Chapitre 3 : Analyse en Composantes Principales

M1 du Master MMAS

James Ledoux

Dépt de mathématiques, Univ. Poitiers

12 juillet 2009

Introduction

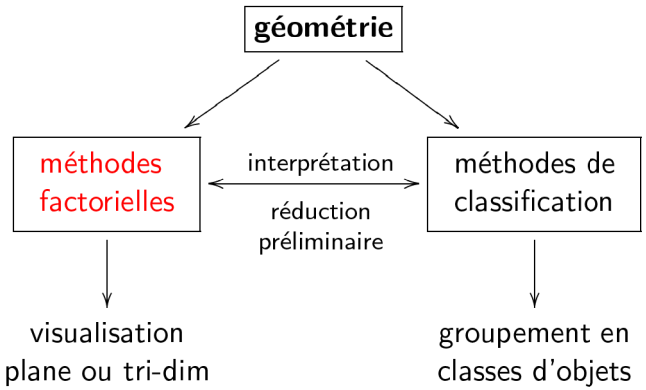
Tableau rectangulaire : individus x variables

- lignes $(i = 1, \ldots, n)$: représentent les n individus ou observations
- colonnes $(j = 1, \ldots, p)$: représentent les p variables

1. p **variables quantitatives** : p mesures numériques pour chaque individu
2. p **variables qualitatives** : p modalités pour chaque individu

   si $p = 2$: **table de contingence** croisant deux variables qualitatives
   - lignes $(i = 1, \ldots, n)$: représentent les n modalités d'une variable qualitative
   - colonnes $(j = 1, \ldots, p)$: représentent les p modalités d'une variable qualitative
3. Cas mixte

Présenter, résumer, structurer les données



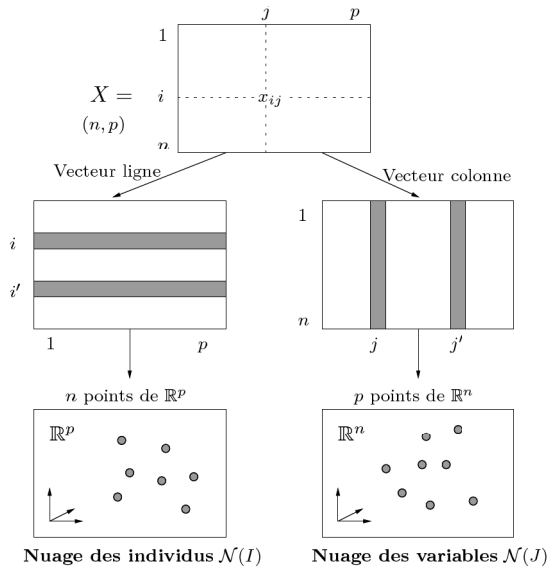
Cas de p variables quantitatives

Tableau : n individus $\times$ p variables

- **lignes** $\equiv$ individus, objets, observations, ...
- **colonnes** $\equiv$ variables quantitatives i.e. à valeurs numériques continues

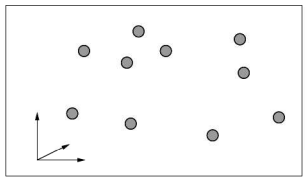
	Denrées alimentaires							
	PAO	PAA	VIO	VIA	POT	LEC	RAI	PLP
AGRI	167	1	163	23	41	8	6	6
SAAG	162	2	141	12	40	12	4	15
PRIN	119	6	69	56	5	5	13	41
CSUP	87	11	63	111	27	3	18	39
CMOY	103	5	68	77	32	4	11	30
EMPL	111	4	72	66	34	6	10	28
OUVR	130	3	76	52	43	7	7	16
INAC	138	7	117	74	53	8	12	20

TABLE 5: Reprise de l'Exemple 9 : table des moyennes annuelles de consommation en frs pour 8 catégorie socio-professionnelles (CSP)



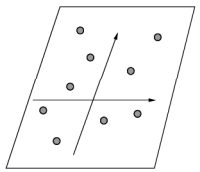
- Définir une distance entre points d'un même nuage :
  - distance entre individus,
  - distance entre variables
- Évaluer les proximités géométriques entre points-lignes et entre points-colonnes.
  - ➔ associations statistiques entre les individus
  - ➔ associations statistiques entre les variables

Analyser le tableau de distances
associé à chaque nuage



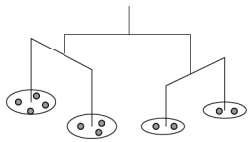
nuage de points

méthodes factorielles



Visualisation dans le meilleur espace **réduit**

méthodes de classification



Regroupements dans tout l'espace

Méthode factorielle

- **Objectif :**
  Fournir des représentations synthétiques d'un grand ensemble de données
- **Entrées :**
  Nuage de points et distance entre ces points
- **Sorties :**
  En général, représentations graphiques planes

Méthodes factorielles

Techniques de réduction de l'espace de visualisation à une représentation plane respectant au mieux les proximités géométriques

Espaces factoriels

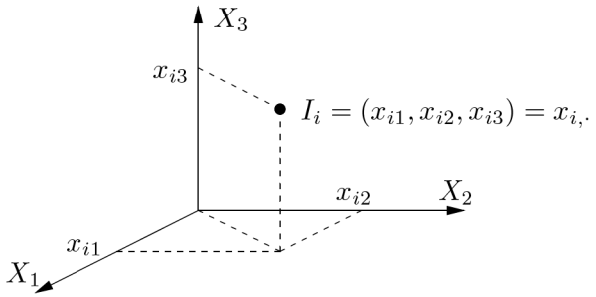
Sous-espaces de dimension 2 (ou 3) qui ajustent au mieux le nuage $\mathcal{N}(I)$ ou le nuage $\mathcal{N}(J)$.

Analyse en composantes principales (ACP)

- p variables X_j sont observées sur n individus I_i

$$p >> 1 \qquad n >> p$$

- **Proximité entre individus** $\equiv$ similitudes globales des grandeurs observées sur les individus
- **Proximité entre variables** $\equiv$ corrélation
- **Objectifs**
  - ➧ Évaluer la ressemblance entre individus
    - ➥ Typologie des individus
  - ➧ Évaluer la liaison entre variables
    Calcul d'un petit nombre de variables synthétiques résumant au mieux les liaisons affines entre les variables originales
    - ➥ Composantes principales



La variable X_j est identifiée à
- la colonne j du tableau de données X
- la « direction » N° j du repère orthonormé de $\mathcal{N}(I)$

- **Mesure de ressemblance entre individus :**

$$d_2(I_i, I_k)^2 = \|x_{i,\cdot} - x_{k,\cdot}\|_2^2 = \sum_{j=1}^{p} (x_{ij} - x_{kj})^2 \tag{23}$$

Remarque : tendance à avantager les variables très dispersées

- **Dans toute la suite**, poids de chaque individu : $\boxed{p_i = \frac{1}{n}}$
  Si « l'individu » représente une classe d'individus

$$0 < p_i < 1 \quad \text{avec} \quad \sum_i p_i = 1$$

- **Centre de gravité du nuage**

$$OG = \sum_{i=1}^{n} \frac{1}{n} OI_i \quad \text{➧} \quad G = (\overline{X}_1, \ldots, \overline{X}_p)$$

- On **centre toujours** le nuage $\mathcal{N}(I)$
  - Géométriquement : $O \to G$ origine du nouveau nuage
  - Au niveau variable : Variable X_j, $j = 1, \ldots, p$ $\Big| \to \Big|$ $X_j - \overline{X}_j \mathbf{1}_n$, $j = 1, \ldots, p$ variables centrées
  - Au niveau tableau des données

$$i = 1, \ldots, n : \qquad x_{ij} - \overline{X}_j \quad j = 1, \ldots, p$$

  - ne modifie pas la distance entre les individus

	Variables centrées : $Y_j := X_j - \overline{X}_j \mathbf{1}$							
	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8
I_1 : AGRI	(39.9	-3.9	66.9	-35.9	6.6	1.4	-4.1	-18.4)
I_2 : SAAG	(34.9	-2.9	44.9	-46.9	5.6	5.4	-6.1	-9.4)
I_3 : PRIN	(-8.1	1.1	-27.1	-2.9	-29.4	-1.6	2.9	16.6)
I_4 : CSUP	(-40.1	6.1	-33.1	52,1	-7.4	-3.6	7.9	14.6)
I_5 : CMOY	(-24.1	0.1	-28.1	18.1	-29.4	-2.6	0.9	5.6)
I_6 : EMPL	(-16.1	-0.9	-24.1	7.1	-0.4	-0.6	-0.1	3.6)
I_7 : OUVR	(2.9	-1.9	-20.1	-6.9	8.6	0.4	-3.1	-8.4)
I_8 : INAC	(10.9	2.1	20.9	15.1	18.6	1.4	1.9	-4.4)
G	(0	0	0	0	0	0	0	0)
	($\overline{C}_1$	$\overline{C}_2$	$\overline{C}_3$	$\overline{C}_4$	$\overline{C}_5$	$\overline{C}_6$	$\overline{C}_7$	$\overline{C}_8$)
Ecart-type	(26.1	2.98	36.29	29.29	13.32	2.64	4.17	11.46)

