Analyse des données

Chapitre 2

Analyse en Composantes Principales

5. Analyse d'un nuage de points

5.1 Ajustement du nuage des individus

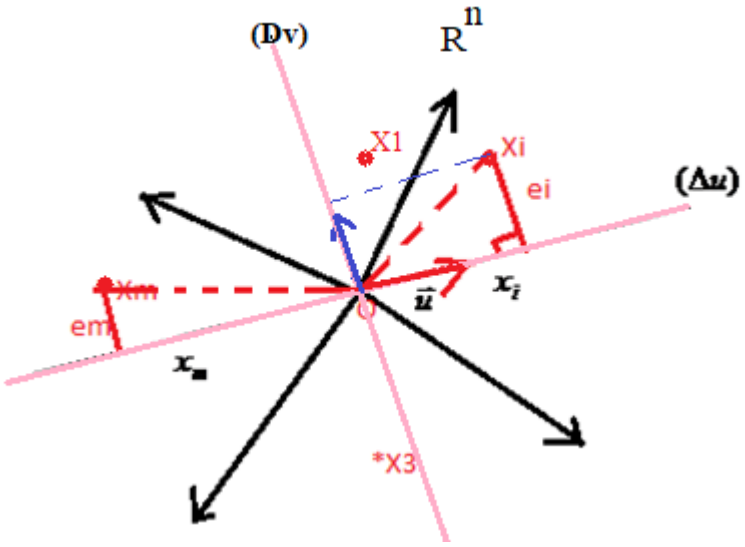
Considérons le nuage des individus défini par :

$$\mathcal{N}(I) = \left\{ \left(X_i, \frac{1}{m} \right) ; X_i \in \Re^n, i = 1, 2, \dots, m \right\}.$$

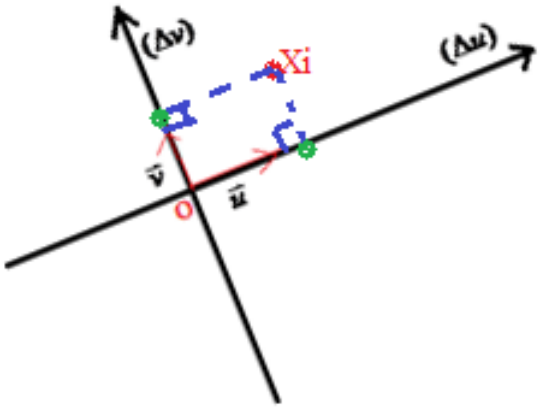
ii) Ajustement du nuage par un plan

Pour déterminer le meilleur plan ajustant le nuage, il suffit de déterminer la deuxième droite (Δv) de vecteur unitaire $\vec{v}$ passant par l'origine et perpendiculaire à (Δu). C'est à dire :

$$\|\vec{v}\|^2 = 1 \quad \text{et} \quad \langle \vec{u}, \vec{v} \rangle = 0.$$



Ce qui revient à projeter chaque individu sur les deux axes.



Analyse des données

La détermination de $\vec{v}$ revient à résoudre le problème suivant :

$$\textbf{Maximiser :} \begin{cases} \vec{v}^t \cdot (X^t \cdot X) \cdot \vec{v} \\ \|\vec{v}\|^2 = \vec{v^t} \cdot \vec{v} = 1 \\ \vec{v}^t \cdot \vec{u} = 0 \end{cases} \quad \textbf{(P).}$$

Dans ce cas, le **Lagrangien** s'écrit comme suit :

$$\boldsymbol{L(\vec{v}, \lambda, \mu) = \vec{v}^t \cdot (X^t \cdot X) \cdot \vec{v} - \lambda \cdot (\vec{v}^t \cdot \vec{v} - 1) - \mu \cdot (\vec{v}^t \cdot \vec{u} - 0).}$$

Où λ et μ sont les multiplicateurs de Lagrange.

La solution est donnée par :

$$Sol = \begin{cases} \dfrac{\partial L}{\partial \vec{v}} = 0 \\ \vec{v}^t \cdot \vec{v} = 1 \\ \vec{v}^t \cdot \vec{u} = 0 \end{cases}.$$

C'est-à-dire :

$$\boldsymbol{Sol = \begin{cases} 2 \cdot (X^t \cdot X) \cdot \vec{v} - 2\lambda \cdot \vec{v} - \mu . \vec{u} = 0 \\ et \\ \vec{v}^t \cdot \vec{v} = 1 \quad i.e \quad \|\vec{v}\|^2 = 1 \quad et \quad \vec{v}^t \cdot \vec{u} = 0 \end{cases}} \quad (2).$$

