Introduction

- Lors de toute étude statistique, il est nécessaire de décrire et explorer les données avant d'en tirer de quelconques lois ou modèles prédictifs
- Dans beaucoup de situations, les données sont trop nombreuses pour pouvoir être visualisables (nombre de caractéristiques trop élevées)
- Il est alors nécessaire d'extraire l'information pertinente qu'elles contiennent ; Les techniques d'ADD répondent à ce besoin

Géométrie des nuages de points

Introduction

- Les analyses des Données est un ensemble de méthodes descriptives ayant pour objectif de résumer et visualiser l'information pertinente contenue dans un grand tableau de données
- Il existe plusieures méthodes pour analyser des données. Dans ce cours on va étudier les deux méthodes suivantes
  i) **L'analyse des composantes principales (ACP)** :
  Elle s'applique sur les variables quantitatives et a pour objectif de repérer et visualiser les ressemblances entre individus
  ii) **Analyse factorielle des correspondances (AFC et AFCM)** :
  Elle s'applique sur les variables qualitaives et a pour objectif de repérer et visualiser les corrèlations multiples entre variables ainsi que réaliser une typologie des individus

Partie I

Analyse en Correspondances Principales

Mourad El Ouali

Introduction

- **Données** : n individus observés sur p variables quantitatives. L'A.C.P. permet d'explorer les liaisons entres variables et les ressemblances entre individus.
- **Résultats** :
  i) Visualisation des individus : Notion de distances entre individus.
  ii) Visualisation des variables : en fonction de leurs corrèlations

Données

Définition : On appelle –variable– un vecteur x de taille n. Chaque coordonnèe x_i correspond à un individu. On s'intéresse ici à des valeurs numériques.

Poids : Chaque individu peut avoir un poids p_i, tel que $p_1 + \cdots + p_n = 1$, notamment quand les individus n'ont pas la même importance (échantillons redressés, données regroupées,...). On a souvent $p = 1/n$.

Géométrie des nuages de points

Exemple de Données Quantitatives

Un tableau de notes attribué à des scolaries dans cinq matiéres

	Maths	Physique	Francais	Anglais	Science
Fatima	6	6	5	5,5	8
Karim	8	8	8	8	9
Ahmed	6	7	11	9,5	11
Meriem	14,5	14,5	15,5	15	8
Fouad	14	14	12	12	10
Naima	11	10	5,5	7	13
Abdelallah	5,5	7	14	11,5	10
Khadija	13	12,5	8,5	9,5	12
Salah	9	9,5	12,5	12	18

Géométrie des nuages de points

En général

On considère un tableau de données numériques où n individus sont décrits sur p variables.

$$X = \begin{array}{c|cccccc} & X^1 & \cdots & X^j & \cdots & X^p \\ e_1 & x_1^1 & \cdots & x_1^j & \cdots & x_n^1 \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ e_i & x_i^1 & \cdots & x_i^j & \cdots & x_i^1 \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ e_n & x_n^1 & \cdots & x_n^j & \cdots & x_n^p \end{array}$$

Individu=Élément de $\mathbb{R}^p$
Variable=Élément de $\mathbb{R}^n$

Principe de L'A.C.P.

- **Interpétation.** Chaque individu est considèré comme un point d'un espace vectoriel $\mathbb{R}^p$ de dimension p. L'ensemble des individus est un nuage de points dans $\mathbb{R}^p$, dont les axe sont les p variables du tableau.

$$e_i = (x_{i1}, \cdots, xij, \cdots, x_{ip})^t$$

- **Principe.** On cherche à réduire le nombre p de variables tout en prèservant au maximum la structure du problème.
- **Méthode.** On cherche une représentation des individus dans un sous-espace de dimension k inférieure á p.

Principe de L'A.C.P.

- On projette le nuage sur un sous-espace vectoriel de dimension k inférieure à p ou la structure de nuage sera préservée le maximum possible
- Autrement dit, on cherche á définir k nouvelles variables combinaisons linéaires des p variables initiales qui feront perdre le minimum possible *d'information*.

1. Ces variables seront appelées : Composantes principales
2. les axes qu'elles déterminent : axes principaux
3. les formes linéaires associées : facteurs principaux

Distance entre deux individus

- Afin de pouvoir considèrer la structure du nuage des individus, il faut définir une distance, qui induira une géométrie.
- La distance la plus simple entre deux points de $\mathbb{R}^p$ est la distance euclidienne définie par

$$d(u,v)^2 = \sum_{j=1}^{p}(u_j - v_j)^2, \quad u,\ v \in \mathbb{R}^p$$

- Si on donne un poids $m_j > 0$ à la variable j alors,

$$d(u,v)^2 = \sum_{j=1}^{p} m_j(u_j - v_j)^2, \quad u,\ v \in \mathbb{R}^p$$

- Eloignement d'un point du nuage par rapport au centre de gravité :

$$d(e_i,g)^2 = \sum_{j=1}^{p}(x_{ij} - \bar{x}_j)^2, \quad u,\ v \in \mathbb{R}^p$$

Formulation matricielle

Matrice poids : On associe aux individus un poids p_i tel que $p_1 + \cdots + p_n = 1$. On appelle matrice poids la matrice diagonale definie comme suit

$$D_p = \begin{pmatrix} p_1 & & & 0 \\ & p_2 & & \\ & & \ddots & \\ 0 & & & p_p \end{pmatrix}$$

Cas uniforme tous les individus ont le même poids $p_i = 1/n$ et $D_p = \frac{1}{n} I_n$

