Analyse de données multivariées
Analyses factorielles
Visualisation et réduction de dimension - statistiques exploratoires

Mireille Batton-Hubert - Institut H.Fayol

UP3: 'Analyse de données et données temporelles'
Majeure Sciences des données- octobre 2020

1

UP3: Analyse de données et données temporelles

Contenu (30 h) :
- Module (15h) statistiques exploratoires : comprendre et faire bon usage des techniques pour décrire, visualiser et analyser les données
- Module séries temporelles (15 h)

- Visualisation et réduction de dimension : analyse en composantes principales (3h) (analyse factorielle) ACP (PCA) # cours + **TP1**
- ACP et décomposition : (3h) # cours + **TP2 :** *TP1 + TP2 : 1 note*

- Cas des variables qualitatives. Analyse factorielle des correspondances (3h) AFC cours + TP3
- Analyse discriminante descriptive (3h). AFD : cours + TP4
- Mise en œuvre sur un cas gros TP (AFD + ACP) (soit 3h en séance de TP + travail personnel) : **TPF** noté

Pas d'examen mais rendus de TP notés : date limite des rendus 28 octobre

2

Définition de l’analyse de données

- On s’intéresse à des données statistiques représentant un échantillon composé de *n* individus décrits par un ensemble de variables *X* de dimension *p* (souvent de grande dimension)
- Ces données sont :
  - -quantitatives sur $\Re$
  - -qualitatives (ordinale ; catégorielle)
- Les techniques de traitement de ces données constituent l’analyse de données multi-variées (multidimension)
  - Analyse en composantes principales : ACP - PCA (Principal Component Analysis)
  - Analyse des correspondances
  - Analyse factorielle discriminante
- Ces méthodes appartiennent aux statistiques multidimensionnelles (regroupe des méthode de description + réduction de dimension et des méthodes de classification)

2

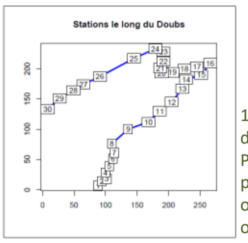
Des exemples de données

- Un échantillon de données avec des individus :
- *Du plus simple* (un échantillon de notes , de mesures d’individus …)

une étude d'un botaniste : dimensions de 15 fleurs d'iris avec 3variables longueur du sépale x1, largeur du sépale x2, longueur du pétale x3

Fleur n°	x_1	x_2	x_3
1	5.1	3.5	1.4
2	4.9	3.0	1.4
3	4.7	3.2	1.3
4	4.6	3.1	1.5
5	5.0	3.6	1.4
6	7.0	3.2	4.7
7	6.4	3.2	4.5
8	6.9	3.1	4.9
9	5.5	2.3	4.0
10	6.5	2.8	4.6
11	6.3	3.3	6.0
12	5.8	2.7	5.1
13	7.1	3.0	5.9
14	6.3	2.9	5.6
15	6.5	3.0	5.8

- *au plus sophistiqué* : des données de capteurs ,
- un suivi de couts / produits financiers
- Échantillon pouvant être en continu



11 variables physico-chimiques : distance a la source, altitude, Pente, débit, pH, dureté de l'eau phosphate, nitrate, ammoniaque, oxygène, demande biologique en oxygène

- **Données temporelles ou non …**
- **De grande dimension (ACP dite de grande dimension…)**

4

Objectifs et organisation

- ✓ Comprendre ce qu'est l'analyse de données multivariées
- ✓ Comment mettre en œuvre ces analyses et leur signification
- ✓ Comment obtenir et interpréter des résultats
- séances : cours (50%) + TD / TP info (50%) (apprentissage des outils et programmation) qui serviront à l'étude de cas

Organisation du cours :
- cours + TP pour chaque technique factorielle : à chaque TP un rendu de l'avancement sur *campus*
Utilisation de 2 TP pour l'étude du cas (3h00) qui sera notée : rendu fait pour le 28 octobre

5

L'enjeu de l'analyse de données

Ces données statistiques : nombreuses variables

- dont aucune n'est parfaite
- La perception d'un phénomène appréhendé comme la combinaison d'un grand nombre de variables
- **Comment faire pour tenir compte de l'ensemble de l'information ?**

Moyens ?

Faire de tableaux croisés (var X×Y) :
Si 10 variables soit 45 tableaux croisés si 100 var 4950 tableaux croisés

Autre méthode : dite des indices exp: indice d'inflation , indice du développement humain, BIP40 avec une somme pondérée mais formule arbitraire avec des pondérations …

Trouver des méthodes pour synthétiser **les variables sans trop les déformer**

6

En fonction du type de variables: différentes méthodes de réduction de dimension

Méthodes factorielles de représentation

- Analyse en composantes principales (ACP) : variables quantitatives
- Analyse factorielle de correspondance (AFC) analyse d'un tableau croisé de deux variables qualitatives (caractère)
- Analyse des correspondances multiples (ACM) plusieurs variables qualitatives
- Analyse factorielle discriminante ou analyse discriminante décisionnelle (analyse discriminante): liaisons existants entre un caractère à expliquer qualitatif et un ensemble de var. quantitatives
- Analyse factorielle multiple (AFM)

7

Analyse en composantes principales (ACP) Principal component analysis (PCA)

Appelée aussi la transformation de Karhunen–Loève (KLT)[1]

Karl Pearson (1901)
Harold Hotelling (1930)

8

1. Objectifs de l'ACP (1)

In *Dictionary of statistics and methodology* » (Vogt, 1993) :
'L'analyse en composantes principales : Ensemble de méthodes permettant de procéder à des transformations linéaires d'un grand nombre de variables intercorrélées de manière à obtenir un nombre relativement limité de composantes non corrélées. Cette approche facilite l'analyse en regroupant les données en des ensembles plus petits et en permettant d'éliminer les problèmes de multicolinéarité entre les variables. L'analyse en composantes principales s'apparente à l'analyse factorielle, mais c'est une technique indépendante qui est souvent utilisée comme première étape à une analyse factorielle ' (*Vogt, 1993, page 177*).

⟹

A pour premier objectif de réduire la dimension de données

Exemple des Iris

Iris - échantillon : *p*= 3 variables - *n* individus = 15

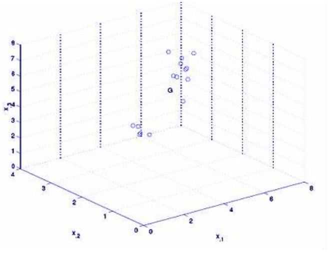
9

1. Objectifs de l'ACP (2)

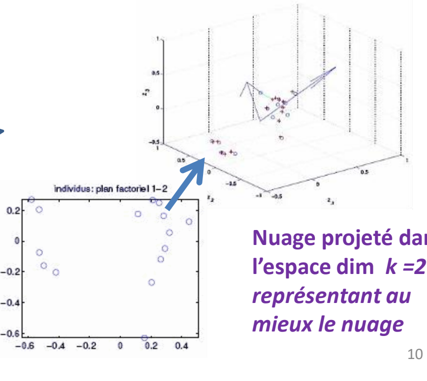
A pour premier objectif de réduire la dimension de données

exemple des Iris - échantillon : *p*= 3 variables - *n* individus = 15

Trouver des axes (~indices) qui respectent la forme du nuage multidimensionnel initial c.a.d la forme des relations entre les variables



Nuage initiale dans l'espace de dim *p = 3*



Nuage projeté dans l'espace dim *k =2 < p* *représentant au mieux le nuage*

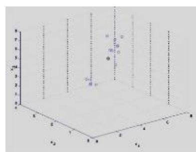
10

1. Objectifs de l'ACP(3)

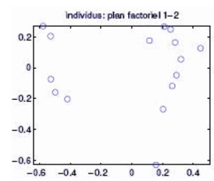
Données : n individus observés sur *p* variables quantitatives.
L'A.C.P. permet d'explorer les liaisons entre variables et les ressemblances entre individus.

