Analyse en composantes principales : ACP

On s'intéresse à un jeu de données multidimensionel, avec n individus observés sur p variables d'intérêt. On se place dans le cas où les données sont quantitatives.

Les données se présentent sous forme d'une matrice X de taille $n \times p$ où $x_{ij} \in \mathbb{R}$ (données quantitatives) est la valeur observée pour le i-ème individu et la j-ème variable.

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

1 Objectif de l'ACP

Exemple. : Voici un jeu de données qui contient les notes de $n = 9$ élèves dans $p = 4$ matières. Cet exemple nous servira à illustrer l'ACP.

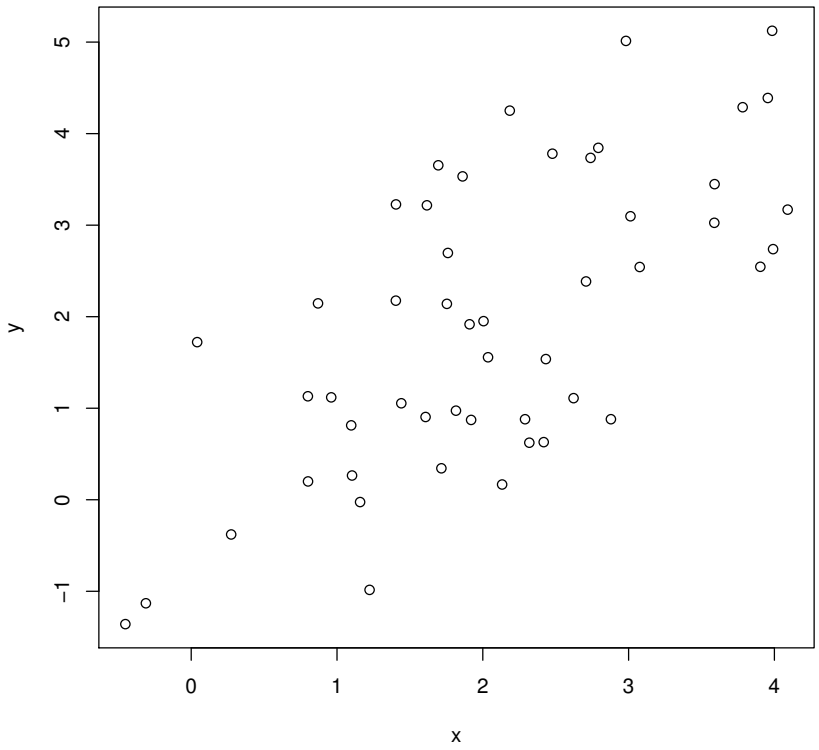
```
> X
     MATH PHYS FRAN ANGL
jean  6.0  6.0  5.0  5.5
alan  8.0  8.0  8.0  8.0
anni  6.0  7.0 11.0  9.5
moni 14.5 14.5 15.5 15.0
didi 14.0 14.0 12.0 12.5
andr 11.0 10.0  5.5  7.0
pier  5.5  7.0 14.0 11.5
brig 13.0 12.5  8.5  9.5
evel  9.0  9.5 12.5 12.0
```

On cherche à extraire de l'information de ce tableau de données. Par exemple on se demande quels individus se ressemblent ? Quelles variables sont liées ?
On peut pour cela faire de l'analyse descriptive univariée (en étudiant chaque variable), ou des statistiques descriptives bivariées (en étudiant les variables 2 à 2).

Lorqu'on étudie simultanément un grand nombre de variables quantitatives (ne serait-ce que 4 dans notre exemple), comment en faire un graphique global ? Quel graphique permet de "généraliser" le nuage de points tracé dans le cas de 2 variables ?
L'objectif de l'ACP est d'extraire l'essentiel de l'information contenue dans le tableau de données pour fournir une représentation se prêtant plus aisément à l'interprétation. Autrement dit l'ACP va fournir un espace de dimension réduite (idéalement le plan parce que c'est ce qui est le plus facilement lisible) qui donne le résumé le plus pertinent possible des données. Mathématiquement, l'ACP est un simple changement de base : passer d'une représentation dans la base des variables initiales à une représentation dans une autre base (on va définir pour cela de nouvelles variables qui seront des combinaisons linéaires des variables initiales et qui "résumeront" au mieux les variables initiales). L'interprétation des graphiques dans cette nouvelle base nous permettra d'obtenir de l'information sur le jeu de données.

