SY19 : Apprentissage non supervisé

Gérard Govaert

Automne 2011

Table des matières

Notations

A^t matrice transposée

$|A|$ déterminant de la matrice A

$\text{diag}(A)$ vecteur colonne défini par la diagonale de A si A est une matrice carrée et matrice diagonale de diagonale A si A est un vecteur

d_p vecteur colonne de dimension n des pondérations p_i

D_p matrice diagonale de dimension n des pondérations p_i

$\mathbb{1}_n$ vecteur colonne de dimension n rempli de 1.

I_n matrice unité de dimension n

U_n matrice carrée de dimension n remplie de 1.

X matrice des données de dimension (n, p)

Chapitre 1

Quelques rappels

1 Algèbre linéaire

Dans ce paragraphe, les notions élémentaires d'algèbre linéaire (espace vectoriel sur un corps K, sous-espace vectoriel, combinaison linéaire de vecteurs, famille libre, famille liée, base, dimension...) seront supposées connues. Dans la suite tous les espaces vectoriels envisagés seront toujours définis sur le corps des réels.

Rappelons quelques notions :

- G et $\mathbf{a}$ étant respectivement un sous-espace vectoriel et un élément de E, l'ensemble $F = \mathbf{a} + G$ des éléments $\boldsymbol{x}$ de E tels que $\boldsymbol{x} = \mathbf{a} + \boldsymbol{y}$ avec $\boldsymbol{y} \in G$ est appelé *variété linéaire*. Le sous-espace vectoriel G est appelé direction de F et si G est de dimension r, F est une variété linéaire de dimension r.
- Une *application linéaire* d'un espace vectoriel E dans un espace vectoriel F est une application f de E dans F vérifiant les propriétés suivantes :

$$\begin{array}{ll} \forall \boldsymbol{x}, \boldsymbol{y} \in E & f(\boldsymbol{x}+\boldsymbol{y}) = f(\boldsymbol{x}) + f(\boldsymbol{y}), \\ \forall \boldsymbol{x} \in E \text{ et } \alpha \in \mathbb{R} & f(\alpha \boldsymbol{x}) = \alpha f(\boldsymbol{x}). \end{array}$$

  Un *endomorphisme* est une application linéaire de E dans E et une *forme linéaire* est une application linéaire de E dans $\mathbb{R}$.
- Un espace vectoriel E se décompose en une *somme directe* de sous-espaces vectoriels $E_1, \ldots, E_k$, notée $E = E1 \oplus \ldots \oplus E_k$, si et seulement si $\forall \boldsymbol{x} \in E$, $\boldsymbol{x}$ s'écrit de manière unique $\boldsymbol{x} = \boldsymbol{x}_1 + \ldots + \boldsymbol{x}_k$ avec $\boldsymbol{x}_i \in E_i$. Lorsque le nombre de sous-espaces de la somme directe se réduit à deux, on parle de *sous-espaces supplémentaires*.

1.1 Un exemple d'espace vectoriel : $\mathbb{R}^p$

Le produit cartésien $\mathbb{R}^p$, ensemble de tous les p-uples de réels, est un exemple d'espace vectoriel très souvent utilisé, en particulier en analyse des données. La dimension de cet espace est p et on peut montrer que tous les espaces vectoriels sur $\mathbb{R}$ de dimension p sont isomorphes à $\mathbb{R}^p$. Les éléments de $\mathbb{R}^p$ sont notés dans la suite sous la forme de « vecteurs colonnes » ou matrice de dimension $(p, 1)$:

$$\boldsymbol{x} = (x_1 \ldots, x_p)'$$

et la base canonique est formée des vecteurs

$$\mathbf{e}_1 = (1, 0, \ldots, 0)', \ldots, \mathbf{e}_p = (0, \ldots, 0, 1)'.$$

La décomposition de $\boldsymbol{x}$ s'écrit donc $\boldsymbol{x} = \sum_{i=1}^{p} x_i \mathbf{e}_i$ et il y a identité entre les coordonnées de $\boldsymbol{x}$ dans la base canonique et les composantes de $\boldsymbol{x}$, élément du produit cartésien $\mathbb{R}^p$. Ceci n'est vrai que pour la base canonique.

1.2 Produit scalaire, orthogonalité, norme et distance

Un *produit scalaire* sur un espace vectoriel E est une application $< ., . >$ de $E \times E$ dans $\mathbb{R}$ véridiant les propriétés suivantes :

$$\begin{array}{lll} \text{- bilinéaire,} & & \\ \text{- symétrique :} & \forall \boldsymbol{x}, \boldsymbol{y} \in E & < \boldsymbol{x}, \boldsymbol{y} >=< \boldsymbol{y}, \boldsymbol{x} >, \\ \text{- définie :} & \forall \boldsymbol{x}, \boldsymbol{y} \in E & < \boldsymbol{x}, \boldsymbol{x} >= 0 \Rightarrow \boldsymbol{x} = 0, \\ \text{- positive :} & \forall \boldsymbol{x}, \boldsymbol{y} \in E & < \boldsymbol{x}, \boldsymbol{x} >\geq 0. \end{array}$$

Le produit scalaire permet de définir la notion d'*othogonalité*.

- Deux vecteurs $\boldsymbol{x}$ et $\boldsymbol{y}$ de E sont *orthogonaux* si leur produit scalaire est nul

$$\boldsymbol{x} \perp \boldsymbol{y} \Longleftrightarrow < \boldsymbol{x}, \boldsymbol{y} >= 0.$$

- Deux sous-espaces vectoriels F et G sont *orthogonaux* si tous les éléments de l'un sont orthogonaux à tous les éléments de l'autre

$$F \perp G \Longleftrightarrow (\forall \boldsymbol{x} \in F, \forall \boldsymbol{y} \in G \quad \boldsymbol{x} \perp \boldsymbol{y}).$$

- Deux variétés linéaires sont orthogonales si les sous-espaces vectoriels qui les définissent le sont. Le sous-espace orthogonal *supplémentaire* $F^{\perp}$ d'un sous-espace vectoriel F est l'ensemble des éléments de E orthogonaux à tous les éléments de F :

$$F^{\perp} = \{\boldsymbol{x} \in E / \forall \boldsymbol{y} \in F \quad \boldsymbol{x} \perp \boldsymbol{y}\}$$

  On peut montrer que les deux sous-espaces F et $F^{\perp}$ sont orthogonaux et supplémentaires. Une décomposition en somme directe de sous-espaces orthogonaux est une décomposition en somme directe de sous-espaces orthogonaux deux à deux.

Une *norme* sur un espace vectoriel E est une application $\|.\|$ de E dans $\mathbb{R}^+$ vérifiant :

$$\begin{array}{ll} \forall \boldsymbol{x} \in E, \forall \lambda \in \mathbb{R} & \|\lambda \boldsymbol{x}\| = |\lambda| \|\boldsymbol{x}\|, \\ \forall \boldsymbol{x} \in E & \|\boldsymbol{x}\| = 0 \Rightarrow \boldsymbol{x} = 0, \\ \forall \boldsymbol{x}, \boldsymbol{y} \in E & \|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|. \end{array}$$

Un vecteur *normé* est un vecteur de norme 1.

Une *distance* sur un ensemble quelconque A une application d de $A \times A$ dans $\mathbb{R}^+$ vérifiant :

$$\begin{array}{ll} \forall \boldsymbol{x}, \boldsymbol{y} \in A & d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x}), \\ \forall \boldsymbol{x}, \boldsymbol{y} \in A & d(\boldsymbol{x}, \boldsymbol{y}) = 0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{y}, \\ \forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in A & d(\boldsymbol{x}, \boldsymbol{y}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z}). \end{array}$$

Lorsque E est muni d'un produit scalaire, on peut montrer que l'application définie par $\|\boldsymbol{x}\| = \sqrt{< \boldsymbol{x}; \boldsymbol{x} >}$ est une norme sur E appelée *norme euclidienne* et

que l'application d définie par $d(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\| = \sqrt{<\boldsymbol{x}-\boldsymbol{y};\boldsymbol{x}-\boldsymbol{y}>}$ est une distance sur E appelée *distance euclidienne*. Un *espace euclidien* est un espace vectoriel réel de dimension finie muni d'un produit scalaire.

Une *base orthonormée* est une base formée de vecteurs normés et orthogonaux deux à deux. Par exemple, on peut facilement montrer que la base canonique est orthonormée pour le produit scalaire usuel.

1.3 Projection linéaire

Ayant une décomposition $E = F \oplus G$ de E en deux sous-espaces supplémentaires, la décomposition unique $\boldsymbol{x} = \boldsymbol{y} + \boldsymbol{z}$ avec $\boldsymbol{y} \in F$ et $\boldsymbol{z} \in G$ permet alors de définir deux applications :

- la première, qui associe au vecteur $\boldsymbol{x}$ de E le vecteur $\boldsymbol{y}$ de F, est appelée projection sur F parallèlement à G ;
- la seconde, qui associe au vecteur $\boldsymbol{x}$ de E le vecteur $\boldsymbol{z}$ de G, est appelée projection sur G parallèlement à F.

On peut montrer qu'une projection est une application linéaire et qu'elle est indempotente ($p \circ p = p$). Réciproquement toute application linéaire indempotente est une projection.

On appelle *projection orthogonale sur un sous-espace vectoriel* F la projection sur F parallèlement à $F^\perp$. On peut étendre cette définition aux variétés linéaires. Si $F = a + G$ est une variété linéaire, $\boldsymbol{x}$ un point quelconque de E et $H = \boldsymbol{x} + G^\perp$, on peut montrer que l'intersection de H et F se réduit à un seul élément qui est appelé *projection orthogonale de $\boldsymbol{x}$ sur la variété linéaire F*.

1.4 Calcul matriciel

Matrice associée à une application linéaire

Si E et F sont de dimensions finies p et n et si $(\mathbf{e}_1,\ldots,\mathbf{e}_p)$ et $(\mathbf{f}_1,\ldots,\mathbf{f}_n)$ sont des bases de E et F, il est possible d'associer à une application linéaire f de E dans F une matrice A de dimension (n,p). Celle-ci est construite en rangeant en colonne les coordonnées des images des vecteurs de la base de E sur la base de F : si la matrice A est notée (a_{ij}), a_{ij} est la i^e coordonnée de $f(\mathbf{e}_j)$.

La relation $\boldsymbol{y} = f(\boldsymbol{x})$ s'écrit alors matriciellement $\boldsymbol{y} = A\boldsymbol{x}$. Ici, pour simplifier l'écriture, les éléments de E et F et leurs vecteurs de coordonnées associées dans les bases correspondantes sont notés de la même façon. Réciproquement, toute application de E dans F se mettant sous la forme $\boldsymbol{y} = A\boldsymbol{x}$ est une application linéaire.

Opérations sur les matrices

Les opérations matricielles de base sont le produit d'une matrice par un réel, la somme de deux matrices, le produit de deux matrices et la transposition d'une matrice.

Les matrices associées aux endomorphismes sont carrées et il est alors possible de définir sur de telles matrices les notions suivantes :

- matrice diagonale : matrice dont tous les termes non diagonaux sont nuls,
- matrice symétrique : matrice égale à sa transposée
- matrice identité I : matrice diagonale dont tous les termes diagonaux sont egaux à 1,

- trace d'une matrice A : $\mathrm{tr}(A)$=somme des termes diagonaux de la matrice A,
- matrice inverse d'une matrice A : matrice A^{-1} vérifiant $A^{-1}A = AA^{-1} = I$.

Pour multiplier les lignes d'une matrice A de dimension (n,p) respectivement par les valeurs $\alpha_1,\ldots,\alpha_n$, il suffit de faire le produit $\mathrm{diag}(\alpha_1,\ldots,\alpha_n).A$.

De même, pour multiplier les colonnes d'une matrice A de dimension (n,p) respectivement par les valeurs $\beta_1,\ldots,\beta_p$, il suffit de faire le produit $A.\,\mathrm{diag}(\beta_1,\ldots,\beta_p)$.

Matrice de changement de base

Si $(\mathbf{e}_j)_{j=1,p}$ et $(\mathbf{f}_j)_{j=1,p}$ sont deux bases d'un espace vectoriel E, et si $\boldsymbol{x}$ et $\boldsymbol{x}'$ sont les vecteurs des coordonnées d'un élément de E dans ces deux bases, on a la relation :

$$\boldsymbol{x} = P\boldsymbol{x}' \text{ et } \boldsymbol{x}' = P^{-1}\boldsymbol{x}$$

où P est une matrice carrée de dimension p, appelée matrice de changement de base. Pour obtenir cette matrice de changement de base, il suffit de ranger en colonne les coordonnées des nouveaux vecteurs de base sur l'ancienne base.

La matrice de passage entre deux bases orthonormées est orthogonale, c'est-à-dire vérifie la relation $P' = P^{-1}$ (sa transposée est aussi son inverse).

Effet sur la matrice associée à un endomorphisme

Si f est un endomorphisme sur E, P la matrice de changement de base, A la matrice associée à f dans la base $(\mathbf{e}_j)$, B la matrice associée à f dans la base $(\mathbf{f}_j)$, alors on a la relation $B = P^{-1}AP$.

Preuve : Soit a un éléments de E. Notons $\boldsymbol{x}$ et $\boldsymbol{y}$ les coordonnées de a et de $f(a)$ dans la première base et $\boldsymbol{x}'$ et $\boldsymbol{y}'$ les coordonnées des mêmes éléments dans la seconde base, nous avons :

$$\boldsymbol{y} = P\boldsymbol{y}' \qquad \boldsymbol{x} = P\boldsymbol{x}' \qquad \boldsymbol{y} = A\boldsymbol{x}$$

d'où

$$P\boldsymbol{y}' = AP\boldsymbol{x}' \qquad \boldsymbol{y}' = P^{-1}AP\boldsymbol{x}' = B\boldsymbol{x} \qquad \text{avec} \quad B = P^{-1}AP.$$

□

Matrice associée à un produit scalaire

Lorsque l'on se place dans l'espace $\mathbb{R}^p$ muni de sa base canonique, on peut montrer que tout produit scalaire $<\boldsymbol{x},\boldsymbol{y}>$ s'écrit sous la forme $\boldsymbol{x}'M\boldsymbol{y}$ où M est une matrice :

- symétrique :	$M' = M$,	
- définie :	$\forall \boldsymbol{x} \in \mathbb{R}^p$	$\boldsymbol{x}'M\boldsymbol{x} = 0 \Rightarrow \boldsymbol{x} = 0$,
- positive :	$\forall \boldsymbol{x} \in \mathbb{R}^p$	$\boldsymbol{x}'M\boldsymbol{x} \geq 0$.

Remarquons que le produit scalaire habituel correspond à la matrice identité. La distance euclidienne et la norme euclidienne définies à partir de ce produit scalaire vérifient alors les relations :

$$||\boldsymbol{x}|| = \sqrt{\boldsymbol{x}'M\boldsymbol{x}} \qquad \text{et} \qquad d(\boldsymbol{x},\boldsymbol{y}) = \sqrt{(\boldsymbol{x}-\boldsymbol{y})'M(\boldsymbol{x}-\boldsymbol{y})}.$$

1.5 Vecteurs et valeurs propres

Définition et propriétés

On appelle vecteur propre d'un endomorphisme f sur E tout élément $\boldsymbol{x}$ de E non nul tel qu'il existe un réel λ vérifiant $f(\boldsymbol{x}) = \lambda\boldsymbol{x}$. Ce réel λ est appelé valeur propre associée au vecteur propre $\boldsymbol{x}$.

Si $\boldsymbol{x}$ est un vecteur propre, les vecteurs $a\boldsymbol{x}$ où a est un réel non nul sont aussi des vecteurs propres et ont même valeur propre.

L'ensemble de tous les vecteurs propres associés à une même valeur propre auquel est ajouté le vecteur nul est un espace vectoriel. Il est appelé espace propre associé à la valeur propre λ et noté E_λ.

Recherche des valeurs propres et vecteurs propres

On suppose dans ce paragraphe que E est de dimension finie. Soient e_i une base de E et A la matrice carrée associée à un endomorphisme f dans cette base, on a alors :

$$\boldsymbol{x} \text{ vecteur propre de } f \Longleftrightarrow \boldsymbol{x} \neq 0 \text{ et } A\boldsymbol{x} = \lambda\boldsymbol{x} \Longleftrightarrow \boldsymbol{x} \neq 0 \text{ et } (A - \lambda I)\boldsymbol{x} = 0$$

Le système de p équations à p inconnues ainsi défini ne doit donc pas être un système de Cramer, sinon la solution unique serait 0. Les solutions λ doivent donc annuler le déterminant de la matrice $(A-\lambda I)$. Il suffit ensuite pour chaque valeur λ réalisant cette condition de trouver les vecteurs $\boldsymbol{x}$ vérifiant le système $A\boldsymbol{x} = \lambda\boldsymbol{x}$.

Application : diagonalisation d'une matrice

Si E est un espace vectoriel de dimension finie muni d'une base $(\mathbf{e}_j)$ et f un endomorphisme sur E dont la matrice associée dans cette base est A , on cherche une nouvelle base $(\mathbf{f}_j)$ telle que la matrice associée à f soit diagonale.

Il est facile de montrer que les vecteurs de la nouvelle base sont nécessairement des vecteurs propres de f et que les termes de la diagonale sont les valeurs propres associées. Réciproquement, si on a une base formée de vecteurs propres de f, la matrice associée à f est diagonale et les valeurs propres sont les termes de la diagonale : diagonaliser une matrice revient donc à trouver une base de vecteurs propres.

Toute matrice carrée n'est pas diagonalisable, mais on peut montrer que certaines familles de matrices carrées le sont.

Matrices symétriques et matrices Q-symétriques

Toute matrice B symétrique possède une base orthonormée (au sens du produit scalaire usuel) de vecteurs propres et est donc diagonalisable. La matrice P de changement de base est orthogonale ($P' = P^{-1}$ ou $P'P = PP' = I$). De plus, si B est positive alors toutes les valeurs propres sont positives ou nulles.

Si Q est une matrice définissant un produit scalaire, toute matrice B Q-symétrique (c'est-à-dire QB symétrique) possède une base Q-orthonormée de vecteurs propres et est donc diagonalisable. La matrice P de changement de base vérifie $P'QP = PP'Q = QPP' = I$ et $P'QBP$ est la matrice diagonale des valeurs propres. De plus, si B est Q-positive (c'est-à-dire QB positive) alors toutes les valeurs propres sont positives ou nulles.

Décomposition spectrale d'une matrice

L'orthogonalité et la norme étant définies à l'aide d'une matrice Q, si B est une matrice Q-symétrique et Q-positive et $(\mathbf{u}_1, \ldots, \mathbf{u}_p)$ une base Q-orthonormée de vecteurs propres de la matrice B rangés suivant l'ordre décroissant des valeurs propres λ_k associés, alors :

- Le vecteur de norme 1 maximisant $< \mathbf{u}, B\mathbf{u} >$ est le vecteur $\mathbf{u}_1$ et la valeur maximisée est λ_1.
- $\forall k, 1 < k \leq 1$, le vecteur de norme 1, orthogonal au sous-espace engendré par les vecteurs $\mathbf{u}_1, \ldots, \mathbf{u}_{k-1}$ maximisant $< \mathbf{u}, B\mathbf{u} >$ est le vecteur $\mathbf{u}_k$ et la valeur maximisée est λ_k.

1.6 Espace euclidien

Rappelons qu'un espace euclidien est un espace vectoriel réel de dimension finie muni d'un produit scalaire.

Théorème de Pythagore

Si $\boldsymbol{x} = \boldsymbol{x}_1 + \ldots + \boldsymbol{x}_k$ est la décomposition d'un élément suivant une somme directe de sous-espaces orthogonaux, on peut montrer

$$\|\boldsymbol{x}\|^2 = \sum_{i=1}^{k} \|\boldsymbol{x}_i\|^2.$$

Si $x_1, \ldots, x_p$ sont les coordonnées d'un vecteur $\boldsymbol{x}$ dans une base orthonormée, on peut en déduire la relation :

$$\|\boldsymbol{x}\|^2 = \sum_{i=1}^{k} (x_i)^2.$$

Projection orthogonale sur un sous-espace vectoriel F étant un sous-espace vectoriel, $\boldsymbol{x}$ un point quelconque de E, $\boldsymbol{y}$ sa projection orthogonale sur F et $\mathbf{t}$ un point quelconque de F, alors on a

$$(\boldsymbol{x} - \mathbf{t}) = (\boldsymbol{x} - \boldsymbol{y}) + (\boldsymbol{y} - \mathbf{t}).$$

Cette relation représente la décomposition de $(\boldsymbol{x} - \mathbf{t})$ sur F et $F^\perp$ (figure 1.1) et le théorème de Pythagore peut donc s'appliquer :

$$\|\boldsymbol{x} - \mathbf{t}\|^2 = \|\boldsymbol{x} - \boldsymbol{y}\|^2 + \|\boldsymbol{y} - \mathbf{t}\|^2$$

$$d^2(\boldsymbol{x}, \mathbf{t}) = d^2(\boldsymbol{x}, \boldsymbol{y}) + d^2(\boldsymbol{y}, \mathbf{t})$$

La quantité $d^2(\boldsymbol{y}, \mathbf{t})$ étant toujours positive, cette relation permet d'affirmer que $\boldsymbol{y}$ est l'élément de F le plus proche de $\boldsymbol{x}$. Finalement, les trois relations suivantes sont équivalentes :

$$\begin{array}{l} y \text{ est la projection orthogonale de } \boldsymbol{x} \text{ sur } F \\ \forall \mathbf{t} \in F \quad (\boldsymbol{x} - \boldsymbol{y}) \perp t \text{ ou } (\boldsymbol{x} - \boldsymbol{y})' M \mathbf{t} = 0 \\ d(\boldsymbol{x}, \boldsymbol{y}) = \inf\{d(\boldsymbol{x}, \mathbf{t}) / \mathbf{t} \in F\} \end{array}$$

La quantité $d(\boldsymbol{x}, \boldsymbol{y})$ est souvent appelée "distance" de $\boldsymbol{x}$ au sous-espace F et notée $d(\boldsymbol{x}, F)$.

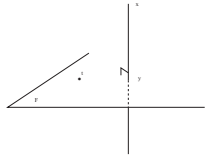


FIGURE 1.1 – Projection sur un plan

Projection orthogonale sur une variété linéaire Soit $\boldsymbol{y}$ la projection orthogonale d'un point $\boldsymbol{x}$ quelconque de E sur la variété linéaire $F = \mathbf{a} + G$. Rappelon que $\boldsymbol{y}$ est l'intersection des variétés linéaires $H = \boldsymbol{x} + G^{\perp}$ et F (fig 1.2. On peut alors étendre les résultats obtenus précédemment :

1. $\boldsymbol{y}$ est la projection orthogonale de $\boldsymbol{x}$ sur F
2. $\forall \mathbf{t}, \mathbf{u} \in F \qquad (\boldsymbol{x} - \boldsymbol{y}) \perp (\mathbf{t} - \mathbf{u})$ ou encore $(\boldsymbol{x} - \boldsymbol{y})' M (\mathbf{t} - \mathbf{u}) = 0$
3. $d(\boldsymbol{x}, \boldsymbol{y}) = \inf\{d(\boldsymbol{x}, \mathbf{t}) / \mathbf{t} \in F\}$

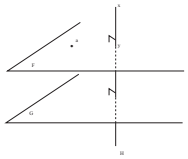


FIGURE 1.2 – Projection sur une variété linéaire

La quantité $d(\boldsymbol{x}, \boldsymbol{y})$ est appelée « distance » de $\boldsymbol{x}$ au sous-espace F et notée $d(\boldsymbol{x}, F)$.

Nuage de points et centre de gravité

Si Ω est un ensemble fini de points d'un espace euclidien $\mathcal{E}$ et si chaque point $\boldsymbol{x}$ de Ω est muni d'une pondération $\mu_{\boldsymbol{x}} > 0$, alors l'ensemble

$$\mathcal{N}(\Omega) = \{(\boldsymbol{x}, \mu_{\boldsymbol{x}}) / \boldsymbol{x} \in \Omega\}$$

est appelé *nuage de points* de $\mathcal{E}$ et son *centre de gravité* est défini par

$$\mathbf{g} = \frac{1}{\mu} \sum_{\boldsymbol{x} \in \Omega} \mu_{\boldsymbol{x}} \boldsymbol{x}$$

où μ est la somme des pondérations $\sum_{\boldsymbol{x}} \mu_{\boldsymbol{x}}$.

Inerties

L'*inertie de* $\mathcal{N}(\Omega)$ *par rapport à un point* $\mathbf{a}$ est définie par

$$I_{\mathbf{a}} = \sum_{\boldsymbol{x} \in \Omega} \mu_{\boldsymbol{x}} d^2(\mathbf{a}, \boldsymbol{x})$$

et l'*inertie du nuage* $\mathcal{N}(\Omega)$ *par rapport à une variété linéaire* F par

$$I_F = \sum_{\boldsymbol{x} \in \Omega} \mu_{\boldsymbol{x}} d^2(\boldsymbol{x}, F)$$

où $d(\boldsymbol{x}, F) = d(\boldsymbol{x}, \boldsymbol{y})$ avec $\boldsymbol{y}$ projection orthogonale de $\boldsymbol{x}$ sur F.

L'inertie $I_{\mathbf{g}}$ du nuage Ω par rapport à son centre de gravité est appelée simplement *Inertie du nuage* et notée I.

