Cours d'analyse des données

Chapitre 1 : Analyse en Composantes Principales (ACP)

Zaineb Smida

5ème année GEA
INSA Lyon

15 octobre 2024

Rappel : Liaison entre deux variables quantitatives

Soient X et Y deux variables quantitatives.
On note x_i (resp. y_i) la valeur observée de X (resp. Y) pour le $i$ème individu de l'échantillon (de taille n)

- Covariance (empirique) de X et Y

$$cov(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

- $cov(X, Y) = 0$: X et Y non corrélées linéairement
- $cov(X, Y) > 0$: X et Y varient dans le même sens
- $cov(X, Y) < 0$: X et Y varient dans le sens opposé

- Coefficient de corrélation linéaire

$$r(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

$r(X,Y) \in [-1,1]$
  - $r = 0$: X et Y non corrélées linéairement
  - r proche de 0 : X et Y faiblement corrélées linéairement
  - $|r|$ proche de 1 : X et Y fortement corrélées linéairement
  - $|r| = 1$: liaison linéaire exacte entre X et Y
- La visualisation de la liaison entre deux variables quantitatives se fait à l'aide d'un nuage de points

Introduction générale

Présentation rapide de l'ACP :

- Cette méthode a pour objectif la description graphique et numérique de données contenues dans un tableau individus×variables de dimension $n \times p$. Les p variables mesurées sur les n individus sont quantitatives. C'est une méthode de "réduction du nombre de variables" permettant des représentations géométriques des n individus et des p variables initiales.
- L'ACP est une méthode factorielle car la réduction du nombre de variables ne se fait pas par la sélection de certaines d'entre elles, mais par la construction de nouvelles variables synthétiques obtenues en combinant les variables initiales à l'aide de facteurs.
- Cette méthode est linéaire car il s'agit de combinaisons linéaires.

Introduction générale

Remarque

- *Cette réduction ne sera possible que si les p variables ne sont pas linéairement indépendantes entre elles (dans le sens où elles ont des coefficient de corrélations linéaires non nuls).*
- *Une spécificité de l'ACP est que cette méthode traite exclusivement des variables quantitatives jouant toutes les mêmes rôle, dans le sens où il n'y en a pas certaines à expliquer par d'autres explicatives.*
- *Les outils utilisés sont ceux de l'algèbre linéaire (dont le calcul matriciel), de l'optimisation et de la statistique descriptive. Les notions les plus imployées sont :*
  - *espaces vectoriels euclidiens, centre de gravité, produit scalaire, norme, distance, inertie,...*
  - *projection, décomposition aux éléments propres,...*
  - *moyenne, variance, écart-type, corrélation linéaire, fréquence,...*

Nature des données pour l'ACP

Les données sont présentées sous la forme d'un tableau brut individus/variables de taille $n \times p$ où n est le nombre d'individus et p est le nombre de variables quantitatives.

On note X le tableau de données et $X^1, X^2, \cdots, X^p$ les p variables quantitatives du tableau.
En pratique, dans la plus part des données réelles, on n'observe non pas une ou deux variables par individus, mais un nombre p souvent élevé de variable.

Notre exemple d'application concerne les 21 régions françaises sauf la Corse (les individus) caractérisées par différents indicateurs (les variables) de la démographie, de l'économie, de la société et des conditions de vie pendant l'année 2002.

Les 7 variables quantitatives considérées pour ce chapitre sont les suivantes :

- POPUL : population de la région (en milliers d'individus)
- TACT : taux d'activité (population active / population totale de la région) en pourcentage
- SUPERF : superficie de la région (en kilomètres carrés)
- NBENTR : nombre d'entreprises de la région
- NBBREV : nombre de brevets déposés au cours de l'année
- CHOM : taux de chômage (en pourcentage)
- TELEPH : nombre de lignes téléphoniques en place dans la région (en milliers)

Exemple

Voir les notes de cours : tableau 1
Ici : $n = 21$ individus et $p = 7$ variables.

- L'étude séparée de chacune des variables est une phase indispensable dans le processus de dépouillement des données.

Exemple

Voir les notes de cours : Boxplot 1
Cette étude "univariée" est tout à fait insuffisante car elle ne tient pas compte des liaisons qui peuvent exister entre les variables, liaisons qui sont souvent l'aspect le plus important. Il est donc préférable d'analyser les données en tenant compte de leur caractère multidimensionnel.

