Analyse en Composantes Principales

Slides 1ère année SN

Jean-Yves Tourneret[(1)] et Axel Carlier[(1)]

(1) Université de Toulouse, ENSEEIHT-IRIT
jyt@n7.fr, http://perso.tesa.prd.fr/jyt/, axel.carlier@toulouse-inp.fr

Année $2023 - 2024$

Bibliographie

Quelques références

- Thierry Foucart, L'analyse des Données - Mode d'emploi, Eyrolles, Paris, 1998.
- Gilbert Saporta, Probabilité, Analyse des Données et Statistique, Technip, Paris, 2nd edition, 2006.
- Ian Jolliffe, Principal Component Analysis, Springer-Verlag, New-York, 2nd edition, 2002.
- Vidéo François Husson, https://www.youtube.com/watch?v=8qw0bNfK4H0

Tableau de données #1

Poids, tailles, âges et notes ($p = 4$ variables) de $n = 10$ individus

	Poids	Taille	Age	Note		Poids	Taille	Age	Note
$\boldsymbol{x}_1$	45	1.50	13	14	$\boldsymbol{x}_6$	60	1.70	14	7
$\boldsymbol{x}_2$	50	1.60	13	16	$\boldsymbol{x}_7$	70	1.60	14	8
$\boldsymbol{x}_3$	50	1.65	13	15	$\boldsymbol{x}_8$	65	1.60	13	13
$\boldsymbol{x}_4$	60	1.70	15	9	$\boldsymbol{x}_9$	60	1.55	15	17
$\boldsymbol{x}_5$	60	1.70	14	10	$\boldsymbol{x}_{10}$	65	1.70	14	11

Thierry Foucart, L'analyse des Données - Mode d'emploi, Eyrolles, Paris, 1998.

Tableau de données #2

- 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

Vidéo François Husson, https://www.youtube.com/watch?v=8qw0bNfK4H0

Problèmes

- Représentation et visualisation des données sous la forme de graphiques simples
- Étude des individus
  - Certains individus se ressemblent-ils ?
  - Peut-on faire un bilan des ressemblances ?
  - Comment construire des groupes d'individus ?
- Étude des variables
  - Certaines variables se ressemblent-elles ?
  - Certains variables sont-elles liées ?
  - Quelles variables sont responsables des groupes d'individus ?

Plan du cours

Résumé

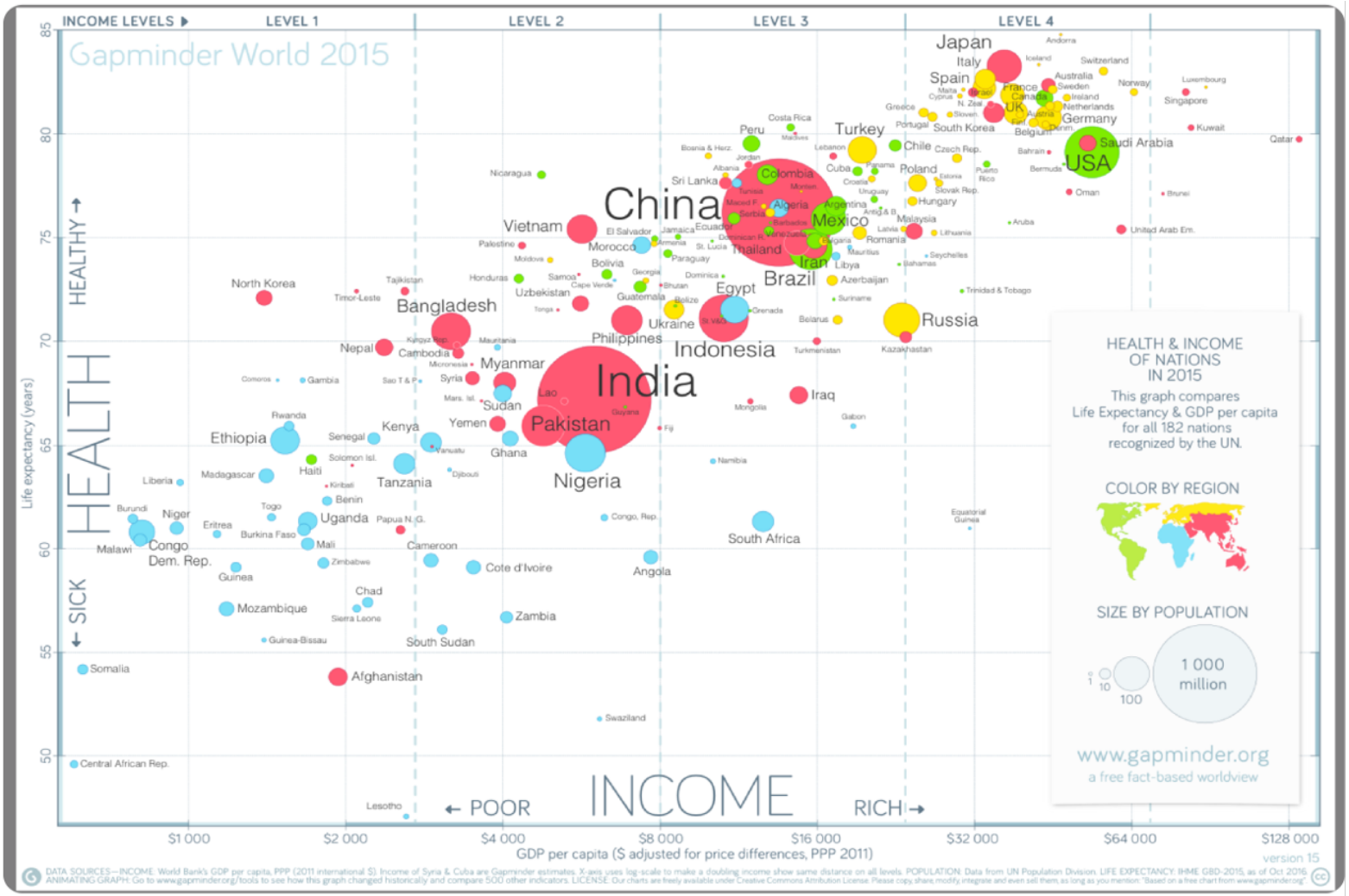
- Visualisation des données
- ACP des individus
- ACP des variables
- Exemples
- Exercice

Comment visualiser les données ?

Outils existants

- en $1D$: représentation axiale
- en $2D$: nuage de points
- en $3D$: plus difficile mais possibilité de tourner autour du nuage de points (visualisation.m)
- en dimension supérieure ?

Évolution de l'espérance de vie et du revenu en fonction du temps



Société Gapminder, https://www.gapminder.org

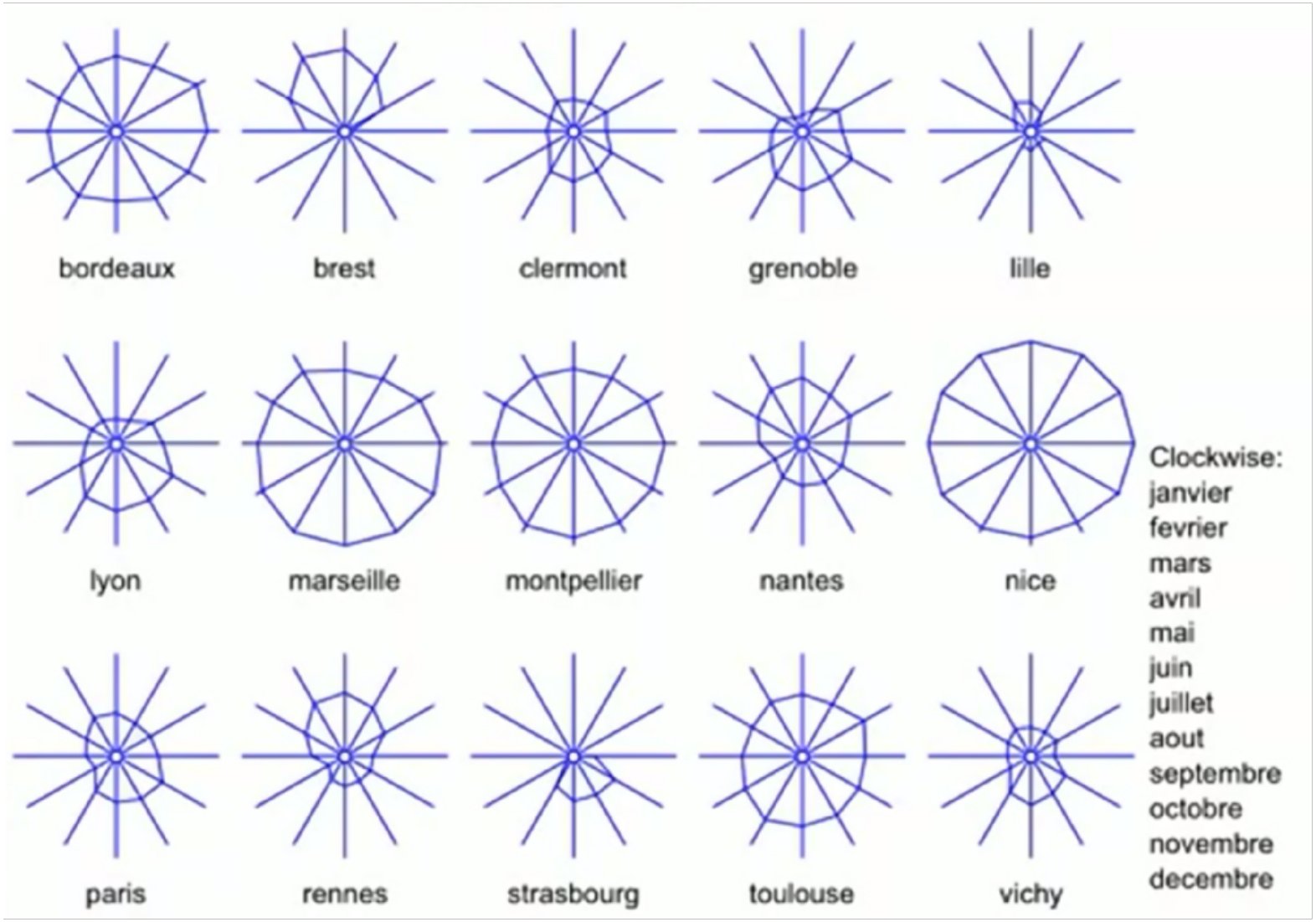
Tableau de données #2

- 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

Vidéo François Husson, https://www.youtube.com/watch?v=8qw0bNfK4H0

Araignées


bordeaux
brest
clermont
grenoble
lille
lyon
marseille
montpellier
nantes
nice
Clockwise:
janvier
fevrier
mars
avril
mai
juin
juillet
aout
septembre
octobre
novembre
decembre
paris
rennes
strasbourg
toulouse
vichy

Joueurs de foot

Quelle est la meilleure projection?

FIGURE: Quel animal ?

Quelle est la meilleure projection?

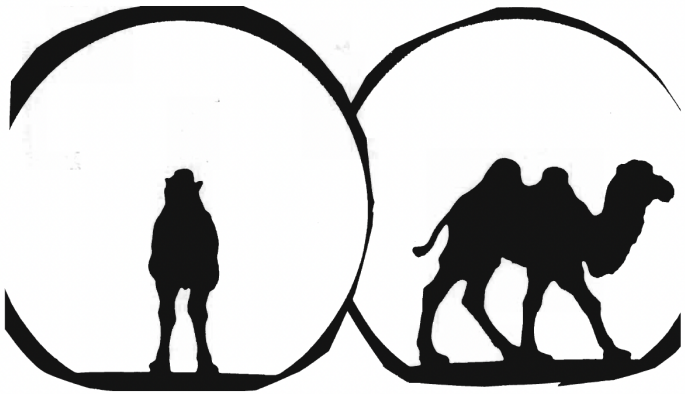
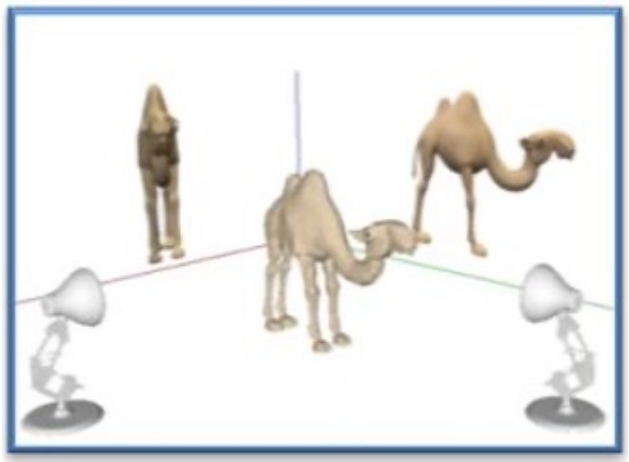
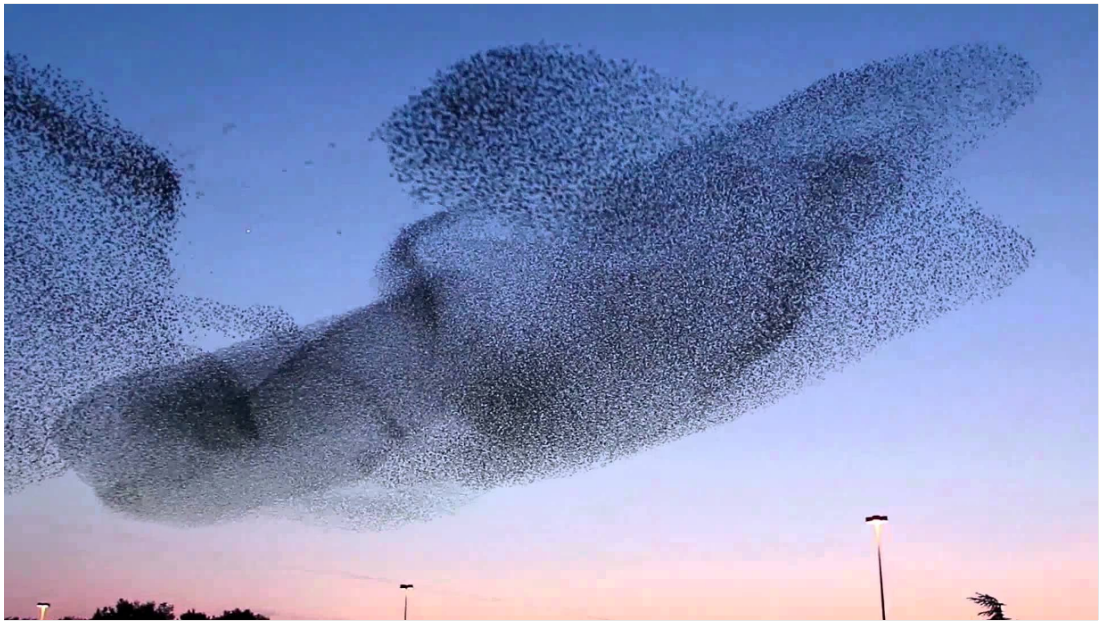


