1er étape : Construction d'un dendrogramme pour une CAH

2ème étape : nous continuer cette étape itérative encore.

2ème étape : Construction d'un dendrogramme jusqu'à ce qu'il ne reste plus et unique groupe réunissant tous les individus.

3ème étape : Construct° dendrogramme : Fin de l'algorithme de clusterings.

Application (Voir page précédente).

Application :

Partie 0

On compte faire l'analyse en composante principale (ACP) de la matrice X composée de 6 individus $\{I_1, I_2, I_3, I_4, I_5, I_6\}$ et 3 variables $\{X_1 ; X_2 ; X_3\}$ est donné comme suite

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

A partir de la matrice X :

1°) Calculons le centre de gravité de X noté G_x

2°) Deduire Y la matrice des données centrées

3) Calculons le centre de gravité de Y, noté G_y

4) Donner la matrice des poids noté D

5) Calculer V la matrice Variance Covariance de X

6°) Calculer l'inertie total de 2 façons differents.

Résolution

1°) Calculons le centre de gravité de X, noté G_x

On a $G_x(\overline{X^1}, \overline{X^2}, \overline{X^3})$

Soit $X(6,3)$ la matrice des données initiales

- Nombre d'individus $n = 6$
- Le compteur des individus $i \{1 \dots 6\}$
- Nombre de variable $P = 3$
- le compteur des variables $j \{1, \dots 3\}$

P_i : le poids de chaque individus $p_i = \frac{1}{6}$

Calcul des moyennes de X

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

$$\overline{X}^1 = \frac{1}{6}\sum_{i=1}^{6} x_i^1 = \frac{1}{6}(2+1+0+2+1+0) = \frac{6}{6} =$$

$$\overline{X}^1 = 1$$

$$\overline{X}^2 = \frac{1}{6}\sum_{i=1}^{6} x_i^2 = \frac{1}{6}(1+1+0+0+0+1) = \frac{3}{6} = \frac{1}{2}$$

$$\overline{X}^3 = \frac{1}{6}\sum_{i=1}^{6} x_i^3 = \frac{1}{6}(0+0+\cancel{0}1+1+1+0) = \frac{3}{6} = \frac{1}{2}$$

Centre de Gravité de X

$$G(\overline{X}_1, \overline{X}_2, \overline{X}_3) = \left(1; \frac{1}{2}; \frac{1}{2}\right)$$

② Déduisons Y la matrice des données centrée et son centre de gravité G_Y

Soit Y la matrice des données centrée

$$y_i^j = x_i^j - \overline{x}^j$$

$$Y = \begin{pmatrix} (2-1) & (1-\frac{1}{2}) & (0-\frac{1}{2}) \\ (1-1) & (1-\frac{1}{2}) & (0-\frac{1}{2}) \\ (0-1) & (0-\frac{1}{2}) & (1-\frac{1}{2}) \\ (2-1) & (2-\frac{1}{2}) & (1-\frac{1}{2}) \\ (1-1) & (1-\frac{1}{2}) & (1-\frac{1}{2}) \\ (0-1) & (0-\frac{1}{2}) & (0-\frac{1}{2}) \end{pmatrix}$$

$$Y = \frac{1}{2}\begin{pmatrix} 2 & 1 & -1 \\ 0 & 1 & -1 \\ -2 & -1 & 1 \\ 2 & -1 & 1 \\ 0 & -1 & 1 \\ -2 & 1 & -1 \end{pmatrix}$$

Calcul des moyennes de Y

$$\overline{y}^1 = \frac{1}{6}\sum y_i^1 = \frac{1}{6}(2+0-2+2+0-2) = 0$$

$$\bar{y}_2 = \frac{1}{6}\sum_{i=1}^{6} y_i^2 = \frac{1}{6}(1+1-1-1-1+1) = 0$$

$$\bar{y}^3 = \frac{1}{6}\sum_{i=1}^{6} y_i^3 = \frac{1}{6}(-1-1+1+1+1-1) = 0$$

Centre de gravité de Y

$$G_y = (\bar{y}^1, \bar{y}^2, \bar{y}^3) = (0; 0; 0)$$

4°) La matrice des poids D

$$D = \begin{pmatrix} p_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & p_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & p_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & p_4 & 0 & 0 \\ 0 & 0 & 0 & 0 & p_5 & 0 \\ 0 & 0 & 0 & 0 & 0 & p_6 \end{pmatrix}$$

$$= \begin{pmatrix} 1/6 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/6 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/6 & \cancel{1/6} & 0 & 0 \\ 0 & 0 & 0 & \cancel{1/6} & 1/6 & 0 \\ 0 & 0 & 0 & 0 & \cancel{1/6} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/6 \end{pmatrix}$$

$$D = \frac{1}{6} I_6$$

5) Calculons la matrice des Variance-Covariance

Soit V la matrice Variance-Covariance noté V.

