Master Droit, Économie, Gestion
Mention Finance
Spécialité Actuariat et gestion du risque

Master de Sciences
Mention Mathématiques et Applications
Spécialité Statistique

Analyse des données

Emmanuel Périnel

Programme

1. ACP – Analyse en composantes principales
2. AFC – Analyse Factorielle des Correspondances
3. ACM – Analyse des Correspondances Multiples
4. CAH – Classification Ascendante Hiérarchique

L'analyse de données brièvement…

Statistique « classique » et analyse de données

- Approche **multidimensionnelle** des individus étudiés
- Synthèse, **résumé** de tableaux de données **volumineux**
- **Peu d'hypothèses** restrictives sur la nature des données
- Représentations graphiques **suggestives**
- Outil mathématique : **géométrie**
- Domaines **d'applications très variés**

L'histoire de l'analyse des données

- Benzecri, Hayashi, Tukey
- Développement « récent »

Quelques références bibliographiques

- **Bouroche J.M. et Saporta G**. (1980). *L'analyse des données*, PUF, Collection Que sais-je ?
- **Cornillon P.A. et autres** (2008). Statistiques avec R, Presses Universitaire de Rennes.
- **Escofier B. et Pagès J**. (2008). *Analyses factorielles simples et multiples,* Objectifs, méthodes et interprétation, 4e édition, Dunod.
- **Husson F., Lê S. et Pagès J.** (2009). *Analyse de données avec R,* Presses Universitaire de Rennes.
- **Lebart L., Morineau A. et Piron M.** (2006). *Statistique exploratoire multidimensionnelle,* Dunod
- **Saporta G.** (2006). *Probabilités, analyses des données et statistiques,* 2e édition, Technip.

1. ACP – Analyse en composantes principales

Plan

1. Tableau de données, exemples
2. Objectifs et principes de base de l'ACP
3. Ajustement du nuage des individus
4. Ajustement du nuage des variables
5. Synthèse entre les deux ajustements
6. ACP normée ou non normée ?
7. Paramétrage avec le logiciel `Rcmdr`
8. Analyse de l'inertie des axes
9. Aides à l'interprétation : qualité de représentation et contribution
10. Méthodologie des variables illustratives

1.1 – Tableaux de données, exemples

Exemple 1. Températures mensuelles de jeux de données

Tableau de données : *Individus* x *Variables*

Ville	janvier	février	mars	avril	mai	juin	juillet	août	septembre	octobre	novembre	décembre
Bordeaux	5,60	6,60	10,30	12,80	15,80	19,30	20,90	21,00	18,60	13,80	9,10	6,20
Brest	6,10	5,80	7,80	9,20	11,60	14,40	15,60	16,00	14,70	12,00	9,00	7,00
Clermont	2,60	3,70	7,50	10,30	13,80	17,30	19,40	19,10	16,20	11,20	6,60	3,60
Grenoble	1,50	3,20	7,70	10,60	14,50	17,80	20,10	19,50	16,70	11,40	6,50	2,30
Lille	2,40	2,90	6,00	8,90	12,40	15,30	17,10	17,10	14,70	10,40	6,10	3,50
Lyon	2,10	3,30	7,70	10,90	14,90	18,50	20,70	20,10	16,90	11,40	6,70	3,10
Marseille	5,50	6,60	10,00	13,00	16,80	20,80	23,30	22,80	19,90	15,00	10,20	6,90
Montpellier	5,60	6,70	9,90	12,80	16,20	20,10	22,70	22,30	19,30	14,60	10,00	6,50
Nantes	5,00	5,30	8,40	10,80	13,90	17,20	18,80	18,60	16,40	12,20	8,20	5,50
Nice	7,50	8,50	10,80	13,30	16,70	20,10	22,70	22,50	20,30	16,00	11,50	8,20
Paris	3,40	4,10	7,60	10,70	14,30	17,50	19,10	18,70	16,00	11,40	7,10	4,30
Rennes	4,80	5,30	7,90	10,10	13,10	16,20	17,90	17,80	15,70	11,60	7,80	5,40
Strasbourg	0,40	1,50	5,60	9,80	14,00	17,20	19,00	18,30	15,10	9,50	4,90	1,30
Toulouse	4,70	5,60	9,20	11,60	14,90	18,70	20,90	20,90	18,30	13,30	8,60	5,50
Vichy	2,40	3,40	7,10	9,90	13,60	17,10	19,30	18,80	16,00	11,00	6,60	3,40

Objectifs

- Comparer les villes du point de vue des températures mensuelles
- Établir des **ressemblances** entre villes, des oppositions ?
  Une **typologie** des villes ?
- Certains mois de l'année sont-ils liés ? Sont-ils **corrélés** entre eux ?
- Peut-on résumer des variables fortement corrélées par des **variables synthétiques** ?

Exemple 2. Analyse sensorielle de 8 bières

- En lignes : 8 bières du marché
- En colonnes : 12 descripteurs sensoriels, une note hédonique
- Chaque case : note moyenne accordée par un jury entraîné

Produit	Mousse	NuanceJaune	Int Odeur	Odeurlevure	Odeur Malt	Efferve	int. Gust.	Amere	Sucree	Acide	SaveurMalt	Sens alcool	Note Hedo
Buckler	3,63	4,88	5,78	3,97	5,34	5,56	5,71	4,94	3,59	2,88	5,13	3,94	4,06
Kro 2,6	6,59	4,78	4,47	3,50	3,91	5,59	5,09	4,41	3,59	3,81	3,78	4,25	4,34
Tourtel	6,88	4,59	5,53	4,53	4,16	5,06	5,13	4,50	3,41	3,69	4,10	3,53	4,13
Leffe	6,47	8,53	8,25	5,78	7,22	6,53	8,16	6,09	4,22	4,00	6,90	6,47	5,06
Kronenbourg	6,72	4,34	5,38	4,41	3,75	4,69	5,53	5,97	3,19	3,81	3,97	4,81	4,38
K	5,22	3,72	4,63	4,13	3,03	6,10	5,31	4,31	3,63	4,13	3,47	4,88	4,56
Heineken	4,41	2,09	4,66	3,88	3,34	5,81	5,94	6,09	3,25	3,75	3,53	4,75	3,66
Kro pur malt	4,74	5,35	6,90	4,19	6,65	5,03	6,67	4,71	4,55	3,55	6,45	4,42	4,91

Objectifs

- Certaines bières présentent-elles un **profil sensoriel similaire** ?
- Établir une **typologie** des produits
- Certains attributs sensoriels ont-ils **corrélés** entre eux ?
- Peut-on résumer des descripteurs sensoriels fortement corrélées entre eux par des **variables synthétiques** ?

Exemple 3. Comparaison de cordages ou raquettes de tennis

- Projection d'une balle sur un tamis

Cordage	Déflexion	Force	Vall	Vret	Ratio V	Pente all	Pente ret	Ratio Pente
167a	-13,86	144,88	49,84	41,49	0,83	-9354,54	8034,75	85,88
266a	-13,91	148,03	49,68	40,22	0,81	-9325,78	8056,26	86,36
546b	-14,10	137,24	49,92	41,84	0,84	-8963,19	8151,42	90,97
659g	-14,50	143,69	49,26	41,49	0,84	-9137,72	8226,57	90,06
VS Touch	-14,13	140,06	49,22	40,18	0,82	-8805,24	8159,37	89,48
693a	-14,31	140,12	49,70	41,86	0,84	-9056,17	8210,71	90,66

- Test sensoriel auprès de joueurs confirmés

Raquette	Raideur	contrôle	Prise effet	Puissance
784a/NOIR/PROTO	6,75	7	6,38	7,25
753a/JAUNE/PHT	6,88	7,13	7,38	7
643a/NATUREL/XL	4,5	6,63	5,75	6,5
301b/XCEL POWER 125	5,33	6,56	6,44	6,78
616j/CONQUEST 125	6	6,11	5,33	6,56
801c/PHTOUR 125	7,22	7,33	7,22	6,67
829 q (125)	7,13	6,63	5,88	4,12
855 i (130)	6	6	6,88	5,93

Exemple 4. Analyse d'échantillons de sol

Code	Altitude	pente en °	Végétation	type de sol	Carbone organique	pH	%argile	%limon	%sable
RTH	1080	5	forêt	ranker sur éboulis	50,3	3,68	8,00	67,30	24,80
THA	1100	4	forêt	alocrisol	21,4	3,62	12,30	36,60	51,10
THB	1060	15	forêt	alocrisol	13,3	3,96	11,00	25,20	63,80
RCA	1120	4	prairie	alocrisol	16,8	4,20	18,40	45,40	36,20
RCD	1170	8	prairie	alocrisol	24,8	4,12	11,10	49,40	39,50
CFA	1070	5	prairie	alocrisol	19,9	4,000	7,000	17,800	75,200
RCB	1080	3	prairie	alocrisol	16,6	4,350	9,700	23,800	66,500
DHS	660	1	forêt	alocrisol	20	3,540	12,900	49,500	37,600
FAC	1300	1	prairie	ranker cryptpodzolique	31,3	3,700	5,000	24,500	70,500
KAC	1350	1	prairie	ranker cryptpodzolique	20,4	3,910	2,900	12,900	84,200
KAT	1350	1	prairie	ranker cryptpodzolique	31,4	3,960	2,700	10,900	86,400
WHS	650	18	forêt	sol podzolique	4,5	3,700	1,100	4,800	94,200
HAC	1080	10	prairie	alocrisol	22,3	4,23	5,00	48,00	47,00

Exemple 5. Typologie d'agence bancaires

Guichet	DateCréation	Nb Clients	Ancien. Moy	Age <12	Age 22-43	Nb Ouv CAV	...	Nb Ferm CAV	% Client risq niv1	% clients Assur v	% Clients CB
10											
12											
15											
17											
19											
21											
23											
25											
27											
29											
⋮											
4110											
4112											
4114											
4121											
4125											
4150											
4161											
4212											
4215											

- 965 agences bancaires
- Variables « clientèles »
- Variables « performance »
- Variables IRIS - Insee

Exemple 6. Perception de publicités sur des parfums

- 400 femmes
- 40 parfums (pub)
- 75 descripteurs

Pub	Addictive_Intriguing	Casual	Cheap	Classic	Conventional	Daring	Easy_Comfortable	Elegant_refined	Expensive	Fashion	Feminine	Fra
Angel	90	19	41	27	6	129	17	56	70	89	143	111
Aromatics Elixir	116	24	24	33	24	162	19	29	51	51	60	133
Avon Imari	45	44	73	54	34	57	53	81	25	88	206	176
Baby Phat Fabulosity	53	14	18	62	13	57	44	137	147	177	227	142
Belong Celine Dion	32	99	45	80	47	20	140	82	36	88	207	155
Chanel 5	66	17	8	249	32	32	30	275	194	145	251	192
Chanel Chance	84	18	28	62	12	143	44	68	99	96	172	165
Chanel Coco Mad	93	13	21	82	11	151	17	107	140	138	193	151
Ck one	53	61	58	44	13	119	44	14	28	96	60	109
Clinique Happy	45	104	36	34	13	43	141	16	20	65	187	159
Coach	53	26	19	98	30	60	33	95	160	209	162	151
DG Light Blue	66	83	27	70	17	81	89	44	79	86	177	128
DG The One	100	20	14	66	21	107	34	111	153	144	215	172
DK Cashmere Mist	72	48	18	103	31	58	58	90	90	87	165	122
DKNY Be Delicious	57	113	26	47	22	38	137	33	31	77	163	140
DKNY Delicious Night	79	37	17	37	24	80	51	60	62	130	198	174
Daisy Marc Jacobs	31	45	89	24	5	85	44	20	31	60	135	115
Diesel Fuel for Life	74	21	56	35	19	173	12	33	56	120	112	122

Extrait du tableau de données

Exemple 7. Comparaison du lait de mammifères

- 13 mammifères
- Composition biologique

LaitMami

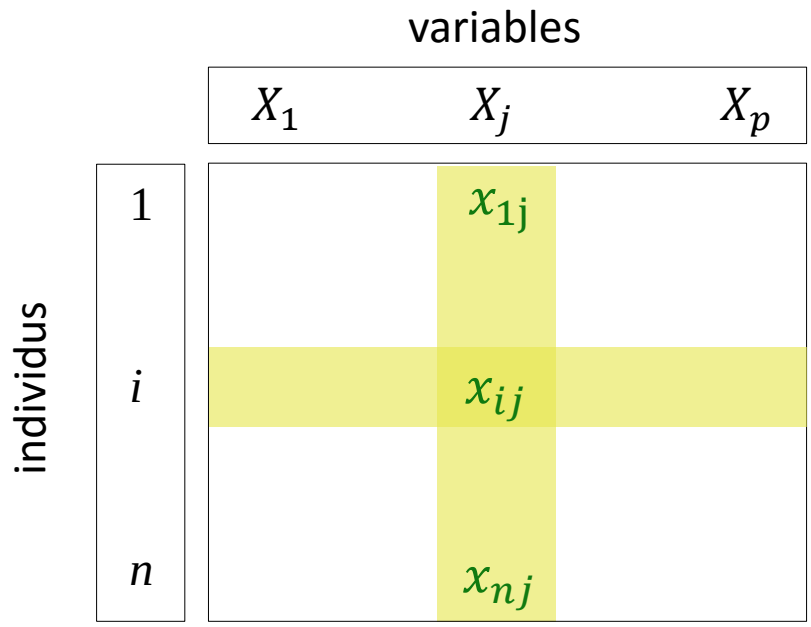
	Eau	Extrait.sec	Matière.grasse	Protéines.totales	Caséine	Lactose	Matières.minérales
Maternel	905	117	35.0	13.0	11.0	67.0	3.0
Jument	925	100	12.5	21.0	11.0	62.5	4.0
Ânesse	925	100	12.5	21.0	11.0	62.5	4.5
Vache	900	130	37.5	32.5	28.5	47.5	9.0
Chèvre	900	140	42.5	37.5	32.5	42.5	9.0
Brebis	860	190	72.5	57.5	47.5	47.5	11.0
Bufflonne	850	180	72.5	47.5	37.5	47.5	9.0
Renne	675	330	180.0	102.5	82.5	27.5	17.5
Truie	850	185	65.0	57.5	27.5	52.5	13.5
Chienne	800	250	95.0	105.0	47.5	40.0	13.0
Chatte	850	200	45.0	95.0	32.5	45.0	11.5
Lapine	720	300	125.0	135.0	95.0	17.5	17.5
Baleine	467	600	440.0	125.0	70.0	18.0	5.0

1.2 – Objectifs et principes de l'ACP

Le tableau de données étudié

Tableau de données sous forme *individus* x *variables*

- Un ensemble de *n* individus *décrits par*
- *p* variables **quantitatives**



Notations

X : matrice de dim (n, p) représentant les valeurs du tableau de données

$x_i = (x_{i1}, \dots, x_{ij}, \dots x_{ip})$ vecteur des valeurs prises par l'individu i

Objectifs

- **Lignes** du tableau : établir un bilan des **ressemblances entre individus**
- **Colonnes** du tableau : réaliser un **bilan des corrélations** entre variables
- **Mise en liaison** des deux études
  Quelles sont les variables caractéristiques d'un groupe d'individus donné ?
- **Construction de variables synthétiques** (composantes principales)

L'ACP est fondée sur des **représentations graphiques suggestives**

Présentation intuitive

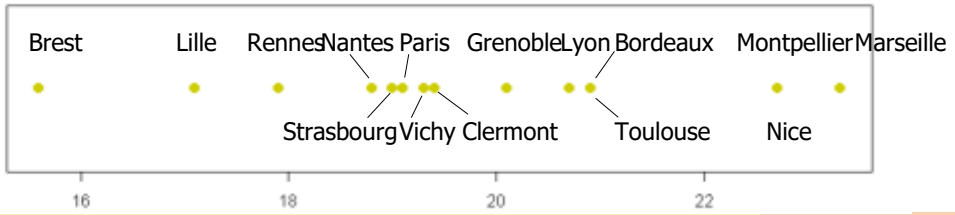
Point de vue des individus

Objectif = résumer par un graphique les similarités entre individus, du point de vue de l'ensemble des *p* variables *simultanément* !

S'il n'y avait **qu'une seule** variable ...

Ville	janvier	février	mars	avril	mai	juin	juillet	août	septembre	octobre	novembre	décembre
Bordeaux	5,80	6,80	10,30	12,80	15,80	19,30	20,90	21,00	18,80	13,80	9,10	6,20
Brest	6,10	5,80	7,80	9,20	11,80	14,40	15,80	16,00	14,70	12,00	9,00	7,00
Clermont	2,80	3,70	7,50	10,30	13,80	17,30	19,40	19,10	16,20	11,20	6,80	3,80
Grenoble	1,50	3,20	7,70	10,80	14,50	17,80	20,10	19,50	16,70	11,40	6,50	2,30
Lille	2,40	2,50	6,00	8,90	12,40	15,30	17,10	17,10	14,70	10,40	6,10	3,50
Lyon	2,10	3,30	7,70	10,90	14,90	18,50	20,70	20,10	16,90	11,40	6,70	3,10
Marseille	5,50	6,60	10,00	13,00	16,80	20,80	23,30	22,80	19,90	15,00	10,20	6,90
Montpellier	5,80	6,70	9,90	12,80	16,20	20,10	22,70	22,30	19,30	14,60	10,00	6,50
Nantes	5,00	5,30	8,40	10,80	13,90	17,20	18,80	18,80	16,40	12,20	8,20	5,50
Nice	7,50	8,50	10,80	13,30	16,70	20,10	22,70	22,50	20,30	16,00	11,50	8,20
Paris	3,40	4,10	7,60	10,70	14,30	17,50	19,10	18,70	16,00	11,40	7,10	4,30
Rennes	4,80	5,30	7,50	10,10	13,10	16,20	17,90	17,80	15,70	11,60	7,80	5,40
Strasbourg	0,40	1,50	5,60	9,80	14,00	17,20	19,00	18,30	15,10	9,50	4,90	1,30
Toulouse	4,70	5,60	9,20	11,80	14,90	18,70	20,90	20,50	18,30	13,30	8,80	5,50
Vichy	2,40	3,40	7,10	9,90	13,80	17,10	19,30	18,80	16,00	11,00	6,80	3,40

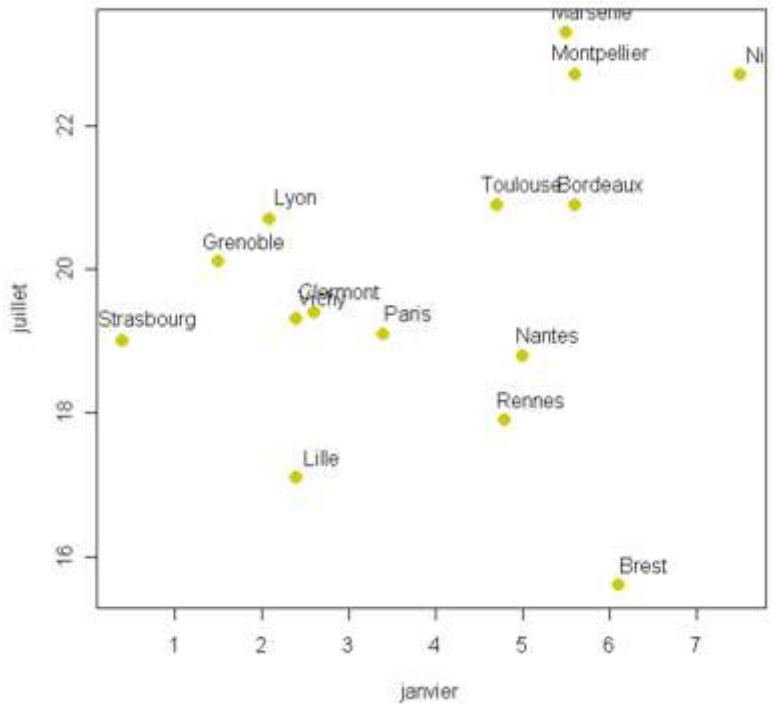
Représentation axiale



S'il y avait **deux** variables ...