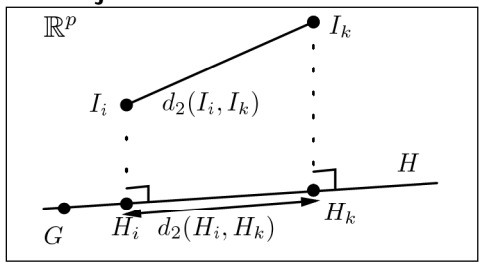
TABLE 6: Nuage centré des individus $\mathcal{N}(I)$ pour la Table 5

Ajuster le nuage des 8 individus par une droite, puis par un plan,
...
représentant le plus fidèlement la proximité entre individus

- **Projection orthogonale** du nuage des individus sur un sous-espace H
- Nuage projeté devra restituer le plus fidèlement les distances (ou ressemblances) entre individus existant dans le nuage d'origine

Cas où H est une droite

Projection réduit les distances



Rechercher une droite H qui maximise :

$$H \mapsto \sum_{i=1}^{n}\sum_{k=1}^{n} d_2(H_i, H_k)^2$$

- **Simplification du problème :** on a la relation

$$\sum_{i=1}^{n}\sum_{k=1}^{n} d_2(H_i, H_k)^2 = 2n \sum_{i=1}^{n} d_2(H_i, G_H)^2$$

$$\Longrightarrow \quad \max_H \left(\sum_{i=1}^{n}\sum_{k=1}^{n} d_2(H_i, H_k)^2 \right) \Longleftrightarrow \boxed{\max_H \left(\sum_{i=1}^{n} d_2(G_H, H_i)^2 \right)}$$

On dit alors que l'on maximise l'inertie du nuage projeté sur la droite H

Définition 18 (Inertie)

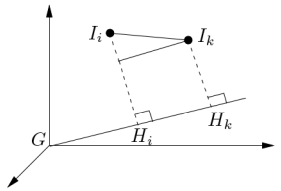
L'inertie du nuage *est définie par*

$$I(\mathcal{N}) = \sum_{i=1}^{n} \frac{1}{n}\, d_2(G, I_i)^2.$$

L'inertie de la projection du nuage sur le sous-espace H *est définie par*

$$I_H(\mathcal{N}) = \sum_{i=1}^{n} \frac{1}{n}\, d_2(G_H, H_i)^2$$

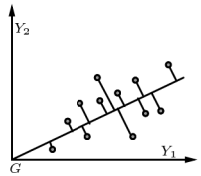
Interprétation de la droite « solution » : $G \equiv G_H \equiv O$



Pythagore : $\sum_i d_2(G, H_i)^2 = \sum_i d_2(G, I_i)^2 - \sum_i d_2(I_i, H_i)^2$

$$\max \sum_i d_2(G, H_i)^2 \Longleftrightarrow \boxed{\min \sum_i d_2(I_i, H_i)^2}$$

$\Longrightarrow$ on recherche la **droite d'allongement maximum** du nuage



Remarque : noter la différence avec le critère utilisé pour la construction de la droite de régression

- **Ajustement du nuage $\mathcal{N}(I)$ par un sous-espace de dimension ≥ 2 :**
  1. La première droite comme précédemment
  2. La seconde droite est choisie, **orthogonale à la première**, de manière à capturer la « seconde » direction d'allongement.
  3. La troisième, **orthogonale aux deux premières**, est choisie de manière à capturer la « troisième » direction d'allongement
  4. ...

Mise en oeuvre

- **Réduction des données**
  - Influence des unités de mesure sur les valeurs des distances
  - Trop de disparité entre les écarts-type

$\Longrightarrow$ **en général, on réduit** le tableau :

$$z_{ij} := \frac{x_{ij} - \overline{X}_j}{s_{X_j}} \quad \forall (i,j) \in I \times J$$

	Variables centrées-réduites							
	Z_1^{PAO}	Z_2^{PAA}	Z_3^{VIO}	Z_4^{VIA}	Z_5^{POT}	Z_6^{LEC}	Z_7^{RAI}	Z_8^{PLP}
AGRI	(1,53	-1,30	1,84	-1,22	0,50	0,52	-0,99	-1,60)
SAAG	(1,34	-0,97	1,24	-1,60	0,42	2,03	-1,47	-0,82)
PRIN	(-0,31	0,38	-0,75	-0,10	-2,20	-0,61	0,69	1,45)
CSUP	(-1,54	2,06	-0,91	1,78	-0,55	-1,37	1,89	1,28)
CMOY	(-0,92	0,04	-0,78	0,62	-0,18	-0,99	0,21	0,49)
EMPL	(-0,62	-0,29	-0,66	0,24	-0,03	-0,24	-0,03	0,32)
OUVR	(0,11	-0,63	-0,55	-0,23	0,65	0,14	-0,75	-0,73)
INAC	(0,42	0,71	0,58	0,52	1,40	0,52	0,45	-0,38)
G	(0	0	0	0	0	0	0	0)
Écart-type	(1	1	1	1	1	1	1	1)

TABLE 7: Nuage centré-réduit des individus $\mathcal{N}(I)$ pour Table 5

- On travaille sur le tableau centré-**réduit** : $Z = (z_{ij})$:
  ACP normée
- On reprend la distance entre individus :

$$d_2(I_i, I_k)^2 = \sum_{j=1}^{p} (z_{ij} - z_{kj})^2$$

- Ajustement par le sous-espace H :

Espace H réalisant $\max \left(\sum_{i=1}^{n} \frac{1}{n} d_2(G, H_i)^2 \right)$

ou encore donnant l'inertie du nuage projeté maximale

- Pour une droite H, on vérifie que

$$I_H(\mathcal{N}) = u_1^\top \left[\frac{1}{n} Z^\top Z \right] u_1 \tag{24}$$

où u_1 est un vecteur directeur unitaire de H

Déterminer une droite H tel que $I_H(\mathcal{N})$ soit maximale

$\Longleftrightarrow$ déterminer le vecteur $u_1 \in \mathbb{R}^p$ unitaire réalisant $\max u_1^\top \left[\frac{1}{n} Z^\top Z \right] u_1$

- **Matrice des corrélations :** $R = (\rho_{x_i,x_j})_{i,j=1}^p$
  On peut vérifier que

$$\left[\frac{1}{n} Z^\top Z \right]_{ij} = \sum_{k=1}^{n} \frac{1}{n} \frac{(x_{ki} - \overline{X}_i)}{s_{X_i}} \frac{(x_{kj} - \overline{X}_j)}{s_{X_j}} = \rho_{X_i,X_j} = R_{ij}$$

Remarque : Notons que si les variables sont seulement centrées, on a $I_H(\mathcal{N}) = u_1^\top \left[\frac{1}{n} Y^\top Y \right] u_1 = u_1^\top (s_{X_i,X_j})_{i,j=1}^p u_1$

	PAO	PAA	VIO	VIA	POT	LEC	RAI	PLP
PAO	1,00							
PAA	-0,77	1,00						
VIO	0,93	-0,6	1,00					
VIA	-0,91	0,9	-0,75	1,00				
POT	0,66	-0,33	0,52	-0,42	1,00			
LEC	0,89	-0,67	0,79	-0,84	0,6	1,00		
RAI	-0,83	0,96	-0,67	0,92	-0,41	-0,82	1,00	
PLP	-0,86	0,77	-0,83	0,72	-0,55	-0,75	0,83	1,00

TABLE 8: Matrice des corrélations pour la Table 5

- R **symétrique, semi-définie positive** alors :

Théorème 3

1. $\mathbb{R}^p$ *admet une base orthonormée de vecteurs propres de* R
2. *Les valeurs propres de* R *sont positives* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$

Résolution du problème avec $\dim H = q$

Théorème 4

Le sous-espace $H^{(q)}$ de dimension $q \leq p$ de $\mathbb{R}^p$ portant l'inertie maximale se calcule par

$$H^{(q-1)} \oplus H_q$$

où H_q est la droite orthogonale à $H^{(q-1)}$ portant l'inertie maximale.

