

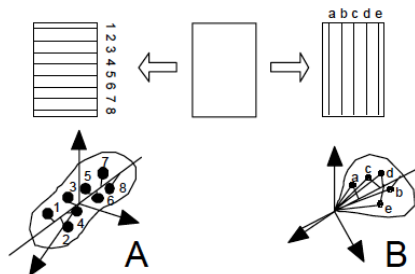
Analyse en Composantes Principales

Anne B Dufour

Octobre 2011

Introduction

Soit X un tableau contenant p variables mesurées sur n individus.



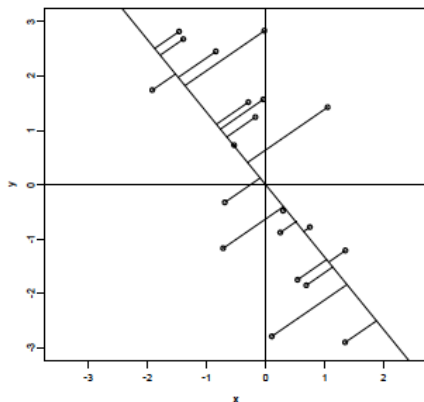
- Situation A : n points-lignes dans \mathbb{R}^p
- Situation B : p points-colonnes dans \mathbb{R}^n

Objectif :

Projection d'un nuage de points sur des axes qui maximisent l'inertie projetée

Énoncé (1)

Deux variables x et y sont mesurées sur n individus.



L'étude de la liaison entre x et y est la recherche d'une droite optimum.
Quel est le critère d'optimisation ?

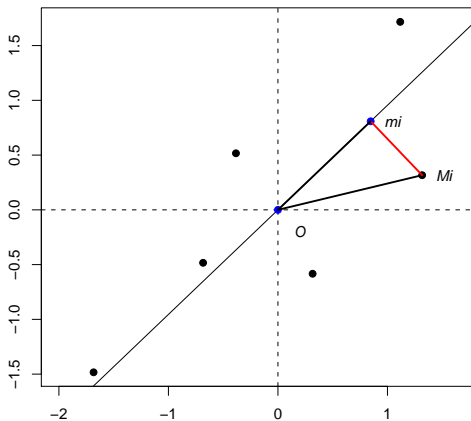
Énoncé (2)

Cette droite minimise

$$\frac{1}{n} \sum_{i=1}^n M_i m_i^2$$

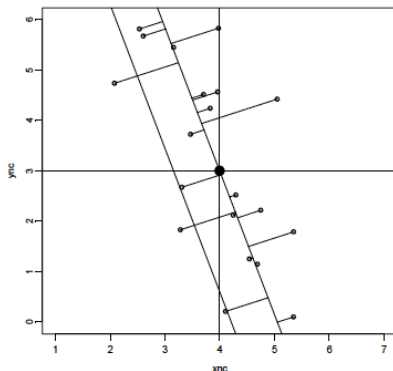
où

- M_i est le point i du plan,
- m_i est la projection orthogonale de M_i sur la droite.



Elle est dite **direction principale** du nuage centré.

Nuage centré ?



La droite qui minimise la moyenne des carrés des distances des points à cette droite passe par le point moyen de coordonnées $(m(\mathbf{x}), m(\mathbf{y}))$ (conséquence du théorème de Huygens).

Théorème de Huygens

Le moment d'inertie d'un nuage de points par rapport à un axe de rotation D' est égale à la somme du moment d'inertie de ce nuage de points par rapport à l'axe de rotation parallèle D passant par le centre de gravité et du moment d'inertie du centre de gravité par rapport à D' .

$$\frac{1}{n} \sum_{i=1}^n \| M_i - m'_i \|^2 = \frac{1}{n} \sum_{i=1}^n \| M_i - m_i \|^2 + \| w \|^2$$

Inertie Totale

- 1 \mathbf{X}_0 est une matrice à n lignes et p colonnes : $\mathbf{X}_0 = [x_{ij} - m(\mathbf{x}^j)]$ où \mathbf{x}^j est le vecteur associé à la variable j .
- 2 Chaque point a un poids $\frac{1}{n}$ (pondération uniforme).