Ville	janvier	février	mars	avril	mai	juin	juillet	août	septembre	octobre	novembre	décembre
Bordeaux	5,80	6,80	10,30	12,80	15,80	19,30	20,90	21,00	18,80	13,80	9,10	6,20
Brest	6,10	5,80	7,80	9,20	11,80	14,40	15,80	16,00	14,70	12,00	9,00	7,00
Clermont	2,80	3,70	7,50	10,30	13,80	17,30	19,40	19,10	16,20	11,20	6,80	3,80
Grenoble	1,50	3,20	7,70	10,80	14,50	17,80	20,10	19,50	16,70	11,40	6,50	2,30
Lille	2,40	2,90	6,00	8,90	12,40	15,30	17,10	17,10	14,70	10,40	6,10	3,50
Lyon	2,10	3,30	7,70	10,90	14,90	18,50	20,70	20,10	16,90	11,40	6,70	3,10
Marseille	5,50	6,80	10,00	13,00	16,80	20,80	23,30	22,80	19,90	15,00	10,20	6,90
Montpellier	5,80	6,70	9,90	12,80	16,20	20,10	22,70	22,30	19,30	14,80	10,00	6,50
Nantes	5,00	5,30	8,40	10,80	13,90	17,20	18,80	18,60	16,40	12,20	8,20	5,50
Nice	7,50	8,50	10,80	13,30	16,70	20,10	22,70	22,50	20,30	16,00	11,50	8,20
Paris	3,40	4,10	7,80	10,70	14,30	17,50	19,10	18,70	16,00	11,40	7,10	4,30
Rennes	4,80	5,30	7,50	10,10	13,10	16,20	17,90	17,80	15,70	11,60	7,80	5,40
Strasbourg	0,40	1,50	5,60	9,80	14,00	17,20	19,00	18,30	15,10	9,50	4,90	1,30
Toulouse	4,70	5,80	9,20	11,80	14,90	18,70	20,90	20,50	18,30	13,30	8,80	5,50
Vichy	2,40	3,40	7,10	9,90	13,80	17,10	19,30	18,80	16,00	11,00	6,80	3,40

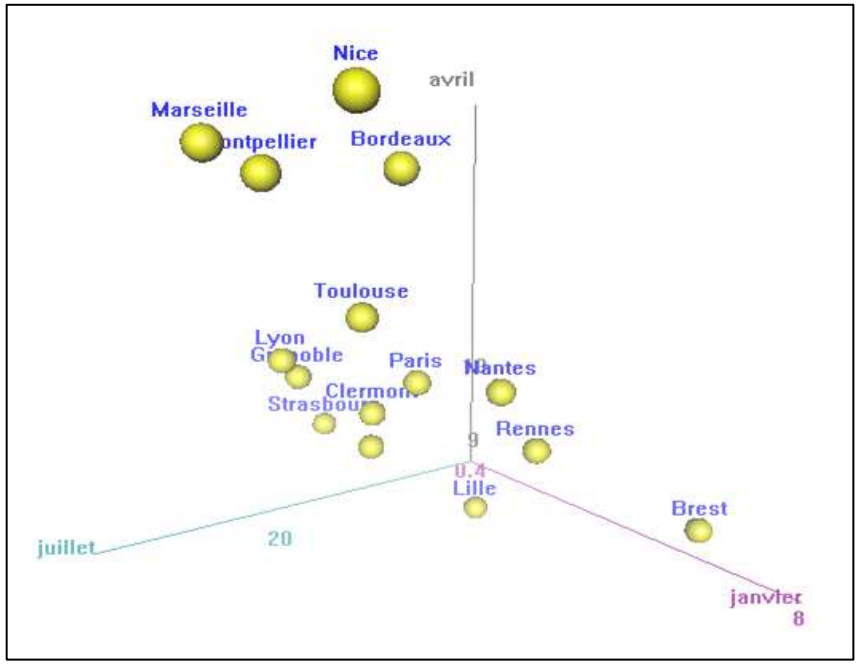
Nuage de points



Pour **trois** variables ...

Ville	janvier	février	mars	avril	mai	juin	juillet	août	septembre	octobre	novembre	décembre
Bordeaux	5,80	6,80	10,30	12,80	15,80	19,30	20,90	21,00	18,80	13,80	9,10	6,20
Brest	8,10	5,90	7,80	9,20	11,80	14,40	15,80	16,00	14,70	12,00	9,90	7,00
Clermont	2,80	3,70	7,50	10,30	13,80	17,30	19,40	19,10	16,20	11,20	6,80	3,80
Grenoble	1,50	3,20	7,70	10,80	14,50	17,80	20,10	19,50	16,70	11,40	6,50	2,30
Lille	2,40	2,90	6,00	8,90	12,40	15,30	17,10	17,10	14,70	10,40	6,10	3,50
Lyon	2,10	3,30	7,70	10,50	14,90	18,50	20,70	20,10	16,90	11,40	6,70	3,10
Marseille	5,50	6,80	10,00	13,00	16,80	20,80	23,30	22,80	19,90	15,00	10,20	6,90
Montpellier	5,80	6,70	9,90	12,80	16,20	20,10	22,70	22,30	19,30	14,80	10,00	6,50
Nantes	5,00	5,30	8,40	10,80	13,90	17,20	18,80	18,80	16,40	12,20	8,20	5,50
Nice	7,50	8,50	10,80	13,30	16,70	20,10	22,70	22,50	20,30	16,00	11,50	8,20
Paris	3,40	4,10	7,80	10,70	14,30	17,50	19,10	18,70	16,00	11,40	7,10	4,30
Rennes	4,80	5,30	7,90	10,10	13,10	16,20	17,90	17,80	15,70	11,80	7,80	5,40
Strasbourg	0,40	1,50	5,80	9,80	14,00	17,20	19,00	18,30	15,10	9,50	4,90	1,30
Toulouse	4,70	5,80	9,20	11,80	14,90	18,70	20,90	20,50	18,30	13,30	8,80	5,50
Vichy	2,40	3,40	7,10	9,90	13,80	17,10	19,30	18,80	16,00	11,00	6,80	3,40

Représentation 3D

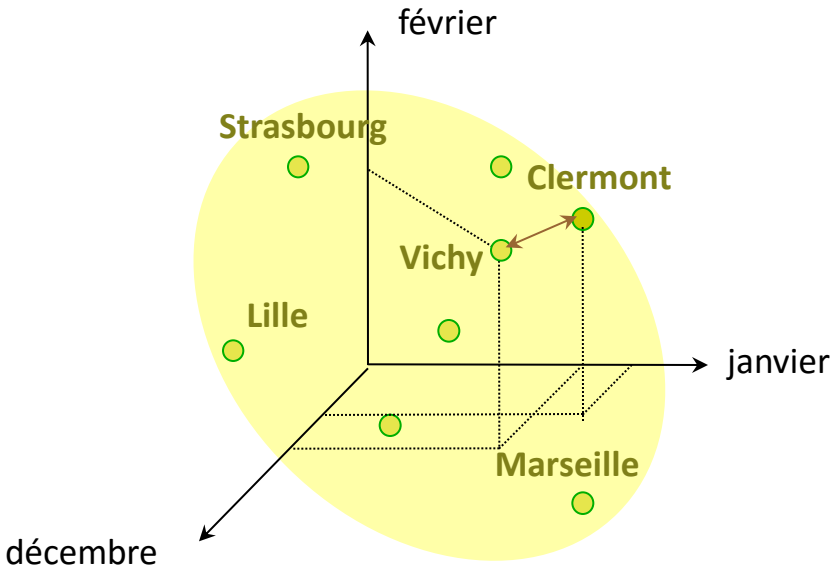


Problème :
Quelle représentation proposer pour plus de trois variables ?

1.3 – Ajustement du nuage des individus

Le nuage des individus

Les individus constituent **un nuage de points** dans un espace comportant autant de dimensions que de variables = inaccessible !



- *Deux individus proches dans ce nuage, prennent des valeurs similaires pour toutes les variables*
- Distance entre 2 individus = **distance euclidienne** usuelle

$$d^2(i,l) = \sum_{j=1}^{p} m_j \left(x_{ij} - x_{lj}\right)^2$$

- m_j : poids de la variable j

Forme d'un nuage - Inertie d'un nuage de point

Mécanique

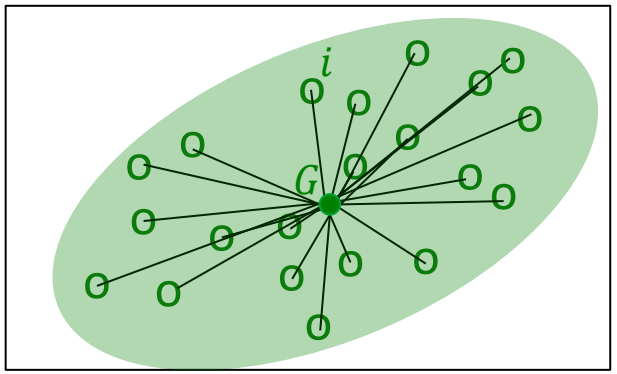
Inertie d'un corps = résistance d'un corps à une mise en rotation autour d'un axe

Statistique

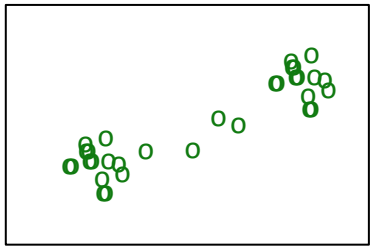
Inertie d'un nuage de points = mesure de **dispersion** des points au sein du nuage
= somme des **carrés des distances** par rapport au centre de gravité *G* du nuage

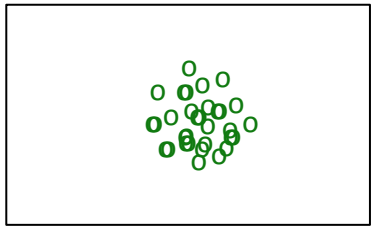
$$I = \sum_{i=1}^{n} m_i \times d^2(G, i)$$

i : individu (i)
G : centre de gravité du nuage
m_i : masse de l'individu (i)

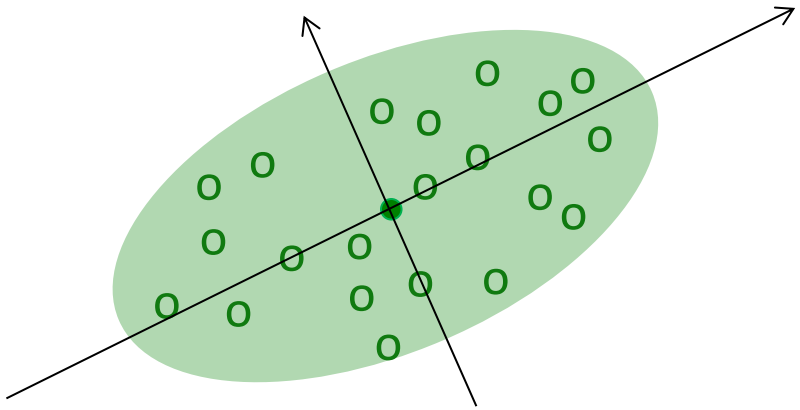


L'inertie est étroitement liée à la **forme du nuage**





Pour étudier les principales différences entre individus, l'ACP va analyser la **forme du nuage**, en rechercher ses principales directions d'allongement (appelées axes factoriels, axes principaux d'inertie)



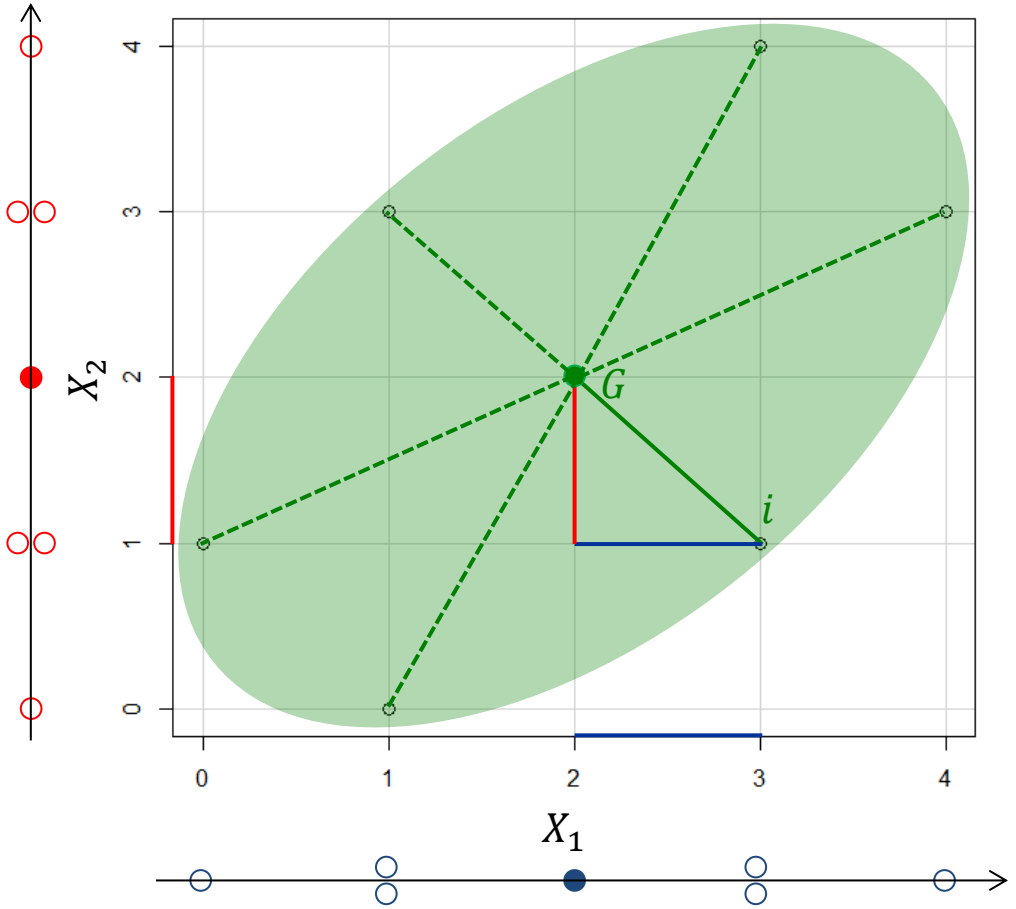
L'Inertie : une mesure de *variance multidimensionnelle*

Illustration en deux dimensions

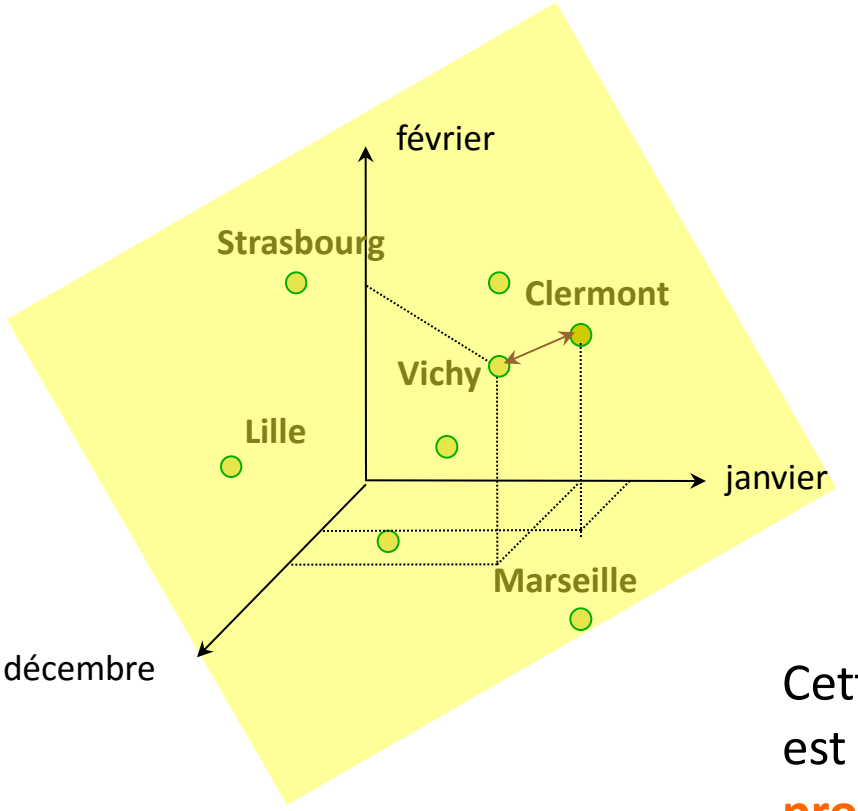
$$I = V(X_1) + V(X_2)$$

Plus généralement

$$I = \sum_{j=1}^{p} V(X_j)$$



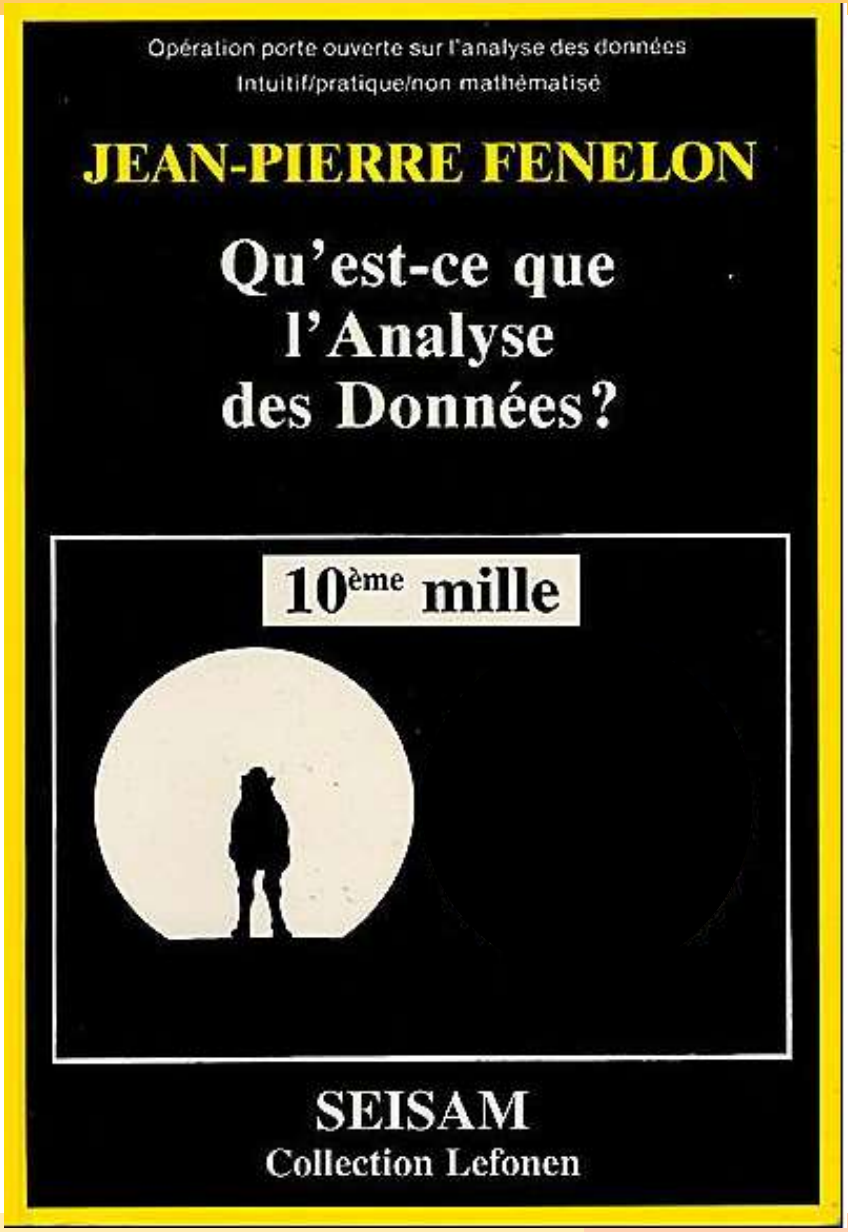
Comment obtenir la « meilleure image » de ce nuage ?



Comment restituer
le plus fidèlement possible
et à l'aide d'un graphique simple
la forme de ce nuage ?

