

Statistique multidimensionnelle Analyse en composantes principales (ACP)

2017-2018

michel.voue@umons.ac.be
Michel Voué



Faculté
des Sciences

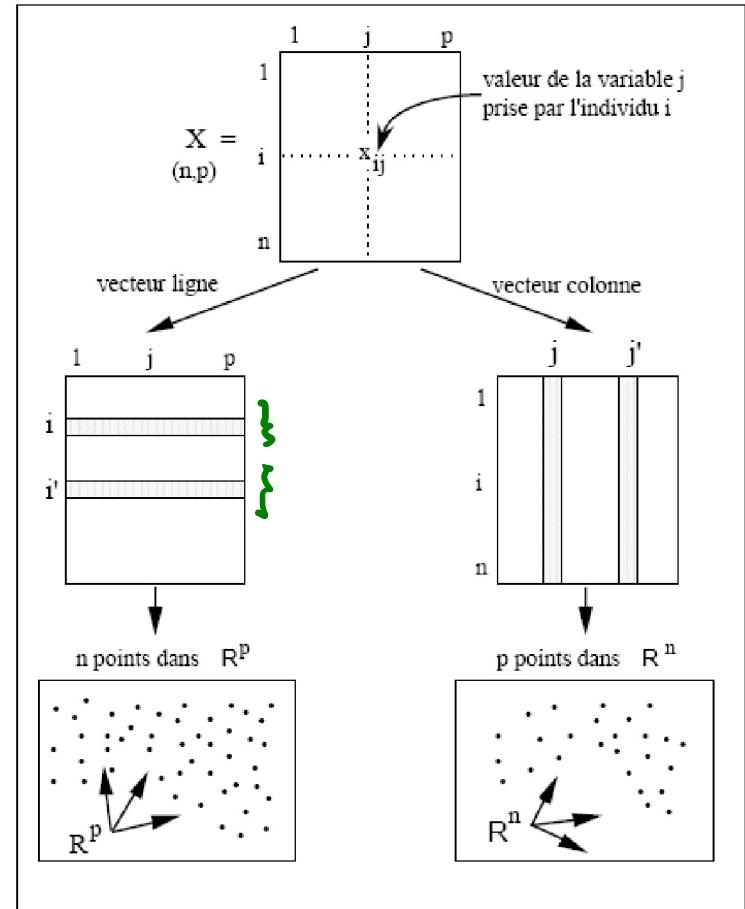
ACP : Historique et domaine d 'application

- Historique : Pearson (1901)
- Domaine d 'application :
 - Tableau rectangulaire R de mesures
 - Colonnes : variables numériques continues (p)
 - Lignes : individus (n)

- Interprétations géométriques :
 - Proximités des individus entre eux

$$d^2(i, i') = \sum_{j=1}^p (r_{ij} - r_{i'j})^2$$

- Corrélation en les variables



Domaines d'application

TABLE 1.1

Some Examples of Datasets

Field	Individuals	Variables	x_{ik}
Ecology	Rivers	Concentration of pollutants	Concentration of pollutant k in river i
Economics	Years	Economic indicators	Indicator value k for year i
Genetics	Patients	Genes	Expression of gene k for patient i
Marketing	Brands	Measures of satisfaction	Value of measure k for brand i
Pedology	Soils	Granulometric composition	Content of component k in soil i
Biology	Animals	Measurements	Measure k for animal i
Sociology	Social classes	Time by activity	Time spent on activity k by individuals from social class i

(Husson et al, 2011)

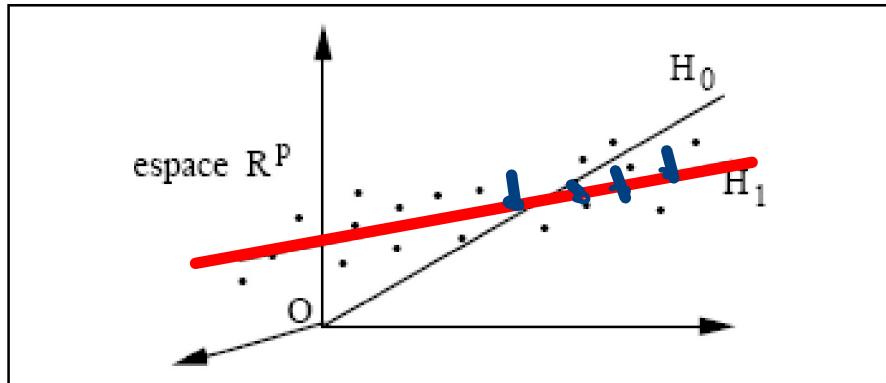
Exemple de données

TABLE 1.2
The Orange Juice Data

	Odour intensity	Odour typicality	Pulp	Intensity of taste	Acidity	Bitterness	Sweetness
Pampryl amb.	2.82	2.53	1.66	3.46	3.15	2.97	2.60
Tropicana amb.	2.76	2.82	1.91	3.23	2.55	2.08	3.32
Fruvita fr.	2.83	2.88	4.00	3.45	2.42	1.76	3.38
Joker amb.	2.76	2.59	1.66	3.37	3.05	2.56	2.80
Tropicana fr.	3.20	3.02	3.69	3.12	2.33	1.97	3.34
Pampryl fr.	3.07	2.73	3.34	3.54	3.31	2.63	2.90

- 16 variables continues (homogénéité des dimensions ?)
 - 6 individus
- (*Husson et al, 2011*)

Analyse du nuage des individus

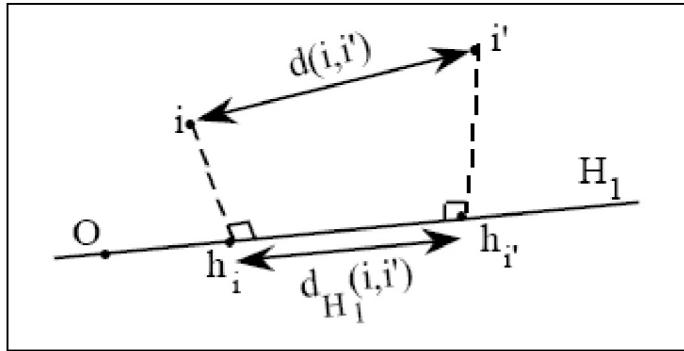


$$\underset{(H)}{\text{Max}} \left\{ \sum_{i=1}^n \sum_{i'=1}^n d_H^2(i, i') \right\}$$

dist. ent
H.

- Analyse générale : Maximiser la somme des carrés des distances à l'origine (H_0)
- ACP : Maximiser la somme des carrés des distances entre *tous les couples d'individus*
- H_1 ne passe plus par l'origine

Principe d'ajustement



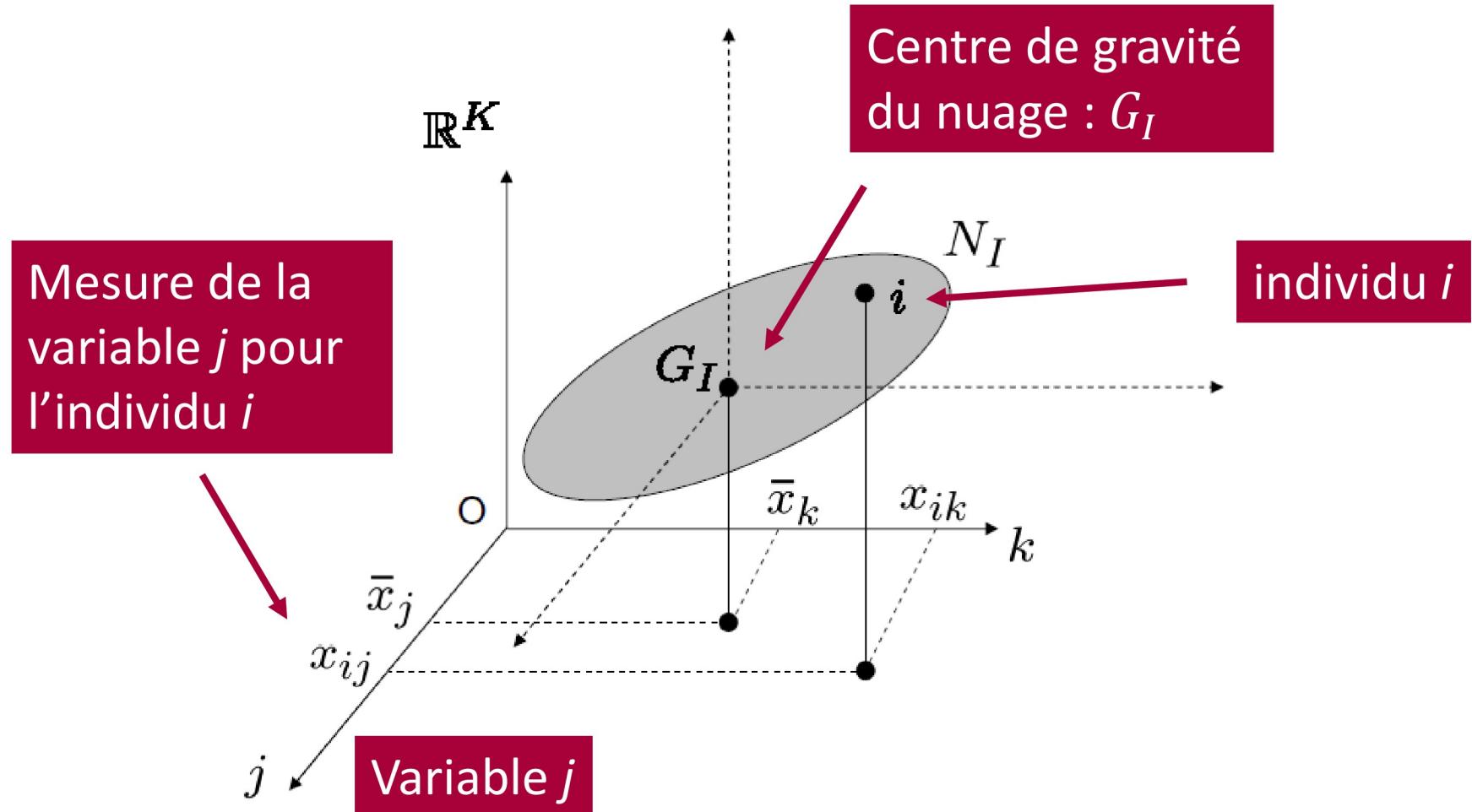
$$\sum_{i,i'=1}^n d^2(i,i') = 2n \sum_{i=1}^n (h_i - \bar{h})^2$$

$$\bar{h} = \frac{1}{n} \sum_{i=1}^n h_i$$

$$\underset{(H)}{\text{Max}} \left\{ \sum_{i=1}^n \sum_{i'=1}^n d_H^2(i,i') \right\} \Leftrightarrow \underset{(H)}{\text{Max}} \left\{ \sum_{i=1}^n d_H^2(i,G) \right\}$$