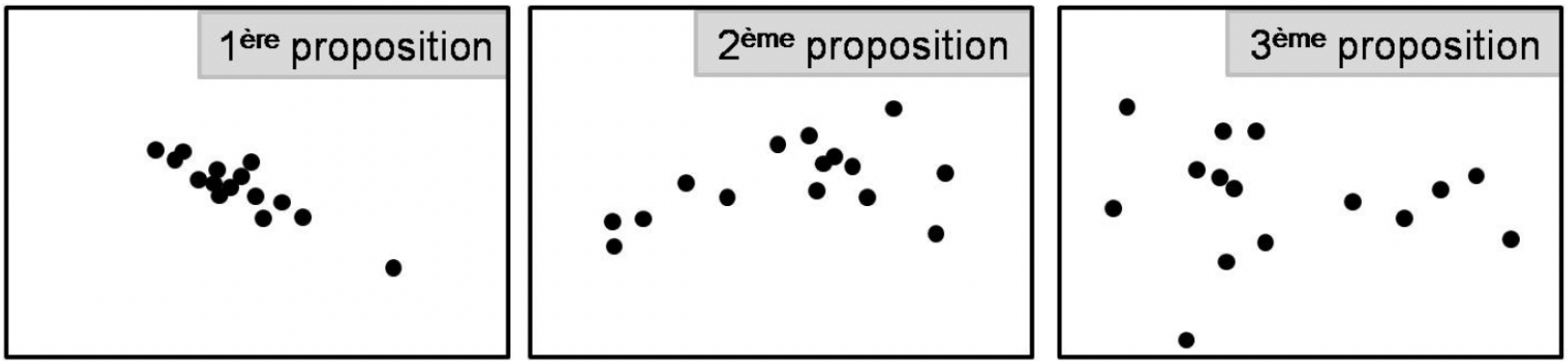
FIGURE – Quel animal ? (*illustration JP Fénelon*)



Vidéo François Husson, https://www.youtube.com/watch?v=8qw0bNfK4H0

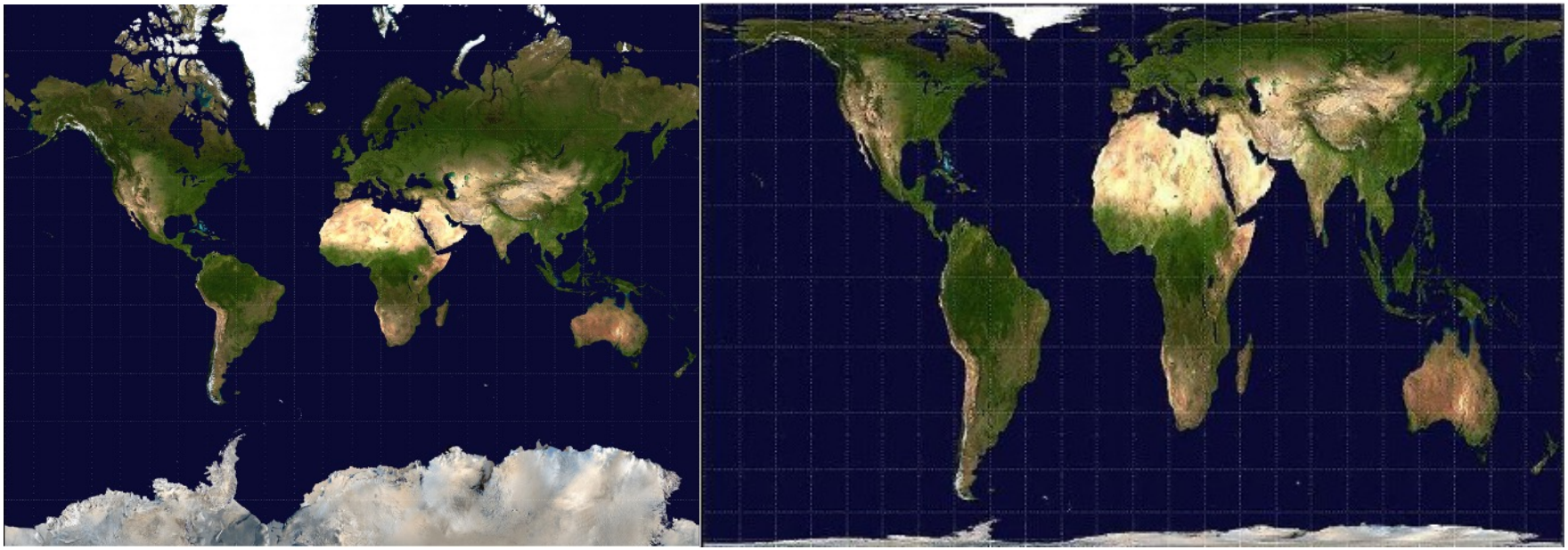
Quelle est la meilleure projection?





Vidéo François Husson, https://www.youtube.com/watch?v=8qw0bNfK4H0

Projections 2D du globe



Projections de Mercator (conservation des angles) et de Arno Peters (conservation des aires)

Plan du cours

Résumé

- Visualisation des données
- ACP des individus
- ACP des variables
- Exemples
- Exercice

Tableau de données #1

Poids, tailles, âges et notes de 10 individus

	Poids	Taille	Age	Note		Poids	Taille	Age	Note
$\boldsymbol{x}_1$	45	1.50	13	14	$\boldsymbol{x}_6$	60	1.70	14	7
$\boldsymbol{x}_2$	50	1.60	13	16	$\boldsymbol{x}_7$	70	1.60	14	8
$\boldsymbol{x}_3$	50	1.65	13	15	$\boldsymbol{x}_8$	65	1.60	13	13
$\boldsymbol{x}_4$	60	1.70	15	9	$\boldsymbol{x}_9$	60	1.55	15	17
$\boldsymbol{x}_5$	60	1.70	14	10	$\boldsymbol{x}_{10}$	65	1.70	14	11

Thierry Foucart, L'analyse des Données - Mode d'emploi, Eyrolles, Paris, 1998.

ACP des individus

Première étape : définition d'une norme

$$\|\boldsymbol{x}\|_{\boldsymbol{M}}^2 = \langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\boldsymbol{M}} = \boldsymbol{x}^T \boldsymbol{M} \boldsymbol{x}$$

$\boldsymbol{M}$ matrice symétrique définie positive de taille $p \times p$

- $\boldsymbol{M} = \boldsymbol{I}_p$

$$d^2(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j=1}^{p} [x(j) - y(j)]^2$$

- $\boldsymbol{M} = \text{diag}\left(\frac{1}{\sigma_1^2}, ..., \frac{1}{\sigma_p^2}\right)$

$$d^2(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j=1}^{p} [x^*(j) - y^*(j)]^2$$

où $x^*(j) = \frac{x(j) - m(j)}{\sigma(j)}$ (données centrées réduites).

Dans la suite, on centre toujours les données (ne change pas la forme du nuage de points) et on réduit parfois les données, ce qui revient à choisir $\boldsymbol{M} = \boldsymbol{I}_p$ après normalisation.

Quand faut-il centrer et réduire les données ?

Centrer

Il faut toujours centrer les données

- Ca ne change pas la forme du nuage de points
- Le nuage est translaté autour de sa valeur moyenne

Réduire

- Indispensable si les unités de mesure des variables sont différentes
- Optionnel si les unités de mesure des variables sont les mêmes

ACP des individus

On cherche un espace de dimension q qui résume au mieux les données.

Deuxième étape : optimisation

$$\text{Minimiser } I_q = \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{y}_i\|^2 \Leftrightarrow \text{Maximiser } J_q = \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{y}_i\|^2$$

où $\boldsymbol{y}_i$ est la projection de $\boldsymbol{x}_i$ dans l'espace de dimension q recherché.

Propriété : les solutions sont emboitées d'où $J(\boldsymbol{u}) = \boldsymbol{u}^T \boldsymbol{\Sigma} \boldsymbol{u}$, *où* $\boldsymbol{u} \in \mathbb{R}^p$ *et* $\boldsymbol{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T$ *est la matrice de covariance des vecteurs* $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$.

- **Détermination des axes principaux**
  Optimisation du Lagrangien

$$L(\boldsymbol{u}) = \boldsymbol{u}^T \boldsymbol{\Sigma}\ \boldsymbol{u} - \lambda(\boldsymbol{u}^T \boldsymbol{u} - 1)$$

ACP des individus

Inertie et composantes principales

- **Nombre d'axes principaux**

  $\mathbf{\Sigma}$ de taille $p \times p$ inversible $\Longrightarrow$ p axes principaux

- **Choix du nombre de vecteurs**

$$I_q = \frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{x}_i\|^2 \left[1 - \frac{\sum_{j=1}^{q} \lambda_j}{\sum_{j=1}^{p} \lambda_j} \right]$$

- **Inertie**

$$\frac{\lambda_j}{\sum_{j=1}^{p} \lambda_j}$$

  est l'inertie du $j$ème axe.

- **Composantes principales**

  Les q nouvelles variables sont appelées composantes principales.

Optimisation sous contraintes égalités

Problème $\mathcal{P}$

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) \text{ s.c. } g_k(\boldsymbol{x}) = 0, \forall k = 1, ..., l$$

s.c. = sous les contraintes.

Conditions de Kuhn et Tucker

Si les fonctions $f, g_1, ..., g_l$ sont différentiables dans un voisinage de la solution $\boldsymbol{x}^*$ et si la matrice $\boldsymbol{G}^* = [\nabla g_1(\boldsymbol{x}^*), ..., \nabla g_l(\boldsymbol{x}^*),]$ est de rang maximal, des conditions nécessaires d'optimalité sont

$$\frac{\partial L}{\partial x_i} = 0, \ \forall i = 1, ..., n \quad \text{et} \quad \frac{\partial L}{\partial \lambda_k} = 0, \ \forall k = 1, ..., l$$

où

$$L(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) - \sum_{k=1}^{l} \lambda_k g_k(\boldsymbol{x}), \quad \lambda_k > 0$$

est le Lagrangien du problème $\mathcal{P}$. Il suffit donc de résoudre ce système de $n + l$ équations à $n + l$ inconnues pour déterminer les solutions potentielles du problème $\mathcal{P}$.
Dans le cas d'une seule contrainte $g(\boldsymbol{x}) = 0$ $(l = 1)$, $\boldsymbol{G}^*$ est de rang 1 si et ssi $\nabla g(\boldsymbol{x}^*) \neq 0$.

