Analyse des données

Introduction aux bases

Aichetou Bouchareb

Outline

1. Introduction
2. Analyse uni-variée et representations graphiques
3. Application sur Iris
4. Représentation graphique des variables
5. Mesures de liaison entre deux ou plusieurs variables
6. Analyse en composantes principales (ACP)

Definition

Qu'est ce qu'une donnée?

- Terme de Mathématique, qui signifie certaines choses ou quantités, qu'on suppose être données ou connues, et dont on se sert pour en trouver d'autres qui sont inconnues, et que l'on cherche.
- Ce qui est connu ou admis comme tel, sur lequel on peut fonder un raisonnement, qui sert de point de départ pour une recherche [**?**].
- Une donnée est un renseignement qui sert de point d'appui [**?**].
- Toute information exploitable.

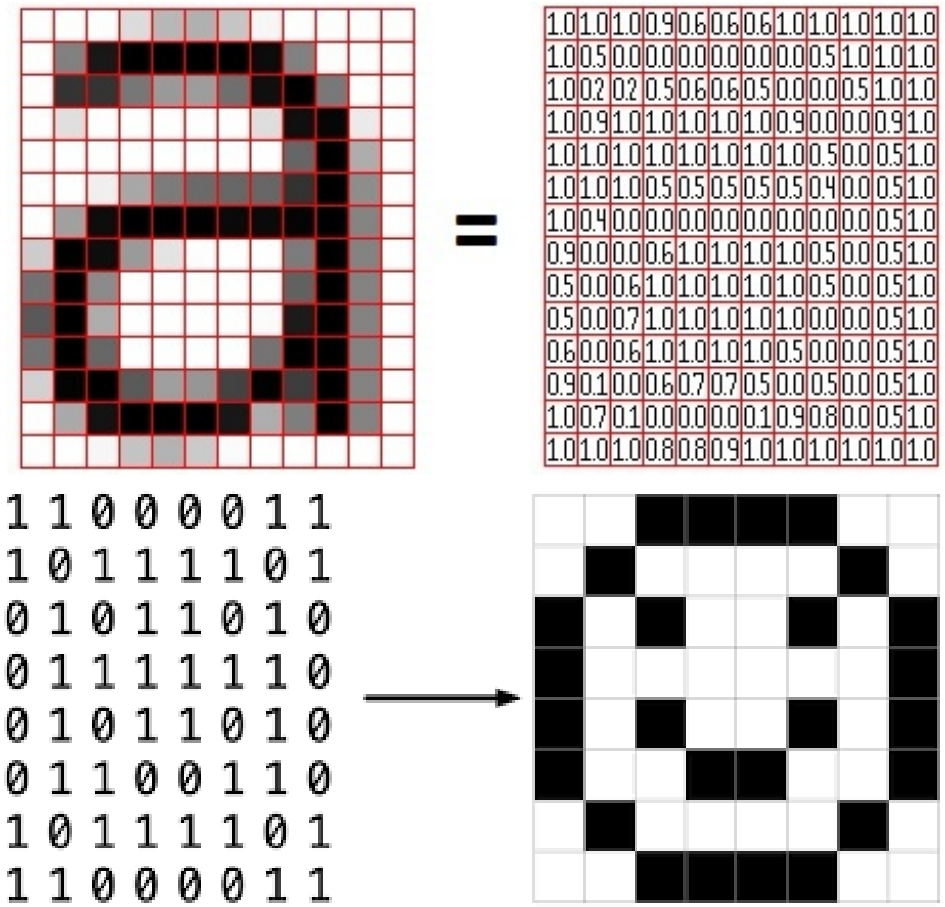
Exemples

- Mon nom est Aichetou.
- Aichetou a les yeux noirs.
- Ahmed a les yeux noirs.
- Ahmed fait 1.87m de taille.
- Alice a les yeux bleus.

$\Rightarrow$

Nom	Taille	Couleur des yeux
Aichetou	??	Noir
Ahmed	1.87	Noir
Alice	??	Bleu

Exemples (2):


1 1 0 0 0 0 1 1
1 0 1 1 1 1 0 1
0 1 0 1 1 0 1 0
0 1 1 1 1 1 1 0
0 1 0 1 1 0 1 0
0 1 1 0 0 1 1 0
1 0 1 1 1 1 0 1
1 1 0 0 0 0 1 1

L'analyse des données : Définitions et Objectifs

Definition

- Un ensemble de méthodes descriptives ayant pour objectif de résumer et visualiser l'information contenue dans un tableau de données.
- L'analyse des données est une technique d'analyse statistique d'un ensemble de données. Elle cherche à décrire des tableaux et à en exhiber des relations pertinentes.

Objectifs

Les principaux objectifs de l'analyse des données incluent :

- **Répondre aux problèmes** posés par des tableaux de grandes dimensions.
- **Résumer, organiser et visualiser les informations** contenues dans des données multidimensionnelles (un grand tableau representé sous forme d'une matrice).
- Décrire des tableaux et faire apparâitre des **relations pertinentes**.

Exemple : le jeu de données Iris

Iris est un ensemble (jeu) de données introduit en 1936 par Ronald Aylmer Fisher comme un exemple pour l'analyse discriminante. Cet ensemble contient 150 exemples de critères observés sur 3 espèces différentes de fleurs d'iris (Setosa, Versicolor, Verginica). Chaque exemple est composé de quatre attributs (longueur et largeur des sépales en mm, longueur et largeur des pétales en mm) et d'une classe (l'espèce).

SepalLength	SepalWidth	PetalLength	PetalWidth	Class
5.4	3.0	4.5	1.5	Iris-versicolor
5.5	4.2	1.4	0.2	Iris-setosa
5.5	3.5	1.3	0.2	Iris-setosa
5.5	2.3	4.0	1.3	Iris-versicolor
5.5	2.4	3.8	1.1	Iris-versicolor
5.5	2.4	3.7	1.0	Iris-versicolor
5.7	2.5	5.0	2.0	Iris-virginica
5.6	2.8	4.9	2.0	Iris-virginica
5.5	2.5	4.0	1.3	Iris-versicolor
5.5	2.6	4.4	1.2	Iris-versicolor
5.6	2.9	3.6	1.3	Iris-versicolor
5.6	3.0	4.5	1.5	Iris-versicolor

Exemples

Les notes attribuées à 9 etudiants (individus) dans 5 matières.

Sujet	Math	Sciences	Français	Latin	Musique
Jean	6	6	5	5,5	8
Aline	8	8	8	8	9
Annie	6	7	11	9,5	11
Monique	14,5	14,5	15,5	15	8
Didier	14	14	12	12	10
André	11	10	5,5	7	13
Pierre	5,5	7	14	11,5	10
Brigitte	13	12,5	8,5	9,5	12
Evelyne	9	9,5	12,5	12	18

Types des données

Definition (Types des données)

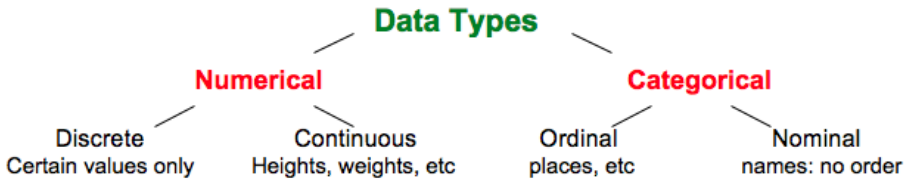
Il existe deux principaux types des données : Données Qualitatives et Données Quantitatives.

Données Qualitatives vs. Données Quantitatives

1. Une donnée quantitative mesure une quantité ou une caractéristique qui peut être mesurée et représentée sous forme numérique. Exemples: taille, age, prix, longueur, surface, temperature, vitesse, ...etc.
2. Une donnée qualitative est une qualité ou une description. Elle représente une caractéristique qui ne peut pas être mesurée ni représentée par une valeur numérique. Exemples : couleur, odeur, taste, genre, profession, ... etc.