Origine en G : analyse générale de X défini par $x_{ij} = r_{ij} - \bar{r}_j$

Centrage des données

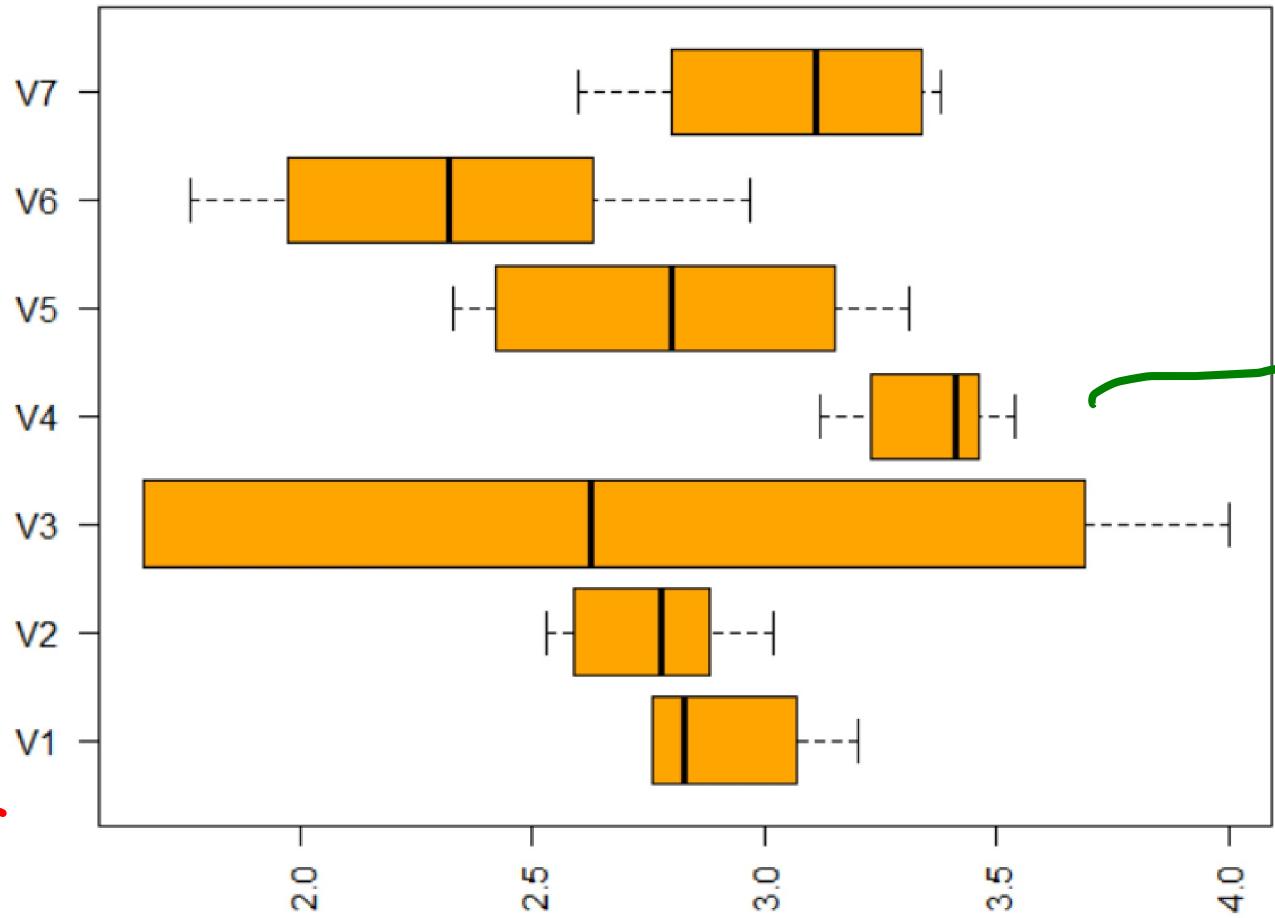
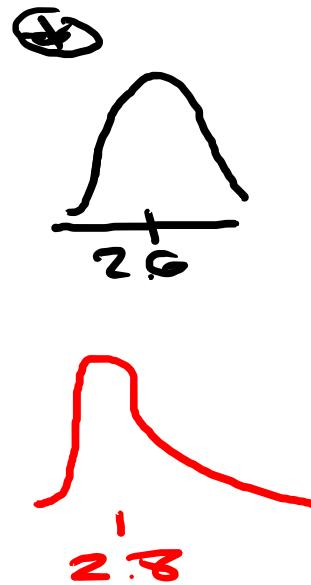


Statistique univariée

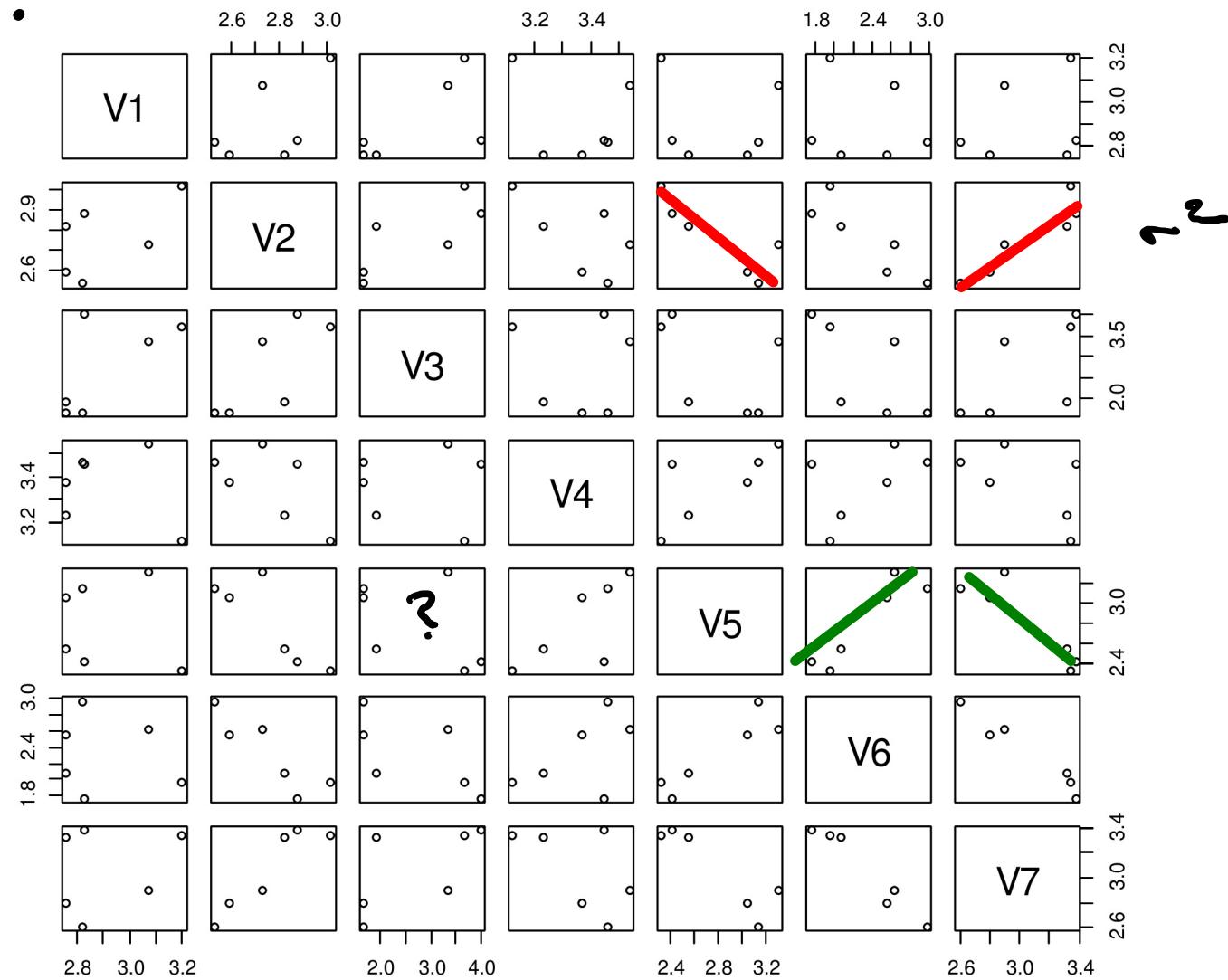
V1	V2	V3	V4
Min. :2.760	Min. :2.530	Min. :1.660	Min. :3.120
1st Qu.:2.775	1st Qu.:2.625	1st Qu.:1.722	1st Qu.:3.265
Median :2.825	Median :2.775	Median :2.625	Median :3.410
Mean :2.907	Mean :2.762	Mean :2.710	Mean :3.362
3rd Qu.:3.010	3rd Qu.:2.865	3rd Qu.:3.603	3rd Qu.:3.458
Max. :3.200	Max. :3.020	Max. :4.000	Max. :3.540
V5	V6	V7	
Min. :2.330	Min. :1.760	Min. :2.600	
1st Qu.:2.453	1st Qu.:1.998	1st Qu.:2.825	
Median :2.800	Median :2.320	Median :3.110	
Mean :2.802	Mean :2.328	Mean :3.057	
3rd Qu.:3.125	3rd Qu.:2.612	3rd Qu.:3.335	
Max. :3.310	Max. :2.970	Max. :3.380	

```
[1] "Odour.intensity"      "Odour.typicality"      "Pulpiness"           "Intensity.of.taste"  
[5] "Acidity"              "Bitterness"             "Sweetness"
```