Pour construire un espace plus petit k<p qui conserve ces propriétés

p=3



k=2



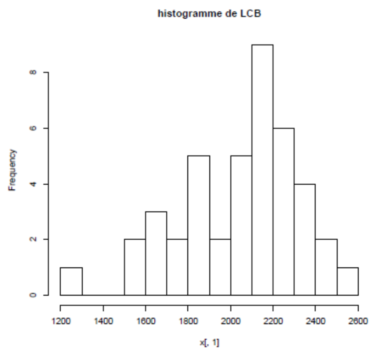
Résultats : dans ce nouvel espace de dimension k
Visualisation des individus : (Notion de distances entre individus)
Visualisation des variables : (en fonction de leurs corrélations)
Interprétation des résultats

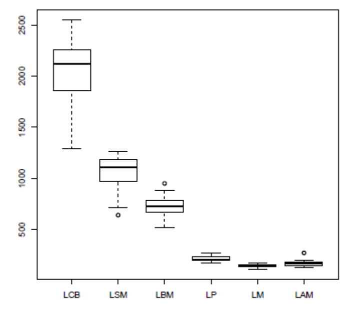
2. Pourquoi : veut-on réduire (1) ?

Les ossements de cranes de 42 canidés, identifiés comme chiens et loups, ont été mesurés. Les variables sont :
LCB : longueur condylo-basale
LSM : longueur de la mâchoire supérieure
LBM : largeur bi-maxillaire
LP : longueur de la carnassière supérieure
LM : longueur 1° molaire supérieure
LAM : largeur 1° molaire supérieure
Source : JAMBU - Classification

TYPE	LCB	LSM	LBM	LP	LM	LAM
BULL-DOG 1	1290	640	950	175	112	138
BULL-DOG 2	1540	740	760	200	142	165
BOXER	1580	710	710	167	125	133
SAINT-BERNARD	2200	1110	880	225	154	180
BULL-MASSIF	1900	930	780	197	132	140
DOGUE ALLEMAND 1	2410	1190	870	210	147	183
DOGUE ALLEMAND 2	2420	1200	850	199	153	176
DOGUE ALLEMAND 3	2550	1260	860	214	150	180
DOGUE ALLEMAND 4	2340	1130	830	213	148	170
SETTER 1	2010	1050	700	198	143	159
SETTER 2	1960	1060	670	185	126	142
SETTER 3	2200	1170	700	198	143	156
GROENDAL	2050	1050	700	190	124	149
BAS-ROUGE	2160	1120	750	196	140	164
BRIARD 1	2280	1220	780	225	142	178
BRIARD 2	2120	1110	730	205	137	166
...	...	...	...	...	...	...

2. Pourquoi : veut-on réduire (2) ?

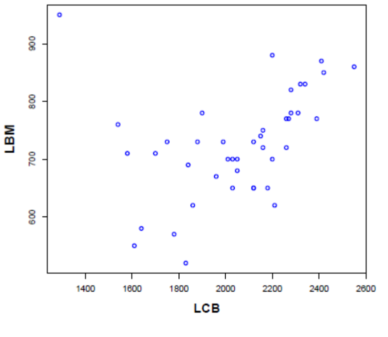


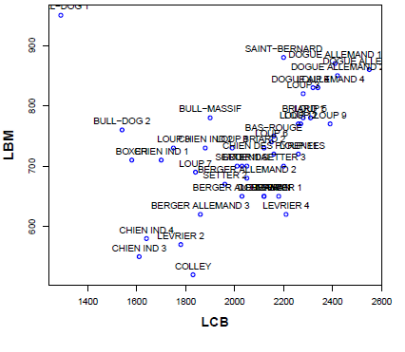


Statistique classique : distribution et fréquences des observations par variables

13

2. Pourquoi : veut-on réduire (3) ?

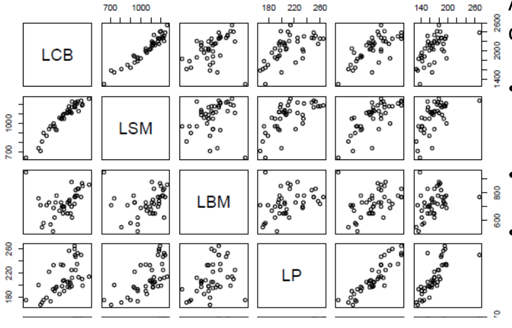




Statistique classique : comparaison de 2 variables : relation entre 2 variables

14

2. Pourquoi : veut-on réduire (4) ?



ACP : Décrire au mieux la variabilité des données et doit permettre de :

- Une réduction des données décrites par *p* variables à un nombre restreint k de descripteurs
- Une visualisation des données (si k =2 ou 3)
- Une interprétation des données : liaisons entre variables

Filtrer l'information (enlever les redondances et le bruit) : compression de données
 Réduire la dimension, en particulier pour la visualisation des données
 Déterminer les individus qui se ressemblent ou, au contraire, s'opposent
 Identifier les variables sur lesquelles sont fondées les ressemblances ou dissemblances (variables discriminantes)
 Obtenir de nouvelles variables non corrélées (décomposition orthogonale)
 Faire une analyse de la corrélation entre toutes les variables à la fois

3. Analyse de données /statistiques

Statistique classique : un nombre restreint de caractères mesurées sur un petit ensemble d'individus

➔ développer des méthodes d'estimation et de tests statistiques

Statistique inférentielle

En réalité

➔ Un individu est observé par un grand nombre de caractères

Analyse de données : étude globale des individus et des variables avec une représentation graphique

Statistique descriptive ou explicative

Différents points de vue de l'analyse :

- Rechercher des ressemblances ou différences entre individus sur plusieurs caractères(profil) : proximité entre individus : regroupement en catégorie homogène ➔ *Classification*
- Description de relations entre caractères : deux caractères sont liés s'ils varient de la même façon pour les différents individus : si tous les caractères ont un rôle identique : mettre en évidence les groupes de caractères qui sont corrélés ou indépendants : individus et var. dans espace géométrique réduit en perdant le moins d'information possible → ***A.D***
- ***Différences entre stats classiques et A.D***

16

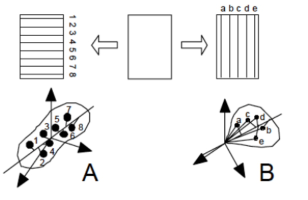
4. Problème de réduction de *dim* (1)

On cherche :

A projeter un ensemble de points $x_{i,j}$ représenté dans un espace de dimension p dans un espace dim $k<p$ qui respecte l'agencement des points du nuage :
ce que l'on cherche → **une projection *(à caractériser)***

Pour représenter (non visible au delà de p=3)
- les individus dans l'espace des variables
situation A : n-points-lignes dans $\mathbb{R}^p$

-les variables dans l'espace des individus
situation B : p-points-lignes dans $\mathbb{R}^n$



➔ **choix d'un espace vectoriel sur $\mathbb{R}^p$ qui sera doté d'un produit scalaire**
Qui permet de représenter les points individus et les variables dans un même espace

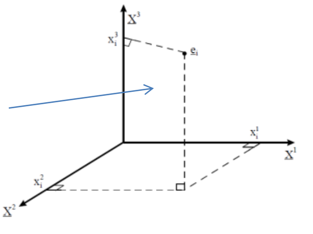
17

4.1 Choix : un espace vectoriel euclidien

Propriété de cet espace vectoriel réel $\mathbb{R}^n$: il est doté de l'addition , de la multiplication par un scalaire qui sont des opérateurs sur les matrices ;
une base canonique $\mathbb{R}^n$ de peut être définie ($e_1, e_2, \ldots e_n$) tout élément X de $\mathbb{R}^n$ est un combinaison linéaire des vecteurs de la base

Soit $\sum_{i=1}^{n} x_i \, e_i$

Individu x_i



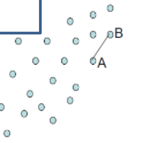
On peut définir un produit scalaire Φ :
Une application de $\mathbb{R}^n \times \mathbb{R}^n$ dans $\mathbb{R}$
Noté $\langle X|Y \rangle_\Phi$ (noté aussi $\langle X|\Phi|Y \rangle$ le nombre réel Φ (X,Y)

On peut définir ainsi
un produit scalaire canonique :
un produit scalaire défini par une matrice diagonale à éléments positifs

