Généralités et Plan

Echantillon

Un échantillon est tout sous ensemble de la population sur lequel sont effectuées les observations (des attributs/descripteurs/features).

Individu ou unité statistique

Il s'agit de tout élément ω de la population ainsi $\omega \in \Omega$ (clients, étudiants, pays, années, appareils, sites Web, automobiles, images…)

Variable ou caractère

Toute application définie sur Ω. Il s'agit d'une caractéristique de la population qui est observée sur l'échantillon traité.

$$\begin{aligned} X \ : \ \Omega &\to \Theta \\ \omega &\mapsto X(\omega) \end{aligned}$$

Si $\Theta \subseteq \mathrm{R}$ alors on dit que le caractère est quantitatif, sinon on parle de caractère qualitatif.
Par exemple, si $X(\omega)$ représente l'âge de l'individu ω alors X est quantitatif. Si $X(\omega)$ représente la catégorie socioprofessionnelle de l'individu ω alors X est qualitatif.

Modalités d'un caractère

Il s'agit des différentes valeurs possibles de ce caractère c.à.d. Θ. Par exemple, Θ = {garçon, fille},
$\Theta = \{1, \ldots, 10\}$
$\Theta =]0, 250]$.

Caractère quantitatif discret ou continu

Si les modalités sont des valeurs isolées alors le caractère est dit quantitatif discret. Si les modalités sont a priori n'importe quel élément d'un intervalle réel, on parle de caractère quantitatif continu.

But de la fouille des données : Synthétiser et structurer l'information contenue dans les données (n individus et p variables) $\Longrightarrow$ deux groupes de méthodes :

I. Méthodes factorielles : dont le but est de réduire le nombre de variables en les résumant par un petit nombre de composantes synthétiques.

Analyse en Composantes Principales (ACP) : cas particulier des l'analyse factorielle pour le cas de variables quantitatives et d'individus de même poids (mass).

Analyse Factorielle de Correspondances (AFC) : cas particulier des l'analyse factorielle pour le cas de variables qualitatives.

II. Méthodes de classification : dont le but est de réduire la taille de l'ensemble des individus en formant des groupes (classes) homogènes.

Classification non hiérarchique (dite aussi classification itérative) : le nombre de classe est fixé.

Classification hiérarchique : le nombre de classes n'est pas forcément fixé en avance.

III. Méthodes de régression : Elles aident à faire ressortir les relations pouvant exister entre les différentes données, telle que la régression linéaire multiple qui décrit les variations d'une variable expliquée (quantitative) associée aux variations de plusieurs variables explicatives (quantitative).

Leçon 2. Analyse Factorielle d'un nuage de points

1 Introduction

Tableau de données : Dans un tableau de données, on y trouve en ligne les individus et en colonne les variables/attributs/features.

	V_1	...	V_j	...	V_p
X_1			:		
:			:		
X_i	X_{i1}	...	X_{ij}	...	X_{ip}
:					
X_n					

Remarques :

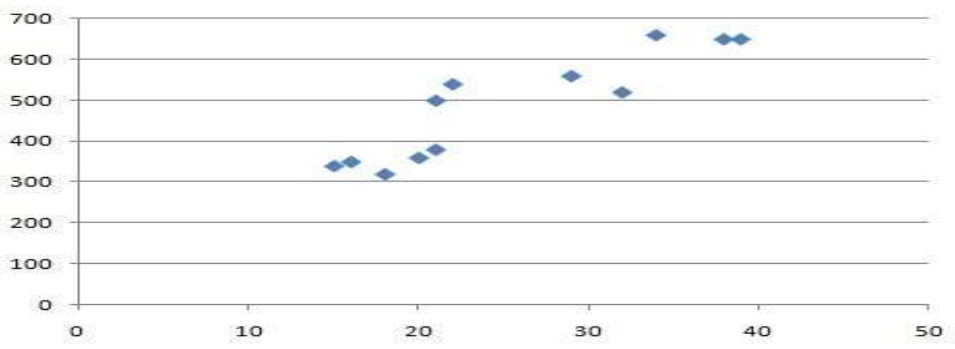
1- Chaque ligne peut être vue comme un point à p coordonnées (c.à.d un point dans un espace de dimension p qui est l'espace des variables) $-\!\!\rightarrow$ nuage de n points.