Types des données

- Les données quantitatives sont appelées aussi données Numériques. Les données quantitatives (numériques) peuvent être discrètes ou continues.
  - Continues : tailles des étudiants en classe (1, 1.5, 1.8, 1.77, 2, ...etc).
  - Discrètes : nombre d'étudiants en classe (10, 15, 2, 3 mais pas 2.5).
- Les données qualitatives sont appelées aussi données Catégorielles. Il n'existe aucun ordre entre les valeurs d'une variable catégorielle. Dans le cas inverse, on parle de données ordinales.
  - Catégorielles ou Nominales : couleur des yeux (noir, bleu, marron, ...)
  - Ordinales : niveaux des étudiants (Passable, Bien, Très bien, ...)



Caractéristiques :

- Support des operations arithmétiques sur les données numériques.
- Une variables numérique peut prendre un nombre (potentiellement) illimité de valeurs.
- Une variables catégorielle peut prendre un nombre limité de valeurs (categories ou modalités).

Le jeu de données

Definition (Un jeu de données)

Un jeu de données est un ensemble de données représentant n individus décrits par p variables (une variable peut être numérique ou catégorielle).

- Selon les types des variables, un jeu de données peut être :
  Homogène
  ou
  Hétérogène
- Le jeu de données peut être avec ou sans valeurs manquantes.

Exemple

Un ensemble de données est souvent représenté par une matrice ou un tableau où les lignes représentent les individus et les colonnes représentent les variables.

- Données hétérogènes avec 3 individus et 3 variables (Nom: catégorielle, Taille : numérique, Couleur des yeux : catégorielle) et 2 valeurs manquantes.

$\rightarrow$

Nom	Taille	Couleur des yeux
Aichetou	??	Noir
Ahmed	1.87	Noir
Alice	??	Bleu

Analyse uni-variée : description unidimensionnelle d'une variable numérique

Il est souvent utile d'effectuer une analyse descriptive visant à décrire un ensemble de données à l'aide d'un ou plusieurs indicateurs typiques.

1. Caractéristiques de tendance centrale
   - Moyenne arithmétique : la moyenne arithmétique est égale à la somme des valeurs divisée par le nombre d'observations: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ ou

     $\overline{x} = \sum_{i=1}^{n} p_i x_i$ pour un ensemble pondéré.
   - La médiane : pour un ensemble ordonné d'observations, la médiane est la valeur qui partage cette série en deux groupes de même effectif. Si la série contient n valeurs rangées dans l'ordre croissant :
     - ★ si n est impair, on prend la $\frac{n+1}{2}$ème valeur pour médiane.
     - ★ si n est pair, on prend pour médiane la moyenne entre la $\frac{n}{2}$ème et la $\frac{n}{2}+1$ ème valeur.

* Exemples : avec un nombre impair d'observations $x_1 \leq x_2 \leq x_3 \leq x_4 \leq x_5$, la médiane est x_3. Avec un nombre pair d'observations $x_1 \leq x_2 \leq x_3 \leq x_4 \leq x_5 \leq x_6$, la médiane est $\dfrac{x_3 + x_4}{2}$.

- Le mode : on appelle *le mode*, la valeur la plus typique (la plus fréquente) d'un ensemble d'observations, c'est-à-dire celle qui apparaît le plus souvent.

Remarque

Le mode peut être utilisé pour résumer des variables catégorielles, alors que la moyenne et la médiane ne peuvent être calculées que pour les variables numériques. Il est aussi utile pour les variables discrètes et pour les variables continues lorsqu'elles sont présentées par intervalles.

3 Caractéristiques de dispersion

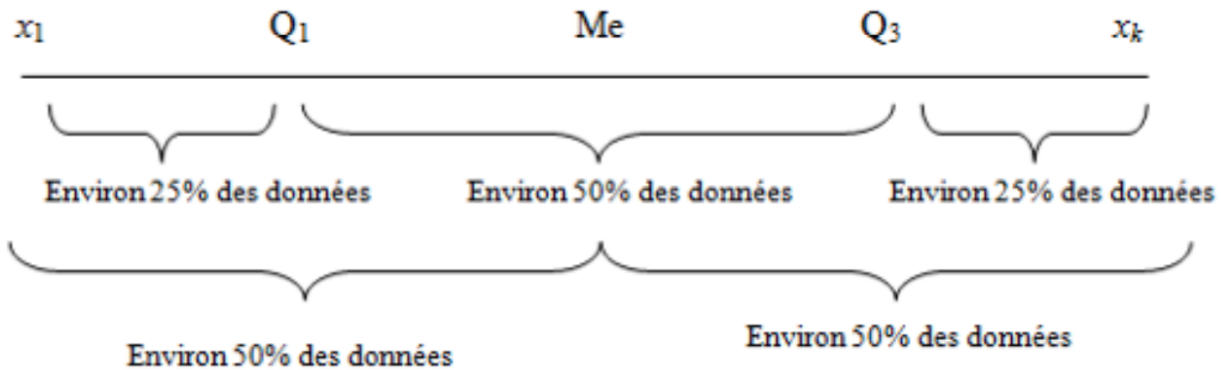
- L'étendu ou l'intervalle de variation est donnée par le rank des valeurs que prend la variable : $w = |x_{max} - x_{min}|$
- La variance et l'écart type la variance est une mesure de la dispersion ou "variation moyenne autour de la moyenne". Il s'agit de la moyenne des carrés des écarts à la moyenne : $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ ou

  $s^2 = \sum_{i=1}^{n} p_i(x_i - \overline{x})^2$. La variance s'exprime en fonction des moyennes par : la moyenne des carrées moins le carré de la moyenne :

  $s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{x})^2$

  L'écart-type (s) est donné par la racine carrée de la variance.
- Coefficient de variation : mesure relative de la dispersion autour de la moyenne, permettant de comparer le degré de variation d'un échantillon à un autre, même si les moyennes sont différentes. Elle est mesurée par le rapport entre l'écart type et la moyenne : $CV = \frac{s}{\overline{x}}$

- Les quartiles : étant donné n observations rangées par ordre croissant. Le premier quartile Q_1 est la plus petite donnée de la liste telle qu'au moins 25% des données soient inférieures ou égales à Q_1. Le troisième quartile Q_3 est la plus petite donnée de la liste telle qu'au moins 75% des données soient inférieures ou égales à Q_3.
  Méthode : Pour Q_1, on calcule $\frac{n}{4}$, puis on détermine le premier entier p supérieur ou égal à $\frac{n}{4}$. Cet entier p est le rang de Q_1. Pour Q_3, on fait de même avec $\frac{3n}{4}$
- L'écart interquartile : l'écart interquartile est mesuré par la différence entre Q_3 et Q_1 : $(Q_3 - Q_1)$.
- Les déciles et percentiles : tandis que les Quartile partagent les données en 4 groupes égaux, les déciles et percentiles partagent les données en 10 et 100 groupes égaux.



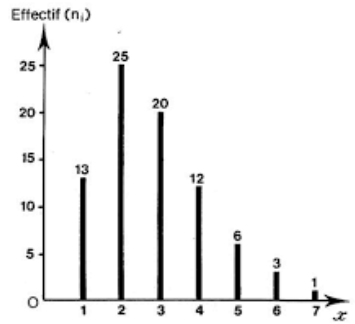
Definitions

- On nomme effectif absolu de la modalité x_i de la variable V_j, le nombre de fois où la variable V_j prends la valeur x_i. Ce simple dénombrement est aussi appelé comptage ou fréquence de la modalité.
- La somme des effectifs absolus pour la variable j est nommée effectif total de la variable.
- Le rapport effectif absolu/effectif total pour chaque modalité est nommé effectif relatif ou proportion ou pourcentage de la modalité.
- Le regroupement des effectifs relatifs est nommé tri-à-plat de la variable (ou table de fréquences).