$$V = Y^t \times D\, Y = Y^t \left(\frac{1}{6}\right) \times Y = \frac{1}{6} Y^t Y$$

$$V = \frac{1}{6}\left(\frac{1}{2}\right) \times \begin{pmatrix} 2 & 0 & -2 & 2 & 0 & -2 \\ 1 & 1 & -1 & -1 & -1 & 1 \\ -1 & -1 & 1 & 1 & 1 & 1 \end{pmatrix} \times$$

$$\left(\frac{1}{2}\right) \begin{pmatrix} 2 & 1 & -1 \\ 0 & 1 & -1 \\ -2 & -1 & 1 \\ 2 & -1 & 1 \\ 0 & -1 & 1 \\ -2 & 1 & -1 \end{pmatrix}$$

$$V = \frac{1}{24} \begin{pmatrix} 16 & 0 & 0 \\ 0 & 6 & -6 \\ 0 & -6 & 6 \end{pmatrix} = \frac{1}{12} \begin{pmatrix} 8 & 0 & 0 \\ 0 & 3 & -3 \\ 0 & -3 & 3 \end{pmatrix}$$

5°) C 12 la matrice des correlation

Con peut chercher d'autre parametre comme

$V(X_1) = \frac{8}{12}$ $\sigma(X_1) = \sqrt{\frac{8}{12}}$

$V(X_2) = \frac{3}{12}$ $\sigma(X_2) = \sqrt{\frac{1}{4}} = \frac{1}{2}$

$V(X_3) = \frac{3}{12} = \frac{1}{4}$ et $\sigma(X_2) = \sqrt{\frac{1}{4}} = \frac{1}{2}$

6°) Calcul de l'inertie total de deux methode differentes.

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

1er Methode

Par definition l'inertie total c'est la somme pondéré des differentes individus autours du centre de gravité

$$I = \sum_{i=1}^{6} P_i \, d^2(i_x, G_x) = \sum_{i=1}^{6} P_i \, d^2(i_y, G_y)$$

$$= \sum_{i=1}^{6} P_i \|i_y\|^2 = \sum_{i=1}^{6} P_i (G_y, G_y)$$

$$I = \frac{1}{6} \sum_{i=1}^{6} \|i_y\|^2 = \left(\frac{1}{6}\right)\left(\frac{1}{4}\left[(4+1+1)\right] + (0+1+1) + (4+1+1)\right) (0+1+1) + (4+1+1)$$

$$I = \frac{6+2+6+6+2+6}{24}$$

$$I = \frac{28}{24} = \frac{7}{6}$$

$$\boxed{I = \frac{7}{6}}$$

$$Y = \frac{1}{2}\begin{pmatrix} 2 & 2 & -1 \\ 0 & 1 & -1 \\ -2 & -1 & 1 \\ -2 & -1 & 1 \\ 2 & -1 & 1 \\ 0 & -1 & 1 \\ -2 & 1 & 1 \end{pmatrix}$$

2eme Methode

L'inertie total (A partir de X) est la somme des variance des variables X_i

$$V = \frac{1}{12}\begin{pmatrix} 8 & 0 & 0 \\ 0 & +3 & -3 \\ 0 & -3 & 3 \end{pmatrix}$$

V est une matrice symetrique semi defini positivement.

$V = P * D * P^{-1}$

$\mathrm{Trace}(V) = \mathrm{Trace}(D)$

$$I = \mathrm{Trace}(V) = V_{11} + V_{22} + V_{33} = \frac{1}{12}(8+3+3)$$

$$I = \frac{14}{12} = \frac{7}{6}$$

L'inertie total c'est la quantité d'information totale qu'on cherche à conservé, situé sur la diagonal de la matrice V.
L'inertie total peut être calculé à partir de valeur propre (somme de valeurs propre)

$$I = \sum_{i=1}^{n} \lambda_i$$

V et D sont semblable et Trace(V) = Trace(D)

$$V = P * D * P^{-1} \quad \text{avec } D = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}$$

