Méthodes d'analyse de données

- Objectifs
- Analyse en composantes principales
- Analyse canonique
- Analyse des correspondances
- Analyse discriminante
- Classification

Objectifs

- **Descriptif**
  - Analyse en Composantes Principales
  - Analyse des Correspondances
- **Explicatif**
  - Régression multiple, Segmentation
- **Décisionnel**
  - Analyse discriminante
  - Méthodes de classification (« clustering »)

Présentation des données /1

- Tableau individus x caractères

caractères

individus	Age	Taille	Poids	Lieu hab.	Sexe	Niveau Satisf.
1						
2						
3						
.						
.			x_{ij}			
.						
n						

Présentation des données /2

- Tableau de contingence

Fréquence d'association entre deux caractères qualitatifs

Arrondissements de Paris

Catégories socio professionnelles	1	2							20
1									
.... ..					x_{ij}				
8									

x_{ij} : quantité d'individus habitant dans le $j^{ième}$ arrondissement et appartenant à la $i^{ième}$ catégorie socio professionnelle

Présentation des données /3

- Tableau de proximité

Même liste d'objets

Liste d'objets (e.g., villes)

	v1	...	...	...	...	...	...	vn
v1								
...								
...				x_{ij}				
vn								

x_{ij} : distance (ressemblance) entre la ville i et la ville j

Caractère quantitatif /1

- Moyenne d'une variable x : $x_1, x_2, \ldots, x_n$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Variance

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

- Ecart type

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Caractère quantitatif /2

- Variable centrée

$$x_i - \bar{x}$$

- Variable centrée réduite

$$\frac{x_i - \bar{x}}{\sigma_x}$$

Liaison entre 2 caractères quantitatifs

x : $x_1, x_2, \ldots, x_n$ y : $y_1, y_2, \ldots, y_n$

- Covariance entre deux variables

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

- Coefficient de corrélation entre deux variables

$$r(x,y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \in [-1;+1]$$

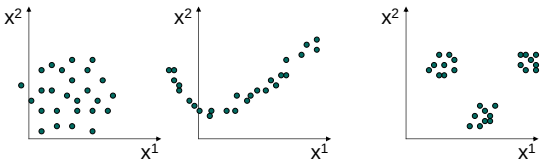
Analyse en Composantes Principales

ACP - Objectif

Décrire les données contenues dans un tableau individus x caractères numériques en réduisant l'espace de représentation

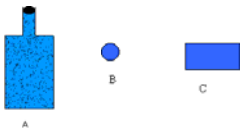
Exemple 2D

- Représentation d'un ensemble d'individus décrit par deux caractères x^1 et x^2



Pas de liaison | Liaison forte | 3 groupes

Exemple 3D



Trouver le plan permettant de conserver au mieux la représentation d'une bouteille

Utilisation de l'ACP

- Repérer des groupes d'individus, homogènes vis à vis de l'ensemble des caractères
- Révéler des différences entre individus ou groupes d'individus, relativement à l'ensemble des caractères
- Mettre en évidence des individus au comportement atypique
- Réduire l'information qui permet de décrire la position d'un individu dans l'ensemble de la population

Données pour l'ACP

- Tableau individus x caractères quantitatifs
- Données centrées ou centrées réduites
- Hypothèse de départ :

  Un nombre limité de caractères est indépendant et les autres peuvent s'en déduire

ACP Principe

- Soient n individus représentés par p caractères

  Par l'ACP, on construit un nouveau jeu de caractères plus réduit par combinaison linéaire des p caractères de départ

  ⇩

  Recherche de nouveaux axes de représentation (**axes principaux**)

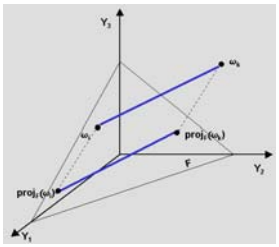
  Construction de nouveaux caractères (**composantes principales**)

  Représentation des données dans le nouvel espace

Représentation des individus

Un individu est représenté dans l'espace des caractères

On recherche un sous espace tel que la distance entre points - individus soit conservée dans la projection sur ce sous-espace. Ainsi, la ressemblance entre individus est conservée dans cette opération de projection



ACP Méthode

- Choix du premier axe

  Choisir la droite sur laquelle les distances entre points (individus) projetés sont les plus grandes possible

  C'est sur cette droite que les points sont les plus dispersés

  La variance de l'ensemble des n points sur cet axe est la plus grande

ACP Méthode

- Choix du deuxième axe

  Même principe que précédemment avec en plus la contrainte d'être perpendiculaire au premier axe

- Choix du troisième axe

  Même principe que précédemment avec en plus la contrainte d'être perpendiculaire au premier axe et au deuxième axe