2 La méthode ACP

2.1 Idée



Voici un nuage de points en dimension 2. On cherche à représenter ce nuage de points dans un espace de dimension 1, ie à projeter ce nuage sur une droite. Pour perdre le moins d'information possible, on va projeter dans la direction où le nuage est le plus étalé/dispersé. On va chercher la droite qui maximise la dispersion du nuage projeté. Cela revient à maximiser ce qu'on appelle l'inertie du nuage projeté.

2.2 Inertie

La matrice X permet de définir deux espaces : l'espace des individus et l'espace des variables.

— Chaque ligne de la matrice X correspond à la description d'un individu.

On note $x_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix} \in \mathbb{R}^p$ le vecteur decrivant le i-ème individu (=ligne i de X)

— Chaque colonne de la matrice X correspond à la description d'une variable.

On note $x^j = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix} \in \mathbb{R}^n$ le vecteur decrivant la j-ème variable (=colonne j de X)

L'ACP consiste à faire un changement de repère. Il nous faut donc définir une distance et une nouvelle origine. Nous détaillons ici le changement de repère pour le nuage des individus. Nous parlerons plus tard des variables.

Nuage des individus :

- Les n lignes x_i de X définissent un nuage de n points de $\mathbb{R}^p$=l'espace des individus.
- On désigne par G le centre de gravité du nuage :

$$G = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} x_{i1} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x_{ip} \end{pmatrix} = \begin{pmatrix} \bar{x}^1 \\ \vdots \\ \bar{x}^p \end{pmatrix} \in \mathbb{R}^p$$

  G est un point de $\mathbb{R}^p$ dont les coordonnées sont les moyennes empiriques des p variables. Le centre de gravité sera l'origine du nouveau repère.
- On munit l'espace des individus de la distance euclidienne classique :

$$d^2(x_i; x_{i'}) = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

- L'inertie mesure la dispersion du nuage de points. Elle est définie par

$$I_T = \frac{1}{n}\sum_{i=1}^{n} d^2(G; x_i)$$

  Remarque. : En général on prend $1/n$ pour donner le même point à chaque individu, mais on pourrait choisir une autre pondération.

Propriété 1. : L'inertie généralise la notion de variance dans le cas de données multidimentionnelles. En effet l'inertie est égale à la somme des variances des p variables, et donc égale à la trace de la matrice de variance-covariance empirique.

Preuve.

$$I_T = \frac{1}{n}\sum_{i=1}^{n} d^2(G; x_i) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} (x_{ij} - \bar{x}^j)^2 = \sum_{j=1}^{p}\frac{1}{n}\sum_{i=1}^{n} (x_{ij} - \bar{x}^j)^2 = \sum_{j=1}^{p} s_j^2 = Tr(S)$$

où s_j est l'écart-type empirique de la variables x^j, et S est la matrice de variance-covariance empirique(voir annexe).

ACP centrée ou ACP normée :

Prendre G comme origine revient à travailler sur le tableau des données centrées :

$$\begin{pmatrix} x_{11} - \bar{x}^1 & x_{12} - \bar{x}^2 & \dots & x_{1p} - \bar{x}^p \\ x_{21} - \bar{x}^1 & x_{22} - \bar{x}^2 & \dots & x_{2p} - \bar{x}^p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}^1 & x_{n2} - \bar{x}^2 & \dots & x_{np} - \bar{x}^p \end{pmatrix}$$

