# Analyse en Composantes Principales
# ACP ou PCA

Prof. Mustapha RACHDI

Université Grenoble Alpes
UFR SHS, BP. 47
38040 Grenoble cedex 09
France
Bureau : C08 au Bât. Michel Dubois
e-mail : mustapha.rachdi@univ-grenoble-alpes.fr

# Introduction

L'Analyse en Composantes Principales (ACP) est sans doute la méthode d'analyse de données la plus connue et la plus utilisée. Proposée dès les années 30 par Hotelling (1933), mais nécessitant d'importants calculs numériques, l'ACP n'est devenue une technique operationnelle qu'à partir des années 60, avec le développement des moyens de calcul informatique.

L'ACP a pour objet de résumer de grands ensembles de données quantitatives. Ces données sont rangées dans un tableau comportant un grand nombre d'individus et/ou un grand nombre de variables et la simple lecture de ce tableau ne permet pas de saisir l'essentiel des informations qu'il contient.

l'ACP synthétise-t-elle les données en construisant un petit nombre de variables nouvelles : les composantes principales. L'essentiel du tableau de données peut alors être saisi rapidement, à l'aide de représentations graphiques établies à partir de ces composantes principales.

# Application 1 :

Si l'on dispose pour chacune des agences d'un réseau bancaire du chiffre d'affaire réalisé pour différents produits (les caractères : livret d'épargne, prêts accordés à court, à moyen ou à long terme, produit boursiers, ...), ou encore d'autres variables comme l'avoir moyen par livret ou le nombre de clients de l'agence, l'ACP répond aux questions suivantes :

1. Existe-t-il des agences ou des groupes d'agences ayant des comportements *atypiques* ? (quelles sont ces agences ? Pour quel type de produit sont-elles *atypiques* ? peut-on classer les agences en groupes dans lesquels les comportements sont homogènes ? ...)
2. Quelles relations existe-t-il entre les différentes variables ? (par exemple : les agences pour lesquelles les avoirs moyens par livret sont élevés sont-elles aussi celles qui drainent une épargne boursière importante ? Cette relation entre avoirs moyens et épargne est-elle vérifiée pour l'ensemble du réseau ?...)

# Application 2 :

En sciences expérimentales, on dispose souvent d'un ensemble de mesures pour une population donnée. Là encore, l'ACP décrit de façon synthétique les individus

1. s'agit-il d'une population homogène ?
2. peut-on distinguer des sous-groupes ?
3. existe-t-il des individus au comportement original ?

et décrit de façon synthétique les relations entre les variables

1. varient-elles de façon concomitante ?
2. certaines variables sont-elles sans lien avec les autres variables ?
3. ...

# Application 3 :

l'évolution des 8 caractéristiques de l'économie française de 1971 à 1994 est synthétisée par 3 composantes principales.

Les proximités entre les années et les corrélations entre les variables sont alors décrites par des graphiques basés sur ces 3 composantes principales. Il s'agit donc, à partir des 3 axes, de résumer l'évolution de la conjoncture économique française de 1971 à 1994, en mettant en évidence :

- les relations entre les variables (quelle relation entre inflation et chômage, entre taux d'intérêt et investissement ?),

et en décrivant

- le cheminement temporel de cette conjoncture (existe-t-il des années de rupture ? pour quelles variables existe-t-il une rupture ces années-là ? après cette rupture, l'économie revient-elle à l'état précédent ? ...).

# Données

Soit $X$ un tableau à $n$ lignes et $p$ colonnes. La ligne $i$ décrit la valeur prise par l'individu $i$ pour $p$ variables quantitatives. Les données sont centrées et réduites, c'est-à-dire que chaque variable a une moyenne nulle et une variance égale à 1.

$$X = \begin{pmatrix} x_1^1 & \dots & x_1^j & \dots & x_1^p \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ x_i^1 & \dots & x_i^j & \dots & x_i^p \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ x_n^1 & \dots & x_n^j & \dots & x_n^p \end{pmatrix} = (X_1, \dots, X_j, \dots, X_p)$$

On note $X_j$ le vecteur colonne constitué par les éléments de la colonne $j$ de $X$. La valeur $x_i^j$ désigne l'élément de $X$ situé à l'intersection de la ligne $i$ et de la colonne $j$, c'est-à-dire la valeur de l'individu $i$ pour la variable $X_j$.

# Position du problème

Il s'agit de synthétiser les données contenues dans le tableau $X$. Pour cela, on construit un petit nombre de variables nouvelles ($C^1, C^2, \ldots$) appelés composantes principales, permettant de saisir l'essentiel du tableau $X$.

Ainsi, à l'étape 1, on détermine une variable synthétique $C^1$, la première composante principale, combinaison linéaire des variables $X_j$ :

$$C^1 = a_1^1 X_1 + \cdots + a_j^1 X_j + \cdots + a_p^1 X_p$$

ce qui signifie que la valeur de $C^1$ pour l'individu $i$ est donnée par :

$$C_i^1 = a_1^1 x_i^1 + \cdots + a_j^1 x_i^j + \cdots + a_p^1 x_i^p.$$

Cette première composante principale ne suffit généralement pas à résumer de façon satisfaisante les données du tableau $X$. On construit aussi une 2ème composante principale, puis une 3ème ...
De façon générale, à l'étape $k$, on construit la composante d'ordre $k$ :

$$C^k = a_1^k X_1 + \cdots + a_j^k X_j + \cdots + a_p^k X_p.$$

# Position du problème

Matriciellement,

$$C^k = X\, a^k$$

où $a^k$ est un vecteur-colonne à $p$ éléments, l'élément d'ordre $j$ étant égal à $a_j^k$. Ce vecteur $a^k$ est appelé facteur d'ordre $k$ (ou $k$ème facteur).

Les facteurs fournissent un système de *poids* pour les variables : certains *poids* $a_j^k$ sont négatifs, d'autres positifs. En fait, ce qui importe n'est pas la valeur de chacun de ces poids, mais le rapport de ces *poids* les uns par rapport aux autres. Si on multiplie un facteur par une constante non nulle, ces rapports sont inchangés. Les facteurs sont donc définis à une constante multiplicative près. On impose aussi une contrainte de normalisation :

$$\sum_{j=1}^{p} (a_j^k)^2 = 1 = \left\| a^k \right\|^2 .$$

Les composantes principales sont des variables de moyenne nulle, puisque les variables d'origine sont centrées. La valeur pour l'individu $i$ de la composante principale $k$ est :

$$C_i^k = a_1^k x_i^1 + \cdots + a_j^k x_i^j + \cdots + a_p^k x_i^p .$$

# Position du problème

Comme pour la régression linéaire, l'analyse en composantes principales (ACP) peut être représentée dans deux espaces :

- des individus
- des variables

# Dans l'espace des individus

Dans cet espace, les $n$ individus forment un nuage de points. Les variables étant centrées, l'origine $O$ du repère est le centre du nuage. La distance utilisée ici est la distance usuelle dans $\mathbb{R}^p$.
L'objet de l'ACP est de décrire de façon synthétique la dispersion du nuage de points.

