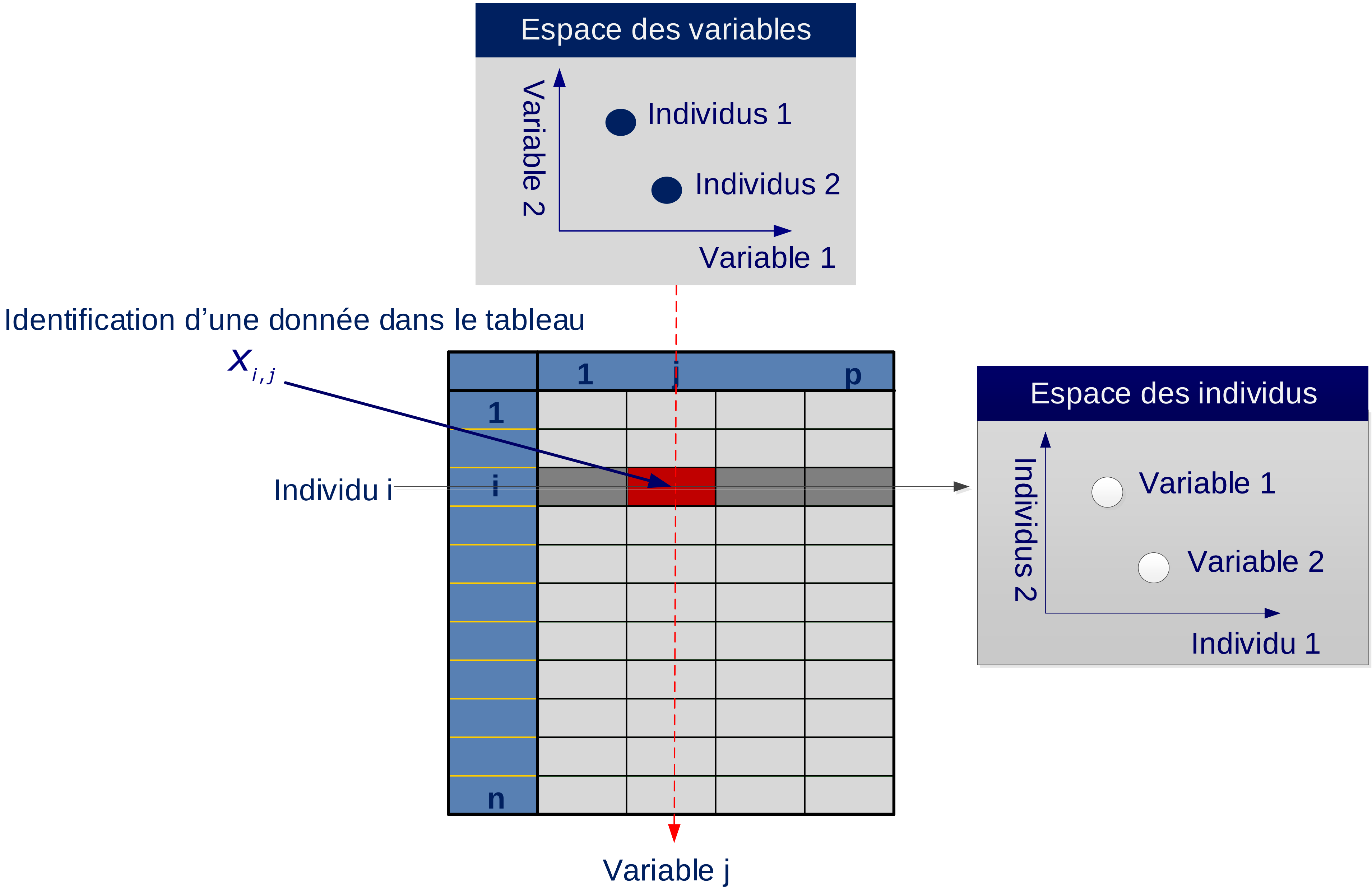
INTRODUCTION A L'ANALYSE FACTORIELLE

PARTIE I

ANALYSE EN COMPOSANTE PRINCIPALE (ACP)


Espace des variables
Variable 2
Individus 1
Individus 2
Variable 1
Identification d'une donnée dans le tableau
$x_{i,j}$
1
j
p
1
i
n
Individu i
Variable j
Espace des individus
Individus 2
Variable 1
Variable 2
Individu 1

- On sait représenter des variables ou des individus dans un espace à deux ou trois dimensions. Au-delà, aucune représentation graphique n'est possible

> Soit un tableau à p variables et à n individus (impossible à représenter graphiquement). L'objectif des analyses factorielles est de trouver des « **espaces de dimensions plus petites** » dans lesquels il est possible d'observer au **mieux les variables et les individus**.

- Les principales techniques

  Analyse non supervisée

  ACP : Analyse en composante principale : variables quantitatives

  AFC : Analyse factorielle des correspondances : variables qualitatives (tableaux de contingence)

  AFCM : Analyse factorielle des correspondances en composante principale : variables qualitatives (tableaux disjonctifs)

  Analyse supervisée

  AFD: Analyse factorielle discriminante

- Historique

  Technique ancienne : Pearson 1900 → Les fondamentaux de l'AD

  Hotteling : 1933 → Présentation de l'ACP

  1940 → Présentation de l'AFC

  Nécessite des calculs intensifs → utilisation de calculateurs

Remarques

➡ Statistique inférentielle

Distribution de probabilité → « fonctions mathématiques » qui possèdent certaines propriétés

Individus → utilisés uniquement pour étudier la distribution de la VA

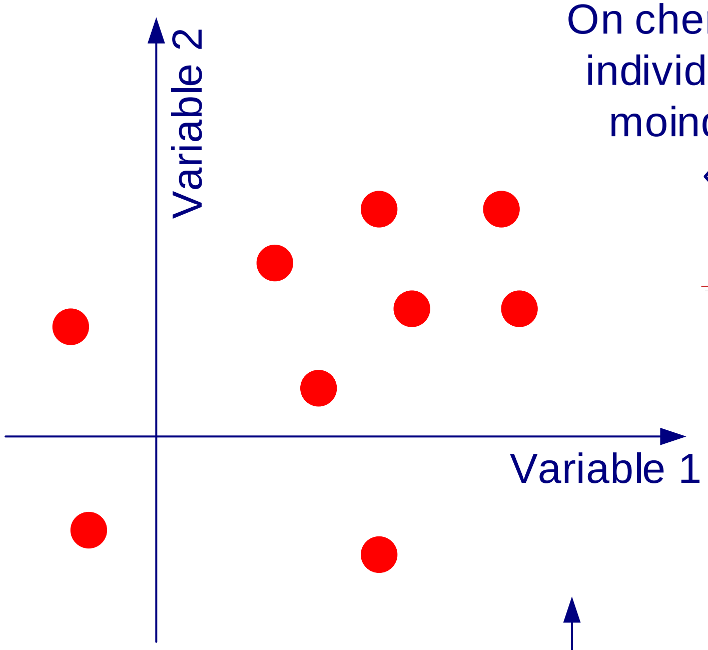
Il n'est pas toujours prouvé que l'on puisse connaître et étudier la distribution de probabilité

➡ Analyse factorielle

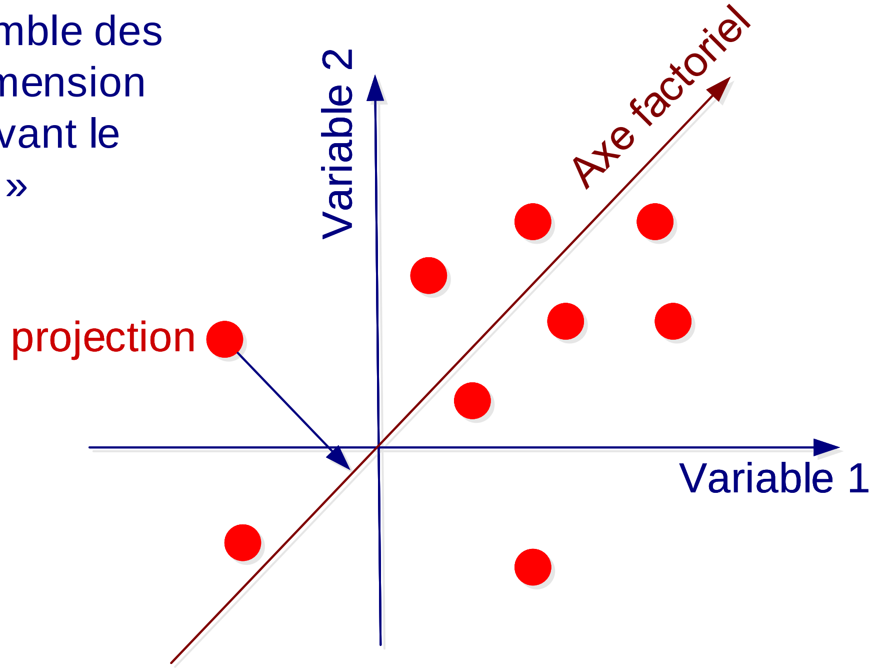
On s'intéresse uniquement aux données c.a.d « au tableau que l'on a sous les yeux »

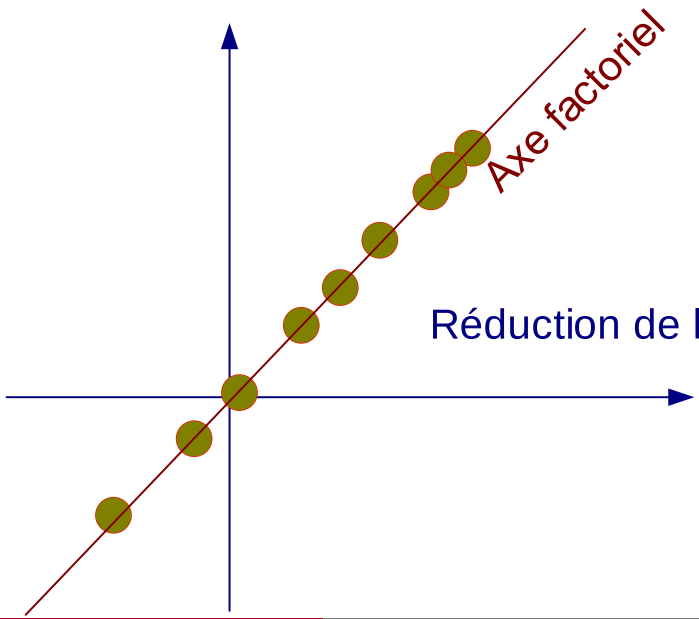
Analyse descriptive

Représentation dans un plan (p = 2)

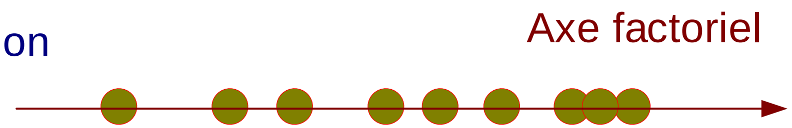


On cherche à représenter l'ensemble des individus dans un espace de dimension moindre (p =1) tout en conservant le « maximum d'information »



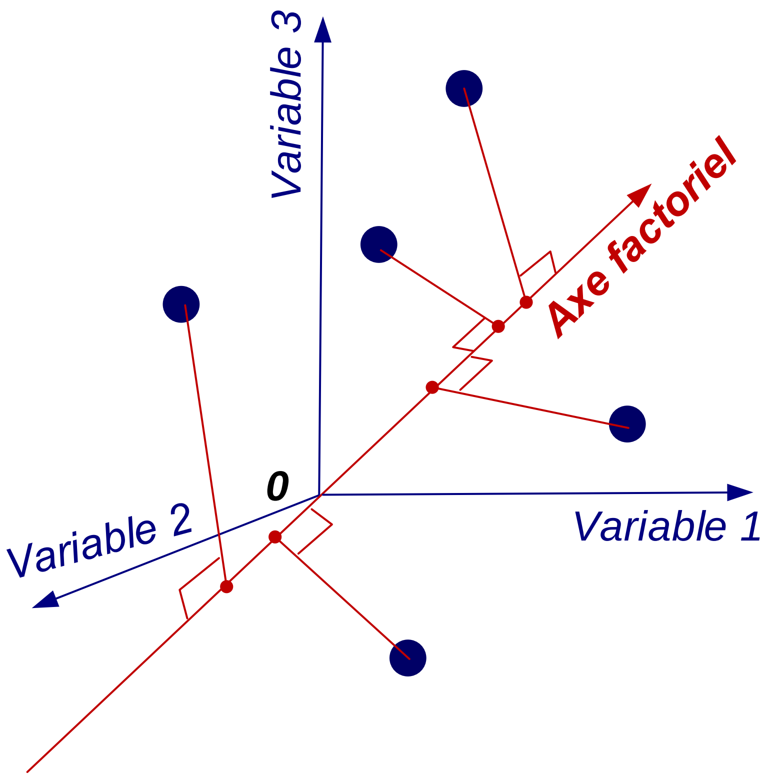


Réduction de la dimension



Extension à p dimensions : nuage de points des individus dans l'espace des variables

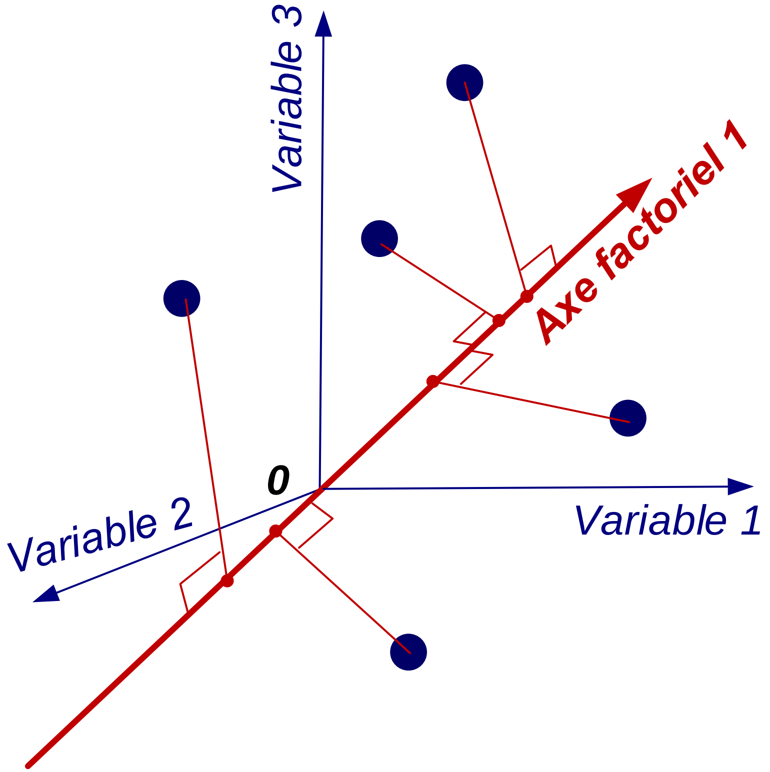
- Si p > 3 pas de représentations possibles
- On va chercher des axes (appelés axes principaux ou axes factoriels) qui sont des « **combinaisons linéaires** » des variables initiales permettant ainsi de décrire l'ensemble des individus en prenant en compte l'ensemble des variables (ou réciproquement). Ces axes factoriels sont des projections



- On va chercher des axes de projection des points (axes qui seront étudiés deux à deux) qui permettront la meilleure « VISUALISATION » <u>du nuage dans des espaces de plus faibles dimensions</u>
- L'AD se caractérise par la présentation de résultats sous forme de graphiques qui vont contenir le maximum d'information du tableau de données initial

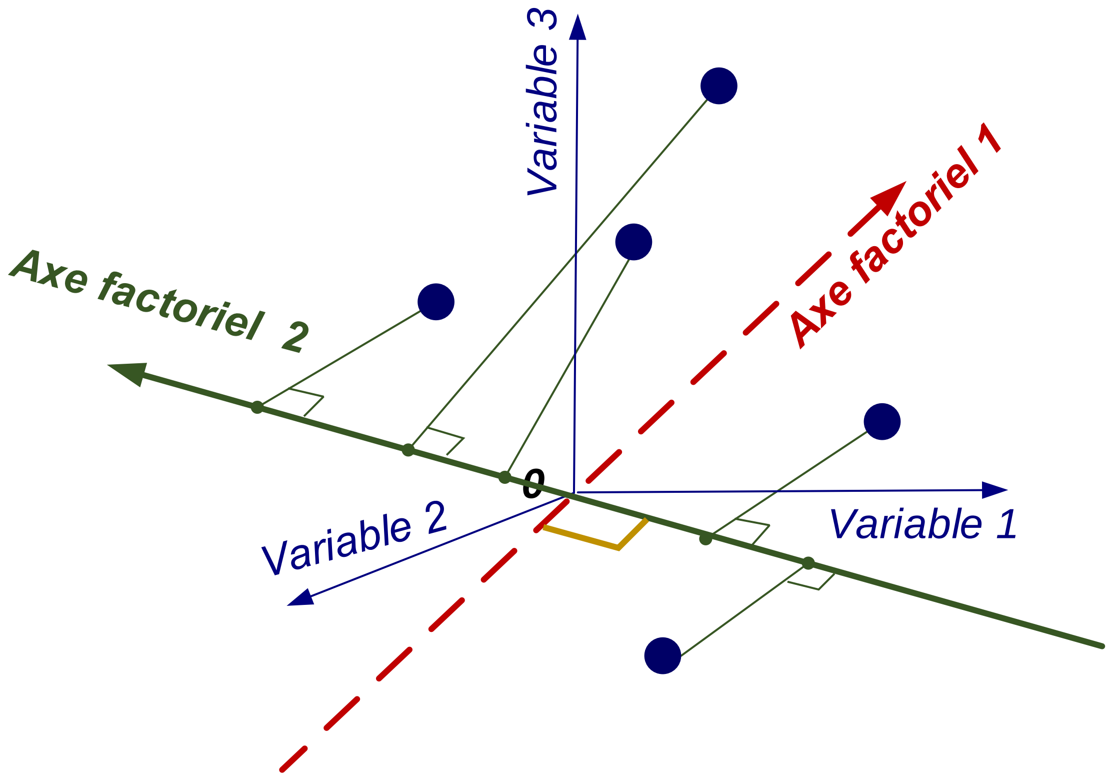
Etape 1

Recherche du premier axe factoriel qui explique au mieux les données initiales



Etape 2… n

Recherche du second axe factoriel PERPENDICULAIRE au premier qui explique au mieux les données initiales...la qualité de la projection sera bien évidemment moindre



Utilisation en analyse de données

Non supervisée : Analyse descriptive

→ Comment se structurent les variables entre elles (corrélation entre les variables)

- Liaisons entre les variables
- Quelles sont celles qui sont associées, celles qui ne le sont pas, quelles sont celles qui vont dans le même sens, quelles sont celles qui s'opposent
- Recherche de « familles de variables » puis sélection au sein de chaque famille

→ Ressemblance entre les individus (distances)

- D'un point de vue calculatoire, l'AF consiste à transformer les p variables initiales toutes plus ou moins corrélées entre elles en p nouvelles **variables non corrélées**

→ Ou et comment se répartissent les individus, quels sont ceux qui ce ressemblent, quels sont ceux qui sont dissemblables

Supervisée : Classification

→ Comment se structurent les variables entre elles

→ Ressemblance entre les individus

→ Classification

- **Estimation de données manquantes (Non linear Iterative PArtial Least Square - NIPALS)**

- **Utilisation en traitement d'image**

  Non supervisée : compression d'images

  Supervisée : reconnaissance de formes (faciale, radio, …)

- ***Enquête sur la consommation de produits alimentaires en fonction de catégories socio-professionnelles***

Les individus : des catégories socio-professionnelles (MA : travailleurs manuels, EM : employés, CA : cadres)- croisées avec leur nombre d'enfants (de 2 à 5)

Les variables : Indices de dépenses annuelles de différents type d'aliments

	pain	fruit	viande	volaille	lait	légume	alcool
MA2	332	354	1437	526	247	428	427
EM2	293	388	1527	567	239	559	258
CA2	372	562	1948	927	235	767	433
MA3	406	341	1507	544	324	563	407
EM3	386	396	1501	558	319	608	363
CA3	438	689	2345	1148	243	843	341
MA4	534	367	1620	638	414	660	407
EM4	460	484	1856	762	400	699	416
CA4	385	621	2366	1149	304	789	282
MA5	655	423	1848	759	495	776	486
EM5	584	548	2056	893	518	995	319
CA5	515	887	2630	1167	561	1097	284

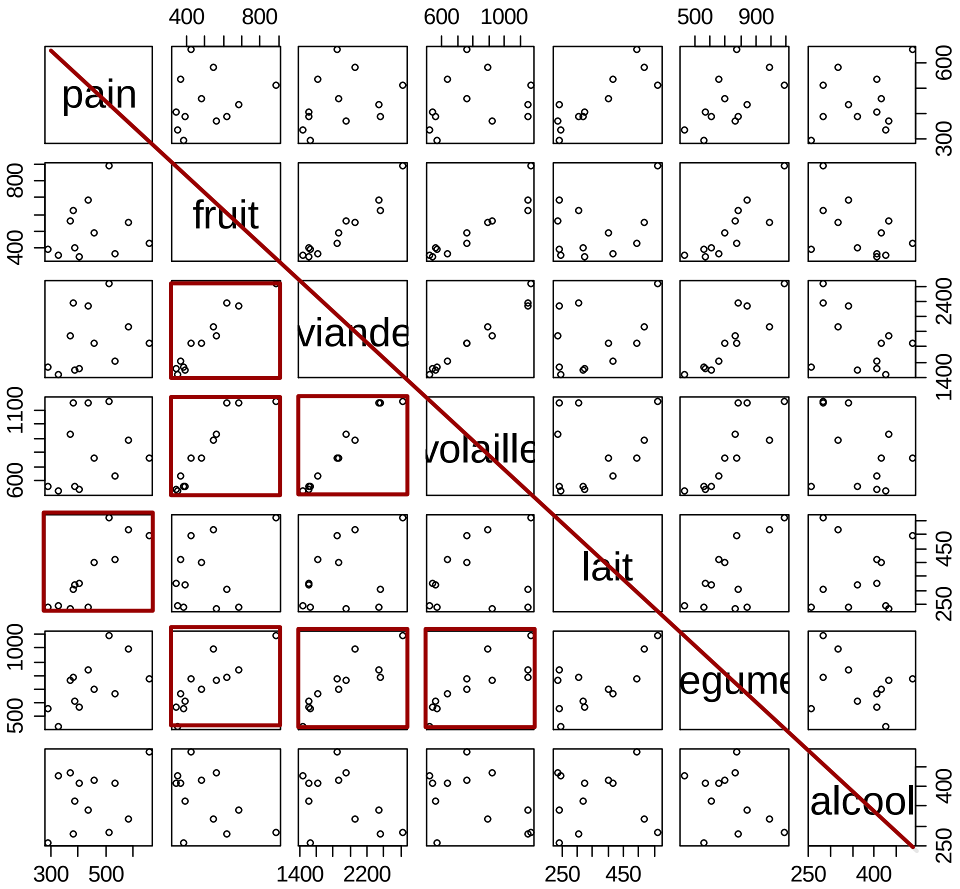
Cette exemple est extrait d'une enquête ancienne (1960) sur la consommation des ménages. Bien qu'elle soit obsolète, il s'agit d'une bon exemple introductif à l'interprétation de l'ACP

Statistiques univariées

	pain	fruit	viande	volaille	lait	legume	alcool
Min.	293	341	1437	526	235	428	258
1st Qu.	381.8	382.8	1522	564.8	246	596.8	310.2
Median	422	453.5	1852	760.5	321.5	733	385
Mean	446.7	505	1887	803.2	358.2	732	368.6
3rd Qu.	519.8	576.8	2128	982.2	434.2	802.5	418.8
Max.	655	887	2630	1167	561	1097	486
sd	107.148	165.092	395.75	249.561	117.127	189.18	71.7818

