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

Analyse en Composantes Principales – ***séance du 11 octobre 1,5h (fin séance du 12/10)***

Cette partie est consacrée à la méthode de l'ACP est composée de 1.5 h de TD puis de TP.
Cette séance se poursuivra par une seconde partie de TP, la semaine suivante après le cours sur complément de l'ACP.
Les rendus de séances du 11 octobre et du 12 octobre constitueront le rendu noté du cours *Analyse de données.*

Vous devez déposer votre TP_ACP par binôme) sur campus : le(s) script(s) développé(s) + le compte rendu (sous campus) en respectant la date limite qui vous est donnée

***Date du* 25/10 /2023**: rendu {1 TD + TP ACP1.1 }

Date du 9/11/2023 Rendu final du {TP ACP1.1 +TPACP1.2}

Partie TD : 1h00

1. Question 1

Nous avons vu en cours que la droite de projection orthogonale doit minimiser la somme des carrés des distances $M_i m_i$, $\frac{1}{n}\sum_{i=1}^{n} M_i m^2{}_i$ ou M_i est le point i dans l'espace d'origine et m_i la projection orthogonale de M_i sur la droite D pour assurer que la distorsion du nuage soit minimale.

Retrouver pourquoi cette droite D doit passer par le point moyen du nuage.

Quelques pistes de résolution

- *Soit deux droites parallèles D et D' : **D** passe par le point moyen **m** et **D'** passe par un point quelconque m' : que vaut $\|M_i - m_i'\|^2$ pour tout point M_i , le projeté orthogonal sur D en m_i' et sur D en m_i, soit pour la somme des carrés des distances $M_i m_i$*
- *A partir du calcul de l'inertie d'une variable statistique par rapport à un point a, définie comme étant la moyenne du carré des distances de la variable X au point a*
- *A partir de la décomposition de la variance pour toute var. aléatoire réelle X, théorème …*

2. Questions 2 : Etude la projection associée à l'ACP en 2D

Dans le cas d'un nuage à 2 dimensions, avec les variables *X* et *Y*, l'ACP nous assure qu'il existe donc deux directions principales données par le vecteur u_1 et le vecteur u_2 qui assurent maximiser l'inertie statistique $Is(u_1)$ et $Is(u_2)$ correspondant aux 2 valeurs propres λ_1 et λ_2 de la matrice de variance-covariance *A*. A chaque valeur propre λ_1 et λ_2 il existe un vecteur propre u_1 et u_2 non nul vérifiant que $A\,u_1 = \lambda_1$ et $A\,u_2 = \lambda_2$.

2.1 Le noyau de l'endomorphisme, soit l'application linéaire dans $\mathbb{R}^2$ dans $\mathbb{R}^{2e}$, est défini par la matrice $A - \lambda\, I_2$. Retrouver les conditions pour lesquelles l'application linéaire du noyau de

l'endomorphisme ne se réduise pas à 0 (*soit encore, que l'application linéaire ne soit pas injective donc bijective*).

2.2 Quelle est l'équation du deuxième degré dont les valeurs propres λ_1 et λ_2 doivent être solution. Donner sa formulation.

2.3 Calculer les deux valeurs propres λ_1 et λ_2 à partir de l'équation question 2.2 . Donner sa formulation.

2.4 Soit (u_1, u_2) une base orthonormée pour le produit scalaire canonique formé pas les vecteurs propres de la matrice *A* , on a $A\,u_1 = \lambda_1$ et $A\,u_2 = \lambda_2$. Pourquoi le vecteur propre *u* défini par $\begin{pmatrix} S^2(Y) - \lambda \\ -Cov(X,Y) \end{pmatrix}$ est un vecteur propre relatif à la valeur propre λ de la matrice *A*. Retrouver la formulation associée.