Exemple 1

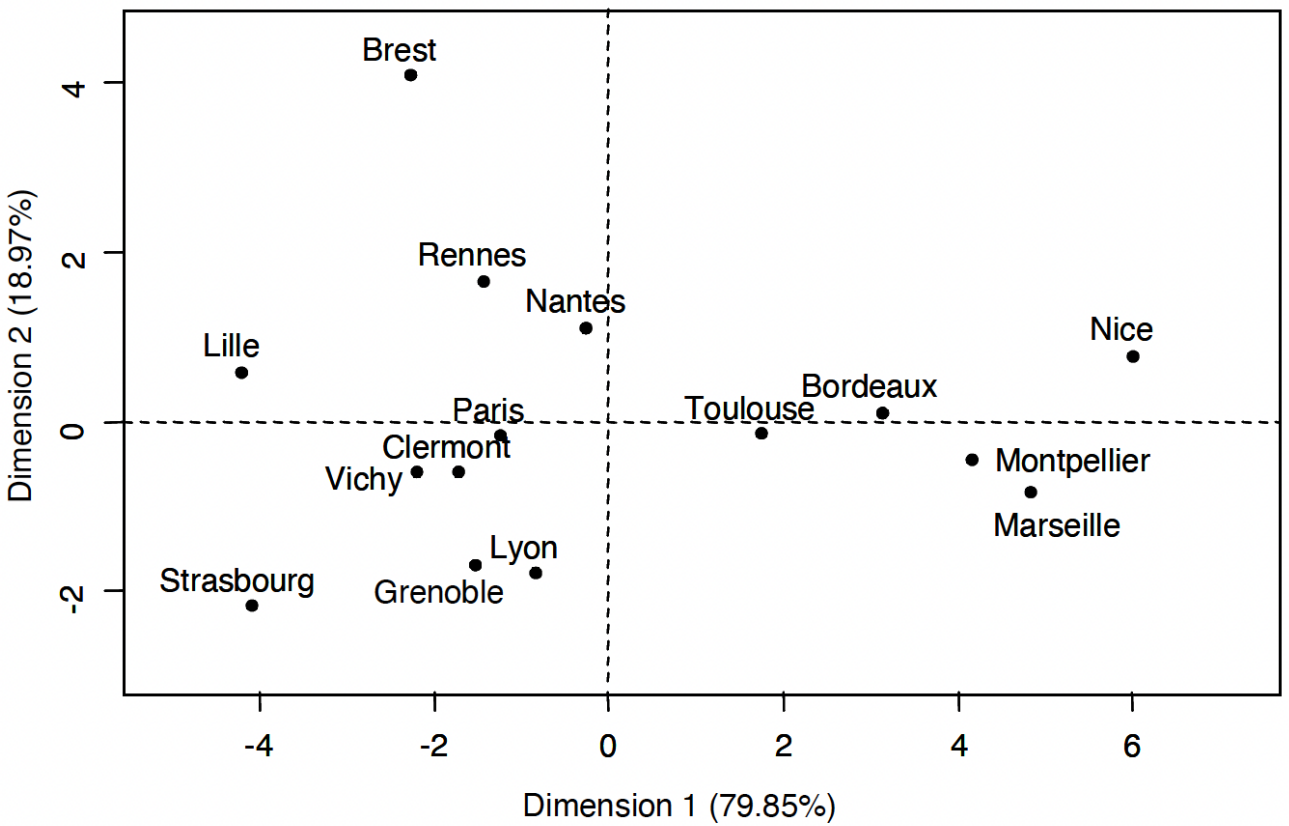
- 15 individus (lignes) : villes de France
- 14 variables (colonnes) :
  - 12 températures mensuelles moyennes (sur 30 ans)
  - 2 variables géographiques (latitude, longitude)

	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

Vidéo François Husson, https://www.youtube.com/watch?v=8qw0bNfK4H0

ACP des individus (centrés et réduits)

Représentation 2D

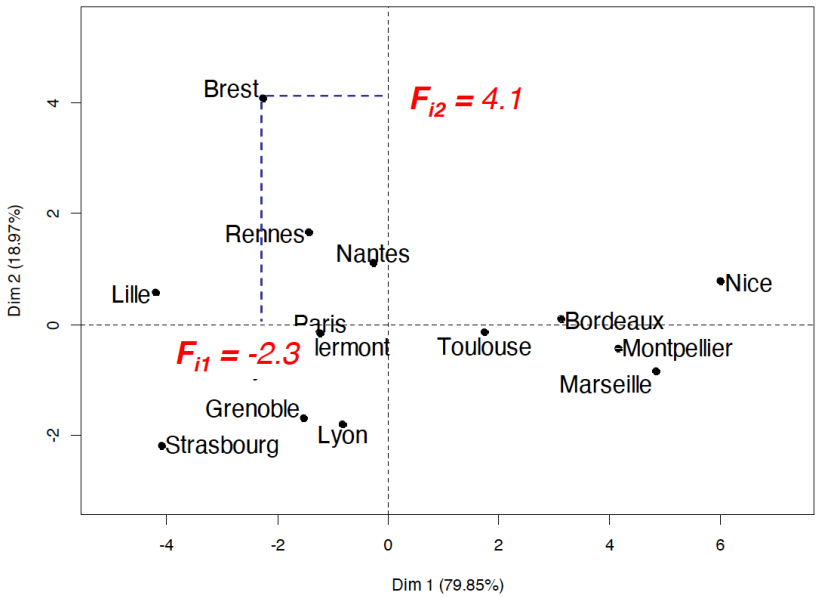


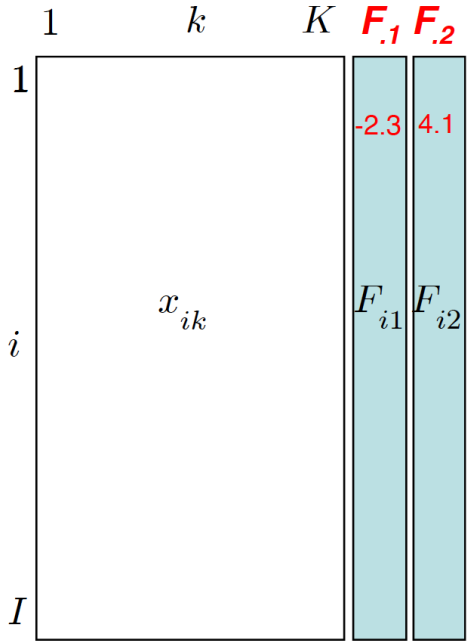
Vidéo François Husson, https://www.youtube.com/watch?v=8qw0bNfK4H0

Cercle des corrélations

Définition

Considérons les coordonnées des individus sur les axes comme des variables



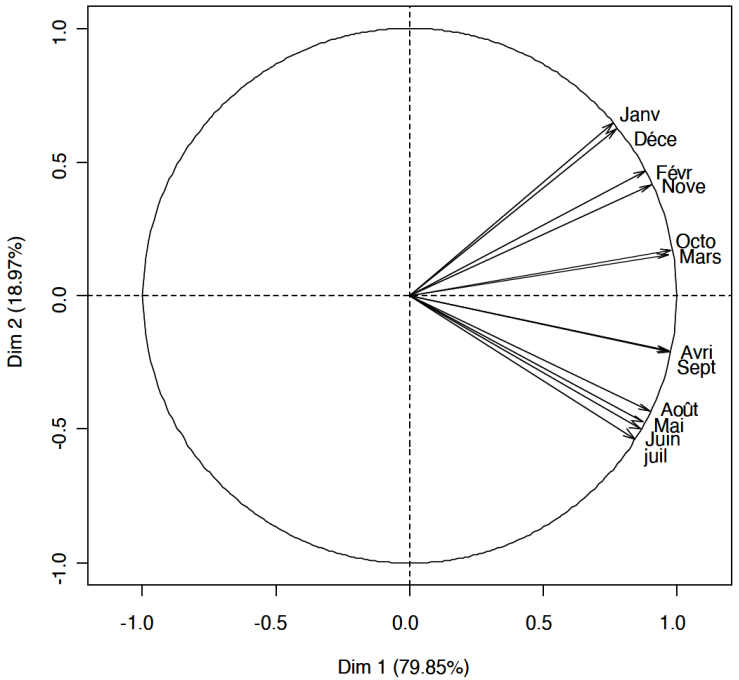


Vidéo François Husson, https://www.youtube.com/watch?v=8qw0bNfK4H0

Cercle des corrélations

Exemple

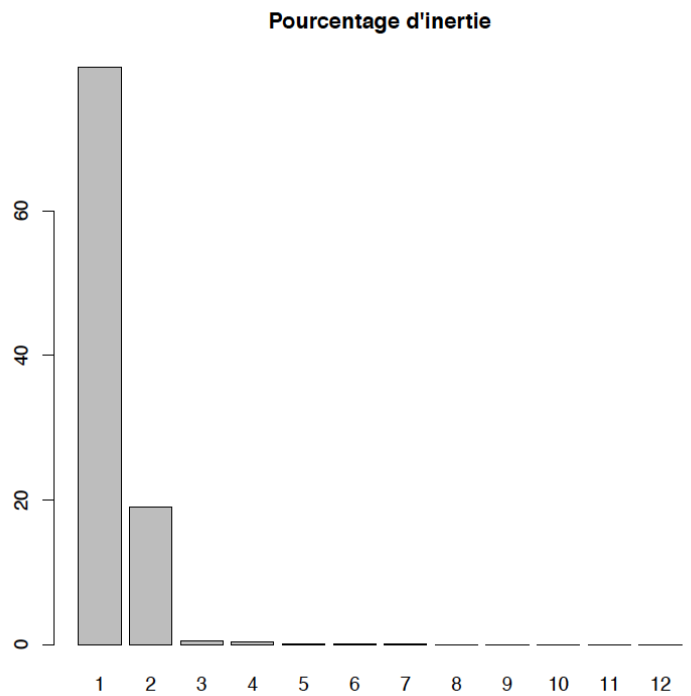
Interprétation du graphe des individus grâce aux variables



Vidéo François Husson, https://www.youtube.com/watch?v=8qw0bNfK4H0

ACP des individus

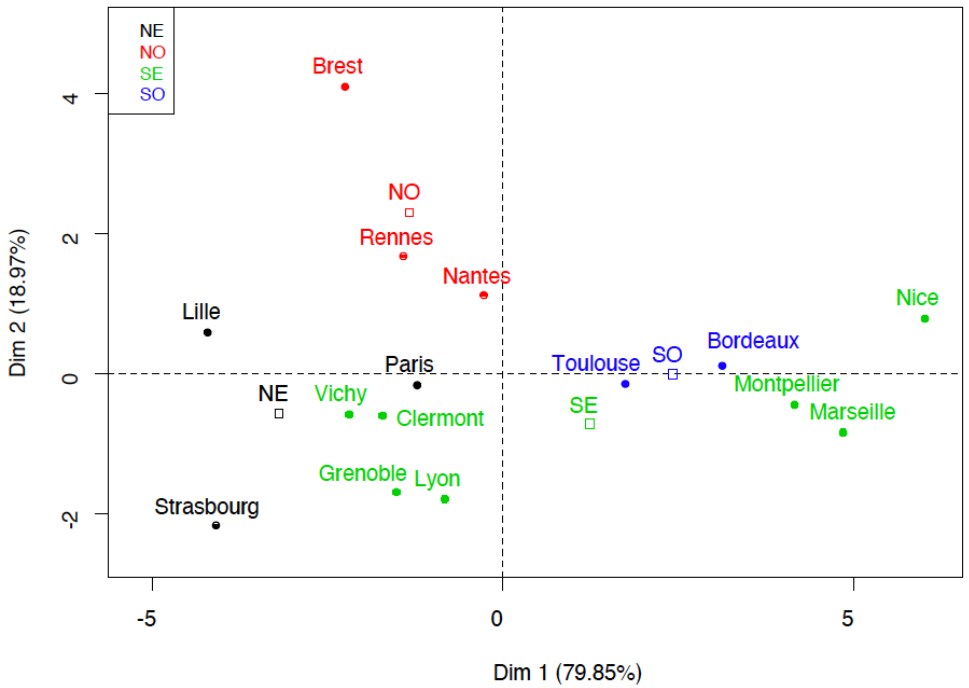
Inerties des axes



Vidéo François Husson, https://www.youtube.com/watch?v=8qw0bNfK4H0

ACP des individus

Variables supplémentaires qualitatives liées à la région : NE, NO, SE, SO



Vidéo François Husson, https://www.youtube.com/watch?v=8qw0bNfK4H0

ACP des individus

Contribution de l'individu $\#i$ à la construction de l'axe s

$$\frac{F_{is}^2}{\sum_{i=1}^n F_{is}^2} = \frac{F_{is}^2}{n\lambda_s}$$

car $\sum_{i=1}^n F_{is}^2 = \sum_{i=1}^n (\boldsymbol{x}_i^T \boldsymbol{u}_s)^2 = n\boldsymbol{u}_s^T \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i^T \boldsymbol{x}_i\right) \boldsymbol{u}_s = n\boldsymbol{u}_s^T \boldsymbol{\Sigma} \boldsymbol{u}_s = n\lambda_s$

Qualité de représentation de l'individu $\#i$ sur l'axe s

$$\cos^2(\theta_{is}) = \frac{F_{is}^2}{\|\boldsymbol{x}_i\|^2}$$

Plan du cours

Résumé

- Visualisation des données
- ACP des individus
- ACP des variables
- Exemples
- Exercice

ACP des variables

- **ACP Normée**

$$v'_j(i) = \frac{x_i(j) - m(j)}{\sqrt{n}\sigma(j)}$$

On a donc

$$\boldsymbol{v}'_j = \begin{pmatrix} \frac{x_1(j)-m(j)}{\sigma(j)\sqrt{n}} \\ \vdots \\ \frac{x_n(j)-m(j)}{\sigma(j)\sqrt{n}} \end{pmatrix} \text{ avec } \|\boldsymbol{v}'_j\| = 1.$$

- **Axes principaux**

Si $\boldsymbol{u}$ est un vecteur propre unitaire de $\boldsymbol{\Sigma} = \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}$ avec la valeur propre λ, alors $\boldsymbol{v} = \frac{\boldsymbol{X}\boldsymbol{u}}{\sqrt{n\lambda}}$ est un vecteur propre unitaire de $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^T$.

