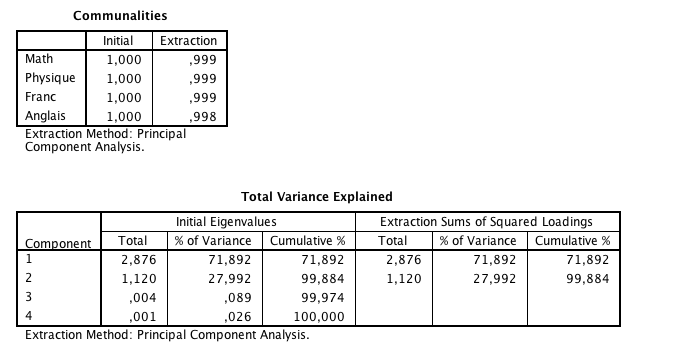
Revenons à notre exemple :

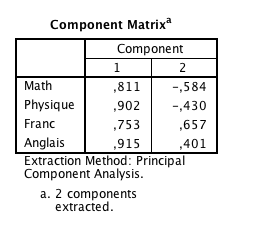
En réalisant une ACP sur la matrice des cotes, nous obtenons les résultats suivants :

Matrice des corrélations :



Cette matrice représente les coefficients de corrélation linéaire entre les différentes variables. Bien sûr, elle est composée de 1 sur sa diagonale (en effet, une variable est toujours 100% corrélée à elle-même) et elle est symétrique (en effet, le coefficient de corrélation linéaire entre la variable j et la variable i est le même que le coefficient de corrélation linéaire entre la variable i et la variable j). La valeur 0,983 pour les variables math-physique indique une forte corrélation (linéaire) positive entre les cotes obtenues en math et les cotes obtenues en physique par les différents individus. La valeur 0,227 pour les variables math-français indique une très faible corrélation positive entre les cotes obtenues en math et les cotes obtenues en français par les différents individus.



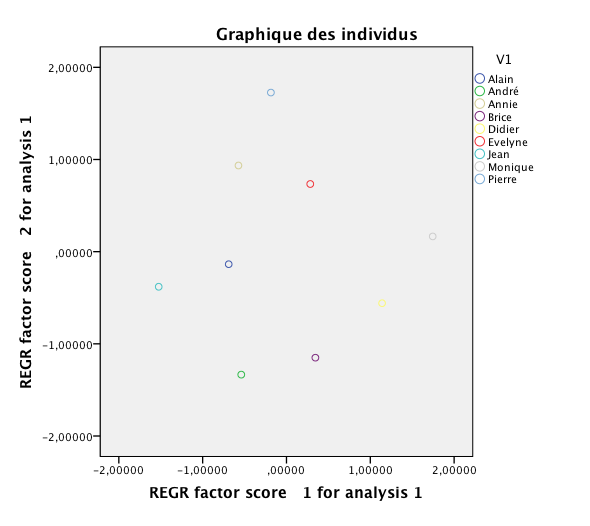


Le tableau ci-dessus montre que deux facteurs (deux vecteurs propres) ont été sélectionnés pour représenter les quatre variables simultanément. Ces facteurs sont des combinaisons linéaires des variables de départ.   
Axe 1 :  
Le premier est calculé (sur les données standardisées) en prenant 0,811\*cote de math (réduite) + 0,902\*cote de physique (réduite)+… . Ainsi, l’axe 1 représente une sorte de « note globale pondérée de l’étudiant ». Plus l’individu (cfr graphique des individus) est représenté « loin » vers l’axe 1 positif, plus sa note globale est bonne (exemple : Monique). Si l’individu est représenté vers le 0 de l’axe 1, il a une note globale plutôt moyenne (exemple : Evelyne). Plus l’individu est représenté « loin » dans l’axe 1 négatif, plus cela signifie que sa note globale est faible (exemple : Jean).

