Analyse en Composantes Principales normée

Ghazi Bel Mufti

belmufti@yahoo.com

ESSAI-3 / DATA MINING

Plan

Introduction I

On considère un tableau de données issu de l'observation de p variables quantitatives : $x^1, ..., x^p$ sur n individus. L'Analyse en Composantes Principlaes (ACP) résume ce tableau $(n \times p)$ par un tableau de plus faible dimension, par exemple $(n \times 2)$, en remplaçant les variables initiales x^j par un petit nombre de variables non corrélées ψ_α, appelées composantes principales, combinaison linéaires des x^j, et résumant au mieux l'information initiale.

Introduction II

- **Individus :** l'ACP trouve la meilleure représentation plane de l'ensemble des individus, la qualité globale de la représentation ainsi que la qualité de représentation de chaque individu. Les individus seront représentés par leur projection sur le plan trouvé. On regroupera alors les groupes d'individus homogènes et on détectera les individus exceptionnels.

- **Variables :** de même, l'ACP permet d'avoir la meilleure représentation plane de l'ensemble des variables. On représen-tera les différentes variables par des points sur un plan puis on analysera les relations entre les différentes variables.

- **Individus - Variables :** l'analyse comparée de la carte des individus et celle des variables permet d'analyser l'influence des variables sur les différents groupes d'individus.

ACP sur matrice de corrélation ou ACP normée I

En divisant chaque variable par son écart-type, on obtient un nouveau tableau Z dont les variables sont toutes centrées réduites. On a :

$$Z = \begin{bmatrix} \ddots & & \iddots \\ & z_i^j = \dfrac{x_i^j - \overline{x}^j}{\sigma_j} & \\ \iddots & & \ddots \end{bmatrix}$$

- On note $\mathcal{M}_Z = \{z_1, \ldots, z_n\}$ ce nuage de points individus muni des poids p_i.
- Le barycentre de ce nuage est confondu avec l'origine O.
- On désigne par $z^1, ..., z^p$ les variables *centrées réduites* obtenues à partir des variables initiales $x^1, ..., x^p$.

ACP sur matrice de corrélation ou ACP normée II

- Dans ce cas la matrice $Z' D_p Z$ est la matrice de corrélation. Donc $R = V_Z$, où V_Z désigne la matrice variance associée au tableau Z.
- On réalise l'ACP sur Z avec $M = I_p$.
- L'inertie totale du nuage est alors égal à p, le nombre de variables.

Aide à l'interprétation de la carte des individus I

- La composante principale Ψ_1 est celle qui reflète au mieux la diversité des individus.
- La meilleure représentation plane du nuage des individus est celle où tout individu est représenté par les coordonnées $(\Psi_{i,1}, \Psi_{i,2})$: c'est la première carte des individus.
- Les variables Ψ_α sont non corrélées entre elles.
- Les variables Ψ_α sont des combinaisons linéaires des variables z^j et sont par conséquent centrées.
- Pour tout $\alpha \leq p$: $Var(\Psi_\alpha) = \lambda_\alpha$.

Aide à l'interprétation de la carte des individus II

- **Contribution "absolue" (CTA) :** La CTA du point *i* à l'inertie des projections sur l'axe α est :

$$CTA(i,\alpha) = \frac{p_i(\Psi_{i,\alpha})^2}{\lambda_\alpha}$$

- **Contribution "relative" (CTR) ou cos carrée :** La CTR indiquant la qualité de représentation du point *i* sur le αème axe est donnée par :

$$CTR(i,\alpha) = cos^2(\theta_{i,\alpha}) = \frac{(\Psi_{i,\alpha})^2}{||z_i||^2}$$

où $\theta_{i,\alpha}$ est l'angle entre z_i et u_α.

... et celle des variables I

- **Contribution "relative" (CTR) ou cos carrée :** Par rapport à l'axe α la CTR indiquant la qualité de représentation du point variable *j* est donnée par le cosinus de l'angle que fait *j* avec sa projection $\hat{Z}^{j,\alpha}$ sur le α ème axe :

$$CTR(j,\alpha) = cos^2(\theta_{j,\alpha}) = \rho^2(z^j, \Psi_\alpha) = (\eta_j^\alpha)^2$$

où $\theta_{j,\alpha}$ est l'angle entre z^j et v_α.