Cette **représentation simplifiée** du nuage
est obtenue à travers une
projection des individus du nuage sur un plan

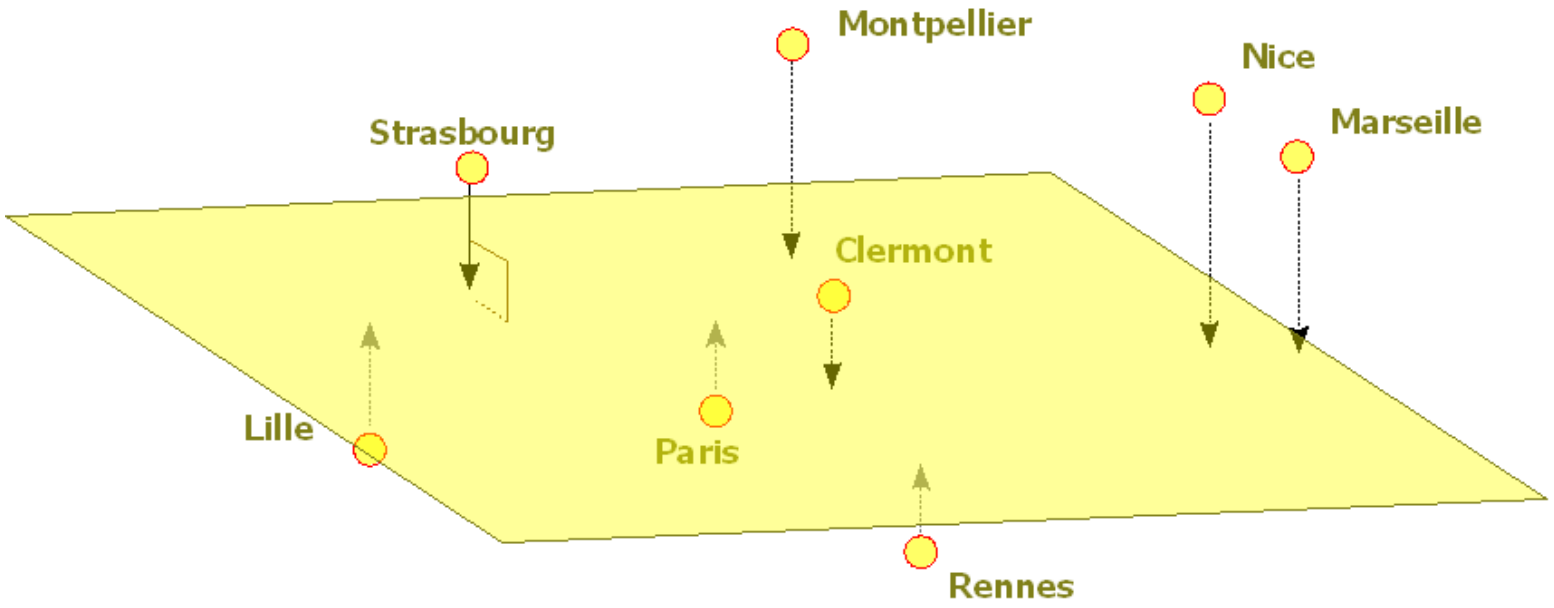
Une projection étrange…



La **meilleure projection** =

- la plus « suggestive »
- La moins déformante
- La vue la plus fidèle de la forme réelle du nuage
- la vue la **plus vaste** de l'objet

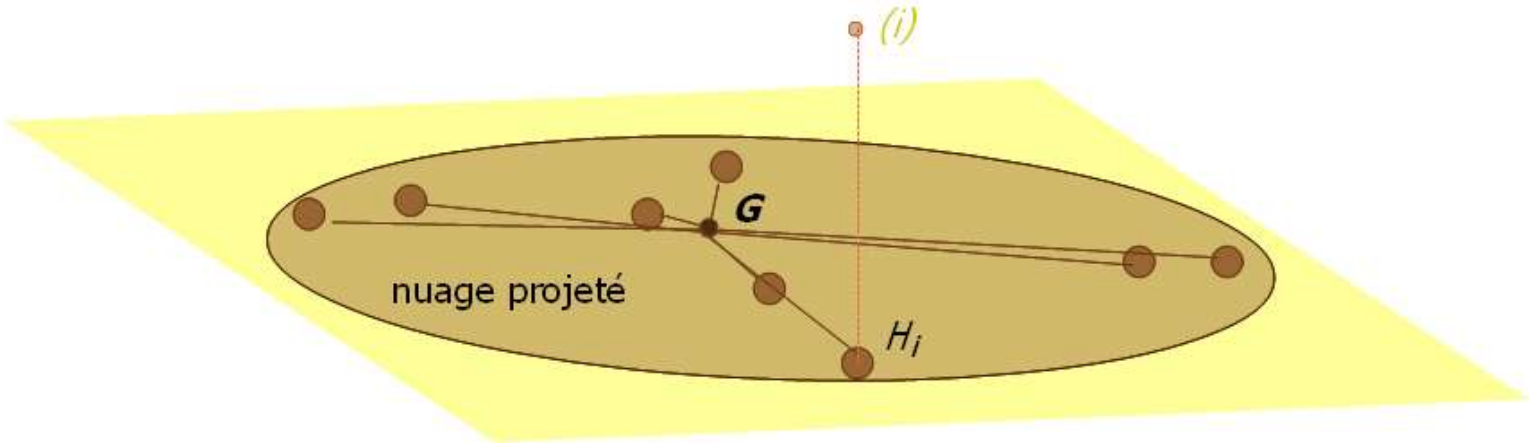
Projection des individus sur le plan



= Projection **orthogonale**

Comment évaluer la qualité de la projection ?

En projection sur le plan, les points ont une **dispersion**, appelée **inertie projetée**
Celle-ci doit être **maximale**

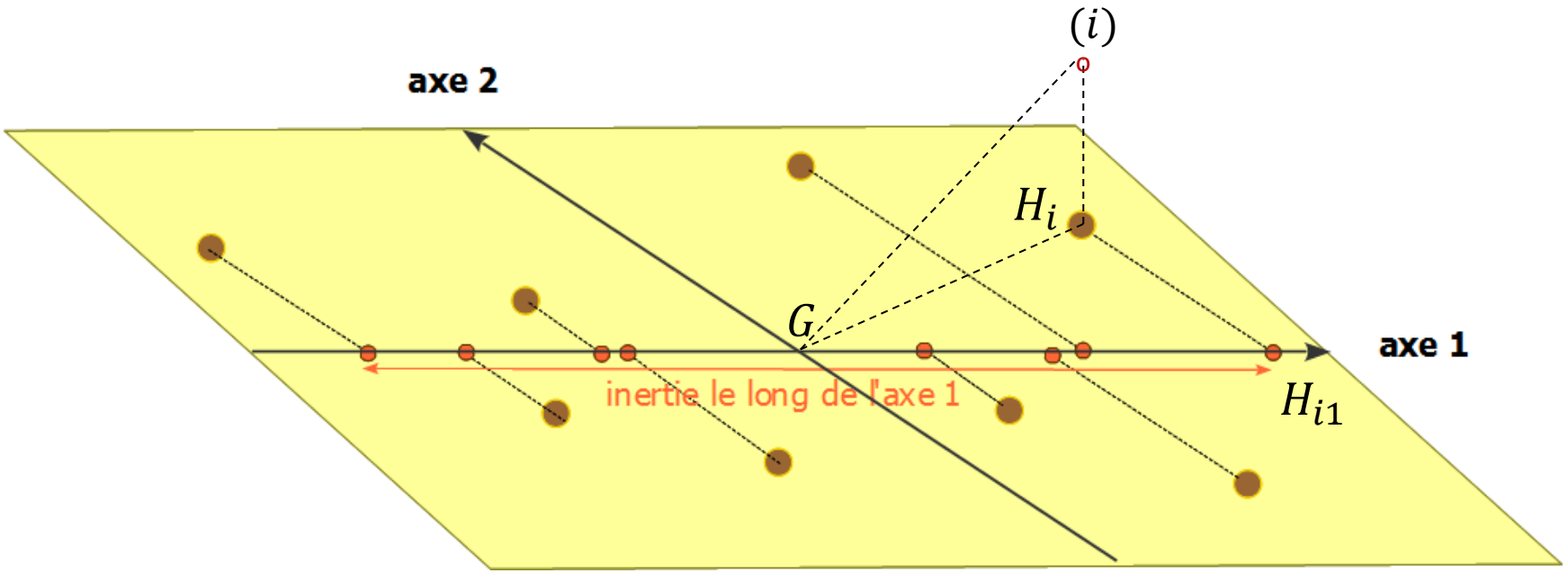


H_i : projection de l'individu (i)
G : centre de gravité du nuage projeté
m_i : masse de l'individu (i)

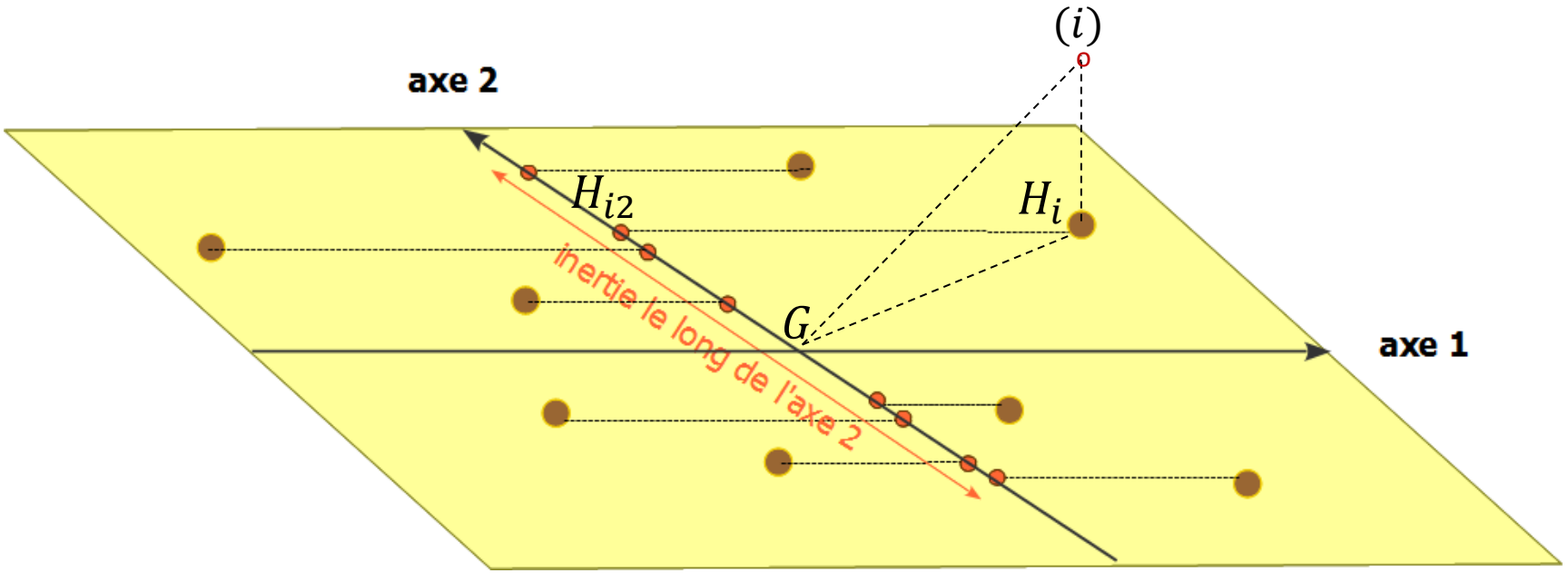
Inertie du nuage projeté

$$I = \sum_{i=1}^{n} m_i \times d^2(G, H_i)$$

Décomposition de l'inertie projetée sur le plan



$$I_1 = \sum_{i=1}^{n} m_i \times d^2(G, H_{i1})$$

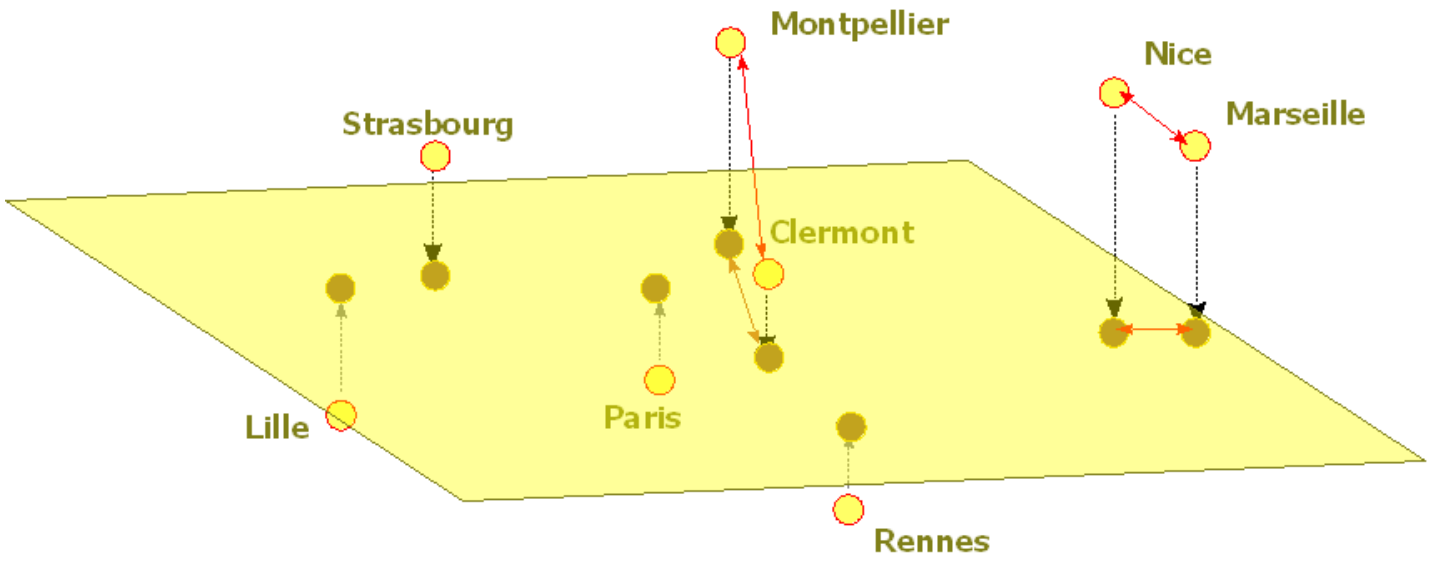


$$I_2 = \sum_{i=1}^{n} m_i \times d^2(G, H_{i2})$$

Inertie du nuage projeté = Inertie le long de l'**axe 1** + Inertie le long de l'**axe 2**

Axe 1 = Axe d'allongement maximal du nuage

Un critère équivalent

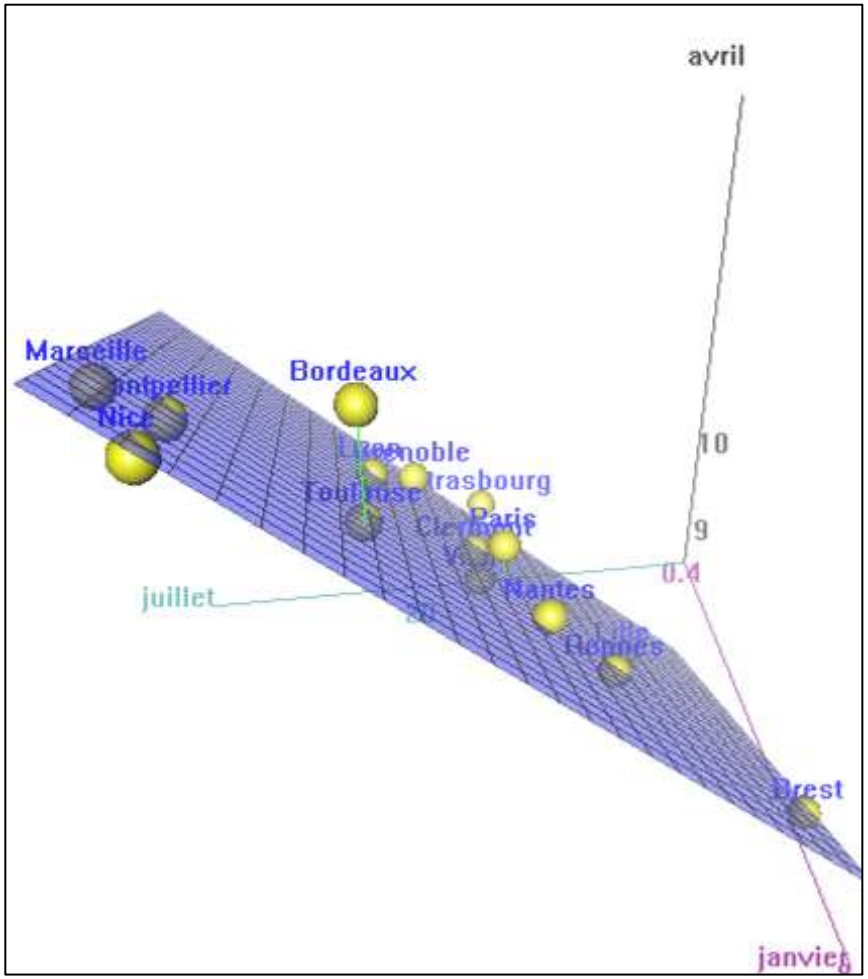


En projection, les distances entre individus sont **les moins « déformées »** possible

= **les plus grandes possible** car la projection réduit toujours les distances

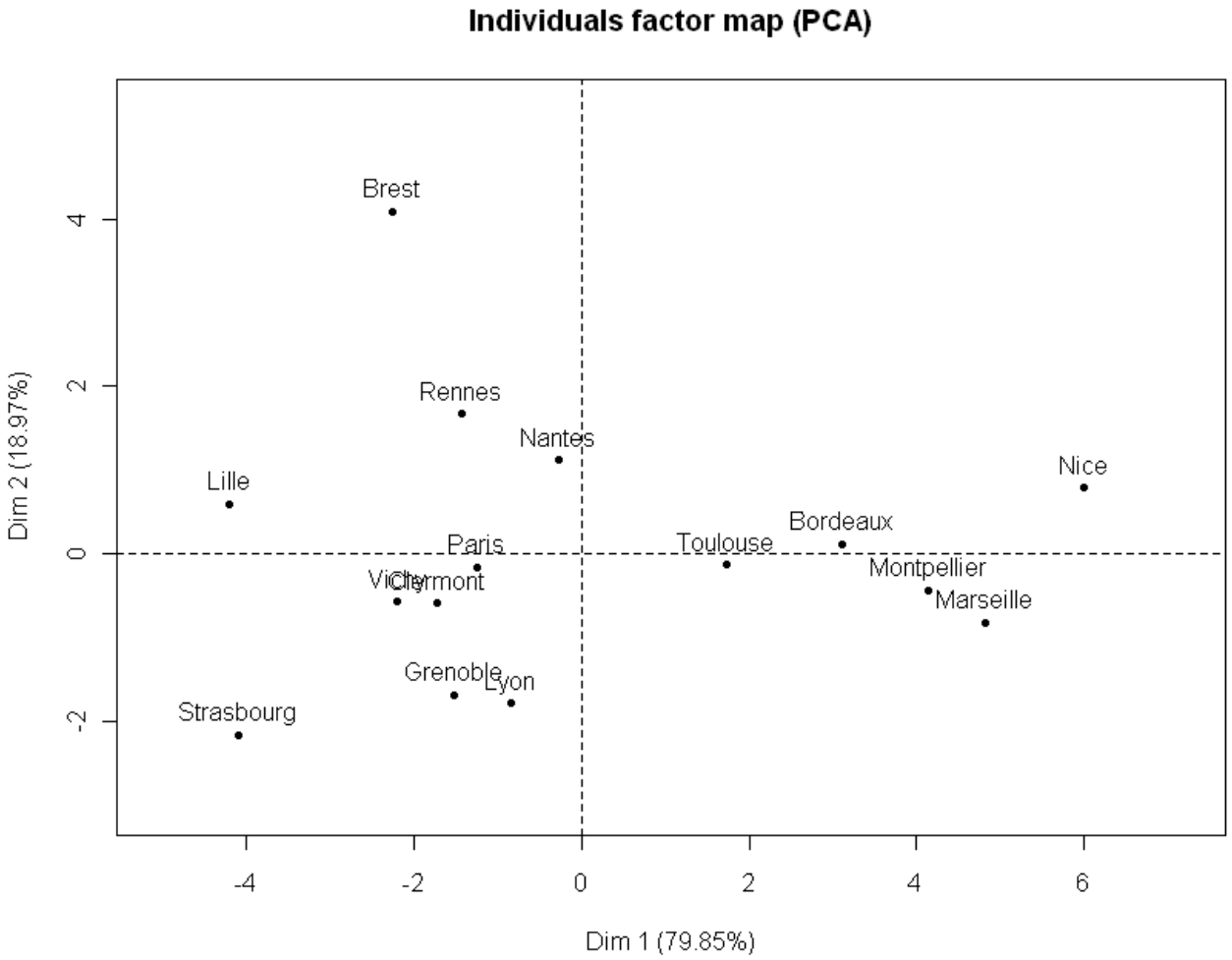
Attention aux **proximités trompeuses !**

Encore un critère équivalent !



