Analyse factorielle d’un nuage de points

GHAZI BEL MUFTI

ghazi.belmufti@gmail.com

ESSAI-1 / ANALYSE DES DONNÉES

Plan

1. Tableau de données ; notations
2. Nuages des individus et nuages des variables
3. Caractéristiques d'un nuage de points
4. Résultat d'une analyse factorielle

Rappels d'algèbre linéaire : espace euclidien, produit scalaire

Soit E un espace vectoriel (e. v.) sur $\mathbb{R}$ de dimension n. Si f est une application de $\mathbb{E} \times \mathbb{E}$ dans $\mathbb{R}$, on dit que :

(1) f est bilinéaire si f est linéaire par rapport à chaque composante

(2) f est symétrique si pour tout x, y de $\mathbb{E}$, $f(x,y) = f(y,x)$

(3) f est définie si pour tout x de $\mathbb{E}$, $f(x,x) = 0 \Leftrightarrow x = 0$

(4) f est positive si pour tout x de $\mathbb{E}$, $f(x,x) \geq 0$

Si f vérifie les conditions (1) à (4), on dit que f est une forme bilinéaire symétrique définie positive ou encore que f est un **produit scalaire**. Dans ce cas, la forme f est associée à :

- une forme quadratique q : $q(x) = f(x,x)$
- une norme euclidienne $|| \; ||$: $||x|| = \sqrt{f(x,x)}$
- une distance euclidienne d : $d(x,y) = ||x - y||$
- une orthogonalité $\perp_f$: x est f-orhtogonal à $y \Leftrightarrow f(x,y) = 0$

• On appelle **espace euclidien** tout espace vectoriel muni d'un produit scalaire.

• Si f est une produit scalaire défini sur $\mathbb{E}$ que l'on munit de la base $\{e_i\}_{1\leq i\leq n}$, alors la matrice M de terme général $f(e_i, e_j)$ pout i et j variant de 1 à n (c'est-à-dire $M = (f(e_i, e_j))_{1\leq i\leq n}$) vérifie :

$$\forall x, y \in \mathbb{E} \quad f(x, y) = x'My$$

• On dit alors que M est la matrice associée au produit scalaire f.

• On en déduit :

$$d^2(x, y) = ||x - y||^2 = q(x - y) = (x - y)'M(x - y)$$

$$x \ \text{ est } M\!-\!\text{orthogonal à } y \ \Leftrightarrow x'My = 0$$

• Lorsque l'on ne précise pas le produit scalaire utilisé, c'est qu'il s'agit du produit scalaire "usuel", c'est-à-dire du produit scalaire, noté $<, >$ défini par :

$$< x, y >= \sum_{i=1}^{n} x_i y_i = x'y = x'Iy$$

• Le produit scalaire usuel a donc pour matrice associée la matrice identité I. Ainsi, un système $\{x_i\}_{1\leq i\leq r}$ sera dit orthonormé si les vecteurs x_i sont normés et orhtogonaux deux à deux pour le produit scalaire usuel, c'est-à-dire si :

$$\forall i,j,\ < x_i, x_j >= x_i' x_j = \delta_{ij}$$

Remarques

1. Un système $\{x_i\}_{1\leq i\leq r}$ est orthonormé si la matrice X dont les vecteurs colonnes sont les x_i, c'est-à-dire si la matrice $X = (x_1, \ldots, x_r)$, est orthogonale, autrement dit si :

$$X'X = I \text{ ou } X^{-1} = X'$$

2. Notation : étant donné un produit scalaire f quelconque, on peut écrire

$$f(x,y) = x'My = y'Mx =< Mx, y >=< x, My >=< x, y >_M$$

Par la suite, M pourra aussi bien désigner le produit scalaire f ou sa forme quadratique associée, ou simplement la matrice M qui sera appelée **métrique** (sur $\mathbb{E}$).

1. Tableau de données ; notations

On note $I = [\![1, n]\!]$ et $J = [\![1, p]\!]$ qui sont les ensembles d'indices désignant respectivement les n individus et les p variables.

$$X = (x_i^j)_{i \in I, j \in J} \in \mathcal{M}_{n,p}(\mathbb{R})$$

Ainsi les valeurs prises par la variable x^j pour les n individus se lisent sur la $j^{ème}$ colonne et les valeurs prises par l'individu i pour les p variables se lisent sur la $i^{ème}$ ligne de X.

