Chapitre 4 – Analyse d'un nuage de points

I Inertie d'un nuage de points

1 Nuage de points des individus

Système des variables : $(x^1, \ldots, x^p)$

Système des individus : $(x_1, \ldots, x_n)$
Chaque individu x_i est muni d'un poids p_i.

Nuage des individus :

$$\mathcal{N} = \left\{ (x_i, p_i) \mid x_i \in \mathbb{R}^p, p_i > 0 \text{ et } \sum_{i=1}^{n} p_i = 1 \right\}$$

Centre de gravité (ou barycentre) du nuage :

$$g = \sum_{i=1}^{n} p_i x_i$$

L'espace $\mathbb{R}^p$ est muni d'un produit scalaire grâce à une matrice Q symétrique et définie positive :

$$\begin{aligned} d^2(u,v) = (u \mid v)_{\mathrm{Q}} &= (\mathrm{U}-\mathrm{V})^t \mathrm{Q} (\mathrm{U}-\mathrm{V}) \\ &= \sum_{i=1}^{p} \sum_{j=1}^{p} (u_i - v_i) \mathrm{Q}_{ij} (u_j - v_j) \end{aligned}$$

Très souvent, Q est diagonale, voire égale à I_p.

2 Inertie du nuage de points par rapport à un point Y

Définition : $\mathrm{I}(\mathrm{Y}) = \sum_{i=1}^{n} p_i d^2(x_i, y)$

C'est une mesure de la dispersion du nuage autour du point considéré Y.

Théorème de Huygens :

$$\mathrm{I}(\mathrm{Y}) = \mathrm{I}(\mathrm{G}) + d^2(\mathrm{G}, \mathrm{Y})$$

L'inertie est donc minimale quand Y = G.

Matrice d'inertie :

$$\mathrm{V}(\mathrm{Y}) = (\mathcal{X} - \mathcal{Y})^t \mathrm{P} (\mathcal{X} - \mathcal{Y})$$

où P est la matrice (diagonale) des poids, $\mathcal{X}$ est la matrice des données (contenant chaque individu en ligne) et $\mathcal{Y}$ est la matrice contenant n fois le vecteur Y en ligne.

Théorème : $\mathrm{I}(\mathrm{Y}) = \mathrm{Tr}\left[\mathrm{V}(\mathrm{Y})\mathrm{Q}\right]$

(Preuve dans le cas Q diagonale uniquement.)

3 Inertie du nuage de points par rapport à une droite Δ

Définition : $\mathrm{I}(\Delta) = \sum_{i=1}^{n} p_i d^2(x_i, \Delta)$

où $d^2(x_i, \Delta) = d^2(x_i, \widehat{x_i})$ en notant $\widehat{x_i}$ le projeté orthogonal de x_i sur Δ.

4 Inertie du nuage de points par rapport à un plan Π

Définition : $\mathrm{I}(\Pi) = \sum_{i=1}^{n} p_i d^2(x_i, \Pi)$

où $d^2(x_i, \Pi) = d^2(x_i, \widehat{x_i})$ en notant $\widehat{x_i}$ le projeté orthogonal de x_i sur Π.

II Ajustement du nuage des individus dans l'espace des variables

Objectif : fournir une image approchée du nuage des individus dans un sous-espace de dimension strictement inférieure à p. (On se placera dans le cadre de l'ACP normée, où les données sont centrées réduites.)

1 Droite d'ajustement

On cherche une droite d_1 de vecteur directeur unitaire u_1 telle que l'inertie du nuage projeté sur cette droite soit maximale.

Théorème : si λ_1 est la plus grande valeur propre de $\mathrm{VQ} = \mathrm{V(G)Q}$ et u_1 un vecteur propre unitaire associé, alors l'axe d_1 défini par u_1 explique la plus forte inertie du nuage.

2 Plan d'ajustement

On cherche une droite d_2 de vecteur directeur unitaire u_2 telle que u_1 et u_2 sont Q-orthogonaux et que l'inertie du nuage projeté sur la droite d_2 soit maximale.

