Analyse des données

Chapitre 2

Analyse en Composantes Principales

4. Caractéristiques numériques

Avant de présenter les quelques caractéristiques de base utiles dans l'analyse des données ou plus précisément dans l'application des méthodes factorielles.

Il faut préciser un point important : Il s'agit d'accorder un poids positif à chaque individu.

C'est-à-dire :

A chaque individu X_i un poids positif $\boldsymbol{w_i}$ lui est affecté.

$$(X_i, \boldsymbol{w_i}).$$

Dans notre cas, ces poids sont choisis tels que : $\sum_{i=1}^{m} w_i = 1$.

Ce poids reflète l'importance de l'individu X_i par rapport aux autres individus.

Dans la pratique, le même poids est accordé à tous les individus, il est égal à $\mathbf{1/m}$.

Ce qui nous conduit à définir une matrice dite **matrice des poids**, elle est définie comme suit :

$D = (d_{ii})_i$ avec $d_{ii} = w_i$ pour tout $i = 1,2, \dots, m$.

$$D = \begin{pmatrix} w_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_m \end{pmatrix} = \begin{pmatrix} \frac{1}{m} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{m} \end{pmatrix}.$$

- ***Moyenne arithmétique***

La moyenne arithmétique $\overline{\boldsymbol{X^k}}$ calculée sur une variable :

$\boldsymbol{X^k = (x_1^k,\ x_2^k,\ \dots, x_j^k\ ,\ x_m^k)^t}$ est donnée d'une manière générale par :

$$\overline{\boldsymbol{X^k}} = \sum_{j=1}^{m} w_j \cdot x_j^k.$$

Dans notre cas :

$$\overline{\boldsymbol{X^k}} = \frac{1}{m}\sum_{j=1}^{m} x_j^k.$$

- ***Centre de gravité***

Le centre de gravité du nuage de points est défini par le vecteur :

$$\boldsymbol{g = \left(\overline{X^1},\ \overline{X^2},\ \dots\ ,\ \overline{X^j},\ \dots\ ,\ \overline{X^n}\right) \in \Re^n}.$$

Où,

$\overline{X^j}$ est la ***moyenne arithmétique*** de la ***variable*** X^j pour tout $j = 1,2,\ldots,n$.

- ***Variance***

  La variance d'une variable X^k est définie d'une manière générale par :

$$Var(X^k) = \sum_{j=1}^{m} w_j \cdot \left(x_j^k - \overline{X^k}\right)^2.$$

  Dans notre cas :

$$Var(X^k) = \frac{1}{m}\sum_{j=1}^{m} \left(x_j^k - \overline{X^k}\right)^2.$$

- ***Ecart-type***

  L'écart-type d'une variable X^k n'est que la racine carrée de sa variance :

$$\sigma(X^k) = S_k = \sqrt{Var(X^k)}.$$

- ***Inertie***

  L'inertie d'un nuage de points est la quantité d'informations contenue dans la matrice de données. Elle est définie par la somme pondérée des variances des n variables :

$$I = \sum_{j=1}^{n} w_j \cdot Var(X^j).$$

  Elle représente la ***dispersion***

5. Mesures d'évaluation (comparaison)

Il s'agit de définir des **mesures** pour pouvoir **comparer** les individus et les variables.
D'après la matrice de l'exemple précédent, nous pouvons voir qu'il y a bien deux types de notions :

- Les individus qui sont présentés par les étudiants
- Les variables sont des informations sur les individus.

Donc,

Comme les individus et les variables définissent deux notions différentes quoiqu'elles soient complémentaires, alors ***deux types de mesures*** seront définis.

- Pour les ***variables***, il s'agit d'évaluer les **liaisons** entre elles
- Pour les ***individus***, il s'agit d'évaluer les **ressemblances** entre eux.

Autrement dit :

Analyse des données

Pour les individus

Il s'agit d'évaluer la ressemblance entre les individus : chercher les individus qui se ***ressemblent*** et ceux qui sont ***différents***. Ceci nous permettra de définir des groupes homogènes d'individus.

