TD : Dimension

EXERCICE 1 :

① On cherche $Y = \sum_{k=1}^{4} \alpha_k X_k$ tel que $Var(Y) = 6.95$.

Par définition des valeurs propres, on sait qu'elles sont égales à la variance de combinaisons linéaires successives.

On remarque $\lambda_3 = 6.95$. Donc en choisissant $\alpha = \alpha_3$, on obtient une combinaison linéaire de $X_1, \dots, X_4$ de variance 6.95.

Et donc $Y = 0.718 X_1 - 0.017 X_2 + 0.077 X_3 - 0.691 X_4$.

et $Var(Y) = 6.95$

② On a $Z = a_0 + \sum_{k=1}^{4} a_k X_k$.

Par définition de la première valeur propre λ_1, λ_1 est égale à la variance maximale des combinaisons linéaires de $X_1, \dots, X_4$. Donc $Var(Z) = \lambda_1 = 16.27$.

Et $a = (a_1, \dots, a_4)^T = \alpha_4$.

Pour trouver a_0, on utilise $\mathbb{E}[Z] = 0$

$$\mathbb{E}[Z] = 0 \iff a_0 + \sum_{k=1}^{4} a_k \mathbb{E}[X_k] = 0$$

$$\iff a_0 + \alpha_1^T \mu = 0$$

$$\iff a_0 + 0.098 \times 1 + (-0.98) \times (-1) = 0$$

$$\iff a_0 = -1.078$$

③ On cherche $M = \begin{bmatrix} b_1 & & & 0 \\ & b_2 & & \\ & & b_3 & \\ 0 & & & b_4 \end{bmatrix}$ tel que $M \Sigma M = R$

où R est la matrice des corrélations de X.

On a $\Sigma = \begin{bmatrix} Var(X_1) & - & Cov(X_1,X_4) \\ | & \diagdown & | \\ Cov(X_1,X_4) & - & Var(X_4) \end{bmatrix}$ et $R = \begin{bmatrix} 1 & - & Corr(X_1,X_4) \\ | & \diagdown & | \\ Corr(X_1,X_4) & - & 1 \end{bmatrix}$

Or $\forall i,j \in \{1,\dots,4\}$, $Corr(X_i,X_j) = \dfrac{Cov(X_i,X_j)}{\sqrt{Var(X_i)Var(X_j)}}$.

On peut donc prendre $M = \begin{bmatrix} Var(X_1) & & 0 \\ & \ddots & \\ 0 & & Var(X_4) \end{bmatrix}^{-\frac{1}{2}}$

EXERCICE 2 :

Soit $X = (X_1, \dots, X_p)^T$

On centre $Z = X - \bar{X} = (X_1 - \bar{X}_1, \dots, X_p - \bar{X}_p)^T$

On réduit $Z_{CR} = D^{-1} Z$ où $D = diag(\sigma_{X_1}, \dots, \sigma_{X_p})$

avec $\forall k \in \{1,\dots,p\}$, σ_{X_k} écart-type de X_k.

La matrice de covariance des données centrées et réduites est

$$\begin{aligned} \Sigma_{Z_{CR}} &= \frac{1}{n} Z_{CR}^{T} Z_{CR} \\ &= \frac{1}{n} (D^{-1}Z)^T (D^{-1}Z) \\ &= \frac{1}{n} Z^T D^{-1} D^{-1} Z \quad \text{(car } D \text{ est diagonale)} \\ &= D^{-1} \left(\frac{1}{n} Z^T Z \right) D^{-1} \end{aligned}$$

Or $\frac{1}{n} Z^T Z$ est la matrice de variance-covariance de X (données centrées) : $\Sigma_{Z_{CR}} = D^{-1} \Sigma_X D^{-1}$

La matrice de corrélation R est donnée par

$$R_{ij} = \frac{Cov(X_i, X_j)}{\sigma_{X_i}\,\sigma_{X_j}} \quad \text{donc} \quad R = D^{-1}\Sigma_X D^{-1} \text{ car}$$

$$Cov(X_i - \bar{X}, X_j - \bar{X}) = Cov(X_i, X_j)$$

Finalement, $\Sigma_{ZCR} = R$.

