Science des données II : module 7

Analyse en Composantes Principales (ACP)

Philippe Grosjean & Guyliann Engels

Université de Mons, Belgique
Laboratoire d'Écologie numérique des Milieux aquatiques

http://biodatascience-course.sciviews.org
sdd@sciviews.org

Analyse en Composantes Principales (ACP)

- C'est la méthode d'ordination de base (la plus simple, la plus rapide à calculer). En anglais : **P**rincipal **C**omponents **A**nalysis ou **PCA**.
- Analyse un tableau à N variables ($N > 3$) constitué de **données quantitatives**.
- Des relations linéaires sont suspectées entres les variables.
- Ces relations conduisent à une répartition des individus (le nuage de points) qui forme 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).
- ***Comment réduire le nombre de variables à représenter graphiquement en perdant le moins d'information possible ?***

UMONS

Rappel : visualisation de données bivariées

- Le nuage de points (scatterplot) est le graphe idéal pour visualiser la distribution des données bivariées.
- Il permet de visualiser également une association entre deux variables.
- Il permet aussi de visualiser comment deux ou plusieurs groupes peuvent être séparés en fonction de ces deux variables.

Exemple

Mesure des pétales de 3 espèces d'iris (jeu de données iris).

UMONS

Rappel : visualisation de données trivariées

- Le nuage de points en pseudo-3D est l'équivalent pour visualiser 3 variables simultanément.
- Il est nécessaire de rendre l'effet de la **troisième dimension** (perspective, variation de taille des objets, ...)
- La possibilité de **faire tourner l'objet 3D virtuel** est indispensable pour concrétiser l'effet 3D et pour le visionner sous différents angles

=> notre esprit est alors capable de reconstituer la disposition spatiale 3D de l'ensemble.

Exemple

Mesure des pétales + longueur des sépales de 3 espèces d'iris (jeu de données iris).

Visualisation de données multivariées : N > 3

- Comment se représenter graphiquement un tableau de données aussi complexe ?
- La matrice de nuages de points peut servir ici, mais dans certaines limites (tous les angles de vue ne sont pas accessibles).

Exemple

Les quatre variables d'iris.

Autre solution

L'**ACP** qui va réduire le nombre de dimensions à 2 ou 3 (donc facilement représentable graphiquement), tout en conservant un maximum de l'information contenue dans le tableau de départ.

ACP : mécanisme (1)

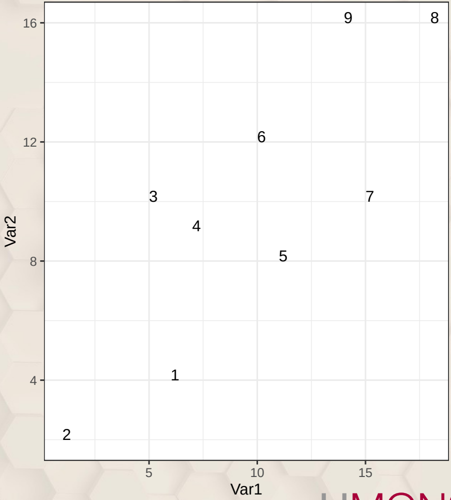
Exemple simple : comment réduire un tableau bivarié en une représentation des individus en une seule dimension (classement sur une droite) ?

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

ACP : mécanisme (2)

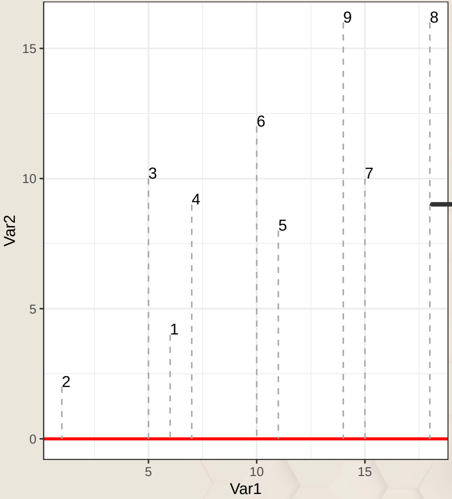
Représentation graphique 2D :

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



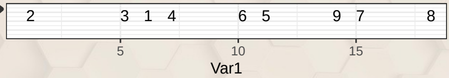
UMONS

ACP : mécanisme (3)



Réduire en 1D ?

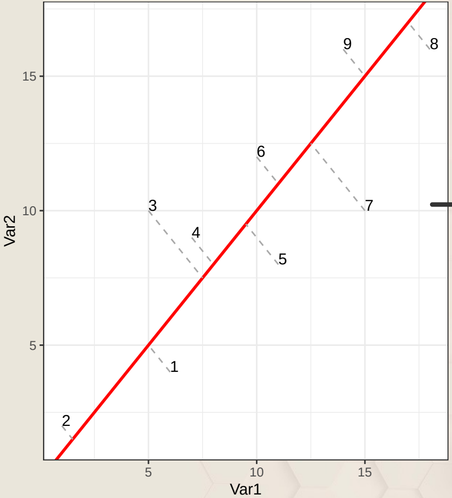
Solution 1 : laisser tomber une variable (ex. : `Var2`)



Mauvaise solution : trop de perte d'information (7 & 9 trop près, 7<->9 / 9<->8, 1<->2 / 1<->3, ...)