Matrice des corrélations

	pain	fruit	viande	volaille	lait	legume	alcool
pain	1	0.19614	0.32127	0.24801	0.85557	0.59311	0.30376
fruit	0.19614	1	0.95948	0.92554	0.33219	0.85625	-0.4863
viande	0.32127	0.95948	1	0.98179	0.37459	0.88108	-0.4372
volaille	0.24801	0.92554	0.98179	1	0.23289	0.82678	-0.4002
lait	0.85557	0.33219	0.37459	0.23289	1	0.6628	0.00688
legume	0.59311	0.85625	0.88108	0.82678	0.6628	1	-0.3565
alcool	0.30376	-0.4863	-0.4372	-0.4002	0.00688	-0.3565	1



- Comment ces informations sont lièes entre elles
- Comment se comportent les différentes catégories socio-professionnelles entre elles
- Comment se comportent les différentes catégories socio-profesionnelles vis-à-vis de la consommation alimentaire

Pour rappel : ACP - Etude de variables quantitatives

- Comment se structurent les variables
  - Liaisons entre les variables
- La ressemblance entre les individus
  - Distance entre les indivdus

Positionnement

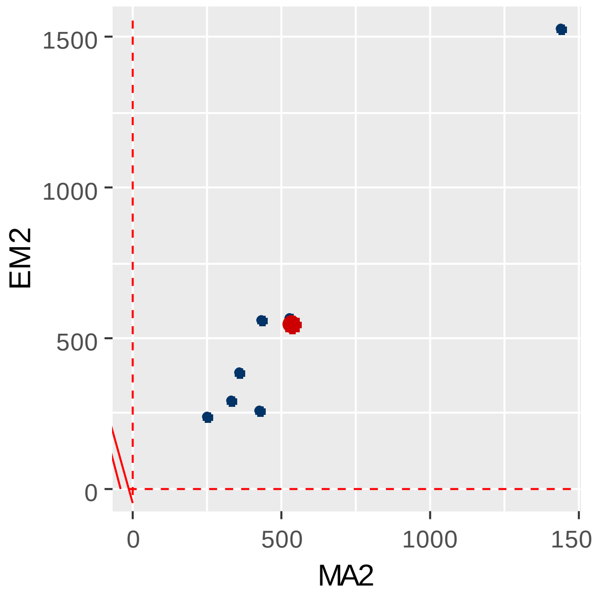
Standardisation

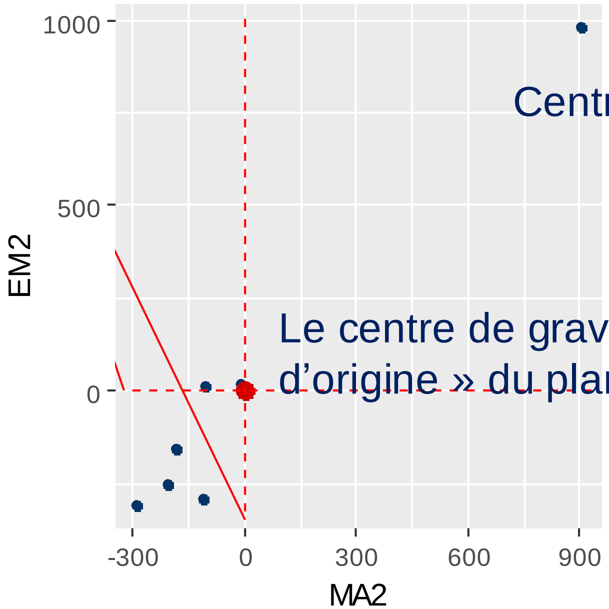
- L'ACP étant une méthode par nature géométrique, le positionnement des individus et des variables dans les plans factoriels s'effectuera de manière directe ou indirecte par la calcul de distances (principalement euclidiennes)
  Les données sont, par nature, dépendantes de leur mesure (métrique), elles n'ont donc pas le même poids
- Pour que chaque données est le même poids c.a.d apporte la « même quantité d'information », on standardise les données
- Modification d'échelle sans modifier la « structure de l'information » (forme du nuage de points)

→ Centrage ou centrage et réduction

$$z_{i,j} = \frac{x_{i,j} - E(x_{.j})}{\sqrt{V(x_{.j})}}$$

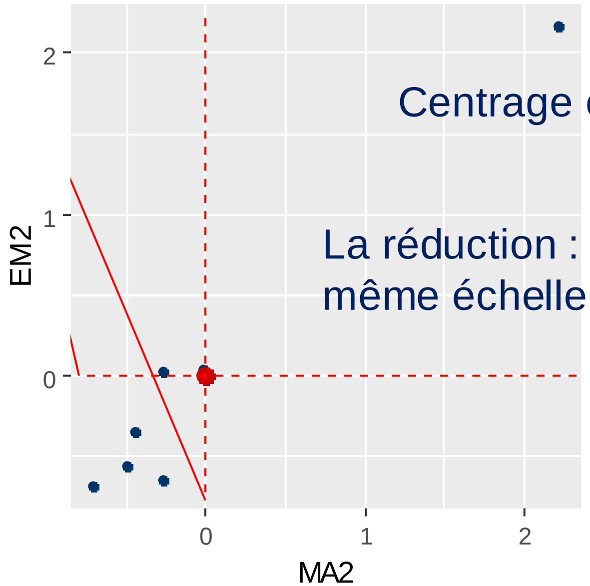
Données initiales





Centrage

Le centre de gravité (la moyenne) devient le « point d'origine » du plan



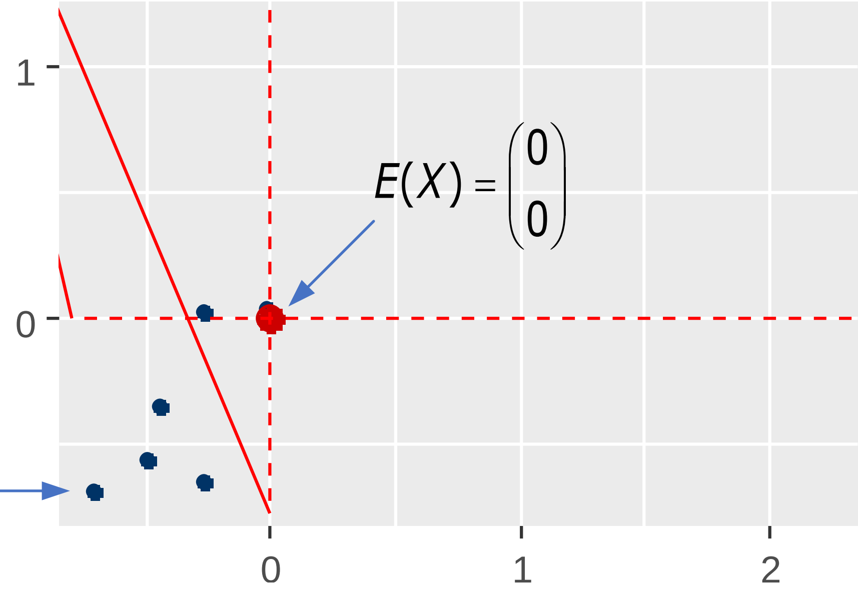
Centrage et réduction

La réduction : ramène toutes les données à la même échelle…

- On effectue un changement de repère (changement de base : cf. cours de géométrie)

	MA2	EM2	distance
Pain	-0,501	-0,562	0,567
fruit	-0,447	-0,352	0,324
viande	2,214	2,165	9,589
volaille	-0,024	0,044	0,002
lait	-0,710	-0,681	0,968
légume	-0,265	0,026	0,071
alcool	-0,267	-0,639	0,480
		Sum	12,000
		Moyenne	1,714

$$M = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$



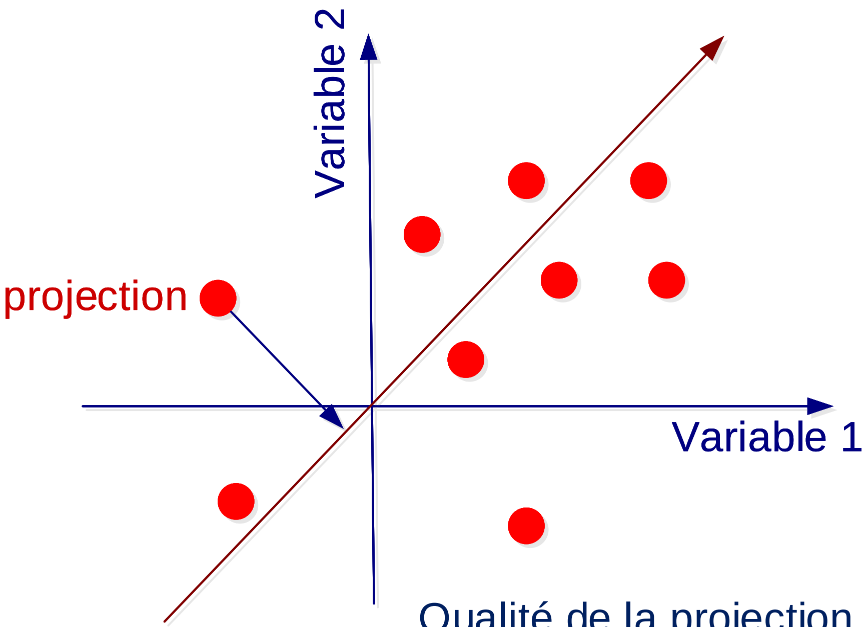
$$\overrightarrow{GM} = \begin{pmatrix} x_1 - E(x_1) \\ x_2 - E(x_2) \end{pmatrix} = \begin{pmatrix} x_1 - 0 \\ x_2 - 0 \end{pmatrix}$$

$$D_{GM}^2 = (x_1 - E(x_1))^2 + (x_2 - E(x_1))^2 = x_1^2 + x_2^2$$

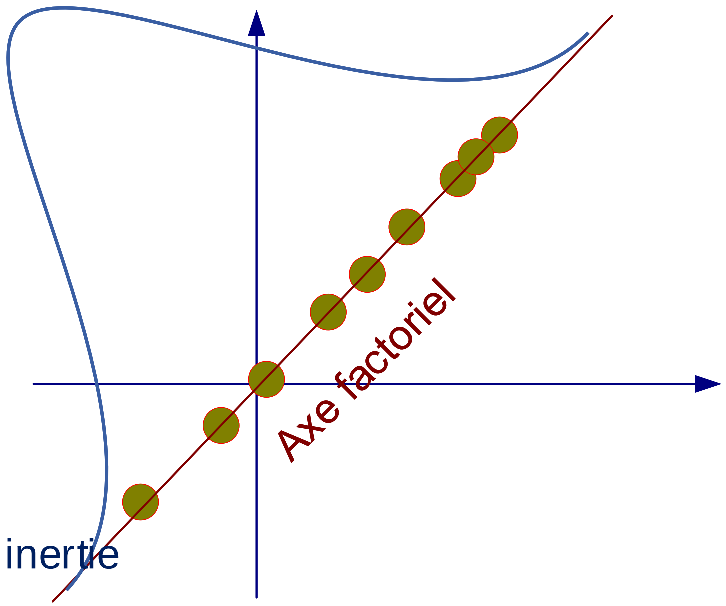
$$\frac{\sum_{i=1}^{n} D_{GM_i}^2}{n} = \frac{\sum_{i=1}^{n} x_{i,1}^2 + x_{i,2}^2}{n} = E(D_{GM}^2) = E(x - E(x))^2$$

La moyenne de la dispersion des points (par rapport au centre de gravité) équivaut à la variance

On cherche donc les axes (cf. ultérieurement) factoriels (= projections) qui préservent au mieux cette variance.



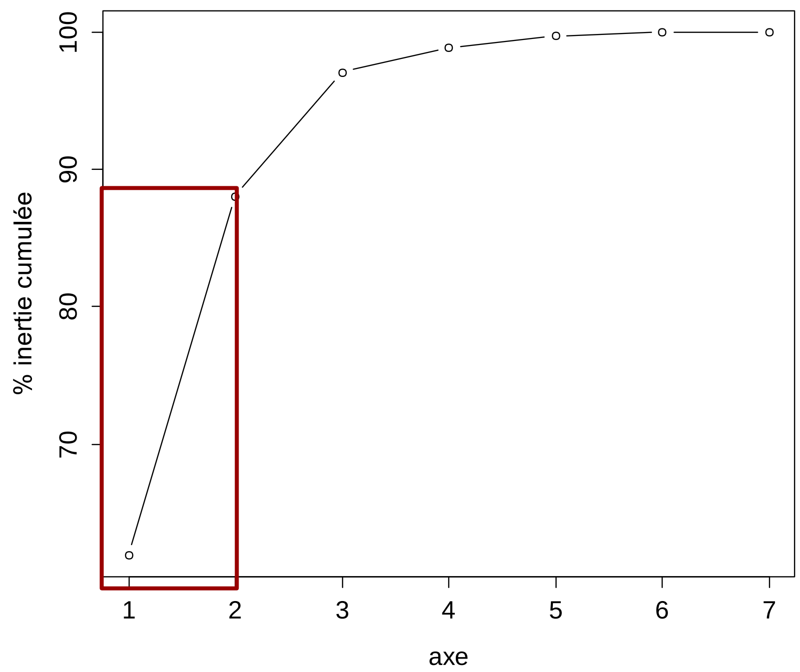
Qualité de la projection s'exprime par l' inertie



L'inertie est le pourcentage de variance expliquée par un axe factoriel (par rapport à la variance initiale du nuage de point) . Il s'agit donc de la quantité d'information exprimée par un axe. C'est la première étape de ACP

Inerties : « Quantités d'information » apportées par les axes factorielles

axe	% inertie cumulée
1	61.9
2	88.1
3	97.1
4	98.9
5	99.7
6	100.0
7	100.0

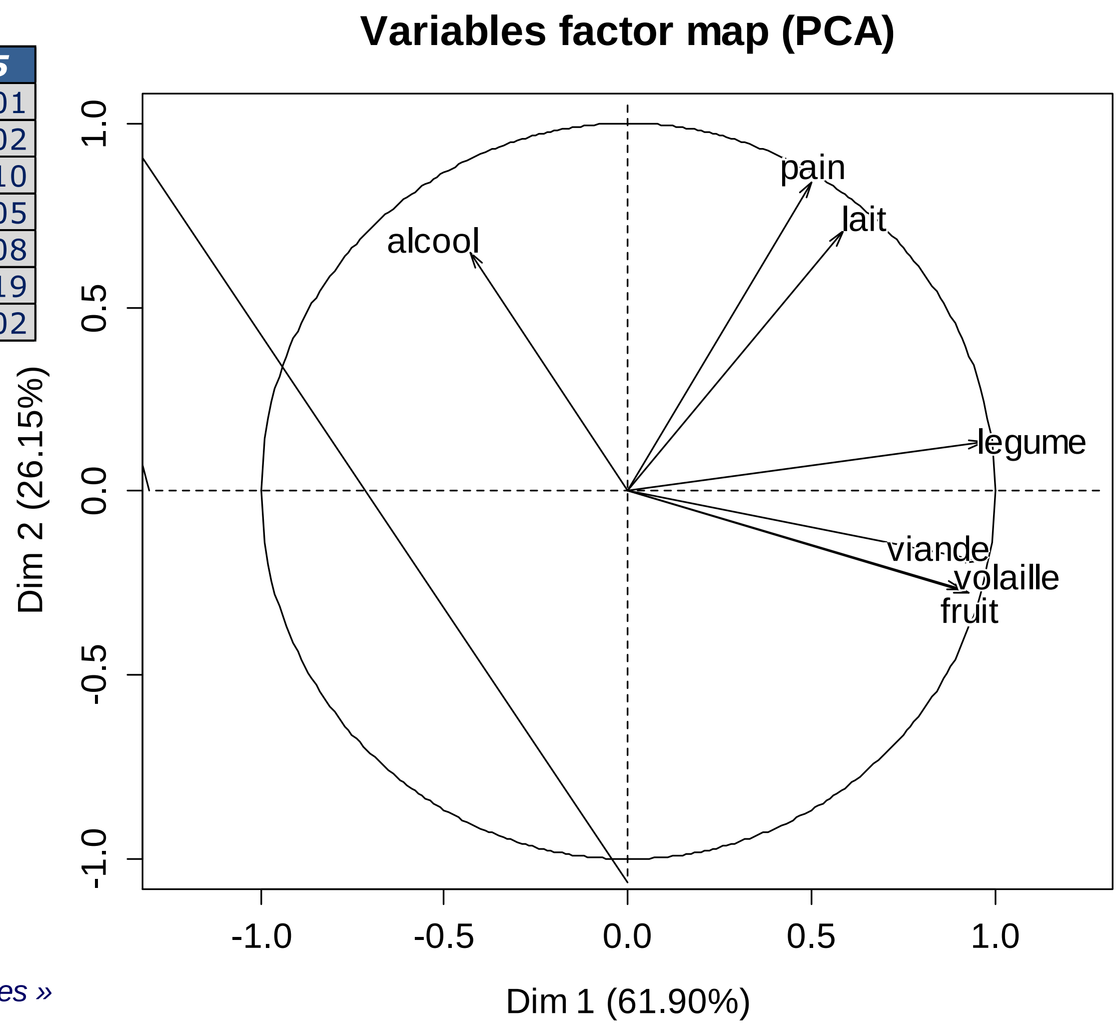


- Les deux premiers axes « expliquent » 88.1 % de la quantité d'information initiale (ensemble des observations).

- Autrement dit, la représentation des variables / individus dans le plan formé par les deux axes factoriels « explique » 88.1% de la « quantité d'information » totale. On est donc passé de 7 variables à deux composantes (axes principaux) qui seront à interpréter en fonction des variables et des individus

Représentation des variables dans le plan factoriel (axe 1 – axe 2)

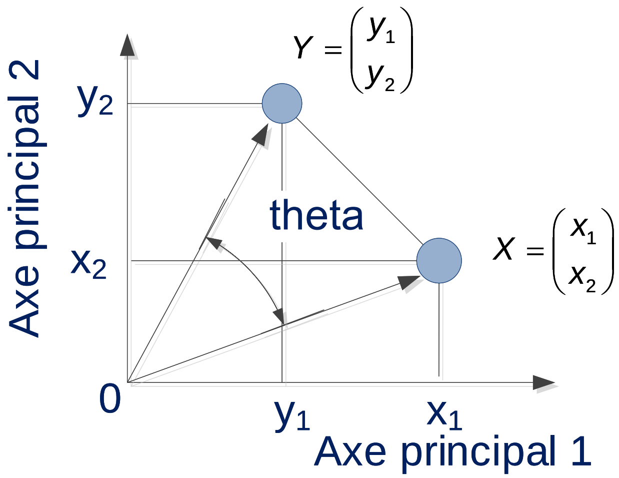
	axes 1	axes 2	axes 3	axes 4	axes 5
pain	0.50	0.84	-0.01	-0.19	0.01
fruit	0.93	-0.28	0.12	0.20	-0.02
viande	0.96	-0.19	0.16	-0.02	0.10
volaille	0.91	-0.27	0.28	-0.12	0.05
lait	0.58	0.71	-0.35	0.16	0.08
legume	0.97	0.13	-0.05	-0.01	-0.19
alcool	-0.43	0.65	0.62	0.11	-0.02



Coordonnées et représentation des variables sur les deux premiers axes factoriels

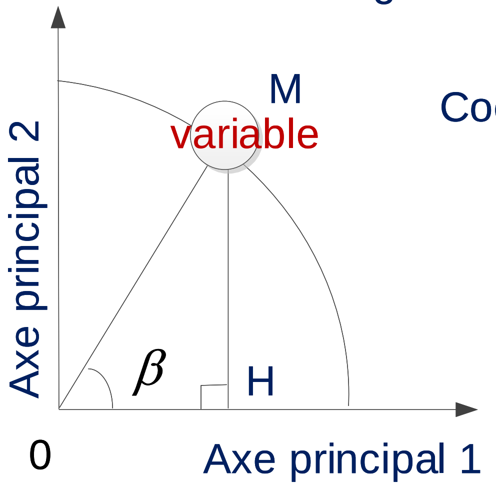
Rmq : On représente généralement les variables par des « flèches »

Qualité de représentation des variables



$$\cos(\theta) = \frac{\langle X, Y \rangle}{\|X\|\|Y\|} \longrightarrow \frac{x_1 y_1 + x_2 y_2}{\sqrt{x_1^2 + x_2^2}\sqrt{y_1^2 + y_2^2}} = \frac{\sum_{i=1}^{2} x_i y_i}{\sqrt{\sum_{i=1}^{2} x_i^2}\sqrt{\sum_{i=1}^{2} y_i^2}} = r$$

Le cosinus de l'angle formé par les vecteurs V1 et V2 correspond à la **corrélation** entre les deux variables. Plus l'angle formé entre deux variables est « petit », meilleure sera la corrélation



Coordonnées des points sur l'axe principal $\cos\beta = \dfrac{OH}{OM}$

$\cos^2\beta = OH^2$ est appelée qualité de représentation

Plus l'angle béta est faible, meilleure est la représentation de la variable sur l'axe factoriel

Qualité de représentation des variables dans le plan

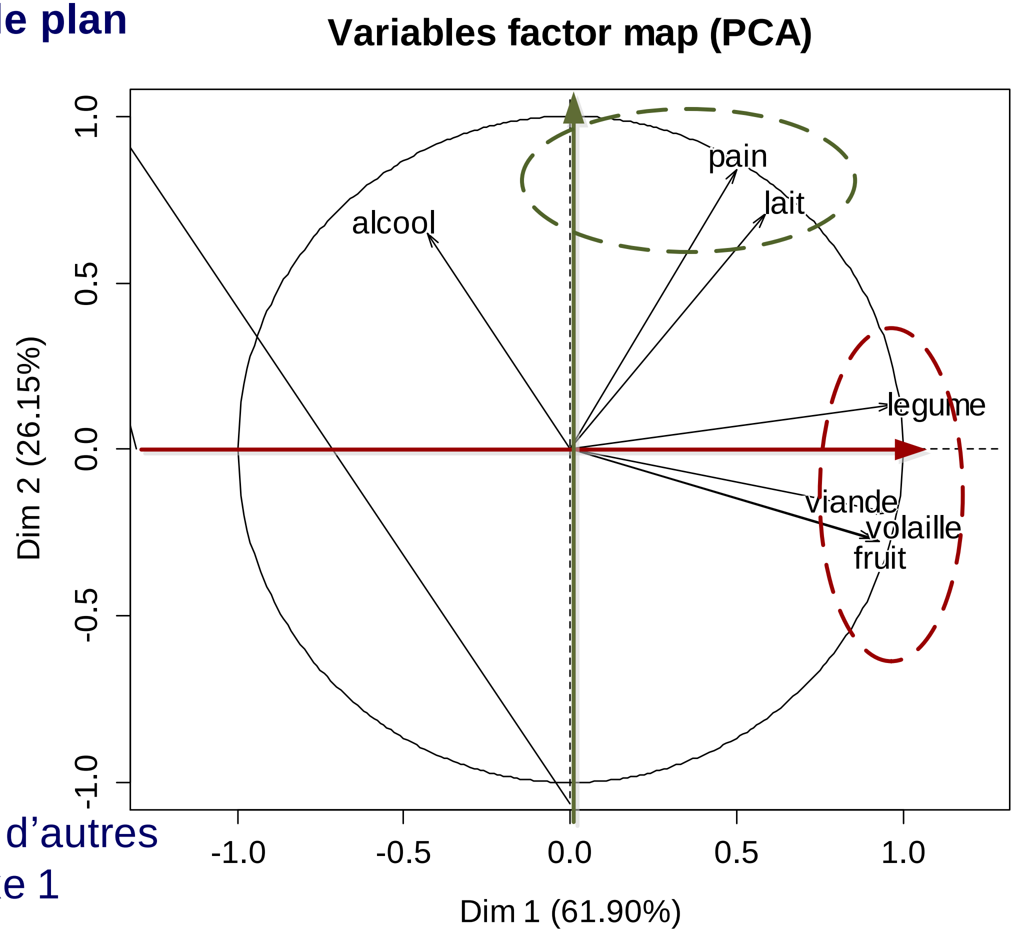
→ Expression en pourcentage

	axe 1	Axe 2	Sum qtl
pain	0.25	0.71	0.96
fruit	0.86	0.08	0.94
viande	0.93	0.04	0.96
volaille	0.83	0.07	0.90
lait	0.34	0.50	0.84
legume	0.94	0.02	0.96
alcool	0.18	0.42	0.60

Exemple de la variable viande

la qualité de représentation est :