**Proposition**

*A l'étape* 1*, l'ACP détermine l'axe* $D_1$ *passant par l'origine selon lequel la dispersion du nuage de points est maximale. Cet axe* $D_1$ *passe au* plus près *du nuage de points, c'est-à-dire est tel que la moyenne des carrés des distances entre les n points et l'axe* $D_1$ *est minimale.*

*Soit* $a^1$ *le vecteur directeur normé de* $D_1$*. Le vecteur* $a^1$ *est alors le vecteur propre normé associé à la valeur propre la plus élevée de la matrice de corrélations entre les variables* $\frac{1}{n}({}^tX\,X)$*.*

# Dans l'espace des individus (suite)

Ainsi, à l'étape 1, l'ACP fournie la meilleure représentation unidimensionnelle possible du nuage de points. Mais l'étape 1 ne suffit pas à décrire complètement le nuage des $n$ points : la dispersion du nuage dans les directions de l'espace orthogonal à $D_1$ n'est pas décrite par l'étape 1.

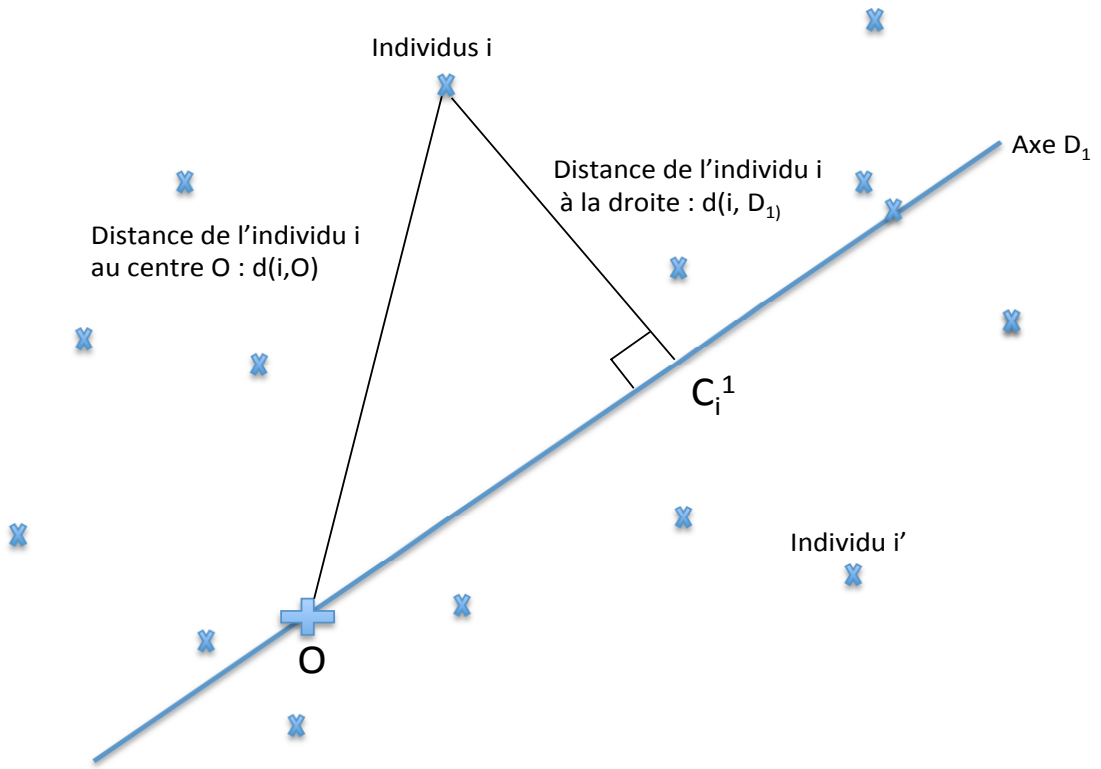


Figure – Géométriquement, projection d'un individu sur l'axe $D_1$.

# Dans l'espace des individus (suite)

- Aussi, à l'étape 2, l'ACP détermine un axe $D_2$, de vecteur directeur normé $a^2$, orthogonal à $a^1$, passant *au plus près* du nuage de points. En procédant comme à l'étape 1, le problème s'écrit alors :

$$\max \left( \frac{1}{n} {}^t(a^2)\, {}^tX\, X a^2 \right)$$

  sous la contrainte d'orthogonalisation : ${}^t(a^2)\, a^1 = 0$, et de normalisation : ${}^t(a^2)\, a^2 = 1$.

- La solution, le vecteur $a^2$, est le vecteur propre normé de $\frac{1}{n}({}^tX\, X)$ associé à sa deuxième valeur propre.
- On continue ainsi de suite, afin de compléter la description du nuage de points donnés par les deux premières étapes.

# Dans l'espace des individus (suite)

## Proposition

- *A l'étape k, l'ACP détermine l'axe $D_k$, passant par l'origine, de vecteur directeur normé $a^k$ orthogonal aux axes $a^r$ $(r < k)$ des étapes précédentes, selon lequel la dispersion du nuage de points est maximale. Cet axe $D_k$ passe au* plus près *du nuage de points, c'est-à-dire est tel que la moyenne des carrés des distances entre les n points et l'axe $D_k$ est minimale.*
- *Le problème s'écrit :*

$$\max \left( \frac{1}{n} {}^t(a^k)\, {}^tX\, X a^k \right)$$

*sous la contrainte d'orthogonalisation : ${}^t(a^k)\, a^r = 0$ pour $r = 1, \ldots, k-1$, et sous la contrainte de normalisation : ${}^t(a^k)\, a^k = 1$.*

- *Le maximum est atteint lorsque $a^k$ est le vecteur propre normé de $\frac{1}{n}({}^tX\, X)$ associé à sa kème valeur propre.*

# Dans l'espace des variabless

La présentation alternative de l'ACP dans l'espace des variables est la suivante : disposant d'un ensemble de $p$ variables, l'ACP construit une variable synthétique résumant le mieux possible ces $p$ variables, c'est-à-dire une variable synthétique la mieux liée linéairement possible aux $p$ variables.