Centre de gravité et tableau centré

Centre de gravité c'est le vecteur g des moyennes arithmétiques de chaque variable

$$g = (\bar{x}^1, \cdots, \bar{x}^p)^t \quad \text{avec } \bar{x}^j = \sum_{i=1}^{n} p_i x_i^j$$

Tableau centré il est obtenu en centrant les variables autour de leur centre de gravité

$$Y = (y_i^j)_{(i,j)\in[n]\times[p]} \quad \text{avec } y_i^j = (x_i^j - \bar{x}^j)$$

Exemple

- On considère le tableau de données suivant

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}$$

$\bar{x}_1 = \frac{1+1+1}{3} = 1$, $\bar{x}_2 = \frac{1+2+3}{3} = 2$ et $g = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$

- Le tableau centé est :

$$Y = \begin{pmatrix} 1-1 & 1-2 \\ 1-1 & 2-2 \\ 1-1 & 3-2 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 0 & 0 \\ 0 & 1 \end{pmatrix}$$

Variance et écart-type

Definition

La variance du vecteur colonne x^j est donnée par

$$var(x^j) = \sigma_{x^j}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^j - \bar{x^j})^2 \quad (1)$$

Généralement on a

$$var(x^j) = \sum_{i=1}^{n} p_i(x_i^j - \bar{x^j})^2 \quad (2)$$

L'écart-type σ est la racine carrée de la variance.

Mesure de la liaison entre de variables

Definition

La variance du vecteur colonne x^j est donnée par, pour $l, t \in [p]$

$$cov(x^l, x^t) = \sigma_{x^l x^t} = \sum_{i=1}^{n} p_i (x_i^l - \bar{x^l})(x_i^t - \bar{x^t}) \tag{3}$$

Le coefficient de corrèlation est donné par

$$cor(x^l, x^t) = r_{x^l x^t} = \frac{\sigma_{x^l x^r}}{\sigma_{x^l} \sigma_{x^t}} \tag{4}$$

- Deux variables x^l et x^t sont linéairement liés ssi $|cor(x^l, x^t)| = 1$.
- Dans le cas $|cor(x^l, x^t)| = 0$ on dit que les variables sont décorélées. Cela ne veut pas dire qu'ils sont indépendantes.

Matrice de variance-covariance

Definition

La matrice de variance-covariance est une matrice carrée de dimension p

$$V = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & . & . & . & \sigma_{1p} \\ \sigma_{21} & \sigma_{22}^2 & & & & \\ . & & . & & & \\ . & & & . & & \\ . & & & & . & \\ \sigma_{p1} & & & & & \sigma_{pp} \end{pmatrix}$$

Où σ_{lt} est la covariance des variables x^l et x^t et σ_j^2 est la variance de la variable x^t.

qui se calcule matriciellement comme $V = Y' D_p Y$

Matrice de corrélation

Definition

La matrice de corrélation est une matrice carrée d'ordre p

$$R = \begin{pmatrix} 1 & r_{12} & . & . & . & r_{1p} \\ r_{21} & 1 & & & & \\ . & & . & & & \\ . & & & . & & \\ . & & & & . & \\ r_{p1} & & & & & 1 \end{pmatrix}$$

Où $r_{lt} = r_{x^l x^t} = \frac{\sigma_{x^l x^r}}{\sigma_{x^l}\sigma_{x^t}}$.

qui se calcule matriciellement comme $R = D_{\frac{1}{\sigma}} V D_{\frac{1}{\sigma}}$

Données centrées réduites

C'est la matrice Z donnée par

$$Z = (z_{ij})_{(i,j)\in[n]\times[p]} \quad \text{avec} \quad z_i^j = \frac{x_i^j - \bar{x^j}}{\sigma_{x^j}} \tag{5}$$

qui se calcule matriciellement comme $Z = YD_{\frac{1}{\sigma}}$

L'A.C.P. sur les données centrées réduites

- La matrice de corrélation est données par,

$$Z' D_p Z = D_{\frac{1}{\sigma}} Y' D_p Y D_{\frac{1}{\sigma}} = D_{\frac{1}{\sigma}} V D_{\frac{1}{\sigma}} = R \qquad (6)$$

- Dans le cas uniforme $p_i = \frac{1}{n}$ on a $R = \frac{1}{n} Z'Z$.
- **Facteurs principaux** ce sont les p vecteurs orthonormés de R c-à-d les vecteurs u_k pour $k \in [p]$ tel que

$$Ru_k = \lambda u_k \quad \text{avec} < u_k, u_l >= 1 \text{ si } k = l \text{ et } 0 \text{ sinon}$$

  dont les valeurs propres verifient $\lambda_1 + \lambda_2 + \cdots + \lambda_p = p$.
- **Composantes principales** elles sont données par $c_k = Zu_k$ pour $k \in [p]$.

L'inertie

- L'inertie I_g mesure la moyenne des carrées des distances entre les individus. Dans la pratique elle est donnée par

$$I_g = tr(V) = \sum_{i=1}^{p} \sigma_i^2 \tag{7}$$

- Dans le cas de données réduites on a

$$I_g = \sum_{i=1}^{p} \lambda_i = tr(R) = p \tag{8}$$

Les axes à retenir

- **Le critère de Kaiser (variables centrées réduites).** On ne retient que les axes associés à des valeurs propres supérieures à 1, c'est-à-dire dont la variance est supérieure à celle des variables d'origine. Une autre interprétation est que la moyenne des valeurs propres étant 1, on ne garde que celles qui sont supérieures à cette moyenne.
- **Qualité de l'analyse** la qualité de la représentation obtenue par k valeurs propres est la proportion de l'inertie expliquée

$$\frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_p} \tag{9}$$