Boxplots



Lien entre les variables



Distance entre les individus

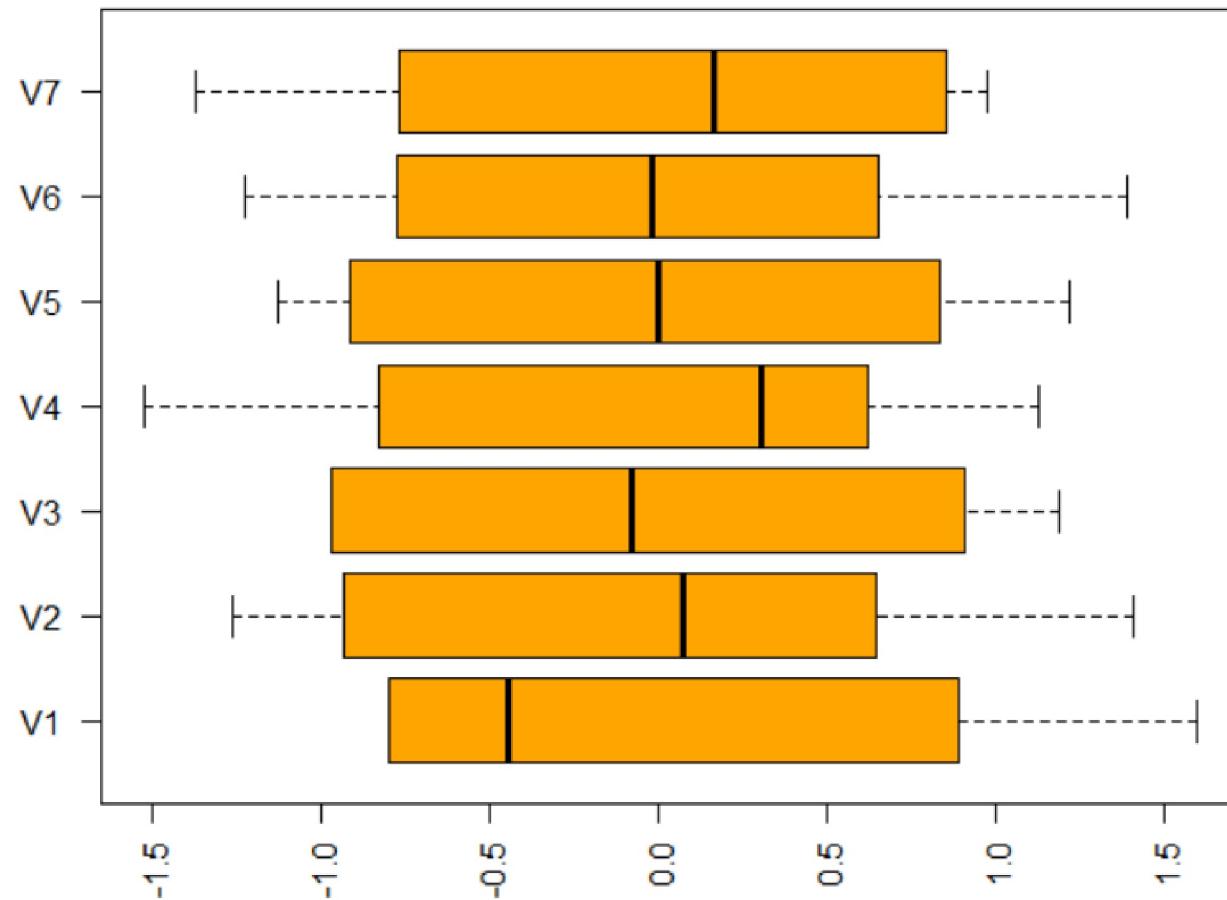
Analyse en composantes principales normée : distance entre 2 points indépendante des unités des variables

$$\left. \begin{aligned} d^2(i, i') &= \sum_{j=1}^p \left(\frac{r_{ij} - r_{i'j}}{s_j \sqrt{n}} \right)^2 \\ s_j^2 &= \frac{1}{n} \sum_{i=1}^n (r_{ij} - \bar{r}_j)^2 \end{aligned} \right\} \Rightarrow x_{ij} = \frac{r_{ij} - r_{i'j}}{s_j \sqrt{n}}$$

Les variables sont *centrées réduites*.

L'écart à la moyenne est mesuré nombre d 'écart-types de la variable j .

Données centrées réduites



Matrice à diagonaliser

Nuage des points-individus dans \Re^p (espace des variables) : translation de l'origine au *centre de gravité* du nuage avec changement éventuel de l'échelle des axes.

Diagonalisation de $\mathbf{C} = \mathbf{X}'\mathbf{X}$ définie par :

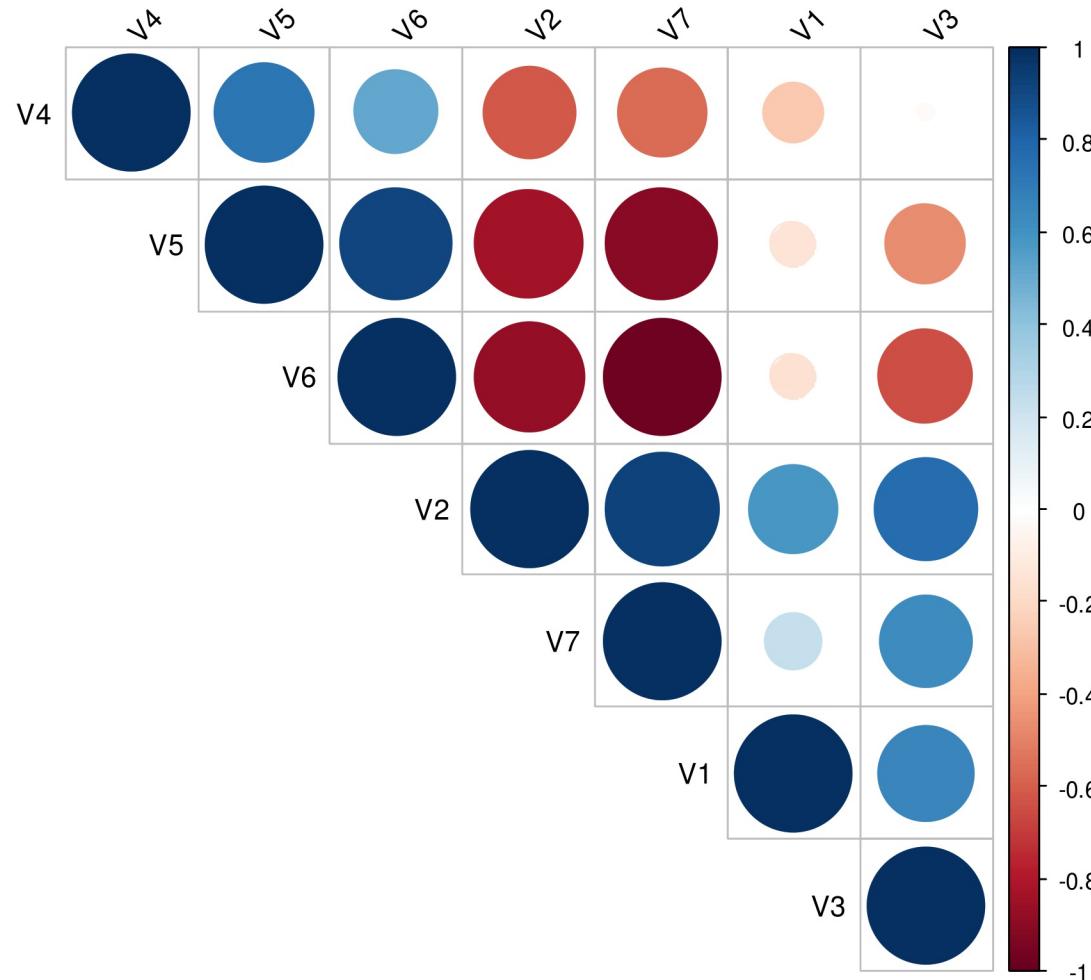
$$c_{jj'} = \sum_{i=1}^n x_{ij} x_{ij'} = \frac{1}{n} \sum_{i=1}^n \frac{(r_{ij} - \bar{r}_j)(r_{ij'} - \bar{r}_{j'})}{s_j s_{j'}} = cor(j, j')$$

C'est la matrice de corrélations

Matrice des corrélations

	V1	V2	V3	V4	V5	V6	V7
V1	1.00	0.58	0.66	-0.27	-0.15	-0.15	0.23
V2	0.58	1.00	0.77	-0.62	-0.84	-0.88	0.92
V3	0.66	0.77	1.00	-0.02	-0.47	-0.64	0.63
V4	-0.27	-0.62	-0.02	1.00	0.73	0.51	-0.57
V5	-0.15	-0.84	-0.47	0.73	1.00	0.91	-0.90
V6	-0.15	-0.88	-0.64	0.51	0.91	1.00	-0.98
V7	0.23	0.92	0.63	-0.57	-0.90	-0.98	1.00

