: 'Apprentissage statistique' – Cours Analyse de données / réduction de dimension -

Mise en place de l'algorithme de l'AFD – ***TP n° 3 : 2 h*** **– octobre 2023**

Travail en groupe TP

Dans cette séance il s'agit de :

- Comprendre ce qu'est l'AFD
- De faire une AFD et de comprendre sa représentativité après la projection (approche géométrique descriptive de l' AFD)
- Appliquer à un cas de données du TP1
- Etudier la qualité de la représentation des individus projetés et des variables initiales dans ce nouvel espace réduit sur R^k (avec k<p)
- Comparer les résultats obtenus avec deux packages existant sous *R ou Pyhton*

Travail à réaliser : Soit avoir un programme sous R ou Python + commentaires

1) Afin d'évaluer le pouvoir discriminant des variables d'origine il est intéressant pour chaque groupe *k* de calculer la valeur de discrimination d'une variable Z_j avant projection. Soit représenter chaque groupe (modalité de la variable qualitative) dans les données.

$d = V^{-1}B$ Valeur de discrimination d'une variable X_j avant projection

Vous devez pour cela calculer les variances intraclasses W_k (pour un groupe *k*) et interclasse B, et V variance totale : calculer et comparer ces 3 variances. Les variables initiales partagent elles bien les groupes dans l'exemple donnée ?

2) Etudiez les matrices V^{-1} et B : dimension, caractéristiques : Pouvez toujours diagonaliser directement $V^{-1}B$?
3) Si vous choisissez de diagonalisation cette matrice, utiliser le complément sur la diagonalisation donnée en cours (technique de régularisation). Apres diagonalisation comme en ACP vous obtiendrez la projection des individus dans un nouvel espace par une méthode factorielle par projetctions.
4) Implémenter l'AFD centrée et tester sur le fichier de données fournies sous campus avec les deux packages.
Vous disposez de deux packages : ade4 : fonction dudi.mix et package : FactorMines : fonction FAMD

5) Comparer les valeurs propres et vecteurs propres des matrices $V^{-1}B$, et de $W^{-1}B$

6) Que remarquez-vous entre la valeur propre λ_s de chaque axe et du rapport de corrélation η_s ?

7) Fournir l'équation de la combinaison linéaire du premier axe factoriel discriminant correspondant à la première valeur propre. (il existe équations des q-1 axes factoriels (q nombre max de groupes))

8) Vous devez disposer d'un script.R pouvant faire une AFD sur données centrées : la qualité de la projection sera étudiée :

- Qualité de la projection d'un nuage par axe a_s
$\frac{\lambda_s}{\sum_{s\in q-1}\lambda_s}$
- Contribution absolue du centre de gravité g_q à l'axe a_s correspondant à λ_s a la valeur propre et vecteur propres V_s

$$\frac{I_k}{n}(a_s' V^{-1} g_k)^2$$

- Contribution relative du centre de gravité g_q à l'axe a_s

$$\frac{I_k}{n}\frac{1}{\lambda_s}(a_s' V^{-1} g_k)^2$$

Calculer ces 3 indicateurs pour chaque axe et chaque groupe

9) Vous devez modifier votre script.R pouvant faire une AFD sur données centrée-réduites *ce qui permettra d'interpréter les contributions partielles de variables aux nouvelles variables cad leurs corrélations.*

9.1 Pour cela vous devez modifier votre code pour centrée et réduire vos données initiales (attention à réduire avec l'écart type intra classes)

9.2 refaire tourner AFD sur ces données standardisées

9.3 Comparer les coefficients de chaque axe factoriel dans les deux cas : existe-il une relation entre les deux coefficients ?

