→ ... ... sont devenues un sep.

séance 7 : 17 - 05 - 2021

Chapitre 3 : Analyse en composantes Principales ACP

→ on a un tableau :

⇒ tableaux rectangulaires X de J variables quantitatives mesurées sur I individus : voir

x_i^j : la valeur de la variable j mesurée sur l'individu i

avec $i = 1, 2 \dots, I$ et $j = 1, 2, \dots J$

- L'individu $\underline{x_i} = (x_i^1, x_i^2, \dots, x_i^J)^t$: c'est un vecteur de J dimensions.

⇒ est un point de l'espace R^J

- La variable $\underline{x^j} = (x_1^j, x_2^j, \dots, x_I^j)^t$: c'est un vecteur de I dimensions.

⇒ est un point de l'espace R^I

* si on s'intéresse à 1 variable le j est constant et les individus changent.

* si on s'intéresse à 1 individu le i est constant et les variables changent.

⇒ donc on peut faire 2 analyses différentes : on peut analyser les individus ou bien analyser les variables.

→ **Analyse directe :**

- c'est une analyse des individus.
- Nuage $N(I)$: des I individus dans l'espace R^J.

→ **Analyse duale :**

- c'est l'analyse des variables.

- Nuage $N(J)$: des J variables dans l'espace R^I.

exemple :

-> analyser les variables, on est dans un espace de dimension.

-> par exemple lorsqu'il s'agit des étudiants et des matières : J c'est le nombre de matières et I c'est le nombre de tous les étudiants par exemple $J = 14$ et $I = 300$

-> analyse directe des individus dans l'espace R^{14} : R^{14} / R^{300}

-> analyse duale des variables dans l'espace R^{300}.

exemple :

Les individus : sont les 22 régions françaises : Paris, Rhône-Alpes, Guyane, etc (ligne)

Les variables : Démographiques, Economiques, Formation, condition de vie ..... (colonne)

-> sur chaque région on mesure un certain nombre de variables : les variables de nature démographiques, les variables de nature economique, les variables qui caractérisent la formation et les variables qui caractérisent les conditions de vie.

Démographiques : (variables démo)

- densité de ~~politien~~ population au Km^2,
- taux d'accroissement de la population, (chaque an de combien augmente la population)
- taux de natalité (= naissances vivantes pour 1000 habitants),
- taux de mortalité (= décès pour 1000 habitants)
- etc

-> tout ce qui caractérise la démographie dans cette régions.

Variables economiques :

- Produit intérieur brut par habitant (qu'est ce que la région produit par habitant)
- taux de chomâge (par région),
- Part de population active travaillant dans le secteur agricole, (%)
- " " " " industriel, (%)
- " " " " des services (%)
- etc.

Variables qui caractérisent la Formations

- proportion des sans diplômes dans la population de plus de 15 ans (% des gens qui n'ont aucun diplôme qui ont plus de 15 ans dans la région).
- proportions des diplômés du supérieur (> bac+2),
- dépenses de la région par habitant pour la formation continu.
- etc

Variables qui caractérisent les conditions de vie :

- proportion des ménages propriétaires de leur logement, (%)
- Nombre de médecins pour 1000 habitants,
- Nombre de lits d'hôpital pour 1000 habitants,
- Nombre de voitures pour 1000 habitants,
- etc.

=> pour le tableau de données : il va avoir cette forme :

-> les variables sont par blocs : si on comme toutes les variables on va avoir plusieurs variables

	Démog	Eco	Formation	condition de vie
1 2 ⋮ 22				

-> on peut avoir plus de variables que d'individus.

-> Quand on fait une enquête, et on met des fichiers à remplir par des gens, on partage l'enquête par paquets de variables : on pose des questions sur l'âge, le nb d'enfants, est ce qu'il est marié ou pas, pour les variables économique : le revenu de la personne, combien il a de comptes bancaires, pour la formation : niveau d'études, pour les variables de type conditions de vie : est ce qu'il est propriétaire de son logement, quelle est la surface de son logement ...

exemple.

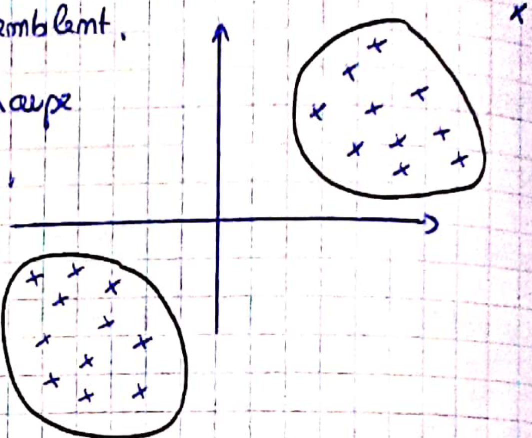
→ on peut avoir comme individus des vaches et pour les variables on prend plusieurs mesures sur les vaches : l'age de la vache, le poids, la taille, la race (qualitative) mais on peut la transformer en variable quantitative en utilisant le Hot coding.

objectif.

Pour le nuage des individus N(I)

- l'objectif d'abord c'est d'etudier le nuage des lignes (des individus) sur un espace de dimension J.
- chaque individu a J dimension, on cherche à la projeter par exemple en 2 dim. en perdant le moins d'information possible
- on met tous les individus dans un plan (étudiants par ex)
- on sait que les étudiants rapprochés se ressemblent.
- ce groupe d'étudiants s'oppose a l'autre groupe
- si un étudiant est loin des autres on l'appelle : observation abhérante càd il est différent des autres observations et il ne se comporte pas comme les autres.

ACP



→ on va déterminer aussi les observations abérrantes

⇒ ça c'est le nuage des individus

① Rechercher les groupes d'individus homogènes.

② Relever les différences entre groupes d'individus.

③ Mettre en évidence des données individus atypiques.
"
observations aberrantes

objectif

Pour le nuage des variables N(J)

on ne dit pas qu'une variable ressemble à une variable.

↳ on dit qu'une variable est corrélée avec une variable.

① Rechercher des groupes de variables corrélées.

② Construire des variables synthétiques non mesurées mais calculées.

Une variable synthétique : c'est par exemple une combinaison linéaire des variables qui représente l'ensemble des variables par exemple, la moyenne c'est une variable synthétique,

il prend la moyenne d'un étudiant c'est une moyenne pondérée sur toutes les matières.
et ça reflète le niveau d'un étudiant donc c'est une variable synthétique.

⇒ dans l'ACP aussi on va essayer de construire des variables synthétiques les plus parlant possibles.

→ ensuite on va s'intéresser à une analyse simultanée ↓

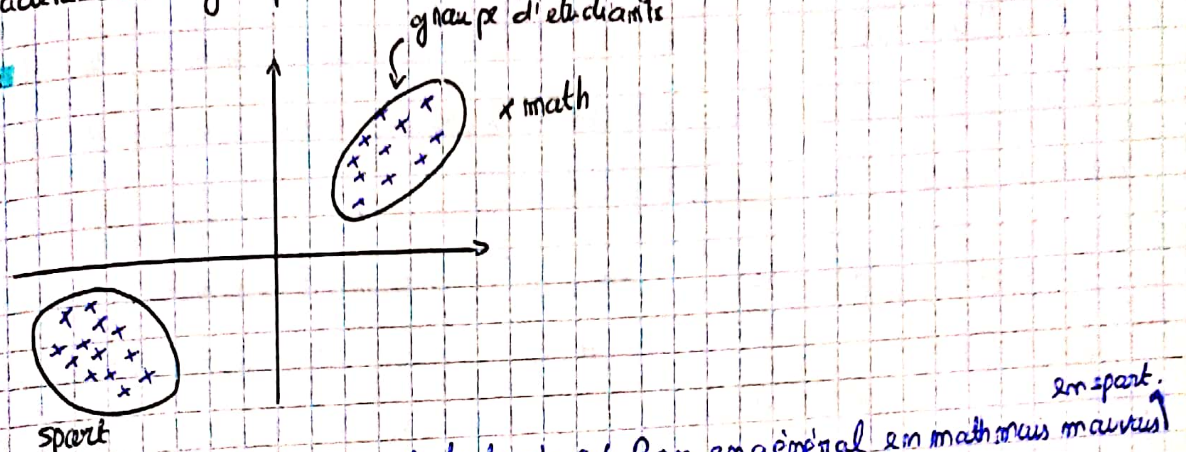
objectif :

Analyse simultanée de N(I) et de N(J) :

c'àd on va essayer d'expliquer les lignes par les colonnes et les colonnes par les lignes.

- caractériser des groupes d'individus par des groupes de variables.

exemple :



→ on deduit que le 1er groupe d'etudiants est Bon en général en math mais mauvais en sport.

→ et le 2ème groupe d'etudiants est Bon en sport mais mauvais en maths.

→ comment atteindre ces objectifs

⇒ **Analyse du nuage des variables N(J).**

- pour chaque variable colonne x^j on calcule :

- sa moyenne : $\overline{x^j} = \frac{\sum_{i=1}^{I} x_i^j}{I}$ (la moyenne de toute la colonne = la moyenne de la variable)

- sa variance : $s_j^2 = \frac{\sum_{i=1}^{I} (x_i^j - \overline{x}^j)^2}{I}$ (variance d'une colonne)

- son écart type : $s_j = \sqrt{s_j^2}$ (écart type d'une colonne)

→ on s'intéresse à un j quelconque et i varie ($j = cte$)

→ ensuite, on effectue la transformation : $x_i^j \longrightarrow z_i^j = \frac{1}{\sqrt{I}} \cdot \frac{x_i^j - \overline{x^j}}{s_j}$

c'àd que chaque variable on va la centrer et la réduire

→ cette partie : $\frac{x_i^j - \overline{x^j}}{s_j}$ c'est centrer et réduire, on ajoute $\frac{1}{\sqrt{I}}$ c'est une camstante qu'am verra sam intérêt par la suite

$\sqrt{I}$ = racine du nb d'individus.

→ donc finalement, puisque chaque x_i^j a été transformé en z_i^j dans le tableau $X = (x_i^j)$ il s'est transformé en un tableau $Z = (z_i^j)$

- l'individu x_i se transforme en un vecteur z_i ; $i = 1, 2 \dots I$. (ligne)
- la variable x^j se transforme en un vecteur z^j ; $j = 1, 2 \dots J$. (colonne)

→ Z a la m̂ dimension que X (I lignes et J colonnes)

l'intérêt du centrage et de la réduction :

on centre les données et on les réduits pour se débarasser des unités

- si on multiplie les données $* 100$ donc m̂ la moyenne sera $* 100$ et la variance sera multiplié $* 10000$ (au carré) par conséquent lorsqu'on prend la racine on revient à $* 100$ (l'écart-type)

$$x_i^j \longrightarrow z_i^j = \frac{1}{\sqrt{I}} \cdot \frac{x_i^j - \overline{x^j}}{s_j} \qquad \frac{(\text{la moyenne} \times 100)}{(\text{l'écart type} \times 100)}$$

→ donc ils se simplifient c à d que si on mesure en m ou en cm on va obtenir la m̂ valeur.

