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

Analyse en Composantes Principales - ***TP 2h00 + travail personnel***

Vous devez déposer votre TP ACP par binôme) sur campus : ***le(s) script(s) développé(s)*** *+ le compte rendu en* ***pdf*** *(sous campus) en respectant la date limite qui vous est donnée* ***soit le 9/11/2023***

Objectifs :

- Programmer les différentes étapes de l'ACP centrée et ACP normée (centrée - réduite),
- 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.
- De visualiser la qualité de la réduction en fonction des propriétés intrinsèques à l'échantillon
- ACP et forme du nuage : étude des espaces R^p et R^n
- *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.*

Il s'agira pour vous de vous constituer un programme qui fonctionne, et efficace pour pouvoir être utilisé dans le TP commun de classification.

Récupérer le début du TP séance précédente, soit les étapes suivantes

1. Retrouver les différentes de l'ACP dans le cadre de dimension quelconque d'une matrice de données X de di$\mathbb{R}^{n\times p}$: donner un pseudo code pour les différentes étapes
2. Implémenter sous R ou sous Python le code obtenu et tester vos résultats pour le cas des 8 points en dimension 2 pour les questions allant de 1 à 7.Pour tester vos scripts vous pouvez aussi utiliser les données founies dans data_PDE20.csv sous campus.

Partie 1 – L'ACP sur l'espace de variables $\mathbb{R}^p$

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 (***fait dans TP1.1***) :
   - la translation du nuage des individus dans l'espace initial $\mathbb{R}^p$ (centrer le nuage)
   - 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) (diagonalisation de matrice variance-covariance).

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[1].

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 permet de définir la qualité de la projection de l'individu Q_i^k, (*k* étant le nombre de composantes principales retenues) définie par : $Q_i^k = \frac{\sum_{j=1}^{k}(c_i^j)^2}{\sum_{j=1}^{p}(c_i^j)^2}$

5. La contribution de l'individu i à l'inertie de l'axe factoriel j, est définie par $\gamma_i^j = \frac{\frac{1}{n}(c_i^j)^2}{\lambda_j}$ .Ce qui permet de calculer sa contribution et sa qualité dans le nouveau sous-espace

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.

Les points 4. à 6. permettent de faire une analyse de la réduction de dimension.

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

Vous disposez de la fonction dudi.pca (). Un certain nombre d'éléments 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_PDE20.csv", sep="")
View(data_PDE19)
data_PDE19t <-read.csv2("~/Seafile/enseignement/ACP_AMV/scriptR/data_PDE20.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

[1] S*tandardisation (centrer-réduire)* est similaire au terme "*normalisation*" en Analyse des données : est la transformation de données qui soustraie à chaque valeur une valeur de référence (la moyenne de l'échantillon) et en la divisant par l'écart-type ce qui permet de rendre les unités compatibles avec une distribution de moyenne 0 et d'écart-type 1.

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. Ceci permet de donner une signification à une composante principale, en la reliant à une des p variables initiales X^j par le coefficient de corrélation linéaire entre une composante c et une variable j.

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

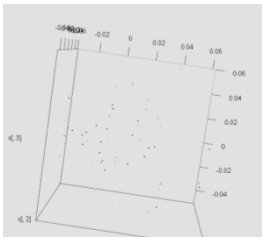
Dans la : library(ade4) fonction dudi.pca ().

Partie 3 : Etude de la forme du nuage initiale et réduction de dimension[2]

8. Nuage isotrope

On veut générer des données pour 3 variables d'un nuage de points proches d'une sphère. Pour cela vous pouvez choisir plusieurs méthodes mais une possibilité est de générer un échantillon de taille n des 3 variables X, Y, Z comme 3 vecteurs gaussiens indépendants de loi $N(0,1)$ puis de récupérer le vecteur V à 3 composantes tel que V/||V||. Car la densité f(v) d'un vecteur gaussien est isotrope et dépend que de v.

Vous pouvez visualiser votre nuage par la :
open3d()
plot3d(x[,1], x[,2], x[,3], col = rainbow(n))



-Vous devez écrire le script qui génère ces données, puis mettre en œuvre l'ACP sur ces données
- En fonction de la taille de l'échantillon, suivre l'évolution de la matrice de covariance ou de corrélation, la cascade des valeurs propres, ainsi que la qualité de la projection des individus. Interpréter.

