

---

# TD : Analyse Multidimensionnelle : ACP ?

---

Cours : Analyse de Données

---

Chouaïb LAGHLAM 2015

---

## **Analyse Multidimensionnelle**

C'est Quoi ? ..... 2

## **Analyse Multidimensionnelle ?**

### **Les 3 méthodes les plus courantes**

### **ACP ?**

t

## Pratique de l'ACP

- ① Choisir les variables actives
- ② Choisir de réduire ou non les variables
- ③ Réaliser l'ACP
- ④ Choisir le nombre de dimensions à interpréter
- ⑤ Interpréter simultanément le graphe des individus et celui des variables
- ⑥ Utiliser les indicateurs pour enrichir l'interprétation
- ⑦ Revenir aux données brutes pour interpréter



# Analyse en composantes principales (ACP)

- ① Données - exemples - notations
- ② Etude des individus
- ③ Etude des variables
- ④ Aides à l'interprétation

## Quel type de données ?

L'ACP s'intéresse à des tableaux de données rectangulaires avec des **individus** en lignes et des **variables quantitatives** en colonnes

	1	$k$	$K$
1			
$i$		$x_{ik}$	
$I$			

Pour la variable  $k$ , on note :

$$\text{la moyenne : } \bar{x}_k = \frac{1}{I} \sum_{i=1}^I x_{ik}$$

$$\text{l'écart-type : } s_k = \sqrt{\frac{1}{I} \sum_{i=1}^I (x_{ik} - \bar{x}_k)^2}$$

**FIGURE :** Tableau de données en ACP

## Exemples

- Analyse sensorielle : note du descripteur  $k$  pour le produit  $i$
- Ecologie : concentration du polluant  $k$  dans la rivière  $i$
- Economie : valeur de l'indicateur  $k$  pour l'année  $i$
- Génétique : expression du gène  $k$  pour le patient  $i$
- Biologie : mesure  $k$  pour l'animal  $i$
- Marketing : valeur d'indice de satisfaction  $k$  pour la marque  $i$
- Sociologie : temps passé à l'activité  $k$  par les individus de la CSP  $i$
- etc.

⇒ Il existe de très nombreux tableaux comme cela



## Les données température

- 15 individus (lignes) : villes de France
- 14 variables (colonnes) :
  - 12 températures mensuelles moyennes (sur 30 ans)
  - 2 variables géographiques (latitude, longitude)

	Janv	Févr	Mars	Avri	Mai	Juin	juil	Août	Sept	Octo	Nove	Déce	Lati	Long
Bordeaux	5.6	6.6	10.3	12.8	15.8	19.3	20.9	21	18.6	13.8	9.1	6.2	44.5	-0.34
Brest	6.1	5.8	7.8	9.2	11.6	14.4	15.6	16	14.7	12	9	7	48.24	-4.29
Clermont	2.6	3.7	7.5	10.3	13.8	17.3	19.4	19.1	16.2	11.2	6.6	3.6	45.47	3.05
Grenoble	1.5	3.2	7.7	10.6	14.5	17.8	20.1	19.5	16.7	11.4	6.5	2.3	45.1	5.43
Lille	2.4	2.9	6	8.9	12.4	15.3	17.1	17.1	14.7	10.4	6.1	3.5	50.38	3.04
Lyon	2.1	3.3	7.7	10.9	14.9	18.5	20.7	20.1	16.9	11.4	6.7	3.1	45.45	4.51
Marseille	5.5	6.6	10	13	16.8	20.8	23.3	22.8	19.9	15	10.2	6.9	43.18	5.24
Montpellier	5.6	6.7	9.9	12.8	16.2	20.1	22.7	22.3	19.3	14.6	10	6.5	43.36	3.53
Nantes	5	5.3	8.4	10.8	13.9	17.2	18.8	18.6	16.4	12.2	8.2	5.5	47.13	-1.33
Nice	7.5	8.5	10.8	13.3	16.7	20.1	22.7	22.5	20.3	16	11.5	8.2	43.42	7.15
Paris	3.4	4.1	7.6	10.7	14.3	17.5	19.1	18.7	16	11.4	7.1	4.3	48.52	2.2
Rennes	4.8	5.3	7.9	10.1	13.1	16.2	17.9	17.8	15.7	11.6	7.8	5.4	48.05	-1.41
Strasbourg	0.4	1.5	5.6	9.8	14	17.2	19	18.3	15.1	9.5	4.9	1.3	48.35	7.45
Toulouse	4.7	5.6	9.2	11.6	14.9	18.7	20.9	20.9	18.3	13.3	8.6	5.5	43.36	1.26
Vichy	2.4	3.4	7.1	9.9	13.6	17.1	19.3	18.8	16	11	6.6	3.4	46.08	3.26



## Problèmes - objectifs

Le tableau peut être vu comme un ensemble de lignes ou un ensemble de colonnes

### Etude des individus

- Quand dit-on que 2 individus se ressemblent du point de vue de l'ensemble des variables ?
- Si beaucoup d'individus, peut-on faire un bilan des ressemblances ?

⇒ construction de groupes d'individus, partition des individus

## Problèmes - objectifs

### Etude des variables

- Recherche des ressemblances entre variables
- Entre variables, on parle plutôt de liaisons
- Liaisons linéaires sont simples, très fréquentes et résument de nombreuses liaisons  $\Rightarrow$  coefficient de corrélation

$\Rightarrow$  visualisation de la matrice des corrélations

$\Rightarrow$  recherche d'un petit nombre d'indicateurs synthétiques pour résumer beaucoup de variables (ex. d'indicateur synthétique a priori : la moyenne, mais ici on recherche des indicateurs synthétiques a posteriori, à partir des données)

## Problèmes - objectifs

### Lien entre les deux études

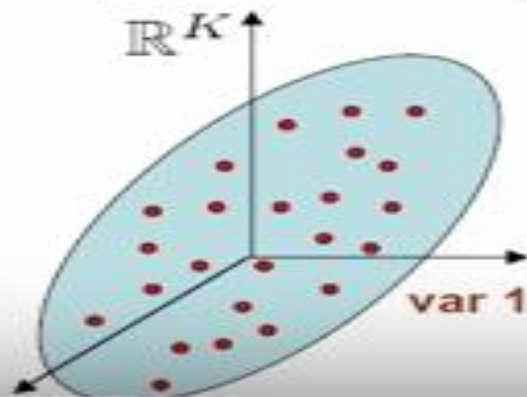
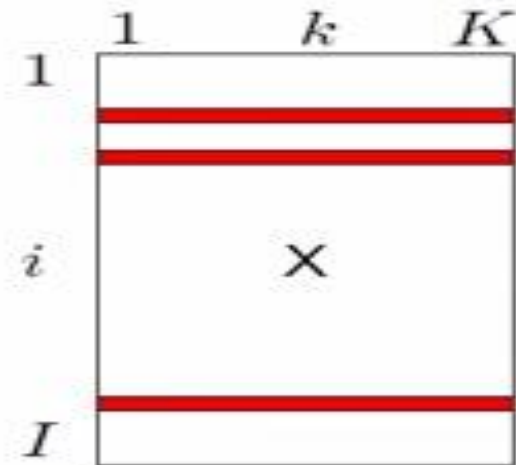
- Caractérisation des classes d'individus par les variables  
⇒ besoin de procédure automatique
- Individus spécifiques pour comprendre les liaisons entre variables  
⇒ utilisation d'individus extrêmes (en terme de variables : langage abstrait mais puissant, revenir aux individus pour voir les choses plus simplement)

