Outils mathématiques pour l'Analyse de Données

Révision

Chargé du cours
Prof. Mustapha Rachdi

Université Grenoble Alpes
Bât. Michel Dubois
UFR SHS, BP. 47
38040 Grenoble cedex 09
France
Bureau : C08 au Bât. Michel Dubois
e-mail : mustapha.rachdi@univ-grenoble-alpes.fr

Démarche en analyse de Données : `FactoMiner`

- Quel type d'information ?
  - Tableau de contingence $\Longrightarrow$ AFC : `CA`
  - Tableau individus-variables $\Longrightarrow$ il faut faire un choix de la méthode (cf. point 3)
- Eléments actifs ? Quels éléments participeront à la construction des axes factoriels ?

  `ind.sup, quanti.sup, quali.sup, row.sup, col.sup`
- Nature des variables actives ?
  - Quantitative $\Longrightarrow$ ACP : `PCA`
  - Qualitative $\Longrightarrow$
    - s'il n'y a que 2 variables , construire le tableau croisé $\Longrightarrow$ AFC : `CA`
    - sinon $\Longrightarrow$ AFCM : `MCA`
- Si ACP, faut-il réduire les variables ?
  oui $\Longrightarrow$ ACP réduite : `scate.unit = TRUE`
- Y a-t-il des données manquantes ?
  si oui $\Longrightarrow$ utiliser le package `missMDA` pour compléter le jeu de données

Une fois qu'on a décidé de la méthode à utiliser

- Lancer l'analyse factorielle : `PCA, CA,` ou `MCA`
- Voir les résultats et construire les graphes : `summary, plot`
- Décrire les axes factoriels par les variables initiales : `dimdesc`
- Faire une classification des individus et décrire les classes : `HCPC`

Quelques extensions que l'on va étudier

- Lorsque certaines variables sont quantitatives et d'autres qualitatives. Comment faire pour qu'elles soient toutes actives ?
  - Découper les variables quantitatives en classes puis `ACM`
  - Analyse factorielle des données mixtes : `FAMD`
- Lorsque les données sont structurées en groupes : plusieurs groupes de variables *quantitatives* ou *qualitatives* ou des *tableaux de contingence* $\Longrightarrow$ Analyse factorielle multiple : `MFA`

Généralités sur les matrices

Ce chapitre rassemble des rappels d'algèbre linéaire utiles à la statistique multidimensionnelle. Les résultas seront énoncés sans démonstration.

Définition

On appelle A, matrice d'ordre (n, p) et de terme général a_{ij}*, un tableau rectangulaire à n lignes et p colonnes :*

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ . & . & \dots & . \\ . & . & . & . \\ . & . & . & . \\ a_{n1} & a_{n2} & \dots & a_{np} \end{pmatrix}$$

Exemple

$$A = \begin{pmatrix} 1 & 0 & 2 & 8 \\ 3 & 1 & 2 & 0 \\ 5 & 7 & 6 & 9 \end{pmatrix} \text{ matrice d'ordre } (3, 4).$$

Cas particuliers

- Si $n = 1$, A est appelée *vecteur-ligne*.
- Si $p = 1$, A est appelée *vecteur-colonne*.
- Si $n = p$, A est appelée *matrice carrée* d'ordre n.

Exemple

$$A = \begin{pmatrix} 1 & 5 & 3 \\ 2 & 4 & 1 \\ 7 & 2 & 8 \end{pmatrix} \text{ matrice carrée d'ordre } 3$$

Définition

Les vecteurs sont dits linéairement indépendants *si la seule combinaison linéaire de ces vecteurs qui soit nulle est la combinaison linéaire nulle.*

Définition

Le rang d'une matrice *A d'ordre* (n, p) *est le nombre maximum de vecteurs-colonne de A linéairement indépendants. C'est aussi le nombre maximum de vecteurs-ligne de A linéairement indépendants.*

Addition de matrices

Définition

Soient A et B deux matrices de même ordre $A = (a_{ij})$ et $B = (b_{ij})$, on pose :

$$A + B = (a_{ij} + b_{ij}).$$

Exemple

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 5 \\ 4 & 0 \end{pmatrix} , \; B = \begin{pmatrix} 0 & -1 \\ 5 & 6 \\ 3 & 4 \end{pmatrix} , \; A + B = \begin{pmatrix} 1 & 1 \\ 8 & 11 \\ 7 & 4 \end{pmatrix} .$$