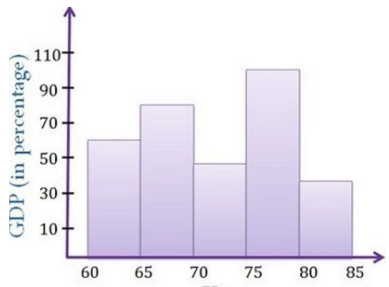
Application sur Iris

Représentation graphique d'une variable numérique

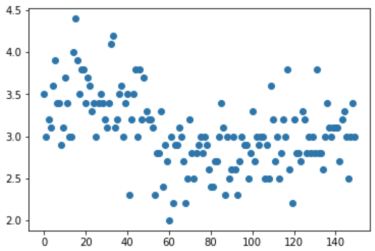
- le diagramme en batons où x représente les valeurs (ou intervalles de valeurs) et sur y nous représentons des batons correspondant aux fréquences des valeurs (ou des intervalles de valeurs).

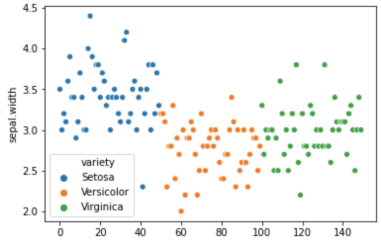


- Histogrammes : similaire au diagramme en bâtons mais l'axe x représente les intervalles de la variable et le y représente le pourcentage (ou effectif relatif) des observations appartenant à l'intervalle.

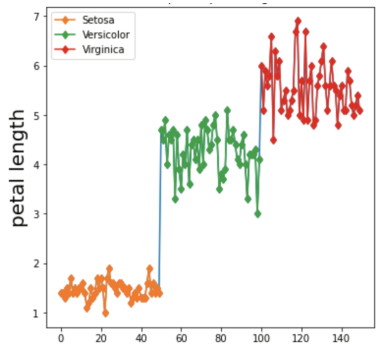


- le nuage de points (scatterplot) où l'axe x représente les indices des valeurs et y représente les valeurs de la variable

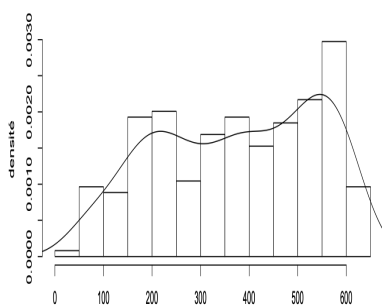


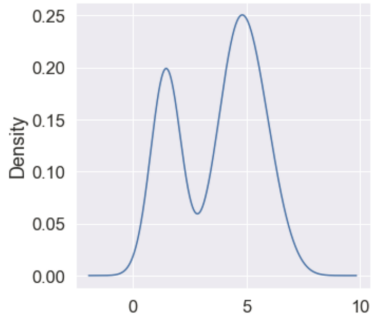


- La courbe des valeurs: similaire au nuage de point mais les observations sont liées par une line. Il est possible d'ordonner les observations pour une meilleur visualisation.

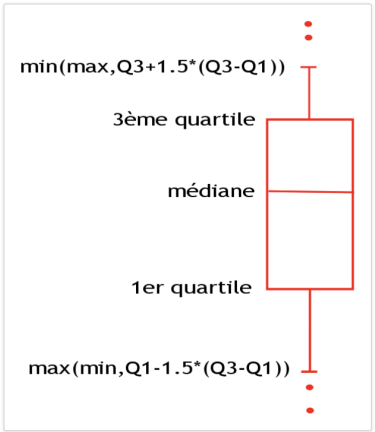


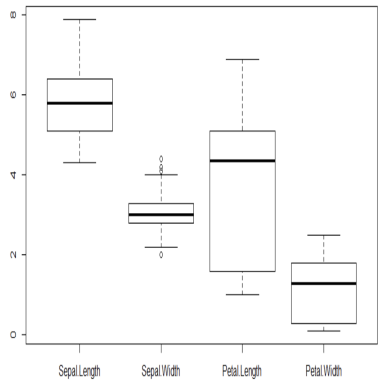
- La fonction de densité : une approximation de la densité de la variable. Cette approximation est plus precise que l'histogramme.





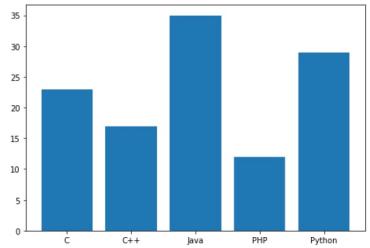
- Boîte à moustaches : représente un résumé de quelques indicateurs de tendance (médiane, quartiles, minimum, maximum ou déciles). La valeur centrale du graphique représente la médiane. Les bords du rectangle sont les quartiles et les extrémités des moustaches sont calculées en utilisant 1.5 fois l'espace interquartile.



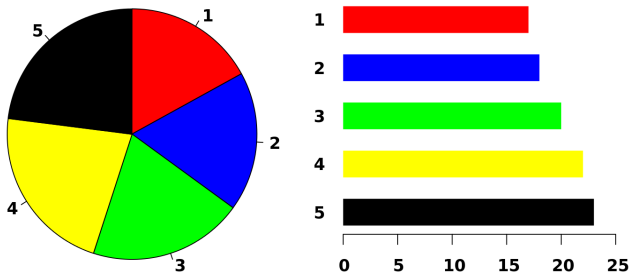


Représentation graphique d'une variable catégorielle

- Le diagramme en batons où x représente les valeurs et sur y nous représentons des batons correspondant aux fréquences des valeurs.



- Diagramme circulaire (camembert) : permet de représenter un petit nombre de valeurs (ou de categories). Le diagramme graphique est formé d'un cercle divisé en secteurs où chaque secteur représente une catégorie particulière. La surface de chaque secteur représente est proportionnelle à l'effectif de la catégorie. Les secteurs sont donc représentés par des angles proportionnels aux fréquences.



Mesures de liaison entre deux variables numériques

- Covariance : mesure de variation simultanée de deux variables X et Y. Cette mesure est définie par :

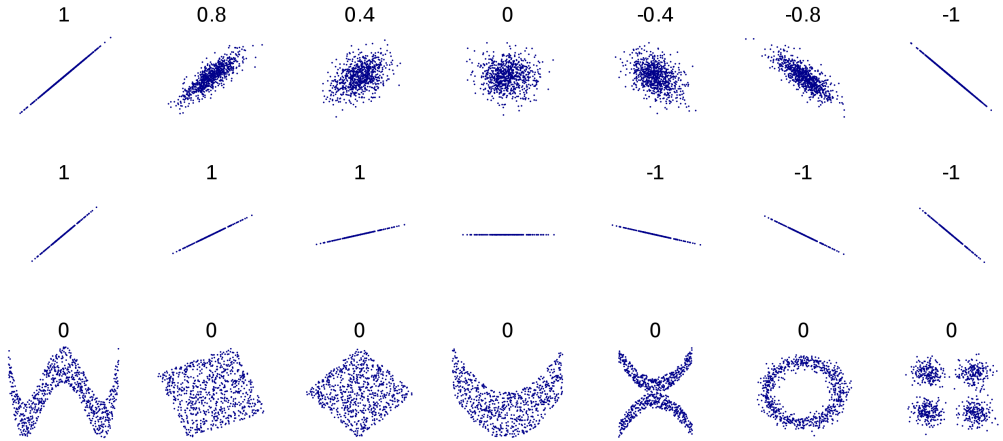
$$S_{(XY)} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

- Coefficient de correlation linéaire (Coefficient de Pearson) : mesure la force de la rélation entre deux variables X et Y.

$$r_{XY} = \frac{S_{XY}}{S_x S_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 (y_i - \overline{y})^2}}$$

où S_X désigne l'écart type de X et S_Y désigne l'écart type de Y.

Mesures de liaison entre plusieurs variables numériques



- Matrice de correlation entre plusieurs variables (matrice des variances et covariances) : la matrice de corrélation mesure les forces d'association entre p variable. Elle est composée des coefficients de corrélations des variables deux à deux.

$$\begin{pmatrix} S_1^2 & S_{12} & S_{13} & \dots & S_{1p} \\ S_{21} & S_2^2 & S_{23} & \dots & S_{2p} \\ . & . & . & \dots & . \\ S_{p1} & S_{p2} & S_{p3} & \dots & S_p^2 \end{pmatrix}$$

Mesures de liaison entre deux variables catégorielles

- Tableau de contingence : le tableau de contingence est un croisement simultané des valeurs de deux variables catégorielles.