- Lorsque l'on considère deux variables simultanément (X^1 et X^2 par exemple), il est facile de représenter, sur un graphique plan, l'ensemble des données. Le simple visuel de l'allure du nuage de points $\{X_i^1, X_i^2), i = 1, \ldots, n\}$ permet d'avoir une idée sur la forme et l'intensité de liaison entre ces deux variables, et de repérer les individus ou les groupes d'individus ayant des caractéristiques voisines.

Exemple

Voir les notes de cours : PLOTS NUAGE DES POINTS

- Si l'on considère trois variables (X^1, X^2 et X^3 par exemple), l'étude visuelle est encore possible en faisant de la géométrie dans l'espace. Les logiciels de statistique proposent ce genre de graphiques interactifs en trois dimensions dans lesquels il est possible de faire tourner les axes pour observer le nuage des points sur toutes ses formes.
- Lorsque l'on considère un nombre p de variables, avec $p \geq 4$, la visualisation directe et totale de toutes les données devient impossible. On peut étudier graphiquement les variables par groupe de 2 : cependant, s'il y a par exemple $p = 11$ variables, cela représentera $p(p-1)/2 = 55$ nuages de points croisant 2 variables à regarder !

Il apparaît donc utile et nécessaire de trouver une manière de visualiser les données multidimensionnelles.

Objectifs de l'ACP

L'ACP a deux objectifs principaux :

1. *Résumer* le tableau de données X par un petit nombre k de nouvelles variables non corrélées entre elles et qui conservent au maximum l'information contenue dans les p variables initiales.

   Intuitivement, on peut dire que ces nouvelles variables sont obtenues en "réunissant" les variables de départ qui sont bien corrélées entre elles.

Exemple

Voir le tableau 3 des notes de statistique descriptive qui donne la matrice des corrélations des variables initiales
POPUL, NBBREV, NBENTR et TELEPH sont très corrélées linéairement positivement entre elles.
CHOM est corrélée linéairement négativement avec TACT.
SUPERF n'est pas corrélée linéairement aux autres variables.

Remarque

- le nombre k de ces nouvelles variables est d'autant plus petit que les corrélations entre les p variables initiales sont importantes.
- comme sous-produit, l'ACP conduit à une visualisation des corrélations entre les variables initiales.

2. *Interpréter* le tableau de données en utilisant les nouvelles variables et des représentations graphiques de type nuages de points.

   L'ACP permet notamment de repérer des individus atypiques ou des groupes d'individus ayant un comportement similaire par rapport aux caractères considérés.

Principe de l'ACP

Notion d'information

L'objectif de "conserver au maximum l'information contenue dans un tableau de données" suppose que l'on définisse mathématiquement la notion d'information. Cette information se fonde sur la variabilité des données et est mesurée par la variance.

Définition

- *l'information apportée par une variable quantitative* X^j *est la variance de* X^j.
- *l'information apportée par un tableau de données* $X = [X^1, X^2, \cdots, X^p]$ *est la somme des variances des variables de* X.

  On l'appelle inertie de X *et on la note* I_X : $I_X = \sum_{j=1}^{p} Var(X^j)$.

Remarque

Inconvénient de la variance : elle dépend de l'unité.

Principe de l'ACP

Variables centrées réduites issues des variables initiales

Soit x^j la variable centrée réduite associée à X^j : $x^j = \frac{X^j - \overline{X^j}}{\sigma_{X^j}}$

Donc $\overline{x^j} = 0$ et $\mathrm{Var}(x^j) = 1$.

Ainsi, les variables centrées réduites apportent toutes la même information (égale à 1).

Le tableau centré réduit est noté x.

Remarque

- $r(x^k, x^j) = r(X^k, X^j)$.
- *Pour nous, l'ACP portera toujours sur des variables centrées réduites. Il s'agit d'un cas particulier de l'ACP appelé parfois* ACP réduite.

Principe de l'ACP

Composantes principales

1. **La première composante principale** c^1 est définie comme une nouvelle variable, combinaison linéaire des variables $x^1, x^2, \ldots, x^p$ s'exprimant sous la forme

$$c^1 = \alpha_1 x^1 + \alpha_2 x^2 + \ldots + \alpha_p x^p \text{ avec } \sum_{j=1}^{p} \alpha_j^2 = 1 \qquad (1)$$

et telle que l'information apportée par c^1 est maximale.
Autrement dit, on cherche les coefficients $\alpha_1, \alpha_2, \ldots, \alpha_p$ tels que c^1 soit de variance maximale.