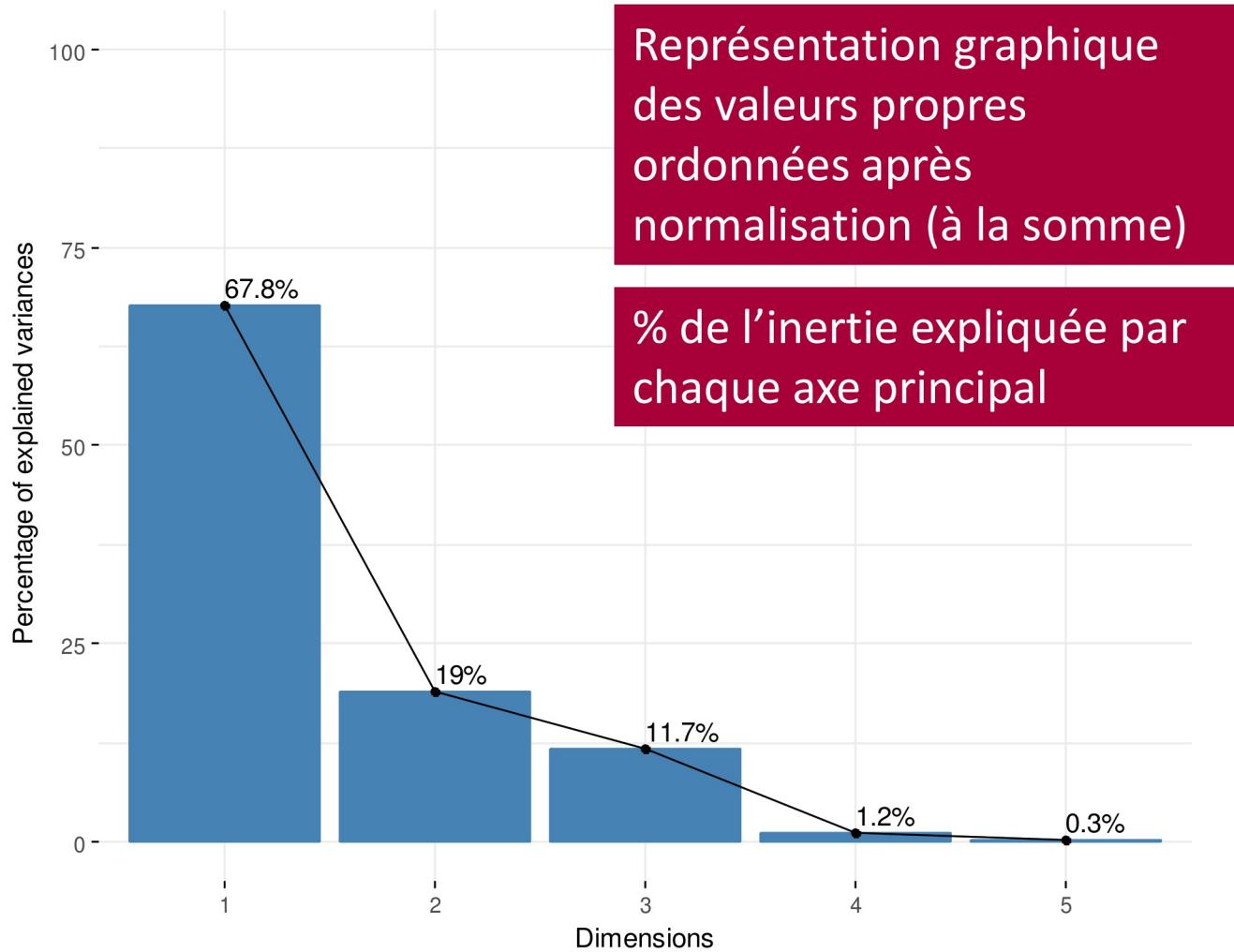
Visualisation de la matrice de corrélation



Valeurs propres de la matrice de corrélation

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	4.74369269	67.7670384	67.76704
Dim.2	1.33328986	19.0469979	86.81404
Dim.3	0.81984115	11.7120164	98.52605
Dim.4	0.08402330	1.2003328	99.72639
Dim.5	0.01915301	0.2736144	100.00000

Scree plot



Choix du nombre d'axes

- Fonction de la taille du problème ($p = 5$ ou $p = 50$)
- A partir du graphe des valeurs propres ordonnées (% de variance expliquée)

Critères :

- Critère de KAISER (contraignant) : A partir de $Inertie = \sum \lambda_i = p$, on sélectionne uniquement les valeurs propres > 1
- Coude dans le graphe en éboulis (pas assez contraignant)

Axes factoriels

Coordonnées des n points-individus sur l'axe factoriel \mathbf{u}_α :

$$\psi_\alpha = Xu_\alpha$$

Nuage centré au centre de gravité :

- Moyenne du facteur nulle :

$$\sum_{i=1}^n \psi_{\alpha i} = 0$$

- Variance :

$$\text{var}(\psi_\alpha) = \lambda_\alpha$$

Coordonnée de l'individu i sur l'axe α :

$$\psi_{\alpha i} = \sum_{j=1}^p u_{\alpha j} x_{ij} = \sum_{j=1}^p u_{\alpha j} \frac{r_{ij} - \bar{r}_j}{s_j \sqrt{n}}$$

Analyse du nuage des points-variables

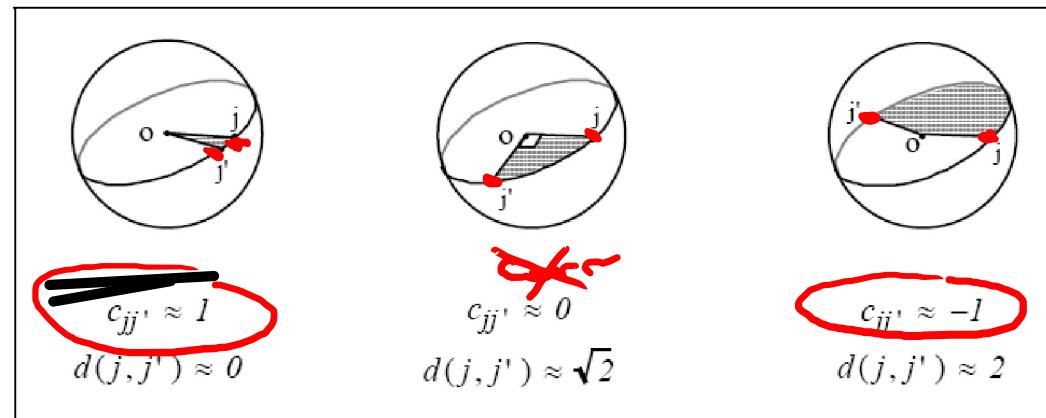
Proximité des variables j et j' ?

$$d^2(j, j') = \sum_{i=1}^n (x_{ij} - x_{ij'})^2 = 2(1 - c_{jj'})$$

$$0 \leq d^2(j, j') \leq 4$$

- Dans \mathbb{R}^n , le cosinus de l'angle entre 2 vecteurs-variables est le coefficient de corrélation entre les 2 variables
- Si les variables sont de variance unité (variables centrées réduites), le cosinus est le produit scalaire

cont



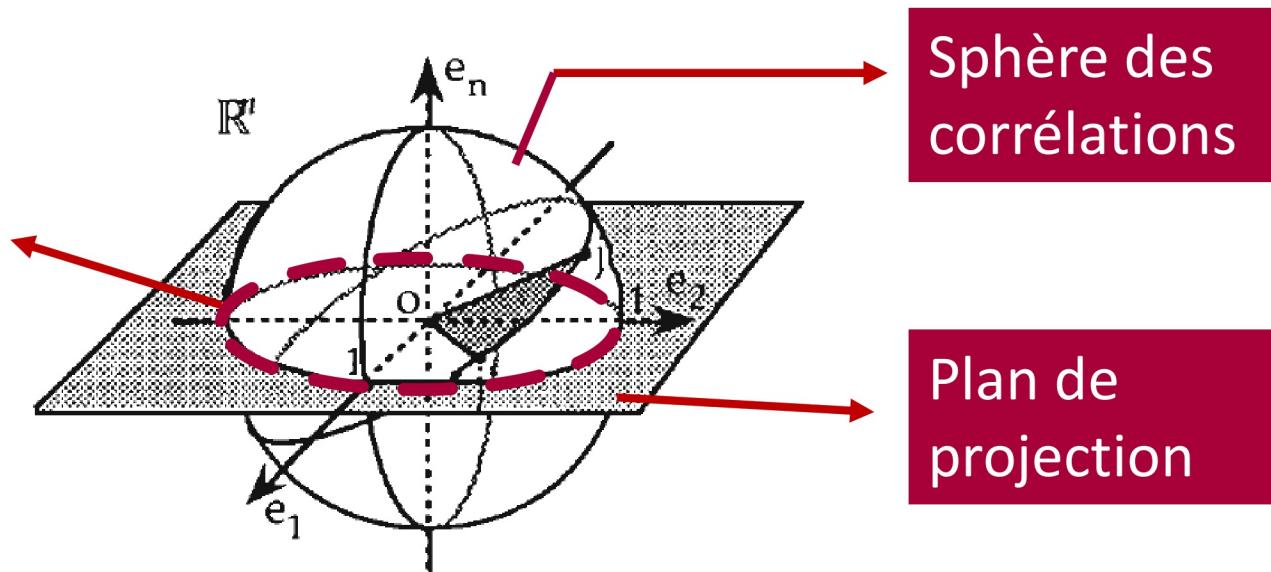
corr -

Distance à l'origine

L'analyse du nuage des points-variables se fait par rapport à l'origine O :

