Analyse de données
Travaux Pratiques

H. Le Capitaine

Polytech Nantes

Lors de ces travaux pratiques, l'objectif est de se familiariser avec d'une part l'environnement de travail R, d'autre part avec les outils mathématiques de l'ACP et de classification. Vous devez rédiger un compte-rendu de votre travail, à déposer sur la plateforme Madoc, en prenant bien soin de détailler l'ensemble des manipulations que vous avez pu faire, le code que vous avez produit et enfin les conclusions auxquelles vous êtes arrivés. Par défaut, vous travaillerez en binôme, et une exception à ce défaut devra être discutée avec les encadrants de TP.
Rappel : Une suite de commandes et leur résultat ne constitue pas un rapport, il faut justifier les choix, analyser les résultats obtenus.

Nous nous intéressons cette année à des statistiques socio-démographiques de pays européens Les données sont fournies dans le fichier `euro.csv`, que vous pouvez télécharger sur Madoc. Un descriptif plus précis de chacune des variables est disponible dans le fichier `readme.md`. Pour mener une analyse sur ces données, nous allons utiliser le logiciel `R`, que vous pouvez utiliser aussi sous Linux, Windows, ou Mac. Trois possibilités s'offrent à vous

- la version console (sous unix) : tapez simplement la commande R dans un terminal,
- Rgui : installée sous windows, elle vous permet d'avoir une interface minimale pour commencer à travailler,
- RStudio : version plus récente, et plus agréable. RStudio doit normalement être maintenant installé sur les machines de l'école. Note pour les machines personnelles: il faut installer R **puis** RStudio.

Avant-propos (1.5h)

Lors de ce TP, vous aurez besoin d'un certain nombre de fonctions disponibles au sein de `R`, quelques unes vous sont présentées ici, et vous en découvrirez par vous même dans la suite du TP.

Cet avant-propos n'est pas à inclure dans le rapport que vous rendrez, mais cela ne vous dispense pas de le faire, ce que vous ferez vous servira par la suite.

1. `matrix(<valeur.defaut>, <n.lignes>, <n.colonnes>)` crée une matrice. Quelle est la matrice obtenue en exécutant `matrix(c(1,2),3,4)` ? Expliquez.
2. `diag(<vecteur>)` ou `diag(<valeur.defaut>, <n.lignes>, <n.colonnes>)` crée une matrice diagonale. Donnez 2 commandes possibles utilisant `diag` permettant d'obtenir une matrice identité de taille 4×4.
3. `diag(<matrice>)` renvoie la diagonale d'une matrice. Calculez la trace d'une matrice aléatoire de taille 4×4 en utilisant cette fonction.
4. L'opérateur `%*%` permet de multiplier deux matrices. Cherchez et expliquez la différence entre `%*%` et `*`.
5. Générez l'ensemble de données $X \cup Y$ suivant. X est un ensemble de 100 observations bi-dimensionnelles suivant une loi normale de paramètre $\mu_x = [1, 3]$ et $\sigma_x = [2, 0.5]$. Y est un ensemble de 100 observations bi-dimensionnelles suivant une loi normale de paramètre $\mu_y = [3, 4]$ et $\sigma_y = [0.7, 0.9]$. Affichez ces données, et colorez les observations provenant de X en rouge, et celles provenant de Y en vert.
6. Écrivez une fonction `distanceMat(x,y,M)` déterminant la distance de Mahalanobis de paramètre M entre deux vecteurs x et y, définie par
$$d_M(x, y) = \sqrt{(x - y)^T M (x - y)}$$
Vérifiez votre fonction avec la matrice $M = Id$.
7. Écrivez une fonction `inertie(X,M)` renvoyant l'inertie du nuage de points X utilisant la distance définie dans la question précédente. Calculez cette inertie sur les données générées précédemment, en utilisant une distance Euclidienne, puis la métrique $M = \sigma^{-2}$. Commentez les valeurs obtenues.

Analyse en composantes principales (4.5h)

1. Dans l'environnement R, chargez les données euro. Assurez-vous qu'elles sont correctement interprétées. Faites en particulier attention au nom des lignes et des colonnes.

2. A l'aide d'une visualisation par boîte à moustache pour chaque variable, faites vos premiers commentaires sur la distribution des valeurs sur chaque variable. Vous pourrez également utiliser d'autres méthodes d'exploration statistique que vous avez étudié l'année dernière.
3. Calculez la matrice de variance-covariance V des données, sans utiliser la fonction `cov`. Que peut-on dire de ces valeurs, en confrontation avec les deux questions précédentes ?
4. Continuez votre analyse en considérant la matrice de corrélation linéaire. Existe-il une corrélation entre les différentes variables descriptives ?
5. A l'aide de la fonction `prcomp`, déterminez les composantes principales. Ces composantes forment-elles une base orthonormée (le prouver numériquement) ? A quoi correspondent les paramètres center et scale? Quel est leur influence sur le résultat ?
6. Grâce à ces composantes, déterminez les coordonnées de chaque pays dans la nouvelle base, et affichez les sur le premier plan factoriel (avec leur nom). Quels sont les pays qui sortent du lot ?
7. Représentez l'ébouli des valeurs propres et pourcentage des variances expliquées sur un même graphique. Combien de composantes retenez-vous ? Justifiez votre réponse.
8. A l'aide de la fonction biplot, observez les projections des individus et les variables initiales. Interprétez ce graphique.
9. Observez d'autres plans factoriels, et commentez.
10. Déterminez quelles sont les variables les mieux représentées par le premier plan factoriel.
11. Calculez la contribution de chacun des individus à la construction des composantes principales. Doit-on supprimer des individus de l'analyse ?
12. La projection des individus sur les composantes correspond-elle, d'une manière ou d'une autre, aux similarités attendues ?

Partitionnement (3h)

Nous continuons d'utiliser le même jeu de données pour cette fois-ci procéder au groupement de pays.

1. Après avoir déterminé les deux matrices de dissimilarités en utilisant respectivement une métrique Euclidienne et une métrique réduite, effectuez une classification ascendante hiérarchique des données fondée pour chaque matrice sur le saut minimum.
2. Représentez le dendrogramme ainsi que la "hauteur" (attribut height) en fonction du nombre de classes. Que représente la "hauteur" ici ? Ou couperiez-vous le dendrogramme ?
3. Caractérisez les classes obtenues en calculant pour chaque classe son centre de gravité et son inertie. Interprétez.
4. Effectuez une autre classification en utilisant le critère de Ward. Commentez les différences de résultats.
5. On considère maintenant l'algorithme des centres mobiles. Utilisez la fonction `kmeans` prenant en paramètres les données X et le nombre de classes désiré k. Cette fonction retourne une liste de classe d'affectation pour chacune des observations. Vous initialiserez de manière aléatoire les k centres. Plusieurs choix sont possibles, expliquez celui que vous prenez.
6. Affichez l'inertie expliquée en fonction du nombre de classes. On désire maintenant trouver un nombre de classes adapté. Pour cela, on fait varier le nombre de classes, et l'on cherche à "optimiser" un critère. Sur quoi un tel critère peut-il se fonder ? Proposer votre critère, et essayer de déterminer de manière **automatique** le nombre de classes optimum au sens de ce critère. Obtient-on les mêmes résultats selon l'algorithme utilisé ?
7. Effectuez une ACP des données et représentez les classes obtenues par CAH et par centres mobiles dans les plans factoriels retenus afin d'inspecter visuellement la qualité ou la représentation de la classification.
8. En vous concentrant sur une zone restreinte de l'espace de votre choix, commentez les proximités des pays par rapport à la similarité que vous-même percevez de ces pays.