2. La **deuxième composante principale** c^2 est définie comme étant une nouvelle variable *non corrélée avec* c^1, combinaison linéaire des variables x^j, $j = 1, \ldots, p$ et de variance maximale.

3. La **troisième composante principale** c^3 est *non corrélée avec* c^1 *et* c^2, combinaison linéaire des x^j et de variance maximale.

   ...

4. La ***p*ème composante principale** c^p est *non corrélée avec* c^1, c^2, ..., c^{p-1}, combinaison linéaire des x^j et de variance maximale.

On a ainsi défini p composantes principales non corrélées entre elles et que l'on peut regrouper dans un tableau de composantes principales noté $C = [c^1, c^2, \cdots, c^p]$.

D'après la définition précédente, $\text{Var}(c^1) \geq \text{Var}(c^2) \geq \cdots \geq \text{Var}(c^p)$.

En effet, chacune des composantes est définie à partir d'un critère de maximisation de variance mais avec une contrainte de plus pour c^2 que pour c^1 (coefficient de corrélation nul entre c^2 et c^1), pour c^3 que pour c^2 (coefficient de corrélation nul entre c^3 et c^2),. . ., pour c^{p-1} que pour c^p (coefficient de corrélation nul entre c^p et c^{p-1}).

De plus, on montrera que l'information apportée par le tableau x se retrouve entièrement reconstituée dans le tableau C. Autrement dit, l'inertie de C est égale à l'inertie de x :

$$I_x = \sum_{j=1}^{p} \mathrm{Var}(x^j) = p = I_C = \sum_{j=1}^{p} \mathrm{Var}(c^j).$$

Mais, alors que chacune des colonnes de x apporte la même information (égale à 1), les colonnes du tableau C apportent une information qui décroît avec le numéro de la colonne.

On comprend dès lors que l'on peut atteindre le premier objectif de l'ACP, c'est-à-dire résumer le tableau x par un tableau contenant moins de colonnes si les dernières composantes principales apportent peu d'information (i.e. sont de faible variance).

Remarque

On peut introduire l'ACP par d'autres critères que la maximisation de la variance. L'approche géométrique notamment (basée sur les projections) est souvent adoptée.

Composantes principales

Calcul des composantes principales

1. On calcule R, la matrice des corrélations des variables $x^1, \cdots, x^p$, définie par :

$$R = \begin{pmatrix} 1 & r(x^1, x^2) & \dots & r(x^1, x^p) \\ r(x^2, x^1) & 1 & \dots & r(x^2, x^p) \\ \dots & \dots & \dots & \dots \\ r(x^p, x^1) & r(x^p, x^2) & \dots & 1 \end{pmatrix} = \frac{1}{n} x' x \quad \text{(cf. démo)}$$

2. On calcule les valeurs propres (*eigenvalues*) et les vecteurs propres (*eigenvectors*) normés de la matrice R, c.à.d. les $\lambda_j \in \mathbb{R}$ et les $v_j \in \mathbb{R}^p$ pour $j = 1, \dots, p$ de norme 1 t.q.

$$R\, v_j = \lambda_j\, v_j.$$

   On trie les valeurs propres λ_j par ordre décroissant :

$$\lambda_1 > \lambda_2 > \cdots > \lambda_p$$

3. On calcule :

$$c^j = x\, v_j$$

Composantes principales

Calcul des composantes principales

Exemple

Voir le tableau ACP1 (colonne "eigenvalue") qui donne : $\lambda_1 = 4,329...$, $\lambda_2 = 1,429...$, $\lambda_3 = 1,012...$ *etc...*
Voir le tableau ACP6 qui donne les 3 premières composantes principales

Composantes principales

Propriétés

Proposition

- $c^j = x\, v_j \hookrightarrow c^j$ *est une combinaison linéaire des variables initiales*
- $\overline{c^j} = 0$,
- $Var(c^j) = \lambda_j$,
- $r(c^j, c^k) = 0$ *pour* $j \neq k$.