En effet $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{u} = \boldsymbol{X}(\lambda\boldsymbol{u})$ et $\|\boldsymbol{X}\boldsymbol{u}\|^2 = \boldsymbol{u}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{u} = n\lambda\boldsymbol{u}^T\boldsymbol{u} = n\lambda$.

- **Nombre d'axes principaux**

Matrice de covariance de taille $N \times N$ avec p axes principaux (p valeurs propres > 0 (non nulles))

ACP des variables

Cosinus entre deux projections

$$\cos(\boldsymbol{v}'_k, \boldsymbol{v}'_j) = \frac{\langle \boldsymbol{v}'_k, \boldsymbol{v}'_j \rangle}{\|\boldsymbol{v}'_k\| \|\boldsymbol{v}'_j\|} = r_{jk}$$

où r_{jk} est le coefficient de corrélation entre les variables $\boldsymbol{v}'_k$ et $\boldsymbol{v}'_j$. Donc

- Si l'angle entre les projections $\boldsymbol{v}'_k$ et $\boldsymbol{v}'_j$ est proche de 0, on a $\cos(\boldsymbol{v}'_k, \boldsymbol{v}'_j) = 1$ et donc les variables k et j sont très liées (positivement corrélées)
- Si l'angle entre les projections $\boldsymbol{v}'_k$ et $\boldsymbol{v}'_j$ est proche de $\frac{\pi}{2}$, on a $\cos(\boldsymbol{v}'_k, \boldsymbol{v}'_j) = 0$ et donc les variables k et j sont peu liées (décorrélées)
- Si l'angle entre les projections $\boldsymbol{v}'_k$ et $\boldsymbol{v}'_j$ est proche de π, on a $\cos(\boldsymbol{v}'_k, \boldsymbol{v}'_j) = -1$ et donc les variables k et j sont très liées (négativement corrélées)

ACP des variables

Contribution de la variable $\boldsymbol{v}_k$ à la construction de l'axe a_s

$$\frac{r^2(\boldsymbol{v}_k, a_s)}{\sum_{k=1}^{p} r^2(\boldsymbol{v}_k, a_s)}$$

Si $\boldsymbol{\phi}_k$ est le vecteur contenant les corrélations entre la variable $\boldsymbol{v}_k$ et les p axes principaux, alors on a

$$\boldsymbol{\phi}_k = \begin{pmatrix} r(\boldsymbol{v}_k, \boldsymbol{a}_1) \\ r(\boldsymbol{v}_k, \boldsymbol{a}_2) \\ \vdots \\ r(\boldsymbol{v}_k, \boldsymbol{a}_p) \end{pmatrix} = \sqrt{\lambda_k}\boldsymbol{u}_k,$$

où $\boldsymbol{u}_k$ est le vecteur propre de $\boldsymbol{\Sigma}$ avec la valeur propre λ_k.

Qualité de représentation de la variable $\boldsymbol{v}_k$ sur l'axe a_s

$$\cos^2(\theta'_{ks}) = \frac{r^2(\boldsymbol{v}_k, a_s)}{\|\boldsymbol{v}_k\|^2} = \frac{\lambda_k u_{ks}^2}{\lambda_k \|\boldsymbol{u}_k\|^2} = \frac{u_{ks}^2}{\|\boldsymbol{u}_k\|^2} = u_{ks}^2.$$

Projections des variables sur les axes de l'ACP

Énoncé

Si $\boldsymbol{\phi}_k$ est le vecteur contenant les corrélations entre la variable $\boldsymbol{v}_k$ (normalisée par $\sqrt{n}$) et les p axes principaux, alors on a

$$\boldsymbol{\phi}_k = \begin{pmatrix} r(\boldsymbol{v}_k, \boldsymbol{a}_1) \\ \vdots \\ r(\boldsymbol{v}_k, \boldsymbol{a}_p) \end{pmatrix} = \sqrt{\lambda_k}\boldsymbol{u}_k,$$

où $\boldsymbol{u}_k$ est le vecteur propre de $\boldsymbol{\Sigma}$ avec la valeur propre λ_k.

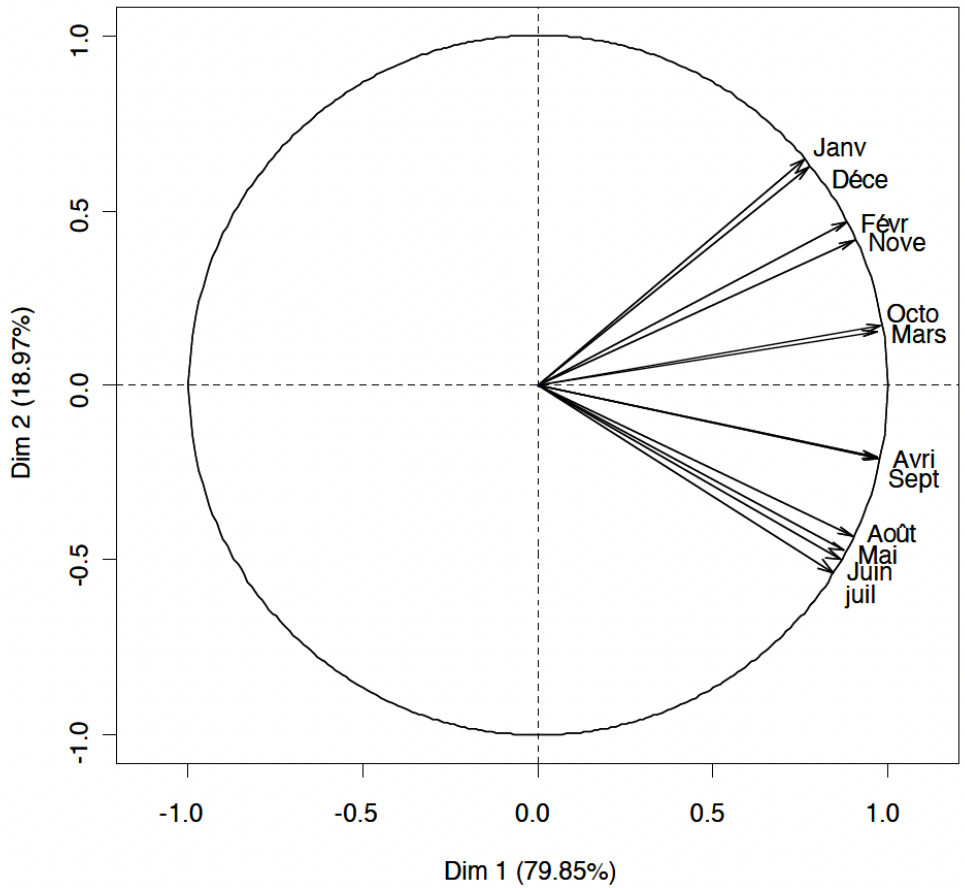
Preuve

Si $\boldsymbol{a}$ est un vecteur propre de $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^T$ (matrice de l'ACP des variables) avec la valeur propre λ, alors $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{v} = \lambda\boldsymbol{v}$, donc $\boldsymbol{X}^T\left(\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^T\right)\boldsymbol{a} = \lambda\boldsymbol{X}^T\boldsymbol{a}$, donc $\boldsymbol{X}^T\boldsymbol{a}$ est un vecteur propre de $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}$ avec la valeur propre λ. Comme $\|\boldsymbol{X}^T\boldsymbol{a}\|^2 = \boldsymbol{a}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{a} = \lambda n$, le vecteur $\boldsymbol{u} = \frac{1}{\sqrt{\lambda}}\left(\frac{\boldsymbol{X}^T}{n}\right)\boldsymbol{v}$ est un vecteur propre unitaire de $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}$ avec la valeur propre λ. Donc

$$\left(\frac{\boldsymbol{X}^T}{n}\right)\boldsymbol{a} = \begin{pmatrix} r(\boldsymbol{v}_k, \boldsymbol{a}_1) \\ \vdots \\ r(\boldsymbol{v}_k, \boldsymbol{a}_p) \end{pmatrix} = \sqrt{\lambda}\boldsymbol{u}$$

ACP des variables

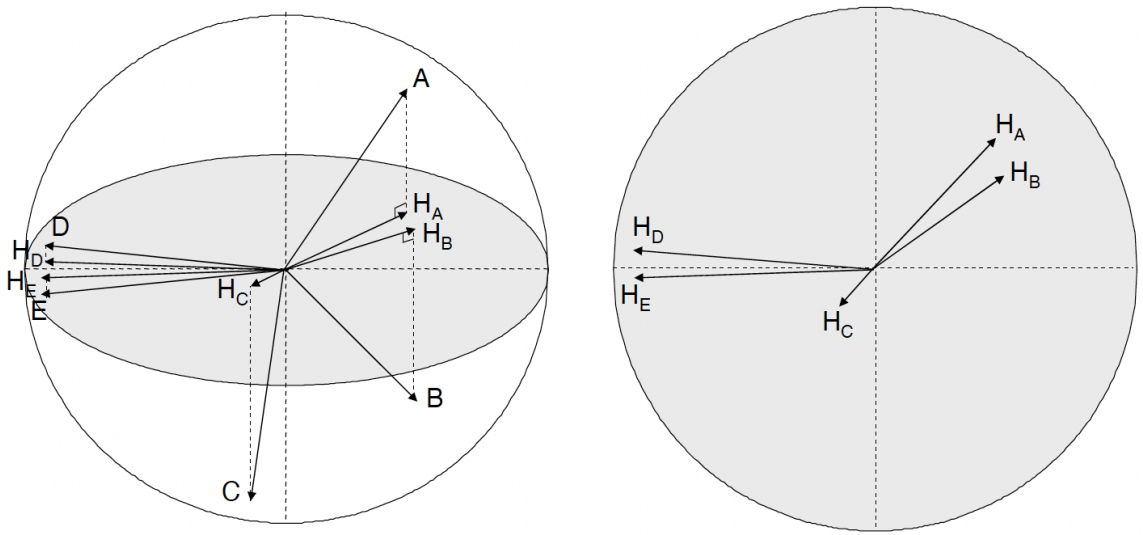
ACP des variables = cercle des corrélations entre F_1, F_2



Vidéo François Husson, https://www.youtube.com/watch?v=8qw0bNfK4H0

Qualité de la projection

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

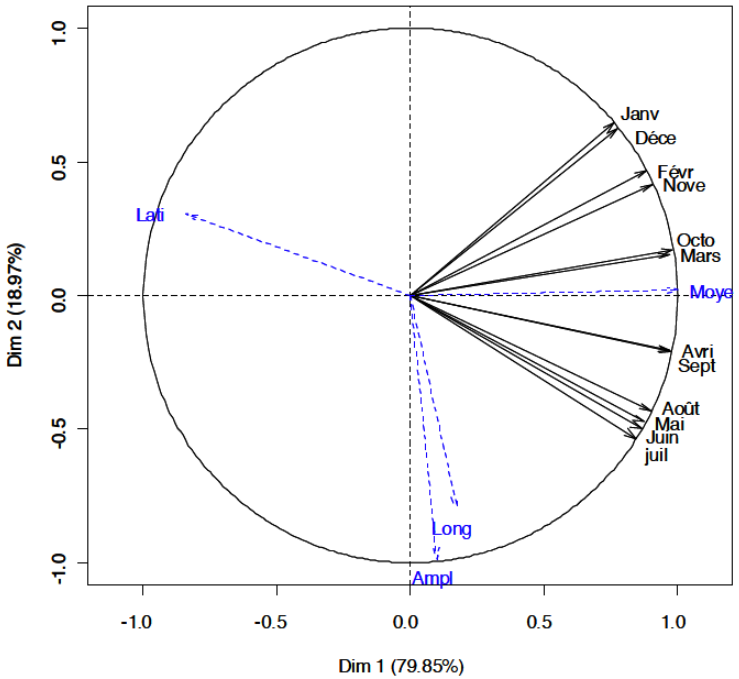


Seules les variables bien projetées peuvent être interprétées !

