Analyse de données

Plan

- Analyse en Composantes Principales
- La statistique descriptive multidimensionnelle
- Exemple illustratif pour l'A.C.P.
- Présentation générale de la méthode
- Analyse Factorielle des Correspondances
- Principe général de l'A.F.C.
- Analyse des Correspondances Multiple
- Définition du tableau de Burt
- Principes de l'A.C.M

Présentation

- Si n individus et seulement 2 variables X et Y, il est facile de représenter l'ensemble des données sur un graphique plan : chaque individu i est un point de coordonnées X_i et Y_j → nuage
- L'allure du nuage renseigne sur l'intensité et la nature de la relation entre X et Y.
- **Si plus de 3 variables**, il faut **trouver de « bonnes » approximations du nuage** pour l'appréhender dans sa globalité.

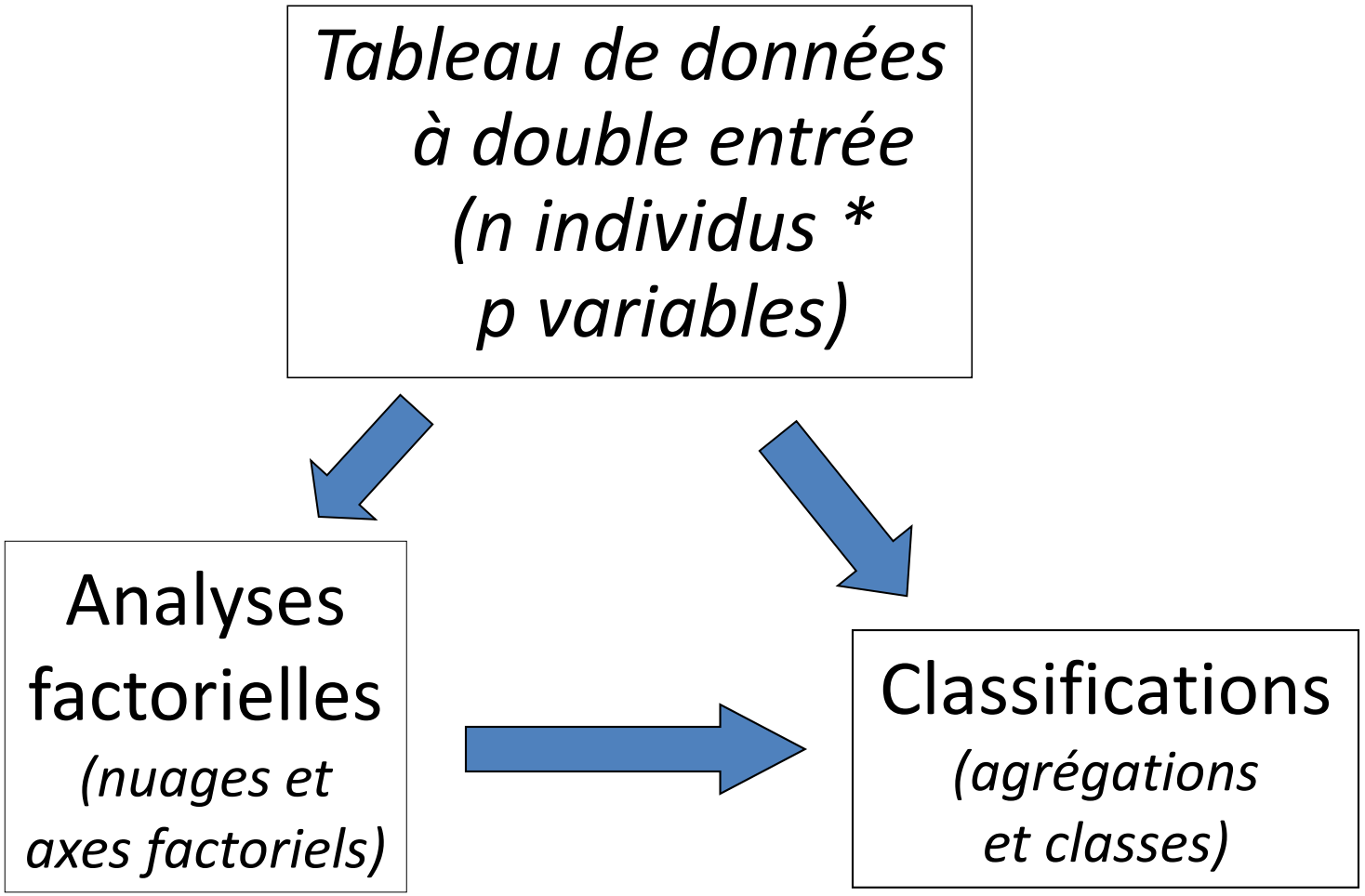
Analyses exploratoires de données

Définition :statistiques descriptives multidimensionnelles (beaucoup de dimensions)

Objectif : extraire l'information principale d'un tableau à double entrée, y compris quand il est très grand

Méthode : consentir une perte … d'information pour gagner … en efficacité

Deux grands types de méthodes



ADD

- L’analyse des données est un sous domaine des statistiques, pour étudier un phénomène déterminé. Ce module (Analyse des données) englobe plusieurs techniques pour aboutir à des résultats. Parmi ces technique il y’a l’ACP, L’AFC, L’AFCM et L’AFD.

Définition

- **Qu'est ce que l'analyse des données (ADD) :**

L'analyse statistique multidimensionnelle, souvent appelée analyse des données est la partie de la statistique qui traite des observations simultanées de plusieurs variables. L'objectif est d'élaborer et de figurer géométriquement dans un espace euclidien de faibles dimensions les informations consignées dans des tableaux statistiques. Il existe plusieurs méthodes adaptées à différents types de données, selon le nombre et la nature, quantitatives ou qualitatives, des variables.

L'analyse des données est une branche de la statistique dont l'objet est de découvrir la structure d'un ensemble de variables, à travers des observations, sans faire des hypothèses à priori sur la structure de ces variables. Vu que la structure des variables ne peut être connue directement du fait de la taille ou de la complexité des données mise en jeu, on fait appel à des méthodes et traitements spécifiques pour analyser et synthétiser les données. Ces analyses sont orientées en fonction des objectifs visés par l'étude statistique. On peut ranger les techniques de cette discipline en deux grandes familles :

Les méthodes factorielles

Les méthodes de classifications

Objectifs

La matrice des données ou tableau des données est le matériau de base de toute analyse de données, elle contient l'information brute la plus complète que l'on puisse obtenir.

Les analyses élémentaires, unidimensionnelles et bidimensionnelles permet de résumer de façon globale les données recueilles.

Élaborer et figurer géométriquement dans un espace euclidien de faible dimension les informations.

L'ADD permet aussi de décrire une partie des variables en le considérant sur le même plan à expliquer.

Objectifs

- Les analyses factorielles tentent de répondre à la question : tenant compte des ressemblances des individus et des liaisons entre variables, est-il possible de résumer toutes les données par un nombre restreint de valeurs sans perte d'information importante ? En effet en cherchant à réduire le nombre de variables décrivant les données, la quantité d'information ne peut être que réduite, au mieux maintenue. La motivation de cette ré-duction du nombre de valeurs vient du fait que des valeurs peu nombreuses sont plus faciles à représenter géométriquement et graphiquement (un des objectifs de l'analyse de données).

La statistique descriptive multidimensionnelle

- On désigne par statistique descriptive multidimensionnelle l'ensemble des méthodes de la statistique descriptive (ou exploratoire) permettant de traiter simultanément un nombre quelconque de variables

Les méthodes les plus classiques de la statistique descriptive multidimensionnelle sont les méthodes factorielles. Elles consistent a rechercher des facteurs en nombre restreint et résumant le mieux possible les données considérées.

