Analyse composante principale

# Introduction

(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>

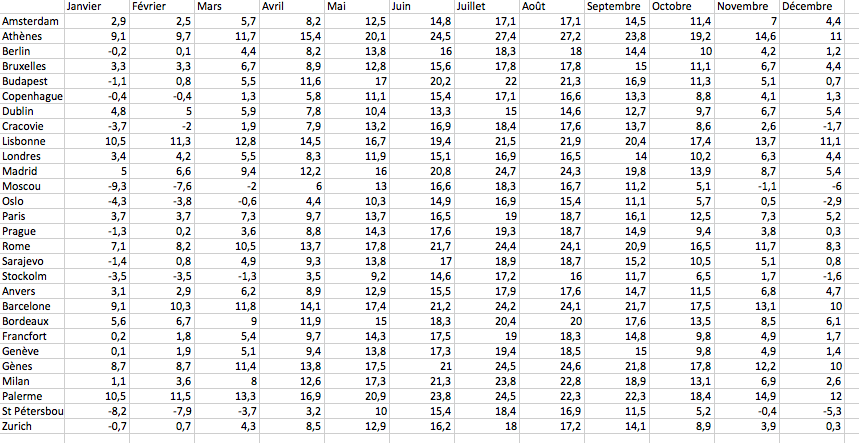
<http://www.showme.com/sh/?h=xGMNCtM>

Lorsqu’on étudie simultanément un nombre important de variables quantitatives (ne serait-ce que 4 !), comment en faire un graphique global ? La difficult ́e vient de ce que les individus ́etudi ́es ne sont plus représentés dans un plan, espace de dimension 2, mais dans un espace de dimension plus importante (par exemple 4). L’objectif de l’Analyse en Composantes Principales est de revenir à un espace de dimension réduite (par exemple 2) en d ́eformant le moins possible la réalité. Il s’agit donc d’obtenir le résumé le plus pertinent possible des données initiales.

C’est la matrice des variances-covariances (ou celle des corrélations) qui va permettre de réaliser ce résumé pertinent, parce qu’on analyse essentiellement la dispersion des donn ́ees considérées. De cette matrice, on va extraire, par un procédé mathématique approprié, les facteurs que l’on recherche, en petit nombre. Ils vont permettre de réaliser les graphiques d ́esir ́es dans cet espace de petite dimension (le nombre de facteurs retenus), en d ́eformant le moins possible la configuration globale des individus telle qu’elle est d ́efinie par l’ensemble des variables initiales (ainsi remplacées par les facteurs).

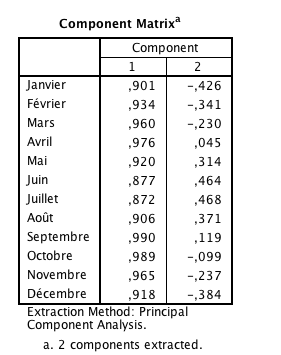
C’est l’interprétation de ces graphiques qui permettra de comprendre la structure des données analysées. Cette interprétation sera guidée par un certain nombre d’indicateurs numériques, appelés aides à l’interprétation, qui sont là pour aider l’utilisateur à faire l’interprétation la plus juste et la plus objective possible.

De ce fait pour utiliser certains algorithme tel que la régression il est nécessaire d’utiliser des variables qui sont plutôt indépendante. Et l’ACP offre ceci. Attention indépendante entres-elles. Et de ce fait l’objectif est également de réduire ce nombre de variables.

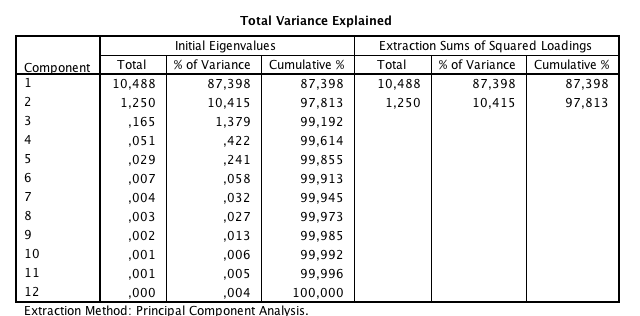


|  |  |
| --- | --- |
| **QUESTION (Analyse en composantes principales)** | / |

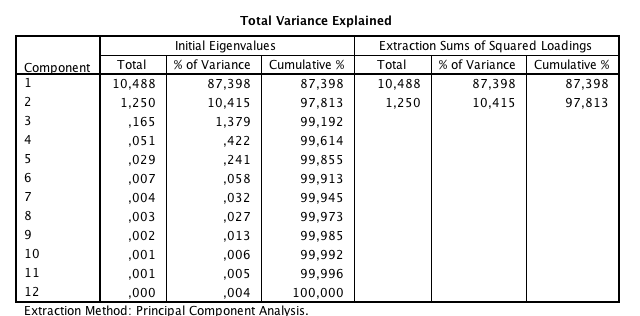
1. **A quoi sert une analyse en composantes principales ?**  
   *Lorsqu’on étudie simultanément un nombre important de variables quantitatives (ne serait-ce que 4 !), La difficulté vient de ce que les individus étudies ne sont plus représentés dans un plan, espace de dimension 2, mais dans un espace de dimension plus importante (par exemple 4). L’objectif de l’Analyse en Composantes Principales est de revenir à un espace de dimension réduite (par exemple 2) en d ́eformant le moins possible la réalité. Il s’agit donc d’obtenir le résumé le plus pertinent possible des données initiales.*

