Séance 2: Modèle Euclidien

Analyse des individus

Sébastien Gadat

Laboratoire de Statistique et Probabilités
UMR 5583 CNRS-UPS

www.lsp.ups-tlse.fr/gadat

Deuxième partie II

Modèle Euclidien

Objectifs

- Analyse du tableau de données X
- Objectifs à préciser
- Possibilité d'analyse des variables colonnes
- Possibilité d'analyse des individus lignes
- Les variables et individus seront représentés dans des espaces euclidiens MULTIDIMENSIONNELS

Nuage des individus lignes

Nuage d'individus X qui synthétise les données :

$$X = \begin{pmatrix} x_1^1 & \dots & x_1^j & \dots & x_1^p \\ \vdots & \ddots & \dots & \dots & \vdots \\ x_i^1 & \dots & x_i^j & \ddots & x_i^p \\ \vdots & \dots & \ddots & \ddots & \vdots \\ x_n^1 & \dots & x_i^j & \dots & x_n^p \end{pmatrix}$$

Dans cet espace inclus dans $\mathbb{R}^p$, on s'attache à reproduire et à analyser graphiquement les distances entre individus

Nuage des variables colonnes

Nuage des variables à partir de X qui synthétise les données :

$$X = \begin{pmatrix} x_1^1 & \dots & x_1^j & \dots & x_1^p \\ \vdots & \ddots & \dots & \dots & \vdots \\ x_i^1 & \dots & x_i^j & \ddots & x_i^p \\ \vdots & \dots & \ddots & \ddots & \vdots \\ x_n^1 & \dots & x_i^j & \dots & x_n^p \end{pmatrix}$$

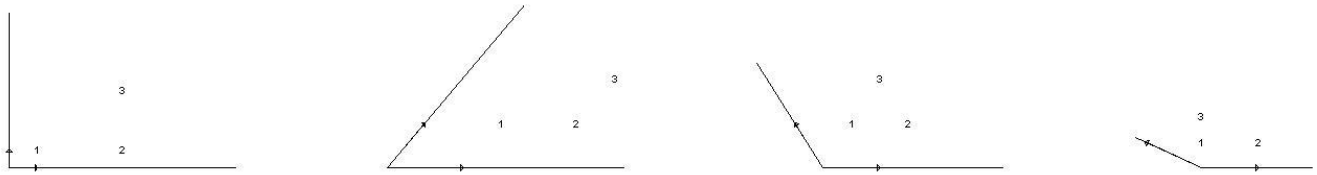
Dans cet espace inclus dans $\mathbb{R}^n$, on s'attache à représenter graphiquement et à analyser les indices entre variables

Rôle de la structure métrique (INDIVIDUS)

On donne le tableau :

$$X = \begin{pmatrix} 1 & 1 \\ 4 & 1 \\ 4 & 4 \end{pmatrix}$$

Selon les choix effectués pour la configuration géométrique des vecteurs de base de $\mathbb{R}^p$, les individus ligne de X sont représentés de façon différente :



Les positions sont dépendantes des choix de vecteurs de la base.

Cas général dans R^p (INDIVIDUS)

La configuration des vecteurs de base $(e_1, e_2, \ldots, e_p)$ est caractérisée par les éléments d'une matrice symétrique positive M. On a la relation

$$M = \begin{pmatrix} m_{1,1} & \ldots & m_{1,j} & \ldots & m_{1,p} \\ \vdots & \ddots & \ldots & \ldots & \vdots \\ m_{i,1} & \ldots & m_{i,j} & \ddots & m_{i,p} \\ \vdots & \ldots & \ddots & \ddots & \vdots \\ m_{p,1} & \ldots & m_{p,j} & \ldots & m_{p,p} \end{pmatrix}$$

avec les relations

$$m_{i,i} = \|e_i\|^2$$

$$m_{i,j} = \langle e_i, e_j \rangle = \|e_i\| \|e_j\| \cos(e_i, e_k)$$

Cas général dans R^p (INDIVIDUS)

Proposition M est diagonale si et seulement si les vecteurs de base $(e_j)_{j=1..p}$ sont orthogonaux pour la métrique euclidienne.
Étant donnés 2 points a et b de $\mathbb{R}^p$, on définit :

- Le produit scalaire entre a et b par

$$\langle a, b \rangle_M = a'Mb = b'Ma =$$

- La norme de a par

$$\|a\|_M^2 =$$

- Le carré de la distance entre a et b

$$d_M(a, b)^2 =$$

- Le cosinus de a et b

$$\cos_M(a, b) =$$

Cas général dans R^p (INDIVIDUS)

Cette matrice M définit dans $\mathbb{R}^p$ une nouvelle géométrie :

- Nouvelles distances
- Nouveaux angles
- Nouvelles isométries
- Nouvelles orthogonalité : deux points a et b seront M-orthogonaux si et seulement si :

$$\langle a, b \rangle_M = 0$$

- Décomposition M-ortogonale sur un espace W :

$$x = x_W + (x - x_W) \qquad \text{où} \qquad \forall u \in W \qquad \langle x - x_W, u \rangle_M = 0$$