Analyse factorielle

- Etude de la position d'un **nuage de points** dans l'espace et **description de sa forme**
- Pour mieux voir :
  - **se placer au milieu du nuage**, c'est-à-dire déplacer l'origine au centre de gravité (= individu fictif « moyen »)
  - **regarder dans les directions d'allongement principal**, c'est-à-dire changer d'axes
- Techniquement, **changer de repère** (→ diagonaliser une matrice)

Analyses factorielles

- Un tronc commun :

  Analyse des proximités
  au sein d'un nuage de points « pesants »
  selon une distance à déterminer

- Plusieurs analyses différentes
  selon la distance choisie :
  - Composantes principales (ACP)
  - Correspondances simples (AFC)
  - Correspondances multiples (ACM)
  - …

Rappels sur les distances

i × <——— D(i,j) ———> × j

- En géométrie :
  Distance euclidienne classique

$$D^2(i,j) = (X_i - X_j)^2 + (Y_i - Y_j)^2$$

(distance du double décimètre)

- En statistique :
  - p variables quantitatives
  - n individus, points d'un espace de dimension p
  - mesure des distances entre couples d'individus
  - la distance euclidienne classique ne convient pas
    → on pondère

Forme générale d'une distance euclidienne

$$D^2(i,j) = \sum \sum M_{ab} (X_{ia} - X_{ja}) (X_{ib} - X_{jb})$$

avec X_{ia} = valeur de la variable a pour l'individu i

et M_{ab} = coefficient de pondération de l'interaction des variables a et b

On peut lui associer une **métrique**, càd une matrice carrée à p lignes et p colonnes contenant les coefficients M_{ab}.

ACP

- Lorsqu'on étudie simultanément un nombre important de variables quantitatives (ne serait-ce que 4 !), comment en faire un graphique global ? La difficulté vient de ce que les individus étudiés ne sont plus représentes dans un plan, espace de dimension 2, mais dans un espace de dimension plus importante (par exemple 4). L'objectif de l'Analyse en Composantes Principales est de revenir a un espace de dimension réduite (par exemple 2) en déformant le moins possible la réalité. Il s'agit donc d'obtenir le résume le plus pertinent possible des données initiales.

ACP

- Faire le bilan des **ressemblances entre individus** et des **liaisons entre variables**
- Rechercher un **nombre limité de « variables » fictives** appelées « composantes principales », non corrélées entre elles et résumant le mieux possible l'information contenu dans le tableau des données brutes

ACP

- **Analyse de la structure de la matrice variance-covariance** c-à-d de la variabilité, dispersion des données.

Excepté si l'une des variables peut s'exprimer comme une fonction d'autres, on a besoin des *p* variables pour prendre en compte toute la variabilité du système

Objectif de l'ACP: ***décrire*** à l'aide de $q < p$ composantes ***un maximum*** de cette variabilité.

• Ce qui permet :

- une réduction des données à *q* nouveaux descripteurs
- une visualisation des données à 2 ou 3 dimensions (si $q = 2$ ou 3)
- une interprétation des données : liaisons inter-variables

• Etape intermédiaire souvent utilisée avant d'autres analyses !

Présentation

	MATH	PHYS	FRAN	ANGL
ALI	6.00	6.00	5.00	5.50
Med	8.00	8.00	8.00	8.00
Amir	6.00	7.00	11.00	9.50
Zaki	14.50	14.50	15.50	15.00
Samy	14.00	14.00	12.00	12.50
Amel	11.00	10.00	5.50	7.00
Neila	5.50	7.00	14.00	11.50
Nada	13.00	12.50	8.50	9.50
Kamel	9.00	9.50	12.50	12.00

ACP

- On sait comment analyser séparément chacune de ces 4 variables, soit en faisant un graphique, soit en calculant des résumés numériques. Nous savons également qu'on peut regarder les liaisons entre 2 variables (par exemple mathématiques et français), soit en faisant un graphique du type nuage de points, soit en calculant leur coefficient de corrélation linéaire, voire en réalisant la régression de l'une sur l'autre.
- Mais, comment faire une étude simultanée des 4 variables, ne serait-ce qu'en réalisant un graphique ? La difficulté vient de ce que les individus (les élèves) ne sont plus représentes dans un plan, espace de dimension 2, mais dans un espace de dimension 4 (chaque étudiant étant caractérisé par les 4 notes qu'il a obtenues). L'objectif de l'Analyse en Composantes Principales est de revenir a un espace de dimension réduite (par exemple, 2) en déformant le moins possible la réalité. Il s'agit donc d'obtenir le résume le plus pertinent des données initiales.

ACP

L'idée à la base de l'analyse en composantes principales est de pouvoir expliquer ou rendre compte de la variance observée dans la masse de données initiales en se limitant à un nombre réduit de composantes, définies comme étant des transformations mathématiques pures et simples des variables initiales.

L'algorithme utilisé pour la détermination de ces composantes obéit à deux contraintes importantes. Tout d'abord, la première composante extraite doit correspondre à un score composite qui **maximise** la proportion de variance expliquée dans les variables initiales. Pour comprendre cette idée il est avantageux de faire une analogie avec la technique de régression multiple.

Dans une analyse de régression multiple nous cherchons à expliquer le maximum de variance possible dans une variable critère (variable dépendante) en déterminant mathématiquement les pondérations optimales des différentes variables prévisionnelles (variables indépendantes).

ACP

l'algorithme utilisé dans l'ACP assure que la composante C1, la première extraite, correspondra à la plus grande proportion possible de variance présente dans les variables initiales. Ainsi, l'analyse en composantes principales nous mettra en présence d'une équation très apparentée à l'équation de

régression classique ayant la forme suivante :

$$C1 = â1\ var1 + â2\ var2 + â3\ var3\ldots + âk\ vark$$

Idéalement, nous aimerions que cette première composante C1 corresponde à une proportion très importante de la variance présente dans nos données initiales; ainsi, 80% ou 70% de variance expliquée à l'aide d'une première composante serait certainement un résultat très apprécié du chercheur.

Cependant la réalité est souvent moins gratifiante et il est fréquent de n'expliquer que 40%, 30%, ou même 20% lors de l'extraction d'une première composante.

La variance restante, inexpliquée par C1, n'est pas laissée de côté dans l'analyse des composantes principales; au contraire, elle est soumise à son tour au même processus d'extraction des composantes. Mais ici, l'algorithme à la base de l'ACP obéit à une deuxième contrainte importante : il cherche à extraire une deuxième composante, **indépendante de la première**, qui expliquerait à son tour la plus grande proportion de variance possible parmi la variance laissée inexpliquée par la composante C1.

ACP : données centrées

- Pour se placer au centre G du nuage E, on retire à chaque variable sa moyenne.
- On passe au tableau X_c des données centrées :

  $X_c = (y_{ij})$ avec $y_{ij} = x_{ij} - x_j$
- Chaque individu a un poids m_i

Opérations sur les matrices

- Multiplication par un scalaire : Soit X une matrice de taille n×p et λ un nombre réel (aussi appelé scalaire), alors la matrice λX est de taille n × p, et a pour coefficients :

$$(\lambda X)_{ij} = \lambda x_{ij} .$$

$$X = \begin{pmatrix} 3 & 2 & 1 \\ 4 & 6 & 1 \\ 8 & 1 & 0 \\ 2 & 1 & 5 \end{pmatrix}, \quad \lambda = 2, \quad 2X = \begin{pmatrix} 6 & 4 & 2 \\ 8 & 12 & 2 \\ 16 & 2 & 0 \\ 4 & 2 & 10 \end{pmatrix}$$

Multiplication

- Multiplication de matrices : Soit X une matrice de taille n × p, et Y une matrice de taille p × q, alors la matrice XY est de taille n × q, et a pour coefficients :

$$(XY)_{ij} = \sum_{k=1}^{p} x_{ik} y_{kj}$$

$$X = \begin{pmatrix} 3 & 2 & 1 \\ 4 & 6 & 1 \end{pmatrix}, \quad Y = \begin{pmatrix} 3 & 1 \\ 4 & 1 \\ 2 & 1 \end{pmatrix}, \quad XY = \begin{pmatrix} 19 & 6 \\ 38 & 11 \end{pmatrix}$$

Formules

- $Var(x) = \frac{1}{n-1}\sum_1^n (x_i - \bar{x})^2$
- Avec n est la dimension du vecteur x

Pour neutraliser le problème des unités on remplace les données d'origine par les données centrées-réduites :

$$X_1^* = \frac{X_1 - \bar{x}_1}{\sigma_1}$$

M

$$X_p^* = \frac{X_p - \bar{x}_p}{\sigma_p}$$

de moyenne 0 et d'écart-type 1.