## Proposition

*A l'étape 1, l'ACP détermine $C^1$ tel que $\sum_{j=1}^{p} R^2(C^1, X_j)$ ait une valeur maximale. Le vecteur $C^1$ est par conséquent le vecteur propre de $\frac{1}{n}(X\,{}^tX)$ associé à sa valeur propre la plus élevée.*

## Remarque

*$C^1$ est donc le meilleur résumé possible de l'ensemble des variables de départ, mais $C^1$ ne décrit pas totalement ces variables et il est nécessaire de calculer d'autres composantes principales.*

# Dans l'espace des variables

Aussi, à l'étape 2, l'ACP détermine $C^2$ qui doit être aussi le meilleur résumé possible des variables $X_j$, mais en complétant les informations fournies par $C^1$ : ceci signifie que $C^2$ doit être non corrélé à $C^1$.

Le problème à l'étape 2 est donc de déterminer $C^2$ tel que $\sum_{j=1}^{p} R^2(C^2, X_j)$ ait une valeur maximale sous la contrainte : $R(C^1, C^2) = 0$.

On continue ainsi de suite et :

**Proposition**

*A l'étape k, l'ACP détermine $C^k$, résumant le mieux possible les variables de départ, et non corrélée aux $(k-1)$ premières composantes principales, c'est-à-dire détermine $C^k$ tel que*

$$\sum_{j=1}^{p} R^2(C^k, X_j) \text{ ait une valeur maximale}$$

*sous les $(k-1)$ contraintes : $R(C^k, C^r) = 0$, pour $r < k$.*

*$C^k$ est alors le vecteur propre de $n^{-1}(X\,{}^tX)$ associé à sa kème valeur propre la plus élevée.*

# Equivalence entre les deux analyses

- Dans l'espace des individus, à l'étape $k$, la solution est le facteur $a^k$, vecteur propre d'ordre $k$ de la matrice $\frac{1}{n}({}^tX\,X)$.
- Dans l'espace des variables, la solution est la composante principale $C^k$ : $k$ème vecteur propre de $\frac{1}{n}(X\,{}^tX)$.
- Il reste à prouver que ces deux solutions sont équivalentes, c'est-à-dire à prouver que les deux approches conduisent aux mêmes résultats.
- Si on note $\beta_k$ la valeur propre d'ordre $k$ de $\frac{1}{n}({}^tX\,X)$, alors $\beta_k$ est aussi la $k$ème valeur propre de $\frac{1}{n}(X\,{}^tX)$, et les deux approches conduisent bien aux mêmes résultats.

# Nombre maximal d'étapes

- Dans l'espace des individus, le nombre maximum d'étapes est $p$, puisqu'après $p$ étapes, on a obtenu une nouvelle base de l'espace des individus : il n'est pas possible de déterminer un vecteur orthogonal aux $p$ vecteurs $u^1, \ldots, u^p$.

- De façon symétrique, le nombre maximal d'étapes dans l'espace des variables est $n$.

- Par conséquent, le nombre maximal d'étapes est le plus petit des nombres $n$ et $p$, soit $\min(n, p)$.

- Généralement ! ! !, $n$ **est supérieur à** $p$ : il y a plus d'observations que de variables, et c'est ce que nous allons supposer ici, pour ne pas alourdir inutilement la notation.

- Le cas où $n$ est inférieur à $p$ s'obtient facilement en considérant alors que le rang de $X$ est $n$ dans la suite de l'exposé. Mais, il y a l'UE "Statistique en Grande dimension" qui étudie rigoureusement cette problématique

# Les valeurs propres

On déduit :

**Proposition**

*$\frac{1}{n}({}^tX\,X)$ et $\frac{1}{n}(X\,{}^tX)$ ont les mêmes valeurs propres $\beta_1, \ldots, \beta_p$. Comme $\frac{1}{n}(X\,{}^tX)$ est une matrice de dimension $n \times n$ qui a le même rang que $\frac{1}{n}({}^tX\,X)$, ses $(n-p)$ dernières valeurs propres sont nulles.*

Deux autres résultats concernant les valeurs propres sont à noter :

**Proposition**

*Si s valeurs propres $\beta_{p-s+1}, \ldots, \beta_p$ sont nulles, alors il existe s relations linéaires exactes entre les variables.*

**Proposition**

$$\sum_{k=1}^{p} \beta_k = \mathit{Trace}\left(\frac{1}{n}({}^tX\,X)\right) = p.$$

# Part de variance expliquée par un axe

- L'ACP comporte $p$ étapes. Chaque étape fournie un *résumé* du tableau $X$ moins intéressant que celui obtenu à l'étape précédente. Comment mesurer la *qualité* du résumé obtenu à l'étape $k$ ?
- A l'étape $k$, le critère de l'ACP dans l'espace des individus est la maximisation sous contrainte de

$$\frac{1}{n}\sum_{i=1}^{n}(C_i^k)^2.$$

- A l'optimum

$$\frac{1}{n}\sum_{i=1}^{n}(C_i^k)^2 = \frac{1}{n}\,{}^t(a^k)\,{}^tX\,X\,a^k = {}^t(a^k)\,\beta_k\,a^k = \beta_k,$$

car $a^k$ est un vecteur normé. Et, $\beta_k$ est donc la variance du nuage expliquée par l'axe $k$, c'est-à-dire la variance de $C^k$.

- Si on considère les $p$ axes déterminés par l'ACP,

$$\sum_{k=1}^{p}\beta_k = p$$

# Part de variance expliquée par un axe

## Proposition

*La part de variance expliquée par l'axe $k$ est égale à* $\frac{\beta_k}{p}$.

## Remarque

1. *Souvent en analyse de données, le terme* inertie *est utilisé pour désigner la variance et on parlera donc de pourcentage d'inertie expliqué par un axe au lieu de part de variance expliquée par un axe.*
2. *De façon symétrique, dans l'espace des variables, on remarque que*

$$\sum_{j=1}^{p} R^2(C^k, X_j) = \frac{{}^t(C^k)\, X\, {}^t X\, C^k}{n\, {}^t(C^k)\, C^k} = \beta_k$$

*et* $\frac{\beta_k}{p}$ *est aussi la moyenne des carrés des corrélations entre les variables d'origine et l'axe $k$.*

# Part de variance expliquée par un axe

**Remarque**

*Si $\beta_k$ a une valeur faible, cela signifie que la variance expliquée par l'axe k est faible. Par conséquent, l'information apportée par l'étape k est peu utile pour résumer les données de départ.*

*La variance expliquée étant décroissante pour des étapes successives, l'information apportée aux étapes d'ordre supérieur à k est, elle aussi, peu utile.*

La conséquence de cette remarque est :

**Proposition**

*Pour résumer le tableau de départ, on n'utilisera donc que les résultats des premières étapes, qui correspondent à des variances expliquées importantes.*

# Graphiques et aides à l'interprétation

- L'information pertinente est donc celle donnée par les premières étapes. Il s'agit maintenant d'analyser simultanément les résultats de ces premières étapes, dans l'espace des individus et dans l'espace des variables.