### Objectifs de l'ACP :

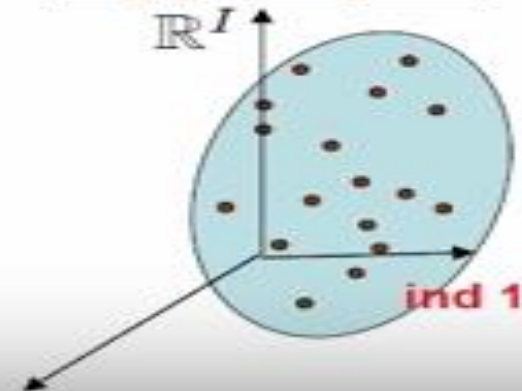
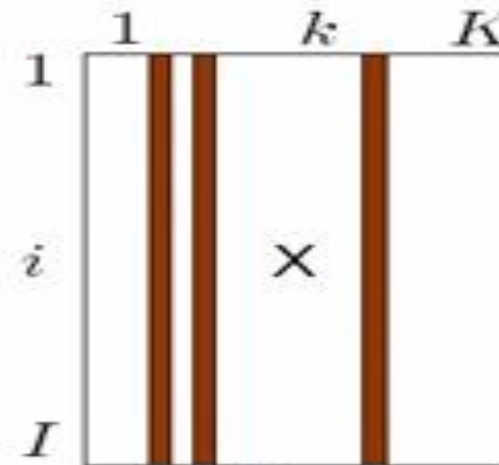
- Descriptif - exploratoire : visualisation de données par graphiques simples
- Synthèse - résumé de grands tableaux individus  $\times$  variables

## Deux nuages de points

### Etude des individus



### Etude des variables



## Le nuage des individus $N'$

1 individu = 1 ligne du tableau  $\Rightarrow$  1 point dans un espace à  $K$  dim

- Si  $K = 1$  : Représentation axiale
- Si  $K = 2$  : Nuage de points
- Si  $K = 3$  : Représentation + difficile en 3D
- Si  $K = 4$  : Impossible à représenter MAIS le concept est simple

Notion de ressemblance : distance (au carré) entre individus  $i$  et  $i'$  :

$$d^2(i, i') = \sum_{k=1}^K (x_{ik} - x_{i'k})^2 \quad (\text{merci Pythagore})$$

Etude des individus  $\equiv$  Etude de la forme du nuage  $N'$



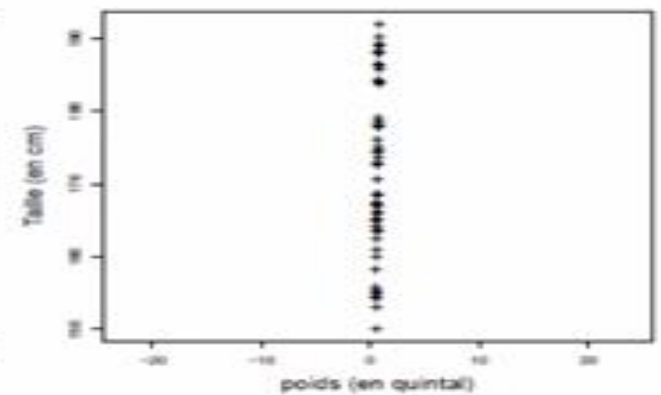
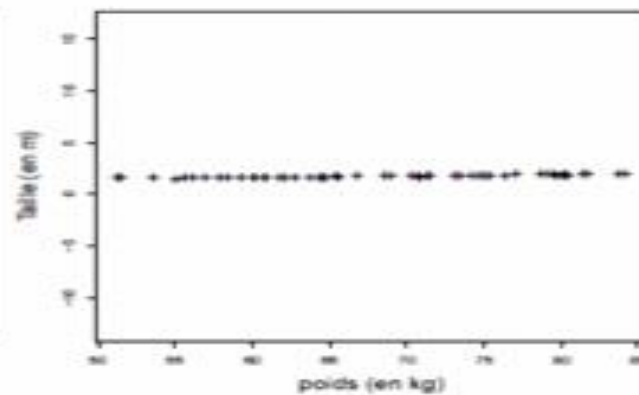
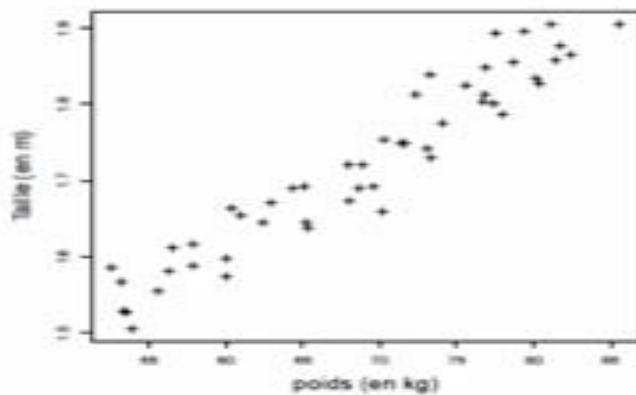
## Le nuage des individus $N'$



- Etudier la structure, *i.e.* la forme du nuage des individus
- Les individus vivent dans  $\mathbb{R}^K$

## Centrage – réduction des données

- Centrer les données ne modifie pas la forme du nuage  
⇒ toujours centrer



- Réduire les données est indispensable si les unités de mesure sont différentes d'une variable à l'autre

$$x_{ik} \mapsto \frac{x_{ik} - \bar{x}_k}{s_k}$$



## Centrage – réduction des données

	Janv	Févr	Mars	Avri	Mai	Juin	juil	Août	Sept	Octo	Nove	Déce
Bordeaux	0.84	0.98	1.40	1.33	0.94	0.85	0.52	0.74	0.90	0.84	0.67	0.72
Brest	1.10	0.54	-0.29	-1.30	-1.95	-1.98	-2.06	-1.83	-1.28	-0.18	0.62	1.14
Clermont	-0.71	-0.63	-0.50	-0.50	-0.44	-0.31	-0.21	-0.24	-0.44	-0.63	-0.76	-0.66
Grenoble	-1.28	-0.90	-0.36	-0.28	0.05	-0.02	0.13	-0.03	-0.16	-0.52	-0.82	-1.35
Lille	-0.81	-1.07	-1.51	-1.52	-1.40	-1.46	-1.33	-1.27	-1.28	-1.09	-1.05	-0.71
Lyon	-0.97	-0.85	-0.36	-0.06	0.32	0.38	0.42	0.27	-0.05	-0.52	-0.70	-0.92
Marseille	0.79	0.98	1.20	1.48	1.63	1.71	1.69	1.66	1.63	1.52	1.30	1.09
Montpellier	0.84	1.03	1.13	1.33	1.22	1.31	1.39	1.41	1.30	1.29	1.19	0.87
Nantes	0.53	0.26	0.11	-0.13	-0.37	-0.37	-0.50	-0.50	-0.33	-0.07	0.16	0.35
Nice	1.82	2.03	1.74	1.70	1.56	1.31	1.39	1.51	1.86	2.08	2.05	1.77
Paris	-0.30	-0.41	-0.43	-0.20	-0.09	-0.19	-0.36	-0.45	-0.55	-0.52	-0.47	-0.29
Rennes	0.43	0.26	-0.23	-0.64	-0.92	-0.94	-0.94	-0.91	-0.72	-0.41	-0.07	0.29
Strasbourg	-1.84	-1.85	-1.78	-0.86	-0.30	-0.37	-0.41	-0.65	-1.06	-1.60	-1.74	-1.87
Toulouse	0.37	0.42	0.65	0.45	0.32	0.50	0.52	0.69	0.74	0.55	0.39	0.35
Vichy	-0.81	-0.79	-0.77	-0.79	-0.57	-0.42	-0.26	-0.39	-0.55	-0.75	-0.76	-0.76

ACP  $\equiv$  Analyse du tableau centré-réduit

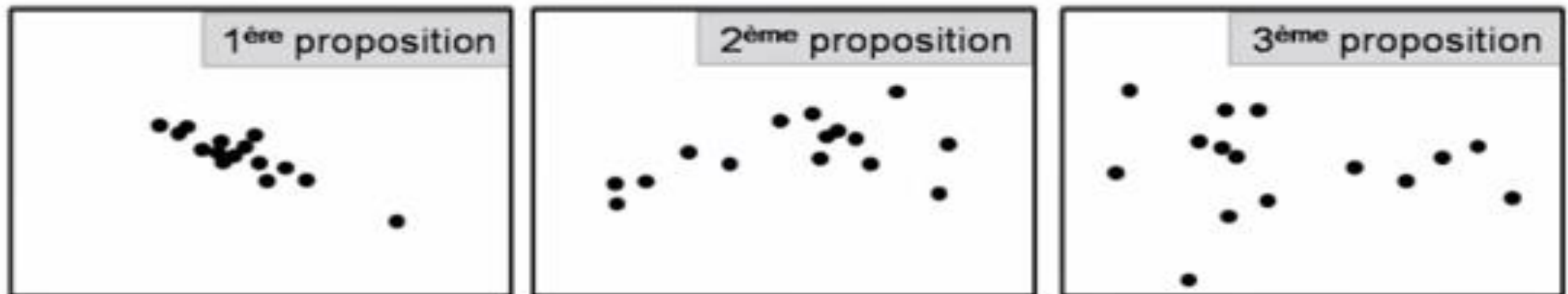
Difficile de voir le nuage  $N'$   $\Rightarrow$  on essaie d'en avoir une image approchée