$$d^2(O, j) = \sum_{i=1}^n x_{ij}^2 = 1$$

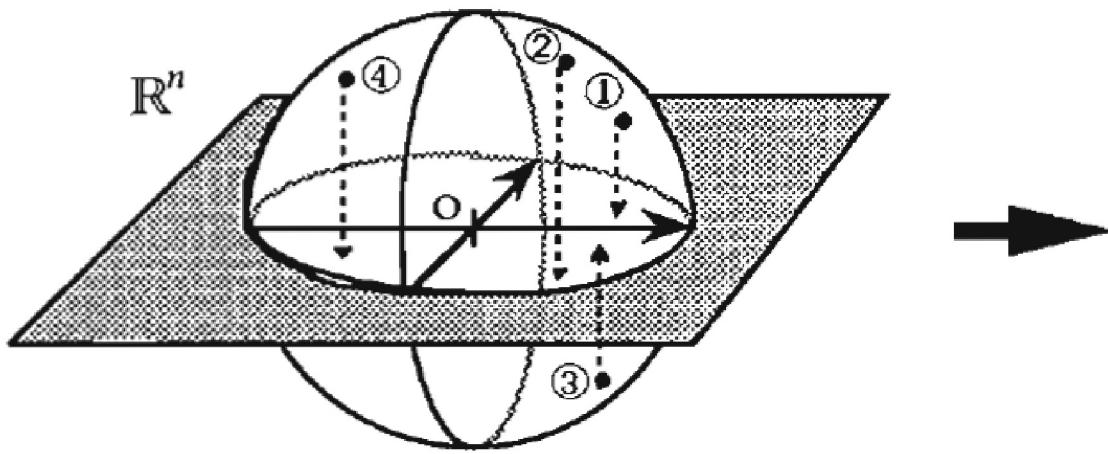
Les points-variables se situent sur une sphère de rayon 1 : c'est la **sphère des corrélations**



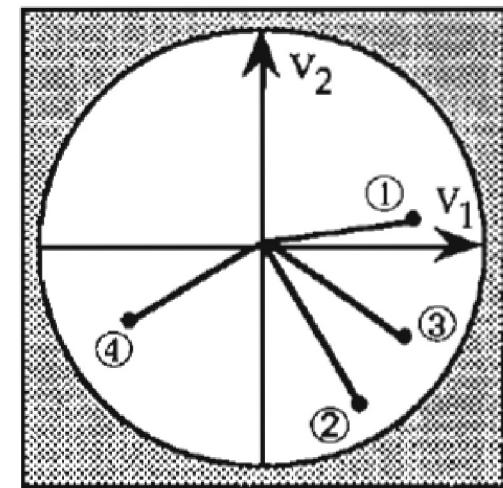
Cercle des corrélations

C'est le résultat de la **projection** de la sphère des corrélations sur un des plans factoriels

projection de 4 variables



Plan factoriel
"cercle des corrélations"



Pour le plan factoriel considéré, les variables « intéressantes » sont celles qui sont **PROCHES** du cercle des corrélation

Axes factoriels ou composantes principales

Dans l'espace des individus (\mathfrak{R}^n), la diagonalisation de la matrice $\mathbf{X}'\mathbf{X}$ est inutile :

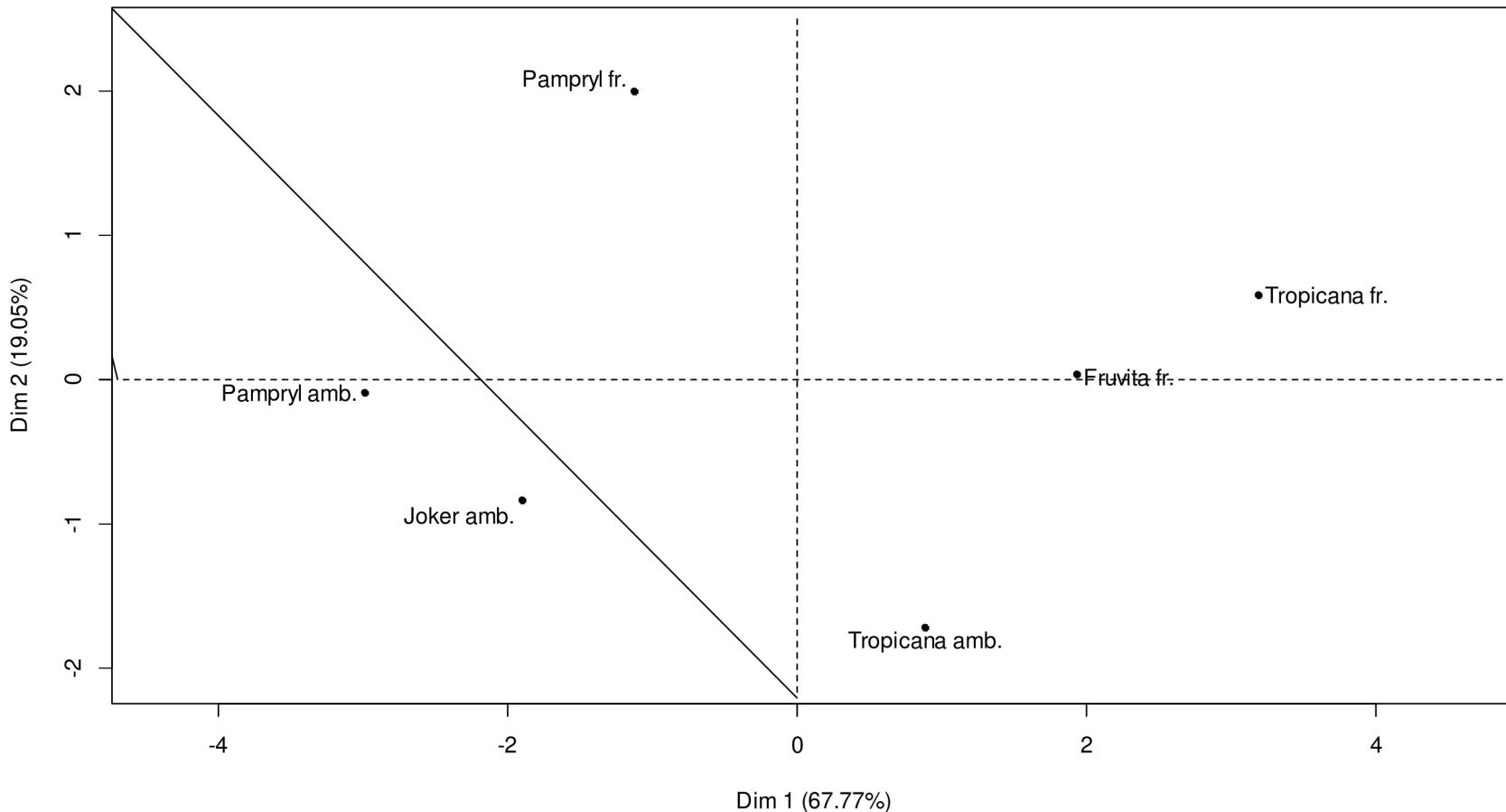
$$\left. \begin{array}{l} v_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \mathbf{X} u_\alpha \\ u_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \mathbf{X}' v_\alpha \end{array} \right\} \xrightarrow{\varphi_\alpha = \mathbf{X}' v_\alpha} \varphi_\alpha = u_\alpha \sqrt{\lambda_\alpha} \longrightarrow \varphi_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \mathbf{X}' \psi_\alpha$$

$$\varphi_{\alpha j} = \sum_{i=1}^n \left(\frac{r_{ij} - \bar{r}_j}{s_j \sqrt{n}} \right) \left(\frac{\psi_{\alpha i}}{\sqrt{\lambda_\alpha}} \right) = cor(j, \psi_\alpha)$$

Coefficient de corrélation de la variable avec le facteur Ψ_α

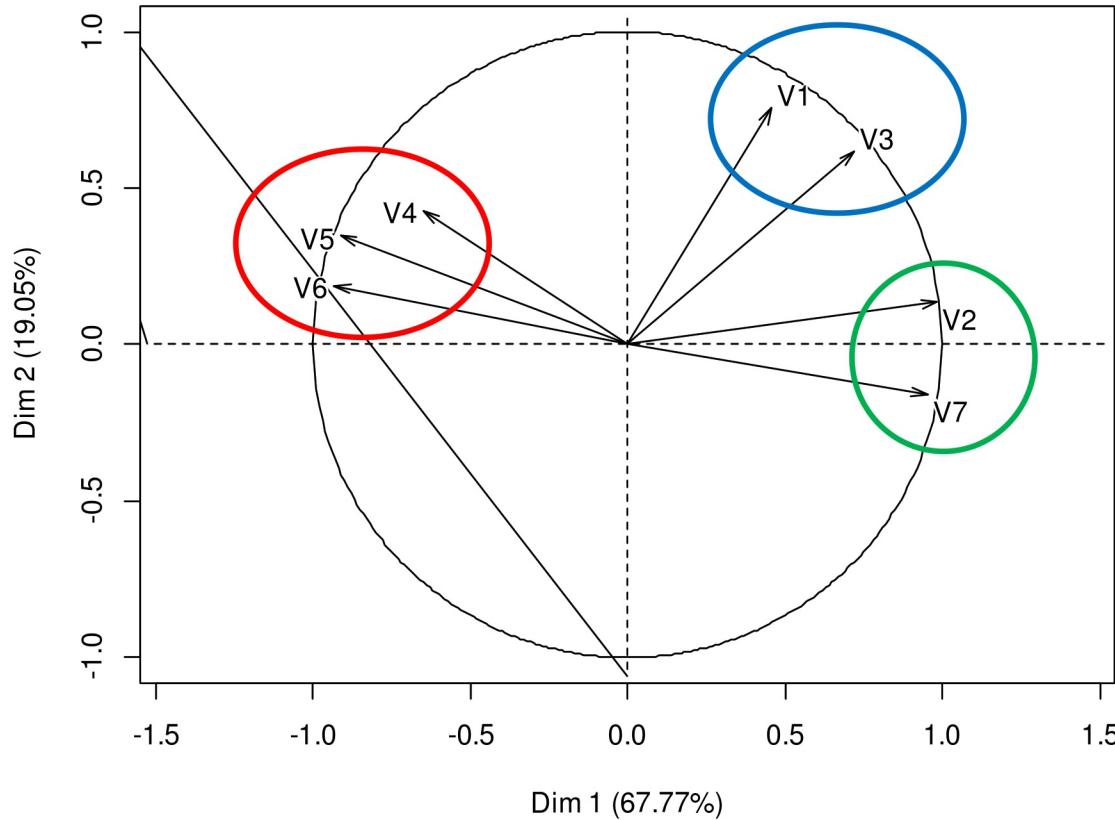
Facteurs (premier plan factoriel)