Propriété

L'addition de deux matrices est associative et commutative :

$$A + (B + C) = (A + B) + C \text{ et } A + B = B + A.$$

Multiplication de matrices : %*%

Définition

Soit λ un scalaire et A une matrice d'ordre (n, p), alors

$$\lambda A = \lambda(a_{ij}) = (\lambda\, a_{ij}).$$

Soit $B = (b_{ij})$ une matrice d'ordre (p, q), le produit AB est défini comme étant la matrice $C = (c_{ij})$ d'ordre (n, q) dont les éléments sont donnés par

$$c_{ij} = \sum_{k=1}^{p} a_{ik}\, b_{kj}.$$

Pour que le produit de deux matrices soit possible, il faut que le nombre de colonnes de celle de gauche soit égal au nombre de lignes de celle de droite.

Multiplication de matrices et transposée : `%*%` et `t`

Exemple

$$A = \begin{pmatrix} 1 & 4 & 2 \\ 0 & 1 & 3 \end{pmatrix} \ , \ B = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & -1 & 2 \end{pmatrix} \ , \ AB = \begin{pmatrix} 1 & 2 & 3 \\ 0 & -2 & 6 \end{pmatrix}.$$

Propriété

La multiplication de matrices est associative mais non commutative :

$$A(BC) = (AB)C \ , \ (A+B)C = AC + BC \ , \ A(B+C) = AB + AC.$$

Définition

Soit A une matrice d'ordre (n, p)*, on appelle* matrice transposée $({}^tA)$ de A *la matrice d'ordre* (p, n) *dont les colonnes sont les lignes de A.*

Propriété

$${}^t(A+B) = {}^tA + {}^tB \qquad et \qquad {}^t(AB) = {}^tB\,{}^tA$$

Trace d'une matrice carrée

Définition

Soit A une matrice carrée d'ordre n de terme général a_{ij}, la trace de *A, notée Trace(A) est la somme de ses éléments diagonaux :*

$$Trace(A) = \sum_{i=1}^{n} a_{ii}.$$

Exemple

Si $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$, *alors* $Trace(A) = 15$.

Propriété

- $Trace(\lambda A) = \lambda\, Trace(A)$.
- $Trace(A + B) = Trace(A) + Trace(B)$.
- $Trace(AB) = Trace(BA)$.

Déterminant des matrices carrées d'ordres 2 et 3

- Soit A une matrice carrée d'ordre 2,

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

Le déterminant est le scalaire noté $det(A)$ ou $|A|$, tel que

$$det(A) = a_{11}a_{22} - a_{12}a_{21}.$$

- Soit A une matrice carrée d'ordre 3,

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}.$$

Le déterminant de A est donné par :

$$\begin{aligned} det(A) &= a_{11}a_{22}a_{33} + a_{21}a_{32}a_{13} + a_{31}a_{12}a_{23} \\ &\quad - a_{13}a_{22}a_{31} - a_{23}a_{32}a_{11} - a_{33}a_{12}a_{21}. \end{aligned}$$

Déterminant des matrices carrées d'ordres 2 et 3

Exemple

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \ , \ B = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$$

$det(A) = -2$ *et* $det(B) = 0$.

Nous ne donnons pas ici le calcul du déterminant d'une matrice d'ordre n $(n > 3)$.

Déterminants et inverses des matrices carrées : `det` et `solve`

Propriété

- $det(AB) = det(A)\, det(B)$.
- $det(I_n) = 1$.
- *Si* D_n *est diagonale d'éléments diagonaux* (d_{ii}), *alors* $det(D_n) = \prod_{i=1}^{n} d_{ii}$.
- $det({}^tA) = det(A)$.
- $det(A) \neq 0$ *si, et seulement si,* A *est inversible.*
- *Si* A *est inversible, alors* $det(A^{-1}) = (det(A))^{-1}$.

