TP Analyse par Composantes Principales

On s'intéresse au climat des différents pays d'Europe. Pour cela, on a recueilli les températures mensuelles (en degrés Celsius) pour les principales capitales européennes ainsi que pour certaines grandes villes. En plus des températures mensuelles, on donne, pour chaque ville, la température moyenne annuelle ainsi que l'amplitude thermique (différence entre la moyenne mensuelle maximum et la moyenne mensuelle minimum d'une ville). On donne également deux variables quantitatives de positionnement : la longitude et la latitude ainsi qu'une variable qualitative : l'appartenance à une région de l'Europe (variable qualitative à quatre modalités : Europe du nord, du sud, de l'est et de l'ouest). Un extrait des données est fourni dans le tableau 1.14.

	Janv	Fév	Mars	Avr	...	Nov	Déc	Moy	Amp	Lat	Lon	Rég
Amsterdam	2.9	2.5	5.7	8.2	...	7.0	4.4	9.9	14.6	52.2	4.5	Ouest
Athènes	9.1	9.7	11.7	15.4	...	14.6	11.0	17.8	18.3	37.6	23.5	Sud
Berlin	-0.2	0.1	4.4	8.2	...	4.2	1.2	9.1	18.5	52.3	13.2	Ouest
Bruxelles	3.3	3.3	6.7	8.9	...	6.7	4.4	10.3	14.4	50.5	4.2	Ouest
Budapest	-1.1	0.8	5.5	11.6	...	5.1	0.7	10.9	23.1	47.3	19.0	Est
Copenhague	-0.4	-0.4	1.3	5.8	...	4.1	1.3	7.8	17.5	55.4	12.3	Nord
Dublin	4.8	5.0	5.9	7.8	...	6.7	5.4	9.3	10.2	53.2	6.1	Nord
Helsinki	-5.8	-6.2	-2.7	3.1	...	0.1	-2.3	4.8	23.4	60.1	25.0	Nord
Kiev	-5.9	-5.0	-0.3	7.4	...	1.2	-3.6	7.1	25.3	50.3	30.3	Est
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

TAB. 1.14 – Données températures : extrait pour quelques capitales, les températures sont mesurées en degrés Celsius.

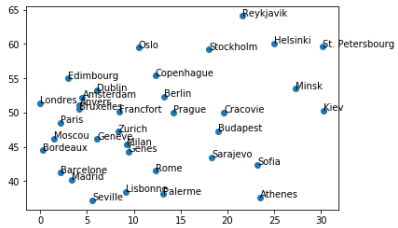
Étude des villes. On souhaite appréhender la variabilité des températures mensuelles d'un pays à l'autre de façon multidimensionnelle, *i.e.* en prenant en compte simultanément les 12 mois de l'année. Un pays sera représenté par le climat de sa capitale, les données des autres villes ne sont pas prises en compte pour éviter d'accorder plus de poids aux pays pour lesquels plusieurs villes sont renseignées. Ainsi, les capitales seront considérées comme des individus actifs tandis que les autres villes seront considérées comme des individus supplémentaires (*i.e.* qui n'interviennent pas dans la construction des axes). Du point de vue multidimensionnel, deux villes sont d'autant plus proches qu'elles présentent le même ensemble de températures mensuelles. Une façon synthétique d'aborder ces données est de mettre en évidence les principaux facteurs de variabilité des capitales. On pourra ainsi répondre à des questions du type : quelles sont les plus grandes disparités entre pays ? Ces facteurs pourront servir de base à la construction d'une typologie sur les pays.

Étude des variables. Chaque variable mesure les températures mensuelles dans

les 23 capitales. La liaison entre les variables n'est appréhendée qu'à partir des capitales (*i.e.* des individus actifs) et non de l'ensemble des villes. Ces liaisons constituent un objectif essentiel dans une telle étude. Deux variables sont corrélées positivement si, globalement, les villes les plus chaudes selon l'une sont les plus chaudes selon l'autre (par exemple, fait-il chaud en août là où il fait chaud en janvier ?). Naturellement, on souhaite obtenir une vision d'ensemble de ces liaisons, sans passer en revue chaque couple de variables.
Cette vision d'ensemble peut se faire par l'intermédiaire de variables synthétiques. La question est alors : peut-on résumer les précipitations mensuelles par un petit nombre de composantes ? Si oui, on examinera les liaisons entre les variables initiales et les variables synthétiques : cet examen indirect est plus commode que l'examen direct (avec 12 variables initiales et 2 variables synthétiques, on examinera 24 liaisons au lieu de $12 \times 11/2 = 66$).
On s'intéresse aux profils de températures des villes donc on prendra comme variables actives uniquement des variables concernant la température (ce qui élimine des variables comme la latitude, la longitude). Pour les autres variables proposées comme supplémentaires (température moyenne annuelle et amplitude annuelle), ce sont des indicateurs synthétiques qu'il sera intéressant de confronter aux composantes principales mais qui n'appartiennent pas non plus au profil proprement dit. En outre, ce sont des variables qui utilisent une information déjà présente dans les autres variables.

Dans ce TP, vous avez le droit d'utiliser toutes les fonctions numpy, matplotlib, et pandas que vous voulez, *mais pas la librairie scikit-learn*

1. Téléchargez le fichier csv de données à l'adresse suivante : http://lipn.univ-paris13.fr/~chevaleyre/pmwiki/files/temperat.csv
2. Créez un python notebook (avec jupyter lab ou google collab) pour importer ces données. Vous pouvez par exemple utiliser la fonction read_csv de pandas
3. Extrayez la matrice des données (par exemple avec la fonction to_numpy de pandas)
4. Affichez sur un graphique le nom des villes. Ces villes devront être positionnées aux coordonnées déterminées par leur latitude et longitude :



5. Eliminez du tableau de données tout ce qui n'est pas temperature mensuelle. Il devrait donc rester 12 colonnes.
6. Affichez la matrice de correlations des variables. Transposez la matrice de donnees et affichez les correlations a nouveau. Que pouvez-vous dire ?
7. Centrez et réduisez vos données (donc retranchez à chaque variable sa moyenne, et divisez la par l'écart type)
8. Re-Codez vous-même l'ACP en extrayant les vecteurs propres et valeurs propres de la matrice de corrélation. Pour cela, vous utiliserez la fonction numpy.linalg.eig, qui extrait les valeurs propres et vecteurs propres. L'ACP devra être calculée à partir des 23 capitales (les 23 premières lignes du tableau).
   Que peut-on dire à propos des valeurs propres observées ?
9. En utilisant les deux axes principaux, calculez les coordonnées de chaque ville en 2D.
10. Affichez les villes comme un nuage de points avec matplotlib (chaque ville doit etre un point, avec son nom dessus)
11. Maintenant, effectuez une ACP pour visualiser non pas les villes mais les mois sur un plan (donc l'ACP doit être lancé sur la matrice transposée des données).
12. Interpretez le résultat