Aide-mémoire ACP

1 Analyse en composantes principales

Matrice de données :

$$X = \begin{pmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{nn} \end{pmatrix}$$

Etape 1 Centrage des données

Matrice centrée :

$$Z = \begin{pmatrix} X_{11} - \overline{X}_1 & \cdots & X_{1p} - \overline{X}_p \\ \vdots & \ddots & \vdots \\ X_{n1} - \overline{X}_1 & \cdots & X_{nn} - \overline{X}_p \end{pmatrix} \tag{1}$$

Remarques

- Les variables centrées ont une moyenne nulle
- Le centre de gravité des données centrée correspond à l'origine du repère

Etape 2 Réduction des données

Matrice centrée-réduite :

$$Z = \begin{pmatrix} \frac{X_{11} - \overline{X}_1}{\sigma_{X_1}} & \cdots & \frac{X_{1p} - \overline{X}_p}{\sigma_{X_p}} \\ \vdots & \ddots & \vdots \\ \frac{X_{n1} - \overline{X}_1}{\sigma_{X_1}} & \cdots & \frac{X_{nn} - \overline{X}_p}{\sigma_{X_p}} \end{pmatrix} \tag{2}$$

avec

$$\sigma^2_{X_j} = var\left(X_j\right) = \frac{1}{n}\sum_{i=1}^{n}\left(X_{ij} - \overline{x}_j\right)^2$$

Remarques

- L'écart type d'une variable centrée-réduite est égal à 1 (de même pour la variance)
- La réduction des données est obligatoire si les variables ne possèdent pas la même unité de mesure
- La réduction a pour objectif d'harmoniser le nuage de points : quand on divise par l'écart type, on compte en nombre d'écart type, du coup toutes les dispersions sont homogénéisées
- Si on a la même unité de mesure pour toutes les variables, le choix de réduire ou non les données est un choix de modèle :
  - si on ne réduit pas les données : une variable à forte variance va « tirer » tout l'effet de l'ACP à elle
  - si on réduit les données : une variable qui n'est qu'un bruit va se retrouver avec une variance apparente égale à une variable informative.

Etape 3 Calcul de la matrice de variance-covariance

La matrice de variance-covariance s'obtient avec la formule

$$\boxed{C = \frac{1}{n} Z^T Z} \tag{3}$$

– Si les données sont centrées (Z est calculée avec la formule 1), la matrice C correspond à la matrice de ***covariance*** :

$$C = \begin{pmatrix} var\left(X_1\right) & cov\left(X_1, X_2\right) & \cdots & cov\left(X_1, X_p\right) \\ cov\left(X_2, X_1\right) & \ddots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ cov\left(X_1, X_p\right) & \cdots & \cdots & var\left(X_p\right) \end{pmatrix}$$

avec

$$\begin{aligned} var\left(X\right) &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \\ cov\left(X, Y\right) &= \frac{1}{n}\sum_{i=1}^{n}\left(Xi - \overline{X}\right)\left(Y_i - \overline{Y}\right) \end{aligned}$$

– Si les données sont en plus réduites (Z est calculée avec la formule 2), la matrice C correspond à la matrice de ***corrélation*** :

$$C = \begin{pmatrix} 1 & corr\left(X_1, X_2\right) & \cdots & corr\left(X_1, X_p\right) \\ corr\left(X_2, X_1\right) & \ddots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ corr\left(X_1, X_p\right) & \cdots & \cdots & 1 \end{pmatrix}$$

avec

$$corr\left(X, Y\right) = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 \sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}} = \frac{cov\left(X, Y\right)}{\sigma_X \sigma_Y}$$

Remarques

– Ces deux matrices sont carrées de taille $p \times p$ (avec p le nombre de variables), symétriques, et réelles. Elles sont donc diagonalisables dans une base orthonormée

Etape 4 Calcul des axes factoriels

Définition Un vecteur u de taille p est un vecteur propre d'une matrice A de taille $p \times p$ s'il existe $\lambda \in \mathbb{C}$ telle que

$$Au = \lambda u$$

λ est la variable propre de A associée à u.

Dans notre cas, la matrice C est symétrique, définie positive, ce qui garantie l'existence de p valeurs propres réelles.

1. Calcul des valeurs propres
   On a

$$Cu = \lambda u \Leftrightarrow \left(C - \lambda I\right) u = 0$$

Les valeurs propres de C sont alors les réels qui vérifient :

$$\det\left(C - \lambda I\right) = \begin{vmatrix} C_{11} - \lambda & C_{12} & \cdots & C_{1p} \\ C_{21} & C_{22} - \lambda & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ C_{p1} & \cdots & \cdots & C_{pp} - \lambda \end{vmatrix} = 0$$

Les valeurs propres λ sont alors les solutions de l'équation ainsi obtenue (dite Equation caractéristique). Cette équation est sous la forme d'un polynôme de degré p en λ. Les valeurs propres λ sont les racines de ce polynôme.

2. Calcul des vecteurs propres
   Un vecteur propre u associé à la valeur propre λ doit vérifier

$$Cu = \lambda u \quad \Leftrightarrow \quad (C - \lambda I)\, u = 0 \;\Leftrightarrow\; (C - \lambda I) \begin{pmatrix} u_1 \\ \vdots \\ u_p \end{pmatrix} = 0$$

   Ce qui donne un système de p équations à p inconnus. Les coordonnées x_i de C sont les solutions de ce système d'équations.

   Les vecteurs propres ainsi obtenus forment une base de $\mathbb{R}^p$ (orthogonaux deux à deux).

