ANALYSE DES DONNEES (Partiel)

Master M1 MMD - MA, 13 mars 2017

Calculatrice autorisée, documents autorisés : 2 feuilles recto-verso.

Barême approximatif : 10 points pour chacun des deux exercices.

Exercice 1

On considère deux nuages de points $\mathcal{N}_1$ et $\mathcal{N}_2$ dans $\mathbb{R}^p$, $p \geq 1$, muni de la métrique identité I_p.

Le nuage $\mathcal{N}_1$ est constitué de n individus $(x_i)_{1\leq i\leq n}$, chaque individu a un poids p_i avec $p_i > 0$ pour tout i et $\displaystyle\sum_{i=1}^{n} p_i = 1$.

De même le nuage $\mathcal{N}_2$ est constitué de $N - n$ individus $(x_i)_{n+1\leq i\leq N}$, chaque individu a un poids q_i avec $q_i > 0$ pour tout i et $\displaystyle\sum_{i=n+1}^{N} q_i = 1$.

Les résultats de l'ACP de $\mathcal{N}_1$, l'espace $\mathbb{R}^p$ étant muni de la métrique identité I_p, sont notés comme suit :

- le α ème vecteur axial factoriel (pour le nuage des individus) est noté u_α,
- la α ème valeur propre est notée λ_α,
- la α ème composante principale du nuage des individus est notée ψ_α,

avec $\alpha \in \{1, \ldots, p\}$.

Les résultats de l'ACP de $\mathcal{N}_2$, l'espace $\mathbb{R}^p$ étant muni de la métrique identité I_p, sont notés comme suit :

- le α ème vecteur axial factoriel (pour le nuage des individus) est noté v_α,
- la α ème valeur propre est notée μ_α,
- la α ème composante principale du nuage des individus est notée ϕ_α,

avec $\alpha \in \{1, \ldots, p\}$.

On note $\mathcal{N}$ l'union des deux nuages, constitué de N individus $(x_i)_{1\leq i\leq N}$, chaque individu a un poids r_i défini par

$$\forall 1 \leq i \leq N, \quad r_i = \begin{cases} p_i/2 & \text{si } 1 \leq i \leq n, \\ q_i/2 & \text{si } n+1 \leq i \leq N. \end{cases}$$

On note G_1 le centre de gravité de $\mathcal{N}_1$, G_2 de $\mathcal{N}_2$ et G de $\mathcal{N}$.

1. **0,5 point** Vérifier que $r_i > 0$ et que $\displaystyle\sum_{i=1}^{N} r_i = 1$.

   La vérification est immédiate.

2. **0,5 point** Comment se situe G par rapport à G_1 et G_2 ?

   On a
   $$G = \sum_{i=1}^{n} \frac{p_i x_i}{2} + \sum_{i=n+1}^{N} \frac{q_i x_i}{2} = \frac{1}{2}(G_1 + G_2).$$

   Donc G est au milieu de (G_1, G_2). 0,25 point seulement

3. **2 points** Exprimer l'inertie totale de $\mathcal{N}$ en fonction de l'inertie totale de $\mathcal{N}_1$ notée $I_T(\mathcal{N}_1)$, de $\mathcal{N}_2$ notée $I_T(\mathcal{N}_2)$ et de $d(G_1, G_2)^2 = \|\overrightarrow{G_1G_2}\|^2$.
   On a

$$\begin{aligned}
I_T(\mathcal{N}) &= \sum_{i=1}^{n} \frac{p_i\|x_i - G\|^2}{2} + \sum_{i=n+1}^{N} \frac{q_i\|x_i - G\|^2}{2}, . \text{ 0,5 point} \\
&= \frac{1}{2}\left(I_T(\mathcal{N}_1) + \|G_1 - G\|^2 + \sum_{i=1}^{n} p_i\langle x_i - G_1, G_1 - G\rangle\right) + \\
&\quad \frac{1}{2}\left(I_T(\mathcal{N}_2) + \|G_2 - G\|^2 + \sum_{i=n+1}^{N} q_i\langle x_i - G_2, G_2 - G\rangle\right), . \text{ 1 point} \\
&= \frac{1}{2}I_T(\mathcal{N}_1) + \frac{1}{2}I_T(\mathcal{N}_1) + \frac{1}{4}d(G_1, G_2)^2. \text{ 0,5 point}
\end{aligned}$$

4. On suppose que $G_1 \neq G_2$ on pose

$$u = \frac{\overrightarrow{G_1G_2}}{\|\overrightarrow{G_1G_2}\|}.$$

   On note V_1, V_2 et V les matrices de variance des nuages $\mathcal{N}_1$, $\mathcal{N}_2$ et $\mathcal{N}$.