Théorème 5

Le sous-espace $H^{(q)}$ ajustant au mieux le nuage est engendré par les q premiers vecteurs propres (orthonormés) de R correspondants aux q plus grandes valeurs propres

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q$$

Axes factoriels

$$R \begin{bmatrix} u_1 & u_2 & \dots & u_l & \cdots & u_p \\ \downarrow & \downarrow & & \downarrow & & \downarrow \\ \lambda_1 & \geq \lambda_2 & \geq \dots \geq & \lambda_l & \geq \cdots \geq & \lambda_p \end{bmatrix}$$

- L'axe factoriel de rang l du nuage des individus, est la droite de vecteur directeur unitaire le vecteur propre normé u_l associé à la valeur propre de rang l, λ_l, de la matrice des corrélations R
- Un plan factoriel : plan engendré par deux axes factoriels (en général de rangs consécutifs)
- L'ensemble de **tous** les axes factoriels définit un changement de repère orthonormé / repère d'origine du nuage centré $\mathcal{N}(I)$

Rappel : le centre de gravité $G = (\overline{X}_1, \overline{X}_2, \dots, \overline{X}_p)$

↪ un pseudo-individu possédant des caractéristiques moyennes / à l'ensemble des variables

- Le premier axe factoriel u_1 capture **la principale direction d'allongement** du nuage des individus
  1. les individus caractéristiques de cette direction sont les plus loin du centre de gravité
     ↪ individus « hors-normes »/ certaines variables
  2. les individus proches en projection sur cet axe devraient être proches dans le nuage original
     ↪ individus ressemblants / certaines variables

- u_2 est la **principale direction d'allongement orthogonale à la première**
  ↪ Information **résiduelle** / premier axe
  1. les individus caractéristiques de cette direction sont les plus loin du centre de gravité dans une direction **orthogonale à** u_1
     ↪ individus « hors-normes »/ certaines variables
  2. les individus proches en projection sur cet axe devraient être proches dans le nuage original
     ↪ individus ressemblants / certaines variables
- u_3 est la **principale direction d'allongement orthogonale aux deux premières**
  ↪ Information résiduelle / 2 premiers axes
- ...

Composantes principales

Rappel : les axes du repère original du nuage $\mathcal{N}(I)$ s'identifiaient aux p variables $X_1, \ldots, X_p$

- L'axe déterminé par le vecteur unitaire u_l devrait correspondre à une nouvelle variable C_l
- Une variable C_l n'est connue qu'à travers ses valeurs pour les n individus
  ➥ **Projection des n individus sur u_l :**

$$C_l = Z\,u_l \qquad l = 1, \ldots, p$$

Définition 19 (Composante principale)

*Le vecteur de $\mathbb{R}^n$, $C_l := Z\,u_l$, est appelé la **composante principale de rang** l.*

$$C_l = Z\,u_l \quad l = 1, \ldots, p \Longleftrightarrow \boxed{C_l = \sum_{j=1}^{p} u_l(j)\; Z_j}$$

- C_l est une combinaison linéaire des variables originales
- La variable C_l est centrée : $\overline{C}_l = 0$.
- La variance de C_l : on vérifie que

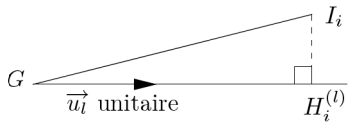
$$s^2_{C_l} = {u_l}^\top R\, u_l = ({u_l}^\top u_l)\lambda_l$$

$$\boxed{s^2_{C_l} = \lambda_l}$$

$$\begin{aligned}
C_1 &= \tfrac{1}{\sqrt{\lambda_1}} \Big[-0,97\, Z^{\mathrm{PAO}} + 0,87\, Z^{\mathrm{PAA}} - 0,87\, Z^{\mathrm{VIO}} + 0,93\, Z^{\mathrm{VIA}} \\
&\quad -0,61\, Z^{\mathrm{POT}} - 0,91\, Z^{\mathrm{LEC}} + 0,93\, Z^{\mathrm{RAI}} + 0,90\, Z^{\mathrm{PLP}} \Big] \\
C_2 &= \tfrac{1}{\sqrt{\lambda_2}} \Big[-0,13\, Z^{\mathrm{PAO}} - 0,41\, Z^{\mathrm{PAA}} - 0,19\, Z^{\mathrm{VIO}} - 0,24\, Z^{\mathrm{VIA}} \\
&\quad -0,70\, Z^{\mathrm{POT}} - 0,12\, Z^{\mathrm{LEC}} - 0,31\, Z^{\mathrm{RAI}} + 0,05\, Z^{\mathrm{PLP}} \Big] \\
C_3 &= \tfrac{1}{\sqrt{\lambda_3}} \Big[0,10\, Z^{\mathrm{PAO}} + 0,21\, Z^{\mathrm{PAA}} + 0,44\, Z^{\mathrm{VIO}} + 0,05\, Z^{\mathrm{VIA}} \\
&\quad -0,36\, Z^{\mathrm{POT}} + 0,02\, Z^{\mathrm{LEC}} + 0,16\, Z^{\mathrm{RAI}} - 0,10\, Z^{\mathrm{PLP}} \Big] \\
C_4 &= \tfrac{1}{\sqrt{\lambda_4}} \Big[0,07\, Z^{\mathrm{PAO}} + 0,12\, Z^{\mathrm{PAA}} - 0,02\, Z^{\mathrm{VIO}} - 0,22\, Z^{\mathrm{VIA}} \\
&\quad -0,04\, Z^{\mathrm{POT}} + 0,29\, Z^{\mathrm{LEC}} + 0,04\, Z^{\mathrm{RAI}} + 0,39\, Z^{\mathrm{PLP}} \Big] \\
C_5 &= \tfrac{1}{\sqrt{\lambda_5}} \Big[0,12\, Z^{\mathrm{PAO}} - 0,11\, Z^{\mathrm{PAA}} + 0,10\, Z^{\mathrm{VIO}} - 0,14\, Z^{\mathrm{VIA}} \\
&\quad +0,07\, Z^{\mathrm{POT}} - 0,27\, Z^{\mathrm{LEC}} + 0,11\, Z^{\mathrm{RAI}} + 0,14\, Z^{\mathrm{PLP}} \Big]
\end{aligned}$$

TABLE 9: Les 5 premières CP en fonction des variables originales centrées-réduites

- Plans factoriels ≡ **images approchées** d'un nuage de points
  ? Indicateurs de la qualité de ces images : globaux (pour le nuage)
  locaux (pour « l'individu »)
- **QLT de représentation d'un individu I_i par sa projection sur l'axe factoriel u_l**



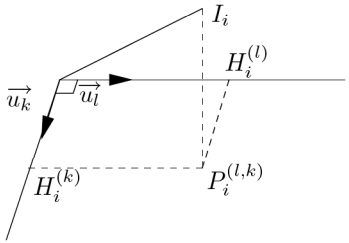
$$\mathrm{QLT}_l(i) \;=\; \frac{d_2(G, H_i^{(l)})^2}{d_2(G, I_i)^2} = \cos(GI_i, u_l)^2$$

Remarque :

$$\mathrm{QLT}_l(i) \;=\; \frac{p_i\, d_2(G, H_i^{(l)})^2}{p_i\, d_2(G, I_i)^2} = \frac{\text{Inertie projetée de } I_i \text{ sur } u_l}{\text{Inertie de } I_i}$$

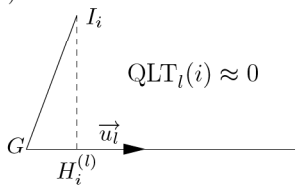
- **QLT de représentation de I_i par sa projection sur un plan (u_l, u_k)**

$$\begin{aligned}\mathrm{QLT}_{l,k}(i) &= \frac{d_2(G, P_i^{(l,k)})^2}{d_2(G, I_i)^2} \\ &= \frac{d_2(G, H_i^{(l)})^2 + d_2(G, H_i^{(k)})^2}{d_2(G, I_i)^2} \\ &= \mathrm{QLT}_l(i) + \mathrm{QLT}_k(i)\end{aligned}$$

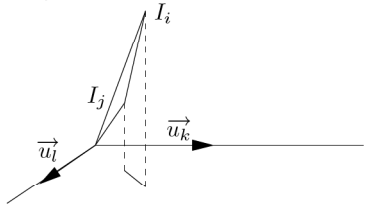


Remarque : $\mathrm{QLT}_{l,k}(i) = \dfrac{\text{Inertie projetée de } I_i \text{ sur } (u_l, u_k)}{\text{Inertie de } I_i}$

- **Plus** $\mathrm{QLT}(i)$ est proche de 1 **plus** le point individu est proche de l'axe ou du plan et alors :
  - la distance de l'individu à G (origine) est lisible sur sa projection sur l'axe
    Elle ne l'est pas si $\mathrm{QLT}(i) \approx 0$



  - la distance entre deux points d'un plan factoriel ne traduit leur distance réelle dans le nuage (i.e. leur ressemblance) que s'ils sont très bien représentés par leur projection sur le plan