(cf. démo)

Remarque

$$I_x = I_C = p = \sum_{j=1}^{p} Var(c^j) = \sum_{j=1}^{p} \lambda_j$$

(cf. démo)

Choix des k composantes principales

Critère de la part d'inertie expliquée

On doit choisir un nombre k suffisant de composantes principales pour résumer l'information (inertie) de départ sans trop en perdre.
Information totale : $I_x = p$
Information apportée par c^j : $\text{Var}(c^j) = \lambda_j$,

part d'inertie expliquée par (c^1) : λ_1/p,
part d'inertie expliquée par (c^1 et c^2) : $(\lambda_1 + \lambda_2)/p$
part d'inertie expliquée par k c. p. $(c^1, c^2, \cdots, c^k)$: $\sum_{j=1}^{k} \lambda_j/p$
part d'inertie expliquée par les p c. p. : $\sum_{j=1}^{p} \lambda_j/p = p/p = 100\%$.

Critère

k est choisi le plus petit possible tel que la part d'inertie expliquée soit suffisamment grande (au moins 80 %).

Choix des k composantes principales

Critère de la part d'inertie expliquée

Exemple

Voir le tableau ACP1 (colonne "cumulative percentage of variance") :
Part d'inertie expliquée par c^1 *:* $4,3296/7 = 0,6185 = 61,85\%$
Part d'inertie expliquée par c^1 *et* c^2 *:*
$(4,3296 + 1,4293)/7 = 0,8227 = 82,27\% > 80\%$

Choix des k composantes principales

Critère de Kaiser

Critère

Les variables initiales ont une variance = 1 (réduites).
Retenir les composantes principales de variance > 1 *car elles apportent plus d'information que les variables initiales :*

$$\hookrightarrow k = \text{ nombre de } \lambda_j > 1.$$

Exemple

$\lambda_1 > 1$, $\lambda_2 > 1$, $\lambda_3 > 1$ *et* $\lambda_4 < 1$. *Par ce critère, on retient les 3 premières composantes principales.*

Choix des k composantes principales

Critère de la différence

On regarde les différences entre valeurs propres successives :

$$\lambda_1 - \lambda_2,\ \lambda_2 - \lambda_3,\ \ldots$$

En général, ces différences diminuent.