Formules

L'**écart type** mesure la dispersion d'une série de valeurs autour de leur moyenne.

- En particulier, si la loi de *X* est uniforme sur un ensemble fini de valeurs, on a :

- $ecart\ type(\rho x) = \sqrt{\frac{1}{n-1}\sum_1^n (x_i - \bar{x})^2}$

Tableau statistique

Variable	Moyenne	Ecart-type	Minimum	Maximum
MATH	9.67	3.37	5.50	14.50
PHYS	9.83	2.99	6.00	14.50
FRAN	10.22	3.47	5.00	15.50
ANGL	10.06	2.81	5.50	15.00

Recap

$$\vec{x}_{.j} = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix} \qquad \bar{x}_{.j} = \frac{1}{n}\sum_{i=1}^{j} x_{ij}$$

$$X_c = \begin{bmatrix} x_{11} - \bar{x}_{.1} & \cdots & x_{1j} - \bar{x}_{.j} & \cdots & x_{1p} - \bar{x}_{.p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} - \bar{x}_{.1} & \cdots & x_{ij} - \bar{x}_{.j} & \cdots & x_{1ip} - \bar{x}_{.p} \\ \vdots & & \vdots & & \vdots \\ x_{n1} - \bar{x}_{.1} & \cdots & x_{nj} - \bar{x}_{.j} & \cdots & x_{np} - \bar{x}_{.p} \end{bmatrix}$$

Covariance

- On appelle covariance de x et y, et on note C_{xy} la quantité :
- $Cov(x,y) = \frac{1}{n-1}\sum_1^n (x_i - \bar{x})(y_i - \bar{y})$

Théorème de Koning-Huygens La variance est égale a :

$$Cov(x,y) = \frac{1}{n-1}\sum_1^n x_i y_i - \frac{n}{n-1}\bar{x}\bar{y}$$

La covariance est la moyenne des produits moins le produit des moyennes.

corrélation

La corrélation de Bravais-Pearson entre les variables $x_{(i)}$ et $x_{(j)}$, notée $r_{(ij)}$, est par définition :

$$r_{(ij)} = \frac{\mathrm{Cov}_{(ij)}}{\sigma_{(i)}\sigma_{(j)}} = \frac{((X - \overline{X})_{(i)})^t (X - \overline{X})_{(j)}}{||(X - \overline{X})_{(i)}|| \cdot ||(X - \overline{X})_{(j)}||}.$$

$$r_p = \frac{\sum_{i=1}^{N} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2}}$$

Corrélation

Remarques

1. $r_{(ii)} = 1$. En effet $r_{(ii)} = \frac{\mathrm{Cov}_{(ii)}}{\sigma^2_{(i)}} = \frac{\mathrm{Var}_{(i)}}{\mathrm{Var}_{(i)}} = 1$.
2. $r_{(ij)}$ représente le cosinus entre les vecteurs $(X - \overline{X})_{(i)}$ et $(X - \overline{X})_{(j)}$.

Coefficient de corrélation

	MATH	PHYS	FRAN	ANGL
MATH	1.00	0.98	0.23	0.51
PHYS	0.98	1.00	0.40	0.65
FRAN	0.23	0.40	1.00	0.95
ANGL	0.51	0.65	0.95	1.00

toutes les corrélations linéaires sont positives (ce qui signifie que toutes les variables varient, en moyenne, dans le même sens), certaines étant très fortes (0.98 et 0.95), d'autres moyennes (0.65 et 0.51), d'autres plutôt faibles (0.40 et 0.23).

Matrice Variances covariances

$$\Sigma_X = \mathrm{var}(\vec{X}) = \mathrm{var}\begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} = \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1X_2) & \cdots & \mathrm{cov}(X_1X_p) \\ \mathrm{cov}(X_2X_1) & \ddots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_PX_1) & \cdots & \cdots & \mathrm{var}(X_p) \end{pmatrix} = \begin{pmatrix} \sigma_{x_1}^2 & \sigma_{x_1x_2} & \cdots & \sigma_{x_1x_p} \\ \sigma_{x_2x_1} & \ddots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{x_px_1} & \cdots & \cdots & \sigma_{x_p}^2 \end{pmatrix}$$

- $Var(x_1) = \frac{1}{n-1}\sum_1^n (x_{i1} - \overline{x_1})^2$

$$Cov(x_1, x_2) = \frac{1}{n-1}\sum_1^n (x_{i1} - \overline{x_1})(x_{i2} - \overline{x_2})$$

Matrice variances covariances

	MATH	PHYS	FRAN	ANGL
MATH	11.39	9.92	2.66	4.82
PHYS	9.92	8.94	4.12	5.48
FRAN	2.66	4.12	12.06	9.29
ANGL	4.82	5.48	9.29	7.91

Valeurs propres

Soit A une matrice carrée n x n et X un vecteur colonne ayant n lignes. λ étant un scalaire. Considérons l'équation suivante:

$$AX = \lambda X$$

Pour X non nul, les valeurs de λ qui vérifient cette équation sont appelées **valeurs propres** de la matrice A. Les vecteurs correspondants sont appelés **vecteurs propres**.

L'équation peut également être écrite sous la forme:

$$(A - \lambda I)X = 0$$

Exemple

$$A = \begin{pmatrix} 5 & -3 \\ 6 & -4 \end{pmatrix}$$

$$det\begin{pmatrix} 5-\lambda & -3 \\ 6 & -4-\lambda \end{pmatrix}=0$$

$$(5-\lambda)(-4-\lambda)+18=0$$

$$\lambda^2-\lambda-2=0$$

$$\lambda=-1 \text{ ou } 2$$

Valeurs propres

FACTEUR	VAL. PR.	PCT. VAR.	PCT. CUM.
1	28.23	0.70	0.70
2	12.03	0.30	1.00
3	0.03	0.00	1.00
4	0.01	0.00	1.00
	40.30	1.00	

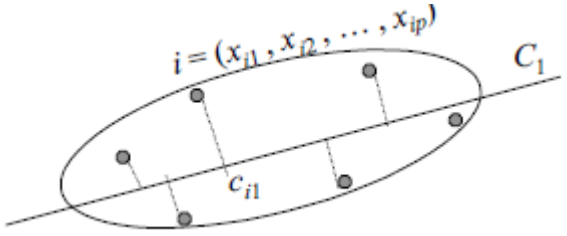
Interprétation

Chaque ligne du tableau ci-dessus correspond a une variable virtuelle dont la colonne val. pr. (valeur propre) fournit la variance (en fait, chaque valeur propre représente la variance du facteur correspondant). La colonne pct. var, ou pourcentage de variance, correspond au pourcentage de variance de chaque ligne par rapport au total. La colonne pct. cum., ou pourcentage cumule, représente le cumul de ces pourcentages.

