

Analyse des données

Travaux pratiques

Séance 3

Application de l'ACP sur un fichier de températures

Objectifs:

Nous souhaitons savoir s'il y a des corrélations entre les profils des températures des différentes capitales en Europe. Pour cela nous allons travailler sur les représentations des individus (les capitales européennes) et des variables (les mois de l'année en termes de températures) fournis par l'ACP. Le tableau fournit un ensemble de valeurs thermiques qui indique selon ce que l'on recherche soit le profil des températures pour une capitale pour l'ensemble des mois de l'année soit le profil des températures pour un mois pour l'ensemble des capitales. Parmi cet ensemble de données, nous allons nous poser la question de savoir s'il existe des corrélations ou non et selon les cas, quels sens leur donner. Bien évidemment, nous nous appuyerons sur notre connaissance géographique et climatique des capitales et villes du continent européen listés dans le tableau. Nous allons jouer le rôle de l'"expert" car seul ce dernier est en mesure/capable de trouver un sens/ une interprétation au modèle obtenu suite à l'ACP.

Travail demandé:

1. Donner la matrice des corrélations entre les variables températures de chaque capitale. Dites pour vous s'il est possible d'en extraire des conclusions.
2. Utiliser l'ACP afin de pouvoir extraire des informations pour obtenir des éléments d'interprétation concernant d'éventuels liens de cause à effet sous-jacent entre telles ou telles observations. Pour cela vous pourrez utiliser le script proposé pour vous aider. Il est censé donner tous les éléments pour vous mener sur le bon chemin c'est-à-dire avoir l'ensemble des grandeurs (valeurs propres, taux d'inertie/axe, les graphes des individus et variables, ... etc.) possibles fournies une fois l'ACP réalisée, permettant ainsi une analyse comparative des profils de températures en Europe!

Afin de vous guider dans votre réflexion, il vous est demandé de répondre aux questions suivantes:

1. Peut-on se borner à travailler sur le premier plan factoriel. En d'autres termes, le plan factoriel porté par les deux premières valeurs propres est-il représentatif de l'inertie totale du nuage de points initial (argumentez)?
2. Quelle signification donnez-vous à l'axe 1. Dit autrement, que représente la variable synthétique portée par cet axe?
3. Même interrogation concernant l'axe 2.
4. Que représente le vecteur mois?
5. Quel est le critère pour que les variables quantitatives actives et supplémentaires apparaissent comme bien représentées dans le graphe donné par l'ACP? Indiquer le(s) vecteur(s) mal représenté(s).
6. Comment interprétez-vous le fait que tous les vecteurs mois soient orientés dans la même direction?
7. Comment expliquez-vous le fait que le vecteur Latitude soit anticorrélé avec certains vecteurs mois et en particulier le vecteur moyenne?
8. Comment interprétez-vous le fait que le vecteur Latitude soit orthogonal au vecteur Amplitude?
9. Comment interprétez-vous le fait que le vecteur *Amplitude* soit en partie orthogonal au vecteur *moyenne*?

Analyse des données

Travaux pratiques

Séance 3

Application de l'ACP sur un fichier de températures

En répondant aux questions précédentes, vous êtes en possession maintenant d'informations qui vont permettre de trouver un sens au graphe des individus. Sur celui-ci vous devez avoir la présence à la fois les capitales et quelques villes (variables supplémentaires) et Vous aurez pris soin de rajouter sur ce graphe les variables qualitatives données.

Comment justifiez-vous maintenant la disposition des capitales

Pour vous aider, voici quelques notes en vrac pour faciliter la compréhension de cet exercice

Description des données et problématique

On va s'intéresser au climat des différents pays d'Europe. On dispose pour cela des températures mensuelles moyennes sur l'année. Elles concernent principalement les capitales mais aussi quelques grandes villes.