Vidéo François Husson, https://www.youtube.com/watch?v=8qw0bNfK4H0

ACP des variables

Variables supplémentaires quantitatives : latitude, longitude, température moyenne, amplitude thermique



Vidéo François Husson, https://www.youtube.com/watch?v=8qw0bNfK4H0

Qualité de représentation – contribution

- Qualité de représentation d'une variable et d'un individu

$\cos^2$ entre une var. et sa projection

```
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
```

$\cos^2$ entre Oi et OH_i

```
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 :

$$Ctr_s(k) = \frac{r(x_{.k}, v_s)^2}{\sum_{k=1}^{K} r(x_{.k}, v_s)^2} \ (\times 100) \qquad Ctr_s(i) = \frac{F_{is}^2}{\sum_{i=1}^{I} F_{is}^2} \ (\times 100)$$

```
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
```

⇒ Eléments avec une forte coordonnée contribuent le plus

Vidéo François Husson, https://www.youtube.com/watch?v=8qw0bNfK4H0

Plan du cours

Résumé

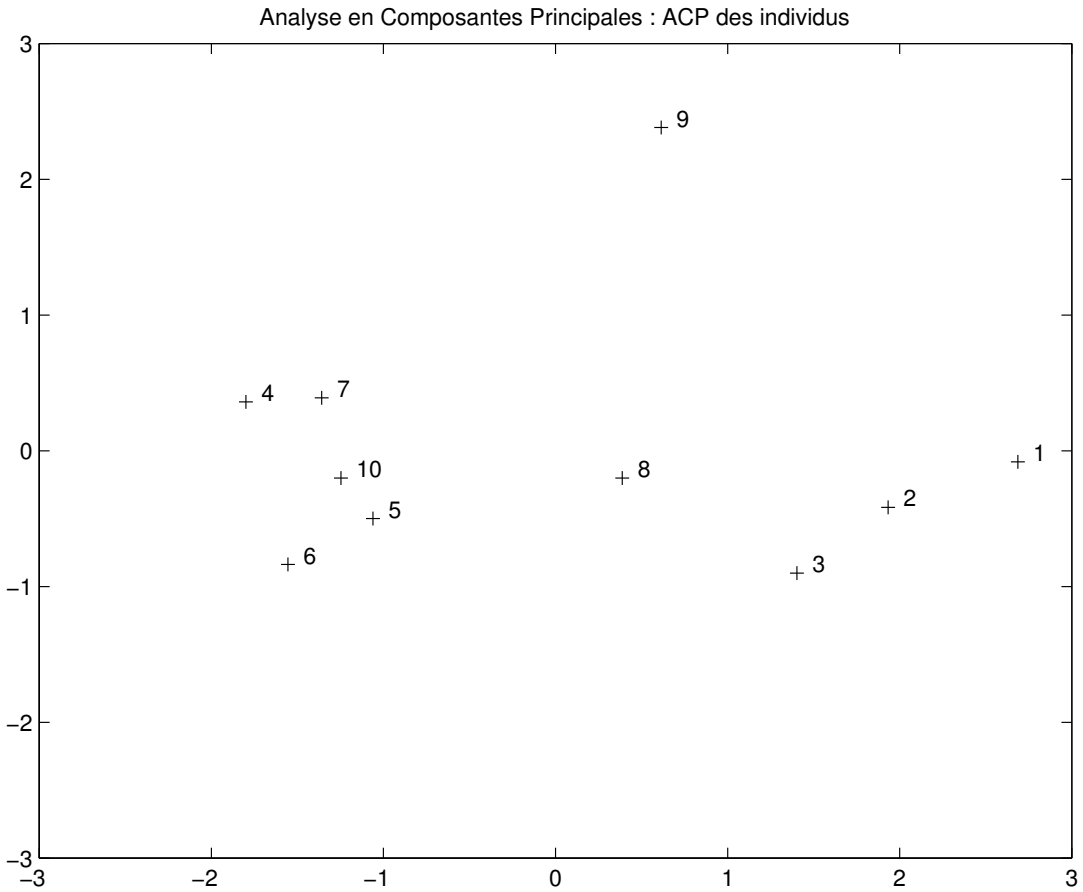
- Visualisation des données
- ACP des individus
- ACP des variables
- Exemples
- Exercice

Exemple 2

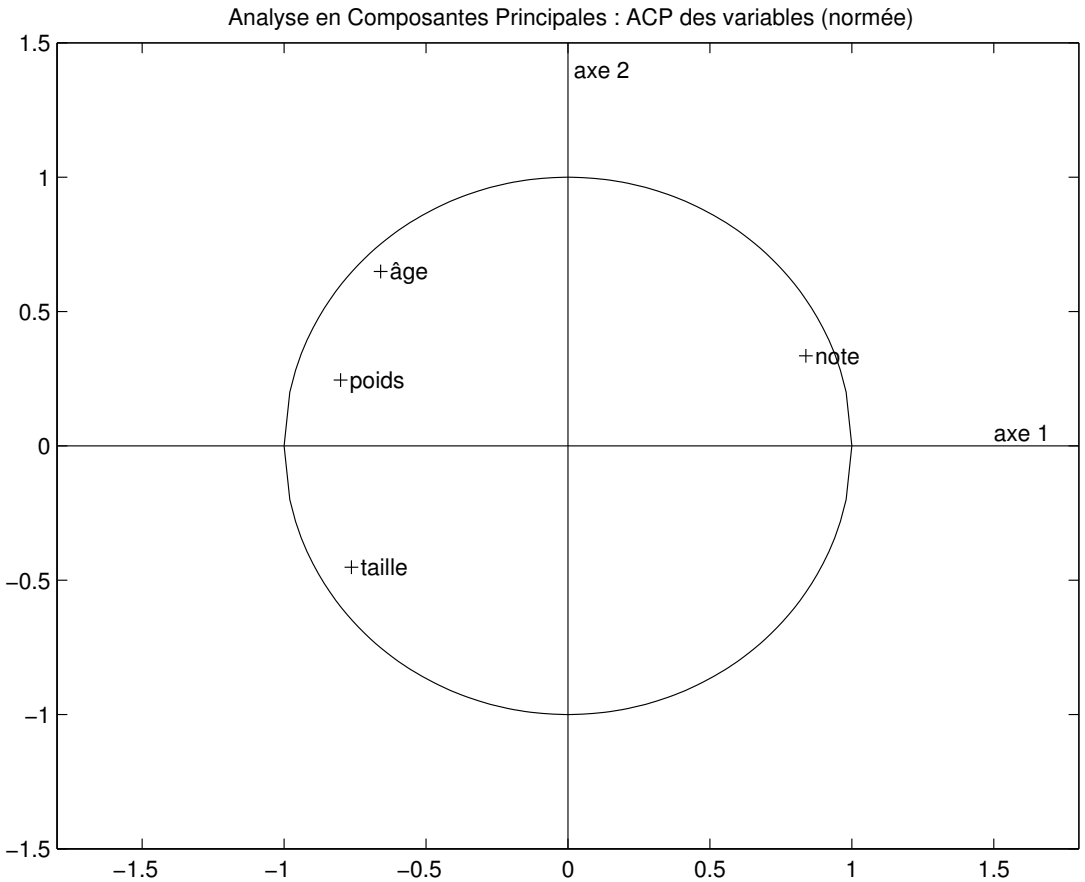
Poids, taille, âge et note de 10 individus

	Poids	Taille	Age	Note		Poids	Taille	Age	Note
$\boldsymbol{x}_1$	45	1.50	13	14	$\boldsymbol{x}_6$	60	1.70	14	7
$\boldsymbol{x}_2$	50	1.60	13	16	$\boldsymbol{x}_7$	70	1.60	14	8
$\boldsymbol{x}_3$	50	1.65	13	15	$\boldsymbol{x}_8$	65	1.60	13	13
$\boldsymbol{x}_4$	60	1.70	15	9	$\boldsymbol{x}_9$	60	1.55	15	17
$\boldsymbol{x}_5$	60	1.70	14	10	$\boldsymbol{x}_{10}$	65	1.70	14	11

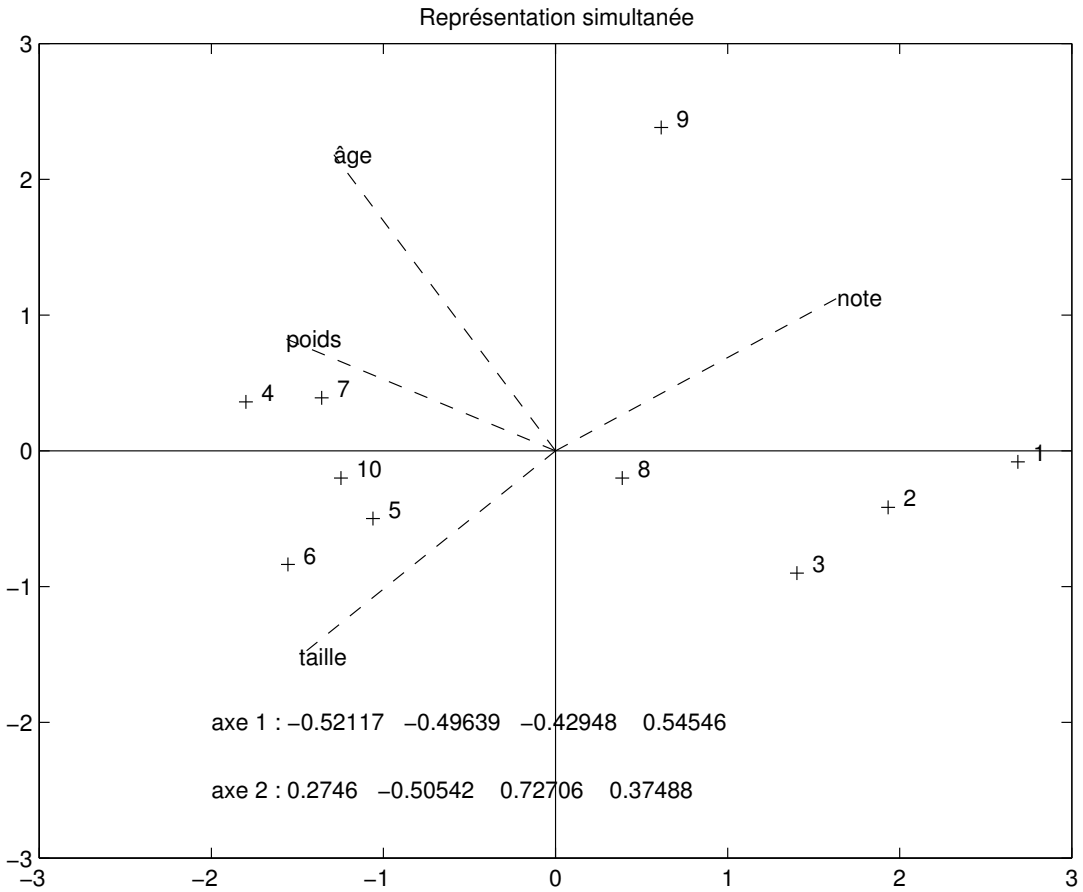
ACP des individus



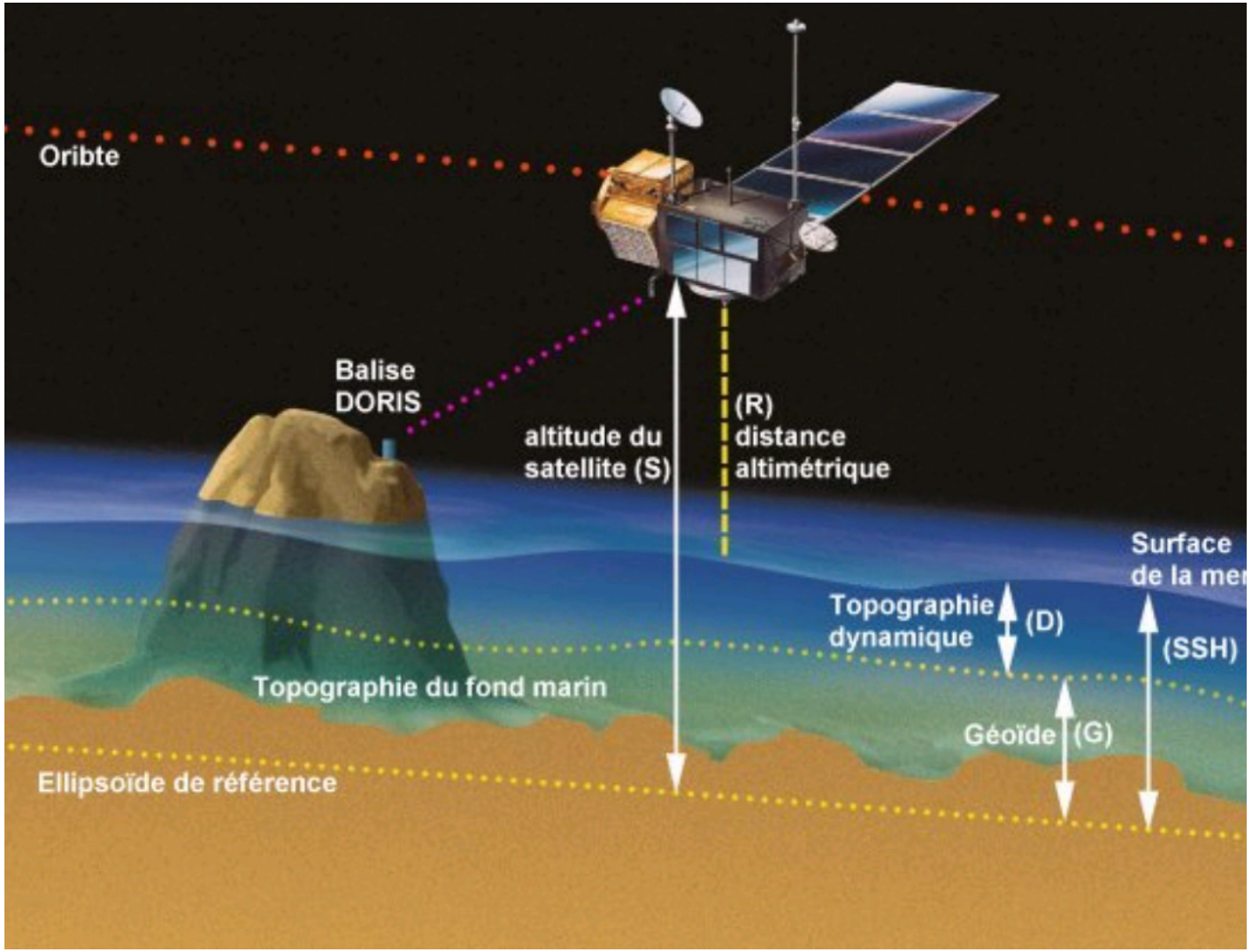
ACP des variables



ACP simultanée



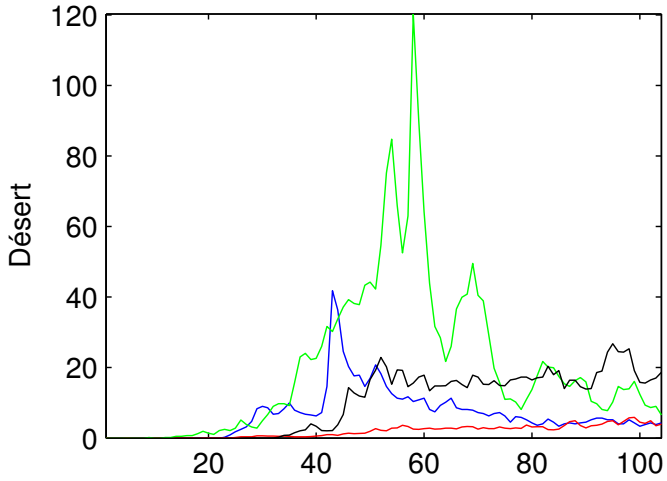
Application à l'altimétrie

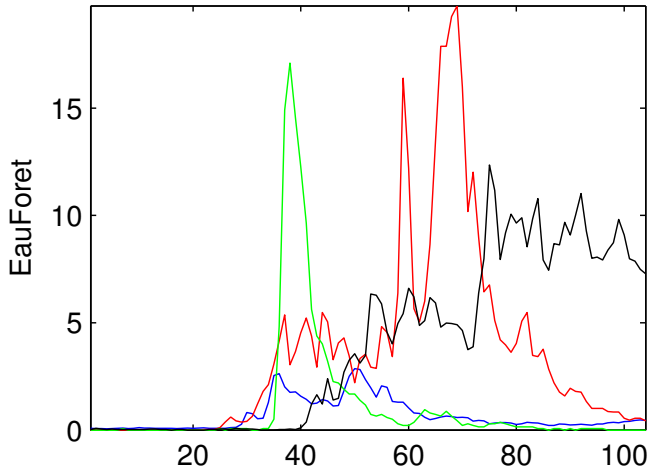


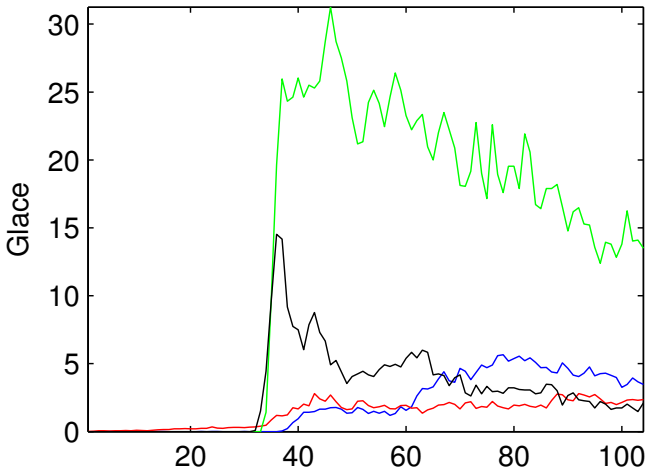
Classification

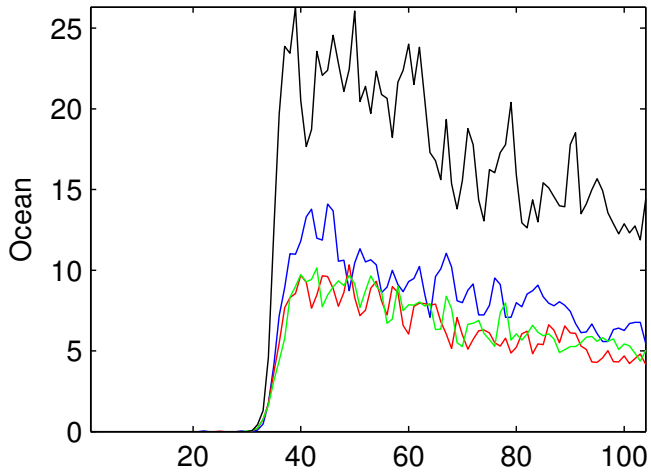
- Classe ω_1 : océans
- Classe ω_2 : déserts (Algérie, Lybie, Afrique du Sud)
- Classe ω_3 : eaux et forêts (Amazonie, Canada, Congo, Russie)
- Classe ω_4 : glaces (glace continentale arctique, glace continentale Groenland, glace mer antarctique, glace mer arctique)

Exemples de formes d'onde

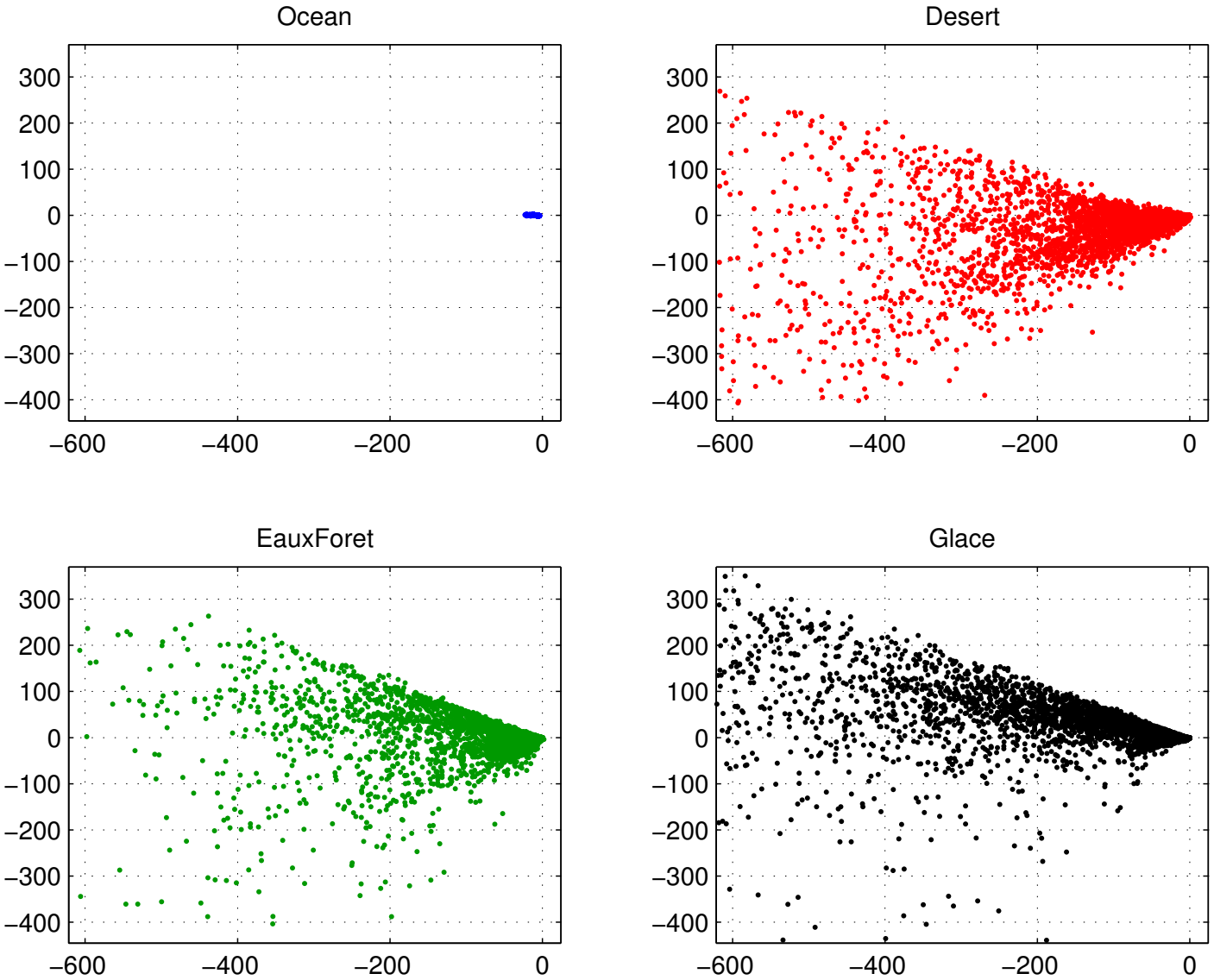




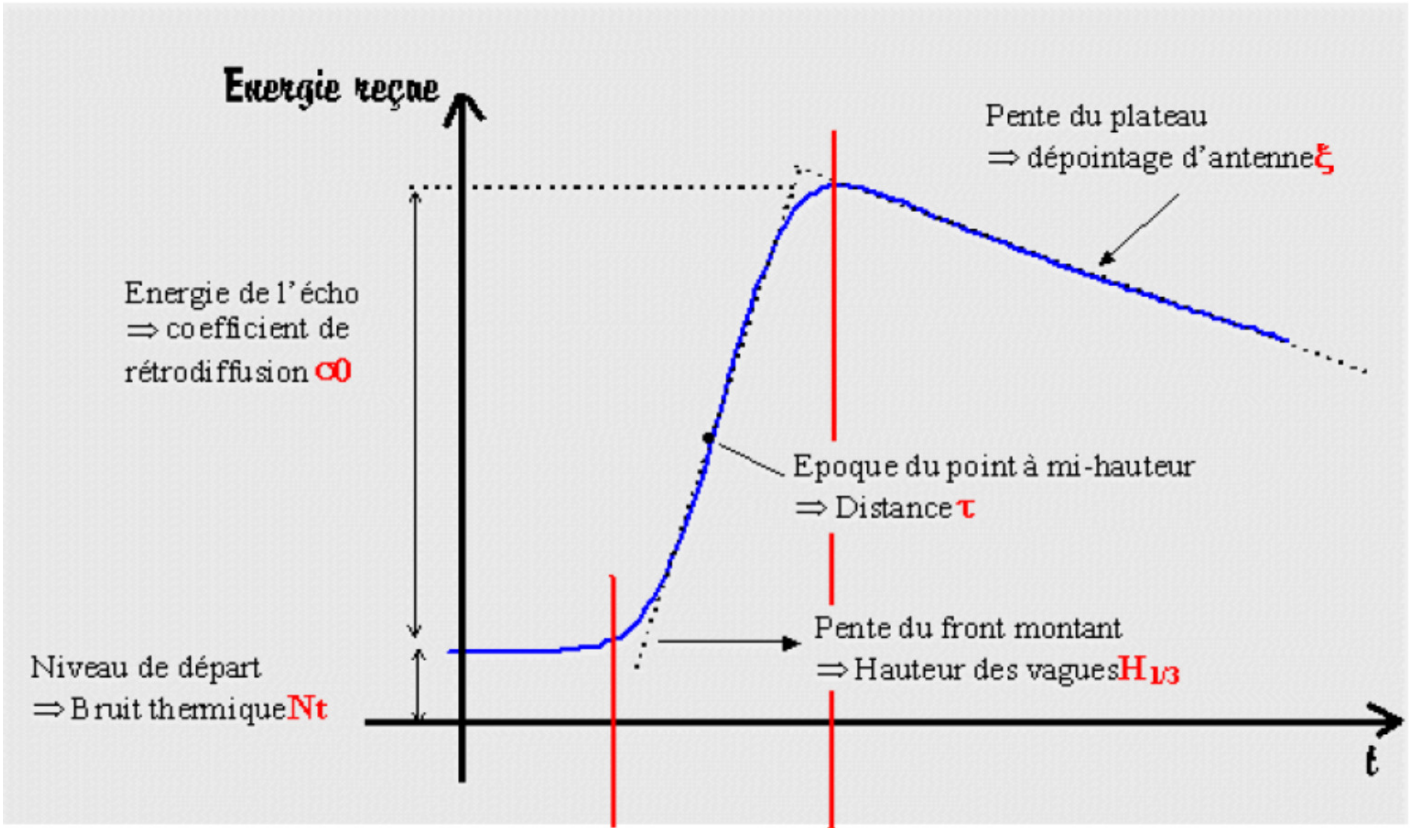




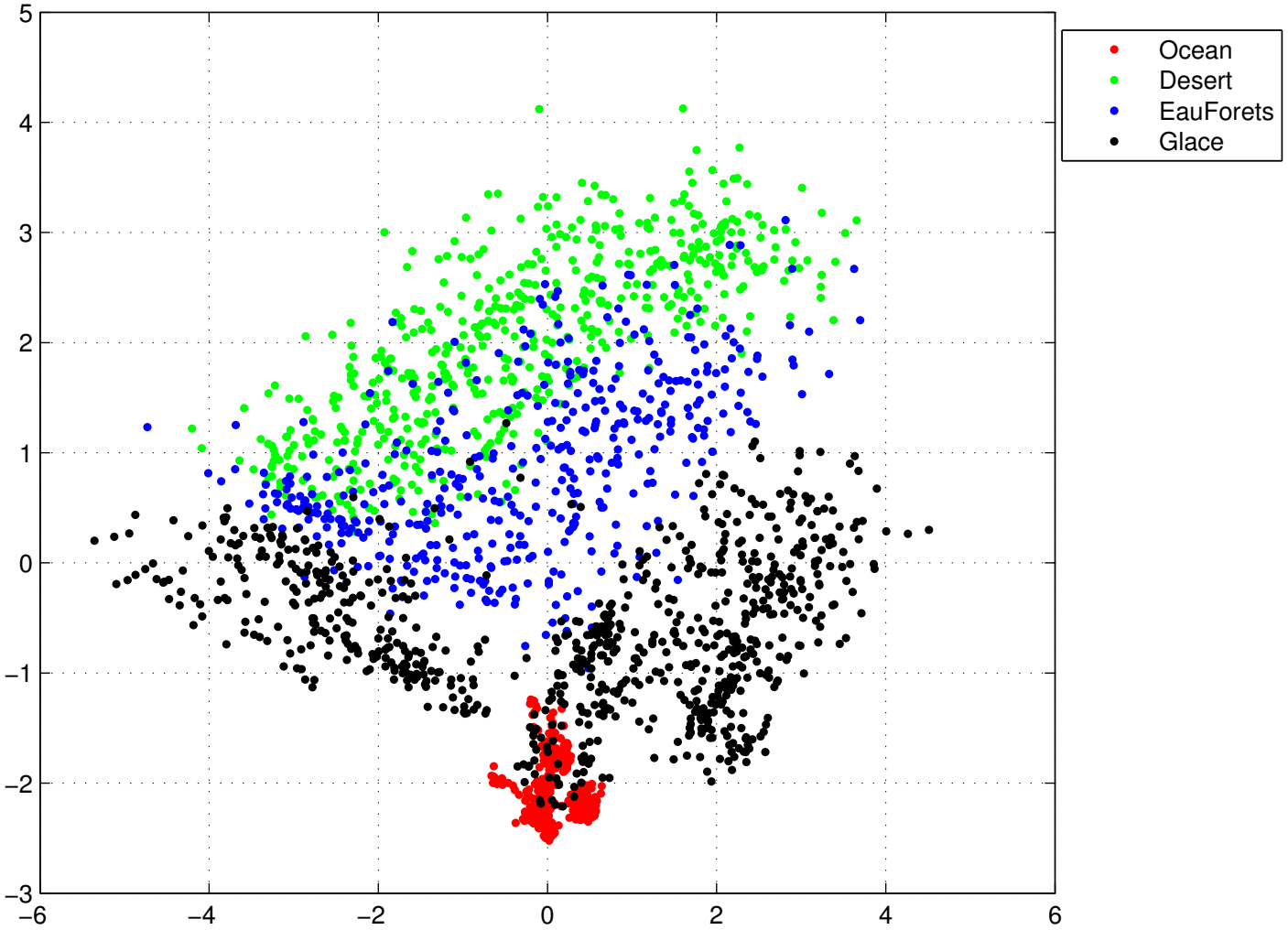
ACP des individus



Modèle de Brown



ACP des individus après extraction de paramètres



Plan du cours

Résumé

- Visualisation des données
- ACP des individus
- ACP des variables
- Exemples
- Exercice

Exercice 1 (inspiré d'un TD de l'université Paris Dauphine)

8 individus et 3 variables

	$\boldsymbol{v}_1$	$\boldsymbol{v}_2$	$\boldsymbol{v}_3$
$\boldsymbol{x}_1$	3	3	3
$\boldsymbol{x}_2$	4	4	1
$\boldsymbol{x}_3$	1	1	7
$\boldsymbol{x}_4$	2	2	5
$\boldsymbol{x}_5$	1	5	3
$\boldsymbol{x}_6$	0	4	5
$\boldsymbol{x}_7$	3	3	3
$\boldsymbol{x}_8$	2	2	5

Exercice 1

Questions

- Déterminer le tableau centré $\boldsymbol{Y}$ associé à $\boldsymbol{X}$.
- Déterminer la matrice de covariance de $\boldsymbol{Y}$ notée $\boldsymbol{\Sigma}$.
- Déterminer les valeurs propres de la matrice $\boldsymbol{\Sigma}$ et les inerties associées. Combien d'axes proposez vous de garder pour l'ACP ? Déterminer les vecteurs propres associés à ces axes en s'assurant que la première composante de ces vecteurs est négative.
- **ACP des individus**
  On donne le tableau suivant