2- Chaque colonne peut être vue comme un point à n coordonnées (c.à.d un point dans un espace de dimensions qui est l'espace des individus) $-\!\!\rightarrow$ nuage de p points.

3- Dans certains cas, les points peuvent être munis de poids (ou masses) m_i (par exemple les notes attribuées à chaque étudiant et chaque note à un coefficient...). En général $m_i = 1$ ou $=1/n$ mais pas tout le temps.

L'étude des deux nuages peut présenter un grand intérêt pour décrire les relations qui existent entre les points.

Exemple : Représentation dans l'espace des variables (variable 1 : température moyenne, variable 2 : durée moyenne d'ensoleillement) des individus (mois : janvier,. . . ,décembre). Dans le tableau de données $n = 12$ et $p = 2$.



Interprétation :

1- Côté variables : forme allongée du nuage, si V_1 croit alors V_2 croit, et vice versa (deux variables positivement corrélées).

2- Côté individus : on distingue deux groupes : les mois assez froids et les mois assez chaud.

Si on avait un nombre important de variables, donc une dimension de l'espace considéré plus élevé (p>2), un tel examen n'est plus possible puisque les individus étudiés ne sont plus représentés dans un plan (espace de dimension 2) mais dans un espace de dimension plus importante.

La solution est l'analyse factorielle qui permet une visualisation de ce nuage de points en le projetant sur un espace de dimension réduite. Cette réduction (diminution du nombre de variables) s'accompagne forcément d'une déformation (erreur due à une perte d'informations suite à la réduction du nombre des attributs) qu'on doit minimiser.

2 Notion d'inertie

On considère un nuage de n points X_i de IR^p munis de poids $m_i \geq 0$ avec :

$$\sum_{i=1}^{n} m_i$$

et

$$\forall i \in \{1,...,n\}, X_i = (X_{i1},...,X_{ip})^t$$

Inertie par rapport à un point

L'inertie du nuage par rapport à un point $Y=(Y_1,\ldots,Y_p)^t$, € IR^p est définie par :

$$I(Y) = \sum_{i=1}^{n} m_i \cdot \mathrm{d}^2(X_i, Y) \quad \text{avec} \quad \mathrm{d}(X_i, Y) = \sqrt{\sum_{j=1}^{p} (X_{ij} - Y_j)^2}$$

d(Xi, Y) est la distance Euclidienne entre le point X_i et Y.

L'inertie par rapport à un point Y indique la dispersion des points X_i par rapport à Y.

Remarque : Le centre de gravité g est le point par rapport auquel l'inertie du nuage est au minimum. Ainsi, le centre de gravité est le meilleur représentant du nuage).

Théorème de Hygens :

$$\text{Si } \; g = \frac{\sum_{i=1}^{n} m_i \cdot X_i}{\sum_{i=1}^{n} m_i} \; ; \quad I(Y) = I(g) + m \cdot \mathrm{d}^2(Y, g) \quad \Longrightarrow \quad I(Y) \geq I(g) \quad \forall Y \in \mathbb{R}^p.$$

Remarque : Si on a un nuage centré, le centre de gravité est l'origine du repère O. Dans ce on dit que le nuage est centré c.à.d X_i est transformé en X_i' tel que :

$$X'_{ij} = X_{ij} - \frac{\sum_{i=1}^{n} m_i \cdot X_{ij}}{\sum_{i=1}^{n} m_i}$$

Dans ce cas :

$$\sum_{i=1}^{n} m_i \cdot X'_{ij} = 0 \quad \Longrightarrow \quad \sum_{i=1}^{n} m_i \cdot X'_i = O$$

et $I(O)$ s'appelle 'inertie du nuage' tout court.

