École Polytechnique de l'Université de Tours
64, Avenue Jean Portalis
37200 TOURS, FRANCE
Tél. +33 (0)2 47 36 14 14
www.polytech.univ-tours.fr

Departement Informatique
4emeannee
2016-2017

Cours - Partie 1

Analyse de données

Encadrants

Jean-Yves RAMEL
jean-yves.ramel@univ-tours.fr

Université Francois-Rabelais, Tours

Auteurs

Jean-Yves RAMEL
jean-yves.ramel@univ-tours.fr
+ Eleves de la promo 2008-2009

Version du 15 février 2017

Table des matières

Table des figures

Plan

— **Introduction**
Qu'est-ce que l'analyse de données ?
Rappels sur les outils.

— **Méthodes adaptées aux problÃĺmes descriptifs :** comment décrire le contenu d'un tableau.
→ **ACP** : méthode de base : analyse en composantes principales.
→ **AFC** : Analyse factorielle des correspondances : lien entre ligne et colonnes.

— **Méthodes adaptées aux problÃĺmes décisionnels**
— Analyse discriminante : reconnaissance des formes / aide au diagnostic.
— Méthode de classification : faire des groupes (classes).

CHAPITRE 1

Cadre général de l'analyse de données

1.1 Qu'est-ce que l'analyse de données ?

1.1.1 Statistiques inférentielles VS statistique descriptives

— L'analyse de données ne doit pas être confondue avec ce que l'on appelle classiquement les statistiques.

— *Les statistiques inférentielles* ont pour objectif l'étude d'une *population mére* à partir d'un *échantillon*.
On suppose que l'échantillon est représentatif.
Questions : comment choisir l'échantillon ? Comment évaluer sa représentativité ? Quelle est la fiabilité des résultats ?

— *Les statistiques descriptives* concernent l'analyse du contenu des données, l'interprétation du contenu et la représentation du contenu.
Le contenu est aussi bien l'échantillon que la population mére.

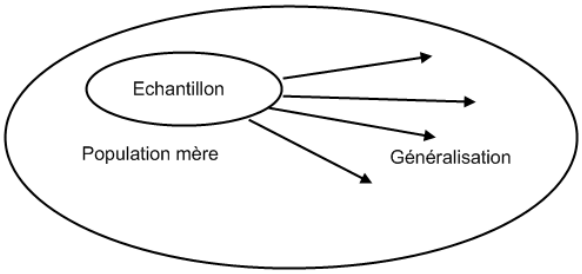


FIGURE 1.1 – Statistiques inférentielles

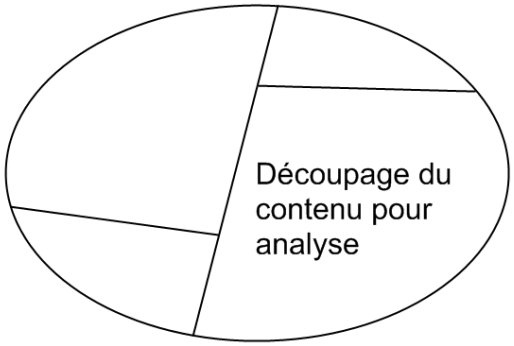


FIGURE 1.2 – Statistiques descriptives

L'analyse de données fait partie des statistiques descriptives.

1.1.1.1 Objectif de l'analyse de données

— Trouver *une représentation optimale* (dessin) selon différents critéres *d'un ensemble d'objets* (lignes du tableau = individus = points) caractérisés par *un ensemble de descripteurs* (colonnes du tableau = variables = coordonnées).

Les techniques d'analyse de données sont donc basées sur la représentation d'objets par un nuage de points dans un espace à n dimensions.

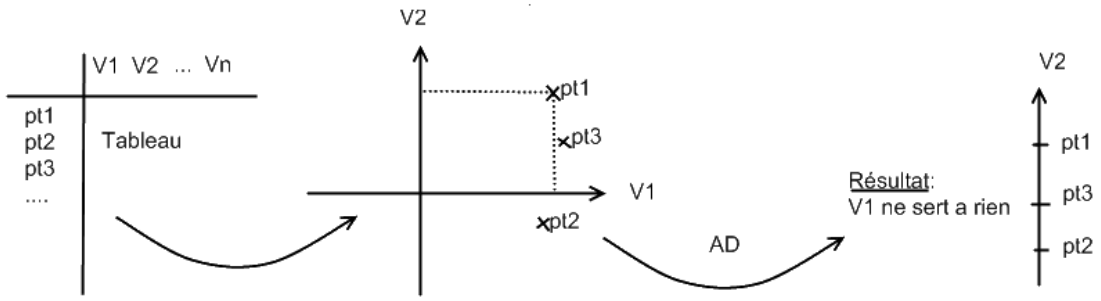


FIGURE 1.3 – Fonctionnement de l'analyse de données

Il s'agit alors de simplifier l'information contenu dans des grands tableaux. Pour cela, on s'autorise à *perdre un peu d'information* afin d'obtenir un *gain de simplification*.
Cela doit permettre de mettre en évidence un "phénoméne", une structuration dans les données (les objets) pour faire de *l'interprétation*.

L'analyse de données propose des approches plus qualitatives que quantitatives. Elle est donc utilisée plutot à titre *exploratoire* ⇒ attention aux fausses interprétations et conclusions.
Une fois la structuration des données mises en évidences, certains méthodes permettent de résoudre des problémes décisionnels ⇒ affecter une classe à un nouvel objet.

Les principales méthodes datent des années 1950.
Peu utilisées au départ, elles sont aujourd'hui trés utilisées dans de *nombreux domaines* (grace aux ordinateurs).
⇒ Applications :

— Big data : gestion d'entrepots de donnees, data architect, data scientist, ...
— Marketing : ciblage, analyse de log, personnalisation d'IHM.
— Santé : aide au diagnostic, RdF.
— Gestion de production, contrÃŧle qualité, perfectionnement de procédés, analyse de trafic et de flux.
— 3D : visualisation de données.

1.1.2 Processus de traitement des données

Un tableau ne peut être mis directement en entrée d'un systéme d'analyse de données.

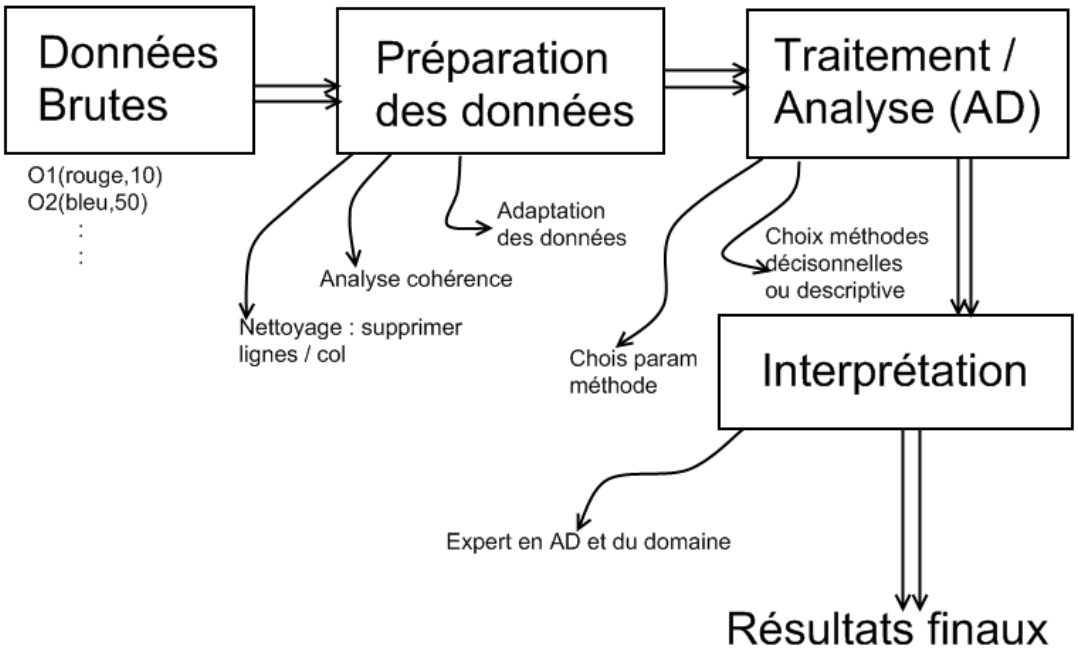


FIGURE 1.4 – Processus de traitement des données

1.2 Typologies des problémes et des méthodes

On a vu deux types de problémes : *descriptifs* contre *décisionnels*.

- **Descriptifs** : géométrie + nuage de points (ACP) + projection dans l'espace (AFC).
- **Décisionnels** : 2 groupes de variables : les variables explicatives et les variables à expliquer. L'objectif est d'expliquer les valeurs prises par les variables à expliquer à l'aide des valeurs prises par les variables explicatives.

$$\overbrace{C}^{\text{variable à expliquer}} = \alpha \overbrace{V_i}^{\text{variable expliquée}} + \beta \cdot V_2 + \dots$$

  → méthode de régression.
  → AF discriminante.

1.3 Types de variables

Les types de variables que l'on peut rencontrer sont :

- **Variables qualitatives** : les variables ne peuvent prendre que des valeurs incomparables sur le plan numérique. Il s'agit la plupart du temps de *modalités textuelles*.
  Exemple : couleur = rouge, vert, bleu, ...
- **Variables quantitatives** : les variables peuvent prendre des valeurs numériques sur lesquelles il est légitime de calculer une moyenne, une variance, un écart-type, ...
  Il peut également être possible d'utiliser plutÃẗt une échelle de mesure ou de créer des intervalles.
  Exemple : taille, poids, ...

— **Variables binomiales** : ne peut prendre que 2 valeurs.
Exemple : 0/1, Vrai/Faux, Homme/Femme, ...

Il est parfois difficile de déterminer le type d'une variable.

On peut transformer le type d'une variable :

Qualitative $\longrightarrow$ Binomiale

Ind	Couleur
o_1	Rouge
o_2	Bleu
...	...