- 93 % sur l'axe 1
- 4 % sur l'axe 2
- 97% dans le plan



→ Cette variable va permettre d'interpréter avec d'autres (légume-0.94, volaille – 0.93 , fruit – 0.86) l'axe 1

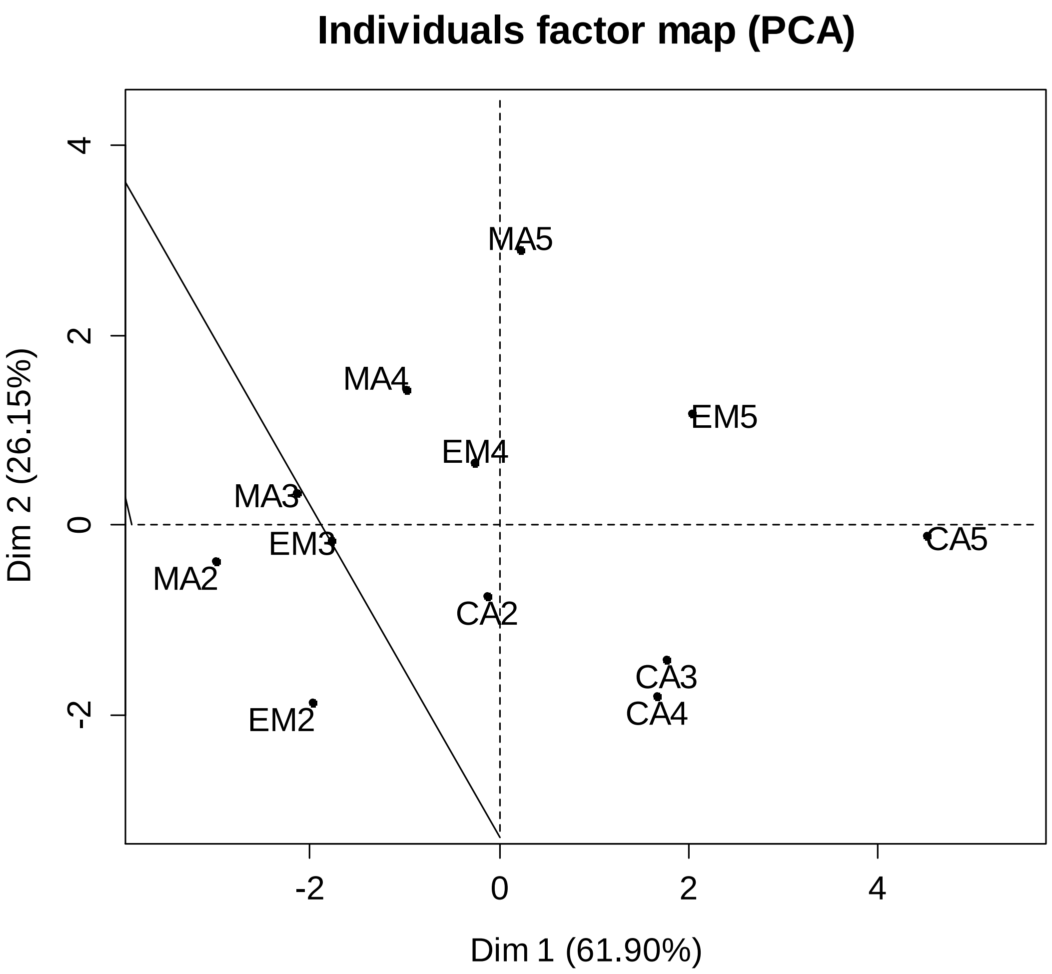
→ L'axe 1 oppose les catégories socio-professionnelles qui consomment préférentiellement ces aliments à ceux qui n'en consomment pas ou peu

→ L'axe 2 oppose les catégories socio-professionnelles qui consomment préférentiellement du pain et du lait à ceux qui n'en consomment pas ou peu

→ L'interprétation des deux axes est **INDEPENDANTE** l'une de l'autre (cf. transparents suivants)

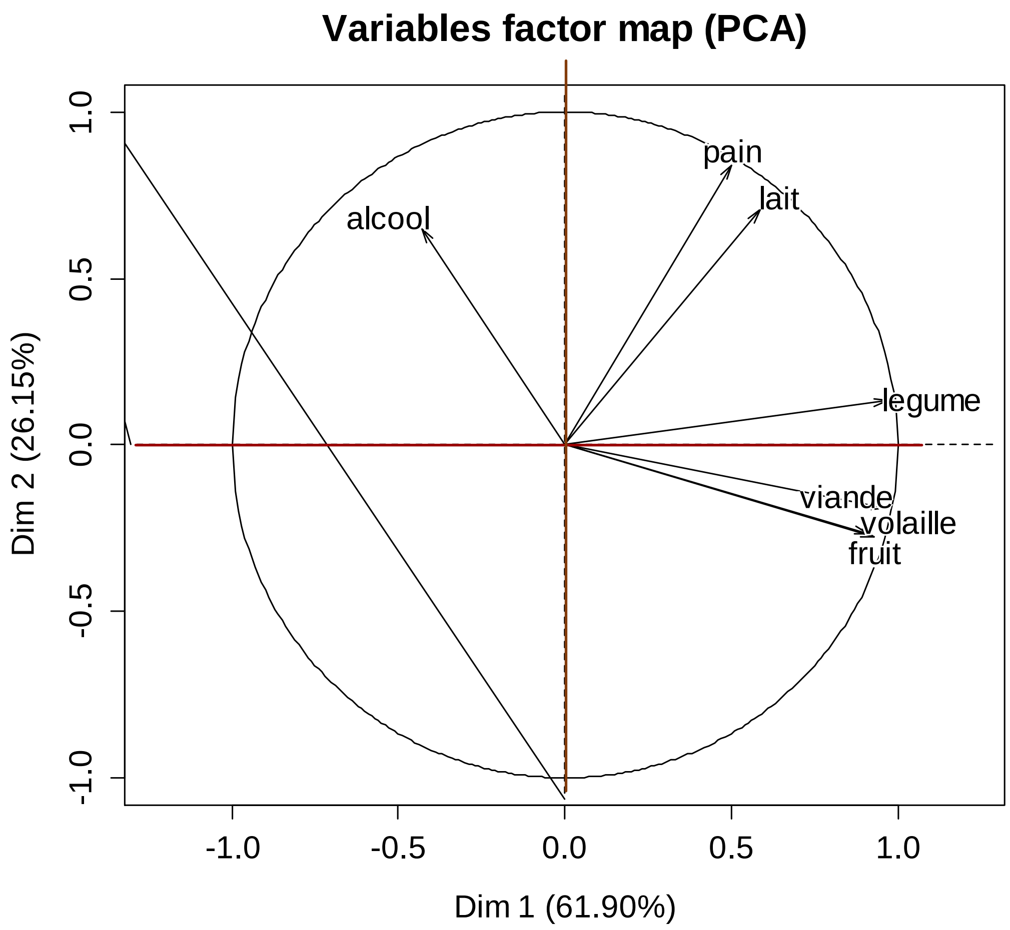
Coordonnées des individus dans le plan factoriel

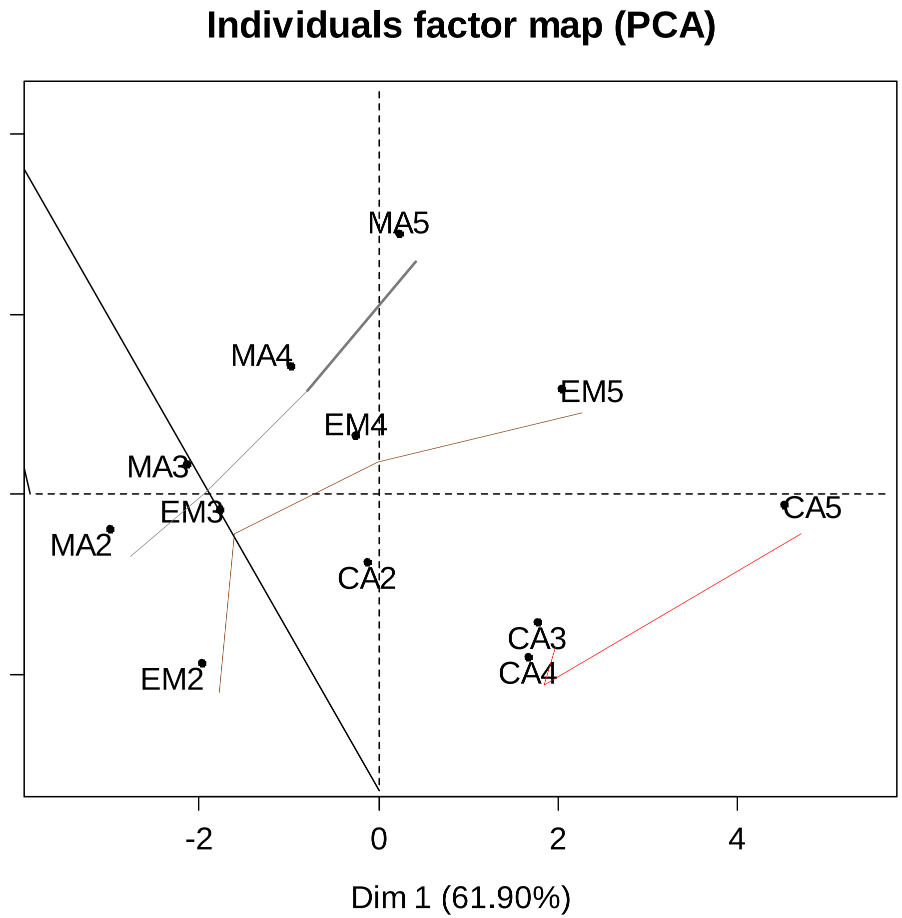
	axes 1	axes 2
MA2	-2.99	-0.38
EM2	-1.97	-1.87
CA2	-0.12	-0.76
MA3	-2.13	0.34
EM3	-1.77	-0.17
CA3	1.77	-1.42
MA4	-0.97	1.43
EM4	-0.26	0.66
CA4	1.67	-1.81
MA5	0.23	2.90
EM5	2.04	1.18
CA5	4.51	-0.11



Qualité de représentation des individus dans le plan factoriel

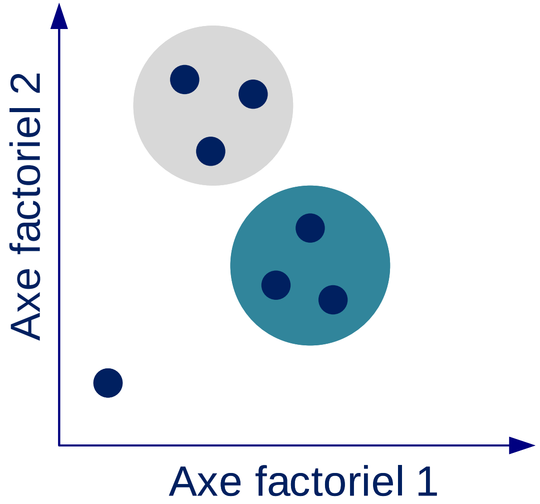
	Dim.1	Dim.2	Somme
MA2	0.94	0.02	0.96
EM2	0.42	0.38	0.80
CA2	0.00	0.19	0.19
MA3	0.97	0.02	0.99
EM3	0.89	0.01	0.90
CA3	0.48	0.31	0.79
MA4	0.30	0.65	0.94
EM4	0.10	0.61	0.70
CA4	0.43	0.50	0.93
MA5	0.01	0.94	0.95
EM5	0.60	0.20	0.81
CA5	0.96	0.00	0.96



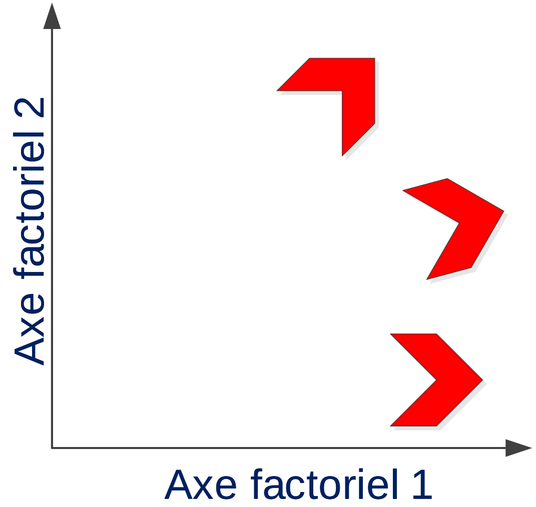


- Comportements de consommation sont différents entre les catégories socio-professionnelles
- Pour chaque catégorie, l'évolution de la consommation est fonction du nombre d'enfants

Représentation des individus dans le plan facoriel

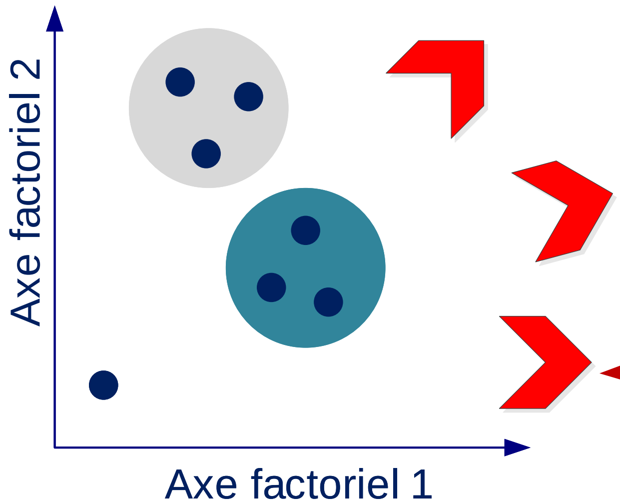


Représentation des variables dans le plan facoriel



Espaces « duals »

Interprétation : positionnement des variables en fonction des individus



Interprétation : positionnement des individus en fonction des variables

Espace des variables

	Var 1	Var 2	Var 3	….
Ind 1	X	Y	Z	….
Ind 2				….
Ind 3				….
….				….

Espace des individus

	Var 1	Var 2	Var 3	….
Ind 1		X		….
Ind 2		Y		….
Ind 3		Z		….
….				….

Par défaut, on travail dans cet espace : les règles de correspondances sont définies ultérieurement

Ajustement du nuage des individus dans l'espace des variables

Ajustement du nuage des variables dans l'espace des individus

Dualité des espaces

Nuage dans R^p ⟷ Nuage dans R^n

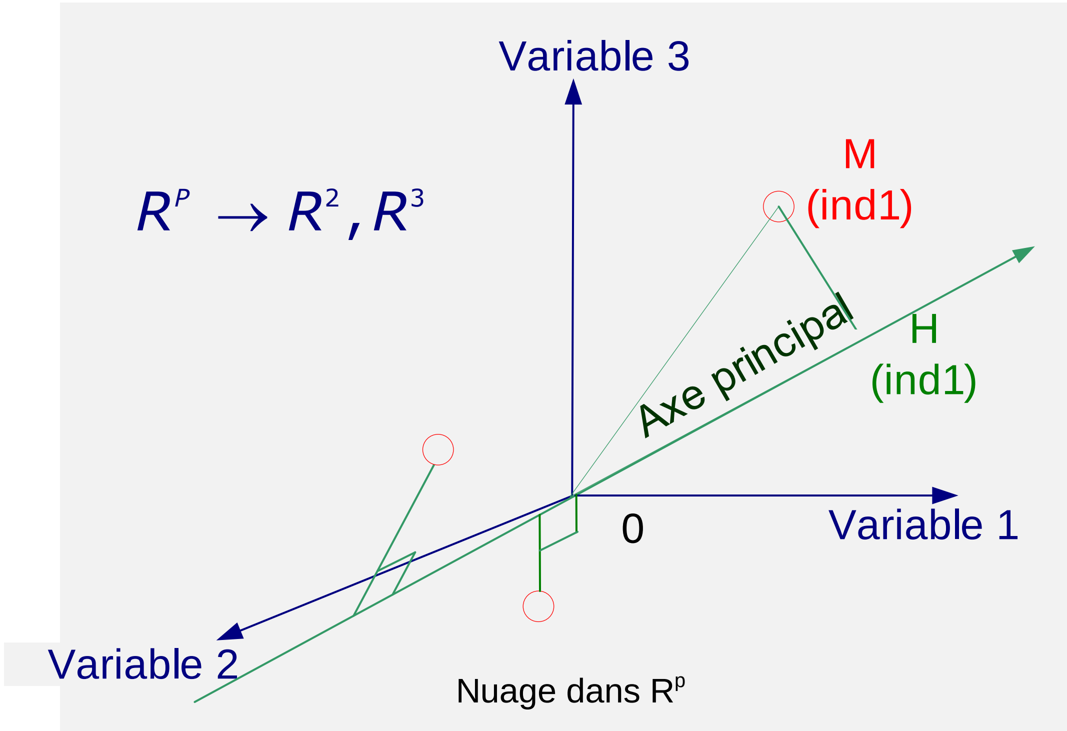
	Var 1	Var 2	Var 3	....
Ind 1	X	Y	Z	....
Ind 2				....
Ind 3				....
....				....

Nuage dans R^p

$$z_{i,1} = \frac{x_{i,1} - E(x_1)}{\sqrt{V(x_1)}}$$

On cherche un premier sous espace vectoriel à une dimension (droite) telle que la distance entre O et H soit la plus grande possible (c.a.d conservation la plus fidèle possible de la distance OM). cela équivaut à rendre d(MH) minimum

→ Représentation des individus dans l'espaces des variables



Conservation « optimale de la dispersion »

$$\max\left[\sum_i d\left(OM_i\right)^2\right] = \min\left[\sum_i d\left(MH_i\right)^2\right]$$

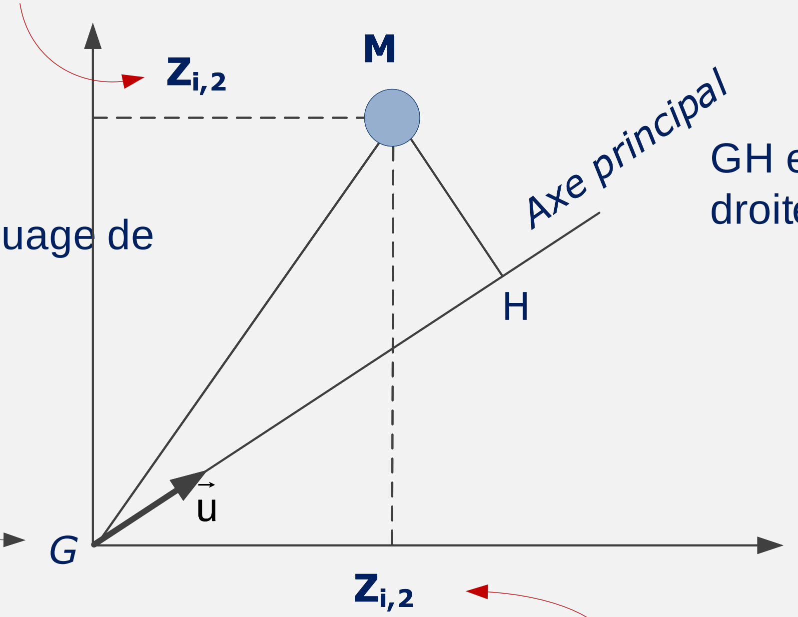
Puis on cherche un second sous espace vectoriel à une dimension (droite) **et perpendiculaire** au premier telle que la distances entre O et H soit la plus grande possible (c.a.d conservation la plus fidèle possible de la distance OM).

On cherche donc des axes de projections qui conservent au maximum l'information **et perpendiculaires entre eux.**

Coordonnée centrée réduite de l'observation i (individus) sur la variable 1

Centre de gravité du nuage de points (moyennes)

$E(z_1)=E(z_2)=0$



GH est donc la projection de OM sur la droite D

Coordonnée centrée réduite de l'observation i (individus) sur la variable 2

Coordonnée du centre de gravité

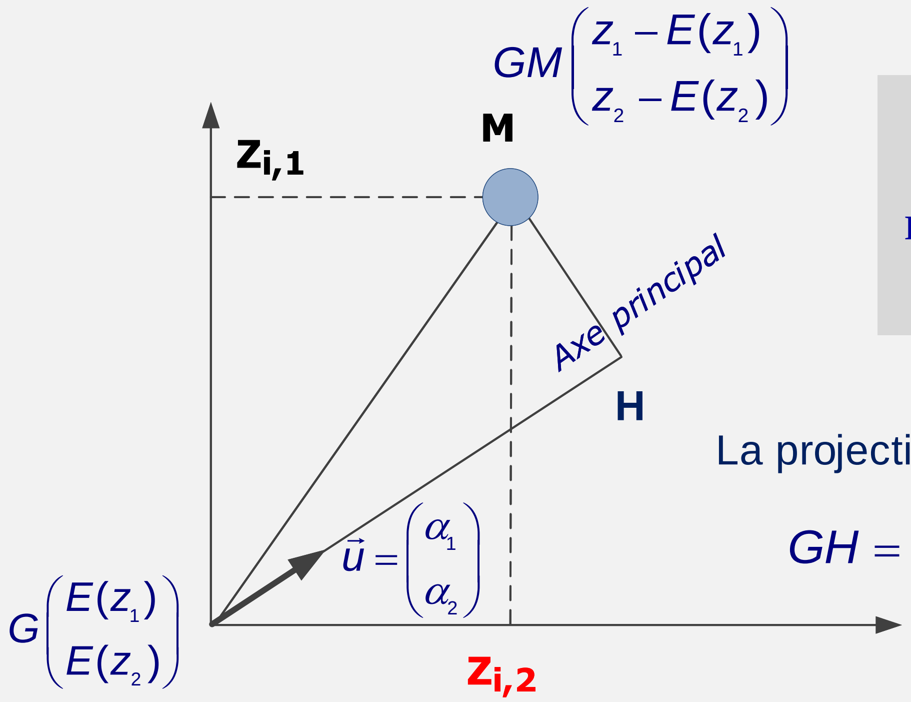
$$G\begin{pmatrix} E(z_1) \\ E(z_2) \end{pmatrix}$$

Coordonnée de M par rapport à G

$$GM\begin{pmatrix} z_1 - E(z_1) \\ z_2 - E(z_2) \end{pmatrix}$$

Soit u le vecteur directeur de de la droite D tel que : $\vec{u} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$

L' approche calculatoire



critère

$$\max\left[\sum_i d\left(GH_i\right)^2\right] = \min\left[\sum_i d\left(MH_i\right)^2\right]$$

La projection est le produit scalaire u.GM

$$GH = \; < u.GM >= (z_1 - E(z_1), z_2 - E(z_2))\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$$

→ Calcul de d(GH)²

$$GH^2 = \left(\alpha_1\left(z_1 - E(x_1)\right) + \alpha_2\left(z_2 - E(z_2)\right)\right)^2$$

$$GH^2 = \alpha_1^2\left(z_1 - E(z_1)\right)^2 + 2\alpha_1\alpha_2\left(z_1 - E(z_1)\right)\left(z_2 - E(z_2)\right) + \alpha_2^2\left(z_2 - E(z_2)\right)^2$$

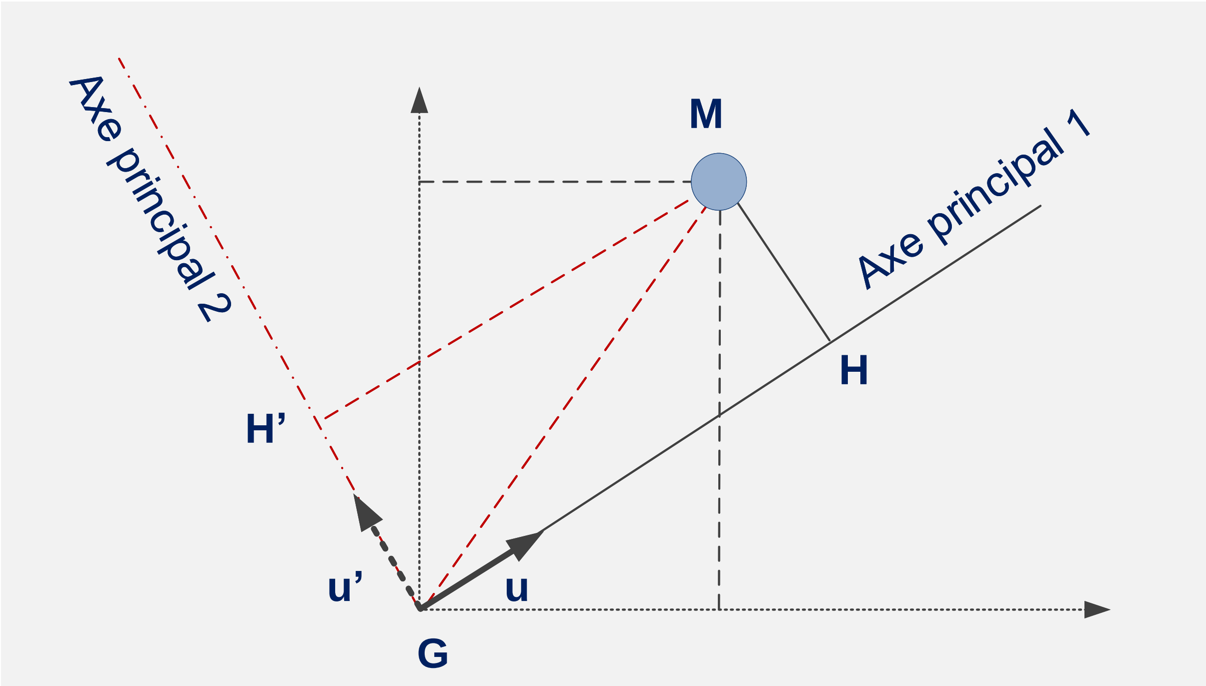
$$GH^2 = \langle u.GM \rangle^2 \quad = \alpha_1^2 V(z_1) + \alpha_2^2 V(z_2) + 2\alpha_1\alpha_2 Cov(z_1 z_2)$$

Sous forme matricielle…..

$$GH^2 = (\alpha_1, \alpha_2)^t \begin{pmatrix} V(z_1) & Cov(z_1 z_2) \\ Cov(z_1 z_2) & V(z_2) \end{pmatrix} (\alpha_1, \alpha_2)$$

$$GH^2 = \quad u^t V u$$

On cherchera ensuite une seconde droite orthogonale à la première qui elle aussi s'ajustera au mieux du nuage de points. Les projections des points sur cette droite « expliquera un peu moins d'informations » que les projections sur le premier axe.