- Cette analyse se fait en établissant des cartes des proximités entre les individus et des cartes des corrélations entre les variables. Ces cartes sont obtenues à partir des résultats des premières étapes, en considérant les projections sur des plans définis par des axes obtenus à deux étapes distinctes.

- Par exemple, si l'on retient les 4 premières étapes, on s'intéressera aux projections du nuage des individus sur le plan défini par les 2 premiers facteurs, ou sur le plan défini par les facteurs 1 et 4, ou encore 2 et 3, ...

- De la même manière les projections des variables sur les plans des composantes principales 1 et 2, ou 1 et 4, ou 2 et 3, ... aideront à la compréhension des phénomènes étudiés.

# Représentation des individus

Pour analyser les résultats fournis par 2 étapes $r$ et $s$, la représentation des individus est obtenue en projetant chaque individu $i$ sur le plan engendré par les facteurs $a^r$ et $a^s$ : on obtient ainsi une description du nuage de points en projection sur le plan défini par les vecteurs $a^r$ et $a^s$.

Les coordonnées de l'individu $i$ sont égales à $C_i^r$ pour l'axe $r$ et $C_i^s$ pour l'axe $s$.

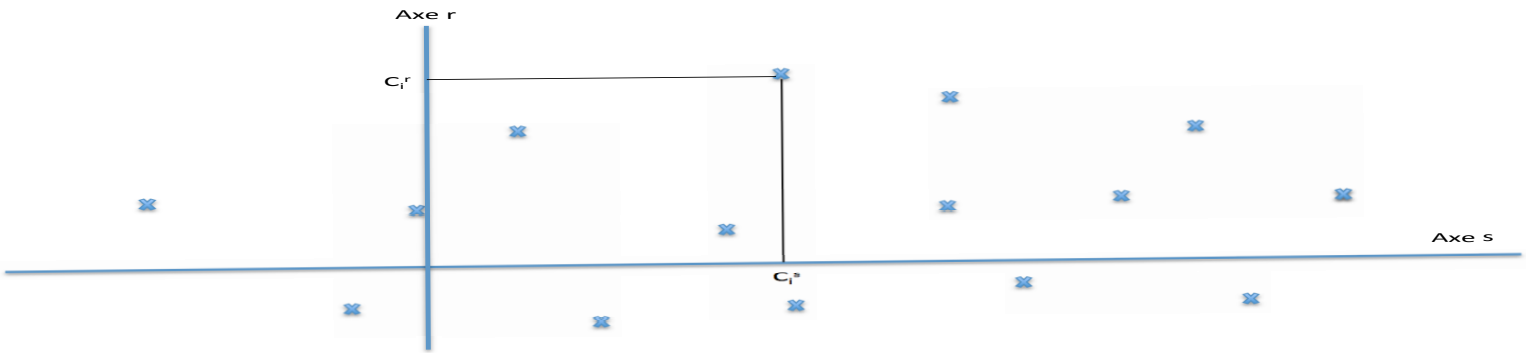


Figure – Représentation d'un seul individu

L'interprétation des résultats consiste à déterminer les caractéristiques du nuage de points décrites par les composantes principales, c'est-à-dire à repérer quelles particularités des individus sont mises en évidence à chaque étape.

# Représentation des individus

Lors de l'interprétation, il faut garder constamment à l'esprit que l'on n'est pas en présence du nuage initial, mais d'une projection de ce nuage et que des erreurs de perspectives peuvent être commises.

Ainsi les projections des individus *i* et *j* sont proches alors que les individus sont très éloignés dans l'espace.

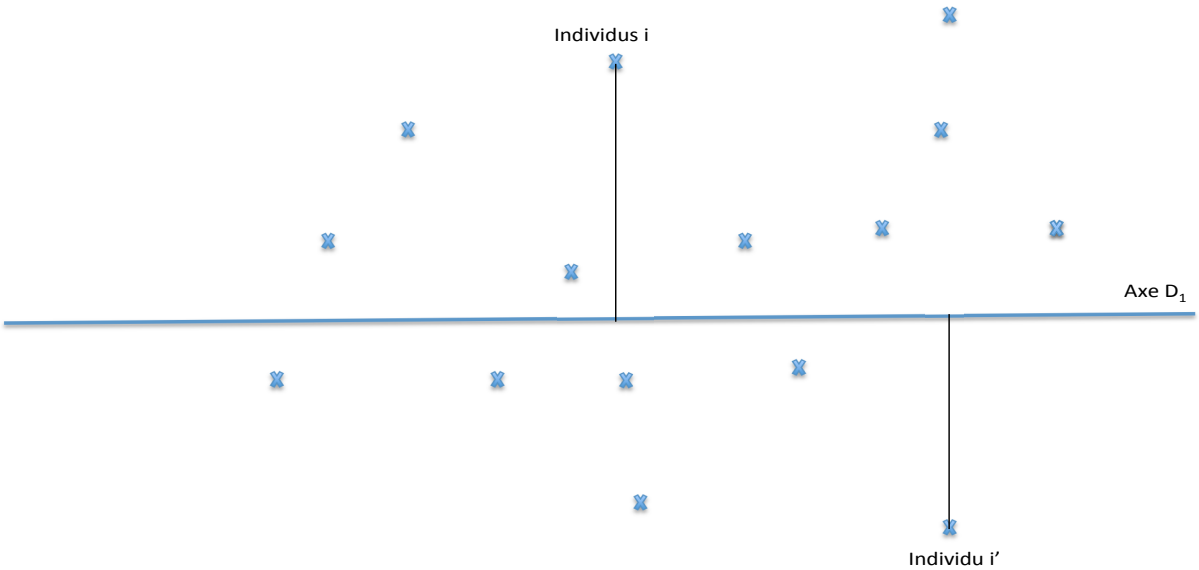


Figure – Les proximiés sur les axes

La lecture des graphiques représentant les individus est facilitée par le calcul de deux aides à l'interprétation : la qualité de représentation d'un point et les contributions des individus à la variance.