Axe 2 :

Le deuxième est calculé (sur les données standardisées) en prenant - 0,584 \* cote de math (réduite) - 0,43 \* cote de physique (réduite)+ 0,657 \* cote de français (réduite)+ 0,401 \* cote d’anglais (réduite). Ainsi, l’axe 2 représente une sorte « d’écart entre les matières littéraires et les matières scientifiques ». Plus l’individu (cfr graphique des individus) est représenté « loin » vers l’axe 2 positif, plus l’écart entre ses notes dans les branches littéraires et les branches scientifiques est grand et ceci, au profit des branches littéraires (exemple : Pierre). Si l’individu est représenté vers le 0 de l’axe 2, ses notes sont alors assez équivalentes dans les branches scientifiques et dans les branches littéraires (exemple : Brice). Plus l’individu est représenté « loin » dans l’axe 2 négatif, plus cela signifie que l’écart entre ses notes dans les branches littéraires et les branches scientifiques est grand et ceci, au profit des branches scientifiques (exemple : Pierre).

En interprétant selon les deux axes, nous pouvons dire que Annie a des notes faibles, inférieures à la moyenne mais qu‘elle a de meilleures notes dans les branches littéraires que dans les branches scientifiques.

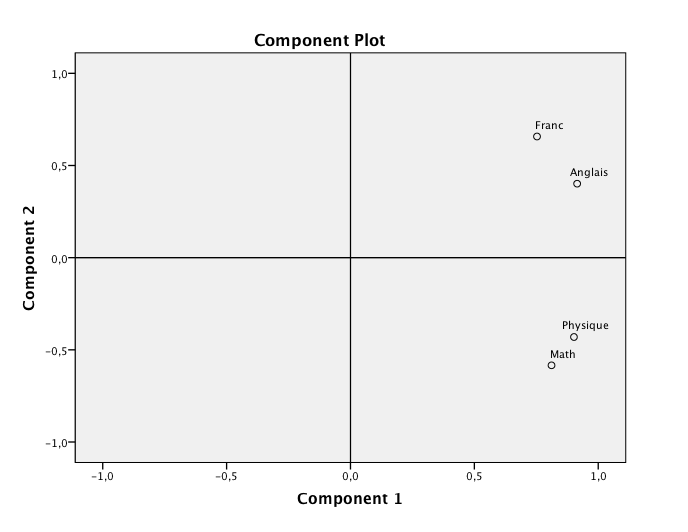


Que peut-on dire des facteurs choisis ?

Le tableau « communalities » signale que 99,9% de l’information contenue dans les notes de math est prise en compte en prenant ces deux facteurs. Idem pour les autres variables. Remarquons que ces valeurs sont souvent plus faibles quand même et qu’elles sont souvent différentes d’une variable à une autre, ce qui n’apparaît pas dans cet exemple. Le graphique des variables (components plot) indique cette information dans la longueur des vecteurs représentés. Il faut imaginer un cercle de centre « croisement des axes » et de rayon 1. Plus la longueur du vecteur représenté s’approche de 1, mieux il est représenté dans la nouvelle configuration. Puisqu’ici, 99,9% de l’information de chaque variable a été extraite, les quatre vecteurs (variables de départ) sont de même longueur et de longueur proche de 1. Si l’information contenue dans l’une de ces variables avait été extraite à 50%, la variable correspondante aurait été représentée sur une longueur 0 ,5.

Le tableau « total variance explained » présente les valeurs propres. En faisant le rapport entre la valeur propre du facteur i et la somme des valeurs propres, nous obtenons le pourcentage de variance expliquée par le facteur i. Exemple première valeur propre 2,876 /(2,876+1,12+0,004+0,001)= 0,7189 donc 71,892% de la variance est expliquée par le facteur 1 (cfr deuxième colonne). Nous constatons aussi, à la lecture de ce tableau (% cumulative variance) que 99,884% de la variance est expliquée par les deux facteurs.

Le graphique ci-dessous permet aussi de connaître la corrélation entre les différentes variables. Celle-ci est représentée par le cosinus de l’angle entre les deux vecteurs. Plus l’angle est petit (vers 0), plus le cosinus de l’angle tend vers 1 donc plus la corrélation est forte et positive (exemple : entre math et physique). Plus l’angle tend vers 90°, plus le cosinus de l’angle tend vers 0 donc plus la corrélation est faible (exemple : entre math et français). Si l’angle se situe entre 90° et 180°, il indique une corrélation négative allant respectivement de très faible (nulle à 90°) à très forte.



Il est possible d’observer la représentation d’une variable sur un axe. Par exemple, un vecteur qui a un angle faible avec l’axe 1 a une bonne représentation sur l’axe 1 (cfr sa projection).