$\rightarrow$

	rouge	bleu
o_1	1	0
o_2	0	1

Quantité $\longrightarrow$ Qualité

Ind	Taille
o_1	$1,70$
o_2	$1,90$

$\rightarrow$

Ind	Taille
o_1	petit
o_2	grand

On peut aussi distinguer les variables *nominales* des variables ordinales :
— **Nominales** : opérateurs de tri non utilisables (seulement le $=$).
— **Ordinales** : opérateurs de tri utilisables ($\leq, \geq, =, \neq$).
Exemple : Pas du tout $\leq$ un peu $\leq$ beaucoup.

1.3.1 Exemples de tableaux

Différents types de tableaux peuvent être étudiés en analyse de données.
— Tableaux individu / variables (numériques ou qualitatives $\Rightarrow$ tableau de modalités).
— Tableaux de contingence $\Rightarrow$ variables / variables $\Rightarrow$ valeurs = effectifs.
— Tableaux de distance ou similarité (individus / individus).

1.4 Représentations des données

Variables binomiales $\Rightarrow$ calcul de pourcentage et proportions.

Variables quantitatives $\Rightarrow$ moyenne, σ, variance, ...

Variables qualitatives $\Rightarrow$ représentation de distributions : histogramme, camembert.

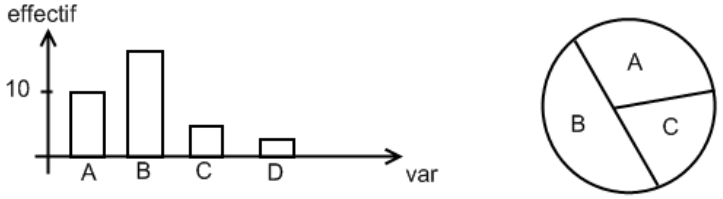


FIGURE 1.5 – Histogramme et camembert

CHAPITRE 2

Rappels mathématiques pour l'analyse de données

2.1 Quelques rappels de statistiques

— **Moyenne et Esperance mathematique :**

$$E(X) = \sum_{i=1}^{n} p_i V_i$$

Si $p_i = \frac{1}{n}$ alors moyenne classique.

— **Variance :**

$$V(X) = E\left[(X - E(X))^2\right] = E\left(X^2\right) - [E(X)]^2$$

X	$(X - E(X))^2$
1	$(-4)^2$
2	$(-3)^2$
3	.
4	.
10	.
3	.
moyenne $E(X) = 5$	$V(X)$

— **Ecart-type :**

$$\sigma(X) = \sqrt{V(X)}$$

— **Covariance :**

$$\text{cov}(X, Y) = E\left((X - E(X))(Y - E(Y))\right)$$

$\text{cov}(X, X) = V(X)$

— **Coefficient de corrélation :**

$$\text{correl}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma(X) \cdot \sigma(Y)}$$

$\text{correl}(X, X) = 1$

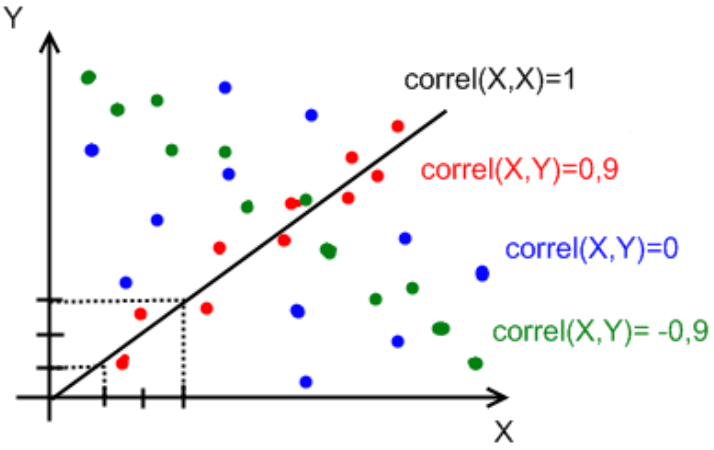


FIGURE 2.1 – Influence du coefficient de corrélation

	V_1	V_2	...	V_p
V_1	1			
V_2		1	0,5	
V_3		0,5	1	
$\vdots$				1

Matrice des correlations : Symétrique, carrée, valeurs $\in [-1;1]$

2.2 Calcul matriciel

Pour faire de l'analyse de données, il faut savoir :

2.2.1 Multiplier deux matrices

$$\underline{A} \in \mathcal{M}_{p,q}(\mathbb{R}) \cdot \underline{B} \in \mathcal{M}_{q,r}(\mathbb{R}) = \underline{C} \in \mathcal{M}_{p,r}(\mathbb{R})$$

2.2.2 Transposer une matrice

- ${}^t\underline{A} = \underline{B}$, on remplace ligne et colonne.
- ${}^t(\underline{A} + \underline{B}) = {}^t\underline{A} + {}^t\underline{B}$
- $(\underline{AB}) = {}^t\underline{B} \cdot {}^t\underline{A}$
- $(\lambda\underline{A}) = \lambda {}^t\underline{A} \quad \lambda \in \mathbb{R}$
- ${}^t\left(\underline{A}^{-1}\right) = ({}^t\underline{A})^{-1}$
- ${}^t({}^t\underline{A}) = \underline{A}$

2.2.3 Calculer la trace d'une matrice

$\underline{A}$ matrice carrée $\in \mathcal{M}_{n,n}(\mathbb{R})$

$$\text{trace}(\underline{A}) = \sum_{i=1}^{n}(a_{ii})$$

$$\underline{A} = \begin{array}{c} \\ i \end{array} \begin{array}{c} \begin{array}{ccc} & j & \\ & \vdots & \end{array} \\ \begin{pmatrix} \dots & a_{ij} & \dots \end{pmatrix} \\ \end{array}$$

2.2.4 Calculer le déterminant d'une matrice

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc$$

$$\det \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} = a \cdot \det \begin{pmatrix} e & f \\ h & i \end{pmatrix} - d \cdot \det \begin{pmatrix} b & c \\ h & i \end{pmatrix} + g \cdot \det \begin{pmatrix} b & c \\ e & f \end{pmatrix}$$

- $\det(\underline{AB}) = \det(\underline{A}) \cdot \det(\underline{B}) = \det(\underline{BA})$
- $\det({}^t\underline{A}) = \det(\underline{A})$
- $\det(\underline{A}) \neq \varnothing \Rightarrow$ A est inversible

2.2.5 Calculer les valeurs propres et vecteurs propres

DÃ©finition : Vecteur propre

On appelle vecteur propre de $\underline{A}$ tout élément $X \in \mathbb{R}^p$ tel que il existe $\lambda \in \mathbb{R}$ vérifiant

$$\underline{A} \cdot \underline{X} = \lambda \cdot \underline{X}$$

- $X \in \mathbb{R}^p$ est un vecteur composé de p valeurs réelles ;
- $\lambda \in \mathbb{R}$ est la valeur propre (*vp*) associée au vecteur propre $\underline{X}$ ($\vec{vp}$) .

Si λ est vp de $\underline{A}$ alors $(\underline{A} - \lambda \underline{Id})$ n'est pas inversible $\Rightarrow \underbrace{\boxed{\det(\underline{A} - \lambda \underline{Id}) = \varnothing}}_{\text{Système d'équations } \lambda\alpha}$ (polynome caractéristique de $\underline{A}$).

$\underline{Id}$ est la matrice identitée $\begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}_{(p \times p)}$

Degré du polynÃ´me = nombre de vp = nombre de lignes de $\underline{A}$ = nombre de colonnes de $\underline{A}$ au maximum (notion de vp double ...).
A tout $\vec{vp}$ $\underline{X}$ correspond une et une seule vp.

Remarques pour les matrices symétriques $(\underline{A})$:

- trace$(\underline{A}) = \sum_{i=1}^{p} \lambda_i =$ somme des vp de $\underline{A}$.
- Si $\underline{A}$ est positive $\Rightarrow {}^t\underline{X} \cdot \underline{A} \cdot \underline{X} \geq 0, \forall \underline{X}$ alors toutes les vp de $\underline{A}$ sont positives.

2.3 Géométrice et espace vectoriel

2.3.1 Produit scalaire

Soit $\underline{X}$ un vecteur de $\mathbb{R}^n$ et $\underline{Y} \in \mathbb{R}^n$.

Produit scalaire de $\underline{X}$ et $\underline{Y}$:

$$ {}^t\underline{X} \cdot \underline{Y} \in \mathbb{R} $$

On le notera parfois $< \underline{X}, \underline{Y} >= \|\underline{X}\| \cdot \|\underline{Y}\| \cdot \cos\sigma$
(Norme de $\underline{X})^2 = {}^t\underline{X} \cdot \underline{X}$.