L'ACP opérera donc toujours (sans qu'on ait besoin de lui spécifier dans la commande sous R) sur les données centrées. Pour simplifier la suite du cours on considérera que le tableau de données X a été centré et donc que $\bar{x}^j = 0, \quad \forall j \in \{1, \dots, p\}$. L'inertie du nuage de points s'écrit alors :

$$I_T = \frac{1}{n}\sum_{i=1}^{n} ||x_i||^2$$

Lorsque les p variables sont hétérogènes (par exemple mesurées dans des unités différentes), on travaillera sur les données également réduites, pour traiter les variables de façon équitable (en

effet, une variable avec une variance forte va contribuer de façon importante à la création d'un axe). On travaille donc sur les données centrées réduites :

$$\frac{x_{ij} - \bar{x}^j}{s_j}$$

C'est ce que l'on appelle l'ACP normée. Dans ce cas, l'inertie du nuage de points vaut $I_T = p$ puisque l'inertie est la somme des variances de p variables et que ces variances sont égales à 1 (les données sont réduites !).

Dans la suite on considère les données centrées mais pas forcément réduites.

2.3 Moindre déformation du nuage de points

Idée : On recherche un sous-espace de $\mathbb{R}^p$ de dimension d ($<< p$) représentant au mieux le nuage de points des n individus. Autrement dit, tel que la représentation des individus dans ce sous-espace de dimension d fasse perdre le moins d'information possible. L'ACP va chercher un sous-espace qui maximise la dispersion ie l'inertie du nuage projetée sur ce sous-espace.

Principe de l'ACP : La méthode consiste à trouver un espace de dimension 1, i.e une droite passant par G telle que l'inergie du nuage de points projetée sur cette droite soit maximale. Puis à chercher une 2-ème droite perpendiculaire à la première telle que l'inertie non prise en compte par le 1er axe soit maximale sur ce 2-ème axe. Et ainsi de suite. Les axes sont perpendiculaires pour que l'information apportée par le 2-ème axe soit complètement nouvelle par rapport à l'information contenue dans le 1er axe.

Outils mathématiques :

- L'ACP s'appuie sur la diagonalisation de la matrice de variances-covariances empirique. La matrice de variance-covariance empirique S est un matrice réelle, symétrique, donc diagonalisable :

  $$S = P\Delta P^T = \sum_{j=1}^{p} \lambda_j u_j u_j^T$$

  où $\Delta = Diag(\lambda_1, \ldots, \lambda_p)$ est une matrice diagonale et les λ_i sont les valeurs propres de S. La matrice P est une matrice orthogonale dont les colonnes sont les vecteurs propres de S.

  Comme S est positive, alors les valeurs propres de S sont toutes ≥ 0. On range les valeurs propres par ordre décroissant :

  $$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$$

  et on suppose que les vecteurs propres u_j sont unitaires (quitte à poser $u_j = u_j/||u_j||$).
- On cherche une droite de vecteur directeur a (supposé unitaire) telle que l'inertie du nuage projeté sur cette droite I_{P_a} soit maximale.

  $$I_{P_a} = \frac{1}{n} \sum_{i=1}^{n} ||P_a(x_i)||^2$$

  La projection de $x_i \in \mathbb{R}^p$ sur la droite de vecteur directeur $a \in \mathbb{R}^p$ s'écrit :

  $$P_a(x_i) = < a, x_i > a$$

  où $< a, x_i >= a^T x_i = x_i^T a$ est le produit scalaire entre le vecteur a et le vecteur x_i

Ainsi

$$\begin{aligned} I_{P_a} = \frac{1}{n}\sum_{i=1}^{n} < a, x_i >^2 &= \frac{1}{n}\sum_{i=1}^{n} a^T x_i x_i^T a \\ &= a^T S a = a^T P \Delta P^T a = a^T \sum_{j=1}^{p} \lambda_j u_j u_j^T a \\ &= \sum_{j=1}^{p} \lambda_j a^T u_j u_j^T a = \sum_{j=1}^{p} \lambda_j < a, u_j >^2 \\ &= \lambda_1 < a, u_1 >^2 + \lambda_2 < a, u_2 >^2 + \dots \end{aligned}$$

Les vecteurs propres u_j forment une base orthonormale, et les valeurs propres λ_j sont rangées par ordre décroissant. I_{P_a} est donc maximale pour $a = u_1$, et l'inertie projeté vaut $I_{P_{u_1}} = \lambda_1$
Par conséquent la meilleur droite pour projeter le nuage est celle de vecteur directeur u_1 qui est le vecteur propre associé à la plus grande valeur propre de S.