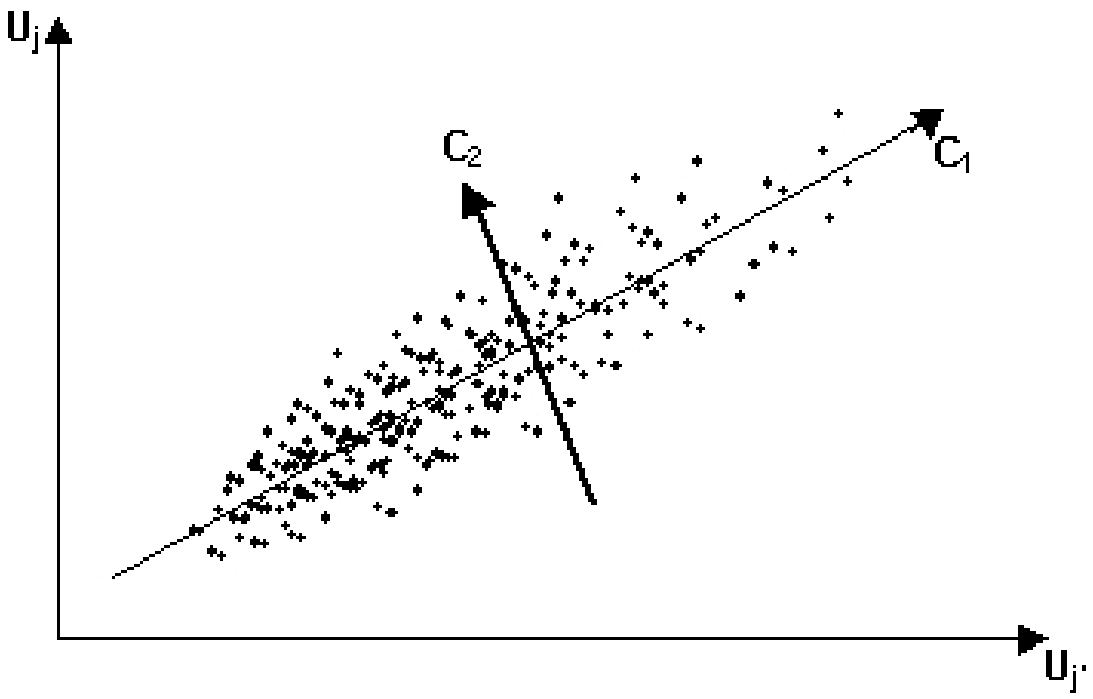
Composante principale

Composantes : C_1 , C_2 ,…, C_k ,…, C_q

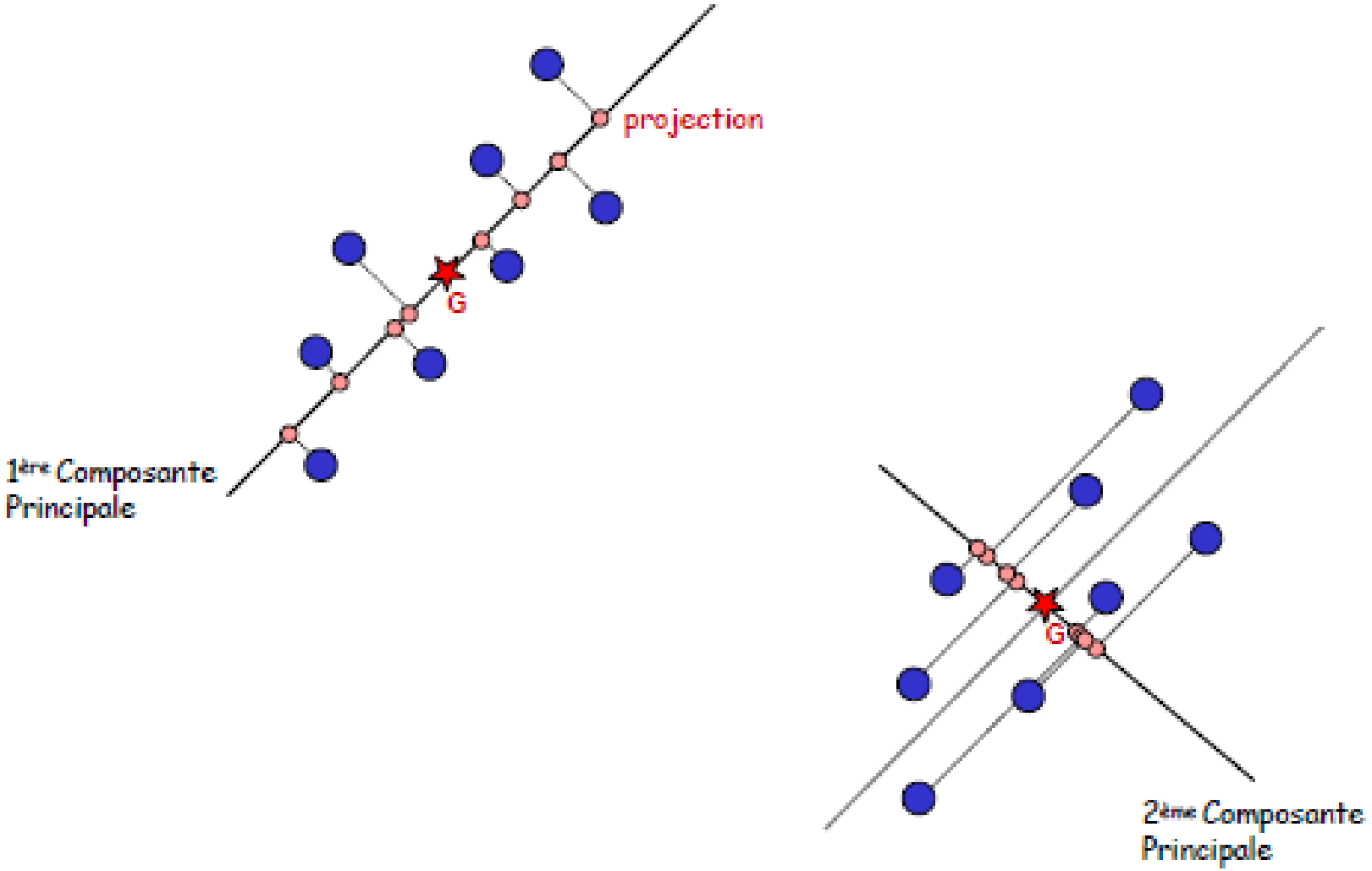
- C_k = nouvelle variable = combinaison linéaire des variables d'origine X_1,…, X_p:
- $C_k = a_{1k} X_1 + a_{2k} X_2 + \ldots\ldots + a_{pk} X_p$ coefficients a_{jk} à déterminer



Composantes principales



Projections



Composantes principales

- Composantes Principales orthogonales (non corrélées) deux à deux
- var(C1) > var(C2) > … > var(Cp)
- var(Ci) = λi

Vecteurs propres

Une matrice hermitienne A à toutes ses valeurs propres réelles et il existe une base orthonormée de $\mathbb{C}^n$ formée par les vecteurs propres de A.

- Notons Λ = diag(λ1, . . . , λn), λi ∈ IR

la matrice diagonale des valeurs propres de A hermitienne et V = [v1, . . . , vn] la matrice dont les colonnes sont les vecteurs propres correspondants.

Donc

$$\forall i, \quad Av^i = \lambda_i v^i \quad \Longleftrightarrow \quad AV = V\Lambda$$

$$\forall i, j \quad v^i {}^* v^j = \delta_{ij} \quad \Longleftrightarrow \quad V^*V = I_n \,,$$

$$\det(A) = \Pi_{i=1}^{n} \lambda_i \; = \lambda_1 \ldots \lambda_n .$$

$$\mathrm{trace}(A) = \sum_{i=1}^{n} a_{ii} \,. \qquad\qquad \mathrm{trace}(A) = \sum_i \lambda_i$$

Def

Techniques utilisées selon leur « origine »

Statistiques

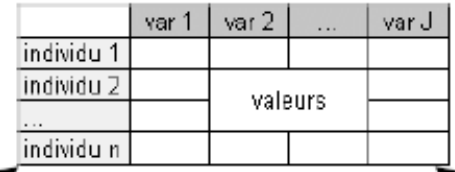
Théorie de l'estimation, tests
Économétrie

Maximum de vraisemblance et moindres carrés
Régression logistique, ...

Analyse de données (Statistique exploratoire)

Description factorielle
Discrimination
Clustering

Méthodes géométriques, probabilités
ACP, ACM, Analyse discriminante, CAH, ...



Informatique (Intelligence artificielle)

Apprentissage symbolique
Reconnaissance de formes

Une étape de l'intelligence artificielle
Réseaux de neurones, algorithmes génétiques...

Informatique (Base de données)

Exploration des bases de données

Volumétrie
Règles d'association, motifs fréquents, ...

Très souvent, ces méthodes reviennent à optimiser les mêmes critères, mais avec des approches / formulations différentes

Définitions

Définitions pré-requises :

Une matrice carrée $A = (a_{ij})$ est dite symétrique, ssi $a_{ij} = a_{ji}$ pour tout i, j.

- Une matrice carrée $Q = (q_{ij})$ est dite orthogonale, ssi $< q_{(i)}, q_{(j)} >= q^t_{(i)} q_{(j)} = 0$, pour tout $i <> j$, et $< q_{(i)}, q_{(i)} >= q^t_{(i)} q_{(i)} = 1$.