9. Nuage non isotrope

On cherche maintenant à voir sur des formes non isotropes si des corrélations plus fortes entre certaines variables changent l'ACP et comment. On génère des données pour 3 variables qui constituent un nuage de points pour lesquelles deux variables sont corrélées fortement. Pour cela vous pouvez choisir plusieurs méthodes mais une possibilité est de générer un échantillon de taille n pour 3 variables X, Y,Z puis appliquer une relation linéaire entre les deux premières X et Y , puis entre Y et Z. Vous pouvez modifier votre corrélation entre deux variables en rajoutant un bruit.

- Tester l'ACP sur les données générées et en fonction de la taille de l'échantillon suivre l'évolution de la matrice de covariance ou de corrélation, la cascade des valeurs propres, ainsi que la qualité de la projection des individus. Interpréter.
  On vous donne le cas de 3 variables générées selon le script suivant :
  x <- sort(rnorm(900))
  y <- rnorm(900)

[2] Cette partie est consacrée à l'étude de l'ACP en fonction du type du nuage de point, la validation et à la stabilité de l'Analyse en Composantes Principales et son approche dite 'dual'. Vous devez disposer des codes de la partie 1 de l'ACP que vous venez de développer. Il s'agit d'étudier la qualité de la réduction en fonction des propriétés intrinsèques à l'échantillon. Ceci fait suite à la mise en œuvre d'une ACP centrée et d'une ACP normée et de sa qualité de réduction pour les variables et les individus.

z <- rnorm(900) + atan2(x, y)

- Que vous donne une telle ACP ? Devez-vous normées les données ?

10. Points extrémaux

Dans le cas d'un nuage dont la forme est non isotrope (choisir votre générateur de nuage) on veut rajouter quelques points extrémaux pour une ou deux variables (vous allez étirer ainsi votre nuage selon cette direction) soit au contraire ôter des points extrémaux (ou quelques points) pour compacter votre nuage selon quelques variables.

- Proposer quelques essais d'ACP centrée ou ACP normée sur ces données, qu'observez-vous ? Que pouvez-vous conclure sur le fait de centrer ou normer les données.

Partie 4 . Etude de la forme du nuage initiale sur la réduction de dimension dans les deux espaces

Il s'agit :

- De mettre en œuvre le volet 'dual', qui sera à la base de la décomposition par ACP espace $\mathbb{R}^n$
- De visualiser la qualité de la réduction en fonction des propriétés intrinsèques à l'échantillon selon ces deux approches $\mathbb{R}^p$ ou $\mathbb{R}^n$

Dans cette partie les données sont celles des deux nuages de l'étape précédente : soit :

Cas 1 : Nuage isotrope sur l'espace $\mathbb{R}^n$

Dans $\mathbb{R}^p$, à partir des données pour 3 variables d'un nuage de points proches d'une sphère dont la génération de l'échantillon de taille *n* des 3 variables X, Y, Z faite avec 3 vecteurs gaussiens indépendants de loi N(0,1) dont le vecteur V à 3 composantes tel que V/||V||. (Densité f(v) d'un vecteur gaussien est isotrope et dépend que de v) : ->>>>récupérer le nuage en $\mathbb{R}^n$

Cas 2 Nuage non isotrope dans $\mathbb{R}^n$

Dans $\mathbb{R}^p$, le nuage de points dans un R^3 pour lesquelles deux variables sont corrélées en générant un échantillon de taille *n* pour 3 variables X, Y,Z et en appliquant une relation linéaire entre les deux premières X et Y , puis entre Y et Z : le cas de 3 variables générées selon le script suivant :
x <- sort(rnorm(900))
y <- rnorm(900)
z <- rnorm(900) + atan2(x, y) ->>>**R**écupérer le nuage en $\mathbb{R}^n$

