Majeure *Science des données* 2020_21

UP3: 'Analyse de données et données temporelles' - Cours 'Analyse de données multivariées -Analyses factorielles'
TP n°1 - Analyse en composantes principales - *TP 1H30*

Cette séance de 1h30 de TP est consacrée à l'Analyse en Composantes Principales.
Objectifs :

- 1) Programmer les différentes étapes de l'ACP centrée et ACP normée (centrée - réduite),
- 2) D'étudier la représentativité de l'ACP avec pour la réduction de dimension 1) le choix de la dimension $m < p$ et 2) d'évaluer la qualité de la projection des individus (projection de $\mathbb{R}^p$ et dans $\mathbb{R}^n$) et la qualité de la projection des variables X^j.
- 3) De visualiser la qualité de la réduction en fonction des propriétés intrinsèques à l'échantillon

Il ne s'agit pas d'utiliser directement les fonctions d'ACP mais bien de programmer les étapes de l'ACP vues en cours dans cette première partie.

La mise en œuvre se fera sur les données anonymes : fichier *data_PDE20.txt* ou au format csv *data_PDE20.txt* .Pour cela vous aurez besoin de télécharger le fichier sur : https://campus.emse.fr/

Vous devez déposer votre TP par trinôme sur campus : le(s) script(s) développé(s) + la fiche de compte rendu remplie (sous campus). >>>>>**Rendus TP pour la prochaine séance de TP**

On ne vous demande pas de faire l'interprétation de votre ACP : elle sera envisagée au TP N°4 ; le jeu de données vous permet seulement de valider et/ou de tester votre programmation.

A noter :
Il s'agira pour vous de vous constituer un programme qui fonctionne, et efficace pour pouvoir être utilisé soit dans les TP suivants, soit sur *l'étude de cas* qui sera à rendre à la fin de cette UP.

Partie 1 - programmer l'ACP sur l'espace de variables

1. Mettre au point un script qui permet successivement de :
- visualiser la matrice X de dimension (n,p) (n individus et p variables)
- Construire les indicateurs statistiques classiques : variance, covariance, écart-type
2. Développer un script qui permet de faire successivement :
- la translation du nuage des individus dans l'espace initial $\mathbb{R}^p$ puis,
- de trouver les hyperplans pour lesquels l'inertie projetée est maximale : les hyperplans sont associés aux p directions de l'espace, chaque axe factoriel est de vecteur propre u (respectivement matrice U) et une valeur propre λ (respectivement matrice Λ).

Vous disposez à ce stade i) d'un script R (soit Python) qui soit paramétrable, ii) ou deux scripts R qui vous permettent de faire ces 2 premières étapes (ACP centrée - ACP normée) puis de faire la partie 2.

Les fonctions suivantes de R seront utiles : (plot, eigen, plot3d) équivalent sous Python .

Partie 2 - qualité de l'ACP

Pour l'évaluation de la qualité de la réduction de données vous devez :

- considérer la qualité de la réduction du nuage
- faire la visualisation de la projection des individus
- faire le calcul de la contribution et la qualité des individus projetés
- Vous devez envisager aussi en parallèle le cas de l'ACP normée.

3. Pour représenter l'inertie expliquée λ_j par les j différentes composantes principales obtenues (j allant de *1 à p*) : en utilisant la fonction barplot() *Creates a bar plot with vertical or horizontal bars , e*t la somme cumulée d'inertie pour les

j de *1 à p*. Vous pourrez également tester les autres règles de choix du nombre de composantes principales à retenir à partir des λ_j pour chaque composante *j*.

4. Calculer les nouvelles coordonnées de l'individu *i* sur chacune des composantes soit C_i^j , qui permettra de définir la qualité de la projection de l'individu Q_i^k, (*k* étant le nombre de composantes principales retenues) définie par : $Q_i^k = \frac{\sum_{j=1}^{k}(c_i^k)^2}{\sum_{j=1}^{p}(c_i^j)^2}$

5. La contribution de l'individu *i* à l'inertie de l'axe factoriel *j*, est définie par $\gamma_i^j = \frac{\frac{1}{n}(c_i^j)^2}{\lambda_j}$

6. Vérifier que votre premier plan factoriel est correct en comparant avec la fonction R, qui résout l'ACP : library(ade4) fonction dudi.pca () par exemple.

7. Pour donner une signification à une composante principale, on la relie à une des *p* variables initiales X^j en calculant un coefficient de corrélation linéaire entre une composante *c* et une variable *j* définit par : $r(c, X^j) = \sqrt{\lambda}\, u_j$ pour l'ACP normée (u_j coordonnées du vecteur *u* pour la *j* variable et λ la valeur propre associée à la composante *c) :* $r(c, X^j) = \frac{\sqrt{\lambda}}{\sqrt{Var(Xj)}} u_j$ pour ACP centrée

Ce qui permet de calculer sa contribution et sa qualité dans le nouveau sous-espace. Les points 4. à 7. permettent de faire une analyse de la réduction de dimension.

8. Représenter graphiquement les nouveaux individus dans le nouveau sous-espace selon les premiers et deuxièmes plans retenus (plot(CP1,CP2) ou plot(CP1,CP3) …)

9. Vous disposez de la fonction dudi.pca (). Un certain nombre de critères de l'ACP est fourni par le frame de cette fonction. Mettre en œuvre sur les données du fichier fourni. Comparer vos résultats obtenus à ceux obtenus par la fonction dudi .pca. ou équivalent sous python.

Sous R :

Quelques fonctions utiles : sample() , rbind() ,st()…, boot()
Lire un fichier sous R :
data_PDE19 <- read.csv("~/Seafile/enseignement/ACP_AMV/scriptR/data_PDE19.prn", sep="")
View(data_PDE19)
data_PDE19t <-read.csv2("~/Seafile/enseignement/ACP_AMV/scriptR/data_PDE19t.txt")
View(data_PDE19t)

Rappels et supports

- On peut tracer de l'inertie expliquée λ_j par les *j* différentes composantes principales obtenues (*j* allant de *1 à p*) : en utilisant la fonction barplot() *Creates a bar plot with vertical or horizontal bars*
- *S*oit les nouvelles coordonnées de l'individu *i* sur chacune des composantes soit C_i^j : la qualité de la projection de l'individu Q_i^k, (*k* étant le nombre de composantes principales retenues) définie par : $Q_i^k = \frac{\sum_{j=1}^{k}(c_i^k)^2}{\sum_{j=1}^{p}(c_i^p)^2}$
- La contribution de l'individu *i* à l'inertie de l'axe factoriel *j*, est définie par $\gamma_i^j = \frac{\frac{1}{n}(c_i^j)^2}{\lambda_j}$
- Une composante principale, est reliée à une variables initiale X^j en calculant un coefficient de corrélation linéaire entre une composante c et une variable j définit par : $r(c, X^j) = \sqrt{\lambda}\, u_j$ pour ACP normée (u_j coordonnées du

vecteur u pour la j variable et λ la valeur propre associée à la composante c $r(c, X^j) = \frac{\sqrt{\lambda}}{\sqrt{Var(Xj)}} u_j$ pour ACP centrée

- les fonctionnalités existantes sous R pour vous familiariser avec ces toolboxes, comme : princomp, boot et des fonctions comme apply, quantile , replicate, … onction dudi.pca ()

Dans la : library(ade4) fonction dudi.pca ().