Valeurs propres et vecteurs propres : `eigen$values` et `eigen$vectors`

Définition

Soit A une matrice carrée. On dit que u est un vecteur propre *de A, si :*

- *u est différent de zéro (le vecteur nul).*
- *il existe λ scalaire, tel que $Au = \lambda u$.*

Le scalaire λ est appelé la valeur propre *de A associée à u. Le vecteur u est également dit* vecteur propre *associé à la valeur propre λ.*

Propriété

- *Si λ est une valeur propre de A, alors λ est racine de $det(A - \lambda I_n)$, qui est un polynôme de degré n.*
- *Si u vecteur propre associé à la valeur propre λ, alors $Au_i = \lambda_i u_i$.*
- *Si λ_i pour $i = 1, \ldots, m$ désignent les valeurs propres de la matrice A, alors $\sum_{i=1}^{m} \lambda_i = Trace(A)$.*

Matrices réelles symétriques

Définition

Une matrice carrée est symétrique*, si elle est égale à sa transposée.*

Exemple

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 0 & 4 \\ 3 & 4 & 8 \end{pmatrix} = {}^tA.$$

Remarque

Le produit d'une matrice par sa transposée donne une matrice symétrique.

Les matrices ${}^tAA$ et $A{}^tA$ sont symétriques même si A n'est pas symétrique.

Matrices réelles symétriques

Propriété

Soit A une matrice réelle symétrique, alors :.

- *A est diagonalisable.*
- *Les valeurs propres de A sont toutes réelles. .*
- *Si x et y sont deux vecteurs propres, alors*

$$ {}^t x\, A\, y = {}^t y\, A\, x. $$

Matrices définies positives

Définition

Une matrice réelle symétrique M est dite définie positive *si*

$$^tX\,M\,X > 0, \quad \textit{pour tout vecteur } X \textit{ non nul.}$$

Théorème

Une matrice M est définie positive si, et seulement si, toutes ses valeurs propres sont strictement positives.

Théorème

Une matrice réelle symétrique A est définie positive si et seulement si il existe une matrice non singulière C telle que $A = C\,^tC$.

Théorème

Si A est une matrice non singulière, alors $^tA\,A$ est définie positive.

Métrique

Définition

Une métrique *est une matrice symétrique définie positive.*

Remarque

La matrice identité I_n est une métrique, d'autres exemples seront donnés plus loin.

Norme

Dans toute la suite M désigne une métrique.

Définition

On appelle norme *sur un espace vectoriel E, une application de E dans $\mathbb{R}_+$, telle que*

$$x \longmapsto \|x\|$$

et vérifiant les propriétés suivantes :

- $\|x\| = 0 \Longleftrightarrow x = 0$.
- $\|ax\| = |a| \ \|x\|$ *pour tout scalaire a et tout x de E.*
- $\|x + y\| < \|x\| + \|y\|$ *pour tout x et y de E (inégalité triangulaire).*

Définition

Soit x un vecteur, sa norme *est définie par*

$$\|x\|_M = \sqrt{{}^t x \, M \, x}.$$

Un vecteur est dit normé si et seulement si sa norme vaut 1.

Produit scalaire

Définition

Soient x et y deux vecteurs et M une métrique.

- *Leur produit scalaire, relativement à M, est défini par :*

$$< x, y >_M = {}^t x\, M\, y.$$

  La notation $< \cdot, \cdot >_M$ permet de préciser la métrique utilisée.
- *La* distance *entre x et y est définie par $d(x, y) = \|x - y\|_M$.*
- *Le vecteur x est M-*orthogonal *à y si, et seulement si, ${}^t x\, M\, y = 0$.*
- *Le* cosinus *de l'angle α formé par ces vecteurs est donné par :*

$$\cos(\alpha) = \frac{< x, y >_M}{\|x\|_M \, \|y\|_M}.$$

Projecteur et projection

Définition

On appelle projecteur *une métrique idempotente.*

Exemple

La matrice nulle et la matrice identité sont des projecteurs.