11. Vous avez développé l'ACP centrée et centrée réduite sur l'espace $\mathbb{R}^p$: on vous demande successivement de :
    1) Mettre en œuvre de l'ACP sur espace $\mathbb{R}^n$sur les deux cas
    2) De comparer avec les résultats dans $\mathbb{R}^n$ sur les deux cas
    3) Retrouver les formules de passage à partir du cours et de vos résultats successifs
    4) Valider les résultats sur vos exemples
    5) Tester sur les deux cas, l'effet de la dimension sur les deux décompositions
    6) Proposer la projection de points supplémentaires (espace sur $\mathbb{R}^p$) et de points colonnes (variables) dans ce nouvel espace $\mathbb{R}^n$. Pour cela vous devez proposer de générer soit des colonnes supplémentaires soit des lignes supplémentaires comme individus lignes :

Cas 1 : un ou deux vecteurs gaussiens V indépendants de loi N(0,1) tel que V/||V|| (densité f(v) d'un vecteur gaussien est isotrope et dépend que de v) permettront ici de créer de nouvelles variables et/ou d'étendre l'échantillon pour placer ensuite ces nouveaux points dans l'espace réduit.
Cas 2 : générer des points supplémentaires pour les 3 variables X,Y,Z et/ou générer une ou deux variables liées aux 3 précédentes . Attention ne pas refaire l'ACP sur ces données car elles sont dites non actives pour la réduction.

A. Partie 3 : (1h30) : ACP sur données réelles environnementales

Dans cette séance il s'agit de d'acquérir un savoir-faire de l'ACP adaptée aux données réelles et leur interprétation. Dans ce TP, les différentes étapes sont de :

- Faire une étude statistique préparatoire des données (moyenne, écart type, médiane, corrélation entre variables) : attention aux données manquantes ou égale à 0
- Mettre en œuvre une ACP avec analyse et interprétation des résultats (avec représentativité de l'ACP)
- Apporter des réponses aux questions posées sur ce cas de décision.

- Données fournies des 6 campagnes: BF2, BF3, CA1, CA2, CA3, CA4. BF pour bruit de fond et CA pour campagne avant installation du site. Un fichier TP4_covC1234_DS19_20.xls sous campus.
- Vous disposez de vos programmes (my_PCA_strandardized.r et my_AFD.r) réalisés dans les TD/TP précédents.
- Vous pouvez utiliser l'outils de r : library(FactoMiner) , library(ade4)

Le travail à rendre se fait sous la forme d'un rapport de type note de synthèse contenant les choix, les traitements et analyses faites sur ce type de données avec les codes sources développés + résultats dans un même fichier.zip (sous format de : noms_binome.zip) sur campus.

Objet d'étude :

Dans le cadre de projet de recherche industrielle, on s'intéresse à la contribution d'un site industriel de traitement de déchets verts par compostage lors de la mise en exploitation, localisé dans la Loire. En effet, un tel processus dans certaines conditions de fonctionnement (entrant important à différentes périodes de l'année, conditions de fermentation anaérobie au lieu de dégradation aérobie, mauvaise gestion du site) peut entrainer l'émission de composés chimiques avec des risques sanitaires potentiels au niveau des populations avoisinantes.
Afin de discriminer la contribution du site par rapport à la présence éventuelle de ces composés avant installation (que l'on appelle bruit de fond) des campagnes de mesure de ces composés ont été effectuées avant (dans le labels les 2 lettres BF) et après la mise en activités du site (lettre CA dans le labels) à différentes périodes de l'année (H pour hiver et E été).

On cherche donc à répondre à certains questionnements comme :