Critère

Trouver u qui maximise u^tVu avec les contraintes suivantes

$$GH^2 = u^tVu \Rightarrow \max(GH^2) = \max(u^tVu)$$

$$\|u\| = 1 \Rightarrow \alpha_1^2 + \alpha_2^2 = 1$$

$$u^tu = 1$$

➜ Recherche du premier axe

- Objectif : Trouver u qui maximise u^tVu avec les contraintes suivantes

$$GH^2 = u^tVu \Rightarrow \max(GH^2) = \max(u^tVu)$$

$$\|u\| = 1 \Rightarrow \alpha_1^2 + \alpha_2^2 = 1$$

$$u^tu = 1$$

Repère normé

➜ Recherche du second axe

- Objectif : Trouver u' qui maximise u'^tVu' avec les contraintes suivantes

$$GH'^2 = u'^t Vu' \Rightarrow \max(GH'^2) = \max(u'^t Vu')$$

$$\|u'\| = 1 \Rightarrow \alpha'^2_1 + \alpha'^2_2 = 1 \longrightarrow u'^t u' = 1$$

Repère normé

➜ Recherche du n ième axe (n <= nombre de variables)

$$d(GH)^2 = \langle u.GM \rangle^2 = \alpha_1^2 V(z_1) + \alpha_2^2 V(z_2) + 2\alpha_1\alpha_2 Cov(z_1 z_2)$$

Sous forme matricielle

$$GH^2 = (\alpha_1, \alpha_2)^t \begin{pmatrix} V(z_1) & Cov(z_1 z_2) \\ Cov(z_1 z_2) & V(z_2) \end{pmatrix} (\alpha_1, \alpha_2)$$

Matrice des variances covariances : V

- **Maximisation**

$$\max\left[\sum_i d\left(GM_i\right)^2\right] = \min\left[\sum_i d\left(MH_i\right)^2\right]$$

La maximisation : trouver les extremums de la dérivée matricielle sous contrainte en utilisant la méthode de Lagrange

$$L = \underbrace{u^tVu}_{\text{Projection}} - \underbrace{\lambda(u^tu - 1)}_{\text{Contrainte de normalité}} \longrightarrow \underbrace{\frac{\partial L}{\partial u} = 2Vu - 2\lambda u}_{\text{la dérivée}} \longrightarrow \frac{\partial L}{\partial u} = 0 \Rightarrow \underbrace{2Vu - 2\lambda u = 0}_{\text{maxima}}$$

$$Vu = \lambda u \leftrightarrow (V - \lambda I)u = 0$$

λ est la valeur propre de V matrice des variances - covariances

u est le vecteur propre de V matrice des variances – covariances

- remarque

$$d(GH)^2 = u^tVu = u^tu\lambda = \lambda \text{ (car } u^tu = 1)$$

La valeur propre correspond donc à la variance expliquée par cet axe = inertie expliquée

La valeur propre est la même dans Rp et Rn mais pas les vecteurs propres

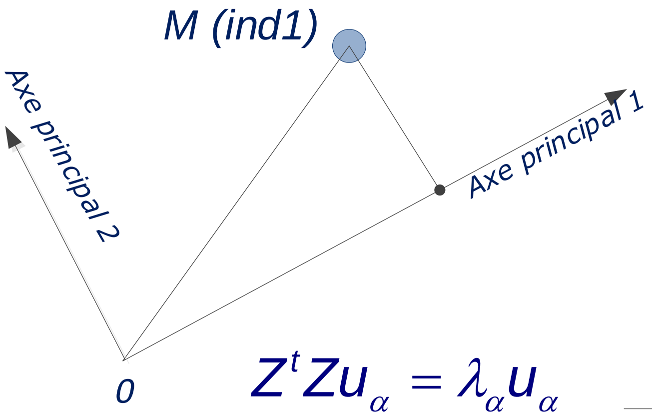
- **Formules de transition des espaces**

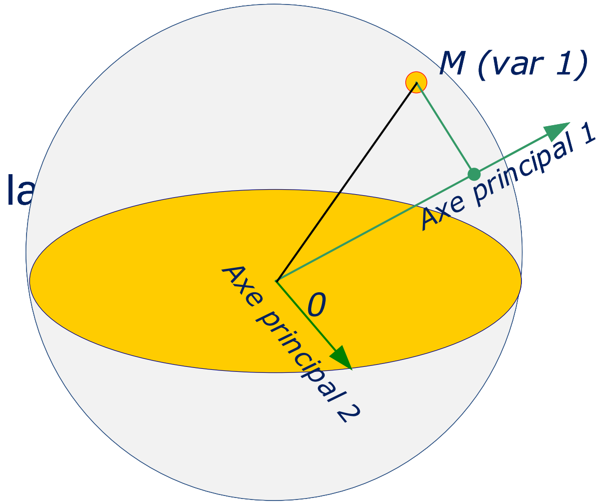
$$Z = \begin{pmatrix} z_{1,1} & \cdots & z_{p,1} \\ \vdots & \ddots & \\ z_{n,1} & \cdots & z_{n,p} \end{pmatrix} \longrightarrow Z^t Z = V$$

Espace des variables

Matrice des corrélation (ACP normée)

Données centrées réduites





$$Z^t Z u_\alpha = \lambda_\alpha u_\alpha \quad \rightleftarrows \quad Z Z^t v_\alpha = \lambda_\alpha v_\alpha$$

Espace des variables

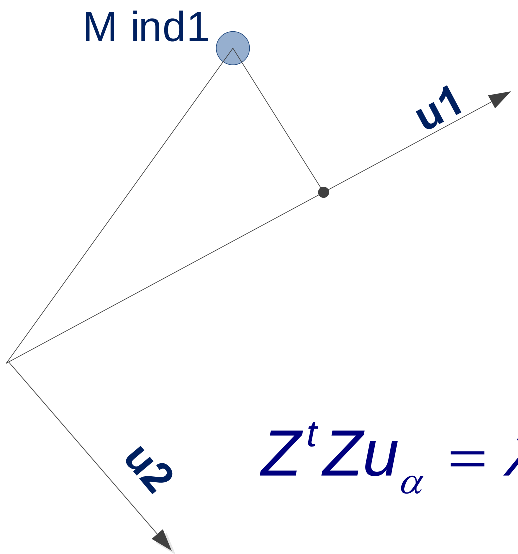
Espace des individus

Dans les deux espaces, Les valeurs propres sont les mêmes. On calcule les coordonnées des vecteurs directeurs dans les deux espaces par les relations de transitions suivantes :

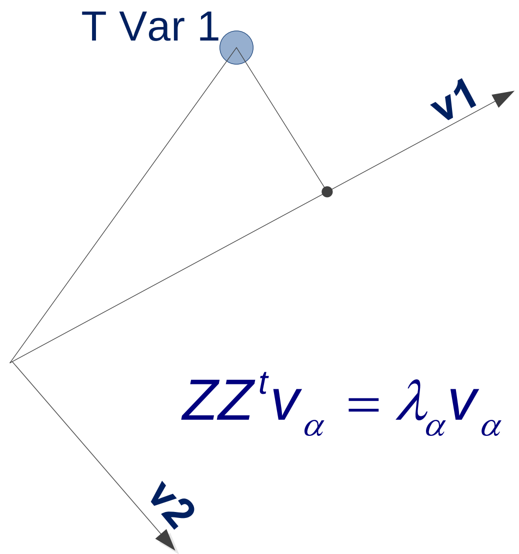
$$v_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} Z u_\alpha$$

$$u_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} Z^t v_\alpha$$

Formule de transition entre les espaces



$$\begin{cases} v_\alpha = \dfrac{1}{\sqrt{\lambda_\alpha}} Z u_\alpha \\ u_\alpha = \dfrac{1}{\sqrt{\lambda_\alpha}} Z^t v_\alpha \end{cases}$$

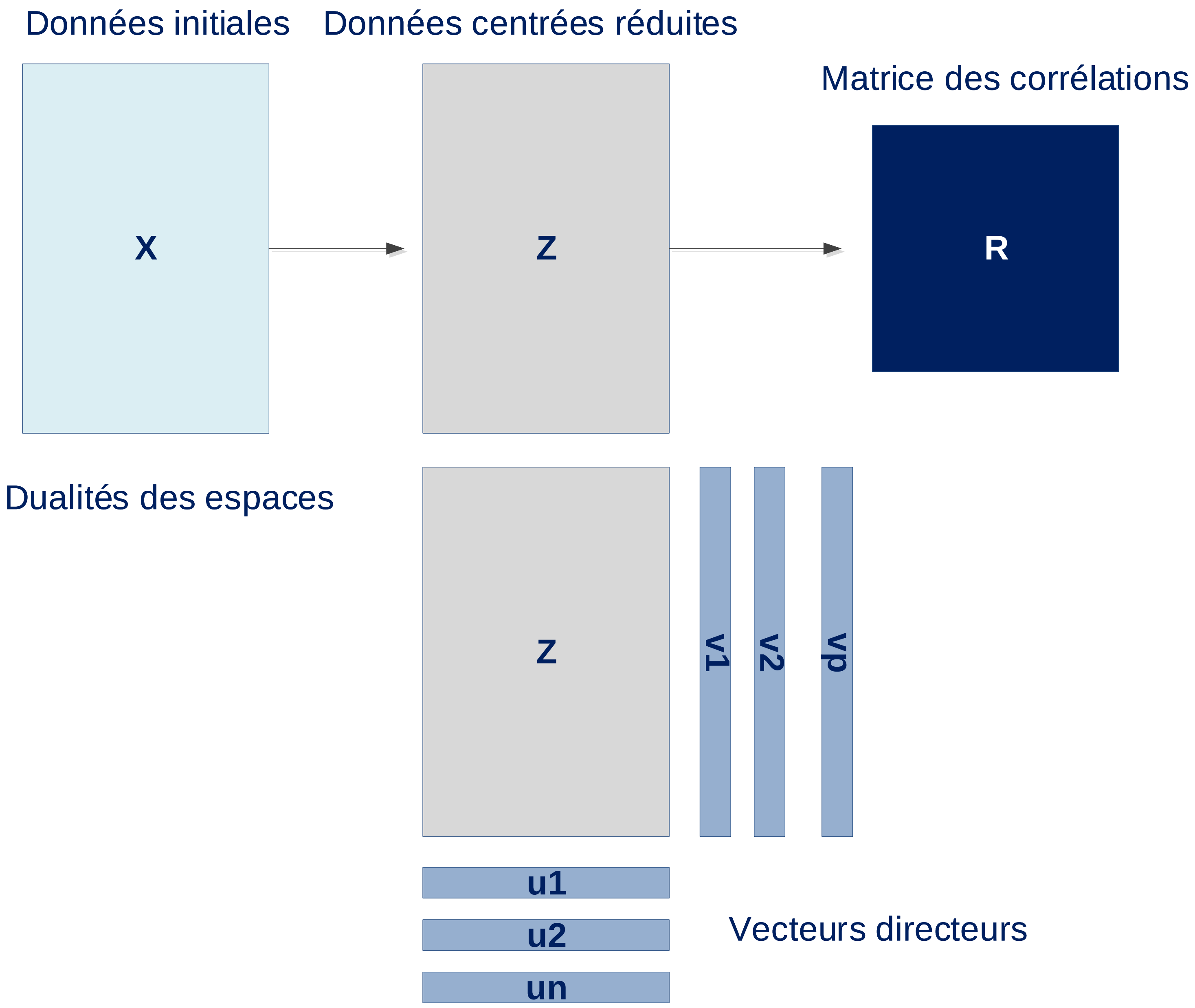


$$Z^t Z u_\alpha = \lambda_\alpha u_\alpha \xrightarrow[\text{la}]{} Z Z^t v_\alpha = \lambda_\alpha v_\alpha$$

Coordonnées des individus dans l'espace des variables

$$u_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} Z^t v_\alpha \Rightarrow \sqrt{\lambda_\alpha} u_\alpha = Z^t v_\alpha \Rightarrow \sqrt{\lambda_\alpha} u_\alpha = \varphi_\alpha$$

Coordonnées des variables dans l'espace des individus


Données initiales
Données centrées réduites
Matrice des corrélations
X
Z
R
Dualités des espaces
Z
v1
v2
vp
u1
u2
un
Vecteurs directeurs

Soit le tableau suivant dans R^2. L'objectif est donc de trouver un axe de projection (une droite qui « explique au mieux l'ensemble des informations »)

	X1	x2
Id 1	0,5	0
Id 2	-0,1	1,2
Id 3	-0,5	0,5
Id 4	-0,3	0,1
Id 5	0	2,5
Id 6	1,6	-0,7
Id 7	2	2
Id 8	2,4	1,2
Id 9	0,5	3,5
Id 10	2,7	-0,9

- Moyennes, variances, standardisation

$$X = \begin{pmatrix} x_{1,1} & x_{1,2} \\ x_{2,1} & x_{2,2} \\ \vdots & \\ x_{n,1} & x_{n,2} \end{pmatrix}$$

Matrice des données

$$\bar{X}_j = \frac{\sum_{i=1}^{n} x_{i,j}}{n}$$ moyenne

$$s_j^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x_{i,j}^2 - \frac{\left(\sum_{i=1}^{n} x_{i,j}\right)^2}{n} \right)$$ variance

Transformations des données

ACP centrée

$$X_c = \begin{pmatrix} \frac{x_{1,1}-\bar{X}_1}{\sqrt{n}} & \frac{x_{1,2}-\bar{X}_2}{\sqrt{n}} \\ \frac{x_{2,1}-\bar{X}_1}{\sqrt{n}} & \frac{x_{2,2}-\bar{X}_2}{\sqrt{n}} \\ \vdots & \\ \frac{x_{n,1}-\bar{X}_1}{\sqrt{n}} & \frac{x_{n,2}-\bar{X}_2}{\sqrt{n}} \end{pmatrix}$$

ACP Normée (standardisation)

$$Z = \begin{pmatrix} \frac{x_{1,1}-\bar{X}_1}{s_1\sqrt{n}} & \frac{x_{1,2}-\bar{X}_2}{s_2\sqrt{n}} \\ \frac{x_{2,1}-\bar{X}_1}{s_1\sqrt{n}} & \frac{x_{2,2}-\bar{X}_2}{s_2\sqrt{n}} \\ \vdots & \\ \frac{x_{n,1}-\bar{X}_1}{s_1\sqrt{n}} & \frac{x_{n,2}-\bar{X}_2}{s_2\sqrt{n}} \end{pmatrix}$$

Matrice des variances covariances

$$X_c^t X_c = \begin{pmatrix} V(X_1) & Cov(X_1, X_2) \\ Cov(X_1, X_2) & V(X_2) \end{pmatrix}$$

Matrice des corrélations

$$Z^t Z = \begin{pmatrix} 1 & r(X_1, X_2) \\ r(X_1, X_2) & 1 \end{pmatrix}$$

$$X = \begin{array}{cc} X_1 & X_2 \\ \left(\begin{array}{cc} 0.5 & 0 \\ -0.1 & 1.2 \\ -0.5 & 0.5 \\ -0.3 & 0.1 \\ 0 & 2.5 \\ 1.6 & -0.7 \\ 2 & 2 \\ 2.4 & 1.2 \\ 0.5 & 3.5 \\ 2.7 & -0.9 \end{array}\right) & \end{array} \quad Z = \begin{pmatrix} -0.107 & -0.221 \\ -0.275 & 0.061 \\ -0.387 & -0.103 \\ -0.331 & -0.197 \\ -0.247 & 0.367 \\ 0.202 & -0.385 \\ 0.314 & 0.249 \\ 0.426 & 0.061 \\ -0.107 & 0.602 \\ 0.510 & -0.432 \end{pmatrix} \quad Z^t Z = \begin{pmatrix} 1.000 & -0.237 \\ -0.237 & 1.000 \end{pmatrix}$$

Calcul des valeurs propres

$$(Z^t Z)u = \lambda u \Rightarrow (Z^t Z - \lambda I)u = 0$$

Pour calculer les valeurs propres, on calcule les valeurs qui annulent le déterminant $\det(Z^t Z - \lambda I) = 0$

$$\det(Z^t Z - \lambda I) = \det\left(\begin{pmatrix} 1.000 & -0.237 \\ -0.237 & 1.000 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) \quad \det(Z^t Z - \lambda I) = \det\begin{pmatrix} 1.000-\lambda & -0.237 \\ -0.237 & 1.000-\lambda \end{pmatrix}$$

$$\det(Z^t Z - \lambda I) = (1.000-\lambda)^2 - 0.237^2 = 0$$

$$\lambda^2 - 2\lambda + 0.943 = 0$$

$$\lambda_{1,2} = \frac{2 \pm \sqrt{4 - 4(0.943)}}{2} \quad \begin{array}{l} \lambda_1 = 1.236 \\ \lambda_2 = 0.763 \end{array} \quad \Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \longrightarrow \Lambda = \begin{pmatrix} 1.236 & 0 \\ 0 & 0.763 \end{pmatrix}$$

Matrice des valeurs propres

Classement des valeurs par ordre décroissant

Calcul des valeurs propres

$$(Z^t Z)u = \lambda u \Rightarrow (Z^t Z - \lambda I)u = 0$$

Pour calculer les valeurs propres, on calcule les valeurs qui annulent le déterminant $\det(Z^t Z - \lambda I) = 0$

$$\det(Z^t Z - \lambda I) = \det\left(\begin{pmatrix} 1.000 & -0.237 \\ -0.237 & 1.000 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

$$\det(Z^t Z - \lambda I) = \det\begin{pmatrix} 1.000 - \lambda & -0.237 \\ -0.237 & 1.000 - \lambda \end{pmatrix}$$

$$\det(Z^t Z - \lambda I) = (1.000 - \lambda)^2 - 0.237^2 = 0$$

$$\lambda^2 - 2\lambda + 0.943 = 0$$

$$\lambda_{1,2} = \frac{2 \pm \sqrt{4 - 4(0.943)}}{2}$$

$$\lambda_1 = 1.236$$
$$\lambda_2 = 0.763$$

Matrice des valeurs propres

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \longrightarrow \Lambda = \begin{pmatrix} 1.236 & 0 \\ 0 & 0.763 \end{pmatrix}$$

Classement des valeurs par ordre décroissant

Calcul des vecteurs propres

$$(Z^t Z)u_1 = \lambda u_1 \Rightarrow (Z^t Z - \lambda_1 I)u_1 = 0$$

On cherche donc $u_1 = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

$$\left(\begin{pmatrix} 1.000 & -0.237 \\ -0.237 & 1.000 \end{pmatrix} - 1.236 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} -0.237 & -0.237 \\ -0.237 & -0.237 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{cases} -0.237x_1 - 0.237x_2 = 0 \\ -0.237x_1 - 0.237x_2 = 0 \end{cases} \Rightarrow x_1 + x_2 = 0$$

$$\|u_1\| = 1 \Rightarrow \sqrt{x_1^2 + x_2^2} = 1$$ → Contrainte de normalité

$$\begin{cases} x_1 + x_2 = 0 \\ \sqrt{x_1^2 + x_2^2} = 1 \end{cases} \Rightarrow x_1 = -x_2 \Rightarrow \sqrt{2x_1^2} = 1 \Rightarrow x_1 = \frac{1}{\sqrt{2}}, x_2 = -\frac{1}{\sqrt{2}} \rightarrow u_1 = \begin{pmatrix} 0.707 \\ -0.707 \end{pmatrix}$$

Première composante principale
(vecteur directeur)

Projection des individus sur le premier axe

$$u_1 = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \qquad \begin{cases} y_{axe1,1} = \alpha_1 z_{11} + \alpha_2 z_{12} \\ y_{axe1,l} = \alpha_1 z_{l1} + \alpha_2 z_{l2} \\ y_{axe1,n} = \alpha_1 z_{n1} + \alpha_2 z_{n2} \end{cases}$$

$$Zu_1 = \begin{pmatrix} -0.107 & -0.221 \\ -0.275 & 0.061 \\ -0.387 & -0.103 \\ -0.331 & -0.197 \\ -0.247 & 0.367 \\ 0.202 & -0.385 \\ 0.314 & 0.249 \\ 0.426 & 0.061 \\ -0.107 & 0.602 \\ 0.510 & -0.432 \end{pmatrix} \begin{pmatrix} 0.707 \\ -0.707 \end{pmatrix} = \begin{pmatrix} -0.081 \\ 0.238 \\ 0.201 \\ 0.094 \\ 0.434 \\ -0.415 \\ -0.046 \\ -0.258 \\ 0.501 \\ -0.667 \end{pmatrix}$$