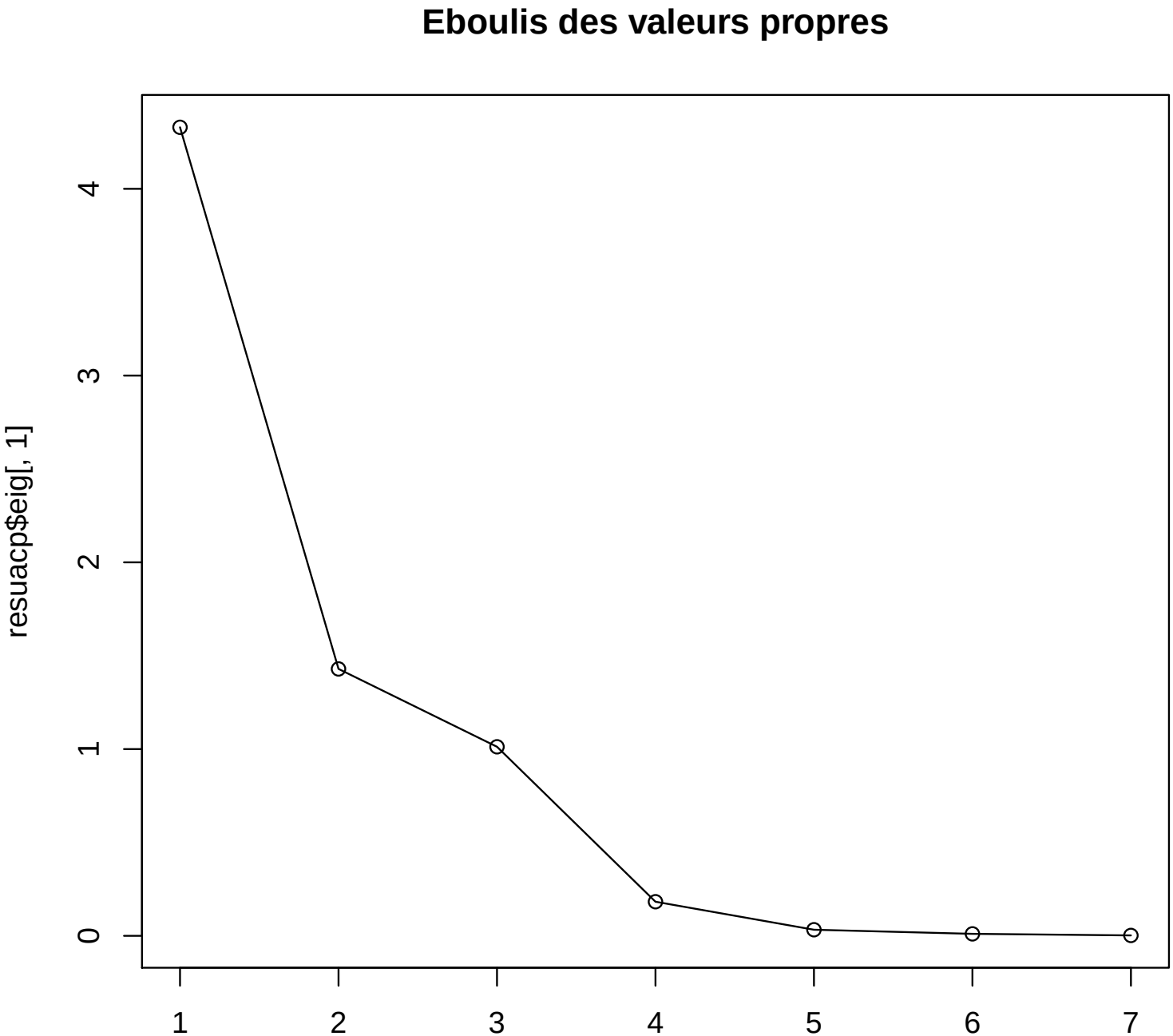
Critère

Retenir les k composantes principales telles que la différence $\lambda_k - \lambda_{k+1}$ soit grande et que les différences $\lambda_j - \lambda_{j+1}$, $j = k+1, \ldots, p-1$ soient faibles.

$\hookrightarrow$ cela revient à s'arrêter avant le "coude" sur l'éboulis des valeurs propres.

Exemple

D'après l'éboulis des valeurs propres, on voit un coude à $k = 4$.
Donc, par ce critère, on retient uniquement les trois premières composantes principales.


Eboulis des valeurs propres
resuacp$eig[, 1]
0
1
2
3
4
1
2
3
4
5
6
7

Interprétation des composantes principales

Présentation du problème

On suppose, qu'après utilisation d'un des critères précédents, on a selectionné k (petit) composantes principales (ou k dimensions ou k facteurs).

Exemple

Après arbitrage entre les différents critères, on décide de retenir les 3 premières composantes principales.

Une des difficultés de l'ACP (et des analyses factorielles en général) est l'interprétation des composantes principales.

L'ACP conduit à une réduction du nombre de variables (de p à k) mais si on connaît la signification des variables initiales, il n'en est pas de même des composantes principales.

Interprétation des composantes principales

Interprétation des coefficients des combinaisons linéaires

$$
\begin{aligned}
c^j &= x\, v_j, \quad j = 1, \ldots, p \\
c^j &= \sum_{\ell=1}^{p} v_{j\ell}\, x^\ell
\end{aligned}
$$

On connaît donc la "composition" de c^j et les variables x^ℓ importantes sont celles associées aux grands coefficients $v_{j\ell}$ (parce qu'elles ont toutes la même variance).

Mais cette méthode est rarement utilisée pour interpréter les c^j (grandeur des coeff. difficilement évaluable).

Interprétation des composantes principales

Étude des corrélations entre composantes principales et variables initiales

- On préfère utiliser les corrélations entre les composantes principales et les variables initiales et interpréter directement ces corrélations

Exemple

Voir ACP3 :

- *c^1 est surtout très corrélée linéairement positivement avec POPUL, NBENTR, NBBREV et TELEPH et assez corrélée positivement avec TACT. Donc c^1 peut s'interprèter comme une composante "Potentiel de développement économique" des régions (plan humain, économique).*
- *c^2 est fortement corrélée linéairement positivement avec CHOM et assez corrélée négativement avec TACT. c^2 est une mesure de l'activité de la région.*
- *c^3 est fortement corrélée linéairement positivement avec la variable SUPERF. c^3 est une mesure de la superficie.*