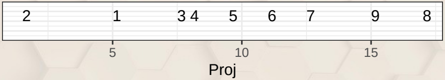
UMONS

ACP : mécanisme (4)



Réduire en 1D ?

Solution 2 : faire un projection sur la droite de "tendance générale"



Meilleure solution : perte minimale d'information
(7 & 9 + éloignés, 7<->9 / 9<->8, 1<->2 / 1<->3, ...)

UMONS

ACP : mécanisme (5)

- **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.
- **Remarque :** on peut calculer la composante 2 comme étant perpendiculaire à la 1 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 le plan.
- Ce mécanisme se généralise facilement à 3, puis à N dimensions.

Préparation des données avant ACP

- Méthode linéaire (combinaison linéaire...) => il faut linéariser les données. **Ex :** allométrie, transformation des données en `log(x)` ou `log(x+1)`
- Centrage : la position dans l'espace importe peu, on s'intéresse à la forme du nuage de points uniquement

=> positionnement du zéro au centre de gravité.

- Réduction : problème d'échelle entre variables ayant des unités différentes...

=> écart type ramené à 1 dans toutes les dimensions.

Remarque

La standardisation (données centrées et réduites) est effectuée automatiquement dans l'analyse lorsqu'on décide de calculer les corrélations. Dans R, on indique `scale = TRUE`.

Démonstration : ACP sur iris

- ACP dans le logiciel.
- Le graphe des éboulis indique la part de variance représenté sur chaque composante principale. Il permet de choisir le nombre de composantes à conserver.
- Générer les cartes correspondant à l'ACP.
- Interpréter les résultats obtenus.

ACP - Rappel de calcul matriciel

- Multiplication matricielle: $\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 11 \\ 5 \end{pmatrix}$

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

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

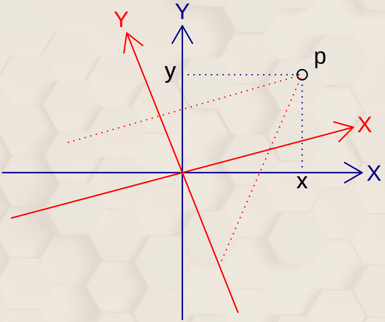
- La constante (4) est une valeur propre et la matrice multipliée (à droite) est la matrice des vecteurs propres.

UMONS

ACP - Rotation d'un système d'axes

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

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).



Exemple numérique simple (1)

ACP sur matrice var/covar sans réduction des données (mais calcul très similaire lorsque les données sont réduites).

Etape 1 : centrage des données

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

Exemple numérique simple (2)

Etape 2 : calcul de la matrice de variance/covariance

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

Exemple numérique simple (3)

Etape 3 : diagonalisation de la matrice var/covar

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

- La **trace** des deux matrices A et B (somme des éléments sur la diagonale) est égale à : $8.2 + 5.8 = 14 = 9 + 5$.
- 8.2 est la **part de variance** exprimée sur le premier axe initial (X)
- 5.8 est la **part de variance** exprimée sur le second axe initial (Y)
- 14 est la **variance totale** du jeu de données
- La matrice diagonale B est la solution exprimant **la plus grande part de variance possible sur le premier axe de l'ACP** : 9, soit 64,3% de la variance totale.
- Les éléments sur la diagonale sont les valeurs propres λ_i !

UMONS

Exemple numérique simple (4)

Etape 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)** effectue la transformation (**rotation des axes**) pour obtenir les **composantes principales**.
- L'angle de rotation se déduit en considérant que cette matrice contient des sin et cos :

$$\begin{pmatrix} 0.894 & -0.447 \\ 0.447 & 0.894 \end{pmatrix} = \begin{pmatrix} \cos(-26.6°) & \sin(-26.6°) \\ -\sin(-26.6°) & \cos(-26.6°) \end{pmatrix}$$

UMONS

Exemple numérique simple (5)

Etape 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 :

Exemple numérique simple (6)

Etape 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.

UMONS

ACP - application à 3 dimensions

- Les calculs restent valables avec 3 variables. Les matrices sont seulement d'autant plus grandes.
- Présentation visuelle : graphe 3D de 3 des variables d'iris.
- Représentation des variables = espace des variables. Approche intuitive en manipulant le graphe 3D.
- Biplot : superposition des deux espaces. Superposition simple des deux = **biplot de distances**. Mise à l'échelle respective (**variables :** $\text{lambda}^{\text{scale}}$, **observations :** $\text{lambda}^{1-\text{scale}}$) = biplot des corrélations.
- Tout ceci se généralise également à n > 3 dimensions.

Exemple

Traitement complet de iris (4 variables)

UMONS