- On cherche ensuite sur quel axe projeter pour maximiser l'inertie non prise en compte par le premier axe. On montre de la même façon qu'il faut projeter sur le vecteur u_2 qui est le vecteur propre associé à la 2-ème plus grande valeur propre de S, et que $I_{P_{u_2}} = \lambda_2$.
- Et ainsi de suite.

L'ACP en pratique :

— On calcule les valeurs propres et vecteurs propres de la matrice de variance-covariance empirique des données.
Remarque. L'ACP normée revient à travailler sur la matrice de corrélation empirique plutôt que sur la matrice de variance-covariances.

— On range les valeurs propres par ordre décroissant : u_1 est le vecteur propre associée à la plus grande valeur propre...

— Les droites engendrées par les vecteurs propres $u_1, u_2, \dots$ sont appelées 1er, 2-ème, ... axes principaux (ou axes propres).

— L'inertie expliquée par le 1er axe principal est égale à la valeur propre λ_1. L'inertie expliquée par le 2ème axe principal est égale à λ_2, et l'inertie expliquée par le 1er plan principal (1er axe-2-ème axe) est égale à $\lambda_1 + \lambda_2$.

— Et ainsi de suite. Le d-ème axe principal est la droite engendrée par le vecteur propre u_d. L'inertie du nuage projeté sur le d-ème axe principal vaut λ_d. Et l'inertie du nuage projeté sur le sous-espace $vect(u_1, \dots, u_d)$ vaut $\sum_{l=1}^{d} \lambda_l$.

2.4 Nombre d'axes à garder ?

D'après la propriété 1, l'inertie total du nuage de points vaut $I_T = Tr(S)$.

Propriété 2. On peut montrer que l'inertie se décompose aussi ainsi :

$$I_T = \lambda_1 + \lambda_2 + \dots + \lambda_p \tag{1}$$

Preuve. L'idée est la suivante :

— l'inertie du nuage I_T est en partie expliquée par le 1er axe. Donc $I_T = I_{P_{u_1}} + I_{P_{u_1}^{\perp}}$
où $P_{u_1}^{\perp}$ est l'inertie non expliquée par ce 1er axe, et ${u_1}^{\perp}$ est l'espace complémentaire orthogonal de la droite dirigée par u_1.

— De même, l'inertie non expliquée par le 1er axe, est en partie expliquée par le 2-ème axe et donc $I_T = I_{P_{u_1}} + I_{P_{u_2}} + I_{P_{u_2}^{\perp}}$
— Ainsi de suite : $I_T = I_{P_{u_1}} + I_{P_{u_2}} + \dots I_{P_{u_p}}$

Or l'inertie expliquée par le d-ème axe vaut $I_{P_{u_d}} = \lambda_d$.

Interprétation : L'inertie totale du nuage ne change pas lorsque l'on change de repère. C'est la répartition de cette dispersion qui est modifiée, chaque axe expliquant une partie de cette dispersion.

Définition. La part d'inertie expliquée par le d-ème axe principal est λ_d / I_T, et celle expliquée par le sous-espace $vect(u_1, \dots, u_d)$ vaut $\frac{\sum_{l=1}^{d} \lambda_l}{I_T}$.