Remarque

On note que $r(c^1, TACT) = 0,72$ *et* $r(c^2, TACT) = -0,59$. *Cela pose un problème d'interprétation car TACT est corrélée à la fois avec* c^1 *et* c^2. *Ce problème sera résolu avec l'interprétation des contributions des variables aux composantes principales retenues (voir paragraphe 8.4).*

- Mais en général, on préfère **représenter** ces corrélations en considérant les composantes principales 2 par 2 et les interpréter graphiquement (possible car k petit $\Rightarrow$ peu de possibilités).
- Les dessins s'inscrivent évidemment dans un carré $[-1, +1] \times [-1, +1]$ (coefficients de corrélation).
- On peut montrer que les points sont toujours dans le cercle centré à l'origine et de rayon 1. On trace souvent ce cercle car il aide à l'interprétation.

Exemple

Voir les figures des notes de cours sur l'ACP ("Variable factor map")

En pratique,

- Pour chaque paire (ou plan) $((c^1, c^2), (c^1, c^3), (c^2, c^3), \ldots (c^{k-1}, c^k))$, représenter les corrélations.
- Tracer le cercle des corrélations.
- Repérer les corrélations fortes, c.a.d. les points proches du cercle. On ne doit pas s'intéresser aux variables initiales trop éloignées du cercle car elles n'interviennent pas ou peu dans le calcul des composantes principales et donc ne servent pas à leur interprétation.
- Interpréter chaque composante en fonction des corrélations fortes avec les variables initiales (positives et négatives).

Interprétation des composantes principales

Contribution d'une variable à un axe

- On peut aussi mesurer la qualité de la représentation d'une variable x^ℓ sur un axe c^j par $r^2(x^\ell, c^j)$ mais cela n'apporte rien de plus que l'étude des corrélations (voir tableau ACP4).
- On peut montrer que $\sum_{\ell=1}^{p} r^2(x^\ell, c^j) = \lambda_j$. On peut donc définir la contribution (en pourcentage) de la variable x^ℓ à la composante c^j par

$$\frac{r^2(x^\ell, c^j)}{\lambda_j} \times 100.$$

- L'étude des contributions permet de mieux associer une variable initiale à une composante principale (cf. TACT dans exemple des régions).

Exemple

Voir tableau ACP5
Seuil = 100/7 =14,29
- Fortes contributions de POPUL, NBBREV, NBENTR et TELEPH à c^1
- Fortes contributions de CHOM et TACT à c^2 $\hookrightarrow$ On interprètera uniquement la corrélation de TACT avec c^2.
- Forte contribution de SUPERF à c^3.

Interprétation des individus

Graphique des individus

On dispose de k nouvelles variables dont on connaît la signification. Pour interpréter le tableau de départ, on représente les individus dans les plans des k composantes principales prises 2 à 2 (= **plans principaux**).

On interprète les graphiques obtenus comme n'importe quel graphique de type nuage de points en tenant compte de l'interprétation des composantes principales. Mais, comme pour le graphique des corrélations des variables, on ne doit pas interpréter des individus mal représentés.

Interprétation des individus

Interprétation géométrique

Remarque

On peut aussi présenter l'ACP par une approche **géométrique** *:*

- *L'ACP correspond à projeter les observations sur un espace $\mathbb{R}^k$ avec k petit tout en essayant de perdre le minimum d'information.*
- *Mais toute projection implique une déformation des distances (toujours plus petites).*
- *Pour interpréter un individu dans les plans principaux, il faut que les distances soient bien conservées (= individu bien représenté). En effet, des points en apparence proches peuvent être fort éloignés dans l'espace sur les autres dimensions laissées de côté par le graphique.*

Interprétation des individus

Mesure de la qualité de représentation des individus

On choisit, pour mesurer cette qualité de représentation, de regarder la **distance à l'origine** de chacun des individus.

- Au départ, un individu i (de vecteur des valeurs des variables x_i) est à une certaine distance de l'origine O :

$$d(x_i, O) = \sqrt{\sum_{j=1}^{p} \left(x_i^j\right)^2} = \sqrt{\sum_{j=1}^{p} \left(c_i^j\right)^2}$$

(on parle aussi de norme de l'individu i : $d(x_i, \mathrm{O}) = \|x_i\|$).

Exemple

Voir tableau ACP9

Par exemple : $d(IdF, O) = 8,62$

- Après projection, la norme devient :
  - sur un espace de dimension k,

$$\|Px_i\|_k = \sqrt{\sum_{j=1}^{k} \left(c_i^j\right)^2}$$

  - sur un axe c^j : $\|Px_i\|_1 = |c_i^j|$.

