Classifications nos supervisées

UFR Environnement - Departement PCMI

PLAN

- CHI- Généralités
- CHII – Analyse en Composante Principale (ACP)
- K-means clustering
- Clustering hiérarchique

CHII- Analyse en Composante Principale (ACP)

CHII- Analyse en Composante Principale (ACP)

Introduction

L'Analyse en Composantes Principales ou ACP (**Principal Component Analysis** ou PCA en anglais) est une méthode qui sera utilisable lorsque :

•Des relations linéaires sont suspectées entre les variables (si elles ne sont pas linéaires, penser à transformer les données auparavant pour les linéariser).

•Ces relations conduisent à une répartition des individus (le nuage de points) qui forment une structure que l'on cherchera à interpréter.

•Pour visualiser cette structure, les données sont simplifiées (réduites) de N variables à n ($n < N$ et $n = 2$ ou 3 généralement). La représentation sous forme d'un nuage de points s'appelle une carte.

•La réduction des dimensions se fait avec une perte minimale d'information au sens de la variance des données.

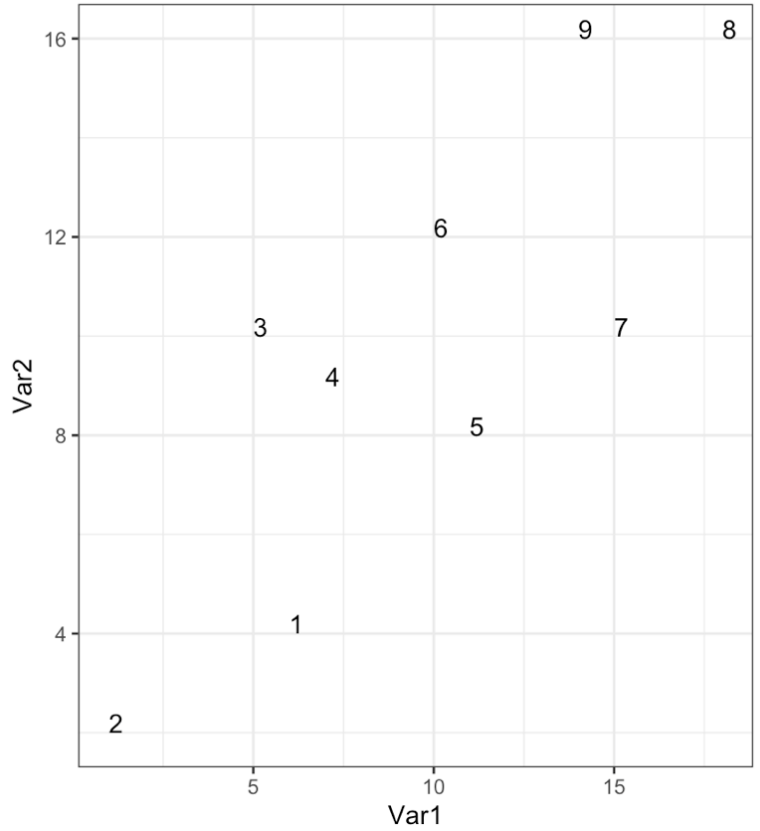
CHII- Analyse en Composante Principale (ACP)

I - ACP : mécanisme

Nous allons partir d'un exemple presque trivial pour illustrer le principe de l'ACP. Comment réduire un tableau bivarié en une représentation des individus en une seule dimension (classement sur une droite) avec perte minimale d'information ?