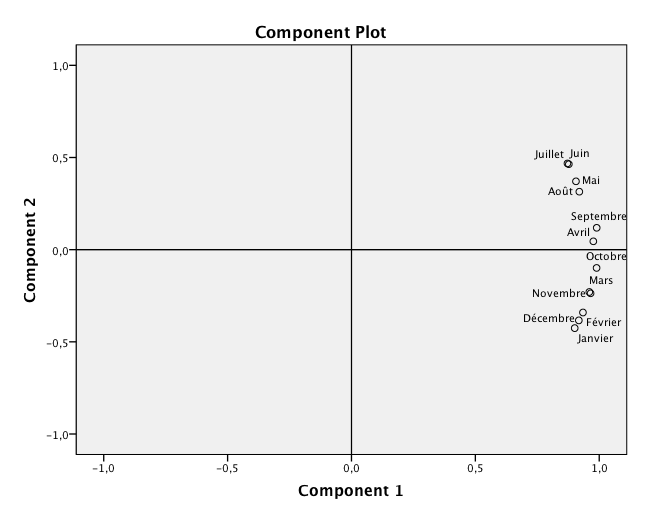
1. **On s ‘intéresse au climat des différents pays d’Europe. Pour cela, on a recueilli les températures moyennes mensuelles (en degré Celsius) pour les principales capitales européennes ainsi que pour certaines grandes villes. Ces données ainsi que les résultats de l’acp sont fournis dans les tableaux et graphiques ci-dessous.**
   1. Donnez l’explication contextuelle (interprétation) des deux nouvelles composantes.   
        
      L’axe 1 représente la température moyenne à l’année pour  
      chaque pays d’europe. Température moyenne sur une année pour un pays d’Europe.

L’axe 2 représente la différence qu’il y a entre la période Octobre-Mars et période Avril-Septembre.

* 1. Quelle part de l’information de départ prennent-elles en compte ?

Il explique 87,813% de la variance que nous voulons expliquer. Cela veut dire qu’avec 2 composantes on arrive dans l’ensemble de la variabilité on en reprends 87% rien qu’avec 2 composantes.





* 1. Enoncez trois indications (idées différentes) données par le vecteur octobre.  
       
     - Due a l’angle formé entre octobre et juin, on peut indiquer que ces deux mois sont très

peu corrélé entre eux. Par contre l’angle entre mars et octobre est très petit donc la corrélation est grande

* Enfin, le mois d’Octobre est fortement représenté par l’axe 1
* Si on fait le rapport entre le cercle que l’on trace et le mois d’octobre on peut remarqué qu’on a un fort pourcentage. On a repris un gros pourcentage d’information du mois d’octobre que les deux composantes contiennent.
  1. Vrai ou faux. Si faux, corrigez la phrase. Si vrai, justifiez.
     1. Pour chacune des villes, la température du mois d’avril est fortement corrélée à la température moyenne de l’année.  
        Vrai. Ce dernier se trouve très proche de l’axe 1.
     2. La variable juin a joué le rôle le plus important dans la composante 2.  
          
        Non, le mois de juillet a joué le même role.
     3. La proximité de deux villes sur la composante 1 implique systématiquement qu’elles ont deux profils de température moyenne semblables au cours de l’année.  
          
        Faux, même si ils ont la même moyenne, ils n’auront peut être pas le même profile sur l’axe 2. Ca veut dire que les distances de températures entres les mois-froids et les mois-chauds seront peut être fortement différentes sur ces deux villes.
     4. S’il fait froid en janvier, il fera froid en juin.  
          
        Faux. C’est plutôt le contraire car il y a un angle entre les 2 très grand. Probablement > 90 degré de ce fait il n’y a aucune corrélation entre les deux.
  2. Ci-dessous les axes pour le graphique des individus. Interprétez la position des deux points.

# La position sur axe 1 d’une ville dans le graphique des individus représentera la température moyenne sur l’année par rapport à la température global de tous les individus. En gros, si il se situe sur 0 en axe1 (Abscisse) cela veut dire qu’il se situe dans la moyenne concernant sa température sur un an face aux autres villes.

# Et sa position sur axe 2 représentera le fait qu’il ai ou non une grande différence concernant les mois froids et les mois chaud en terme de température. Un rapprochement vers 0 sur l’axe2 = hiver froid et été très chaud