→ l'idée c'est de ramener les données à la m̂ unité qui est sans unité en fait.

→ les grandes données deviennent petites. Parceque toute les données multipliées fois 100 on va se débarasser de ce 100 ⇒ les données varieront à côté de 0.

exemple : on prend 3 données

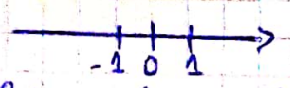
1	100
2	200
3	300

la moyenne = 2

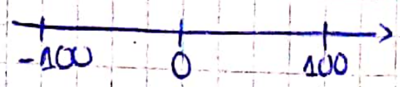
donc on enlève le 2 pour ces données

$1 - 2 = -1$
$2 - 2 = 0$
$3 - 2 = 1$

↑ les données centrées



moyenne = 200
centrer les données :
$100 - 200 = -100$
$200 - 200 = 0$
$300 - 200 = 100$



le 1 signifie qu'il a +1 que la moyenne et le -1 signifie -1 que la moyenne.

. on calcule la variance = la somme des carrées²

$$(-100-0)^2+(0-0)^2+(100-0)^2 = 20\,000 \Rightarrow \text{la variance} = \frac{20000}{2}$$

. la variance pour les autres données (1, 2, 3)

$$(-1-0)^2+(0-0)^2+(1-0)^2 = 2$$

$$\text{la variance} = \frac{2}{2} = 1 \Rightarrow \text{l'écart type} = 1$$

→ si on divise les données centrées par l'écart type:

$$\begin{array}{l} -1 \Rightarrow -1 \\ 0 \Rightarrow 0 \\ 1 \Rightarrow 1 \end{array}$$

les données centrées réduites

→ la variance de (100, 200, 300)

$$= \frac{20\,000}{2} = 10\,000 \Rightarrow \text{l'écart-type} = 100 \Rightarrow s = 100$$

⇒ réduire les données

$$-100 \Rightarrow \frac{-100}{100} = -1$$

$$0 \Rightarrow 0$$

$$100 \Rightarrow \frac{100}{100} = 1$$

⇒ c'est exactement comme les premiers (1, 2, 3)

⇒ donc si on multiplie les données par 100 ça va donner la m̂ chose.

⇒ toutes les données deviennent dans un intervalle bien défini ⇒ m̂ échelle de grandeur.

⇒ c'est ça l'utilité du centrage et de la réduction.

exemple :

→ pour plusieurs clients on a des variables :

clients	Revenu	Epargne
1	-	-
2	-	-
3		
⋮		

→ les valeurs de l'Epargne sont plus petites que les valeurs du Revenu.

→ lorsqu'on centre et on réduit chaque colonne on va avoir la m̂ échelle de valeurs

→ par exemple les valeurs du Revenu vont varier entre -2 et +2 et mêmes les valeurs de l'Epargne vont varier entre -2 et +2 ⇒ c'est pas les mêmes valeurs mais ils vont avoir la même échelle

pourquoi faire : c'est parce que après on va fabriquer des variables synthétiques (c'est une moyenne pondérée de chaque ligne) donc lorsque vous avez le Revenu est mille fois plus grand que l'Epargne ça fait des dispersions, mais lorsque vous avez des choses du même degré, même échelle de valeur vous faites par exemple $C_1 \times$ Revenu + $C_2 \times$ Epargne sur la même échelle de valeurs donc ils deviennent comparables. C'est pour cela qu'on réduit les données et on les centres.

Analyse des nuages des variables N(J). Effet de la transformation

→ la variable x^j se transforme en vecteur Z^j.

$$Z^j = \begin{pmatrix} Z_1^j \\ Z_2^j \\ \vdots \\ Z_I^j \end{pmatrix} \quad \text{avec } Z_i^j = \frac{1}{\sqrt{I}} \cdot \frac{x_i^j - \overline{x^j}}{s_j}$$

→ si on calcule la moyenne des Z_i^j : $\frac{\text{la somme}}{I}$ ⇒ $\boxed{\text{la moyenne} = 0}$

$$\text{la moyenne} = \frac{\sum_{i=1}^{I} Z_i^j}{I} = 0$$

$$= \frac{1}{\sqrt{I}} \cdot \frac{\sum_{i=1}^{I} x_i^j - \sum_{i=1}^{I} \overline{x^j}}{s_j} = \frac{1}{\sqrt{I}} \cdot \frac{I.\overline{x^j} - I.\overline{x^j}}{s_j} = 0$$

$s_j = cte$

→ la moyenne = 0 ⇒ la variable est centrée c.à.d que la moyenne devient = 0

→ et la variance ⇒ $\frac{\sum_{i=1}^{I} (Z_i^j)^2}{I} = S_{Z_j}^2 = \cancel{1}$

car $$\frac{\sum_{i=1}^{I} (Z_i^j)^2}{I} = \frac{1}{I} \sum_{i=1}^{I} \frac{(x_i^j - \overline{x^j})^2}{s_j^2} = \frac{s_j^2}{s_j^2} = \cancel{1}$$

$$\frac{\sum_{i=1}^{I} (x_i^j - \overline{x^j})^2}{I} = s_j^2$$

* on a : $$Z_i^j = \frac{1}{\sqrt{I}} \cdot \frac{x_i^j - \overline{x^j}}{s_j}$$

avec $$s_j^2 = \frac{1}{I} \cdot \sum_{i=1}^{I} (x_i^j - \overline{x^j})^2$$

→ la variance de Z^j

$$S_Z^2 = \frac{\sum_{i=1}^{I}(Z_i^j)^2}{I} = \frac{1}{I}\frac{1}{I}\sum_{i=1}^{I}\frac{(x_i^j - \overline{x^j})^2}{S_j^2}$$

parce que la moyenne $= 0$

$$= \frac{1}{I^2}\cdot\frac{1}{S_j^2}\cdot\sum_{i=1}^{I}(x_i^j - \overline{x^j})^2$$

$$= \frac{1}{I}$$

la norme de Z^j :

$$\|Z^j\|^2 = \sum_{i=1}^{I}(Z_i^j)^2 = 1$$

① les points variables (colonnes) se trouvent tous sur la sphère de centre O et de rayon 1 dans l'espace dual $\mathbb{R}^I$: $\|Z^j\| = 1$

→ sphère de dim I ⇒ hypersphère
→ si on est en 3 dim ⇒ sphère

② La corrélation entre deux variables j et j' :

$$r(j,j') = \sum_{i=1}^{I} Z_i^j Z_i^{j'} = \langle \underline{Z}^j, \underline{Z}_i^{j'} \rangle = \cos(\underline{Z}^j, \underline{Z}_i^{j'})$$

* la division sur ~~A_I~~ $\sqrt{I}$ ⇒ pourque la norme devient $= 1$

* et aussi :

la corrélation entre 2 variables X (note de maths) et Y (note de physique) et on a N étudiants. Chaque étudiant on a sa note de maths et de physique.

Z^1	Z^2
X	Y
x_1	y_1
x_2	y_2
⋮	⋮
x_N	y_N

⇒ la corrélation entre X et Y c'est la covariance :

$$\rightarrow corr(X,Y) = \frac{cov(X,Y)}{S_X S_Y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{S_X \cdot S_Y}$$

→ S_X : écart type de X ; S_Y : écart type de Y.

→ la corrélation est toujours entre -1 et +1. Ça mesure la liaison linéaire entre X et Y.

→ si on calcule la corrélation entre Z^1 et Z^2 on trouve :

$$\frac{\sum_i (Z_i^1 - 0)(Z_i^2 - 0)}{S_{Z^1}\cdot S_{Z^2}}$$

⇒ la moyenne de Z^1 et de $Z^2 = 0$ car ces 2 variables ont été centrées et réduites.

→ le produit scalaire entre Z^1 et Z^2.

$$\langle Z^1 . Z^2 \rangle = \sum_{i=1}^{I} Z_i^1 . Z_i^2$$

$$= \sum_{i=1}^{I} \left[\frac{(x_i^1 - \bar{x}^1)}{S_1} . \frac{1}{\sqrt{I}} . \frac{1}{\sqrt{I}} . \frac{(x_i^2 - \bar{x}^2)}{S_2} \right]$$

$$= \sum_{i=1}^{I} \frac{1}{I} . \frac{(x_i^1 - \bar{x}^1)(x_i^2 - \bar{x}^2)}{S_1 . S_2}$$ ⇒ c'est exactement la corrélation entre X et Y

parce que la covariance $= \frac{(x_i^1 - \bar{x}^1)(x_i^2 - \bar{x}^2)}{I}$

donc le produit scalaire devient la corrélation et la norme = 1.

$$\langle Z^1 , Z^2 \rangle = \underbrace{\| Z^1 \|}_{=1} . \underbrace{\| Z^2 \|}_{=1} . \cos(Z^1, Z^2) = \cos(Z^1, Z^2) = \cos(X, Y)$$

→ donc le produit scalaire c'est une corrélation

→ la corrélation entre 2 variables c'est le cosinus entre ces 2 vecteurs.

→ la norme de chaque vecteur = 1.

→ Toutes les variables se trouvent dans une sphère de norme 1

→ les angles entre les vecteurs (variables) c'est la corrélation.

- $\cos(Z^5, Z^6)$ = corrélation entre Z^5 et Z^6
- les variables Z^3 et Z^4 sont très corrélées parce que l'angle entre eux est très petit et donc le cosinus est grand.
- les variables Z^6 et Z^2 ($\cos(Z^6, Z^2) = -1$) sont très corrélées négativement. corrélation négative très forte.
  (par exemple la variable math et la variable sport)

* si on a 3 individus chaque variable a 3 dimensions (voir p 16)

→ ~~corr~~ 2 variables sont corrélées négativement ça veut dire que lorsque l'une augmente l'autre va diminuer.

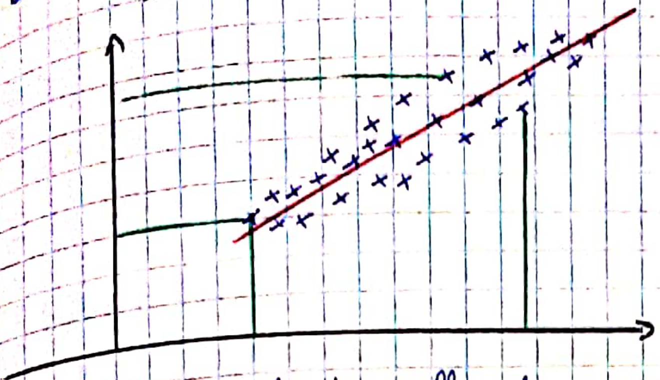
$\boxed{\text{dim } I}$ ⇒ on est dans un espace de dim I.

