TP : Analyse en composantes principales

Moyennes mensuelles des températures pour 15 villes françaises

1 Description des données

On dispose, pour 15 villes de France, des moyennes mensuelles de températures calculées sur 30 ans (de 1931 à 1960). Ces données sont rassemblées dans la Table ci-jointe qui croise les 15 villes (lignes) et les 12 mois de l'année (colonnes). Dans ce tableau, les 2 dernières colonnes représentent la latitude et la longitude de chaque ville.

Toutes ces données sont regroupées dans le fichier **temper.npz** sur le site web du cours. Le fichier contient 3 variables :

data Matrice de taille 15*14 contenant les températures moyennes des 15 villes françaises sur 12 mois ainsi que la latitude et longitude de ces villes (2 dernière colonnes).

varname Liste contenant les noms des variables associées à chaque colonne de **data**.

villes Liste contenant les noms des villes associées à chaque ligne de **data**.

Le but de l'étude est de comparer les températures mensuelles des différentes villes. L'analyse en composantes principales (ACP) sur ce tableau devra préciser les points suivants. Il s'agit en autre de

- réaliser une typologie des villes, c'est-à-dire celles qui se ressemblent et celles qui diffèrent du point de vue des températures mensuelles ;
- proposer un bilan des liaisons entre les variables et, dans la mesure du possible, résumer approximativement l'ensemble des variables par un petit nombre de variables synthétiques, c'est-à-dire non pas extraites du tableau mais les combinant ;
- étudier si les ressemblances ou les dissemblances correspondent à des proximités ou des éloignements géographiques.

2 Question de cours

1. Comment mesure-t-on la dissemblance entre les individus ?
2. Le nuage est toujours centré. Pourquoi ? Quel est l'effet du centrage sur l'analyse du nuage des individus ? Quel type de liaisons évalue-t-on dans le nuage des variables ?
3. Montrer que si on ne réduit pas les variables, on accorde aux variables un poids égal à leur écart-type. La liste des écart-types permet-elle, a priori, de prévoir une différence significative entre une ACP normée et une ACP non-normée ?

3 Analyse des données

Lors du TP vous aurez besoin des bibliothèque Python `numpy` `pylab` et `scipy`. Il vous est conseillé de les importer dès le début avec le code suivant :

```
import numpy as np
import pylab as pl
import scipy as sp
```

Vous pourrez ensuite accéder aux fonctions de ces modules à l'aide du point (par exemple `np.zeros(10)` pour la fonction `zeros` de numpy). Dans la suite du TP, pour chaque question la liste des fonctions `numpy/pylab` nécessaires est donné entre parenthèses.

1. Charger les données en mémoire et les interpréter (`np.load`).
2. Visualiser les villes sur le plan 2d contenant latitude et longitude (`pl.plot,pl.text`).
3. Tracer l'évolution de température pour les différentes villes (`pl.plot,pl.legend`). Que remarquez vous ?
4. Centrer et réduire les données et tracer une fois que plus l'évolution des températures (`np.mean,np.std`). peut-on reconnaître des groupes de villes ?

4 Analyse en composantes principales

Tous les calculs suivant se feront sur des données centrée et réduites.

1. Analyser les statistiques descriptives de chacune des 16 variables avant normalisation (`np.mean,np.std`).
2. Calculer et interpréter la matrice des corrélations (`np.cov,pl.imshow`). Quelles sont les variables les plus corrélées ?
3. Analyser les valeurs propres (`np.linalg.eigh,pl.stem`). Quel est le pourcentage d'inertie expliquée par chaque axe ? Combien d'axes peut-on conserver ?
4. Tracer les vecteurs propres principaux et les interpréter (`pl.plot`).
5. Quelle sont les composantes principales ? Étudier la représentation des individus sur le premier axe et le second axe (`np.dot`). Quels sont les points dont la contribution est la plus grande ? Quelles sont leurs coordonnées ?
6. Visualiser les villes sur le plan 2d contenant les deux axes principaux.

5 Données de caractères manuscrits (bonus)

1. Charger les données et les visualiser (fichier **pcadigit.npz**).
2. Effectuer l'analyse en composante principale.
3. Visualiser leur projection dans un sous espace 2D.
4. Visualiser les directions principales sous la forme d'image.