Théorèmes de Huygens

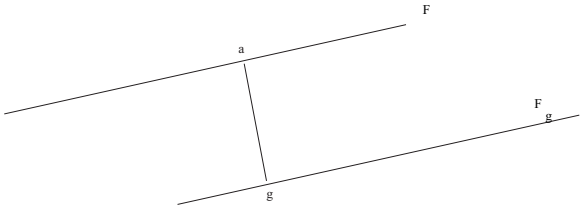
- Version 1 :

$$I_{\mathbf{a}} = I_{\mathbf{g}} + \mu d^2(\mathbf{a}, \mathbf{g}) \qquad \forall \mathbf{a} \in \mathbb{R}^p$$

  Le centre de gravité est donc le point d'inertie minimum.
- Version 2 :

$$I_F = I_{F_{\mathbf{g}}} + \mu d^2(\mathbf{a}, \mathbf{g}) \qquad \forall \quad \text{variété linéaire} \quad F$$

  où $F_{\mathbf{g}}$ est la variété linéaire parallèle à F passant par $\mathbf{g}$ et $\mathbf{a}$ la projection orthogonale de $\mathbf{g}$ sur F.



Le sous-espace affine parallèle à F d'inertie minimum est donc $F_{\mathbf{g}}$.

Inertie expliquée

Les propriétés d'optimalité du centre de gravité vis à vis de l'inertie conduisent souvent à placer celui-ci à l'origine à l'aide d'une translation. On dit alors que le nuage est centré. C'est ce que l'on supposera dans ce paragraphe.

Si $\mathbb{R}^p = F \oplus F^{\perp}$ est une décomposition de $\mathcal{E}$ en 2 sous-espaces vectoriels supplémentaires orthogonaux, on peut alors montrer que l'inertie I se décompose suivant la relation

$$I = I_F + I_{F^{\perp}}.$$

En outre, l'inertie $I_{F^{\perp}}$, inertie du nuage par rapport à $F^{\perp}$, peut s'interpréter comme l'inertie du nuage des points projetés orthogonalement sur F. Pour cette raison, cette inertie est aussi appelée *inertie expliquée* par le sous-espace vectoriel F. On peut alors montrer la décomposition suivante :

$$A = B \oplus C \quad \text{et} \quad B \perp C \Rightarrow I_{A^{\perp}} = I_{B^{\perp}} + I_{C^{\perp}}.$$

2 Probabilités

2.1 Variables aléatoires

Expérience aléatoire

On appelle *expérience aléatoire* une expérience qui, répétée plusieurs fois dans des conditions opératoires identiques, produit des résultats qui peuvent être différents. Mathématiquement, la notion d'expérience aléatoire $\mathcal{E}$ se formalise en définissant :

1. un ensemble fondamental Ω définissant l'ensemble des résultats possibles de $\mathcal{E}$, appelés *événements élémentaires* ;
2. un ensemble $\mathcal{A}$ de parties de Ω, appelées *événements*. Un événement aléatoire correspond à une affirmation qui peut être vraie ou fausse suivant le résultat de l'expérience aléatoire.
3. une fonction $\mathbb{P} : \mathcal{A} \to [0,1]$, appelée *mesure* ou *distribution* de probabilité, qui à tout événement A associe un nombre $\mathbb{P}(A)$ appelé probabilité de cet événement.

Variable aléatoire

Une variable aléatoire (v.a.) est une grandeur numérique dont la valeur est fonction du résultat d'une expérience aléatoire. Mathématiquement, cette notion se formalise par une fonction

$$\begin{array}{rl} X : & \Omega \to \mathbb{R} \\ & \omega \to X(\omega). \end{array}$$

On notera $V_X = X(\Omega)$ l'ensemble des valeurs prises par la v.a. X. On parle de v.a. *discrète* lorsque V_X est fini ou dénombrable. Dans le cas contraire, la v.a. X est dite *continue*.

Fonction de variables aléatoires

Si X est une v.a. et φ une fonction de $\mathbb{R}$ dans $\mathbb{R}$, on définit une nouvelle v.a., notée $\varphi(X)$, en associant à chaque ω le réel $\varphi(X(\omega))$. On peut ainsi définir les variables aléatoires aX, $X+b$ et X^2.

Si X et Y sont deux v.a. définies sur le même espace fondamental Ω et φ une fonction de $\mathbb{R}^2$ dans $\mathbb{R}$, on définit une nouvelle v.a., notée $\varphi(X)$, en associant à chaque ω le réel $\varphi((X(\omega), Y(\omega))$. On peut ainsi définir la v.a. $X+Y$.

Loi de probabilité d'une v.a.

Soit B un intervalle de $\mathbb{R}$. On peut définir la probabilité que la v.a. X prenne sa valeur dans B comme

$$\mathbb{P}_X(B) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in B\}) = \mathbb{P}(X^{-1}(B)),$$

quantité notée simplement $\mathbb{P}(X \in B)$. La donnée de $\mathbb{P}_X(B)$ pour tout intervalle B définit la *loi (ou distribution) de probabilité* de X.

Mathématiquement, c'est une fonction de $\mathcal{B}(\mathbb{R})$ dans $[0,1]$, $\mathcal{B}(\mathbb{R})$ étant l'ensemble des intervalles ou unions dénombrables d'intervalles de $\mathbb{R}$, appelé *tribu*

borélienne. La fonction $\mathbb{P}_X$ est une mesure de probabilité sur l'espace probabilisable $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, appelée mesure image de $\mathbb{P}$ par X. Pour décrire complètement $\mathbb{P}_X$, il suffit de donner les probabilités pour des intervalles de la forme $]-\infty, x]$ pour tout $x \in \mathbb{R}$. On appelle *fonction de répartition* de X la fonction

$$\begin{array}{rl} F_X : & \mathbb{R} \to [0,1] \\ & x \to \mathbb{P}_X(]-\infty, x]), \end{array}$$

ce que l'on note $F_X(x) = \mathbb{P}(X \leq x)$.

Une loi de probabilité peut également être définie :

- dans le cas discret par la *fonction de probabilité* p_X qui à chaque élément de V_X associe sa probabilité :

$$\begin{array}{rl} p_X : & \mathbb{R} \to [0,1] \\ & x \to \mathbb{P}_X(\{x\}) \end{array}$$

  et qui vérifie

$$\forall B \in \mathcal{B}(\mathbb{R}),\ \mathbb{P}_X(B) = \sum_{x \in B} p_X(x)$$

- et dans le cas continu, par la *fonction de densité de probabilité* qui vérifie

$$\forall B \in \mathcal{B}(\mathbb{R}), \quad \mathbb{P}_X(B) = \int_B f_X(t)dt.$$

Espérance mathématique

L'espérance mathématique d'une variable aléatoire réelle, qui représente la « valeur moyenne » prise par cette variable aléatoire, est définie par

$$\mathbb{E}(X) = \begin{cases} \displaystyle\sum_{x \in V_X} x p_X(x) & \text{si } X \text{ est une v.a. discrète,} \\ \displaystyle\int_{\mathbb{R}} x f_X(x)dx & \text{si } X \text{ est une v.a. continue} \end{cases}$$

si ces quantités existent. Dans le contraire, X n'a pas d'espérance mathématique. L'espérance est un opérateur linéaire :

$$\mathbb{E}(aX) = a\mathbb{E}(X) \qquad \mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$

Variance

La variance, qui est une mesure de dispersion de la v.a. autour de son espérance, est définie par

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2)\right] = \mathbb{E}[(X)^2] - (\mathbb{E}[X])^2.$$

La racine carrée de la variance est appelée *écart-type* de la v.a. X et notée σ. La variance, étant une espérance, peut ne pas être définie. La variance vérifie la propriété suivante

$$\mathrm{Var}(aX) = a^2 \mathrm{Var}(X).$$

Covariance

La covariance entre entre deux variables aléatoires X et Y est définie par

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

et que cette covariance vérifie les propriétés suivantes

- $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$
- $\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$
- $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$
- $\mathrm{Cov}(aX,bY) = ab\,\mathrm{Cov}(X,Y)$.
- Inégalité de Cauchy-Schwarz : $[\mathrm{Cov}(X,Y)]^2 \leq \mathrm{Var}(X)\,\mathrm{Var}(Y)$ (égalité ssi $X - \mathbb{E}(X) = k(Y - \mathbb{E}(Y))$.

Corrélation

Le coefficient de corrélation ρ_{ij} entre deux variables aléatoires X_j et $X_{j'}$ est défini par

$$\rho_{jj'} = \frac{\mathrm{Cov}(X_i, X_{j'})}{\sqrt{\mathrm{Var}(X_j)\,\mathrm{Var}(X_{j'})}} = \frac{\mathrm{Cov}(X_j, X_{j'})}{\sigma_i \sigma_{j'}}.$$

2.2 Vecteurs aléatoires

Définition

La notion de vecteur aléatoire (réel), ou variable aléatoire vectorielle, généralise celle de variable aléatoire présentée au paragraphe précédent. On appelle vecteur aléatoire (réel) un vecteur de $\mathbb{R}^p$ dont les composants sont fonctions du résultat d'une expérience aléatoire $\mathcal{E}$. Il s'agit donc d'une application :

$$\begin{aligned} X : \Omega &\to \mathbb{R}^p \\ \omega &\mapsto X(\omega) = (X_1(\omega), \ldots, X_p(\omega))'. \end{aligned}$$

Un *vecteur aléatoire* (réel) est donc un vecteur $\boldsymbol{X} = (X_1, \ldots, X_p)'$ dont les composantes X_j sont des variables aléatoires réelles.

Fonction de vecteurs aléatoires

Si $\boldsymbol{X}$ est un vecteur aléatoire de dimension p et φ une fonction de $\mathbb{R}^p$ dans $\mathbb{R}^q$, on définit un nouveau vecteur aléatoire de dimension q, notée $\varphi(\boldsymbol{X})$, en associant à chaque ω le vecteur $\varphi(\boldsymbol{X}(\omega))$. Si A est une matrice de dimension (p,q) et $\mathbf{b}$ un vecteur de dimension p, on peut ainsi définir les vecteurs aléatoires $A'X$ et $X + \mathbf{b}$.

Si $\boldsymbol{X}$ et $\boldsymbol{Y}$ sont deux vecteurs aléatoires de dimension p définies sur le même espace fondamental Ω et φ une fonction de $\mathbb{R}^p \times \mathbb{R}^p$ dans $\mathbb{R}^p$, on définit un nouveau vecteur aléatoire, notée $\varphi(\boldsymbol{X}, \boldsymbol{Y})$, en associant à chaque ω le vecteur $\varphi((\boldsymbol{X}(\omega), \boldsymbol{Y}(\omega))$. On peut ainsi définir le vecteur aléatoire v.a. $\boldsymbol{X} + \boldsymbol{Y}$.

Loi jointe

La loi de probabilité du vecteur aléatoire $\boldsymbol{X} = (X_1, \ldots, X_p)'$, appelée loi jointe, est définie par :

$$\mathbb{P}(\boldsymbol{X} \in A) = \mathbb{P}_{\boldsymbol{X}}(A) = \mathbb{P}(\omega \in \Omega | (X_1(\omega), \ldots, X_p(\omega)) \in A).$$

Cette loi de probabilité du vecteur aléatoire $\boldsymbol{X}$ peut être décrite par sa *fonction de répartition* définie pour tout $\boldsymbol{x} = (x_1, \ldots, x_p)'$ de $\mathbb{R}^p$ par

$$F_{\boldsymbol{X}}(x_1, \ldots, x_p) = \mathbb{P}(X_1 \leq x_1; \ldots; X_p \leq x_p).$$

Lorsque cette fonction est dérivable par rapport à chaque variable, on peut définir la *densité de probabilité* de $\boldsymbol{X}$ par

$$f_{\boldsymbol{X}}(x_1, \ldots, x_p) = \frac{\partial^p F_{\boldsymbol{X}}(x_1, \ldots, x_p)}{\partial x_1 \partial x_2 \ldots \partial x_p}.$$

Par la suite, nous noterons indifféremment $F_{\boldsymbol{X}}(x_1, \ldots, x_p)$ ou $F_{\boldsymbol{X}}(\boldsymbol{x})$, $f_{\boldsymbol{X}}(x_1, \ldots, x_p)$ ou $f_{\boldsymbol{X}}(\boldsymbol{x})$, etc.

Si $f_{\boldsymbol{X}}$ existe, on a :

$$\mathbb{P}(\boldsymbol{X} \in A) = \int_A f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x},$$

pour tout $A \subseteq \mathbb{R}^p$ pour lequel cette intégrale est définie.

Lorsque le vecteur aléatoire $\boldsymbol{X}$ est discret, c'est-à-dire lorsque les composantes X_j sont des variables aléatoires discrètes, on peut définir la fonction de probabilité de $\boldsymbol{X}$ (équivalent de la fonction de densité) par

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}(X_1 = x_1, \ldots, X_p = x_p).$$

On a alors $\mathbb{P}(\boldsymbol{X} \in A) = \sum_{\boldsymbol{x} \in A} f_{\boldsymbol{X}}(\boldsymbol{x})$ pour tout $A \subseteq \mathbb{R}^p$)

Dans la suite et s'il n'y a pas d'ambiguïté, on notera simplement p, f et F les fonctions $p_{\boldsymbol{X}}$, $f_{\boldsymbol{X}}$ et $F_{\boldsymbol{X}}$.

Lois marginales

Tout sous-vecteur du vecteur aléatoire $\boldsymbol{X}$, c'est-à-dire tout sous-ensemble de l'ensemble des variables aléatoires $X_1, \ldots, X_p$ est lui-même un vecteur aléatoire. La loi d'un tel vecteur aléatoire est appelée loi marginale. Si $X_{j_1}, \ldots, X_{j_q}$ est ce sous-ensemble, la loi marginale sera notée $f_{j_1,\ldots,j_q}$.

Si cet ensemble se réduit à une seule variable et

- si $\boldsymbol{X}$ est discret, la loi de X_J est définie par la probabilité élémentaire

$$f_j(x_j) = \sum_{x_1 \in V_1, \ldots, x_{j-1} \in V_{j-1}, x_{j+1} \in V_{j+1}, \ldots, \ldots, x_p \in V_p} f(\boldsymbol{x})$$

- et si $\boldsymbol{X}$ est continu, la loi de X_J est définie par la densité

$$f_j(x_j) = \int_{\mathbb{R}^{p-1}} f(\boldsymbol{x}) dx_1 \ldots dx_{j-1} dx_{j+1} \ldots dx_p.$$

Espérance

L'espérance du vecteur aléatoire $\boldsymbol{X}$ est le vecteur des espérances des variables aléatoires X_j :

$$\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{X}) = (\mathbb{E}(X_1), \ldots, E(X_p))'.$$

$\boldsymbol{X}$ et $\boldsymbol{Y}$ étant des vecteurs aléatoires de dimension p, on a les propriétés suivantes.

Proposition 1

$$\mathbb{E}(\boldsymbol{X}+\boldsymbol{Y}) = \mathbb{E}(\boldsymbol{X}) + \mathbb{E}(\boldsymbol{Y})$$

Proposition 2 *Pour toute matrice* $A \in \mathcal{M}_{q,p}(\mathbb{R})$ *et tout vecteur* $\mathbf{b} \in \mathbb{R}^q$ *constants, on a*

$$\mathbb{E}(A\boldsymbol{X}+\mathbf{b}) = A\,\mathbb{E}(\boldsymbol{X}) + \mathbf{b}.$$

En particulier, si $\mathbf{u} \in \mathbb{R}^p$, $\mathbb{E}(\mathbf{u}'\boldsymbol{X}) = \mathbf{u}'\mathbb{E}(\boldsymbol{X})$.

Espérance d'une fonction réelle d'un vecteur aléatoire Si φ est une fonction de $\mathbb{R}^p$ dans $\mathbb{R}$, on a

$$\mathbb{E}(\varphi(\boldsymbol{X})) = \int_{\mathbb{R}^p} \varphi(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x}$$

pour un vecteur aléatoire continu et

$$\mathbb{E}(\varphi(\boldsymbol{X})) = \sum_{\boldsymbol{x} \in V_1 \times \ldots \times V_p} \varphi(\boldsymbol{x}) f(\boldsymbol{x})$$

pour un vecteur aléatoire discret.

Comme pour les variables aléatoires, ce résultat permet de calculer l'espérance d'un vecteur aléatoire $\varphi(\mathbf{X})$ sans avoir besoin de calculer sa loi.

Matrice de Variance

La variance du vecteur aléatoire $\boldsymbol{X}$, souvent appelée *matrice de variance*, est la matrice carrée symétrique Σ de dimension p de terme général

$$\begin{aligned} \sigma_{jj'} &= \mathrm{Cov}(X_j, X_{j'}) \\ &= \mathbb{E}[(X_j - \mathbb{E}(X_{j'}))(X_j - \mathbb{E}(X_{j'}))] \\ &= \mathbb{E}(X_j X_{j'}) - \mathbb{E}(X_j)\mathbb{E}(X_{j'}). \end{aligned}$$

En particulier, $\sigma_{jj} = \mathrm{Var}(X_j)$.

On peut écrire matriciellement

$$\Sigma = \mathrm{Var}(\boldsymbol{X}) = \mathbb{E}([\boldsymbol{X} - \mathbb{E}(\boldsymbol{X})][\boldsymbol{X} - \mathbb{E}(\boldsymbol{X})]').$$

Proposition 3 *Pour toute matrice* $A \in \mathcal{M}_{q,p}(\mathbb{R})$ *et tout vecteur* $\mathbf{b} \in \mathbb{R}^q$ *constants, on a*

$$\mathrm{Var}(A\boldsymbol{X}+\mathbf{b}) = A\,\mathrm{Var}(\boldsymbol{X})A' = A\Sigma A'.$$

En particulier, on a donc pour tout $\mathbf{u}$ et $\mathbf{v} \in \mathbb{R}^p$

$$\mathrm{Var}(\mathbf{u}'\boldsymbol{X}) = \mathbf{u}'\Sigma\mathbf{u} \quad \text{et} \quad \mathrm{Cov}(\mathbf{u}'\boldsymbol{X}, \mathbf{v}'\boldsymbol{X}) = \mathbf{u}'\Sigma\mathbf{v},$$

ce qui montre que la matrice Σ est définie positive ($\mathbf{u}'\Sigma\mathbf{u} > 0, \forall \mathbf{u} \neq 0$), sauf s'il existe une relation $\mathbf{u}'\boldsymbol{X} = c$ pour un vecteur constant $\mathbf{u}$ et une constante c, auquel cas $\mathrm{Var}(\mathbf{u}'\boldsymbol{X}) = 0$.

On rappelle que toute matrice d'ordre p symétrique et définie positive a p valeurs propres strictement positives. La matrice Σ est donc inversible.

Matrice de corrélation La matrice de corrélation R d'un vecteur aléatoire est la matrice dont le terme général est la corrélation $\rho^{jj'}$. Toutes les valeurs sont donc comprises entre -1 et +1 et les termes de la diagonale sont égaux à 1. Si on note D la matrice diagonale $\mathrm{diag}(\sigma_1, \ldots, \sigma_p)$, on obtient les relations $\Sigma = DRD$ et $R = D^{-1}\Sigma D^{-1}$.

Indépendance de variables aléatoires

Les composantes $X_1, \ldots, X_p$ du vecteur aléatoire $\boldsymbol{X}$ sont *indépendantes* si la loi jointe du vecteur aléatoire $\boldsymbol{X} = (X_1, \ldots, X_p)'$ s'exprime comme le produit des lois marginales, c'est-à-dire si et seulement si :

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{j=1}^{p} f_{X_j}(x_j).$$

Etant donnée une variable aléatoire Z à valeurs dans Ω, les composantes $X_1, \ldots, X_p$ du vecteur aléatoires $\boldsymbol{X}$ sont *indépendantes conditionnellement à* Z ssi

$$f_{bsX}(\boldsymbol{x}|Z=z) = \prod_{j=1}^{p} f_{X_j}(x_j|Z=z), \quad \forall \boldsymbol{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p, \forall z \in \Omega.$$

On a les propriétés suivantes :

- $X_1, \ldots, X_p$ indépendantes $\Rightarrow$ tout sous-ensemble des v. a. est indépendant ; en particulier les v. a. $X_1, \ldots, X_p$ sont indépendantes 2 à 2
- Attention, la réciproque est fausse : l'indépendance 2 à 2 n'entraîne pas l'indépendance
- $X_1, \ldots, X_p$ indépendantes $\Longrightarrow \mathbb{E}(X_1 \ldots X_p) = \mathbb{E}(X_1) \ldots \mathbb{E}(X_p)$
- X_j et $X_{j'}$ indépendantes $\Longrightarrow \mathrm{Cov}(X_j, X_{j'}) = 0$
- $X_1, \ldots, X_p$ indépendantes $\Longrightarrow \mathrm{Var}(\sum_{j=1}^p X_j) = \sum_{j=1}^p \mathrm{Var}(X_j)$
- La matrice de variance sera diagonale si les variables X_j sont indépendantes 2 à 2. La réciproque est fausse.

Transformation d'un vecteur aléatoire

Soit $\mathbf{U}$ un vecteur aléatoire de dimension p, φ une application bijective de $\mathbb{R}^p$ dans $\mathbb{R}^p$, et $\boldsymbol{X} = \varphi(\mathbf{U})$. La densité de $\boldsymbol{X}$ s'obtient en fonction de celle de $\mathbf{U}$ par l'expression :

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{f_{\mathbf{U}}(\varphi^{-1}(\boldsymbol{x}))}{|\det J_\varphi|},$$

où $\det J_\varphi$ est le jacobien de la transformation défini par

$$\det J_\varphi = \begin{vmatrix} \frac{\partial x_1}{\partial u_1} & \cdots & \frac{\partial x_1}{\partial u_p} \\ \vdots & & \vdots \\ \frac{\partial x_p}{\partial u_1} & \cdots & \frac{\partial x_p}{\partial u_p} \end{vmatrix}.$$

2.3 Loi normale multidimensionnelle

Définition

Soient $U_1, \ldots, U_p$ p v.a. réelles normales, centrées-réduites et indépendantes, et $\mathbf{U} = (U_1, \ldots, U_p)'$. On appelle *loi normale à p dimensions* la loi suivie par $\boldsymbol{X} = \boldsymbol{\mu} + B\mathbf{U}$, où $\boldsymbol{\mu} \in \mathbb{R}^p$ et $B \in \mathcal{M}_{p,p}(\mathbb{R})$ sont des constantes.

La densité de $\mathbf{U}$ est

$$f(\mathbf{u}) = \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}\mathbf{u}'\mathbf{u}\right).$$

On peut calculer la densité de $\boldsymbol{X}$ par la méthode rappelée dans la section 2.2. On vérifie que $\det J_\varphi = \det B$. On en déduit :

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\det B|} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'(BB')^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

Or, $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ et $\Sigma = \mathrm{Var}(\boldsymbol{X}) = BB'$, d'où $|\det B| = (\det \Sigma)^{1/2}$. On obtient finalement l'expression usuelle de la densité de $\boldsymbol{X}$:

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}(\det \Sigma)^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

On note
$bsX \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$.

Propriétés

1. Dans le cas $p = 1$, on retrouve l'expression de la loi normale monodimensionnelle, avec $\sigma^2 = \Sigma$.
2. La matrice Σ est diagonale ssi les variables $X_1, \ldots, X_p$ sont indépendantes.
3. Tout sous-vecteur d'un vecteur aléatoire gaussien suit une loi normale. En particulier, ses composantes sont toutes gaussiennes.
4. Les courbes isodensité ont pour équation $(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = c$, où c est une constante. Ce sont des ellipsoïdes de centre $\boldsymbol{\mu}$. Lorsque la matrice Σ est diagonale, les axes de ces ellipsoïdes sont parallèles aux axes de coordonnées. Lorsque Σ est scalaire, ce sont des hypersphères.
5. Soient $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, $\mathbf{b} \in \mathbb{R}^q$ un vecteur constant, et $A \in \mathcal{M}_{q,p}(\mathbb{R})$ une matrice constante. Alors, $\boldsymbol{Y} = A\,\boldsymbol{X} + \mathbf{b} \sim \mathcal{N}(A\,\boldsymbol{\mu} + \mathbf{b}, A\Sigma A')$.

Estimation des paramètres

Soit $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ un échantillon indépendant et identiquement distribué (iid) de $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$. Les estimateurs du maximum de vraisemblance de $\boldsymbol{\mu}$ et de Σ sont :

$$\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i = (\overline{X}_1, \ldots, \overline{X}_p)',$$

où $\overline{X}_j$ est la moyenne empirique de l'échantillon $X_{1j}, \ldots, X_{nj}$, et

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}})'.$$

La matrice $\widehat{\Sigma}$ est appelée *matrice de variance empirique*.

L'estimateur $\widehat{\boldsymbol{\mu}}$ est sans biais et $\widehat{\boldsymbol{\mu}} \sim \mathcal{N}(\boldsymbol{\mu}, \frac{1}{n}\Sigma)$. En revanche, $\mathbb{E}(\widehat{\Sigma}) = \frac{n-1}{n}\Sigma$. Pour obtenir un estimateur sans biais de Σ, on définit la *matrice de variance empirique corrigée* par

$$S = \frac{1}{n-1}\sum_{i=1}^{n} (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}})'.$$

3 Statistiques

3.1 Les données individus-variables

On considère ici des données correspondant à un ensemble de n *individus* pour lesquels on connaît la valeur de p *variables*. On notera $X = (x_{ij})$ la matrice réelle à n lignes et p colonnes associée à ces données et $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^t$ et $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^t$ les vecteurs colonnes associés aux individus et aux variables.

En statistique, un tel tableau de données peut être vu comme la réalisation d'un échantillon de taille n d'un vecteur aléatoire de dimension p . Ce vecteur aléatoire de dimension p défini par les variables aléatoires $X_1, \ldots, X_p$ sera noté $\boldsymbol{X} = (X_1, \ldots, X_p)'$.