L'ensemble des n points forme un nuage dont l'**inertie** autour du centre de gravité vaut :

$$I_T = \frac{1}{n} \sum_{i=1}^n \| M_i \|^2 = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^p (x_{ij} - m(\mathbf{x}^j))^2$$

C'est la variabilité totale de la position des points dans l'espace.

Exemple en dimension 2

La droite passe par le point moyen. Donc on place l'origine au centre de gravité en utilisant les nouvelles coordonnées centrées :

$$\begin{cases} X_i = x_i - m(\mathbf{x}) \\ Y_i = y_i - m(\mathbf{y}) \end{cases}$$

$$\begin{aligned} I_T &= \frac{1}{n} \sum_{i=1}^n (X_i^2 + Y_i^2) \\ &= \frac{1}{n} \sum_{i=1}^n \left((x_i - m(\mathbf{x}))^2 + (y_i - m(\mathbf{y}))^2 \right) \\ &= \frac{1}{n} \sum_{i=1}^n (x_i - m(\mathbf{x}))^2 + \frac{1}{n} \sum_{i=1}^n (y_i - m(\mathbf{y}))^2 \\ &= v(\mathbf{x}) + v(\mathbf{y}) \end{aligned}$$

Décomposition de l'inertie totale

Quand on prend dans \mathbb{R}^p un vecteur unitaire \mathbf{u} , il définit un axe. Le point M_i se projette sur cet axe en m_i . On a :

$$M_i = m_i + (M_i - m_i) \text{ et } \| M_i \|^2 = \| m_i \|^2 + \| M_i - m_i \|^2$$

$$I_T = I_S(\mathbf{u}) + I_M(\mathbf{u}) = \frac{1}{n} \sum_{i=1}^n \| m_i \|^2 + \frac{1}{n} \sum_{i=1}^n \| M_i - m_i \|^2$$

$$I_T = \underbrace{I_S(\mathbf{u})}_{\text{à maximiser}} + \underbrace{I_M(\mathbf{u})}_{\text{à minimiser}}$$

Recherche du vecteur directeur \mathbf{u}

La matrice \mathbf{X}_0 contient les n points centrés.

$$\mathbf{X}_0 = \begin{bmatrix} x_1 - m(\mathbf{x}) & y_1 - m(\mathbf{y}) \\ \vdots & \vdots \\ x_n - m(\mathbf{x}) & y_n - m(\mathbf{y}) \end{bmatrix}$$

Le vecteur \mathbf{u} recherché est unitaire. On l'écrit sous la forme $\mathbf{u} = \begin{bmatrix} a \\ b \end{bmatrix}$
avec $a^2 + b^2 = 1$.

Recherche du vecteur directeur \mathbf{u}

L'inertie statistique ou inertie projetée est :

$$\begin{aligned} I_S(\mathbf{u}) &= \frac{1}{n} \sum_{i=1}^n \|m_i\|^2 = \frac{1}{n} (\mathbf{X}_0 \mathbf{u})^T \mathbf{X}_0 \mathbf{u} = \mathbf{u}^T \left(\frac{1}{n} \mathbf{X}_0^T \mathbf{X}_0 \right) \mathbf{u} \\ &= [a \quad b] \begin{bmatrix} v(\mathbf{x}) & c(\mathbf{x}, \mathbf{y}) \\ c(\mathbf{x}, \mathbf{y}) & v(\mathbf{y}) \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} \\ &= v(\mathbf{x})a^2 + 2c(\mathbf{x}, \mathbf{y})ab + v(\mathbf{y})b^2 \end{aligned}$$