On se contente en général d'un sous-espace de dimension d si ce sous espace explique un pourcentage d'inertie proche de 1. Idéalement, $d = 2$ ce qui permet d'interpréter des graphiques en dimension 2. Lorsque $d > 2$ s'intéresser aux deux premiers axes principaux permet déjà d'extraire de l'information du tableau de données.

```
Importance of components:
                          Comp.1    Comp.2       Comp.3       Comp.4
Standard deviation     5.3136495 3.4685078 0.1806433317 0.1029208059
Proportion of Variance 0.7004669 0.2984607 0.0008095538 0.0002627896
Cumulative Proportion  0.7004669 0.9989277 0.9997372104 1.0000000000
```

Ainsi le 1er axe explique 70%*, le 2ème axe* 29%*. Les 2 premiers axes restituent à eux seuls la quasi-totalité du nuage, ce qui permet de négliger les deux autres axes principaux. Par conséquent, des graphiques en dimension 2 (dans le 1er plan principal) résument presque parfaitement les données qui sont en dimension 4.*

Remarque. Lorsque les variables sont peu corrélées, les valeurs propres décroissent régulièrement et l'ACP présente alors peu d'intérêt. A l'inverse, lorsqu'il existe une structure sur les données, l'ACP arrive à extraire cette information et on observe des ruptures dans la décroissance des valeurs propres. (voir le graphe d'éboulis ou scree-graph en TP)

3 Représentations graphiques

3.1 Représentations et interpréations

Remarque. L'orientation des axes est complétement arbitraire : le signe des coordonnées des individus sur un axe n'as donc pas de signification. En revanche, la comparaison des signes est intéressante.

Interprétation du cercle de corrélation :
Le premier axe est corrélé positivement et assez fortement, avec chacune des 4 variables initiales : plus un élève obtient de bonnes notes dans chacune des 4 disciplines, plus son score est positif sur l'axe 1 ; réciproquement, plus ses notes sont mauvaises, plus son score est négatif.
En ce qui concerne l'axe 2, il oppose, d'une part, le français et l'anglais (corrélations positives), d'autre part, les mathématiques et la physique (corrélations négatives). Il s'agit donc d'un axe d'opposition entre disciplines littéraires et disciplines scientifiques.