X \ Y	d_1	...	d_k	...	d_s	total
c_1	n_{11}	...	n_{1k}	...	n_{1s}	$n_{1\bullet}$
⋮	⋮		⋮		⋮	⋮
c_h	n_{h1}	...	n_{hk}	...	n_{hs}	$n_{h\bullet}$
⋮	⋮		⋮		⋮	⋮
c_r	n_{r1}	...	n_{rk}	...	n_{rs}	$n_{r\bullet}$
total	$n_{\bullet 1}$	...	$n_{\bullet k}$	...	$n_{\bullet s}$	n

où chaque n_{ij} représente l'effectif de co-occurence simultanée de la valeur i pour la première variable et la valeur j de la deuxième variable. Les $n_{i.}$ s'appellent les marges en lignes et les $n_{.j}$ s'appellent les marges en colonnes. On peut également remplacer les effectifs par les fréquences ($\frac{n_{ij}}{n}$).

- Tableau de profils lignes : on appelle le tableau de profils lignes le tableau des fréquences relatives $\frac{n_{ij}}{n_{i.}}$.
- Tableau de profils colonnes : on appelle le tableau de profils colonnes le tableau des fréquences relatives $\frac{n_{ij}}{n_{.j}}$.

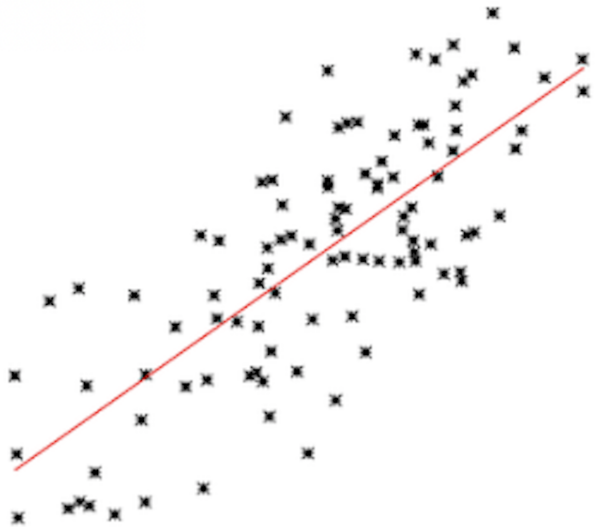
Analyse en composantes principales (ACP)

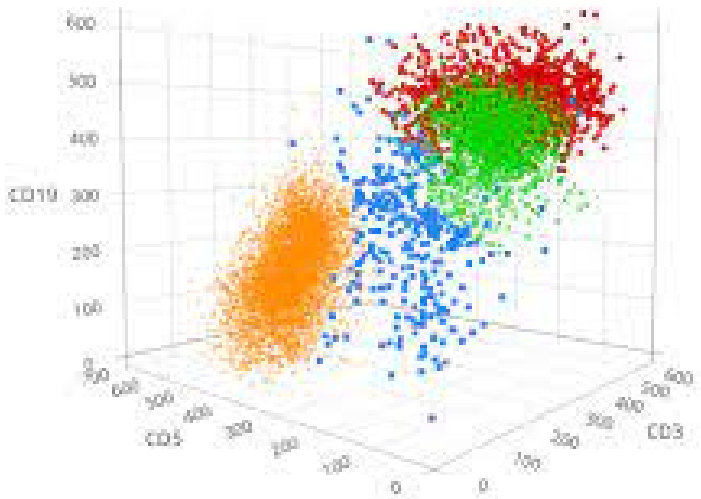
Definition

L'analyse en composantes principales est une méthode qui s'applique sur des données quantitatives. Il s'agit de l'une des méthode descriptives multidimensionnelles appelées méthodes factorielles.

Problématique

Difficulté à mettre en évidence les relations globales entre les variables dès qu'on a plus de 3 variables car impossibles à visualiser.





Solution

Condenser l'information du tableau de manière à retirer les relations vraiment caractéristiques (proximités entre variables et individus) en limitant la perte d'information $\Rightarrow$ déterminer un sous-espace de dimension $q < p$ (q nouveaux axes) sur lequel on peut faire une projection compréhensible et faible.

Ce nouveau (sous-espace) est appelé espace factoriel.
Les axes sont appelés axes factoriels (axes principaux). Les nouvelles variables sont appelés composantes principales.

Les données

Les données sont les mesures effectuées pour n individus $\{u_1, u_2, \ldots, u_i, \ldots, u_n\}$ sur p variables quantitatives $\{v_1, v_2, \ldots, v_j, \ldots, v_p\}$ représentées par un tableau des données noté $\mathbf{X}$ de la forme suivante:

$$\mathbf{X} = \begin{array}{c} \\ u_1 \\ u_2 \\ . \\ u_i \\ . \\ u_n \end{array} \begin{array}{c} \begin{array}{cccccc} v_1 & v_2 & \ldots & v_j & \ldots & v_p \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1j} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2j} & \ldots & x_{2p} \\ . & . & \ldots & . & \ldots & . \\ x_{i1} & x_{i2} & \ldots & x_{ij} & \ldots & x_{ip} \\ . & . & \ldots & . & \ldots & . \\ x_{n1} & x_{n2} & \ldots & x_{nj} & \ldots & x_{np} \end{bmatrix} \end{array}$$

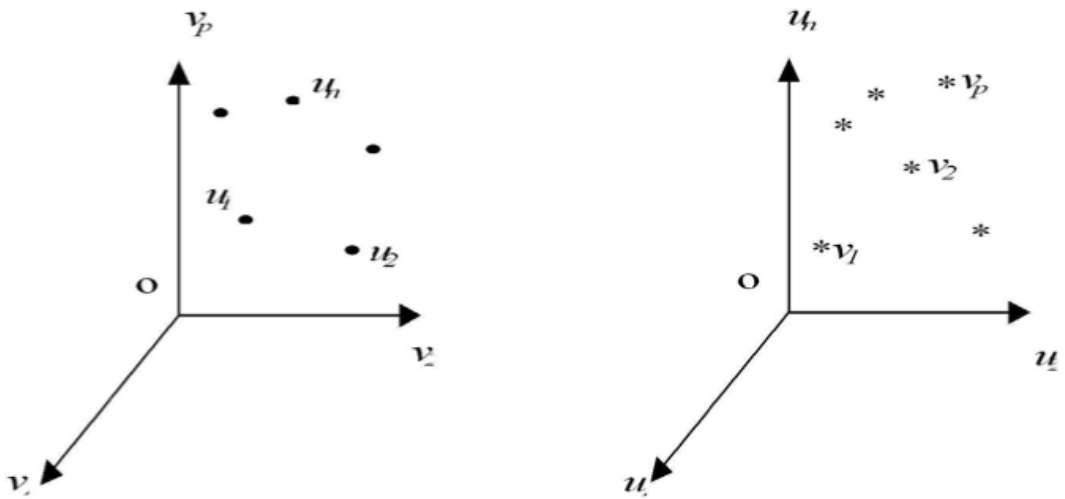
1. Vue individus : on peut représenter chaque individu par le vecteur

$u_i \in \mathbb{R}^p$ de ses mesures sur les p variables : $u_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ . \\ . \\ x_{ip} \end{pmatrix}$

L'ensemble des points qui représentent les individus est appelé "nuage des individus".

2. Vue variables : on peut représenter chaque variable par un vecteur $v_j \in \mathbb{R}^n$ dont les composantes sont les valeurs de la variable pour les n individus: $v_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ . \\ . \\ x_{nj} \end{pmatrix}$

L'ensemble des points qui représentent les variables est appelé "nuage des variables".



