

Exercice 1 : ACP sur la matrice des distances

On observe un p -vecteur aléatoire quantitatif sur n individus, soit \mathbf{X}_i , $i = 1, \dots, n$ la i -ème observations de dimension p . On note $X_{i,k}$ la k -ème composante du vecteur \mathbf{X}_i , $k = 1, \dots, p$. Notons $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)'$ la matrice $(n \times p)$ des observations. On suppose que les composantes de \mathbf{X}_i sont centrées. Soit \mathbf{D} la matrice diagonale des poids $p_i = 1/n$. On munit \mathbb{R}^p d'une métrique \mathbf{M} ; $\|x\|_{\mathbf{M}}^2 = x'\mathbf{M}x$, $x \in \mathbb{R}^p$. Soit $\mathcal{D} = (d_{ij}^2)_{i,j=1,\dots,n}$ la matrice $n \times n$ des carrés des distances entre les n individus (d_{ij} est la distance entre \mathbf{X}_i et \mathbf{X}_j , $d_{ii} = 0$) :

$$d_{ij}^2 = (\mathbf{X}_i - \mathbf{X}_j)'\mathbf{M}(\mathbf{X}_i - \mathbf{X}_j) = \|\mathbf{X}_i - \mathbf{X}_j\|_{\mathbf{M}}^2$$

Posons

$$d_{i.}^2 = \sum_{j=1}^n p_j d_{ij}^2, \quad d_{.j}^2 = \sum_{i=1}^n p_i d_{ij}^2, \quad d_{..}^2 = \sum_{i=1}^n p_i d_{i.}^2.$$

Soit $I_g = \sum_{i=1}^n p_i \|\mathbf{X}_i\|_{\mathbf{M}}^2$, dit inertie du nuage de points des observations.

1. Montrer que la matrice de variance-covariance empirique des \mathbf{X}_i est $\mathbf{S} = \mathbf{X}'\mathbf{D}\mathbf{X}$.
2. Montrer que $\forall i = 1, \dots, n$, $d_{i.}^2 = \|\mathbf{X}_i\|_{\mathbf{M}}^2 + I_g$.
3. En déduire que $d_{..}^2 = 2I_g$
4. Posons $\mathbf{W} = (w_{ij} = \langle \mathbf{X}_i, \mathbf{X}_j \rangle_{\mathbf{M}} = \mathbf{X}_i'\mathbf{M}\mathbf{X}_j)_{i,j}$ la matrice des produits scalaires, montrer que

$$w_{ij} = -\frac{(d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d_{..}^2)}{2}$$

5. Exprimer \mathbf{W} en fonction de \mathcal{D}
6. Supposons dans la suite que $\mathbf{M} = \mathbf{I}_p$ et que l'ACP du nuage de \mathbf{X} donne p axes principaux normés $(u_k)_{k=1,\dots,p}$ de valeurs propres correspondants λ_k . Notons v_k les composantes principales associées. Montrer que $\mathbf{X}\mathbf{S}u_k = \lambda_k v_k$. Que peut-on en déduire ?
7. Montrer que, toujours si $\mathbf{M} = \mathbf{I}_p$, v_k est également vecteur propre de $\mathbf{W}\mathbf{D}$.
8. Soit le vecteur $f_k \in \mathbb{R}^n$ dont la composante numéro i est $f_{ik} = \sqrt{p_i} v_{ik}$. En déduire que la matrice $\mathbf{W}\mathbf{D}$ admet pour vecteur propre f_k avec valeur propre associé à λ_k .
9. Montrer que le vecteur $(\sqrt{p_i})_{i=1,\dots,n}$ est vecteur propre de $\mathbf{W}\mathbf{D}$ associé à la valeur propre 0.
10. Montrer que $\sum_{i=1}^n f_{ik}^2 = \lambda_k$ et pour tout $k \neq l$, $\sum_{i=1}^n f_{ik} f_{il} = 0$.
11. Application sous Python : Soit un nuage de points de 3 individus tel que

$$d_{12}^2 = d_{23}^2 = 1, \quad d_{13}^2 = 2, \quad p_i = 1/3, \quad i = 1, \dots, 3$$

Déterminer $\mathbf{W}\mathbf{D}$, les valeurs propres λ_k et les vecteurs propres f_k associés.