I	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$	$\boldsymbol{x}_4$	$\boldsymbol{x}_5$	$\boldsymbol{x}_6$	$\boldsymbol{x}_7$	$\boldsymbol{x}_8$
CP_1	-1.225			1.225	-1.225	1.225	-1.225	1.225
CP_2	-0.7071	-0.7071	-0.7071	-0.7071	2.1213	2.1213	-0.7071	-0.7071
CT_1	4.167			4.167	4.167	4.167	4.167	4.167
C_1^2	0.75			0.75	0.25	0.25	0.75	0.75

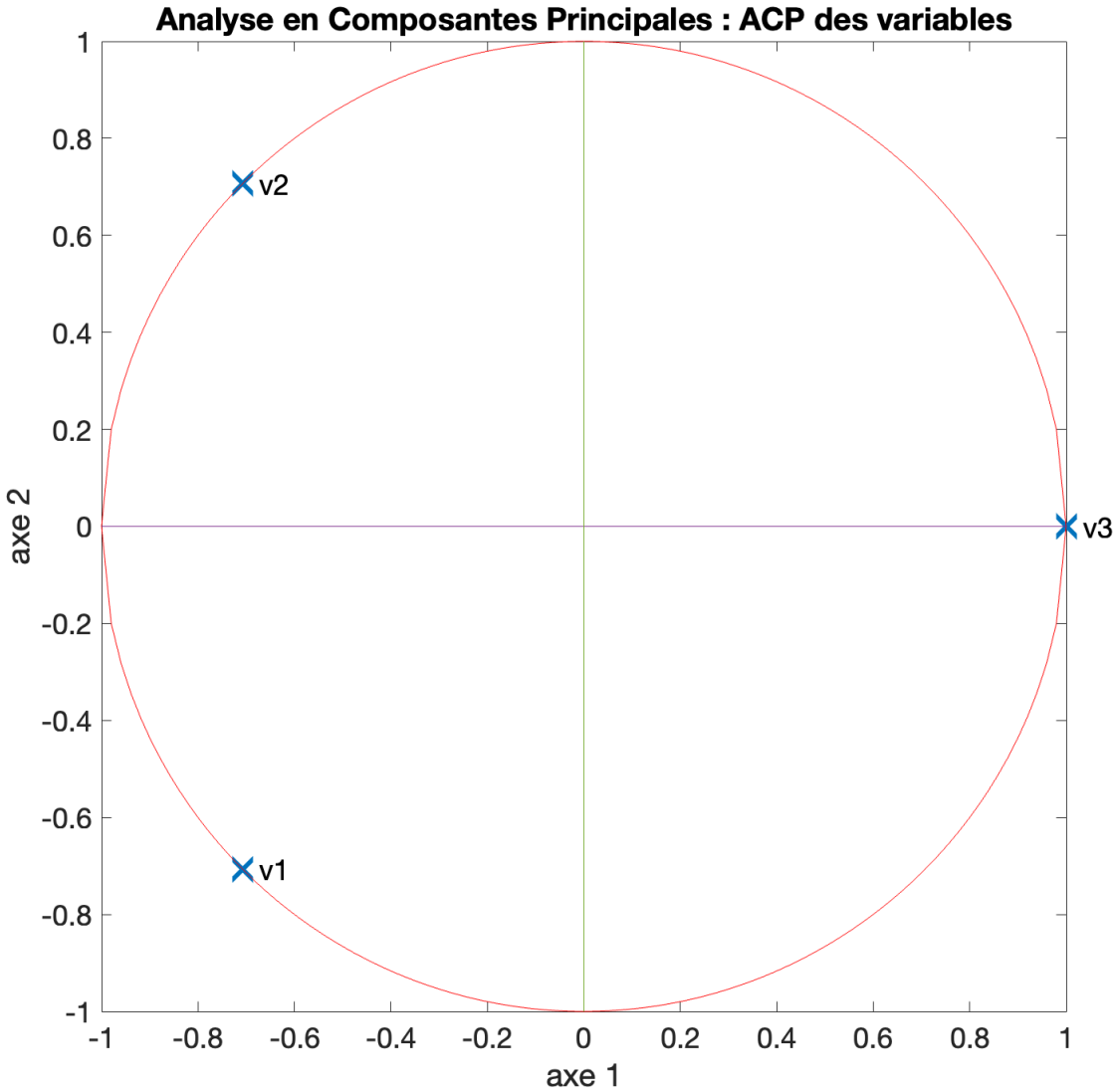
  où I = "Individus", CP_i = "Projection de l'individu sur sur l'axe i", CT_1 ="Contribution sur l'axe 1" et C_1^2 = "Cosinus carré de la représentation sur l'axe 1".
  - Compléter les données manquantes de ce tableau
  - Représenter l'ACP de ces 8 individus.
  - Quels individus sont les mieux représentés sur l'axe 1 ?

Exercice 1

Questions (suite)

- **ACP des variables**
  - Déterminer les deux premiers axes factoriels de l'ACP des variables (avec les données centrées non réduites). Déterminer ensuite les projections de la première variable sur ces deux axes factoriels.
  - Calculer la la contribution de la variable $\boldsymbol{v}_1$ à l'inertie de l'axe 1 (noté $\boldsymbol{a}_1$).
  - Calculer la qualité de représentation de la variable $\boldsymbol{v}_2$ sur l'axe 2 (noté $\boldsymbol{a}_2$).
  - L'ACP des variables de ce tableau de données centrées réduites est représenté sur la figure ci-dessous. Pourquoi les projections des trois variables sont-elles toutes situées sur le cercle unité ? Interpréter la signification des deux axes principaux.

ACP des variables



Exercice 1

Réponses

- Les moyennes des variables sont $\bar{\boldsymbol{v}}_1 = 2$, $\bar{\boldsymbol{v}}_2 = 3$ et $\bar{\boldsymbol{v}}_3 = 4$. Le tableau centré est donc

	$\boldsymbol{v}_1$	$\boldsymbol{v}_2$	$\boldsymbol{v}_3$
$\boldsymbol{y}_1$	1	0	-1
$\boldsymbol{y}_2$	2	1	-3
$\boldsymbol{y}_3$	-1	-2	3
$\boldsymbol{y}_4$	0	-1	1
$\boldsymbol{y}_5$	-1	2	-1
$\boldsymbol{y}_6$	-2	1	1
$\boldsymbol{y}_7$	1	0	-1
$\boldsymbol{y}_8$	0	-1	1

- La matrice de covariance de $\boldsymbol{Y}$ est $\boldsymbol{\Sigma} = \frac{1}{8}\boldsymbol{Y}^T\boldsymbol{Y}$. Des calculs élémentaires permettent d'obtenir

$$\boldsymbol{\Sigma} = \begin{pmatrix} \frac{3}{2} & 0 & -\frac{3}{2} \\ 0 & \frac{3}{2} & -\frac{3}{2} \\ -\frac{3}{2} & -\frac{3}{2} & 3 \end{pmatrix} = \frac{3}{2}\begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ -1 & -1 & 2 \end{pmatrix} = \frac{3}{2}\boldsymbol{\Sigma}'$$

- On préfère calculer les valeurs propres de $\boldsymbol{\Sigma}'$ et les multiplier par $\frac{3}{2}$ pour avoir celles de $\boldsymbol{\Sigma}$. On doit alors résoudre

$$\begin{vmatrix} 1-\lambda & 0 & -1 \\ 0 & 1-\lambda & -1 \\ -1 & -1 & 1-\lambda \end{vmatrix} = 0 \Leftrightarrow (1-\lambda)\lambda(\lambda-3) = 0.$$

Exercice 1

- Les valeurs propres de $\mathbf{\Sigma}$ sont donc $\mu_1 = 9/2$, $\mu_2 = 3/2$ et $\mu_3 = 0$. Les inerties associées sont $3/4$, $1/4$ et 0. On fera donc une ACP avec les deux axes associés aux valeurs propres non triviales $\mu_1 = 9/2$ et $\mu_2 = 3/2$ qui contiennent 100% de l'information. Des calculs simples permettent d'obtenir les trois vecteurs propres

$$\boldsymbol{u}_1 = \begin{pmatrix} -1/\sqrt{6} \\ -1/\sqrt{6} \\ 2/\sqrt{6} \end{pmatrix}, \boldsymbol{u}_2 = \begin{pmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{pmatrix}, \boldsymbol{u}_3 = \begin{pmatrix} -1/\sqrt{3} \\ -1/\sqrt{3} \\ -1/\sqrt{3} \end{pmatrix}.$$

- Les composantes principales sont les projections des individus sur les vecteurs propres de l'ACP. Pour les vecteurs $\boldsymbol{u}_2$ et $\boldsymbol{u}_3$, on a