→ si le I est très grand par exemple 300 étudiants : on ne peut rien voir dans un espace de dim 300

⇒ on doit prendre toutes les variables et les projeter sur un plan, donc il faut chercher un bon plan pas n'importe quel plan : c'est le plan qui ne déforme pas beaucoup les points z^j (plan de projection)

→ plus le plan est étalé plus on a perdu bcp d'informations.

exemple :
dans un plan on a les points suivants :



→ on ne veut plus travailler dans un plan de 2 dim, on veut travailler sur 1 seule dimension. Càd on réduit les dimensions.

→ donc la meilleure droite sur laquelle on projeter c'est la droite en rouge.

⇒ Chaque point, on le projète perpendiculairement sur cette droite, la distance entre les points devient maintenue

⇒ Par contre si on projète sur l'axe x on aura une déformation

⇒ si on projète sur l'axe y c'est une déformation.

→ il faut chercher une droite tq en projetant dessus on perd le moins possible cad que les points restent les plus étalés possibles sur cette droite.

→ c'est la m̂ idée lorsqu'on parle de plan, on cherche un bon plan de projection.

p 18 :

→ on a des points : les z^j

→ z^1 se projète en t^1, z^2 se projète en t^2

→ les ~~plans~~ points étaient à l'intérieur d'une sphère donc après la projection ils deviennent à l'intérieur d'un cercle (cercle de corrélation)

→ si le point appartient à la surface de la sphère après la projection il va appartenir au périmètre du cercle.

→ les points projetés doivent être les plus étalés possibles.

→ si la projection des points se trouve en O ⇒ c'est très mauvais.

→ les points doivent être les plus dispersés possible et la distance % O doit être la plus grande possible pour garder le max d'information.

Comment projeter les points pour que l'étalement soit le plus grand possible?

→ il faut chercher un plan tel que en projetant les points, les projections doivent être les plus étalées possible.

→ les projections des z^j vont être dans un cercle.

p 19.

→ les points doivent être loin de O. S'ils sont sur le cercle c'est extraordinaire

→ l'étalement global des projections ⇒ la somme des distances des t^j % O,
cette quantité doit être la plus grande possible.

⇒ l'idée est de chercher un plan tq la somme des normes des t^j soit la plus grande possible.

définir une mesure d'étalement : $\sum_{j=1}^{J} d^2(O, t^j) = \sum_{j=1}^{J} \|t^j\|^2$ Inertie du nuage J sur le Plan E

doit être la plus grande possible.

→ c'est la dispersion des points variables autour de O dans la direction de E qui est appelé inertie et noté Im^E.

plan de projection

Propriété d'additivité de l'inertie (Pythagore).

si E et F sont deux sous espaces orthogonaux engendrant le sous espace E+F on

$$Im^{E+F} = Im^E + Im^F$$

- on projete z^j sur le plan, la projection de z^j sur le plan est y^j.
- $\|y^j\|^2 = \|t^j\|^2 + \|w^j\|^2$ pythagore
- ça veut dire que l'inertie des z^j on peut la calculer % à des droites orthogonales

on cherche le meilleur vecteur de projection, on calcule l'inertie sur cette droite, ensuite on cherche un autre vecteur orthogonale sur le premier et on calcule l'inertie sur ce vecteur. Donc au lieu de chercher le meilleur plan de projection, on cherche la meilleure droite de projection et on projete sur cette droite, ensuite on cherche une droite perpendiculaire sur la 1ère droite, on projete, ensuite on cherche une $3^{ème}$ droite perpendiculaire sur les 2 premières droites et ainsi de suite. Parce que l'inertie s'additionne sur cettes droites là. ⚠

→ Finalement on a ramené la projection sur un plan ou sur un espace quelconque à la projection sur des vecteurs unitaires. À condition que les vecteurs unitaires que je choisisse soient orthogonaux les uns sur les autres.

⇒ l'inertie globale = la somme des inerties.

* on doit d'abord chercher le premier vecteur sur lequel on va projeter par

∘ l'inertie le long d'un axe dirigé par un vecteur unitaire $\underline{v}$ a une expression matricielle simple: on cherche v:

$$\|t^j\|^2 = (\underline{z}^{j'}\underline{v})^2 = (\underline{v}'\underline{z}^j)(\underline{z}^{j'}\underline{v}) = \underline{v}'(\underline{z}^j\underline{z}^{j'})\underline{v}$$

$\underline{v}$: le premier vecteur sur lequel on va projeter

t^j : projection de z^j sur v

preuve $t^j = \langle z^j, v\rangle$ produit scalaire (nb réel)
$= (z^j)^t . v$

v' ⇒ transposé ⇓ v^t

$\overline{\overline{z}}$ $t^j = \langle z^j, v\rangle$ vecteur unitaire sur lequel on va projeter z^j

donc $\|t^j\|^2 = \left((z^j)^t . v\right)^2 = \left((z^j)^t . v\right)\left((z^j)^t . v\right) =$

on sait que $\langle a, b\rangle = \langle b, a\rangle = \sum a_i b_i$

↳ $\|t^j\|^2 = (v^t . z^j)\left((z^j)^t . v\right) = v^t . (z^j . (z^j)^t) . v$

donc $$\sum_{j=1}^{J} \|t^j\|^2 = \sum_{j=1}^{J} \underline{v}'(\underline{z}^j\underline{z}^{j'})\underline{v} = \underline{v}'\left(\sum_{j=1}^{J} \underline{z}^j . \underline{z}^{j'}\right)\underline{v}$$ Inertie sur v

or $$\sum_{j=1}^{J} \underline{z}^j\underline{z}^{j'} = \left(\underline{z}^1 \dots \underline{z}^j \dots \underline{z}^J\right)\begin{pmatrix} \underline{z}^1 \\ \vdots \\ \underline{z}^j \\ \vdots \\ \underline{z}^J \end{pmatrix} = \underline{Z}\underline{Z}'$$

$$Z = \begin{pmatrix} | & & | & & | \\ z^1 & \dots & z^j & \dots & z^J \\ | & & | & & | \end{pmatrix}$$

donc $\underbrace{Im(v)}_{\text{l'inertie sur le vecteur } v} = v^t (Z Z^t) v$

$$Z^t = \begin{pmatrix} \text{---} (z^1)^t \text{---} \\ \text{---} (z^2)^t \text{---} \\ \vdots \\ \text{---} (z^J)^t \text{---} \end{pmatrix}$$

$$Im^v = \underline{v}'\underline{Z}\,\underline{Z}'\underline{v}$$ Inertie sur v

$\Gamma = \underline{Z}\underline{Z}'$: matrice d'inertie du nuage des variables elle est symétrique définie positive. donc $Im^v = \underline{v}' . \Gamma . \underline{v}$

• Γ est symétrique car si on la transpose on va trouver la m̂ chose.

• Γ est positive : càd que tous ses valeurs propres sont positives ou nulles. (conséquence

↳ comment le montrer ?

dimension de Γ est : ~~(I,I)~~ $I \times I$ (I, I) matrice carrée

" Z est : (I, J) avec I c'est le nb d'individus

" Z' est : (J, I)

Γ : Symétrique $\Rightarrow$ $\Gamma = Z . Z^t$

$\Gamma^t = (Z . Z^t)^t = (Z^t)^t . Z^t = Z Z^t = \Gamma$

donc $\Gamma^t = \Gamma$ $\Rightarrow$ Γ est symétrique

Γ : positive. une matrice A est définie positive ssi $\boxed{x^t A x \geqslant 0}$

on prend x de dim (I, I) : $x^t \Gamma x = x^t Z Z^t x = x^t Z^t Z x$

$= (Z x)^t \underbrace{Z x}_{a} = \underbrace{a^t a}$ $\geqslant 0$

donc $x^t \Gamma x \geqslant 0$ et donc Γ est positive.

c'est une norme au carré

→ on cherche v qui rend $I n^v$ la plus grande possible. càd l'étalement le long de v est le plus grand possible

→ Z : connue, v : inconnu

chercher v : $\|v\| = 1$ (vecteur unitaire)

et $v^t \Gamma v$ maximal

$$\boxed{v^* = \underset{\substack{v \\ \|v\|=1}}{\arg\max}\; v^t \Gamma v}$$

le problème mathématique
on cherche le v qui maximise cette qté

$$\max \left(\left(\frac{v}{\|v\|}\right)^t \Gamma \left(\frac{v}{\|v\|}\right) \right) = \max \left(\frac{v^t \Gamma v}{\|v\|^2} \right) = \max_{v} \left(\frac{v^t \Gamma v}{v^t v} \right)$$

on cherche v tq $\dfrac{v^t \Gamma v}{v^t v}$ maximale ⚠ $\Gamma = Z Z^t$ connue

on cherche la solution du problème : x : tq : $\dfrac{x^t A x}{x^t x}$ maximale

$\Rightarrow$ on va trouver que c'est le vecteur propre de A qui est associé à la plus grande valeur propre

v_1 = le vecteur propre de Γ associé à la plus grande valeur propre

→ on va montrer que :

* le vecteur (meilleur axe d'inertie) recherché est le vecteur propre v_1 associé à la plus grande valeur propre λ_1 de la matrice d'inertie $\Gamma = ZZ'$

→ si on a une matrice symétrique donc tous ses vecteurs propres sont orthogonaux.

→ une matrice symétrique est diagonalisable dans une base orthonormée

* donc le $2^{ème}$ axe d'inertie est le vecteur propre v_2 associé à la $2^{ème}$ plus grande valeur λ_2 de la matrice $\Gamma = ZZ'$

* etc

Le meilleur plan d'étalement du nuage est généré par les 2 premiers vecteurs propres de ~~l'association~~ l'associés aux 2 plus grandes valeurs propres.

Résumé : pour trouver le meilleur plan de projection :

→ on prend la matrice X : on la centre et on la réduit : $X_i^{j*} = \frac{x_i^j - \bar{x}^j}{s_j}$

① calculer X^* : matrice centrée réduite

② calculer $\Gamma = \frac{1}{I} X^* X^{*t} = ZZ'$ $\qquad \boxed{Z = \frac{X^*}{\sqrt{I}}}$

③ rechercher les premiers vecteurs et valeurs propres de Γ ce sont les axes d'inertie.
→ associés aux plus grandes valeurs propres.
→ sur lesquels on va projeter

$X \rightarrow Z \rightarrow \Gamma = ZZ^t \rightarrow$ diagonaliser Γ

$\lambda_1 \geq \lambda_2 \dots \geq \lambda_I$ (valeurs propres)

↓

vecteurs propres : $v_1, v_2 \dots v_I$

→ le meilleur plan (v_1, v_2)

→ dim 3 → (v_1, v_2, v_3)

→ il y a des algos qui permettent de trouver v_1 et v_2 sans diagonaliser la matrice.