## Ajustement du nuage des individus

L'ACP vise à fournir une image simplifiée de  $N'$  la + fidèle possible  
 $\iff$  Trouver le sous-espace qui résume au mieux les données

Qualité d'une image :

- Restitue fidèlement la forme générale du nuage (*animation*)



## Ajustement du nuage des individus

L'ACP vise à fournir une image simplifiée de  $N'$  la + fidèle possible  
 $\iff$  Trouver le sous-espace qui résume au mieux les données

Qualité d'une image :

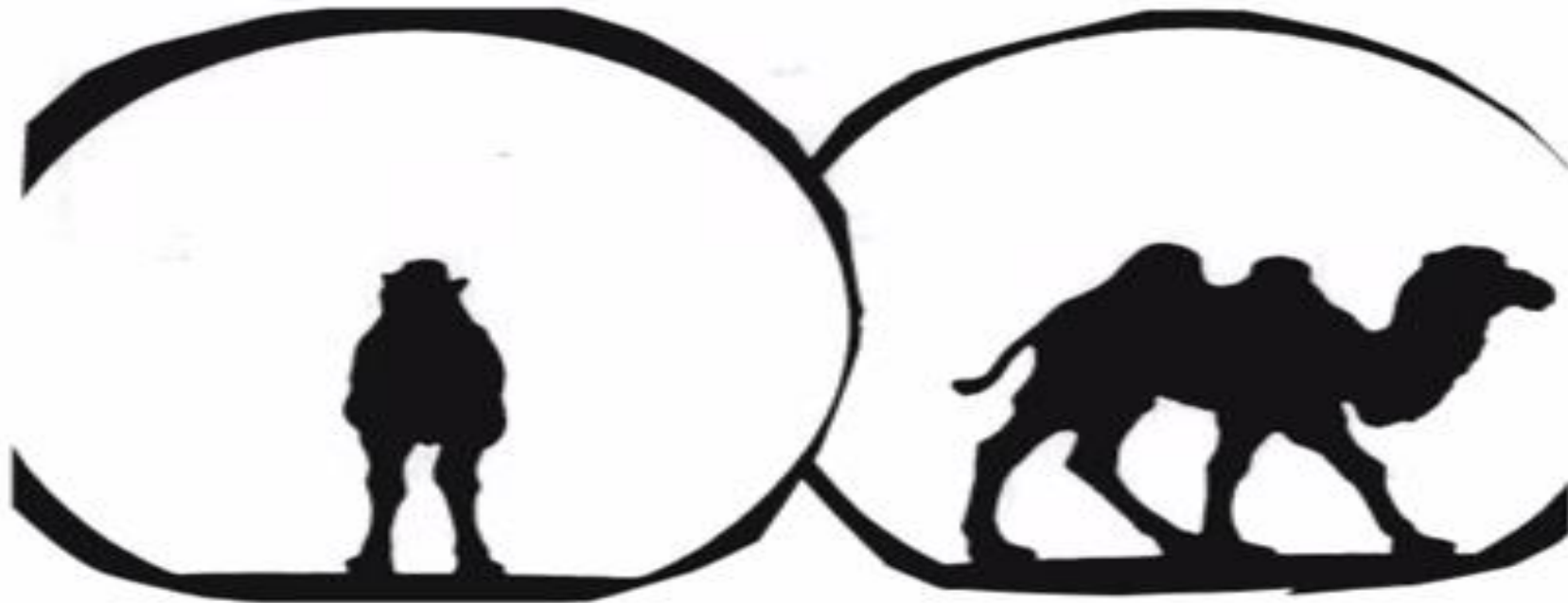
- Restitue fidèlement la forme générale du nuage (*animation*)
- Meilleure représentation de la diversité, de la variabilité
- Ne perturbe pas les distances entre individus

Comment quantifier la qualité d'une image ?

A l'aide de la notion de dispersion ou variabilité appelée  
**Inertie**

Inertie  $\equiv$  variance généralisée à plusieurs dimensions

## Ajustement du nuage des individus



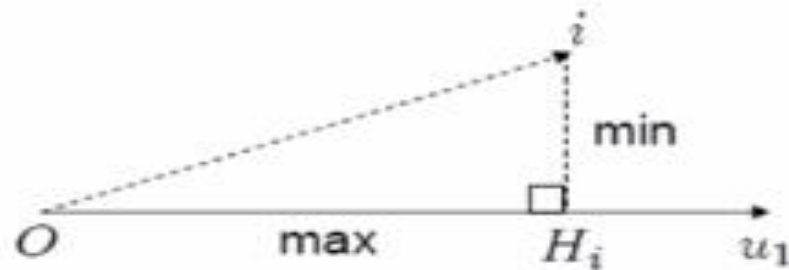
**FIGURE:** Quel animal ? (*illustration JP Fénelon*)



## Ajustement du nuage des individus

Comment trouver la meilleure image approchée du nuage ?

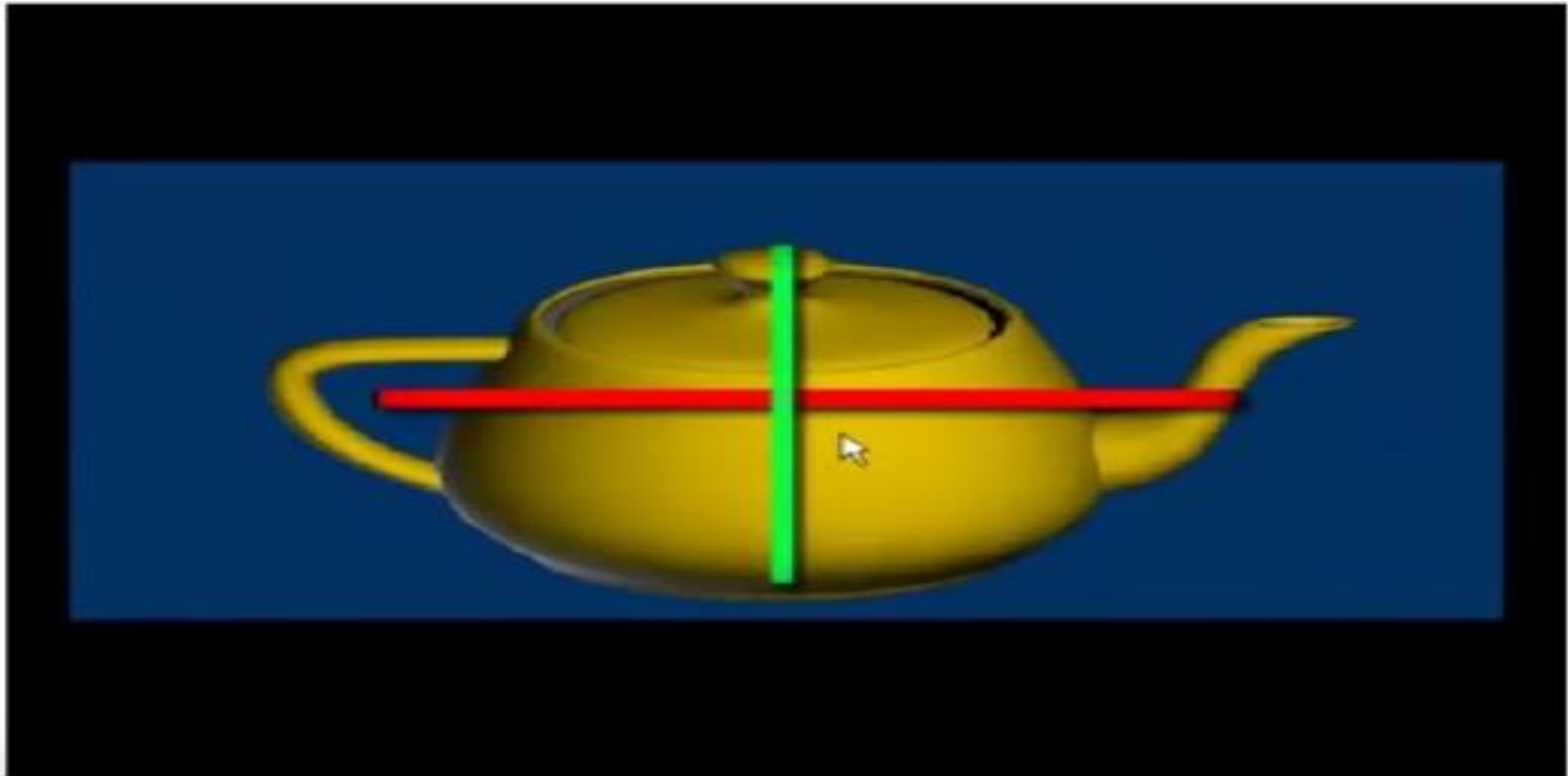
- 1 Trouver l'axe (facteur) qui déforme le moins possible le nuage