Pour chaque individu, on compare sa norme de départ au carré avec sa norme après projection au carré en calculant le rapport des 2, soit sur l'axe c^j :

$$\frac{(c_i^j)^2}{\sum_{\ell=1}^{p} \left(x_i^\ell\right)^2} = \frac{(c_i^j)^2}{\sum_{\ell=1}^{p} \left(c_i^\ell\right)^2} = cos^2\theta_i^j.$$

Remarque

Pour un individu i fixé, $\sum_{j=1}^{p} cos^2\theta_i^j = \sum_{j=1}^{p} \frac{\left(c_i^j\right)^2}{\sum_{\ell=1}^{p}\left(c_i^\ell\right)^2} = 1.$

Valeurs seuils indicatives :

- un individu i est bien représenté sur c^1 si $cos^2\theta_i^1 > 0,5$,
- un individu i est bien représenté sur c^2 si $cos^2\theta_i^2 > 0,25$,
- un individu i est bien représenté sur c^3 si $cos^2\theta_i^3 > 0,15$.

$\hookrightarrow$ On ne commente que les individus bien représentés (axe par axe).

Exemple

- *Tableau ACP7 et figure "Individuals factor map" : une région est bien représentée sur C1 si cos2>0,5. Donc les régions bien représentées sur c^1 sont : Auvergne, Champagne-Ardennes,Ile de France, Picardie, Poitou-Charentes et Rhône-Alpes.*
  U, E, I, D, T, et R peuvent être interprétés. On repère sur le graphique que l'Ile de France (I) est seule à droite sur l'axe 1. Ce qui signifie qu'elle a c^1 élevée et donc que Idf se caractérise par : popul, nbbrev, nbentr et teleph : élevées (car toutes les corr. sont positives). La région Rhône-Alpes (R) est aussi relativement à droite sur le graphique par rapport aux autres régions. Donc l'idf et Rhône-Alpes s'opposent à l'Auvergne, la Bourgogne, la Picardie et Poitou-Charentes qui ont un potentiel de développement démographique et économique peu élevé.
  Difficulté de l'interprétation car régions agglomérés à cause de I (refaire l'analyse sans I).

Exemple

- *Tableau ACP7 et "Individuals factor map", deuxième axe : une région est bien représentée sur c^2 si cos2>0,25 Donc les régions bien représentées sur c^2 sont : Alsace, Aquitaine, Basse- Normandie, Bourgogne, Bretagne, Franche-Comté, Languedoc-Roussillon, Limousin, Nord-Pas-de-Calais et Provence-Côte d'Azur. A, S, F sont opposées à G, P, Z et Q.*
  A, S et F se caractérisent par un tact élevé et un chômage faible alors que G, P, Z et Q se caractérisent par un chômage élevé et un taux d'activité faible. Enfin, N, O et P sont des régions moyennement économiquement dynamiques.
- *Tableau ACP7 et "Individuals factor map", troisième axe uniquement : une région est bien représentée sur c^3 si cos2>0,15 Donc les régions bien représentées sur c^3 sont : Aquitaine, Bourgogne, Centre, Haute-Normandie, Midi-Pyrénées, Nord-Pas de Calais, Pays de Loire, Picardie et Rhône-Alpes.*
  P, H et D : superficie faible alors que C, Q, M et R : superficie relativement élevée. O et Y : superficie moyenne.

Complément : contribution d'un individu à un axe

- La contribution (en pourcentage) de l'individu i à la composante c^j est définie par

$$\frac{\frac{1}{n} c_i^{j2}}{\lambda_j} \times 100.$$

- Une forte contribution est une contribution $> \frac{1}{n} \times 100.$
- Un individu ayant une forte contribution modifie l'analyse. On a donc intérêt à le porter en individu supplémentaire.

Exemple

Seuil = 100/21 = 4,76
Voir tableau ACP8
On remarque la très forte contribution de l'IdF à c^1 (79,9%)

Conclusion

L'ACP est une méthode statistique applicable :

- à un tableau individus / variables,
- pour p variables quantitatives,
- $p > 3$,
- certaines variables bien corrélées entre elles.

Remarque

Si $R = I_p$, alors l'ACP ne sert à rien !