18

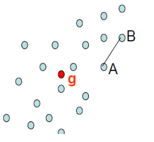
4.2 Formalisation du problème

Construire un critère permettant de construire puis d'évaluer la projection :
- Conserver au mieux les caractéristiques du nuage :
  soit la forme du nuage
  qui est liée à la distance entre les individus
- Se doter d'une distance entre individus



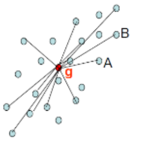
➔ **Distance euclidienne dans un espace vectoriel sur $\mathbb{R}^p$**

$$d^2(X_{A.}, X_{B.}) = \sum_{j=1}^{p}\left(X_A^j - X_B^j\right)^2$$



➔ **Caractériser le nuage de points (dans espace des individus)**

centre de gravité du nuage : $g = \left(\frac{1}{n}\right)\sum_{i=1}^{n} X_i$

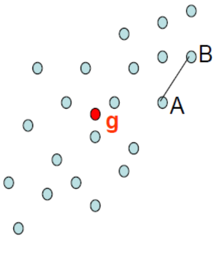


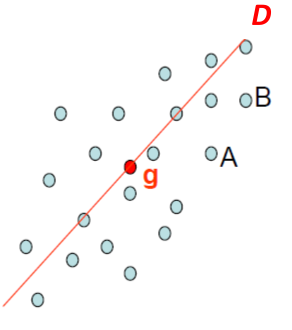
Inertie totale du nuage : $I_T = \sum_{i=1}^{n}\frac{1}{n}d^2(g, X_i)$

$$I_T = \sum_{i=1}^{p}\left(\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - g_j)^2\right)$$

4.3 Formalisation du problème

Définir une projection tel que :
Ces sous-espaces représentant au mieux ce nuage de point en respectant 2 critères : le **critère de proximité** et la **fidélité des distances (Inertie du nuage et variance)**.





Projection sur une droite : projeté orthogonal

Les données sont fournies par une matrice X
De n individus décrits par p variables

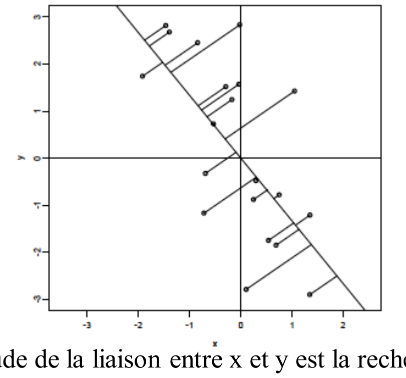
$$\begin{pmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix}$$

21

4.4 Enoncé (1)

Soit deux variables *X1* et *X2* mesurées sur n individus.

$$\begin{pmatrix} X_{11} & X_{12} \\ \vdots & \vdots \\ X_{n1} & X_{n2} \end{pmatrix}$$



Deux variables X1 et X2 notées : X et Y mesurées sur *n* individus

L'étude de la liaison entre x et y est la recherche d'une droite optimum.
Quel est le critère d'optimisation ?

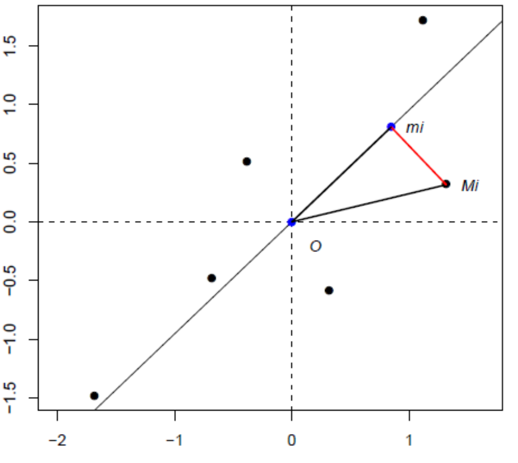
22

4.4 Enoncé (2)

Cette droite minimise

$$\frac{1}{n}\sum_{i=1}^{n} M_i m^2{}_i$$

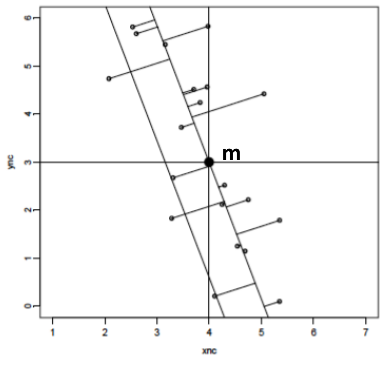
où
- M_i est le point *i* du plan,
- m_i est la projection orthogonale de M_i sur la droite.



Minimisation de la distorsion entre le point d'origine M_i et son projeté m_i
Minimisation de la moyenne des carrés des distances des points à cette droite D

23

4.5 Pourquoi cette droite doit passer par le centre de gravité - point moyen du nuage ?



($m(x)m(y)$) le point moyen

D et *D'* : deux droites parallèles :
D passe par le point moyen m et
D' passe par un point quelconque *m'*

Que vaut :

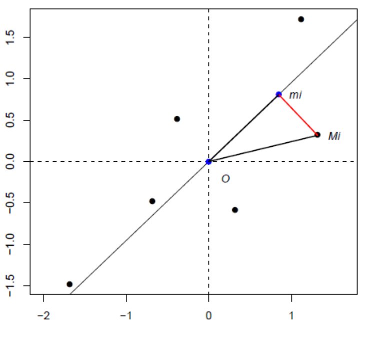
$$\|M_i - m_i'\|^2$$

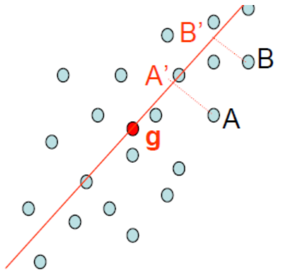
A retrouver …

24

4.6 Soit un problème de minimisation

Recherche un sous-espace qui passe par G centre de gravité du nuage (point moyen)

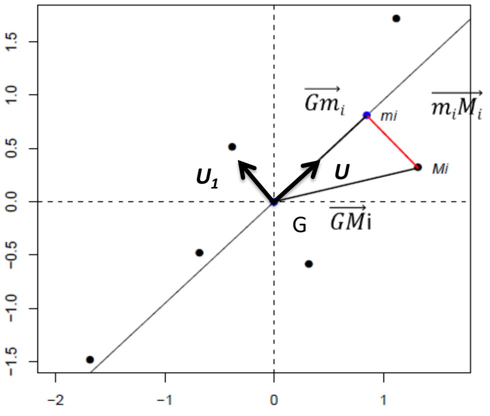




Minimisation de la distorsion entre le point d'origine M_i et son projeté m_i *(inertie)*
Minimisation de la moyenne des carrés des distances des points à cette droite D (distance)

26

4.7 Formulation de la projection sur une droite D



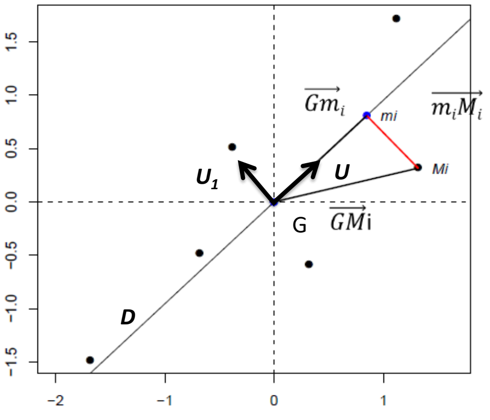
ACP_projection.pdf

27

4.7 Formulation de la projection sur une droite D

$u_1 = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$, avec $\alpha^2 + \beta^2 = 1$.

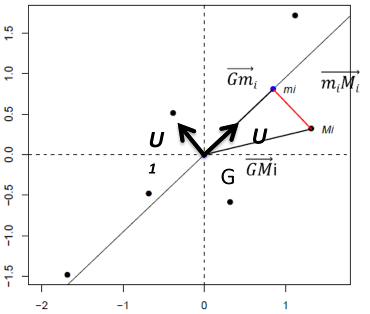
$u = \begin{pmatrix} \beta \\ -\alpha \end{pmatrix}$.