Démonstration :

$$I(Y) = \sum_{i=1}^{n} m_i \cdot \mathrm{d}^2(X_i, Y) = \sum_{i=1}^{n} m_i \cdot \sum_{j=1}^{p} (X_{ij} - g_j + g_j - Y_j)^2 = \sum_{i=1}^{n} m_i \cdot \sum_{j=1}^{p} (X_{ij} - g_j)^2 + \sum_{i=1}^{n} m_i \cdot \sum_{j=1}^{p} (g_j - Y_j)^2 +$$
$$2 \cdot \sum_{i=1}^{n} m_i \cdot \sum_{j=1}^{p} (X_{ij} - g_j)(g_j - Y_j) \text{ or } \sum_{j=1}^{p} \left(\sum_{i=1}^{n} m_i \cdot (X_{ij} - g_j) \right) (g_j - Y_j) = 0 \text{ car } \sum_{i=1}^{n} m_i \cdot X_{ij} - m \cdot g_j = 0.$$

Inertie par rapport à un axe Δ passant par l'origine *O*

L'inertie du nuage par rapport à un axe Δ passant par l'origine O est définie par :

$$I(\Delta) = \sum_{i=1}^{n} m_i \cdot \mathrm{d}^2(X_i, \Delta) \quad \text{avec} \quad \mathrm{d}(X_i, \Delta) = \mathrm{d}(X_i, \mathrm{Proj}_{\Delta}(X_i))$$

L'inertie par rapport à un axe indique la dispersion des points X_i du nuage par rapport à cet axe.

Remarque : $I(\Delta) = 0$ si et seulement si tous les points sont alignés sur Δ.

Inertie expliquée par un axe Δ passant par l'origine O

L'inertie expliquée par un axe Δ passant par l'origine est définie par :

$$I_E(\Delta) = \sum_{i=1}^{n} m_i \cdot \mathrm{d}^2\big(\mathrm{Proj}_\Delta(X_i), O\big) = \sum_{i=1}^{n} m_i \cdot \|\mathrm{Proj}_\Delta(X_i)\|_2^2$$

L'inertie expliquée par un axe Δ passant par l'origine indique dans quelle mesure l'axe Δ ajuste-t-il la forme du nuage.

D'après le théorème de Pythagore :

$$I(O) = I(\Delta) + I_E(\Delta)$$

Donc $I(O) - I(\Delta) = I_E(\Delta)$ $\Longrightarrow$ plus les points sont proches de Δ plus l'inertie expliquée est grande.

Inertie par rapport à un sous espace vectoriel P de dimension $k < p$

L'inertie d'un nuage par rapport à un sous espace vectoriel P de dimension $k < p$ est définie par :

$$I(P) = \sum_{i=1}^{n} m_i \cdot \mathrm{d}^2\big(X_i, P\big) = \sum_{i=1}^{n} m_i \cdot \mathrm{d}^2\big(X_i, \mathrm{Proj}_P(X_i)\big)$$

Rappel : P est un sous espace vectoriel ssi P est non vide et $\forall u, v \in P$ et $\lambda \in \mathrm{R}$, $\lambda u + v \in P$.

Inertie expliquée par un sous espace vectoriel *P*

Lrinertie expliquée par un sous espace vectoriel *P* est définie par :

$$I_E(P) = \sum_{i=1}^{n} m_i \cdot \mathrm{d}^2\big(\mathrm{Proj}_P(X_i), O\big) = \sum_{i=1}^{n} m_i \cdot \|\mathrm{Proj}_P(X_i)\|_2^2$$

L'inertie expliquée par un sous espace vectoriel *P* indique dans quelle mesure le sous espace vectoriel *P* ajuste-t-il la forme du nuage.

D'après le théorème de Pythagore : $I(O) = I(P) + I_E(P)$.