2.3.2 Dérivation vectorielle

$$ \underline{a} \in \mathbb{R}^p = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} \qquad \underline{x} \in \mathbb{R}^p = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} $$

a_j constante, $\forall j$ $\qquad$ x_j variable, $\forall j$

On appelle combinaison linéaire de $\underline{x}$ et $\underline{a}$:

$$ \begin{aligned} g &= {}^t\underline{a} \cdot \underline{x} = \sum_{j=1}^{p} (a_j \cdot x_j) \\ \frac{\partial g}{\partial x} &= \begin{pmatrix} \frac{\partial g}{\partial x_1} \\ \frac{\partial g}{\partial x_2} \\ \vdots \\ \frac{\partial g}{\partial x_p} \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \underline{a} \\ \frac{\partial \left(\underline{a}\,{}^t\underline{x}\right)}{\partial d\underline{x}} &= \underline{a} \end{aligned} $$

Forme quadratique d'une matrice $\underline{A}$:

$$ \begin{aligned} F_q &= \left({}^t\underline{x} \cdot \underline{A} \cdot \underline{x}\right) \text{ avec } \underline{A} \in \mathcal{M}_{(p\times p)}(\mathbb{R}) \\ F_q &= \sum_{i=1}^{p}\sum_{j=1}^{p} (a_{ij} \cdot x_i \cdot x_j) \\ \frac{\partial F_q}{\partial x} &= \begin{pmatrix} \frac{\partial F_q}{\partial x_1} \\ \frac{\partial F_q}{\partial x_2} \\ \vdots \\ \frac{\partial F_q}{\partial x_p} \end{pmatrix} = \underline{A} \cdot \underline{x} + {}^t\underline{A} \cdot \underline{x} \\ \frac{\partial \left({}^t\underline{x} \cdot \underline{A} \cdot \underline{x}\right)}{d \cdot \underline{x}} &= 2 \cdot \underline{A} \cdot \underline{x} \text{ lorsque } \underline{A} \text{ est symetrique} \end{aligned} $$

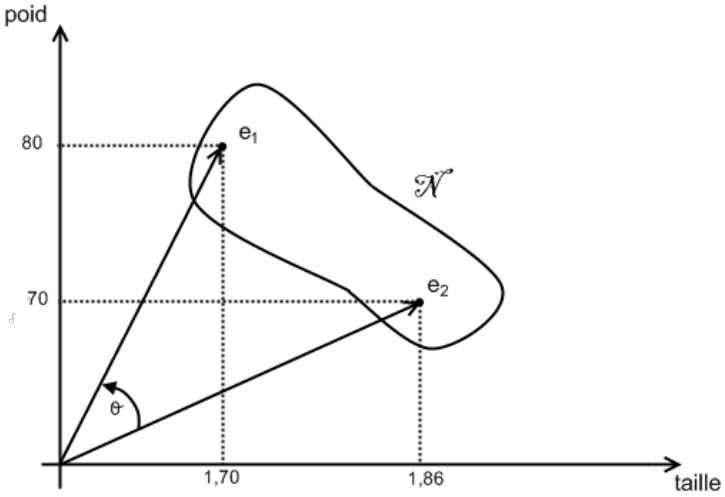


FIGURE 2.2 – Distance entre deux individus

2.4 Notion de métriques

2.4.1 Dans l'espace des individus

On a n individus $= n$ lignes. Comment mesurer la distance entre deux individus ?

	x	y
e_1	1,86	80
e_2	1,70	70

$x =$ taille, $y =$ poids.

$$d(e_1, e_2) = \sqrt{(1,86 - 1,70)^2 + (80 - 70)^2} = \sqrt{0,03 + 100}$$

Problème : $\sqrt{\underbrace{(1,86 - 1,70)^2}_{\text{négligeable}} + \underbrace{(80 - 70)^2}_{\text{prépondérante}}}$.

Solution : il faut normaliser les variables $\Rightarrow$ normalisation par l'écart-type $\sigma_{\text{poids}} = 5$ et $\sigma_{\text{taille}} = 0,08$.

$$d(e_1, e_2) = \sqrt{\left(\frac{1,86 - 1,70}{\sigma_{\text{taille}}}\right)^2 + \left(\frac{80 - 70}{\sigma_{\text{poids}}}\right)^2} = \sqrt{4 + 4}$$

Cette idée de pondération des variables se généralise via la notion de *métrique*. On peut utiliser une notation matricielle :

$$\underline{M} = \begin{pmatrix} m_{11} & m_{12} & \dots & \dots & m_{1p} \\ m_{21} & & & & \\ \vdots & & \ddots & & \\ m_{p1} & & & & m_{pp} \end{pmatrix}_{(p \times p)}$$

Et calculer une distance pondérée par

$$\begin{aligned} d^2_M(\underline{e_1}, \underline{e_2}) &= \\ &= {}^t(\underline{e_2} - \underline{e_1}) \cdot \underline{M} \cdot (\underline{e_2} - \underline{e_1}) \\ &= \sum_{j=1}^{p} \sum_{k=1}^{p} \left[m_{jk} \left(x_{1k} - x_{2k} \right) \left(x_{ij} - x_{2j} \right) \right] \\ \underline{M} &= \begin{pmatrix} \frac{1}{\sigma^2_{\text{taille}}} & 0 \\ 0 & \frac{1}{\sigma^2_{\text{poids}}} \end{pmatrix}_{(2\times 2)} \end{aligned}$$

La plupart du temps, il n'y a des valeurs non nulles que sur la diagonale de $\underline{M}$.

Distance euclidienne classique

$$\underline{M} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}_{(2\times 2)}$$

Si M est symétrique et définie positive, c'est-à-dire si $({}^t x \cdot M \cdot x) \geq 0$, on a une métrique.

2.4.2 Dans l'espace des variables

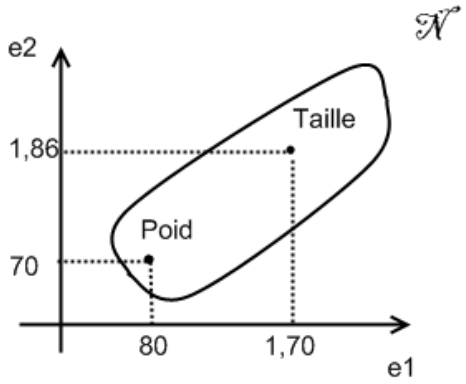


FIGURE 2.3 – Espace des variables

Pondération des individus (p_i à l'individu i) :

$$\begin{pmatrix} p_1 & & 0 \\ & \ddots & \\ 0 & & p_n \end{pmatrix}_{(n\times n)}$$

Souvent, on utilise la pondération :

$$\sum_{i=1}^{n} p_i = 1$$

$$\Rightarrow p_i = \frac{1}{n}$$

2.4.3 Métrique et produit scalaire

On peut combiner produit scalaire et métrique. Par exemple :

$$\begin{aligned} \underline{D} &= \frac{1}{n} Id_{(n\times n)} \\ &= \begin{pmatrix} \frac{1}{n} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{n} \end{pmatrix} \end{aligned}$$

On peut faire le produit scalaire suivant $\underline{D}$:

$$\begin{aligned} < x_j, x_k >_{\underline{D}} &= {}^t\underline{x}_j \cdot \underline{D} \cdot \underline{x}_k \quad \text{avec } \underline{x}_j \text{ vecteur correspondant à une variable} \\ &= \frac{1}{n}\left({}^t\underline{x}_j \cdot \underline{x}_k\right) \in \mathbb{R} \end{aligned}$$

Il est possible de travailler avec *des variables centrées*. Ce sont des variables avec une moyenne égale à 0. On soustrait la moyenne de la variable à chaque valeur de la variable.

V_1 non centrée	V_1 centrée
2	-2
4	0
6	2
moy $= 4$	moy $= 0$

$$\begin{aligned} < \underline{x_j}, \underline{x_k} >_D &= \frac{1}{n}\sum_{i=1}^{n}\left(\left(x_{1j} - \underbrace{\bar{x}_k}_{\text{moyenne de } \underline{x}_j}\right)(x_{1j} - \bar{x}_k)\right) \\ &= \text{covariance}\left(\underline{x}_j, \underline{x}_k\right) \\ &= E\left((X - E(X))\,(Y - E(Y))\right) \end{aligned}$$

$$\begin{aligned} \|\underline{x}_j\|_D^2 &= < \underline{x}_j, \underline{x}_j >_D \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[(x_{ij} - \bar{x}_j)^2\right] \quad \left(\text{variance}\left(\underline{x}_j\right)\right) \\ \Rightarrow \|\underline{x}_j\|_D &= \sigma_{\underline{x}_j} \quad \text{(écart type)} \end{aligned}$$

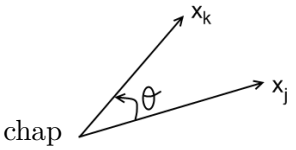


FIGURE 2.4 – $\cos\theta$

$$\begin{aligned} \cos\theta &= \frac{< \underline{x}_j, \underline{x}_k >_D}{\|\underline{x}_j\|_D \times \|\underline{x}_k\|_D} \\ &= \frac{\text{covariance}\left(\underline{x}_j, \underline{x}_k\right)}{\sigma_{\underline{x}_j} \times \sigma_{\underline{x}_k}} \\ &= \text{coefficient de correlation}\left(\underline{x}_j, \underline{x}_k\right) \end{aligned}$$

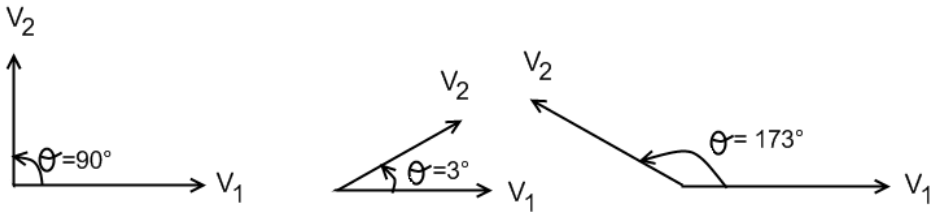


FIGURE 2.5 – $\text{correl}(V_1, V_2) = \cos\theta = 0$, $\text{correl}(V_1, V_2) \simeq 1$, $\text{correl} \simeq -1$

CHAPITRE 3

Analyse en composantes principales (ACP)

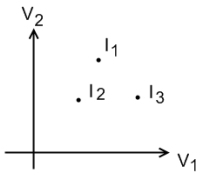
Crée en 1933 par Hotteling, l'ACP permet l'étude d'un phénomène quantitatif. C'est la représentation sous forme graphique d'un tableau de données. C'est une méthode descriptive.

3.1 Problématique

Question : comment analyser le contenu d'un grand tableau ?

— Etude algébrique : je calcule les moyennes, les écart-types, etc. sur les variables.
$\rightarrow$ analyse et conclusion ?

— Représentations graphiques partielles : je prend les variables 2 à 2 $\rightarrow$ représentation XY



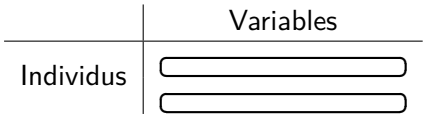
$\rightarrow$ conclusion ?

— ACP : mélange des deux approches.

3.1.1 Objectifs de l'ACP

3.1.1.1 Vision analytique

- A propos de 2 *individus*, on essai d'étudier leurs ressemblances.
  2 individus se ressemblent s'ils possèdent des valeurs similaires pour l'ensemble des *variables*.