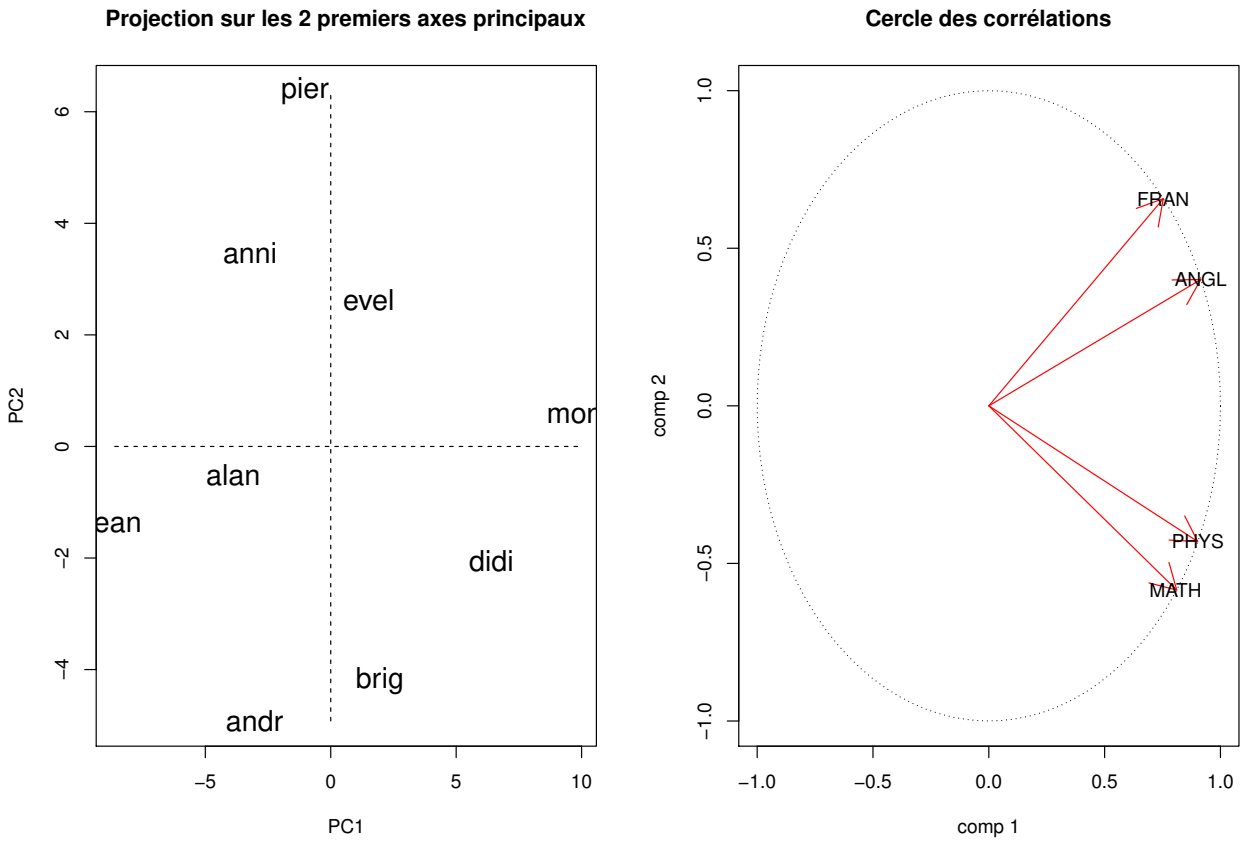


FIGURE 1 – à gauche : représentation des individus dans le 1er plan principal et à droite : cercle de corrélation des variables.

Interprétation des individus :
L'axe 1 représente le résultat des élèves sur l'ensemble des notes. Ainsi Monique a de bons résultats (nettement meilleur que la moyenne) partout, et Jean des résultats nettement moins bon que la moyenne partout.
L'axe 2 représente les différences entre matière scientifique et littéraire : Pierre a résultats les plus contrastés en faveur des disciplines littéraires ; pour André c'est l'inverse ; Monique et Alain ont un score voisin de 0 sur l'axe 2 car ils ont des résultats très homogènes dans les 4 disciplines (mais à des niveaux très distincts, ce qu'a déjà révélé l'axe 1).

3.2 Représentation des individus dans les nouveaux axes

Pour obtenir la représentation des individus dans les plans définis par les nouveaux axes, il suffit de caluler les coordonnées des individus dans les nouveaux axes, ie les projections des individus sur chaque axe.
La projection du i-ème individu sur le 1er axe principal est donné par : $P_{u_1}(x_i) =< u_1, x_i > u_1$

Définition. Le vecteur $C^1 = \begin{pmatrix} < u_1, x_1 > \\ \vdots \\ < u_1, x_n > \end{pmatrix} = Xu_1$ donne les coordonnées des n individus sur le 1er axe principal. Ce vecteur est appelé 1ère composante principale.

La d-ème composante principale $C^d = Xu_d$ donne les coordonnées des n individus sur le d-ème axe principal.

3.3 Qualité de la représentation d'un individu

Attention : deux individus dont les projections sont proches sur un axe peuvent ne pas être proches dans l'espace. Pour interpréter correctement la proximité des individus dans le plan projeté (ou un sous-espace de dimension supérieure), il faut s'assurer que les individus sont bien représentés dans ce plan (ou sous-espace).