Pour les variables

Il s'agit de chercher les variables qui sont ***corrélées*** **positivement** ou bien **négativement** ce qui nous conduit à ***réduire*** les variables initiales à un nombre très petit de nouvelles variables dites ***composantes principales***.

Donc, chaque ***composante principale*** représente un ***groupe*** de ***variables corrélées***.

5.1 Rappels

- ***Produit scalaire***

  Soient u et v deux vecteurs quelconques de $\Re^n$, alors leur ***produit scalaire*** noté $\langle u, v\rangle$ ou bien $u \cdot v$ est défini par :

  $$\langle \boldsymbol{u}, \boldsymbol{v}\rangle = \|\boldsymbol{u}\| \cdot \|\boldsymbol{v}\| \cdot \boldsymbol{cos}(\boldsymbol{u}, \boldsymbol{v}) = \langle \boldsymbol{v}, \boldsymbol{u}\rangle.$$

  Le produit scalaire de u et v peut être défini directement en fonction des coordonnées des vecteurs en question par la relation suivante :

  $$\langle u, v\rangle = u^t \cdot v = u \cdot v^t = \sum_{i=1}^{n} u_i \cdot v_i.$$

  Remarque

  - Si les deux vecteurs u et v sont ***orthogonaux***, alors leur ***produit scalaire*** est ***nul.***
  - Si les deux vecteurs u et v sont ***unitaires*** c'est-à-dire : $\|\boldsymbol{u}\| = \|\boldsymbol{v}\| = \mathbf{1}$, alors : $\langle \boldsymbol{u}, \boldsymbol{v}\rangle = \boldsymbol{cos}(\boldsymbol{u}, \boldsymbol{v})$.
  - D'une manière plus générale, le produit scalaire se définit relativement à une matrice comme suit :

    $$\langle \boldsymbol{u}, \boldsymbol{v}\rangle_{\boldsymbol{M}} = \boldsymbol{u}^{\boldsymbol{t}} \cdot \boldsymbol{M} \cdot \boldsymbol{v}$$

- ***Dérivation matricielle***

  - ***Dérivée d'une matrice par rapport à une variable***

    Nous faisons recours à ce type de dérivation quand nous avons des problèmes d'optimisation à résoudre.

    Soit A une matrice à n lignes et p colonnes dépendant d'un paramètre t. C'est-à-dire : $A(t) = (a_{ij}(t))_{i,j}$ pour i allant de 1 à n et $j = 1, 2, \ldots, p$. Alors, la dérivée de A par rapport à la variable t est définie pour tout i, j par :

Analyse des données

$$\frac{dA(t)}{dt} = \frac{d(a_{ij}(t))}{dt}.$$

En posant pour tout i, j :

$$\frac{d(a_{ij}(t))}{dt} = b_{ij}(t).$$

Alors,

$$\frac{dA(t)}{dt} = B(t).$$

Exemple

Considérons la matrice carrée A d'ordre 2 définie par :

$$A(t) = \begin{pmatrix} cos\,(t) & -sin\,(t) \\ sin\,(t) & cos\,(t) \end{pmatrix}.$$

Alors : la dérivée de A par rapport à la variable t est donnée par :

$$\frac{dA(t)}{dt} = \begin{pmatrix} \frac{d(cos(t))}{dt} & \frac{d(-sin(t))}{dt} \\ \frac{d(sin(t))}{dt} & \frac{d(cos(t))}{dt} \end{pmatrix}.$$

Ce qui donne :

$$\frac{dA}{dt} = \begin{pmatrix} sin\,(t) & cos\,(t) \\ -cos\,(t) & sin\,(t) \end{pmatrix} = B(t).$$