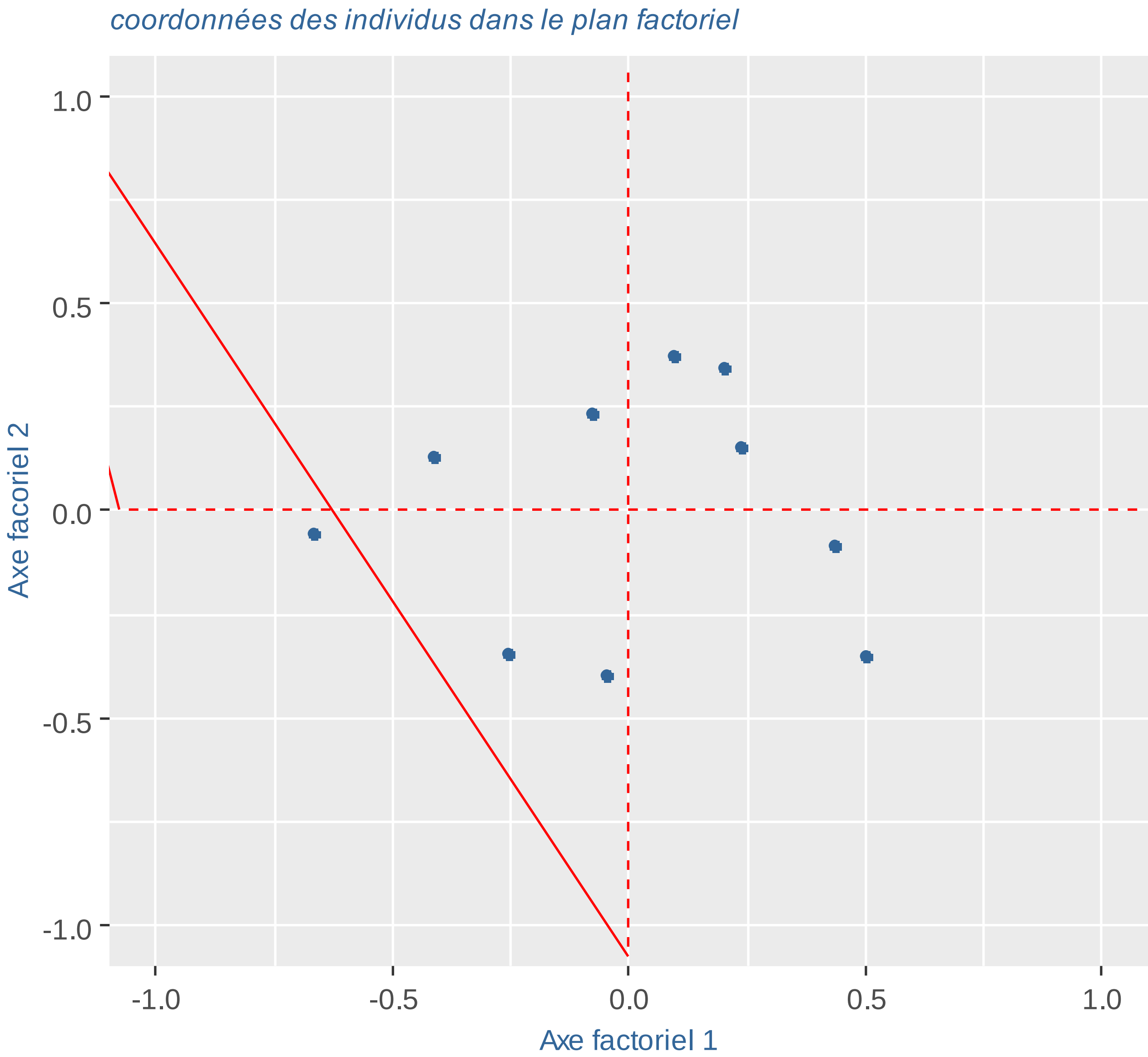
Projection des individus sur le deuxième axe

$$u_1 = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \qquad \begin{cases} y_{axe2,1} = \beta_1 z_{11} + \beta_2 z_{12} \\ y_{axe2,l} = \beta_1 z_{l1} + \beta_2 z_{l2} \\ y_{axe2,n} = \beta_1 z_{n1} + \beta_2 z_{n2} \end{cases}$$

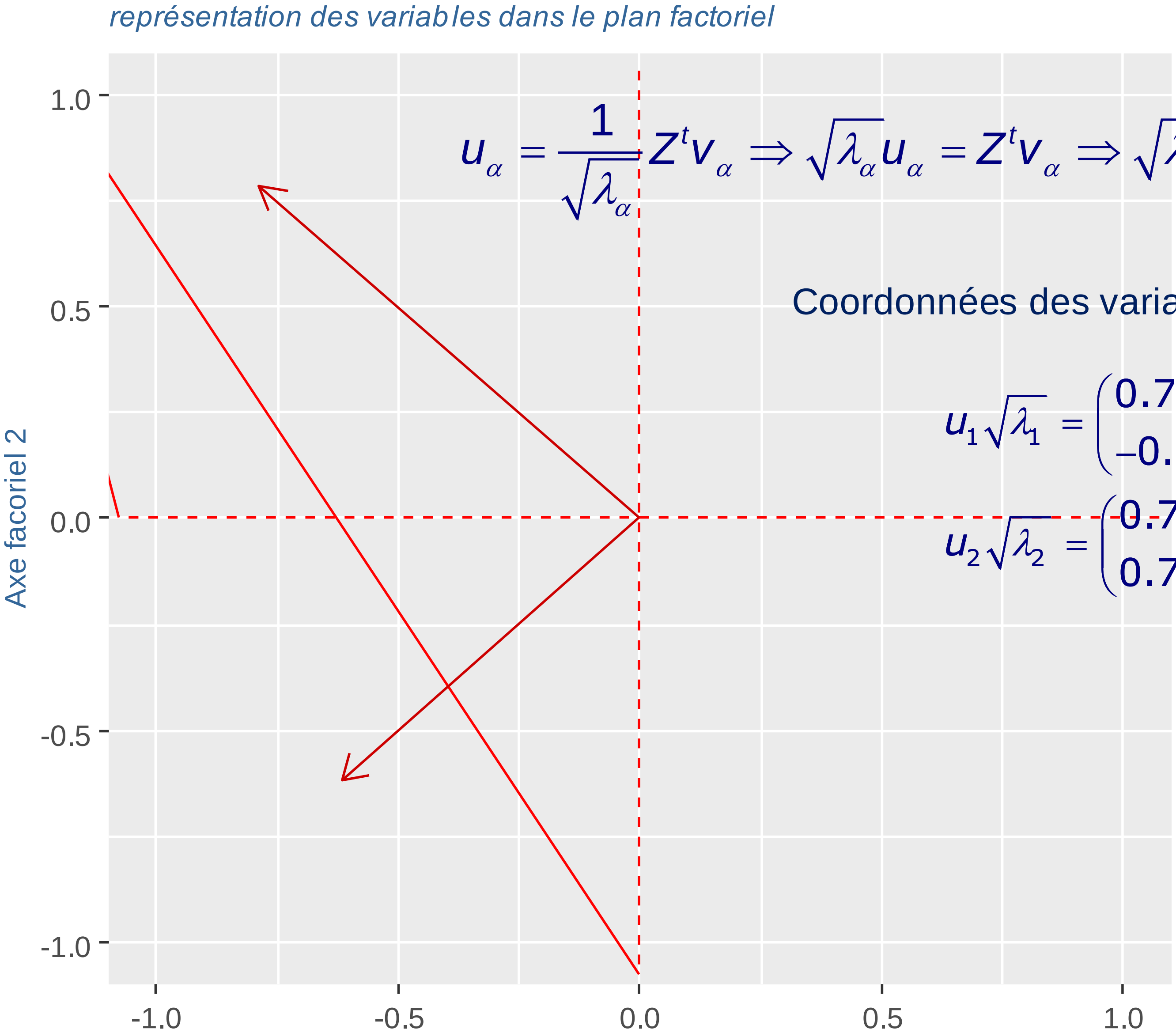
$$Zu_2 = \begin{pmatrix} -0.107 & -0.221 \\ -0.275 & 0.061 \\ -0.387 & -0.103 \\ -0.331 & -0.197 \\ -0.247 & 0.367 \\ 0.202 & -0.385 \\ 0.314 & 0.249 \\ 0.426 & 0.061 \\ -0.107 & 0.602 \\ 0.510 & -0.432 \end{pmatrix} \begin{pmatrix} 0.707 \\ 0.707 \end{pmatrix} = \begin{pmatrix} -0.232 \\ -0.151 \\ -0.347 \\ -0.374 \\ 0.085 \\ -0.130 \\ 0.398 \\ 0.345 \\ 0.350 \\ 0.055 \end{pmatrix}$$

Les coordonnées des individus = combinaisons linéaires des variables initiales

Représentation des individus dans le plan factoriel



Représentation des variables dans le plan factoriel



$$u_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} Z^t v_\alpha \Rightarrow \sqrt{\lambda_\alpha} u_\alpha = Z^t v_\alpha \Rightarrow \sqrt{\lambda_\alpha} u_\alpha = \varphi_\alpha$$

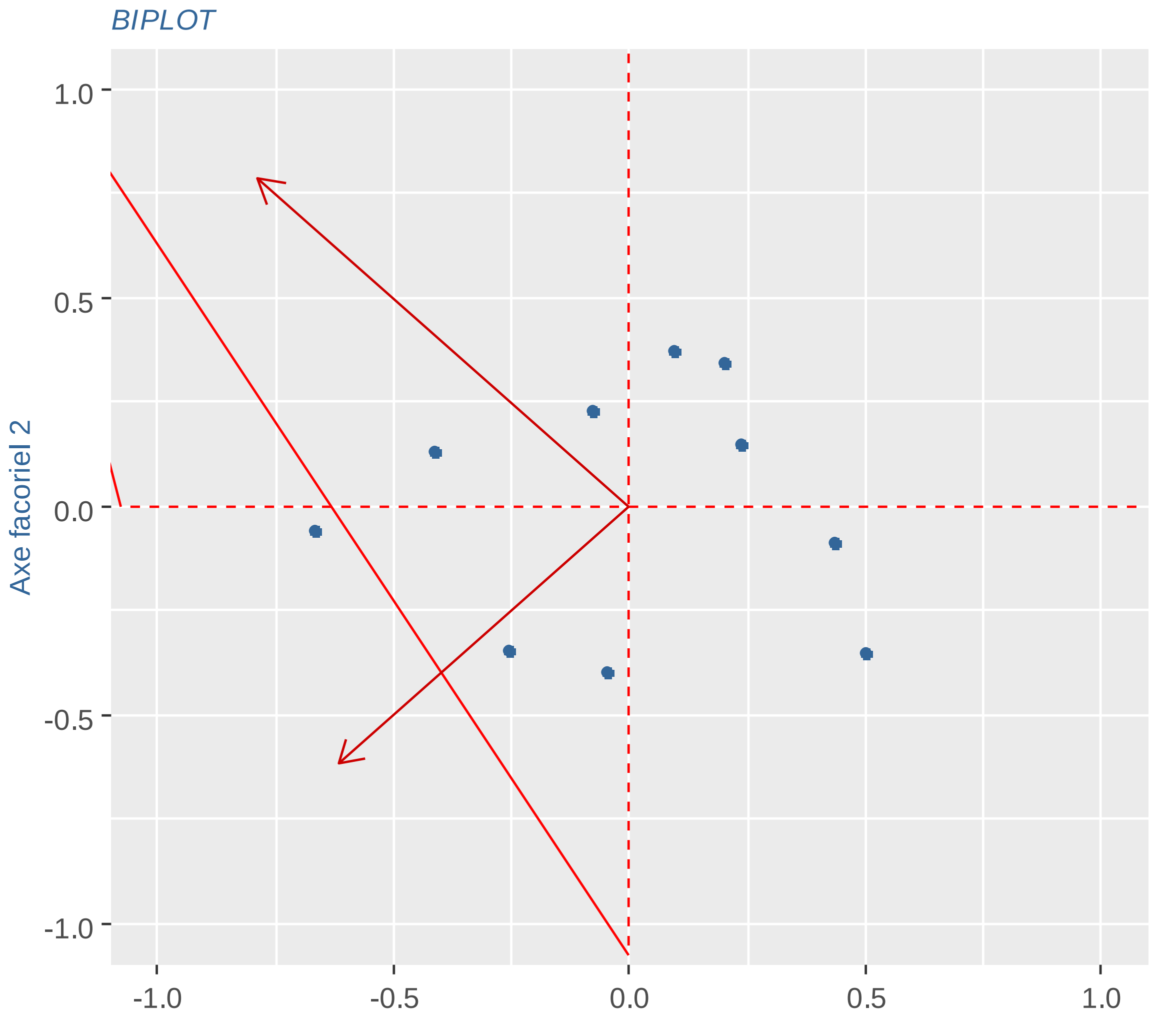
Coordonnées des variables dans l'espace des individus

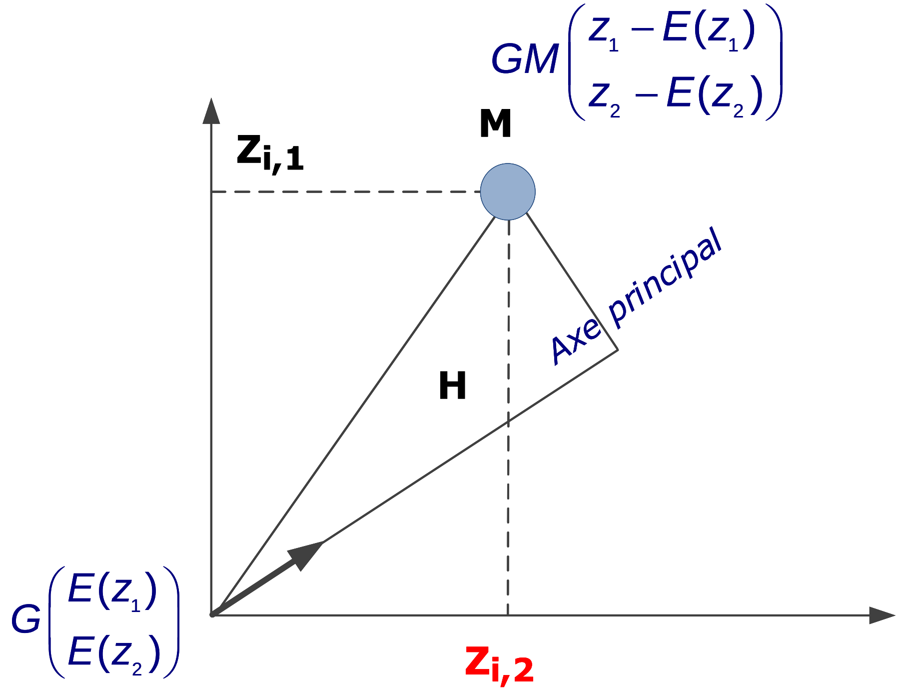
$$u_1 \sqrt{\lambda_1} = \begin{pmatrix} 0.707 \\ -0.707 \end{pmatrix} \sqrt{1.236} = \begin{pmatrix} 0.786 \\ -0.786 \end{pmatrix} = Z^t v_1$$

$$u_2 \sqrt{\lambda_2} = \begin{pmatrix} 0.707 \\ 0.707 \end{pmatrix} \sqrt{0.763} = \begin{pmatrix} 0.617 \\ 0.617 \end{pmatrix} = Z^t v_2$$

Représentation des variables et des individus

Les espaces des individus et des variables sont duals. En toute rigueur, on ne devrait pas effectuer de superposition des graphique, mais elle revêt un aspect pratique permettant de visualiser le « positionnement » des individus par rapport aux variables et réciproquement.





$$\begin{pmatrix} 1.236 & 0 \\ 0 & 0.763 \end{pmatrix} \longrightarrow \sum GH^2 = u^t V u = u^t (Z^t Z) u$$

$$\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \quad \begin{pmatrix} 0.707 & -0.707 \\ 0.707 & 0.707 \end{pmatrix} \begin{pmatrix} 1 & 0.236 \\ 0.236 & 1 \end{pmatrix} \begin{pmatrix} 0.707 & 0.707 \\ -0.707 & 0.707 \end{pmatrix}$$

Les variances expliquées par les axes correspondent aux valeurs propres

- Les valeurs propres représentent la part de variance (de dispersion) expliquée par les axes. Plus cette part de variance est importante (appelée inertie) , meilleure sera « l'information » apportée par cet axe

  On remarque $\lambda_j \geq 0$ avec $\lambda_1 \geq \lambda_j \geq \ldots \geq \lambda_p$

Inertie totale

$$\mathrm{I}_0 = \sum_{j=1}^{p} \lambda_j$$

Inertie expliquée par un axe factoriel

$$\mathrm{I}(\Delta_k) = \frac{\lambda_k}{\sum_{j=1}^{p} \lambda_j}, j = 1,\ldots,k,\ldots p$$

- L'inertie totale représente la dimension Rp. Elle correspond donc au nombre de variables initiales

Exemple suite

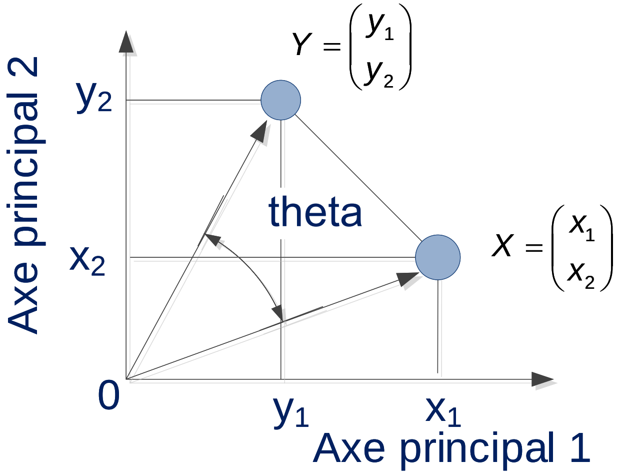
Inertie totale $I_t = 1.236 + 0.763 = 2$

$$I_1 = \frac{1.236}{2} = 0.618$$

$$I_2 = \frac{0.763}{2} = 0.362$$

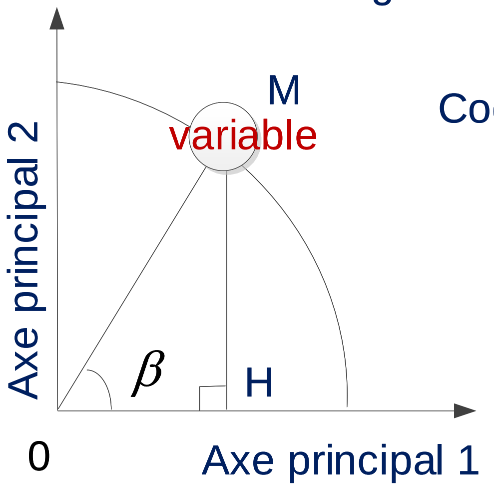
Le premier axe factoriel explique 61.8 % de la dispersion totale du nuage de point et le second 36.2%

Qualité de représentation des variables (démarche analogue pour les individus)



$$\cos(\theta) = \frac{< X, Y >}{\|X\|\|Y\|} \longrightarrow \frac{x_1 y_1 + x_2 y_2}{\sqrt{x_1^2 + x_2^2}\sqrt{y_1^2 + y_2^2}} = \frac{\sum_{i=1}^{2} x_i y_i}{\sqrt{\sum_{i=1}^{2} x_i^2}\sqrt{\sum_{i=1}^{2} y_i^2}} = r$$

Le cosinus de l'angle formé par les vecteurs V1 et V2 correspond à la **corrélation** entre les deux variables. Plus l'angle formé entre deux variables est « petit », meilleure sera la corrélation



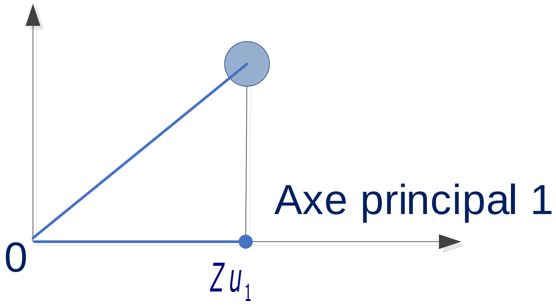
Coordonnées des points sur l'axe principal $\cos\beta = \frac{OH}{OM}$

$\cos^2 \beta = OH^2$ est appelée qualité de représentation

Plus l'angle béta est faible, meilleure est la représentation de la variable sur l'axe principal

- **Contribution relative d'un individu à l'élaboration des axes factoriels (démarche analogue pour les variables)**

« Dispersion » d'un individus sur l'axe factoriel



$$Ctr_{ind} = \frac{(Zu_\alpha)^2}{\lambda_\alpha}$$

Variance expliquée par l'axe

Premier exemple (suite)

$$Zu_\alpha = \begin{pmatrix} 0.0809 & -0.2316 \\ -0.2375 & -0.1511 \\ -0.2005 & -0.3468 \\ -0.0944 & -0.3736 \\ -0.4338 & 0.0848 \\ 0.4153 & -0.1298 \\ 0.0459 & 0.3982 \\ 0.2582 & 0.3446 \\ -0.5008 & 0.3501 \\ 0.6667 & 0.0551 \end{pmatrix}$$

$$\frac{(Zu_\alpha)^2}{\lambda_\alpha} = \begin{array}{c c c} & \text{Axe 1} \quad \text{Axe 2} & \text{Ind 1} \\ & \begin{pmatrix} 0.0053 & 0.0703 \\ 0.0456 & 0.0299 \\ 0.0325 & 0.1575 \\ 0.0072 & 0.1829 \\ 0.1521 & 0.0094 \\ 0.1395 & 0.0221 \\ 0.0017 & 0.2078 \\ 0.0539 & 0.1556 \\ 0.2028 & 0.1606 \\ 0.3594 & 0.0040 \end{pmatrix} & \downarrow \\ & \text{Alpha = 1} \quad \text{Alpha = 2} & \text{Ind n} \end{array}$$

Le choix du nombre d'axes

→ *Règles empiriques*

- *Critère de Cattel*
- *Critère de Kaiser*

→ *Critères statistiques*

- *Intervalle de confiance d'Anderson*

Règles empiriques

- *On se réfère à l'histogramme des décroissances des valeurs propres pour y déceler un variation « brutale de la pente »*
- *Si les données sont structurées (variables corrélées entre elles), le nuage a une forme irrégulière, et certain axes seront susceptibles d'avoir une inertie « importante » par rapport aux autres. On observera une décroissance inégale de la pente.*
- *Si les données sont peu structurées (variables faiblement corrélées entre elles), le nuage à une forme régulière. Dans ce cas, les valeurs propres ont une décroissance régulière et en générale, l'analyse factorielle ne fournit pas de résultats intéressants.*

- Le critère de Cattel

On ne retient que les axes dont les valeurs propres sont supérieures à 1

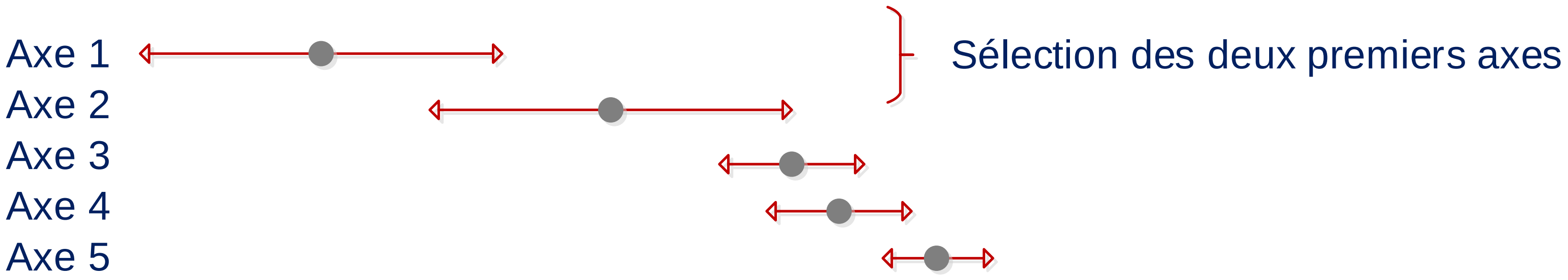
Attention cependant : Si une valeur propre est très importante par rapport aux autres, les autres valeurs seront donc très petites ce qui peut conduire à sous estimer l'importance de la prise en compte d'un autre axe factoriel.

- Intervalle de confiance d'Anderson : intervalles de confiance de variance expliquées par les axes factoriels

$$\lambda_i \in \left[\hat{\lambda}_i \pm \left(1 - z_{1-\alpha}\sqrt{2/n-1}\right)\right]$$ n : nombre d'individus

L'ampleur de l'IC donne une indication sur la « stabilité » de la valeur propres

Le chevauchement de deux IC indiquera l'égalité des valeurs propres



Bootstrap sur les individus

élève	notes par matière				
	math	phys	litt	angl	mus
1	17	14	18	14	12
2	09	13	15	16	18
.					
i	x_{i1}	x_{i2}	x_{ij}		x_{i5}
.					
N	19	15	09	12	06

bootstrap

$\mathbf{X}^{*1} =$

élève	notes par matière				
	math	phys	litt	angl	mus
1	08	11	19	17	15
2	09	13	15	16	18
.					
i	x_{i1}	x_{i2}	x_{ij}		x_{i5}
.					
N	17	14	18	14	12

$$\frac{\lambda_i^{*b}}{\sum_{i=1}^{k} \lambda_i^{*b}} \qquad \frac{\lambda_i^{*1}}{\sum_{i=1}^{k} \lambda_i^{*1}}$$

...