$(iH_i)^2$  petit avec  $H_i \in \text{axe} \Leftrightarrow$   
 $(OH_i)^2$  grand (Pythagore)  
 $\Rightarrow$  on veut  $\sum_i (OH_i)^2$  grand

- 2 Trouver le meilleur plan : maximiser  $\sum_i (OH_i)^2$  avec  $H_i \in \text{plan}$   
 Meilleur plan contient le meilleur axe : on cherche  $u_2 \perp u_1$  et maximisant  $\sum_i (OH_i)^2$
- 3 on peut chercher un 3ème axe, etc. d'inertie maximum

## Illustration de la construction des axes en ACP



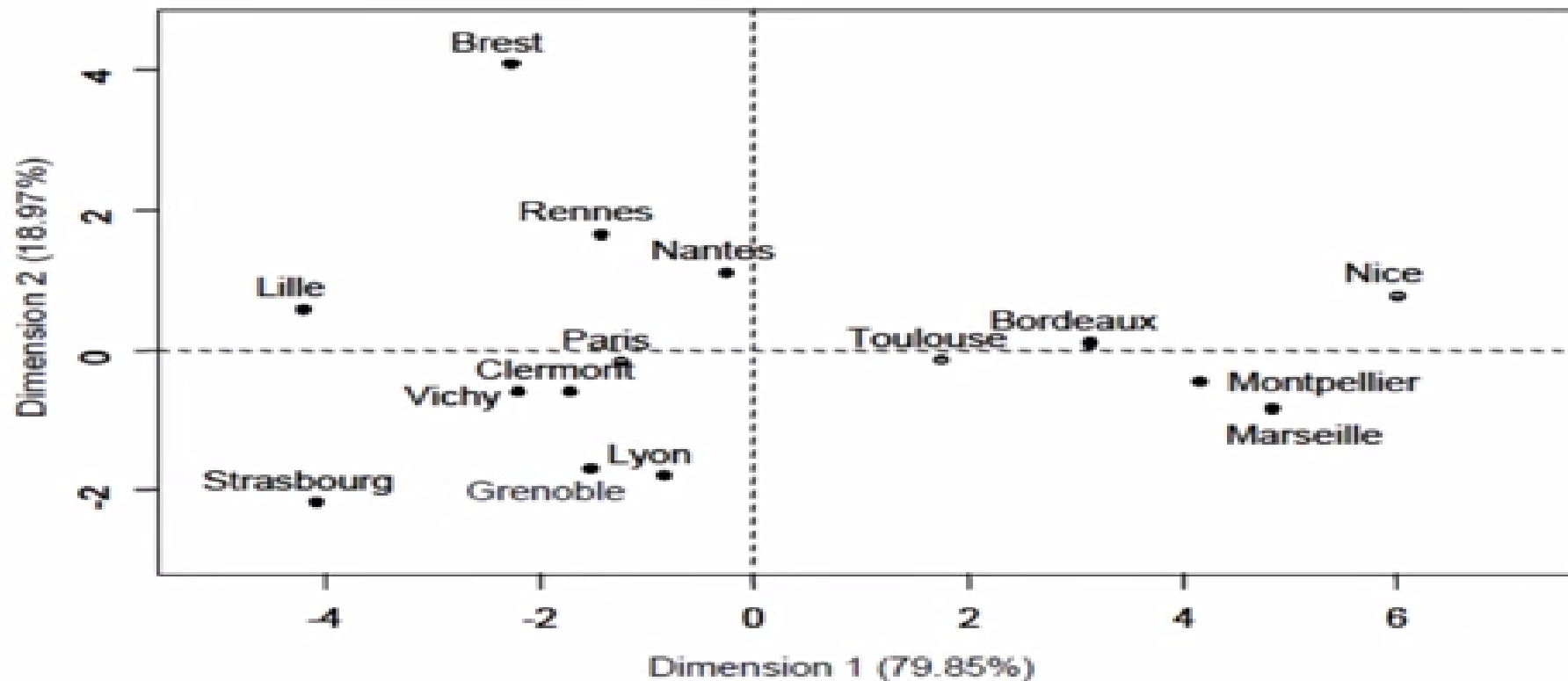
## Les données température

- 15 individus (lignes) : villes de France
- 14 variables (colonnes) :
  - 12 températures mensuelles moyennes (sur 30 ans)
  - 2 variables géographiques (latitude, longitude)

	Janv	Févr	Mars	Avri	Mai	Juin	juil	Août	Sept	Octo	Nov	Déce	Lati	Long
Bordeaux	5.6	6.6	10.3	12.8	15.8	19.3	20.9	21	18.6	13.8	9.1	6.2	44.5	-0.34
Brest	6.1	5.8	7.8	9.2	11.6	14.4	15.6	16	14.7	12	9	7	48.24	-4.29
Clermont	2.6	3.7	7.5	10.3	13.8	17.3	19.4	19.1	16.2	11.2	6.6	3.6	45.47	3.05
Grenoble	1.5	3.2	7.7	10.6	14.5	17.8	20.1	19.5	16.7	11.4	6.5	2.3	45.1	5.43
Lille	2.4	2.9	6	8.9	12.4	15.3	17.1	17.1	14.7	10.4	6.1	3.5	50.38	3.04
Lyon	2.1	3.3	7.7	10.9	14.9	18.5	20.7	20.1	16.9	11.4	6.7	3.1	45.45	4.51
Marseille	5.5	6.6	10	13	16.8	20.8	23.3	22.8	19.9	15	10.2	6.9	43.18	5.24
Montpellier	5.6	6.7	9.9	12.8	16.2	20.1	22.7	22.3	19.3	14.6	10	6.5	43.36	3.53
Nantes	5	5.3	8.4	10.8	13.9	17.2	18.8	18.6	16.4	12.2	8.2	5.5	47.13	-1.33
Nice	7.5	8.5	10.8	13.3	16.7	20.1	22.7	22.5	20.3	16	11.5	8.2	43.42	7.15
Paris	3.4	4.1	7.6	10.7	14.3	17.5	19.1	18.7	16	11.4	7.1	4.3	48.52	2.2
Rennes	4.8	5.3	7.9	10.1	13.1	16.2	17.9	17.8	15.7	11.6	7.8	5.4	48.05	-1.41
Strasbourg	0.4	1.5	5.6	9.8	14	17.2	19	18.3	15.1	9.5	4.9	1.3	48.35	7.45
Toulouse	4.7	5.6	9.2	11.6	14.9	18.7	20.9	20.9	18.3	13.3	8.6	5.5	43.36	1.26
Vichy	2.4	3.4	7.1	9.9	13.6	17.1	19.3	18.8	16	11	6.6	3.4	46.08	3.26