Un vecteur : α X + β Y qui est :
-combinaison linéaire de 2 variables centrées
- minimise: $\|\alpha X + \beta Y\|^2_{D^{\perp}_n}$

28

4.8 Décomposition de l'Inertie du nuage (a)



Inertie Totale du nuage I_T

$$I_T = \frac{1}{n}\sum_{i=1}^{i=n} \|\overrightarrow{GM_i}\|^2 = \frac{1}{n}\sum_{i=1}^{i=n}(x_{i0}^{\ 2} + y_{i0}^{\ 2})$$

Inertie statistique I_s

Du nuage de points dans R^p par rapport à une direction Δ de R^p définie par un vecteur unitaire u:

$$I_S(u) = \frac{1}{n}\sum_{i=1}^{i=n} \|\overrightarrow{Gm_i}\|^2$$

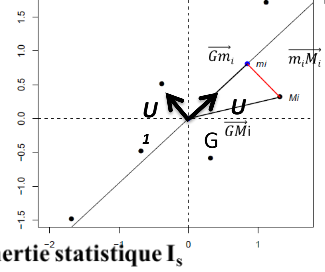
Avec :Matrice des variables centrées

$$Z = \begin{pmatrix} x_{10} & y_{10} \\ \vdots & \vdots \\ x_{n0} & yn0 \end{pmatrix}$$

Avec $\overrightarrow{Gm_i}$ le projeté orthogonal de $\overrightarrow{GM_i}$ sur D

29

4.8 Décomposition de l'Inertie du nuage (b)



Inertie Totale du nuage I_T

$$I_T = \frac{1}{n}\sum_{i=1}^{i=n} \|\overrightarrow{GM_i}\|^2 = \frac{1}{n}\sum_{i=1}^{i=n}(x_{i0}^{\ 2} + y_{i0}^{\ 2})$$

Inertie statistique I_s

Du nuage de points dans R^p par rapport à une direction Δ de R^p définie par un vecteur unitaire u:
avec $\overrightarrow{Gm_i}$ le projeté orthogonal de $\overrightarrow{GM_i}$ sur u

$$I_S(u) = \frac{1}{n}\sum_{i=1}^{i=n} \|\overrightarrow{Gm_i}\|^2$$

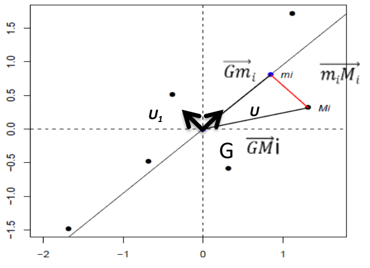
Inertie Mécanique IM

Du nuage de points dans R^p par rapport à une direction Δ de R^p définie par un vecteur unitaire u:
avec $\overrightarrow{Gm_i}$ le projeté orthogonal de $\overrightarrow{GM_i}$ sur u

$$I_M(u) = \frac{1}{n}\sum_{i=1}^{i=n} \|\overrightarrow{m_i M_i}\|^2$$

30

4.8 Décomposition de l'inertie totale (c)



Inertie Totale du nuage I_T

$$I_T = \frac{1}{n}\sum_{i=1}^{i=n} \|\overrightarrow{GM_i}\|^2 = \frac{1}{n}\sum_{i=1}^{i=n}(x_{i0}{}^2 + y_{i0}{}^2)$$

Inertie statistique I_s

Du nuage de points dans R^p par rapport à une direction Δ de R^p définie par un vecteur unitaire u:

avec $\overrightarrow{Gm_i}$ le projeté orthogonal de $\overrightarrow{GM_i}$ sur u

$$I_S(u) = \frac{1}{n}\sum_{i=1}^{i=n} \|\overrightarrow{Gm_i}\|^2$$

Inertie Mécanique IM

$$I_M(u) = \frac{1}{n}\sum_{i=1}^{i=n} \|\overrightarrow{m_i M_i}\|^2$$

Théorème de Pythagore :

$$\|\overrightarrow{GM_i}\|^2 = \|\overrightarrow{Gm_i}\|^2 + \|\overrightarrow{m_i M_i}\|^2$$

$$\underbrace{I_M(u)}_{\text{à minimiser}} = I_T - \underbrace{I_S(u)}_{\text{à maximiser}}.$$

31

Liens entre géométrie et variables statistiques (dans $\mathbb{R}^2 \rightarrow \mathbb{R}^p$) (1)

une variable statistique quantitative bidimensionnelle (*X*, *Y*) à valeurs dans R ², définie dans une population Ω de taille *n*

R ² est l'**espace des individus**
La variable statistique est représentée par un nuage de points dans R ² et chaque point du nuage statistique représente un individu de la population Ω.

Les *n* valeurs *X* (ω) de *X* pour les *n* individus de la population peuvent être considérées comme les *n* valeurs *X* (ω) de *X* pour les *n* individus de la population peuvent être considérées comme les coordonnées d'un vecteur de R *ⁿ*.

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{pmatrix}. \qquad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}.$$

L'espace *E* = R ^{n} : l'**espace des variables**
Chaque élément de *E* peut être considéré comme les valeurs d'une variable statistique quantitative réelle définie sur Ω.

32

Liens entre géométrie et variables statistiques (dans $\mathbb{R}^2 \to \mathbb{R}^p$) (2)

Produit scalaire

Dans cet espace des variables, la matrice $D = In$, où In est la matrice unité à n lignes et n colonnes, définit un **produit scalaire** :

$$< X \mid Y >_{D_{\frac{1}{n}}} = < X \mid D_{\frac{1}{n}} \mid Y > = \sum_i \frac{1}{n} x_i y_i = \frac{1}{n} \sum_i x_i y_i = \frac{1}{n} < X \mid Y >$$

Moyenne d'une variable statistique.

$$1_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

La moyenne de la variable statistique X :

$$\overline{X} = \frac{1}{n} \sum_\omega X(\omega) = \frac{1}{n} \sum_i x_i = \frac{1}{n} \sum_i x_i \times 1 = < X \mid D_{\frac{1}{n}} \mid 1_n >. = < X \mid 1_n >_{D_{\frac{1}{n}}}$$

Soit X_0 la variable centrée correspondant à X : pour chaque individu ω de population, sa valeur est $X(\omega) - \overline{X}$:

$$X_0 = \begin{pmatrix} x_1 - \overline{X} \\ \vdots \\ x_i - \overline{X} \\ \vdots \\ x_n - \overline{X} \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{pmatrix} - \overline{X} \begin{pmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{pmatrix} = X - \overline{X}\, 1_n.$$

$$\boxed{X = X_0 + \overline{X}\, 1_n = X_0 + < X \mid 1_n >_{D_{\frac{1}{n}}} 1_n}$$

33

Liens entre géométrie et variables statistiques (dans $\mathbb{R}^2 \to \mathbb{R}^p$) (3)

Variance d'une variable statistique

$$s^2(X) = \overline{X_0^2} = \frac{1}{n} \sum_i (x_i - \overline{X})^2 = < X_0 \mid D_{\frac{1}{n}} \mid X_0 > = \| X_0 \|^2$$

$$\boxed{s^2(X) = \| X_0 \|^2}$$

Covariance

$$Cov(X, Y) = \frac{1}{n} \sum_i (x_i - \overline{X})(y_i - \overline{Y}) = < X_0 \mid D_{\frac{1}{n}} \mid Y_0 > = < X_0 \mid Y_0 >_{D_{\frac{1}{n}}}$$

$$\boxed{Cov(X, Y) = < X_0 \mid D_{\frac{1}{n}} \mid Y_0 > = < X_0 \mid Y_0 >_{D_{\frac{1}{n}}}}$$

Coefficient de corrélation linéaire

$$r_{XY} = \frac{Cov[X,Y]}{s[X]\, s[Y]} = \frac{\left\langle X_0 \middle| D_{\frac{1}{n}} \middle| Y_0 \right\rangle}{\|X_0\|_\Phi \|Y_0\|_\Phi} = cos(X_0, Y_0)$$

$$\boxed{r_{XY} = cos(X_0, Y_0)}$$

34

4.9 Inertie statistique (= inertie projetée)

Pour le vecteur unitaire **u** qui définit l'axe factoriel D $u = \begin{pmatrix} \beta \\ -\alpha \end{pmatrix}$.