- ***Dérivée d'une fonction par rapport à une matrice***

Soit f une fonction des éléments de la matrice $A = (a_{ij})_{i,j}$ pour i allant de 1 à n et $j = 1, 2, \ldots, p$. Alors, pour tout $i = 1, 2, \ldots, n$ et pour tout $j = 1, 2, \ldots, p$, nous avons :

$$\frac{\partial f(A)}{\partial A} = \frac{\partial f(A)}{\partial(a_{ij})}.$$

Exemples

1) Soit A une matrice carrée d'ordre n : $A = (a_{ij})_{1 \leq i,j \leq n}$. Alors : la trace de la matrice A définie par :

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

Nous définit une fonction matricielle comme suit :

$$f(A) = trace(A).$$

Par suite, pour tout $i, j = 1,2, \ldots, n$ nous avons :

$$\frac{\partial f(A)}{\partial A} = \frac{\partial f(A)}{\partial(a_{ij})} = \frac{\partial(\sum_{i=1}^{n} a_{ii})}{\partial(a_{ij})}.$$

Or :

$$\sum_{i=1}^{n} a_{ii} = a_{11} + a_{22} + \cdots + a_{nn}.$$

Donc, pour tout $i, j = 1, 2, \ldots, n$, nous avons :

$$\frac{\boldsymbol{\partial}\ (\sum_{i=1}^{n} \boldsymbol{a_{ii}})}{\boldsymbol{\partial}\ (\boldsymbol{a_{ij}})} = \begin{cases} \mathbf{1} & \boldsymbol{Si: i = j} \\ \mathbf{0} & \boldsymbol{Si: i \neq j} \end{cases}$$

Ce qui donne :

$$\frac{\boldsymbol{\partial f}(A)}{\boldsymbol{\partial A}} = \begin{pmatrix} \mathbf{1} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & \mathbf{1} \end{pmatrix} = \boldsymbol{Id}.$$

2) Soient X et Y deux vecteurs de $\Re^n$, et f une fonction réelle définie par :

$$f(X, Y) = \langle X, Y \rangle = X^t \cdot Y.$$

C'est-à-dire :

$$f(X, Y) = \sum_{i=1}^{n} x_i y_i.$$

Par suite, la ***dérivée*** de f par rapport au **vecteur $\boldsymbol{X}$** est le **vecteur colonne** constitué des **dérivées partielles** $\frac{\partial f(X,Y)}{\partial x_i}$ pour tout $i = 1, 2, \ldots, n$, défini comme suit :

$$\frac{\partial f(X, Y)}{\partial X} = \left(\frac{\partial f(X, Y)}{\partial x_1}, \ldots, \frac{\partial f(X, Y)}{\partial x_i} \ldots, \frac{\partial f(X, Y)}{\partial x_n}\right)^t.$$

C'est-à-dire :

$$\frac{\boldsymbol{\partial f}(X, Y)}{\boldsymbol{\partial X}} = \begin{pmatrix} \frac{\boldsymbol{\partial}\ (\sum_{i=1}^{n} \boldsymbol{x_i y_i})}{\boldsymbol{\partial\ x_1}} \\ \frac{\boldsymbol{\partial}\ (\sum_{i=1}^{n} \boldsymbol{x_i y_i})}{\boldsymbol{\partial\ x_2}} \\ \vdots \\ \frac{\boldsymbol{\partial}\ (\sum_{i=1}^{n} \boldsymbol{x_i y_i})}{\boldsymbol{\partial\ x_n}} \end{pmatrix} = \begin{pmatrix} \frac{\boldsymbol{\partial}\ (\boldsymbol{x_1 y_1} + \boldsymbol{x_2 y_2} + \ldots + \boldsymbol{x_n y_n})}{\boldsymbol{\partial\ x_1}} \\ \frac{\boldsymbol{\partial}\ (\boldsymbol{x_1 y_1} + \boldsymbol{x_2 y_2} + \ldots + \boldsymbol{x_n y_n})}{\boldsymbol{\partial\ x_2}} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \frac{\boldsymbol{\partial}\ (\boldsymbol{x_1 y_1} + \boldsymbol{x_2 y_2} + \ldots + \boldsymbol{x_n y_n})}{\boldsymbol{\partial\ x_n}} \end{pmatrix}.$$