Le meilleur plan passe en moyenne **au plus près** de l'ensemble des individus du nuage

Exemple « températures » : meilleur plan de projection des individus



La recherche des axes factoriels

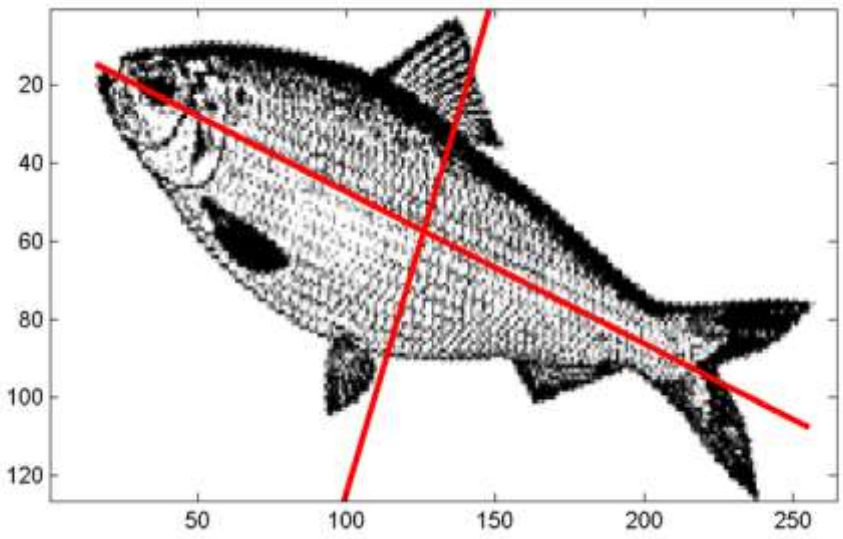
- L'ACP recherche une suite de directions orthogonales 2 à 2
  = **axes factoriels**, dimensions

- **Nombre d'axes factoriels** $= \min(n, p)$
  $p =$ le nombre de variables,
  $n =$ le nombre d'individus

- **Décroissance de l'inertie**
  Plus les axes sont de rang élevé = plus leur inertie est faible

- **Problème mathématique**
  Diagonalisation d'une matrice de variance-covariance ou d'inertie
  (matrice de corrélations pour des données centrées-réduites)
  On extrait :
  - les **vecteurs propres** (directions ou axes factoriels)
  - les **valeurs propres** (inerties associées aux axes)

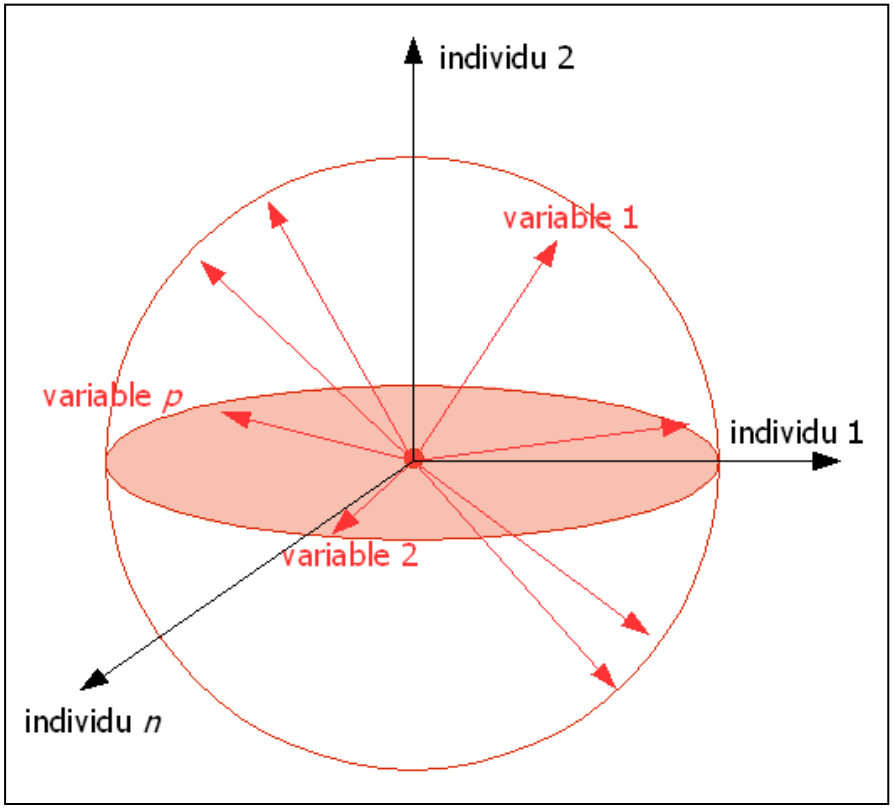
L'ACP d'un nuage à deux dimensions



Les deux axes d'une ACP sur la photo d'un poisson (Wikipédia)

1.4 – Ajustement du nuage des variables

Allure du nuage des variables Les variables constituent **un nuage de points** dans un espace comportant autant de dimensions que d'individus

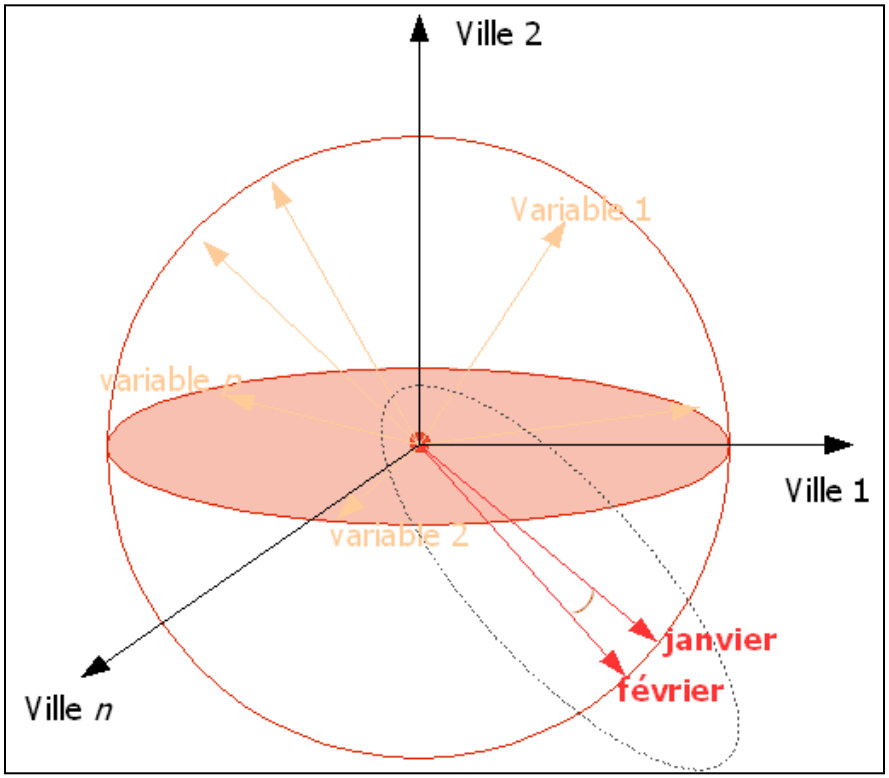


Dans ce nuage

- Les variables sont affectées de masses m_j
- La **longueur** d'une variable x_j est égale à son **écart-type**
- Si les données sont **centrées- réduites** : les variables sont toutes de longueur 1
- Nuage = *hypersphère* de rayon 1

Mesure de liaison entre deux variables

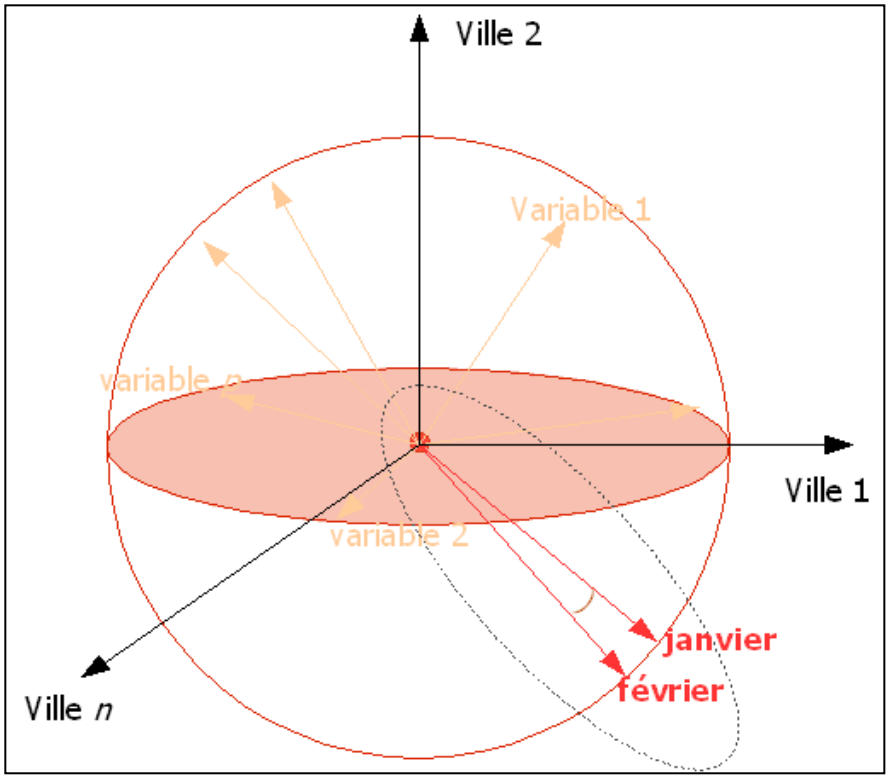
Le **cosinus de l'angle** formé par les deux variables est égale à la **corrélation linéaire** entre les deux variables

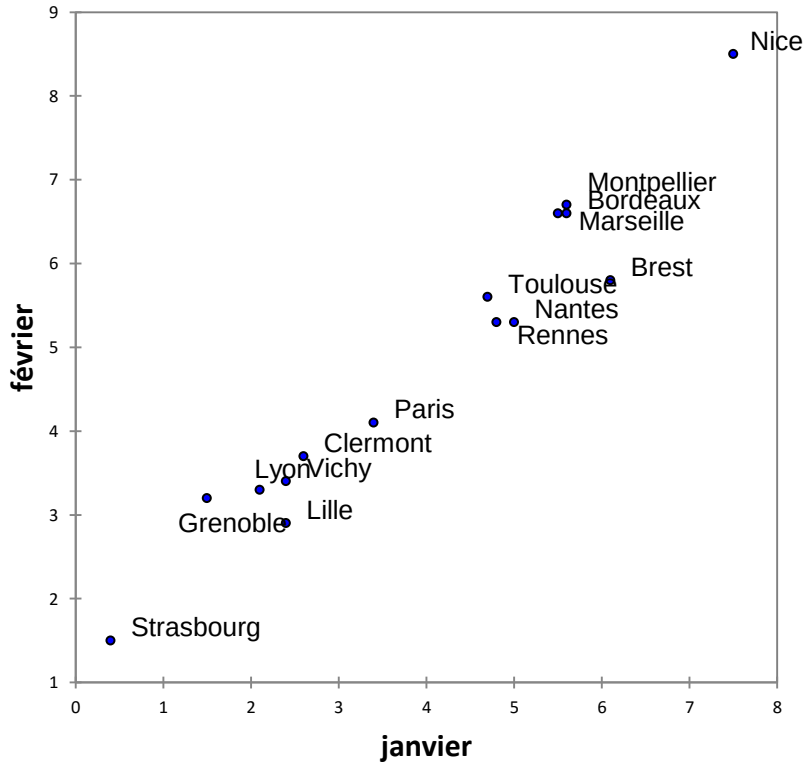


$$r(X_1, X_2) = \cos\left(\widehat{X_1 X_2}\right)$$

L'ACP étudie des liaisons de nature **linéaire**

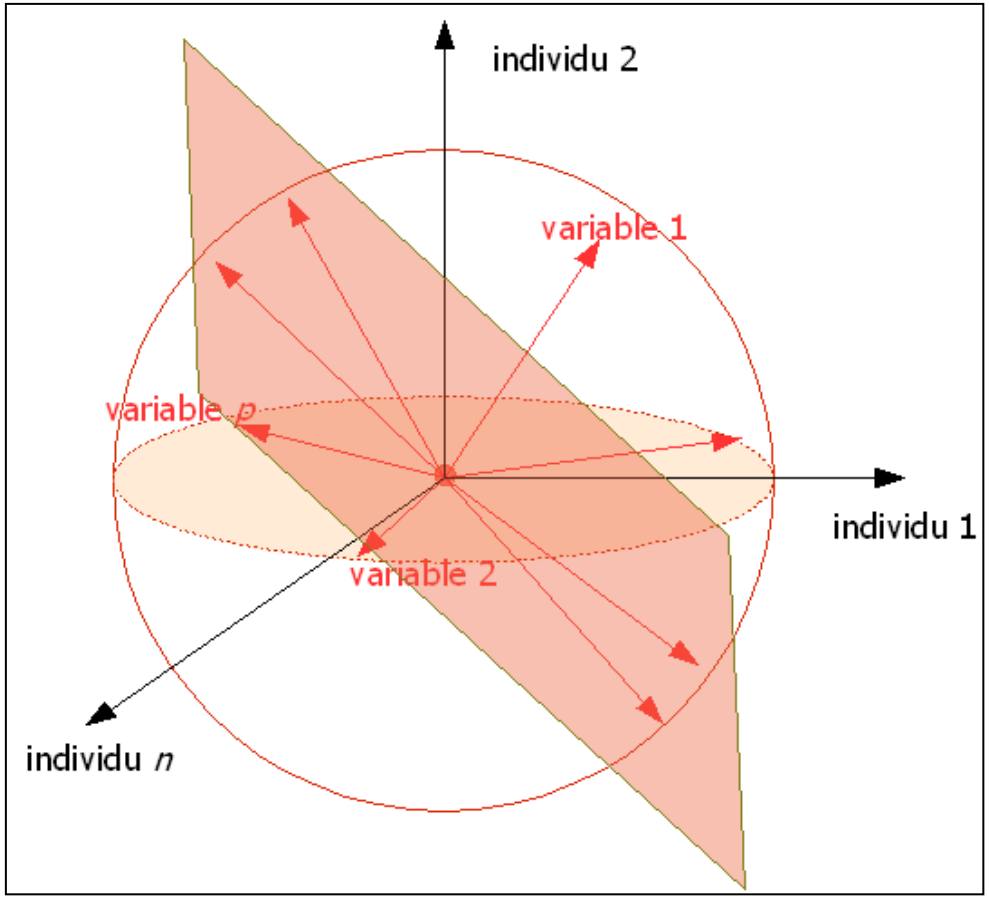
Représentation usuelle d'une corrélation linéaire entre deux variables





Intensité de la liaison linéaire
= **allongement** du nuage de points

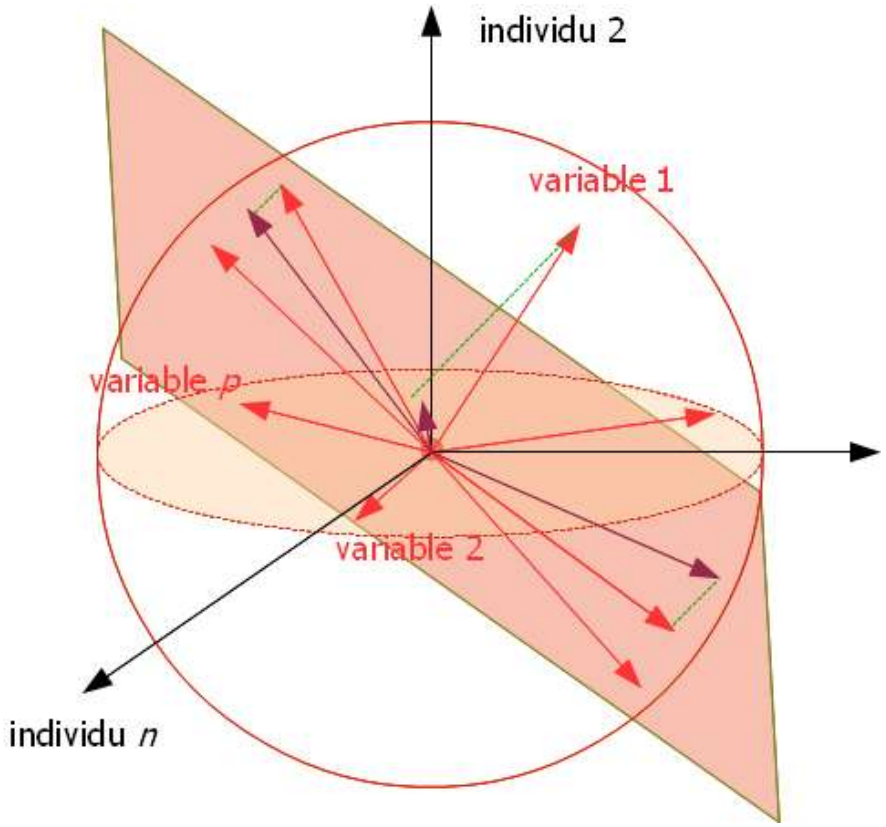
Recherche du meilleur plan de projection



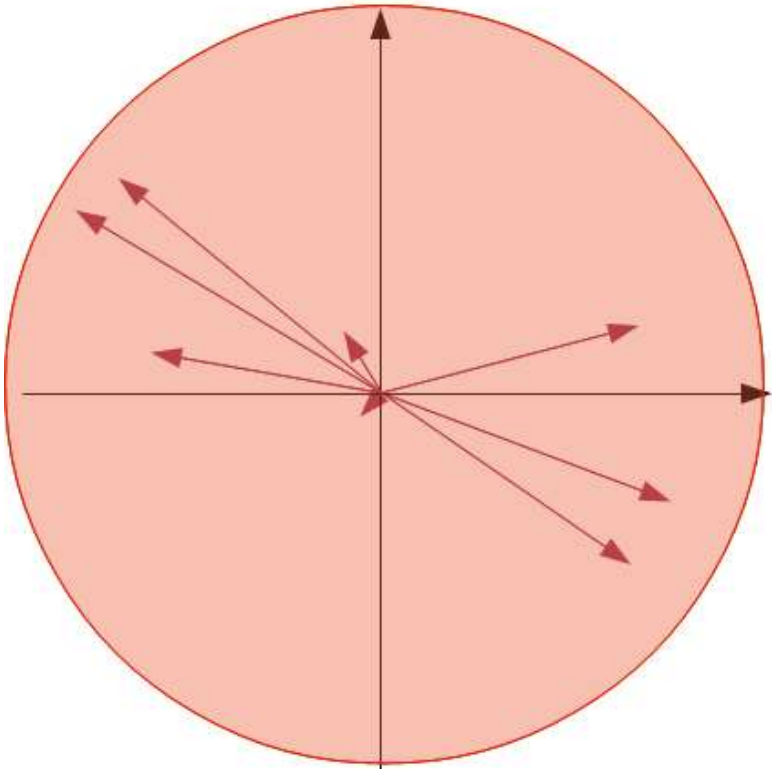
- Critère d'**inertie maximale** en projection,
  ou
- Les **angles** entre vecteurs sont les **moins déformés** possible