3.2 Quelques statistiques élémentaires

Description monodimensionnelle

On peut associer à chaque variable j la moyenne empirique :

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij},$$

la variance empirique

$$s_j^2 = \frac{1}{n}\sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2$$

et l'écart-type empirique

$$s_j = \sqrt{s_j^2}.$$

Description bidimensionnelle

On peut associer à chaque couple de variables j et j' la covariance empirique

$$s_{jj'} = \frac{1}{n}\sum_{i=1}^{n} (x_{ij} - \overline{x}_j).(x_{ij'} - \overline{x}_{j'})$$

et la corrélation empirique

$$r_{jj'} = \frac{s_{jj'}}{s_j s_{j'}}$$

Remarquons que l'on a $s_{jj} = (s_j)^2$ et $r_{jj} = 1$.

Description multidimensionnelle

On peut associer à l'ensemble des variables la matrice de covariance S de l'échantillon qui regroupent l'ensemble des covariances pour tous les couples de variables :

$$S = (s_{jj'})_{j,j'=1,\dots,p} = \frac{1}{n}(X - 1_n\overline{\boldsymbol{x}}^t)^t(X - 1_n\overline{\boldsymbol{x}}^t) = \frac{1}{n}Y^tY$$

où $\mathbb{1}_n$ est la matrice de dimension $(n, 1)$ remplie de 1 et Y est la matrice centrée associée à X, et la matrice de corrélation R de l'échantillon qui regroupent l'ensemble des corrélations pour tous les couples de variables :

$$R = (r_{jj'})_{j,j'=1,\dots,p} = D_{1/s_j}SD_{1/s_j}$$

où D_{1/s_j} est la matrice diagonale définie par les valeurs $(1/s_1, \dots, 1/s_p)$.

Remarquons que l'on utilise aussi la matrice de covariance empirique

$$S^* = (s_{jj'})_{j,j'=1,\dots,p} = \frac{1}{n-1}(X - 1_n\overline{\boldsymbol{x}}^t)^t(X - 1_n\overline{\boldsymbol{x}}^t) = \frac{1}{n-1}Y^tY.$$

Chapitre 2

L'analyse en composantes principales

1 Introduction

Les *méthodes factorielles* ont pour objectif de visualiser, et plus généralement, d'analyser des données multidimensionnelles, c'est-à-dire des données regroupant souvent un grand nombre de variables. La prise en compte simultanée de ces variables est un problème difficile ; heureusement, l'information apportée par ces variables est souvent redondante et toutes ces méthodes vont exploiter cette caractéristique pour tenter de remplacer les variables initiales par un nombre réduit de nouvelles variables sans perdre trop d'information. Remarquons que la construction de variables synthétiques, qui consiste à résumer plusieurs variables par une seule, est une démarche habituelle (moyenne à l'école, QI, répartition des hommes politiques sur un axe droite-gauche) . Il y a mieux à faire. C'est ce qu'ont proposé les psychologues américain Spearman, Burt et Thurstone en caractérisant les résultats à de nombreux tests psychologiques par un facteur général d'aptitude et un nombre très limité de facteurs spécifiques comme la mémoire ou l'intelligence.

Lorsque les variables sont toutes quantitatives, l'analyse en composantes principales (ACP) va chercher à résoudre ce problème en considérant que les nouvelles variables sont des combinaisons linéaires des variables initiales et, qu'en plus, elles doivent être non corrélées linéairement. Si l'on représente les données initiales à l'aide d'un nuage de points, on peut montrer que ce problème revient à chercher les droites, les plans et de manière plus générale les variétés linéaires proches du nuage initial. Nous utiliserons ce point de vue géométrique dans ce chapitre. Cette méthode a d'abord été développée par K.Pearson (1900) pour deux variables, puis par H. Hotelling (1933) qui l'a étendue à un nombre quelconque de variables. L'ouvrage de Jackson (1991) constitue un panorama très complet et assez récent de l'ACP.

Les méthodes factorielles, dont l'ACP est l'exemple le plus connu, varient suivant la forme des données mais utilisent toutes les mêmes bases mathématiques. Il faut les distinguer des méthodes regroupées sous le terme « factor analysis » par les anglo-saxons qui sont des méthodes de statistiques inférentielles s'appuyant sur un modèle statistique et qui sont assez peu utilisées en

France. En dehors de l'ACP destinée aux tableaux de variables quantitatives, les principales méthodes factorielles sont l'analyse factorielle des correspondances (AFC) pour les tableaux de contingence, l'analyse des correspondances multiples (ACM) pour les tableaux de variables qualitatives, l'analyse factorielle d'un tableau de distances (AFTD) pour les tableaux de proximités et l'analyse factorielle discriminante qui permet de mettre en évidence les différences entre des individus issus de plusieurs classes.

2 Les données

Dans tout ce chapitre, on supposera que les données sont définies par une matrice X de dimension (n,p) correspondant à la mesure de p variables quantitatives effectuées sur n individus. Rappelons que l'on peut associer à chaque individu i un vecteur $\boldsymbol{x}_i$ de dimension p et à chaque variable j un vecteur $\mathbf{x}_j$ de dimension n.

On suppose en outre qu'à chaque individu est associée la pondération $p_i = 1/n$ et à chaque variable la pondération $q_j = 1$. On notera D_p la matrice diagonale $\text{diag}(p_1,\ldots,p_n) = \frac{1}{n}I_n$ et M la matrice diagonale $\text{diag}(q_1,\ldots,q_j) = I_p$. L'utilisation de pondérations générales p_i et q_j permettra, par exemple, d'étendre sans difficulté l'analyse en composantes principales à l'analyse des correspondances.

On peut alors définir le nuage

$$\mathcal{N}(\Omega) = \{(\boldsymbol{x}_i, p_i), i = 1,\ldots,n\}$$

inclus dans $\mathbb{R}^p$ muni de la métrique euclidienne M.

De façon symétrique, les variables peuvent être représentées par le nuage :

$$\mathcal{N}(V) = \{(\mathbf{x}_j, q_j), j = 1,\ldots,p\}$$

inclus dans $\mathbb{R}^n$ muni de la métrique euclidienne D_p souvent appelée « métrique des poids ».

Pour simplifier les calculs, on supposera dans la suite que le nuage des individus est centré, c'est-à-dire que son centre de gravité est à l'origine ou encore que la moyenne de chaque variable est nulle ; on dit alors que le tableau est *centré en colonne*. Si ce n'est pas le cas, il est facile de s'y ramener en soustrayant à chaque colonne sa moyenne. Centrer en colonne revient, dans l'espace des individus, à prendre comme nouvelle origine le centre de gravité. On peut alors montrer que l'inertie du nuage des individus et l'inertie portée par un axe Δ_u de vecteur unitaire $\mathbf{u}$ s'écrivent respectivement

$$I = \text{tr}\,(SM) \quad \text{et} \quad I_{\Delta_u^+} = u^t MSMu$$

où S est la matrice de variance empirique

$$S = X^t D_p X.$$

3 Axes principaux d'inertie

3.1 Formulation mathématique

L'objectif est d'obtenir une représentation fidèle du nuage $N(\Omega)$ de $\mathbb{R}^p$ en le projetant sur un espace de faible dimension. Pour ceci, on cherche à minimiser

les « écarts » entre les points de $N(\Omega)$ et leurs projections. Les espaces de représentation choisis sont les variétés linéaires (droite, plan,...). La formulation mathématique de l'ACP est alors la suivante : *Trouver la variété linéaire E_k de dimension k ($k < p$) tel que I_{E_k}, l'inertie du nuage $N(\Omega)$ par rapport à E_k, soit minimum.*

Rappelons que l'on a

$$I_{E_k} = \frac{1}{n}\sum_i d^2(\boldsymbol{x}_i, E_k).$$

En utilisant le théorème de Huygens (2), on peut en déduire que l'espace E_k minimisant I_{E_k} contient nécessairement le centre de gravité du nuage $N(\Omega)$, c'est-à-dire ici l'origine O puisqu'on a supposé le tableau X centré en colonne. E_k est un donc un sous-espace vectoriel. D'autre part, nous savons que dans ce cas, l'inertie totale du nuage I se décompose en une somme $I_{E_k} + I_{E_k^\perp}$ où $I_{E_k^\perp}$ est l'inertie expliquée par E_k. En conséquence, le problème peut s'écrire maintenant : *Trouver le sous-espace vectoriel E_k de dimension k ($k < p$) tel que l'inertie expliquée $I_{E_k^\perp}$ par E_k soit maximum.*

3.2 Résultats préalables

Théorème 1 ([Emboîtement des solutions) *Si E_{k-1} est un sous-espace vectoriel optimal de dimension $k-1$, alors la recherche d'un sous-espace optimal de dimension k peut se faire parmi l'ensemble des sous-espaces vectoriels de dimension k contenant E_{k-1}.*

Preuve : Soit F_k un sous-espace quelconque de dimension k de $\mathbb{R}^p$.
Le sous-espace $F_k \cap E_{k-1}^\perp$ ne peut être réduit au vecteur nul sinon le sous-espace $F_k \oplus E_{k-1}^\perp$ serait de dimension $p+1$. Il existe donc $\mathbf{v} \neq 0 \in F_k \cap E_{k-1}^\perp$. Soit Δv l'axe correspondant et G l'espace supplémentaire M-orthogonal à Δv dans F_k (on a donc $F_k = G \oplus \Delta v$).

Si on note $H = E_{k-1} \oplus \Delta v$, on a

$$I_{F_k^\perp} = I_{G^\perp} + I_{\Delta v^\perp} \text{ car } G \perp \Delta v$$

$$I_{H^\perp} = I_{E_{k-1}^\perp} + I_{\Delta v^\perp} \text{ car } E_{k-1} \perp \Delta v.$$

Mais par hypothèse, E_{k-1} est optimal. On a donc :

$$I_{E_{k-1}^\perp} \geq I_{G^\perp} \Rightarrow I_{H^\perp} \geq I_{F_k^\perp}$$

On peut donc restreindre la recherche d'un sous-espace optimal aux sous-espaces contenant E_{k-1}.

□

Remarquons qu'on n'affirme pas dans ce théorème l'existence d'espaces optimaux.

Théorème 2 *La recherche d'un sous-espace vectoriel optimal E de dimension k contenant un sous-espace F de dimension $k-1$ est équivalente à la recherche d'un axe Δv M-orthogonal à F et maximisant $I_{\Delta v^\perp}$.*

Preuve : On a une décomposition $E = F \oplus \Delta v$ avec $\Delta v \perp F$. On a donc $I_{E^\perp} = I_{F^\perp} + I_{\Delta v^\perp}$. Maximiser $I_{E^\perp}$ est donc équivalent à maximiser $I_{\Delta v^\perp}$.

□

3.3 Résolution du problème

On suppose dans la suite que les vecteurs $\mathbf{u}_j$ sont unitaires. En outre, on sait que pour tout vecteur unitaire $\mathbf{u}$, $I_{\Delta u^\perp}$ est égale à $< \mathbf{u}, SM\mathbf{u} >_M$ ($= \mathbf{u}^t MSM\mathbf{u}$).

À partir des deux théorèmes précédents, il est alors facile de voir que le problème de l'ACP se ramène au problème suivant :

- rechercher un axe Δu_1 maximisant l'inertie $I_{\Delta u_1^\perp} =< \mathbf{u}_1, SM\mathbf{u}_1 >_M$, on note $E_1 = \Delta u_1$;
- rechercher un axe Δu_2, M-orthogonal à E_1 maximisant l'inertie $I_{\Delta u_2^\perp} =< \mathbf{u}_2, SM\mathbf{u}_2 >_M$, on note $E_2 = E_1 \oplus \Delta u_2$;
- ...
- Rechercher un axe Δu_k, M-orthogonal à E_{k-1} maximisant l'inertie $I_{\Delta u_k^\perp} =< \mathbf{u}_k, SM\mathbf{u}_k >_M$, on note $E_k = E_{k-1} \oplus \Delta u_k$.

En posant $B = SM$ et $Q = M$, le théorème de décomposition spectrale d'une matrice (1.5) fournit une réponse à notre problème : les vecteurs propres normés de la matrice SM ordonnés suivant les valeurs propres décroissantes fournissent les axes $\Delta u_1, \ldots, \Delta u_k$, appelés *axes factoriels* ou encore *axes principaux d'inertie* et les inerties $I_{\Delta u_k^\perp}$ portées ou expliquées par ces axes sont égales aux valeurs propres λ_k.

Les espaces E_k sont donc solutions du problème et on obtient du même coup toutes les solutions pour les dimensions inférieures à k.

Par ailleurs, en utilisant les propriétés de la décomposition en valeurs propres et vecteurs propres et si on note U la matrice des vecteurs propres normés rangés en colonne et L la matrice diagonale des valeurs propres rangés dans le même ordre, on a

$$UMU^t = U^tMU = MUU^t = I \quad \text{et} \quad SMU = UL.$$

3.4 Résultats pratiques

Si $\mathbf{u}_1, \ldots, \mathbf{u}_p$ sont les vecteurs propres normés ordonnés suivant les valeurs propres décroissantes de la matrice SM, la solution pour les différentes valeurs de k est la suivante :

- $k = 1 : E_1 = \Delta \mathbf{u}_1$;
- $k = 2 : E_2 = E_1 \oplus \Delta \mathbf{u}_2$;
- ...
- $k : E_k = E_{k-1} \oplus \Delta \mathbf{u}_k$.

On a en outre $I_{\Delta \mathbf{u}_k^\perp} = \lambda_k$.

3.5 Inerties expliquées

Proposition 3 $I_{E_k^\perp} = \lambda_1 + \ldots + \lambda_k$

Preuve : Les vecteurs propres $\mathbf{u}_\alpha$ sont M-orthogonaux (matrice SM M-symétrique). L'espace E_k se décompose donc en une somme directe de sous-espaces M-orthogonaux $\Delta_{\mathbf{u}_\alpha}$, on sait alors que

$$I_{E_k^\perp} = \sum_{\alpha=1}^{k} I_{\Delta \mathbf{u}_\alpha^\perp}$$

Puisque $I_{\Delta \mathbf{u}_\alpha^\perp} = \lambda_\alpha$, le résultat est démontré.

□

Remarque 1 *En prenant $k = p$, on retrouve $I = trace(S)$. De plus, si r est le rang de la matrice $\mathbf{x}$ ($r \leq \min(p, n)$), on a*

$$\lambda_1, \ldots, \lambda_r > 0 \text{ et } \lambda_{r+1}, \ldots, \lambda_p = 0.$$

et par suite

$$I_{E_r^\perp} = I.$$

Finalement, le nuage est exactement dans le sous-espace vectoriel E_r engendré par les r premiers axes factoriels.

La qualité de représentation de l'ensemble initial Ω sur le sous-espace E_k peut être mesurée par le *pourcentage d'inertie* pris en compte par E_k :

$$\frac{\lambda_1 + \ldots + \lambda_k}{\text{trace}(S)} 100.$$

Cet indicateur est souvent utilisé pour choisir le nombre d'axes à retenir.

4 Composantes principales

4.1 Définition

Rappelons que le problème de départ était d'obtenir une représentation du nuage $N(\Omega)$ dans des espaces de dimension réduit. On connaît maintenant les axes définissant ces espaces. Pour pouvoir obtenir les différentes représentations, il suffit de déterminer les coordonnées de la projection de tous les points du nuage sur chaque axe factoriel. On notera $c_{1\alpha}, \ldots, c_{n\alpha}$ les n coordonnées ainsi obtenues avec l'axe α, $\mathbf{c}_\alpha$ le vecteur $(c_{1\alpha}, \ldots, c_{n\alpha})'$, appelé α^e composante principale et C la matrice obtenue en rangeant en colonne les vecteurs $\mathbf{c}_\alpha$. On peut alors obtenir la projection du nuage $N(\Omega)$ dans un plan factoriel quelconque $(\mathbf{u}_\alpha, \mathbf{u}_\beta)$ grâce aux composantes principales $\mathbf{c}_\alpha$ et $\mathbf{c}_\beta$. Par exemple, la représentation dans le premier plan factoriel est obtenue grâce à $\mathbf{c}_1$ et $\mathbf{c}_2$.

Pour $\alpha > r$, λ_α, et donc $I_{\Delta \mathbf{u}_\alpha^\perp}$, est nul ; ce qui entraîne que $\mathbf{c}_\alpha = 0$. Enfin, en exprimant l'inertie expliquée par l'axe α dans la relation $\lambda_\alpha = I_{\Delta \mathbf{u}_\alpha^\perp}$, on obtient la relation

$$\lambda_\alpha = \frac{1}{n} \sum_{i \in \Omega} (c_{i\alpha})^2.$$

4.2 Calcul des composantes principales

Proposition 4 *les composantes principales vérifient*

$$\mathbf{c}_\alpha = XM\mathbf{u}_\alpha$$

qui s'exprime matriciellement

$$C = XMU.$$

Preuve : La nouvelle base est orthonormée : il suffit donc de projeter les $\mathbf{x}_i$ sur les vecteurs de base :

$$c_{i\alpha} = < \boldsymbol{x}_i, \mathbf{u}_\alpha >_M = \boldsymbol{x}_i^t M \mathbf{u}_\alpha$$

$$\mathbf{c}_\alpha = XM\mathbf{u}_\alpha$$

$$C = XMU.$$

□

On peut aussi démontrer cette proposition de la manière suivante.
Preuve : Les composantes principales peuvent être obtenues aussi par changement de base. Si on note $\mathbf{c}_i$ les vecteurs lignes transposés de C, on obtient

$$\boldsymbol{x}_i = U\mathbf{c}_i$$

$$U^t M \boldsymbol{x}_i = U^t MU\mathbf{c}_i = \mathbf{c}_i \text{ car } U^t MU = I$$

$$\mathbf{c}_i^t = \boldsymbol{x}_i^t MU$$

et donc

$$C = XMU$$

□

4.3 Composantes principales : nouvelles variables

Une composante principale associe à chaque individu $\mathbf{x}_i$ de Ω un nombre réel. On peut donc la considérer comme une nouvelle variable. Comme les variables initiales $\mathbf{x}_j$, cette variable appartient à l'espace $\mathbb{R}^n$. Quelques propriétés de ces nouvelles variables peuvent alors être établies :

Proposition 5 *Les composantes principales sont des combinaisons linéaires des variables $\mathbf{x}_j$.*

Preuve : On a $\mathbf{c}_\alpha = XM\mathbf{u}_\alpha = X(M\mathbf{u}_\alpha) = \sum_{j=1}^p a_{\alpha j}\mathbf{x}_j$ si on note $\mathbf{a}_\alpha$ le vecteur $M\mathbf{u}_\alpha$.

□

Proposition 6 *Les composantes principales $\mathbf{c}_\alpha$ sont centrées, de variance λ_α et non corrélées 2 à 2.*

Preuve : Une combinaison linéaire de variables centrées est centrée.

$$\begin{aligned} \text{Cov}(\mathbf{c}_\alpha, \mathbf{c}_\beta) &= < \mathbf{c}_\alpha, \mathbf{c}_\beta >= \mathbf{c}_\alpha^t D_p \mathbf{c}_\beta \\ &= \mathbf{u}_\alpha^t MX^t D_p MX\mathbf{u}_\beta = \mathbf{u}_\alpha^t M(X^t D_p X)M\mathbf{u}_\beta \\ &= \mathbf{u}_\alpha^t MSM\mathbf{u}_\beta = \mathbf{u}_\alpha^t M(S\mathbf{u}_\beta) = \lambda_\beta \mathbf{u}_\alpha^t M\mathbf{u}_\beta = \lambda_\beta < \mathbf{u}_\alpha, \mathbf{u}_\beta > . \end{aligned}$$

$$\text{On en déduit } \begin{cases} Var(\mathbf{c}_\alpha) = \lambda_\alpha & \text{si } \alpha = \beta \\ Cov(\mathbf{c}_\alpha, \mathbf{c}^\beta) = 0 & \text{si } \alpha \neq \beta. \end{cases}$$

□

Dans la nouvelle base, la matrice de variance est donc diagonale : l'ACP revient à diagonaliser la matrice de variance. On peut ainsi poser le problème de l'ACP de manière différente : trouver k nouvelles variables, combinaisons linéaires normés des p variables centrées initiales, non corrélées deux à deux et de variance maximum.

Proposition 7 *Les composantes principales $\mathbf{c}_\alpha$ sont vecteurs propres de la matrice WD_p associées aux valeurs propres λ_α où*

$$W = XMX^t$$

est la la matrice des produits scalaires associés aux vecteurs individus $\boldsymbol{x}_i$.

Preuve : les $\mathbf{u}_\alpha$ et les λ_α étant vecteurs propres de la matrice SM, on peut en déduire

$$SM\mathbf{u}_\alpha = \lambda_\alpha \mathbf{u}_\alpha,$$
$$(X^t D_p X)M\mathbf{u}_\alpha = \lambda_\alpha \mathbf{u}_\alpha.$$

On multiplie à gauche par XM :

$$(XM)X^t D_p XM\mathbf{u}_\alpha = \lambda_\alpha (XM)\mathbf{u}_\alpha,$$
$$XMX^t D_p (XM\mathbf{u}_\alpha) = \lambda_\alpha (XM\mathbf{u}_\alpha),$$
$$XMX^t D_p \mathbf{c}_\alpha = \lambda_\alpha \mathbf{c}_\alpha.$$

□

Si on avait posé directement le problème en terme de recherche de variables, nous aurions obtenu ces variables comme vecteurs propres de la matrice WD_p.

5 Représentation des variables

Dans l'espace des variables, les composantes principales normées $\mathbf{v}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}\mathbf{c}_\alpha$ forment un système de vecteurs orthonormés (une base si $n \geq p$). Dans ce système, les coordonnées des variables initiales normées sont alors simplement les corrélations. La représentation des p variables initiales dans ce système permet de visualiser les liens entre les variables initiales et les liens entre les composantes principales et les variables initiales. Cette représentation est utilisée pour donner une « interprétation » aux axes. Le calcul des ces coordonnées vérifie donc

$$\mathrm{Cor}(\alpha, j) = \mathrm{Cov}\left(\frac{1}{\sigma_j}\mathbf{x}_j, \frac{1}{\sqrt{\lambda_\alpha}}\mathbf{c}_\alpha\right) = \frac{1}{\sigma_j}\frac{1}{\sqrt{\lambda_\alpha}}\mathbf{x}_j^t D_p \mathbf{c}_\alpha.$$

Chapitre 3

Analyse des correspondances

L'étude des liens existant entre deux variables qualitatives s'appuie généralement sur le tableau de contingence qui leur est associé.

1 Tableau de contingence

Pour simplifier les notations, on notera $\sum_i$, $\sum_j$ et $\sum_{i,j}$ pour $\sum_{i\in I}$, $\sum_{j\in J}$ et $\sum_{(i,j)\in I\times J}$.

1.1 Définition

Si I et J représentent les modalités des deux variables qualitatives I et J à r et s modalités, connaissant les valeurs prises par ces deux variables sur un échantillon de taille n, le tableau de contingence est alors défini par la matrice $(n_{ij})_{i,j\in I\times J}$ de dimension (r,s) où n_{ij} représente le nombre d'individus prenant à la fois la modalité i et la modalité j.

On définit alors le tableau des *fréquences relatives* $F=(f_{ij}=n_{ij}/n)$, les *fréquences marginales* $f_{i.}=\sum_j f_{ij}$ et $f_{.j}=\sum_i f_{ij}$, les *lois marginales* $f_I=(f_{1.},\dots f_{r.})'$ et $f_J=(f_{.1},,\dots f_{.s})'$ et les *profils* en ligne et en colonne $f_J^i=\left(\frac{f_{i1}}{f_{i.}},\dots,\frac{f_{is}}{f_{i.}}\right)'$ et $f_I^j=\left(\frac{f_{1j}}{f_{.j}},\dots,\frac{f_{rj}}{f_{.j}}\right)'$.

Remarquons que l'on a $n=\sum_{i,j}n_{ij}$ et $\sum_j f_{.j}=\sum_i f_{i.}=1$.

1.2 Nuages associés

À partir du tableau de contingence, on peut alors définir le *nuage des profils en lignes* $\mathcal{N}(I)=\{(f_J^i,f_{i.}),i\in I\}$ dans R^r muni de la métrique du χ^2, métrique quadratique définie par la matrice diagonale $D_{\frac{1}{f_J}}$ de terme général $\frac{1}{f_{.j}}$ et vérifiant $d^2(f_J^i,f_J^{i'})=\sum_j\frac{1}{f_{.j}}(\frac{f_{ij}}{f_{i.}}-\frac{f_{i'j}}{f_{i'.}})^2$. Ce nuage possède des propriétés intéressantes : son centre de gravité est la loi marginale f_J et il appartient à l'hyperplan orthogonal à f_J passant par f_J.¨Pour vérifier cette propriéte, il suffit de calculer le produit scalaire $<f_J^i-f_J,f_J>$ et de constater qu'il s'annule pour tout i.

De façon symétrique, on défint le *nuage des profils en colonnes* $\mathcal{N}(J)=\{(f_I^j,f_{.j}),j\in J\}$ dans R^r muni de la métrique du χ^2 définie cette fois par

par la matrice diagonale $D_{\frac{1}{f_I}}$ de terme général $\frac{1}{f_{i.}}$ et vérifiant $d^2(f_I^j,f_I^{j'})=\sum_i\frac{1}{f_{i.}}(\frac{f_{ij}}{f_{.j}}-\frac{f_{ij'}}{f_{.j'}})^2$. Ce nuage possède évidemment les mêmes propriétés que le nuage $\mathcal{N}(I)$.