# Qualité de représentation d'un point

Dans l'espace des individus, on dispose de deux bases orthonormées :

- la base d'origine. Dans cette base, les coordonnées de l'individu $i$ sont : $(x_i^1, \ldots, x_i^j, \ldots, x_i^p)$.
- la base constituée par les $p$ facteurs. Dans cette base, les coordonnées de l'individu $i$ sont : $(C_i^1, \ldots, C_i^k, \ldots, C_i^p)$.

Le carré de la distance de l'individu $i$ au centre du nuage est égal à :

- $\sum_{j=1}^{p} (x_i^j)^2$, en effectuant les calculs avec la première base.
- $\sum_{k=1}^{p} (C_i^k)^2$, en effectuant les calculs avec la seconde base.

Donc, $\sum_{j=1}^{p} (x_i^j)^2 = \sum_{k=1}^{p} (C_i^k)^2$ : un individu est bien représenté sur un axe $r$ si $(C_i^r)^2$ a une valeur importante par rapport à $\sum_{j=1}^{p} (x_i^j)^2$.

# Qualité de représentation d'un point

## Proposition

*La qualité de représentation de l'individu $i$ sur l'axe $r$ est donc mesurée par*

$$\frac{(C_i^r)^2}{\sum_{j=1}^{p}(x_i^j)^2}$$

## Remarque

1. *Cette qualité de représentation est égale au carré du cosinus de l'angle entre le vecteur représentatif du point $i$ et le vecteur $a^r$ .*
2. *Pour un individu donné, la somme de ses qualités de représentation pour les $p$ axes est égale à* 100%.
3. *Un individu est* bien *expliqué par un axe pour lequel sa qualité de représentation est élevée.*
4. *Souvent, pour un axe $r$ donné, la qualité de représentation d'un point est faible lorsque la projection de ce point est proche de l'origine ($C_i^r$ a alors une valeur faible). Cette qualité de représentation est souvent forte lorsque la projection est éloignée de l'origine ($C_i^r$ a alors une valeur forte).*

# Contributions des individus à la variance

La variance expliquée à l'étape $r$ est égale à

$$\frac{1}{n}\sum_{i=1}^{n}(C_i^r)^2 = \beta_r.$$

La part de cette variance due à l'individu $i$ est $\frac{1}{n\beta_r}(C_i^r)^2$.

**Proposition**

*La contribution de l'individu $i$ à la variance de l'axe $r$ est donc mesurée par*

$$\frac{(C_i^r)^2}{n\beta_r}$$

1. Pour un axe donné, la somme des contributions de tous les individus est égale à 100%.
2. Si la contribution d'un individu à un axe donné est importante, ceci signifie que cet individu joue un rôle important dans la construction de cet axe, rôle qu'il convient d'analyser lors de l'interprétation des résultats.

# Représentation des variables

- Lorsqu'on s'intéresse aux résultats fournis par les étapes $r$ et $s$, on représente chaque variable par sa projection sur le plan défini par les composantes principales d'ordre $r$ et $s$ normées à 1.
- La projection de $X_j$ sur l'axe engendré par la composante principale d'ordre $r$ normé à 1 est donnée par

$$\frac{\frac{1}{n}\,{}^t(X_j)\,C^r}{\sigma(C^r)} = \frac{cov(X_j, C^r)}{\sigma(C^r)} = R(X_j, C^r)$$

  où $\sigma(C^r)$ désigne l'écart-type de $C^r$.
- La coordonnée de la variable $X_j$ pour l'axe $r$ est donc $R(X_j, C^r)$.
- Ainsi, les coordonnées de $X_j$ dans la base des composantes principales normées sont : $\left(R(X_j, C^1), \ldots, R(X_j, C^p)\right)$.

Il s'ensuit, puisque $X_j$ est un caractère normé que :

**Proposition**

$$\sum_{k=1}^{p} R^2(X_j, C^k) = 1.$$

# Représentation des variables

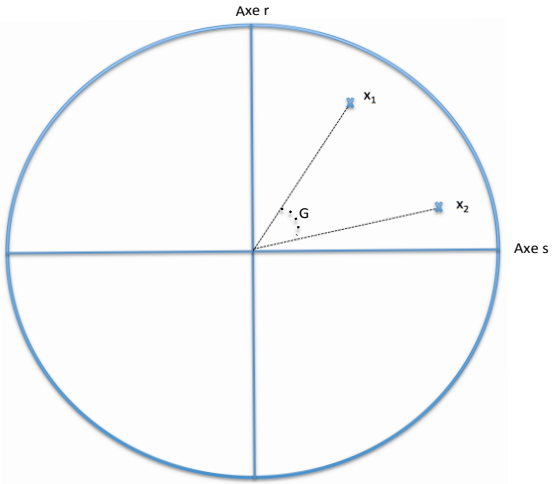


Figure – Angle entre deux variables dans le cercle des corrélations

**Proposition**

*Si deux variables $X_j$ et $X_{j'}$ sont proches du bord du cercle, alors l'angle au centre G entre ces deux variables est proche de l'angle que font ces variables entre elles dans l'espace des variables, et le cosinus de cet angle G est approximativement égal à leur coefficient de corrélation.*

# Représentation des variables

## Remarque

- *Si on ne considère que 2 composantes principales $C^r$ et $C^s$*

$$R^2(X_j, C^r) + R^2(X_j, C^s) \leq 1.$$

*La somme des carrés des coordonnées de la variable $X_j$ est inférieure ou égale à 1, c'est-à-dire que le point représentatif de $X_j$ est situé à l'intérieur d'un cercle de rayon 1 et de centre l'origine.*

- *Si le point $X_j$ est proche du bord du cercle, ceci signifie que la variable $X_j$ est très proche du plan défini par les composantes principales r et s, puisque les coefficients de corrélations avec les autres composantes principales sont alors très faibles.*

## Proposition

*Si deux variables $X_j$ et $X_{j'}$ sont proches du bord du cercle, alors l'angle au centre G entre ces deux variables est proche de l'angle que font ces variables entre elles dans l'espace des variables, et le cosinus de cet angle G est approximativement égal à leur coefficient de corrélation.*