En multipliant l'équation (2) par $\vec{u}^t$, nous obtenons :

$$\boldsymbol{2 \cdot \vec{u}^t \cdot (X^t \cdot X) \cdot v - 2 \cdot \lambda \, \vec{u}^t \cdot \vec{v} - \mu \cdot \vec{u}^t \cdot \vec{u} = 0.}$$

C'est-à-dire :

$$\boldsymbol{2 \cdot \vec{u}^t \cdot (X^t \cdot X) \cdot v = \mu.}$$

Or :

$$(\boldsymbol{X^t \cdot X}) \cdot \vec{u} = \boldsymbol{\lambda \vec{u}} \qquad \text{c'est-à-dire} \qquad \vec{u}^t \cdot (X^t \cdot X) = \lambda \cdot \vec{u}^t$$

Alors :

$$\boldsymbol{\mu = 2\lambda \cdot \vec{u}^t . \vec{v} = 0.}$$

Donc,

$$\boldsymbol{\frac{\partial L}{\partial \vec{v}} = 0 \Leftrightarrow 2 \cdot (X^t \cdot X) \cdot \vec{v} - 2\lambda \cdot \vec{v} = 0 \Leftrightarrow (X^t \cdot X) \cdot \vec{v} = \lambda \cdot \vec{v}}$$

C'est-à-dire :

$\vec{v}$ est un vecteur propre de la matrice $\boldsymbol{X^t \cdot X}$ associé à la valeur propre $\boldsymbol{\lambda}$.

D'où,

le ***maximum*** de : $\vec{v}^t \cdot (X^t \cdot X) \cdot \vec{v}$

Correspond à la

2eme grande valeur propre de : $X^t \cdot X$.

Ainsi,

le ***meilleur plan*** est constitué du :

1er axe factoriel (Δu) et le **second axe factoriel** (Δv).

Question

Comment définir les nouvelles variables à partir de ces axes principaux ?

Réponse

Ces nouvelles variables se déterminent tout simplement par la projection de tous les individus.

6. Analyse en composantes principales

Définition 1

A chaque axe principal (factoriel) (Δu_j) est associé une **variable** C^j appelée ***composante principale***. Sa dimension est égale à m (nombre d'individus).

Pour tout j,

$(\Delta u_j) \longrightarrow C^j$

Définition 2

Une ***composante principale*** n'est que ***le vecteur constitué des projections des individus*** sur ***l'axe principal*** associé à cette nouvelle variable.

Par exemple,

$$C^1 = (x_1, x_2, \dots, x_i, \dots, \dots, x_m)^t.$$

Où,

x_i **Represente la projection de l'individu** X_i sur le premier axe (Δu_1).

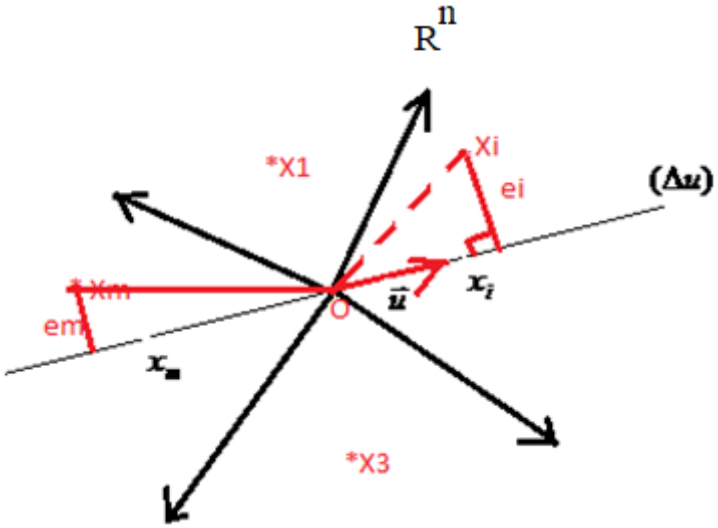
Pour tout : $i = 1, 2, \dots, m$.

6.1 Calcul explicite des composantes principales

La détermination explicite des composantes principales se fait moyennant le produit scalaire.