  Avec l'ACP, on mesure la ressemble avec *la distance euclidienne*.

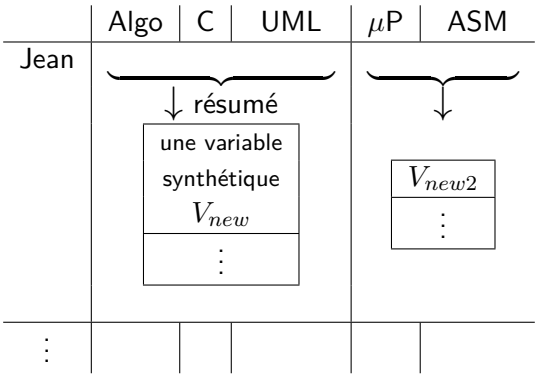
- A propos de 2 variables, on essaie d'évaluer leur liaison.
  Avec l'ACP, on utilise *le coefficient de corrélation* (parfois la covariance).

L'ACP est l'étude exploratrice visant à répondre aux questions suivantes :

— Quels sont les individus qui se ressemblent ? Quels sont ceux qui sont différents ? Existe-t'il des groupes d'individus ?

— Quelles sont les variables liées positivement ? Quelles sont les variables liées négativement ? Existe-t'il des groupes de variables ?

Un autre aspect est d'essayer de résumer l'ensemble des variables pour un petit nombre de *variables synthétiques* appelées *composantes principales*.



Cette variable synthétique est *une combinaison linéaire* de toutes les variables initiales :

$$V_{new} = \boxed{\alpha} V_1 + \boxed{\beta} V_2 + \ldots + \boxed{\gamma} V_p$$

Données par l'ACP

Si $\gamma = 0$ alors V_p ne contribue pas à la variable V_{new} et inversement.

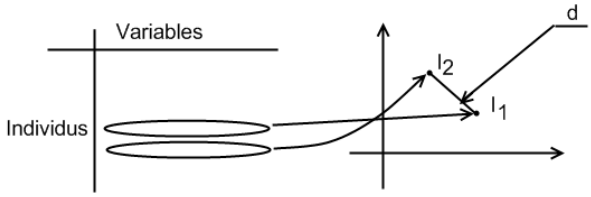


FIGURE 3.1 – Distance euclidienne entre deux variables

3.1.1.2 Vision géométrique

L'ACP permet de rechercher une représentation des n individus dans un sous espace de $\mathbb{R}^p$ de dimension $k \ll p$ (souvent $k = 2$ ou 3).

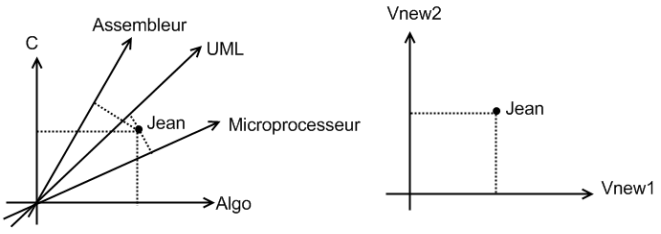


FIGURE 3.2 – Vision géométrique de l'ACP

Pour cela on définit de nouvelles variables synthétiques (combinaisons linéaires des variables initiales) de faÃğon à perdre le moins d'information possible. Ces nouvelles variables sont appelées *composantes*

principales. Les axes qu'elles déterminent sont les *axes factoriels* (ou axes principaux. Les formes linéaires normées à 1 de ses axes sont appelés *facteurs principaux*.

L'ACP revient à ajuster les nuages des n points individus par un sous espace vectoriel de $\mathbb{R}^p$ en utilisant la distance euclidienne. Cela revient à effectuer des projections orthogonales des points individus dans un sous espace. Le sous espace est choisi de manière à perdre le moins d'information possible.

Exemple 3.1

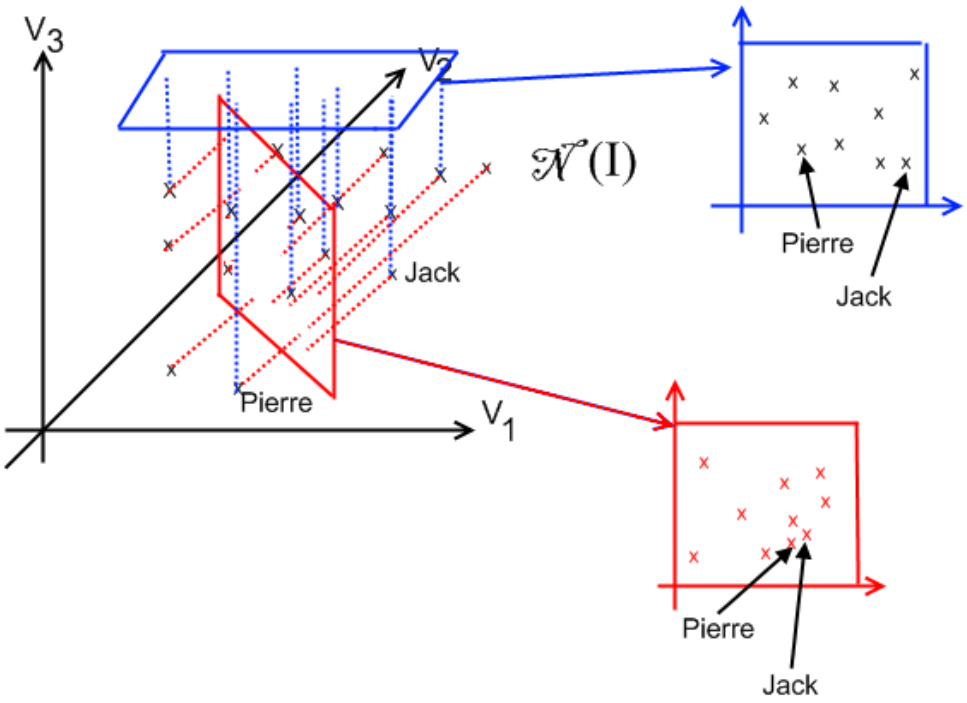


FIGURE 3.3 – Exemple de projection avec un espace initial de dimension 3

Les points doivent bouger le moins possible lors de la projection.

Remarque : la méthode des moindres carrés est l'ACP qui permet de passer d'un espace en deux dimensions à un espace à une dimension.

3.2 Notations

L'ACP travaille sur un *tableau individus / variables quantitatives*. C'est une matrice $\underline{X}_{(n\times p)}$ oÃź les n lignes sont les individus et les p colonnes les variables :

$$\begin{pmatrix} x_{11} & x_{12} & \dots & \dots & x_{1p} \\ x_{21} & \ddots & & & \vdots \\ x_{31} & & x_{ij} \in \mathbb{R} & & \vdots \\ \vdots & & & \ddots & \vdots \\ x_{n1} & \dots & \dots & \dots & x_{np} \end{pmatrix} \leftarrow \text{individu 1}$$

— Les variables sont des vecteurs :

$$\underline{x}_j = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix}_{(n\times 1)}$$

Suite de vecteurs :

$$\underline{X} = \left[\underline{x}_1, \underline{x}_2, \dots, \underline{x}_p\right] \qquad \underline{x}_j \in \mathbb{R}^n$$

— Les individus sont des vecteurs :

$$\underline{L}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix} \qquad L_i \in \mathbb{R}^p$$

Dans $\underline{X}$, on a :

$${}^t\underline{L}_i = [x_{i1}, x_{i2}, \dots, x_{ip}]$$

$$\underline{X} = \begin{pmatrix} {}^t\underline{L}_1 \\ {}^t\underline{L}_2 \\ \vdots \\ {}^t\underline{L}_n \end{pmatrix}$$

$\underline{X}$ est un ensemble d'individus.

3.3 Nuage et inertie

3.3.1 Inertie du nuage

On note $\mathcal{N}(L)$ le nuage formé de $\underline{L}_i$ points individus, $\underline{L}_i \in \mathbb{R}^p$.

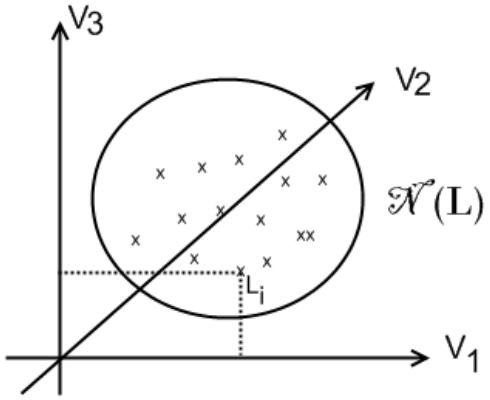


FIGURE 3.4 – Nuage $\mathcal{N}(L)$ des n points individus

On peut associer un poids (une masse) m_i à chaque individu.
Masse du nuage $\mathcal{N}(L)$:

$$m = \sum_{i=1}^{n} m_i$$

$$\underline{D} = \begin{pmatrix} m_1 & & & \\ & m_2 & & 0 \\ & & \ddots & \\ & 0 & & \\ & & & m_n \end{pmatrix}_{(n\times n)}$$

Centre de gravité de $\mathcal{N}(L)$:

$$\underline{G} = \frac{1}{m}\sum_{i=1}^{n}(m_i \cdot \underline{L}_i) = \begin{pmatrix} \bar{x_1} \\ \bar{x_2} \\ \vdots \\ \bar{x_p} \end{pmatrix}$$

DÃ©finition : **Inertie du nuage**

On appelle inertie du nuage $\mathcal{N}(L)$ par rapport au point $A \in \mathbb{R}^p$, le réel

$$I_n(A) = \sum_{i=1}^{n}\left[\|\underline{L}_i - \underline{A}\|^2 \cdot m_i\right]$$

C'est la somme de toutes les distances au carré du point A à l'ensemble des points du nuage (pondérés par leur masse).