Diagonalisation de V

$$V = P \times D \times P^{-1} \quad I = \sum V(\bar{X}) = \text{Trace}(D)$$

$$= \sum \lambda_i$$

D : matrice des valeurs propre
P : matrice des vecteur propre

7) Calculons la matrice de corrélation et interpreter les corrélations plus élevé

Soit R la matrice de corrélation

$$R = Z^t \times D \times Z = Z^t \times \left(\frac{1}{6}\right) \times Z = \frac{1}{6} Z^t Z$$

où Z la matrice des donnée centrée

$$Z_i^j = \frac{x_i^j - \bar{x}^j}{\sigma(x^j)} = \frac{y_i^j}{\sigma(x^j)}$$

$$V = \frac{1}{12}\begin{pmatrix} 8 & 0 & 0 \\ 0 & 3 & -3 \\ 0 & -3 & 3 \end{pmatrix}$$

$V(x_1) = \frac{8}{12}$; $V(x_2) = \frac{3}{12}$ $V(x_3) = \frac{3}{12}$

$\sigma(x_1) = \sqrt{\frac{8}{12}} = 0{,}816$; $\sigma(x_2) = \sqrt{\frac{3}{12}} = 0{,}5$

$\sigma(x_3) = \sqrt{\frac{3}{12}} = 0{,}5$

$$Y = \begin{pmatrix} 1 & 0{,}5 & -0{,}5 \\ 0 & 0{,}5 & -0{,}5 \\ -1 & 0{,}5 & 0{,}5 \\ 1 & 0{,}5 & 0{,}5 \\ 0 & 0{,}5 & 0{,}5 \\ -1 & 0{,}5 & -0{,}5 \end{pmatrix} ; \quad Z = \begin{pmatrix} \frac{1}{0{,}816} & (0{,}5/0{,}5) & -(0{,}5/0{,}5) \\ 1/0{,}816 & (0{,}5/0{,}5) & 0 \\ -1/0{,}816 & (0{,}5/0{,}5) & \\ 1/0{,}816 & (0{,}5/0{,}5) & \\ 0 & (0{,}5/0{,}5) & \\ -1/0{,}816 & (0{,}5/0{,}5) & \end{pmatrix}$$

$$Z = \begin{pmatrix} \frac{1}{0{,}816} & \frac{0{,}5}{0{,}5} & \frac{0{,}5}{0{,}5} \\ 0 & \frac{0{,}5}{0{,}5} & -\frac{0{,}5}{0{,}5} \\ -\frac{1}{0{,}816} & \frac{0{,}5}{0{,}5} & \frac{0{,}5}{0{,}5} \\ \frac{1}{0{,}816} & -\frac{0{,}5}{0{,}5} & \frac{0{,}5}{0{,}5} \\ 0 & -\frac{0{,}5}{0{,}5} & \frac{0{,}5}{0{,}5} \\ -\frac{1}{0{,}86} & \frac{0{,}5}{0{,}5} & -\frac{0{,}5}{0{,}5} \end{pmatrix}$$

$$Z = \begin{pmatrix} 1{,}22 & 1 & -1 \\ 0 & 1 & -1 \\ -1{,}22 & 1 & 1 \\ 1{,}22 & -1 & 1 \\ 0 & -1 & 1 \\ -1{,}22 & 1 & 1 \end{pmatrix}$$

$$R = Z^t \times D \times Z = Z^t \left(\frac{1}{6}\right) Z = \frac{1}{6} Z^t \times Z$$

$$R = \frac{1}{6} \begin{pmatrix} 1{,}22 & 0 & -1{,}22 & 1{,}22 & 0 & -1{,}22 \\ 1 & 1 & -1 & -1 & -1 & 1 \\ -1 & -1 & 1 & 1 & 1 & 1 \end{pmatrix} \times$$

$$X = \begin{pmatrix} 1,22 & 1 & -1 \\ 0 & 1 & -1 \\ -1,22 & -1 & 1 \\ 1,22 & -1 & 1 \\ 0 & -1 & 1 \\ -1,22 & 1 & 1 \end{pmatrix}$$

$$R = \frac{1}{6}\begin{pmatrix} 6 & 0 & 0 \\ 0 & 6 & -6 \\ 0 & -6 & 6 \end{pmatrix} = \begin{array}{c} X_1 \\ X_2 \\ X_3 \end{array}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{pmatrix}$$