- **Qualité de représentation du nuage $\mathcal{N}(I)$ par les $q \leq p$ premiers axes factoriels $u_1, \ldots, u_q$**

$$\begin{aligned}\mathrm{QLT}_{u_1,\ldots,u_q}(\mathcal{N}(I)) &= \frac{\text{Inertie de } \mathcal{N}(I) \text{ projeté sur les } q \text{ premiers axes}}{\text{Inertie de } \mathcal{N}(I)} \\ &= \frac{\sum_{l=1}^{q}\left[\sum_{i=1}^{n} p_i\, d_2(G, H_i^{(l)})^2\right]}{\text{Inertie de } \mathcal{N}(I)} \\ &= \frac{\sum_{l=1}^{q} u_l{}^{\top} R u_l}{\text{Inertie de } \mathcal{N}(I)} \text{ avec (24)} \\ &= \frac{\sum_{l=1}^{q}\lambda_l}{\sum_{l=1}^{p}\lambda_l} = \frac{\sum_{l=1}^{q}\lambda_l}{p} \qquad (25)\end{aligned}$$

Pourcentage d'inertie « expliquée » par les q premiers axes factoriels

Numéro	Valeur propre	Pourcentage d'inertie	Pourcentage cumulé
1	6,2079	77,60	77,60
2	0,8797	11,00	88,60
3	0,4160	5,20	93,79
4	0,3065	3,83	97,63
5	0,1684	2,11	99,73
6	0,0181	0,23	99,96
7	0,0034	0,04	100,00
8	0,0000	0,00	100,00
	8		

TABLE 10: Pourcentage d'inertie expliquée par chacun des axes factoriels

- L'axe factoriel u_l rend maximum
  (sous la contrainte d'être orthogonal aux précédents)
  l'inertie projetée du nuage sur une droite
- **Décomposition de cette inertie point par point :**

$$\text{Contr}_l(i) = \frac{p_i \; d_2(G, H_i^{(l)})^2}{\sum_i p_i \; d_2(G, H_i^{(l)})^2}$$

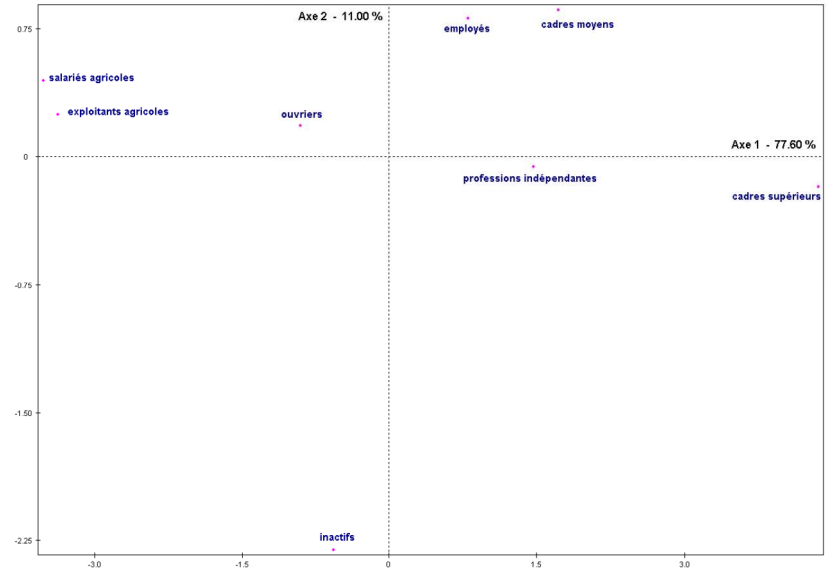
Remarques :

- le dénominateur $I_H(\mathcal{N}) = \sum_i p_i \; d_2(G, H_i^{(l)})^2 = \lambda_l$ (cf Page 25)
- $\sum_i \text{Contr}_l(i) = 1$
- **Quand $p_i = 1/n$ alors**
  $\text{Contr}_l(i) \propto$ (coordonnées)^2 du point sur l'axe
  $\propto \text{QLT}_l(i)$
  $\Longrightarrow$ **aucune information supplémentaire par rapport à l'information graphique**

- Si $p_i \neq 1/n$ alors
  1. Repérer les individus qui contribuent très fortement
  2. Si un seul individu admet une contribution très forte :
     - vérifier si pas d'erreur de mesure
     - il peut être utile de refaire l'analyse sans lui et de le rajouter comme **individu supplémentaire**

	PLAN	AXE1			AXE2		
	QLT(1,2)	COORD	QLT	CTR	COORD	QLT	CTR
Exploit. agri.	0,88	-3,37	0,88	22,9	0,25	0,00	0,9
Salar. agri.	0,91	-3,52	0,90	25,0	0,45	0,01	2,8
Prof. indép.	0,57	1,47	0,57	4,4	-0,06	0,00	0,0
Cadres sup.	0,94	4,36	0,94	38,3	-0,18	0,05	0,4
Cadres moy.	0,94	1,72	0,75	5,9	0,86	0,19	10,4
Employés	0,86	0,86	0,43	1,3	-0,81	0,43	9,3
Ouvriers	0,37	-0,90	0,36	1,6	0,18	0,01	0,5
Inactifs	0,99	-0,56	0,06	0,60	-2,31	0,93	75,6

TABLE 11: Nuage des individus projetés sur le premier plan factoriel



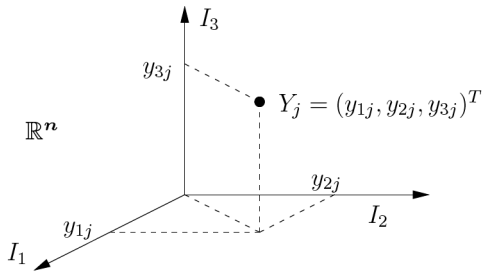


FIGURE 18: Nuage des variables centrées

Produit scalaire euclidien :

$$\begin{aligned}\langle Y_j, Y_k\rangle_{1/s^2} &= \sum_{i=1}^{n} \frac{1}{n} y_{ij}\, y_{ik} = \frac{1}{n} {Y_j}^\top Y_k = s_{Y_j,Y_k} = s_{X_j,X_k}\\ \|Y_j\|^2_{1/s^2} &= s^2_{Y_j} = s^2_{X_j}\\ d^2(Y_j, Y_k) &= \|Y_j - Y_k\|^2_{1/s^2} = \langle Y_j - Y_k, Y_j - Y_k\rangle_{1/s^2}\end{aligned}$$

Les variables sont centrées-**réduites** : $Z_j = Y_j/s_{Y_j}$

- **Chaque point-variable Z_j est de variance unité** :

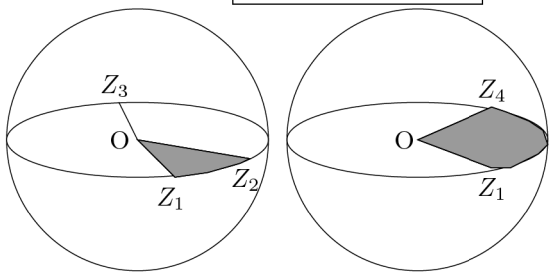
$$\|Z_j\|^2_{1/s^2} = 1$$

  i.e. se trouve sur la **sphère unité** (des corrélations)

- $\langle Z_j, Z_k\rangle_{1/s^2} = \dfrac{s_{Y_j,Y_k}}{s_{Y_j} s_{Y_k}} = \rho_{Y_j,Y_k} = \rho_{X_j,X_k}$
- $\cos(Z_j, Z_k) := \dfrac{\langle Z_j, Z_k\rangle_{1/s^2}}{\|Z_j\|_{1/s^2}\|Z_j\|_{1/s^2}} = \langle Z_j, Z_k\rangle_{1/s^2}$

$$\boxed{\cos(Z_j, Z_k) = \rho_{z_j,z_k}}$$

- **Lecture graphique de la liaison** (corrélation) linéaire entre deux point-variables : **les angles** : $\boxed{\cos(Z_j, Z_k) = \rho_{z_j\, z_k}}$



$\rho_{Z_1,Z_2} \approx 1$	$\rho_{Z_1,Z_3} \approx -1$	$\rho_{Z_1,Z_4} \approx 0$
$\cos(Z_1, Z_2) \approx 0$	$\cos(Z_1, Z_3) \approx \pi$	$\cos(Z_1, Z_4) \approx \pi/2$
$\widehat{(Z_1, Z_2)} \approx 0$	$\widehat{(Z_1, Z_3)} \approx \pi$	$\widehat{(Z_1, Z_4)} \approx \pi/2$

- Les variables Z_1, Z_2 et Z_1, Z_3 sont fortement corrélées
- Les variables Z_1 et Z_4 sont non-corrélées
  ↪ tout repère orthonormé est un repère à base de « variables » non-corrélées deux à deux.