Théorème : si λ_2 est la deuxième plus grande valeur propre de VQ et u_2 un vecteur propre unitaire associé et Q-orthogonal à u_1, alors l'axe d_2 défini par u_2 explique la plus forte inertie du nuage.

3 Sous-espace d'ajustement

On construit ainsi la suite $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_q$ des valeurs propres de VQ, et les vecteurs propres unitaires associés $u_1, \ldots, u_q$ deux-à-deux Q-orthogonaux.

Inertie expliquée par les axes :
L'inertie totale est $\mathrm{I(O)} = \mathrm{Tr}\,(\mathrm{VQ}) = \lambda_1 + \ldots \lambda_q$.
L'inertie expliquée par l'axe d_j est $\mathrm{I}(d_j^{\perp}) = \lambda_j$.
En pourcentage, cela représente $\lambda_j / (\lambda_1 + \cdots + \lambda_q)$.

Coordonnées des individus :
La coordonnée de l'individu x_i sur l'axe d_j est obtenue par le produit scalaire $(x^i)^t \mathrm{QU}_j$.

Composantes principales :
Le produit $\mathcal{X}\mathrm{QU}_j$ donne le vecteur C_j contenant les coordonnées des individus sur d_j. C'est une nouvelle variable, appelée *composante principale*.

Propriétés :

1. Les composantes principales sont des vecteurs propres de la matrice $\mathcal{X}\ \mathrm{Q}\ \mathcal{X}^t\ \mathrm{P}$, associés aux mêmes valeurs propres $\lambda_1 \geqslant \ldots \geqslant \lambda_q$.
2. $\forall i \qquad \mathrm{Var}\ (\mathrm{C}_i) = \lambda_i$
3. $\forall i \neq j \qquad \mathrm{Cov}\ (\mathrm{C}_i, \mathrm{C}_j) = 0$
4. $\forall i \qquad \mathrm{C}_i$ est centrée et $\|\mathrm{C}_i\| = \sqrt{\lambda_i}$

Coordonnées des variables :
La coordonnée de la variable x^k (centrée réduite) sur la composante principale C_j (nouvelle variable) est obtenue par le produit scalaire

$$(\mathrm{D}_j)_k = \frac{1}{\|\mathrm{C}_j\|}\,(\mathrm{C}^j)^t\mathrm{P}x^k = \frac{1}{\sqrt{\lambda_j}}\mathrm{Cov}\ (\mathrm{C}_j, x^k)$$

C'est donc le coefficient de corrélation associé :

$$(\mathrm{D}_j)_k = \mathrm{Corr}(\mathrm{C}_j, x^k)$$

Facteurs principaux :
Le vecteur $\mathrm{D}_j = (\mathrm{Corr}(\mathrm{C}_j, x^k))_{k=1\ldots p}$ contient donc les coordonnées des variables sur le j^{e} axe factoriel d_j. C'est un nouvel individu, appelé *facteur principal*.

Qualités de représentation :
La qualité de représentation d'un élément k sur l'axe j (mesurée par le cosinus au carré) est égale au rapport entre l'inertie de la projection de l'élément sur l'axe et de l'inertie totale de l'élément, c'est-à-dire

$$\mathrm{qlt}_j(x_k) = \frac{\|\widehat{x}_k\|^2}{\|x_k\|^2} = \frac{((\mathrm{C}_j)_k)^2}{\sum\limits_{i=1}^{p}((\mathrm{C}_i)_k)^2}$$

et $\mathrm{qlt}_j(x^k) = \cos^2(\mathrm{C}_j, x^k) = \mathrm{Corr}(\mathrm{C}_j, x^k)^2 = ((\mathrm{D}_j)_k)^2$

III Ajustement du nuage des variables dans celui des individus

En effectuant la même analyse pour le nuage de variables, on obtient les relations suivantes, dites de conjugaison :

	Analyse des individus	Analyse des variables
Espace		
Métrique		
Poids		
Relation de diagonalisation		
Conditions sur la norme		
Axe principal		
Inertie de l'axe		
Inertie totale		
Composante principale		
Qualité de la représentation		
Propriétés et relations		