Inertie statistique est : $I_S(u) = \frac{1}{n} \sum_{i=1}^{i=n} \| \overrightarrow{Gm_i} \|^2$

avec $\overrightarrow{Gm_i} = < \overrightarrow{GM_i} \mid u > u = (\beta x_{i0} - \alpha y_{i0}) \begin{pmatrix} \beta \\ -\alpha \end{pmatrix}$

$$I_S(u) = \frac{1}{n} \sum_{i=1}^{i=n} (\beta x_{i0} - \alpha y_{i0})^2 = \beta^2 \times \frac{1}{n} \sum_{i=1}^{i=n} x_{i0}^2 + \alpha^2 \times \frac{1}{n} \sum_{i=1}^{i=n} y_{i0}^2 - 2\alpha\beta \times \frac{1}{n} \sum_{i=1}^{i=n} x_{i0} y_{i0}$$

OR : $\frac{1}{n} \sum_{i=1}^{i=n} x_{i0}^2 = s^2(X)$, $\frac{1}{n} \sum_{i=1}^{i=n} y_{i0}^2 = s^2(Y)$, $\frac{1}{n} \sum_{i=1}^{i=n} x_{i0} y_{i0} = Cov(X, Y)$,

$$\boxed{\begin{aligned} I_S(u) &= \beta^2 s^2(X) + \alpha^2 s^2(Y) - 2\alpha\beta\, Cov(X, Y) \\ &= (\beta \; -\alpha) \begin{pmatrix} s^2[X] & Cov[X,Y] \\ Cov[X,Y] & s^2[Y] \end{pmatrix} \begin{pmatrix} \beta \\ -\alpha \end{pmatrix} \\ &= {}^t u\, A\, u \end{aligned}}$$

35

4.10. Inertie statistique *Is* et matrice de variance-covariance *A* et recherche du vecteur **u**

$$I_S(u) = \frac{1}{n} \sum_{i=1}^{i=n} (\beta x_{i0} - \alpha y_{i0})^2 = \beta^2 \times \frac{1}{n} \sum_{i=1}^{i=n} x_{i0}^2 + \alpha^2 \times \frac{1}{n} \sum_{i=1}^{i=n} y_{i0}^2 - 2\alpha\beta \times \frac{1}{n} \sum_{i=1}^{i=n} x_{i0} y_{i0}$$

Et :

$$\boxed{\begin{aligned} I_S(u) &= \beta^2 s^2(X) + \alpha^2 s^2(Y) - 2\alpha\beta\, Cov(X, Y) \\ &= (\beta \; -\alpha) \begin{pmatrix} s^2[X] & Cov[X,Y] \\ Cov[X,Y] & s^2[Y] \end{pmatrix} \begin{pmatrix} \beta \\ -\alpha \end{pmatrix} \\ &= {}^t u\, A\, u \end{aligned}}$$

36

4.11 Zoom sur la matrice de covariance *A*

La **matrice de variance-covariance**

$$A = \begin{pmatrix} s^2[X] & Cov[X,Y] \\ Cov[X,Y] & s^2[Y] \end{pmatrix}$$

$$A = \frac{1}{n}\,{}^tZ\,Z = {}^tZ\,D_{\frac{1}{n}}\,Z$$

Matrice symétrique réelle qui si est diagonalisable sera décomposable
Par : $P^{-1}A\,P = D \iff A = P\,D\,P^{-1}$ et $A\,P = PD$
Pour lesquels il existe des valeurs propres λ et vecteurs propres V associés avec $AV = \lambda V \iff (A - \lambda I)V = 0$
Donc un polynôme caractéristique : $det((A - \lambda I) = 0$
Qui représente (application linéaire de R ² dans R ² dont le noyau de l'endomorphisme est défini par la matrice $A - \lambda\,\boldsymbol{I}2$.

37

4.11 Valeurs propres et vecteurs propres de *A* *(1)*

Les valeurs propres sont solution de :

$$det((A - \lambda I_2) = 0$$

$$\lambda^2 - (s^2(X) + s^2(Y))\,\lambda + s^2(X)\,s^2(Y) - (Cov\,(X, Y))^2 = 0$$

$$P_A(\lambda) = (-1)^d\lambda^d + (-1)^{d-1}\,tr(A)^{d-1} + det(A)$$

Le discriminant de cette équation est :

$$(s^2(X) + s^2(Y))^2 - 4\,(s^2(X)\,s^2(Y) - (Cov\,(X, Y))^2) = (s^2(X) - s^2(Y))^2 + 4\,(Cov\,(X, Y))^2 \geq 0$$

Le polynôme caractéristiques (pour n=2) a 2 racines donc 2 valeurs propres distinctes ($\lambda 1$ et $\lambda 2$) avec :

- la somme des valeurs propres est la **trace** de la matrice, somme des éléments de la première diagonale $\lambda_1 + \lambda_2 = s^2(X) + s^2(Y) \geq 0$
- le produit de ces valeurs propres est le **déterminant** de la matrice

$$\lambda_1\,\lambda_2 = s^2(X)\,s^2(Y) - (Cov\,(X, Y))^2 \geq 0$$

38

4.12 Valeurs propres et vecteurs propres de *A (2)*

λ1 et λ2 les **valeurs propres de la matrice des variances - covariances**, rangées par ordre décroissant :

$$\lambda 1 > \lambda 2 > 0$$

$$\boxed{\begin{aligned}\lambda_1 &= \frac{1}{2}\left(s^2(X) + s^2(Y) + \sqrt{\left(s^2(X) - s^2(Y)\right)^2 + 4\left(Cov(X,Y)\right)^2} \right)\\ \lambda_2 &= \frac{1}{2}\left(s^2(X) + s^2(Y) - \sqrt{\left(s^2(X) - s^2(Y)\right)^2 + 4\left(Cov(X,Y)\right)^2} \right)\end{aligned}}$$

R ² possède une **base propre orthonormée**, une base $\{u1, u2\}$ orthonormée

$$A\, u_1 = \lambda_1 u_1 \, et \, A\, u_2 = \lambda_2 u_2$$

et $$\| u_1 \|^2 = 1, \| u_2 \|^2 = 1, < u_1 \mid u_2 > = 0.$$

Pour un valeur propre λ : on peut trouver un vecteur propre et u vecteur propre normée :

$$\begin{pmatrix} s^2(Y) - \lambda \\ -Cov(X,Y) \end{pmatrix} \qquad u = \frac{1}{\sqrt{\left(s^2(Y)-\lambda\right)^2 + \left(Cov(X,Y)\right)^2}} \begin{pmatrix} s^2(Y) - \lambda \\ -Cov(X,Y) \end{pmatrix}$$

39

4.11 Valeurs propres et vecteurs propres de *A (3)*

Pour une valeur propre λ : on peut trouver u un vecteur propre normée :

$$u = \frac{1}{\sqrt{\left(s^2(Y)-\lambda\right)^2 + \left(Cov(X,Y)\right)^2}} \begin{pmatrix} s^2(Y) - \lambda \\ -Cov(X,Y) \end{pmatrix}$$

Et définir deux vecteurs u1 et u2 formant une base orthonormée dans R^2 :

$$\boxed{\begin{aligned}u_1 &= \frac{1}{\sqrt{\left(s^2(Y)-\lambda_1\right)^2 + \left(Cov(X,Y)\right)^2}} \begin{pmatrix} s^2(Y) - \lambda_1 \\ -Cov(X,Y) \end{pmatrix}\\ u_2 &= \frac{1}{\sqrt{\left(s^2(Y)-\lambda_2\right)^2 + \left(Cov(X,Y)\right)^2}} \begin{pmatrix} s^2(Y) - \lambda_2 \\ -Cov(X,Y) \end{pmatrix}\end{aligned}}$$

La matrice V formée par les deux vecteurs propres u1 et u2 est une matrice orthogonale ou : $V^{-1} = {}^tV$

On peut alors obtenir $\boxed{\begin{aligned}\Lambda &= V\, A\, {}^tV \\ A &= {}^tV\, \Lambda\, V\end{aligned}}$ *(diagonalisation de la matrice A)*