En effet,

Analyse des données



Pour tout individu $\boldsymbol{X_i = (x_i^1, x_i^2, \ldots, x_i^j, \ldots, x_i^n) \in \mathfrak{R}^n}$, $i = 1, 2, \ldots, m$, sa projection sur le premier axe (Δu_1) généré par le vecteur $\boldsymbol{u_1}$ est donnée par :

$$\boldsymbol{x_i = \langle X_i, u_1 \rangle.}$$

Alors :

$$\boldsymbol{C^1 = \begin{pmatrix} \langle X_1, u_1 \rangle \\ \vdots \\ \langle X_i, u_1 \rangle \\ \vdots \\ \langle X_m, u_1 \rangle \end{pmatrix}.}$$

Ou encore :

$$\boldsymbol{C^1 = (\langle X_1, u_1 \rangle, \langle X_2, u_1 \rangle, \ldots, \langle X_i, u_1 \rangle, \ldots, \ldots, \langle X_m, u_1 \rangle)^t.}$$

D'autre part,

$\boldsymbol{u_1}$ est un vecteur de $\boldsymbol{\mathfrak{R}^n}$, alors : $\boldsymbol{u_1 = (u_{11}, u_{21}, \ldots, u_{n1})^t}$.

Donc,

$$\boldsymbol{x_i = x_i^1 u_{11} + x_i^2 u_{21} + \ldots, x_i^j u_{i1} + \ldots, x_i^n u_{n1} = \sum_{j=1}^{n} x_i^j \cdot u_{j1}.}$$

Par conséquent,

$$\boldsymbol{C^1 = \begin{pmatrix} \sum_{j=1}^{n} x_1^j \cdot u_{j1} \\ \vdots \\ \sum_{j=1}^{n} x_i^j \cdot u_{j1} \\ \vdots \\ \sum_{j=1}^{n} x_m^j \cdot u_{j1} \end{pmatrix} = \begin{pmatrix} x_1^1 u_{11} + x_1^2 u_{21} + \cdots + x_1^n u_{n1} \\ \vdots \\ \vdots \\ \vdots \\ x_m^1 u_{11} + x_m^2 u_{21} + \cdots + x_m^n u_{n1} \end{pmatrix}.}$$

Que nous pouvons écrire sous la forme :

$$C^1 = \begin{pmatrix} \sum_{j=1}^{n} x_1^j \cdot u_{j1} \\ \vdots \\ \sum_{j=1}^{n} x_i^j \cdot u_{j1} \\ \vdots \\ \sum_{j=1}^{n} x_m^j \cdot u_{j1} \end{pmatrix} = u_{11} \begin{pmatrix} x_1^1 \\ \vdots \\ x_i^1 \\ \vdots \\ x_m^1 \end{pmatrix} + u_{21} \begin{pmatrix} x_1^2 \\ \vdots \\ x_i^2 \\ \vdots \\ x_m^2 \end{pmatrix} + \cdots + u_{n1} \begin{pmatrix} x_1^n \\ \vdots \\ x_i^n \\ \vdots \\ x_m^n \end{pmatrix}.$$

Autrement dit :

$$C^1 = u_{11}X^1 + u_{21}X^2 + \cdots + u_{n1}X^n.$$

C'est-à-dire :

$$C^1 = \sum_{j=1}^{n} u_{j1}X^j = Xu_1.$$

De même,

$$C^2 = u_{12}X^1 + u_{22}X^2 + \cdots + u_{n2}X^n.$$

C'est-à-dire :

$$C^2 = \sum_{j=1}^{n} u_{j2}X^j = Xu_2.$$

Ou, u_2 est le vecteur générateur du deuxième axe (Δu_2).

Conclusion
Les composantes *principales sont des combinaisons linéaires des variables initiales.*

En résumé

Les étapes à suivre pour réaliser une ***analyse en composantes principales*** sont données comme suit :

1. ***Centrer*** et ***normer*** (réduire) les variables

   $X \rightarrow Z$ **(matrice centrée-réduite).**

   C'est-à-dire :

$$Z = \begin{pmatrix} z_1^1 & \cdots & z_1^j & \cdots & z_1^n \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ z_i^1 & \cdots & z_i^j & \cdots & z_i^n \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ z_m^1 & \cdots & z_m^j & \cdots & z_m^n \end{pmatrix}.$$

   Avec :

Analyse des données

$$z_i^j = \frac{y_i^j}{\sigma_j}. \quad \textbf{Et} \quad y_i^j = x_i^j - \bar{X}^j.$$

2. ***Déterminer*** la matrice de ***corrélation***

$$R = \begin{pmatrix} 1 & r_1^2 & \cdots & \cdots & r_1^n \\ r_2^1 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & r_{m-1}^n \\ r_m^1 & \cdots & \cdots & r_m^{n-1} & 1 \end{pmatrix}.$$

Avec

$$r_i^j = r_{ij} = \rho_{X^i X^j} = \frac{cov(X^i, X^j)}{\sigma_i \sigma_j}.$$

3. ***Déterminer*** les ***q premières grandes valeurs propres*** de la matrice ***R*** : $\lambda_1, \lambda_2, \ldots, \lambda_q$.

q Représente la dimension du sous espace à retenir.

q n'est que le nombre de composantes principales i.e les nouvelles variables.