Objectif

Objectifs

1. Objectif général:
   - D'écrire (descriptif): visualisation de données par graphiques simples
   - Synthèse : résumé de grands tableaux individus $\times$ variables
2. En particulier, l'objectif de l'ACP est de résumer l'information portant sur les variables et sur les individus en donnant :
   - une représentation géométrique des individus permettant de faire émerger leur structure sous forme de groupes d'individus qui se ressemblent.
   - une représentations des variables permettent de faire émerger les ressemblances et liaisons linéaires (corrélations) entre variables.

$\Rightarrow$ **visualisation et étude de la matrice des corrélations.**

Attention

Réaliser une ACP n'est pas une fin en soi. L'ACP sert à mieux connaître et parler des données. On pourra aussi, se servir des représentations fournies par l'ACP pour illustrer certains résultats dans un un context exploratoire.

Exemples de jeux des données

1. Iris
2. Données températures ()

   - 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

Le nuage des individus

1 individu $\equiv$ 1 ligne du tableau $\equiv$ 1 point dans un espace à p dimensions.

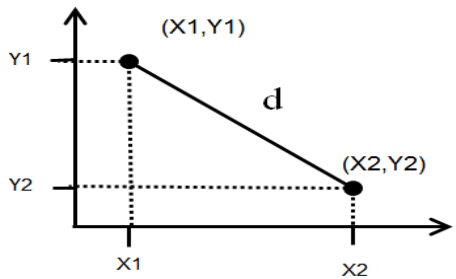
Rappel

L'un des objectifs est d'extraire des groupes d'individus similaire $\Rightarrow$ il faut choisir une distance entre deux points de l'espace des individus.

Mesure de distance

La distance utilisée par l'ACP dans l'espace des individus est la distance euclidienne classique :

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



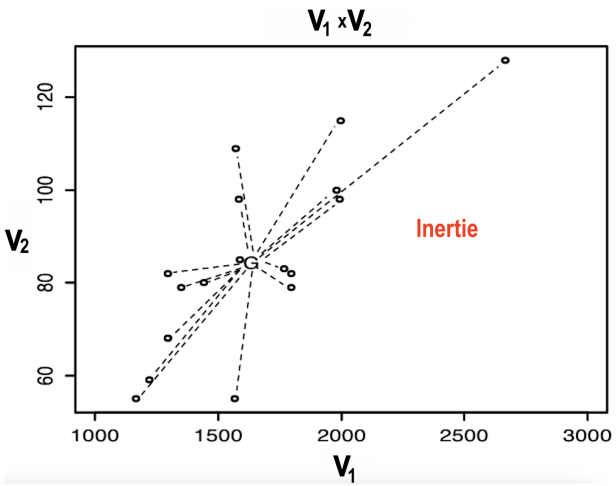
$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

L'inertie du nuage de points

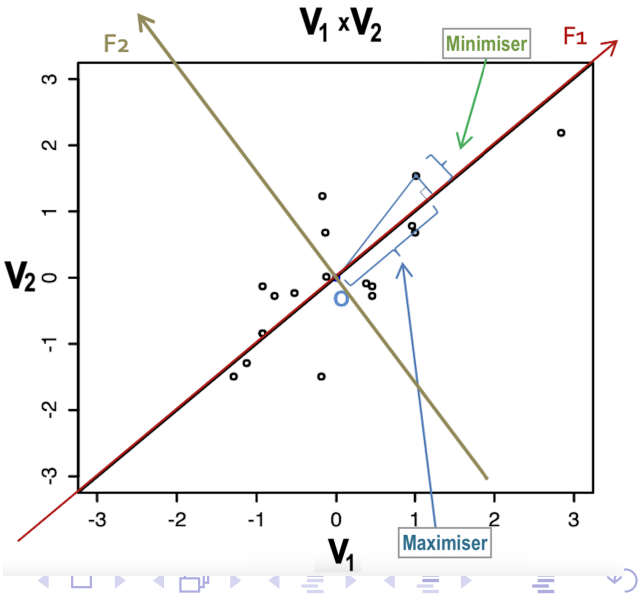
Definition (L'inertie)

L'inertie (inertie du nuage des individus par rapport au centre de gravité) est une quantification de l'information sur l'ensemble des distances : écart par rapport au barycentre multidimensionnel **G**. C'est une variance multidimensionnelle et elle est donnée par :

$$I_G = \sum_{i=1}^{n} \sum_{j=1}^{p} (x_{ij} - \overline{x}_j)^2$$



On cherche des axes (composantes principales) qui maximisent l'écart global des points par rapport à l'origine. Chaque axe porte un pourcentage de l'inertie. ⇒



L'inertie du nuage de points

On peut remarquer que : $I_G = \frac{1}{n}\sum_{i=1}^{n} {}^t(u_{ci})u_{ci}$ où u_{ci} est le vecteur centré de l'individu i et ${}^t(.)$ signifie le transposé. On peut aussi remarquer que I_G peut s'écrire comme la somme des variances des variables :

$$I_G = \sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2 = \sum_{j=1}^{p} Var(v_j)$$

Ainsi, l'inertie totale est égale à la trace de la matrice de covariance.

$$\mathbf{\Sigma} = \begin{bmatrix} \mathrm{Var}(v_1) & \mathrm{Cov}(v_1,v_2) & \dots & \mathrm{Cov}(v_1,v_j) & \dots & \mathrm{Cov}(v_1,v_p) \\ \mathrm{Cov}(v_2,v_1) & \mathrm{Var}(v_2) & \dots & \mathrm{Cov}(v_2,v_j) & \dots & \mathrm{Cov}(v_2,v_p) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathrm{Cov}(v_j,v_1) & \mathrm{Cov}(v_j,v_2) & \dots & \mathrm{Var}(v_j) & \dots & \mathrm{Cov}(v_j,v_p) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(v_p,v_1) & \mathrm{Cov}(v_p,v_2) & \dots & \mathrm{Cov}(v_p,v_j) & \dots & \mathrm{Var}(v_p) \end{bmatrix}$$

Quand les moyennes sont nulles (données centrées), la matrice de covariance peut aussi s'écrire :

$$\Sigma = \frac{1}{n}{}^t X X$$

Centrage et réduction des données

Il est coutume d'effectuer un centrage et une réduction des données avant d'appliquer une ACP. Dans ce cas, on parle d'ACP normée.

1. Centrer les données ne modifie pas la forme du nuage $\Rightarrow$ toujours centrer pour que le barycentre $\mathbb{G}$ soit situé à l'origine : $x_{ij} \leftarrow x_{ij} - \overline{x}_j$ où $\overline{x}_j$ est la moyenne de la variable v_j

$$X = \begin{bmatrix} x_{11} - \bar{x}_1 & . & . & x_{1j} - \bar{x}_j & . & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & . & . & . & . & x_{2p} - \bar{x}_p \\ . & . & . & . & . & . \\ . & . & . & x_{ij} - \bar{x}_j & . & . \\ . & . & . & . & . & . \\ x_{n1} - \bar{x}_1 & . & . & x_{nj} - \bar{x}_j & . & x_{np} - \bar{x}_p \end{bmatrix}$$

2. Réduire les données est indispensable si les unités de mesure sont différentes d'une variable à l'autre (difference dans les échelles des variables): $x_{ij} \leftarrow \dfrac{x_{ij} - \overline{x}_j}{s_j}$ où s_j est l'écart type de la variable v_j.

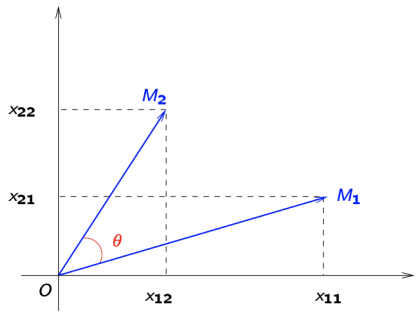
$$X = \begin{bmatrix} (x_{11} - \bar{x}_1)/\sigma_1 & . & . & (x_{1j} - \bar{x}_j)/\sigma_j & . & (x_{1p} - \bar{x}_p)/\sigma_p \\ (x_{21} - \bar{x}_1)/\sigma_1 & . & . & . & . & (x_{2p} - \bar{x}_p)/\sigma_p \\ . & . & . & . & . & . \\ . & . & . & (x_{ij} - \bar{x}_j)/\sigma_j & . & . \\ . & . & . & . & . & . \\ (x_{n1} - \bar{x}_1)/\sigma_1 & . & . & (x_{nj} - \bar{x}_j)/\sigma_j & . & (x_{np} - \bar{x}_p)/\sigma_p \end{bmatrix}$$

Nuage des variables

Une variable est considérée comme un vecteur de $\mathbf{R}^n \Rightarrow$ p vecteurs dans $\mathbf{R}^n$ où chaque axe est associé à un individu.