Projections sur un sous-espace

- Le principe de l'ACP est de trouver un axe *u*, issu d'une combinaison linéaire des $X_{.j}$, tel que:

la variance du nuage autour de cet axe soit maximale. Nous cherchons donc le vecteur *u* tel que la projection orthogonale du nuage sur *u* ait une variance maximale. Soit *C* la matrice de covariance ou de corrélation précédemment calculée. La projection de l'échantillon des *X* sur *u* s'écrit :

$$u(X) = X \cdot u$$

Projection

$$\text{Covariances} = 1/n \cdot \bar{X}^t \cdot \bar{X}$$

$$\pi_u(X)^t \cdot 1/n \cdot \pi_u(X) = u^t \cdot \underbrace{X^t \cdot 1/n \cdot X}_{C} \cdot u$$

Corrélations variables-facteurs

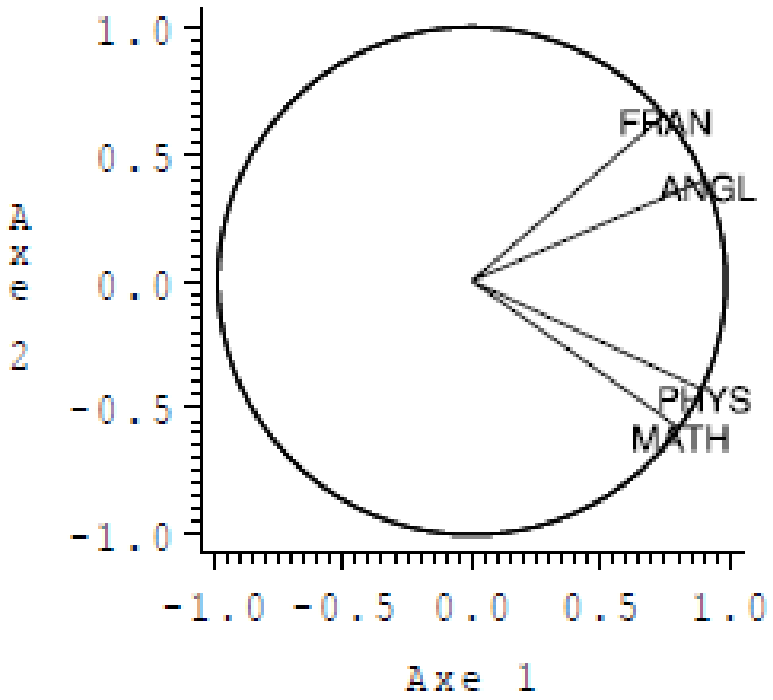
	Math	Phys	Fran	Angl
MATH	0.81	-0.58	0.01	-0.02
PHYS	0.90	-0.43	-0.03	0.02
FRAN	0.75	0.66	-0.02	-0.01
ANGL	0.91	0.40	0.05	0.01

Composantes principales

- Les vecteurs propres sont les coefficients à affecter aux variables initiales pour obtenir les composantes principales.

- La direction du premier axe factoriel est définie par le vecteur propre associé à la plus grande valeur propre de la matrice des variances-covariances.

Projection deux axes



Interprétations

on voit que le premier facteur est corrélé positivement, et assez fortement, avec chacune des 4 variables initiales : plus un étudiant obtient de bonnes notes dans chacune des 4 disciplines, plus il a un score élevé sur l'axe 1 ; réciproquement, plus ses notes sont mauvaises, plus son score est négatif ; l'axe 1 représente donc, en quelques sortes, le résultat global (dans l'ensemble des 4 disciplines considérées) des étudiants. En ce qui concerne l'axe 2, il oppose, d'une part, le français et l'anglais (corrélations positives), d'autre part, les mathématiques et la physique (corrélations négatives). Il s'agit donc d'un axe d'opposition entre disciplines littéraires et disciplines scientifiques, surtout marqué par l'opposition entre le français et les mathématiques. Cette interprétation, qui est déjà assez claire, peut être précisée avec graphiques et tableaux relatifs aux individus.

Résultats sur les individus

	POIDS	FACT1	FACT2	CONTG	CONT1	CONT2	COSCA1	COSCA2
• ali	0.11	-8.61	-1.41	20.99	29.19	1.83	0.97	0.03
• med	0.11	-3.88	-0.50	4.22	5.92	0.23	0.98	0.02
• amir	0.11	-3.21	3.47	6.17	4.06	11.11	0.46	0.54
• zaki	0.11	9.85	0.60	26.86	38.19	0.33	1.00	0.00
• sami	0.11	6.41	-2.05	12.48	16.15	3.87	0.91	0.09
• amel	0.11	-3.03	-4.92	9.22	3.62	22.37	0.28	0.72
• neila	0.11	-1.03	6.38	11.51	0.41	37.56	0.03	0.97
• nada	0.11	1.95	-4.20	5.93	1.50	16.29	0.18	0.82
• kamel	0.11	1.55	2.63	2.63	0.95	6.41	0.25	0.73

- Le Poids= 1/9
- Les 2 colonnes suivantes fournissent les coordonnées des individus (les étudiants) sur les deux premiers axes.
- Coordonnées des individus ; contributions ; cosinus carres

exemple

- Considérons la matrice : $A = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 3 & 0 \\ 2 & -4 & 2 \end{pmatrix}.$

Valeurs propres

$$\lambda_1 = 3, \quad \lambda_2 = 2, \quad \lambda_3 = 1.$$

- Ainsi *A* qui est de taille 3, a 3 valeurs propres distinctes, donc est diagonalisable.
- Si nous voulons diagonaliser *A*, nous avons besoin de déterminer les vecteurs propres correspondants. Il y a par exemple :

Suite exemple

$$v_1 = \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}, \quad v_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad v_3 = \begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix}.$$

$$A = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 3 & 0 \\ 2 & -4 & 2 \end{pmatrix}.$$

Maintenant soit *P* la matrice ayant ces vecteurs propres comme colonnes :

$$P = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ -2 & 1 & -2 \end{pmatrix}.$$

Alors « *P* diagonalise *A* », comme le montre un simple calcul :

$$P^{-1}AP = \begin{pmatrix} 0 & 1 & 0 \\ 2 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 2 & 0 \\ 0 & 3 & 0 \\ 2 & -4 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ -2 & 1 & -2 \end{pmatrix} = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Composantes principales

- $C1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$ = 1ère composante principale
- $C2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p$ = 2ème composante principale
- …C
- $p = a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p$ = pème composante principale

Les constantes $a_{i1}, a_{i2}, \ldots, a_{ip}$ sont les éléments du **vecteur propre** de la composante Ci

Notions de maths

- On repère dans le plan un point M par son abscisse x et son ordonnée y. Ceci pour dire que le vecteur OM s'écrit comme :

  $$OM = x\,i + y\,j$$

- Si maintenant, nous changeons d'axes, et que nous prenons deux nouveaux vecteurs directeurs k et l dans ce nouveau système d'axes, le point M a pour nouvelles coordonnées u et v. Ceci pour dire que le vecteur OM s'écrit comme :

  $$OM = u\,k + v\,l$$

Mathématique

- On passe facilement des coordonnées de M dans le système (i , j) aux coordonnées de M dans le système (k , l) c'est à dire de (x, y) à (u, v) en disant qu'il existe quatre nombres α, α', β et β' tels que :

  $k = \alpha i + \beta j$

  $l = \alpha' i + \beta' j$

- Dans ces conditions, on a :

$OM = u k + v l = u(\alpha i + \beta j) + v(\alpha' i + \beta' j)$

Mathématique

- soit :

$$OM = (u\alpha + v\alpha') i + (u\beta + v\beta') j$$

Autrement dit, entre les deux systèmes de coordonnées, on a la relation :

$$x = u\alpha + v\alpha'$$

$$y = u\beta + v\beta'$$

Il faut maintenant s'assurer que les distances calculées dans les deux systèmes d'axes sont les mêmes : autrement dit les coefficients α, α', β, β' ne peuvent pas être choisis n'importe comment : Si les coordonnées du point M sont (x, y) dans un premier système d'axes et (u, v) dans un second système d'axes on montre facilement qu'il existe deux nombres α et β tels que l'on ait :

$$\alpha^2 + \beta^2 = 1$$

$$x = u\alpha + v\beta$$

$$y = u\beta - v\alpha$$

Suite

dans le système d'axes (O, i , j), les coordonnées de M sont (x, y) et OM^2 est égal à $x^2 + y^2$ dans le système d'axes (O, k , l), les coordonnées de M sont (u, v) et OM^2 est égal à $u^2 + v^2$