Conséquence

- On passe d'un espace de **caractères initiaux corrélés** à un espace de dimension réduite de **caractères « artificiels » indépendants**
- L'ACP est une méthode factorielle

  La réduction du nombre de caractères résulte non pas d'une sélection de certains d'entre eux, mais d'une construction de nouveaux caractères obtenus par combinaison linéaire des caractères initiaux

Lien entre les p caractères pris 2 à 2

Evalué par la matrice de covariance des p caractères

$$V = \begin{pmatrix} s_1^2 & s_{12} & \cdot & \cdot & s_{1p} \\ & s_2^2 & & & \cdot \\ & & \cdot & & \cdot \\ & & & \cdot & \cdot \\ & & & & s_p^2 \end{pmatrix}$$

ou la matrice de corrélation des p caractères

$$R = \begin{pmatrix} 1 & r_{12} & \cdot & \cdot & r_{1p} \\ & 1 & & & \cdot \\ & & \cdot & & \cdot \\ & & & \cdot & \cdot \\ & & & & 1 \end{pmatrix}$$

Obtention des axes principaux

- Les axes principaux sont les vecteurs propres de la matrice des coefficients de corrélation linéaire

  Remarques :

  – Les axes principaux sont orthogonaux deux à deux

  – Le premier axe correspond à la valeur propre la plus élevée Le deuxième axe correspond à la deuxième valeur la plus élevée, etc…

Mesure de la qualité d'un axe

- Pourcentage d'inertie ou part de la variance expliquée par un axe i

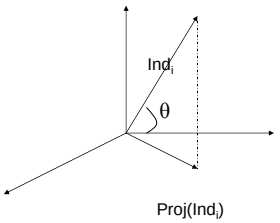
$$\frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i}$$

- Pourcentage d'inertie ou part de la variance expliquée par le sous espace des j premiers axes principaux

$$\frac{\sum_{i=1}^{j} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$$

Représentation des individus

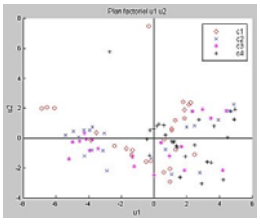
$\cos^2(\theta)$ est un indicateur de l'erreur de perspective commise en représentant l'individu i par sa projection



Diagrammes de dispersion

- Les individus sont représentés sous forme de nuages de points, dans des plans ou espaces factoriels de deux ou trois composantes principales.

Ce type de représentation permet :
-de situer chaque individu par rapport aux composantes principales
- d'identifier les individus ayant des comportements similaires, opposés ou les cas isolés



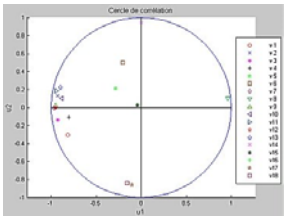
Représentation des caractères

- La corrélation entre les composantes principales et les caractères initiaux indique leur lien
- On représente chaque caractère par un point dont les coordonnées sont ses corrélations avec les axes principaux
  - Deux caractères proches géométriquement et éloignés de l'origine des axes sont très dépendants positivement
  - Deux caractères diamétralement opposés géométriquement et éloignés de l'origine des axes sont très dépendants négativement

Cercles de corrélation

- Les caractères sont représentés dans des sous espaces associés à deux composantes principales, munis d'un cercle de rayon 1 pour aider à l'interprétation.

Ce type de graphique permet de visualiser :
- les relations existant entre les variables elles-mêmes,
- la force de ces relations,
- les composantes principales qui les expliquent le mieux.



Interprétation de l'ACP

- Donner un sens aux axes factoriels découverts (les composantes principales)
- Déterminer la nature des relations existantes
  - entre ces axes et les individus
    Cette indication est donnée par les diagrammes de dispersion, qui expriment la contribution de chacun des individus à la variance des axes.
  - entre ces axes et les variables
    Cette indication est donnée par les cercles de corrélation, qui expriment en terme de coefficient de corrélation la contribution des variables aux axes factoriels.

Interprétation des cercles de corrélation

- Basée sur l'observation visuelle du graphique
  - Plus un caractère est proche de la périphérie du cercle et d'un des axes (i.e., plus sa coordonnée sur cet axe est proche de 1 ou -1), plus il est corrélé à cet axe (plus cette composante principale explique cette variable).
  - Plus deux variables sont proches sur le graphe, plus leur lien est fort.
  - Un caractère proche de l'origine des axes est peu corrélé

Interprétation des individus

- A l'aide d'indicateurs numériques :
  - le pourcentage d'information initiale expliqué par chaque composante principale
  - la qualité de la représentation des individus sur chaque axe (cos^2)
  - la contribution de chaque individu à chaque axe ou par rapport à tout le nuage