Remarques :

1. Les variables sont centrées mais la nouvelle origine du nuage n'est pas le centre de gravité du nuage original des points variables
   Le centrage des variables

$$\Longrightarrow \forall j \qquad 0 = \frac{1}{n}\sum_{i=1}^{n} y_{ij} = \langle Y_j, \mathbf{1}_n\rangle_{1/s^2} \quad (Y_j \perp \text{au vecteur } \mathbf{1}_n)$$

   Tous les Y_j (Z_j) sont ainsi dans l'hyperplan orthogonal à la première bissectrice dans le $\mathcal{N}(J)$ original
2. Les variables sont de plus réduites
   $\Longrightarrow$ chaque point est à distance 1 de l'origine
   $\Longrightarrow$ distance à l'origine n'a aucun intérêt

- **Projection orthogonale** du nuage des variables sur un sous-espace H
  Le **Nuage projeté devra restituer le plus fidèlement les liaisons (corrélations) entre les points-variables** du nuage d'origine

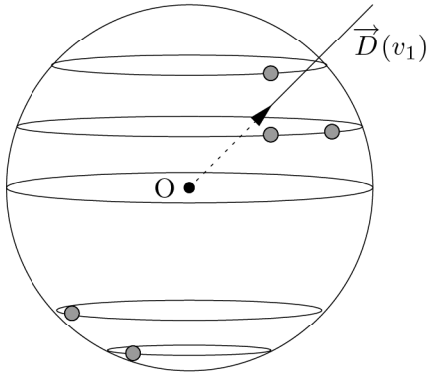
Cas où H est une droite

- **Rechercher une variable** v_1
  - de variance unité (norme 1)
  - qui maximise la somme des carrés des corrélations entre les variables originales et la nouvelle variable $\sum_{j=1}^{p} \rho_{Z_j,v_1}{}^2$

$$\sum_{j=1}^{p} \rho_{z_j,v} = \sum_{j=1}^{p} \cos(Z_j, v_1)^2 = \sum_{j=1}^{p} \langle Z_j, v_1\rangle^2_{1/s^2} = \sum_{j=1}^{p} d_{1/s^2}(O, H_j)^2$$

où H_j est la projection orthogonale de Z_j pour $\langle\cdot,\cdot\rangle_{1/s^2}$ sur la droite engendrée par v_1

Interprétation



v_1 direction de plus forte concentration de points

- **Rechercher un vecteur unitaire v_1 de la droite H (pour $\langle\cdot,\cdot\rangle_{1/s^2}$) qui réalise**

$$\max\left(\sum_{j=1}^{p} d_{1/s^2}(O, H_j)^2\right)$$

On vérifie que

$$\sum_{j=1}^{p} d_{1/s^2}(O, H_j)^2 = v_1^\top \left[\frac{1}{n} ZZ^\top\right] v_1$$

- Déterminer le vecteur $v_1 \in \mathbb{R}^n$ tel que

$$\langle v_1, v_1\rangle_{1/s^2} = 1 \quad \text{et} \quad v_1 := \arg\max v^\top \left[\frac{1}{n} ZZ^\top\right] v$$

La matrice $ZZ^\top$ est symétrique et semi-définie positive pour $\langle\cdot,\cdot\rangle_{1/s^2}$, alors

Théorème 6

1. $\mathbb{R}^n$ *admet une base orthonormée pour* $\langle\cdot,\cdot\rangle_{1/s^2}$ *de vecteurs propres de* $ZZ^\top$
2. *les valeurs propres de* $ZZ^\top$ *sont positives.*

Théorème 7

Le sous-espace $H^{(q)}$ *de dimension* q *de* $\mathbb{R}^n$ *portant « l'inertie » maximale se calcule par*

$$H^{(q-1)} \oplus H_q$$

où H_q *est la droite orthogonale, relativement à* $\langle\cdot,\cdot\rangle_{1/s^2}$, *à* $H^{(q-1)}$ *portant « l'inertie » maximale.*

Théorème 8

le sous-espace $H^{(q)}$ ajustant au mieux le nuage est engendré par les q premiers vecteurs propres (variance unité) $\{v_1, \ldots, v_q\}$ correspondants aux q plus grandes valeurs propres de $ZZ^\top/n$

$$\mu_1 \geq \mu_2 \geq \cdots \geq \mu_q \geq 0$$

Proposition 2

La matrice $\frac{1}{n}ZZ^\top$ et la matrice des corrélations $R = \frac{1}{n}Z^\top Z$ ont les même valeurs propres non-nulles (de même multiplicité)

➥ au plus p ($n >> p$) valeurs propres.

Remarque : Comme $n > p$, le calcul des éléments spectraux des deux matrices de la proposition sera réalisé sur la « petite » matrice des corrélations R (voir Prop 3) ou de covariance si les données ne sont pas réduites.

Axes factoriels

- Soit C_l une composante principale i.e.

$$C_l = Zu_l \qquad \left[\begin{array}{l} u_l \text{ axe factoriel de rang } l \text{ de } I(\mathcal{I}(I)) \\ Ru_l = \lambda_l u_l \text{ et } {u_l}^\top u_l = 1 \end{array} \right.$$

alors

$$\left\{ \begin{array}{rcl} \frac{1}{n}[ZZ^\top]\ C_l & = & \frac{1}{n}ZZ^\top\ Zu_l = Z\left[\frac{1}{n}Z^\top Z\right]\ u_l = ZRu_l = \lambda_l\ Zu_l \\ & = & \lambda_l\ C_l \\ \text{et } s^2_{c_l} & = & \lambda_l \end{array} \right.$$

➥ vecteur unitaire v_l pour $\langle \cdot, \cdot \rangle_{1/s^2}$: $\boxed{v_l = \frac{1}{\sqrt{\lambda_l}} C_l}$

Définition 20 (variable principale)

*On appelle **axe factoriel (ou variable principale) de rang** l du nuage des variables, les vecteurs*

$$\boxed{v_l = \frac{1}{\sqrt{\lambda_l}} C_l} \quad l = 1, \ldots, p$$

où C_l est la composante principale de rang l

Remarques :

- $(v_l)_l$ famille orthonormée pour $\langle \cdot, \cdot \rangle_{1/s^2}$

  orthogonalité $\longleftrightarrow$ non-corrélation
  normé $\longleftrightarrow$ variance unité

- $v_l = \frac{1}{\sqrt{\lambda_l}} \sum_{j=1}^{p} u_l(j)\ Z_j$

Les p points-variables du nuage $\mathcal{N}(J)$ sont projetés sur l'axe factoriel v_l :

$$F_l := \left(\langle Z_j, v_l \rangle_{1/s^2} \right)_{j=1}^{p} = \left(\rho_{Z_j, v_l} \right)_{j=1}^{p} \in \mathbb{R}^p$$

- $|F_l(j)| \leq 1$ pour tout $j = 1, \ldots, p$
- Pour toute composante principale C_l :

$$\begin{array}{rcl} \rho_{Z_j, c_l} & = & \rho_{Z_j, v_l} = \rho_{X_j, v_l} = \rho_{X_j, C_l} \\ & = & F_l(j) \end{array}$$

$$\boxed{\rho_{X_j, C_l} = F_l(j)} \qquad \forall j = 1, \ldots, p$$

Abus de langage : on parle indifféremment de variables ou de composantes principales

Bilan graphique des liaisons : le cercle des corrélations

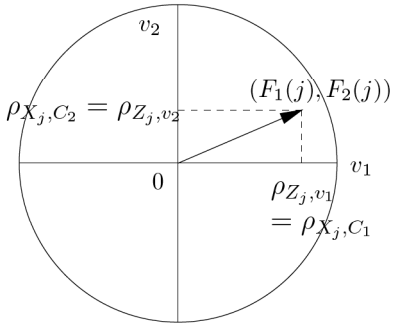


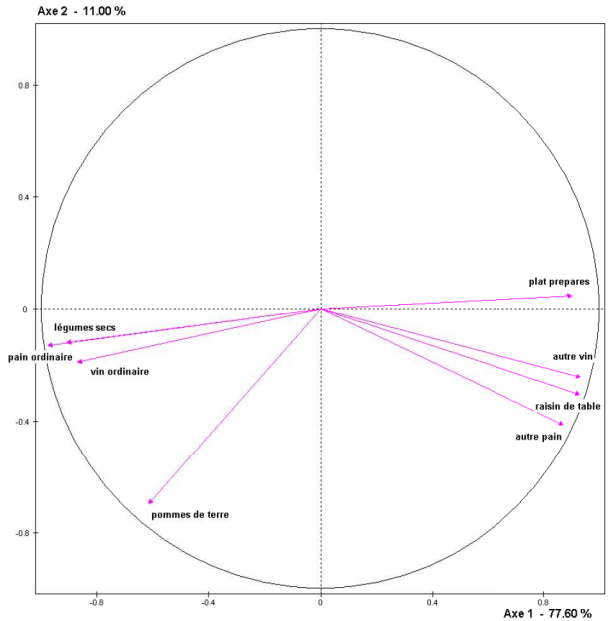
FIGURE 19: « Plan factoriel » 1-2

Liaison entre les variables originales et les variables/composantes principales

Bilan des liaisons linéaires

	Coordonnées sur les variables principales v_j $j = 1, \ldots, 5$				
Variables	v_1	v_2	v_3	v_4	v_5
PAO	-0,97	-0,13	0,10	0,07	0,12
PAA	0,87	-0,41	0,21	0,12	-0,11
VIO	-0,87	-0,19	0,44	-0,02	0,10
VIA	0,93	-0,24	0,05	-0,22	-0,14
POT	-0,61	-0,70	-0,36	-0,04	0,07
LEC	-0,91	-0,12	0,02	0,29	-0,27
RAI	0,93	-0,31	0,16	0,04	0,11
PLP	0,90	0,05	-0,10	0,39	0,14