- Ces deux quantités doivent coïncider :

$x^2+y^2= u^2+v^2$

soit :

$(u\alpha + v\alpha')^2+(u\beta + v\beta')^2= u^2+v^2$

En développant, on trouve :

$u^2(\alpha^2+\beta^2) + v^2(\alpha'^2+\beta'^2) + 2uv(\alpha\alpha'+\beta\beta') = u^2+v^2$

suite

Cette relation doit être vraie pour tous les points M du plan, c'est à dire pour toutes valeurs de u et v, en particulier :

$u = 1$ et $v = 0 \Longrightarrow \alpha^2 + \beta^2 = 1$

$u = 0$ et $v = 1 \Longrightarrow \alpha'^2 + \beta'^2 = 1$

La relation précédente s'écrit :

$$u^2+v^2+2uv(\alpha\alpha'+\beta\beta') = u^2+v^2$$

soit, en simplifiant et en choisissant $u = 1$ et $v = 1$

$$\alpha\alpha'+\beta\beta' = 0$$

En conclusion, les quatre nombres α, β, α', β' sont tels que

$$\alpha^2 + \beta^2 = 1$$

$$\alpha'^2 + \beta'^2 = 1$$

$$\alpha\alpha' + \beta\beta' = 0$$

Ce résultat admet une réciproque :

Suite

- Supposons qu'aucun des quatre nombres α, β, α', β' n'est nul. Appelons r le rapport α β et ρ le rapport β' α' . la dernière ligne s' écrit : $\alpha\beta'(r + \rho) = 0$. Ce qui indique que $r = -\rho$ et donc que $r^2 = \rho^2$.
- les deux premières lignes s'écrivent :

$$\beta^2(1 + r^2) = 1$$

$$\alpha'^2(1 + \rho^2) = 1$$

Suite

Ceci montre que $\beta^2 = \alpha'^2$ on obtient donc deux solutions selon que

Première solution $\beta = \alpha'$ et compte tenu de l'égalité entre r et $-\rho$,

$\beta = \alpha'$

$\beta' = -\alpha$

Seconde solution $\beta = -\alpha'$ et compte tenu de l'égalité entre r et ρ,

$\beta = -\alpha'$

$\beta' = \alpha$

nous trouvons deux solutions pour les relations entre (x, y) d'une part et (u, v) d'autre part :

Relation de type 1 : On se fixe deux nombres α et β tels que

$\alpha^2 + \beta^2 = 1$ et on a :

$x = u\alpha + v\beta$

$y = u\beta - v\alpha$

Relation de type 2 : On se fixe deux nombres α et β tels que $\alpha^2 + \beta^2 = 1$ et on a :

Suite

$$x = u\alpha + v\beta$$
$$y = -u\beta + v\alpha$$

Savoir calculer x et y à partir de u et v c'est bien, mais on peut aussi aller dans l'autre sens :

on montre facilement (exercice) que si

$$\alpha^2 + \beta^2 = 1$$
$$x = u\alpha + v\beta$$
$$y = u\beta - v\alpha$$

alors

$$\alpha^2 + \beta^2 = 1$$
$$u = x\alpha + y\beta$$
$$v = x\beta - y\alpha$$

- on est maintenant capable de calculer les coordonnées des nouveaux points dans n'importe quel système d'axes.

Exercice

Soit la matrice

$$\begin{matrix} 13 & 1 & -9 \\ 1 & 2 & 2 \\ -9 & 2 & 10 \end{matrix}$$

Calculer les valeurs propres, les vecteurs propres?

Est-ce qu'on considère les nouvelles coordonnées sur deux axes ou un axe?

Solution

- Valeurs propres :

20,6394103
4,3605897
-6,3697E-19

- Vecteurs Propres:

0,76048398	-0,02880728	-0,64871738
0,49384849	0,67433352	0,54898814
2	-3,5	2,5

il y a 3 valeurs propres, deux sont positives, la troisième est nulle ou négative. Chaque valeur propre correspond à un axe du graph :

— sur le premier axe, les coordonnées des points, x_b, x_c, x_d doivent être proportionnelles aux composantes du vecteur propre, c'est à dire qu'il existe un nombre α_1 tel que l'on aura :

$x_b = \alpha_1 * (0, 760483981)$

$x_c = \alpha_1 * (-0, 028807286)$

$x_d = \alpha_1 * (-0, 648717381)$

— α_1 doit être tel que la somme $x^2_b+x^2_c+x^2_d$ doit être égale à la première valeur propre. Soit :

$20, 63941029 = \alpha^2_1 * [(0, 760483981)^2 + (-0, 028807286)^2 + (-0, 648717381)^2]$

Avec $[(0, 760483981)^2 + (-0, 028807286)^2 + (-0, 648717381)^2] = 1$

$20, 63941029 = \alpha^2_1$ ou $\alpha_1 = 4, 543061775$

Suite

- Avec ce résultat, on obtient :

  $x_b = 3, 454925706$

  $x_c = -0, 130873278$

  $x_d = -2, 947163139$

Appliquons le même raisonnement pour le second axe :

Suite

— la seconde valeur propre est de 4, 360589709 tandis que les coordonnées y_b, y_c, y_d des trois variables doivent être proportionnelles à 0, 49384849 0, 6743335180 et 0, 548988141. Il existe donc un nombre α_2 tel que

$y_b = \alpha_2 * 0, 49384849$

$y_c = \alpha_2 * 0, 6743335180$

$y_d = \alpha_2 * 0, 548988141$

α_2 doit être tel que $y^2_b + y^2_c + y^2_d = 4, 360589709$, égal à la seconde valeur propre. Soit :

$\alpha^2_2 * [0, 493848492 + 0, 67433351802 + 0, 5489881412] = 4, 360589709$

Mais, $[0, 493848492 + 0, 67433351802 + 0, 5489881412] = 1$, il apparait que l'on a $\alpha^2_2 = 4, 360589709$, soit $\alpha_2 = 2, 088202507$

Il ne reste plus qu'à calculer les valeurs de y_b, y_c, y_d :

$y_b = 0, 49384849$

$y_c = 0, 674333518$

$y_d = 0, 548988141$

La troisième valeur propre est nulle, toutes les coordonnées sur le troisième axe sont nulles, ce qui veut dire que le graphique se trouve dans un plan.

Suite

points	Abscisse	Ordonnée
A	0	0
B	3,45492571	1,03125565
C	-0,13087328	1,40814494
D	-2,94716314	1,14639841

Exercice:

Trouver les coefficients nécessaires pour l’exemple des notes vue au cours.

Pondération des individus

Il est possible que les individus statistiques n'aient pas la même importance : si les individus statistiques sont par exemple les notes de modules, il faut accorder plus d'importance aux modules de spécialités (coefficients plus grand). On va donc mettre en place une pondération non uniforme des individus.

Les individus ont toujours un poids ; lorsque les individus ont la même importance, leurs poids sont identiques (uniformes) et cette étape peut être négligée.

Projections

- **Les vecteurs propres étant orthogonaux deux à deux , ils constituent une base orthonormée dans laquelle on peut représenter les vecteurs initiaux.**
- **Les coordonnées des vecteurs initiaux dans la nouvelle base sont données par leurs projections sur les vecteurs propres.**

$$F_{\alpha i} = X_i u_\alpha$$

ou $F = XU$

$$v = \frac{1}{\sqrt{\lambda}} Xu$$ est un vecteur propre de la matrice XX'

Les CONT décrivent les contributions des individus à l'inertie des axes

$$CONT_{\alpha i} = \frac{m_i}{\lambda_\alpha} F_{\alpha i}^2$$

$$avec \sum_{i=1}^{n} CONT_{\alpha i} = 1$$

Par construction, les individus les plus contributeurs sont excentrés.