- **Communalité :** on appelle communalité d'une variable relativement à 2 ou plusieurs facteurs, la somme des cos2 de cette variable sur ces facteurs. Si l'on retient les deux premiers facteurs,

$$Com(j,(1,2)) = \rho^2(z^j, \Psi_1) + \rho^2(z^j, \Psi_2)$$

Si cette communalité est proche de 1 alors la variable est bien représentée dans le premier plan.

... et celle des variables II

- η_j^α est la coordonnée factorielle du point variable *j* sur l'axe α.
- η_j^α est le coefficient de corrélation entre la variable *j* et la composante principale Ψ_α (i.e $\rho^2(z^j, \Psi_\alpha)$) :

  $\rightarrow$ Une variable qui a une forte coordonnée sur un axe α est fortment correllée à la composante Ψ_α.

- $(\eta_j^\alpha)^2$ est le cos2 du point variable *j* sur l'axe α.

Inertie et choix du nombre d'axes

- **Inertie totale :** L'inertie totale du nuage des individus $\mathcal{N}$

$$\mathcal{I}(\mathcal{N}) = \sum_{j=1}^{p} \lambda_j = Tr(R) = p$$

- **Taux d'inertie :** La qualité globale de la représentation du nuage $\mathcal{N}$ sur le s.e. principal engendré par $(u_1, ..., u_q)$ est mesurée par le taux d'inertie cumulée par ce s.e., elle vaut

$$\frac{\lambda_1 + ... + \lambda_q}{\mathcal{I}(\mathcal{N})} = \frac{\lambda_1 + ... + \lambda_q}{\sum_{j=1}^{p} \lambda_j}$$

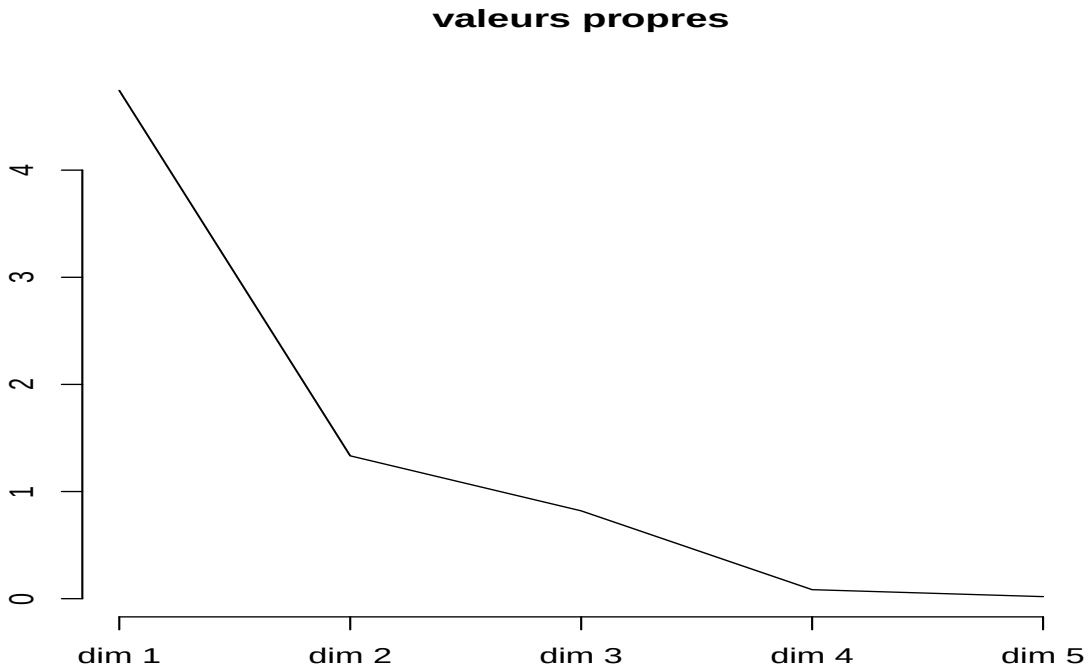
Le taux d'inertie absorbée par le premier plan est donnée par :

$$\frac{\lambda_1 + \lambda_2}{\sum_{j=1}^{p} \lambda_j} = \frac{\lambda_1 + \lambda_2}{p}$$

Nombre d'axes à retenir I

- **Critère de Kaiser.** Il consiste à ne garder, dans une ACP normée, que les axes dont la valeur propre est supérieure à 1 (i.e. l'inertie moyenne).
- **Taux d'inertie cumulé.** Son appréciation doit tenir compte du nombre de variables et du nombre d'individus : un taux d'inertie relatif à un axe de 10% peut être une valeur importante si le tableau possède 100 variables et faible s'il n'en a que 10.