$\mathbf{X}^{*B} =$

élève	notes par matière				
	math	phys	litt	angl	mus
1	09	13	15	16	18
2					
.					
i	x_{i1}	x_{i2}	x_{ij}		x_{i5}
.					
N	08	11	19	17	15

$$\frac{\lambda_i^{*B}}{\sum_{i=1}^{k} \lambda_i^{*B}}$$

→ *Statistique d'intérêt S(x)*

- *% d'inertie expliquée par les axes factoriels*
- *Vecteurs propres*

→ *Effectuer B boostrap par rééchantillonnage*

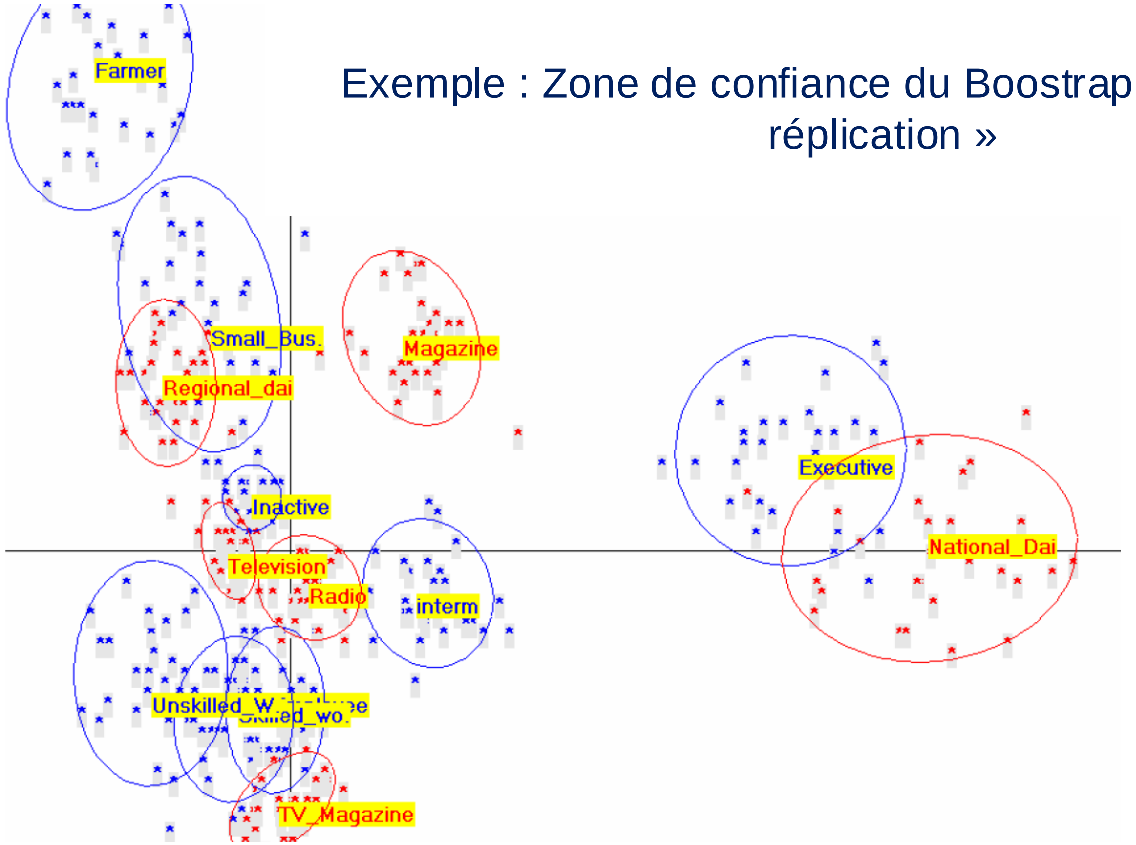
- *Calcul de la statistique d'intérêt sur chaque échantillon*

$$\hat{se}_B(I\%) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\frac{\lambda_i^{*b}}{\sum_{i=1}^{k}\lambda_i^{*b}} - \frac{1}{B}\sum_{b=1}^{B}\frac{\lambda_i^{*b}}{\sum_{i=1}^{k}\lambda_i^{*b}}\right)^2}$$

$$\hat{se}_B(I\%) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\vec{u}_i^{*b} - \frac{1}{B}\sum_{b=1}^{B}\vec{u}_i^{*b}\right)^2}$$

- *Estimation des Intervalles de confiance*

- Bootstrap partiel
  On effectue B rééchantillonges : L'ensemble des individus bootstrappés sont considérés comme des individus supplémentaires. On recalcule alors le positionnement des variables.
- Bootstrap sur les variables (à effectuer que si le nombre de variables est important).
- Jacknife sur les variables permet de voir l'influence d'une variable sur le calcul d'une valeur propre et donc des inerties variables



Exemple : Zone de confiance du Boostrap : « Ellipses de réplication »

Individus et variables supplémentaires

→ L'introduction d'individus et de variables supplémentaires vise à enrichir l'analyse. Il s'agit d'éléments « passif » et illustratifs qui ne participent pas à la construction des axes mais dont la représentation dans les plans d'analyse peut apporter des indications supplémentaires et aider à l'interprétation de l'analyse factorielle

- Soit Z_+ le tableau d'individus supplémentaires (données centrées et réduites). Le calcul des coordonnées des individus sera obtenu simplement par le produit : Z_+u, u étant la matrice des vecteur propres dans R^p (espace des variables)

- Soit Z'_+ le tableau des variables supplémentaires (données centrées et réduites). Le calcul des coordonnées des variables sera obtenu simplement par le produit : Z'_+v ,v étant la matrice des vecteur propres dans R^n (espace des individus)

Variables catégorielles

→ Il est aussi possible de positionner des variables catégorielles (qualitatives) en effectuant la moyennes des coordonnées sur les axes pour chaque catégorie. Cette technique peut être utilisée comme une première approche pour « visualiser » différents groupes. Il existe cependant des techniques dédiées lorsque l'on souhaite discriminer et prédire différentes sous population au sein d'un tableau de données (Analyse factorielle discriminante)

La démarche

- Effectuer dans un premier temps une statistique descriptive univariée (moyennes, écart types, boxplot, histogramme,…..

  Permet de voir la cohérence des données (distributions, points aberrants ou critiques, ...)

- Effectuer une statistique descriptive bivariée (corrélations…etc) qui permet de voir les liaisons entre les variables

- Effectuer l'analyse en composante principale

  - Sélectionner le nombre d'axes à étudier (Boostrap, Kaiser, Anderson)
  - Bien analyser les contributions et les qualités de représentation (étape la plus sensible !)
  - Interpréter les graphiques
  - Positionner les variables et/ ou individus supplémentaires

  Analyser et intreprèter les résultats avec toute la prudence qu'il se doit !!!!!

  S'en tenir uniquement à ce que l'on observe !!!!!!!!!

- L'ACP est très sensible au points abhérents qui peuvent complètement déformer les représentations.
  2 solutions :
  éliminer les individus qui provoquent cette déformation (après validation)
  utiliser d'autres approches (ex : robust PCA)

ACP et diminution de la dimensionalité

La démarche

Tableau des données initiales $\xrightarrow{\text{ACP}}$ Tableau des données axes factoriels (*scores*)

Tableau des données estimées ⟵ Sélection des k < p axes factoriels

Estimation des données initiales

↓

Qualité de la reconstruction ?

n : nombre de lignes (individus)

p : nombre de colonnes (variables)

→ ACP

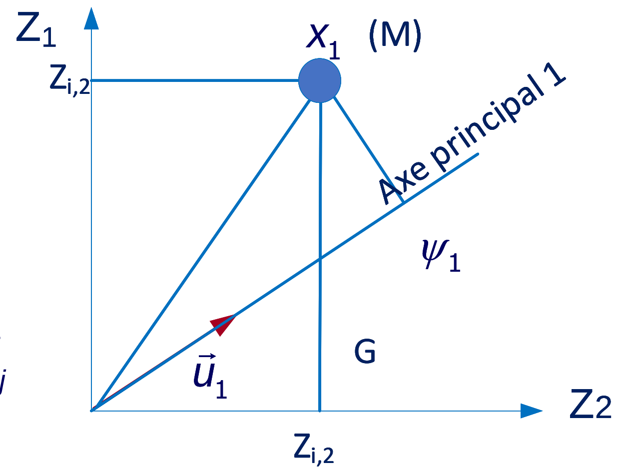
Z : matrice des données centrées

$\vec{u}_1$: vecteur directeur de la première composante ('loading')

Z : matrice des données (n x p)

Z_j : individus : variable j (n x 1)

u_j : vecteur directeur : composante j (p x 1) (*loading*)



Coordonnées des individus sur la première composante (n x 1) $\Psi_1 = Zu_1$

Coordonnées des individus sur la j ième composante $\Psi_j = Zu_j$

Si l'on connait Ψ il est possible d'estimer Z

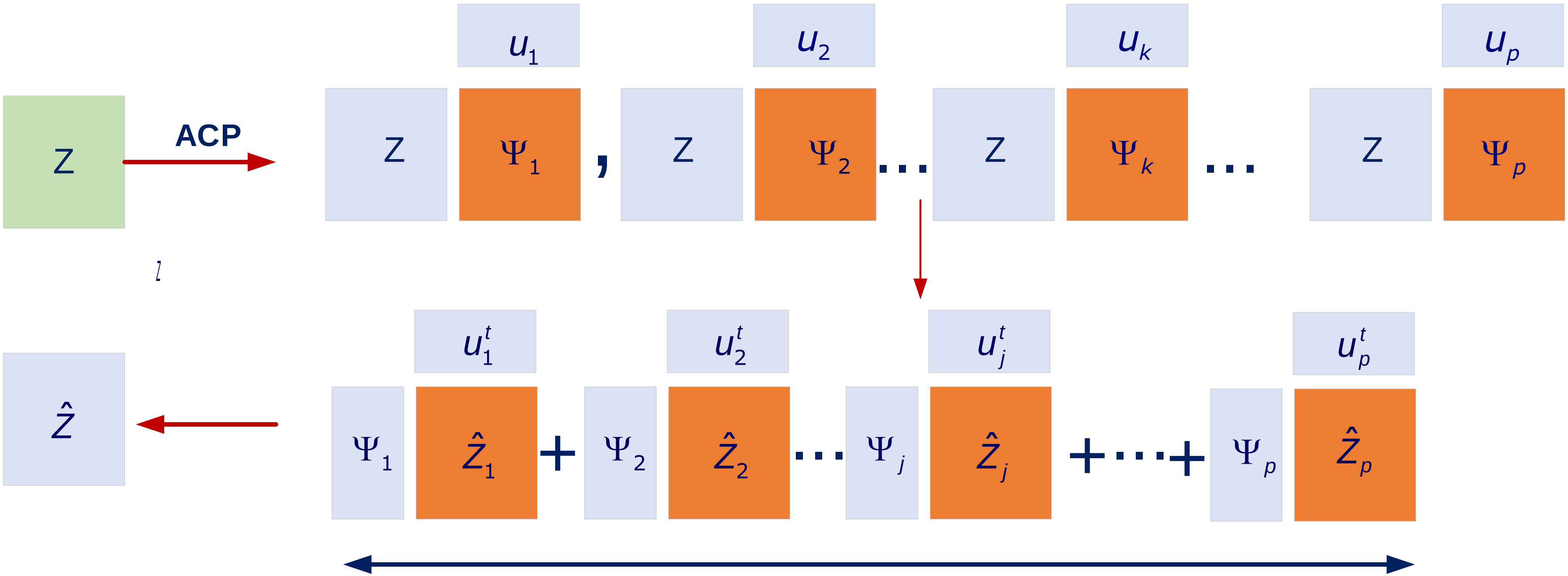
Estimation de Z à l'aide de la première composante $\hat{Z}_1 = \Psi_1 u_1^t$

Estimation de Z à l'aide de la j ième composante $\hat{Z}_j = \Psi_j u_j^t$

⟶ n x p

p variables

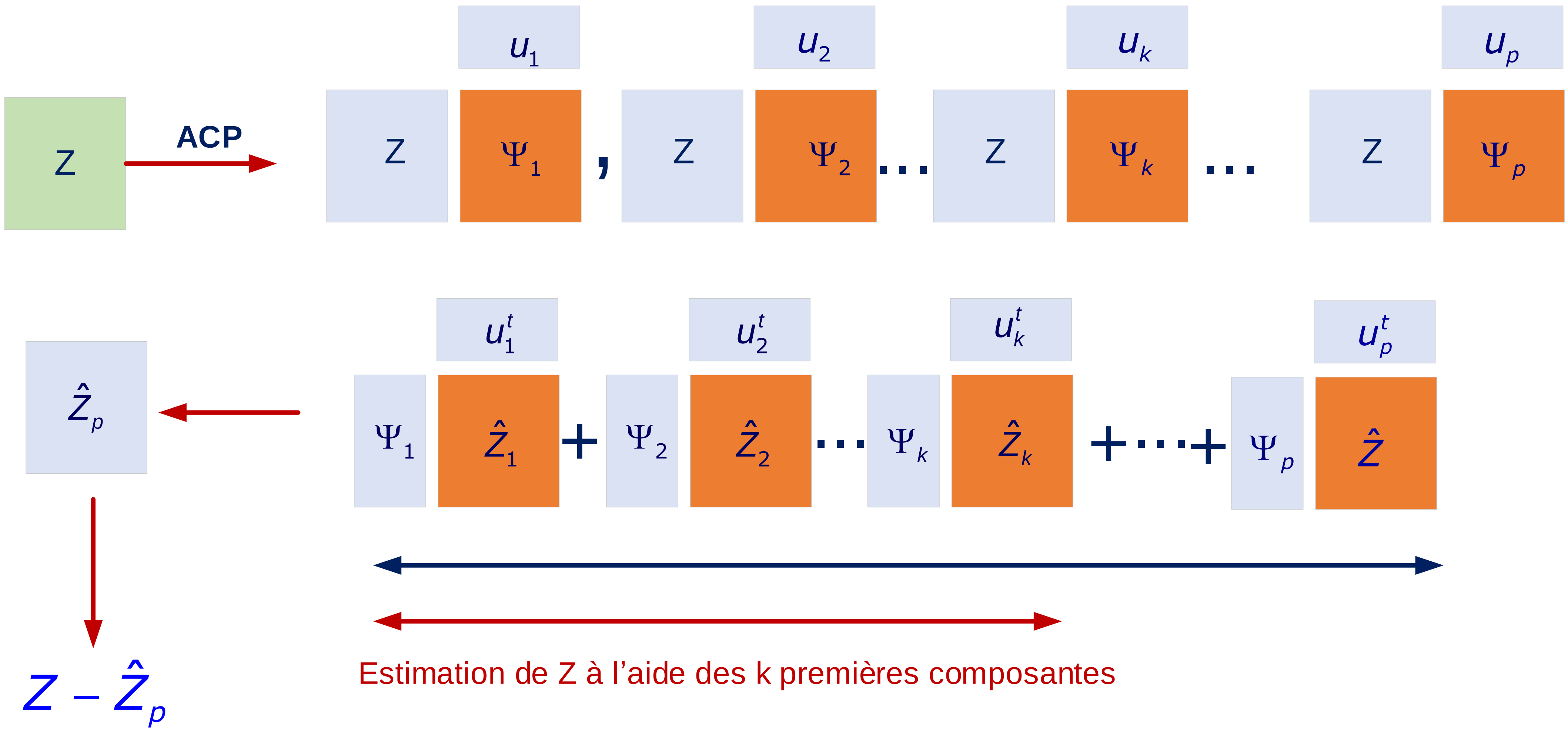
Coordonnées individus sur les axes factoriels



Estimation de Z à l'aide de toutes les composantes = p

$$Z - \hat{Z} \approx 0$$

Coordonnées individus sur les axes factoriels



Etude histopathologie : Etude morphologique de cellules : detection de cancer du sein : n = 589 patients

Analyse de 30 cellules

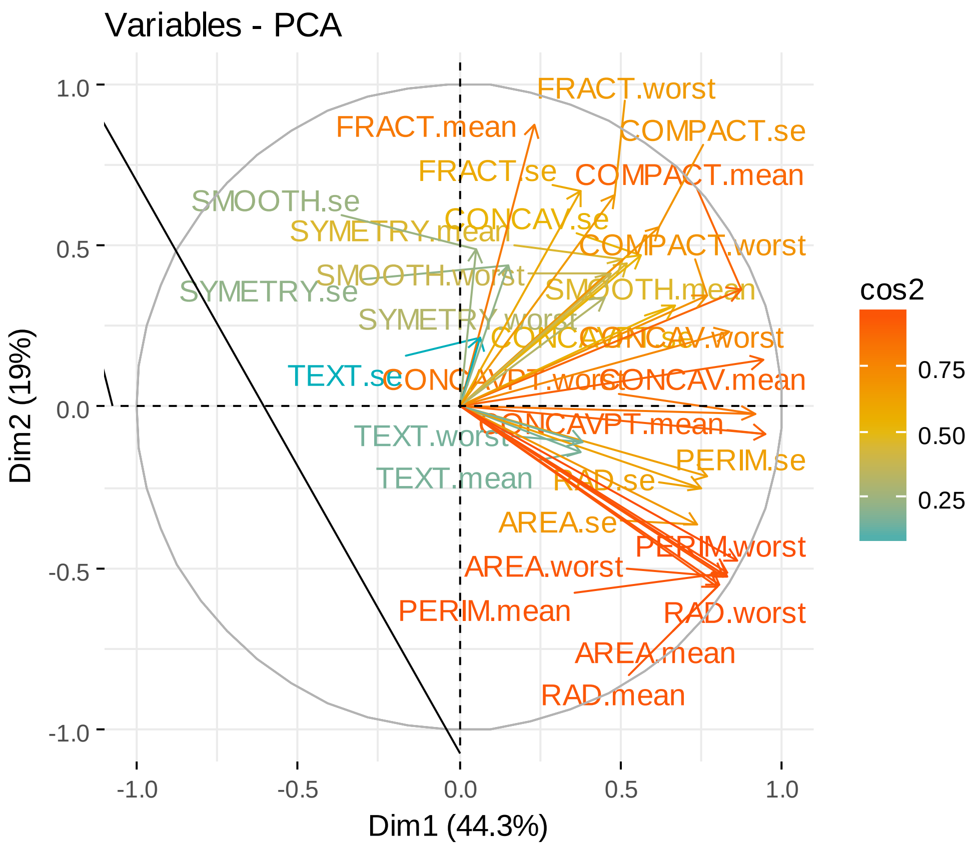
10 critères

- Angle
- Texture
- Périmétre
- Surface
- Facteur de lissage
- Compacité
- Concavité
- Concavité VPT
- Symétrie
- Dimension fractale

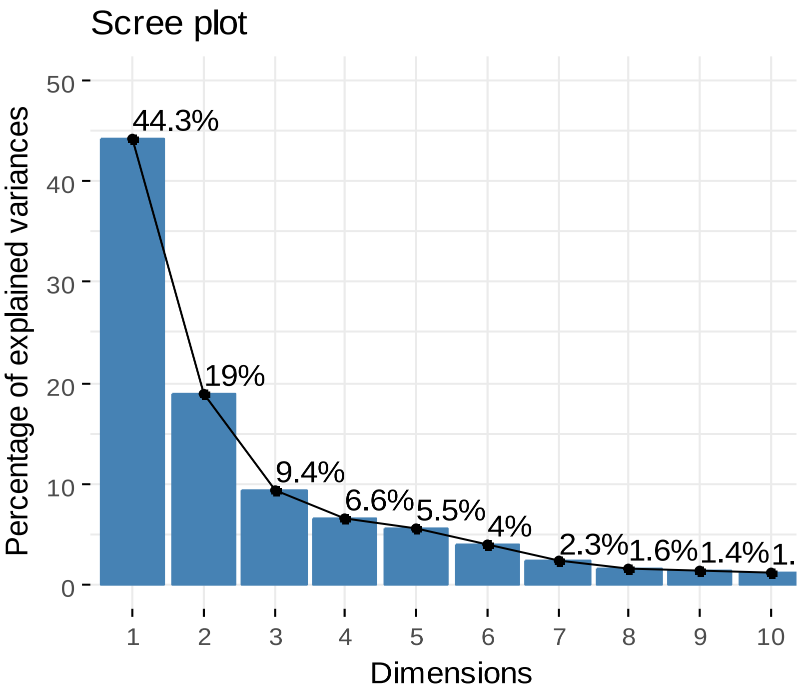
3 items

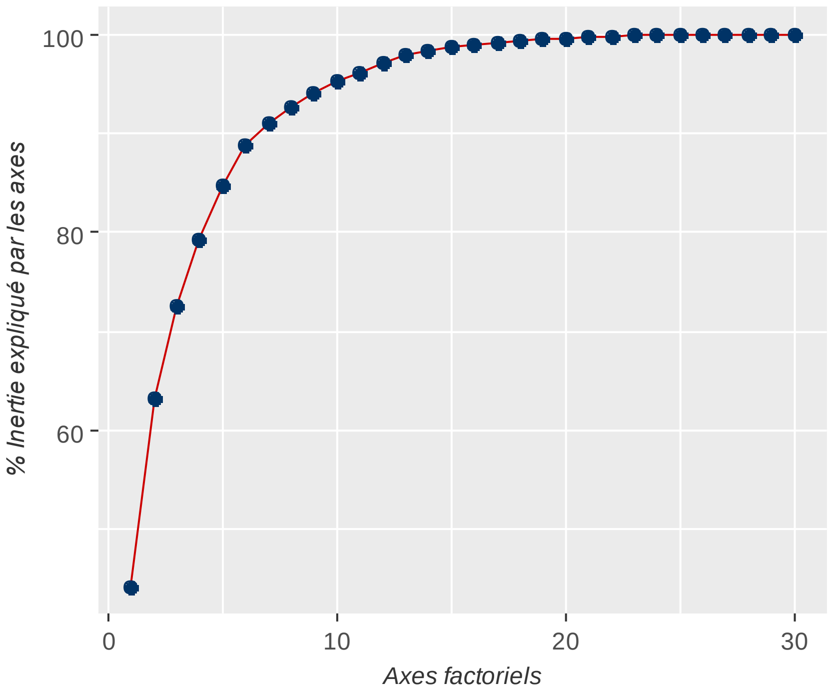
- Moyenne
- Déviation standard
- La plus mauvaise mesure

30 variables
17670 données



	Vp	% var	% var cum
comp1	13,282	44,272	44,272
comp 2	5,691	18,971	63,243
comp 3	2,818	9,393	72,636
comp 4	1,981	6,602	79,239
comp 5	1,649	5,496	84,734
comp 6	1,207	4,025	88,759
comp 7	0,675	2,251	91,010
comp 8	0,477	1,589	92,598
comp 9	0,417	1,390	93,988
comp 10	0,351	1,169	95,157
comp 11	0,294	0,980	96,137
comp 12	0,261	0,871	97,007
comp 13	0,241	0,805	97,812
comp 14	0,157	0,523	98,335
comp 15	0,094	0,314	98,649
comp 16	0,080	0,266	98,915
comp 17	0,059	0,198	99,113
comp 18	0,053	0,175	99,288
comp 19	0,049	0,165	99,453
comp 20	0,031	0,104	99,557
comp 21	0,030	0,100	99,657
comp 22	0,027	0,091	99,749
comp 23	0,024	0,081	99,830
comp 24	0,018	0,060	99,890
comp 25	0,015	0,052	99,942
comp 26	0,008	0,027	99,969
comp 27	0,007	0,023	99,992
comp 28	0,002	0,005	99,997
comp 29	0,001	0,002	100,000
comp 30	0,000	0,000	100,000