4. ***Déterminer*** les ***q axes principaux*** $F_1, F_2, \ldots, F_q$ engendrés par les ***q*** vecteurs propres normés associés aux valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_q$.

Ces ***axes principaux (factoriels)*** sont orthogonaux deux à deux et ils ne sont pas corrélés.

5. ***Détermination*** des q ***composantes principales*** $C^1, \ldots, C^q$ associées aux q axes principaux.

$$C^i = Z.u_i, \; i = 1, 2, \ldots, q.$$

Où, u_i est le vecteur propre normé générateur de l'axe principal (Δu_i).

Exemple

Soit le tableau de données suivant :

$$X = \begin{pmatrix} 4 & 5 \\ 6 & 7 \\ 8 & 0 \end{pmatrix}.$$

Alors :

X =

I \ *J*	X^1	X^2
X_1	**4**	**5**
X_2	**6**	**7**
X_3	**8**	**0**

Analyse des données

1. Centrage des données :

$$g = \left(\frac{4+6+8}{3}, \frac{5+7+0}{3}\right) = (\mathbf{6}, \mathbf{4}).$$

Alors :

Y =

I \ J_C	X^1	X^2
X_1	**4 - 6**	**5 - 4**
X_2	**6 – 6**	**7 - 4**
X_3	**8 – 6**	**0 - 4**

=

I \ J_C	Y^1	Y^2
X_1	**-2**	**1**
X_2	**0**	**3**
X_3	**2**	**-4**

Les écarts sont donnés par :

$$\sigma_1 = \sqrt{var(Y^1)} = \sqrt{\frac{1}{3}\sum_{i=1}^{3}(y_i^1)^2} = \sqrt{\frac{1}{3}(4+4)} = \sqrt{\frac{8}{3}}.$$

$$\sigma_2 = \sqrt{var(Y^2)} = \sqrt{\frac{1}{3}\sum_{i=1}^{3}(y_i^2)^2} = \sqrt{\frac{1}{3}(1+9+16)} = \sqrt{\frac{26}{3}}.$$

D'où, la ***Matrice centrée –réduite des données*** :

Z =

I \ J_{CR}	Z^1	Z^2
X_1	$\mathbf{-2/\sigma_1}$	$\mathbf{1/\sigma_2}$
X_2	$\mathbf{0/\sigma_1}$	$\mathbf{3/\sigma_2}$
X_3	$\mathbf{2/\sigma_1}$	$\mathbf{-4/\sigma_2}$

I \ J_{CR}	Z^1	Z^2
X_1	$-\sqrt{3/2}$	$\sqrt{\mathbf{3/26}}$
X_2	**0**	$\mathbf{3\sqrt{3/26}}$
X_3	$\sqrt{\mathbf{3/2}}$	$\mathbf{-4\sqrt{3/26}}$

2. Les matrices des variances-covariances et des corrélations sont données par :

$$\boldsymbol{V = \frac{1}{3}(Y^t \times Y)} \qquad \textbf{et} \qquad \boldsymbol{R = \frac{1}{3}(Z^t \times Z)}.$$

Ce qui donnent :

$$V = \frac{1}{3}\begin{pmatrix} -2 & 0 & 2 \\ 1 & 3 & -4 \end{pmatrix} \times \begin{pmatrix} -2 & 1 \\ 0 & 3 \\ 2 & -4 \end{pmatrix} = \mathbf{\frac{1}{3}\begin{pmatrix} 8 & -10 \\ -10 & 26 \end{pmatrix}}.$$

Et

$$R = \frac{1}{3}\begin{pmatrix} 3 & -15/2\sqrt{13} \\ -15/2\sqrt{13} & 3 \end{pmatrix} = \mathbf{\begin{pmatrix} 1 & -5/2\sqrt{13} \\ -5/2\sqrt{13} & 1 \end{pmatrix}}.$$

Cette dernière matrice peut être présentée comme suit :