# Représentation des variables

## Remarque

- *Soient deux variables proches du bord du cercle.*
  - *Si elles sont proches l'une de l'autre, alors elles sont très fortement corrélées.*
  - *Au contraire, si l'angle au centre est un angle droit, ces variables ont une corrélation nulle.*
  - *Enfin, si ces deux variables sont opposées par rapport à l'origine, elles ont une corrélation proche de $-1$.*
- *La représentation des variables permet ainsi de saisir les relations linéaires existant entre ces variables. En d'autres termes, les cercles de corrélations des axes principaux décrivent l'essentiel de la matrice des corrélations entre les variables.*

# Calcul des coefficients de corrélation

Les coefficients de corrélation entre les variables de départ et les composantes principales sont calculés facilement :

**Proposition**

*Le vecteur constitué par les corrélations des p variables $X_j$ avec la composante principale $C^r$ est égal au facteur d'ordre $a^r$ multiplié par $\sqrt{\beta_r}$ :*

$$a^r \sqrt{\beta_r}$$

# Exemple d'application

- Dans le but d'illustrer les calculs effectués par l'ACP, de permettre une meilleure compréhension de la façon dont les résultats sont obtenus, et donc une meilleure compréhension des règles d'interprétation de ces résultats, voici, donc, un exemple numérique simple.
- Considérons le tableau de données suivant :

| Individu | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $Z_1$ | 1.00 | 2.00 | 3.00 | 4.00 | 9.00 |
| $Z_2$ | 5.00 | 10.00 | 8.00 | 8.00 | 12.00 |

- Il est évident que, comme les deux variables $Z_1$ et $Z_2$ forment un plan dans l'espace des individus, l'ACP comportera 2 étapes seulement. La 1ère étape consiste à calculer les moyennes et les écart-types de chacune des deux variables ; soit $\bar{Z}_1 = 3.8$, $\bar{Z}_2 = 8.6$, $\sigma_{Z_1} = 2.79$ et $\sigma_{Z_2} = 2.33$. Les variables centrées-réduites sont donc :

| Individu | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $X_1$ | -1.005 | -0.0646 | -0.0287 | 0.072 | 1.867 |
| $X_2$ | -1.543 | -0.600 | -0.257 | -0.257 | 1.458 |

# Exemple d'application

- La matrice des corrélations entre les variabels s'écrit :

$$\frac{1}{n}\,{}^{t}X\,X = \begin{pmatrix} 1 & 0.788 \\ 0.788 & 1 \end{pmatrix}$$

- Les facteurs principaux sont les vecteurs propres normés de cette matrice de corrélations. Le premier facteur, associé à la valeur propre 1.788 est $\begin{pmatrix} 0.707 \\ 0.707 \end{pmatrix}$, et le 2ème facteur est $\begin{pmatrix} 0.707 \\ -0.707 \end{pmatrix}$, correspondant à la valeur propre 0.212
- Les composantes principales sont donc :

$$C^1 = 0.707\,X_1 + 0.707\,X_2 = Xa^1$$

et

$$C^2 = 0.707\,X_1 - 0.707\,X_2 = Xa^2$$

- Les pourcentages de variance expliquée sont :

$$\text{1er axe : } 1.788/2 = 89.4\% \text{ et, 2ème axe : } 0.212/2 = 10.6\%$$

# Exemple d'application

- Dans le but de tracer les cercles de corrélation, on calcule les corrélations entre les composantes principales et les variables d'origine : on les obtient à partir des facteurs et des vecteurs propres.

$$\begin{pmatrix} R(C^1, X_1) \\ R(C^1, X_2) \end{pmatrix} = \sqrt{1.788} \begin{pmatrix} 0.707 \\ 0.707 \end{pmatrix} = \begin{pmatrix} 0.946 \\ 0.946 \end{pmatrix}$$

et

$$\begin{pmatrix} R(C^2, X_1) \\ R(C^2, X_2) \end{pmatrix} = \sqrt{0.212} \begin{pmatrix} 0.707 \\ -0.707 \end{pmatrix} = \begin{pmatrix} 0.324 \\ -0.324 \end{pmatrix}$$

Donc, les cooronnées de $X_1$ sont $\begin{pmatrix} 0.946 \\ 0.324 \end{pmatrix}$ et les coordonnées de $X_2$ sont $\begin{pmatrix} 0.946 \\ -0.324 \end{pmatrix}$. D'où :

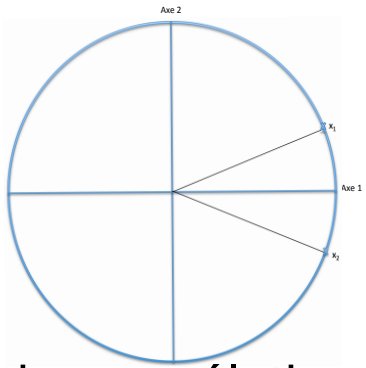


Figure – Cercle des corrélations

# Exemple d'application

- *Interprétation* :
  - Les variables $X_1$ et $X_2$ sont parfaitement représentées sur le cercle des corrélations (cf. Fig. 5), car ce cercle est situé dans le plan des 2 variables $X_1$ et $X_2$ et se traduit, algébriquement, par $(0.946)^2 + (0.324)^2 = 1$.
  - Aussi, l'angle au centre entre les deux variables est égale à $38^\circ$ et on retrouve la valeur du coefficient de corrélation entre $X_1$ et $X_2$ à partir de Fig. 5 : $R(X_1, X_2) = \cos(38^\circ) = 0.79$.
  - Les variables $X_1$ et $X_2$ sont positivement et fortement corrélées avec la 1ère composante principale.
  - Au contraire, $X_1$ et $X_2$ sont assez faiblement corrélées avec la seconde composante principale et s'opposent sur ce second axe.