1.3 Équivalence distributionnelle

Si on remplace dans le tableau de contingence deux colonnes ayant le même profil, c.-à-d. proportionnelles, par leur somme, les distances entres les profils du nuage $\mathcal{N}(I)$ et du nuage $\mathcal{N}(J)$ ne sont pas modifiées. La propriété est évidemment encore vraie si on regroupe deux lignes de même profil.

2 Analyse des correspondances

2.1 Définition

A.C.P des nuage $\mathcal{N}(I)$ et $\mathcal{N}(J)$

2.2 Quelques résultats préalables

Afin de faciliter l'application des résultats matriciels de l'ACP et d'obtenir les formules classiques de l'AFC, nous donnons quelques formules matricielles très pratiques. Si on note 1_r et 1_s les vecteurs colonne de dimensions r et s composés uniquement de 1 et D_{f_I} la matrice diagonale de dimension (n,n) dont le terme général est $f_{i.}$, on peut montrer les relations

$$F'.1_r=f_J \qquad F.1_s=f_I,$$

$$D_{\frac{1}{f_J}}f_J=1_s \qquad D_{\frac{1}{f_I}}f_I=1_r,$$

et

$$1_s'D_{f_I}1_r=1.$$

2.3 ACP de $N(I)$

Axes factoriels

Calcul Ce sont les vecteurs propres normés $\mathbf{u}_\alpha$ de la matrice $Z=SD_{\frac{1}{f_J}}$ où S est la matrice de variance. Sachant que le tableau centré associé au nuage des profils en ligne s'écrit $X=D_{\frac{1}{f_I}}F-1_nf_J'$, on obtient $S=F'D_{\frac{1}{f_I}}F-f_Jf_J'$ et finalement

$$Z=F'D_{\frac{1}{f_I}}FD_{\frac{1}{f_J}}-f_Jf_J'D_{\frac{1}{f_J}}.$$

Lien avec les vecteurs et valeurs propres de $Z'=F'D_{\frac{1}{f_I}}FD_{\frac{1}{f_J}}$ On peut facilement montrer que f_J est un axe factoriel associé à une valeur propre nulle ; on peut en déduire :

- que l'axe factoriel trivial f_J est aussi vecteur propre de la matrice Z', mais cette fois avec la valeur propre 1 ;
- que les axes factoriels non triviaux sont aussi vecteurs propres de Z' avec les mêmes valeurs propres.

Cette propriété est quelquefois utilisée pour calculer les axes factoriels. La diagonalisation de la matrice Z' permet généralement de retrouver les axes factoriels en supprimant simplement le vecteur propre correspondant à la valeur propre 1 (axe factoriel trivial). Toutefois, lorsque la valeur propre 1 est multiple (cas rare mais se produisant lorsque le tableau F a une structure en blocs diagonaux), ce calcul ne conduit pas aux bons résultats. Il est donc conseillé de diagonaliser directement la matrice Z.

Propriétés

$$\sum_j \frac{1}{f_{.j}}(u_\alpha^j)^2 = 1, \quad \sum_j \frac{1}{f_{.j}} u_\alpha^j u_\beta^j = 0, \quad \sum_j u_\alpha^j = 0$$

Preuve : les deux premières propriétés sont les propriétés habituelles des axes factoriels d'une ACP. La dernière, vraie pour les axes factoriels non triviaux, provient de l'orthogonalité de ces axes avec f_J :

$$0 =< \mathbf{u}_\alpha, f_J >= \sum_j u_\alpha^j 1/f_{.j} f_{.j} = \sum_j u_\alpha^j.$$

Facteurs

Rappelons que les facteurs sont les opérateurs de projections sur les axes factoriels $\mathbf{u}_\alpha$. On a

$$\mathbf{a}_\alpha = D_{\frac{1}{f_J}} \mathbf{u}_\alpha$$

À partir des propriétés précédentes des axes factoriels, on peut en déduire

$$\sum_j f_{.j}(a_\alpha^j)^2 = 1, \quad \sum_j f_{.j} a_\alpha^j a_\beta^j = 0, \quad \sum_j f_{.j} a_\alpha^j = 0.$$

Si on considère l'espace J comme un ensemble muni de la distribution de probabilité f_J, les facteurs peuvent ainsi être interprétés comme un ensemble de fonctions sur J d'espérance nulle, de variance unitaire et non corrélées.

On peut aussi montrer que les facteurs non triviaux $\mathbf{a}_\alpha$ sont les vecteurs propres normées de $D_{\frac{1}{f_J}} F' D_{\frac{1}{f_I}} F$ vérifiant $\sum_j f_{.j} a_\alpha^j = 0$.

Composantes principales

Elles sont définies à partir des axes factoriels par la relation habituelle de l'ACP :

$$\mathbf{c}_\alpha = D_{\frac{1}{f_I}} F D_{\frac{1}{f_J}} \mathbf{u}_\alpha = D_{\frac{1}{f_I}} F \mathbf{a}_\alpha.$$

On peut en déduire la relation suivante entre les facteurs et les composantes principales :

$$\mathbf{a}_\alpha = \frac{1}{\lambda_\alpha} D_{\frac{1}{f_J}} F' \mathbf{c}_\alpha.$$

2.4 ACP du nuage $N(J)$

On effectue de manière symétrique l'A.C.P du nuage $N(J)$. Les axes factoriels $\mathbf{v}_\alpha$ sont les vecteurs propres normés $\mathbf{v}_\alpha$ de la matrice

$$F D_{\frac{1}{f_J}} F D_{\frac{1}{f_I}} - f_I f_I' D_{\frac{1}{f_I}}.$$

Comme précédemment, les axes factoriels non triviaux sont les vecteurs propres de

$$F D_{\frac{1}{f_J}} F D_{\frac{1}{f_I}},$$

et vérifient les relations

$$\sum_i \frac{1}{f_{i.}}(v_\alpha^i)^2 = 1, \quad \sum_i \frac{1}{f_{i.}} v_\alpha^i v_\beta^i = 0, \quad \sum_i v_\alpha^i = 0.$$

Les facteurs $\mathbf{b}_\alpha$ sont définis à partir des axes par la relation

$$\mathbf{b}_\alpha = D_{\frac{1}{f_I}} \mathbf{v}_\alpha,$$

et vérifient

$$\sum_i f_{i.}(b_\alpha^i)^2 = 1, \quad \sum_j f_{.j} b_\alpha^i b_\beta^i = 0, \quad \sum_j f_{.j} b_\alpha^i = 0.$$

Les facteurs non triviaux sont les vecteurs propres de $D_{\frac{1}{f_I}} F D_{\frac{1}{f_J}} F'$ vérifiant $\sum_i f_{i.} b_\alpha^j = 0$.

Les composantes principales $\mathbf{d}_\alpha$ sont définies à partir des axes $\mathbf{v}_\alpha$ par la relation

$$\mathbf{d}_\alpha = D_{\frac{1}{f_J}} F' D_{\frac{1}{f_I}} \mathbf{v}_\alpha = D_{\frac{1}{f_J}} F \mathbf{b}_\alpha$$

et vérifient

$$\mathbf{b}_\alpha = \frac{1}{\lambda_\alpha} D_{\frac{1}{f_I}} F \mathbf{d}_\alpha.$$

Chapitre 4

Positionnement multidimensionnel

1 Introduction

Lorsque les données sont fournies sous la forme d'un ensemble d'individus mesurés par un ensemble de variables, l'analyse en composantes principales et les méthodes qui en sont issues comme l'analyse des correspondances et l'analyse des correspondances multiples fournissent une représentation fidèle des données dans des espaces euclidiens de faible dimension permettant, par exemple, de visualiser les données sur un plan.

L'analyse des proximités, encore appelée positionnement multidimensionnel (*multidimensional scaling*) ou analyse ordinale (en écologie par exemple), a aussi pour objectif d'obtenir une représentation fidèle des données dans des espaces euclidiens de faible dimension, souvent le plan, mais cette fois à partir d'un tableau de proximités entre les individus. Historiquement, ces méthodes ont été développées et proposées dans la revue Psychometrika dans les années 1950 par Torgerson et Shepard. Parmi les références portant sur l'analyse des proximités, on peut citer les deux ouvrages récents Borg and Groenen (1997) et Cox and Cox (1994).

De manière générale, dans tout ce chapitre, les résultats seront bien sûr toujours déterminés aux isométries près (translations, rotations, symétries,...).

2 Tableaux de proximités

Un tableau de proximité est un tableau carré de nombres mesurant une ressemblance ou une dissemblance entre les éléments d'un ensemble Ω. On peut citer par exemple les tableaux de distances géographiques, les tableaux de distances routières, les tableaux de durées du trajet par le train, les tableaux de corrélations entre variables.

2.1 Types de proximités

Une *distance* d sur un ensemble Ω est une application de $\Omega \times \Omega$ dans $\mathbb{R}^+$ vérifiant les propriétés suivantes :

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \Omega \qquad d(x, \boldsymbol{y}) = 0 \Longleftrightarrow \boldsymbol{x} = \boldsymbol{y} \quad \text{(séparation)}$$

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \Omega \qquad d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x}) \qquad \text{(symétrie)}$$

$$\forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \Omega \qquad d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) d(\boldsymbol{y}, \boldsymbol{z}) \qquad \text{(inégalité triangulaire)}$$

En analyse des données, il n'est pas toujours nécessaire d'avoir toutes ces propriétés et les mesures suivantes sont souvent suffisantes.

Une *mesure de dissimilarité* sur un ensemble Ω est une fonction d de $\Omega \times \Omega$ dans $\mathbb{R}^+$ vérifiant

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \Omega \qquad d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$$

$$\forall \boldsymbol{x} \in \Omega \qquad d(\boldsymbol{x}, \boldsymbol{x}) = 0.$$

Une *mesure de similarité* sur un ensemble Ω est une fonction s de $\Omega \times \Omega$ dans $\mathbb{R}^+$ vérifiant

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \Omega \qquad s(\boldsymbol{x}, \boldsymbol{y}) = s(\boldsymbol{y}, \boldsymbol{x})$$

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \Omega, \boldsymbol{x} \neq \boldsymbol{y} \qquad s(\boldsymbol{x}, \boldsymbol{x}) = s_{max} \quad \text{avec} \quad s_{max} \geq s(\boldsymbol{x}, \boldsymbol{y}).$$

Remarquons qu'il est facile de transformer un indice de similarité s en un indice de dissimilarité en posant $d(\boldsymbol{x}, \boldsymbol{y}) = s_{max} - s(\boldsymbol{x}, \boldsymbol{y})$.

Enfin, terminons en citant la distance ultramétrique, fondamentale pour l'étude de la classification hiérarchique.

Une *ultramétrique* sur un ensemble Ω est une fonction d de $\Omega \times \Omega$ dans $\mathbb{R}^+$ vérifiant

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \Omega \quad d(\boldsymbol{x}, \boldsymbol{y}) = 0 \Longleftrightarrow \boldsymbol{x} = \boldsymbol{y} \qquad \text{(séparation)}$$

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \Omega \quad d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x}) \qquad \text{(symétrie)}$$

$$\forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \Omega \quad d(\boldsymbol{x}, \boldsymbol{z}) \leq \max(d(\boldsymbol{x}, \boldsymbol{y}), d(\boldsymbol{y}, \boldsymbol{z})) \qquad \text{(inégalité ultramétrique)}$$

Il est facile de montrer que la propriété d'inégalité ultramétrique entraîne la propriété d'inégalité triangulaire. Une ultramétrique est donc une distance.

2.2 Constitution d'un tableau de proximités

Un tableau de proximités peut être issu directement du recueil des données, par exemple les tableaux de distances routières, ou peuvent être obtenus à partir d'autres tableaux. À partir de variables quantitatives, il est possible d'utiliser toutes les distances définies sur $\mathbb{R}^p$.

- Distance euclidienne : $d^2(\boldsymbol{x}, \boldsymbol{y}) = \sum_j (\boldsymbol{x}_j - \boldsymbol{y}_j)^2 = (\boldsymbol{x} - \boldsymbol{y})' I (\boldsymbol{x} - \boldsymbol{y})$
- Distance euclidienne pondérée : $d^2(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} - \boldsymbol{y})' D (\boldsymbol{x} - \boldsymbol{y})$
- distance de Mahalanobis : $d^2(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} - \boldsymbol{y})' S^{-1} (\boldsymbol{x} - \boldsymbol{y})$ où S est la matrice de variance empirique
- distance « city-block » ou distance L^1 : $d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j=1}^p |x_j - y_j|$
- distance de Chebychev ou distance L^∞ : $d(\boldsymbol{x}, \boldsymbol{y}) = \max_j |x_j - y_j|$.
- distance de Minkowski L_p : $d(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{j=1}^p (x_j - y_j)^p \right)^{1/p}$ (il s'agit de la généralisation des distances précédentes : L_1=city-block, L_2=euclidienne, L_∞=chebychev).

Une distance entre variables peut être définie à partir de la corrélation : $d = 1 - r^2$.

À partir de variables qualitatives nominales, il est possible d'utilisée la distance du χ^2 ou, plus simplement, la distance $d = 1-$ proportion de modalités communes. Cette dernière peut être généralisée en en utilisant une table de ressemblance entre modalités.

À partir de variables qualitatives ordinales , il est possible d'utiliser la distance euclidienne sur les rangs renormalisés entre 0 et 1.

Pour les variables binaires, si a, b, c et d représente le nombre de fois où les individus $\mathbf{x}$ et $\mathbf{y}$ ont répondu respectivement $(1,1)$, $(1,0)$, $(0,1)$ et $(0,0)$ aux variables binaires, alors toute une série de mesures de proximité ont été proposées, par exemple :

- $d(\boldsymbol{x},\boldsymbol{y}) = \frac{2a}{2a+b+c}$ (Csekanowski, Sorensen, Dice) ;
- $d(\boldsymbol{x},\boldsymbol{y}) = \frac{a-(b+c)+d}{a+b+c+d}$ (Hamman) ;
- $d(\boldsymbol{x},\boldsymbol{y}) = \frac{a}{a+b+c}$ (Jaccard) ;
- $d(\boldsymbol{x},\boldsymbol{y}) = \frac{a}{a+b}$ (Kulezynsk) ;
- $d(\boldsymbol{x},\boldsymbol{y}) = \frac{a}{[(a+b)(a+c)]^{1/2}}$ (Ochiai).

2.3 Transformation

Il existe de nombreux moyens de passer d'un type de proximités à un autre. Par exemple, la relation $s_{ii'} = (r_{ii'} + r_{i'i})/2$ permet de symétriser une proximité qui ne l'était pas ; la relation $d^2_{ii'} = s_{ii} - 2s_{ii'} + s_{i'i'}$ (Mardia et al. (1979)) permet de transformer une mesure de similarité en distance euclidienne ; la transformation $d_{ii'} + c$ où c est le maximum des valeurs $|d_{ij} + d_{jk} - d_{ik}|$ permet de transformer une dissimilarité en distance.

2.4 Utilisation

Les mesures de proximités peuvent être intégrées dans les méthodes (ACP, ACM, méthode des centres-mobiles, discrimination linéaire ou quadratique,...) ou peut être la donnée de base de la méthode (AFTD, MDS, classification hiérarchique). Lorsqu'il n'existe pas de méthode adaptée à un type de données, il est toujours possible de définir une proximité cohérente avec les données et d'appliquer ce dernier type de méthodes.

3 Le problème

Étant donnée une matrice de dissimilarités Δ sur n individus, l'objectif est de déterminer une représentation X de dimension p donnée telle que la distance euclidienne associée $D(X)$ à cette représentation soit proche de la dissimilarité initiale Δ.

Avant d'étudier les méthodes proposées pour résoudre ce problème, le paragraphe suivant est consacré à l'étude de quelques propriétés des distances euclidiennes.

4 Distances euclidiennes

On notera dans la suite $Q_n = I_n - \frac{1}{n}U_n$ la matrice associée à la projection orthogonale sur le vecteur $(1,...,1)'$ dans l'espace $\mathbb{R}^n$. Cette matrice correspond à l'opérateur de centrage en colonne d'une matrice de dimension (n,p).

4.1 Équivalence entre distances euclidiennes et produits scalaires

On considère X la matrice de données centrée associée à un ensemble de n points de $\mathbb{R}^p$; on note $D^2 = D^2(X)$ la matrice des distances au carré entre les n points, $W = W(X)$ la matrice XX' des produits scalaires et $\mathbf{h} = (\|\boldsymbol{x}_1\|^2, \ldots, \|\boldsymbol{x}_n\|^2)' = \mathrm{diag}(W)$.

Proposition 1 *D^2 est une fonction de W*

Preuve : Il suffit de développer :

$$d^2_{ii'} =< \boldsymbol{x}_i - \boldsymbol{x}_{i'}, \boldsymbol{x}_i - \boldsymbol{x}_{i'} >= \|\boldsymbol{x}_i\|^2 - 2w_{ii'} + \|\boldsymbol{x}_{i'}\|^2$$

que l'on peut noter matriciellement

$$D^2 = \mathbf{h}\mathbb{1}'_n - 2W + \mathbb{1}_n\mathbf{h}' = \mathrm{diag}(W)\mathbb{1}'_n - 2W + \mathbb{1}_n\,\mathrm{diag}(W)'.$$

□

Proposition 2 *W est une fonction de D^2.*

Preuve : Partant de la relation précédente, on obtient

$$-\frac{1}{2}Q_nD^2Q_n = -\frac{1}{2}Q_n\mathbf{h}\mathbb{1}'_nQ_n + Q_nXX'Q_n - \frac{1}{2}Q_n\mathbb{1}_n\mathbf{h}'Qn.$$

Sachant que $Q\mathbb{1}_n = 0$ et que $QX = X$ puisque la matrice X est déjà centrée, on en déduit

$$-\frac{1}{2}Q_nD^2Q_n = XX' = W.$$

□

L'expression $Q_nD^2Q_n$ est quelquefois appelée double-centrage.

Conséquence Les deux propositions précédentes montrent qu'il existe une bijection entre la matrice des distances et la matrice des produits scalaires. On notera $D^2 = \varphi(W)$ et $W = \varphi^{-1}(D^2)$ les fonctions associées.

4.2 Matrice de distances euclidiennes

On dira qu'une matrice de dissimilarités Δ est une matrice de distances euclidiennes si et seulement s'il existe une représentation X des n individus dans un espace $\mathbb{R}^p$ telle que la distance euclidienne associée soit la distance Δ, c'est-à-dire, en notant $D(X)$ la matrice des distances euclidiennes associées à un tableau X, si et seulement si il existe X tel que $D(X) = \Delta$.

4.3 CNS pour qu'une matrice de dissimilarités soit euclidienne

Proposition 3 *Une matrice de dissimilarités Δ est euclidienne si et seulement si $W = -\frac{1}{2}Q_n\Delta^2Q_n$ est une matrice semi-définie positive (SDP). En outre, la représentation associée est contenue dans un espace de dimension $p \leq n-1$.*

Preuve : La proposition directe est immédiate car dans ce cas W est une matrice de produits scalaires et est donc semi-définie positive.

Réciproque : soient V la matrice des vecteurs propres normés (au sens de $\frac{1}{n}I$) de $\frac{1}{n}W$ et L la matrice diagonale formée des valeurs propres. Ces matrices vérifient $\frac{1}{n}WV = VL$, $VV' = \frac{1}{n}I$ et $VLV' = W$. W étant SDP, les valeurs propres de $\frac{1}{n}W$ sont positives et on peut définir $X = V\sqrt{L}$.

La matrice $VL = \frac{1}{n}WV$ est de la forme Q_nA ; elle est donc centrée en colonne. On peut alors en déduire que V et donc X sont aussi centrées en colonne. La matrice des produits scalaires associée à la représentation X s'écrit donc XX' et on a

$$XX' = (V\sqrt{L})(V\sqrt{L})' = VLV' = W.$$

c'est-à-dire

$$W(X) = \varphi^{-1}(\Delta^2)$$

et donc

$$D^2(X) = \Delta^2$$

ce qui montre que Δ est une matrice de distance euclidienne.

Le rang de la matrice W est inférieur ou égal à $n-1$. La dimension de la représentation associée X, égale au nombre de valeurs propres non nulles de WD^{-p}, est donc inférieure ou égale à $n-1$.

□

5 Analyse factorielle d'un tableau de distances

Cette méthode est historiquement la première technique de positionnement multidimensionnel et a été développée par Togerson (1952). Elle est aussi connue sous les noms d'analyse du triple (Benzecri (1973)), de codage en composantes principales, de *p*rincipal coordinate analysis ou encore de *c*lassical scaling.

5.1 $W = -\frac{1}{2}Q_n\Delta^2Q_n$ est SDP

Nous venons de voir que dans ce cas, il existait une représentation euclidienne X exacte de dimension $\leq n-1$. Pour obtenir une représentation de dimension p fixée, il suffit alors d'utiliser l'ACP sur X et de retenir les p premiers axes. Mais comme les composantes principales sont les vecteurs propres ordonnés de norme λ_α de $\frac{1}{n}W$, il suffit dans la construction de X d'ordonner les vecteurs propres pour que la matrice des composantes principales C ne soit rien d'autre que X.

En pratique, il faudra donc :

1. calculer la matrice $W = \varphi^{-1}(\Delta^2) = -\frac{1}{2}Q_n\Delta^2Q_n$,

2. diagonaliser la matrice $\frac{1}{n}W$,
3. ordonner les valeurs propres et vecteurs propres et normer les vecteurs propres (au sens de $\frac{1}{n}I$ (si les vecteurs propres étaient normés au sens habituel, il suffit de les multiplier par $\sqrt{n}$),
4. calculer les composantes principales $C = V\sqrt{L}$ où L et V sont les matrices associées à ces valeurs propres et vecteurs propres
5. utiliser œs résultats comme pour une ACP classique (pourcentage d'inertie, choix du nombre d'axes,...).

La vérification de l'hypothèse $W = -\frac{1}{2}Q_n\Delta^2Q_n$ est SDP se fait a posteriori ; il faut et il suffit que toutes les valeurs propres sont positives ou nulles.

5.2 $W = -\frac{1}{2}Q_nD^2Q_n$ n'est pas SDP

Lorsqu'il existe des valeurs propres négatives, plusieurs stratégies peuvent être envisagées :

Application directe de l'AFTD

L'AFTD est utilisée normalement comme s'il existait une représentation euclidienne et seules les composantes principales associées aux valeurs propres positives sont utilisées. Les résultats seront en pratique assez bons si les valeurs propres négatives sont petites (en valeur absolue). Toutefois, la définition du pourcentage d'inertie expliquée par un axe ne convient plus puisque la somme des valeurs propres positives est supérieure à la somme totale des valeurs propres. Généralement, la somme des valeurs propres est remplacée par la somme des valeurs absolues des valeurs propres.

Transformation de la dissimilarité en distance

Il existe différents moyens. Par exemple, On peut en additionnant une certaine constante à la dissimilarité initiale la transformer en une distance. On peut alors appliquer sur cette distance l'AFTD. En pratique cette méthode ne donne pas toujours de très bons résultats.

5.3 L'AFTD dans R

L'AFTD peut être réalisée à l'aide de la fonction `cmdscale` qui permet, en particulier, l'introduction de la constante signalée dans le paragraphe précédent.

5.4 Un exemple

Nous avons appliqué l'AFTD aux données d'Ekman portant sur la couleur. Comme il s'agit d'un tableau de similarité, la première chose à faire est de transformer en un tableau de dissimilarités. Voilà le tableau obtenu :

	L434	L445	L465	L472	L490	L504	L537	L555	L584	L600	L610	L628	L651	L674
L434	0.00	0.14	0.6	0.6	0.82	0.94	0.9	1.0	0.98	0.93	0.9	0.88	0.87	0.84
L445	0.14	0.00	0.5	0.6	0.78	0.91	0.9	0.9	0.98	0.96	0.9	0.89	0.87	0.86
L465	0.58	0.50	0.0	0.2	0.53	0.83	0.9	0.9	0.98	0.99	1.0	0.99	0.95	0.97
L472	0.58	0.56	0.2	0.0	0.46	0.75	0.9	0.9	0.98	0.99	1.0	0.99	0.98	0.96
L490	0.82	0.78	0.5	0.5	0.00	0.39	0.7	0.7	0.93	0.98	1.0	0.99	0.98	1.00
L504	0.94	0.91	0.8	0.8	0.39	0.00	0.4	0.6	0.86	0.92	1.0	0.98	0.98	0.99
L537	0.93	0.93	0.9	0.9	0.69	0.38	0.0	0.3	0.78	0.86	0.9	0.98	0.98	1.00
L555	0.96	0.93	0.9	0.9	0.74	0.55	0.3	0.0	0.67	0.81	1.0	0.97	0.98	0.98
L584	0.98	0.98	1.0	1.0	0.93	0.86	0.8	0.7	0.00	0.42	0.6	0.73	0.80	0.77
L600	0.93	0.96	1.0	1.0	0.98	0.92	0.9	0.8	0.42	0.00	0.3	0.50	0.59	0.72
L610	0.91	0.93	1.0	1.0	0.98	0.98	0.9	1.0	0.63	0.26	0.0	0.24	0.38	0.45
L628	0.88	0.89	1.0	1.0	0.99	0.98	1.0	1.0	0.73	0.50	0.2	0.00	0.15	0.32
L651	0.87	0.87	0.9	1.0	0.98	0.98	1.0	1.0	0.80	0.59	0.4	0.15	0.00	0.24
L674	0.84	0.86	1.0	1.0	1.00	0.99	1.0	1.0	0.77	0.72	0.4	0.32	0.24	0.00

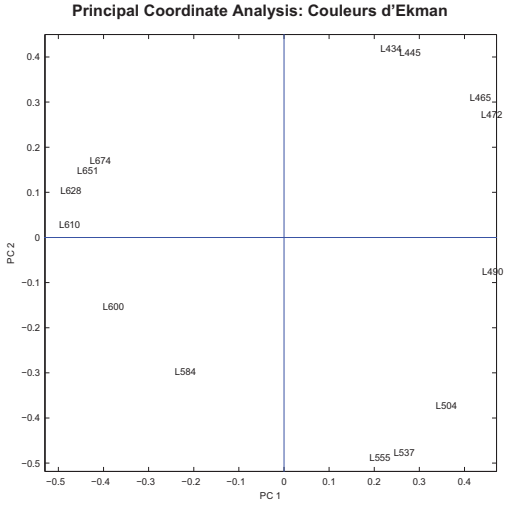
Voici les valeurs propres fournies :

α	1	2	3	4	5	6	7	8	9	10	11	12	13	14
λ_α	0.14	0.06	0.03	0.027	0.011	0.007	0.003	0.0023	0.0013	0.0003	0.0001	0	−0.002	−0.003
%	45.2	29.7	10.1	8.5	3.6	2.3	0.95	0.72	0.42	0.09	0.03	0	−0.61	−1.1
	45.2	74.9	85.0	93.5	97.1	99.5	100.4	101.1	101.6	101.7	101.7	101.7	101.1	100

les coordonnées des 5 premières composantes principales :

Ind.	*PC1*	*PC2*	*PC3*	*PC4*	*PC5*
L434	0.21	0.42	0.23	−0.18	−0.09
L445	0.26	0.41	0.20	−0.17	−0.04
L465	0.41	0.31	−0.01	0.18	0.07
L472	0.44	0.27	−0.07	0.20	0.06
L490	0.44	−0.08	−0.27	0.16	−0.03
L504	0.34	−0.37	−0.23	−0.06	−0.09
L537	0.24	−0.48	0.03	−0.21	−0.06
L555	0.19	−0.49	0.15	−0.15	0.11
L584	−0.24	−0.30	0.28	0.22	0.17
L600	−0.40	−0.15	0.19	0.21	−0.16
L610	−0.50	0.03	−0.03	0.11	−0.13
L628	−0.50	0.10	−0.13	−0.05	−0.04
L651	−0.46	0.15	−0.19	−0.11	0.02
L674	−0.43	0.17	−0.16	−0.16	0.20

et la représentation obtenue dans le premier plan :



On peut remarquer qu'il y a des valeurs propres négatives (la matrice de dissimilarité initiale n'est pas euclidienne) mais qu'elles sont très petites et ne sont par génantes. Par ailleurs, la représentation dans le premier plan fournit une très bonne représentation des données.