TABLE 12: Projection du nuage des variables sur les cinq premières variables principales



- **QLT de représentation de Z_j par sa projection sur l'axe v_l**

$$\text{QLT}_l(j) = \cos(Z_j, v_l)^2 = {\rho_{Z_j,v_l}}^2 = {\rho_{Z_j,C_l}}^2$$

où C_l est la CP de rang l

➥ Plus la direction du vecteur Z_j est proche de celle de la nouvelle variable v_l (plus leur corrélation est grande), meilleure est la qualité

- **QLT de représentation de Z_j par ses projections sur le cercle (v_1, v_2)**

$$\begin{aligned}\text{QLT}_{v_1,v_2}(j) &= {\rho_{Z_j,v_1}}^2 + {\rho_{Z_j,v_2}}^2 \\ &= {\rho_{Z_j,c_1}}^2 + {\rho_{Z_j,C_2}}^2\end{aligned}$$

➥ Graphiquement, plus l'extrémité du vecteur représentant la variable Z_j est proche du cercle, meilleure est la qualité

- **QLT de représentation du nuage par les q premières variables principales**

$$\mathrm{QLT}_{v_1,\ldots,v_q}(\mathcal{N}(J)) = \frac{\sum_{l=1}^{q}\left[\sum_{j=1}^{p}\rho_{Z_j,v_l}{}^2\right]}{\sum_{l=1}^{p}\sum_{j=1}^{p}\rho_{Z_j,v_l}{}^2} = \frac{\sum_{l=1}^{q}\lambda_l}{\sum_{l=1}^{p}\lambda_l} = \frac{\sum_{l=1}^{q}\lambda_l}{p} \text{(voir (26))}$$

Remarque :

$$\mathrm{QLT}_{v_1,\ldots,v_q}(\mathcal{N}(J)) = \mathrm{QLT}_{u_1,\ldots,u_q}(\mathcal{N}(I))$$

On vérifie que

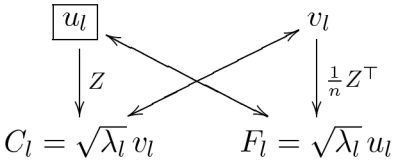
$$\boxed{F_l = \sqrt{\lambda_l}\, u_l}$$

où u_l est l'axe factoriel du nuage des individus de rang l

- Comme $\|u_l\|_2 = 1$ et d'après la relation ci-dessus, v_l est tel que

$$\sum_{j=1}^{p}\rho_{Z_j,v_l}{}^2 = \lambda_l \quad (26)$$

Proposition 3 (Relations de transition)



Résultats d'une ACP normée

- liste de valeurs propres de la matrice des corrélations
- liste d'axes factoriels
  ➥ liste de graphiques ≡ plans factoriels
  bilan des ressemblances entre individus
- liste de variables principales
  ➥ liste de cercles des corrélations
  bilan des liaisons entre les variables
- Indicateurs de qualité de la représentation
  (de contribution des éléments à la construction des axes factoriels)

Deux phases dans l'interprétation d'une ACP

Phase 1 (partie objective)
Bilan des inerties associées aux différents axes factoriels et des corrélations associées aux différentes composantes principales
➥ fondé seulement sur des indices numériques

Phase 2 (partie subjective)
Interprétation proprement dite des axes et composantes principales
➥ fondée en grande partie sur les **connaissances** du problème étudié c.a.d **extérieures** au tableau de données

Éléments pour l'interprétation

Proposition 4

$1 \leq \lambda_1 \leq p =$ *nombre de variables*

- $\boldsymbol{\lambda_1 = 1}$ lorsque toutes les variables initiales sont non-corrélées deux à deux

  $\Longrightarrow$ ACP aucun intérêt

  Si $\boldsymbol{\lambda_1 \approx 1}$, l'ACP du tableau ne présente pas d'intérêt
- $\boldsymbol{\lambda_1 = p}$ lorsqu'il existe une liaison linéaire parfaite entre toutes les variables

$$j = 1, \ldots, p \qquad \rho_{x_j, v_1} = \pm 1$$

➥ Plus λ_1 est grande plus v_1 résume les variables et plus v_1 risque d'être intéressante

- Plus d'encadrement pour les valeurs propres suivantes mais la valeur 1 reste un point de repère :

  ⇨ une valeur propre ≤ 1 indique que la variable principale associée synthétise moins d'information qu'une variable isolée

$$\left(\sum_{j=1}^{p} {\rho_{X_j, v_l}}^2 = \lambda_l \text{ et } {\rho_{X_l, X_l}}^2 = 1\right)$$

$\Longrightarrow$ Prudence pour interpréter v_l telle que $\lambda_l \approx 1$

Choix de la dimension de l'espace de visualisation des nuages i.e. nombre de valeurs propres à retenir

- $\text{QLT}_{u_1,\ldots,u_q}(\mathcal{N}(I)) = \text{QLT}_{v_1,\ldots,v_q}(\mathcal{N}(J)) = \dfrac{\sum_1^q \lambda_l}{\sum_1^p \lambda_l}$

Choisir q telle que $\text{QLT}_{u_1,\ldots,u_q} \geq$ seuil fixé

Règle intéressante pour une utilisation de simple réduction de dimension

- **Règle de Kaiser**.
  Ne retenir que les composantes associées à des valeurs propres supérieures à 1
  ➥ tendance à surestimer le nombre de CP pertinentes

- **Diagramme des valeurs propres** : Rechercher la présence d'un « coude »

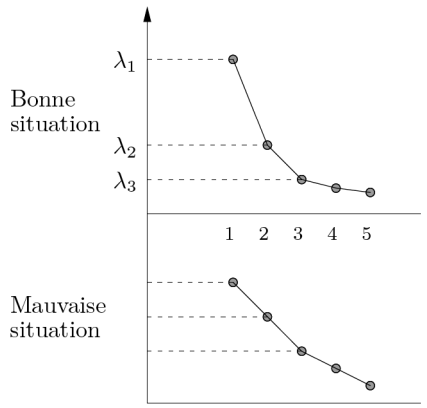


FIGURE 20: graphique représentant la décroissance des valeurs propres

- Si l'ACP est utilisée pour extraire une information la plus riche possible

  ➥ Ne retenir que des variables ou axes interprétables

- Les axes et CP sont considérés dans l'ordre décroissant des valeurs propres.
- Garder à l'esprit que l'axe au CP d'ordre $l > 1$ traduit les tendances résiduelles non prise en compte par les précédents
- L'ordre proposé $\longrightarrow$ phase de découverte

Approfondissement trop lié aux données

Contribution des individus. Coordonnées des individus actifs

Intérêt d'un axe $\equiv$ nombre d'individus concernés

Liste des contributions

Repérer si un ou un très petit nombre d'individus ont une contribution très supérieure à la moyenne.
Si $p_i = 1/n$ alors points loin du centre de gravité et proche de l'axe

- éventuellement remise en cause du champ de l'étude (si l'axe est dans les premiers)
- si après quelques axes alors phénomènes ponctuels / tendances générales

Etude des plans factoriels

- Allure générale de la répartition des individus
  Toute plage de faible densité ou forte densité doit être décelée
- Aider le choix d'individu-type
  ➜ très bonne qualité de représentation par les axes concernés et moyenne pour les autres

Coordonnées des variables actives

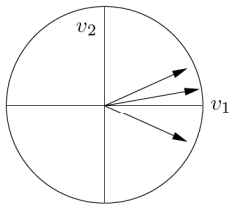
Rappel en ACP normée : projection d'une variable sur un axe $\equiv$ QLT de représentation et $\propto$ à sa contribution

➥ Étude graphique des cercles

Interprétation par composante

- Recenser les variables actives les plus corrélées à chaque composante
  Les var. corrélées > 0 ont une coord. > 0 et celles corrélées < 0 une coord. < 0
- Rechercher un dénominateur commun qui relie les variables situées du même côté et oppose les variables situées de part et d'autre de l'origine

Remarque : facteur « taille » (v_1 en général)



toutes les variables sont corrélées positivement entre elles

- Si pour un individu, une variable prend une valeur forte, toutes les autres prennent également des valeurs fortes
- Si de plus les corrélations sont toutes du même ordre alors

$$v_1 \propto \frac{1}{p}\sum_{j=1}^{p} Z_j$$

et v_2 différencie alors les individus de « taille » semblable : **« facteur de forme »**

Interprétation par cercle

- Rappel :
  proximité de l'extrémité d'un vecteur du cercle $\equiv$ qualité de la représentation
- Les vecteurs joignant l'origine aux points variables permettent de visualiser les **angles** mesurant la liaison entre variables.