# Exemple d'application

- Notons bien que :

$$R^2(C^1, X_1) + R^2(C^1, X_2) = 1.788 \quad \text{et} \quad R^2(C^2, X_1) + R^2(C^2, X_2) = 0.212$$

Comme

$$C^1 = 0.707\, X_1 + 0.707\, X_2 \quad \text{et} \quad C^2 = 0.707\, X_1 - 0.707\, X_2$$

on obtient le tableau des coordonnées des individus :

| Individu | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $C^1$ | -1.802 | -0.032 | -0.385 | -0.136 | 2.351 |
| $C^2$ | 0.381 | -0.881 | -0.021 | 0.233 | 0.289 |

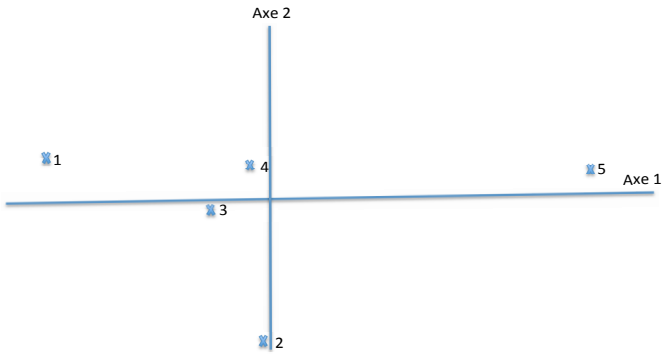


Figure – Représentation des individus sur le 1er plan principal

# Exemple d'application

- *Interprétation* :
  - Les individus 1 et 5 sont ceux qui contribuent le plus fortement à la variance sur le premier axe. De plus, sur cet axe, ces individus s'opposent, puisque l'individu 1 et à la gauche de l'axe et l'individu 5 à la droite de l'axe. Donc, l'interprétation est : le 1er axe est lié fortement aux variables $X_1$ et $X_2$ (cf. le cercle des corrélations) : les individus 1 et 5 sont ceux qui connaissent des valeurs extrêmes à la fois pour $X_1$ et $X_2$ (petites pour $X_1$ et grandes pour $X_2$).
  - L'individu 3 est presque parfaitement représenté sur l'axe 1 : sa position correspond aux valeurs qu'il prend pour $X_1$ et $X_2$ i.e., légèrement en dessous de la moyenne pour chacune des deux variables.
  - Remarquons que, c'est surtout l'individu 2 qui contribue à la variance de l'axe 2. En fait, cet axe est lié positivement à $X_1$ et négativement à $X_2$ (cf. cercle des corrélations), et la position de l'individu 2 est due à la faible valeur qu'il prend pour $X_1$ par rapport à la forte valeur prise par $X_2$.
  - A l'opposé, l'individu 4, bien représenté sur l'axe 2, doit sa position à une valeur de $X_1$ relativement forte par rapport à la valeur prise par $X_2$.

# Formules de reconstitution
## Reconstitution du tableau de données

Soit $C$ la matrice $n \times p$ dont les $p$ colonnes sont les $p$ composantes principales $C^1, \ldots, C^p$ et soit $A$ la matrice $p \times p$ dont les $p$ colonnes sont les facteurs $a^1, \ldots, a^p$.

On montre le résultat suivant, qui nous permet de reconstituer les données $X$ à partir des composantes et des facteurs principaux.

**Proposition**

$$X = \sum_{k=1}^{p} C^k \, {}^t(a^k)$$

*ce qui s'écrit aussi*

$$X = \sum_{k=1}^{p} \sqrt{\beta_k} \frac{C^k}{\sigma(C^k)} \, {}^t(a^k),$$

*où $\sigma(C^k)$ désigne l'écart-type de $C^k$.*

# Remarques

- Cette dernière formule montre comment $X$ peut être reconstituée à partir d'éléments dont les normes sont les mêmes d'une étape à l'autre (à l'étape $k$, la composante principale normée d'ordre $k$ et le facteur normé d'ordre $k$), en pondérant ces éléments par les racines carrées des valeurs propres.
- En effectuant une ACP, on n'utilise que les résultats des $r$ premières étapes, ce qui revient à considérer que les $r$ premières étapes apportent une information suffisante pour l'analyse du tableau $X$. Comme les dernières valeurs propres sont petites, l'information résiduelle

$$\sum_{k=r+1}^{p} \sqrt{\beta_k} \frac{C^k}{\sigma(C^k)} {}^t(a^k)$$

est négligeable et donc $X \approx \sum_{k=1}^{r} \sqrt{\beta_k} \frac{C^k}{\sigma(C^k)} {}^t(a^k)$ est très peu différent de $X$.

# Reconstruction de la matrice de corrélations

De la même manière, on peut reconstituer la matrice de corrélations $\frac{1}{n}({}^tX\,X)$ à partir des $p$ facteurs :

$$\frac{1}{n}({}^tX\,X) = A\left(\frac{1}{n}\,{}^tC\,C\right)\,{}^tA$$

où $\frac{1}{n}\,{}^tC\,C$ est la matrice de variances-covariances entre les composantes principales. On déduit également que

**Proposition**

$$\frac{1}{n}({}^tX\,X) = \sum_{k=1}^{p} \beta_k\, a^k\,{}^t(a^k).$$

**Remarque**

*La connaissance des premiers facteurs permet une bonne reconstitution de la matrice des corrélations si les valeurs propres suivantes ont de faibles valeurs. Graphiquement, ceci signifie que les cercles de corrélations obtenus à partir des premières composantes principales décrivent l'essentiel des corrélations entre les variables.*

# Variables et individus supplémentaires

Dans certains cas, il peut être intéressant de faire figurer sur les graphiques un ou plusieurs individus, ou une ou plusieurs variables qui ne figurent pas dans le tableau de départ.

Les coordonnées de cet individu supplémentaire ou de cette variable supplémentaire se calculent de la même manière que pour les individus ou variables d'origine, par leurs produits scalaires avec les facteurs dans le cas des individus ou par leurs coefficients de corrélation avec les composantes principales, dans le cas des variables.

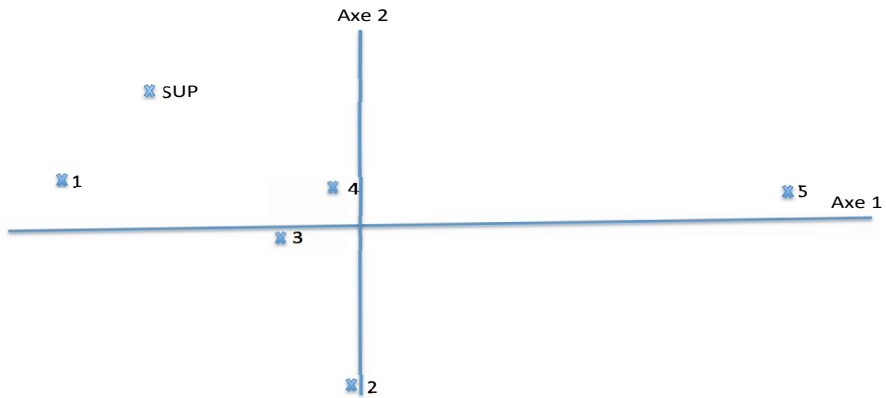


Figure – Représentation d'un individus supplémentaire

# Exemple d'application

- Reprenons l'exemple numérique traité dans ce chapitre, et considérons l'individu supplémentaire qui prend la valeur 5 pour $x_1$ et la valeur 3 pour $x_2$. Autrement dit, les valeurs prises pour les variables centrées et réduites $X_1$ et $X_2$ valent respectivement 0.43 et $-2.40$. Les coordonnées de cet individu sont alors :

$$\begin{aligned} \text{pour } C^1 &: (0.707)(0.43) + (0.707)(-2.40) = -1.39. \\ \text{pour } C^2 &: (0.707)(0.43) + (-0.707)(-2.40) = 2.00. \end{aligned}$$

- La qualité de représentation de cet individu vaut :

$$\begin{aligned} \text{pour } C^1 &: \frac{(-1.39)^2}{(0.43)^2+(-2.40)^2} = 0.33. \\ \text{pour } C^2 &: \frac{(2.00)^2}{(0.43)^2+(-2.40)^2} = 0.67. \end{aligned}$$

  La contribution d'un individu supplémentaire à la variance des axes est évidemment nulle.