40

4.12 Recherche des axes principaux (1)

Posons , pour un vecteur normé u, $v = V\, u$

Dans R2, rapporté a la base $\{u1; u2\}$ notons $v = \binom{v1}{v2}$

On a : ${}^t v = {}^t u \, {}^t V$

$$\|v\|^2 = {}^t v v = {}^t u \, {}^t V \, V \, u = {}^t u \, u = \|u\|^2 = 1$$

u est normé et *v* aussi

$$Is\,(u) = {}^t u \, A \, u = {}^t u \, {}^t V \, \Lambda \, V \, u = {}^t v \, \Lambda \, v$$

$$Is(u) = {}^t v \, \Lambda \, v = (v_1 \quad v_2) \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \lambda_1 \; {v_1}^2 + \lambda_2 \; {v_2}^2$$

Recherche de la droite D orthogonale se ramène a
Maximiser $\lambda_1 \; {v_1}^2 + \lambda_2 \; {v_2}^2$ sous la contrainte ${v_1}^2 + {v_2}^2 = 1$ avec $\lambda 1 > \lambda 2 > 0$

41

4.12 Recherche des axes principaux (2)

Recherche de la droite D orthogonale :

Maximiser $\lambda_1 \; {v_1}^2 + \lambda_2 \; {v_2}^2$ sous la contrainte ${v_1}^2 + {v_2}^2 = 1$ avec $\lambda 1 > \lambda 2 > 0$

$$\begin{aligned} Is(u) &= \lambda_1 \; {v_1}^2 + \lambda_2 \; {v_2}^2 \\ &= \lambda_1 (1 - {v_2}^2) + \lambda_2 \; {v_2}^2 \\ &= \lambda_1 - (\lambda_1 - \lambda_2) {v_2}^2 \end{aligned}$$

La valeur $\lambda_1 - (\lambda_1 - \lambda_2){v_2}^2$ avec $\lambda 1 > \lambda 2$ atteint son maximum $\lambda 1$
lorsque l'on prend $v_2 = 0$ donc $|v_1| = 1$

La direction du premier axe factoriel est prise par le vecteur *v* de coordonnées $\binom{1}{0}$dans la base (u1;u2) (vecteurs propres de A): $v = u_1$
Et $Is(u1) = \lambda 1$

Théorème : la direction du premier axe factoriel est définie par le vecteur propre associé a la plus grande valeur propre de la matrice de variances-covariances

42

4.12 Recherche des axes principaux (3)

Recherche des autres axes factoriels :
Corollaire : la direction perpendiculaire au premier axe factoriel définit le deuxième axe factoriel il est défini par le deuxième vecteur propre associé à la deuxième valeur propre de la matrice variance-covariance

La direction du deuxième axe factoriel est prise par le vecteur *v* de coordonnées $\binom{0}{1}$dans la base (u1;u2) (vecteurs propres de A): $v=u2$
Et $Is(u2) = \lambda 2$

La somme des valeurs propres est égale à la somme des variances de chaque variable = trace de la matrice de variance -covariance A

$$I_T = Is(u_1\ ) + Is(u_2)$$

Pour dimension p :
Le premier vecteur propre normé u_1 est un vecteur de R^p qui maximise l'inertie projetée
Le 2ème vecteur propre norme u_2 est un vecteur de R^p orthogonal à u_1 qui maximise à nouveau l'inertie projetée . Ainsi de suite pour $u_3 \dots u_r$ axes suivants ou r représente le rang de la matrice diagonalisée

43

4.13 En résumé l' ACP est basée sur:

- Une matrice de covariance/ corrélation
- Une décomposition de la matrice *A* en :
- une matrice de valeurs propres ordonnées et
- une matrice de vecteurs propres associés, ordonnés selon un critère de conservation maximale de l'inertie du nuage
- Une visualisation dans un espace $\mathbb{R}^k \quad k \ll p$

Lors de cette réduction de dimension : choix de la dimension k<<p ?

44

5. Choix des *k* premières composantes principales (*dim <p*)

Choix des *r premières composantes principales*
$k << p$ réduction de la dimension
objectif : garder un maximum d'information des données initiales

Mesure de cette information : le % de variance expliquée

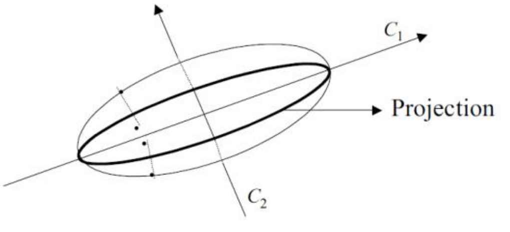
Pourquoi ce critère :

- **La conservation de la variance initiale du nuage comme hypothèse de la décomposition**
- **Mais une interprétation géométrique →**

Géométriquement : revient à projeter les données dans un sous-espace de dimension k, centré sur g, reprenant les k premiers axes principaux d'allongement du nuage

45

5.1 Exemple : données initiales à 3 dimensions distribuées dans un 'ballon de rugby'



Plus le nuage est aplati sur *C1 , C2 ⇒ moins de variance sur la 3iè dimension.*
⇒ % de variance expliquée par *C1 , C2*

En général :
- Le % de variance expliquée par *C1 , C2 , … , Cr = mesure d'aplatissement du nuage* sur le sous-espace des composantes (à k *dim.).*
Plus ce % est grand, meilleure est la représentation des données dans le sous-espace

- Les composantes principales sont entièrement déterminées par la matrice **A variance- covariance**
=> toute modification de A **-> modification des composantes**

46

5.2 Axes factoriels et taux d'inertie expliquée

Le taux d'inertie expliqué par chaque composante *i*

$$TI_i = \frac{\lambda_i}{\lambda_1 + \lambda_2 + \ \dots \lambda_p}$$

Le taux d'inertie expliqué par k composantes (soit par l'hyperplan de dimension *k*)

$$PI_{1..k} = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \ \ \lambda_i}$$

$\dfrac{\lambda_1 + \lambda_2}{\sum_{i=1}^{p} \lambda_i}$ est la part d'inertie expliquée par le premier plan principal.

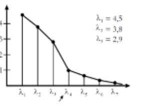
47

5.3 Différentes règles existes (a)

1) Un seuil de pourcentage d'inertie est fourni

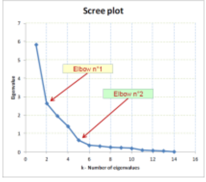
- L'inertie cumulée permet de choisir par rapport à un seuil fourni entre 80 à 90%
  - Souffre du caractère a priori

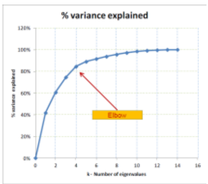
Composante	Valeur propre « Eigenvalue »	Pourcentage de variance	Pourcentage de variance cumulée
C_1	3.31217	47.3	47.3
C_2	2.61662	37.4	84.7
C_3	.57780	8.3	93.0
C_4	.23840	3.4	96.4
C_5	.13526	1.9	98.3
C_6	.08297	1.2	99.5
C_7	.03678	.5	100.0
Total :	7.00000	100.00	



Construire l'histogramme des inerties

2) la courbe de décroissance des valeurs propres λ_k (règle de Cattell).
L'idée est de détecter les « coudes » (les « cassures ») signalant un changement de structure :
- intéressante car permet de dépasser l'arbitraire purement numérique.
- mais compliquée à mettre en œuvre, car soumise à l' appréciation





Combinaisons des deux indicateurs pour choisir (*ici k=4*):

{Scree plot + % explained variance}

48

5.3 Différentes règles existes (b)

3) Règle de Kaiser - Guttman

pour ACP normée : données initiales sont centrées et réduites

ACP normée, la somme des valeurs propres est égale au nombre de variables

Kaiser critical value	1

Axis	Eigen value	% explained	% cumulated
1	5.838210	41.70%	41.70%
2	2.640156	18.86%	60.56%
3	1.953466	13.95%	74.51%
4	1.385635	9.90%	84.41%
5	0.634600	4.53%	88.94%
6	0.353217	2.52%	91.47%
7	0.310052	2.21%	93.68%
8	0.252763	1.81%	95.49%
9	0.228203	1.63%	97.12%
10	0.189341	1.35%	98.47%
11	0.092301	0.66%	99.13%
12	0.069035	0.49%	99.62%
13	0.047970	0.34%	99.96%
14	0.005051	0.04%	100.00%
Tot.	14	-	-