1. Produit scalaire
   - La norme d'un vecteur correspond à sa longueur
   - Le produit scalaire de deux vecteurs prend en compte longueurs et l'angle qu'ils forment.

$$\langle \overrightarrow{OM}_1, \overrightarrow{OM}_2 \rangle = \|\overrightarrow{OM}_1\| \times \|\overrightarrow{OM}_2\| \cos(\theta) = x_{11}x_{12} + x_{21}x_{22}$$



$$\overrightarrow{OM}_1 = \begin{pmatrix} x_{11} \\ x_{21} \end{pmatrix}$$

$$\overrightarrow{OM}_2 = \begin{pmatrix} x_{12} \\ x_{22} \end{pmatrix}$$

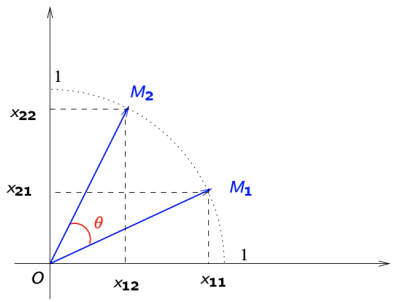
Norme : $\|\overrightarrow{OM}_1\| = \sqrt{x_{11}^2 + x_{21}^2}$

$\Rightarrow$ Pour des vecteurs de norme 1, le produit scalaire donne une mesure de l'angle (via le cosinus)

$\Rightarrow$

$$\langle \overrightarrow{OM}_1, \overrightarrow{OM}_2 \rangle = \cos(\theta) = x_{11}x_{12} + x_{21}x_{22}$$

$$\cos(\alpha) = \frac{\langle \overrightarrow{ou_i}, \overrightarrow{ou_{i'}} \rangle}{\|\overrightarrow{ou_i}\| \, \|\overrightarrow{ou_{i'}}\|} = \frac{\sum_{j=1}^{p} x_{ij}x_{i'j}}{\sqrt{\sum_{j=1}^{p} x_{ij}^2 \sum_{j=1}^{p} x_{i'j}^2}} = \frac{{}^tU_i \, U_{i'}}{\sqrt{({}^tU_i \, U_i) \, ({}^tU_{i'} \, U_{i'})}}.$$



$$\overrightarrow{OM}_1 = \begin{pmatrix} x_{11} \\ x_{21} \end{pmatrix}$$

$$\overrightarrow{OM}_2 = \begin{pmatrix} x_{12} \\ x_{22} \end{pmatrix}$$

Norme : $\|\overrightarrow{OM}_1\| = \|\overrightarrow{OM}_2\| = 1$

2. Le coefficient de corrélation entre deux variables X et Y peut s'écrire sous la forme du produit scalaire des deux colonnes centrées-réduites associées (à $\frac{1}{n}$ près) :

$$cor(X,Y) = r_{XY} = \frac{\sigma_{XY}}{\sigma_x \sigma_y}$$

où

$$\sigma_{(XY)} = S_{(XY)} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

On peut donc écrire :

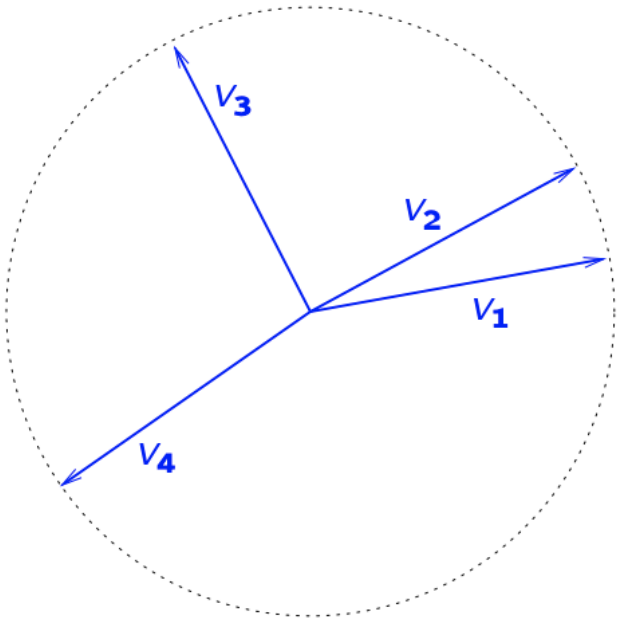
$$cor(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{\sigma_X}\right)\left(\frac{y_i - \overline{y}}{\sigma_Y}\right)$$

Quand la matrice des données est centrée-réduite alors les colonnes (variables) ont la même norme (1) et le même écart type (1), on a :

$$cor(X,Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i = \frac{1}{n} cos(X,Y)$$

Donc:

- Les p colonnes sont alors sur une (hyper)sphère (de rayon 1).
- L'angle formé par deux vecteurs colonnes (représentant les variables) correspond à leur corrélation (plus l'angle est petit, plus la corrélation est grande).



$\mathbf{cor}(V_1, V_2) \approx 1$

$\mathbf{cor}(V_1, V_4) \approx \mathbf{cor}(V_2, V_4) \approx -1$

$\mathbf{cor}(V_1, V_3) \approx \mathbf{cor}(V_2, V_3) \approx \mathbf{cor}(V_4, V_3) \approx 0$

Etude de la matrice de covariance

Rappel : Inertie, composantes, . . .

On cherche des axes (composantes principales) qui maximisent l'écart global des points par rapport à l'origine. Chaque axe porte un pourcentage de l'inertie.

La matrice Σ est symétrique définie positive. On peut trouver que:

1. L'axe premier a comme vecteur directeur unitaire le premier vecteur propre associé $\mathbf{u}_1$ à λ_1 la plus grande valeur propre de la matrice de covariance ($\Sigma = \frac{1}{n} {}^t X X$).
2. L'inertie portée par le premier axe correspond à λ_1 la plus grande valeur propre de la matrice Σ. On parle souvent de pourcentage d'information (d'inertie) expliqué par le premier axe ($\frac{\lambda_1}{I_G} \times 100$).
3. L'inertie portée par les autres axes (vecteurs propres $\mathbf{u}_k$) correspondent aux valeurs propres λ_k de Σ dans leur ordre décroissant.

4. On utilise souvent de pourcentage d'information expliqué par l'axe au lieu de la valeur propre ($\frac{\lambda_k}{I_G} \times 100$).
5. Le vecteur des coordonnées des n points du nuage des individus sur le premier axe est : $C_1 = X\mathbf{u}_1$.
   Il s'agit du vecteur des valeurs de la première composante principale sur les n individus. Cet vecteur est centré (de moyenne nulle) et de variance égale à λ_1. Les composantes principales sont les nouvelles variables créées.
6. La projection du nuage sur le k^{eme} axes est donnée par la k^{eme} composante : $C_k = X\mathbf{u}_k$.
7. Plus la fraction d'inertie expliquée par un axe est proche de 1, plus la projection des points sur cet axe est proche de points originaux.
8. Par definition, toutes les composantes sont centrées et de variances égales aux valeurs propres correspondantes et leur covariance est nulle.
9. L'inertie totale est égale à la somme de toutes les valeurs propres
   $I_G = \sum_{k=1}^{p} I_k = \sum_{k=1}^{p} \lambda_k$

9. L'inertie portée par q axes factoriels est égale à la somme des valeurs propres correspondantes à ces axes: $I = \sum_{k=1}^{q} \lambda_k$

Calcul pratique de l'ACP

- Diagonaliser $\Sigma = {}^tXX$ sous la forme $\Sigma = UD^tU$ où D est la matrice diagonale des valeurs propres (ordre décroissant).
- Les $\mathbf{u}_i$ sont dans les colonnes de U.
- Les composantes principales sont dans la matrice $C = XU$. La i^{eme} composante principale est la combinaison linéaire des variables avec les poids contenus dans la i^{eme} colonne de U.