Le cercle des corrélations

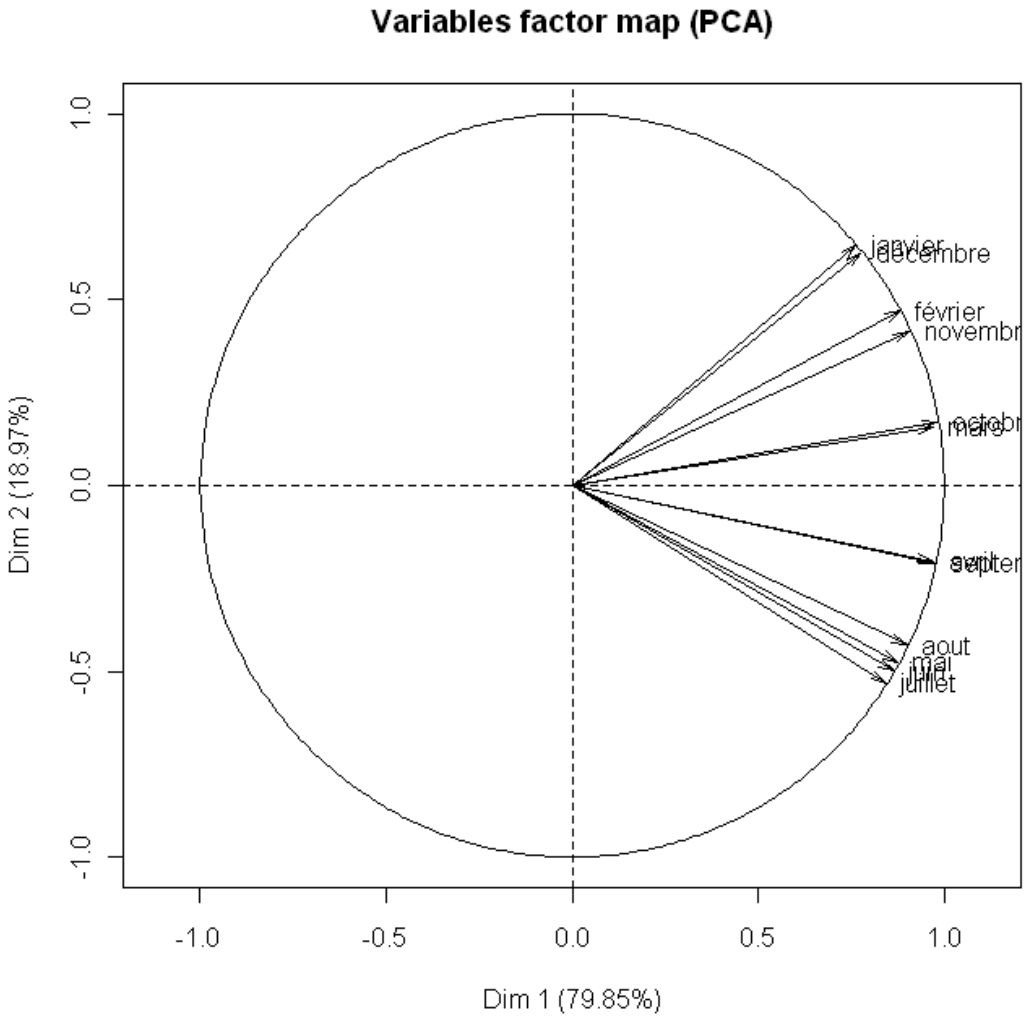
Projection orthogonale des variables



Le cercle des corrélations

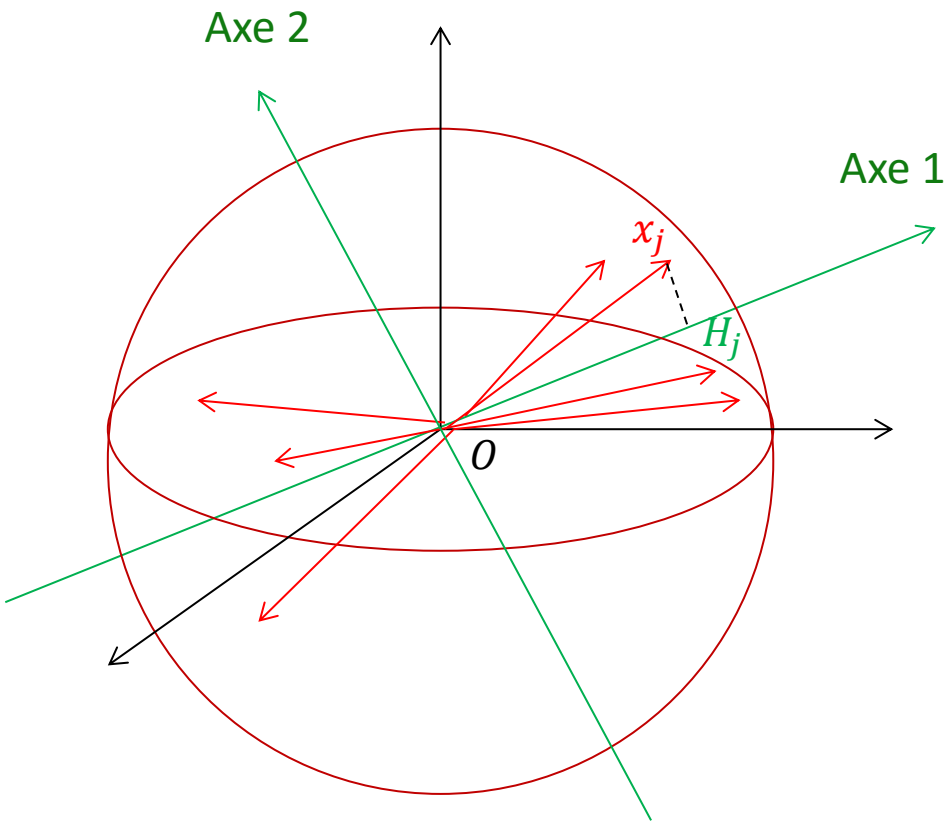


Exemple « températures » : meilleur plan de projection des variables



- Inertie du plan (1,2) =
  inertie de l'axe 1 + inertie de l'axe 2
- Angles entre les variables ?
- Interprétation des axes 1 et 2 ?

Notion de composante principale

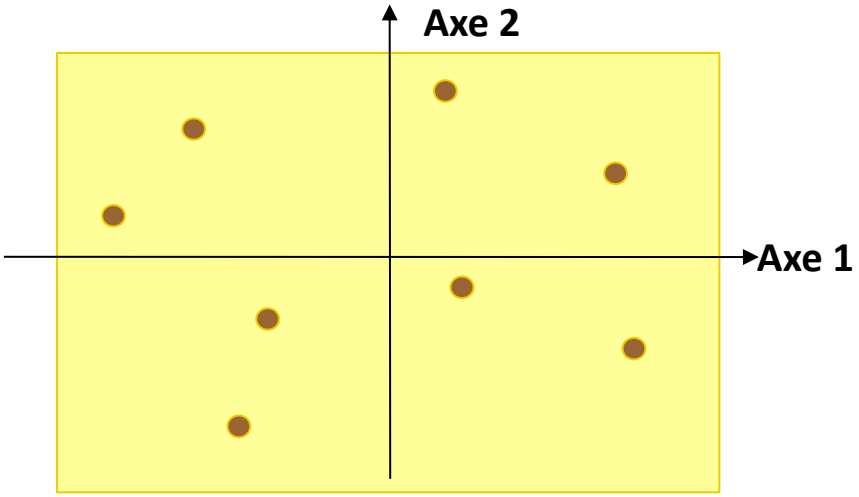


Les variables étant *centrées réduites*

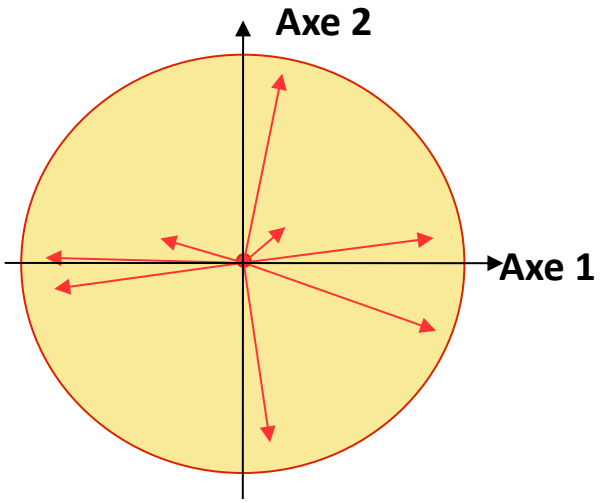
- La **projection** d'une variable sur un axe est égale à leur **corrélation**
- Rechercher le meilleur axe (d'inertie maxi) = rechercher la **combinaison linéaire** la plus liée à l'ensemble des variables
- La variable engendrant le 1er axe est celle qui **synthétise** au mieux les variables initiales
- ACP recherche une suite de variables synthétiques, les **composantes principales**

1.5 – Synthèse entre les deux ajustements

Meilleur plan des individus

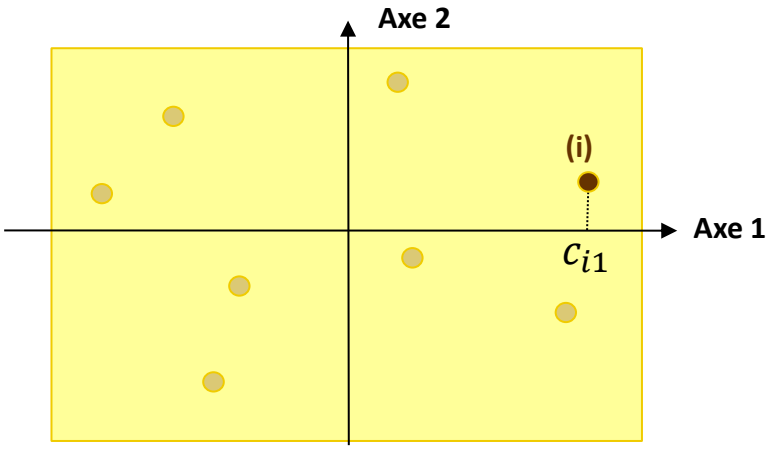


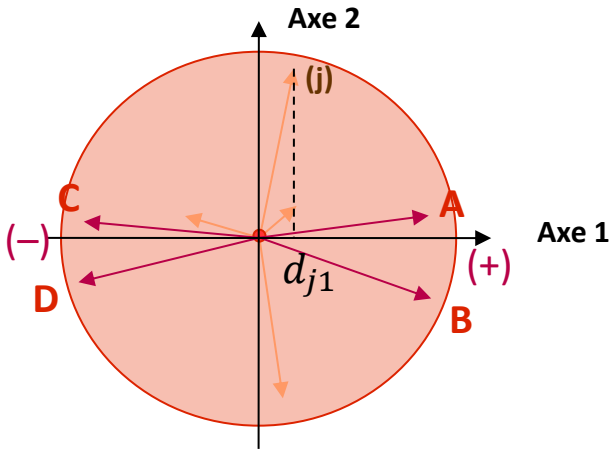
Meilleur plan des variables



- Chaque ajustement fournit **le même nombre d'axes factoriels**
- Ces axes factoriels ont la **même inertie**
- Les **coordonnées des individus** sur un axe sont reliées aux **coordonnées des variables** sur ce même axe (formules de transition)

Il est possible d'**interpréter conjointement** les deux représentations !





$$c_{ik} = \frac{1}{\sqrt{I_k}} \sum_{j=1}^{p} \left(\frac{x_{ij} - \bar{x}_j}{s_j} \right) d_{jk}$$

Formules de transition

$$d_{jk} = \frac{1}{\sqrt{I_k}} \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_{ij} - \bar{x}_j}{s_j} \right) c_{ik}$$

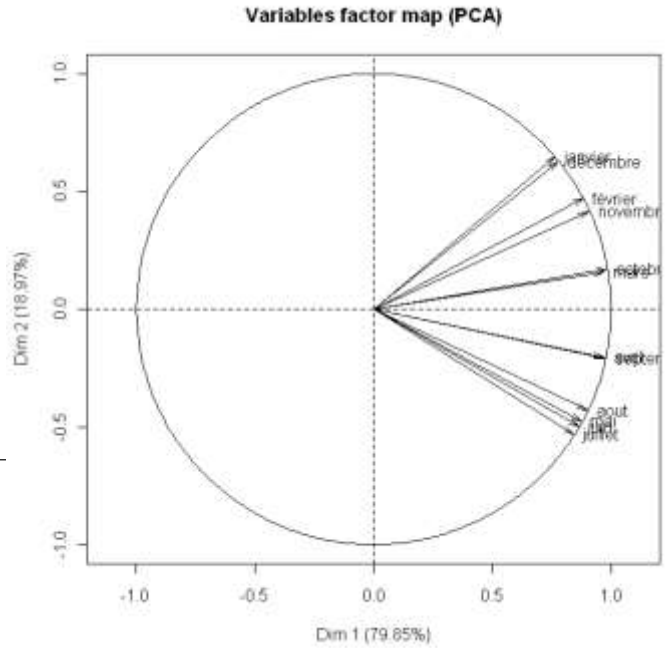
Règle de **lecture directionnelle :** l'individu ***(i)*** prend des valeurs :

- **plus élevées que la moyenne** pour les variables allant dans sa direction et fortement liées à cet axe
- **moins élevées que la moyenne** pour les variables allant en direction opposée et fortement liées à cet axe

(i) prend des valeurs élevées pour **A** et **B**

(i) prend des valeurs faibles pour **C** et **D**

Illustration : exemple « températures »



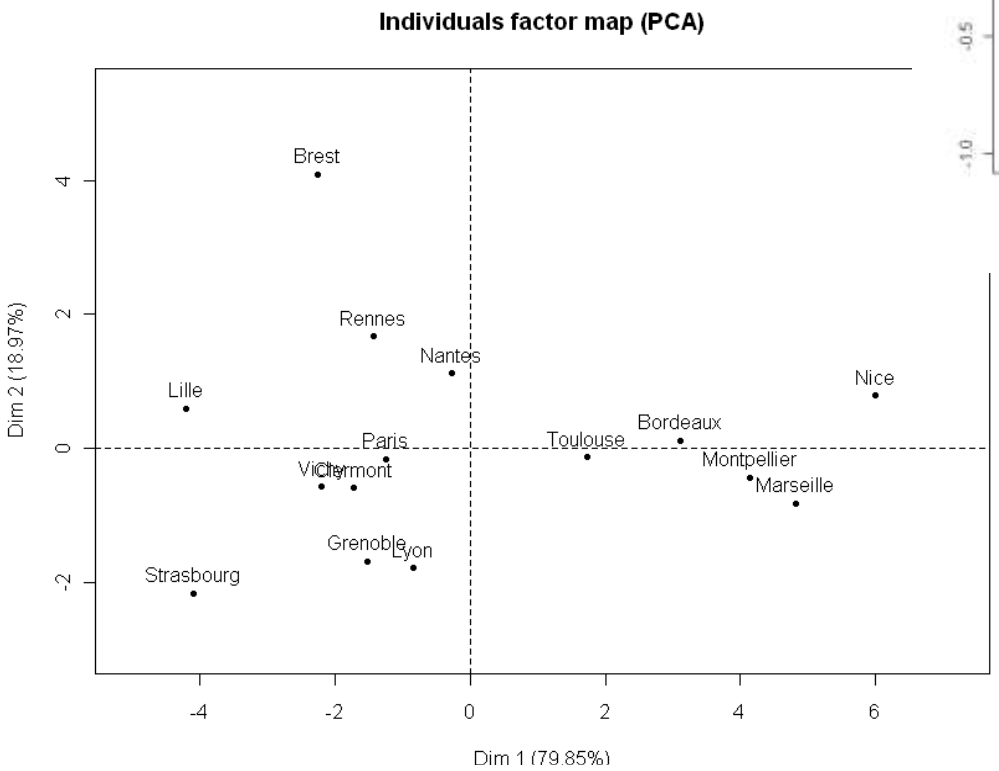
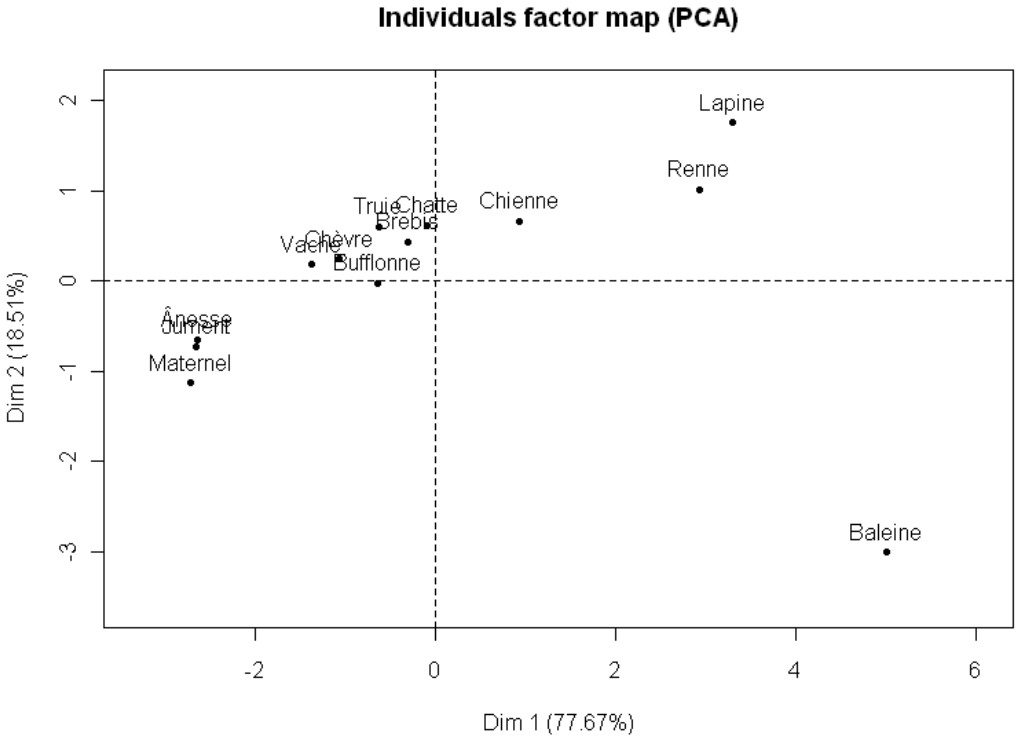
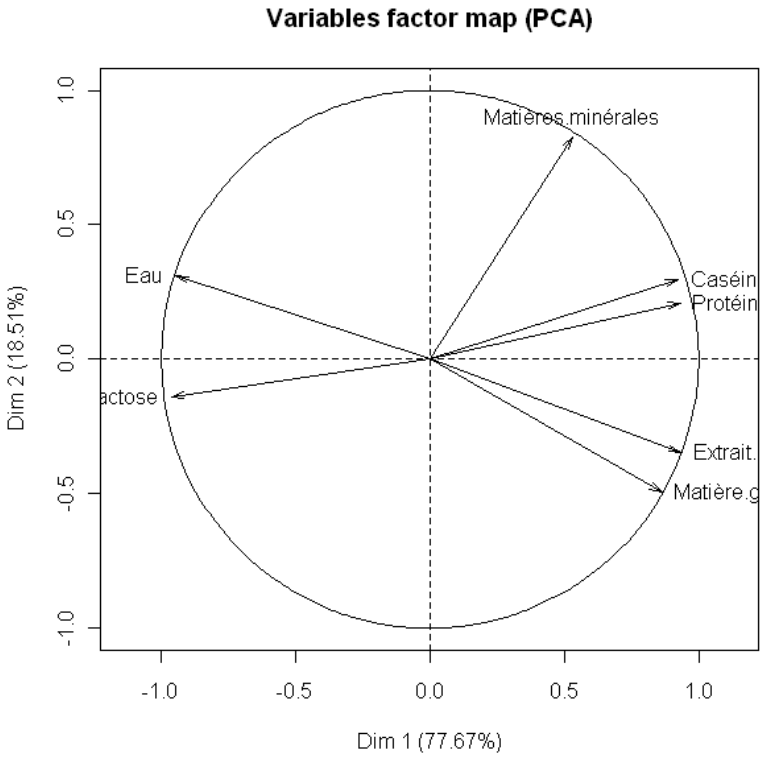


Illustration : exemple « lait »





LaitMami

	Eau	Extrait.sec	Matière.grasse	Protéines.totales	Caséine	Lactose	Matières.minérales
Maternel	905	117	35.0	13.0	11.0	67.0	3.0
Jument	925	100	12.5	21.0	11.0	62.5	4.0
Ânesse	925	100	12.5	21.0	11.0	62.5	4.5
Vache	900	130	37.5	32.5	28.5	47.5	9.0
Chèvre	900	140	42.5	37.5	32.5	42.5	9.0
Brebis	860	190	72.5	57.5	47.5	47.5	11.0
Bufflonne	850	180	72.5	47.5	37.5	47.5	9.0
Renne	675	330	180.0	102.5	82.5	27.5	17.5
Truie	850	185	65.0	57.5	27.5	52.5	13.5
Chienne	800	250	95.0	105.0	47.5	40.0	13.0
Chatte	850	200	45.0	95.0	32.5	45.0	11.5
Lapine	720	300	125.0	135.0	95.0	17.5	17.5
Baleine	467	600	440.0	125.0	70.0	18.0	5.0

Moyenne	817.46	217.07	95.00	65.38	41.07	44.42	9.80
écart-type	129.85	135.93	113.59	42.04	27.01	15.78	4.82

LaitMami

	Z.Caséine	Z.Eau	Z.Extrait.sec	Z.Lactose	Z.Matière.grasse	Z.Matières.minérales	Z.Protéines.totales
Maternel	-1.1132653	0.6741624	-0.7362398	1.43029417	-0.5282066	-1.4099073	-1.2459709
Jument	-1.1132653	0.8281889	-0.8613043	1.14520998	-0.7262840	-1.2028023	-1.0556905
Ânesse	-1.1132653	0.8281889	-0.8613043	1.14520998	-0.7262840	-1.0992498	-1.0556905
Vache	-0.4655214	0.6356557	-0.6406022	0.19492936	-0.5061980	-0.1672771	-0.7821624
Chèvre	-0.3174657	0.6356557	-0.5670348	-0.12183085	-0.4621807	-0.1672771	-0.6632371
Brebis	0.2377434	0.3276026	-0.1991978	0.19492936	-0.1980775	0.2469329	-0.1875360
Bufflonne	-0.1323960	0.2505894	-0.2727652	0.19492936	-0.1980775	-0.1672771	-0.4253865
Renne	1.5332311	-1.0971430	0.8307456	-1.07211147	0.7482926	1.5931156	0.8827914
Truie	-0.5025354	0.2505894	-0.2359815	0.51168957	-0.2641033	0.7646955	-0.1875360
Chienne	0.2377434	-0.1344770	0.2422065	-0.28021095	0.0000000	0.6611430	0.9422541
Chatte	-0.3174657	0.2505894	-0.1256305	0.03654925	-0.4401721	0.3504854	0.7044035
Lapine	1.9959053	-0.7505832	0.6100434	-1.70563189	0.2641033	1.5931156	1.6558057
Baleine	1.0705569	-2.6990191	2.8170650	-1.67395587	3.0371878	-0.9956973	1.4179552

Notion de composante principale

- Un axe factoriel une nouvelle **variable « synthétique »** appelée composante principale
- Une composante principale est construite comme une **combinaison linéaire** des variables initiales

$$CP1 = 0.76\ \textit{Janvier} + 0.88\ \textit{Février} + \ldots + 0.77\ \textit{Décembre}$$

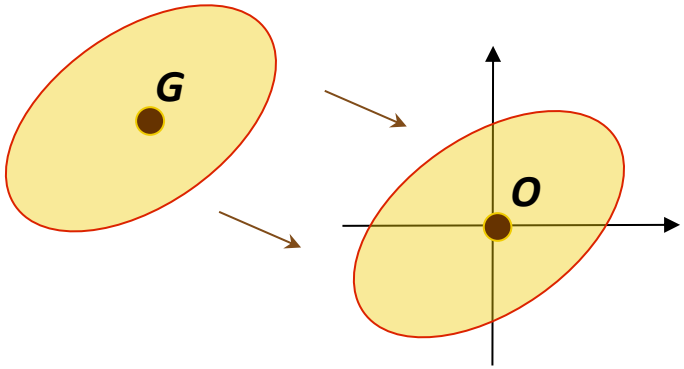
- Une composante principale est un vecteur renfermant les **coordonnées des individus** le long d'un axe donné

1.6 – ACP normée ou non normée ?