6 Méthodes non linéaires

Il est possible de montrer que la solution obtenue par l'AFTD minimise le critère $\sum_{i,i'}(\delta_{ii'}^2 - d_{ii'}^2)$ sous la contrainte que la représentation est de dimension p fixée et qu'en plus $d_{ii'} \leq \delta_{ii'}$ pour tous les couples d'individus i, i'. Cette méthode revient à projeter dans un espace de faible dimension une représentation parfaite dans un espace de grande dimension de la distance initiale et donc finalement à effectuer une transformation linéaire des données initiales. Les méthodes développées dans la suite n'imposeront plus que la représentation obtenue soit une projection linéaire. L'objectif de ces méthodes sera donc de trouver une représentation euclidienne X dans un espace de dimension fixée k telle que la distance euclidienne D associée minimise une fonction d'écart entre Δ et D appelée *S*tress.

6.1 Fonctions Stress

Plusieurs fonctions ont été proposées :

$$Stress_1(X) = \frac{\sum_{i<i'}(\delta_{ii'} - d_{ii'})^2}{\sum_{i<i'} d_{ii'}^2}$$

$$Stress_2(X) = \frac{\sum_{i<i'} w_{ii'}(\delta_{ii'} - d_{ii'})^2}{\sum_{i<i'} w_{ii'} d_{ii'}^2}$$

où les $w_{ii'}$ sont des pondérations données a priori (ces pondérations permettent de prendre en compte, par exemple, la présence de données manquantes).

$$Stress_3(X) = \frac{1}{\sum_{i<i'} \delta_{ii'}} \sum_{i<i'} \frac{(\delta_{ii'} - d_{ii'})^2}{\delta_{ii'}}$$

Tous ces critères sont normalisés de manière à être invariants pour des rotations, translations et changements d'échelles. Remarquons que le dernier critère prend en compte de manière plus importante les erreurs commises sur les petites distances.

6.2 Optimisation

Il n'existe pas d'algorithme permettant de résoudre en toute généralité ce type de problème et le plus souvent, les méthodes proposées sont des méthodes d'optimisation itératives qui font simplement décroître le critère et conduisent donc à des optima locaux du critère. On peut citer les méthodes suivantes :

- méthodes de gradient ;
- méthode SMACOF (la plus efficace à ce jour) ;
- méthode de Newton.

6.3 Projection de Sammon

La projection de Sammon, très utilisée dans le monde de la Rdf, utilise le critère $Stress_3$ et la méthode de Newton. En R, la fonction `sammon` est disponible dans le module `MASS`.

6.4 Remarques

- Le choix du nombre de dimension se fait généralement, comme pour l'ACP, en étudiant la décroissance du critère en fonction de la dimension (méthode du coude). Toutefois, contrairement à l'AFTD, les calculs doivent être recommencés pour chaque dimension et les solutions ne sont pas emboîtées.
- Cette approche ne pose le problème des valeurs propres négatives comme pour l'AFTD ; quelque soit la dissimilarité initiale, une solution est obtenue. Toutefois, la méthode ne garantit pas l'optimum global et donc l'unicité de la solution. Généralement les logiciels prennent comme point de départ les résultats obtenus par l'AFTD.
- En dehors du critère minimisé, un certain nombre d'outils permettent d'analyser les résultats. On peut citer, par exemple, le graphique représentant les couples δ_{ij}, d_{ij}.

7 Méthodes non métriques ou ordinales

7.1 Généralisation

L'approche précédente peut être étendue en relâchant les contraintes du problème. L'idée sous-jacente est qu'en relâchant le lien entre la dissimilarité et la distance obtenue, le résultat soit plus fidèle. Pour ceci, une fonction supplémentaire f est introduite dans le critère de la façon suivante :

$$Stress(X, f) = \frac{\sum_{i<i'}(f(\delta_{ii'}) - d_{ii'})^2}{\sum_{i<i'} d_{ii'}^2}.$$

L'objectif est alors de déterminer le couple (X, f) minimisant ce critère. Plusieurs situations ont été envisagées ; par exemple

- f est une fonction linéaire $f(d_{ii'}) = \alpha d_{ii'} + \beta$
- f est une fonction exponentielle $f(d_{ii'}) = e^{\alpha d_{ii'} + \beta}$
- f est simplement une fonction monotone croissante : le critère ne prend en compte que l'ordre induit sur tous les couples d'individus par la dissimilarité initiale.

La solution du problème est obtenue par optimisation alternée :

- pour f fixée, on cherche la meilleure représentation X ; pour cela, il suffit d'appliquer l'une des méthodes précédentes à la dissimilarité $f(\Delta)$;
- pour X fixée, on cherche la meilleure fonction f ; il s'agit alors d'un problème de régression.

7.2 Projection de Kruskal

Dans cette méthode, développée par Shepard et Kruskal et connue sous le nom de *Non metric muldidimensional scaling*, la fonction f est simplement monotone croissante et l'algorithme de régression est un algorithme original appelé

régression isotonique. En R, la fonction correspondante `isoMDS` est disponible dans le module `MASS`.

Comme pour les méthodes précédentes, des outils d'analyse, comme le diagramme de Shepard, ont été développés.

8 Quelques remarques

8.1 Dissimilarités initiales

La dissimilarité initiale Δ peut recouvrir de nombreuses situations. En particulier, ces méthodes peuvent être utilisées pour pour étudier les liens existant entre les variables, par exemple en partant d'une distance entre variables définie à partir des corrélations.

8.2 Autres méthodes

On peut citer quelques méthodes voisines : par exemple, l'analyse procustéenne permet de comparer deux tableaux de dissimilarités et si il y a plus de deux tableaux de dissimilarités, les méthode de dépliage (*unfolding method*) permettent de comparer ces différents tableaux et la méthode Indscal (*I*ndividual differences) permet de représenter simultanément les tableaux et les individus sur lesquels portent ces dissimilarités.

Chapitre 5

La méthode des k-means

1 Introduction

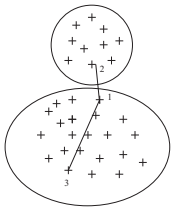
L'objectif de la classification automatique est l'organisation en classes homogènes des éléments d'un ensemble Ω. Les structure de la classification recherchée peut être variée mais nous nous limitons ici à la recherche d'une partition de Ω.

Pour définir cette notion de classes homogènes, on utilise le plus souvent une mesure de similarité (ou de dissimilarité) sur Ω. Par exemple, si d est une mesure de dissimilarité sur Ω, on peut caractériser cette homogénéité en imposant aux classes de la partition recherchée de vérifier la propriété suivante :

$$\forall x, y \in \text{ même classe et } \forall z, t \in \text{ classes différentes } \Rightarrow d(x, y) < d(z, t).$$

Cette propriété signifie simplement que l'on cherche à obtenir des classes telles que deux points d'une même classe se ressemblent plus que deux points de classes différentes.

En pratique, cet objectif est inutilisable. Par exemple sur la figure 1, alors qu'on « distingue clairement » deux classes, la distance entre les deux points 1 et 3 situés dans une même classe est supérieure à la distance entre les deux points 1 et 2 pourtant classés séparément.



On remplace généralement cette condition trop exigeante par une fonction numérique qui mesurera la qualité d'homogénéité d'une partition. Cette fonction est appelée généralement *critère*. Le problème peut paraître alors très simple. En effet, par exemple, dans le cas de la recherche d'une partition, il suffit de chercher parmi l'ensemble fini de toutes les partitions celle qui optimise le critère

numérique. Malheureusement, le nombre de ces partitions étant très grand, leur énumération est impossible dans un temps raisonnable (explosion combinatoire). On utilise alors des heuristiques qui donnent, non pas la meilleure solution, mais une « bonne solution », c'est-à-dire une solution proche de la solution optimale. On parle alors d'optimisation locale. Lorsqu'il existe une structure d'ordre sur l'ensemble Ω et que celle-ci doit être respectée par la partition, il existe un algorithme de programmation dynamique, appelé algorithme de Fisher, qui fournit la solution optimale.

2 Le critère d'inertie intra-classe

Dans tout ce chapitre, on supposera que l'ensemble Ω à classifier correspond à un ensemble de n individus mesurés par p variables quantitatives. L'ensemble Ω peut être vu comme un nuage de points dans $\mathbb{R}^p$ muni des pondérations $\frac{1}{n}$ et de la distance euclidienne usuelle.

Dans cette situation, le critère de classification le plus utilisé est certainement le critère d'inertie intra-classe.

De manière analogue à la définition de la matrice de variance qui peut alors s'écrire dans cette situation

$$S = \frac{1}{n}(X - \mathbb{1}_n\overline{\boldsymbol{x}})^t(X - \mathbb{1}_n\overline{\boldsymbol{x}}) = \frac{1}{n}\sum_i(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^t$$

et à celle de l'inertie $I = \frac{1}{n}\sum_i d^2(\boldsymbol{x}_i, \overline{\boldsymbol{x}})$ qui vérifie $I = \text{tr}(S)$, il est possible d'associer à une partition $P = (P_1, \ldots, P_K)$ de Ω en g classes

- la matrice de variance intra-classe

$$S_W = \frac{1}{n}\sum_k n_k S_k$$

où S_k est la matrice de variance de chaque classe ($S_k = \frac{1}{n_k}(X_k - \mathbb{1}_n\overline{\boldsymbol{x}}_k)^t(X_k - \mathbb{1}_n\overline{\boldsymbol{x}}_k)$), X_k est la matrice X réduite aux lignes correspondant à la classe k et $\overline{\boldsymbol{x}}_k$ est le centre de gravité de la classe k, et

- l'inertie intra-classe

$$I_W = \sum_k I(P_k)$$

où $I(P_k) = \frac{1}{n}\sum_{i \in P_k} d^2(\boldsymbol{x}_i, \overline{\boldsymbol{x}})$ est l'inertie de la classe k.

Remarquons que ces deux notions sont reliées par la relation

$$I_W = \text{tr}(S_W).$$

L'inertie intra-classe peut alors être utilisé comme critère de classification : une partition sera d'autant plus homogène que l'inertie intra-classe sera proche de 0 ; en particulier, ce critère sera nul si tous les points de chaque classe sont concentrés en un même point.

3 L'algorithme des k-means

L'algorithme des k-means MacQueen (1967), encore connue sous le nom de méthode des centres mobiles ou méthode de réallocation-centragecentres-mobiles peut alors se définir ainsi :

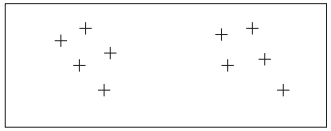
1. Tirage au hasard de g points de Ω qui forment les centres initiaux des g classes.
2. Tant que non convergence
   (a) construction de la partition suivante en affectant chaque point de Ω à la classe dont il est le plus près du centre (en cas d'égalité, l'affectation se fait à la classe de plus petit indice).
   (b) les centres de gravité de la partition qui vient d'être calculée deviennent les nouveaux centres.

Si $L = (\lambda_1, \ldots, \lambda_K)$ représente un K-uple de $\mathbb{R}^p$ et $P = (P_1, \ldots, P_K)$ une partition de Ω en K classes, la suite construite par l'algorithme peut être notée sous la forme :

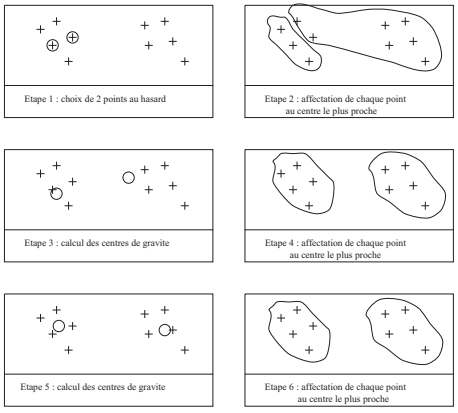
$$L^0 \to P^1 \to L^1 \to P^2 \to L^2 \to \ldots \to P^n \to L^n \to \ldots$$

4 Exemple

Les données sont constituées d'un ensemble Ω de 10 points du plan.



L'algorithme des centres-mobiles peut alors se résumer de la façon suivante :



La poursuite de cet algorithme ne changera plus les résultats : l'algorithme a convergé. Remarquons que la classification obtenue correspond effectivement à la structure en deux classes observable visuellement. Nous allons maintenant définir et étudier les propriétés de cet algorithme.

5 Convergence

La qualité d'un couple partition-centres est mesurée par la somme des inerties des classes par rapport à leur centre :

$$C(P, L) = \sum_k I(P_k, \boldsymbol{\lambda}_k) = \frac{1}{n} \sum_k \sum_{\boldsymbol{x} \in P_k} d^2(\boldsymbol{x}, \boldsymbol{\lambda}_k)$$

où $P = (P_1, P_2, \ldots, P_g)$ et $L = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_g)$.

On peut montrer qu'à chacune des deux étapes de l'algorithme, on améliore le critère C. Plus précisément, on a les relations suivantes :

Proposition 1

$$C(P^{n+1}, L^n) \leq C(P^n, L^n)$$
$$C(P^{n+1}, L^{n+1}) \leq C(P^{n+1}, L^n)$$

Preuve : Le critère $C(P, L)$ peut s'écrire :

$$C(P, L) = \frac{1}{n} \sum_{\boldsymbol{x} \in \Omega} d^2(\boldsymbol{x}, \boldsymbol{\lambda}_{k(\mathbf{x})})$$

où $k(\boldsymbol{x})$ est le numéro de la classe à laquelle appartient $\boldsymbol{x}$ dans la partition P.

Lorsque l'on compare les expressions $C(P^{n+1}, L^n)$ et $C(P^n, L^n)$, les centres des classes ne bougent pas et comme P^{n+1} est construit en associant chaque point de Ω au meilleur centre, la relation 1 est vraie.

Le critère $C(P, L)$ s'écrit aussi :

$$C(P, L) = \sum_k I(P_k, \boldsymbol{\lambda}_k).$$

L^{n+1} est par définition de l'algorithme des centres mobiles formé des g centres de gravité des classes de P^{n+1}. Or, la propriété d'optimalité du centre de gravité (voir théorème de Huygens) entraîne l'inégalité

$$I(P_k^{n+1}, \boldsymbol{\lambda}_k^{n+1}) \leq I(P_k^{n+1}, \boldsymbol{\lambda}_k^n).$$

L'inéquation 1 est donc démontrée.

□

Corollaire 2 *La suite numérique $C(P^n, L^n)$ est une suite stationnaire.*

Preuve : Les deux inégalités de la propriété 1 entraîne la décroissante de la suite $C(P^n, L^n)$. Le nombre de partitions en K classes d'un ensemble fini est fini. En outre, l'ensemble contenant les éléments L^n, formés par construction de centres de classes d'un ensemble fini est aussi fini. Par conséquent, la suite $C(P^n, L^n)$ est une suite décroissante qui ne peut prendre qu'un ensemble fini de valeurs. Elle est donc stationnaire.

□

Proposition 3 *La suite (P^n, L^n) est une suite stationnaire.*

Remarquons tout d'abord que la stationnarité de $C(P^n, L^n)$ n'entraîne pas forcément la stationnarité de (P^n, L^n). En effet, il serait tout à fait possible d'avoir une suite de partitions et de centres ayant la forme suivante :

$$\ldots, P, L, P', L', P, L, \ldots, P, L, P', L', \ldots$$

avec

$$P \neq P' \text{ et } L \neq L' \text{ et } C(P, L) = C(P', L) = C(P, L').$$

Preuve : $C(P^n, L^n)$ est stationnaire, c'est-à-dire :

$$\exists N \text{ t.q } \forall n > N \qquad C(P^n, L^n) = C(P^{n+1}, L^{n+1}).$$

Cette relation entraîne d'après la propriété 1

$$\exists N \text{ t.q } \forall n > N \qquad C(P^n, L^n) = C(P^{n+1}, L^n) = C(P^{n+1}, L^{n+1}).$$

Sachant que pour tout k on a nécessairement $I(P_k^{n+1}, \lambda_k^n) \geq I(P_k^{n+1}, \lambda_k^{n+1})$ (propriété du centre de gravité), l'égalité précédente entraîne les égalités $I(P_k^{n+1}, \lambda_k^n) \geq I(P_k^{n+1}, \lambda_k^{n+1})$ pour tout k et comme le centre de gravité est l'unique point de $\mathbb{R}^p$ minimisant l'inertie de P_k^{n+1}, on obtient $\boldsymbol{\lambda}_k^{n+1} = \boldsymbol{\lambda}_k^n$ et donc $L^{n+1} = L^n$.
Comme par construction, P^n est définie de manière unique à partir de L^n, l'égalité

$$L^{n+1} = L^n$$

entraîne aussi l'égalité

$$P^{n+1} = P^n.$$

□

Remarque : Finalement, si notre objectif initial avait été de trouver le couple (P, L) minimisant le critère C, l'algorithme des centres-mobiles ne fournit pas nécessairement le meilleur résultat, mais simplement une suite de couples dont la valeur du critère va en décroissant. On parle alors d'« optimisation locale ».

Plus précisément, l'algorithme des centres-mobiles est un algorithme d'optimisation alternée. En effet, il est facile de montrer que les deux étapes de l'algorithme des centres mobiles vérifie les deux définitions suivantes :

- recherche de la partition : minimisation de $C(P, L)$ avec L fixé ;
- recherche des centres : minimisation de $C(P, L)$ avec P fixée.

En pratique, la convergence est atteinte très vite (souvent moins de 10 itérations même avec des données de taille importante).

6 Lien avec le critère d'inertie intra-classe

Puisque L^n est fonction de P^n, il est possible d'exprimer le critère $C(P^n, L^n)$ uniquement en fonction de P^n :

$$C(P^n, L^n) = \sum_k I(P_k^n, \boldsymbol{\lambda}_k^n) = \sum_k I(P_k^n)$$

puisque $\boldsymbol{\lambda}_k^n$ est le centre de gravité de la classe P_k^n. Et en conséquence

$$C(P^n, L^n) = I_W(P^n).$$

Finalement, l'algorithme des centres mobiles défini de manière algorithmique se révèle être un algorithme dont l'objectif est la recherche de la partition en g classes minimisant le critère d'inertie intra-classe.

7 Exploitation des optima locaux

Sachant que suivant les points de départ choisis, les résultat seront différents, il reste à exploiter ces différentes résultats. Plusieurs solutions ont été proposées : On fait différents essais de l'algorithme en tirant au hasard plusieurs initialisations. Plusieurs stratégies sont alors possibles. Soit retenir la meilleure partition, c'est-à-dire celle qui optimise le critère, soit utiliser l'ensemble des résultats pour en déduire les groupes stables (« méthode des formes fortes ») ; On sélectionne une « bonne » initialisation à l'aide d'informations supplémentaires ou à l'aide d'une procédure automatique (points les plus éloignés les uns des autres, zones de forte densité...). Il faut toutefois faire un compromis entre le temps nécessaire à la recherche de la configuration initiale et celui nécessaire à l'algorithme proprement dit ; Il est aussi possible d'utiliser un certain nombre de méthodes stochastiques comme le recuit simulé qui, sans garantir l'optimum global, possèdent des propriétés de convergence asymptotique.

8 Nombre de classes

Le critère n'est pas indépendant du nombre de classes. Par exemple, la partition en n classes où chaque point forme une classe a un critère d'inertie intra-classe nul et est donc, de ce point de vue, la partition optimale ce qui est sans intérêt. Il est donc nécessaire de fixer a priori le nombre de classes. Si ce nombre de classes n'est pas connu, plusieurs solutions permettant de résoudre ce problème très difficile sont utilisées. Par exemple, on recherche la meilleure partition pour plusieurs nombres de classes et on étudie la décroissance du critère en fonction du nombre de classes pour sélectionner le nombre de classes (« méthode du coude »). Une autre procédure consiste à pénaliser le critère de classification par une fonction dépendant du nombre de classes rendant ainsi le critère « indépendant » de ce nombre de classes. Il est aussi possible d'ajouter des contraintes supplémentaires portant, par exemple sur le nombre d'individus par classe ou sur le volume d'une classe. C'est l'option retenue par la méthode Isodata. D'autres approches enfin utilisent les tests statistiques.

Chapitre 6

Modèles probabilistes en classification

1 Introduction

Rappelons que l'objectif de la classification automatique est la recherche de classes « homogènes ». Nous avons vu que cet objectif conduisait à des algorithmes souvent conçus d'un point de vue heuristique et utilisant des critères métriques. Ainsi, les deux algorithmes de classification sans doute les plus utilisés, l'algorithme des centres-mobiles (ou *k-means*) pour la recherche de partitions et l'algorithme de classification ascendante hiérarchique de Ward pour la recherche de hiérarchies utilisent tous deux l'inertie intraclasse d'une partition, c'est-à-dire la somme des inerties de chaque classe. La difficulté de cette approche est la justification du choix de la métrique et du critère utilisés. Pour mettre en place de telles solutions, il est donc nécessaire de choisir d'une part, une métrique mesurant la dissimilarité entre les objets de l'ensemble à classifier et d'autre part, un critère défini à partir de cette métrique mesurant de degré de cohésion et de séparation des classes.

Tout ceci a conduit depuis quelques années à une évolution de l'approche algorithmique, heuristique et géométrique vers une approche plus statistique qui utilise des *modèles probabilistes de classification* pour formaliser l'idée intuitive de la notion de classe naturelle.

Cette approche permet d'analyser de manière précise et de donner une interprétation statistique à certains critères métriques dont les différentes variantes n'étaient pas toujours bien claires (comme par exemple les critères d'inertie $\mathrm{tr}(S_W)$ et $|S_W|$ définis à partir de la matrice de variance intraclasse S_W) et permet en outre de proposer de nouvelles variantes répondant à des hypothèses précises. Elle fournit aussi un cadre formel pour proposer des solutions à des problèmes difficiles comme la détermination du nombre de classes ou encore la validation de la structure de classification obtenue.

Remarquons enfin que, très souvent l'ensemble à classifier n'est qu'un échantillon d'une population beaucoup plus grande et que les conclusions obtenues à partir de la classification de cet échantillon sont souvent étendues à toute la population. Dans ce cas, la classification n'a pas de sens sans un recours à un modèle probabiliste permettant de justifier cette inférence.

2 Approches probabilistes de la classification

L'hypothèse faite dans toute approche probabiliste de la classification automatique est de considérer que les données forment un échantillon aléatoire $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, issu d'une population, et de s'appuyer sur l'analyse de la distribution de probabilité de cette population pour définir une classification. Différentes approches probabilistes de la classification ont été envisagées parmi lesquelles on peut distinguer les approches paramétriques et les approches non paramétriques.

2.1 Approches paramétriques

Une première approche consiste à faire des hypothèses sur la distribution de probabilité induisant une classification et formalisant ainsi la notion de classes « naturelles ». On peut distinguer plusieurs types de modèles paramétriques de la classification ; les plus importants étant les modèles de mélange fini de lois de probabilités, les modèles fonctionnels à effet fixe et les processus ponctuels.

Modèles de mélange fini

Les modèles de mélange fini, qui supposent que chaque classe est caractérisée par une distribution de probabilité, sont des modèles très souples permettant de prendre en compte des situations variées comme la présence de population hétérogène ou d'éléments aberrants. Grâce à l'algorithme d'estimation EM, particulièrement adapté à cette situation, les modèles de mélange ont fait l'objet de nombreux développements en statistique. Leur utilisation pour la classification a été considérée par de nombreux auteurs. Cette approche se justifie pour plusieurs raisons : elle correspond souvent à l'idée intuitive que l'on peut se faire d'une population composée de plusieurs classes ; elle possède des liens forts avec des méthodes de références comme l'algorithme des *k-means* ; enfin, elle est capable de prendre en compte de manière assez naturelle de nombreuses situations particulières. C'est l'approche qui va être développée dans ce chapitre.