EXERCICE 3 :

① Par construction, $Var(C)$ est maximale si $C = \alpha^T X$ où α est le vecteur propre associé à la première (plus grande) valeur propre de Σ. Donc $\alpha = v_1$.

Et donc $C = -\frac{1}{\sqrt{2}} X_2 + \frac{1}{\sqrt{2}} X_3 \quad (c_1 = 0)$.

On vérifie que $\sum_{j=1}^{3} c_j^2 = 1 \quad \left(0^2 + \left(\frac{1}{\sqrt{2}}\right)^2 + \left(\frac{1}{\sqrt{2}}\right)^2 = 1\right)$

D'après la théorie de l'ACP, $Var(C) = \lambda_1 = 3$.

② La proportion de variance expliquée par la première composante, et donc par C est

$$\frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} = \frac{3}{3+2+1} = 0.5 \quad (50\,\%)$$

③ La décomposition spectrale de Σ est $\Sigma = VDV^T$ où $V = [v_1 | v_2 | v_3]$ et $D = \begin{bmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \lambda_3 \end{bmatrix} = \begin{bmatrix} 3 & & 0 \\ & 2 & \\ 0 & & 1 \end{bmatrix}$

On a donc $\Sigma = \begin{bmatrix} 0 & 1 & 0 \\ -1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 1/\sqrt{2} & 0 & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 3 & & 0 \\ & 2 & \\ 0 & & 1 \end{bmatrix} \begin{bmatrix} 0 & -1/\sqrt{2} & 1/\sqrt{2} \\ 1 & 0 & 0 \\ 0 & 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$

EXERCICE 4 :

① La somme des éléments d'une même ligne indique le nombre de patients survivants (ligne 1) ou décédés (ligne 2) et la somme des éléments d'une même colonne indique le nombre de patients selon l'importance des troubles cognitifs en phase initiale. Il s'agit donc d'un tableau de contingence et l'Analyse Factorielle des Correspondances (AFC) est la méthode la plus adaptée.

② Les éléments du tableau des fréquences relatives sont donnés par $f_{ij} = \frac{k_{ij}}{k_{..}}$. Ce qui donne le tableau suivant :

	0	1	2	Total
0 - survie	0.55	0.31	0.07	0.93
1 - décès	0.04	0.02	0.01	0.07
Total	0.59	0.33	0.08	1

③ Le profil moyen des lignes correspond à la répartition des patients selon la présence de troubles cognitifs.

	0	1	2	Total
Profil-moyen	59	33	8	100

④ Le profil moyen des colonnes correspond à la répartition des patients selon l'issue observée, survie ou décès.

	Profil-moyen
0-survie	93
1-décès	7
Total	100

⑤ En AFC, on obtient $\min(n,p)-1$ axes factoriels non triviaux où n désigne le nombre de lignes et p désigne le nombre de colonnes du tableau de contingence.
Ici, $n=2$ et $p=3$. On obtiendra donc 1 axe.

⑥. Le test du χ^2 permet de détecter une éventuelle indépendance entre les lignes et les colonnes. On a $I_G = \frac{S}{k_{..}}$
où S est la statistique de test et $k_{..}$ est l'effectif total du tableau. Donc $S = I_G\, k_{oo} = 0.0050 \times 196 = 0.98$.

EXERCICE 5

①. a. Oui après codage en disjonctif complet des var. qualitatives.
b. Oui après codage en disjonctif complet des var. ordinales mais pas idéale car pas de prise en compte de l'ordre.

c. Non, on n'a pas les réponses individuelles selon l'aggrégation.

d. Oui, après découpage en classe des var. quantitatives
et codage en disjonctif complet.

2. a. Oui, par définition.

b. Non, autant de colonnes que de modalités.

c. Non, autant de colonnes que de modalités.

d. Oui (cf. b)

3. a. Oui
b. Oui
} par définition des distances.

c. Non
d. Oui
} la distance utilisée en ACM est pondérée par la fréquence des modalités.

4. a. Oui
b. Oui
} par définition des distances.

c. Non
d. Non
} si les obs qui ont les modalités sont les mêmes, les modalités auront une distance faible.

5. a. Oui
b. Oui
} c'est comme ça que le biplot est construit.

c. Non, pas nécessairement.

d. Oui
e. Oui
} par construction.