Définition. La qualité de projection du i-ème individu sur le 1er axe principal est donnée par le carré du cosinus de l'angle

$$cos^2_{u_1}(x_i) = \frac{< x_i, u_1 >^2}{||x_i||^2||u_1||^2} = \frac{(C_i^1)^2}{||x_i||^2} \quad \text{car } u_1 \text{ est unitaire.}$$

Par orthogonalité des vecteurs u_d, la qualité de projection sur le plan formé par les vecteurs (u_1, u_2) correspond à la somme $cos^2_{u_1+u_2}(x_i) = cos^2_{u_1}(x_i) + cos^2_{u_2}(x_i)$. C'est d'ailleurs pour cette proprieté que l'on s'interesse au carré du cosinus plutôt qu'au cosinus.
Si pour un individu i, $cos^2_{u_1+u_2}(x_i)$ est grand (proche de 1) alors la représentation de l'individu dans le 1er plan principal est bonne. Au contraire si $cos^2_{u_1+u_2}(x_i)$ est proche de 0, alors la représentation de l'individu est mauvaise. Attention alors à l'interprétation d'une quelconque proximité de ce point avec un autre individu.

3.4 Contribution d'un individu à un axe

Les individus qui contribuent de manière excessive à la construction des axes sont source d'instabilité. Pour savoir si un individu contribue à un axe, on évalue la part de l'inertie expliquée par l'individu.
L'inertie projetée sur le 1er axe vaut : $I_{P_{u_1}} = \lambda_1$. Or

$$I_{P_{u_1}} = \frac{1}{n}\sum_{i=1}^{n} ||P_{u_1}(x_i)||^2 = \frac{1}{n}\sum_{i=1}^{n} < x_i, u_1 >^2 = \frac{1}{n}\sum_{i=1}^{n}(C_i^1)^2$$

Définition. La contribution relative d'un individu i à l'axe 1 vaut donc :

$$CTB_{u_1}(x_i) = \frac{(C_i^1)^2/n}{\lambda_1}$$

On parle en général de contribution excessive lorsqu'elle dépasse 25% de l'inertie. Mais cela dépend fortement du nombre d'individus n. Si n est très elevée (des milliers d'individus), une contribution de 5% sera considérée comme forte. Dans tous les cas, il faut faire attention aux points qui contribuent beaucoup à un axe ($CTB_{u_d}(x_i)$ forte) mais qui y sont mal représentés ($cos^2_{u_d}(x_i)$ faible).

4 Représentation des variables

4.1 Les composantes principales sont les "nouvelles" variables

Les composantes principales donnent les coordonnées des individus dans le nouveau repère. En fait, elles correspondent à de nouvelles variables "résumant" les anciennes variables. En effet, les composantes principales s'écrivent comme combinaison linéaire des anciennes variables :

Vérifions le pour la 1-ère composante principale :

$$C^1 = Xu_1 = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} u_1^1 \\ \vdots \\ u_1^p \end{pmatrix} = \begin{pmatrix} x_{11}u_1^1 + \dots + x_{1p}u_1^p \\ \vdots \\ x_{n1}u_1^1 + \dots + x_{np}u_1^p \end{pmatrix}$$
$$= x^1u_1^1 + \dots + x^pu_1^p$$

De même la d-ème composante principale s'écrit combinaison linéaire des p variables :

$$C^d = Xu_d = x^1u_d^1 + \dots + x^pu_d^p$$

Propriété 3. Les composantes principales sont de "nouvelles" variables centrées, de variance $var(C^d) = \lambda_d$, et non-corrélées 2 à 2.

4.2 Cercle de corrélations

On pourrait traiter individus et variables de façon symétrique et tracer les points-variables dans un nouveau repère en procédant de la même manière (maximiser l'inertie...). Mais ndividus et variables n'ont pas la même signification.
Les composantes principales correspondent à de nouvelles variables, et ce qui est intéressant, est de voir comment les anciennes variables sont liées aux nouvelles variables. Une variable x^j expliquera d'autant mieux un axe principal qu'elle sera fortement corrélée avec la composante principale correspondant à cet axe.
Pour cela on calcule les corrélations des anciennes variables x^j avec les nouvelles i.e avec les composantes principales C^d .
En traçant $(corr(C^1, x^j), corr(C^2, x^j))$ pour toutes les variables $(j = 1, \dots, p)$ on obtient ce qui s'appelle le cercle de corrélation.