Un axe est retenu si sa valeur propre est >= 1

4) Règle de Karlis - Saporta - Spinaki

$$\lambda > 2\sqrt{\frac{p-1}{n-1}}$$

n	47
p	14

Karlis et al. critical value	2.063

Axis	Eigen value	% explained	% cumulated
1	5.838210	41.70%	41.70%
2	2.640156	18.86%	60.56%
3	1.953466	13.95%	74.51%
4	1.385635	9.90%	84.41%
5	0.634600	4.53%	88.94%
6	0.353217	2.52%	91.47%
7	0.310052	2.21%	93.68%
8	0.252763	1.81%	95.49%
9	0.228203	1.63%	97.12%
10	0.189341	1.35%	98.47%
11	0.092301	0.66%	99.13%
12	0.069035	0.49%	99.62%
13	0.047970	0.34%	99.96%
14	0.005051	0.04%	100.00%

Seuil = la moyenne des valeurs λ propres + 2 fois leur écart-type

Une règle plus restrictive

6. ACP $\rightarrow$construire de nouvelles variables (a)

Une projection : construire de nouvelles variables

Recherche des p composantes principales

Composantes : $C_1, C_2, \ldots, C_k, \ldots, C_p$

C_k = nouvelle variable = combinaison linéaire des variables d'origine $X_1, \ldots, X_p$:

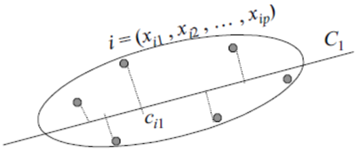
$C_k = a_{1k} X_1 + a_{2k} X_2 + \ldots\ldots + a_{pk} X_p$ coefficients a_{jk} à déterminer

tel que les C_k soient: - 2 à 2 non corrélées,

- de variance maximale,
- d'importance décroissante.

C_1 = 1ère composante principale doit être de variance maximale

Géométriquement : C_1 détermine une nouvelle direction dans le nuage de points qui suit l'axe d'allongement (étirement) maximal du nuage.



c_{i1} = coordonnée du point i sur l'axe C_1

projection de $\mathbf{x}_i$ sur C_1

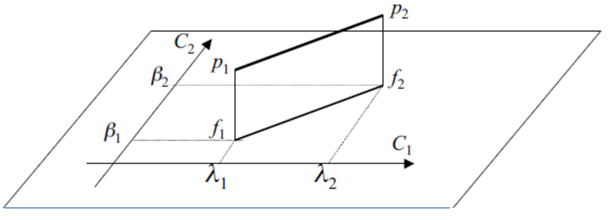
$$c_{i1} = \sum_{j=1}^{p} a_{1j} x_{ij}$$

C_1 de variance maximale les projections c_{i1} sont les plus dispersées possible.

50

6. ACP → construire de nouvelles variables (b)

*C*1 = droite passant par le centre de gravité réalisant le meilleur ajustement possible du nuage c-à-d : qui conserve au mieux la distance entre les points (après projection)
=> droite de projection assurant une distorsion minimale.
*C*2 = 2ème composante, orthogonale à *C*1 et de variance maximale.
Géométriquement : *C*2 détermine une droite perpendiculaire à *C*1 (au point *g*), suivant un axe (perpendiculaire au 1er) d'allongement maximum.
=> *C*1 et *C*2 déterminent le plan principal : le meilleur plan de projection (de distorsion minimum).



*C*1 est telle que la moyenne des *d2* (λ*i* , λ*i'*) max.
*C*2 est ⊥ à *C*1 et telle que la moyenne des *d2* (β*i* , β*i'*) max.
=> *C*1 et *C*2 déterminent le plan tel que *d2* (**f***i* , **f***i'*) soit maximum.
=> *C*3 est la droite ⊥ à *C*1 et *C*2 (par *g*) telle que la variance des coord. soit maximum ...

51

7. Quelques termes de l'ACP et notations

Nouvelles variables seront appelées : « **composantes principales** » : max p
A chaque axe est associé une variable appelée composante principale.

La composante c1 est le vecteur renfermant les coordonnées des projections des individus sur l'axe 1. La composante c2 est le vecteur renfermant les cordonnées des projections des individus sur l'axe 2.
Pour obtenir ces coordonnées, on écrit que chaque composante principale est une combinaison linéaire des variables initiales.
$C_k = a_{1k} X_1 + a_{2k} X_2 + \ldots\ldots + a_{pk} X_p$ coefficients a_{jk}

les axes qu'elles déterminent : « **axes principaux** » : il y a p axes , de direction principale donnée par le vecteur propre associé: le premier est associé a l'axe principal de direction donnée par u1 et associé à la plus grande valeur propre λ1, ...

les formes linéaires associées : « **facteurs principaux** »

52

8.1 Projection des individus selon leurs coordonnées - les composantes principales

La k ème composante principale c_k fournit les coordonnées des n individus sur le k ème axe principal. Si on désire une représentation plane des individus, la meilleure sera celle réalisée grâce aux deux premières composantes principales.

Si u_k est le vecteur propre de rang k, les coordonnées des projections des *n* points sont obtenus :

$$\mathbf{l}_k = \begin{bmatrix} l_{1k} \\ \vdots \\ l_{nk} \end{bmatrix} = \begin{bmatrix} \langle M_1 | \mathbf{u}_k \rangle \\ \vdots \\ \langle M_n | \mathbf{u}_k \rangle \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{p} (x_{1j} - \bar{x}_j) u_{jk} \\ \vdots \\ \sum_{j=1}^{p} (x_{nj} - \bar{x}_j) u_{jk} \end{bmatrix}$$

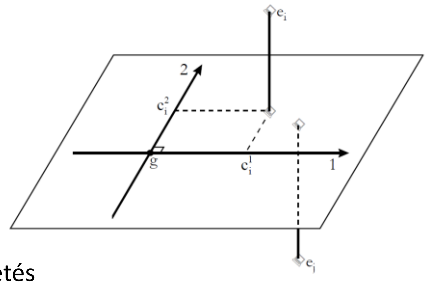
soit en écriture matricielle : $l_k = Z\ u_k$

V : matrice de vecteurs propres (chaque colonne est un vecteur propre u_k)

Z matrice des coordonnées des variables centrées

L_i^k : coordonnées de l'individu M_i sur la composante C_K (X: matrice des données dans R^p

l_k axe principal composé des n individus projetés



8. 2 Axes principaux et composante principale l_k

$\mathbf{u}_k$ est appelé **axe principal** de rang k.
$\mathbf{l}_k$ est appelé vecteur des **coordonnées** sur l'axe principal.
C'est une variable artificielle de moyenne nulle et de variance λ_k.

$$\begin{aligned} m(\mathbf{l}_k) &= \frac{1}{n}\sum_{i=1}^{n} l_{ik} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - \bar{x}_j)u_{jk} \\ &= \sum_{j=1}^{p} u_{jk}\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j) = 0 \end{aligned}$$

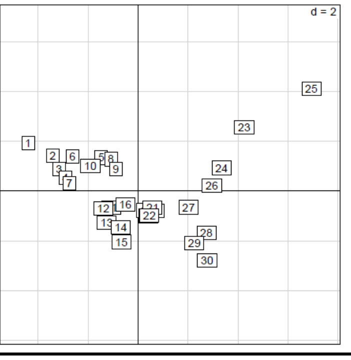
$$\begin{aligned} v(\mathbf{l}_k) &= \frac{1}{n}\sum_{i=1}^{n} l_{ik}^2 = \frac{1}{n}(\mathbf{X}\mathbf{u}_k)^T\mathbf{X}\mathbf{u}_k = \mathbf{u}_k^T\mathbf{C}\mathbf{u}_k \\ &= \lambda_k \mathbf{u}_k^T\mathbf{u}_k = \lambda_k \end{aligned}$$

54

8.3 individus projetés sur couple d'axes principaux

La représentation du nuage projeté sur un couple d'axes principaux est appelée carte factorielle. C'est une manière de voir l'information multidimensionnelle.