$$\boldsymbol{y}_2^T \boldsymbol{u}_2 = \begin{pmatrix} 2 \\ 1 \\ -3 \end{pmatrix} \cdot \begin{pmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{pmatrix} = -\frac{1}{\sqrt{2}} \approx -0.7071, \ \boldsymbol{y}_3^T \boldsymbol{u}_2 = \begin{pmatrix} -1 \\ -2 \\ 3 \end{pmatrix} \cdot \begin{pmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{pmatrix} = -\frac{1}{\sqrt{2}} \approx -0.7071$$

$$\boldsymbol{y}_2^T \boldsymbol{u}_1 = \begin{pmatrix} 2 \\ 1 \\ -3 \end{pmatrix} \cdot \begin{pmatrix} -1/\sqrt{6} \\ -1/\sqrt{6} \\ 2/\sqrt{6} \end{pmatrix} = \frac{-9}{\sqrt{6}} \approx -3.674, \ \boldsymbol{y}_3^T \boldsymbol{u}_1 = \begin{pmatrix} -1 \\ -2 \\ 3 \end{pmatrix} \cdot \begin{pmatrix} -1/\sqrt{6} \\ -1/\sqrt{6} \\ 2/\sqrt{6} \end{pmatrix} = \frac{9}{\sqrt{6}} \approx 3.674$$

Par ailleurs

$$CT_1(2) = \frac{F_{21}^2}{\sum_{i=1}^n F_{i1}^2} = \frac{F_{21}^2}{n\mu_1} = \frac{(3.674)^2}{8 \times \frac{9}{2}} = CT_1(3) \approx 37.5\%$$

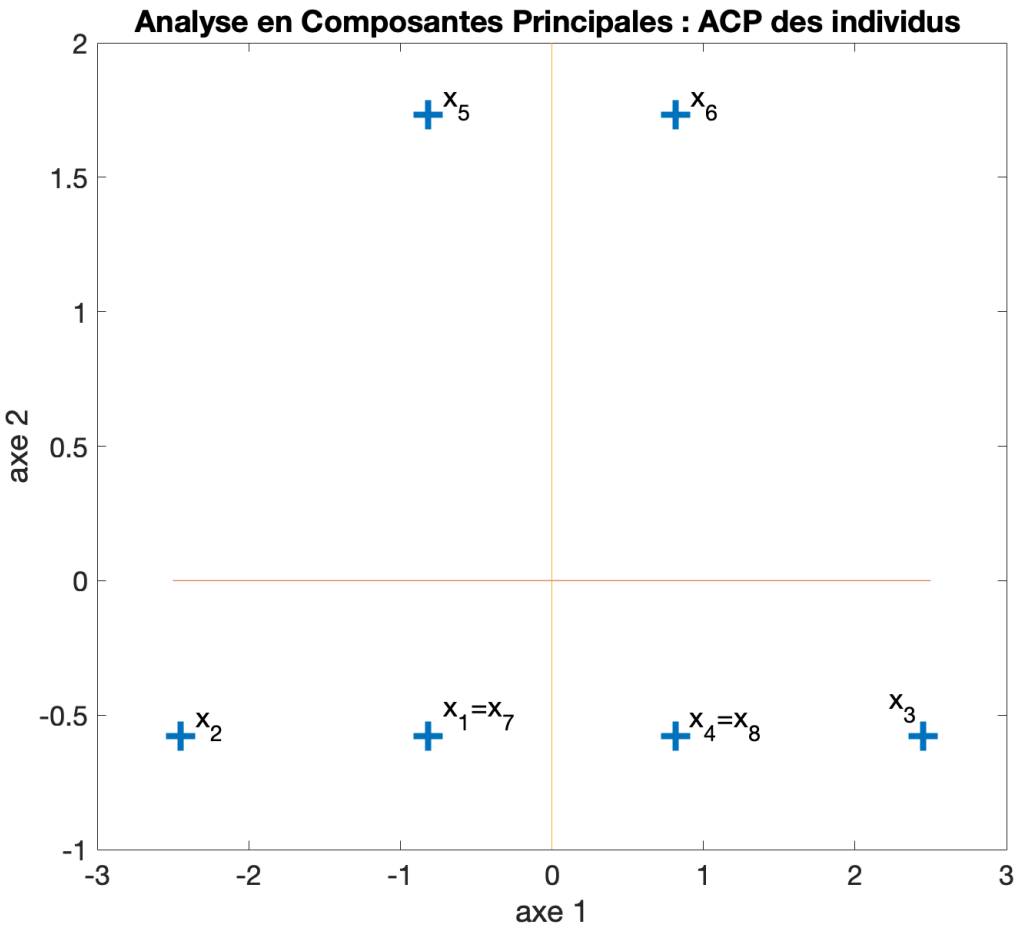
et

$$\mathsf{C}_1^2(2) = \mathsf{C}_1^2(3) = \frac{(\boldsymbol{y}_2^T \boldsymbol{u}_1)^2}{\|\boldsymbol{y}_2\|^2} = \frac{-\left(\frac{9}{\sqrt{6}}\right)^2}{14} = \frac{27}{28} \approx 0.964$$

Les individus $\boldsymbol{x}_2$ et $\boldsymbol{x}_3$ sont donc les mieux représentés sur l'axe 1.

Exercice 1

- L'ACP des 8 individus $\boldsymbol{x}_i$ est représentée ci-dessous



Exercice 1

- D'après le cours, les vecteurs propres $\boldsymbol{a}_i$ de l'ACP des variables peuvent s'obtenir à partir des vecteur propres $\boldsymbol{u}_i$ de l'ACP des individus à l'aide de la relation $\boldsymbol{a}_i = \frac{\boldsymbol{Y}\boldsymbol{u}_i}{\sqrt{n\mu_i}}$. En pratique, il suffit de calculer les vecteurs $\boldsymbol{Y}\boldsymbol{u}_i$ et de les normaliser. On obtient alors (en s'assurant que la première composante est négative)

$$\boldsymbol{a}_1 = \frac{1}{2\sqrt{6}} \begin{pmatrix} -1 \\ -3 \\ 3 \\ 1 \\ -1 \\ 1 \\ -1 \\ 1 \end{pmatrix} \text{ et } \boldsymbol{a}_2 = \frac{1}{2\sqrt{6}} \begin{pmatrix} -1 \\ -1 \\ -1 \\ -1 \\ 3 \\ 3 \\ -1 \\ -1 \end{pmatrix}$$

- Les projections de la variable k (normalisée par $\sqrt{n}$ mais pas réduite) sur tous les axes factoriels sont définies par

$$\boldsymbol{\phi}_k = \begin{pmatrix} r(\boldsymbol{v}_k, \boldsymbol{a}_1) \\ r(\boldsymbol{v}_k, \boldsymbol{a}_2) \\ r(\boldsymbol{v}_k, \boldsymbol{a}_3) \end{pmatrix} = \sqrt{\mu_k}\boldsymbol{u}_k$$

Les projections de la première variable (normalisée par $\sqrt{n}$ mais pas réduite) sur les deux axes factoriels sont donc

$$(\sqrt{\mu_1}u_{11}, \sqrt{\mu_2}u_{21}) = \left(\sqrt{\frac{9}{2}} \times \frac{-1}{\sqrt{6}}, \sqrt{\frac{3}{2}} \times \frac{-1}{\sqrt{2}}\right) = \left(-\frac{\sqrt{3}}{2}, -\frac{\sqrt{3}}{2}\right) \approx (-0.87, -0.87).$$

On remarquera que la calcul direct du produit scalaire entre $\boldsymbol{v}_1$ et $\boldsymbol{a}_1$ est
$(\boldsymbol{v}_1/\sqrt{8})^T \boldsymbol{a}_1 = \frac{1}{2\sqrt{6\times 8}}[(1)(-1) + (2)(-3) + \ldots + (0)(1)] = -\frac{\sqrt{3}}{2}$, ce qui est cohérent.

Exercice 1

- La contribution de la variable $\boldsymbol{v}_1$ à l'inertie de l'axe 1 est définie par

$$\frac{r^2(\boldsymbol{v}_1, \boldsymbol{a}_1)}{\sum_{k=1}^{p} r^2(\boldsymbol{v}_k, \boldsymbol{a}_1)} = \frac{\mu_1 u_{11}^2}{\mu_1 \|\boldsymbol{u}_1\|^2} = u_{11}^2 = \frac{1}{6}.$$