9.4 Interprétation des variables par rapport aux nouveaux axes factoriels. Dans cette dernière étape vous devez évaluer la représentation des *p* variables projetées, selon les nouveaux axes, soit les variables initiales Zj selon les axes factoriels $\underline{a_s}$ associés à la valeur propre et notamment pour l'axe factoriel principal et le second axe factoriel. Une question de l'AFD réside dans la signification des axes discriminants qui représentent les nouvelles variables obtenues par combinaison linéaire des variables initiales. On étudie en général les corrélations avec les variables initiales sur le de cercle de corrélation (AFD centrée réduite). Pour cela, on calcule les coefficients de corrélation entre le kième axe discriminant a_k et la j^{ième} variable initiale Z_j : c'est la j^{ième} composante du k^{ième} vecteur propre v_k multipliée par la racine de λ^k :

$$r(a_k, Z_j) = \sqrt{\lambda_k}\ (v_k)_j$$

Calculer ces coefficients de corrélation avec les différentes variables.

9.5 Projeter les p variables Z_j pour j [1..p] dans le nouvel espace des *k* axes retenus et tracer les p variables dans le plan factoriel retenu (axe factoriel1,axe factoriel 2) sur le cercle de corrélation.

FactoMineR Pagès, J., «Analyse Factorielle de Données Mixtes », Revue de Statistique Appliquée, Vol : 52, Issue : 4, pp. 93, 111, 2004
ADE4 Hill, M., Smith, A., «Principal Component Analysis of taxonomic Data withmulti - State discrete Characters », Taxon, 25, pp. 249 255,1976
(PCAmixdata)Kiers , H.A.L., « Simple structure in Component AnalysisTechniques for mixtures of qualitative and quantitative variables », Psychometrika , 56, pp. 197 - 212, 1991

Pour charger un package #chargement du package library(FactoMineR)
#lancement de la procédure FAMD(data,ncp=2)
#affichage des résultats
Print(summary(afdm.datar))

Annexes

- A partir des valeurs propres et des vecteurs propres: le pouvoir discriminant (rapport de corrélation = η = SCE/SCT) de chaque axe factoriel a_s est évalué en calculant les nouvelles coordonnées z_{ij} des individus x_{ij} à partir de l'équation de l'axe a_s puis les barycentres conditionnels des groupes projetés sur cet axe soit :

(SCT = SCE + SCR) $\Leftrightarrow$

$$\sum_i (z_i - \bar{z})^2 = \sum_k n_k (\bar{z}_k - \bar{z})^2 + \sum_k \sum_i (z_{ik} - \bar{z}_k)^2$$

A partir du calcul de η_s de chacun des axes factoriels : on peut choisir le nombre d'axes à retenir.

- Diagonalisation de la matrice $\mathbf{V^{-1}B\,a} = \lambda\,\mathbf{a}$

Avec **a** vecteur propre de $\mathbf{V^{-1}B}$, associé à la plus grande valeur propre λ

Posons **C** tel que **B = C^t C** et $\mathbf{a = V^{-1}C\,v}$

On a que : $\mathbf{B\,a} = \lambda\,\mathbf{V\,a}$

On obtient alors : $\mathbf{C\,C^t V^{-1} C\,v} = \lambda\,\mathbf{C\,v}$

Il suffit alors de diagonaliser la matrice $\mathbf{C^t V^{-1} C}$ qui est d'ordre q le nombre de classes (nombre de modalités de la var Y) et qui donne pour la valeur propre λ le vetceur propre V ; puis de revenir au vecteur à l'aide de $\mathbf{a = V^{-1}C\,v}$

En règle générale, il y a q -1 valeurs propres donc q-1 axes discriminants. C'est le cas si N> K > q et si les variables ne sont pas liées linéairement.

La matrice **B**

$$\mathbf{B} = \frac{1}{N}\sum_{c\in q} I_c\,(g_c - G)\;\;(g_c - G)^{\,t}$$

La matrice **C** a pour élément : $c_{kc} = \sqrt{\frac{Ic}{N}}\;(g_{ck} - G_k)$ et de dim (k,q).