Station	Var1	Var2
1	6	4
2	1	2
3	5	10
4	7	9
5	11	8
6	10	12
7	15	10
8	18	16
9	14	16

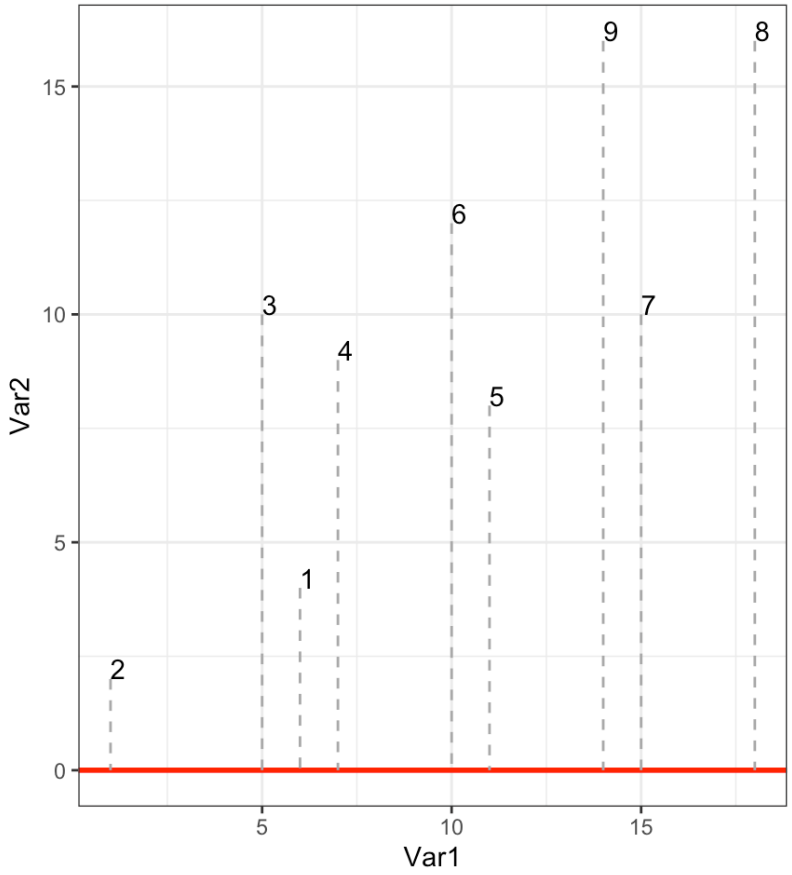
Unereprésentation graphique 2D de ces données



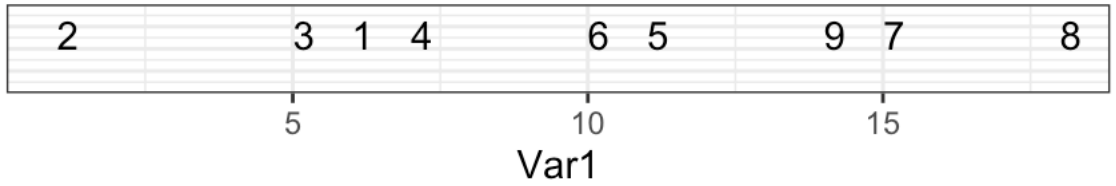
CHII- Analyse en Composante Principale (ACP)

I - ACP : mécanisme

Si nous réduisons à une seule dimension en laissant tomber une des deux variables, voici ce que cela donne (ici on ne garde que Var1, donc, on projette les points sur l'axe des abscisses).



Au final, nous avons ordonné nos individus en une dimension comme suit :