Interprétation « conjointe »

La coordonnée de l'individu N° i sur l'axe u_l

$$C_l(i) = \langle Z_{i\,.}, u_l \rangle_2 = \sum_{j=1}^{p} \left[\frac{x_{ij} - \overline{X}_j}{s_{X_j}}\right] u_l(j)$$

et $u_l = \frac{1}{\sqrt{\lambda_l}} F_l$ (cf Prop 3) avec

$F_l(j)$ = coord. de X_j sur la var. principale v_l $= \rho_{X_j,v_l}$

$$\Longrightarrow \boxed{C_l(i) = \frac{1}{\sqrt{\lambda_l}} \sum_{j-1}^{p} \left[\frac{x_{ij} - \overline{X}_j}{s_{X_i}}\right] \rho_{X_j,v_l}}$$

$C_l(i) >> 0$: pour l'individu i on a, pour certaines variables X_j :

$$\left[\frac{x_{ij} - \overline{X}_j}{s_{X_j}}\right] \rho_{X_j,v_l} >> 0$$

et pour ces variables X_j on a

$$\begin{cases} \text{soit} & x_{ij} - \overline{X}_j >> 0 \ \underline{\text{et}} \ \rho_{X_j,v_l} \approx 1 \\ \text{soit} & x_{ij} - \overline{X}_j << 0 \ \underline{\text{et}} \ \rho_{X_j,v_l} \approx -1 \end{cases}$$

$$\Longleftrightarrow \begin{cases} \text{soit} & x_{ij} >> \overline{X}_j \ \underline{\text{et}} \ \rho_{X_j,v_l} \approx 1 \\ \text{soit} & x_{ij} << \overline{X}_j \ \underline{\text{et}} \ \rho_{X_j,v_l} \approx -1 \end{cases}$$

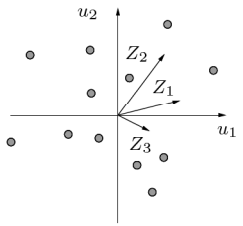
Ainsi un individu très éloigné du pseudo-individu G selon u_l, l'est, en général, par des scores « extrêmes » pour les variables dont les directions dans le cercle des corrélations, sont très proches de la direction de la variable principale v_l.
De plus, un score positif signifie que les variables concernées « sont du même coté » que l'individu sur l'axe u_l

Un représentation simultanée

Après analyse du nuage des individus $\mathcal{N}(I)$

- anciens axes unitaires $\{e_j\}_{j=1}^p$ (directions associées aux variables originales)
$$e_j = (0, \ldots, 0, 1, 0, \ldots, 0)^\top.$$
- nouveaux axes unitaires : $\{u_l,\ l = 1, \ldots, p\}$ axes factoriels

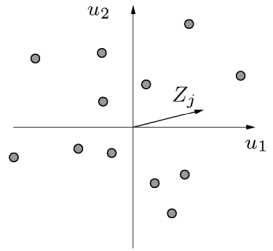
Représentation simultanée ≡ représentation des directions associées aux variables originales sur les plans factoriels de $\mathcal{N}(I)$

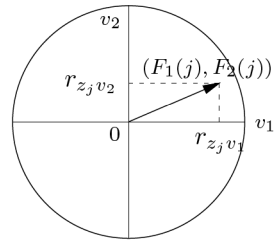


↪ « écrasement » du repère orthonormé initial sur les plans factoriels

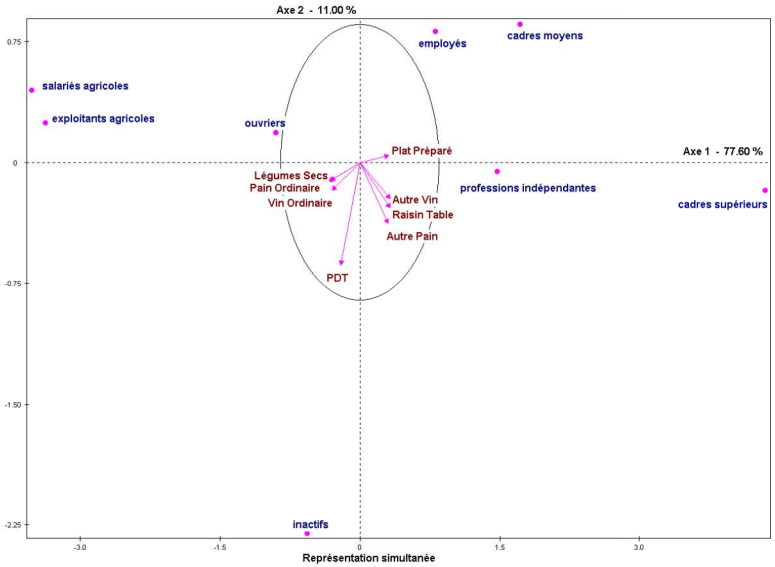
- **Projection de e_j sur u_l :** $e_j{}^\top u_l = u_l(j)$
- **Projection du point variable Z_j sur la variable principale v_l**

avec Prop 3 $F_l(j) = \sqrt{\lambda_l}\, u_l(j) \Longrightarrow (F_l(j))_l = \left(\sqrt{\lambda_l}\, e_j{}^\top\, u_l\right)_l$





- Le nuage des extrémités (des projections) des e_j ds $\mathbb{R}^p$ et le nuage des extrémités (des projections) des Z_j dans $\mathbb{R}^n$ ne diffèrent l'un de l'autre que par une dilatation sur chaque axe de rapport $\sqrt{\lambda_l}$.
- Si les valeurs propres sont presque égales alors la déformation est faible.



Interprétation à partir de la représentation simultanée

- Étudier les positions respectives de deux individus / à l'**ensemble** de toutes les variables
- Étudier les positions respectives de deux variables / à l'**ensemble** de tous les individus
  Enfin, la distance entre deux vecteurs variables ne peut pas être interprétée en termes de corrélation

Individus supplémentaires

- But :
  Vérifier sur ces individus des hypothèses formulées après une ACP sur les individus actifs.
- Mise en oeuvre :
  Positionnement / au centre de gravité du nuage initial en appliquant la même transformation de centrage réduction que pour les individus actifs
  Puis projection sur les axes factoriels

Variables supplémentaires

- Expliquer certaines CP ou affiner les interprétations
- Valider a posteriori
- Et/ou suggérer le réexamen d'une CP délaissée sur la seule vue des variables actives

Variables continues supplémentaires

Dans le nuage de variables :

distance $\equiv$ corrélation

- On calcule la moyenne et l'écart-type des nouvelles variables
  ➥ position dans la sphère unité des variables
- Coordonnées $\equiv$ coefficients de corrélations entre la variable et les CP

Variables nominales supplémentaires

- Variables à m modalités
  ➥ m groupes d'individus
- Ces m groupes sont traités comme des individus supplémentaires, chaque groupe étant représenté en général par son centre de gravité

ACP et l'approximation du tableau de données

- Une décomposition en valeurs singulières de Z

$$\left.\begin{array}{l} u_l \text{ axe factoriel de } \mathcal{N}(I) \\ v_l \text{ axe factoriel de } \mathcal{N}(J) \end{array}\right] \rightarrow \lambda_l$$

La composante principale $C_l = Zu_l = \sqrt{\lambda_l} v_l$

$$\Longrightarrow Zu_l{u_l}^\top = \sqrt{\lambda_l}\ v_l{u_l}^\top$$

$$\Longrightarrow Z\,[\sum_{l=1}^{p} u_l{u_l}^\top] = \sum_{l=1}^{p} \sqrt{\lambda_l}\ v_l{u_l}^\top$$

Si $U = [u_1 \cdots u_l \cdots u_p]$, alors $\sum_{l=1}^{p} u_l{u_l}^\top = UU^\top = I$

$$\boxed{Z = \sum_{l=1}^{p} \sqrt{\lambda_l}\ v_l{u_l}^\top}$$

- **Conservation des $q < p$ premières valeurs propres :**

$$\boxed{Z \approx Z(q) := \sum_{l=1}^{q} \sqrt{\lambda_l}\ v_l{u_l}^\top} \quad \text{en } O(qp^2)$$

Théorème 9 (Un théorème d'Eckard-Young)

Les q premiers termes de la décomposition définit une matrice $Z(q)$ qui est la meilleure matrice $n \times p$ (au sens de l'inertie ou encore des moindres carrés) de rang q approchant Z : pour toute matrice A de rang q

$$\|Z - Z(q)\|_{\text{In}} \leq \|Z - A\|_{\text{In}} \tag{27}$$

avec $\|M\|^2_{\text{In}} := \text{trace}(M^\top M)/n$. Enfin l'erreur commise vaut (l'inertie résiduelle)

$$\|Z - Z(q)\|^2_{\text{In}} = \sum_{l=q+1}^{p} \lambda_l$$

Golub, G. H. et Van Loan, C. F. (1996).
Matrix Computations. Ed. 3.
Johns Hopkins University Press, Baltimore.