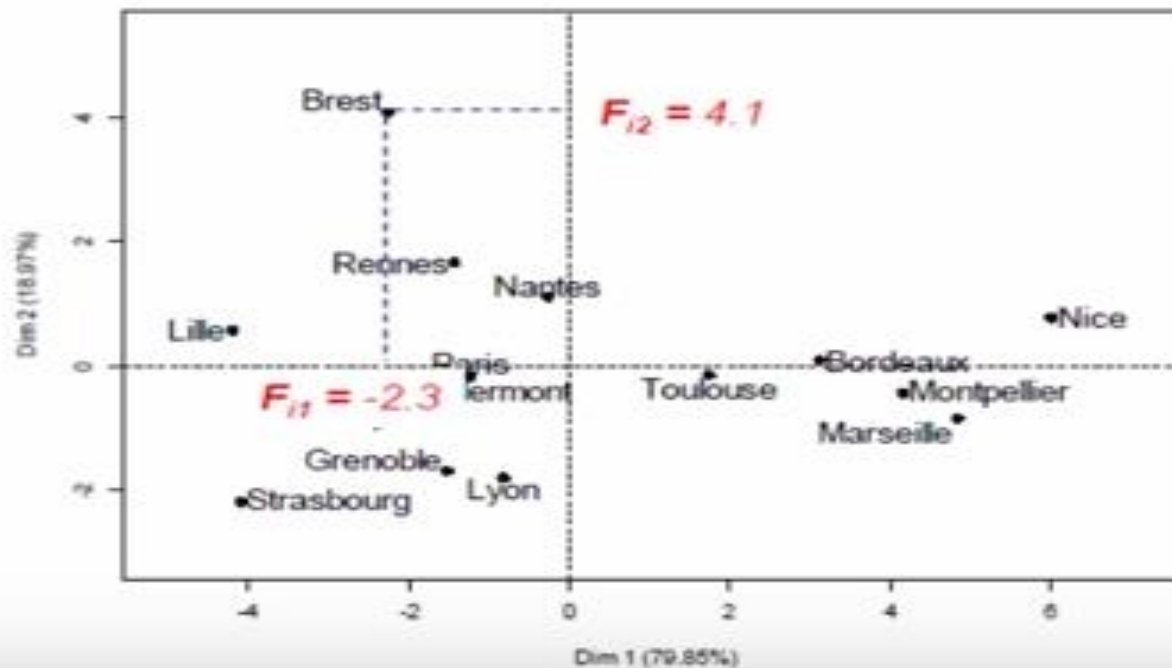
## Exemple : graphe des individus



Comment interpréter les axes ? Qu'est-ce qui oppose Lille à Nice ?  
 ⇒ Besoin de variables pour interpréter ces dimensions de variabilité

# Interprétation du graphe des individus grâce aux variables

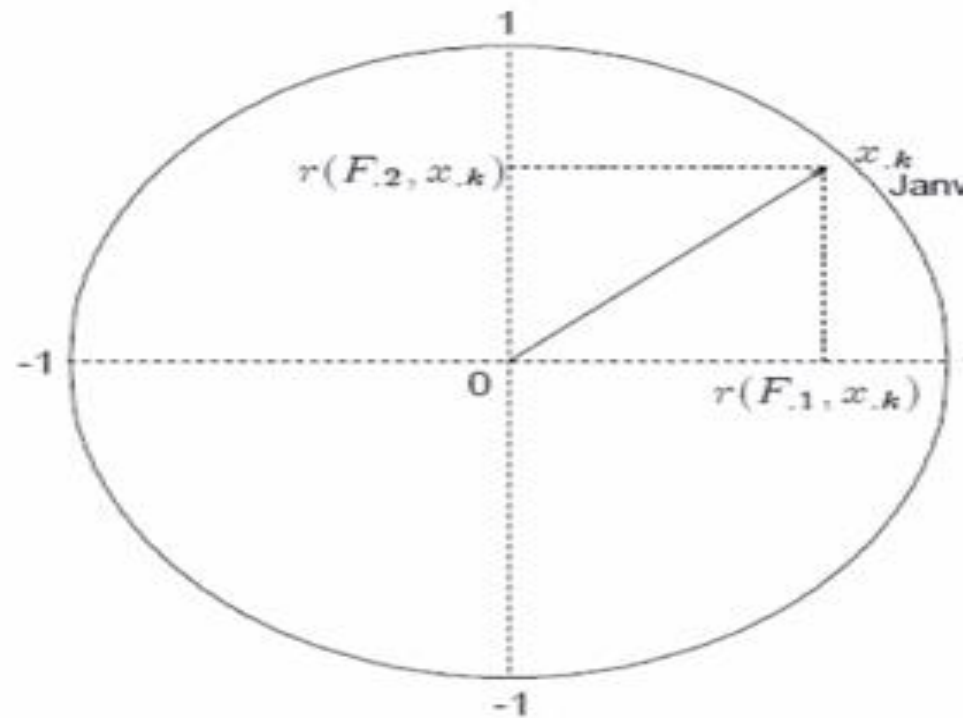
Considérons les coordonnées des individus sur les axes comme des variables



	1	$k$	$K$	$F_{.1}$	$F_{.2}$
1	$x_{ik}$			-2.3	4.1
$i$				$F_{i1}$	$F_{i2}$
$I$					

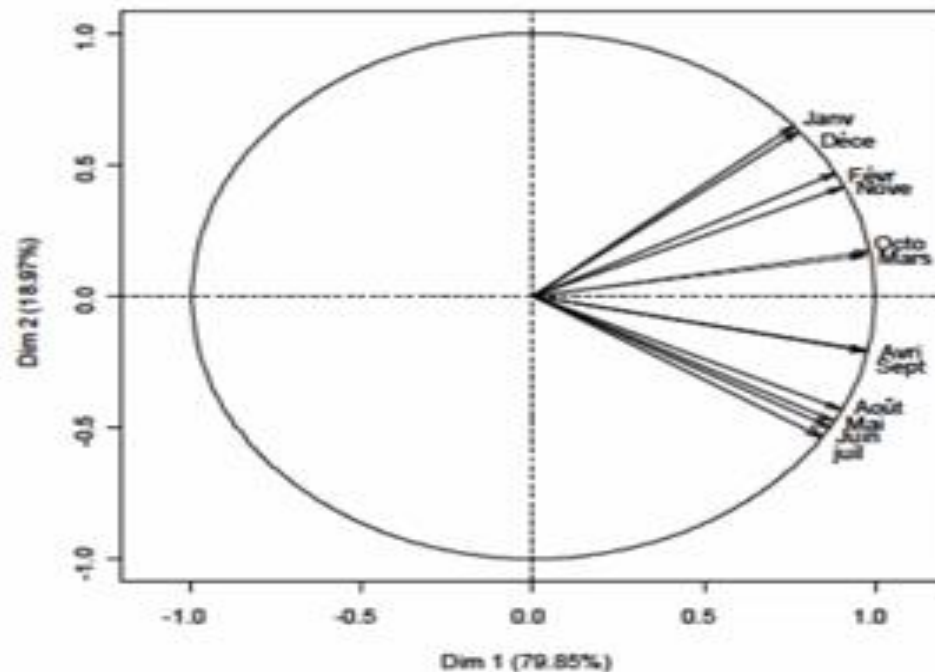
## Interprétation du graphe des individus grâce aux variables

- Corrélations entre la variable  $x_{.k}$  et  $F_{.1}$  (et  $F_{.2}$ )



⇒ Cercle des corrélations

## Interprétation du graphe des individus grâce aux variables



Toutes les variables sont corrélées à  $F_1$ .