- Les valeurs propres λ_k, $k \in \{1, \dots p\}$ obtenues doivent être triées en ordre décroissant
- Les vecteurs propres unitaires u_k obtenus définissent les **axes factoriels** D_k associés aux valeurs propres λ_k. L'axe factoriel qui correspond à la meilleure valeur propre est nommé **axe principal**.
- Une valeur propre λ_k définit l'**inertie** des individus par rapport à l'axe factoriel D_k (c-à-d, la dispersion des projections des individus sur l'axe D_k).

Etape 4 Interprétation de l'ACP

Qualité de l'ACP

- Le **pourcentage d'inertie** expliqué par les h premiers axes factoriels est :

$$\frac{\sum_{j=1}^{h} \lambda_j}{\sum_{i=1}^{p} \lambda_i}$$

Composantes principales. Les projections des individus sur un axe factoriel D_k muni du vecteur unitaire u_k constitue la **composante principale** c^k. c^k est le vecteur de $\mathbb{R}^n$ qui donne les coordonnées des individus sur l'axe principal D_k.

Contribution des individus. La contribution de l'individu e_i à la composante c^k (à l'axe D_k) est définie par :

$$\boxed{\frac{\frac{1}{n}\left(c_i^k\right)^2}{\lambda_k}}$$

Cercle de corrélation. Il est obtenu en représentant les variables initiales par un vecteur dans les plans factoriels. Les coordonnées des variables sont les coefficients de corrélation de ces variables avec les composantes principales. Le cercle de corrélation permet d'étudier les corrélations entre les variables initiales et les composantes principales ainsi que la corrélation de ces variables les unes par rapport aux autres.

2 Rappels

2.1 Calcul du déterminant d'une matrice

2.1.1 Cas d'une matrice de dimension 2

Le déterminant de la matrice de dimension 2 : $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ est :

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

2.1.2 Cas d'une matrice de dimension > 2

Développement suivant une colonne : Soit une matrice carrée d'ordre n. On a, pour tout indice de colonne j fixé,

$$\det(A) = \sum_{i=1}^{n} a_{ij} (-1)^{i+j} \Delta_{ij}$$

avec Δ_{ij} est le déterminant de la matrice A dont on a supprimé la $i^{ème}$ ligne et la $j^{ème}$ colonne :

$$\Delta_{ij} = \begin{vmatrix} a_{1,1} & \cdots & a_{1,j-1} & a_{1,j+1} & \cdots & a_{1,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{i-1,1} & \cdots & a_{i-1,j-1} & a_{i-1,j+1} & \cdots & a_{i-1,n} \\ a_{i+1,1} & \cdots & a_{i+1,j-1} & a_{i+1,j+1} & \cdots & a_{i+1,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n,1} & \cdots & a_{n,j-1} & a_{n,j-+1} & \cdots & a_{n,n} \end{vmatrix}$$

Développement suivant une ligne : Soit une matrice carrée d'ordre n. On a, pour tout indice de ligne i fixé,

$$\det(A) = \sum_{j=1}^{n} a_{ij} (-1)^{i+j} \Delta_{ij}$$

2.1.3 Exemple

On veut calculer le déterminant de cette matrice :

$$\begin{vmatrix} -1 & 2 & 5 \\ 1 & 2 & 3 \\ -2 & 8 & 10 \end{vmatrix}$$

on peut procéder par exemple en développant sur la première ligne :

$$\begin{aligned} \begin{vmatrix} \mathbf{-1} & \mathbf{2} & \mathbf{5} \\ 1 & 2 & 3 \\ -2 & 8 & 10 \end{vmatrix} &= \mathbf{-1} \begin{vmatrix} 2 & 3 \\ 8 & 10 \end{vmatrix} \mathbf{- 2} \begin{vmatrix} 1 & 3 \\ -2 & 10 \end{vmatrix} \mathbf{+ 5} \begin{vmatrix} 1 & 2 \\ -2 & 8 \end{vmatrix} \\ &= -1(2 \times 10 - 3 \times 8) - 2(1 \times 10 - 3 \times (-2)) + 5(1 \times 8 - 2 \times (-2)) \\ &= -1(-4) - 2(16) + 5(12) \\ &= 4 - 32 + 60 = \mathbf{32} \end{aligned}$$

Ou bien en développement sur la deuxième colonne :

$$\begin{aligned} \begin{vmatrix} -1 & \mathbf{2} & 5 \\ 1 & \mathbf{2} & 3 \\ -2 & \mathbf{8} & 10 \end{vmatrix} &= \mathbf{-2} \begin{vmatrix} 1 & 3 \\ -2 & 10 \end{vmatrix} \mathbf{+ 2} \begin{vmatrix} -1 & 5 \\ -2 & 10 \end{vmatrix} \mathbf{- 8} \begin{vmatrix} -1 & 5 \\ 1 & 3 \end{vmatrix} \\ &= -2(1 \times 10 - 3 \times (-2)) + 2(-1 \times 10 - 5 \times (-2)) - 8(-1 \times 3 - 5 \times 1) \\ &= -2(16) + 0 - 8(-8) \\ &= -32 + 64 = \mathbf{32} \end{aligned}$$

2.2 Calcul des racines d'un polynôme de degré 2

Soit $ax^2 + bx + c$ un polynôme de degré 2. Le discriminant est $\Delta = b^2 - 4ac$

– Si $\Delta > 0$, alors le polynôme admet deux racines réelles :

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \text{ et } x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}$$

– Si $\Delta = 0$, alors le polynôme admet une racine double :

$$x_1 = x_2 = \frac{-b}{2a}$$