→ car diagonaliser → trouver I vecteurs propres (si I est grand donc on aura besoin de bcp de temps)

→ les éléments de la matrice Z : $Z_i^j = \frac{x_i^j - \bar{x}^j}{s_j}$

→ chaque vecteur propre a une valeur propre

→ on prend le vecteur propre associé à la valeur propre la plus grande. Ensuite on prend le vecteur propre qui lui est orthogonale qui est associé à la 2ème plus grande valeur propre ⇒ pour trouver le plan de projection

* les variables sont sur une sphère de centre G. si on était en 3 dim, chaque variable aura 3 individus : individu 1, individu 2, individu 3.

* on cherche le meilleur plan sur lequel les variables seront les plus étalées possibles

les variables : $z^1 \quad z^2 \quad \dots \quad z^J$

→ le plan de projection est constitué de (v_1, v_2)

→ v_1 : vecteur propre associé à la plus grande valeur propre λ_1

→ v_2 : vecteur propre associé à la 2ème plus grande valeur propre λ_2

→ mnt on doit projeter les variables z^j sur v_1

↳ on projète z^1 sur v_1, z^2 sur v_1 ..., z^J sur v_1

on note $G_\alpha(j) = (z^j)^t . v_\alpha$: projection de z^j sur v_α
= coordonnée de j sur l'axe α

$\alpha = 1 \dots I \mid j = 1 \dots J$

car on a I vecteurs propres et I valeurs propres.

$$\underline{G}_\alpha = \begin{pmatrix} G_\alpha(1) \\ G_\alpha(2) \\ \vdots \\ G_\alpha(J) \end{pmatrix} = \begin{pmatrix} \underline{z}^{1'} \\ \vdots \\ \underline{z}^{J'} \end{pmatrix} \underline{v}_\alpha = z' \underline{v}_\alpha \quad \Rightarrow \quad \underline{G}_\alpha = \underline{Z}' . \underline{v}_\alpha$$

$$G_\alpha(j) = G_\alpha(z^j)$$

projection de toute les variables sur l'axe α.

→ si on veut avoir les coordonnées de toutes les vari projetées sur v_α on :

$\underline{G}_\alpha = \underline{Z}^t . \underline{v}_\alpha$

→ on peut donc avoir les coordonnées de toutes les var sur v_1 et sur v_2

p33

on obtient qlq chose comme ça (voir).

→ on a toutes les variables des régions françaises projetées sur le meilleure plan de projection par ex : % des ménages propriétaires, le revenu par tête d'habitant, % de personnes seuls, formation continue ....

→ projection sur le plan (v_1, v_2)

chaque variable a 2 composantes (G_1, G_2)

→ plus les variables sont petites et proches du cercle ⇒ donc lorsqu'on a projeté on a rien perdu (sur le cercle les variables sont bien projetées) plus on s'éloigne du cercle plus c'est mieux.

les variables : revenu par tête d'habitant, PIB/habitant sont corrélées.

→ plus les gens gagnent d'argent plus le revenu par habitant augmente

=> les variables à droite caractérisent la richesse => corrélées positivement

→ les variables à gauche caractérisent la pauvreté : (% de ménages RMIS :

→ qui prennent des allocations familiales, taux de chômage, % de ménages

> 5 personnes → plus il y a de gens → plus on est pauvre → revenu par tête d'hab diminue, % sans diplôme) ces var sont corrélées positivement entre eux.

→ les variables à gauche sont corrélées négativement avec les var à droite.

→ lorsqu'on projette on perd de l'information. Pour savoir combien on a perdu de l'information, c'est avec les valeurs propres.

* on suppose qu'on a 25 variables : $J = 25$

* et on a 22 régions : $I = 22$

* la matrice Γ est de dim $I \times I = 22 \times 22$

donc on a 22 valeurs propres

λ_1 = première valeur propre

λ_2

⋮

→ sur l'axe (λ_1) on a le plus grand étalement

le pourcentage du 1er axe : $\dfrac{\lambda_1}{\underbrace{\lambda_1 + \dots + \lambda_{22}}_{\text{trace}}} \times 100$ pourcentage d'inertie expliqué par le 1er axe

→ chaque axe explique un pourcentage

ex : 1er axe valeur propre = 16,7 % d'inertie = 52 %

le premier axe explique 52% d'information

~~plan~~ $\cancel{(22 \ldots)}$ valeur propre $\lambda_2 = 4,68$ % d'inertie = 14 %

2ème axe

donc le plan (v_1, v_2) explique 66% d'information = premier plan

de la matrice de données qu'on avait au début

→ on a perdu 34% de l'information

→ si on projète sur un espace de dim 3 % d'inertie cumulé = 77%

→ espace de dim 5 % d'inertie cumulé = 91,20%

$\frac{5}{22}$ explique 91,20% de l'information.
de var

→ c'est l'ACP de nuage des variables.

→ par la suite il faudra faire l'ACP des nuages des individus.

* on s'interesse au vecteur des lignes.
* chaque individu Z_i du nuage $N(I)$ est une ligne du tableau de données. c'est un vecteur de $\mathbb{R}^d$.
* donc au lieu de refaire toute l'analyse, il suffit de transposer la matrice Z, transposer le tableau de données et refaire l'analyse dans Z' au lieu de Z.

→ les variables vont jouer le rôle des individus.

→ les individus vont jouer le rôle des variables.

$V = Z'Z$, matrice d'inertie du nuage.

↳ qu'il faut la diagonaliser pour trouver les valeurs propres.

Télécharger Tanagra : logiciel de data mining (voir la documentation)

↳ ACP, clustering, K-means, ...

TD 1

exercice 2 : arbres de décisions :

Rappel : Pour un noeud t de l'arbre contenant les k probabilités (fréquences) d'un label de K classes t :

p_1	p_2	-----	p_k

$$\text{Entropie}(t) = -\sum_{l=1}^{k} p_l * \log(p_l)$$ le logarithme est calculé en base 2

$$\text{Erreur}(t) = 1 - \max(p_1, p_2, \ldots, p_k)$$

$$\text{Gini}(t) = 1 - \sum_{l=1}^{k} p_l^2$$

A remarquer qu'après le partage d'un noeud t en deux noeuds binaires gauche : tg et droite : td, la mesure d'un indice après ce partage par exemple pour Gini (après partage) = fg * (Gini(tg)) + fd * (Gini(td)) où fg et fd représentent les poids (en termes de fréquences d'éléments de la partie gauche et la partie droite) respectivement.

exo : on considère le fichier d'apprentissage du tableau suivant pour une classification binaire :

Instance	a_1	a_2	a_3	La classe
1	T	T	1	+
2	T	T	6	+
3	T	F	5	-
4	F	F	4	+
5	F	T	7	-
6	F	T	3	-
7	F	F	8	-
8	T	F	7	+
9	F	T	5	-

T : true

F : false

→ le gain est calculé à partir de l'entropie ou à partir de Gini.

→ 3 mesures pour construire un arbre de décision : gini, entropie, erreur.

setosa virginica

noeud t : 1/3 | 1/3 | 1/3 versicolor

$$Gini(t) = 1 - \left(\frac{1}{3}\right)^2 \times 3 = 1 - \frac{1}{3} = \frac{2}{3}$$

$$Entropie(t) = -\frac{1}{3}\log\frac{1}{3} - \frac{1}{3}\log\frac{1}{3} - \frac{1}{3}\log\frac{1}{3} = -\log\frac{1}{3} = \log 3$$

$$Erreur(t) = 1 - \frac{1}{3} = \frac{2}{3}$$

↳ càd que chaque noeud on l'affecte à la classe majoritaire, donc c'est l'une parmi les 3, donc on se trompe sur les 2 autres.

→ on se trompe 2/3 de fois.

① Ecrire les formules de l'entropie, de l'erreur et de l'indice de gini en fonction de K dans le cas de l'équiprobabilité.

‖ même probabilités

• équiprobabilité ⇒ $\boxed{P_i = \frac{1}{K}}$ $1 \leq i \leq K$

donc :

• $Gini(t) = 1 - \frac{1}{K^2} \times K = 1 - \frac{1}{K}$

• $Entropie(t) = - \sum_{j=1}^{K} \frac{1}{K} \log_2\left(\frac{1}{K}\right) = \log_2(K)$

• $Erreur(t) = 1 - \frac{1}{K}$

$K = 2$. malade _ non malade   Entropie $(t) = \log_2(2) = 1$

→ càd que pour envoyer un message de longueur 2 c'est 1 bit.

→ entropie : donne le nb moyen de bit qu'il faut pour envoyer des messages qui ont pour chacun une probabilité P_j

Equiprobabilité :
$$\text{Gini} = 1 - \frac{1}{K}$$
$$\text{Entropie} = \log_2(K)$$
$$\text{Erreur} = 1 - \frac{1}{K}$$

② Ecrire les formules de l'entropie, de l'erreur et de l'indice de Gini pour le cas où le label est binaire (deux classes) $(p, 1-p)$. Tracer les 3 formules trouvées sur un seul graphique en fonction de p (commentaire).

$K = 2$

$+$ p | $1-p$ $-$

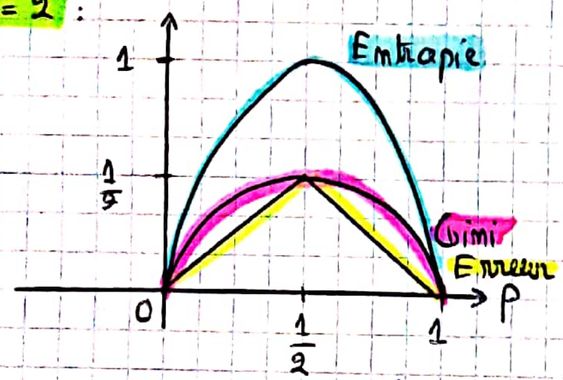
→ $\text{Gini} = 1 - p^2 - (1-p)^2 = 1 - p^2 - 1 - p^2 + 2p = 2p - 2p^2$

$\text{Gini} = 2p(1-p)$

→ $\text{Entropie} = -p\log p - (1-p)\log(1-p)$

→ $\text{Erreur} = 1 - \max(p, 1-p)$

cas $K = 2$:

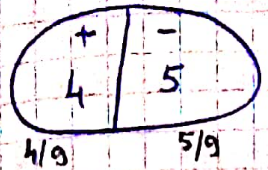


→ lorsque $p = 0$ ou $p = 1$ l'impureté est nulle, les 3 mesures s'annullent.

→ les 3 mesures prennent leur maximum dans $1/2$.

pour mesurer l'impureté dans un noeud.

③ quelle est l'entropie de cette collection d'instances % à la classe positive.

