

Master 1 EDSB

Data Mining

TD 2 : ACP

Christelle Reynès/ Myrtille Vivien



I - Influence du centrage/réduction : les Poissons d'Amiard

Rappels du cours : les données

Le jeu de données est constitué de 23 individus, des mulets (sortes de rougets), qui ont été répartis dans 3 bassins contaminés de la même façon par une source radioactive. Les temps d'exposition à la radioactivité diffèrent pour les poissons des différents bassins.

Sur les poissons, on mesure la radioactivité pour 9 différents organes :

ry (des yeux), **rb** (des branchies), **ro** (des opercules), **rn** (des nageoires), **rf** (du foie), **rtd** (du tube digestif), **rr** (des reins), **re** (des écailles), **rm** (des muscles).

On mesure également 7 variables de taille :

LO (longueur), **LS** (Longueur standard), **P** (poids), **LT** (largeur de la tête), **LA** (largeur), **LM** (largeur du museau), **DY** (diamètre des yeux).

Constitution des jeux de données transformées :

1. Importer les données et nommer *amiard* la matrice obtenue.
2. Construire un jeu de données centré et non réduit et nommer *amiardC* la matrice obtenue.
3. Construire un jeu de données centré et réduit et nommer *amiardCr* la matrice obtenue.

Analyses univariées et bivariées :

1. Calculer la moyenne et l'écart-type de chaque variable pour la matrice *amiard*. Commenter.
2. Observer la distribution des variables avant transformation.
3. Faire de même pour les deux matrices transformées. Commenter.
4. Visualiser les liens bivariés (graphique et corrélation linéaire). Repérer quelques liens forts.

Analyse en Composantes Principales :

1. Appliquer l'ACP centrée-réduite à la matrice *Amiard* (fonction **princomp**) et conserver l'objet obtenu (*resacpCr* par exemple).
2. Explorer les différents résultats produits par la fonction. Retrouver les résultats vus en cours.
3. Etudier ensuite les différents éléments de l'objet obtenu.
4. Pour apprendre à manipuler quelque peu ces sorties, reconstruire le premier plan en faisant apparaître les trois groupes d'individus en différentes couleurs. Idem pour ce qui est de la représentation des variables.
5. Enfin, même si cela n'a pas forcément d'intérêt pour ce jeu de données (d'après l'écoulement des valeurs propres) représenter en 3 dimensions les 3 premiers axes de l'ACP (fonction **plot3d** du package **rgl**).
6. Recommencer les deux dernières opérations pour l'ACP centrée-non réduite.

II - Mise en œuvre et interprétation d'une ACP : jeu de données « Villes »

Ce second jeu de données présente des informations concernant des villes françaises :

- *varpop* : taux annuel moyen de variation de la population 1990 – 1999
- *longitude*
- *latitude*
- *altitude*
- *soleil* : ensoleillement en nombre d'heures par an
- *prixm2* : prix moyen de l'immobilier (vente), en €/m²
- *revenu* : revenus moyens par ménages (2004)
- *txchom* : taux de chômage (2005)
- *population* : nombre d'habitants en 2004
- *txcadres* : proportion de la population active constituée de cadres et professions intellectuelles
- *ouvriers* : proportion de la population active constituée d'ouvriers
- *pasdiplom* : proportion de la population des plus de 15 ans, non scolarisée et sans diplôme
- *nivbac* : proportion de la population des plus de 15 ans, non scolarisée et ayant un niveau bac ou brevet professionnel
- *nivsupbac2* : proportion de la population des plus de 15 ans, non scolarisée et ayant un niveau d'étude supérieur à bac+2
- *sarko1* : résultats obtenus par Nicolas Sarkozy au premier tour de l'élection présidentielle 2007, en% des votes exprimés non nuls.
- *sarko2* : résultats obtenus par Nicolas Sarkozy au second tour de l'élection présidentielle 2007, en% des votes exprimés non nuls.
- *lepen1* : résultats obtenus par Jean-Marie Le Pen au premier tour de l'élection présidentielle 2007, en % des votes exprimés non nuls.
- *sego1* : résultats obtenus par Ségolène Royal au premier tour de l'élection présidentielle 2007, en% des votes exprimés non nuls.
- *bayrou1* : résultats obtenus par François Bayrou au premier tour de l'élection présidentielle 2007, en % des votes exprimés non nuls.

Exploration des données :

1. Importer le fichier « **villes** ».
2. Utiliser le contenu de la première colonne comme noms des lignes.
3. Construire la matrice centrée et réduite.
4. Utiliser la fonction **pairs()** pour visualiser les relations bivariées entre les variables.

Analyse en composantes principales :

1. Appliquer l'ACP centrée-réduite aux données.
2. Explorer et commenter les différentes sorties.
3. Choisir un nombre optimal de composantes.