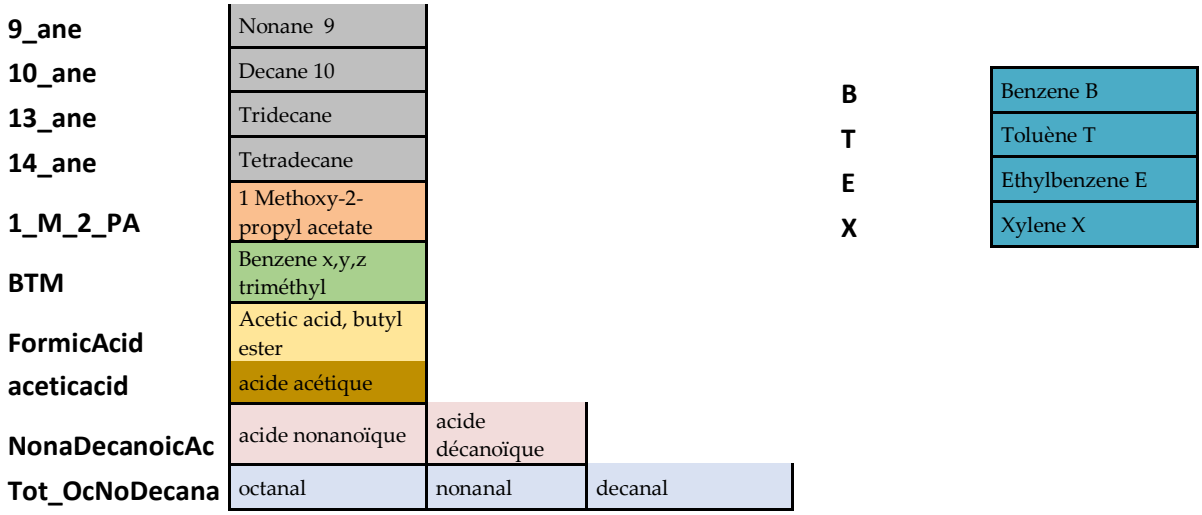
- la localisation des *m* points de mesure autour du site, montre-elle des regroupements de comportement (composés chimiques atmosphériques d'origine industrielle, automobile, milieu urbain, milieu rural…)?
- existe-il une différence entre les campagnes hiver/ été ?
- existe- il une signature entre les individus avant et après la mise en activité du site ?

Description des données fournies

- **Données réelles**

On dispose de 6 campagnes de mesure effectuées sur m points de mesure, pour un certain nombre de COV (composés organiques volatiles) : Liste des p variables (p=14) : p composés (ou familles de composés) : familles de composés COV

B T E X 9_ane 10_ane 13_ane 14_ane 1_M_2_PA BTM FormicAcid aceticacid NonaDecanoicAc Tot_OcNoDecana



Effectués sur m points (localisation) donnés ci -dessous (plusieurs mesures sur certains mêmes points) :

P19	P21	P18	P17	P20	P10	P02	P01	P05	P06	P07	P03	P08
P13	P14	P15	P09	P16	P11	P19	P21	P18	P20	P02	P05	P06
P07	P03	P08	P13	P14	P15	P04	P09	P16	P11	P10	P12	P19
P21	P18	P17	P20	P05	P06	P07	P03	P08	P13	P14	P15	P04
P09	P16	P10	P02	P01	P12	P22	P19	P21	P18	P17	P20	P05
P06	P07	P03	P08	P13	P14	P15	P04	P09	P16	P32	P33	P10
P29	P11	P26	P30	P27	P28	P25	P02	P01	P12	P19	P21	P17
P20	P05	P06	P07	P03	P08	P13	P14	P15	P09	P16	P32	P29
P11	P26	P27	P28	P25	P02	P01	P12	P19	P21	P18	P17	P20
P05	P06	P07	P03	P08	P13	P14	P15	P04	P09	P16	P32	P33
P29	P11	P26	P30	P27	P28	P25	P02	P01	P12			

Données météorologiques durant les campagnes de mesures

Campagne	Date de début	Date de fin	nombre de jours	Période	T° moyenne (°C)	Cumul de pluie (mm)	% de temps sans vent	Principales directions des vents (%)			% de vents supérieur à 2 m.s-1
BF1	09/05/2007	16/10/2007	160	Estivale	16,6	386	30,6	N : 23,7%	NNW : 6,4%	SE : 4,4%	17,57
BF2	16/10/2007	05/11/2007	20	Hivernale	7,6	2	31,1	N : 41%	NNW : 16,3%	NW : 4,2%	23,2
BF3	25/01/2008	08/02/2008	14	Hivernale	2,9	11,9	52,5	N : 12,5%	SSE : 6%	S : 5,4%	17,8
CA1	15/02/2009	05/03/2009	18	Hivernale	4,6	6	49,5	N :19,5%	SE : 11,7%	NNW : 11,7%	8,0
CA2	09/09/2009	23/09/2009	14	Estivale	15,6	11	47,4	NNW : 26,3%	N : 17,8%	E : 2,35%	9,0
CA3	02/03/2010	16/03/2010	14	Hivernale	1,2	2,5	36,5	N : 30,4%	NNW : 24,9%	NW : 4%	25,6
CA4	05/07/2010	16/07/2010	11	Estivale	23,2	28,7	54,9	N : 19,3%	SE : 8,3%	NNW : 3,3%	7,3