Nombre d'axes à retenir II



- **Critère du coude (*scree test*).** Il consiste à retenir les axes dont les valeurs propres se situent avant le *coude*.

Individu supplémentaire

Soit i_s un individu supplémentaire. On suppose que l'on a centré et réduit ses coordonnées.

- **Coordonnées sur l'axe α :**

$$\begin{aligned}\Psi_{i_s,\alpha} &= < z_s, u_\alpha > \\ &= \frac{1}{\sqrt{\lambda_\alpha}} < z_s, \eta^\alpha > \\ &= \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^{p} z_s^j \eta_j^\alpha\end{aligned}$$

Variable quantitative supplémentaire

Soit x^s une variable supplémentaire et z^s la variable centrée réduite correspondante.

- **Intérêt :** éclairage aux variables quantitatives actives par les relations qu'elles ont avec une variable supplémentaire.
- **Exemple :** les variables actives sont des réponses à une échelle mesurant la satisfaction par un produit et la variable supplémentaire est l'âge de la personne enquêtée.
- **Coordonnées sur l'axe α :**

$$\begin{aligned}
\eta_s^\alpha &= < z^s, v_\alpha >_{D_p} \\
&= \frac{1}{\sqrt{\lambda_\alpha}} < z^s, \psi_\alpha >_{D_p} \\
&= \frac{1}{\sqrt{\lambda_\alpha}} \frac{1}{n} \sum_{i=1}^{n} z_i^s \psi_{i,\alpha} \\
&= \rho^2(z^s, \psi_\alpha)
\end{aligned}$$

Variable qualitative supplémentaire I

- **Intérêt :** en représentant chaque modalité de la variable qualitative au barycentre des individus qui la possèdent on apporte une information d'une autre nature sur les groupes d'individus.
- **Exemples :** le secteur d'activité s'il s'agit d'entreprises ou le sexe, s'il s'agit de clients...
- **Représentation sur la carte des individus :**
  Soit x^s une variable supplémentaire qualitative à m modalités.
  1. On calcule les centres de gravité $g^1, \dots g^m$ des m groupes engendrés par les m modalités :

$$g^l = (\overline{x}^{l,1}, \dots, \overline{x}^{l,p}), \quad 1 \leq l \leq m$$

où pour $j \in \{1, \dots, p\}$, $\overline{x}^{l,j} = \sum_{i \ tq \ x_i^s = l} p_i x_i^j$ est la moyenne de la variable x^j calculée sur les individus possédant la l ème modalité de x^s.

Variable qualitative supplémentaire II

2. Puis $g^1, \ldots g^m$ sont traités comme des individus supplémentaires :

   $\rightarrow$ chaque modalité de x^s est représentée, sur la carte des individus, au braycentre des individus qui la possède