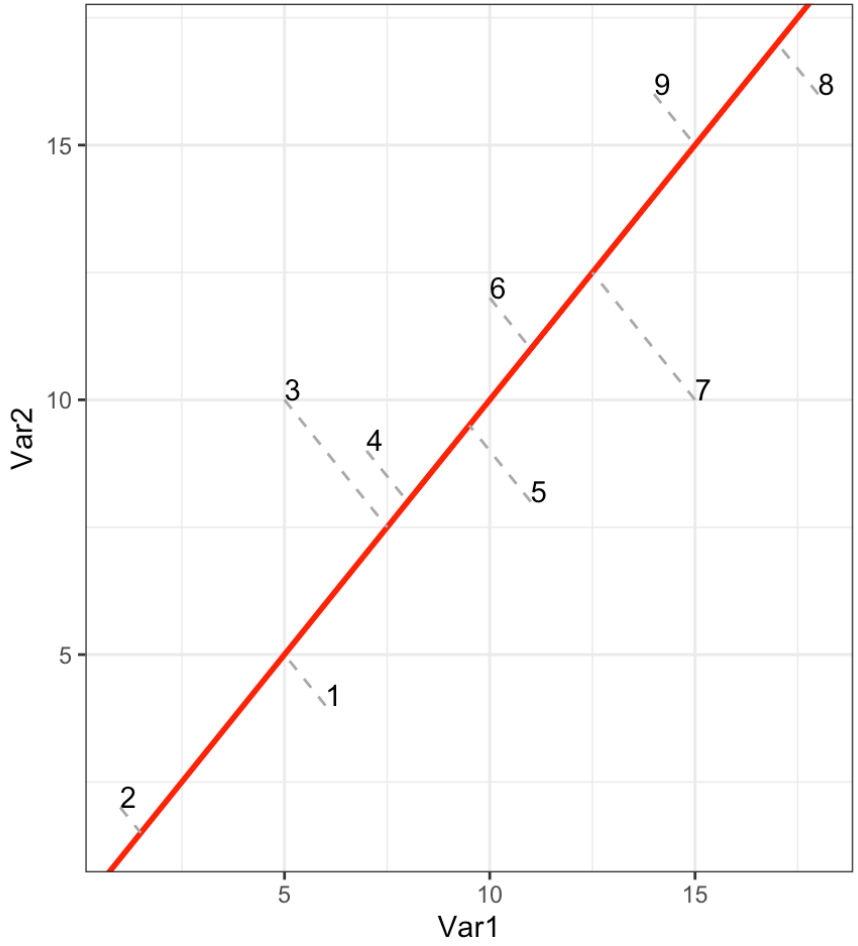


il y a trop de perte d'information. Regardez l'écart entre 7 et 9 sur le graphique en deux dimensions et dans celui à une dimension : les points sont trop près.

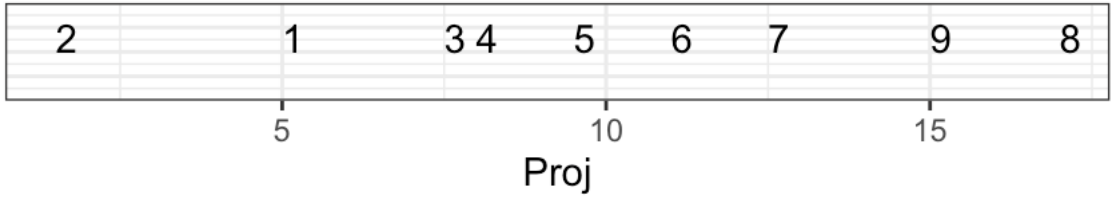
CHII- Analyse en Composante Principale (ACP)

I - ACP : mécanisme

Une autre solution serait de projeter le long de la droite de “tendance générale”, c’est-à-dire le long de l’axe de plus grand allongement du nuage de points.



Cela donne ceci en une seule dimension



C’est une bien meilleure solution, car la perte d’information est ici minimale.

CHII- Analyse en Composante Principale (ACP)

I - ACP : mécanisme

L'ACP effectue précisément la projection que nous venons d'imaginer.

- La droite de projection est appelée **composante principale 1**. La composante principale 1 présente la plus grande variabilité possible sur un seul axe.
- Ensuite on calcule la composante 2 comme étant orthogonale (c.-à-d., perpendiculaire) à PC1 et présentant la plus grande variabilité non encore capturée par la composante 1.
- Le mécanisme revient à projeter les points sur des axes orientés différemment dans l'espace à N dimensions (pour N variables initiales).

En effet, mathématiquement, ce mécanisme se généralise facilement à trois, puis à N dimensions.