Exemples de métriques M (INDIVIDUS)

Quelques métriques M utilisées souvent en analyse des données :

- Cas de variables quantitatives non hétérogènes :

$$M=$$

- Cas de variables quantitatives hétérogènes (Variables réduites)

$$M=$$

- Cas de variables quantitatives hétérogènes (Mahalanobis)

$$M=$$

- Cas de variables quantitatives hétérogènes (Joreskog)

$$M=$$

- Cas de variables qualitatives :
  - Effectifs marginaux de chacune des modalités : $y_{+j}=$
  - $M=$

Utilisation des poids sur les individus

L'interprétation géométrique des indices incite fortement à utiliser la métrique diagonale des poids $D = Diag(p_i)$.
La variance d'une variable colonne . . . du tableau de données peut s'interpréter comme le carré de la longueur d'un vecteur de composantes . . . pour la métrique de $\mathbb{R}^n$ pondérée par . . .

- Produit scalaire entre A et B :

$$\langle A, B \rangle_D =$$

- Longueur d'une variable A :

$$\|A\|_D^2 =$$

- Variance d'une variable A :

$$Var(A) =$$

Coefficient de corrélation multiple

Ce coefficient est utilisé pour quantifier la variabilité linéaire d'une variable x par rapport aux variables $x^1, \dots x^p$. Il est défini par

- $$R(x; \{x^1, \dots x^p\}) = \sup_{a_1, \dots a_p} \rho(x; \sum_{i=1}^{p} a_i x^i)$$
- R vaut 1 si et seulement si x est combinaison linéaire des x^i
- $R = 0$ implique x est non corrélée avec tous les x^i
- $0 \leq R \leq 1$
- Calcul de R : si A désigne la matrice de projection orthogonale sur l'espace engendré par les x^i, alors
$$R^2 = \frac{(x' - \bar{x})A(x' - \bar{x})}{\|x' - \bar{x}\|^2}$$
- Si x est centrée et X désigne la matrice des individus décrits par les p variables x^i, alors
$$R^2 = \frac{x'X(X'X)^{-1}X'x}{x'x}$$

Inertie du nuage de points

- Une variable descriptive x, n individus
- L'inertie des individus est définie par

$$I(x_i) = Var(x) =$$

- L'inertie est d'autant plus grande que le nuage de points est "étalé.

Inertie du nuage de points

- X nuage de points décrivant n individus et p variables
- Métrique M, matrice de taille $p \times p$
- g centre de gravité du nuage défini par

$$g =$$

- Inertie du nuage X donné par

$$I(X) = I(X, g) = \sum_{i=1}^{n} p_i d_M(x_i, g)^2$$

- L'inertie dépend :
  - de M
  - des poids p_i
  - de X

Illustration

Pour le nuage des individus lignes du tableau Y ci-dessous

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

Donner :

- l'inertie par rapport à l'origine
- l'inertie totale (par rapport au centre de gravité du nuage)

On étudiera les questions dans le cas de l'équipondération et dans le cas d'une pondération $(2, 1, 4, 1, 2)/10$.

Autres expressions de l'inertie

- On peut également calculer l'inertie du nuage de points par rapport à n'importe quel point de $\mathbb{R}^p$ *via* :

$$I(X, y) =$$

- On a également

$$I(X) = \frac{1}{2} \sum_{i,j} p_i p_j d_M(x_i, x_j)^2$$

- En formulation matricielle :

$$I(X) = Tr(MV) = Tr(VM)$$

-

$$I(X) = Tr(XMX'D) = Tr(DXMX')$$

Décomposition de l'inertie

- Si l'espace $\mathbb{R}^p$ se décompose en

$$x_i = x_W + x_{W^\perp}$$

alors

$$I(X) = I(X_{W_M}) + I(X_{W^{\perp_M}})$$

- Si on a une partition des individus en groupes $\mathcal{G}_k$, on définit
  - Poids P_k du groupe k
  - Centre de gravité g_k
  - Inertie du groupe k
- On a alors

$$I(X) =$$

]

Décomposition de l'inertie

- On peut définir l'inertie à partir des poids et distances inter points
- Si $\bar{D}^2$ désigne la moyenne des carrés des distances inter-points :

$$I_{totale} = \frac{1}{2}\bar{D}^2 =$$

- L'inertie intra classe vaut alors

$$I_{intra} = \frac{1}{2}\sum_k P_k \bar{D}_k^2$$

avec $\bar{D}_k^2$ moyenne des carrés des distances dans le groupe k

- L'inertie inter est définie à partir de la formule de reconstitution

$$I_{inter} =$$

- En regroupant deux classes C_1 et C_2 de poids P_1 et P_2 en une seule classe $C_1 \cup C_2$ de poids $P_1 + P_2$, le gain intra est :

$$\text{Gain} = \frac{P_1 P_2}{P_1 + P_2}\left(\bar{D}_{12}^2 - \frac{\bar{D}_1^2 + \bar{D}_2^2}{2}\right) = \frac{P_1 P_2}{P_1 + P_2}\|g_1 - g_2\|^2$$