Recherche du vecteur directeur u

La matrice $\begin{bmatrix} v(\mathbf{x}) & c(\mathbf{x}, \mathbf{y}) \\ c(\mathbf{x}, \mathbf{y}) & v(\mathbf{y}) \end{bmatrix}$ est dite **matrice de variances-covariances** des deux variables. On la note

$$\mathbf{C} = \frac{1}{n} \mathbf{X}_0^T \mathbf{X}_0$$

Elle est symétrique. Son polynôme caractéristique s'écrit :

$$|\mathbf{C} - \lambda \mathbf{I}_2| = \begin{vmatrix} v(\mathbf{x}) - \lambda & c(\mathbf{x}, \mathbf{y}) \\ c(\mathbf{x}, \mathbf{y}) & v(\mathbf{y}) - \lambda \end{vmatrix} \quad (1)$$

$$= \lambda^2 - \lambda(v(\mathbf{x}) + v(\mathbf{y})) + v(\mathbf{x})v(\mathbf{y}) - c^2(\mathbf{x}, \mathbf{y}) \quad (2)$$

Recherche du vecteur directeur u

- Le polynôme caractéristique a toujours deux racines donc C a toujours deux valeurs propres et deux vecteurs propres.
- Les valeurs propres sont en général distinctes. On les note λ_1 et λ_2 .
 - ▶ $\lambda_1 + \lambda_2 = v(\mathbf{x}) + v(\mathbf{y})$
 - ▶ $\lambda_1 \times \lambda_2 = v(\mathbf{x}) \times v(\mathbf{y}) - c^2(\mathbf{x}, \mathbf{y})$

Toute matrice symétrique admet une base de vecteurs propres orthogonaux. Donc,

$$C = U\Lambda U^T$$

$$\text{avec } U = \begin{bmatrix} u_{11} & u_{21} \\ u_{12} & u_{22} \end{bmatrix} \text{ et } \Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

Recherche du vecteur directeur \mathbf{u}

$$\begin{aligned}
 I_S(\mathbf{u}) &= \mathbf{u}^T \mathbf{C} \mathbf{u} \\
 &= [a \quad b] \begin{bmatrix} u_{11} & u_{21} \\ u_{12} & u_{22} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} \\
 &= [\alpha \quad \beta] \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \\
 &= \lambda_1 \alpha^2 + \lambda_2 \beta^2 \leq \lambda_1 \alpha^2 + \lambda_1 \beta^2 = \lambda_1
 \end{aligned}$$

$\begin{bmatrix} \alpha \\ \beta \end{bmatrix}$ représente les coordonnées du vecteur \mathbf{u} dans la base des vecteurs propres.

Recherche du vecteur directeur u

L'inertie ne peut dépasser la première valeur propre et l'atteint pour $\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, donc le premier vecteur propre.

Conclusion dans le cas de 2 variables :

L'axe principal d'un nuage bivarié est le premier vecteur propre de la matrice de variance-covariance des deux variables.

Généralisation à p variables

$\mathbf{C} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ admet une base de p vecteurs propres orthonormés.

- Le premier vecteur propre normé \mathbf{u}_1 est un vecteur de \mathbb{R}^p qui maximise l'inertie projetée.
- Le deuxième vecteur propre normé \mathbf{u}_2 est un vecteur de \mathbb{R}^p , orthogonal à \mathbf{u}_1 , qui maximise à nouveau l'inertie projetée.
- et ainsi de suite pour $\mathbf{u}_3 \dots \mathbf{u}_r$ axes suivants où r représente le rang de la matrice diagonalisée.

Coordonnées des projections

Si \mathbf{u}_k est le vecteur propre de rang k , les coordonnées des projections des n points sont obtenus simplement par :

$$\mathbf{l}_k = \begin{bmatrix} l_{1k} \\ \vdots \\ l_{nk} \end{bmatrix} = \begin{bmatrix} \langle M_1 | \mathbf{u}_k \rangle \\ \vdots \\ \langle M_n | \mathbf{u}_k \rangle \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^p (x_{1j} - m(\mathbf{x}^j)) u_{jk} \\ \vdots \\ \sum_{j=1}^p (x_{nj} - m(\mathbf{x}^j)) u_{jk} \end{bmatrix}$$

soit en écriture matricielle : $\mathbf{l}_k = \mathbf{X}_0 \mathbf{u}_k$.