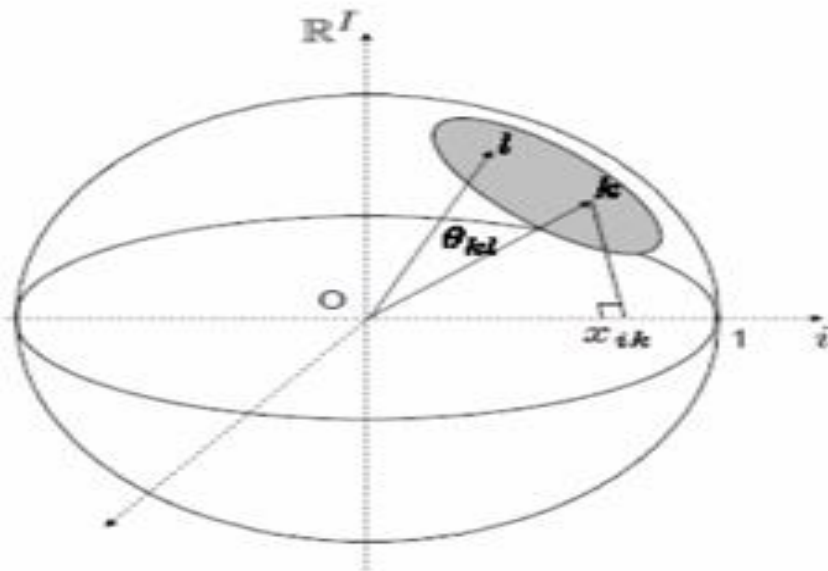
Comment interpréter le 1er axe ?

Comment interpréter le 2ème ?

Principaux facteurs de variabilité :

- 1 - villes chaudes et froides ;
- 2 - à  $T^\circ$  moyenne constante : l'amplitude thermique

## Nuage des variables $N^K$



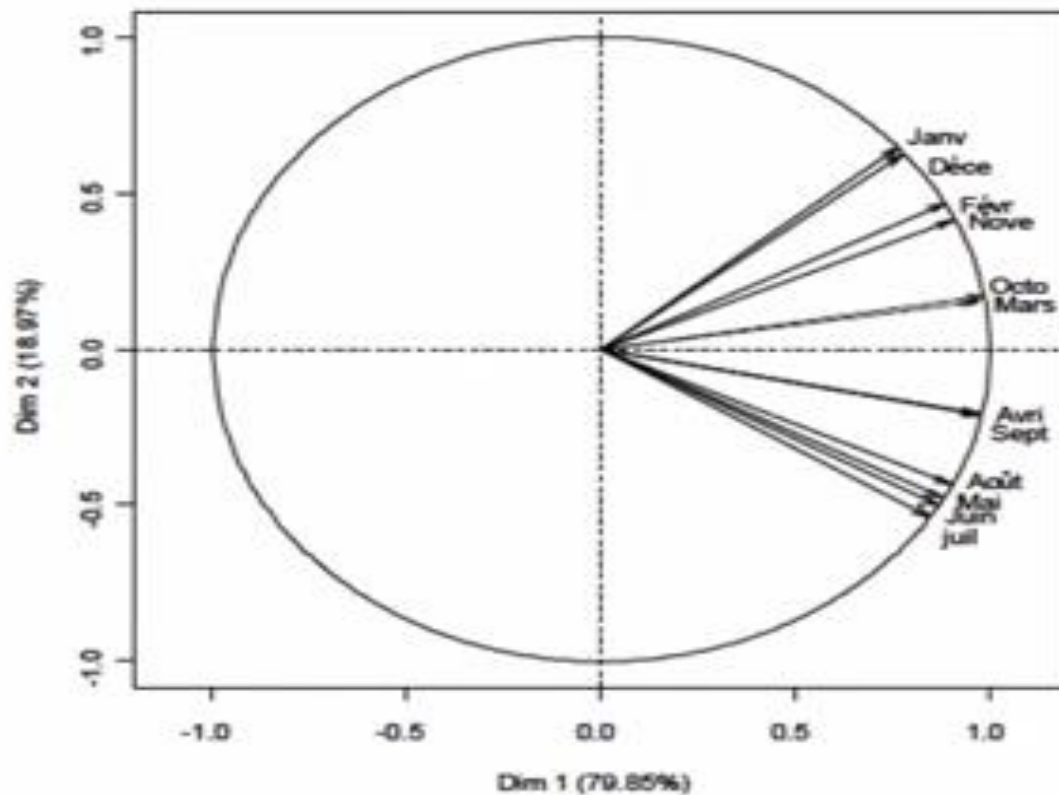
1 variable = 1 point dans un espace à  $I$  dimensions

$$\begin{aligned}\cos(\theta_{kl}) &= \frac{\langle X_{.k}, X_{.l} \rangle}{\|X_{.k}\| \|X_{.l}\|} \\ &= \frac{\sum_{i=1}^I x_{ik} x_{il}}{\sqrt{\sum_{i=1}^I x_{ik}^2} \sqrt{\sum_{i=1}^I x_{il}^2}}\end{aligned}$$

Comme les variables sont **centrées** :  $\cos(\theta_{kl}) = r(x_{.k}, x_{.l})$

Si variables **réduites**  $\Rightarrow$  points sur une hypersphère de rayon **1**

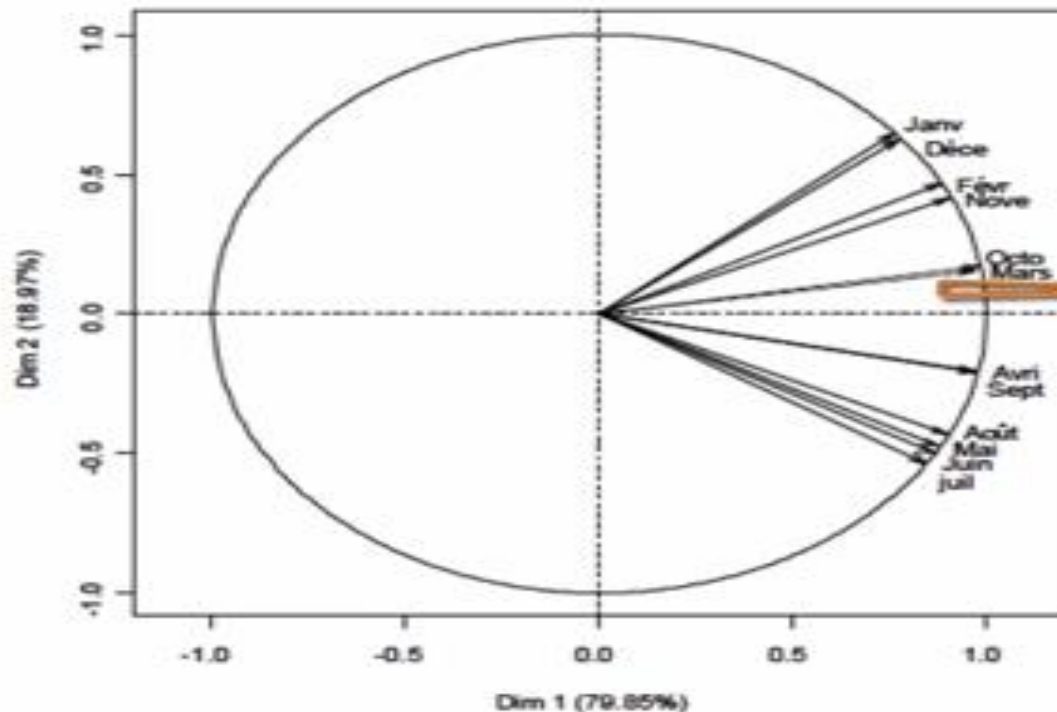
## Ajustement du nuage des variables



⇒ Même représentation que précédemment !!!!



## Ajustement du nuage des variables



⇒ Même représentation que précédemment !!!!