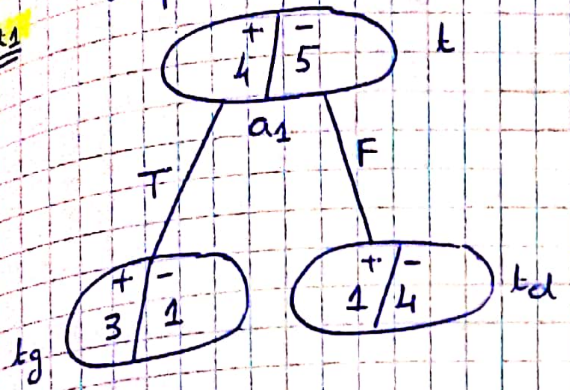


$$\text{Entropie}(t) = -\left(\frac{4}{9}\log\frac{4}{9} + \frac{5}{9}\log\frac{5}{9}\right) = 0,9911$$

b) quels sont les gains d'information « information gain » correspondants à a_1 et a_2 respectivement.

on split

a_1



$$\text{Entropie}(t_g) = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4}$$

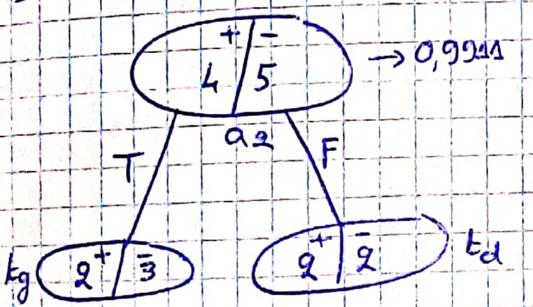
$$\text{Entropie}(t_d) = -\frac{1}{5}\log\frac{1}{5} - \frac{4}{5}\log\frac{4}{5}$$

$$\text{Entropie}(\text{niveau } 1) = \frac{4}{9}\,\text{Entropie}(t_g) + \frac{5}{9}\,\text{Entropie}(t_d) = 0,7616$$

$\Rightarrow$ gain $= 0,9911 - 0,7616 = 0,2295$

entropie diminue lorsqu'on a splité sur a_1.

a_2 on split sur a_2



$$\text{Entropie} = \frac{5}{9}\,\text{Entropie}(t_g) + \frac{4}{9}\,\text{Entropie}(t_d) = 0,9838$$

gain $= 0,9911 - 0,9838 = 0,073$ très petit

$\rightarrow$ donc sur a_1 on gagne plus

gain$(a_1) = 0,2295$

gain$(a_2) = 0,073$

→ l'attribut qui permet de gagner plus en pureté c'est a_1.

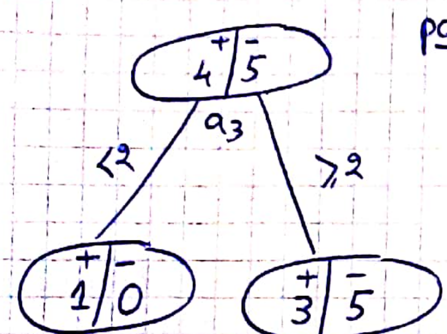
⑤ Pour a_3 qui est un attribut continu, calculer le gain d'information pour chaque partage possible. Un partage sur une valeur de a_3 correspond à la valeur moyenne de 2 valeurs consécutives de a_3 trié.

moyenne — entropie lorsqu'on partage par 2

a_3

a_3	label	split	Entropie	Gain
1	+	2	0,8484	0,1427
3	-	3,5		
4	+	4,5	0,9834	0,0072
5	-	5,5		
5	-			
6	+	6,5		
7	+	7,5		
7	-			
8	-	—		

← le plus grand gain c'est lorsqu'on partage par 2 donc c'est celui-là qu'on va utiliser



partage par 2

on calcul l'entropie droite et gauche

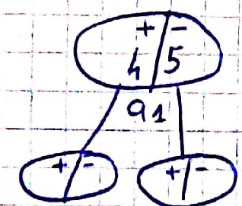
ensuite l'entropie du partage

gain (a_1) = 0,2295

gain (a_2) = 0,073

gain (a_3) = 0,1427 gain du meilleur partage de a_3

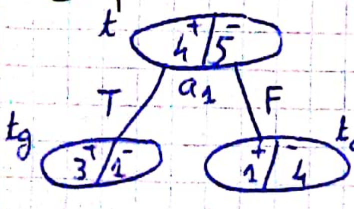
⇒ le meilleur attribut = celui qui a donné le meilleur gain = a_1



donc on split sur a_1 au début

pour les autres nœuds on doit recalculer les entropies

⑥ quel est le meilleur partage parmi a_1 et a_2 selon le taux d'erreur.



• on affecte à la classe majoritaire, donc on se trompe 1/4

Erreur (t_g) = $\frac{1}{4}$ (classe majoritaire c'est le +)

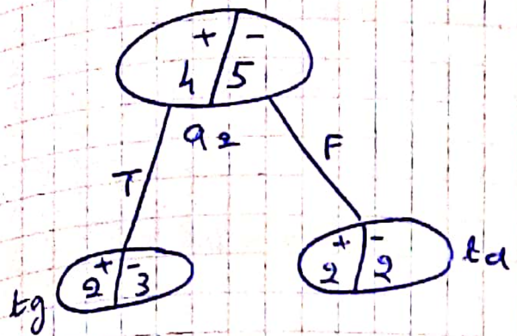
Erreur (t_d) = $\frac{1}{5}$ (classe majoritaire c'est -)

$Erreur(t) = \frac{4}{9}$

$$Erreur(\underset{a_1}{niveau}) = \frac{4}{9} \times \frac{1}{4} + \frac{5}{9} \times \frac{1}{5} = \frac{2}{9}$$

$\underline{\underline{a_2}}$ $Erreur(a_2) = \frac{4}{9}$

→ donc c'est toujours a_1 qui est le meilleur.



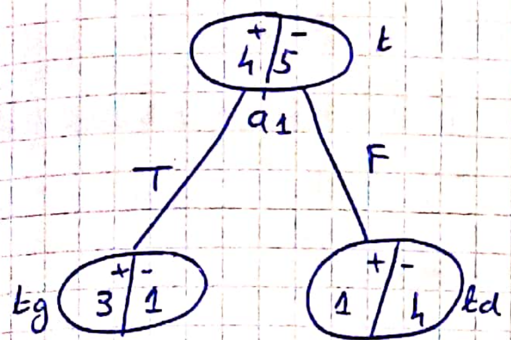
$Erreur(t_g) = \frac{2}{5}$

$Erreur(t_d) = \frac{1}{2}$

$Erreur(a_2) = \frac{4}{9} \times \frac{2}{5} + \frac{5}{9} \times \frac{1}{2} = ?$ vérifier

(1) Quel est le meilleur partage parmi a_1 et a_2 selon l'indice de gini.

$\underline{a_1}$



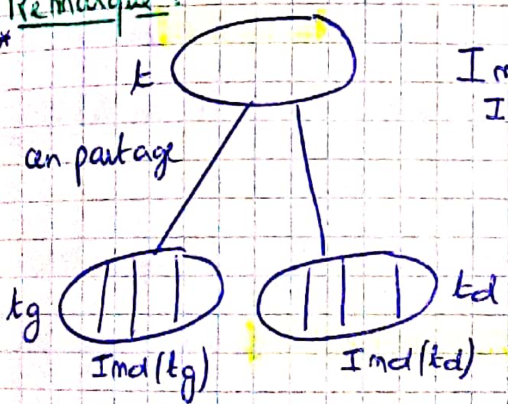
$Gini(t) = 1 - \left(\frac{4}{9}\right)^2 - \left(\frac{5}{9}\right)^2$

$Gini(t_g) = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2$

$Gini(t_d) = 1 - \left(\frac{1}{5}\right)^2 - \left(\frac{4}{5}\right)^2$

$$Gini(a_1) = \frac{4}{9} Gini(t_g) + \frac{5}{9} Gini(t_d)$$

* Remarque :



Indice / Ind(t) <
- gini
- erreur
- entropie

$$Indice(niveau\ 1) = \underset{\text{fréquence gauche}}{f_g}\ Indice(t_g) + f_d\ Indice(t_d)$$

NB

gain = Indice(t) − Indice(niveau 1) ⩾ 0 ⚠

$\cancel{f(\lambda t_1) + f(\lambda t_2) \geq f(\lambda t_1 + \lambda t_2)}$ convexité

$f[(\lambda t_1) + (1-\lambda) t_2] \leq \lambda f(t_1) + (1-\lambda) f(t_2)$ convexité

séance ⑧ : 24 - 05 - 2021

→ on tortue les données pour en tirer les plus importants.

→ on garde le signal et on laisse tomber tout ce qui est bruit.

→ le point départ c'est un tableau (p2) : on a des lignes : I lignes, si on veut les représenter, ils seront dans un espace de dim J. par exemple chaque ligne représente les ~~8~~ notes d'un étudiant, cad que chaque étudiant est dans l'espace des notes et chaque matière est dans l'espace des individus (I individus).

→ **l'Analyse duale** : analyse des variables dans l'espace des individus.

- chaque variable est dans un espace R^I

→ **l'Analyse directe** : analyse des individus dans l'espace des variables.

* la dernière fois on a fait : Analyse duale individus variables

- on est parti d'un tableau X de dim $I \times J$, c'est un tableau (x_i^j)
- on crée un tableau X^* : $X \longrightarrow X^*$ de dim $I \times J$

$$x_i^{*j} = \frac{x_i^j - \bar{x}^j}{s_j} \quad \text{centrage et réduction}$$

pour avoir des variables normées : $X^* \longrightarrow Z = \frac{1}{\sqrt{n}} X^*$ avec $Z_i^j = \frac{x_i^{*j}}{\sqrt{n}}$

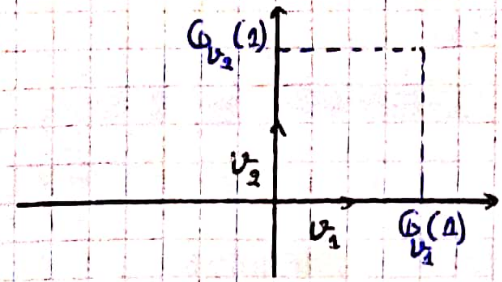
Z est un tableau de dim $I \times J$.

- à partir de Z on fabrique la matrice d'inertie $\Gamma = Z Z^t$ de dim $I \times I$
- avec la matrice Γ on va trouver les vecteurs propres, ce sont les vecteurs sur lesquels on va projeter les variables.
- on cherche les valeurs propres $\lambda_1 \geq \lambda_2 \geq \lambda_3 \dots \geq \lambda_r$ et les vecteurs propres associés $V_1, V_2, \dots, V_r$ avec $r = rang(Z)$ on ne peut pas aller plus que le rang car après le rang les valeurs propres sont nulles

$$r = rang(Z) \leq \min(I, J)$$

- une fois on a les vecteurs propres, on commence à projeter les variables.
- par exemple on projete la variable 1 sur V_1. $G_{v_1}(1) = (\underline{Z}^1)^t . v_1$

$G_{v_2}(1) = (Z^1)^t . v_2$. Donc on a 2 coordonnées de la variable Z^1, coordonnées sur v_1 et coordonnées sur v_2. Donc on peut représenter la variable Z^1 sur le plan (v_1, v_2)



→ on fait la m̂ chose pour toutes les variables

→ les angles entre les variables représentent les corrélations.