Si $A = 0$, on a

$$\begin{aligned} I_n(0) &= \sum_{i=1}^{n} \left[\|L_i\|^2 \cdot m_i \right] \\ &= \sum_{i=1}^{n} \underbrace{< \underline{L_i}, \underline{L_i} >_D}_{\text{Produit scalaire}} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{p} \left[(x_{ij})^2 \cdot m_i \right] \end{aligned}$$

Sous forme matricielle, on obtient :

$$I_n(0) = \text{trace}(V) \text{ avec } V = {}^t\underline{X} D \underline{X}$$

V est appelée matrice d'inertie de $\mathcal{N}(L)$.

$$\underset{n \text{ individus}}{\underline{X}} = \begin{pmatrix} L_1 \\ L_2 \\ \vdots \\ L_n \end{pmatrix}$$

$$\underset{p \text{ variables}}{\underline{X}} = \begin{pmatrix} \underline{x}_1 & \underline{x}_2 & \dots & \underline{x}_p \end{pmatrix}$$

L'objectif de l'ACP est de trouver un sous espace de $\mathbb{R}^p$ dans lequel le nuage $\mathcal{N}(L)$ sera *bien projeté*[1]. La mesure de la qualité de la projection sera évaluée en comparant l'inertie du nuage avant projection avec l'inertie du nuage après projection.

3.3.2 Ajustement du nuage N dans un sous espace

3.3.2.1 Projection d'un point sur une droite

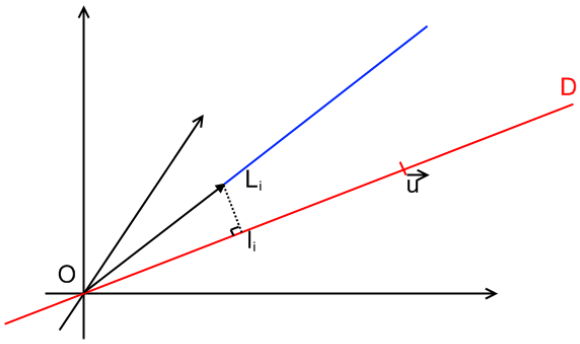


FIGURE 3.5 – Projection d'un point sur une droite

1. cf. schéma 3.3 p. 21

l_i : coordonnées de L_i sur l'axe définie par $\vec{u}$.

$$\begin{aligned} \text{Si } \|\vec{u}\| &= 1 \\ \|\vec{Ol_i}\| &= \|\vec{OL_i}\| - \|\vec{u}\| \cos\theta \\ &= {}^t\underline{L_i} \cdot \underline{u} \\ &= < \vec{OL_i}; \vec{u} > \end{aligned}$$

3.3.2.2 Projection du nuage sur une droite

→ Ajustement de $\mathcal{N}(L)$ dans un sous espace de dimension 1

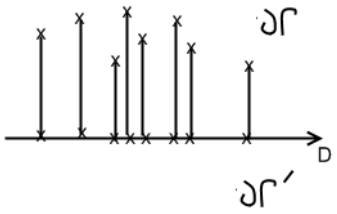


FIGURE 3.6 – Projection du nuage sur une droite

Bon ajustement ⇒ $\|l_i\vec{L_i}\|$ proche de 0 : minimum.
On cherche

$$\begin{aligned} \underset{\vec{u}}{\text{Minimum de}} \sum_{i=1}^{n} \|l_i L_i\|^2 &= \underset{\vec{u}}{\text{Maximum}} \left(\sum_{i=1}^{n} \|\vec{Ol_i}\|^2 \right) \\ &= \underset{\vec{u}}{\text{Max}} \left[\sum_{i=1}^{n} \left({}^tL_i u \right) \left({}^tL_i u \right) \right] \left(\Rightarrow {}^t(L_i u) = \left({}^tL_i u \right) \right) \\ &= \underset{\vec{u}}{\text{Max}} \sum_{i=1}^{n} \left[\underbrace{{}^t\left({}^tL_i u \right)}_{} \left({}^tL_i u \right) \right] \\ &= \underset{\vec{u}}{\text{Max}} \sum_{i=1}^{n} \left[\left({}^t u \cdot L_i \right) \left({}^tL_i u \right) \right] \\ &= \underset{\vec{u}}{\text{Max}} \left[{}^t u {}^t X X u \right] \end{aligned}$$

Pythagore : $\underbrace{\|OL_i\|^2}_{\text{constante}} = \underbrace{\|l_i L_i\|^2}_{\text{minimum}} + \underbrace{\|Ol_i\|^2}_{\text{maximum}}$

Donc l'ACP sur un espace de dimension 1 revient à chercher $\vec{u}$ tel que $\underset{\vec{u}}{\text{Max}} \left({}^t u {}^t X X u \right)$ et $\|\vec{u}\| = 1 \Rightarrow {}^t u \cdot u = 1$.
Méthode pour résoudre ce "double" problème : méthode du multiplicateur de Lagrange.
On note

$$\underbrace{\Phi}_{\text{on cherche le max}} = {}^t u {}^t X X u - \underbrace{\lambda}_{\text{multiplicateur de Lagrange}} \underbrace{\left({}^t u u - 1 \right)}_{\text{il faut Ãğa } \min=0}$$

$\vec{u}$ est la variable de $\Phi(\vec{u})$.
On cherche un Max d'une fonction de $\vec{u}$ ⇒ dérivée $= 0$.

$$\frac{\delta\Phi}{\delta u} = \overbrace{\frac{\delta \left({}^t u {}^t X X u \right)}{\delta u}}^{2 {}^t X X u} - \lambda \overbrace{\frac{\delta \left({}^t u u - 1 \right)}{\delta u}}^{\frac{\delta \left({}^t u u \right)}{\delta u} = 2u} = 2 {}^t X X u - 2\lambda u$$

$$\frac{\delta\Phi(u)}{\delta u} = 0 \Rightarrow {}^tXXu = \lambda u, \quad \lambda \in \mathbb{R}$$

$$AX = \lambda X$$

Pour avoir $\underset{\vec{u}}{\text{Max}}\ (\sum_{i=1}^{n} \|l_i L_i\|^2)$ il faut prendre $\vec{u}$ tel que u soit vecteur propre de la matrice ${}^tXX$.
Il faut prendre la plus grande valeur propre de ${}^tXX$ et le vecteur propre $\vec{u}$ associé.
Inertie du nouveau nuage obtenu lorsque $m_i = 1$:

$$\begin{aligned} {}^tu{}^tXXu &= \sum_{i=1}^{n} Ol_i^2 \\ &= \lambda_1 \text{ plus grande valeur propre de } {}^tXX \end{aligned}$$

3.3.2.3 Ajustement du nuage sur un plan (dimension 2)

- On a trouvé $\vec{u}$, il faut $\vec{v}$ tel que $\vec{u} \perp \vec{v}$ et $\|\vec{v}\| = 1$.
- Une démonstration similaire à celle de la section 3.3.2.2 p. 25 (avec multiplicateur de Lagrange) permet de montrer qu'il faut prendre

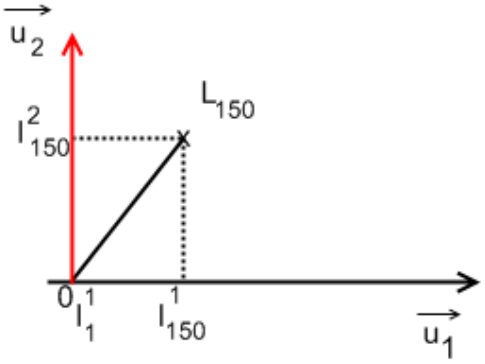
$$\vec{v} = {}^t\underline{XXv} = \mu\underline{V}, \quad \mu \in \mathbb{R}$$

μ est la deuxième plus grande valeur propre de la même matrice ${}^tXX = \lambda_2$ et $\vec{v}$ le vecteur propre associé à λ_2.

3.3.2.4 Ajustement de N par un sous espace de dimension k

Le raisonnement précédent se généralise par un espace de dimension $k < p$.

La base orthonormée de dimension k ajustant au mieux le nuage $\mathcal{N}(L) \in \mathbb{R}^p$ est constituée par les k vecteurs propres associés aux k plus grande valeurs propres de ${}^tXX$.

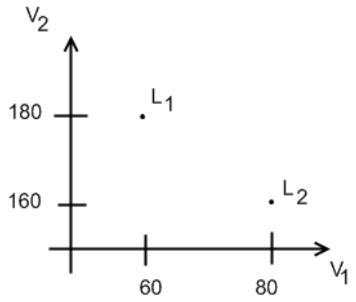


$$\mathcal{N}' \longrightarrow \underline{X} = n \underbrace{\begin{pmatrix} l_1^1 & l_1^2 \\ l_2^1 & \vdots \\ \vdots & \vdots \\ l_{150}^1 & l_{150}^2 \end{pmatrix}}_{\longleftrightarrow\ k=2} \overset{\text{ACP}}{\longleftarrow} \underline{X} = n \begin{pmatrix} p \\ \boxed{} \\ \text{tableau de départ} \end{pmatrix} \mathcal{N}(L)$$

3.3.3 Ajustement du nuage de p points variables dans un sous espace de Rn

3.3.3.1 Rappels

$$L_1 \begin{pmatrix} 180 \\ 60 \end{pmatrix} L_2 \begin{pmatrix} 160 \\ 80 \end{pmatrix} \longrightarrow \underline{X} = \begin{pmatrix} 180 & 60 \\ 160 & 80 \end{pmatrix}$$



$$\underline{x}_1 \begin{pmatrix} 180 \\ 60 \end{pmatrix} \underline{x}_1 \begin{pmatrix} 60 \\ 80 \end{pmatrix} \longrightarrow \underline{X} = \begin{pmatrix} \boxed{} & \boxed{} \end{pmatrix}$$

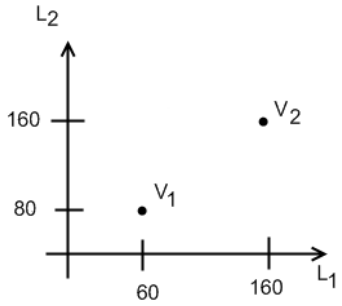


FIGURE 3.7 – Nuage des points Variables

3.3.3.2 Ajustement du nouveau nuage des points Variables

On utilise la même technique que pour les individus ⇒ projection orthogonale des points, maximisation de l'inertie.