- Les cosinus carrés (cos^2) décrivent les qualités de représentation des individus

$$COS_{\alpha i}^2 = \frac{F_{\alpha i}^2}{X_i^2}$$

Un COS^2 proche de 0 implique une mauvaise représentation de l'individu.
Un COS^2 proche de 1 implique une bonne représentation de l'individu.

Tous les points d'une photo uni ou bi-dimensionnelle ne sont pas visibles avec la même précision. En conséquence, on ne pourra interpréter la position relative de deux points projetés m_{l1} et m_{l2} que si elle reflète bien la position des points M_{l1} et M_{l2} de l'espace $(\mathbb{R}^p, M)$. De façon plus pratique, on mesurera la proximité relative entre m_l et M_l par le carré du cosinus du M-angle entre les vecteurs $\overrightarrow{Om_l}$ et $\overrightarrow{OM_l}$

$$\cos^2\theta_l = \frac{\| \overrightarrow{Om_l} \|_M^2}{\| \overrightarrow{OM_l} \|_M^2} = \frac{\| \overrightarrow{Om_l} \|_M^2}{X_l M X_l'} = \frac{\| \overrightarrow{Om_l} \|_M^2}{\mathbb{W}_l^l} = \frac{\| \overrightarrow{Om_l} \|_M^2}{\sum_{j=1}^r (C_l^j)^2}.$$

On dira que l'individu l est bien représenté par m_l si cette expression, appelée aussi *contribution relative de l'axe ou du plan factoriel à la représentation de l'individu* l, est voisine de 1, mal représentée si elle est voisine de 0.

Analyse Factorielle des Correspondances

L'A.F.C. est, en fait, une Analyse en Composantes Principales (A.C.P.) particulière, réalisée sur les profils associes a la table de contingence croisant les deux variables considérées. Plus précisément, l'A.F.C. consiste a réaliser une A.C.P. sur les profils-lignes et une autre sur les profils-colonnes. Les résultats graphiques de ces deux analyses sont ensuite superposes pour produire un graphique (éventuellement plusieurs) de type nuage de points, dans lequel sont réunies les modalités des deux variables considérées, ce qui permet d‘étudier les correspondances entre ces modalités, autrement dit la liaison entre les deux variables.

L'analyse des correspondances (croisement de deux variables qualitatives)

- Analyse dédiée à des tableaux croisant 2 variables qualitatives (notes obtenues x présence)
- Tableau de contingence

		Variable 2			Profil moyen
		Modalité 1	Modalité 2	Modalité 3	
Variable 1	Modalité 1	k_{11}	k_{12}	k_{13}	$k_{1.}$
	Modalité 2	k_{21}	k_{22}	k_{23}	$k_{2.}$
	Modalité 3	k_{31}	k_{32}	k_{33}	$k_{3.}$
	Modalité 4	k_{41}	k_{42}	k_{43}	$k_{4.}$
Profil moyen		$k_{.1}$	$k_{.2}$	$k_{.3}$	n

Suite

- L’analyse des correspondances va consister à étudier la répartition de chaque classe de la variable 1 suivant les modalités de la variable 2 (et inversement).
- On parle alors de profils lignes (lorsqu’on étudie les classes de la variable 1) et de profils colonnes (lorsqu’on étudie les classes de la variable 2).

Suite

Tableau des fréquences tel que fij = kij/n

		Variable 2			Profil moyen
		Modalité 1	Modalité 2	Modalité 3	
Variable 1	Modalité 1	f_{11}	f_{12}	f_{13}	$f_{1.}$
	Modalité 2	f_{21}	f_{22}	f_{23}	$f_{2.}$
	Modalité 3	f_{31}	f_{32}	f_{33}	$f_{3.}$
	Modalité 4	f_{41}	f_{42}	f_{43}	$f_{4.}$
Profil moyen		$f_{.1}$	$f_{.2}$	$f_{.3}$	1

Tableau des contributions

		Variable 2			
		Modalité 1	Modalité 2	Modalité 3	
Variable 1	Modalité 1				
	Modalité 2				
	Modalité 3		$c_{ij} = \frac{\left(f_{ij} - f_i f_j\right)}{f_i f_j}$		
	Modalité 4				

Suite

- Les fortes valeurs sont intéressantes, puisqu'elles dénotent une valeur "inattendue" par rapport à la structure générale du tableau ; en lecture rapide de tableaux, les raisons de ces fortes valeurs sont à étudier.
- Profils lignes tel que fij/fi. = kij/ki.

		Variable 2			Masse
		Modalité 1	Modalité 2	Modalité 3	
Variable 1	Modalité 1	$f_{11}/f_{1.}$	$f_{12}/f_{1.}$	$f_{13}/f_{1.}$	1
	Modalité 2	$f_{21}/f_{2.}$	$f_{22}/f_{2.}$	$f_{23}/f_{2.}$	1
	Modalité 3	$f_{31}/f_{3.}$	$f_{32}/f_{3.}$	$f_{33}/f_{3.}$	1
	Modalité 4	$f_{41}/f_{4.}$	$f_{42}/f_{4.}$	$f_{43}/f_{4.}$	1

suite

- Les profils-lignes donnent, pour chaque modalité de la variable 1, la répartition des modalités d
- Profils colonnes fij/f.j = k.j/k.j

		Variable 2		
		Modalité 1	Modalité 2	Modalité 3
Variable 1	Modalité 1	$f_{11}/f_{.1}$	$f_{12}/f_{.2}$	$f_{13}/f_{.3}$
	Modalité 2	$f_{21}/f_{.1}$	$f_{22}/f_{.2}$	$f_{23}/f_{.3}$
	Modalité 3	$f_{31}/f_{.1}$	$f_{32}/f_{.2}$	$f_{33}/f_{.3}$
	Modalité 4	$f_{41}/f_{.1}$	$f_{42}/f_{.2}$	$f_{43}/f_{.3}$
Masse		1	1	1

Les profils-colonnes donnent, pour chaque modalité de la variable 2, la répartition des modalités de la variable 1.

Tableau des fréquences théoriques

- Les deux variables sont indépendantes si :

$$f_{ij} = f_{i.}\, f_{.j}$$

- Alors, pour chaque modalité de la variable 1, le produit de la fréquence de chaque modalité de la variable 2 par la fréquence de la variable 1 est constant.
- Réciproquement, pour chaque modalité de la variable 2, le produit de la fréquence de chaque modalité de la variable 1 par la fréquence de la variable 2 est constant.

Suite

		Variable 2			Profil moyen
		Modalité 1	Modalité 2	Modalité 3	
Variable 1	Modalité 1	$f_{1.}\ f_{.1}$	$f_{1.}\ f_{.2}$	$f_{1.}\ f_{.3}$	$f_{1.}$
	Modalité 2	$f_{2.}\ f_{.1}$	$f_{2.}\ f_{.1}$	$f_{2.}\ f_{.3}$	$f_{2.}$
	Modalité 3	$f_{3.}\ f_{.1}$	$f_{3.}\ f_{.1}$	$f_{3.}\ f_{.3}$	$f_{3.}$
	Modalité 4	$f_{4.}\ f_{.1}$	$f_{4.}\ f_{.1}$	$f_{4.}\ f_{.3}$	$f_{4.}$
Profil moyen		$f_{.1}$	$f_{.2}$	$f_{.3}$	1

- Le cœur de l'AFC est de représenter les similitudes entre les différentes modalités d'une même variable, c'est-à-dire à représenter les proximités entre les profils et le profil moyen. Il faut donc considérer le nuage centré sur son centre de gravité.

- Le nuage des n lignes dans l'espace des p colonnes
  · Comme $\sum_{j=1}^{p} \frac{f_{ij}}{f_{i.}} = 1$, le nuage est même situé dans un sous-espace à p-1 dimensions
  · Le centre (de gravité) du nuage de points composé des f.j
- Le nuage des p colonnes dans l'espace des n lignes
  · Comme $\sum_{i=1}^{n} \frac{f_{ij}}{f_{.j}} = 1$, le nuage est même situé dans un sous-espace à n-1 dimensions
  · Le centre (de gravité) du nuage de points composé des fi..
- Exemple avec 3 variables initiales : le nuage de points est contenu dans un espace à 2 dimensions, centré sur le centre de gravité G.