Individuals factor map (PCA)



Variables

Variables factor map (PCA)



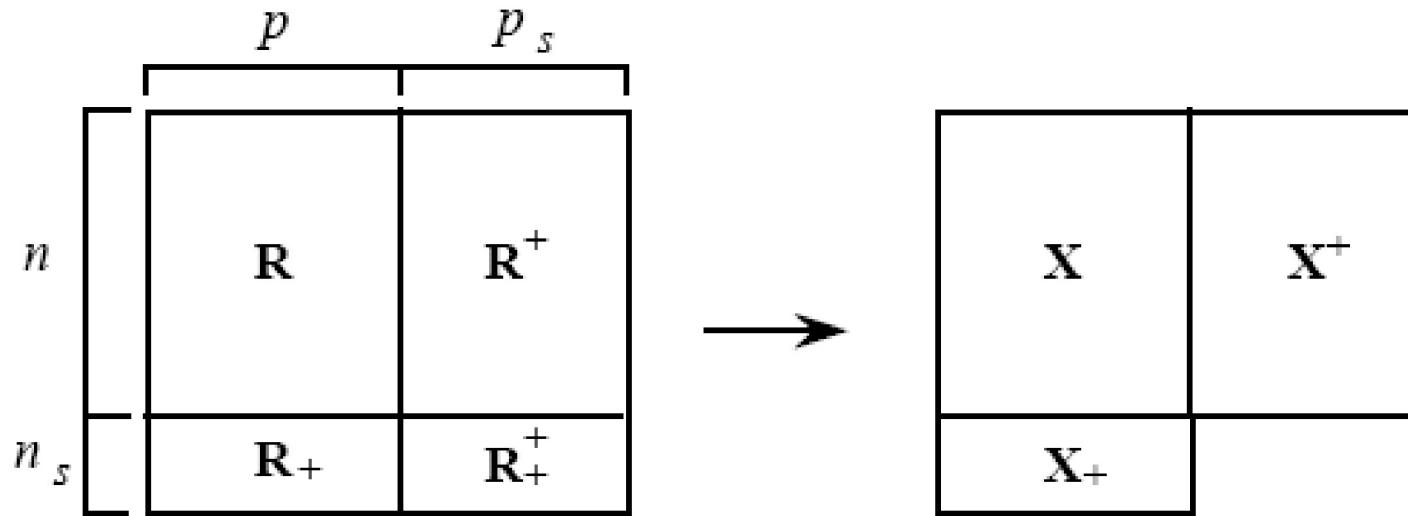
[1] "Odour.intensity"
[5] "Acidity"

"Odour.typicality"
"Bitterness"

"Pulpiness"
"Sweetness"

"Intensity.of.taste"

Individus et variables supplémentaires



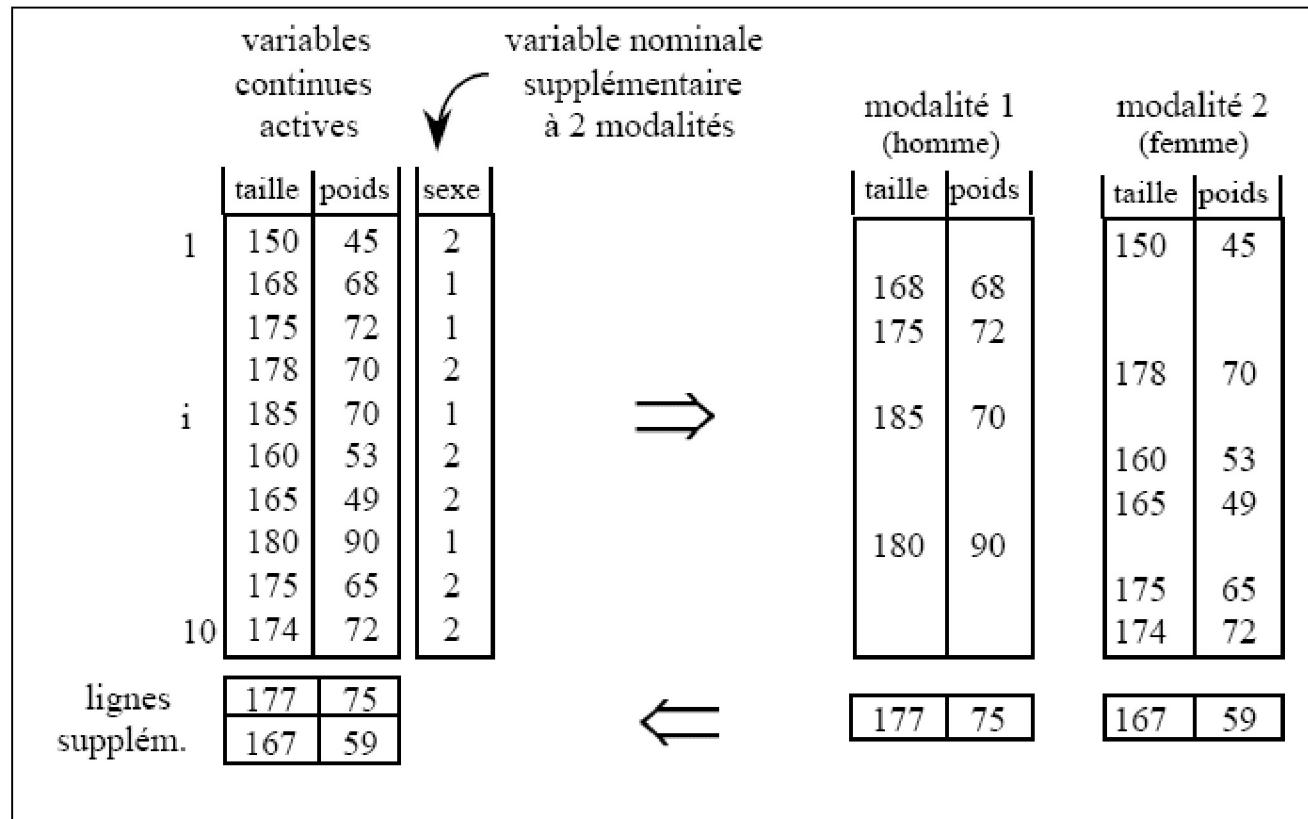
Supplémentaires
=

Ne sont pas utilisés pour
déterminer les axes
Uniquement projetés !

$$x_{+ij} = \frac{r_{+ij} - \bar{r}_j}{s_j \sqrt{n}} \longrightarrow X_+ u_\alpha$$

$$x_{ij}^+ = \frac{r_{ij}^+ - \bar{r}_i^+}{s_j^+ \sqrt{n}} \longrightarrow X^+ v_\alpha$$

Variables NOMINALES supplémentaires

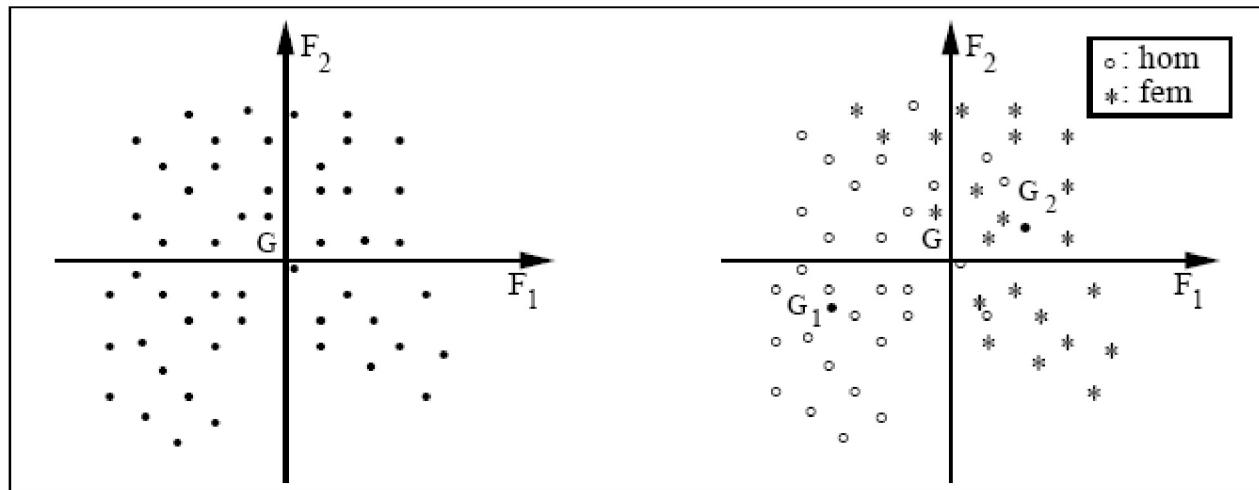


Variable NOMINALE à m modalités : m groupes d'individus

Représentation des variables nominales

On considère les m groupes comme m individus supplémentaires

On positionne les centres de gravité des m groupes



Représentation simultanée : biplot

A) Représentation séparée des 2 nuages

- Espace des variables : $\{G, u_1, \dots, u_\alpha, \dots, u_p\}$

Meilleure visualisation approchée des distances entre individus

- Espace des individus : $\{O, v_1, \dots, v_\alpha, \dots, v_n\}$

Synthèse graphique de la matrice des corrélations

La superposition des plans factoriels est dénuée de sens : ne pas interpréter les distances entre individus et variables

