Plan du chapitre 5, partie 1

Introduction I

Objectif

Les *méthodes factorielles* ont pour objectif de visualiser, et plus généralement, de traiter des données multidimensionnelles, c'est-à-dire des données regroupant souvent un grand nombre de variables. La prise en compte simultanée de ces variables est un problème difficile ; heureusement, l'information apportée par ces variables est souvent redondante et toutes ces méthodes vont exploiter cette caractéristique pour tenter de remplacer les variables initiales par un nombre réduit de nouvelles variables sans perdre trop d'information.

Principes

Par exemple, lorsque les variables sont toutes quantitatives, l'analyse en composantes principales (ACP) va chercher à résoudre ce problème en considérant que les nouvelles variables sont des combinaisons linéaires des variables initiales et, qu'en plus, elles doivent être non corrélées linéairement. On passe d'un tableau original X à un tableau synthétique avec le même nombre de lignes mais un nombre de colonnes réduit C.

Historique

Cette méthode a d'abord été développée par K.Pearson (1900) pour deux variables, puis par H. Hotelling (1933) qui l'a étendue à un nombre quelconque de variables.

Un exemple de données

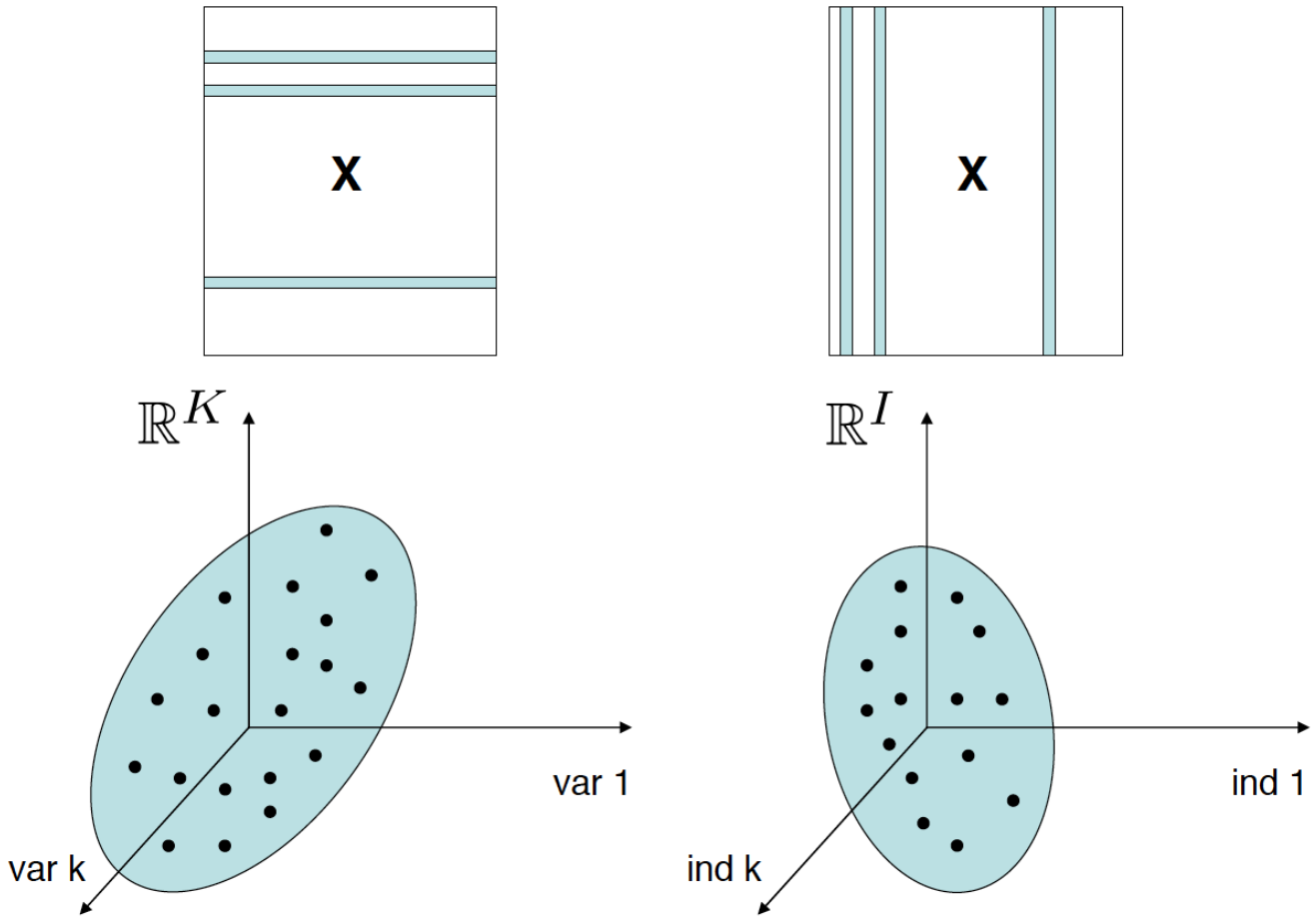
Les données **decathlon** (du package `FactoMineR` :