Démarche

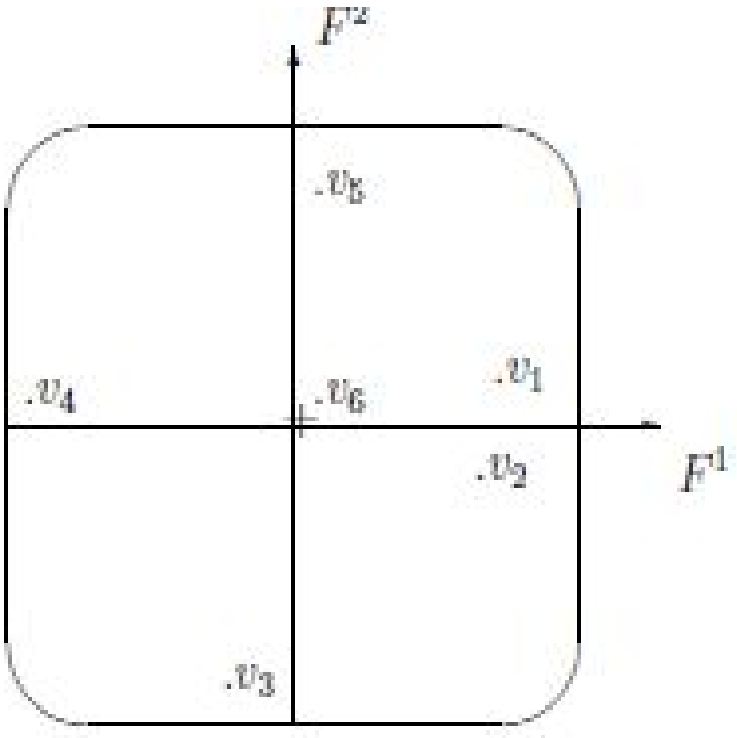
1. Variable à retenir : dans une analyse d'une carte des variables, on ne s'intéresse qu'aux variables bien représentées sur cette carte (i.e. aux variables proches du cercle de corrélation).

2. Variable-axe : les variables fortement corrélées avec un facteur vont contribuer à la définition de cet axe.

3. Variable-variable :

- Un angle faible entre 2 variables indique une forte corrélations entre elles.
- 2 points variables diamétralement opposés indiquent une parfaite corrélation négative (*i.e.* $\rho(j, j') \simeq -1$) entre ces variables.
- Des directions presque orthogonales indiquent une faible corrélation linéaire entre j et j'.

Exemple I



Exemple II

- la variable v_6 est à exclure de l'étude car proche de l'origine
- la première composante principale est fortement corrélée aux variables v_1, v_2 et v_4 car celles-ci ont de forte coordonnées sur cet axe
- elle est par contre très peu corrélée à v_3 et v_5

  $\rightarrow$ on dit que cette première composante oppose la variable v_4 aux variables v_1 et v_2
- la deuxième composante oppose la variable v_3 à la variable v_5

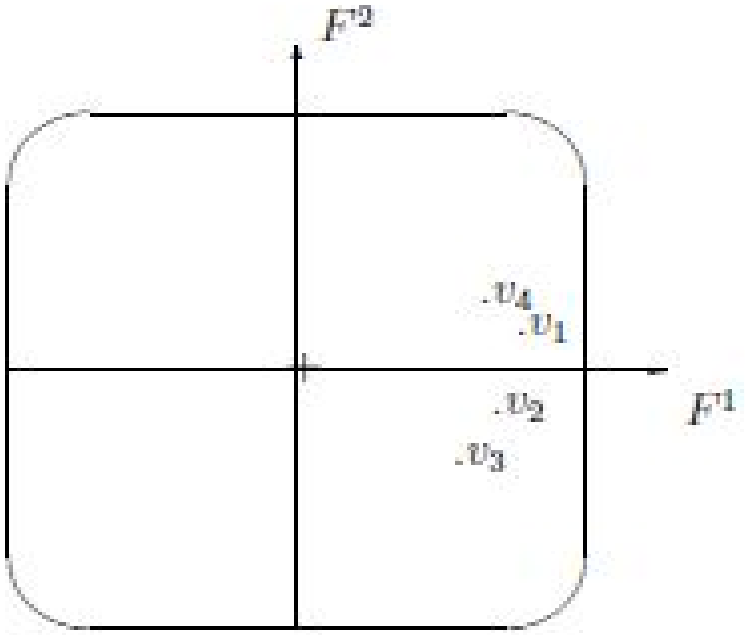
Rotations

Pour aider à l'interprétation, il peut être commode, une fois le nombre de facteurs déterminé, d'effectuer une rotation des axes :

- La rotation (la méthode VARIMAX, ...) permet de se rapprocher d'une *structure simple* :
  - ⋆ une composante est fortement corrélée avec quelques variables et peu corrélée avec les autres.
  - ⋆ une variable est corrélée avec une seule composante.
- Dans ce cas, l'information restituée par le plan factoriel reste la même mais celle restituée par les axes change.