Représentation simultanée : biplot

B) Justification de la représentation simultanée

Points variables → Direction des variables

Dans l'espace des variables : 2 systèmes d'axes $\{e_i\}_{i=1,p}$ et $\{u_\alpha\}_{\alpha=1,p}$
Projection (ligne supplémentaire) de e_i sur u_α :

$$e'_i u_\alpha = u_{\alpha j} \text{ car } e'_i = (0,0,\dots, \underbrace{1}_i, 0, \dots, 0)$$

Représentation simultanée



Écrasement du repère orthonormé des axes d'origine sur le plan factoriel du nuage des individus

Pas de corrélation entre les variables dans cette représentation

Éléments pour l 'interprétation

1. Examen de l 'inertie (variance)
 2. Éléments contribuant à construire les facteurs
-
- A) La valeur propre associée à un axe est la variance des coordonnées des points-individus sur l 'axe correspondant : **indice de dispersion du nuage dans la direction définie par l 'axe.**
 - B) ACP normée : somme des inerties = nombre de variables (inertie moyenne = 1). **On s 'intéresse à des axes dont l 'inertie est notablement supérieure à la moyenne.**
 - C) Pourcentage d 'inertie des axes : « **pouvoir explicatif** »
 - D) Qualité de la représentation :
$$\tau_k = \sum_{i=1,k} \lambda_i / \sum_{i=1,n} \lambda_i$$
 - E) Variables fortement corrélées à un axe : définition de l 'axe
 - F) Variables proches du cercle des corrélations
 - G) Contribution de l 'individu à l 'inertie de l 'axe : $Cr_\alpha(i) = m_i \psi_{\alpha i}^2 / \lambda_\alpha$

Exemple d 'application

Lecture du tableau (voir planches plus loin)

(16 variables continues actives)

Les 27 "individus" (qui sont en réalité dans le cadre de cet exemple des groupes d'individus) sont repérés par un identificateur en 4 caractères :

- le 1er caractère est l'âge du groupe (1=jeune, 2=moyen, 3=âgé)
- le 2ème caractère est ici toujours égal à 1 (car il s'agit ici d'une sélection d'hommes actifs)
- le 3ème est le niveau d'éducation (1= primaire, 2=secondaire, 3=supérieur)
- le 4ème est le type d'agglomération (1=communes rurales; 2=villes moyennes; 3=villes importantes; 4=agglomération parisienne; 5,6,7 = groupes mixtes).

(On trouvera des libellés plus détaillés des variables dans le tableau 1.2 - 2 ci-après.)

On lit par exemple sur la première ligne du tableau 1.2 - 1 que le groupe '1111' (jeunes, actifs, peu instruits, ruraux) consacre en moyenne par jour **463.8** minutes au "sommeil", **23.8** minutes à des activités regroupées sous la rubrique "repos", **107.3** minutes pour les "repas chez soi", etc

Exemple : variables actives

IDEN - LIBELLE	MOYENNE	ECART- TYPE	MINIMUM	MAXIMUM
Variables actives				
Somm - Sommeil	458.91	16.47	433.10	515.60
Repo - Repos	44.63	8.90	23.80	63.10
Reps - Repas chez soi	89.18	8.90	74.20	107.30
Repr - Repas restaurant	13.87	7.82	.30	31.60
Trar - Travail rémunéré	286.27	46.75	208.80	380.60
Ména - Ménage	27.90	9.29	12.90	52.10
Visi - Visite à amis	27.64	13.26	6.50	55.60
Jard - Jardinage, Bricolage	58.49	27.39	4.00	112.90
Lois - Loisirs extérieur	11.42	5.95	1.40	25.60
Disq - Disque cassette	2.54	2.32	.00	8.70
Lect - Lecture livre	7.95	5.47	.00	19.80
Cour - Courses démarches	40.99	9.47	23.30	67.60
Prom - Promenade	9.06	3.88	1.40	17.60
A pi - Déplacement a pied	12.66	5.01	6.40	24.60
Voit - Déplacement en Voiture	58.38	11.29	29.40	81.40
Fréq - Fréquentation Média	140.58	32.56	82.40	225.80

Exemple : variables supplémentaires

Variables continues supplémentaires

Autr - Autres activités	12.71	5.70	2.10	25.90	
Domi - Total Domicile	928.73	49.92	826.00	1034.00	
Tdep - Total Déplacement	88.45	14.65	67.50	122.10	
Habitudes Cinema	.14	.14	.00	.60	
Habitudes Radio.	1.92	.23	1.49	2.64	
Habitudes Télévision	3.20	.37	2.13	3.90	
Habitudes Presse Quotidienne	.18	.14	.03	.53	
Habitudes Presse magazine	3.56	.74	2.00	5.31	
Habitudes Hebdomadaires News	.31	.18	.00	.67	

Exemple : Composantes principales

"nouvelles"

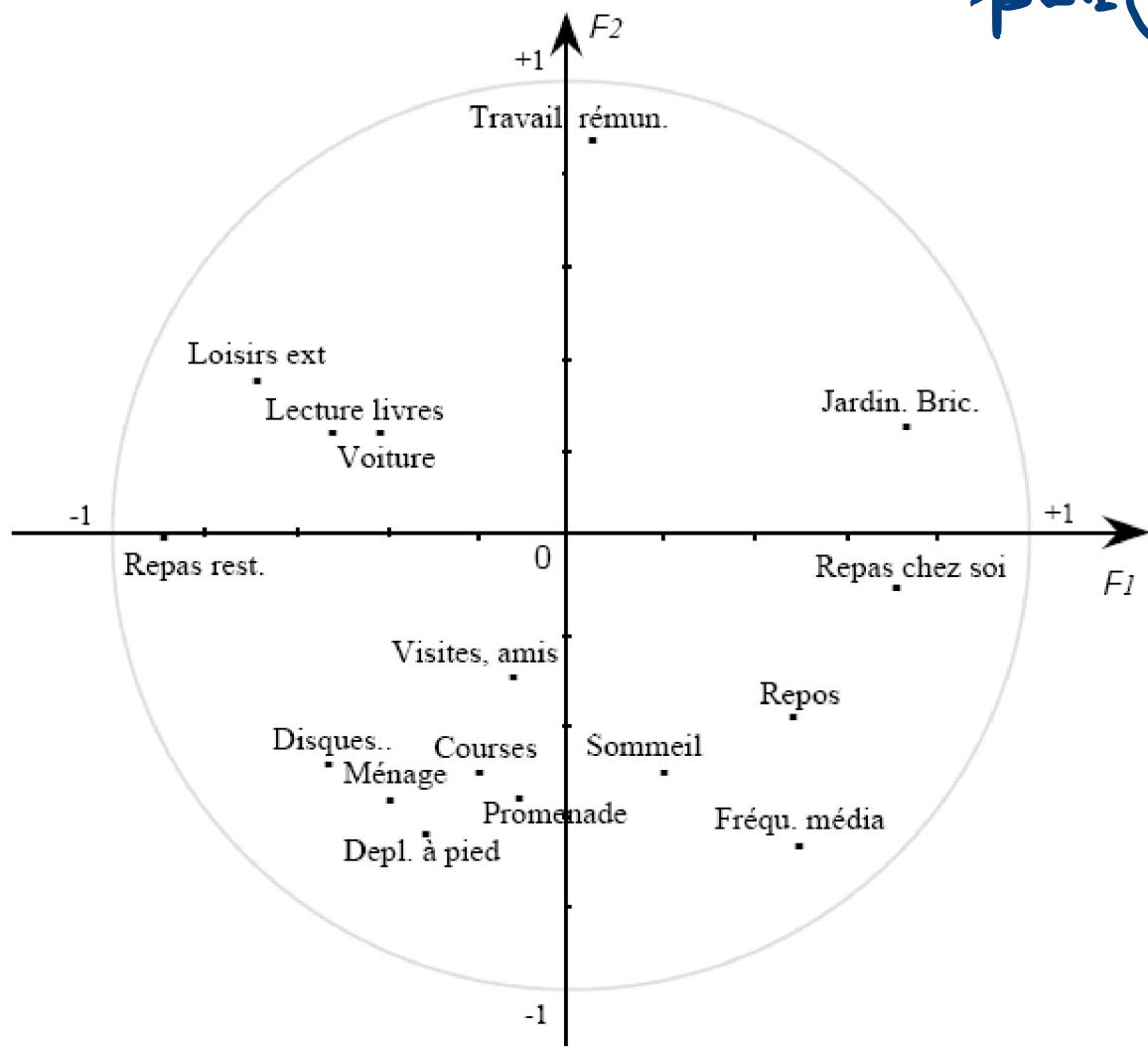
VARIABLES	COORDONNEES			ANCIENS AXES UNIT.		
	1	2	3	1	2	3
Sommeil	.22	-.52	.18	.11	-.27	.13
Repos	.46	-.40	-.17	.23	-.21	-.12
Repas chez soi	.67	-.15	-.23	.34	-.08	-.17
Repas restaurant	-.84	.00	-.07	-.43	.00	-.05
Travail rémunéré	.05	.88	-.34	.03	.46	-.24
Ménage	-.40	-.57	-.08	-.20	-.30	-.06
Visite à amis	-.13	-.33	.73	-.07	-.17	.52
Jardinage, Bricolage	.76	.22	.35	.39	.11	.25
Loisirs extérieur	-.72	.30	.30	-.37	.16	.21
Disque cassette	-.53	-.53	.01	-.27	-.27	.01
Lecture livre	-.54	.24	-.50	-.27	.12	-.36
Courses démarches	-.21	-.54	.11	-.11	-.28	.08
Promenade	-.10	-.58	.04	-.05	-.30	.03
A pied	-.37	-.62	-.57	-.19	-.33	-.40
En Voiture	-.41	.22	.65	-.21	.11	.46
Fréquentation Média	.49	-.68	-.05	.25	-.36	-.03