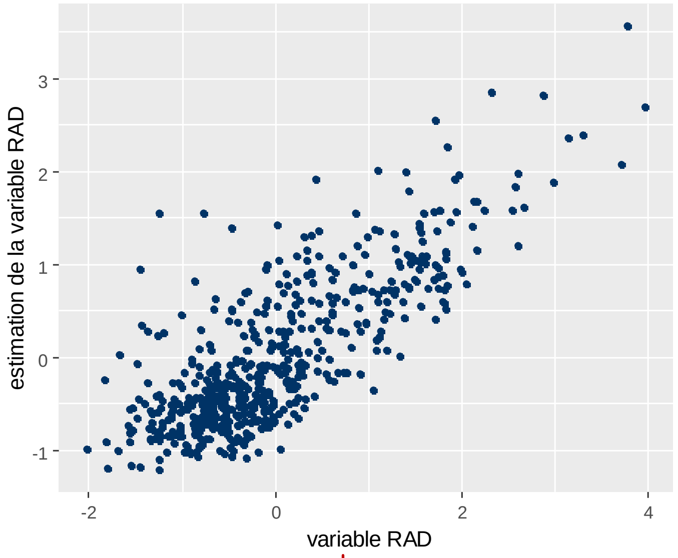


les 10 premiers axes principaux expliquent 95 % de la variance

On peut donc « reconstruire » le tableau initial avec 10 axes factoriels en consentant « une perte globale de 5% »

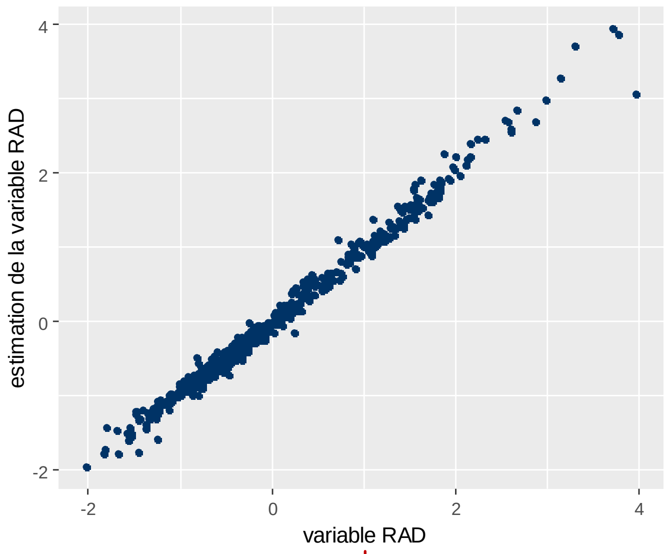
Il s’agit d’une compression statistique

Reconstruction



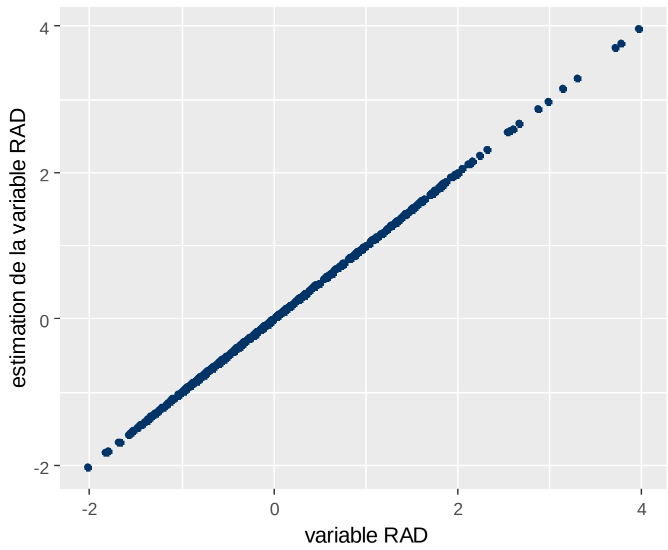
1 axe factoriel

42.27 %



10 axes factoriels

95.15 %



30 axes factoriels

100 %

1 octet = 8 bits = 256 niveaux de gris

Chaque image = 321 x 261 = 83781 pixels

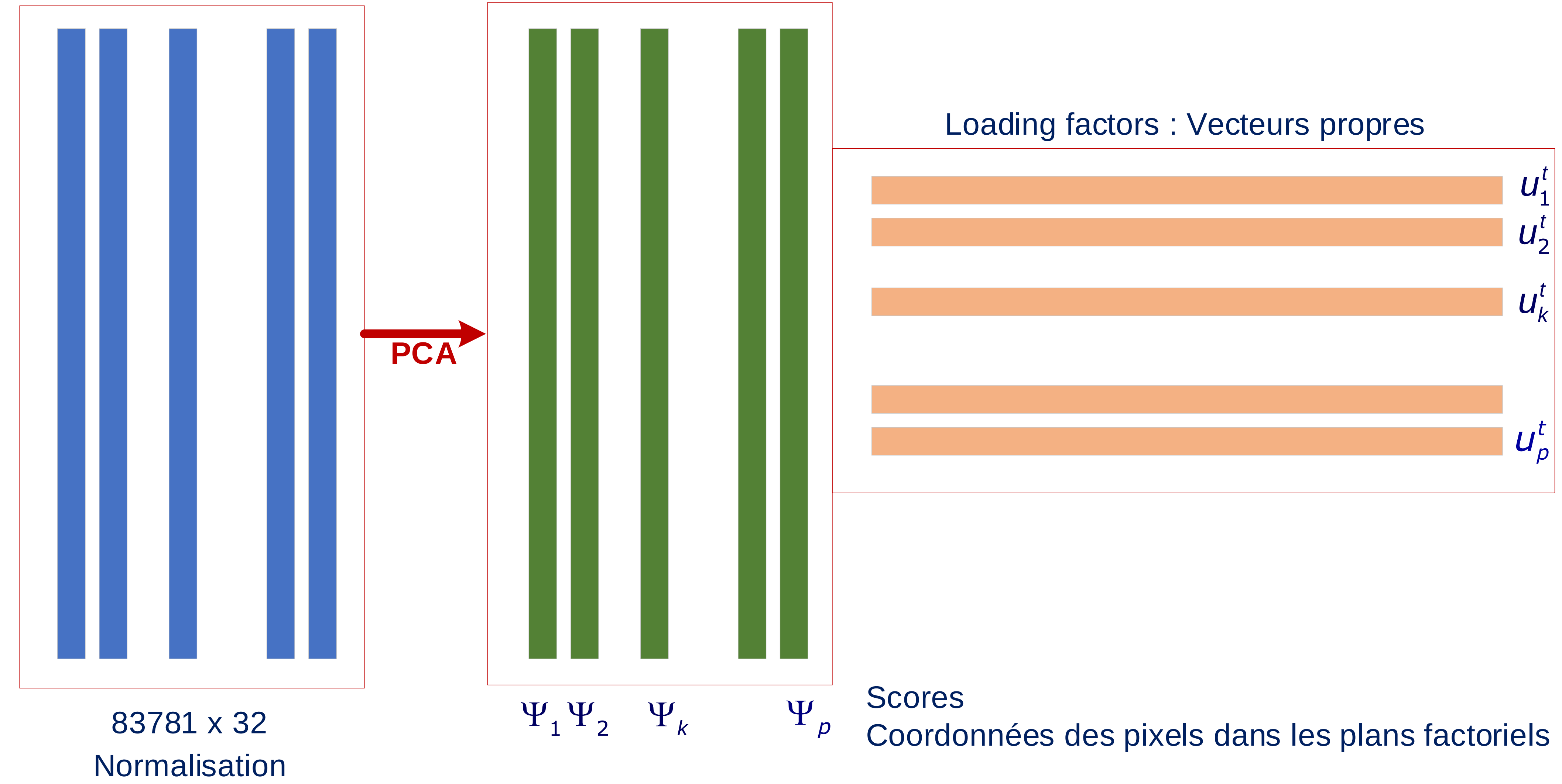
Stockage / image = (83781) * 1 / 1024 = 81,81 Ko

Stockage d'une série d'images

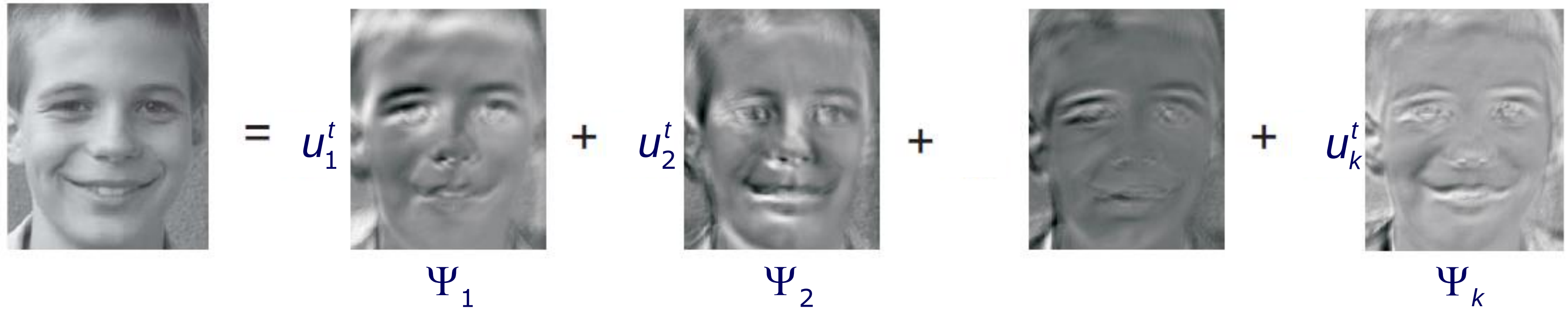
(32 * 81.81) /1024 = 2.55 Mo

- Prétraitement : Recadrage + Egalisation
- Transformation PCA

Transformation 1D de chaque image



Approximation d'une image avec 4 axes factoriels



Approximation des images avec 4 axes factoriels

- Image originale

- Image reconstituée avec 4 axes factoriels

Facteur de compression (Fc) = 4/32 =0.125

Taux de compression = 1 – Fc = 0.875

- A titre d'exemple :

Une image radiologique nécessite (ERLM)

- Écran (2k) 2048 pixel/ligne x 1080 pixel /colonnes
- Un codage sur 16 bits (4096 niveaux de gris)
- Soit 33,75 Mo

ANALYSE EN COMPOSANTE PRINCIPALE

ACP et estimation des données manquantes: le NIPALS

Une brève introduction aux rotations Varimax et Quartimax

Régression en composante principale

ACP

n : nombre de lignes (individus)

p : nombre de colonnes (variables)

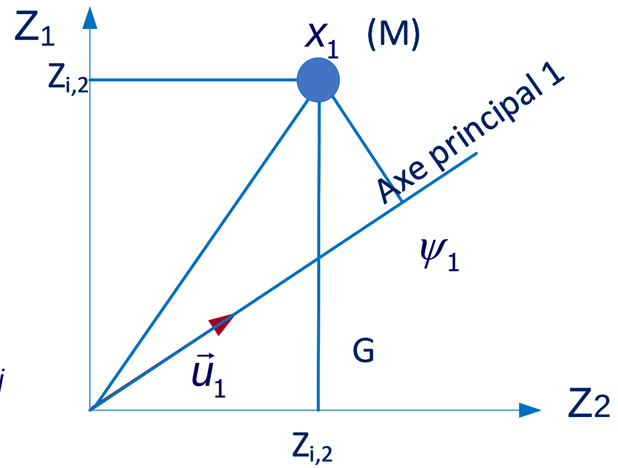
Z : matrice des données centrées

$\vec{u}_1$: vecteur directeur de la première composante ('loading')

Z : matrice des données (n x p)

Z_j : individus : variable j (n x 1)

u_j : vecteur directeur : composante j (p x 1) (*loading*)



Coordonnées des individus sur la première composante (n x 1) $\Psi_1 = Zu_1$

Coordonnées des individus sur la j ième composante $\Psi_j = Zu_j$

Si l'on connait Ψ il est possible d'estimer Z

Estimation de Z à l'aide de la première composante $\hat{Z}_1 = \Psi_1 u_1^t$

Estimation de Z à l'aide de la j ième composante $\hat{Z}_j = \Psi_j u_j^t$

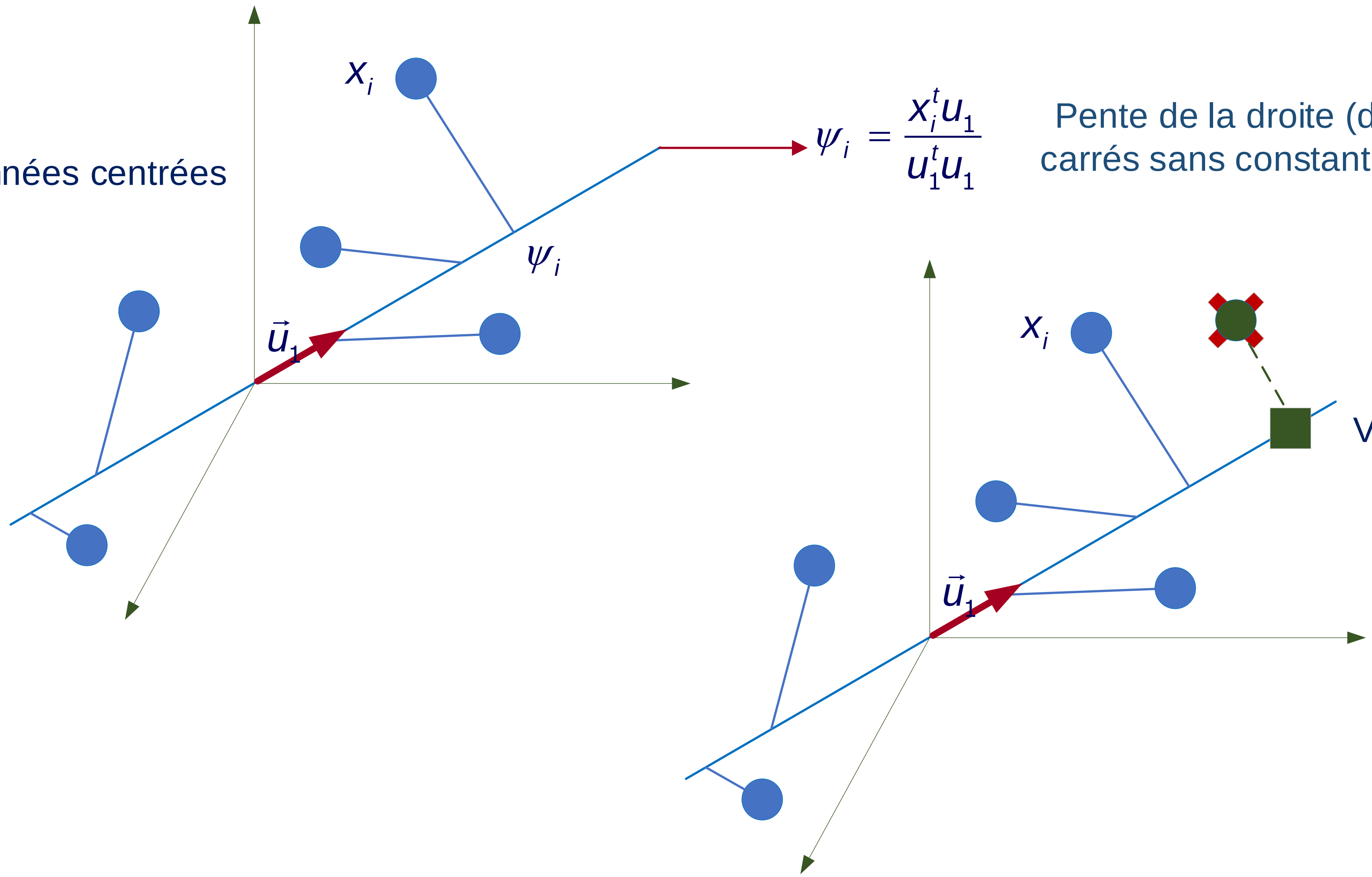
→ n x p

- NIPALS (**N**on linear **I**térative **PA**rtial **L**east **S**quare)

Si données centrées

$$\psi_i = \frac{x_i^t u_1}{u_1^t u_1}$$

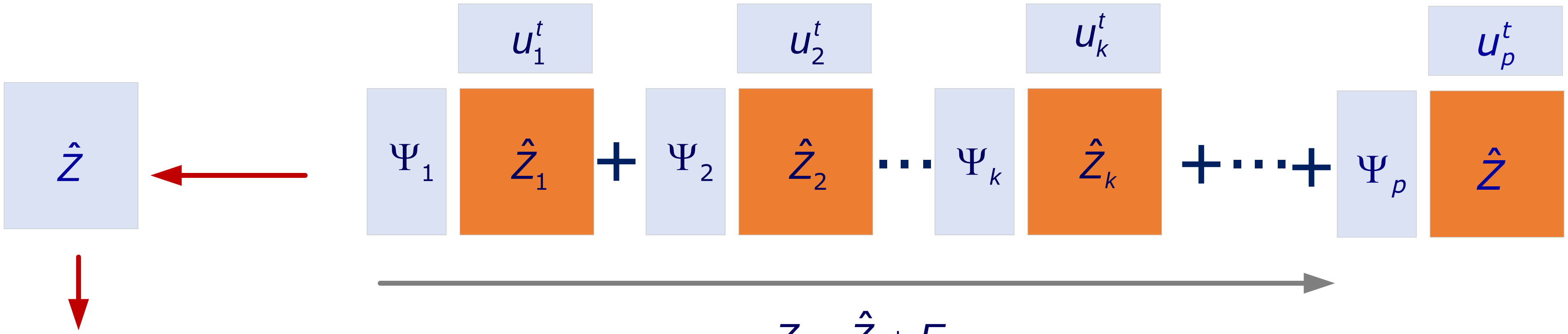
Pente de la droite (des moindres carrés sans constante de x_i sur u_1)



Valeur manquante

Si données manquantes ψ_i est calculé sur les données disponibles

Z Tableau initiale des données (p variables)



$$Z = \hat{Z} + E$$

Approximation de Z

$$J = \left\| Z_{nxp} - \underbrace{\Psi_{nxk} u_{kxn}^t}_{\hat{Z}} \right\| \longrightarrow J = \sum_{i=1}^{n} \sum_{j=1}^{p} \left(x_{ij} - \sum_{k=1}^{k} \Psi_{ik} u_{jk}^t \right)^2$$

L'estimation s'effectue en minimisant la fonction objective

$$\min\left(J\left(\Psi, u\right)\right)$$

- Approche calculatoire Méthode incrémentale

étape 1 → $\hat{Z}_1 = \Psi_1 u_1^t + E_1$

$\hat{Z}_2 = \Psi_1 u_1^t + \Psi_2 u_2^t + E_2$

$\hat{Z}_j = \Psi_1 u_1^t + \Psi_2 u_2^t + \ldots + \Psi_j u_j^t + E_j$

étape k → $\hat{Z}_k = \Psi_1 u_1^t + \Psi_2 u_2^t + \ldots + \Psi_j u_j^t + \ldots + \Psi_k u_k^t + E_j$

$$J_k < \ldots < J_i < \ldots < J_2 < J_1$$

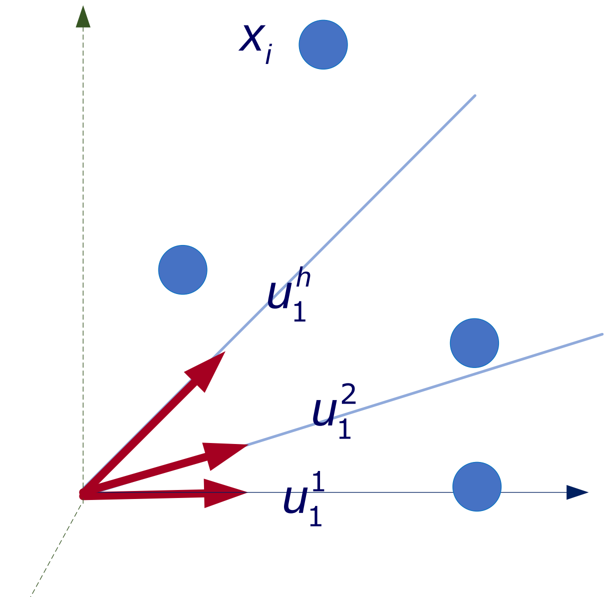
- A chaque étape on calcule successivement Ψ_j et u_j^t par itérations successives jusqu'à convergence

Exemple : étape 1

$$J_1 = \sum_{i=1}^{n}\sum_{j=1}^{p}\left(x_{ij} - \Psi_{i1}u_{j1}^t\right)^2 \longrightarrow \begin{cases} \dfrac{\partial^2 J_1}{\partial \Psi_{i1}} = 0 \Rightarrow u_{j1} = \dfrac{\sum_i \left(x_{ij} \times \Psi_{i1}\right)}{\sum_i \Psi_{i1}^2} \\ \dfrac{\partial^2 J_1}{\partial u_{j1}} = 0 \Rightarrow \Psi_{i1} = \dfrac{\sum_i \left(x_{ij} \times u_{j1}\right)}{\sum_i u_{j1}} \end{cases}$$

Estimation axe1 (vecteurs propres et coordonnées des points

n itérations jusqu'à convergence



Les vecteurs propres sont tous perpendiculaires entre eux

n = 1 : axe 1 est « confondue » avec la variable 1

Estimation de Ψ_1^1 et u_1^1 avec les données diponibles

Calcul de J_1^1

n = 2 : à partir de u_1^1 et Ψ_1^1

Estimation de Ψ_1^2 et [illegible] avec les données diponibles

Calcul de J_1^2

n = h : à partir de u_1^2 et Ψ_1^2

Estimation de u_1^h et Ψ_1^h avec les données diponibles

Calcul de J_1^h

On normalise à chaque itération les vecteurs directeurs

itérer jusqu'à convergence $J_1^{h-1} - J_1^h < seuil$

On réalise le même processus à chaque étape (étape = prise en compte des autres variables 1...k < p)

