Fiche 2 - AD - M1 MAS

Tran Viet Chi, `chi.tran@math.univ-lille1.fr`, bureau 316 (bâtiment M3).

Exercice 1 (Calcul des axes principaux et des facteurs principaux, sur un exemple simple)
On considère :

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

1. Caculer la matrice d'inertie du nuage des individus. Quelle est l'inertie totale du nuage ?

2. Donner les axes principaux et les facteurs principaux.

3. Quelle est la droite qui maximise l'inertie expliquée ? Quelle est la part d'inertie expliquée par cette droite ?

4. Quelles sont les contributions de chaque individu ? Que peut-on dire de la qualité de leur représentation ?

Exercice 2 (Théorème d'Huygens)
On considère une population de six individus caractérisés par trois variables :

$$X = \begin{pmatrix} 8 & 1 & 0 \\ 4 & 6 & 6 \\ 6 & 8 & 7 \\ 10 & 4 & 7 \\ 8 & 2 & 5 \\ 0 & 3 & 6 \end{pmatrix}.$$

On muni l'ensemble de ces individus de la partition suivante : $P = \{\{1,2\},\{3,4\},\{5,6\}\}$. Chaque individu a le poids $1/6$.

1. Calculer la matrice d'inertie et l'inertie totale du nuage de ces 6 individus.

2. Donner les centres de gravité dans chacune des 3 classes.

3. Pour $a \subset \{1,\dots 6\}$, on définit l'inertie du sous-nuage composé des individus dans a par

$$I(a) = \sum_{i\in a} \frac{1}{6}\|x_i - g(a)\|^2,$$

où $g(a)$ est le centre de gravité de la classe a. Quelle est la part de l'inertie totale qui est due au sous-ensemble a ? L'exprimer en fonction de $I(a)$.

4. Soit $s \in \mathbb{R}^3$. Appliquer le théorème de Huygens pour calculer

$$\sum_{i\in a} \frac{1}{6}\|x_i - s\|^2.$$

5. Montrer que si on pose $m_a = \mathrm{Card}(a)/6$,

$$I(a) = \frac{m_a}{36} \sum_{i,i'\in a} \|x_i - x_{i'}\|^2.$$

6. Soient a et $b \subset \{1,\dots 6\}$ tels que $a \cap b = \emptyset$. Montrer que

$$\frac{1}{36}\sum_{i\in a}\sum_{i'\in b} \|x_i - x_{i'}\|^2 = m_a I(b) + m_b I(a) + m_a m_b \|g(a) - g(b)\|^2.$$

7. On définit l'inertie inter entre a et b comme

$$I(a,b) = I(a \cup b) - I(a) - I(b).$$

Montrer que

$$\begin{aligned} I(a,b) =& m_a \|g(a) - g(a \cup b)\|^2 + m_b \|g(b) - g(a \cup b)\|^2 \\ =& \frac{m_a m_b}{m_a + m_b} \|g(a) - g(b)\|^2. \end{aligned}$$

8. On définit par $D = \text{diag}\big(1/6\big)$ la matrice des poids des individus, et on introduit :

$$\begin{aligned} V =& \sum_{i=1}^{6} \frac{1}{6}(x_i - g)(x_i - g)^T \\ W =& \sum_{a \in P} \sum_{i \in a} \frac{1}{6}(x_i - g(a))(x_i - g(a))^T \\ B =& \sum_{a \in P} m_a (g(a) - g)(g(a) - g)^T. \end{aligned}$$

Montrer que $I = \text{tr}(VD)$.

10. On définit les inerties intra et inter par

$$I_{intra}(P) = \text{tr}(WD), \qquad I_{inter}(P) = \text{tr}(BD).$$

Quelle est la relation entre I_{intra} et les quantités $I(a)$ de questions précédentes ? Idem pour $I_{inter}(P)$ et les quantités $I(a,b)$ des questions précédentes ?

11. Calculer les inerties intra et inter pour le tableau donné et la partition P. Vérifier que $I = I_{intra}(P) + I_{inter}(P)$.

12. Rappeler la définition des variances intra et inter classes, pour la partition P, et les écrire en fonction des matrices V, B et W.

Exercice 3 (ACP sur les températures en France)
On dispose de données (issues de Escofier et Pagès 1998) sur les températures mensuelles moyennes de 15 grandes villes françaises entre 1931 et 1960 (voir annexe). On dispose également des latitudes, longitudes de ces villes. L'analyse factorielle de ce tableau a été réalisée avec SAS (procédure princomp, complétée par des descriptions - proc means et proc corr).

1. Commenter le tableau : décrire le nuage des points, qu'est-ce qui se dégage de l'observation des données 'brutes' ?

2. Examiner la matrice de corrélation, qu'on notera V.

3. Les valeurs propres et vecteurs propres de V ont été calculés. Commenter.

4. Quels sont les deux premiers facteurs principaux ? Pour les interpréter, relever, pour les individus et les variables, ceux qui contribuent le plus à l'inertie de ces axes et ceux qui sont le mieux représentés, en indiquant bien s'ils sont du côté positif ou du côté négatif de l'axe principal.

5. A l'aide des variables supplémentaires, confirmer l'interprétation des facteurs principaux.