- 41 athlètes (en ligne) sur lesquels on a mesuré
- 13 variables
  - les performances sur 10 concours (100m, le saut en longueur, etc)
  - 2 variables continues : la rang et le nombre de points obtenus
  - 1 variable catégorielle qui indique l'événement pendant lequel les mesures précédentes ont été faites

```r
data(decathlon)
head(decathlon)
```

```
##           100m Long.jump Shot.put High.jump  400m 110m.hurdle Discus
## SEBRLE   11.04      7.58    14.83      2.07 49.81       14.69  43.75
## CLAY     10.76      7.40    14.26      1.86 49.37       14.05  50.72
## KARPOV   11.02      7.30    14.77      2.04 48.37       14.09  48.95
## BERNARD  11.02      7.23    14.25      1.92 48.93       14.99  40.87
## YURKOV   11.34      7.09    15.19      2.10 50.42       15.31  46.26
## WARNERS  11.11      7.60    14.31      1.98 48.68       14.23  41.10
##          Pole.vault Javeline 1500m Rank Points Competition
## SEBRLE         5.02    63.19 291.7    1   8217    Decastar
## CLAY           4.92    60.15 301.5    2   8122    Decastar
## KARPOV         4.92    50.31 300.2    3   8099    Decastar
## BERNARD        5.32    62.77 280.1    4   8067    Decastar
## YURKOV         4.72    63.44 276.4    5   8036    Decastar
## WARNERS        4.92    51.77 278.1    6   8030    Decastar
```

Deux nuages de points



D'après

`FactoMineR`

Le nuage des individus

Le tableau de données X est une $n \times p$ matrice réelle :

- chaque lignes X_i décrit un individu par p variables
- chaque colonnes X^j décrit une variable par n individus

Centrage de la matrice X

Le nuage des individus est centré autour du centre de gravité du nuage (ou vecteur des moyennes empiriques) :

$$\bar{X} = \frac{1}{n} X_i = \begin{pmatrix} X_{i1} \\ X_{i2} \\ \dots \\ X_{ip} \end{pmatrix}$$

Sans perte de généralité nous supposerons que ce vecteur moyenne est le vecteur nul (il suffit de centrer la matrice X originale).

Variance des vecteurs, matrices ect

Souvent, on travaille également sur des variables re-normalisées, c'est-à-dire que l'on a divisées par leur écart-type

Matrice de variance-covariance et des produits scalaires

Quand les variables ont été centrées, on définit

- la matrice de variance-covariance par

$$S = \frac{1}{n} X^\top X$$

- la matrice des produits scalaire par

$$W = X X^\top$$

Les matrices S et W sont des matrices symétriques définies positives (toutes les valeurs propres sont strictement positives)

Interprétation pour S

- les termes de la diagonale sont les variances empiriques :

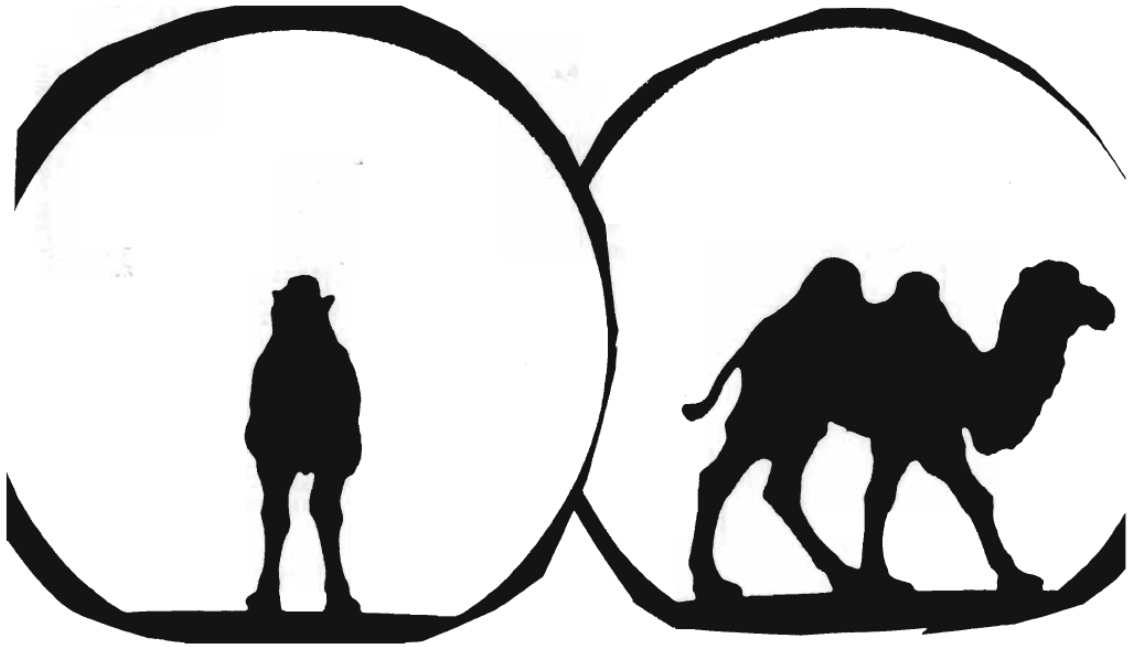
$$\hat{\sigma}_j^2 = \frac{1}{n} \sum_i x_{ij}^2$$