$r(X_1, X_2) = r(X_1, X_3) = 0$

$\Rightarrow X_1$ est independant de X_2 et X_3

$r(X_2, X_3) = -1 \Rightarrow X_2$ et X_3 sont fortement corrélés negativement

Partie II

1) Choix entre ACP normé et ACP non normée
2) Normalisation de V
3) polynome caracteristique de X
4) ~~Calcul de l'inertie à partir des valeurs propres~~
4) Valeurs propres
5) Calcul de l'inertie à partir des valeurs propres.
6) Fiabilité d'une ACP
7) Choix du nombre des axes à retenir pour une ACP

__Exemple :__

Soit V. la matrice Variance-Covariance

$$V = \frac{1}{12}\begin{pmatrix} 8 & 0 & 0 \\ 0 & 3 & -3 \\ 0 & -3 & 3 \end{pmatrix} \qquad V(X_1) = \frac{8}{12}$$

$$V(X_2) = V(X_3) = \frac{3}{12} \qquad \sigma(X_1) = \sqrt{\frac{8}{12}}$$

$$\sigma(X_2) = \sigma(X_3) = \sqrt{3/12} =$$

$$S = \frac{\max(\sigma(X_j))}{\max(\sigma(X_j))} = \frac{\sigma(X_1)}{\sigma(X_2)} = \sqrt{\frac{8/12}{3/12}}$$

$$= 2,62 < 5$$

Conclusion $S < 5 \rightarrow$ données homogène

On choisie une ACP non normée
La matrice variance covariance V sera notre forme quadratique normé.

Le polynôme caracterique de V

$$V = \frac{1}{12}\begin{pmatrix} 8 & 0 & 0 \\ 0 & 3 & -3 \\ 0 & -3 & 3 \end{pmatrix}$$

$$P_V(\lambda) = \det(V - \lambda I_3)$$

$$\det \frac{1}{12}\begin{pmatrix} 8-\lambda & 0 & 0 \\ 0 & 3-\lambda & -3 \\ 0 & -3 & 3-\lambda \end{pmatrix}$$

$\lambda_1 = 2/3$. $\lambda_2 = \frac{1}{2}$ $\lambda_3 = 0$

$$I = \sum_{j=1}^{3} \lambda_j = \frac{2}{3} + \frac{1}{2} + 0 = \frac{7}{6}$$

$$I = \sum_{j=1}^{p} V(\bar{x}) = \text{Trace}(D) = \sum_{j=1}^{p} \lambda_j = \frac{7}{6}$$

$V(\bar{X}_1) = \frac{2}{3}$ et $V(\bar{X}_2) = V(\bar{X}_3) = \frac{1}{4}$

Fiabilité de l'ACP

$(\lambda_1 / I) * 100 \longrightarrow C_1$

$(\lambda_2 / I) \times 100 \longrightarrow C_2$

$(\lambda_3 I) \times 100 \longrightarrow C_3$

Gain d'inertie

$$\text{Gain} = \left[\left(\lambda_1 + \lambda_2 \right) / I \right] \times 100$$

ACP fiable $\longrightarrow$ Gain d'intertie $> 75\%$

$\lambda_1 = 2/3$; $\lambda_2 = 1/2$; $\lambda_3 = 0$

$$\text{Gain d'info} = \left[(\lambda_1 + \lambda_2) / I \right] \times 100$$

$$= \left[\left(\frac{2}{3} + \frac{1}{2} \right) \times \frac{7}{6} \right] \times 100 = 100\%$$

Gain d'info = 100%

Conclusion Gain info > 75%

d'ou l'ACP est fiable (credible)

Choix du nombre des axe à retenir 3 criteres

1er critère : Critere de Kaiser :

$$\bar{\lambda} = \frac{1}{p}\sum_{j=1}^{p} \lambda_j = \frac{1}{3}(\lambda_1 + \lambda_2 + \lambda_3)$$

$$= \frac{1}{3}\left(\frac{2}{3} + \frac{1}{2} + 0\right) = \frac{7}{18} = 0{,}388$$

Si $\lambda_j > \bar{\lambda}$ alors cette axe est à retenir

$\lambda_1 = \frac{2}{3} > \bar{\lambda} = 0{,}378$

λ_1 $0{,}66 > 0{,}388 \Rightarrow$ l'axe est à retenir

$\lambda_2 = \frac{1}{2} = 0{,}5 > \bar{\lambda} = 0{,}388$

l'axe n°2 est à retenir

$\lambda_3 = 0 < \bar{\lambda} = 0{,}388$ est à rejeté

2ème critere : Critere de Coude