Exercice 2 : Application sur données réelles avec Python

On considère 11 pôles de dépenses d'un Etat (répartitions des dépenses en pourcentages) entre plusieurs années successives. On note X la matrice des données dont les pôles de dépenses (en colonne) : PVP : pouvoirs publics ; AGR : agriculture ; CMI : commerce et industrie ; TRA : travail ; LOG : logement et aménagement du territoire ; EDU : éducation ; ACS : action sociale ; ACO : anciens combattants ; DEF : défense ; DET : dette ; DIV : divers.

1. Effectuer sur ces données une Analyse en Composantes Principales.
2. Combien d'axes retiendriez-vous pour cette analyse ? pourquoi ?
3. Donner une interprétation globale des dépenses sur les axes retenus.

Exercice 3 : Convergence des vecteurs propres d'une matrice de variance-covariance empirique

Soit $\mathbf{X}_1, \dots, \mathbf{X}_n$ des p -vecteurs aléatoires Gaussiens i.i.d. d'espérances nulles et de matrices de variance-covariance $\Sigma = \mathbf{I}_p$. Soit $\hat{\mathbf{S}}$ la matrice de variance-covariance empirique des $\hat{\mathbf{X}}_i$, $i = 1, \dots, n$ et sa décomposition spectrale $\hat{\mathbf{S}} = \hat{\boldsymbol{\beta}} \hat{\boldsymbol{\Lambda}} \hat{\boldsymbol{\beta}}'$. Ici, $\hat{\boldsymbol{\Lambda}}$ est une matrice diagonale dont les p éléments diagonaux sont bien ordonnés et $\hat{\boldsymbol{\beta}} \in \mathcal{SO}_p$. Nous rappelons que par le théorème central limite multivarié, nous avons que $\sqrt{n}(\hat{\mathbf{S}} - \mathbf{I}_p) = O_P(1)$.

1. Montrer que $\sqrt{n}(\hat{\boldsymbol{\Lambda}} - \mathbf{I}_p) = O_P(1)$ pour $n \rightarrow \infty$.
2. Montrer que pour toute matrice $\boldsymbol{\Theta} \in \mathcal{SO}_p$, $\boldsymbol{\Theta} \sqrt{n}(\hat{\mathbf{S}} - \mathbf{I}_p) \boldsymbol{\Theta}' \stackrel{\mathcal{D}}{=} \sqrt{n}(\hat{\mathbf{S}} - \mathbf{I}_p)$.
3. Montrer qu'il existe $\tilde{\boldsymbol{\Theta}} \in \mathcal{SO}_p$ tel que $\tilde{\boldsymbol{\Theta}} \sqrt{n}(\hat{\boldsymbol{\Lambda}} - \mathbf{I}_p) \tilde{\boldsymbol{\Theta}}'$ et $\sqrt{n}(\hat{\boldsymbol{\Lambda}} - \mathbf{I}_p)$ ne convergent pas vers la même distribution lorsque $n \rightarrow \infty$.

Indice : Rappelez-vous que les valeurs propres de $\hat{\boldsymbol{\Lambda}}$ sont *ordonnées*.

4. Montrer que les trois premières questions impliquent qu'il n'existe pas de matrice $\boldsymbol{\beta}$ telle que $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + o_P(1)$ pour $n \rightarrow \infty$.
5. En considérant le résultat obtenu au point 4, est-il pertinent de donner une interprétation du type de celle donnée dans l'exercice 2.3 aux différents vecteurs propres contenus dans $\hat{\boldsymbol{\beta}}$?