Axes principaux et coordonnées

- \mathbf{u}_k est appelé **axe principal** de rang k .
- \mathbf{l}_k est appelé vecteur des **coordonnées** sur l'axe principal. C'est une variable artificielle de moyenne nulle et de variance λ_k .

$$\begin{aligned}
 m(\mathbf{l}_k) &= \frac{1}{n} \sum_{i=1}^n l_{ik} = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^p (x_{ij} - m(\mathbf{x}^j)) u_{jk} \\
 &= \sum_{j=1}^p u_{jk} \frac{1}{n} \sum_{i=1}^n (x_{ij} - m(\mathbf{x}^j)) = 0
 \end{aligned}$$

$$\begin{aligned}
 v(\mathbf{l}_k) &= \frac{1}{n} \sum_{i=1}^n l_{ik}^2 = \frac{1}{n} (\mathbf{X}_0 \mathbf{u}_k)^T \mathbf{X}_0 \mathbf{u}_k = \mathbf{u}_k^T \mathbf{C} \mathbf{u}_k \\
 &= \lambda_k \mathbf{u}_k^T \mathbf{u}_k = \lambda_k
 \end{aligned}$$

Le graphe des valeurs propres

- La coordonnées sur l'axe de rang k est donc centrée de variance λ_k .
- La somme des valeurs propres est l'inertie totale.

$$\sum_{j=1}^p \lambda_j = \sum_{j=1}^p v(\mathbf{x}^j)$$

où \mathbf{x}^j est la variable j du tableau \mathbf{X} .

Le graphe des valeurs propres exprime la manière dont la variabilité des données se répartit dans l'espace.

C'est une représentation en bâtons avec k sur l'axe horizontal et λ_k sur l'axe vertical . En anglais, on parle de screeplot.

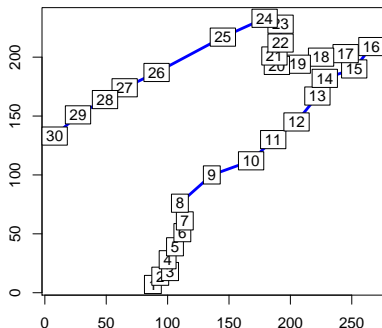
La carte factorielle

La représentation du nuage projeté sur un couple d'axes principaux est appelée **carte factorielle**. C'est une manière de voir l'information multidimensionnelle.

La carte factorielle des axes 1 et 2 est dite premier plan factoriel et représente la part maximale de la variabilité. Chaque point i est positionné par ses deux coordonnées (l_{i1}, l_{i2}) .

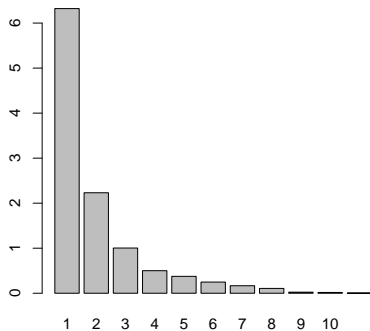
Exemple écologique

Stations le long du Doubs

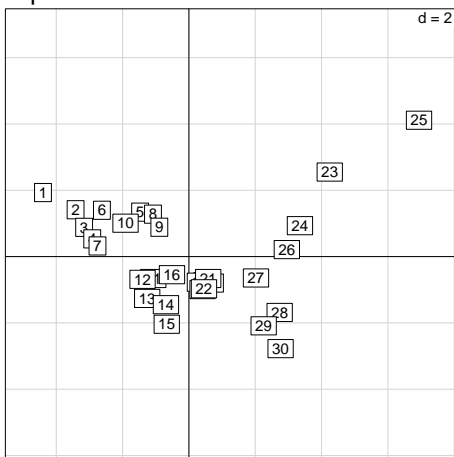


11 variables physico-chimiques :
 distance à la source, altitude, pente
 débit, pH, dureté de l'eau
 phosphate, nitrate, ammoniacque
 oxygène, demande biologique en
 oxygène

Graphe des valeurs propres

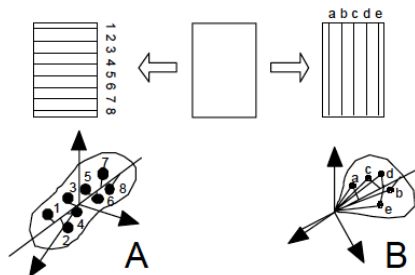


Représentation des individus - stations



Retour à l'énoncé du problème

On vient d'étudier la projection d'un nuage de n points sur des axes qui maximisent l'inertie projetée (situation A).