Une ACP normée = réalisée sur des données **centrées - réduites**

Le centrage

Il est réalisé de façon **systématique** en ACP
Translation du centre de gravité du nuage sur l'origine



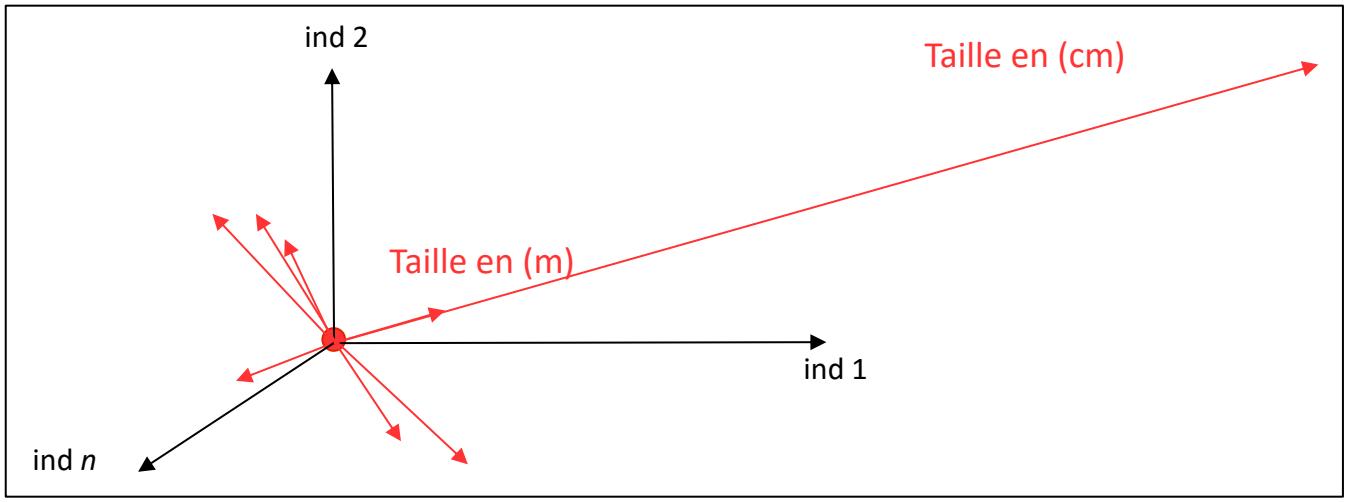
Ville	janvier	Janvier Centré
Bordeaux	5,60	1,63
Brest	6,10	2,13
Clermont	2,60	-1,37
Grenoble	1,50	-2,47
Lille	2,40	-1,57
Lyon	2,10	-1,87
Marseille	5,50	1,53
Montpellier	5,60	1,63
Nantes	5,00	1,03
Nice	7,50	3,53
Paris	3,40	-0,57
Rennes	4,80	0,83
Strasbourg	0,40	-3,57
Toulouse	4,70	0,73
Vichy	2,40	-1,57
moyenne	*3,97*	*0*
écart-type	*1,94*	*1,94*

La réduction

- Nuage des variables : longueur d'une variable est égale à son écart-type
- Plus la variable a un **écart-type élevé**, plus elle apporte de l'inertie en projection plus elle « **attire les axes** »
- Or, l'écart type dépend directement de l'unité de mesure...

Ville	janvier	Janvier Centré	Janvier CR
Bordeaux	5,60	1,63	0,84
Brest	6,10	2,13	1,1
Clermont	2,60	-1,37	-0,71
Grenoble	1,50	-2,47	-1,28
Lille	2,40	-1,57	-0,81
Lyon	2,10	-1,87	-0,97
Marseille	5,50	1,53	0,79
Montpellier	5,60	1,63	0,84
Nantes	5,00	1,03	0,53
Nice	7,50	3,53	1,82
Paris	3,40	-0,57	-0,3
Rennes	4,80	0,83	0,43
Strasbourg	0,40	-3,57	-1,84
Toulouse	4,70	0,73	0,37
Vichy	2,40	-1,57	-0,81
moyenne	*3,97*	*0*	*0*
écart-type	*1,94*	*1,94*	*1*

Illustration



- Pour éviter d'accorder une plus grande importance aux variables exprimées arbitrairement avec de plus grandes valeurs, on réduit les variables
- Chaque variable a le même écart-type = 1 (donc la même longueur)

Lorsque les variables sont exprimées dans des **unités de mesure différentes**,

→ **réduction systématique** des données

En cas d'unités de mesure identiques ?

- Pas de règle systématique
- Cela se discute au cas par cas

Ville	janvier	février	mars	avril	mai	juin	juillet	août	septembre	octobre	novembre	décembre
Bordeaux	5,80	6,80	10,30	12,80	15,80	19,30	20,90	21,00	18,80	13,80	9,10	6,20
Brest	6,10	5,80	7,80	9,20	11,80	14,40	15,80	16,00	14,70	12,00	9,00	7,00
Clermont	2,80	3,70	7,50	10,30	13,80	17,30	19,40	19,10	16,20	11,20	6,80	3,80
Grenoble	1,50	3,20	7,70	10,80	14,50	17,80	20,10	19,50	16,70	11,40	6,50	2,30
Lille	2,40	2,90	6,00	8,90	12,40	15,30	17,10	17,10	14,70	10,40	6,10	3,50
Lyon	2,10	3,30	7,70	10,90	14,90	18,50	20,70	20,10	16,90	11,40	6,70	3,10
Marseille	5,50	6,80	10,00	13,00	16,80	20,80	23,30	22,80	19,90	15,00	10,20	6,90
Montpellier	5,80	6,70	9,90	12,80	16,20	20,10	22,70	22,30	19,30	14,80	10,00	6,50
Nantes	5,00	5,30	8,40	10,80	13,90	17,20	18,80	18,80	16,40	12,20	8,20	5,50
Nice	7,50	8,50	10,80	13,30	16,70	20,10	22,70	22,50	20,30	16,00	11,50	8,20
Paris	3,40	4,10	7,80	10,70	14,30	17,50	19,10	18,70	16,00	11,40	7,10	4,30
Rennes	4,80	5,30	7,90	10,10	13,10	16,20	17,90	17,80	15,70	11,80	7,80	5,40
Strasbourg	0,40	1,50	5,80	9,80	14,00	17,20	19,00	18,30	15,10	9,50	4,90	1,30
Toulouse	4,70	5,80	9,20	11,80	14,90	18,70	20,90	20,90	18,30	13,30	8,80	5,50
Vichy	2,40	3,40	7,10	9,90	13,80	17,10	19,30	18,80	16,00	11,00	6,80	3,40
écart-type	1,94	1,81	1,48	1,37	1,45	1,73	2,08	1,94	1,79	1,77	1,74	1,89

Réduction : consiste à accorder une même importance à chaque variable

Non réduction : accorde plus d'importance aux variables de forte dispersion

Bilan sur la réduction des données en ACP

- **Réduire** ou **normer** donne la même dispersion, une même importance, à chaque variable (dans l'espace, elles ont même longueur : 1)

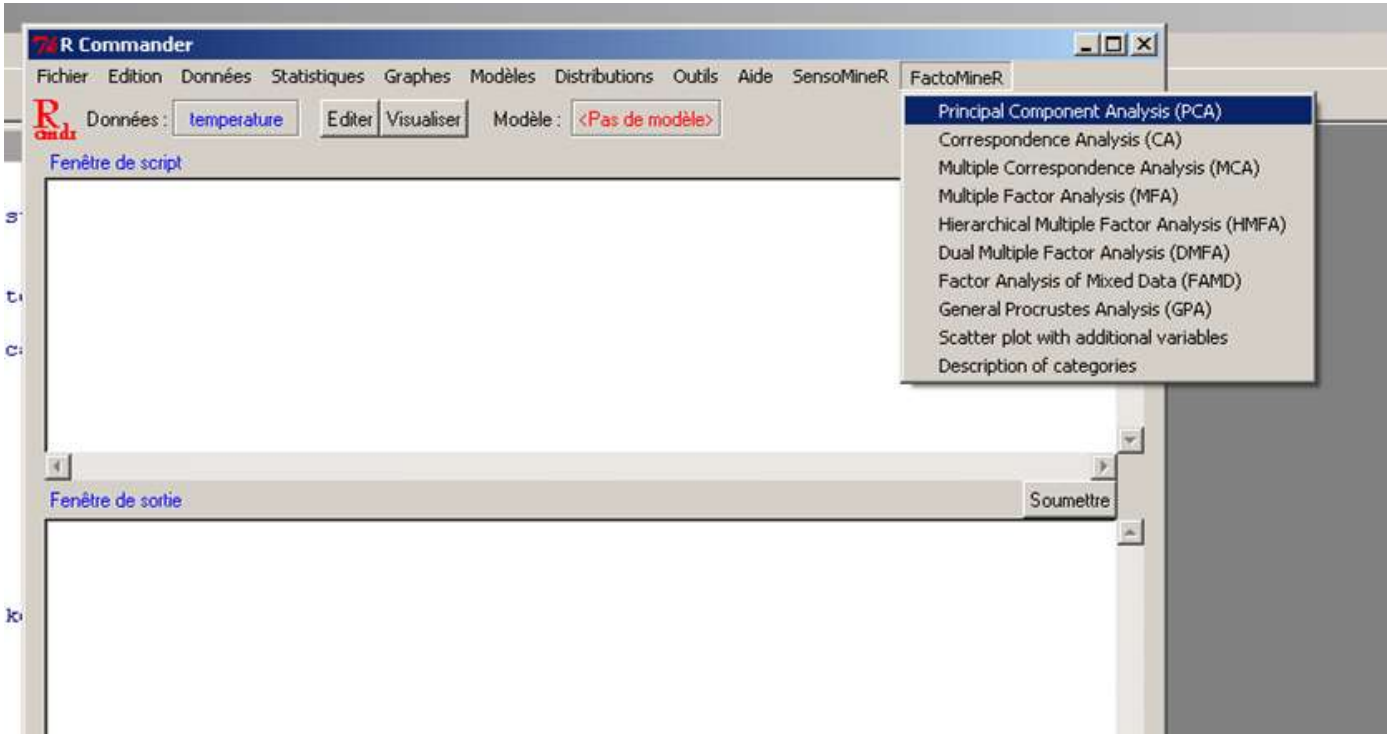
  On dit que l'on réalise une **ACP normée**

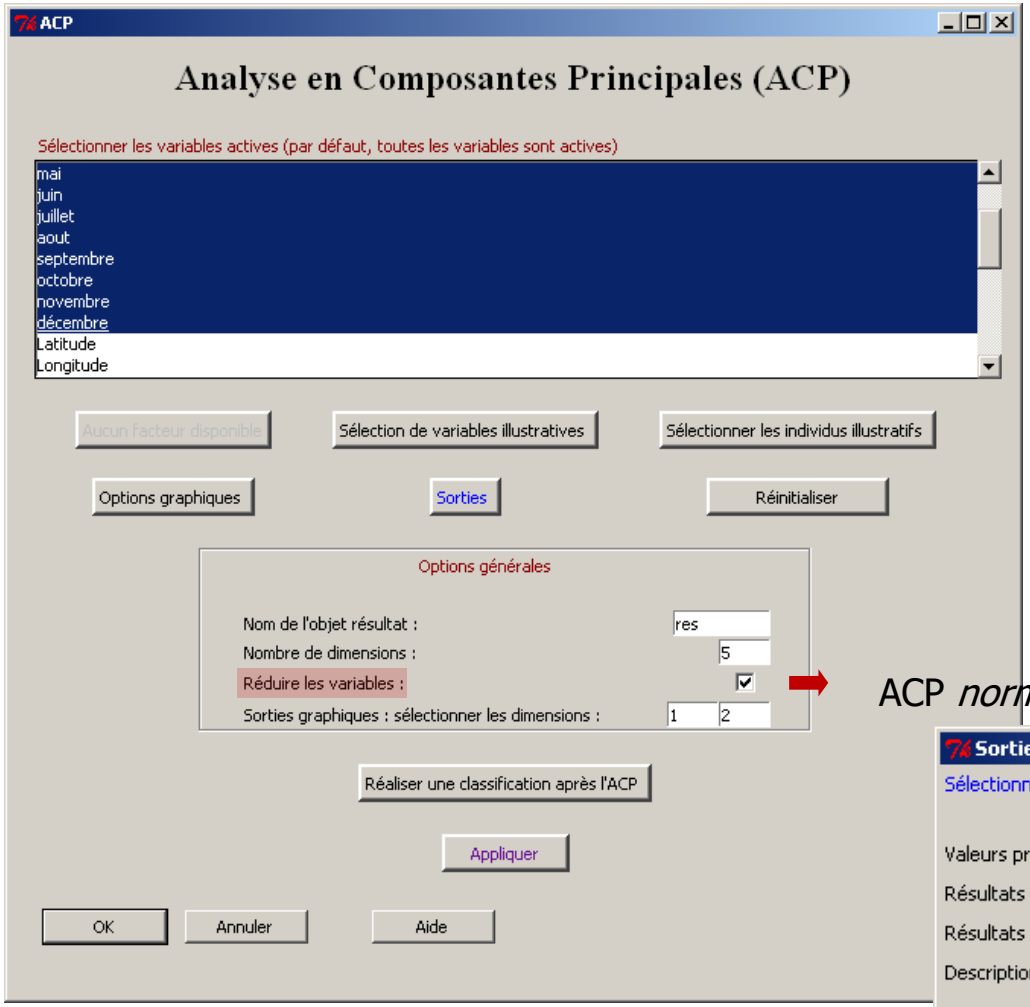
- **Ne pas réduire** ou **ne pas normer** laisse à chaque variable son écart-type initial
  Conduit à accorder à chaque variable une importance proportionnelle à son écart-type

  On dit que l'on réalise une **ACP non normée**

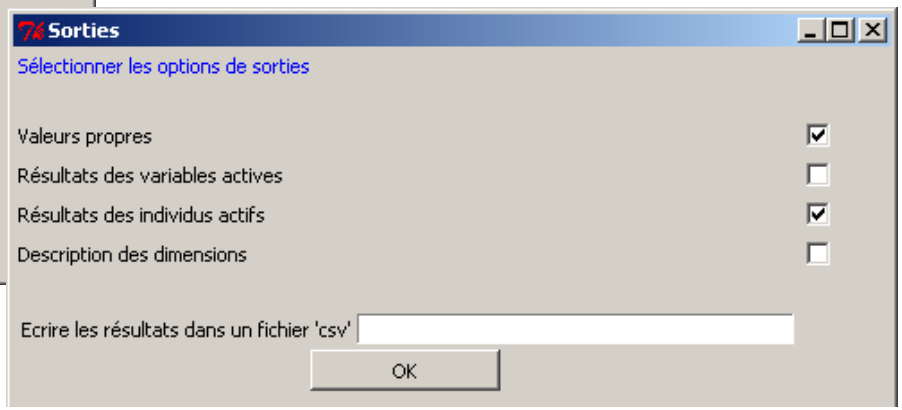
1.7 – Paramétrage avec FactoMineR

Utilisation du package **FactoMineR** intégré à *Rcmdr* (logiciel R)





ACP *normée* ou *non normée*



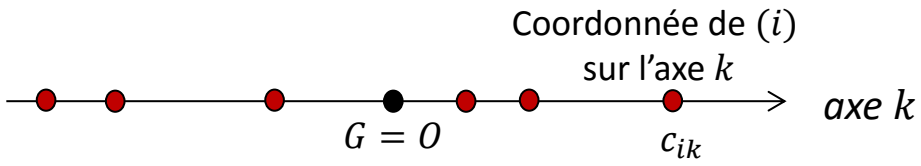
1.8 – Analyse de l'inertie des axes factoriels

Le tableau des inerties (ou valeurs propres)

```
> res$eig
```

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	9.581780e+00	7.984816e+01	79.84816
comp 2	2.276418e+00	1.897015e+01	98.81832
comp 3	7.001440e-02	5.834534e-01	99.40177
comp 4	3.967473e-02	3.306228e-01	99.73239
comp 5	1.404529e-02	1.170441e-01	99.84944
comp 6	7.981537e-03	6.651281e-02	99.91595
comp 7	6.049255e-03	5.041046e-02	99.96636
comp 8	1.746893e-03	1.455744e-02	99.98092
comp 9	1.492178e-03	1.243482e-02	99.99335
comp 10	4.921334e-04	4.101112e-03	99.99745
comp 11	2.858357e-04	2.381964e-03	99.99984
comp 12	1.976210e-05	1.646842e-04	100.00000

Inertie d'un axe = mesure de dispersion *des individus le long de l'axe*



$$I_k = \sum_{i=1}^{n} m_i \times {c_{ik}}^2$$

Inertie = *variance des coordonnées !*

- eigenvalue (valeur propre) = Inertie brute
- percentage of variance = Pourcentage d'inertie
- percentage of variance cumulative = Pourcentage cumulé
- comp 1 = composante 1 (« Axe 1 », « dimension 1 »)

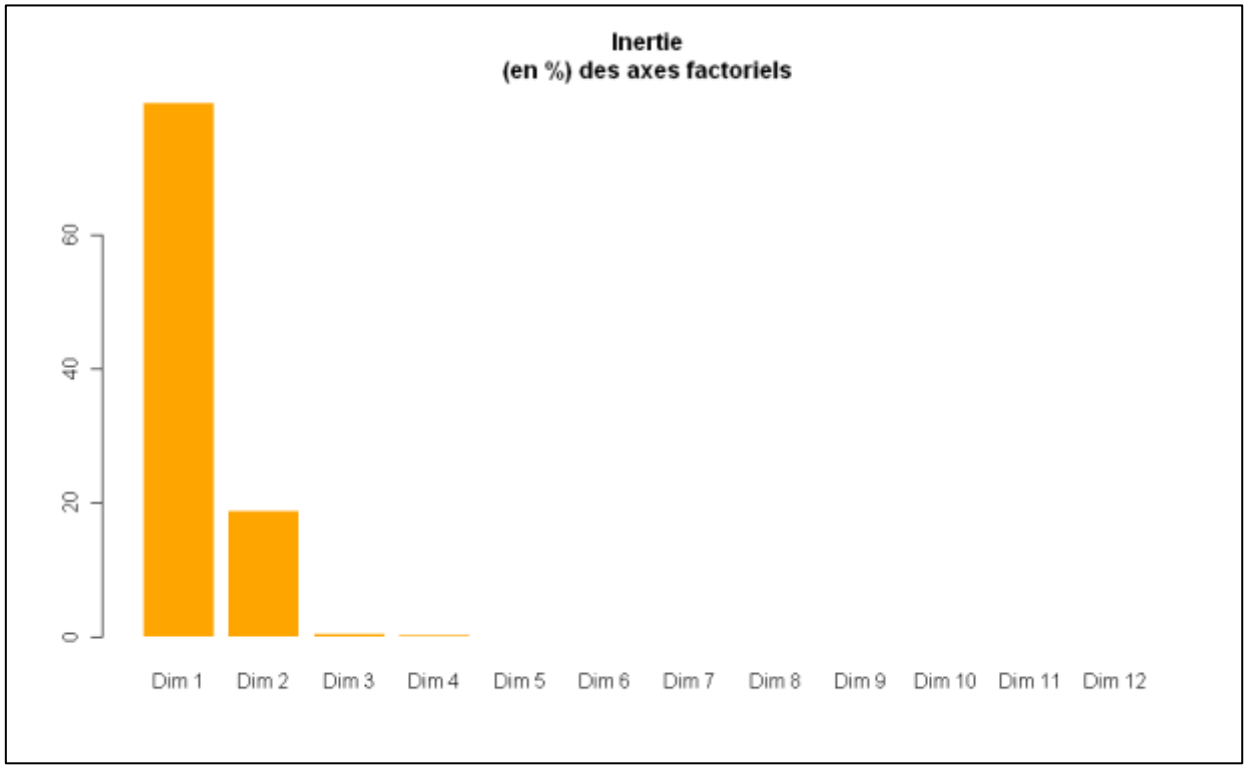
```
> res$eig
```

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	9.581780e+00	7.984816e+01	79.84816
comp 2	2.276418e+00	1.897015e+01	98.81832
comp 3	7.001440e-02	5.834534e-01	99.40177
comp 4	3.967473e-02	3.306228e-01	99.73239
comp 5	1.404529e-02	1.170441e-01	99.84944
comp 6	7.981537e-03	6.651281e-02	99.91595
comp 7	6.049255e-03	5.041046e-02	99.96636
comp 8	1.746893e-03	1.455744e-02	99.98092
comp 9	1.492178e-03	1.243482e-02	99.99335
comp 10	4.921334e-04	4.101112e-03	99.99745
comp 11	2.858357e-04	2.381964e-03	99.99984
comp 12	1.976210e-05	1.646842e-04	100.00000

Somme =	12.00

- Inertie totale du tableau de données = somme des valeurs propres = somme des variances des X_j
- ACP <u>normée</u> : inertie totale = nombre de variables

Diagramme de décroissance de l'inertie : « éboulis » des valeurs propres



Commande R

```
barplot (res$eig[,2], names=paste("Dim",1 :length(res$eig[,2])),
main="Inertie (en %) des axes factoriels", col="orange", border="white")
```

Combien d'axes retenir ?