- les termes hors diagonales sont les covariances empiriques

$$\hat{\rho}_{jk}^2 = \frac{1}{n} \sum_i x_{ij} x_{ik}$$

Principe de l'ACP

On va chercher un sous-espace de $\mathbb{R}^k$ (de dimension plus petite) mais dans lequel on garde le maximum d'information



D'après `FactoMineR`

Rappels sur les projections

Attention

- les vecteurs X_i sont en analyse des données des vecteurs lignes, en algèbre linéaire, on considère plutôt $X_i^\top$.
- on suppose $\|\boldsymbol{u}_j\|^2 = 1$.

Projection sur l'axe $\boldsymbol{u}_1$

La projection vectorielle du vecteur $X_i^\top$ sur la droite vectorielle de vecteur directeur $\boldsymbol{u}$ est défini par

$$c_{i1}\boldsymbol{u}_1$$

où $c_{i1} =< X_i^\top, \boldsymbol{u}_1 >= X_i\boldsymbol{u}_1$ est la coordonnée de X_i dans la base $\{\boldsymbol{u}_1\}$.

Projection sur le sous espace vectoriel de base $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d$

La projection vectorielle du vecteur X_i sur le s.e.v. de base $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d$ est défini par le vecteur

$$c_{i1}\boldsymbol{u}_1 + \ldots + c_{id}\boldsymbol{u}_d$$

où $c_{ik} =< X_i^\top, \boldsymbol{u}_k >= X_i\boldsymbol{u}_k$ est la $k$ième coordonnées de $X_i^\top$ dans base.

Projection et acp

On cherche des vecteurs $\boldsymbol{u}_j$ qui

- maximisent la variance des données projetées, pour garder le plus d'information possible
- minimisent la distance entre les variables de départ et leurs projections, pour avoir le moins de distorsion possible

Trouvons le sous-espace (1)

Trouvons le premier axe $\boldsymbol{u}_1$ pour lequel la variance de $X\boldsymbol{u}_1$ est maximale

$$\boldsymbol{u}_1 = \text{argmax}_{\boldsymbol{u}} \mathbb{V}(X\boldsymbol{u}) \text{ avec } \|\boldsymbol{u}\|^2 = 1.$$

Mais

$$\mathbb{V}(X\boldsymbol{u}) = \frac{1}{n}(X\boldsymbol{u})^\top (X\boldsymbol{u}) = \boldsymbol{u}^\top (\frac{1}{n} X^\top X)\boldsymbol{u} = \boldsymbol{u}^\top S\boldsymbol{u}$$

Variance projetée sur un axe I

On cherche à trouver $\boldsymbol{v}$ tel que la projection des individus de X sur le vecteur $\boldsymbol{v}$ (projection orthogonale) soit maximale :

$$\begin{cases} \max_{\boldsymbol{v}} \boldsymbol{v}^t S \boldsymbol{v}, \\ \boldsymbol{v}^t \boldsymbol{v} = 1. \end{cases}$$

avec $S = \frac{1}{n} X^t X$
Si l'on exprime $\boldsymbol{v}$ dans la base (orthonormée) des vecteurs propres de S,

$$\boldsymbol{v} = \sum_{j=1}^{p} \alpha_j \boldsymbol{u}_j$$

alors le problème précédent devient

$$\begin{cases} \max_{\alpha_1,\ldots,\alpha_d} (\sum_{j=1}^{p} \alpha_j \boldsymbol{u}_j)^t U D U^t (\sum_{j=1}^{p} \alpha_j \boldsymbol{u}_j), \\ \sum_j \alpha_j^2 = 1. \end{cases}$$

$$\begin{cases} \max_{\alpha_1,\ldots,\alpha_d} (\sum_{j=1}^{p} \alpha_j^2 \lambda_j), \\ \sum_j \alpha_j^2 = 1. \end{cases}$$

où λ_j est la jème valeur propre.