Ce qui donne :

Analyse des données

$$\boldsymbol{\frac{\partial f(X,Y)}{\partial X} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = Y.}$$

De même, nous pouvons définir la dérivée de f par rapport au vecteur Y :

$$\boldsymbol{\frac{\partial f(X,Y)}{\partial Y} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = X.}$$

3) Soit f une fonction vectorielle définie pour tout X et Y de $\Re^n$ par :

$$f(X) = Y.$$

Alors, la dérivée de f par rapport au vecteur X est défini par :

$$\frac{\partial f(X)}{\partial X} = \frac{\partial Y}{\partial X} = \left(\frac{\partial y_1}{\partial X}, \dots, \frac{\partial y_\mathrm{i}}{\partial X} \dots, \frac{\partial y_\mathrm{n}}{\partial X}\right)^t.$$

Or pour tout i fixé, nous avons :

$$\frac{\partial y_\mathrm{i}}{\partial X} = \left(\frac{\partial y_\mathrm{i}}{\partial x_1}, \dots, \frac{\partial y_\mathrm{i}}{\partial x_\mathrm{i}} \dots, \frac{\partial y_\mathrm{n}}{\partial x_\mathrm{n}}\right).$$

Par conséquent, la dérivée de f par rapport au vecteur X est donnée par :

$$\frac{\partial f(X)}{\partial X} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \dots & \frac{\partial y_1}{\partial x_j} & \dots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \dots & \vdots & \dots & \vdots \\ \frac{\partial y_i}{\partial x_1} & \vdots & \frac{\partial y_i}{\partial x_j} & \dots & \frac{\partial y_i}{\partial x_n} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \dots & \frac{\partial y_n}{\partial x_j} & \dots & \frac{\partial y_n}{\partial x_n} \end{pmatrix}.$$

Exercice

Montrer que :

$$\frac{\partial(X^t \cdot M \cdot Y)}{\partial M} = XY^t.$$

$$\frac{\partial(X^t \cdot M \cdot X)}{\partial X} = MX + M^t.X = (M + M^t)X.$$

Si : $\boldsymbol{M}$ est symétrique, alors :

$$\frac{\partial(X^t \cdot M \cdot X)}{\partial X} = 2MX.$$

5.2 Mesures de liaison

Deux types de mesures :

Pour les individus

La proximité entre les individus se mesure moyennant la ***distance Euclidienne.***

$$d(X_i, X_j) = \left(\sum_{k=1}^{n} (x_i^k - x_j^k)^2 \right)^{1/2}.$$

Pour les variables

Il s'agit de mesurer la dispersion du nuage de points en jugeant de la liaison entre les variables par les mesures suivantes :

- ***Covariance***

  Elle permet de préciser le sens de la liaison entre les variables, elle est définie par :

  $$\boldsymbol{Cov(X,Y) = \frac{1}{m} \cdot \sum_{j=1}^{m} (x_j - \overline{X}) \cdot (y_j - \overline{Y})}.$$

  Ou, X et Y deux variables de $\Re^m$.

Remarque

La covariance de deux variables X et Y n'est que le produit scalaire des deux variables centrées relativement à la matrice des poids.

$$\left\langle \tilde{X}, \tilde{Y} \right\rangle_M$$

$$Cov\ (X, Y) = \langle \tilde{X}, \tilde{Y} \rangle_M.$$

Ou, M est la matrice des poids.

- ***Coefficient de corrélation***

  Ce n'est qu'une normalisation de la covariance. Il permet de mesurer l'intensité de la liaison linéaire entre les variables, il est défini par :

  $$\boldsymbol{\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)} \cdot \sqrt{Var(Y)}}}.$$

Analyse des données

Remarque
Le coefficient de corrélation prend ses valeurs entre -1 et 1. Plus précisément, le coefficient de corrélation mesure le cosinus de l'angle inscrit entre les variables.