2.5 Calculer alors la norme de ce vecteur *u*.

2.6 Quel est le vecteur normé relatif à la valeur propre λ , soit le vecteur u

2.7 Donner la formule associée à chacun des vecteurs u_1 et u_2 pour leurs valeurs propres respectives λ_1 et λ_2 (on pourra également vérifier que leur produit scalaire est nul) . Donner la matrice V des coordonnées des vecteurs u_1 et u_2

2.8 Le vecteur propre $\begin{pmatrix} S^2(Y) - \lambda \\ -Cov(X,Y) \end{pmatrix}$ pris pour la valeur propre λ de la matrice A est-il le seul vecteur propre possible ?

Partie TP1.1 : 1h30

3. Mise en œuvre sur un échantillon de données : axe principal d'un nuage de point et projection dans le nouvel espace

Soit le nuage de 6 points de $\mathbb{R}^2$ suivant :

0	1	0	1	1	0
0	1	1	0	1	0

Chaque ligne du tableau correspond aux 2 var. :

- première ligne : var. X

- deuxième ligne var. Y .(penser à transposer ce tableau de données)

Chaque point est représenté par ses coordonnées : $M_1(0,0)$, $M_2(1,1)$ ….

3.1 Statistiques nécessaires à l'ACP

- Ecrire la matrice de données *X*
- Comme rien n'est précisé quel est le poids statistique de chaque point *i* et donner la matrice diagonale D des poids des individus
- Calculer la moyenne de chacune des var. X et Y (si possible de façon matricielle)
- Centrer les données pour obtenir la matrice de données centrées Z de dimension (6,2)
- Calculer la matrice de variance-covariances, que valent $S^2(X)$ et $S^2(Y)$ et cov(X,Y)

3.2 Calculer les valeurs propres λ_1 et λ_2 . Vérifier que la somme des valeurs propres de la matrice de covariance *A* est la trace de la matrice *A* et sachant que le déterminant de la matrice A est égal au produit des valeurs propres $\lambda_1 \times \lambda_2$.

3.3 A partir des éléments du TD , donner l'expression et la valeur des 2 vecteurs propres u_1 et u_2 correspondant aux 2 valeurs propres λ_1 et λ_2 .

3.4 Vérifier que $I_s(u) = {}^t u\, A\, u = \lambda\ {}^t u\, u\ = \lambda \|u\|^2 = \lambda$ pour un axe principal de droite **D** de vecteur propre u pour l'inertie statistique expliquée par les 2 axes principaux.

3.5 Vérifier que la somme de l'inertie totale du nuage est à la somme de l'inertie statistique portée par chacun des 2 axes principaux (trace de la matrice de covariance) : $I_T = I_s(u_1) + I_s(u_2)$

3.6 Faire le graphique des 2 axes principaux D1 et D2.

3.7 Calculer le taux d'inertie expliqué par chaque composante

4. Coordonnées factorielles et composantes principales

Calculer les coordonnées factorielles de chacun des 8 individus, soit les 8 vecteurs $\overrightarrow{GM_i}$, dans la base propre orthonormée {u1,u2} à partir de la matrice des coordonnées *V* des vecteurs propres. Pour chaque point *i* le projeté de Gm_i est défini par le produit scalaire entre $\overrightarrow{GM_i}$, $et\ u$, soit les coordonnées factorielles du nuage de points dans $\mathbb{R}^2$ muni de la base {u1,u2} :

$$\left(\langle \overrightarrow{GM_i} | u_1 \rangle\ \langle \overrightarrow{GM_i} | u_2 \rangle\right) = {}^t \begin{pmatrix} \langle \overrightarrow{GM_i} | u_1 \rangle \\ \langle \overrightarrow{GM_i} | u_2 \rangle \end{pmatrix} = {}^t \left({}^t V \right) \begin{pmatrix} x_{i0} \\ y_{i0} \end{pmatrix}) = (x_{i0}\ \ y_{i0})\, V \ \ pour\ tout\ i \in \{1, .. n\}$$

5.Calculer la qualité des projections individus

6. Calculer la contribution des individus à une composante

7. Quels sont les points les mieux projetés et ceux qui contribuer le plus à une composante.

Partie TP : (qui sera utilisée et étendue pour la seconde partie du TP (séance 12/10 : 1h30h TP)

1. A partir de ces étapes 3.1 a 3.6 , 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.