— Projection d'un point Variable sur une droite $= {}^t\underline{x}_x \cdot \underline{v}$

— Comme auparavant, on veut

$$\underset{v}{\text{Max}} \left(\sum_{j=1}^{p} \underbrace{\left[{}^t\left({}^t x_j \cdot v \right) \left({}^t x_j \cdot v \right) \right]}_{{}^t v \underline{x}_j \cdot {}^t \underline{x}_j \cdot \underline{v}} \right) \quad \text{sous la contrainte } {}^t v \cdot v = 1$$

$$\varphi = \underset{v}{\text{Max}} \left[{}^t v X \cdot {}^t X \cdot v \right]$$

La contrainte sur l'expression φ est similaire à la section 3.3.2 p. 24 sauf qu'on a $X \cdot {}^tX$ au lieu de ${}^tX \cdot X$.

⇒ **Le meilleur sous espace de projection des points Variables sera obtenu en prenant les k vecteurs propres associés aux k plus grandes valeurs propres de $(X \cdot {}^tX)$.**

3.3.3.3 Relations entre points Variables et points Individus après projection

On a :

$$\begin{aligned} &\text{Dans } \mathbb{R}^p, \quad {}^t\underline{X} \cdot \underline{X} \cdot \underline{u}_\alpha = \lambda_\alpha \underline{u}_\alpha \quad \lambda_\alpha \in \mathbb{R} \\ &\text{Dans } \mathbb{R}^n, \quad \underline{X} \cdot {}^t\underline{X} \cdot v_\alpha = \mu_\alpha \underline{v}_\alpha \quad \mu_\alpha \in \mathbb{R} \end{aligned} \tag{3.1}$$

$$\begin{aligned} (3.1) \quad &\Rightarrow \quad X \cdot {}^tX \cdot \underbrace{X u_\alpha}_{z_\alpha} = X \cdot \lambda_\alpha \cdot u_\alpha = \lambda_\alpha \underbrace{X \cdot u_\alpha}_{z_\alpha} \\ &\Rightarrow \quad X \cdot {}^tX z_\alpha = \lambda_\alpha \cdot z_\alpha \\ &\Rightarrow \quad \lambda_\alpha \text{ est vp de } X \cdot {}^tX \end{aligned}$$

— En fait, on peut montrer que $\lambda_\alpha = \mu_\alpha \quad \forall \alpha$

— ${}^tXX$ et $X \cdot {}^tX$ ont les mêmes valeurs propres.

— On a ${}^tXX \cdot z_\alpha = \lambda_\alpha \cdot z_\alpha$ ⇒ Question 2 : est-ce que $\underline{z}_\alpha = \underline{u}_\alpha$?
On sait que

$$\begin{aligned} \|u_\alpha\| &= 1 \\ \|z_\alpha\| &= \|X \cdot u_\alpha\| \end{aligned}$$

Or

$$\begin{aligned} \|X \cdot u_\alpha\|^2 &= {}^t(Xu_\alpha)\,(Xu_\alpha) \\ &= {}^t u_\alpha {}^t X X u_\alpha \\ &= \lambda_\alpha \underbrace{{}^t u_\alpha \cdot u_\alpha}_{1} \\ &= \lambda_\alpha \end{aligned}$$

On en déduit que

$$v_\alpha = \frac{\overset{(n\times p)}{X}\ u_\alpha}{\sqrt{\lambda_\alpha}} \quad \text{et} \quad u_\alpha = \frac{\overset{(p\times n)}{{}^t X}\ v_\alpha}{\sqrt{\lambda_\alpha}}$$

— $\vec{vp}$ pour points Variables dans $\mathbb{R}^n$
— $\vec{vp}$ pour points Individus dans $\mathbb{R}^p$

$\begin{matrix} ({}^tXX)_{(p\times p)} & \to \\ (X{}^tX)_{(n\times n)} & \to \end{matrix}$ mêmes valeurs propres non nulles.

Coordonnées des points individus dans le nouveau repère ?

1. J'ai $\underline{X} \Rightarrow$ variables quantitatives : tableau individu / variable.
2. On fait une ACP :
   — En fonction de p et n, on calcule $\begin{matrix} \text{soit } {}^tXX \\ \text{soit } X{}^tX \end{matrix} \Rightarrow V$
   — On cherche les vp et $\vec{vp}$ de $\underline{V}$, on trouve des couples $\lambda_\alpha, \vec{u_\alpha}$ α de 1 à $\min(n,p)$
   — On calcule les projections des points individus sur les axes factoriels définis par les $\vec{u_\alpha}$.
   Pour projeter $\underline{L}_1$ sur $\vec{u_1}$, je fais ${}^t\underline{L}_1 \cdot \underline{u}_1 =$ Abs de L_1 sur $\vec{u1} = OL_1$

$$\begin{pmatrix} \alpha \\ \beta \\ \vdots \\ \gamma \end{pmatrix} \vec{u_1} \text{ 1er axe factoriel } \in \mathbb{R}^p \qquad \overset{\vec{u_2}}{\begin{pmatrix} \\ \\ \\ \end{pmatrix}} \quad \dots$$

$$\underline{X} \cdot \underline{u}_2 = \begin{pmatrix} \boxed{L_1} \\ \boxed{L_2} \\ \vdots \\ \boxed{L_n} \end{pmatrix} \underbrace{\begin{pmatrix} l_1^1 \\ l_2^1 \\ \vdots \\ l_n^1 \end{pmatrix}}_{\text{1ère CP}} \qquad \begin{pmatrix} l_1^2 \\ l_2^2 \\ \vdots \\ l_n^2 \end{pmatrix} \quad \dots \quad \begin{pmatrix} \\ \\ \\ \end{pmatrix}$$

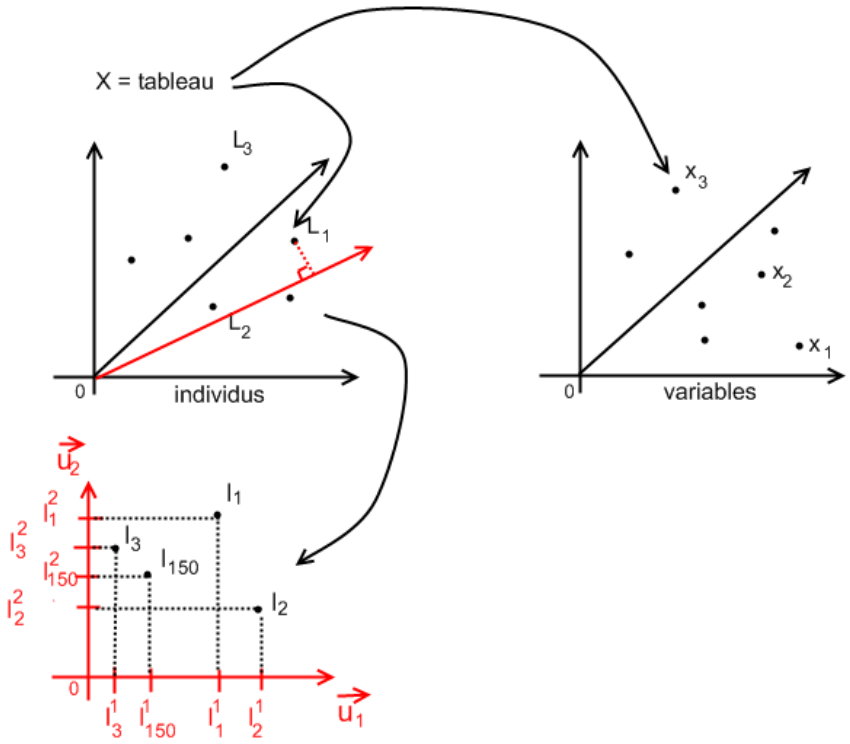


FIGURE 3.8 – Lien entre le tableau X et les nuages Individus et Variables

— On dessine le nouveau nuage en 1D, 2D ou 3D des points individus
— Formule de transition :
  — à partir des $\vec{u_i}$ j'obtiens les $\vec{v_i}$
  — ${}^tX \cdot v_i =$ abscisses des points Variables sur $\vec{v_i}$

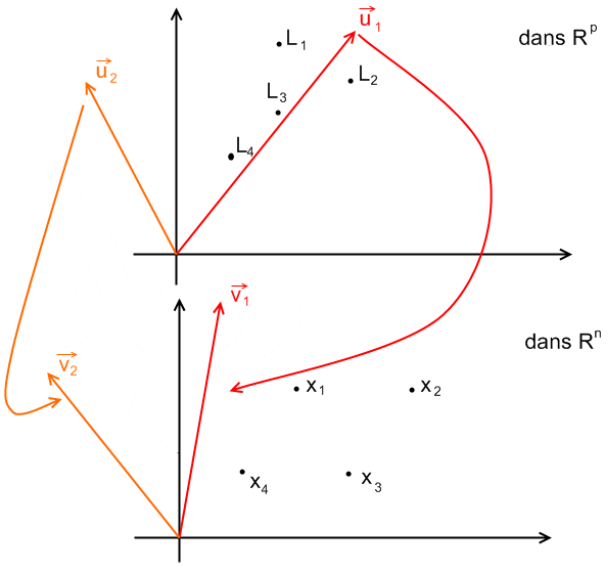


FIGURE 3.9 – Transition $\vec{u_i}$ vers $\vec{v_i}$

$$Ol_i = \alpha V_1^i + \beta V_2^i + \ldots + \gamma V_p^i$$

3.3.4 Qualité des représentations obtenues

3.3.4.1 Indice global de qualité

L'ACP utilise l'inertie comme mesure de la qualité des représentations.
On peut définir

$$\zeta_q = \frac{\sum_{\alpha=1}^{q} \lambda_\alpha}{\sum_{\alpha=1}^{p} \lambda_\alpha} \quad 0 \leq q \leq p$$

On a vu que

$$\begin{aligned} \text{inertie} &= \text{trace}(\underbrace{V}_{{}^tXDX}) \\ &= \sum_{\alpha=1}^{p} \lambda_\alpha \longleftarrow \text{ inertie du nuage initial} \end{aligned}$$

$$\rightarrow \zeta_1 = \frac{\overbrace{\lambda_1}^{\text{1er plan. Inertie du nuage obtenu sur l'axe 1}} + \overbrace{\lambda_2}^{\text{2ème axe}}}{\boxed{\lambda_1 + \lambda_2 + \ldots + \lambda_p}}$$

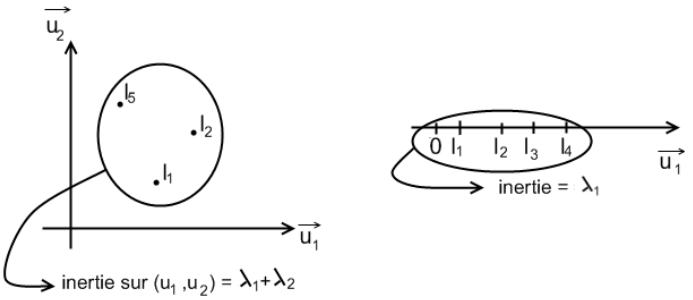


FIGURE 3.10 – Indice global de qualité

ζ_q représente la perte (ou conservation) d'information qui a lieu lors de la projection. ζ_q est aussi appelé taux d'inertie total expliqué.