Analyse des données

$\boldsymbol{R}$ =

Corr	X^1	X^2
X^1	**1**	$-\frac{5}{2\sqrt{13}}$
X^2	$-\frac{5}{2\sqrt{13}}$	**1**

3. Détermination des axes principaux

– ***Détermination des valeurs propres*** Il suffit de résoudre l'équation :

$$P_R(\lambda) = det(R - \lambda.Id) = 0.$$

Nous avons :

$$P_R(\lambda) = det(R - \lambda.Id) = \begin{vmatrix} 1-\lambda & -\frac{5}{2\sqrt{13}} \\ -\frac{5}{2\sqrt{13}} & 1-\lambda \end{vmatrix}.$$

C'est-à-dire :

$$P_R(\lambda) = (1-\lambda)^2 - \left(\frac{5}{2\sqrt{13}}\right)^2.$$

Par suite,

$$P_R(\lambda) = 0 \Leftrightarrow \left(1 - \lambda - \frac{5}{2\sqrt{13}}\right) \times \left(1 - \lambda + \frac{5}{2\sqrt{13}}\right) = 0.$$

Ce qui donne :

$$\begin{cases} \lambda_1 = 1 - \frac{5}{2\sqrt{13}} \approx 0.31 \\ \lambda_2 = 1 + \frac{5}{2\sqrt{13}} \approx \mathbf{1.69} \end{cases}.$$

– ***Détermination des vecteurs propres***

i) ***Pour*** $\lambda_1 = 1 + \frac{5}{2\sqrt{13}} \approx 1.69.$

Soit $v_1 = (x, y)^t$ un vecteur propre non nul, associé à la valeur propre $\boldsymbol{\lambda_1}$. Alors :

$$R \times v_1 = \lambda_1 \times v_1.$$

C'est-à-dire :

$$\begin{cases} x - 0.69y = (1.69)x \\ -0.69x + y = (1.69)y \end{cases}.$$

Ce qui est équivalent à :

Analyse des données

$$\begin{cases}(-0.69)\cdot x-(0.69)y=0\\(-0.69)\cdot x+(-0.69)y=0\end{cases}$$

La résolution du système donne :

$$y=-x.$$

Par suite,

$$v_1=(x,-x)^t=x(1,-1)^t \quad \text{avec } x\in\Re^*.$$

Donc,

$$v_1=(1,-1)^t.$$

Par conséquent,

Le ***1er axe principal*** est engendré par le vecteur unitaire : $u_1=\frac{1}{\|v_1\|}\cdot v_1$.

C'est-à-dire :

$$u_1=\left(\frac{1}{\sqrt{2}},-\frac{1}{\sqrt{2}}\right)^t.$$

ii) Pour $\lambda_2=1-\frac{5}{2\sqrt{13}}\approx 0.31$.

Soit $v_2=(x,y)^t$ un vecteur propre associé à la valeur propre $\boldsymbol{\lambda_2}$. Alors :

$$R\times v_2=\lambda_2\times v_2.$$

Ce qui est équivalent à :

$$\begin{cases}(0.69)x-(0.69)y=0\\(-0.69)x+(0.69)y=0\end{cases}$$

La résolution donne :

$$x=y.$$

Ce qui donne :

$$v_2=(x,x)^t=x(1,1)^t \quad \text{avec } x \text{ quelconque dans } \Re^*.$$

Donc,

$$v_2=(1,1)^t.$$

Par conséquent,

Le ***2ème*** **axe principal** *est engendré par le vecteur unitaire :* $u_2=\frac{1}{\|v_2\|}v_2$.

C'est-à-dire :

$$u_2=\left(\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}\right)^t.$$

4. Détermination des composantes principales

   Les composantes principales sont définies par :

$$C^i=Z.u_i,\ i=1,2.$$

Analyse des données

Avec :

$$Z = \begin{pmatrix} -\sqrt{3/2} & \sqrt{\mathbf{3/26}} \\ 0 & 3\sqrt{\mathbf{3/26}} \\ \sqrt{3/2} & -\mathbf{4}\sqrt{\mathbf{3/26}} \end{pmatrix}.$$

Faites les calculs, déterminer les composantes principales pour appliquer ce qui suit.

6.2 Propriétés des composantes principales

Les composantes principales vérifient les propriétés suivantes :

1) La variance d'une composante principale est égale à l'inertie portée par l'axe principal qui lui est associé c'est à dire pour chaque composante C^i.

$$Var(C^i) = \lambda_i,\ i = 1, 2, ..., q\ .$$

Où q représente la dimension du sous espace ajustant le nuage de points.

2) Les composantes principales sont ***non corrélées deux à deux*** c'est-à-dire les axes associés sont orthogonaux deux à deux.

$Cov(C^i, C^j) = 0$ pour tout i *et* j dans $\{1, 2, ..., q\}$.

Vous pouvez les vérifier sur les composantes de l'exemple.