- Considérons maintenant la variable supplémentaire $x_3$ :

| Individu | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $x_3$ | 4.00 | 2.00 | 5.00 | 4.00 | 7.00 |

# Exemple d'application

La moyenne de cette variable est égale à 4.4 et son écart-type à 1.62. La variable centrée-réduite correspondante $X_3$ prend alors les valeurs suivantes :

| Individu | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $X_3$ | -0.25 | -1.48 | 0.38 | -0.25 | 1.60 |

et la corrélation de $X_3$ avec la première composante principale est égale à 0.62, tandis que la corrélation de $X_3$ avec la seconde composante principale est égale à 0.70.

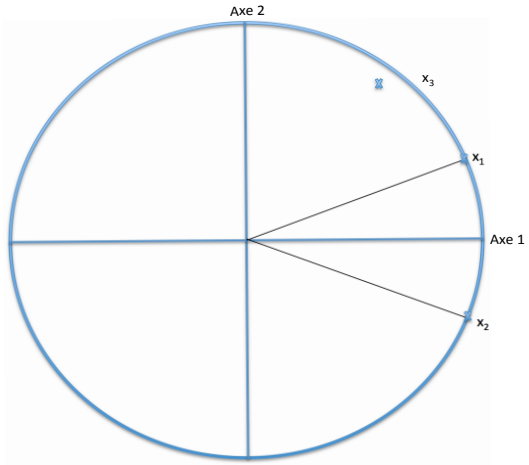


Figure – Représentation d'une variable supplémentaire

# L'ACP de variables non réduites

Une des hypothèses formulées au début de ce chapitre est que les variables de départ (les colonnes du tabeau) sont réduites, de sorte que l'ACP présentée tout au long du chapitre est l'ACP normée, ou encore ACP sur matrice des corrélations.

Il est aussi possible d'effectuer une ACP sur des variables dont les variances ne sont pas égales à 1 : il s'agit alors d'une ACP non normée ou ACP sur matrice de variances-covariances.

# L'ACP de variables non réduites

## Définition

*Principe de l'ACP non normée : dans l'espace des individus, le critère est le même en ACP non normée qu'en ACP normée, mais il s'applique à un nuage de points différent, puisque les colonnes de X sont alors non normées.*
*Dans l'espace des variables, le critère de l'étape k, c'est-à-dire la maximisation de :*

$$\sum_{j=1}^{p} R^2(C^k, X_j)$$

*est alors remplacé par la maximisation de :*

$$\sum_{j=1}^{p} cov^2(C^k, X_j).$$

*Le facteur correspondant est le kème vecteur propre de* $\frac{1}{n}({}^tX\,X)$*, qui est alors la matrice de variances-covariances entre les p variables.*
*Les graphiques sont de même type et suivent les mêmes règles d'interprétation pour les deux types d'ACP.*

# Changement d'échelle des variables

Contrairement à l'ACP normée, l'ACP non normée est sensible aux changements d'échelle des variables : les premières composantes principales sont fortement influencées par les variables à variance élevée, comme le montre l'exemple suivant.

## Exemple

*Considérons deux variables $X_1$ et $X_2$, de même variance égale à* 1*, et telles que*

$$cov(X_1, X_2) = 0.5$$

*$X_1$ s'exprime en milliers d'euros et $X_2$ en tonnes. Par conséquent, la matrice à diagonaliser est*

$$\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

*et le premier facteur est*

$$\begin{pmatrix} 0.707 \\ 0.707 \end{pmatrix}$$

*et donc*

$$C^1 = 0.707\, X_1 + 0.707\, X_2$$

# Changement d'échelle des variables

## Exemple

*Exprimons maintenant la variable $X_1$ en centaines d'euros : on obtient une variable*

$$X_{1N} = 10\, X_1$$

*la matrice de variances-covariances devient*

$$\begin{pmatrix} 100 & 5 \\ 5 & 1 \end{pmatrix}$$

*et le premier facteur est*

$$\begin{pmatrix} 0.9987 \\ 0.0509 \end{pmatrix}$$

*Par conséquent*

$$C^1 = 0.9987\, X_{1N} + 0.0509\, X_2 \text{ soit } C^1 = 0.9987\, X_1 + 0.0509\, X_2.$$

*Cette nouvelle composante est presque entièrement déterminée par $X_1$ et dépend très peu de $X_2$.*

# L'ACP dans une métrique euclidienne quelconque

- L'ACP peut se généraliser au cas où les distances entre les individus sont calculées grâce à une métrique euclidienne quelconque.
- Cette ACP dans une métrique euclidienne $M$ se ramène à une ACP non normée moyennant une transformation des données de départ, comme cela sera montré dans le cas de l'analyse factorielle des correspondances (AFC) et dans le cas de l'analyse factorielle discriminante (AFCM).

# Pondération des individus

Une hypothèse implicite de ce chapitre est que tous les individus ont le même poids, mais il est possible d'effectuer une ACP en associant à chaque individu un poids.

Soit $p_i$ le poids associé à l'individu $i$ :

$$0 < p_i < 1 \text{ et } \sum_{i=1}^{n} p_i = 1.$$

Soit $D$ la matrice diagonale des poids :

$$D = \begin{pmatrix} p_1 & 0 & \dots & \dots & \dots & 0 \\ 0 & p_2 & 0 & \dots & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & p_i & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & p_n \end{pmatrix}$$

En reprenant la démarche du paragraphe 3.1, le problème de l'ACP s'écrit alors à l'étape $k$, dans l'espace des individus :

$$\text{Maximiser} \sum_{i=1}^{n} p_i \, (C_i^k)^2$$