Facteur taille

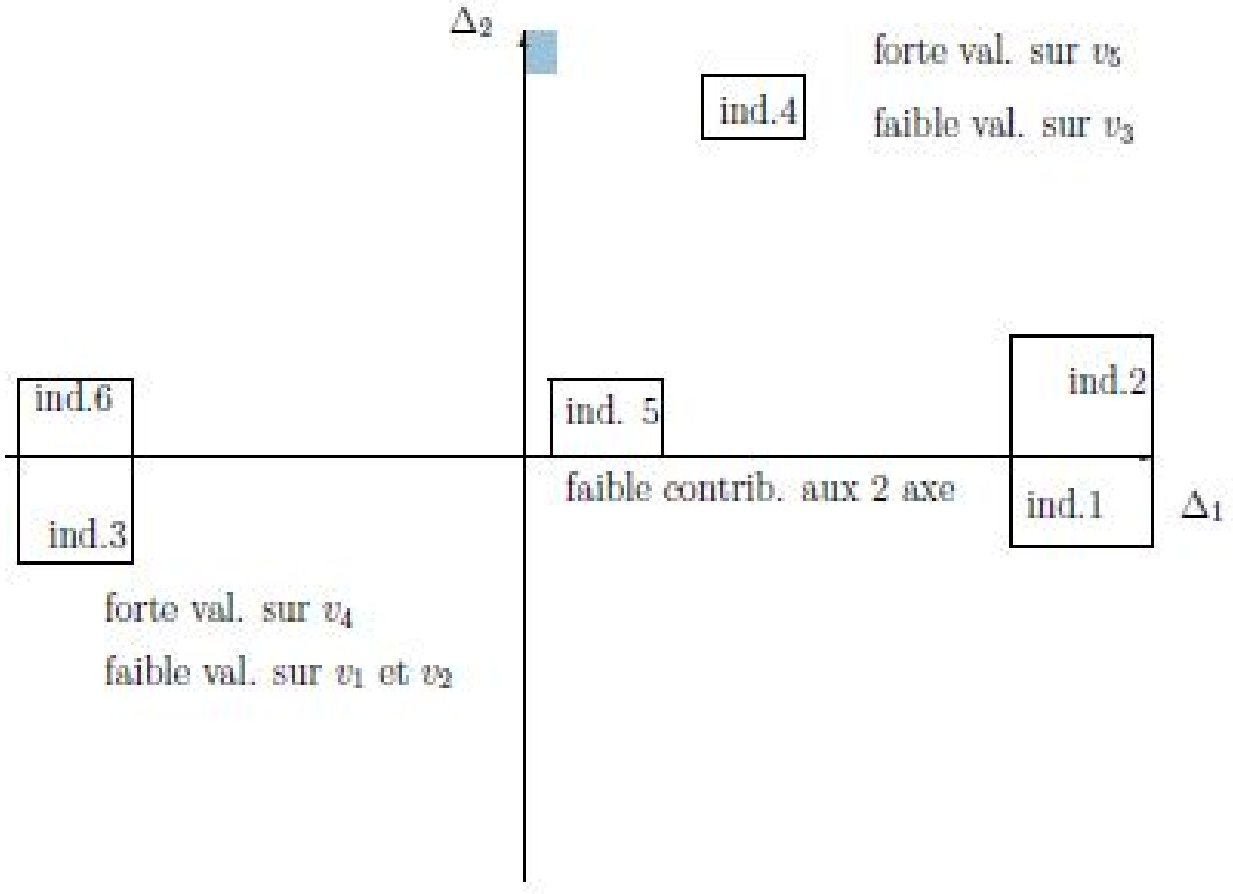
- Les variables peuvent être toutes du même coté d'un axe : une telle disposition apparaît lorsque toutes les variables sont corrélées positivement entre elles.
- Cette caractéristiques apparaît le plus souvent sur le premier axe que l'on appelle alors *facteur taille*.



Démarche

1. Exclure de l'analyse les individus mal représentés sur le plan étudié (i.e. cosinus carré faible).
2. Regrouper les individus assez proches.
3. Donner les spécificités de chaque groupe grâce à l'analyse de la carte des variables associée.

Exemple


Δ_2
forte val. sur v_5
ind.4
faible val. sur v_3
ind.2
ind.6
ind. 5
faible contrib. aux 2 axe
ind.1
Δ_1
ind.3
forte val. sur v_4
faible val. sur v_1 et v_2