   (a) **3 points** Montrer que

$$V = \frac{1}{2}(V_1 + V_2) + \frac{\|\overrightarrow{G_1G_2}\|^2}{4}uu'.$$

   On a 0,5 point

$$V = \sum_{i=1}^{n} \frac{p_i}{2}(x_i - G)(x_i - G)' + \sum_{i=n+1}^{N} \frac{q_i}{2}(x_i - G)(x_i - G)',$$

   Or pour tout $1 \leq i \leq n$, on a 0,5 point

$$\begin{aligned}
(x_i - G)(x_i - G)' &= (x_i - G_1 + G_1 - G)(x_i - G_1 + G_1 - G)', \\
&= (x_i - G_1)(x_i - G_1)' + (x_i - G_1)(G_1 - G)' + \\
&\quad (G_1 - G)(x_i - G_1)' + (G_1 - G)(G_1 - G)'.
\end{aligned}$$

   Puisque

$$\sum_{i=1}^{n} p_i(x_i - G_1) = 0,$$

   on en déduit que 0,5 point

$$\sum_{i=1}^{n} \frac{p_i}{2}(x_i - G)(x_i - G)' = \frac{1}{2}(V_1 + (G_1 - G)(G_1 - G)').$$

   De même 1 point

$$\sum_{i=n+1}^{N} \frac{q_i}{2}(x_i - G)(x_i - G)' = \frac{1}{2}(V2 + (G_2 - G)(G_2 - G)').$$

Par ailleurs, G étant au milieu de (G_1, G_2), on a 0,5 point

$$(G_1 - G)(G_1 - G)' = (G_2 - G)(G_2 - G)' = \frac{||\overrightarrow{G_1G_2}||^2}{4}uu'$$

Par conséquent

$$V = \frac{1}{2}(V_1 + V_2) + \frac{||\overrightarrow{G_1G_2}||^2}{4}uu'.$$

On fait l'hypothèse, dans les deux dernières questions, que pour tout $1 \leq \alpha \leq p, \quad u_\alpha = v_\alpha$**.**

(b) On suppose que $u = u_1 = v_1$

i. **1 point** Montrer que u est un axe factoriel de $\mathcal{N}$. Est ce le premier axe factoriel ?
On a 0,5 point

$$Vu = (\frac{\lambda_1 + \mu_1}{2} + \frac{||\overrightarrow{G_1G_2}||^2}{4})u.$$

Donc u est un vecteur propre unitaire de V, c'est un axe factoriel. Soit $\alpha > 1$, on a 0,5 point

$$Vu_\alpha = \frac{\lambda_\alpha + \mu_\alpha}{2} \leq \frac{\lambda_1 + \mu_1}{2} < \frac{\lambda_1 + \mu_1}{2} + \frac{||\overrightarrow{G_1G_2}||^2}{4}.$$

On en déduit que u est le premier axe factoriel.

ii. **1 point** Exprimer la composante principale H associée à l'axe factoriel u en fonction de ψ_1 et de ϕ_1.
On obtient

$$\forall 1 \leq i \leq N, \quad H_i = \begin{cases} \phi_1 - \frac{d(G_1,G_2)}{2} & \text{si } 1 \leq i \leq n, \\ \psi_1 + \frac{d(G_1,G_2)}{2} & \text{si } n+1 \leq i \leq N. \end{cases}$$

(c) On suppose que $u \in \ker V_1 \cap \ker V_2$.

i. **1 point** Montrer que u est un axe factoriel de $\mathcal{N}$ et déterminer une condition sur $d(G_1, G_2)$ pour que u soit le premier axe factoriel de $\mathcal{N}$.
On a 0,5 point

$$Vu = \frac{||\overrightarrow{G_1G_2}||^2}{4}u.$$

Soit $\alpha \geq 1$, on a

$$Vu_\alpha = \frac{\lambda_\alpha + \mu_\alpha}{2}u_\alpha.$$

Donc 0,5 point si $\frac{||\overrightarrow{G_1G_2}||^2}{4}) \geq \frac{\lambda_1+\mu_1}{2}$, alors u est le premier axe factoriel de $\mathcal{N}$

ii. **1 point** Montrer que $P_u(\mathcal{N})$, où P_u est la projection orthogonale sur la droite passant par G de vecteur directeur u, se réduit à deux points. Calculer l'inertie totale du nuage $P_u(\mathcal{N})$,
Puisque $u \in \ker V_1$, la projection de $\mathcal{N}_1$ sur la droite de vecteur directeur u, se réduit au point G_1. Donc $P_u(\mathcal{N})$ se réduit aux points G_1 et G_2. 0,5 point On a 0,5 point

$$I_T(P_u(\mathcal{N}) = \frac{1}{2}(d(G, G_1)^2 + d(G, G_2)^2 = \frac{||\overrightarrow{G_1G_2}||^2}{4}.$$