3 Analyse factorielle

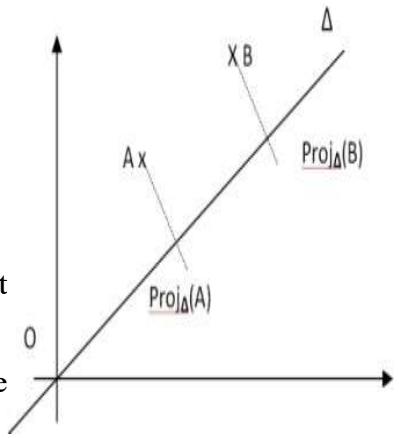
L'analyse factorielle est une opération géométrique qui consiste à partir d'un nuage de *n* points de IR^p munis de masses ≥ 0, d'élaborer une représentation la moins déformée possible de ce nuage dans un espace de dimension réduite.

Exemple : $n = 2$ et $p = 2$

Soient deux points *A* et *B* de IR^2. On veut représenter ce nuage de 2 points sur un espace de dimension 1 (une droite) avec le minimum de déformation (c.à.d en conservant la même distance entre *A* et *B)*.

Donc on va projeter *A* et *B* sur une droite Δ qui est parallèle à (*AB*) et qui passe par exemple par l'origine *O*.

Dans ce cas $d(A, B) = d(\mathrm{Proj}_\Delta(A), \mathrm{Proj}_\Delta(B))$ → Même distance donc la perte d'information est nulle



Dans le cas $n = 3$ et $p = 2$, soient trois points non alignés de IR^2, il est impossible de projeter ces trois points sans déformation $\Rightarrow$ il faudrait définir un critère de mesure pour déterminer l'axe qui provoque une déformation (perte) faible.

Première solution : Considérer la droite parallèle au plus long côté ; on conservera dans ce cas une seule distance et pas les autres.
Deuxième solution : droite (OG) avec G le centre de gravité du triangle ABC.

Remarque : Soit P un sous espace vectoriel de dimension $< p$, on a :

$$\mathrm{d}(A, B) \geq \mathrm{d}(\mathrm{Proj}_P(A), \mathrm{Proj}_P(B)).$$

Formalisation du problème

L'analyse factorielle consiste à trouver le s.e.v P qui minimise :

$$\sum_{i=1}^{n}\sum_{j=1}^{n} m_i \cdot m_j \cdot \left(\mathrm{d}^2(X_i, X_j) - \mathrm{d}^2\left(\mathrm{Proj}_P(X_i), \mathrm{Proj}_P(X_j)\right)\right)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} m_i \cdot m_j \cdot \mathrm{d}^2(X_i, X_j) - \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{n} m_i \cdot m_j \cdot \mathrm{d}^2\left(\mathrm{Proj}_P(X_i), \mathrm{Proj}_P(X_j)\right)}_{\text{à maximiser}}$$

Or si les données sont centrées,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} m_i \cdot m_j \cdot \mathrm{d}^2\left(\mathrm{Proj}_P(X_i), \mathrm{Proj}_P(X_j)\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} m_i \cdot m_j \cdot \left(\mathrm{d}^2\left(\mathrm{Proj}_P(X_i), O\right) + \mathrm{d}^2\left(\mathrm{Proj}_P(X_j), O\right)\right)$$

$$= 2m \cdot \sum_{i=1}^{n} m_i \cdot \|\mathrm{Proj}_P(X_i)\|_2^2$$

puisque $\sum_{i=1}^{n}\sum_{j=1}^{n} m_i \cdot m_j \cdot \sum_{l=1}^{p} \left((Proj_P(X_i))_l - O_l\right)\left(O_l - (Proj_P(X_j))_l\right) = 0$ *car* $\sum_{i=1}^{n} m_i Proj_P(X_i) = O.$