En complément, se rajoute comme information supplémentaire

- La température moyenne annuelle
- L'amplitude thermique, (écart entre la température mensuelle max et min).
- Deux variables quantitatives de positionnement (la longitude et la latitude)
- Une variable quantitative d'appartenance à une des quatre régions d'Europe (quatre modalités : est, ouest, nord et sud)

Quelques éléments pour faciliter l'analyse - choix des éléments actifs

Étude des villes

L'objectif est d'appréhender la variabilité des températures mensuelles d'un pays à l'autre de façon multidimensionnelle, c'est-à-dire en prenant en compte simultanément les 12 mois de l'année.

Un pays est représenté par le climat de sa capitale. Elle sera vue comme un candidat actif. Les autres villes du pays s'il y en a, ne seront pas prise en compte afin de ne pas donner un poids élevé à ce pays. Ces villes seront vues comme des individus supplémentaires. Ils ne participeront pas à la construction des axes.

Du point de vue multidimensionnelle, deux villes sont d'autant plus proches qu'elles présentent le même ensemble de températures mensuelles.

On va mettre en évidence les principaux facteurs de variabilité des capitales ce qui nous permettra de construire une typologie sur les pays.

Bien évidemment, on souhaite obtenir **une vision d'ensemble de ces liaisons** sans passer en revue **chaque couple de variables** (on fait allusion ici à la matrice des corrélations)

Analyse des données

Travaux pratiques

Séance 3

Application de l'ACP sur un fichier de températures

Étude des variables (étude sur le graphe des variables)

Il est clair que chaque variable donne les températures mensuelles des 23 capitales.

L'objectif essentiel du travail est d'étudier les liaisons entre celles-ci.

Rappel, on prend les individus actifs seulement !

Deux variables (ici représentées par des vecteurs) sont corrélées positivement (même direction) si globalement les villes les plus chaudes selon l'une (vecteur août par exemple) sont les plus chaudes selon l'autre (vecteur janvier par exemple). Dit autrement, fait-il chaud en août là où il fait chaud en janvier !

Cette vision d'ensemble va se faire par l'intermédiaire des variables synthétiques (**axes factoriels**). La question est donc : peut-on résumer les températures mensuelles par un petit nombre de composants.

Si cela est possible, nous examinerons les liaisons entre les variables initiales et les variables synthétiques.

Il va sans dire que ce travail est plus commode que l'examen direct à partir de la matrice des corrélations. Ici, 12 variables initiales et 2 variables synthétiques (le plan factoriel) nous donnent 24 liaisons à étudier plutôt que $12 \times 11/2 = 66$ liaisons ! (1/2 matrice des corrélations).

On s'intéresse aux profils de températures des capitales basés sur les 12 variables actives.

Les variables supplémentaires comme température moyenne annuelle et amplitude annuelle seront confrontées aux composantes principales mais ne sont pas intégrées au profil (ensemble des températures annuelles pour une capitale).

Quelles observations peut-on faire une fois les deux graphes obtenus?

Le premier facteur est prépondérant. Il exprime à lui seul 82,9% de l'inertie totale!

Le second facteur est relativement important puisqu'il exprime 15,4% de l'inertie totale.

Ces deux facteurs expriment 98,3% de l'inertie totale ce qui justifie de s'y limiter.

Cela veut dire qu'à partir de ces 2 variables « synthétiques » (portées par les axes) dits principaux, on résume presque toute l'information contenue dans les douze variables initiales !

C'est un cas d'école car le résumé apporté par l'ACP est presque parfait, on dit parfait!

Une conséquence de cela : dans le premier plan factoriel, les variables et les individus sont très bien projetés ce qui veut dire que :

- 1- La proximité de deux individus (capitales) dans ce plan factoriel traduit une proximité dans l'espace à 12 dimensions.
- 2- L'angle entre deux variables (mois) dans le plan donne une très bonne approximation de l'angle dans l'espace réel à 12 dimensions.

On observe que toutes les variables actives ont une coordonnée (projection sur l'axe 1) de même signe. Les vecteurs mois sont presque colinéaires. On dit que l'on est en présence d'un effet taille.