$$\forall (i, j) \in I \times J, x_i = \begin{pmatrix} x_i^1 \\ \vdots \\ x_i^p \end{pmatrix} \in \mathbb{R}^p \quad x^j = \begin{pmatrix} x_1^j \\ \vdots \\ x_n^j \end{pmatrix} \in \mathbb{R}^n$$

2. Nuages des individus et nuages des variables

A la matrice X, on associe deux nuages de points : celui des individus et celui des variables.

$$\mathcal{M}_X = \{x_1, \dots, x_n\} \subseteq \mathbb{R}^p \quad \text{nuage des individus}$$
$$\mathcal{N}_X = \{x^1, \dots, x^p\} \subseteq \mathbb{R}^n \quad \text{nuage des variables}$$

• Le nuage $\mathcal{M}_X$ est muni de la métrique M de $\mathbb{R}^p$, et chaque individu i est muni d'une masse p_i telle que $\sum_i p_i = 1$ (souvent $p_i = \frac{1}{n}$).

• Le nuage $\mathcal{N}_X$ est muni de la métrique D_p de $\mathbb{R}^n$ avec

$$D_p = \mathrm{Diag}(p_i)_{1 \leq i \leq n}.$$

3. Caractéristiques d'un nuage de points

3.1 Centre de gravité. Le centre de gravité du nuage des individus affecté des poids p_i est le point g centre de gravité de $\mathcal{M}_X$:

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

La $j^{ème}$ coordonnée de g est donnée par : $g_j = \sum_{i=1}^{n} p_i x_i^j = \overline{x^j}$.

Ainsi g_j est la moyenne de la variable x^j et les coordonnées de g sont les moyennes des p variables.

Remarque : On vérifie facilement que $g = X' D_p 1_n$ où 1_n est le vecteur de $\mathbb{R}^n$ dont toutes les coordonnées sont égales à 1.

3.2 Matrice centrée. On centre toujours le nuage des individus sur le centre de gravité g, c'est-à-dire, on construit un nouveau tableau Y tel que

$$\forall (i,j) \in I \times J, y_i^j = x_i^j - \overline{x^j}$$

soit

$$\forall i \in I, y_i = x_i - g$$

3.3 Matrice Variance. Par définition, la matrice variance V, des p variables pour les n individus est une matrice carrée d'ordre p et de terme général $v_{j,j'}$ donnée par :

$$\forall (j,j') \in [\![1,p]\!], v_{j,j'} = cov(x^j, x^{j'}) = \sum_{i=1}^{n} p_i(x_i^j - g_j)(x_i^{j'} - g_{j'}) = < y^j, y^{j'} >_{D_p} .$$

En notations matricielles, on a :

$$V = Y' D_p Y = X' D_p X - gg'$$

3.4 Le tableau centré réduit. En divisant chaque variable par son écart-type, on obtient un nouveau tableau Z dont les variables sont toutes centrées réduites. On a :

$$\forall (i,j) \in I \times J, z_i^j = \frac{x_i^j - \overline{x^j}}{\sqrt{v_{jj}}}$$

D'où $Z = Y\Delta$ où $\Delta = diag(\frac{1}{\sqrt{v_{11}}}, \ldots, \frac{1}{\sqrt{v_{pp}}})$.

3.5 Matrice de corrélation. La matrice $Z' D_p Z$ est la matrice de corrélation. Donc $R = V_Z$, où V_Z désigne la matrice variance associée au tableau Z. On a :

$$R = \Delta V \Delta$$

4. Résultat d'une analyse factorielle

• L'analyse factorielle d'un tableau X a pour but de trouver un sous-espace E_k, de dimension k, tel que l'inertie totale de la projection de $\mathcal{M}_Y$ sur E_k soit maximale.

• On montre que les sous-espaces E_α solution de ce problème sont emboîtés lorsque k varie.

$$E_1 \subset E_2 \ldots \subset E_p = E$$

• La droite E_1 est générée par un vecteur normé u_1. Le plan E_2 est générée par le couple (u_1, u_2) où u_2 est normé et orthogonal à u_1. Donc $E_1 \subset E_2$. Le sous-espace E_k est généré par $(u_1, \ldots, u_k)$ où u_k normé est orthogonal à E_{k-1} généré par $(u_1, \ldots, u_{k-1})$.

• Le vecteur u_k $(1 \leq k \leq p)$ s'appelle le k ème **vecteur axial factoriel** et la droite dirigée par u_k, notée Δu_k, est appelée $k^{ème}$ axe factoriel.

• Les vecteurs u_k $(1 \leq k \leq p)$ forment un système orthonormé de E : l'analyse factorielle revient à un changement de base.