- aide pour interpréter les individus
- représentation optimale du nuage des variables
- visualisation de la matrice des corrélations

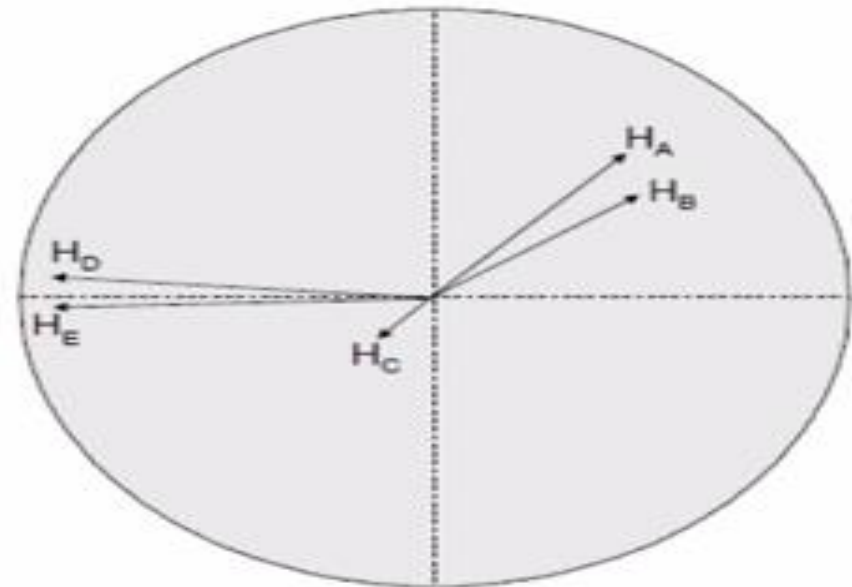
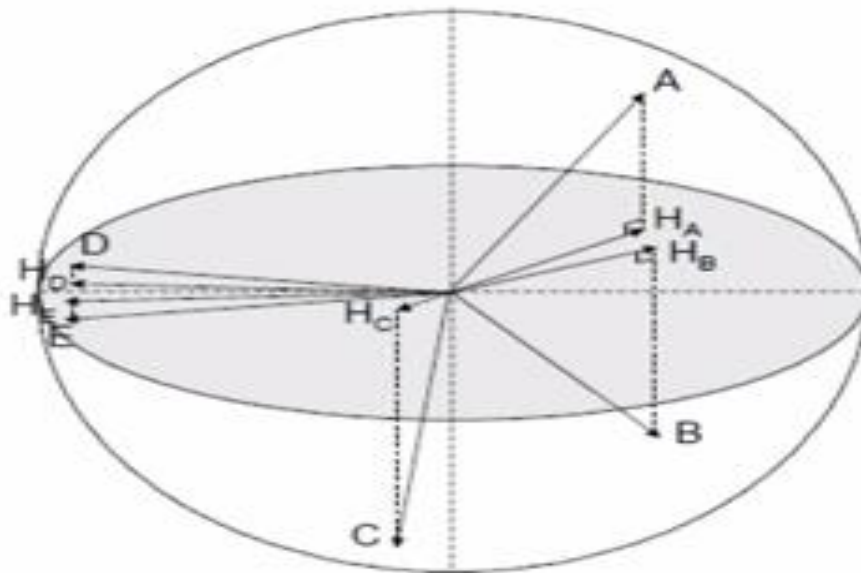
**INCROYABLE !! EXTRAORDINAIRE !!!**



## Projections...

$$r(A, B) = \cos(\theta_{A,B})$$

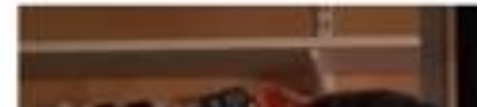
$\cos(\theta_{A,B}) \approx \cos(\theta_{H_A, H_B})$  si les variables sont bien projetées



Seules les variables bien projetées peuvent être interprétées !

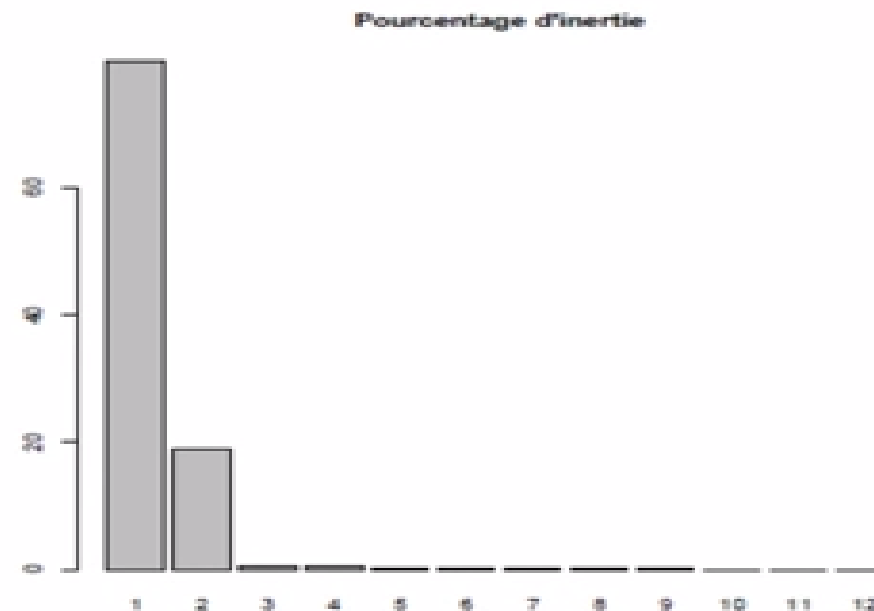
# Analyse en Composantes Principales (ACP)

- ① Données - Exemples
- ② Etude des individus
- ③ Etude des variables
- ④ Aides à l'interprétation



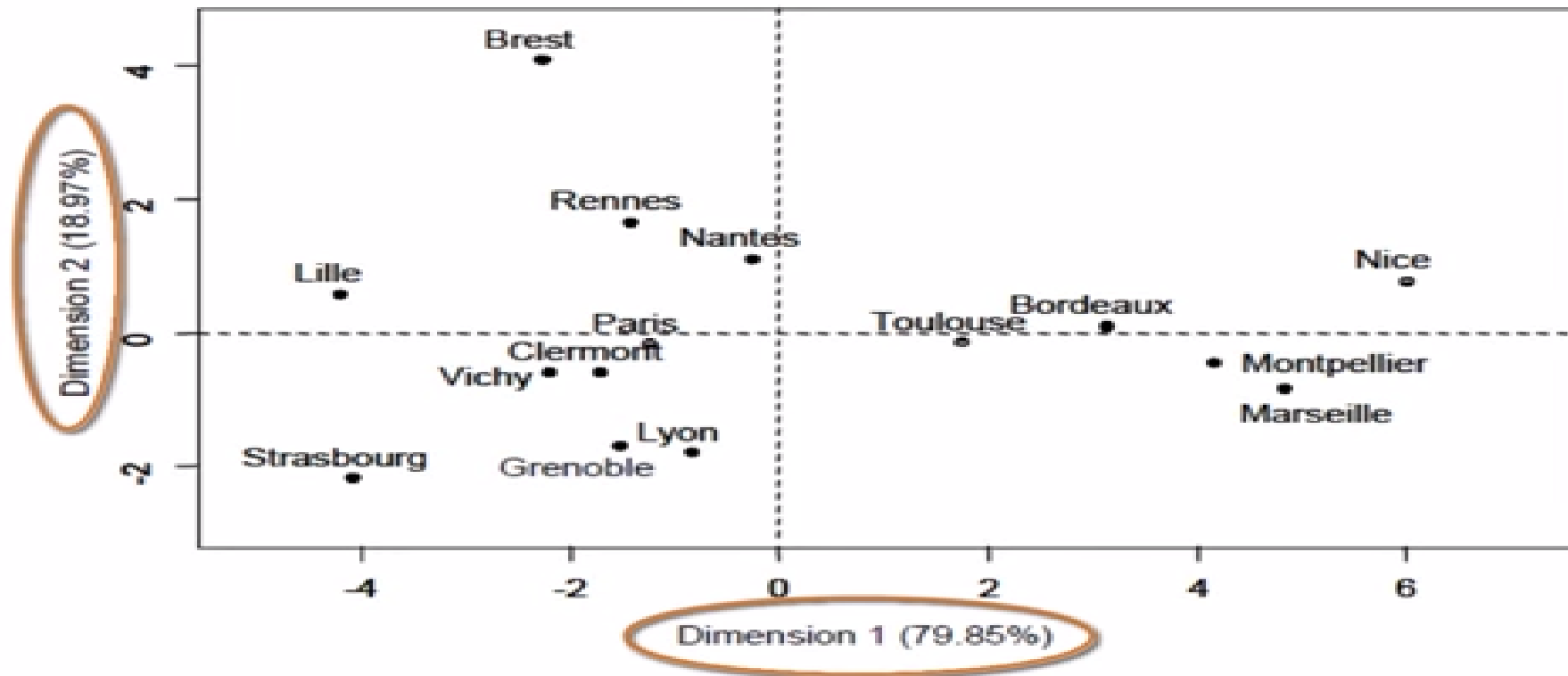
## Pourcentage d'inertie

- Pourcentage d'information (d'inertie) expliqué par chaque axe



⇒ Choix d'un nombre de dimensions à interpréter

## Exemple : graphe des individus



Comment interpréter les axes ? Qu'est-ce qui oppose Lille à Nice ?  
 ⇒ Besoin de variables pour interpréter ces dimensions de variabilité