Définition

Soit x un vecteur quelconque et $\mathcal{D}_u$ la droite engendrée par le vecteur u. Le point

$$x_h = \frac{{}^t x\, M\, u}{\|u\|} u$$

est la projection orthogonale *du point x sur $\mathcal{D}_u$.*

Définition

Une application $g : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$*, est dit une forme bilinéaire si, et seulement si, pour tout* $(u, v, w) \in \mathbb{R}^3$ *et pour tout* $\lambda \in \mathbb{R}$ *:*

$$\begin{aligned} g(u+v, w) &= g(u, w) + g(v, w) \\ g(u, v+w) &= g(u, v) + g(u, w) \\ g(\lambda u, v) &= g(u, \lambda v) = \lambda g(u, v) \end{aligned}$$

A toute forme bilinéaire, on peut associer une matrice M, telle que :

$$g(u, v) = {}^t u \, M \, v$$

Formes bilinéaires

Définition

Une forme bilinéaire g est dite symétrique si et seulement si, pour tout $u, v \in \mathbb{R}^n$:

$$g(u, v) = g(v, u).$$

Remarque

Si g est symétrique, alors la matrice M associée à g est symétrique.

Définition

On appelle forme quadratique associée à la forme bilinéaire g, l'application

$$g : \begin{array}{lll} \mathbb{R}^n & \longrightarrow \mathbb{R} \\ u & \longmapsto f(u) = g(u,u) = {}^tuMu \end{array}$$

M étant une matrice carrée de dimension n, qui s'écrit $(M_{ij})_{1\leq i,j\leq n}$ *avec* $M_{ij} = M_{ji}$.

Donc

$${}^tuMu = \sum_{i=1}^{n} u_i^2 M_{ii} + 2\sum_{i=1}^{n}\sum_{i<j} u_i M_{ij} u_j$$

Formes quadratiques

Formes quadratiques définies positives La forme quadratique ${}^tuMu$ est définie positive si, pour tout vecteur $u \in (\mathbb{R}^n)^*$,

$$ {}^tuMu > 0. $$

Formes quadratiques semi-définies positives La forme quadratique ${}^tuMu$ est semi-définie positive si, pour tout vecteur $u \in (\mathbb{R}^n)^*$,

$$ {}^tuMu \geq 0. $$

Tableau de données multidimensionnelles

En général les observations de p variables sur n individus sont présentées sous le forme d'une matrice X d'ordre (n, p) de terme général x_i^j où x_i^j désigne la valeur prise par la variable numéro j sur l'individu numéro i.

Exemple

$$X = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad et \quad X_2 = \begin{pmatrix} 0 \\ 2 \\ 0 \end{pmatrix}$$

- Chaque variable X_j est considérée comme un vecteur d'un espace à n dimensions appelé *espace des variables*.
- De même, chaque individu noté e_i est un élément d'un espace vectoriel appelé *espace des individus*.

Matrice de poids

Définition

On supposera qu'à chaque individu e_i est associé un poids p_i positif tel que $\sum_{i=1}^{n} p_i = 1$.

On notera $D_n = \begin{pmatrix} p_1 & & 0 \\ & \ddots & \\ 0 & & p_n \end{pmatrix}$ la matrice diagonale de poids.

Définition

Soit $\mathbb{I}_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ *le vecteur dont les n composantes sont égaux à* 1. *Le* centre de gravité *du nuage des individus est le vecteur :* $g = {}^tX\, D_n\, \mathbb{I}_n$.

Matrice de variance-covariance et de corrélation

Définition

La matrice de variance-covariance V *est définie par :*

$$V = {}^tX\, D_n\, X - g\, {}^tg.$$

Si les poids sont centrés i.e., $g = 0$*, alors la matrice de variance-covariance devient :*

$$V = {}^tX\, D\, X$$

Remarque

La matrice de variance-covariance est une matrice carrée et symétrique.

Définition

Si on note $D_{1/s}$ *la matrice diagonale des inverses des écart-types, la matrice de* corrélation*, notée* R *est définie par*

$$R = D_{\frac{1}{s}}\, V\, D_{\frac{1}{s}}.$$

Utilisation de la métrique statistique

- Pour mesurer la distance entre individus ou entre variables, il faut définir un produit scalaire, donc une métrique.
- La métrique que l'on utilise dans le cas de variables, est la matrice des poids D_n.
- Pour mesurer la proximité entre individus, on utilise une des métriques suivantes :
  - la métrique euclidienne classique : c'est le cas où $M = I_n$, c'est-à-dire celle qui revient à utiliser le produit scalaire usuel. Par conséquent, on a la distance euclidienne :

$$d^2(x, y) = \sum_i (x_i - y_i)^2.$$

  - La métrique de Mahalanobis correspond à $M = V^{-1}$.