La carte factorielle des axes 1 et 2 est dite premier plan factoriel et représente la part maximale de la variabilité. Chaque point i est positionne par ses deux coordonnées ($l_{i1;}\ l_{i2}$).



9.1 Qualité de la projection des individus : representation des individus

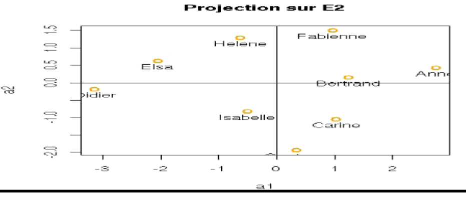
Qualité de la représentation de l'individu i sur l'espace E à k dimensions :

$$Qik = \frac{\sum_{j=1}^{k} l_i^{j2}}{\sum_{j=1}^{p} l_i^{j2}}$$

Remarque : c'est le cosinus carré de l'angle entre xi et sa projection sur l'axe j.

Qualité de la représentation sur sur Ek

Contribution de l'inertie d'un axe par un individu



9.2 Contribution d'un individu i à une composante j

La contribution d'un individu i , observation, quantifie l'importance de i dans la définition d'un vecteur propre j

$$Ct_{i\ j} = {n^{-1}(l_i^{\ j})^2}/{\lambda_j}$$

57

9.3 Variables et espace réduit R^p

2/ Interprétation des composantes principales

corrélations avec les variables initiales

	C_1	C_2	C_3	…
X_1	r_{11}	r_{12}	r_{13}	…
X_2	r_{21}	r_{22}	23	…
⋮	⋮	⋮	⋮	…
X_p	r_{p1}	r_{p2}	r_{p3}	…

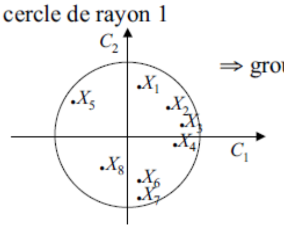
repérer les variables très corrélées

($r \approx 1$ ou $r \approx -1$)

Interprétation des 2 premières composantes C_1 , C_2 : <u>cercle des corrélations</u> :

C_1 et C_2 étant non-corrélées, on a $r^2(c_1, x_j) + r^2(c_2, x_j) \leq 1$

=> chaque variable représentée par les coordonnées : $(r(c_1, x_j), r(c_2, x_j))$ est dans un cercle de rayon 1



⇒ groupes de variables liées ou opposées

si proches de la circonférence, bien représentées par les 2 composantes !

9.4 Qualité et Contribution d'un variable à une composante

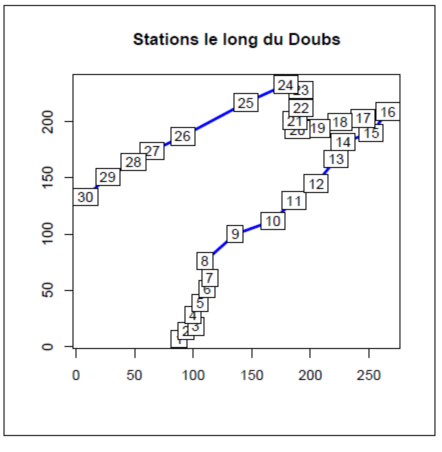
La qualité la représentation de la variable i sur un axe k est égal à la corrélation entre la variable i et le vecteur propre k
Pour une ACP réduite = au carré de la coordonnée

$$Q_{v\,ik} = {(C_{v\,ik})^2}/{(l_{v\,k})^2}$$

La contribution d'une variable i à un axe k est l'importance dans la définition du vecteur propre associé à l'axe *k* et de valeur propre λ_k
La contribution d'une variable i est égale à la valeur au carré des éléments des vecteurs propres

59

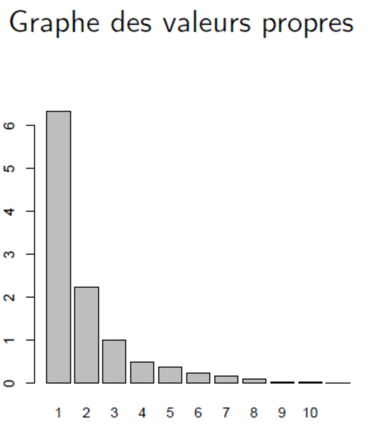
10. Un exemple (1): données initiales



11 variables physico-chimiques : distance à la source, altitude, pente
débit, pH, dureté de l'eau
phosphate, nitrate, ammoniaque
oxygène, demande biologique en oxygène

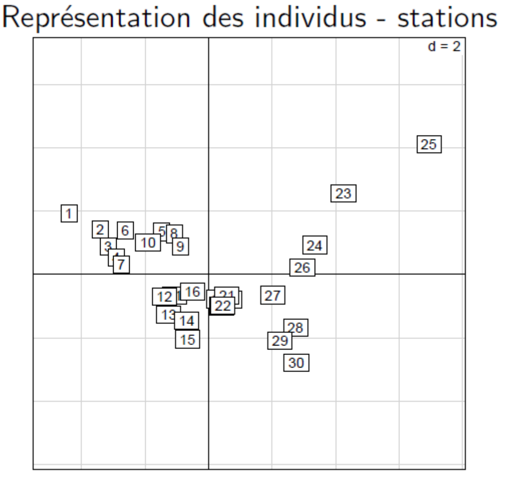
60

10. Un exemple(2): valeurs propres



61

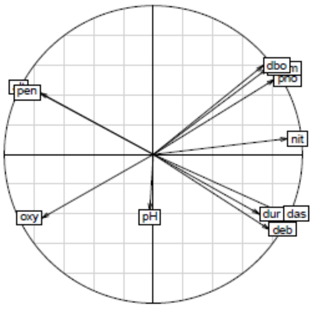
10. Un exemple(3): projections des individus sur les 2 axes factoriels



62

10. Cas de données normées corrélation entre une variable Xj et une composante k …

Représentation des variables physico-chimiques



63

11. Interprétation : un fil conducteur

- Etapes d'interprétation: Considérer l'analyse dans son ensemble

Soit répondre aux questions suivantes :

- A-t-on pu réduire la dimension du problème ?
- Combien de vecteurs propres sont importants ?
- Pourcentage de l'inertie initiale retenue ?
- Quelles sont les observations(variables) qui contribuent le plus à définir une composante ?
- Quelles sont les observations(variables) qui sont bien représentées ?
- Quels sont les signes affectant les coordonnées des observations (variables)
- Quelle est la position des différents facteurs sur les différents plans
- Y a t il des regroupements des tendances au niveau des observations si oui quelles en sont les explications ?
- Identifier de individus bizarres *(un point extrême , écarté du nuage de points peut avoir une forte contribution*) :individu standard ou à exclure ?

Quelque recommandation dans la partie visualisation : les distances sur les plans factoriels sont des distances, projetées et après réduction !!

64

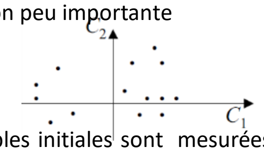
11. Interprétation : quelques repères

Représentation observations (variables) n'est valable que si le % de variance expliquée par *C1,..Ck est* suffisamment grand

- vérifier si les proximités se maintiennent dans d'autres plans de projection: *C1 - C3 , C2 - C3 , …*
- individus les mieux représentés: points proches du plan (projection peu importante

des groupes d'individus présentant des similitudes:

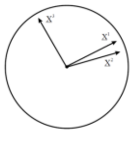
Attention aux proximités abusives dues aux projections



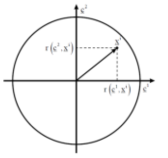
Représentation des variables

Les « proximités » entre les composantes principales et les variables initiales sont mesurées par les covariances, et surtout les corrélations, r(c j , xi) coefficient de corrélation linéaire entre la composante C_j et l'individu $Z_{i.}$

r(C_j , Z_i) est le coefficient de corrélation linéaire C_j et l'individu $Z_{i.}$



X^1 et X^2 ont une corrélation proche de 1.
X^1 et X^3 ont une corrélation proche de 0.



CERCLE DES CORRÉLATIONS

65