4.3 Qualité de la représentation des variables

Une variable x^j sera bien représentée dans le nouveau repère si elle est proche du cercle de corrélation (corrélations proche de 1, ie pas à l'intérieur du cercle). Si ce n'est pas le cas, gare aux interprétations !
Sur le cercle de corrélation, on peut également interpréter les positions des variables x^j (si ces variables sont bien représentées) les unes par rapport aux autres en terme de corrélations (puisqu'on peut interpréter le cosinus d'un angle en terme de corrélation, voir Annexe) Si l'angle entre deux variables est proche de 90°, alors les variables sont tres faiblement correlees. Si l'angle entre deux variables est proche de 0° (respectivement 180°), alors les variables sont fortement corrélées positivement (respectivement négativement)

Remarque. se souvenir tout de même que l'on parle ici de corrélation linéaire.

5 Limites de l'ACP

Les coefficients de corrélation linéaire sont adaptés pour dérire un lien linéaire entre des variables aléatoires. C'est pourquoi l'ACP est souvent peu satisfaisante s'il existe des liaisons non linéaires entre variables (extension possible : ACP non-linéaire (à noyau))
Les corrélations empiriques ne sont pas fiables par rapport aux observations aberrantes, et donc les résultats de l'ACP non plus.

6 Annexe

6.1 Matrice de covariance et de corrélation empirique

Soient p variables quantitatives $x^1, \dots, x^p \in \mathbb{R}^n$.
On définit respectivement la moyenne empirique et la variance empirique de la variable x^j par

$$\bar{x}^j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \qquad s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}^j)^2$$

On définit respectivement la covariance empirique entre les variables x^j et $x^{j'}$ et la corrélation linéaire empirique par

$$s_{jj'} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}^j)(x_{ij'} - \bar{x}^{j'}) \qquad r_{jj'} = \frac{s_{jj'}}{\sqrt{s_j^2 s_{j'}^2}}$$

La matrice de variance-covariance des p variables est définie par :

$$S = \begin{pmatrix} s_1^2 & s_{12} & \dots & s_{1p} \\ s_{21} & s_2^2 & \dots & s_{2p} \\ \vdots & & & \\ s_{p1} & s_{p2} & \dots & s_p^2 \end{pmatrix}$$

Remarque. Si les variables sont réduites, alors S est aussi la matrice de corrélation.

On suppose pour simplifier que les variables x^j ont été centrées, ie $\bar{x}^j = 0, \quad \forall j$, ce qui simplifie les définitions du dessus (sinon remplacer X par le tableau des données centrées).

Propriété 4. On peut montrer (par simple écriture matricielle) que

$$S = \frac{1}{n}X^T X = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$$

où X^T est la transposé de la matrice X et x_i^T est la transposé de x_i (x_i est le vecteur décrivant le i-ème individu)

Propriété 5. S est une matrice positive ie

$$\forall y \in \mathbb{R}^p, \quad y^T S y = \frac{1}{n} y^T X^T X y = \frac{1}{n}||Xy||^2 \geq 0$$

6.2 Correspondance entre statistique et géométrie

Soient p variables quantitatives $x^1, \dots, x^p \in \mathbb{R}^n$. Pour simplifier, on suppose à nouveau que ces variables sont centrées.
Les méthodes type ACP (dites méthode factorielle) et leur représentations utilisent la correspondance entre statistiques empiriques et géométrie. On a équivalence entre variance empirique et norme, covariance empirique et produit scalaire, corrélation empirique et cosinus.

$$n\, s_j^2 = \sum_{i=1}^{n} x_{ij}^2 = ||x^j||^2$$

$$n\, s_{jj'} = \sum_{i=1}^{n} x_{ij} x_{ij'} = < x^j, x^{j'} >$$

$$n\, r_{jj'} = \frac{< x^j, x^{j'} >}{||x^j||||x^{j'}||} = cos(x^j, x^{j'})$$

en voyant la variable x^j comme un point dans le nuage de points de l'espace des variables .