Les modèles fonctionnels à effet fixe

Les modèles fonctionnels à effet fixe caractérisés par l'équation :

$$\text{Données} = \text{Structure} + \text{Erreur}$$

où la structure est inconnue mais fixe et l'erreur est aléatoire, peuvent être appliqués à la classification en choisissant une structure adaptée. Si les données sont des vecteurs $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ de $\mathbb{R}^p$, le modèle $\boldsymbol{x}_i = \boldsymbol{y}_i + \boldsymbol{\varepsilon}_i$, où on impose aux $\boldsymbol{y}_i$ d'appartenir à un ensemble de g centres $\{\mathbf{a}_1, \ldots, \mathbf{a}_g\}$ et aux erreurs $\boldsymbol{\varepsilon}_i$ de suivre une loi normale centrée de même variance, en est l'exemple le plus simple. On peut aussi appliquer ce type de modèles à des données de similarité et supposer, par exemple, que la dissimilarité d entre deux objets de l'ensemble à classifier s'écrit sous la forme $d(a,b) = \delta(a,b) + \varepsilon(a,b)$ où δ est une distance ultramétrique. Degens montre ainsi que la classification ascendante hiérarchique du lien moyen est un maximum local de la vraisemblance de ce modèle lorsque l'erreur est gaussienne.

Les processus ponctuels

En statistique spatiale, les données qui peuvent être, par exemple, la répartition des arbres dans une forêt ou des étoiles dans l'espace, sont considérées comme des semis de point issus de processus ponctuels. Certains de ces processus correspondent à une organisation en agrégats et peuvent être considérés comme des modèles probabilistes associés à une classification. Le plus utilisé est le processus de Neyman-Scott qui peut être interprété comme une génération des données en trois étapes : (a) g points $\mathbf{a}_1, \ldots, \mathbf{a}_g$ sont tirés au hasard suivant une distribution uniforme sur une région convexe ; (b) les tailles $n_1, \ldots, n_g$ des classes sont tirées au hasard, par exemple à l'aide d'une distribution de Poisson ; (c) pour chaque classe k, n_k points sont tirés au hasard en utilisant une distribution sphérique centrée en $\mathbf{a}_k$, par exemple une distribution gaussienne de moyenne $\mathbf{a}_k$.

2.2 Approches non paramétriques

Cette seconde approche regroupe les méthodes probabilistes de classification qui ne supposent aucune hypothèse sur la distribution de probabilités. Ces méthodes peuvent être très variées mais vont toutes s'appuyer sur la forme de cette distribution. Lorsque les données sont continues, cette distribution est caractérisée par sa fonction de densité et ces méthodes utilisent alors cette fonction de densité pour définir la notion de classes, par exemple des classes de forte densité ou des classes modales. Hartigan définit ainsi une classe de forte densité comme un sous-ensemble connexe de points de densité supérieure à un certain seuil et, en faisant varier ce seuil, il obtient un arbre hiérarchique de classes. La présence de plusieurs maxima de la densité peut être interprétée comme la présence de données hétérogènes et donc de classes. La recherche de ces maxima et l'affectation des points de l'espace de référence à chacun d'entre eux permet alors de définir les classes modales.

L'application de ces méthodes nécessite évidemment l'estimation de la distribution inconnue à partir des données. Les méthodes les plus courantes s'appuient sur une estimation non paramétrique de la densité comme l'estimation par la méthode des plus proches voisins, par la méthode des noyaux ou même simplement à l'aide d'un histogramme. Cette démarche a donné lieu à de nombreux algorithmes et des liens avec des algorithmes classiques comme l'algorithme de classification hiérarchique du lien minimal ont pu être établis. Nous ne détaillerons pas plus ces méthodes dans ce chapitre.

On pourrait aussi ranger dans ces approches non paramétriques les algorithmes de classification hiérarchique de Lerman définis à partir de la notion de vraisemblance du lien.

2.3 Validation

Une autre utilisation importante des outils probabilistes en classification concerne la validation des résultats. En effet, tout algorithme de classification fournissant toujours un résultat, il est nécessaire de savoir si ce résultat correspond à une véritable structure ou s'il est simplement le fait du hasard ; pour ceci, des outils de validation statistique ont été développés. Lorsque les résultats ont été obtenus à l'aide des modèles paramétriques précédents, une approche

naturelle consiste à s'appuyer sur ces modèles pour définir de tels outils en vérifiant, par exemple, la normalité d'une classe dans le cas des modèles de mélanges gaussiens. Il existe aussi d'autres outils statistiques de validation qui sont indépendants des algorithmes de classification. On peut citer, par exemple, l'utilisation du modèle des graphes aléatoires de Ling pour tester la significativité d'une classe ; B. Van Cutsem et B. Ycart ont aussi étudié les structures de classification produites par différents algorithmes sous des hypothèses de non-classifiabilité aléatoire et ont proposé une batterie de tests explicites de non-classifiabilité. Pour avoir plus d'information, on pourra se reporter aux travaux de Bock qui donne une synthèse bibliographique détaillée de ces problèmes de validation.

2.4 Notations

Dans tout ce chapitre, nous supposerons que les données se présentent sous la forme d'un échantillon $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ où chaque individu est mesuré par un vecteur $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$. Les données sont ainsi caractérisées par une matrice de dimension (n, p) que nous noterons aussi $\mathbf{x}$ définie par les nombres x_{ij} où i décrit un ensemble I de n individus et j décrit un ensemble de p variables qui pourront être continues ou qualitatives.

L'objectif sera donc la recherche d'une partition $\mathbf{z}$ en g classes de l'ensemble I. Le nombre de classes g sera supposé connu.

Nous utiliserons les notations $\mathbf{z} = (z_1, \ldots, z_n)$, où $z_i \in \{1, \ldots, g\}$ indique la classe de l'objet i et $\mathbf{z} = (z_{11}, \ldots, z_{ng})$ avec $z_{ik} = 1$ si i appartient à la classe k et $z_{ik} = 0$ sinon. Dans ce dernier cas, la classification sera représentée par une matrice $\mathbf{z}$ vérifiant $z_{ik} \in \{0, 1\}$ et $\sum_{k=1}^{g} z_{ik} = 1$. Ainsi, la classe k correspond à l'ensemble des objets i tel que $z_i = k$ ou encore $z_{ik} = 1$ et $z_{i\ell} = 0\ \forall \ell \neq k$. Enfin, on notera n_k le cardinal de la classe k.

Par exemple, si l'ensemble I est constitué de 5 éléments, pour la partition $\mathbf{z}$ constituée des 2 classes $\{1, 3, 4\}$ et $\{2, 5\}$, on aura :

$$\mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 1 \\ 1 \\ 2 \end{pmatrix} \quad \text{notée aussi} \quad \mathbf{z} = \begin{pmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \\ z_{31} & z_{32} \\ z_{41} & z_{52} \\ z_{51} & z_{52} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$$

et $n_1 = 3$ et $n_2 = 2$.

3 Le modèle de mélange

3.1 Introduction

Depuis leur utilisation par Newcomb en 1886 pour la détection de points aberrants, puis par Pearson en 1894 pour l'identification de deux populations de crabes, les mélanges finis de distributions ont permis de modéliser une grande variété de phénomènes aléatoires. Ces modèles supposent que les mesures sont effectuées sur un ensemble d'individus provenant de différentes classes dont l'origine est inconnue. Une étude portant sur la migration des passereaux permet

d'illustrer cette situation : des mesures rapides, pour ne pas perturber les oiseaux, sont effectuées ; par exemple, nous disposons de la longueur de l'aile mais pas du sexe, plus difficile à identifier. Or l'étude statistique doit tenir compte de l'origine, mâle ou femelle, des oiseaux. Les données sont reportées dans le tableau 6.1 et la figure 6.1 représente graphiquement les fréquences.

Longueur	82	83	84	85	86	87	88	89	90
Fréquence	5	3	12	36	55	45	21	13	15
Longueur	91	92	93	94	95	96	97	98	
Fréquence	34	59	48	16	12	6	0	1	

TABLE 6.1 – Données passereau

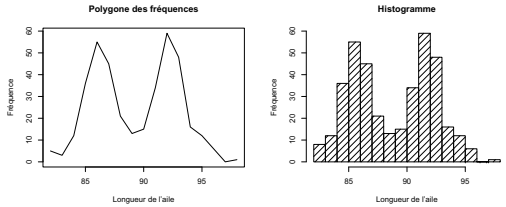


FIGURE 6.1 – Longueurs des ailes en mm de 381 passereaux

Ce type de modèle permet donc de représenter l'hétérogénéité d'une population et sera tout particulièrement adapté au problème de la classification. Ce domaine a fait l'objet de très nombreux travaux ; le livre récent de McLachlan and Peel (2000) constitue une référence très détaillée de ce domaine qui s'est beaucoup développé ces dernières années. Dans ce paragraphe, nous rappelons succinctement le modèle et les problèmes posés par l'estimation de ses paramètres.

3.2 Le modèle

Si on reprend l'exemple précédent, les données peuvent être modélisées à l'aide d'un couple de variables aléatoires (X, Z) où X est la variable aléatoire continue associée à la longueur de l'aile et Z la variable aléatoire discrète associée au sexe. La distribution de probabilité d'un tel couple est définie par une fonction $f(x, z)$ vérifiant

$$P(X \in I, Z \in A) = \sum_{z \in A} \int_I f(x, z)dx$$

où I est un intervalle réel et A est un sous-ensemble de l'ensemble {mâle, femelle}. Remarquons que la fonction $f(x, z)$ n'est ni une densité ni une probabilité ; mais $f(., z)$ pour z fixée est une densité et $f(x, .)$ pour x fixée est une probabilité. Ces deux fonctions définissent les distributions conditionnelles.

Si l'on cherche à ajuster ce modèle, comme on ne dispose que d'un échantillon de la variable aléatoire X, les valeurs de la variable Z étant manquantes,

l'estimation des paramètres du modèle devra se faire à partir de la loi de X, c'est-à-dire d'une loi marginale qui peut être obtenue à partir de la loi du couple (X, Z) par la relation :

$$f(x) = \sum_{z=1}^{2} f(x, z) = \sum_{z=1}^{2} p(z)f(x/z) = \sum_{z=1}^{2} \pi_z f_z(x)$$

où $\pi_z = P(Z = z)$, f_z est la densité de X conditionnellement à $Z = z$ et les valeurs de la variable Z ont été codées 1 et 2. La loi marginale de X est appelée *distribution de mélange fini à 2 composants*.

De manière plus générale, dans la suite les données disponibles pourront être des vecteurs de p mesures et la variable Z pourra prendre un nombre fini quelconque g de valeurs codées $1, \ldots, g$; les données $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ constitueront donc un échantillon de n réalisations indépendantes d'un vecteur aléatoire $\boldsymbol{X}$ à valeur dans $\mathbb{R}^p$ dont la fonction de densité peut s'écrire sous la forme :

$$f(\boldsymbol{x}) = \sum_{z=1}^{g} \pi_z f_z(\boldsymbol{x})$$

où g est le nombre de composants, les f_z sont les densités de chacun des composants et les π_z sont les proportions du mélange ($\pi_z \in]0, 1[$ $\forall z$ et $\sum_z \pi_z = 1$).

Une interprétation de ce modèle de mélange consiste à considérer que, connaissant les proportions $\pi_1, \ldots, \pi_g$ et les distributions f_k de chaque classe, les données sont générées suivant le mécanisme suivant :

- z : chaque individu est rangé dans une classe suivant les probabilités $\pi_1, \ldots, \pi_g$;
- $\boldsymbol{x}$: chaque $\boldsymbol{x}_i$ suit la loi de probabilité associée à la classe à laquelle il appartient.

Généralement, on suppose en plus que les densités f_k des composants appartiennent à une famille paramétrée $f(., \alpha)$; la densité du mélange peut alors s'écrire :

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_z \pi_z f(\boldsymbol{x}; \alpha_z), \quad \forall i \in I,$$

plus souvent notée

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_k \pi_k f(\boldsymbol{x}, \alpha_k),$$

où $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_g, \alpha_1, \ldots, \alpha_g)$ est le paramètre du modèle.

Dans l'exemple des passereaux, il est certain qu'il existe une variable sexe ; la seule inconnue est la valeur de cette variable pour les individus de l'échantillon. En pratique, on utilisera le modèle de mélange dans des situations où l'existence même d'une telle variable n'est pas sûre. Par exemple, dans une étude portant sur les programmes de vaccination contre les oreillons, une enquête a fourni la log-concentration d'anticorps de 385 enfants non vaccinés contre les oreillons, tous âgés de 14 ans. L'histogramme de cette distribution est fourni dans la figure 6.2. Un mode important apparaît autour de la valeur 3 et un second mode, moins net, semble aussi apparaître autour de la valeur 0. Pour ce type de données, il est connu qu'une population homogène aurait du conduire à une distribution sensiblement gaussienne. Comme l'immunisation peut être obtenue par vaccination, ce qui n'est pas le cas dans cet échantillon, ou, naturellement, par contact avec le virus, une explication raisonnable des deux modes

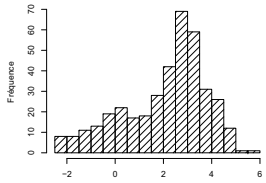


FIGURE 6.2 – Histogramme de la log-concentration d'anticorps contre les oreillons de 385 enfants de 14 ans

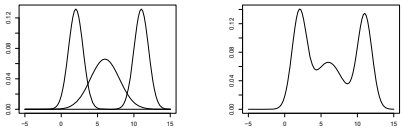
serait que la population est un mélange de deux groupes : les enfants immunisés naturellement et les enfants non immunisés. Contrairement à l'exemple précédent, cette fois les deux groupes sont moins séparés mais surtout, l'existence de deux groupes, qui avait une signification physique incontestable dans le premier cas, n'est maintenant qu'une hypothèse de travail suggérée par les données et qui n'est pas directement observable.

3.3 Exemples

La densité d'un modèle de mélange de deux densités normales dans $\mathbb{R}$ s'écrit

$$f(x;\pi,\mu_1,\mu_2,\sigma_1^2,\sigma_2^2) = \pi\varphi(x;\mu_1,\sigma_1^2) + (1-\pi)\varphi(x;\mu_2,\sigma_2^2)$$

où $\varphi(.;\mu,\sigma^2)$ est la densité de la loi normale univariée de moyenne μ et de variance σ^2. La figure 6.3 (a) représente les fonctions de densité de trois lois normales monodimensionnelles et figure 6.3(b) représente la fonction de densité mélange ainsi obtenue. Dans la figure 6.4, nous avons reporté la loi mélange



(a) Les composantes (b) Le mélange

FIGURE 6.3 – Mélange gaussien dans $\mathbb{R}$

correspondant aux paramètre suivants :
- (a) : $p1 = 0.5$, $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 2$, $\sigma_2 = 1$
- (b) : $p1 = 0.25$, $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 3$, $\sigma_2 = 1$
- (c) : $p1 = 0.8$, $\mu_1 = 1$, $\sigma_1 = 1$, $\mu_2 = 1$, $\sigma_2 = 4$
- (c) : $p1 = 0.6$, $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 2$, $\sigma_2 = 2$

- (d) : $p1 = 0.9$, $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 2.5$, $\sigma_2 = 0.2$
- (e) : $p1 = 0.6$, $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 2.5$, $\sigma_2 = 1$.

Ces exemples permettent d'illustrer la richesse des situations qui peuvent être modélisées par une loi mélange. Un dernier exemple de mélange gaussien, cette

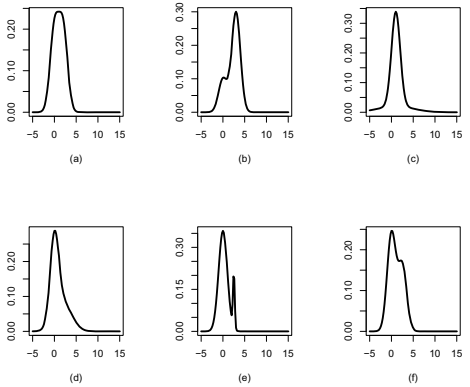


FIGURE 6.4 – Mélanges gaussiens dans $\mathbb{R}$

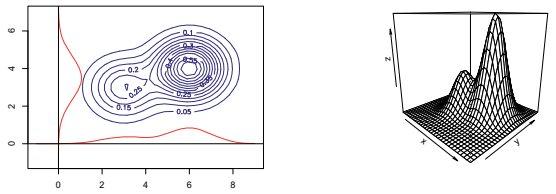
fois dans $\mathbb{R}^2$ est donné dans la figure 6.5.

3.4 Estimation des paramètres

L'estimation des paramètres du modèle de mélange a fait l'objet de nombreuses approches depuis les travaux de Pearson qui, pour estimer les 5 paramètres d'un modèle de mélanges gaussiens unidimensionnel à 2 composants $(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\pi)$ par la méthode des moments, fut conduit à résoudre des équations polynomiales de degré 9. Depuis, ce problème a fait l'objet de nombreux travaux et différentes méthodes d'estimation ont été envisagées : en dehors de la méthode des moments déjà citée, on retrouve des méthodes graphiques, la méthode du maximum de vraisemblance et des approches bayésiennes. La méthode la plus utilisée aujourd'hui est sans doute celle du maximum de vraisemblance à l'aide de l'algorithme EM. Avant de développer cette approche, nous allons évoquer quelques difficultés posées par l'estimation des paramètres d'un modèle de mélange.

3.5 Nombre de composants

Dans un certain nombre de situations, comme celle des oiseaux du paragraphe 3.1 où la notion de composant a une signification physique bien précise,



(a) Lignes de niveau (b) Graphe de densité en perspective

FIGURE 6.5 – Mélange gaussien dans $\mathbb{R}^2$

le nombre de composants peut être parfaitement déterminé mais, le plus souvent, ce nombre inconnu est un paramètre supplémentaire qu'il faut aussi estimer.

Remarquons que, si l'on considère le nombre de composants comme un paramètre supplémentaire, le modèle de mélange peut être vu comme un compromis semi-paramétrique entre un problème d'estimation paramétrique classique quand le nombre de composant est égal à une constante fixée et un problème d'estimation non paramétrique, ici par la méthode des noyaux, quand le nombre de composants est égal à la taille de l'échantillon.

Nous supposerons dans la suite que le nombre g de composants est connu et nous verrons plus loin les solutions proposées pour effectuer ce choix difficile.

3.6 Identifiabilité

Pour que ce problème ait un intérêt, il est nécessaire que la densité du mélange soit identifiable, c'est-à-dire que deux mélanges ayant la même densité correspondent exactement aux mêmes paramètres. De nombreux travaux ont été menés sur ce problème. Plusieurs difficultés apparaissent. La première est due à la numérotation des classes ; par exemple, dans le cas d'un mélange de deux composants, les paramètres $(\pi_1, \pi_2, \alpha_1, \alpha_2)$ et $(\pi_2, \pi_1, \alpha_2, \alpha_1)$, bien que différents, conduisent évidemment à la même densité : la densité d'un mélange n'est donc jamais identifiable ! Les difficultés entraînées par cette situation dépendront des algorithmes d'estimation. Par exemple, avec l'algorithme EM que nous utiliserons, cette difficulté n'est pas gênante – ce qui n'est pas le cas pour l'approche bayésienne où cette situation est connue sous le nom de « *switching problem* ». La seconde difficulté, celle-ci beaucoup plus gênante, peut provenir de la forme même des densités des composants. On peut ainsi facilement vérifier qu'un mélange de lois uniformes ou de lois binomiales n'est pas identifiable. Par contre, les mélanges gaussiens, exponentiels et de Poisson sont identifiables.

3.7 Estimation du maximum de vraisemblance

Rappelons que cette méthode, très largement utilisée, consiste à maximiser la log-vraisemblance

$$L(\boldsymbol{\theta}; \mathbf{x}) = \ln \left(\prod_i f(\boldsymbol{x}_i; \boldsymbol{\theta}) \right) = \sum_i \ln \left(\sum_k \pi_k f(\boldsymbol{x}_i; \boldsymbol{\alpha}_k) \right).$$

L'annulation des dérivées partielles conduit aux *équations de vraisemblance*.

Dans le cas du modèle de mélange, cette approche pose plusieurs difficultés :

- la résolution des équations de vraisemblance ainsi obtenues ne conduit pas à une solution analytique et une méthode de maximisation itérative, comme la méthode de Newton-Raphson ou l'algorithme EM de Dempster, Laird et Rubin (1977) est nécessaire ;
- il existe souvent de nombreux maxima locaux de la fonction de vraisemblance et les algorithmes précédents convergent vers un des ces maxima ce qui ne garantit pas l'obtention du maximum global de vraisemblance ;
- enfin, souvent la fonction de vraisemblance n'est pas bornée supérieurement : il n'y a donc pas de maximum de vraisemblance. Par exemple, dans le cas d'un modèle gaussien monodimensionnel, cette situation se produit si la variance d'un des composants tend vers 0. On traitera ce type de problème en empêchant l'algorithme d'atteindre ces solutions inintéressantes.

4 Algorithme EM

Lorsque la maximisation de la vraisemblance ne conduit pas à une solution analytique, l'algorithme EM est une méthode de maximisation itérative souvent simple à mettre en place. L'idée fondamentale est de considérer que les données observées $\boldsymbol{x}$ ne correspondent qu'à une connaissance partielle de données $\mathbf{y}$ inconnues, appelées *données complétées*, et que la maximisation de la vraisemblance associée à ces nouvelles données conduit à une solution explicite.

4.1 Données complétées et vraisemblance complétée

Le seule hypothèse qui est faite sur le lien entre les données complétées et les données observées est qu'elles sont reliées par une fonction $\mathbf{x} = h(\mathbf{y})$. Ces données complétées peuvent prendre, par exemple, la forme $\mathbf{y} = (\mathbf{x}, \mathbf{z})$; $\mathbf{z}$ est alors appelée *information manquante*. Cette notion de données complétées peut avoir une véritable signification pour le modèle comme ce sera le cas pour le modèle de mélange, ou être, au contraire, complètement artificielle. La vraisemblance $f(\mathbf{y}; \boldsymbol{\theta})$ calculée à partir de ces données complétées est appelée *vraisemblance complétée* ou, dans le cas du modèle de mélange, *vraisemblance classifiante*. Partant de la relation

$$f(\mathbf{y}; \boldsymbol{\theta}) = f(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta}) = f(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}) f(\mathbf{x}; \boldsymbol{\theta}) \quad \text{où} \quad \mathbf{x} = h(\mathbf{y})$$

qui s'écrit encore

$$f(\mathbf{x}; \boldsymbol{\theta}) = f(\mathbf{y}; \boldsymbol{\theta}) / f(\mathbf{y}/\mathbf{x}; \boldsymbol{\theta}) \quad \forall \mathbf{y} \in h^{-1}(\mathbf{x})$$

on obtient la relation :

$$L(\boldsymbol{\theta};\mathbf{x}) = L(\boldsymbol{\theta};\mathbf{y}) - \log f(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}) \quad \forall \mathbf{y} \in h^{-1}(\mathbf{x}) \tag{6.1}$$

entre la log-vraisemblance initiale $L(\boldsymbol{\theta};\mathbf{x})$ et la log-vraisemblance complétée $L(\boldsymbol{\theta};\mathbf{y})$.

4.2 Principe

La vraisemblance $L(\boldsymbol{\theta};\mathbf{y})$, supposée simple à maximiser, n'est pas calculable puisque $\mathbf{y}$ est inconnu. La procédure itérative va alors s'appuyer sur la maximisation de l'espérance conditionnelle de la vraisemblance $L(\boldsymbol{\theta};\mathbf{Y})$ pour une valeur du paramètre courant $\boldsymbol{\theta}^{(c)}$. En effet, en calculant l'espérance conditionnelle des 2 membres de la relation 6.1, on obtient la relation fondamentale de l'algorithme EM :

$$L(\boldsymbol{\theta};\mathbf{x}) = \underbrace{E(L(\boldsymbol{\theta};\mathbf{Y})|\mathbf{x},\boldsymbol{\theta}^c)}_{Q(\boldsymbol{\theta},\boldsymbol{\theta}^c)} - \underbrace{E(\log f(\mathbf{Y}|\mathbf{x},\boldsymbol{\theta})|\mathbf{x},\boldsymbol{\theta}^c))}_{H(\boldsymbol{\theta},\boldsymbol{\theta}^c)}$$

et en itérant à partir d'une valeur initiale $\boldsymbol{\theta}^0$, la relation $\boldsymbol{\theta}^{c+1} = \text{Argmax}\ Q(\boldsymbol{\theta},\boldsymbol{\theta}^c)$, on définit un algorithme faisant croitre la vraisemblance car, si on note $\Delta L = L(\boldsymbol{\theta}^{c+1};\mathbf{x}) - L(\boldsymbol{\theta}^c;\mathbf{x})$, on obtient

$$\Delta L = \underbrace{\big(Q(\boldsymbol{\theta}^{c+1},\boldsymbol{\theta}^c) - Q(\boldsymbol{\theta}^c,\boldsymbol{\theta}^c)\big)}_{\geq 0 \quad \text{par construction}} - \underbrace{\big(H(\boldsymbol{\theta}^{c+1},\boldsymbol{\theta}^c) - H(\boldsymbol{\theta}^c,\boldsymbol{\theta}^c)\big)}_{\leq 0 \quad \text{Inégalité de Jensen}} \geq 0.$$

Une itération de l'algorithme EM ainsi défini se décompose en deux étapes :
- étape E (estimation) : calcul de Q à partir de $\boldsymbol{\theta}^c$
- étape M (maximisation) : détermination de $\boldsymbol{\theta}^{c+1}$.