Sommeil	1.00																
Repos	.21	1.00															
Repas c.	.21	.10	1.00														
Repas r.	-.08	-.30	-.53	1.00													
Travail	-.52	-.28	-.02	-.01	1.00												
Ménage	.20	.08	-.01	.39	-.46	1.00											
Visites	.27	-.08	-.07	.10	-.47	.15	1.00										
Jardin.	-.09	.19	.43	-.64	.08	-.37	-.02	1.00									
Loisirs	-.17	-.61	-.55	.52	.10	-.01	.12	-.39	1.00								
Disques	.07	-.17	-.15	.52	-.46	.50	.30	-.42	.25	1.00							
Lecture	-.44	-.21	-.15	.38	-.24	.08	-.36	-.51	.27	-.01	1.00						
Courses	-.04	.18	-.17	-.03	-.56	.23	.24	-.24	-.01	.08	.18	1.00					
Promen.	.00	.09	.04	-.02	-.45	.27	.18	-.01	-.05	.40	-.03	.48	1.00				
A pied	.17	.15	-.14	.28	-.38	.49	-.18	-.62	-.09	.48	.27	.37	.30	1.00			
Voiture	-.19	-.22	-.55	.21	-.15	.10	.27	.03	.44	-.09	.15	.23	-.11	-.33	1.00		
Fréq.med	.40	.42	.37	-.44	-.62	.05	.01	.18	-.45	.07	-.38	.30	.28	-.33	1.00		
		I Somm	Repo	Reps	Repr	Trar	Ména	Visi	Jard	Lois	Disq	Lect	Cour	Prom	A pi	Voit	Fréq



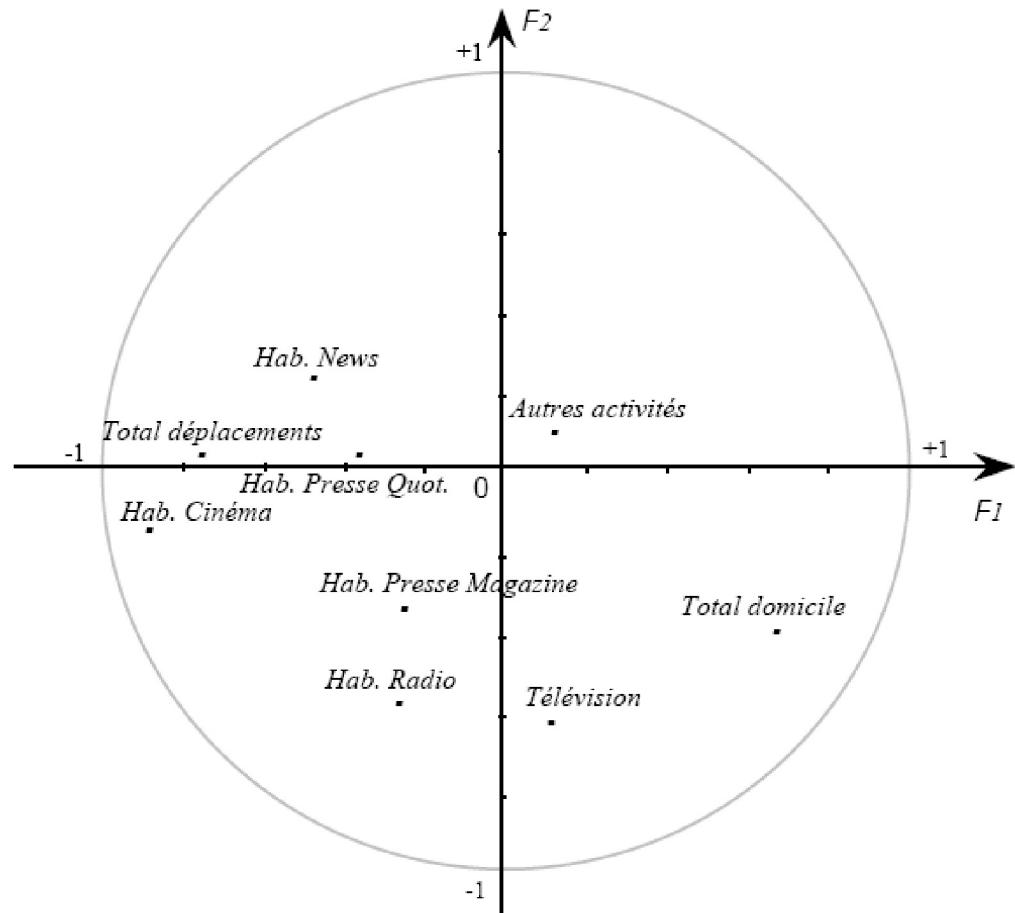
Exemple : Cercle des corrélations

$$+ \rho_{F_1 F_2} (\tau_1 \tau_2)$$



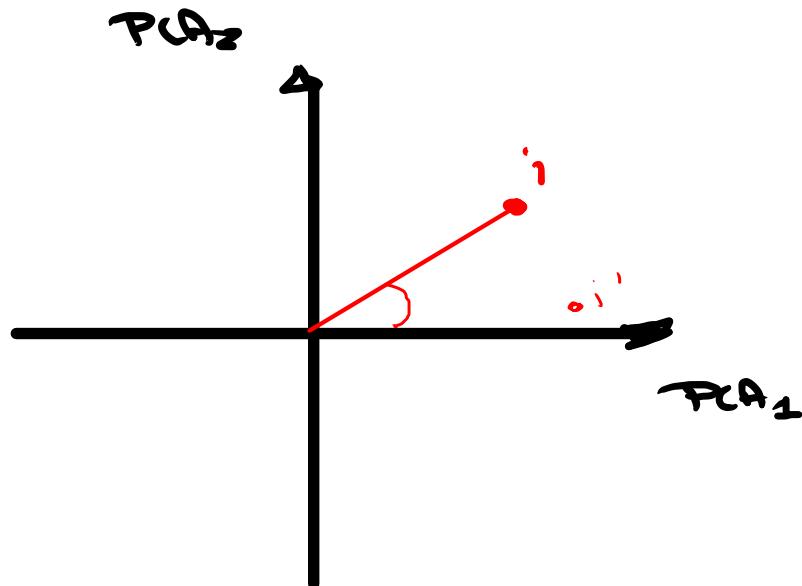
Exemples : variables supplémentaires

VARIABLES	COORDONNEES		
	1	2	3
Autres activités	.08	.16	.04
Total Domicile	.67	-.50	-.21
Total Déplacement	-.72	.05	.14
Habitudes Cinéma	-.87	-.11	-.14
Habitudes Radio.	-.27	-.57	.07
Habitudes Télévision	.04	-.55	.34
Habitudes Presse Quot	-.39	.01	-.70
Habitudes Presse mag	-.24	-.38	-.26
Habitudes Hebdo-News	-.46	.20	-.48



Exemple : individus

INDIVIDUS		COORDONNEES		CONTRIBUT.		COSIN. CARRI	
IDENTIF.	DISTO	1	2	1	2	1	2
1111	19.89	2.01	.85	3.8	.7	.20	.04
1115	47.51	2.26	-5.11	4.9	26.4	.11	.55
1121	10.55	-.71	1.01	.5	1.0	.05	.10
1122	13.29	-1.86	-.64	3.3	.4	.26	.03
1123	14.49	-1.28	-1.01	1.6	3.3	.11	.23
1124	19.06	-2.72	-2.93	7.1	8.7	.39	.45
1136	10.68	-.56	1.97	.3	3.9	.03	.36
1133	27.04	-4.21	-.30	17.0	.1	.66	.00
1134	25.35	-4.29	-.91	17.6	.8	.73	.03
2111	12.86	1.91	2.12	3.5	4.5	.28	.35
2112	17.27	1.43	-1.68	2.0	2.8	.12	.16
2117	10.89	1.03	-2.16	1.0	4.7	.10	.43
2121	10.96	1.27	2.55	1.5	6.6	.15	.59
2122	7.92	.62	-.21	.4	.0	.05	.01
2123	8.33	.30	-.33	.1	.1	.01	.01
2124	15.54	-.12	2.06	.0	4.3	.00	.27
2131	7.39	.55	2.03	.3	4.2	.04	.56
2132	24.45	-1.17	3.53	1.3	12.6	.06	.51
2133	7.85	-1.63	-.11	2.5	.0	.34	.00
2134	17.19	-2.54	1.36	6.2	1.9	.37	.11
3116	16.19	2.68	.96	6.9	.9	.45	.06
3117	15.96	2.43	-1.84	5.7	3.4	.37	.21
3121	13.00	1.90	2.11	3.4	4.5	.28	.34
3122	17.31	2.12	-.95	4.3	.9	.26	.05
3123	10.26	.56	-1.74	.3	3.1	.03	.30
3136	9.09	1.56	.09	2.3	.0	.27	.00
3137	21.68	-1.55	.08	2.3	.0	.11	.00



Exemple : variables nominales

MODALITES		VALEURS-TEST		COORDONNEES	
IDEN - LIBELLE	EFFECT.	1	2	1	2
. AGE					
A-35 - Jeunes	9	-2.3	-1.6	-1.26	-.87
A+35 - Age-Moy	11	.3	1.8	.15	.83
A+50 - Ages	7	2.1	-.3	1.39	-.18
. Education					
prim - primaire	7	3.0	-1.5	1.96	-.98
seco - secondaire	11	.0	-.2	.01	-.08
supe - superieur	9	-2.8	1.6	-1.54	.86
. Agglomération (EXTRAITS)					
AGG1 - de 20 000	6	1.6	2.5	1.15	1.78
AGG3 - Plus de 100 000	5	-1.5	-1.1	-1.25	-.86
AGG4 - Paris	4	-2.6	-.1	-2.42	-.11