On s'intéresse maintenant à l'ensemble de p points dans \mathbb{R}^n (situation B).

Énoncé (1)

Le point de vue de cette analyse a été proposé par Hotelling (1933) dans le cas où les données sont centrées réduites.

Soit \mathbf{y} une variable quelconque. On peut calculer sa corrélation avec chacune des variables de départ \mathbf{x}^j ($j = 1, p$). Le lien entre \mathbf{y} et \mathbf{X} peut se mesurer par la relation :

$$\mathcal{L}(\mathbf{y}, \mathbf{X}) = \sum_{j=1}^p r^2(\mathbf{y}, \mathbf{x}^j)$$

Objectif : Trouver une variable \mathbf{y} qui optimise cette quantité.

Le problème reste inchangé si on suppose \mathbf{y} de moyenne nulle et de variance 1.

Enoncé (2)

$r^2(\mathbf{y}, \mathbf{x}^k)$ est :

- le carré de la norme du projeté de \mathbf{y} sur le vecteur \mathbf{x}_{\bullet}^k , vecteur centré réduit,
- le carré de la norme du projeté de \mathbf{x}_{\bullet}^k sur le vecteur \mathbf{y} .

Le lien est alors l'inertie projetée du nuage des variables sur \mathbf{y} en pensant que le poids de chaque variable est 1 et que le produit scalaire de \mathbb{R}^n est $\mathbf{D} = \frac{1}{n} \mathbf{I}_n$.

C'est donc le même problème que précédemment.

Données centrées réduites

Reprenant les résultats de la représentation des individus lorsque les variables du tableau \mathbf{X} sont centrées réduites. Il est alors noté \mathbf{X}_{\bullet} .
On cherche les axes principaux, vecteurs propres de la matrice $\frac{1}{n}\mathbf{X}_{\bullet}^T\mathbf{X}_{\bullet}$.
On est passé de la matrice de variance-covariance \mathbf{C} à la matrice des corrélations \mathbf{R} .

$$\mathbf{R} = \frac{1}{n}\mathbf{X}_{\bullet}^T\mathbf{X}_{\bullet} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \text{ et } \mathbf{L} = \mathbf{X}_{\bullet}\mathbf{U}$$

Recherche du vecteur y

$$\mathcal{L}(\mathbf{y}, \mathbf{X}) = \sum_{k=1}^p r^2(\mathbf{y}, \mathbf{x}_{\bullet}^k) = 1 \langle \mathbf{y} | \mathbf{x}_{\bullet}^k \rangle^2$$

$$\mathcal{L}(\mathbf{y}, \mathbf{X}) = \left(\frac{1}{n} \mathbf{X}_{\bullet}^T \mathbf{y} \right)^T \left(\frac{1}{n} \mathbf{X}_{\bullet}^T \mathbf{y} \right) = \frac{1}{n^2} \mathbf{y}^T \mathbf{X}_{\bullet} \mathbf{X}_{\bullet}^T \mathbf{y}$$

On cherche \mathbf{y} normé pour le produit scalaire \mathbf{D} sous la contrainte

$$\frac{1}{n} \sum_{i=1}^n y_i^2 = 1.$$

Recherche du vecteur y

On note que $\frac{1}{n} \sum_{i=1}^n y_i^2$ peut s'écrire $= \sum_{i=1}^n \left(\frac{1}{\sqrt{n}} y_i \right)^2$.