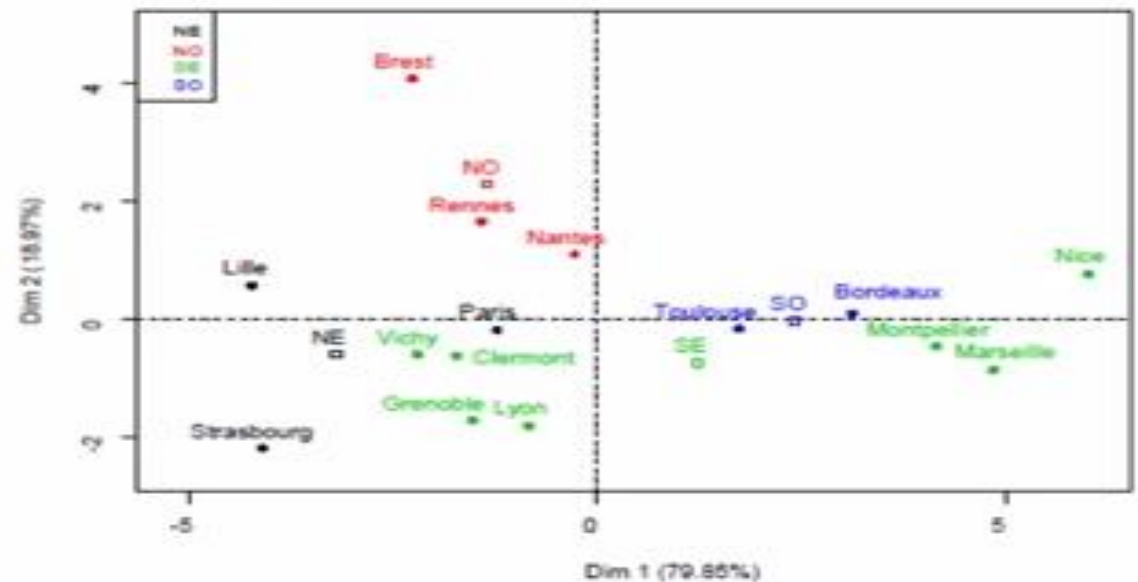
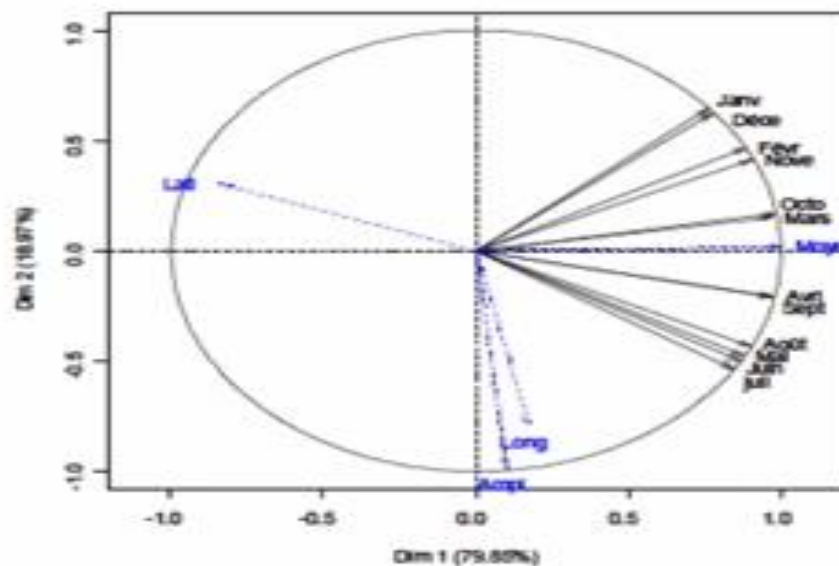
## Pourcentage d'inertie si indépendance entre variables

nbind	Nombre de variables												
	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

## Information supplémentaire

- Pour les variables quantitatives : projection des variables
- Pour les modalités : projection au barycentre des individus qui prennent cette modalité



⇒ Information supp. ne participe pas à la construction des axes



## Qualité de représentation – contribution

- Qualité de représentation d'une **variable** et d'un **individu**  
 $\cos^2$  entre une var. et sa projection       $\cos^2$  entre  $Oi$  et  $OH_i$

```
round(res.pca$var$cos2,2)
```

	Dim.1	Dim.2	Dim.3
Janv	0.58	0.42	0.00
Févr	0.78	0.22	0.00

```
round(res.pca$ind$cos2,2)
```

	Dim.1	Dim.2	Dim.3
Bordeaux	0.95	0.00	0.05
Brest	0.23	0.76	0.00

⇒ Seuls les éléments bien projetés peuvent être interprétés

- Contribution d'1 **var.** et d'1 **individu** à la construction de l'axe  $s$  :

$$Ctr_s(k) = \frac{r(x_k, v_s)^2}{\sum_{k=1}^K r(x_k, v_s)^2}$$

$$Ctr_s(i) = \frac{F_{is}^2}{\sum_{i=1}^I F_{is}^2}$$

```
round(res.pca$var$contrib,2)
```

	Dim.1	Dim.2	Dim.3
Janv	6.05	18.24	0.66
Févr	8.09	9.67	1.61

```
round(res.pca$ind$contrib,2)
```

	Dim.1	Dim.2	Dim.3
Bordeaux	6.78	0.03	49.48
Brest	3.58	49.07	1.26

⇒ Éléments avec une forte coordonnée contribuent le plus



## Description des dimensions

Par les variables quantitatives :

- calcul des corrélations entre chaque variable et la dimension  $s$
- tri des coefficients de corrélation (significatifs)

```
> dimdesc(res.pca)
```

```
$Dim.1$quanti
```

	correlation	p.value
Moye	0.9997097	0.000000e+00
Octo	0.9801599	1.609672e-10
Sept	0.9740289	9.130414e-10
Avri	0.9693357	2.657670e-09
Mars	0.9687704	2.988670e-09
Nove	0.9037531	3.834950e-06
...		
juil	0.8415346	8.385040e-05
Déce	0.7743349	7.017832e-04
Janv	0.7612384	9.784512e-04
Lati	-0.8389348	9.259113e-05

```
$Dim.2$quanti
```

	correlation	p.value
Janv	0.6443379	9.519348e-03
Déce	0.6242957	1.285835e-02
juil	-0.5314197	4.148657e-02
Long	-0.7922192	4.298867e-04
Ampl	-0.9856753	1.963381e-11



## Description des dimensions

Par les variables qualitatives :

- Analyse de variance des coordonnées des individus sur l'axe  $s$  (variable  $Y$ ) expliqués par la variable qualitative
  - un test  $F$  par variable
  - un test  $t$  de Student par modalité pour comparer la moyenne de la modalité avec la moyenne générale

```
> dimdesc(res.pca)
$Dim.2$quali
          R2    p.value
Région 0.6009012 0.01467946
```

```
$Dim.2$category
      Estimate    p.value
NO    2.0503647 0.003530801
SE   -0.9738852 0.047120253
```



## Pratique de l'ACP

- ① Choisir les variables actives
- ② Choisir de réduire ou non les variables
- ③ Réaliser l'ACP
- ④ Choisir le nombre de dimensions à interpréter
- ⑤ Interpréter simultanément le graphe des individus et celui des variables
- ⑥ Utiliser les indicateurs pour enrichir l'interprétation
- ⑦ Revenir aux données brutes pour interpréter