4.3 Propriétés

Sous certaines conditions de régularité, il a été établi que l'algorithme EM assure une convergence vers un maximum local de la vraisemblance. Il a un bon comportement pratique mais peut être toutefois assez lent dans certaines situations ; c'est le cas, par exemple, si les classes sont très mélangées. Cet algorithme, proposé par Dempster, Laird et Rubin dans un papier célèbre, souvent simple à mettre en place, est devenu populaire et a fait l'objet de nombreux travaux que l'on pourra trouver dans l'ouvrage très complet de McLachlan and Krishnan (1997).

4.4 Application au modèle de mélange

Pour le modèle de mélange, les données complétées sont obtenues en ajoutant le composant d'origine $\mathbf{z}_i$ de chaque individu de l'échantillon :

$$\mathbf{y} = (\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n) = ((\boldsymbol{x}_1,z_1),\ldots,(\boldsymbol{x}_n,z_n)).$$

Si on code $z_i = (z_{i1},\ldots,z_{ig})$ où z_{ik} est égal à 1 si i appartient au composant k et 0 sinon, on obtient les relations suivantes :

$$f(\boldsymbol{y}_i;\boldsymbol{\theta}) = f(\boldsymbol{x}_i,z_i;\boldsymbol{\theta}) = \pi_{z_i} f(\boldsymbol{x}_i;\alpha_{z_i})$$

$$\begin{aligned} L(\boldsymbol{\theta};\mathbf{y}) &= \log(f(\mathbf{y};\boldsymbol{\theta})) = \sum_i f(\boldsymbol{y}_i;\boldsymbol{\theta}) = \sum_i \log\left(\pi_{z_i} f(\boldsymbol{x}_i;\alpha_{z_i})\right) \\ &= \sum_{i,k} z_{ik} \log\left(\pi_k f(\boldsymbol{x}_i;\alpha_k)\right), \end{aligned}$$

Sachant que $\mathbf{x}$ est fixé, on a $\mathbf{Y} = (\mathbf{Z},\mathbf{x})$ et la fonction Q peut s'écrire :

$$\begin{aligned} Q(\boldsymbol{\theta},\boldsymbol{\theta}') &= E(L(\boldsymbol{\theta};\mathbf{Y})|\mathbf{x},\boldsymbol{\theta}') = \sum_{i,k} E(Z_{ik}|\mathbf{x},\boldsymbol{\theta}') \log \pi_k f(\boldsymbol{x}_i;\alpha_k) \\ &= \sum_{i,k} t_{ik} \log\left(\pi_k f(\boldsymbol{x}_i;\alpha_k)\right) \quad \text{« vraisemblance pondérée »} \end{aligned}$$

où $t_{ik} = E(Z_{ik}|\mathbf{x},\boldsymbol{\theta}') = P(Z_{ik}=1|\mathbf{x},\boldsymbol{\theta}')$ sont les probabilités d'appartenance a posteriori. En notant, $\boldsymbol{\theta}' = (\pi'_1,\ldots,\pi'_g,\alpha'_1,\ldots,\alpha'_g)$, on obtient

$$\begin{aligned} t_{ik} &= P(Z_{ik}=1|X_i=x_i;\boldsymbol{\theta}') = \frac{P(Z_{ik}=1,X_i=x_i;\boldsymbol{\theta}')}{P(X_i=x_i;\boldsymbol{\theta}')} \\ &= \frac{P(X_i=x_i|Z_{ik}=1;\boldsymbol{\theta}')P(Z_{ik}=1;\boldsymbol{\theta}')}{P(X_i=x_i;\boldsymbol{\theta}')} = \frac{\pi'_k f(x_i;\alpha'_k)}{\sum_{k'}\pi'_{k'} f(x_i;\alpha'_{k'})} \end{aligned}$$

L'algorithme EM prend alors la forme suivante :
- initialisation : choix arbitraire d'une solution initiale $\boldsymbol{\theta}^{(0)}$;
- répétition jusqu'à la convergence des 2 étapes suivantes :
  - étape E (estimation) : calcul des probabilités d'appartenance des $\mathbf{x}_i$ aux classes conditionnellement au paramètre courant :

$$t_{ik}^{(c)} = \frac{\pi_k^{(c)} f(\boldsymbol{x}_i,\alpha_k^{(c)})}{(\sum_\ell \pi_\ell^{(c)} f(\boldsymbol{x}_i,\alpha_\ell^{(c)}))};$$

  - étape M (maximisation) : maximisation de la vraisemblance conditionnellement aux $t_{ik}^{(c)}$; les proportions sont alors obtenues simplement par la relation $\pi_k^{(c+1)} = 1/n \sum_i t_{ik}^{(c)}$ alors que les paramètres $\alpha_k^{(c+1)}$ sont obtenus en résolvant des équations de vraisemblance qui dépendent du modèle de mélange retenu.

4.5 Exemple des mélanges gaussiens monodimensionnel à 2 composants

Dans cette situation, l'algorithme EM s'écrit :
- Initialisation de $\pi_1^{(0)}$, $\pi_2^{(0)}$, $\mu_1^{(0)}$, $\mu_1^{(0)}$, $(\sigma_1^2)^{(0)}$ et $(\sigma_2^2)^{(0)}$;
- étape E : calcul des $t_{ik}^{(c)}$ pour $i = 1\ldots,n$

$$t_{i1}^{(c)} = \frac{\pi_1^{(c)}\varphi(x_i,\mu_1^{(c)},(\sigma_1^2)^{(c)})}{\pi_1^{(c)}\varphi(x_i,\mu_1^{(c)},(\sigma_1^2)^{(c)}) + \pi_2^{(c)}\varphi(x_i,\mu_1^{(c)},(\sigma_1^2)^{(c)})}$$

$$t_{i2}^{(c)} = \frac{\pi_2^{(c)}\varphi(x_i,\mu_2^{(c)},(\sigma_2^2)^{(c)})}{\pi_1^{(c)}\varphi(x_i,\mu_1^{(c)},(\sigma_1^2)^{(c)}) + \pi_2^{(c)}\varphi(x_i,\mu_1^{(c)},(\sigma_1^2)^{(c)})}$$

– étape M :

$$\pi_1^{(c+1)} = \frac{\sum_i t_{i1}^{(c)}}{n} \quad \text{et} \quad \pi_2^{(c+1)} = 1 - \pi_1^{(c+1)}$$

$$\mu_1^{(c+1)} = \frac{\sum_i t_{i1}^{(c)} x_i}{\sum_i t_{i1}^{(c)}} \quad \text{et} \quad \mu_2^{(c+1)} = \frac{\sum_i t_{i2}^{(c)} x_i}{\sum_i t_{i2}^{(c)}}$$

$$(\sigma_1^2)^{(c+1)} = \frac{\sum_i t_{i1}^{(c)} (x_i - \mu_1^{(c+1)})^2}{\sum_i t_{i1}^{(c)}} \quad \text{et} \quad (\sigma_2^2)^{(c+1)} = \frac{\sum_i t_{i2}^{(c)} (x_i - \mu_2^{(c+1)})^2}{\sum_i t_{i2}^{(c)}}.$$

L'application de cet algorithme aux exemples précédents fournit les résultats reportés dans le tableau 6.2 et la figure 6.6.

	Paramètres	π_1	π_2	μ_1	μ_2	σ_1^2	σ_2^2	nb d'itér.
Passereaux	initiaux	0.50	0.50	85	95	1	1	
	obtenus	0.49	0.51	86.1	92.3	2.2	2.5	54
Oreillons	initiaux	0.50	0.50	−3	5	1	1	
	obtenus	0.30	0.70	−0.07	2.98	1.35	0.79	221

TABLE 6.2 – Résultats numériques de l'algorithme EM

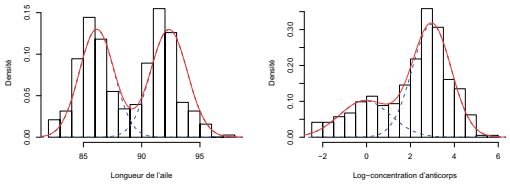


FIGURE 6.6 – Densités mélange obtenues

5 Classification et modèle de mélange

5.1 Les deux approches

L'utilisation des modèles de mélange pour obtenir une partition des données initiales peut se faire de deux manières :

- la première, appelée *approche mélange*, estime les paramètres du modèle puis détermine la partition en rangeant chaque individu dans la classe maximisant la probabilité *a posteriori* t_{ik} calculée à partir des paramètres estimés ; cette affectation est connue sous le nom de la méthode du MAP ou maximum *a posteriori* ;
- la seconde, appelée *approche classification*, consiste à rechercher une partition de l'échantillon de telle sorte que chaque classe k soit assimilable à un sous-échantillon issue de la loi $f(., \alpha_k)$. Il s'agit donc d'estimer simultanément les paramètres du modèle et la partition recherchée.

Dans la suite de cette section, nous précisons le critère optimisé par cette dernière approche et l'algorithme d'optimisation généralement utilisé dans cette situation. Nous faisons ensuite une rapide comparaison des deux approches et étudions les liens que peut avoir ce type de méthodes avec les approches métriques plus classiques de la classification. Nous terminons cette section sur l'interprétation que l'on peut faire du modèle de mélange en terme de classification floue.

5.2 La vraisemblance classifiante

L'introduction de la partition $\mathbf{z}$ dans le critère de vraisemblance n'est pas immédiate et plusieurs propositions ont été faites : Scott et Symons définissent le critère :

$$L_{CR}(\boldsymbol{\theta}, \mathbf{z}) = \sum_{i,k} z_{ik} \log f(\boldsymbol{x}_i, \alpha_k)$$

dans lequel les proportions n'apparaissent pas. Symons, remarquant que ce critère a tendance à donner des classes de mêmes proportions, le modifie pour finalement utiliser la log-vraisemblance complétée (ou classifiante) définie précédemment :

$$L_C(\boldsymbol{\theta}, \mathbf{z}) = \sum_{i,k} z_{ik} \log (\pi_k f(\boldsymbol{x}_i, \alpha_k))$$

liée au critère précédent par la relation :

$$L_C(\boldsymbol{\theta}, \mathbf{z}) = L_{CR}(\boldsymbol{\theta}, \mathbf{z}) + \sum_k \#z_k \log \pi_k$$

où $\#z_k$ est le cardinal de la classe k. La quantité $\sum_k \#z_k \log \pi_k$ représente un terme de pénalité qui disparaît si on impose aux proportions d'être toutes identiques. Le critère $L_{CR}(\boldsymbol{\theta}, \mathbf{z})$ apparaît donc comme une variante de la vraisemblance classifiante restreinte à un modèle de mélange où les classes ont toutes la même proportion.

5.3 L'algorithme CEM

Pour maximiser la vraisemblance classifiante, il est possible d'utiliser une version classifiante de l'algorithme EM obtenue en lui ajoutant une étape de classification. On obtient ainsi l'algorithme de classification très général appelé CEM (classification EM) (Celeux and Govaert, 1992) défini de la manière suivante :

- étape 0 : choix arbitraire d'une solution initiale $\theta^{(0)}$;
- étape E : calcul des $t_{ik}^{(c)}$ comme dans l'algorithme EM ;
- étape C : la partition $\mathbf{z}^{(c)}$ est obtenue en rangeant chaque $\boldsymbol{x}_i$ dans la classe maximisant $t_{ik}^{(c)}$ (MAP) ;
- étape M : maximisation de la vraisemblance conditionnellement aux $z_{ik}^{(c+1)}$: les estimations du maximum de vraisemblance des π_k et des α_k sont obtenues en utilisant les classes de la partition $\mathbf{z}^{(c+1)}$ comme sous-échantillons. Les proportions sont alors fournies par la formule $\pi_k^{(c+1)} = \frac{1}{n} \#z_k^{(c)}$, le calcul des $\alpha_k^{(c+1)}$ dépendant du modèle de mélange retenu.

On retrouve ici un algorithme d'optimisation alternée de type nuées dynamiques (Diday, E. et Collaborateurs, 1979) où les étapes E et C correspondent à l'étape d'affectation et l'étape M à l'étape de représentation.

On peut montrer que cet algorithme itératif est stationnaire et fait croître à chaque itération la vraisemblance complétée sous des conditions très générales.

5.4 Comparaison des deux approches

L'approche classification, déterminant à chaque itération les paramètres à l'aide d'échantillons tronqués du modèle de mélange, fournit une estimation biaisée et inconsistante car le nombre de paramètres à estimer croît avec la taille de l'échantillon. Différents auteurs ont étudié ce problème et ont montré qu'il est généralement préférable d'utiliser l'approche mélange.

Toutefois, lorsque les classes sont bien séparées et les effectifs relativement petits, l'approche classification peut fournir de meilleurs résultats. D'autre part, l'algorithme CEM est beaucoup plus rapide que l'algorithme EM et son utilisation peut être nécessaire lorsque des contraintes de temps de calcul sont imposées, pour un fonctionnement en temps réel par exemple, ou sur des données de très grande taille.

Enfin, l'approche classification a l'avantage de pouvoir présenter de nombreux algorithmes de classification comme des cas particuliers de l'algorithme CEM et de pouvoir les englober dans une approche probabiliste de la classification. Ainsi, par exemple, nous verrons dans le paragraphe 6.4 que l'algorithme des centres-mobiles correspond à un cas particulier simple de l'algorithme CEM. Nous montrerons en particulier que les critères optimisés, l'inertie intraclasse pour les données continues et le critère d'information pour les données qualitatives, correspondent à la vraisemblance classifiante associée à un modèle de mélange particulier.

5.5 Lien avec la classification floue

Dans les méthodes de classification floue, l'appartenance, vraie ou fausse, d'un objet à une classe est remplacée par un degré d'appartenance. Formellement, une classification floue sera caractérisée par une matrice $\mathbf{c}$ de terme général c_{ik} vérifiant $c_{ik} \in [0,1]$ et $\sum_k c_{ik} = 1$. La méthode des « k-moyennes floues » de Bezdek (1981), l'une des plus répandues, consiste à minimiser le critère :

$$W(\mathbf{c}) = \sum_{i,k} c_{ik}^{\gamma} d^2(\boldsymbol{x}_i, \mathbf{g}_k)$$

où $\gamma > 1$ est un coefficient permettant de régler le degré de flou, $\mathbf{g}_k$ est le centre de la classe et d est la distance euclidienne. Il est nécessaire d'imposer à γ d'être différent de 1, sinon la fonction W est minimale pour des valeurs de $c_{ik} = 0$ ou 1 et on retrouve le critère habituel d'inertie intraclasse. Les valeurs généralement conseillées se situent entre 1 et 2. La minimisation de ce critère se fait à l'aide d'un algorithme qui alterne les deux étapes suivantes :

1. calcul des centres : $\mathbf{g}_k = \sum_i c_{ik}^{\gamma} \boldsymbol{x}_i / \sum_i c_{ik}^{\gamma}$;
2. calcul de la partition floue : $c_{ik} = \frac{D_i}{\|\boldsymbol{x}_i - \mathbf{g}_k\|^{\frac{2}{\gamma-1}}}$ où $D_i = \sum_{\ell} \frac{1}{\|\boldsymbol{x}_i - \mathbf{g}_\ell\|^{\frac{2}{\gamma-1}}}$.

La validation d'une telle approche avec, en particulier le choix du coefficient γ, est assez délicate.

L'estimation des paramètres d'un modèle de mélange est une autre façon d'aborder, et de manière plus naturelle, ce problème. En effet, l'estimation des probabilités *a posteriori* t_{ik} d'appartenance des objets à chaque classe fournit directement une classification floue et l'algorithme EM, appliqué au modèle de mélange, peut être considéré comme un algorithme de classification floue.

Hathaway (1986) a montré que la recherche d'une partition floue et du paramètre θ, effectuée à l'aide d'une optimisation alternée d'un critère de classification floue, conduit exactement aux deux étapes de l'algorithme EM qui peut donc être considéré comme un algorithme de classification floue. Il montre en particulier que l'algorithme EM maximise de manière alternée le critère :

$$W(\mathbf{c}, \theta) = L_C(\theta, \mathbf{c}) + H(\mathbf{c})$$

où L_C est la fonction de log-vraisemblance complétée dans laquelle on a remplacé la partition $\mathbf{z}$ par la partition floue $\mathbf{c}$:

$$L_C(\theta, \mathbf{c}) = \sum_{i,k} c_{ik} \log\left(\pi_k f(\boldsymbol{x}_i; \alpha_k)\right)$$

et H est la fonction d'entropie :

$$H(\mathbf{c}) = -\sum_i \sum_k c_{ik} \log c_{ik}.$$

Il est facile de vérifier que, si l'on supprime le terme d'entropie du critère W, on obtient à chaque étape des partitions « dures ». L'algorithme ainsi obtenu est alors simplement l'algorithme CEM : la différence entre l'algorithme EM et l'algorithme CEM est la présence du terme d'entropie. Si, à la convergence de EM, les composants sont très séparés, la partition floue $\mathbf{z}(\theta)$ est proche d'une partition et on a $H(\mathbf{z}(\theta)) \approx 0$ et $L(\theta) = W(\mathbf{z}(\theta), \theta) = L_C(\theta, \mathbf{z}(\theta)) + H(\mathbf{z}(\theta)) \approx L_C(\theta, \mathbf{z}(\theta))$.

6 Mélange gaussien multidimensionnel

On suppose maintenant que les données $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ sont dans l'espace $\mathbb{R}^p$.

6.1 Définition

Avec le modèle de mélange gaussien multidimensionnel, chaque classe est modélisée par une distribution normale. La densité du mélange s'écrit alors

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_k \pi_k \varphi(\boldsymbol{x}, \boldsymbol{\mu}_k, \Sigma_k), \quad \forall \boldsymbol{x} \in \mathbb{R}^p$$

où φ est la densité de la loi normale multidimensionnelle :

$$\varphi(\boldsymbol{x}, \boldsymbol{\mu}_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)' \Sigma_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k)\right)$$

et $\boldsymbol{\theta}$ est le vecteur $(\pi_1, \ldots, \pi_g, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g, \Sigma_1, \ldots, \Sigma_g)$ formé des proportions π_k et des paramètres $\boldsymbol{\mu}_k$ et Σ_k qui sont respectivement le vecteur moyenne et

la matrice de variance-covariance de la classe k. Si on note $d^2_{\Sigma_k^{-1}}$ la distance quadratique définie par Σ_k^{-1}, φ s'écrit aussi

$$\varphi(\boldsymbol{x}, \boldsymbol{\mu}_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} d^2_{\Sigma_k^{-1}}(\boldsymbol{x}, \boldsymbol{\mu}_k)\right).$$

Ainsi, les classes associées aux composants du mélange sont ellipsoïdales, centrées à la moyenne $\boldsymbol{\mu}_k$ et les matrices de variance-covariance Σ_k déterminent leurs caractéristiques géométriques.

Remarquons que dans certaines situations, ce modèle pourra être simplifié en imposant aux proportions d'êtres toutes égales à $\frac{1}{g}$.

6.2 Les algorithmes EM et CEM

Dans ce paragraphe, on explicite ce que deviennent les algorithmes EM et CEM décrits précédemment pour un modèle de mélange de lois normales multidimensionnelles.

Étape E

L'étape E des algorithmes EM et CEM ne pose pas de problème particulier et conduit au calcul des probabilités d'appartenance des $\boldsymbol{x}_i$ aux classes conditionnellement au paramètre courant :

$$t_{ik}^{(c)} = \frac{\pi_k^{(c)} \varphi(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(c)}, \Sigma_k^{(c)})}{\sum_\ell \pi_\ell^{(c)} \varphi(\boldsymbol{x}_i, \boldsymbol{\mu}_\ell^{(c)}, \Sigma_\ell^{(c)})}.$$

Étape de classification

Rappelons que cette étape supplémentaire de classification n'est définie que pour l'algorithme CEM. La partition $\mathbf{z}^{(c+1)}$ est obtenue en rangeant chaque $\boldsymbol{x}_i$ dans la classe maximisant $t_{ik}^{(c)}$:

$$z_{ik}^{(c+1)} = \begin{cases} 1 & \text{si } k = \text{argmax}_{k=1,\ldots,g}\, t_{ik}^{(c)} \\ 0 & \text{sinon.} \end{cases}$$

Chaque $\boldsymbol{x}_i$ est donc rangé dans la classe qui maximise $\pi_k \varphi(\boldsymbol{x}_i, \boldsymbol{\mu}_k, \Sigma_k)$ ou, de manière équivalente, qui minimise $-2\log(\pi_k \varphi(\boldsymbol{x}_i, \boldsymbol{\mu}_k, \Sigma_k))$ ou encore :

$$d^2_{\Sigma_k^{-1}}(\boldsymbol{x}_i, \boldsymbol{\mu}_k) + \log|\Sigma_k| - 2\log(\pi_k). \tag{6.2}$$

Étape M

Pour unifier la présentation des algorithmes EM et CEM, on utilisera la matrice

$$\mathbf{c} = (c_{ik}, i = 1, \ldots, n; k = 1, \ldots, g)$$

avec $0 \leqslant c_{ik} \leqslant 1$ et $\sum_k c_{ik} = 1$. Pour l'algorithme EM, $\mathbf{c}$ définit une partition floue et nous avons $c_{ik} = t_{ik}^{(c)}$ pour $1 \leqslant i \leqslant n$ et $1 \leqslant k \leqslant g$. Pour l'algorithme CEM, $\mathbf{c}$ est la partition $\mathbf{z}^{(c)}$ obtenue à l'étape C et nous avons $c_{ik} = 1$ si $\boldsymbol{x}_i$ appartient à la classe k et 0 sinon ($1 \leqslant i \leqslant n$, $1 \leqslant k \leqslant g$).

On notera dans les deux situations

$$n_k = \sum_i c_{ik},$$

$$\bar{\boldsymbol{x}}_k = \frac{1}{n_k} \sum_i c_{ik} \boldsymbol{x}_i,$$

$$S_k = \frac{1}{n_k} \sum_i c_{ik} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)'$$

et

$$S_W = \frac{1}{n} \sum_k n_k S_k.$$

Lorsque $\mathbf{c}$ est une partition, remarquons que n_k, $\bar{\boldsymbol{x}}_k$ et S_k correspondent simplement au cardinal, au vecteur moyenne et à la matrice de variance-covariance de la classe k. La matrice S_W est alors la matrice de variance-covariance intraclasse.

Avec cette convention, dans tous les cas l'étape M consiste à maximiser pour $\mathbf{c}$ fixé la fonction $L_c(\boldsymbol{\theta}, \mathbf{c}) = \sum_{i,k} c_{ik} \log(\pi_k \varphi(\boldsymbol{x}_i, \boldsymbol{\mu}_k, \Sigma_k))$ qui s'écrit pour le modèle de mélange gaussien multidimensionnel

$$-\frac{np}{2}\log 2\pi - \frac{1}{2}\sum_{i,k} c_{ik}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)'\Sigma_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k) - \frac{1}{2}\sum_k n_k \log|\Sigma_k| + \sum_k n_k \log \pi_k.$$

Il est facile de montrer que le paramètre $\boldsymbol{\mu}_k$ est alors nécessairement le vecteur moyenne $\bar{\boldsymbol{x}}_k$ et que les proportions, si elles sont libres, vérifient $\pi_k = n_k/n$.

Les paramètres Σ_k doivent alors minimiser la fonction

$$F(\Sigma_1, \ldots, \Sigma_g) = \sum_k n_k \left(\text{tr}(S_k \Sigma_k^{-1}) + \log|\Sigma_k|\right) \tag{6.3}$$

qui est liée à la fonction L_c par la relation

$$L_c(\boldsymbol{\theta}, \mathbf{c}) = -\frac{1}{2} F(\Sigma_1, \ldots, \Sigma_g) + \sum_k n_k \log \pi_k - \frac{np}{2}\log 2\pi.$$

On peut alors montrer que les matrices optimales Σ_k sont simplement les matrices S_k et que la vraisemblance classifiante maximisée s'écrit

$$L_c(\boldsymbol{\theta}, \mathbf{c}) = -\frac{1}{2}\sum_{k=1}^{g} n_k \log|S_k| + \sum_{k=1}^{g} n_k \log \pi_k - \frac{np}{2}(\log 2\pi + 1).$$

6.3 Modèle parcimonieux

Lorsque la taille de l'échantillon est faible ou lorsque la dimension de l'espace est grande, il devient nécessaire de diminuer le nombre de paramètres afin d'obtenir des modèles plus parcimonieux. Pour ceci, il est possible de reparamétriser les matrices de variance-covariance des composants du mélange en s'appuyant sur leur décomposition en valeurs propres et vecteurs propres :

$$\Sigma_k = D_k B_k D_k'$$

où D_k est la matrice des vecteurs propres et B_k la matrice diagonale composée des valeurs propres. Pour obtenir une décomposition unique, les valeurs propres sont ordonnées suivant leurs valeurs décroissantes. Ensuite, chaque matrice B_k peut être elle-même décomposée en un nombre réel λ_k et une matrice A_k tel que $B_k = \lambda_k A_k$ avec $|A_k| = 1$. Chaque matrice de variance-covariance est donc finalement décomposée sous la forme

$$\Sigma_k = \lambda_k D_k A_k D_k' \tag{6.4}$$

où A_k, matrice diagonale de déterminant 1, avec des valeurs allant en décroissant, caractérise la *forme* de la classe, D_k, matrice orthogonale, caractérise l'*orientation* de la classe et λ_k, nombre réel positif, représente le *volume* de la classe. Par exemple, lorsque les données sont dans un plan, D est une matrice de rotation définie par un angle $\boldsymbol{\theta}$ et A est une matrice diagonale de termes diagonaux a et $1/a$ ($a \neq 0$). La figure 6.7 représente alors l'ellipse d'équidensité de cette distribution en fonction des valeurs $\boldsymbol{\theta}$, λ et a.