CHII- Analyse en Composante Principale (ACP)

II- ACP : Calcul matriciel

La rotation optimale des axes vers les PC1 à PCN se résout par un calcul matriciel. Nous allons maintenant le détailler. Mais auparavant, nous devons nous rafraîchir l'esprit concernant quelques notions.

- Vecteurs propres et valeurs propres (il en existe autant qu'il y a de colonnes dans la matrice de départ) :

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 12 \\ 8 \end{pmatrix} = 4 \times \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

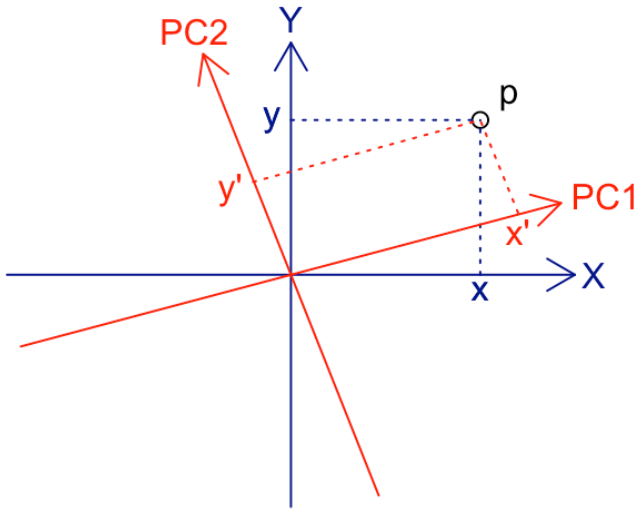
- La constante (4) est une **valeur propre** et la matrice multipliée (à droite) est la matrice des **vecteurs propres**
- La rotation d'un système d'axes à deux dimensions d'un angle α peut se représenter sous forme d'un calcul matriciel :

$$\begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix} \times \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x' \\ y' \end{pmatrix}$$

CHII- Analyse en Composante Principale (ACP)

II- ACP : Calcul matriciel

Dans le cas particulier de l'ACP, la matrice de transformation qui effectue la rotation voulue pour obtenir les axes principaux est **la matrice rassemblant tous les vecteurs propres calculés après diagonalisation de la matrice de corrélation ou de variance/covariance** (réduction ou non, respectivement). Le schéma suivant visualise la rotation depuis les axes initiaux X et Y (variables de départ) en bleu royal vers les PC1, PC2 en rouge. Un individu p est représenté par les coordonnées {x, y} dans le système d'axes initial XY. Les nouvelles coordonnées {x', y'} sont recalculées par projection sur les nouveaux axes PC1-PC2. Les flèches bleues sont représentées dans l'espace des variables, tandis que les points reprojetés sur PC1-PC2 sont représentés dans l'espace des individus selon les coordonnées primes en rouge.



CHII- Analyse en Composante Principale (ACP)

II- ACP : Calcul matriciel (Résolution numérique simple)

Effectuons une ACP sur matrice var/covar sans réduction des données (mais calcul très similaire lorsque les données sont réduites) sur un exemple numérique simple.

Étape 1 : centrage des données.

$$\underbrace{\begin{pmatrix} 2 & 1 \\ 3 & 4 \\ 5 & 0 \\ 7 & 6 \\ 9 & 2 \end{pmatrix}}_{\text{Tableau brut}} \xrightarrow{\text{centrage}} \underbrace{\begin{pmatrix} -3.2 & -1.8 \\ -2.2 & 1.4 \\ -0.2 & -2.6 \\ 1.8 & 3.4 \\ 3.8 & -0.6 \end{pmatrix}}_{\text{Tableau centré (X)}}$$

Étape 2 : calcul de la matrice de variance/covariance.

$$\underbrace{\begin{pmatrix} -3.2 & -1.8 \\ -2.2 & 1.4 \\ -0.2 & -2.6 \\ 1.8 & 3.4 \\ 3.8 & -0.6 \end{pmatrix}}_{\text{Tableau centré (X)}} \xrightarrow{\text{var/covar}} \underbrace{\begin{pmatrix} 8.2 & 1.6 \\ 1.6 & 5.8 \end{pmatrix}}_{\text{Matrice carrée (A)}}$$