3.3.4.2 Indice ponctuel de qualité

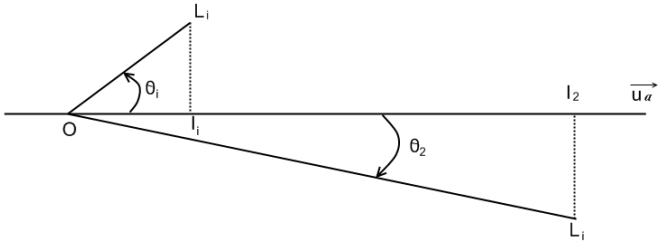


FIGURE 3.11 – Indice ponctuel de qualité

On a, via produit scalaire, $\|Ol_i\| = \|OL_i\| \cos\theta_i$. On met au carré et on obtient $\cos^2\theta_i = \frac{\|Ol_i\|^2}{\|OL_i\|^2}$. $cos^2\theta_i$ est appelé indice ponctuel de qualité de la représentation d'un individu L_i sur l'axe u_α. Plus cet indice est proche de 1, mieux est représenté l'individu sur l'axe u_α.

On peut calculer la qualité de la représentation d'un individu dans un sous espace :

$$\text{Qualité } (L_i) = \sum_{\alpha=1}^{q} \left[\cos^2\theta_i\right]$$

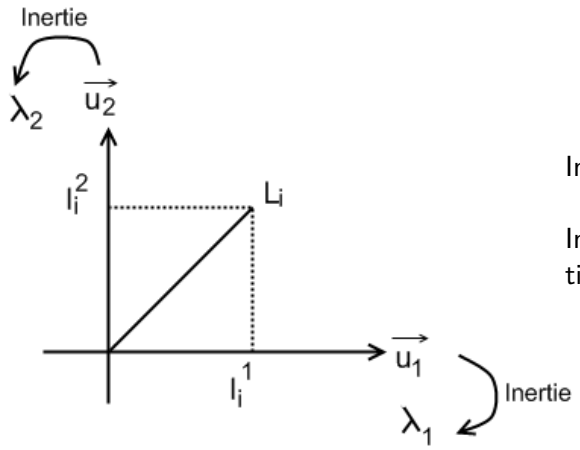
$\rightarrow$ Il faut Qualité (L_i) proche de 80% pour une bonne interprétation.

3.3.4.3 Contribution d'un point

L'inertie expliquée par l'axe u_α est λ_α. On a donc la relation :

$$\lambda_\alpha = \text{Inertie} = \sum_{i=1}^{n} m_i \left(OL_i^\alpha\right)^2$$

Formule de l'inertie appliquée au nuage projeté sur $\vec{u_\alpha}$.



Inertie sur le plan $(\vec{u_1}, \vec{u_2}) = \lambda_1 + \lambda_2$

Inertie du nuage initial avant projection $= \lambda_1 + \lambda_2 + \ldots + \lambda_p$

La contribution du point L_i à cette inertie peut s'exprimer par

$$\frac{m_i \left(OL_i^\alpha\right)^2}{\lambda_\alpha} = \text{ctr}_\alpha(L_i)$$

$$0 \leq \text{ctr}_\alpha(L_i) \leq 1$$

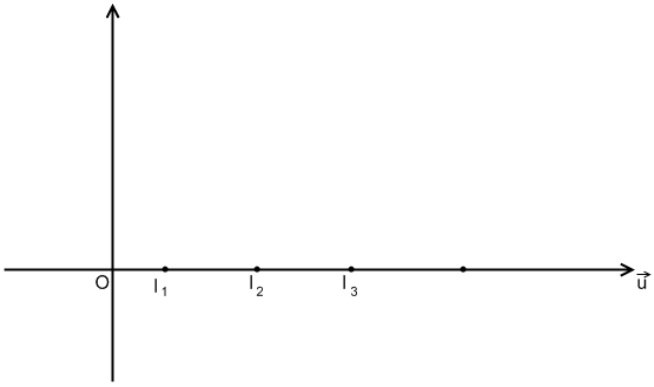


FIGURE 3.12 – $\text{ctr}_\alpha(L_3) = 0.95$

— Les points les plus éloignés de l'origine sont ceux qui contribuent le plus.
Un point ayant une contribution trop élevée peut être un facteur d'instabilité ($\geq 50\%$).
— l'axe trouvé représente bien un point et mal tous les autres ;
— il peut être intéressant de refaire l'ACP sans ce point.

3.3.5 Reconstruction du tableau X

— Abscisse des n points individus sur l'axe $u_\alpha = \underline{X} \cdot \underline{u_\alpha}$
— Abscisse des n points Variables sur l'axe $v_\alpha = {}^t\underline{X} \cdot \underline{v_\alpha}$

On aimerait trouver une relation entre $\underline{X}$ et $\underline{u_\alpha}$, v_α, λ_α.
$\Rightarrow$ On peut montrer que

$$\underline{X}_{(n\times p)} = \sum_{\alpha=1}^{p} \left(\sqrt{\lambda_\alpha}\, v_{\alpha_{(n\times 1)}} \cdot \left({}^t u_{\alpha(1\times p)}\right)\right)$$

Utilisation de cette formule : théorème de Eckart Young

On peut utiliser cette formule pour obtenir une reconstruction approchée de $\underline{X}$ nommée $\underline{X}^*$

$$\underline{X}^* = \sum_{\alpha=1}^{q} \left[\sqrt{\lambda_\alpha} v_\alpha \cdot \left({}^t u_\alpha\right)\right] \qquad \text{avec } q \leq p$$

$\underline{X}^*$ est la meilleure approximation de $\underline{X}$ au sens des moindres carrés.

$$\underline{X}^* = \sqrt{\lambda_1} \cdot \underset{(1000\times 1)}{v_1} \cdot \begin{pmatrix} {}^t u_1 \\ {}_{(1\times 10)} \end{pmatrix} \qquad \text{pour } q = 1$$

CHAPITRE 4

L'ACP, l'ACP centrée et l'ACP normée

Pour l'instant, on a vu l'ACP générale qui possède certaines limitations car elle ne tient pas compte de *l'hétérogénéité* des données concernant aussi bien : $\begin{cases} \text{les moyennes des variables} \\ \text{la dispertion des variables} \end{cases}$

Population	% CO_2
10 000	0.58
$\vdots$	$\vdots$
10 000	0.7

4.1 Prise en compte de la dispersion des moyennes

On part d'une matrice de données brutes : tableau initial $= \underline{R}_{(n\times p)}$.
Il est possible de centrer les variables

$$\begin{aligned} x_{ij} &= (r_{ij} - \bar{r_j}) \\ \Rightarrow \bar{x_j} &= 0 \quad \forall j \end{aligned}$$

On peut aussi associer un poids aux individus. En général, on prend $m_i = \frac{1}{n}$.
Cela revient à faire :

$$x_{ij} = \frac{(r_{ij} - \bar{r_j})}{\sqrt{n}}$$

De manière à obtenir :

$$ {}^t\underline{XX} = \begin{array}{c} \\ j \end{array} \begin{pmatrix} & l & \\ & \vdots & \\ \cdots & \underbrace{\frac{1}{n}\sum_{i=1}^{n}(r_{ij} - \bar{r_j})(r_{il} - \bar{r_l})}_{\text{cov}(\text{Var}_j, \text{Var}_l)} & \end{pmatrix} = \text{On a modifié } \underline{R} \text{ pour faire en sorte que } {}^t\underline{XX} \text{ soit la matrice des variances - covariances}$$

$$\Rightarrow \textit{ACP.}$$

$${}^tR \times \underbrace{\begin{pmatrix} \frac{1}{n} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{n} \end{pmatrix}}_{D} \times R \qquad {}^tXX = \begin{pmatrix} \text{Var}(1) & \text{cov}(1,2) & \cdots & \text{cov}(1,p) \\ & \text{Var}(2) & & \vdots \\ & & \ddots & \\ & & & \text{Var}(p) \end{pmatrix}$$

$$\text{cov}\,(\text{Var}_j, \text{Var}_j) = \text{variance}\,(\text{variance Var}_j)$$

$$\underline{R} = \begin{array}{c} \\ \text{toto} \\ \text{titi} \\ \vdots \\ \text{tata} \end{array} \begin{array}{c} \begin{array}{cc} \text{poids} & \text{taille} \end{array} \\ \begin{pmatrix} 80 & 1.8 \\ 70 & 1.6 \\ \vdots & \vdots \\ 65 & 1.7 \\ 70 & 1.6 \end{pmatrix} \end{array} \qquad \underline{X} = \begin{pmatrix} \frac{10}{\sqrt{5}} & \frac{0.2}{\sqrt{5}} \\ \frac{0}{\sqrt{5}} & \frac{0}{\sqrt{5}} \\ \vdots & \vdots \\ \frac{-5}{\sqrt{5}} & \frac{0.1}{\sqrt{5}} \end{pmatrix}$$

$$\begin{array}{c} \\ \text{toto} \\ \text{titi} \\ \vdots \\ \text{tata} \end{array} \begin{array}{c} \begin{array}{cc} \text{poids} & \text{taille} \end{array} \\ \begin{pmatrix} \frac{10}{\sqrt{5}} & \frac{0.2}{\sqrt{5}} \\ \frac{0}{\sqrt{5}} & \frac{0}{\sqrt{5}} \\ \vdots & \vdots \\ \frac{-5}{\sqrt{5}} & \frac{0.1}{\sqrt{5}} \end{pmatrix} \end{array}$$

$${}^t\underline{X}\underline{X} = \begin{array}{c} \\ \text{poids} \\ \text{taille} \end{array} \begin{array}{c} \begin{array}{cccc} \text{toto} & \text{titi} & \cdots & \text{tata} \end{array} \\ \begin{pmatrix} \frac{10}{\sqrt{5}} & \frac{0}{\sqrt{5}} & \cdots & \frac{-5}{\sqrt{5}} \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix} \end{array} \begin{array}{c} \begin{array}{cc} \text{poids} & \text{taille} \end{array} \\ \begin{pmatrix} \text{variance(poids)} & \text{cov(poids, taille)} \\ \text{cov(poids, taille)} & \text{variance(taille)} \end{pmatrix} \end{array}$$