→ et donc on peut représenter toutes les variables sur le plan (v_1, v_2) et c'est le meilleur plan de projection, càd qu'en projetant les variables sur ce plan on perd le moins d'informations possible, on compresse les informations qu'on avait dans un espace de dim I dans un espace de dim 2 et (v_1, v_2) c'est le meilleur espace pour ceci.

→ ensuite on a toutes les var représentées sur un espace de dim 2 au lieu qu'ils soient représentées dans un espace de dim I qui peut être assez grand (6M clients, 900 étudiants ...)

→ les projections des variables se trouvent dans un cercle unitaire ⚠

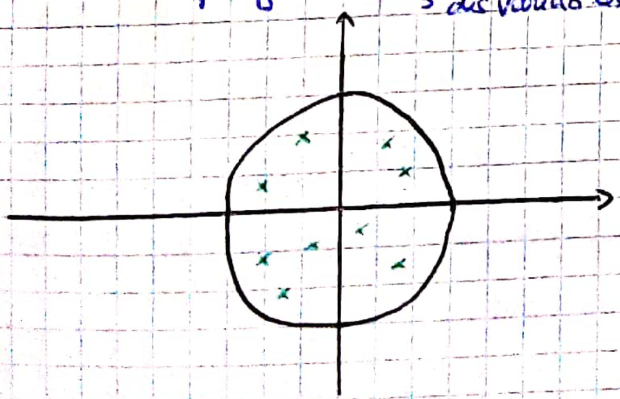
$G_{v_\alpha}(j)$: projection de Z^j sur v_α

$$= (Z^j)^t . v_\alpha = \frac{(Z^j)^t . v_\alpha}{\|Z^j\| . \|v_\alpha\|}$$ car $\|Z^j\| = \|v_\alpha\| = 1$ vecteurs unitaires

$$G_{v_\alpha}(j) = \cos$$

donc $-1 \le G_{v_\alpha}(j) \le 1$

donc toutes projections des variables seront à l'intérieur d'un cercle unitaire.



22 régions : lignes

variables économiques : colonnes

→ ona la projection de toutes les variables sur le plan (v_1, v_2).

→ on remarque que le revenu par habitant est très corrélé avec PIB/habitant, car l'angle entre eux est très petit = corrélation.

→ la variable parc auto / habitant est corrélé avec la mortalité, plus il y a des automobiles plus il y a des accidents.

→ plus il y a des automobiles plus le % personne active dans l'industrie est grand, donc c'est corrélé.

→ Tout ce qui se trouve proche du cercle, lorsqu'on la projeter, il était pratiquement déjà proche du cercle, on a rien perdu lorsqu'on les a projeter.

→ on était dans un espace de dim 22 et mnt on est dans un espace de dim 2.

→ sur espace de dim 2 : l'inertie cumulé = 66,74 % donc on a perdu 33,26 % de l'information, càd les 2 premières valeurs propres représentent 66,74 % de l'information. On peut aller jusqu'à 22 (dim 22)

→ lorsque les variables sont proche du cercle après la projections ils sont tjrs proches du cercle.

Analyse du nuage des individus $N(I)$:

- chaque individu est un vecteur $Z_i \in \mathbb{R}^J$, donc il peut être représenté dans un espace de dim J avec J le nb de variables.
- au lieu de refaire tout dès le début, il suffit de transposer la matrice Z, les individues deviennent les variables et les variables deviennent les individus
- la matrice d'inertie : $V = Z'Z$ de dim $J \times J$
- il faut diagonaliser V pour trouver ses vecteurs propres et c'est sur ces vecteurs propres qu'on va projeter les individus.

Résultat :

$\Gamma = ZZ'$ et $V = Z'Z$. possèdent les m̂ valeurs propres non nulles

matrice d'inertie des variables — matrice d'inertie des individus

* les valeurs propres nulles de Γ sont aussi dans V.

→ les vecteurs propres de V : $u_1, u_2 \dots \dots$ et les vecteurs propres de Γ : $v_1, v_2 \dots$

sont reliés par les formule :

$$u_\alpha = \frac{Z^t . v_\alpha}{\sqrt{\lambda_\alpha}} \quad \text{et} \quad v_\alpha = \frac{Z . u_\alpha}{\sqrt{\lambda_\alpha}}$$

→ on peut diagonaliser soit Γ soit V et déduire les vecteurs propres de l'autre et les valeurs propres sont les m̂.

⇒ $\Gamma = ZZ^t$ de dim $I \times I$ — Z de dim $I \times J$

$V = Z^t Z$ de dim $J \times J$

⇒ si on a plus de variables que d'individus on diagonalise Γ.

⇒ si on a plus d'individus que de variables il vaut mieux diagonaliser V.

En général : le nb d'individus dépasse le nb de variables.

* 300 étudiant et 15 matières

* des millions de malades sur lesquels on mesure des tests médicaux.

⇒ En général on diagonalise V.

preuve :

$\Gamma = ZZ^t$ et $V = Z^t Z$

on suppose qu'on a diagonaliser Γ donc on connaît les v_α

donc $\Gamma . v_\alpha = \lambda_\alpha v_\alpha$ avec λ_α valeur propre associé à v_α.

donc $ZZ^t = \Gamma$ ⇒ $ZZ^t v_\alpha = \lambda_\alpha v_\alpha$

⇒ $Z^t (ZZ^t v_\alpha) = Z^t (\lambda_\alpha v_\alpha)$

⇒ $\underbrace{Z^t Z}Z^t v_\alpha = Z^t \lambda_\alpha v_\alpha$

⇒ $V . \underbrace{Z^t v_\alpha} = \lambda_\alpha \underbrace{Z^t v_\alpha}$

donc le vecteur $Z^t v_\alpha$ est un vecteur propre de V donc si v_α est un vecteur propre de Γ alors $Z^t v_\alpha$ est un vecteur propre de V + ils ont la m̂ valeur propre.

→ V a des vecteurs propres car elle aussi est symétrique définie positive. Donc elle est diagonalisable dans un espace orthonormé.

→ Donc il suffit de normer le vecteur $Z^t v_\alpha$.

→ on suppose que $u_\alpha = \frac{Z^t v_\alpha}{\|Z^t . v_\alpha\|}$ pour que le vecteur soit unitaire

$$\|Z^t . v_\alpha\|^2 = v_\alpha^t Z Z^t v_\alpha = v_\alpha^t . \Gamma . v_\alpha = v_\alpha^t \Gamma v_\alpha = v_\alpha^t \lambda_\alpha v_\alpha = \lambda_\alpha \underbrace{v_\alpha^t v_\alpha}_{=1}$$

$$\Rightarrow \|Z^t . v_\alpha\|^2 = \lambda_\alpha$$

donc $u_\alpha = \frac{Z^t . v_\alpha}{\sqrt{\lambda_\alpha}}$ d'où le résultat.

on peut faire la m̂ démarche pour trouver que $v_\alpha = \frac{Z . u_\alpha}{\sqrt{\lambda_\alpha}}$

→ on ne cherche que le vecteur de Γ ou de V ça dépend de la dimension.

→ les valeurs propres sont les m̂, donc on n'a pas besoin de diagonaliser l'autre matrice

→ et les vecteurs propres se déduisent les uns des autres.

* mnt qu'on a les vecteurs u_α on va projeter les individus sur ce vecteur. Parce que le meilleur plan de projection c'est :

→ le vecteur (meilleur axe d'inertie) recherché est le vecteur propre u_1 associé à la plus grande valeur propre λ_1 de la matrice d'inertie $V = Z^t Z$

→ le 2ème axe d'inertie est le vecteur propre u_2 associé à la 2ème grande valeur λ_2 de la matrice d'inertie $V = Z^t Z$

→ etc

⇒ le meilleur plan d'étalement du nuage N(I) est généré par les deux premiers vecteurs propres de V associés aux deux plus grandes valeurs propres.

$V = Z^t Z$. $J \times J$: $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_n$

$$\underbrace{u_1 , u_2}_{\text{meilleur plan de projection}} \ldots \quad u_n$$

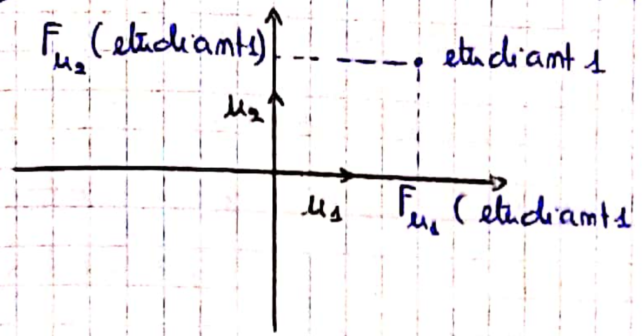
→ si on veut par exemple projeter un étudiant sur u_1 :

F_{u_1} (etudiant 1) = $(\text{etudiant } 1)^t \cdot u_1$

F_{u_2} (etudiant ~~2~~) = $(\text{etudiant } 1)^t \cdot u_2$

donc on a 2 coordonnées de l'étudiant 1 . donc on va le présenter

par ex notes de math et physique



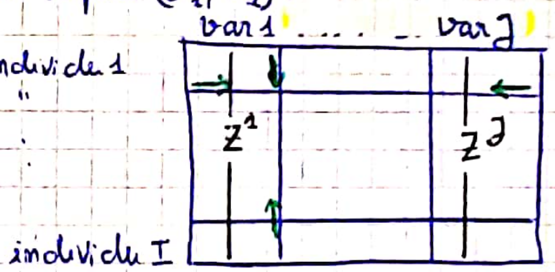
→ on peut mnt représenter tous les étudiants dans un plan.

→ puisque la norme de l'étudiant 1 ≠ 1 donc la valeur F_{u_1} (etudiant) n'est pas toujours à l'intérieur d'un cercle

⇒ on peut donc représenter à la fois les étudiants dans un espace (u_1, u_2) et on peut aussi représenter les variables dans un espace (v_1, v_2).

Résumé :

après transformation : on a la matrice Z =



on projete la Var 1 sur v_1 et sur v_2

↳ $G_{v_1}(1)$ et $G_{v_2}(1)$

→ on a projeter toutes les variables sur v_1

$$G_\alpha(j) = (Z^j)^t \cdot v_\alpha$$

$G_{v_1}(1)$ --- $G_{v_1}(2)$ --- $G_{v_1}(J)$

$G_{v_2}(1)$ $G_{v_2}(2)$ --- $G_{v_2}(J)$

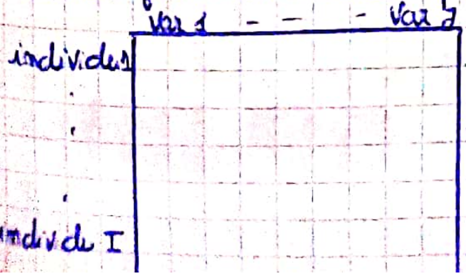
→ : compresser.