Étape 3 : diagonalisation de la matrice var/covar

$$\underbrace{\begin{pmatrix} 8.2 & 1.6 \\ 1.6 & 5.8 \end{pmatrix}}_{\text{Matrice carrée (A)}} \xrightarrow{\text{diagonalisation}} \underbrace{\begin{pmatrix} 9 & 0 \\ 0 & 5 \end{pmatrix}}_{\text{Matrice diagonalisée (B)}}$$

La **trace** des deux matrices A et B à : 8.2 + 5.8 = 14

8.2 est la **part de variance** exprimée sur le premier axe initial (X)

14 est la **variance totale** du jeu de données

La matrice diagonale B est la solution exprimant **la plus grande part plus grand. e part de variance possible sur le premier axe de l'ACP** : 9, soit 64,3% de la variance totale.

Les éléments sur la diagonales sont les valeurs propres « eigenvalues »

CHII- Analyse en Composante Principale (ACP)

II- ACP : Calcul matriciel (Résolution numérique simple)

Étape 4 : calcul de la matrice de rotation des axes (en utilisant la propriété des valeurs propres A.U=B.U).

$$\underbrace{\begin{pmatrix} 8.2 & 1.6 \\ 1.6 & 5.8 \end{pmatrix}}_{\text{Matrice A}} \times \mathrm{U} = \underbrace{\begin{pmatrix} 9 & 0 \\ 0 & 5 \end{pmatrix}}_{\text{Matrice B}} \times \mathrm{U} \rightarrow \mathrm{U} = \underbrace{\begin{pmatrix} 0.894 & -0.447 \\ 0.447 & 0.894 \end{pmatrix}}_{\text{Matrice des vecteur propres (U)}}$$

La **matrice des vecteurs propres (U)** (“eigenvectors” en anglais) effectue la transformation (**rotation des axes**) pour obtenir les **composantes principales**.

L’angle de rotation se déduit en considérant que cette matrice contient des sinus et cosinus d’angles de rotation des axes : $\begin{pmatrix} 0.894 & -0.447 \\ 0.447 & 0.894 \end{pmatrix} = \begin{pmatrix} \cos(-26.6^\circ) & \sin(-26.6^\circ) \\ -\sin(-26.6^\circ) & \cos(-26.6^\circ) \end{pmatrix}$

Étape 5 : représentation dans l’espace des variables. C’est une représentation dans un cercle de la matrice des vecteurs propres U sous forme de vecteurs.

CHII- Analyse en Composante Principale (ACP)

II- ACP : Calcul matriciel (Résolution numérique simple)

Étape 6 : représentation dans l'espace des individus. On recalcule les coordonnées des individus dans le système d'axe après rotation.

$$\underbrace{\begin{pmatrix} -3.2 & -1.8 \\ -2.2 & 1.4 \\ -0.2 & -2.6 \\ 1.8 & 3.4 \\ 3.8 & -0.6 \end{pmatrix}}_{\text{Tableau centré (X)}} \times \underbrace{\begin{pmatrix} 0.894 & -0.447 \\ 0.447 & 0.894 \end{pmatrix}}_{\text{Matrice des vecteur propres (U)}} \xrightarrow{\text{X.U=X'}} \underbrace{\begin{pmatrix} -3.58 & 0.00 \\ -1.34 & 2.24 \\ -1.34 & -2.24 \\ 3.13 & 2.24 \\ 3.13 & -2.24 \end{pmatrix}}_{\text{Tableau avec rotation (X')}}$$

Ensuite, on représente ces individus à l'aide d'un graphique en nuage de points.

Tous ces calculs se généralisent facilement à trois, puis à N dimensions.

Reference: consulter le document PCA de James Scott ppour plus de detail