Calcul des coordonnées des k < p variables

Calcul des coordonnés des individus sur l'axe

Algorithme NIPAL pour données complètes

1 $\hat{X}_0 = \hat{X}$

2 $for\ k = 1,2,....,K$

a $t_h = colonne\ 1\ de\ X_{k\text{-}1}$

b $réitérer\ jusqu'à\ convergence\ de\ p_k$

i $p_k = \hat{X}_{k-1}^{'} t_k / t_k^{'} t_k \quad (RL\ 1)$

ii $Normalisation\ de\ p_h$

iii $t_k = \hat{X}_{k-1} p_k / p_k^{'} p_k \quad (RL\ 2)$

3 $\hat{X}_k = \hat{X}_{k-1} - t_k p_k$

Relation entre les approches (relations cycliques)

$$\frac{1}{n-1} X^{'} X p_1 = \lambda_1 p_1$$

$$\frac{1}{n-1} X X^{'} t_1 = \lambda_1 t_1$$

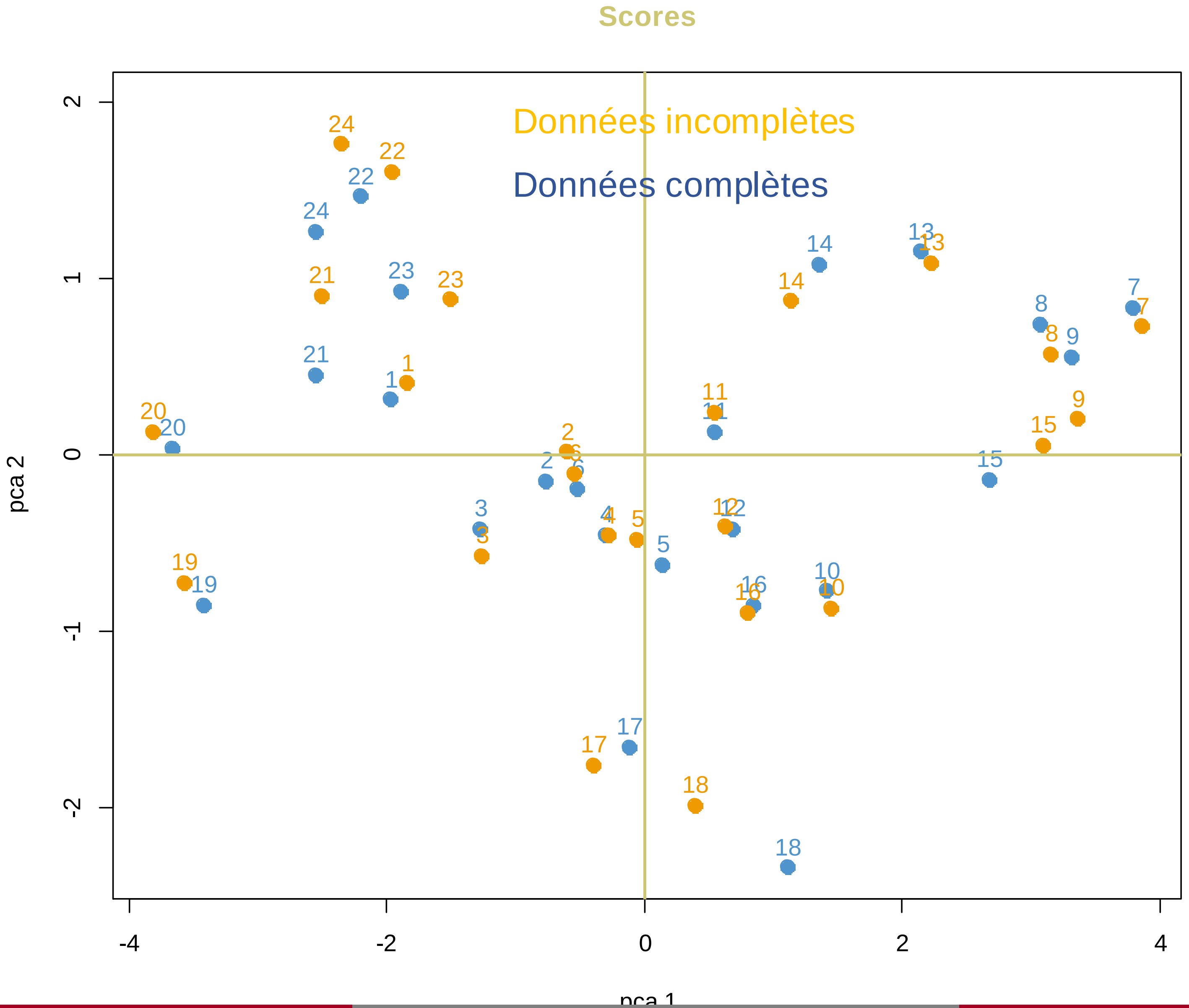
Données complètes

	Cylindree	Puissance	Vitese	Poids	Longueur	Largeur
Honda civic	1396	90	174	850	369	166
Renault 19	1721	92	180	965	415	169
Fiat Tipo	1580	83	170	970	395	170
Peugeot 405	1769	90	180	1080	440	169
Renault 21	2068	88	180	1135	446	170
Citroen BX	1769	90	182	1060	424	168
BMW 530i	2986	188	226	1510	472	175
Rover 827i	2675	177	222	1365	469	175
Renault 25	2548	182	226	1350	471	180
Opel Omega	1998	122	190	1255	473	177
Peugeot 405 Break	1905	125	194	1120	439	171
Ford Sierra	1993	115	185	1190	451	172
BMW 325iX	2494	171	208	1600	432	164
Audi 90 Quattro	1994	160	214	1220	439	169
Ford Scorpio	2933	150	200	1345	466	176
Renault Espace	1995	120	177	1265	436	177
Nissan Vanette	1952	87	144	1430	436	169
VW Caravelle	2109	112	149	1320	457	184
Ford Fiesta	1117	50	135	810	371	162
Fiat Uno	1116	58	145	780	364	155
Peugeot 205	1580	80	159	880	370	156
Peugeot 205 Rallye	1294	103	189	805	370	157
Seat Ibiza SX I	1461	100	181	925	363	161
Citroen AX Sport	1294	95	184	730	350	160

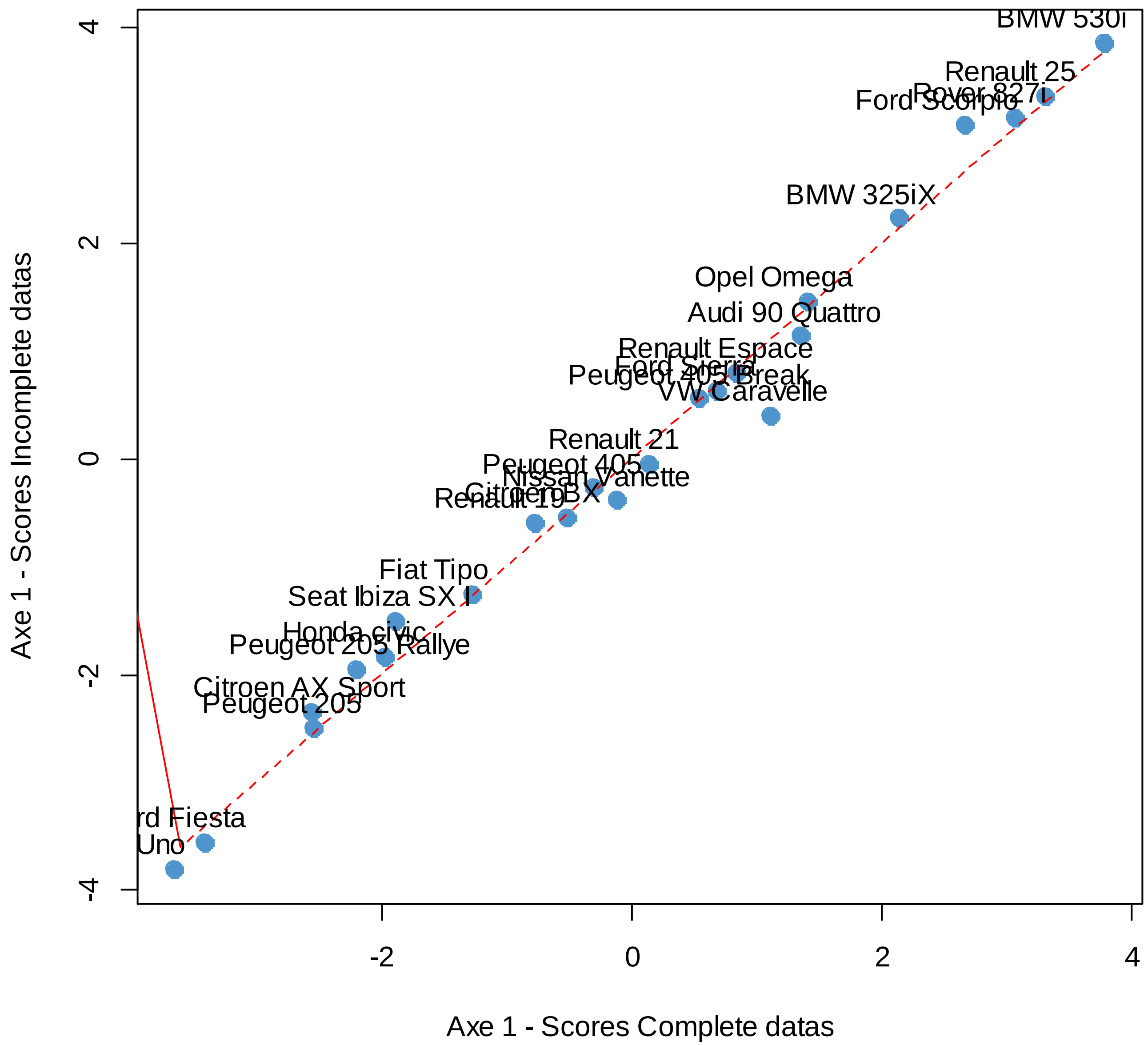
Données incomplètes

	Cylindree	Puissance	Vitese	Poids	Longueur	Largeur
Honda civic	NA	90	174	850	369	166
Renault 19	1721	NA	180	965	415	169
Fiat Tipo	1580	83	NA	970	395	170
Peugeot 405	1769	90	180	NA	440	169
Renault 21	2068	88	180	1135	NA	170
Citroen BX	1769	90	182	1060	424	NA
BMW 530i	NA	188	226	1510	472	175
Rover 827i	2675	NA	222	1365	469	175
Renault 25	2548	182	NA	1350	471	180
Opel Omega	1998	122	190	NA	473	177
Peugeot 405 Break	1905	125	194	1120	NA	171
Ford Sierra	1993	115	185	1190	451	NA
BMW 325iX	NA	171	208	1600	432	164
Audi 90 Quattro	1994	NA	214	1220	439	169
Ford Scorpio	2933	150	NA	1345	466	176
Renault Espace	1995	120	177	NA	436	177
Nissan Vanette	1952	87	144	1430	NA	169
VW Caravelle	2109	112	149	1320	457	NA
Ford Fiesta	NA	50	135	810	371	162
Fiat Uno	1116	NA	145	780	364	155
Peugeot 205	1580	80	NA	880	370	156
Peugeot 205 Rallye	1294	103	189	NA	370	157
Seat Ibiza SX I	1461	100	181	925	NA	161
Citroen AX Sport	1294	95	184	730	350	NA

Positionnement des données complètes et incomplètes dans le plan (axe 1 axe 2)



Variation de la position des données sur le premier axe factoriel



Données complétes

	Cylindree	Puissance	Vitese	Poids	Longueur	Largeur
Honda civic	1396	90	174	850	369	166
Renault 19	1721	92	180	965	415	169
Fiat Tipo	1580	83	170	970	395	170
Peugeot 405	1769	90	180	1080	440	169
Renault 21	2068	88	180	1135	446	170
Citroen BX	1769	90	182	1060	424	168
BMW 530i	2986	188	226	1510	472	175
Rover 827i	2675	177	222	1365	469	175
Renault 25	2548	182	226	1350	471	180
Opel Omega	1998	122	190	1255	473	177
Peugeot 405 Break	1905	125	194	1120	439	171
Ford Sierra	1993	115	185	1190	451	172
BMW 325iX	2494	171	208	1600	432	164
Audi 90 Quattro	1994	160	214	1220	439	169
Ford Scorpio	2933	150	200	1345	466	176
Renault Espace	1995	120	177	1265	436	177
Nissan Vanette	1952	87	144	1430	436	169
VW Caravelle	2109	112	149	1320	457	184
Ford Fiesta	1117	50	135	810	371	162
Fiat Uno	1116	58	145	780	364	155
Peugeot 205	1580	80	159	880	370	156
Peugeot 205 Rallye	1294	103	189	805	370	157
Seat Ibiza SX I	1461	100	181	925	363	161
Citroen AX Sport	1294	95	184	730	350	160

Estimation effectuée à partir des données incomplètes

	Cylindree	Puissance	Vitese	Poids	Longueur	Largeur
Honda civic	1302.566	90.00000	174.0000	850.0000	369.0000	166.0000
Renault 19	1721.000	98.74165	180.0000	965.0000	415.0000	169.0000
Fiat Tipo	1580.000	83.00000	165.7565	970.0000	395.0000	170.0000
Peugeot 405	1769.000	90.00000	180.0000	1069.4370	440.0000	169.0000
Renault 21	2068.000	88.00000	180.0000	1135.0000	437.1509	170.0000
Citroen BX	1769.000	90.00000	182.0000	1060.0000	424.0000	166.2133
BMW 530i	2752.654	188.00000	226.0000	1510.0000	472.0000	175.0000
Rover 827i	2675.000	173.10778	222.0000	1365.0000	469.0000	175.0000
Renault 25	2548.000	182.00000	215.0906	1350.0000	471.0000	180.0000
Opel Omega	1998.000	122.00000	190.0000	1190.6156	473.0000	177.0000
Peugeot 405 Break	1905.000	125.00000	194.0000	1120.0000	422.1847	171.0000
Ford Sierra	1993.000	115.00000	185.0000	1190.0000	451.0000	169.7205
BMW 325iX	2643.484	171.00000	208.0000	1600.0000	432.0000	164.0000
Audi 90 Quattro	1994.000	117.97577	214.0000	1220.0000	439.0000	169.0000
Ford Scorpio	2933.000	150.00000	238.6293	1345.0000	466.0000	176.0000
Renault Espace	1995.000	120.00000	177.0000	1186.7816	436.0000	177.0000
Nissan Vanette	1952.000	87.00000	144.0000	1430.0000	428.3082	169.0000
VW Caravelle	2109.000	112.00000	149.0000	1320.0000	457.0000	171.7323
Ford Fiesta	1115.230	50.00000	135.0000	810.0000	371.0000	162.0000
Fiat Uno	1116.000	53.70638	145.0000	780.0000	364.0000	155.0000
Peugeot 205	1580.000	80.00000	164.2647	880.0000	370.0000	156.0000
Peugeot 205 Rallye	1294.000	103.00000	189.0000	787.4955	370.0000	157.0000
Seat Ibiza SX I	1461.000	100.00000	181.0000	925.0000	382.2857	161.0000
Citroen AX Sport	1294.000	95.00000	184.0000	730.0000	350.0000	159.3122

- « Optimisation » de la projection des variables sur les axes : **augmenter la corrélation des variables par rapport aux axes tout en préservant l'orthogonalité des axes**

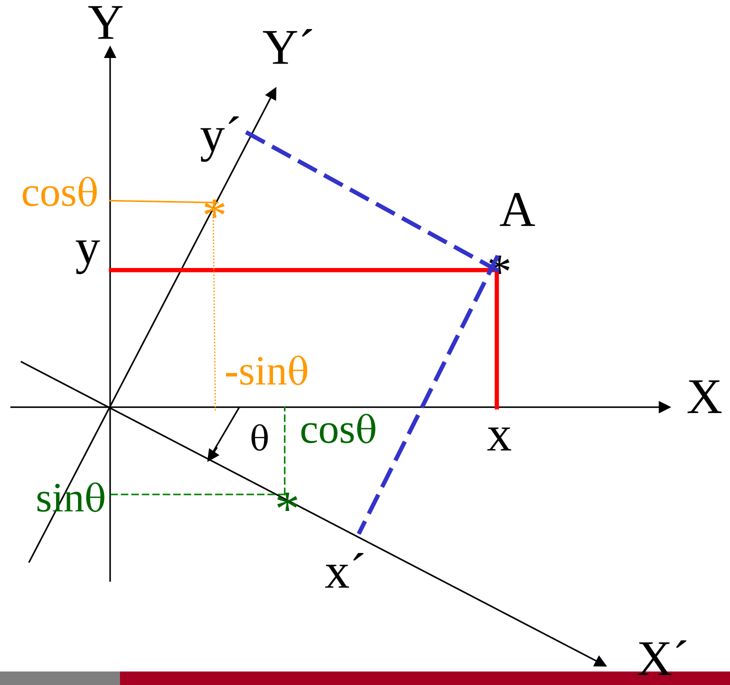
On cherche une décomposition de la matrice des corrélations de la forme : $Z'Z = R = \Lambda\Lambda' + \Psi$

$$\begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \\ \lambda_{31} & \lambda_{32} \end{pmatrix} \begin{pmatrix} \lambda_{11} & \lambda_{21} & \lambda_{31} \\ \lambda_{12} & \lambda_{22} & \lambda_{32} \end{pmatrix} + \begin{pmatrix} \psi & 0 & 0 \\ 0 & \psi & 0 \\ 0 & 0 & \psi \end{pmatrix}$$

$$TT' = I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}$$

On « insére » une matrice orthogonale qui est une matrice de rotation



$$Z'Z = R = \Lambda\Lambda' + \Psi$$

$$= \begin{bmatrix} 1 & Cor(X_1,X_2) & Cor(X_1,X_3) \\ \ldots & 1 & Cor(X_2,X_3) \\ \ldots & \ldots & Var(X_3) \end{bmatrix} = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \\ \lambda_{31} & \lambda_{32} \end{bmatrix} TT' \begin{bmatrix} \lambda_{11} & \lambda_{21} & \lambda_{31} \\ \lambda_{12} & \lambda_{22} & \lambda_{32} \end{bmatrix} + \begin{bmatrix} \psi_1 & 0 & 0 \\ 0 & \psi_2 & 0 \\ 0 & 0 & \psi_3 \end{bmatrix}$$

Nouvelle matrice après rotation

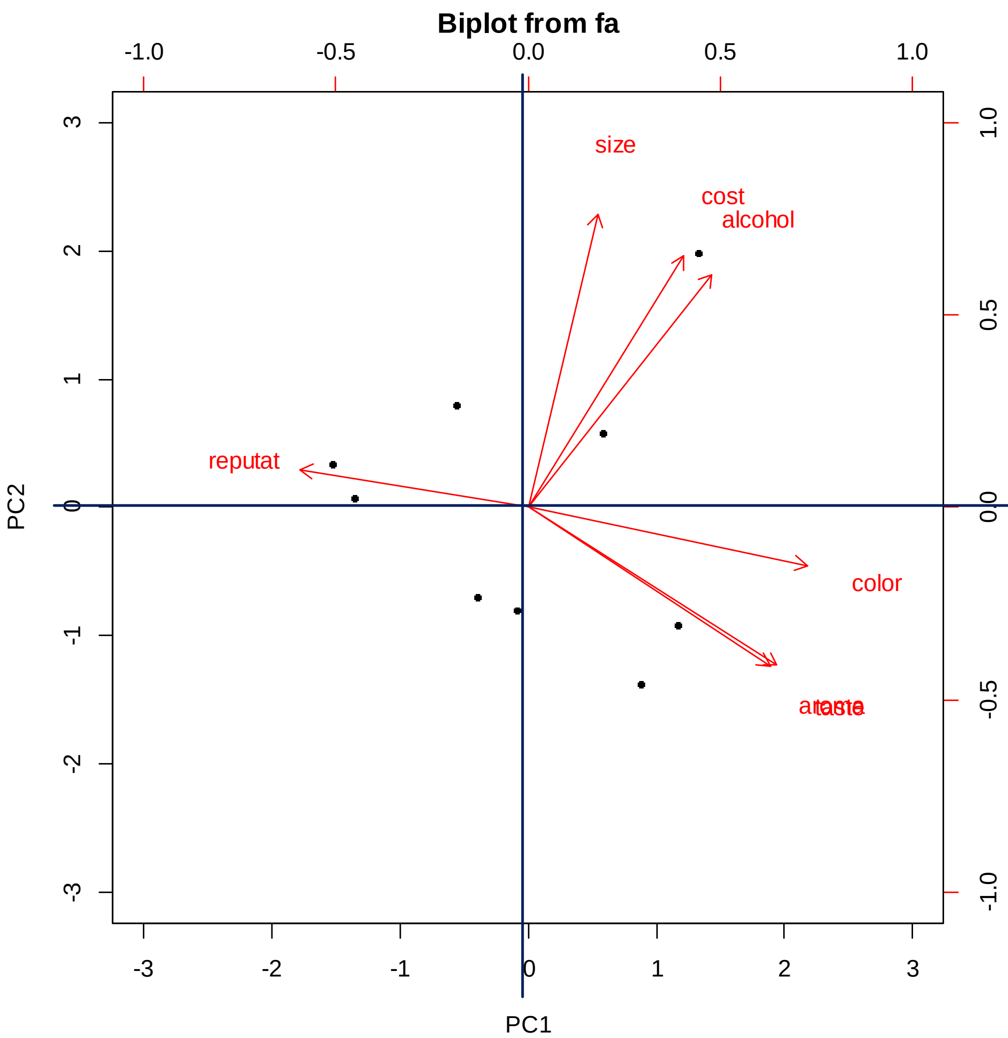
$$\Gamma = \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \\ \gamma_{31} & \gamma_{32} \end{bmatrix} = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \\ \lambda_{31} & \lambda_{32} \end{bmatrix} \underbrace{\begin{bmatrix} \cos(\theta) & \sin(\theta \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}}_{T}$$

Resultats

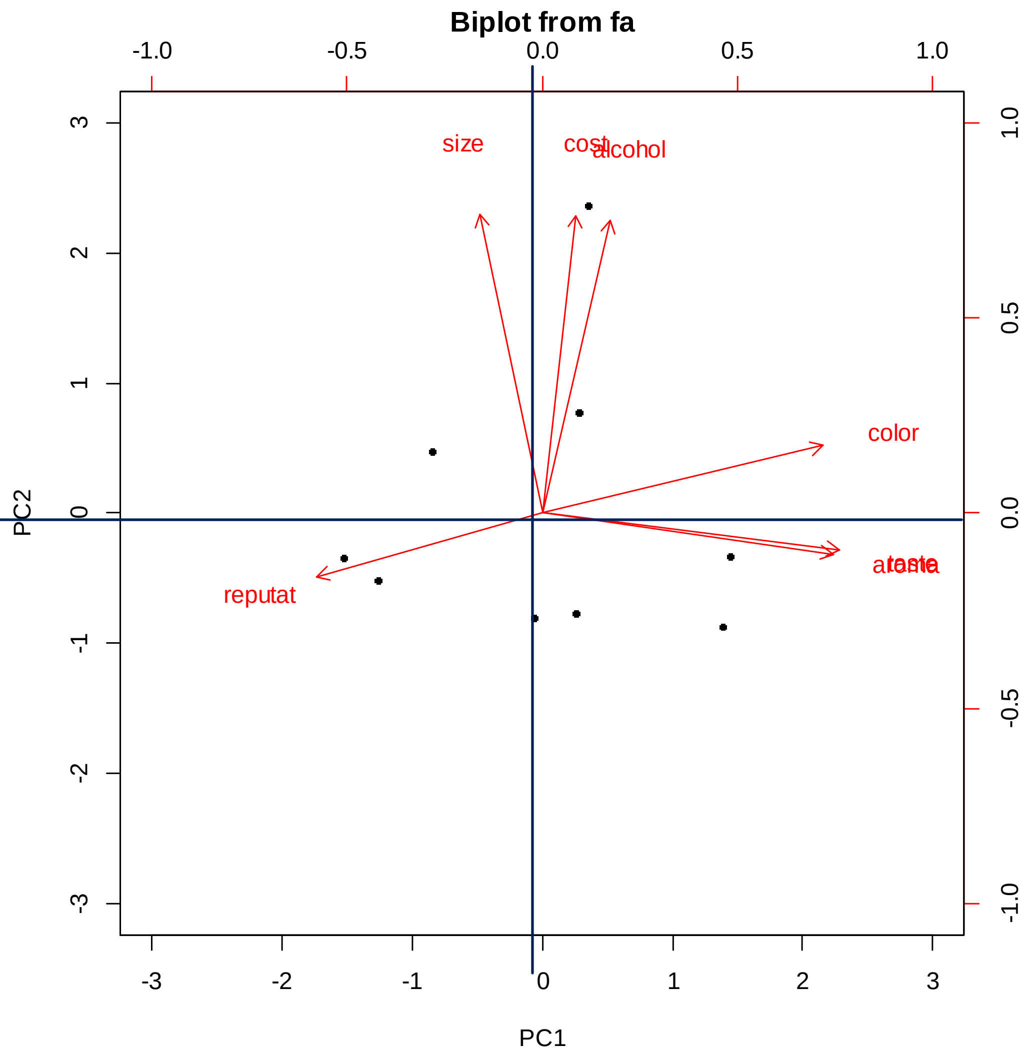
Pour chaque colonne de Γ, les $|\gamma_{ij}|$ sont proches de 0 ou 1 : Varimax

Pour chaque ligne de Γ il y a un $|\gamma_{ij}|$ proche 1 et tous les autres proches de 0 : Quartimax

ACP standard



VARIMAX



- On utilise la régression en composante principale lorsque :
  - les variables explicatives sont très fortement corrélées. Dans ce cadre, les variables risquent d'être colinéairee et engendrer des résultats « instables »
  - Le nombre de variables explicatives est supérieur au nombre d'observations
- Conséquences
  - Coefficients de régressions estimés peuvent être très élevés
  - Le signe des coefficients peuvent être contraire à l'intuition
  - Coefficients de régressions instables

La régression en composante principale consiste à effectuer la régression sur les composantes principales des variables explicatives

- Procédure de calcul

1. Centrer le données
   - Soit Z tableau des variables explicatives (centrée)
   - Soit Y la variable à expliquer et Yc la même variable centrèe
2. Calcul des valeurs propres et vecteur propres de la matrice des corrélations de $Z^t Z$
   - Soit U matrice des vecteurs propres
   - Soit Λ matrice des valeurs propres (matrice diagonale)
3. Elimination des composantes principales de faible inertie (dépend de la nature du tableau)
4. Calcul des coordonnées des individus sur les composantes retenues (= scores)
   - Score = U * Z
5. effectuer la régression $\mathrm{B} = \left[score^t \times score\right]^{-1} \times score^t \times Y_c$

   $(X^t X)^{-1} X^t Y = B$ Il s'agit donc d'une simple régression
6. calculer les valeurs estimées de Y
7. comparaison des résultats avec la régression linéaire multiple (modèle plein)