En effet,

$$\frac{Cov(X,Y)}{\sqrt{Var(X)}\cdot\sqrt{Var(Y)}} = \frac{\langle\tilde{X},\tilde{Y}\rangle_M}{\|\tilde{X}\|_M\cdot\|\tilde{Y}\|_M} = Cos(\theta_{\tilde{X},\tilde{Y}}).$$

Définition

Deux variables sont dites ***décorrélées*** si leur coefficient est nul, et ***linéairement liées*** si leur coefficient de corrélation est de ***module égal à 1***.

Exemple

Soit la matrice de données.

$$X = \begin{pmatrix} 1 & 6 \\ 3 & 2 \end{pmatrix}.$$

Alors : les variables sont :

$$\boldsymbol{X^1 = (1,3)^t \quad \text{et} \quad X^2 = (6,2)^t.}$$

Alors :

- Le ***centre de gravité*** du nuage est donné par :

$$g = (\bar{X}^1, \bar{X}^2) = \left(\frac{1+3}{2}, \frac{6+2}{2}\right) = (2,4).$$

La matrice ***centrée*** est la matrice :

$$\boldsymbol{Y = \begin{pmatrix} 1-2 & 6-4 \\ 3-2 & 2-4 \end{pmatrix}, Y = \begin{pmatrix} -1 & 2 \\ 1 & -2 \end{pmatrix}.}$$

- Les ***variances*** sont données par :

$$Var(X^1) = \frac{1}{2}\cdot\sum_{j=1}^{2}\left(x_j^k - \overline{X^1}\right)^2 = (1/2) * (1 + 1) = 1.$$

$$Var(X^2) = \frac{1}{2}\cdot\sum_{j=1}^{2}\left(x_j^k - \overline{X^2}\right)^2 = (1/2) * (4 + 4) = 4.$$

D'où,

$$\boldsymbol{Cov(X^1,\ X^2) = \frac{1}{2}\times(-2-2) = -2 < 0.}$$

(Produit scalaire relativement à la matrice des poids)
Sens de la liaison entre les deux variables

Donc,

Analyse des données

Les deux variables en question sont de sens contraire.

D'autre part, nous avons :

$$\boldsymbol{\rho}_{X^1,X^2} = \frac{Cov(\boldsymbol{X^1},\boldsymbol{X^2})}{\boldsymbol{\sigma}_{X^1}\cdot\boldsymbol{\sigma}_{X^2}} = \frac{Cov(\boldsymbol{X},\boldsymbol{Y})}{\sqrt{Var(\boldsymbol{X})}\cdot\sqrt{Var(\boldsymbol{Y})}} = \frac{\mathbf{-2}}{\mathbf{1*2}} = \mathbf{-1}.$$

Donc,

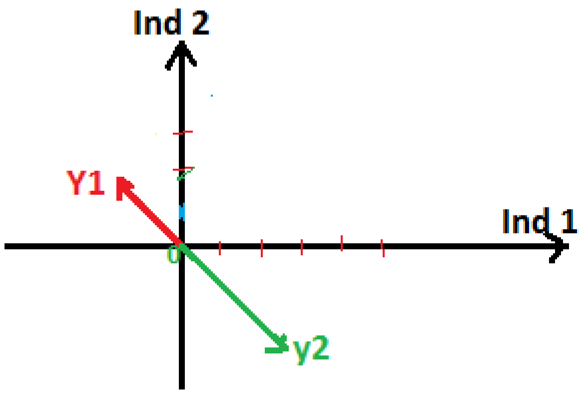
L'angle entre les deux variables est égal à pi.

C'est-à-dire :

Les deux variables sont colinéaires mais de sens contraire

Remarque

Ces résultats peuvent être interprétés graphiquement :



Presentation des variables dans l'espace des individus.