(v_1, v_2) : meilleur plan

→ donc on a les coordonnées de toutes les variables sur l'axe 1 et sur l'axe 2

donc on peut les représenter sur un plan

→ on fait la m̂ chose pour les individus



— $F_{u_1}(1)$, $F_{u_2}(1)$

— $F_{u_1}(2)$, $F_{u_2}(2)$

⋮

— $F_{u_1}(I)$, $F_{u_2}(I)$

$F_{u_1}(1)$	$F_{u_2}(1)$
$F_{u_1}(2)$	$F_{u_2}(2)$
⋮	⋮
$F_{u_1}(I)$	$F_{u_2}(I)$

→ on a projeté tous les individus sur (u_1, u_2)

→ la projection de tous les individus sur u_1 et la projection de tous les individus sur u_2

→ on compresse la matrice : au lieu d'avoir J coordonnées on a que 2 coord
↳ chaque individu n'a que 2 coordonnées au lieu de 2.

→ la m̂ chose pour les variables : on compresse la matrice de 2 côtés et au lieu d'avoir pour chaque variable I coordonnées on a que 2 coordonnées

⇒ Donc on peut représenter les individus dans un espace de dim 2 et représente les variables sur un espace de dim 2.

→ on peut ajouter v_3 et u_3, mais à partir de 3 on ne comprend plus rien, c'est un peu difficile d'imaginer ça

→ Par contre, le meilleur plan ça sera facile de projeter.

ACP ⇒ prendre ce qui est principale et laisser tomber ce qui est secondaire

Commentaires : sur le graphique

- les lignes représentent les individus on a I individus.
- les colonnes représentent les variables on a J variables.
  ex : les étudiants (individus) et les matières (variables)
- on peut faire une analyse duale et une analyse directe.
- notre but est d'avoir le meilleur plan de projection pour les variables et pour les individus
- pour le plan de projection des variables, on obtient les 2 vecteurs propres v_1 et v_2 ce sont les vecteurs propres de la matrice $ZZ^t = r$
  → les coordonnées des variables sont : $G_{v_1}(1)$ et $G_{v_2}(1)$ coordonnées de la variable 1 sur les vecteurs propres v_1 et v_2.
- chaque variable avait I dimension au début, ~~ensuite~~ mnt elle a que 2 dim.
- on projete les individus sur les vecteurs u_1 et u_2. Au début un individu avait J variables et mnt il a que 2 coordonnées. $F_{v_1}(1)$, $F_{v_2}(1)$

p 38

→ on projète sur le plan(u_α, u_β).

→ les individus ce sont les régions .

p39

. la coordonnée du point individus i sur l'axe α ($\vec{u_\alpha}$) est donné par le produit scalaire : $F_\alpha(i) = \underline{Z}_i^t . \underline{u}_\alpha$

. En empilant ces equations on obtient le vecteur formé par toutes les coordonnées des variables sur l'axe α :

$$\underline{F}_\alpha = \begin{pmatrix} F_\alpha(1) \\ F_\alpha(2) \\ \vdots \\ F_\alpha(i) \\ \vdots \\ F_\alpha(I) \end{pmatrix} = \begin{pmatrix} \underline{Z}_1^t \\ \vdots \\ \underline{Z}_i^t \\ \vdots \\ \underline{Z}_I^t \end{pmatrix} \underline{u}_\alpha = Z . \underline{u}_\alpha$$

$$\underline{F}_\alpha = Z . \underline{u}_\alpha \qquad 1 \leq \alpha \leq r = \text{rang}$$

projection de toutes les individus sur l'axe α.

p40

Remarque

on peut écrire aussi : $\underline{F}_\alpha = Z . \underline{u}_\alpha = (\underline{Z}^1, \underline{Z}^2, \dots \underline{Z}^j, \dots \underline{Z}^J) \begin{pmatrix} u_\alpha^1 \\ u_\alpha^2 \\ \vdots \\ u_\alpha^j \\ \vdots \\ u_\alpha^J \end{pmatrix}$

$$\underline{F}_\alpha = \sum_j^J \underline{u_\alpha^j \underline{Z}^j}$$

car $Z = (\underline{Z}^1, \underline{Z}^2, \dots, \underline{Z}^J) = \begin{pmatrix} \underline{Z}_1 \\ \underline{Z}_2 \\ \vdots \\ \underline{Z}_I \end{pmatrix}$

$$F_\alpha = Z^1 u_\alpha^1 + Z^2 u_\alpha^2 + \dots \underline{Z}^J . u_\alpha^J$$

⇒ une composante principale $\underline{\alpha}$ c'est une combinaison linéaire des variables.

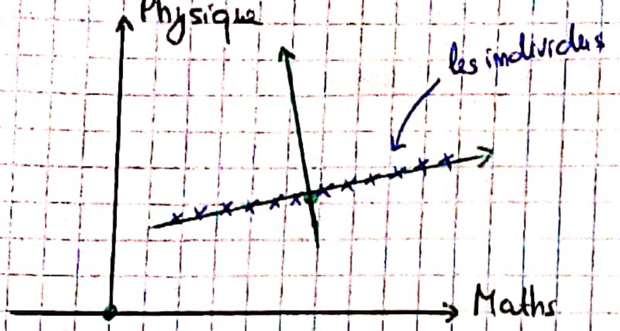
F_α : est une combinaison linéaire des variables , une sorte de variable de synthèse appelée :

- Facteur principal α ou
- Composante principale α

→ au maximum on a ~~r~~ au maximum $r = \text{rang}(Z)$ de E_u (car on ne peut avoir que r vecteurs propres).

exemple

* supposons qu'on a 2 variables : maths et physique. Le tableau de données contient les étudiants (les lignes) et les matières (les colonnes).



→ chaque individu a deux coordonnées (maths et PC).

→ la 1ère composante principale

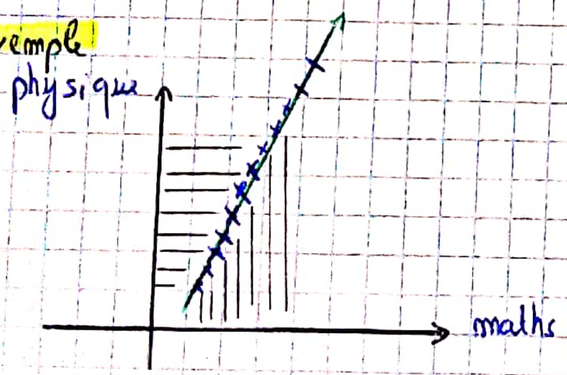
- centrer : ça veut dire changer le centre de gravité des individus
- chercher les vecteurs propres ça veut dire chercher comment projeter les individus dans l'espace des variables.

→ on cherche donc l'axe sur lequel les individus sont les plus étalés possible (l'axe en vert) donc on a 1 seule composante principale on n'a pas besoin de la 2ème dans cet exemple : si on diagonalise la matrice on va trouver 1 seule valeur propre. l'autre sera nulle, elle ne représente rien du tout, càd qu'elle ne contient aucune information.

→ Lorsqu'on va faire la projection, tout individu va se trouver sur l'axe en vert

→ donc il y a une seule composante principale, on peut ajouter une autre ⊥ mais elle ne contient rien. Tout se projette sur l'axe en vert.

exemple



→ si on fait l'ACP on obtiendra l'axe en vert.

→ est ce qu'il vaut mieux projeter les individus sur les maths ou sur la physique ?

⇒ Physique, car en projetant sur l'axe de la physique les individus seront les plus étalés possibles.

- on cherche l'axe sur lequel on a plus d'étalement
- pour l'ACP → on trouve que le meilleur axe c'est celui qui est en vert car on ne perd rien en projetant sur cet axe.
- Par contre en projetant sur les maths on perd qlq chose
- Lorsqu'on projète sur l'axe de physique c'est vrai qu'on a plus d'étalement mais perd un peu d'information.

l'ACP ⇒ cherche les meilleurs ~~et~~ axes d'inertie, sur lesquels on a plus d'étalement.

ACP : démonstration avec python (voir le fichier .ipynb) ACP_SUR_IRIS.ipynb

→ ACP des iris, ce sont des fleurs, chaque fleur a 4 coordonnées, on avait 150 fleurs et 3 catégories de fleurs

→ la 1ère chose qu'on fait c'est d'importer les biblios dans lesquelles on a besoin.

→ uci : c'est un site contenant plusieurs fichiers de données. Δ

→ url = " " : l'url du fichier iris dans le site uci. [affectation]

→ on lit le fichier de données avec pandas + on lui donne les noms des variables : long pétal .... + la classe.

→ df.head() : pour voir les 5 premières observations.

→ ensuite on définit les variables features = ['sepal length', ...]

↳ on constitue la matrice X, car on fait l'ACP donc on oublie le target (la classe), on l'utilise juste pour la couleur.

→ on le met dans X : X = df.loc[:, features].values

→ on met le target dans Y : y = df.loc[:, ['target']].values

↳ on l'utilise juste pour le coloriage.

X = x

→ ensuite on centre et on réduit x = StandardScaler().fit_transform(X)

→ on a besoin de 2 composantes principales : pca = PCA(n_components = 2)

F_1 et F_2 car on fait de l'ACP sur les individus.

→ on calcule v_1 et v_2 : principalComponents = pca.fit_transform(X)

on les projète.

→ on met F_1 dans la composante principale 1 et F_2 dans la composante principale 2.

```
principalDf = pd.DataFrame(data = principalComponents, columns = [
                    'principal component 1', 'principal component 2'
```

→ on affiche les 5 premières composantes principales. `principalDf.head(5)`

individus projetés sur les

→ si on fait `principalDf.shape` ⇒ (150, 2)

nb de lignes et colonnes

⇒ on obtient 2 colonnes et 150 lignes

→ les 2 colonnes sont les 2 composantes principales.

→ les coordonnées des individus sur les 2 composantes principales.

→ càd qu'on projeter toutes les fleurs qui avaient 4 variables sur la 1ère compo principale et sur la 2ème composante principale. Mnt elles ont que 2 coordonnées.

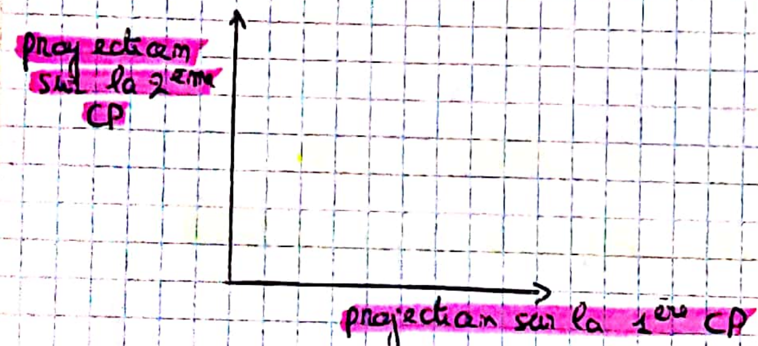
`df[['target']].head()` catégorie pour les 5 premières fleurs.