4.2 Prise en compte de l'hétérogénéité des valeurs (ecart-types)

— De maniere similaire a auparavant, on peut poser $x_{ij} = \frac{1}{\sqrt{n}} \times \frac{(r_{ij} - \bar{r_j})}{\sigma_j}$ afin de normer les valeurs des variables (division par écart-type de la variable j).

— Les termes contenus dans ${}^t\underline{X}\underline{X}$ sont égaux à

$$\frac{1}{n} \sum_{i=1}^{n} \left(\frac{(r_{ij} - \bar{r_j})}{\sigma_j} \frac{(r_{ij} - \bar{r_j})}{\sigma_l} \right) = t_{jl}$$

$${}^t\underline{X}\underline{X} = \begin{array}{c} \\ j \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{c} l \end{array} \\ \begin{pmatrix} 1 & \vdots & \\ \cdots & t_{jl} & \\ & & 1 \end{pmatrix}_{p \times p} \end{array}$$

Symétrique. $t_{jl} \in [-1; 1]$. Que des 1 sur la diagonale.

${}^t\underline{X}\underline{X}$ est la matrice des corrélations entre Variables. t_{jl} est le coefficient de corrélation $(\text{Var}_j, \text{Var}_l)$.

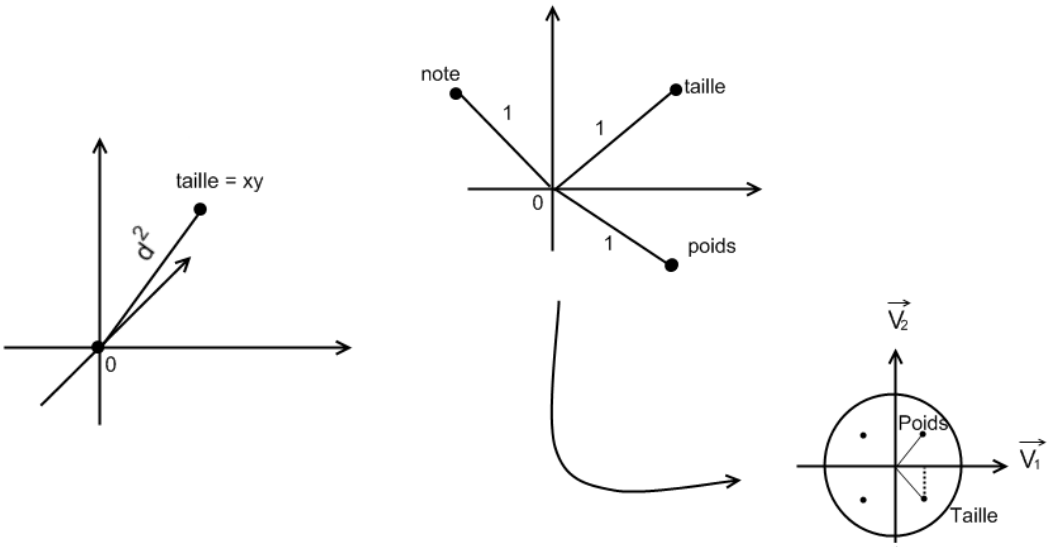
Inertie du nuage avant projection (avec des variables normées) :

$$\text{trace}\left({}^t\underline{X}\underline{X}\right) = p = \text{Nombre de variable}$$

$\Rightarrow$ *ACP normée*, que l'on fait la plupart du temps.

4.3 Quelques détails sur l'ACP normée

On avait $\underline{R}$, on norme les Variables, on obtient $\underline{X}$.



$$\begin{aligned} d^2\left(O, \underline{X}_j\right) &= \sum_{i=1}^{n} (x_{ij} - 0)^2 \\ &= \sum i = 1^n \left(\frac{r_{ij} - \bar{r_j}}{\sqrt{n} \cdot \sigma_j}\right)^2 \qquad (\text{variance} = \sigma^2) \\ &= 1 \end{aligned}$$

Le nuage des points Variables est très particulier. Tous les points Variables sont sur une sphère de rayon 1 et de centre O. Les projections de ces points sur un plan sont toutes dans un cercle de rayon 1 : *cercle des corrélations*.
Les points proche de O sont mal représentés.

- L'angle (le cosinus de l'angle) entre deux variables $\underline{x}_j$ et $\underline{x}_k$ est égal au coefficient de corrélation entre ces variables.
- L'abscisse d'un point Variable sur un axe factoriel donne le coefficient de corrélation entre la variable initial et *la composante principale* (variable synthétique créée, combinaison linéaire des variables initiales $= X \cdot u_\alpha$) représentée par cet axe.
  $\rightarrow$ Utilse pour donner un sens aux axes factoriels.

4.4 Interprétation du nuage des individus (après projection)

- Le sens donné aux axes factoriels obtenus lors de la projection des points Variables permet une interprétation du nuage des points individus.
- La contribution d'un individu à l'inertie expliquée par un axe est proportionnelle au carré de sa coordonnée sur cet axe.

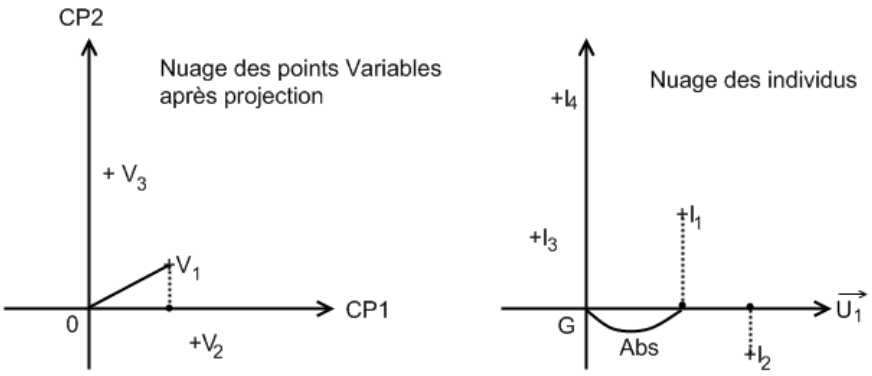


FIGURE 4.1 – Nuage des points Variables et nuage des individus

$$\text{ctr}_\alpha = \underbrace{\frac{1}{n}}_{m_i \text{ en ACP normée}} \frac{(Ol_i)^2_\alpha}{\lambda_\alpha}$$

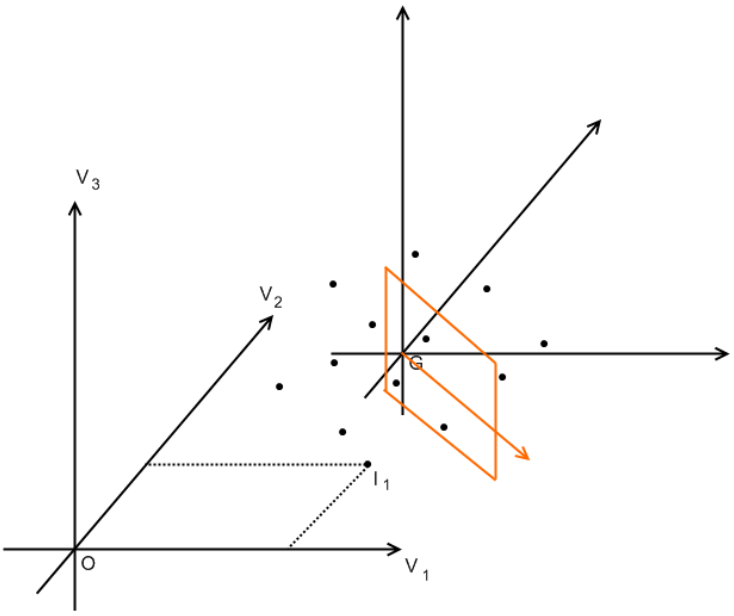


FIGURE 4.2 – Changement de repère

$$X = \begin{array}{c} \\ I_1 \\ \\ G \end{array} \begin{array}{c} \begin{array}{ccc} V_1 & V_2 & V_3 \end{array} \\ \begin{pmatrix} \square\square\square & & \\ & & \\ & & \end{pmatrix} \end{array}$$

— Un point individu proche de l'origine (G) du repère aura des valeurs proches de la moyenne de la population.

— Inversement, un individu éloigné de l'origine aura des valeurs éloignées de la moyenne de la population pour ce qui concerne les variables représentées par cet axe.

Index

Analyse de données

Departement Informatique
4emeannee
2016-2017

Cours - Partie 1

Résumé : Cours d'analyse de données

Mots clefs : analyse de données

Abstract:

Keywords:

Encadrants

Jean-Yves RAMEL
jean-yves.ramel@univ-tours.fr

Université Francois-Rabelais, Tours

Auteurs

Jean-Yves RAMEL
jean-yves.ramel@univ-tours.fr
+ Eleves de la promo 2008-2009