4.1 Axes factoriels u_α $(1 \leq \alpha \leq p)$. u_α est solution de l'équation aux valeurs propres et vecteurs propres :

$$\begin{cases} VM\,u_\alpha &= \lambda_\alpha u_\alpha \\ (u_\alpha)' M u_\beta &= \delta_\alpha^\beta \end{cases}$$

avec $\delta_\alpha^\beta = 0$ si $\alpha \neq \beta$ et 1 sinon
L'inertie de la projection de $\mathcal{M}_Y$ sur Δu_α vaut λ_α.

4.2 Les composantes principales Ψ_α $(1 \leq \alpha \leq p)$. Par définition

$$\Psi_\alpha = \begin{pmatrix} \Psi_{1,\alpha} \\ \vdots \\ \Psi_{n,\alpha} \end{pmatrix} = YMu_\alpha$$

est appelée $\alpha^{ème}$ **composante principale**. On a :

$$\Psi_{i,\alpha} = (y_i)' M u_\alpha$$

D'autre part, on a :

$$VMu_\alpha = Y' D_p Y M u_\alpha = Y' D_p \Psi_\alpha = \lambda_\alpha u_\alpha$$

D'où, en multipliant par YM à gauche :

$$YMY'D_p\Psi_\alpha = \lambda_\alpha YMu_\alpha = \lambda_\alpha \Psi_\alpha$$

Autrement dit, en posant $W = YMY'$, le vecteur Ψ_α est vecteur propre de WD_p relatif à la valeur propre λ_α. Les Ψ_α sont donc solution de :

$$\begin{cases} WD_p\Psi_\alpha & = \lambda_\alpha \Psi_\alpha \\ (\Psi_\alpha)'D_p\Psi_\beta & = \lambda_\alpha \delta_\alpha^\beta \end{cases}$$

En effet :

$$\begin{aligned} < \Psi_\alpha, \Psi_\beta >_{D_p} &= \Psi'_\alpha D_p \Psi_\beta \\ &= u'_\alpha MY'D_pYMu_\beta \\ &= u'_\alpha MVMu_\beta \\ &= \lambda_\beta < u_\alpha, u_\beta >_M \end{aligned}$$

4.3 Inertie

• Soit $I_T = trace(VM)$ l'inertie totale du nuage. Le taux d'inertie expliqué par le $\alpha^{ème}$ axe factoriel, noté τ_α est la quantité

$$\tau_\alpha = \frac{\lambda_\alpha}{I_T} = \frac{\lambda_\alpha}{\sum_{i=1}^{p} \lambda_i}$$

• Le taux d'inertie expliqué par E_α, noté $\tau_{1\ldots\alpha}$ est la quantité

$$\tau_{1\ldots\alpha} = \frac{\lambda_1 + \ldots + \lambda_\alpha}{I_T} = \sum_{i=1}^{\alpha} \tau_i$$

4.4 Représentation des variables

- On suppose la matrice V de rang r. Comme V et VM ont même rang (M étant inversible), les valeurs propres $\lambda_{r+1}, \ldots, \lambda_p$ sont nulles.
- Un individu i a $p - r$ coordonnées nulles donc est caractérisé par r valeurs $\Psi_{i,1}, \ldots, \Psi_{i,r}$ au lieu des p coordonnées initiales.

E.V. engendré par les variables.

On a vu que :

$$\forall(\alpha, \beta) \in [\![1, p]\!]^2, < \Psi_\alpha, \Psi_\beta >_{D_p} = \begin{cases} \lambda_\alpha & si\ \alpha = \beta \\ 0 & si\ \alpha \neq \beta \end{cases}$$

En posant :

$$\forall \alpha \in [\![1, r]\!],\ v_\alpha = \frac{\Psi_\alpha}{\sqrt{\lambda_\alpha}},$$

On en déduit que :

$(v_1, \ldots, v_r)$ est une base D_p-orthonormale de $Vect(y^1, \ldots, y^p) = ImY$

.

Coordonnées des variables.

On se place dans la base $(v_1, \ldots, v_r)$ pour représenter les variables.
La α ème coordonnée de la variable centrée y^j est donnée par :

$$\eta_j^\alpha =< y^j, \frac{\Psi_\alpha}{\sqrt{\lambda_\alpha}} >_{D_p} .$$

On a :

$$\eta^\alpha = \begin{pmatrix} \eta_1^\alpha \\ \vdots \\ \eta_p^\alpha \end{pmatrix} = \sqrt{\lambda_\alpha} u_\alpha$$