Exercice 2

On considère le tableau de données, noté X, qui est défini par :

$$X = \begin{pmatrix} 2 & 1 & -1 & 3 & -3 \\ 0 & 3 & -1 & 2 & -2 \\ 1 & 2 & -1 & 1 & -1 \end{pmatrix}$$

où la $i^{\text{ème}}$ ligne désigne l'individu x_i et la $j^{\text{ème}}$ colonne désigne la variable x^j.
Chaque individu possède un poids égal à 1/3. On considère les résultats de l'ACP du tableau X lorsque $\mathbb{R}^5$ est muni de la métrique identité.

1. **0,5 point** Déterminer le tableau centré Y.
   On a
   $$Y = \begin{pmatrix} 1 & -1 & 0 & 1 & -1 \\ -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}$$
2. **1 point** Sans calculer la matrice variance V du tableau X, combien existe-t-il d'axes factoriels non triviaux ?
   Le rang de Y est 2 puisque la colonne 1 est l'opposé de la colonne 2, la colonne 4 est l'opposé de la colonne 5 et la colonne 3 est nulle. Enfin la première et la cinquième colonne ne sont pas colinéaires. Donc le rang de Y est 2 donc celui de V aussi. Il y a donc deux axes factoriels non triviaux.
3. **1 point** Calculer V. En déduire l'inertie totale du nuage étudié.
   On a 0,5 point
   $$V = \frac{1}{3}Y'Y = \frac{1}{3}\begin{pmatrix} 2 & -2 & 0 & 1 & -1 \\ -2 & 2 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 2 & -2 \\ -1 & 1 & 0 & -2 & 2 \end{pmatrix}$$
   L'inertie totale du nuage étudié est la trace de V soit $8/3$. 0,5 point
4. On pose $u = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 1 \\ -1 \end{pmatrix}$.
   (a) **0,5 point** Vérifier que u est un vecteur propre de V associé à la valeur propre 2.
   On a
   $$Vu = 2u.$$
   (b) **1 point** En déduire les valeurs propres de V.
   Puisqu'il n'y a que deux axes factoriels non triviaux, il y a trois valeurs propres nulle, 2 et une dernière valeur propre. 0,5 point
   Or la trace vaut $8/3$ donc la dernière valeur propre est $2/3$ 1 point

(c) **2 points** Déterminer, pour chacun des axes factoriels non triviaux, l'unique vecteur axial factoriel qui le dirige et dont la première coordonnée est positive.

1 point Pour la valeur propre 2, on a $u_1 = \frac{1}{2}u$ et pour la valeur propre $2/3$, on a $u_2 = \frac{1}{2}\begin{pmatrix} 1 \\ -1 \\ 0 \\ -1 \\ 1 \end{pmatrix}$.

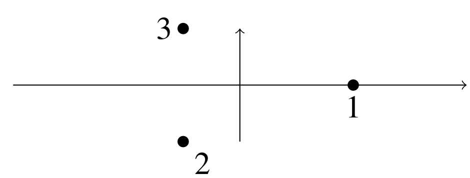
1 point Pour obtenir u_2, on cherche un vecteur unitaire orthogonal à u ainsi qu'à une base du noyau de V ou de Y soit $(\begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix})$.

Enlever la moitié des points si oubli de normaliser.

5. **1 point** Calculer les deux premières composantes principales , notée Ψ_1 et Ψ_2.

   On a
$$\Psi_1 = Yu_1 = \begin{pmatrix} 2 \\ -1 \\ -1 \end{pmatrix} \text{ et } \Psi_2 = Yu_2 = \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}.$$

6. **1 point** Représenter les trois individus dans le plan factoriel constitué des deux premiers axes.



7. **1 point** Calculer la contribution relative de chaque individu à l'inertie du premier axe.

   La contribution de l'individu 1 à l'axe 1 est $2/3$, à l'axe 2 est 0.
   La contribution de l'individu 2 à l'axe 1 est $1/6$, à l'axe 2 est $1/2$.
   La contribution de l'individu 3 à l'axe 1 est $1/6$, à l'axe 2 est $1/2$.

8. **1 point** Représenter les 5 variables dans le plan factoriel constitué des deux premiers axes.