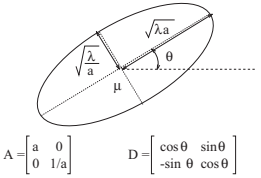


FIGURE 6.7 – Paramétrage d'une classe gaussienne dans le plan.

Le modèle de mélange est finalement paramétré par les centres $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g$, les proportions $\pi_1, \ldots, \pi_g$, les volumes $\lambda_1, \ldots, \lambda_g$, les formes $A_1, \ldots, A_g$ et les orientations $D_1, \ldots, D_g$ de chaque classe. En utilisant cette paramétrisation, il est alors possible de proposer des situations intermédiaires entre des hypothèses restrictives (matrices de variance-covariance proportionnelles à la matrice identité ou matrices de variance-covariance identiques pour toutes les classes) et les hypothèses très générales (aucune contrainte) pour obtenir des modèles plus ou moins parcimonieux facilement interprétables et qui pourront être adaptés à des problèmes variés de classification.

La famille générale

Tout d'abord, en supposant que les volumes λ_k, les orientations D_k et les formes A_k des classes sont égales ou non, huit modèles peuvent être obtenus. Par exemple, on peut supposer des volumes différents et imposer aux formes et aux orientations d'être égales en prenant $A_k = A$ (A inconnue) et $D_k = D$ (D inconnue) pour $k = 1, \ldots, g$. On notera ce modèle $[\lambda_k DAD']$. Avec cette convention, le modèle $[\lambda D_k A D_k']$ correspondra à un modèle de mélange avec des volumes égaux, des formes égales et des orientations différentes.

La famille diagonale

Une autre famille intéressante consiste à supposer que les matrices de variance-covariances Σ_k sont diagonales. Dans la paramétrisation (6.4), cela signifie que la

matrice d'orientation D_k est une matrice de permutation. On écrira $\Sigma_k = \lambda_k B_k$ où B_k est une matrice diagonale avec $|B_k| = 1$. Dans cette situation, le modèle de mélange gaussien avec des matrices de variance-covariance diagonales peut être vu comme un modèle élégant permettant de pondérer les variables. Il conduit à un algorithme de pondérations adaptatives supposant suivant la situation les mêmes poids pour chaque classe ou des pondérations différentes par classes. Cette paramétrisation conduit aux quatre modèles suivant : $[\lambda B]$, $[\lambda_k B]$, $[\lambda B_k]$ et $[\lambda_k B_k]$.

La famille sphérique

La dernière famille consiste à supposer des formes sphériques, c'est-à-dire à supposer que $A_k = I$ où I est la matrice identité. Dans une telle situation, deux modèles sont possibles : $[\lambda I]$ et $[\lambda_k I]$.

On obtient ainsi 14 modèles différents ayant une interprétation géométrique simple illustrée dans la figure (6.8). Cette figure représente les ellipses d'équi-

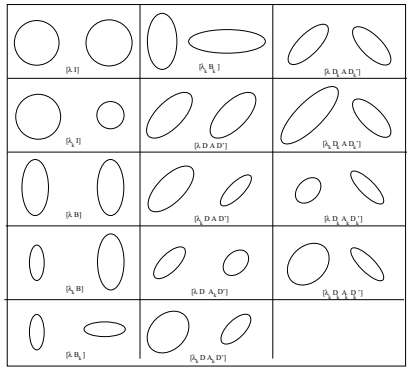


FIGURE 6.8 – Les 14 modèles

densité des composants des différents modèles obtenus pour deux classes dans le plan. En tenant compte de la possibilité d'imposer ou non aux proportions π_k d'être égales ou non, on obtient finalement 28 modèles.

Remarquons que cette paramétrisation met en évidence deux notions souvent confondues sous l'appellation un peu floue de taille : la proportion des individus présents dans une classe et le volume que représente la place occupée par une classe dans l'espace. En particulier, il est possible d'avoir des classes de faible volume et de grande proportion et, vice-versa, des classes de grand volume et de faible proportion.

6.4 Algorithmes associés aux modèles parcimonieux

Les seules répercussions du choix du modèle parcimonieux concernent la forme particulière que peut prendre la fonction d'affectation de l'étape C de

l'algorithme CEM et le calcul des matrices de variance-covariance Σ_k de l'étape M. Nous étudions maintenant quelques uns de ces modèles parcimonieux.

Modèle $[\lambda I]$: forme sphérique et volumes identiques

Il s'agit de la situation la plus simple correspondant à des classes ayant toutes la même distribution normale sphérique. L'ensemble des paramètres $\Sigma_1, \ldots, \Sigma_g$ se réduit au réel λ et la fonction F définie dans la relation (6.3) devient :

$$F(\lambda) = n\left(\frac{1}{\lambda}\operatorname{tr}(S_W) + p\log\lambda\right).$$

On obtient alors à l'étape M

$$\lambda = \frac{\operatorname{tr}(S_W)}{p}$$

et la vraisemblance classifiante s'écrit :

$$L_c(\boldsymbol{\theta}, \mathbf{c}) = -\frac{np}{2}\log\left(\frac{\operatorname{tr}(S_W)}{p}\right) + \sum_k n_k \log\pi_k - \frac{np}{2}(\log 2\pi + 1) \qquad (6.5)$$

Pour l'algorithme CEM et lorsque les proportions sont égales, la formule (6.2) montre que l'affectation des individus aux classes se fait simplement en utilisant la distance euclidienne habituelle $d^2(\boldsymbol{x}_i, \boldsymbol{\mu}_k)$ et la formule (6.5) montre que la maximisation de la vraisemblance classifiante est équivalente à la minimisation du critère de variance-covariance intraclasse $\operatorname{tr}(S_W)$; l'algorithme CEM est alors simplement l'algorithme des centres mobiles (*k-means*). Finalement, utiliser le critère d'inertie revient à supposer que les classes sont sphériques, de même proportion et de même volume.

Modèle $[\lambda_k I]$: forme sphérique et volumes différents On reprend le modèle précédent en le modifiant légèrement pour prendre en compte l'existence de classes pouvant avoir des volumes différents. Formellement, les modèles $[\lambda I]$ et $[\lambda_k I]$ ne semblent pas très différents et l'augmentation du nombre de paramètres est assez faible (voir la table 6.3). En fait, ces deux modèles peuvent conduire à des classifications très différentes.

L'ensemble des paramètres $\Sigma_1, \ldots, \Sigma_g$ se réduit cette fois au vecteur $(\lambda_1, \ldots, \lambda_g)$ et la fonction F devient

$$F(\lambda_1, \ldots, \lambda_g) = \sum_k n_k\left(\frac{1}{\lambda_k}\operatorname{tr}(S_k) + p\log\lambda_k\right).$$

On obtient alors à l'étape M

$$\lambda_k = \frac{\operatorname{tr}(S_k)}{p}$$

et la vraisemblance classifiante s'écrit

$$L_c(\boldsymbol{\theta}, \mathbf{c}) = -\frac{p}{2}\sum_k n_k \log\left(\frac{\operatorname{tr}(S_k)}{p}\right) + \sum_k n_k\log\pi_k - \frac{np}{2}(\log 2\pi + 1). \qquad (6.6)$$

Pour l'algorithme CEM et lorsque les proportions sont égales, la formule (6.2) montre que l'affectation des individus aux classes se fait en utilisant la distance :

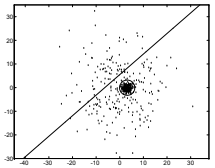
$$\frac{1}{\lambda_k}d^2(\boldsymbol{x}_i, \boldsymbol{\mu}_k) + p\log(\lambda_k)$$

et la relation (6.6) montre que le critère minimisé s'écrit

$$\sum_k n_k \log\operatorname{tr}(S_k). \qquad (6.7)$$

La distance d'un point au centre d'une classe a été modifiée par une quantité qui dépend du volume de la classe. Cette modification a des répercussions importantes ; par exemple, les surfaces séparatrices, qui étaient dans le cas précédent des hyperplans, deviennent des hypersphères. Ce modèle permet de reconnaître des situations comme celle de la figure (6.9) sans aucune difficulté. Dans cet

FIGURE 6.9 – Exemple de classes de volumes différents



exemple, les deux classes ont été simulées suivant deux lois normales sphériques avec les mêmes proportions mais avec des volumes très différents. Le résultat obtenu avec le critère d'inertie intraclasse classique correspond à la séparation de la population par la droite et n'a donc aucun rapport avec la partition simulée. Avec le modèle à volume variable, la partition obtenue, indiquée par le cercle est très proche de la classification initiale.

Remarquons que, sans l'aide du modèle de mélange, il aurait été difficile de proposer la distance et le critère utilisés dans cette approche à partir d'une simple interprétation métrique.

Si dans la situation précédente des classes sphériques de même volume, l'algorithme CEM avec des proportions fixes conduisaient vers le critère $\operatorname{tr}(S_W)$, certainement l'un des critères de classification les plus anciens et les plus utilisés, on remarque cette fois que l'on obtient le critère $\sum_k n_k \log\operatorname{tr}(S_k)$ beaucoup moins connu.

Modèle $[\lambda B]$: formes diagonales identiques

La matrice de variance-covariance de chaque classe a maintenant la forme $\Sigma_k = \lambda B$ où B est une matrice diagonale de déterminant 1 ; il est possible de simplifier ce modèle en prenant comme paramètre la matrice $A = \lambda B$ où cette fois la matrice A est une matrice diagonale quelconque. L'ensemble des paramètres $\Sigma_1, \ldots, \Sigma_g$ se réduit donc à la matrice A et la fonction F devient

$$F(A) = n\left(\operatorname{tr}(S_W A^{-1}) + \log|A|\right).$$

On peut montrer qu'à l'étape M la matrice A minimisant F est

$$A = \operatorname{diag}(S_W)$$

où $\mathrm{diag}(S_W)$ est la matrice diagonale obtenue en ne conservant que la diagonale de S_W. La vraisemblance classifiante s'écrit alors

$$L_c(\boldsymbol{\theta}, \mathbf{c}) = -\frac{n}{2}\log|\,\mathrm{diag}(S_W)| + \sum_k n_k \log \pi_k - \frac{np}{2}(\log 2\pi + 1). \qquad (6.8)$$

Pour l'algorithme CEM et lorsque les proportions sont égales, la formule (6.2) montre que l'affectation des individus aux classes se fait en utilisant la distance $d^2_{B^{-1}}(\mathbf{x}_i, \mu_k)$ qui correspond à une distance euclidienne avec des pondérations sur les variables et la relation (6.8) montre que le critère minimisé s'écrit $|\,\mathrm{diag}(S_W)|$.

Modèle $[\Sigma]$: Formes identiques

La matrice de variance-covariance de chaque classe a maintenant la forme $\Sigma_k = \Sigma$ et la fonction F devient

$$F(\Sigma) = n\left(\mathrm{tr}(S_W\Sigma^{-1}) + \log|\Sigma|\right).$$

La matrice optimale Σ obtenue à l'étape M vérifie donc

$$\Sigma = S_W$$

et la vraisemblance classifiante à maximiser s'écrit alors

$$L_c(\boldsymbol{\theta}, \mathbf{c}) = -\frac{1}{2}\left(np + n\log|S_W|\right) + \sum_k n_k \log \pi_k - \frac{np}{2}(\log 2\pi + 1). \qquad (6.9)$$

Lorsque les proportions sont égales, la relation (6.9) montre que le critère minimisé s'écrit $|S_W|$.

Conclusions

La table 6.3 résume les principales caractéristiques des 14 modèles. Dans cette table, la première colonne spécifie le modèle. La seconde donne le nombre de paramètres réels à estimer. La troisième colonne indique si l'étape M peut être obtenue de manière explicite (Exp.) – c'était le cas des quatre exemples précédemment décrits – ou si cette étape nécessite une procédure itérative (PI). La dernière colonne donne le critère à minimiser obtenu dans le cas de proportions égales et pour l'algorithme CEM.

À partir de cette table 6.3, il peut être remarqué que les critères obtenus pour les modèles $[\lambda DAD']$, $[\lambda D_k A_k D'_k]$ et $[\lambda_k D_k A_k D'_k]$ sont des critères de classification classiques proposés sans aucune référence à des modèles statistiques. Par exemple, $\mathrm{tr}(S_W)$ est le critère des *k-means*, $|S_W|$ a été proposé par Friedman et Rubin en 1967 et $\sum_{k=1}^{g} n_k \log|S_k|$ a été proposé par Scott et Symons en 1971. Au contraire les modèles $[\lambda_k DAD']$, $[\lambda_k DA_k D']$ et $[\lambda_k D_k AD'_k]$, qui permettent des volumes différents pour chaque classe, conduisent à des critères moins usuels.

Plusieurs remarques peuvent être faites.

- Les huit modèles de la famille générale conduisent à des résultats invariants pour toute transformation linéaire des données.
- Les trois critères obtenus pour les modèles de la famille diagonale $[\lambda B]$, $[\lambda B_k]$ et $[\lambda_k B_k]$ sont des adaptations simples des critères correspondant de la famille générale et les quatre modèles de cette famille diagonale sont invariants pour toute normalisation des variables mais pas pour toute transformation linéaire.

TABLE 6.3 – Quelques caractéristiques des 14 modèles. On a $\alpha = gd + g - 1$ dans le cas des proportions libres et $\alpha = gd$ dans le cas des proportions égales, et $\beta = \frac{d(d+1)}{2}$. Ω_k est la matrice diagonale des valeurs propres de S_k.

modèle	nombre de paramètres	étape M	critère
$[\lambda DAD']$	$\alpha + \beta$	Exp.	$\lvert S_W\rvert$
$[\lambda_k DAD']$	$\alpha + \beta + g - 1$	PI	-
$[\lambda DA_k D']$	$\alpha + \beta + (g-1)(d-1)$	PI	-
$[\lambda_k DA_k D']$	$\alpha + \beta + (g-1)d$	PI	-
$[\lambda D_k AD'_k]$	$\alpha + g\beta - (g-1)d$	Exp.	$\lvert\sum_k n_k\Omega_k\rvert$
$[\lambda_k D_k AD'_k]$	$\alpha + g\beta - (g-1)(d-1)$	PI	-
$[\lambda D_k A_k D'_k]$	$\alpha + g\beta - (g-1)$	Exp.	$\sum_k n_k\lvert S_k\rvert^{\frac{1}{d}}$
$[\lambda_k D_k A_k D'_k]$	$\alpha + g\beta$	Exp.	$\sum_k n_k \log\lvert S_k\rvert$
$[\lambda B]$	$\alpha + d$	Exp.	$\lvert\,\mathrm{diag}(S_W)\rvert$
$[\lambda_k B]$	$\alpha + d + g - 1$	PI	-
$[\lambda B_k]$	$\alpha + gd - g + 1$	Exp.	$\sum_k n_k\lvert\,\mathrm{diag}(S_k)\rvert^{\frac{1}{d}}$
$[\lambda_k B_k]$	$\alpha + gd$	Exp.	$\sum_k n_k \log\mathrm{diag}(S_k)$
$[\lambda I]$	$\alpha + 1$	Exp.	$\mathrm{tr}(S_W)$
$[\lambda_k I]$	$\alpha + d$	Exp.	$\sum_k n_k \log\mathrm{tr}(S_k)$

- Les deux modèles sphériques dont invariants pour toute transformation isométrique.
- Enfin, pour les modèles sphériques et diagonaux, on retrouve l'hypothèse des « classes latentes » qui suppose que les variables initiales sont indépendantes conditionnellement à la connaissance du composant.

7 Mise en œuvre

Il existe plusieurs logiciels permettant d'utiliser les méthodes étudiées dans ce chapitre ; on peut citer en particulier le logiciel MIXMOD[1]. Dans cette section, on passe en revue rapidement les problèmes que peut poser leur mise en œuvre.

7.1 Choix du modèle et du nombre de classes

Les méthodes de classification automatique sont souvent justifiées de façon heuristique et le choix de la « bonne » méthode ou du « bon » nombre de classes est alors un problème difficile et souvent mal posé. L'utilisation de méthodes de classification s'appuyant sur les modèles de mélange permet de placer ce problème dans le cadre plus général de la sélection de modèles probabilistes.

Dans le cadre bayésien, la recherche du modèle le plus probable conduit à des critères de sélection de modèles très utilisés comme le critère BIC de Schwarz constitués de deux termes : le premier est la vraisemblance qui a tendance à choisir le modèle le plus complexe et le second est une terme de pénalisation,

1. http ://www-math.univ-fcomte.fr/mixmod

fonction croissante du nombre de paramètres du modèle. On peut citer en particulier le critère ICL qui, en prenant en compte l'objectif de classification, fournit généralement de bonnes solutions.

7.2 Stratégies d'utilisation

La maximisation du critère de vraisemblance par l'algorithme EM ou de la vraisemblance classifiante par l'algorithme CEM conduit à chaque fois à la construction d'une suite de solutions faisant croître le critère vers un maximum local qui dépend donc de la position initiale choisie par l'algorithme. La stratégie généralement retenue pour obtenir une « bonne » solution consiste à répéter l'algorithme à partir de plusieurs positions initiales et à retenir la meilleure. On pourra se reporter, par exemple, au texte de Biernacki et al. (2003), où des stratégies plus fines et assez performantes, incluant une première phase consistant à lancer de nombreuses fois l'algorithme sans attendre la convergence complète, ont été étudiées.

Annexe A

Quelques résultats

1 Trois minimisations classiques

Proposition 1 *La fonction*

$$f(x) = \frac{a}{x} + b\log x, \quad a, b, x > 0$$

atteint son minimum pour $x = \frac{a}{b}$

Preuve :

$$f'(x) = -\frac{a}{x^2} + \frac{b}{x} = \frac{-a+bx}{x^2}$$
$$f''(x) = \frac{2a}{x^3} - \frac{b}{x^2} = \frac{2a-bx}{x^3}$$

En $x = \frac{a}{b}$, la dérivée s'annule et la dérivée seconde est positive ; la proposition est donc démontrée.

□

Proposition 2 *La fonction*

$$f(x_1, \ldots, x_p) = \sum_{i=1}^{p} a_i \log x_i$$

sous les contraintes $x_i > 0$, $a_i > 0$ $\forall i$ *et* $\sum_i x_i = 1$ *atteint son maximum pour* $x_i = \frac{a_i}{\sum_j a_j}$

Preuve : En utilisant les multiplicateurs de Lagrange, on se ramène à la maximisation de

$$g(x_1, \ldots, x_p) = \sum_i a_i \log x_i + \lambda(\sum x_i - 1).$$

L'annulation des dérivées partielles

$$g'_{x_j}(x_1, \ldots, x_p) = \frac{a_j}{x_j} + \lambda = 0 \quad \forall j$$

entraîne

$$x_j = -\frac{a_j}{\lambda} \quad \forall j.$$

En utilisant la contrainte $\sum_j x_j = 1$, on en déduit $\lambda = -\sum_j a_j$ et donc le résultat attendu.

□

Proposition 3 *La fonction*

$$f(x_1, \ldots, x_p) = \sum_{i=1}^{p} x_i$$

sous les contraintes $x_i > 0$ $\forall i$ *et* $\prod_i x_i = 1$ *atteint son minimum pour* $x_i = 1 \quad \forall i$

Preuve : En utilisant les multiplicateurs de Lagrange, on se ramène à la minimisation de

$$g(x_1, \ldots, x_p) = \sum_i x_i + \lambda(\prod x_i - 1).$$

L'annulation des dérivées partielles

$$g'_{x_j}(x_1, \ldots, x_p) = x_j + \lambda\frac{\prod x_i}{x_j} = x_j + \frac{\lambda}{x_j} = 0 \quad \forall j$$

entraîne

$$x_j^2 = -\lambda \quad \forall j$$

et en utilisant la contrainte $\prod_j x_j = 1$, on en déduit $\lambda = -1$ et donc le résultat attendu.

□

2 Minimisations matricielles

Proposition 4 *La matrice symétique définie positive* M *de dimension* $p \times p$ *et de déterminant* $|M| = 1$ *minimisant* $\mathrm{tr}(M)$ *est la matrice identité* I *et la valeur minimisée est égale à* p.

Preuve : Si on note $\lambda_1, \ldots, \lambda_p$ les p valeurs propres de la matrice symétique M, le problème se ramène à la minimisation de $\sum_i \lambda_i$ sous la contrainte $\prod \lambda_i = 1$. Sachant que toutes les valeurs propres sont > 0, la proposition précédente entraîne donc que les valeurs propres sont toutes égales à 1 et le résultat en découle alors directement.

□

Corollaire 5 *La matrice symétique définie positive* M *de dimension* $p \times p$ *et de déterminant* $|M| = 1$ *minimisant* $\mathrm{tr}(M^{-1})$ *est la matrice identité* I *et la valeur minimisée est égale à* p.

Preuve : Il suffit de poser $N = M^{-1}$. La proposition précédente permet d'en déduire que $N = I$ et donc le résultat.

□

Corollaire 6 *La matrice symétique* M *de dimension* $p \times p$ *de déterminant* $|M| = 1$ *minimisant* $tr(QM^{-1})$ *où* Q *est une matrice symétique définie positive est*

$$M = \frac{Q}{|Q|^{\frac{1}{p}}},$$

et la valeur minimisée est égale à $p|Q|^{\frac{1}{p}}$.

Preuve : Il suffit de poser $N = |Q|^{-\frac{1}{p}} Q M^{-1}$ et d'utiliser la proposition 4

□

Corollaire 7 *La matrice diagonale M de dimension $p \times p$ de déterminant $|M| = 1$ minimisant $tr(QM^{-1})$ où Q est une matrice symétrique positive est*

$$M = \frac{diag(Q)}{|diag(Q)|^{\frac{1}{d}}},$$

et la valeur minimisée est $p|diag(Q)|^{\frac{1}{d}}$.

Preuve : Si M est une matrice diagonale, M^{-1} est aussi une matrice diagonale et nous avons $\mathrm{tr}(QM^{-1}) = \mathrm{tr}(\mathrm{diag}(Q)M^{-1})$. Le corollaire précédent permet alors de conclure.

□

Proposition 8 *La matrice symétrique M de dimension $p \times p$ minimisant $tr(M^{-1}) + \alpha \ln |M|$ où α est un réel positif est $M = \frac{1}{\alpha} I$.*

Preuve : On peut écrire $M = d.N$ avec $d = |M|^p$ et $|N| = 1$. nous obtenons alors

$$\mathrm{tr}(M^{-1}) + \alpha \ln |M| = \frac{1}{d} \mathrm{tr}(N^{-1}) + \alpha p \ \ln d.$$

On peut donc d'abord chercher la matrice N de déterminant 1 minimisant $\mathrm{tr}(N^{-1})$. En utilisant le corollaire 5, on obtient $N = I$. Il suffit alors de minimiser

$$\frac{p}{d} + \alpha p \ \ln d = p(\frac{1}{d} + \alpha \ \ln d).$$

La proposition 1 permet alors d'en déduire $d = \frac{1}{\alpha}$ ce qui conduit au résultat attendu.

□

Corollaire 9 *La matrice symétrique M de dimension $p \times p$ minimisant $\mathrm{tr}(QM^{-1}) + \alpha \ln |M|$ où Q est une matrice symétrique positive et α est un réel positif est $M = \frac{1}{\alpha} Q$.*

Preuve : En posant $N = Q^{-1}M$, on obtient

$$\mathrm{tr}(QM^{-1}) + \alpha \ln |M| = \mathrm{tr}(N^{-1}) + \alpha \ln |N| + cste.$$

En utilisant la proposition précédente, on obtient $N = \frac{1}{\alpha} I$ ce qui conduit au résultat attendu.

□

Corollaire 10 *La matrice diagonale M de dimension $p \times p$ minimisant $\mathrm{tr}(QM^{-1}) + \alpha \ln |M|$ où Q est une matrice symétrique positive et α est un réel positif est $M = \frac{1}{\alpha} \mathrm{diag}(Q)$.*

Preuve : Si M est une matrice diagonale, M^{-1} est aussi une matrice diagonale et nous avons $\mathrm{tr}(QM^{-1}) = \mathrm{tr}(\mathrm{diag}(Q)M^{-1})$. Le corollaire précédent permet alors de conclure.

□

Bibliographie

Benzecri, J.-P. (1973). *L'analyse des données tome 1 : la taxinomie*. Dunod, Paris.

Bezdek, J. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York.

Biernacki, C., Celeux, G., and Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics and Data Analysis*, 41 :561–575.

Borg, I. and Groenen, P. (1997). *Modern Multidimensional Scaling*. Springer, New York.

Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14(3) :315–332.

Cox, T. and Cox, M. (1994). *Multidimensional Scaling*. Chapman and Hall, London.

Diday, E. et Collaborateurs (1979). *Optimisation et classification automatique*. INRIA, Rocquencourt.

Hathaway, R. J. (1986). Another interpretation of the em algorithm for mixture distributions. *Statistics & Probability Letters*, 4 :53–56.

Jackson (1991). *A User's Guide to Principal Components*. Wiley, New York.

MacQueen, J. B. (1967). Some methods for classification and analysis of cluster analysis. In LeCam, L. M. and Neyman, J., editors, *Proceedings of 5th Berkeley Symposium on Mathematics, Statistics and Probability*, pages 281–297, CA. University of California Press.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press, New York.

McLachlan, G. J. and Krishnan, K. (1997). *The EM Algorithm*. Wiley, New York.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.