- **Données qualitatives (seront traitées dans le TP sur AFD)**

Vous disposez aussi de 4 variables qualitative que sont : TYPE : environnement du site soit urbain, industriel, rural, le site de compostage ; la SAISON: (hiver - été) ; Campagne : (4 campagnes en été et 2 campagne hivernale) ; Localisation : un label de point.

Vous devez choisir une stratégie de traitement de données pour essayer de répondre aux question posées.

Etape 1) Pour les données réelles : matrice X(n×p) dans R^p

Vous disposez de données qui sont dans des ordres de grandeurs différentes (des concentrations (ng/m3)): pour l'ACP vous devez au moins centrer vos données ; si vous voulez vous affranchir du problème des échelles vous devez réduire vos données (pour visualiser sur le cercle de corrélation mais pas seulement… il est recommandé de centrer et de réduire les données initiales).
Les données disponibles présentent des données manquantes,soit : en remplaçant quelques valeurs même si vous biaiserez votre approche, mais vous conservez les variables les plus échantillonnées que possibles ; soit, vous pouvez aussi prendre une approche complémentaire en faisant une ACP sur des données avec moins de variables mais complètes et comparer .Vous disposez de données pour chaque campagne de mesure d'un échantillon de mesure effectuées en ces n= nombre des points localisés autour du site et p variables = nombre de types de composés mesurés :
- un individu Xi. est un vecteur ligne, Xij une mesure des j=1 à p composés, en un point donné pour une période de temps donnée. chaque nouvelle campagne de mesure effectuée au niveau de la même station de mesure est un re échantillonnage en ce point : il constitue un nouvel individu si l'on décide que cela ne constitue pas une redondance d'échantillonnage : n = nombre de campagne × nombre de points de mesures individus au total

Vous devez **successivement de réaliser les étapes suivantes :**

1) **Le traitement statistique des données**

Il permettra d'évaluer la variabilité :
- sur l'ensemble de l'échantillon
- sur les 6 campagnes différentes
- pour les campagnes de mesure en hiver et les campagnes en été
- sur les deux campagnes avant ouverture du site (BF) et après ouverture du site (CA)

L'affichage des corrélations possibles entre les différentes variables : pour chacune des 4 périodes (hiver, été, avant activité, après activité)
Des traitements statistiques que pouvez-vous en déduire sur les différents périodes hiver/été et avant et après ouvertures du site ?

2) **Réduction de dimension et ACP**

- On cherche à savoir si l'on peut identifier une réduction du nombre de variables par ACP: soit rechercher de composantes principales et voir si les individus se regroupent ou pas selon ce nouvel espace R (q<p avec p=14). Vous devez mettre en œuvre l'ACP en justifiant les résultats (Inertie expliquée/inertie totale, qualité de la réduction de dimension et de la qualité des projections de individus sur les 14 variables quantitatives. Vous pouvez avancer des éléments de réponse sur les points suivants :

- Quelles variables sont les mieux expliquées ?
- Quelles variables sont regroupées dans les différentes composantes ?
- Les individus sont-ils bien projetés ?
- Pouvez donner un sens aux composantes ?

- On recherche une signature des composants pour chaque période (été, hiver, 1 avant_activité, 1 apres_activité) ; une réduction du nombre de variables par ACP serait-elle une méthode adaptée pour tenter de répondre ? Compte tenu de vos résultats de cette étape : quels sont vos principaux constats, quelles propositions de traitements faites-vous pour chaque période, quelles sont les variables marqueurs ?

(Ce cas d'étude sera repris pour le dernier TP en AFD)