- La quantité d'inertie d'un axe traduit l'intérêt de celui-ci
- Seules les premières dimensions sont en général intéressantes à interpréter
- Combien d'axes (de dimensions) est-il pertinent de retenir?

Le critère de Kaiser : l'inertie moyenne

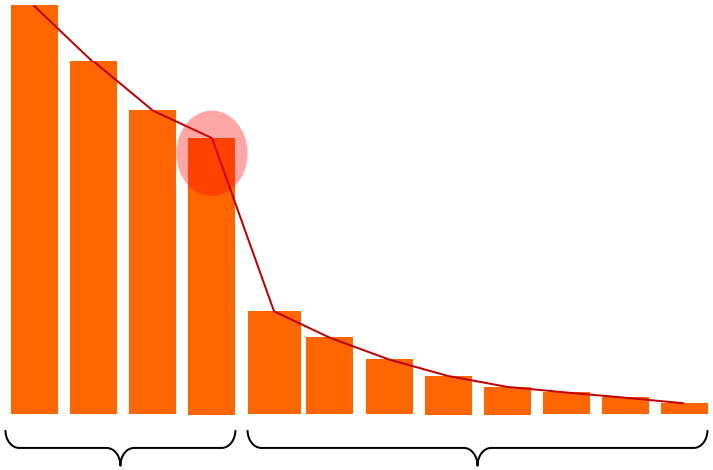
On retient les axes dont l'inertie est supérieure à l'inertie moyenne
Application très simple en ACP normée !

$$\text{Inertie moyenne} = \frac{\text{Inertie totale}}{\text{nombre d'axes}} = 1$$

On ne retient que les axes d'inertie supérieure à 1

Application...

Scree – test de Cattell : la recherche d'un **coude**



On repère, dans l'éboulis des valeurs propres, une rupture de pente, un coude

On conserve les axes d'inertie situés *avant le coude*

Une version numérique

Calcul des différences premières $\varepsilon_k = (\lambda_k - \lambda_{k+1})$

Calcul des différences secondes $\delta_k = (\varepsilon_k - \varepsilon_{k+1})$

Lorsque $\delta_k < 0$: on retient les valeurs propres $\lambda_1, \ldots, \lambda_{k+1}$

Qu'obtiendrait-on dans une situation purement aléatoire ?

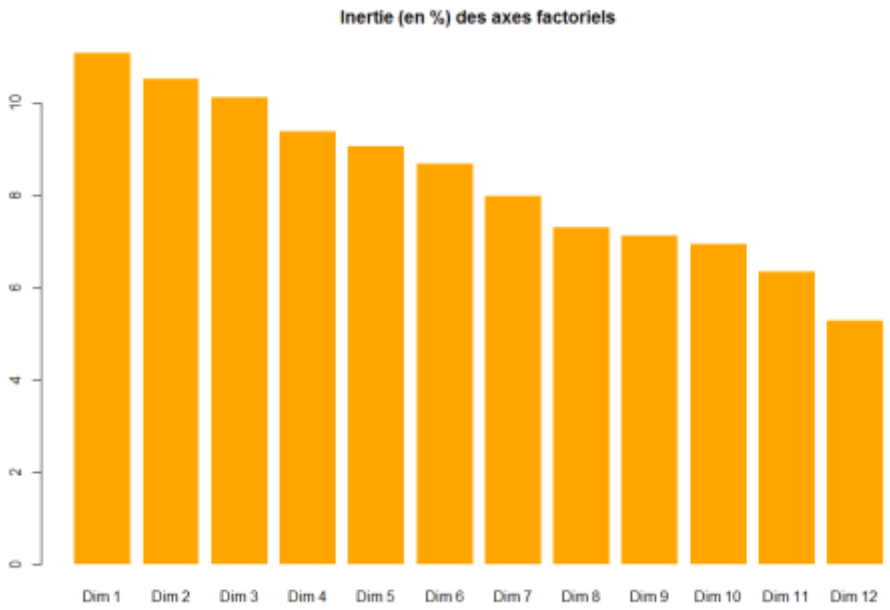


Tableau de données aléatoires pour 12 variables

Aucune structure n'apparaît
Pas de rupture de pente
Typique d'une variabilité aléatoire

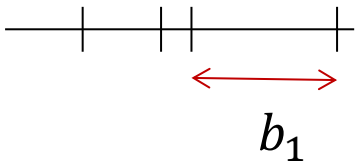
Le modèle du « bâton brisé » : *broken stick*

On ne retient que les axes dont le pourcentage d'inertie est supérieure à celui donné par le **modèle aléatoire** du bâton brisé (*Frontier*, 1976)

p =	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	75,00	61,11	52,08	45,67	40,83	37,04	33,97	31,43	29,29	27,45	25,86	24,46	23,23	22,12	21,13	20,23	19,42	18,67	17,99
	25,00	27,78	27,08	25,67	24,17	22,76	21,47	20,32	19,29	18,36	17,53	16,77	16,08	15,45	14,88	14,35	13,86	13,41	12,99
		11,11	14,58	15,67	15,83	15,61	15,22	14,77	14,29	13,82	13,36	12,92	12,51	12,12	11,75	11,41	11,08	10,78	10,49
			6,25	9,00	10,68	10,85	11,06	11,06	10,96	10,79	10,58	10,36	10,13	9,90	9,67	9,45	9,23	9,02	8,82
				4,00	6,11	7,28	7,93	8,28	8,46	8,51	8,50	8,44	8,34	8,23	8,11	7,98	7,84	7,71	7,57
					2,78	4,42	5,43	6,06	6,46	6,70	6,83	6,90	6,92	6,90	6,86	6,80	6,73	6,65	6,57
						2,04	3,35	4,21	4,79	5,18	5,44	5,62	5,73	5,79	5,82	5,82	5,81	5,78	5,74
							1,56	2,62	3,36	3,88	4,25	4,52	4,71	4,84	4,92	4,98	5,01	5,03	5,02
								1,23	2,11	2,75	3,21	3,56	3,81	4,00	4,14	4,25	4,32	4,37	4,40
									1,00	1,74	2,29	2,70	3,02	3,26	3,45	3,59	3,70	3,78	3,84
										0,83	1,45	1,93	2,30	2,60	2,82	3,00	3,15	3,26	3,34
											0,69	1,23	1,65	1,99	2,26	2,47	2,64	2,78	2,89
												0,59	1,06	1,43	1,73	1,98	2,18	2,34	2,47
													0,51	0,92	1,25	1,53	1,75	1,93	2,09
														0,44	0,81	1,11	1,35	1,56	1,73
															0,39	0,71	0,98	1,21	1,40
																0,35	0,64	0,88	1,09
																	0,31	0,57	0,79
																		0,28	0,51
																			0,25

Frontier 1976

Taille moyenne attendue (en pourcentage de la taille totale) pour le plus grand morceau d'un bâton brisé aléatoirement en 5 morceaux



$$b_k = \frac{1}{p}\sum_{i=k}^{p}\frac{1}{i}$$

Évaluer le pourcentage d'inertie des axes 1 et 2

(d'après Husson F.)

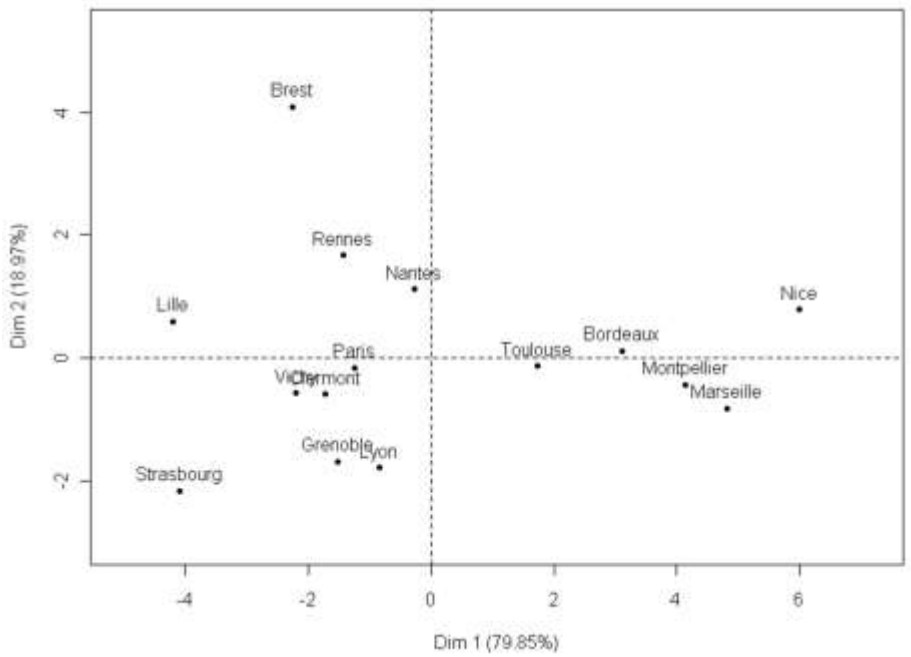
	Nombre de variables												
nbind	4	5	6	7	8	9	10	11	12	13	14	15	16
5	96.5	93.1	90.2	87.6	85.5	83.4	81.9	80.7	79.4	78.1	77.4	76.6	75.5
6	93.3	88.6	84.8	81.5	79.1	76.9	75.1	73.2	72.2	70.8	69.8	68.7	68.0
7	90.5	84.9	80.9	77.4	74.4	72.0	70.1	68.3	67.0	65.3	64.3	63.2	62.2
8	88.1	82.3	77.2	73.8	70.7	68.2	66.1	64.0	62.8	61.2	60.0	59.0	58.0
9	86.1	79.5	74.8	70.7	67.4	65.1	62.9	61.1	59.4	57.9	56.5	55.4	54.3
10	84.5	77.5	72.3	68.2	65.0	62.4	60.1	58.3	56.5	55.1	53.7	52.5	51.5
11	82.8	75.7	70.3	66.3	62.9	60.1	58.0	56.0	54.4	52.7	51.3	50.1	49.2
12	81.5	74.0	68.6	64.4	61.2	58.3	55.8	54.0	52.4	50.9	49.3	48.2	47.2
13	80.0	72.5	67.2	62.9	59.4	56.7	54.4	52.2	50.5	48.9	47.7	46.6	45.4
14	79.0	71.5	65.7	61.5	58.1	55.1	52.8	50.8	49.0	47.5	46.2	45.0	44.0
15	78.1	70.3	64.6	60.3	57.0	53.9	51.5	49.4	47.8	46.1	44.9	43.6	42.5
16	77.3	69.4	63.5	59.2	55.6	52.9	50.3	48.3	46.6	45.2	43.6	42.4	41.4
17	76.5	68.4	62.6	58.2	54.7	51.8	49.3	47.1	45.5	44.0	42.6	41.4	40.3
18	75.5	67.6	61.8	57.1	53.7	50.8	48.4	46.3	44.6	43.0	41.6	40.4	39.3
19	75.1	67.0	60.9	56.5	52.8	49.9	47.4	45.5	43.7	42.1	40.7	39.6	38.4
20	74.1	66.1	60.1	55.6	52.1	49.1	46.6	44.7	42.9	41.3	39.8	38.7	37.5
25	72.0	63.3	57.1	52.5	48.9	46.0	43.4	41.4	39.6	38.1	36.7	35.5	34.5
30	69.8	61.1	55.1	50.3	46.7	43.6	41.1	39.1	37.3	35.7	34.4	33.2	32.1
35	68.5	59.6	53.3	48.6	44.9	41.9	39.5	37.4	35.6	34.0	32.7	31.6	30.4
40	67.5	58.3	52.0	47.3	43.4	40.5	38.0	36.0	34.1	32.7	31.3	30.1	29.1
45	66.4	57.1	50.8	46.1	42.4	39.3	36.9	34.8	33.1	31.5	30.2	29.0	27.9
50	65.6	56.3	49.9	45.2	41.4	38.4	35.9	33.9	32.1	30.5	29.2	28.1	27.0
100	60.9	51.4	44.9	40.0	36.3	33.3	31.0	28.9	27.2	25.8	24.5	23.3	22.3

TABLE : Quantile à 95 % du pourcentage d'inertie des 2 premières dimensions de 10000 PCA obtenue avec des variables indépendantes

1.9 – Aides à l'interprétation : coord., cos2 et contribution

Coordonnées des individus

À la base du calcul de plusieurs grandeurs : inertie, cos2, contributions




```
> res$ind
$coord
                Dim.1      Dim.2
Bordeaux     3.1207071  0.1092968
Brest       -2.2680052  4.0933073
Clermont    -1.7259361 -0.5925322
Grenoble    -1.5292581 -1.6879482
Lille       -4.2168252  0.5952014
Lyon        -0.8349399 -1.7882279
Marseille    4.8327218 -0.8288031
Montpellier  4.1473020 -0.4353508
Nantes      -0.2812894  1.1145634
Nice         6.0070350  0.7893084
Paris       -1.2419376 -0.1563459
Rennes      -1.4386528  1.6711221
Strasbourg  -4.1055997 -2.1722526
Toulouse     1.7361660 -0.1361262
Vichy       -2.2014879 -0.5752126
```

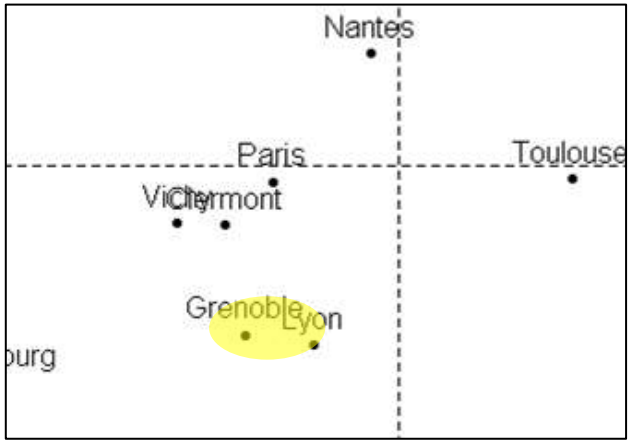
Application

Calcul de l'inertie de l'axe 1

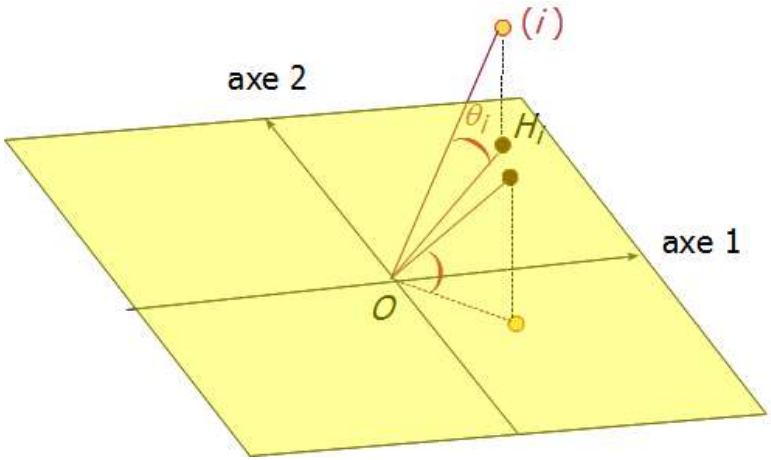
- Peut-on faire confiance aux proximités observées entre produits ?
- Peut-on associer à chaque axe un ou des individus représentatifs ?

Consulter les **aides à l'interprétation**

Qualité de représentation : le cos²

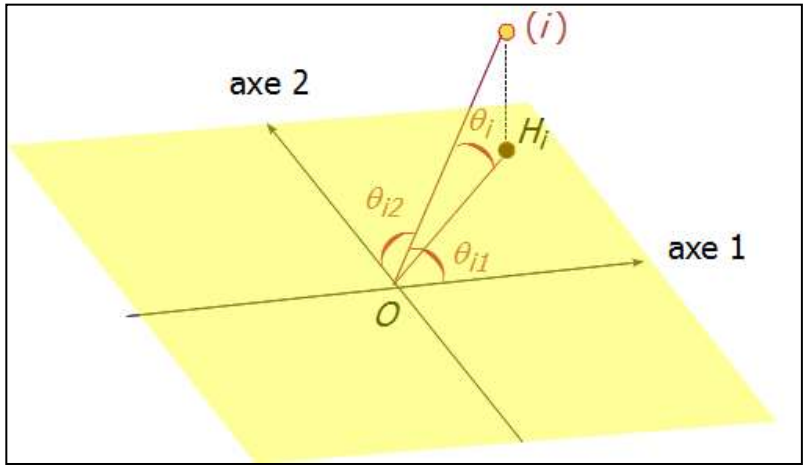


- *La proximité entre Grenoble et Lyon est-elle réelle ? trompeuse ?*
- *Quelle est la qualité de représentation de chacun des individus sur chaque axe ? sur le plan de projection ?*



Règle d'interprétation

- La proximité entre deux points sur le plan est fidèle à la réalité si leur qualité de représentation sur ce plan est bonne
- La qualité de représentation est mesurée par le *cos* ² de l'angle formé avec le plan
- « Mesure de proximité » entre un point et le plan



Qualité de représentation de l'individu (*i*) sur le plan (1,2)

$$QLT_{(1,2)}(i) = cos^2(\theta_i) = \left(\frac{OH_i}{Oi}\right)^2$$

$$QLT_{(1,2)}(i) = QLT_{axe\ 1}(i) + QLT_{axe\ 2}(i)$$
$$QLT_{(1,2)}(i) = cos^2(\theta_{i1}) + cos^2(\theta_{i2})$$

$cos2

	Dim.1	Dim.2
Bordeaux	0.94668773	0.001161224
Brest	0.23436246	0.763393814
Clermont	0.87988441	0.103705112
Grenoble	0.42894041	0.522580994
Lille	0.97152116	0.019355705
Lyon	0.17813711	0.817127272
Marseille	0.96419529	0.028358560
Montpellier	0.98575843	0.010862202
Nantes	0.05645333	0.886324192
Nice	0.98005143	0.016920844
Paris	0.88935998	0.014094539
Rennes	0.41985296	0.566502170
Strasbourg	0.77565410	0.217137845
Toulouse	0.95255524	0.005855863
Vichy	0.92150642	0.062910418

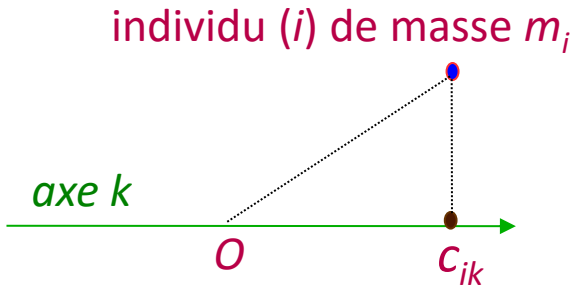
Propriété

$$\sum_{k=1}^{K} QLT_k(i) = 1$$

Contribution d'un individu à un axe : CTR

Contribution à l'inertie d'un individu (*i*)

Importance prise par un individu (*i*) dans la construction de l'axe (*k*)
Elle est liée à
- son éloignement de l'origine sur l'axe (*k*) : c_{ik}
- sa masse m_i



$$CTR_k(i) = m_i \times {c_{ik}}^2$$

La contribution est en général exprimée **en % de l'inertie I_k** de l'axe (*k*)

$$CTR_k(i) = \frac{m_i \times {c_{ik}}^2}{I_k}$$

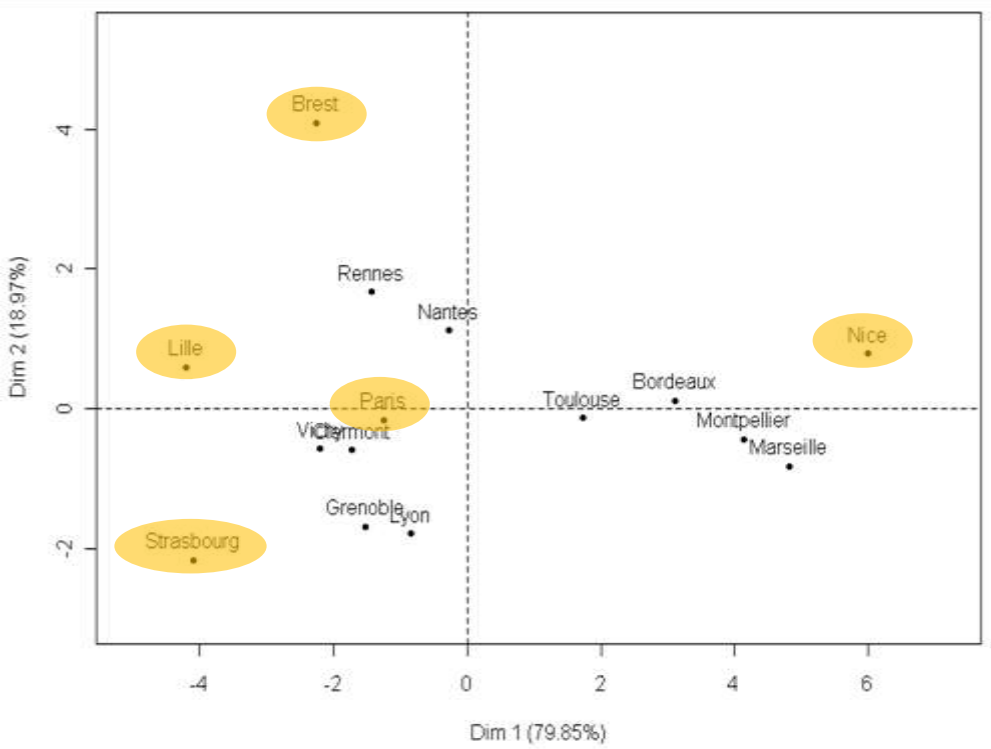
avec

$$I_k = \sum_{i=1}^{n} m_i \times {c_{ik}}^2$$

La CTR dépend de l'éloignement au carré !