Dans ce cas, maximiser cette quantité revient à maximiser l'inertie expliquée par le s.e.v P : $I_E(P)$. Soient (U_1, U_2, ..., U_k) une base orthonormale de P et Δ_j l'axe de vecteur directeur U_j, pour $j = 1, \ldots k$.

$$\mathrm{Proj}_P(X_i) = \sum_{j=1}^{k} < X_i, U_j > U_j$$

Dans ce cas

$$\|\mathrm{Proj}_P(X_i)\|_2^2 = \sum_{j=1}^{k} \|\mathrm{Proj}_{\Delta_j}(X_i)\|_2^2$$

puisque $\|x + y\|^2 = \|x\|^2 + \|y\|^2 + 2 < x, y >$

Ainsi,

$$I_E(P) = \sum_{j=1}^{k} I_E(\Delta_j)$$

Pour trouver P, on doit chercher k axes Δ_i avec :

Δ_1 tel que Δ_1 passe par O et $I_E(\Delta_1)$ est maximale
Δ_2 tel que Δ_2 passe par O, $\Delta_1 \perp \Delta_2$ et $I_E(\Delta_2)$ est maximale
.
Δ_k tel que Δ_k passe par O, $\Delta_1 \perp \Delta_k$, $\Delta_2 \perp \Delta_k$ etc. et $I_E(\Delta_k)$ est maximale

Si la dimension de P n'est pas fixée, on fixe un seuil (c.à.d un pourcentage de l'inertie du nuage) et on cherche la valeur de k minimale tel que $I_E(P) \geq seuil$ (exemple seuil= 80% de $I(O)$).

Conclusion : Dans les deux cas (k *fi*xé ou non) on a à maximiser l'inertie expliquée par un axe Δ passant par l'origine.

Soit U le vecteur directeur de Δ tel que $\|U\| = 1$.

$$I_E(\Delta) = \sum_{i=1}^{n} m_i \cdot \|\mathrm{Proj}_\Delta(X_i)\|^2 = \sum_{i=1}^{n} m_i \cdot < X_i, U >^2 \|U\|^2 = \sum_{i=1}^{n} m_i \cdot < X_i, U > \cdot < X_i, U >$$

Comme X_i et U appartenant à IR^p sont des vecteurs colonnes,

$$I_E(\Delta) = \sum_{i=1}^{n} m_i \cdot {}^tU \cdot X_i \cdot {}^tX_i \cdot U = {}^tU \cdot \left(\sum_{i=1}^{n} m_i \cdot X_i \cdot {}^tX_i \right) \cdot U = {}^tU \cdot V \cdot U \text{ avec } V = {}^tX \cdot M \cdot X \text{ de taille } (p,p)$$

$$X = \begin{pmatrix} X_{11} & \dots & X_{1p} \\ \vdots & & \vdots \\ X_{i1} & \dots & X_{ip} \\ \vdots & & \vdots \\ X_{n1} & \dots & X_{np} \end{pmatrix} \begin{matrix} \leftarrow X_1 \\ \vdots \\ \leftarrow X_i \\ \vdots \\ \leftarrow X_n \end{matrix} \text{ de taille } (n,p) \quad \text{et} \quad M = \begin{pmatrix} m_1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & m_n \end{pmatrix} \text{ de taille } (n,n)$$

V est définie positive et V est symétrique (V^t=V) donc V est diagonalisable dans une base orthonormale (U_1, ..., U_p) et $V = P \cdot D \, {}^tP$ avec P la matrice des vecteurs propres (U_1, ..., U_p) et D la matrice des valeurs propres (λ_1, . . . , λ_p) tel que $\lambda_1 > \lambda_2 > ... > \lambda_p$ et $trace(V) = trace(D)$ = somme(λ_i, i=1…p).

Soit $\lambda_1 \geq \lambda_2 \geq . . . \geq \lambda_p$, si $U = U_1$ alors $I_E(\Delta) = {U_1}^t.V.U_1 = \lambda_1 \|U_1\|^2 = \lambda_1 = \max 1 \leq i \leq p(\lambda_i)$ et si $U = U_2$ alors $I_E(\Delta) = \lambda_2$ etc.