- La qualité de représentation de la variable $\boldsymbol{v}_2$ sur l'axe $\boldsymbol{a}_2$ est

$$\cos^2(\theta) = \frac{r^2(\boldsymbol{v}_2, \boldsymbol{a}_2)}{\|\boldsymbol{v}_2\|^2} = \frac{\mu_2 u_{22}^2}{\frac{1}{8} \times 12} = \frac{\frac{3}{2} \times \frac{1}{2}}{\frac{3}{2}} = \frac{1}{2}.$$

où on a pris soin de normaliser la variable $\boldsymbol{v}_2$ de manière à ce qu'elle soit sur l'hypersphère

$$\boldsymbol{v}_2 = \frac{1}{\sqrt{8}} \begin{pmatrix} 0 \\ 1 \\ -2 \\ -1 \\ 2 \\ 1 \\ 0 \\ -1 \end{pmatrix}$$

- Les projections des trois variables sont situées sur le cercle unité car la troisième valeur propre est nulle. Le plan constitué des deux premiers axes principaux contient 100% de l'information. Le premier axe principal oppose la variable $\boldsymbol{v}_3$ aux deux autres variables $\boldsymbol{v}_1$ et $\boldsymbol{v}_2$. Le second axe principal oppose les variables $\boldsymbol{v}_1$ et $\boldsymbol{v}_2$.

Que faut-il savoir ?

ACP des individus

- Déterminer les axes principaux et projeter les individus sur ces axes principaux
- Déterminer le pouvoir de représentation (l'inertie) de chaque axe
- Représenter des individus supplémentaires
- Analyser la qualité de représentation et la contribution de chaque individu

Qualité de représentation de l'individu i sur l'axe s : $\frac{F_{is}^2}{n\lambda_s}$

et

Contribution de individu i sur l'axe s : $\frac{F_{is}^2}{\|\boldsymbol{x}_i\|^2}$

Que faut-il savoir ?

ACP des variables

- Déterminer les axes principaux et projeter les variables sur ces axes principaux

$$\frac{\boldsymbol{X}\boldsymbol{u}_k}{\|\boldsymbol{X}\boldsymbol{u}_k\|}$$

et

$$\boldsymbol{\phi}_k = \begin{pmatrix} r(\boldsymbol{v}_k, \boldsymbol{a}_1) \\ \vdots \\ r(\boldsymbol{v}_k, \boldsymbol{a}_p) \end{pmatrix} = \sqrt{\lambda_k}\boldsymbol{u}_k$$

- Représenter des variables supplémentaires
- Analyser la qualité de représentation et la contribution de chaque variable

Qualité de représentation de la variable k sur l'axe s : $\frac{r^2(\boldsymbol{v}_k, \boldsymbol{a}_s)}{\sum_{k=1}^{p} r^2(\boldsymbol{v}_k, \boldsymbol{a}_s)}$

et

Contribution de la variable k sur l'axe s : $\cos^2(\theta'_{ks}) = u_{ks}^2$

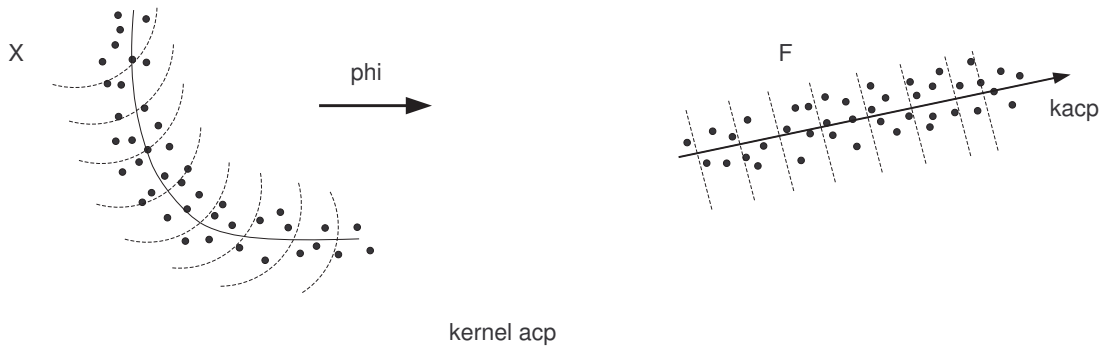
Pour aller plus loin

ACP à noyaux (Kernel PCA)

Plongement dans un nouvel espace de représentation à l'aide d'une application non-linéaire

$$\begin{array}{rccc} \Phi : & \mathcal{X} & \longrightarrow & \mathcal{F} \\ & \boldsymbol{x} & \longmapsto & \Phi(\boldsymbol{x}) \end{array}$$

On applique l'une des méthodes précédentes aux données transformées $\Phi(\boldsymbol{x}_i)$.



- Bernhard Schölkopf, Alex Smola and Klaus Robert Müller, Nonlinear Component Analysis as a Kernel Eigenvalue Problem, Neural computation, vol. 10, no. 5, pp. 1299-1319, 1998.

t-SNE : une autre méthode de réduction de dimension

t-distributed stochastic neighbor embedding (t-SNE)

- Construction d'une loi de probabilité P_1 pour les vecteurs (de grande dimension) d'une base de données, de manière à ce que les objets similaires aient une forte probabilité.
- Recherche d'une loi de probabilité P_2 pour les projections de ces vecteurs dans un espace de dimension réduite de manière à ce que P_1 et P_2 soient proches.

- Laurens van der Maaten and Geoffrey Hinton, Vizualizing Data using t-SNE, Journal of Machine Learning Research, vol. 9, pp. 2579-2605, 2008.

MNIST Dataset

tSNE of MNIST Dataset