Choix du nombre des axes

Il existe plusieurs méthodes pour choisir le nombre d'axes à retenir. Parmi les méthodes les plus utilisées :

1. Par partie d'inertie : la qualité globale d'un espace de représentation de dimension q est mesurée par la part d'inertie expliquée : $r_q = \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k}$.

La valeur de q est choisie de telle que r_q soit supérieure à une valeur seuil fixée a priori par l'utilisateur (exemple $r_q \geq 70\%$).

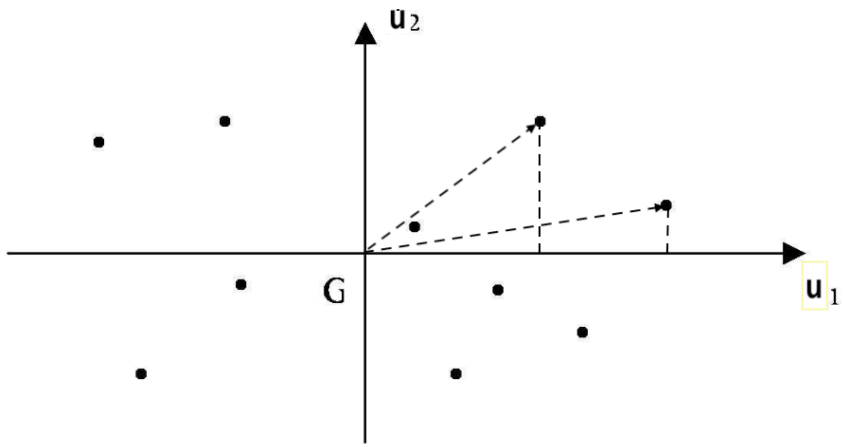
2. Utiliser le point d'inflexion de l'inertie cumulée ou quand l'ajout de plus d'axes n'augmente pas l'inertie de façon significative.
3. Scree plot : étudier la courbe de décroissance des valeurs propres, détecter les coudes (les cassures) signalant un changement de structure et choisir les axes correspondants aux valeurs propres avant le changement de structure.
4. Critère de Kaiser: on retient les axes dont l'inertie est supérieure à l'inertie moyenne $\frac{I_G}{p}$. Kaiser en ACP normée: $\frac{I_G}{p} = 1$ et on ne retiendra que les axes associés à des valeurs propre supérieures à 1.
5. Règle de Karlis–Saporta-Spinaki : choisir les axes vérifiant $\lambda > 1 + 2\sqrt{\frac{p-1}{n-1}}$.

Outline

1. Introduction
2. Analyse uni-variée et representations graphiques
3. Application sur Iris
4. Représentation graphique des variables
5. Mesures de liaison entre deux ou plusieurs variables
6. Analyse en composantes principales (ACP)
   - Représentations graphiques
   - Interpretation

Représentation des individus dans le nouvel espace

Il s'agit de la représentation des individus comme un nuage de points dans un des plans $(\mathbf{u}_i, \mathbf{u}_j)$.



Deux notions sont à prendre en compte pour interpreter les projections :

1. Les qualités de représentation: pour interpréter correctement les proximités entre les projections des individus, il faut donc s'assurer que ces individus sont bien représentés $\rightarrow$ il faut que l'angle entre le vecteur et l'axe soit petit. La qualité de représentation d'un individus u_i est donnée par : $QLT(i,k) = cos^2(individu\, u_i,\ \ axe\ \ \mathbf{u}_k) = \dfrac{(C_k^i)^2}{\sum_{j=1}^{p}(C_k^i)^2}$

$$QLT(i,k) = cos^2(individu\ u_i,\ \ axe\ \ \mathbf{u}_k) = \frac{(C_k^i)^2}{\sum_{k=1}^{p}(C_k^i)^2}$$

- Si la quantité $cos^2(individu\ u_i,\ \ axe\ \ \mathbf{u}_k)$ est proche de 1, l'individu est bien représenté par cet axe. Par contre, si cette quantité est est proche de 0, l'individu est très mal représenté par cet axe.
- Lorsque des points projections des individus sont éloignés sur un axe (ou sur un plan), on peut assurer que les points représentants ces individus sont éloignés dans l'espace. En revanche, deux individus dont les projections sont proches sur un axe (ou sur un plan) peuvent ne pas être proches dans l'espace original.
  Par contre, si deux individus sont bien représentés en projection sur un axe (ou sur un plan) et ont des projections proches, alors on pourra dire que ces deux individus sont proches dans l'espace original.
- La qualité de représentation sur un plan (ou un sous espace) est égale a la somme des cosinus carrés des angles entre l'individus et les axes engendrant le plan (ou le sous espace).

2 Les contributions des individus

- La contribution absolue d'un individu à un axe: un individu contribuera d'autant plus à la confection d'un axe, que sa projection sur cet axe sera éloignée du centre de gravité du nuage. Inversement, un individu dont la projection sur un axe sera proche du centre de gravité contribuera faiblement à l'inertie portée par cette axe.
- La contribution relative d'un individu à un axe : la contribution relative d'un individu à un axe est donnée par : $CTR(i,k) = \frac{1}{n}\frac{(C_k^i)^2}{\lambda_k}$

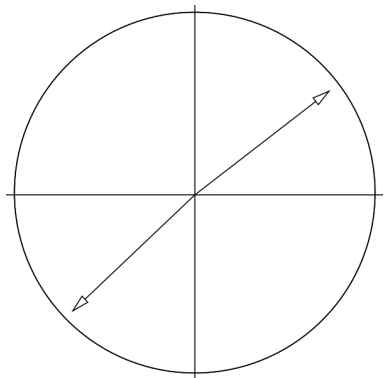
Ces contributions peuvent être utilisées pour interpréter les axes principaux avec les individus.

Le cercle de corrélations: correlations entre composantes et variables initiales

- Les coordonnées d'un variable sur un axe $\mathbf{u}_k$ est donnée par : $r(V_j, \mathbf{u}_k) = \sqrt{\lambda_k}\mathbf{u}_{k_j}$
- Chaque ancienne variable possède une corrélation avec les nouvelles variables. Ces corrélation sont utilisées pour interpréter les nouvelles variables en fonctions des anciennes. En les représentant dans un cercle de corrélations, on donne une signification aux composantes principales en fonction des variables initiales. Dans le cas fréquent des données centrées réduites, on a :

$$cor(V_j, \mathbf{u}_k) = \sqrt{\lambda_k} C_k^j$$

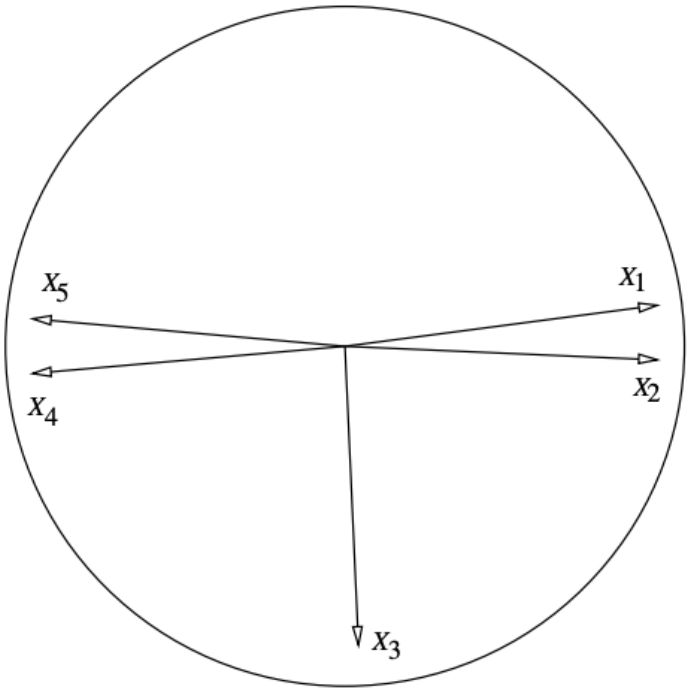
Dans le cas des données non réduites, on a $cor(V_j, \mathbf{u}_k) = \sqrt{\frac{\lambda_k}{\sigma_j}} C_k^j$.