Ainsi les axes Δ_1, Δ_2,. . . , Δ_k qu'on cherche seront tels que Δ_1 de vecteur directeur U_1 qui est normé et représente le vecteur propre de V correspondant à la plus grande valeur propre $\lambda_1 > 0$ et dans ce cas $I_E(\Delta_1) = \lambda_1$. Pour l'axe Δ_2 il est de vecteur directeur U_2 tel que $U_1 \perp U_2$ et représente le vecteur propre de V correspondant à la deuxième valeur propre $\lambda_2 > 0$ et dans ce cas $I_E(\Delta_2) = \lambda_2$ etc.

Dans ca cas, la proportion de l'inertie apportée par Δ_1 est :

$$\frac{\lambda_1}{\sum_{i=1}^{k} \lambda_i} \text{ et } \sum_{i=1}^{k} \lambda_i = I_E(P)$$

Avec P le s.e.v de dim k. Les axes Δ_1, Δ_2,. . . , Δ_k s'appellent les axes factoriels.

Remarque : Si k n'est pas donné, on prend Δ_1 et on vérifie si λ_1 est supérieur ou égal au seul. Si oui, on peut projeter sans un seul axe (k=1) sinon on ajoute Δ_2 et on vérifie si $\lambda_1 + \lambda_2$ est supérieur ou égal au seul. Si oui, on peut projeter sans deux axes (k=2)…

4 Projection et interprétation des résultats

Les axes factoriels sont ainsi déterminés et on peut calculer les nouvelles coordonnées des points X_i dans le s.e.v. P de dimension k. La coordonnée de X_i sur l'axe Δ_j de vecteur directeur U_j est définie par :

$$< X_i, U_j >= \mu_j(X_i)$$

On définie la contribution du point X_i à l'inertie du nuage par :

$$\text{contr}(X_i) = \frac{m_i \cdot \|X_i\|^2}{I(O)} = \frac{m_i \cdot \|X_i\|^2}{\sum_{i=1}^{n} m_i \cdot \text{d}^2(X_i, O)} = \frac{m_i \cdot \|X_i\|^2}{\sum_{i=1}^{n} m_i \cdot \|X_i\|^2}$$

Ce paramètre indique quels sont les points qui jouent un rôle important dans l'analyse. Ainsi pour chaque point Xi et chaque axe Δj on peut définir 'contrj (Xi)' la contribution du point Xi à l'inertie expliquée par l'axe Δj par :

$$\mathrm{contr}_j(X_i) = \frac{m_i \cdot (\mu_j(X_i))^2}{\sum_{i=1}^{n} m_i \cdot (\mu_j(X_i))^2} = \frac{m_i \cdot (\mu_j(X_i))^2}{I_E(\Delta_j)} = \frac{m_i \cdot (\mu_j(X_i))^2}{\lambda_j}$$

Ce paramètre permet d'interpréter le contenu d'un axe en identifiant les points qui ont le plus contribués à son positionnement. On a donc :

$$\sum_{i=1}^{n} \mathrm{contr}_j(X_i) = 1$$

Pour chaque point X_i et chaque axe Δ_j de vecteur directeur U_j, on définit la part de l'inertie du point X_i restituée par l'axe Δ_j par :

$$\cos_j^2(X_i) = \frac{(\mu_j(X_i))^2}{\|X_i\|^2}$$

C'est le cosinus carré de l'angle formé par le vecteur U_j et le point X_i. Ce paramètre indique la qualité de représentation du point X_i sur l'axe Δ_j.

Pour chaque point X_i et pour le s.e.v. engendré par les k premiers axes, on définit la qualité de représentation du point X_i sur ce s.e.v. par :

$$QLT_k(X_i) = \sum_{j=1}^{k} \cos_j^2(X_i)$$