Plan du cours d'ANALYSE des DONNÉES

1 Introduction

1.1 Présentation

Exemple 1.1. Jeux de données.

- Données : beaucoup de données issues d'une enquête, d'un recensement (pas d'un sondage). Pas de modèles a priori.
- But : description à l'aide d'indices, de liaison entre variables.
- Conséquence : compréhension et explication, prévision.

1.2 Rappels d'analyse unidimensionnelle

Définition 1.1. • Tableau : n individus x p variables : $X = (x_i^j)_{1 \leq i \leq n, 1 \leq j \leq p}$.

- Variables quantitatives et qualitatives.
- Variables qualitatives : modalités, tableau de contingence : fréquence d'invidus possèdant le caractère i et j.
- Variables quantitatives : effectifs, fréquences, histogramme, polygône des fréquences cumulées, médiane, quantiles, moyenne arithmétique, variance et écart-type.

1.3 Rappels sur les espaces euclidiens

Définition 1.2. • Produit scalaire sur e.v. de dimension finie.

- Espace euclidien.

Théorème 1.1 (Cauchy-Schwartz). • $|(x|y)|^2 \leq (x|x).(y|y)$.

- Si $|(x|y)|^2 = (x|x).(y|y) \Longleftrightarrow x$ et y sont liés.

Conséquence 1.1. • $||x|| = \sqrt{(x|x)}$ est une norme sur E et $||x+y|| \leq ||x|| + ||y||$.

- $\cos(x,y) = \dfrac{(x|y)}{||x|| ||y||} \in [-1, 1]$.

- $d(x,y) = ||y-x|| = \sqrt{\sum_{i=1}^{n} m_i (y_i - x_i)^2}$, distance euclidienne sur E.

Définition 1.3. • Orthogonalité.Produit scalaire sur e.v. de dimension finie.

- Famille orthogonale, orthonormale.

Théorème 1.2 (Pythagore). • x et y sont orthogonaux $\Longleftrightarrow ||x+y||^2 = ||x||^2 + ||y||^2$.

Définition 1.4. • Distance d'un vecteur à sous s.e.v.

- Projection orthogonale dans un espace de dimension finie.

Propriété 1.1. • Si $a' = P_F(a)$ alors $a - a' \in F^{\perp}$.

- $||x||^2 = ||P_F(x)||^2 + ||P_{F^\perp}(x)||^2 \Rightarrow ||P_F(x)|| \leq ||x||$.

Théorème 1.3 (Diagonalisation des matrices symétriques). Soit M matrice de $M_n(\mathbb{R})$, symétrique. Alors M est diagonalisable dans $\mathbb{R}$ et il existe P matrice orthogonale (${}^tP.P = P.{}^tP = I_n$) telle que $D = {}^tP.M.P$ où D est une matrice diagonale constituée des valeurs propres réelles de M.

Corollaire 1.1. Si $M = {}^tX.X$ alors M diagonalisable dans $\mathbb{R}_+$.

Exemple 1.2. Ecriture matricielle de la projection orthogonale dans E ($dimE = n$) sur un s.e.v. F de base $(b_1, .., b_p)$ pour le produit scalaire associé à la matrice M :

$$P = B.({}^tB.M.B)^{-1}.{}^tB.M$$

2 Analyse bidimensionnelle

2.1 Régression entre deux variables quantitatives

Définition 2.1. Soit X et Y deux variables quantitatives.

- Nuage de point.
- Point moyen du nuage.
- Forme du nuage.

Définition 2.2. Soit X et Y deux variables quantitatives.

- Covariance de X et Y.
- Coefficient de corrélation.
- Méthode des moindres carrés.
- Équation de la droite de régression linéaire de Y par X.
- Équation de la droite de régression linéaire de X par Y.

Extensions: Légitimité de la régression, régression polynômiale, logarithmique, moindres carrés pondérés,...

2.2 Généralisation de la régression

1. La regression linéaire entre 2 variables comme une minimisation d'une distance $\Longrightarrow$ Utilisation d'une projection orthogonale et écriture matricielle.
2. Même chose avec la régression d'une variable par rapport à p variables.
3. Même chose pour la régression non-linéaire.

2.3 Lien entre deux variables qualitatives

On suppose le tableau de contingence $(n_{ij})_{1\leq i\leq k,\ 1\leq j\leq l}$. On calcule les $n_{i.}$ et les $n_{.j}$. Lorsqu'il y a "indépendance" alors $f_{ij} \sim f_{i.}f_{.j}$. On calcule

$$D^2 = \sum_{j=1}^{l}\sum_{i=1}^{k} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}.$$

Lorsque les deux variables sont indépendantes, alors

$$D^2 \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2_{(k-1)(l-1)}.$$

Grossièrement, plus D^2 devient très grand lorsqu'il y a corrélation entre les deux variables.

2.4 Lien entre variable quantitative et qualitative

On suppose que X est une variables quantitatives et Y est qualitatives avec k modalités. On suppose que l'on a n individus.
On note $\bar{X}_i$ la moyenne des x_i pour la modalité i. Il n'y a pas de corrélation entre X et Y lorsque $\bar{X}_1 \simeq \bar{X}_2 \simeq .. \simeq \bar{X}_k$.On définit ainsi un rapport de corrélation empirique η tel que:

$$\eta^2 = \frac{\sigma^2_{\bar{X}_i}}{\sigma^2_X}.$$

Grossièrement, plus η est proche de 0, moins il y a de corrélation entre X et Y.
Méthode empirique: regarder les distributions avec boxplots de X pour chaque modalité. Si les boxplots sont à peu près les mêmes, alors pas de corrélation entre X et Y.

3 Analyse en composantes principales

3.1 Présentation

On a un tableau avec n individus et p variables quantitatives.

1. Généralisation du nuage de points pour p variables.

2. Possibilité de regarder les variables 2 par 2 mais insuffisant.

3. On cherche un tableau de rang k ($k = 2$ ou $k = 3$ le plus souvent) qui soit le plus proche du tableau X (de rang $\leq p$) $\Longrightarrow$ On cherche à projeter les données dans un s.e.v., tout en gardant le maximum d'information possible.

3.2 Mise en place formelle

On considère:

- Le tableau X de n individus et p variables quantitatives.
- La matrice des poids D d'ordre n si l'on a affecté un poids différent à des individus; sinon, $D = \frac{1}{n} I_n$.
- On pose g le vecteur avec les moyennes des x_i pour chaque variable.
- Le tableau recentré Y.
- M la matrice du produit scalaire utilisé.
- La matrice des poids E d'ordre p si l'on affecte un poids différent aux variables; par défaut, E est une matrice diagonale avec l'inverse des écart-types sur la diagonale.
- La matrice des covariances V d'ordre p: $V = {}^tY.D.Y$. V est symétrique.
- La matrice des corrélations R travaillant sur les variables réduites: $R = {}^tZ.D.Z$ avec $Z = Y.E$. R est symétrique.

3.3 Inertie et projection

3.4 Axes, composantes et facteurs principaux

3.5 Interprétation d'une ACP