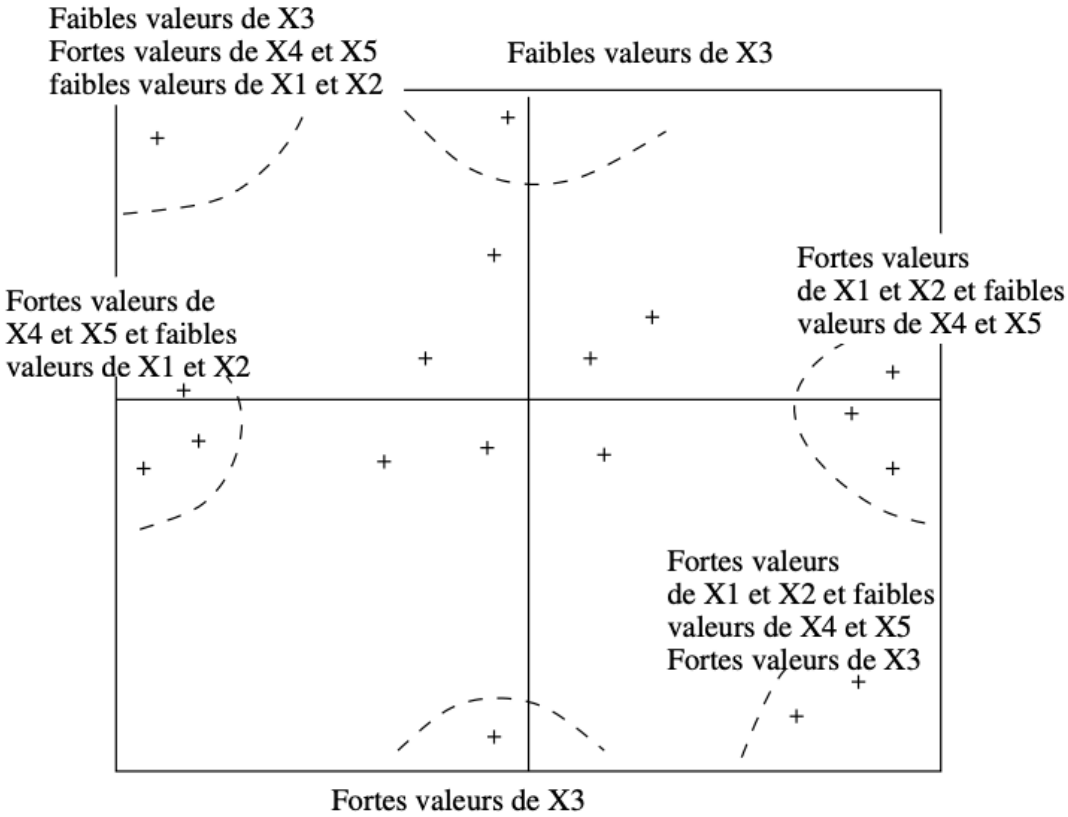


Outline

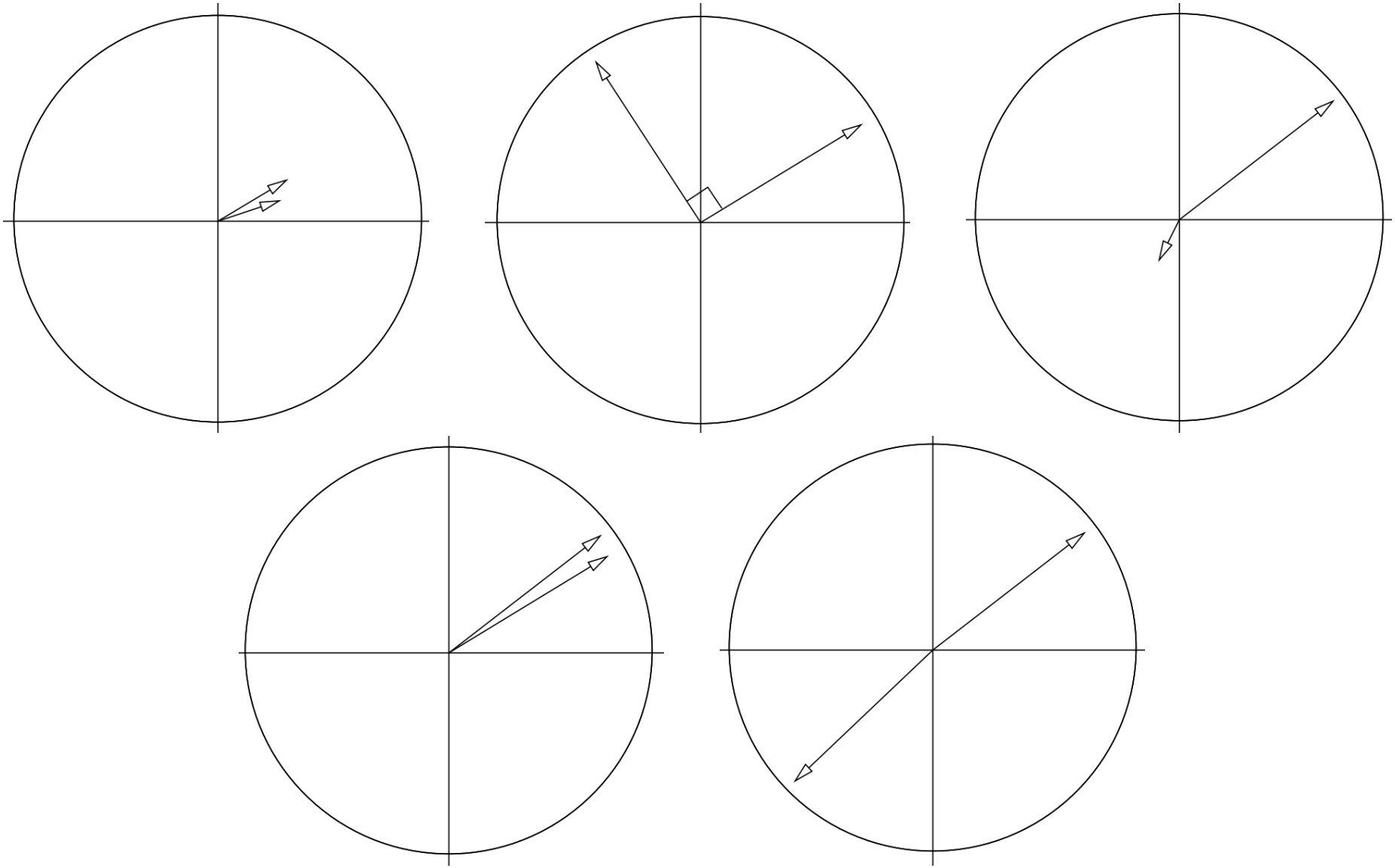
1. Introduction
2. Analyse uni-variée et representations graphiques
3. Application sur Iris
4. Représentation graphique des variables
5. Mesures de liaison entre deux ou plusieurs variables
6. Analyse en composantes principales (ACP)
   - Représentations graphiques
   - Interpretation

Interpretation





Interpretation



En plus . . .

ACP normée ou non

- Lorsque X est centré Σ est la matrice de covariance. Dans ce cas, l'ACP (non normée) est basée sur la matrice de covariances.
- Lorsque X est centré-réduit, l'ACP (normée) est basée sur la matrice de corrélations.

Variables supplémentaires

Il est possible de projeter des variables qui n'ont pas servi à la création de l'ACP (et étudier leur corrélation avec les axes principaux). Ces variables sont appelés variables supplémentaires.[a]

[a] Exercice : voir comment faire.