→ mnt on va concaténer les composantes principales avec leur target.

```
finalDf = pd.concat([principalDf, df[['target']]], axis = 1)
finalDf.head()
```

⇒ afficher les 5 fleurs : leurs 2 composantes principales et leur target.

→ Mnt on va faire la projection sur F_1 et F_2 (voir la suite)



* chaque axe a une valeur propre

$u_1 \quad \lambda_1$

$u_2 \quad \lambda_2$

$\vdots \quad \vdots$

$u_j \quad \lambda_j$

l'axe u_1 explique le max d'information $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_j$

→ le pourcentage : $\dfrac{\lambda_1}{\sum \lambda_j} \times 100\%$

$\dfrac{\lambda_2}{\sum \lambda_j} \times 100\%$

→ chaque axe explique un pourcentage

pca.explained_variance_ratio_

⇒ array([0,72770452, 0,23030523])

- pourcentage d'information expliqué par la 1ère composante principale
- pourcentage d'information expliqué par la 2ème composante principale.

⇒ le premier plan (u_1, u_2) explique 0,95% de l'information.

→ donc on a pas perdu bcp d'info en projetant sur un espace de 2 dim

de 4 dim à 2 dim sans perdre bcp d'info

ACP : avec Knime :

voir TP4

ACP : avec Tanagra

voir TP4 + voir documentation de Tanagra.

Exercice p 41 :

→ Prouver que la variable F_α est la moyenne nulle $\left(\dfrac{\sum_{i=1}^{I} F_\alpha(i)}{I} = 0 \right)$

F_α : coordonnées de tous les individus sur l'axe α.

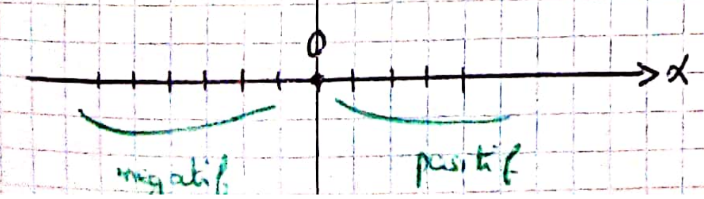
F_α est de dimension I

$F_\alpha(1), F_\alpha(2) \ldots F_\alpha(I)$

→ on veut montrer que la moyenne des F_α est nulle pour n'importe quel facteur (vecteur)

et l'invariance $= \dfrac{\lambda_\alpha}{I}$

→ F_α : projection des individus sur l'axe α.



il faut avoir des coords négatifs et d'autres positifs pour que la moyenne soit nulle.

α fixé

$$\sum_{i=1}^{I} F_\alpha(i) = 0$$

$$\overline{F}_\alpha = \frac{\sum_{i=1}^{I} F_\alpha(I)}{I} = 0 \quad \text{à montrer}$$

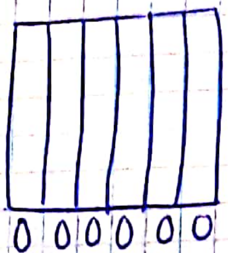
c.à.d que si on projète les individus sur n'importe quel axe on aura des coordonnées positifs et d'autres négatifs $\Rightarrow$ c'est toujours centré sur 0.

$$\rightarrow \overline{F}_\alpha = \frac{\sum_{i=1}^{I} F_\alpha(I)}{I} = \frac{\sum_{i=1}^{I} Z_i^t u_\alpha}{I} = \frac{\left(\sum_{i=1}^{I} z_i\right)^t u_\alpha}{I}$$

$F_\alpha(i) = Z_i^t . u_\alpha$ et $\sum_{i=0}^{I} Z_i = 0$ car Z est centré

donc

$\rightarrow \overline{F}_\alpha = 0$



on fait la somme sur la 1ère var = 0

sur la 2ème var = 0 ...

et on somme donc = 0

$\rightarrow$ Prouver que la variable F_α est de Variance $\frac{\sum_{i=1}^{I} (F_\alpha(i))^2}{I} = \frac{\|F_\alpha\|^2}{I} = \frac{\lambda_\alpha}{I}$

$$Var(F_\alpha) = \frac{\sum (F_\alpha(i) - \overset{\overset{0}{\shortparallel}}{\overline{F}_\alpha})^2}{I} = \frac{\sum_{i=1}^{I} (F_\alpha(i))^2}{I} = \frac{\|F_\alpha\|^2}{I} = \frac{F_\alpha^t F_\alpha}{I}$$

$$F_\alpha = Z . u_\alpha$$

$$\text{donc } var(F_\alpha) = \frac{(Z.u_\alpha)^t (Z.u_\alpha)}{I} = \frac{\overset{1}{\shortparallel}\ u_\alpha^t \overbrace{Z^t Z u_\alpha}^{V.u_\alpha = \lambda_\alpha}}{I} = \frac{\lambda_\alpha}{I}$$

Fin

$\rightarrow u_1$ vecteur propre sur lequel la projection est plus étalée, ça veut dire que la variance est la plus grande la variance qui correspond à $F_{u_1} = \frac{\lambda_1}{I}$ et λ_1 est la plus grande valeur propre.

$\rightarrow u_2$: à la 2ème plus grande variance. $F_{u_2} = \frac{\lambda_2}{I}$

Remarque

$$CTR_\alpha(i) = \frac{(F_\alpha(i))^2}{\lambda_\alpha} \times 100$$

contribution en % de l'individu i ($i = 1 \ldots I$) à l'inertie λ_α expliquée par l'axe α.

car $\sum = \lambda_\alpha$

p42 :

$\Rightarrow$ les individus projetés sur (u_1, u_2)

graphique des individus — graphique des variables

→ projection des individus.

→ projection des variables.

Commentaires :

→ les individus à droite sont caractérisés par les variables à droite.

→ les individus à gauche sont caractérisés par les variables à gauche.

→ Si un individu se trouve à côté d'une variable. Alors il est caractérisé par cette variable

[ex : si on a des étudiants à côté du maths alors ces étudiants sont bons en maths

→ les variables à droite caractérisent les régions à droite et ce sont les variables de richesse (PIB / hab, nb de voiture ...)

→ à gauche ce sont les régions pauvres

→ les régions à droite sont les régions les plus riches et les régions à gauche sont les plus pauvres.

→ donc les régions les plus pauvres de la France sont les régions d'Outre-Mer, ne se trouvent pas en France, ce sont des colonies françaises comme la Guyane, la Reunion ... sont les régions les plus pauvres car elles sont caractérisées par les variables de pauvreté

→ les régions ayant un fort PIB sont : Languedoc - Roussillon, Rhône - Alpes...

p. 43 Formules de transitions :

pour tout variable j : $G_\alpha(j) = \frac{1}{\sqrt{\lambda_\alpha}} \cdot \sum_{i=1}^{I} z_i^j F_\alpha(i)$ u_α

v_α

pour tout variable i : $F_\alpha(i) = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^{J} z_i^j G_\alpha(j)$

→ coordonnées des variables en fonction des coordonnées des individus.

→ " des individus " variables.

→ chaque variable se trouve au barycentre de tous les individus, chaque individu est pondéré par ~~son~~ ses variables qui le caractérise par exemple ceux qui sont bons en maths vont être pondérés par leur coordonnées en maths, vont être proche des maths

→ une variable se trouve au milieu des individus qui sont forts dans cette variable

une matière se trouve au milieu des étudiants forts dans cette matière.

un individu se trouve au milieu des variables où il est fort.

voir démo p 44.

Qualité de représentation des individus

- par exemple on a un individu Z_i et on va le projeter sur un axe u_α, plus il est proche du plan, mieux il est représenté.
- plus le θ est grand, moins il est représenté
- si $\theta = \frac{\pi}{2}$ on va tomber dans le centre de gravité.
- plus θ est petit, plus il va être sur l'axe α

⇒ la qualité de représentation : $CO2 = \cos^2\theta$

θ : l'angle que fait Z_i avec l'axe α, l'angle que fait l'individu avec l'axe α.

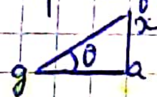
si on est dans un plan :

→ càd qu'on projète un individu sur le plan et on veut voir comment il est représenté

→ quelle est sa représentation sur un plan.

→ par ex : est ce qu'un étudiant est bien représenté lorsqu'on le projète sur le plan factoriel.

⇒ c'est l'angle qu'il fait avec le plan. Plus l'angle est petit, plus il est bien représenté



$$QLT_{\alpha,\beta} = \cos^2\theta = \frac{\|ga\|^2}{\|gx\|^2} = \frac{\|ga_1\|^2 + \|ga_2\|^2}{\|gx\|^2} = \frac{\|ga_1\|^2}{\|gx\|^2} + \frac{\|ga_2\|^2}{\|gx\|^2}$$

qualité de représentat° sur l'axe α et β

$$QLT_{\alpha,\beta} = CO2_\alpha + CO2_\beta$$

→ le logiciel Tanagra permet de calculer la qualité de représentation

p54 :

pourquoi $z^j = \sum_\alpha G_\alpha(j)\, v_\alpha$?

$\underline{Z}^j = \sum_\alpha G_\alpha(j)\underline{v}_\alpha$ c'est la représentation graphique d'une variable quelconque ~~sur l'axe~~ des vecteurs propres dans la base

on a démontré que $\underline{v}_\alpha = \dfrac{Z \cdot \underline{u}_\alpha}{\sqrt{\lambda_\alpha}}$

$$\underline{v}_\alpha = \frac{F_\alpha}{\sqrt{\lambda_\alpha}}$$

donc $\underline{Z}^j = \sum_\alpha G_\alpha(j) \cdot \dfrac{F_\alpha}{\sqrt{\lambda_\alpha}}$

→ chaque variable peut être exprimée par ses coordonnées sur les axes des individus et ses coordonnées, donc c'est une combinaison linéaire des coordonnées des individus correspondant à cette variable.

→ si on prend la coordonnée $i \longrightarrow F_\alpha(i)$

Exo :

① vérifier que $\underline{Z}_i = \sum_\alpha F_\alpha(i)\, \underline{u}_\alpha$

→ produit scalaire de $Z_i\, u_\alpha = \sum_\alpha F_\alpha(i)$

→ càd la représentation d'un individu dans la base de ses vecteurs propres

et $\underline{Z}^j = \sum_\alpha G_\alpha(j) v_\alpha$: représentation d'une variable sur la base de ce vecteur propre

② montrer que $\underline{Z} = \sum_\alpha \sqrt{\lambda_\alpha}\, \underline{v}_\alpha\, \underline{u}_\alpha^t$ (décomposition en valeurs singulières de $\underline{Z}$)