Variance projetée sur un axe II

Solution

L'équation donne donc un barycentre sur la demi droite des réels positifs entre λ_1 et λ_p. La valeur max du barycentre estλ_1, et elle est obtenue pour $\alpha_1 = 1$ et $\alpha_j = 0\,, \forall j \neq 1$ (car tous les λ_j sont positifs). Le vecteur solution est donc le vecteur propre de S associé à la plus grand valeur propre λ_1. La projection des X_i sur $\boldsymbol{u}_1$ est la première composante principale

$$C^1 = (c_{11}, ..., c_{n1})^t$$

Variance projetée sur un sous espace vectorielle

On admettra que le sous espace vectoriel de dimension k qui maximise la variance projetée est le sous espace vectoriel engendré par les k premiers vecteur propres de S.

Composantes principales

Les projections des X_i sur les vecteurs propres $\boldsymbol{u}_j$ sont les composantes principales :

$$C = \left(C^1 ... C^p\right) = \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ c_{n1} & c_{n2} & \dots & c_{np} \end{pmatrix}$$

Relation entre données originales et composantes principales

$$C = XU$$

Formule de reconstitution

$$X = \sum_j C^j u_j$$

La dernière relation montre que l'on peut reconstituer le tableau initial avec les composantes principales et les axes principaux. Cette relation est appelée formule de reconstitution. Si on se limite aux k $(k < p)$ premiers termes, on obtient une approximation du tableau initial :

Qualité globale

La qualité globale de représentation de l'ensemble initial X sur les k premières composantes principales est mesuré comme le pourcentage de variance expliquée :

$$\frac{\lambda_1 + \ldots + \lambda_k}{\text{trace}(S)} 100.$$

Contribution relative d'un axe à un individu

Sachant que l'inertie totale du nuage est $\frac{1}{n}\sum_{i=1}^{p}\|X_i\|^2$, la quantité $\frac{1}{n}\|X_i\|^2$ représente la part d'inertie apportée par chaque X_i.
Après projection sur l'axe $\boldsymbol{u}_\alpha$, l'inertie restante est donc $\frac{1}{n}(c_\alpha^i)^2$. Chacun des termes $\frac{1}{n}(c_\alpha^i)^2$ représente donc la part de l'inertie initial $\frac{1}{n}\|X_i\|^2$ qu'apportait l'individu i, conservée par l'axe α. Le rapport de ces deux quantités est appelée *contribution relative* du α axe factoriel à l'individu i et elle est notée $COR(i,\alpha)$:

$$COR(i,\alpha) = \frac{(c_\alpha^i)2}{\|X_i\|^2}.$$

Cette quantité représente aussi le carré du cosinus de l'angle formé par l'individu $ X_i$ et par le vecteur $\boldsymbol{u}_\alpha$. Si $COR(i,\alpha)$ est proche de 1, l'individu est bien représenté par cet axe, si $COR(i,\alpha)$ est au contraire proche de 0, l'individu est très mal représenté par cet axe.

$$QLT(i,k) = \frac{\sum_{\alpha=1}^{k}(c_\alpha^i)^2}{\|X_i\|^2} = \sum_{i=1}^{k} COR(i,\alpha).$$

Contribution relative d'un individu à un axe

En partant de la relation $\lambda_\alpha = \frac{1}{n}\sum_{i=1}^{n}(c_\alpha^i)^2$, on peut décomposer λ_α, l'inertie conservée par l'axe $\boldsymbol{u}_\alpha$, selon les individus. On définit alors la contribution relative de l'individu i à l'axe α , notée $CTR(i,\alpha)$: c'est la part d'inertie du αaxe pris en compte (ou expliquée) par l'individu i. Nous avons :

$$CTR(i,\alpha) = \frac{1}{n}\frac{(c_\alpha^i)^2}{\lambda_\alpha}.$$

Cercle des corrélations

- Chaque ancienne variable possède une corrélation avec les nouvelles variables.
- Ces corrélation sont utilisées pour interpréter les nouvelles variables en fonctions des anciennes.