On pose donc $\mathbf{z} = \frac{1}{\sqrt{n}} \mathbf{y}$ et on cherche \mathbf{z} normé pour le produit scalaire ordinaire maximisant :

$$\frac{1}{n} \mathbf{z}^T \mathbf{X} \bullet \mathbf{X} \bullet \mathbf{z}$$

On pose $\mathbf{S} = \frac{1}{n} \mathbf{X} \bullet \mathbf{X}^T$. C'est une matrice symétrique qui admet une base de vecteurs propres orthogonaux :

$$\mathbf{S} = \mathbf{V} \Lambda \mathbf{V}^T$$

Vecteurs propres de S

$$\mathbf{S}\mathbf{L} = \frac{1}{n} \mathbf{X} \bullet \mathbf{X} \bullet^T \mathbf{X} \bullet \mathbf{U} = \mathbf{X} \bullet \mathbf{R}\mathbf{U} = \mathbf{X} \bullet \mathbf{U}\mathbf{\Lambda} = \mathbf{L}\mathbf{\Lambda}$$

- Les matrices \mathbf{S} et \mathbf{R} ont les mêmes valeurs propres non nulles.
- Les vecteurs que l'on cherche sont connus : à une constante de normalisation près, les coordonnées des individus sur les axes principaux.

$$\mathbf{V} = \mathbf{L}\mathbf{\Lambda}^{-1/2}$$

Solution et représentation graphique

$\mathcal{L}(\mathbf{y}, \mathbf{X})$ ne peut pas dépasser λ_1 et l'atteint pour la variable normée $\frac{1}{\sqrt{\lambda_1}}\mathbf{I}_1$.

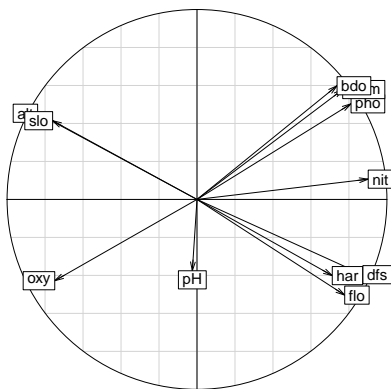
Ce vecteur est appelé **première composante principale**. C'est la variable qui est la plus corrélée avec toutes les variables du tableau.

Quand on projette les variables sur les composantes principales, on obtient les coordonnées des variables.

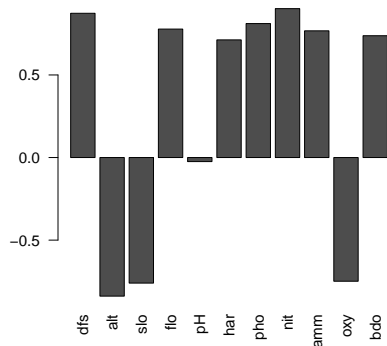
$$\mathbf{K} = \frac{1}{n} \mathbf{X}_\bullet^T \mathbf{L} \Lambda^{-1/2} = \frac{1}{n} \mathbf{X}_\bullet^T \mathbf{X}_\bullet \mathbf{U} \Lambda^{-1/2} = \mathbf{R} \mathbf{U} \Lambda^{-1/2} = \mathbf{U} \Lambda^{1/2}$$

La représentation liant variables et composantes principales est appelée **cercle des corrélations**.

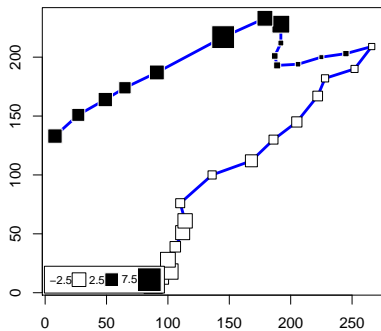
Représentation des variables physico-chimiques



Exemple écologique (5)



Pollution de l'eau le long du Doubs



Conclusion

L'Analyse en Composantes Principales dite A.C.P. est l'étude du triplet $(\mathbf{X}, \mathbf{I}_p, \frac{1}{n}\mathbf{I}_n)$.

- Si les données sont centrées, on parle d' **ACP centrée**.
- Si les données sont normées, on parle d' **ACP normée**.

Pour des compléments d'informations, consulter les fiches de cours : BS4, BS5.