ACP ou transformation de Karhunen-Loeve

Exemple 11 (Réduction de canaux : image du satellite SPOT)

- Pour une région donnée, le satellite SPOT fournit 12 images (NB) issues de 12 canaux correspondant à des longueurs d'ondes différentes. Une image est constituée de $293 \times 282 = 82626$ pixels. Pour chaque pixel et chaque longueur d'onde, l'amplitude est codée de 0 à 255.
- Un pixel correspond à un individu et la matrice X des données admet 12 colonnes correspondant aux 12 canaux. On applique une ACP normée sur cette matrice et les valeurs des composantes principales sont recodées de 0 à 255 et on retransforme chaque composante principale en une matrice « image ».
- La première composante principale donne une image NB qui maximise « le contraste ».

Remarque : idée similaire à partir d'une décomposition (RedGreenBlue) d'une image

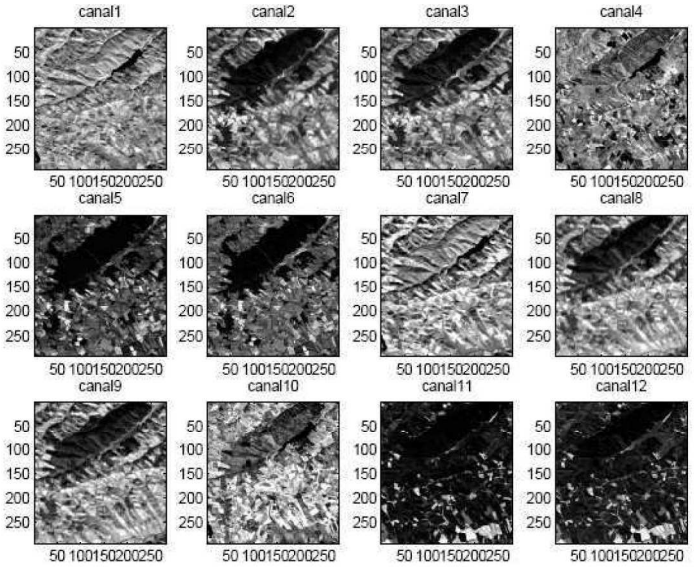


FIGURE 21: Une image via les 12 canaux

Résultats

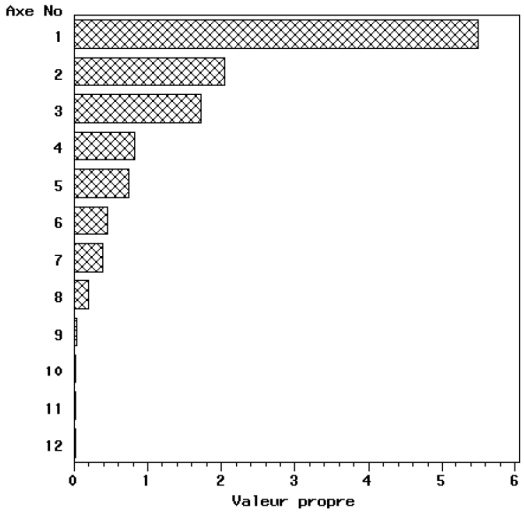


FIGURE 22: Diagramme des valeurs propres

Numéro	Valeur propre	Pourcentage d'inertie	Pourcentage cumulé
1	5.5027	45.86	45.86
2	2.0452	17.05	62.91
3	1.7215	14.35	77.26
4	0.8298	6.92	84.21
5	0.7369	6.14	90.35
6	0.4618	3.85	94.18
7	0.3965	3.31	97.39
8	0.2031	1.70	99.09
9	0.0404	0.34	99.43
10	0.0245	0.21	99.64
11	0.0233	0,20	99.84
12	0.0145	0.12	99.96
	≈ 12		

TABLE 13: Pourcentage d'inertie expliquée par chacun des axes factoriels

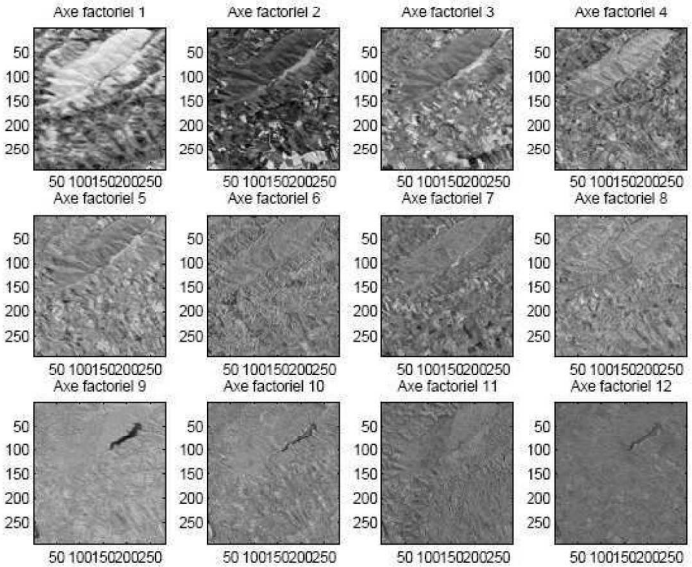


FIGURE 23: les 12 canaux ACP

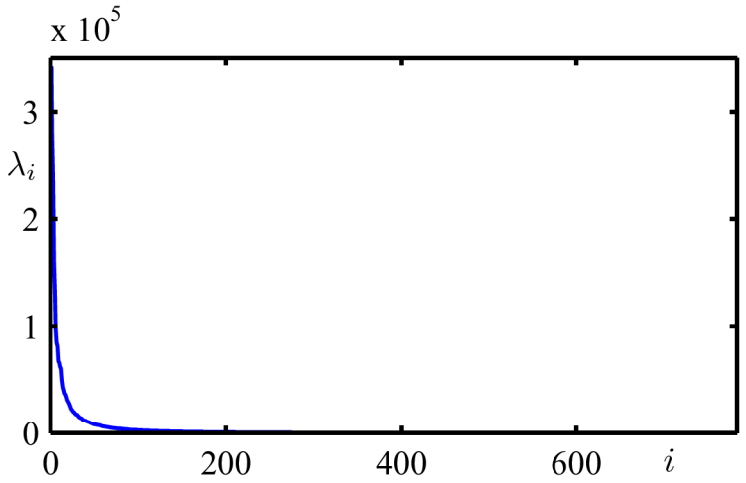


FIGURE 24: « Courbe » donnant les valeurs des valeurs propres

Exemple 12 (Chiffres manuscrits)

- Un jeu de données utilisés dans [LBBH98] est constitué de 70000 images de chiffres manuscrits. Il est lui même extrait d'une base du National Institute of Standard and Technology.
- On se restreint ici aux images du chiffre 3 de 28×28 pixels avec un codage en niveau de gris.
- Chaque individu est donc une image ou un point de $\mathbb{R}^{784}$.

LECUN Y. ET AL. (1998).
Gradient-Based Learning Applied to Document Recognition
Proceedings of the IEEE. Vol. 86, 2278–2324.

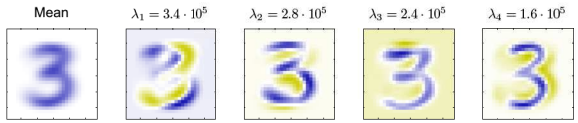


FIGURE 25: Le centre de gravité et les quatre premiers vecteurs et valeurs propres. Les valeurs bleues correspondent à des valeurs positives, les jaunes à des valeurs négatives et 0 à blanc.

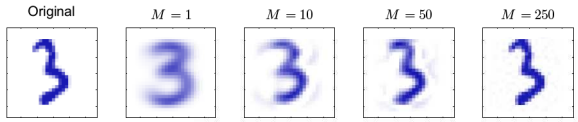


FIGURE 26: Exemple de reconstruction d'un chiffre 3 suivant les valeurs de p