$contrib

	Dim.1	Dim.2
Bordeaux	6.7759249	0.03498418
Brest	3.5789091	49.06878939
Clermont	2.0725832	1.02820712
Grenoble	1.6271372	8.34401167
Lille	12.3718247	1.03749158
Lyon	0.4850349	9.36488350
Marseille	16.2497301	2.01168233
Montpellier	11.9672370	0.55505413
Nantes	0.0550515	3.63802974
Nice	25.1063100	1.82452633
Paris	1.0731542	0.07158624
Rennes	1.4400400	8.17848800
Strasbourg	11.7277788	13.81902127
Toulouse	2.0972252	0.05426751
Vichy	3.3720591	0.96897702



Propriété

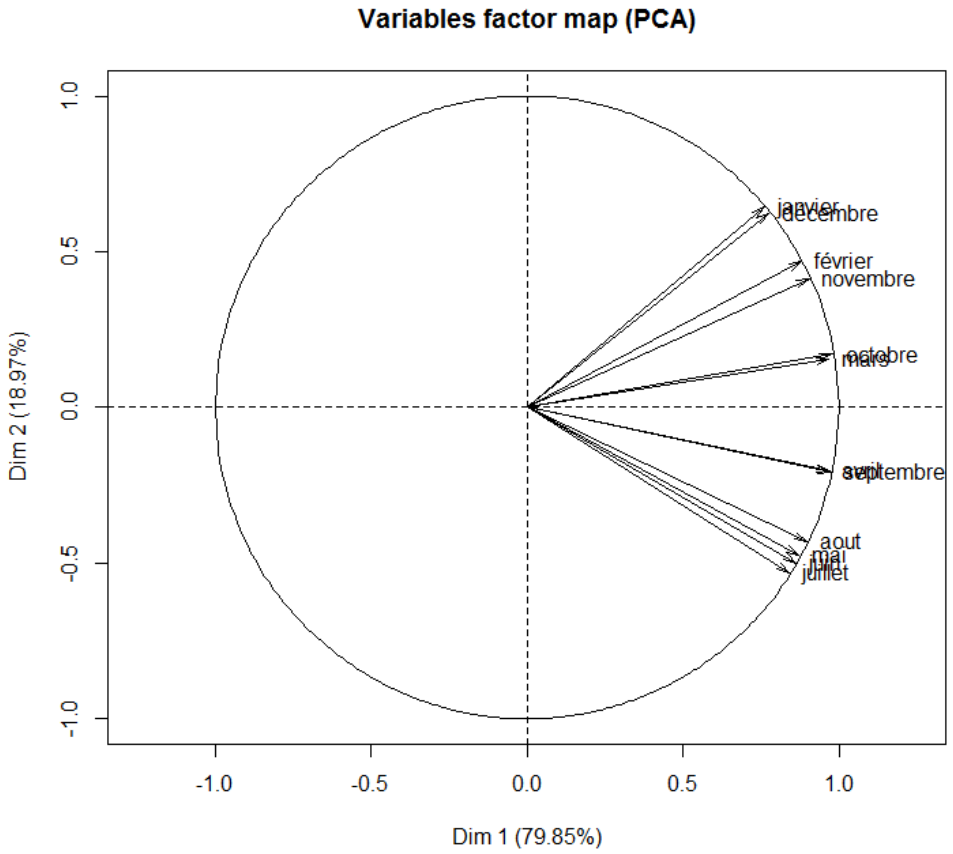
$$\sum_{i=1}^{n} CTR_k(i) = 1$$

Coordonnées des variables

En ACP normée, la c*oordonnée* d_{jk} d'une variable X_j sur l'axe (k) représente **sa corrélation** avec l'axe

$$d_{jk} = \mathrm{r}\,(\text{axe}\,(k), X_j)$$

Pour **interpréter un axe**, on retiendra les variables les plus corrélées à l'axe



En ACP normée

```
$coord
             Dim.1
janvier   0.7612384
février   0.8804578
mars      0.9687704
avril     0.9693357
mai       0.8727646
juin      0.8635747
juillet   0.8415346
aout      0.8986059
septembre 0.9740289
octobre   0.9801599
novembre  0.9037531
décembre  0.7743349

$cor
             Dim.1
janvier   0.7612384
février   0.8804578
mars      0.9687704
avril     0.9693357
mai       0.8727646
juin      0.8635747
juillet   0.8415346
aout      0.8986059
septembre 0.9740289
octobre   0.9801599
novembre  0.9037531
décembre  0.7743349
```

En ACP non normée

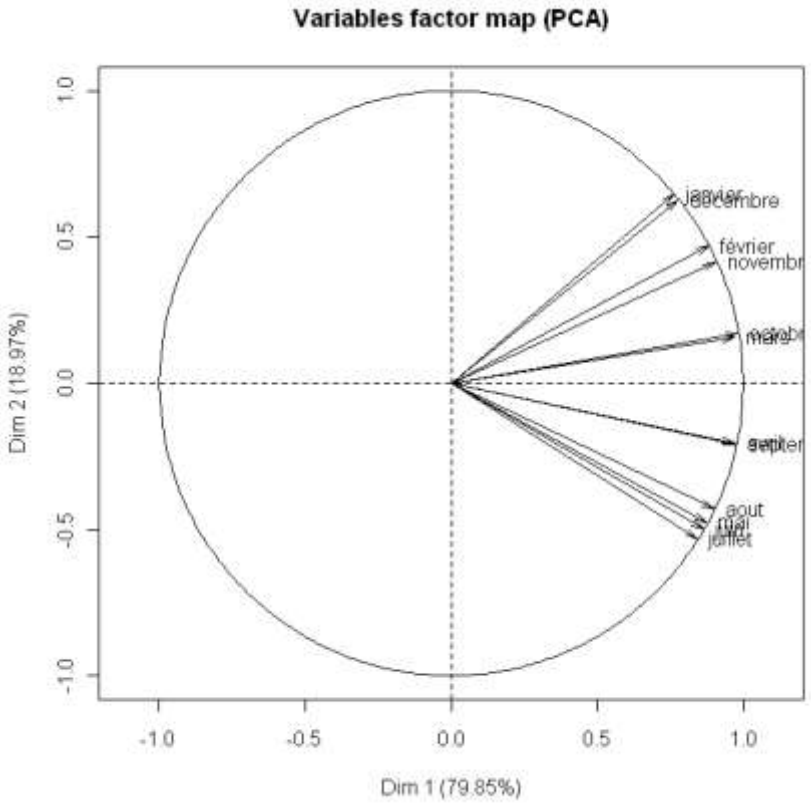
```
$coord
             Dim.1
janvier   1.497028
février   1.602768
mars      1.431510
avril     1.317004
mai       1.254992
juin      1.480660
juillet   1.714372
aout      1.734002
septembre 1.733409
octobre   1.739256
novembre  1.588737
décembre  1.486788

$cor
             Dim.1
janvier   0.7719694
février   0.8879332
mars      0.9690470
avril     0.9635356
mai       0.8633813
juin      0.8545429
juillet   0.8335342
aout      0.8920842
septembre 0.9708544
octobre   0.9837772
novembre  0.9117095
décembre  0.7858115
```

Qualité de représentation et contribution d'une variable

Définitions = idem que pour les individus

- **Qualité de représentation** d'une variable sur un axe : cos² de l'angle avec l'axe
  Conséquence : la qualité sur le plan est liée à sa **proximité** avec le **bord du cercle**
- **Contribution** d'une variable à un axe
  = masse x coordonnée ²



Qualité des variables sur le plan (1,2) ?
Contributions des variables aux axes 1 et 2 ?

Exemple de calcul

$cos2

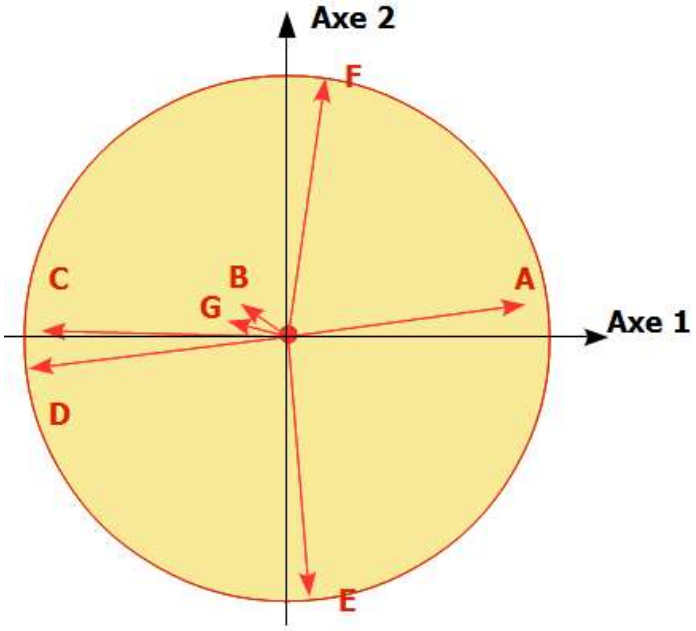
	Dim.1	Dim.2
janvier	0.5794838	0.41517130
février	0.7752059	0.22003470
mars	0.9385160	0.02433862
avril	0.9396117	0.04148299
mai	0.7617180	0.22534881
juin	0.7457613	0.24934646
juillet	0.7081804	0.28240688
aout	0.8074926	0.18484719
septembre	0.9487322	0.04330619
octobre	0.9607134	0.02905542
novembre	0.8167697	0.17133472
décembre	0.5995945	0.38974512

$contrib

	Dim.1	Dim.2
janvier	6.047768	18.237917
février	8.090417	9.665829
mars	9.794799	1.069163
avril	9.806233	1.822292
mai	7.949651	9.899270
juin	7.783119	10.953455
juillet	7.390907	12.405755
aout	8.427376	8.120089
septembre	9.901420	1.902383
octobre	10.026461	1.276365
novembre	8.524196	7.526504
décembre	6.257653	17.120979

- Contribution de janvier à l'axe 1 ?
- Qualité de représentation de janvier sur l'axe 1 ?

Règle de lecture du cercle des corrélations



- Qualité de représentation
- Variables liées à l'axe 1 ? l'axe 2 ?
- r(*C,D*), r(*A,D*), r(*A,E*), r(*A*, *B*), r(*B*,*G*) ?

Quand le cercle des corrélations ne permet pas une lecture aisée des liaisons entre variables, on peut toujours se reporter à la matrice des corrélations !

Matrice des corrélations

Non fournie automatiquement par l'ACP sous `FactoMineR`

Obtenue par le menu : *Statistique – Résumés – Matrice de corrélations*

```
> round(cor(temp[,c("aout","avril","décembre","février","janvier","juillet","juin",
+   "mai","mars","novembre","octobre","septembre")], use="complete.obs"),4)
```

	aout	avril	décembre	février	janvier	juillet	juin	mai	mars	novembre	octobre	septembre
aout	1.0000	0.9490	0.4302	0.5880	0.4051	0.9906	0.9887	0.9803	0.7981	0.6366	0.8121	0.9701
avril	0.9490	1.0000	0.6193	0.7612	0.6115	0.9123	0.9422	0.9526	0.9195	0.7820	0.9053	0.9784
décembre	0.4302	0.6193	1.0000	0.9701	0.9939	0.3239	0.3613	0.3807	0.8336	0.9609	0.8657	0.6224
février	0.5880	0.7612	0.9701	1.0000	0.9735	0.4903	0.5244	0.5466	0.9311	0.9861	0.9403	0.7597
janvier	0.4051	0.6115	0.9939	0.9735	1.0000	0.2969	0.3389	0.3626	0.8353	0.9509	0.8505	0.6037
juillet	0.9906	0.9123	0.3239	0.4903	0.2969	1.0000	0.9915	0.9812	0.7215	0.5472	0.7387	0.9329
juin	0.9887	0.9422	0.3613	0.5244	0.3389	0.9915	1.0000	0.9938	0.7567	0.5725	0.7567	0.9404
mai	0.9803	0.9526	0.3807	0.5466	0.3626	0.9812	0.9938	1.0000	0.7667	0.5910	0.7702	0.9421
mars	0.7981	0.9195	0.8336	0.9311	0.8353	0.7215	0.7567	0.7667	1.0000	0.9273	0.9685	0.9106
novembre	0.6366	0.7820	0.9609	0.9861	0.9509	0.5472	0.5725	0.5910	0.9273	1.0000	0.9643	0.7961
octobre	0.8121	0.9053	0.8657	0.9403	0.8505	0.7387	0.7567	0.7702	0.9685	0.9643	1.0000	0.9256
septembre	0.9701	0.9784	0.6224	0.7597	0.6037	0.9329	0.9404	0.9421	0.9106	0.7961	0.9256	1.0000

1.10 – Méthodologie des variables illustratives

Variables actives (ou de base)

- Elles participent à la construction des axes
- Variables à partir desquelles sont calculées les ressemblances entre individus

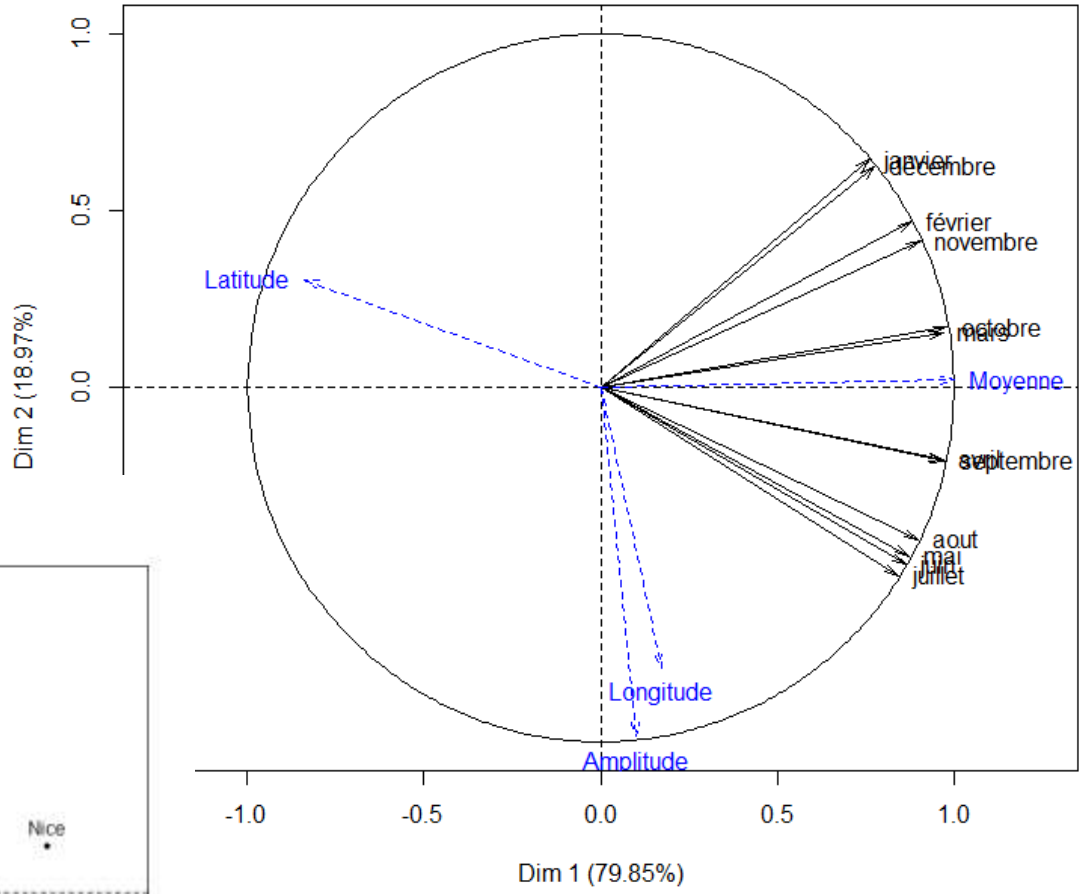
Variables illustratives (ou supplémentaires)

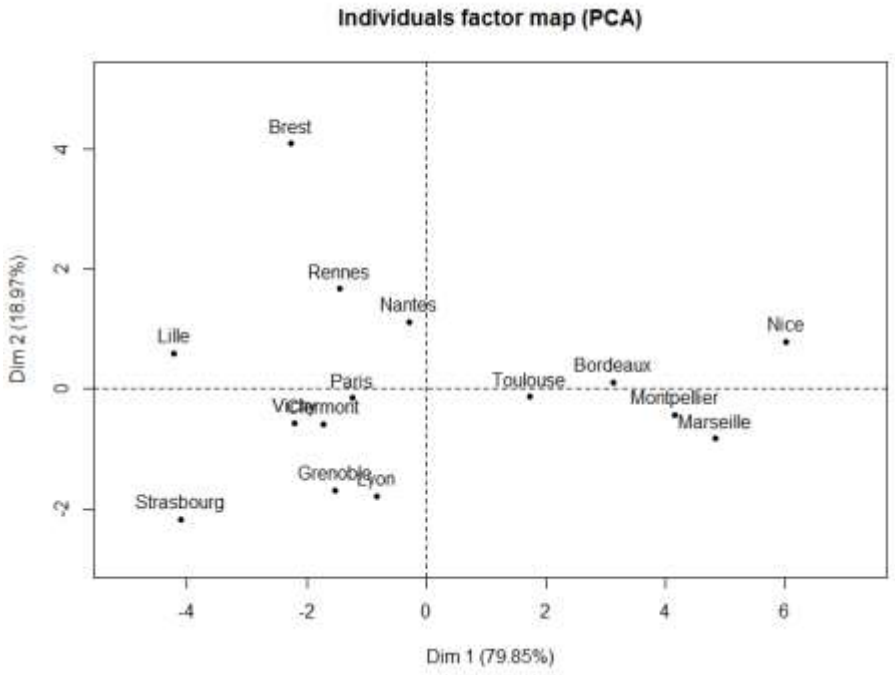
Elles ne participent pas à la construction des axes.
On peut les projeter *a posteriori* sur les graphes

1. pour faciliter l'interprétation des axes
2. pour mettre en relation deux ensembles de variables

En ACP, les variables illustratives = *qualitatives* et / ou *quantitatives*

On peut gérer également des **individus supplémentaires**


janvier
décembre
février
novembre
Latitude
octobre
mars
Moyenne
septembre
août
mai
juillet
Longitude
Amplitude
1.0
0.5
0.0
-1.0
-0.5
0.0
0.5
1.0
Dim 2 (18.97%)
Dim 1 (79.85%)


Individuals factor map (PCA)
Brest
Rennes
Nantes
Lille
Nice
Bordeaux
Toulouse
Paris
Montpellier
Marseille
Grenoble
Lyon
Strasbourg
4
2
0
-2
-4
-2
0
2
4
6
Dim 2 (18.97%)
Dim 1 (79.85%)