- Quelle distance utiliser ?
- La distance euclidienne entre des points-lignes (respectivement colonnes) réalisée dans un tableau de données brutes traduirait la différence d'effectif entre deux modalités de la variable 1 (respectivement variable 2).
- La distance euclidienne entre profils-lignes (respectivement colonnes) traduirait bien la ressemblance entre deux modalités de la variable 1 (respectivement variable 2) sans tenir compte des effectifs totaux de ces deux modalités. Mais cette distance favorise les colonnes qui ont une fréquence élevée.
- Pour palier cela, on pondère chaque écart par l'inverse de l'effectif de la colonne (profils-lignes) ou de la ligne (profils-colonnes). Cette distance est appelée distance du χ^2:

Suite

$$d^2_{\chi^2}(j,j') = \sum_{i=1}^{n} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2$$

pour les profils-lignes

$$d^2_{\chi^2}(i,i') = \sum_{j=1}^{p} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$$

pour les profils-colonnes

Propriétés du χ^2

- Propriétés de la distance du χ^2

· Equivalence distributionnelle : on peut agréger deux modalités d'une même variable sans changer les distances entre modalités de cette variable, ni aux distances entre modalité de l'autre variable (on peut par exemple agréger les départements d'une même région)

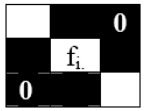
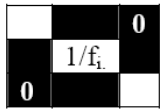
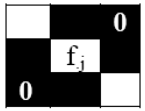
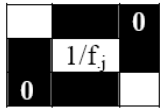
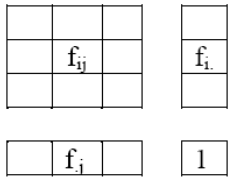
· Relations quasi-barycentriques : les modalités de la variable 1 sont des centres de gravité pour les modalités de la variable 2 prises ensemble, et réciproquement. En d'autres termes, l'ensemble des modalités de la variable 2 est contenu dans "l'enveloppe" des modalités de la variable 1, et réciproquement

Manière simple

La distance du X^2 peut être calculée par

$$\frac{(n_{\ell h} - \frac{n_{\ell +} n_{+h}}{n})^2}{\frac{n_{\ell +} n_{+h}}{n}}$$

Notations

D_n	matrice des marges-lignes	
D_n^{-1}	inverse de D_n	
D_p	matrice des marges-colonnes	
D_p^{-1}	inverse de D_p	
F	matrice des fréquences	

- Analyse du nuage des points-lignes

$$\begin{cases} \underset{u}{Max} \sum_{i} \overbrace{f_{i.}\, d^2_{\chi^2}}^{\substack{\text{distance du } \chi^2 \text{ des projections sur l'axe } u \text{ par rapport à l'origine pondérée par les fréquences des lignes}}} \quad (i, O) \\ s.c. \quad \underbrace{u' D_p^{-1} u}_{\substack{\text{vecteur unitaire pour la métrique utilisée}}} \end{cases}$$

- On pondère les modalités par leurs fréquences afin de ne pas privilégier les classes de faible effectif.
- Cela revient à résoudre le programme :

$$\begin{cases} Max\, u' \underbrace{\left(D_p^{-1} F'\right)}_{\text{Pr ofils-colonnes}} \underbrace{\left(D_n^{-1}\right)}_{\text{Métrique}} \underbrace{\left(D_p^{-1} F'\right)'}_{\text{Transposée des profils-colonnes}} u = u'\left(D_p^{-1} F'\right) D_n^{-1} \left(F D_p^{-1}\right) u = u'\, D_p^{-1} F'\, D_n^{-1} F D_p^{-1} u \\ s.c.\ \underbrace{u'\, D_p^{-1} u}_{\substack{\text{vecteur unitaire} \\ \text{pour la métrique} \\ \text{utilisée}}} \end{cases}$$

Suite

- En excluant la valeur propre triviale unitaire (analyse par rapport au barycentre), cela revient à diagonaliser la matrice :

$$S = F' D_n^{-1} F D_p^{-1}$$

- de terme général :

$$s_{jj'} = \sum_{i=1}^{n} \frac{f_{ij} \; f_{ij'}}{f_{i.} \; f_{.j'}}$$

Le tableau de référence

- Comment calculer « la probabilité d'observer un tableau » ?
- Pour répondre à cette question, on peut construire un tableau théorique,

Suite

$$\frac{N_{\bullet j}}{n} = \widehat{p}_{\bullet j} \to \frac{153}{180} = 0,85$$

$$\underbrace{0,85 \times 0,30}_{\widehat{p}_{ij} = \widehat{p}_{i\bullet}\widehat{p}_{\bullet j}} \times 180 = \frac{153 \times 55}{180} = 46,75$$

O	reçu	refusé	
Jury 1	50	5	55
Jury 2	47	14	61
Jury 3	56	8	64
	153	27	180

T	reçu	refusé	
Jury 1	46,75	8,25	30,56%
Jury 2	51,85	9,15	33,89 %
Jury 3	54,40	9,60	35,56 %
	85 %	15 %	180

Exemple 2

	université	Ecoles prépa	autres	Total
Sci	13	2	5	20
Math	20	2	8	30
Tech	10	5	5	20
Lettre	7	1	22	30
Total	50	10	40	100

Suite

M[1,1]= (20*50)/100

Matrice théorique

10	2	8
15	3	12
10	2	8
15	3	12

- Soit T un tableau de probabilité de même dimension. On appelle distance du χ^2 entre les tableaux O et T la quantité

$$D(O,T) = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(O_{ij} - T_{ij})^2}{T_{ij}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(N_{ij} - n\,\widehat{p}_{ij})^2}{n\,\widehat{p}_{ij}}$$

$O_{ij} = N_{ij}$ les effectifs observés

$T_{ij} = n\,\widehat{p}_{ij}$ les effectifs théoriques sous hypothèse d'indépendance

n l'effectif total

$\widehat{p}_{ij}$ la probabilité estimée sous hypothèse d'indépendance

Khi 2

$$
\begin{aligned}
D(O,T) &= \frac{(50-46{,}75)^2}{46{,}75} + \frac{(5-8{,}25)^2}{8{,}25} + \frac{(47-51{,}85)^2}{51{,}85} + \frac{(14-9{,}15)^2}{9{,}15} + \frac{(56-54{,}40)^2}{54{,}40} + \frac{(8-9{,}60)^2}{9{,}60} \\
&= 0,2259 + 1,2803 + 0,4537 + 2,5708 + 0,0471 + 0,2667 \\
&= 4,84
\end{aligned}
$$

Distance du χ^2 est-elle grande ?

La matrice des écarts à l'indépendance

13	**2**	**5**		**10**	**2**	**8**		**3**	**0**	**-3**
20	**2**	**8**		**15**	**3**	**12**		**5**	**-1**	**-4**
10	**5**	**5**	-	**10**	**2**	**8**	=	**0**	**3**	**-3**
7	**1**	**22**		**15**	**3**	**12**		**-8**	**-2**	**10**

Comment exprimer simplement R

On décompose la matrice des écarts à l'indépendance en une somme de matrices..

$R = T1 + T2$

.. Chacune de ces matrices étant mise en facteur (le produit d'un vecteur ligne et d'un vecteur colonne).

$$T_1 = C_1 L_1$$

$$R = T_1 + T_2 = C_1L_1 + C_2L_2$$

$$\begin{pmatrix} 3 & 0 & -3 \\ 5 & -1 & -4 \\ 0 & 3 & -3 \\ -8 & -2 & 10 \end{pmatrix} = \begin{pmatrix} 1 & 1 & -2 \\ 1 & 1 & -2 \\ 2 & 2 & -4 \\ -4 & -4 & 8 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 2 \\ -4 \end{pmatrix} + \begin{pmatrix} 2 & -1 & -1 \\ 4 & -2 & -2 \\ -2 & 1 & 1 \\ -4 & 2 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ -1 \\ -2 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 1 & -2 \end{pmatrix} \qquad \begin{pmatrix} 2 & -1 & -1 \end{pmatrix}$$

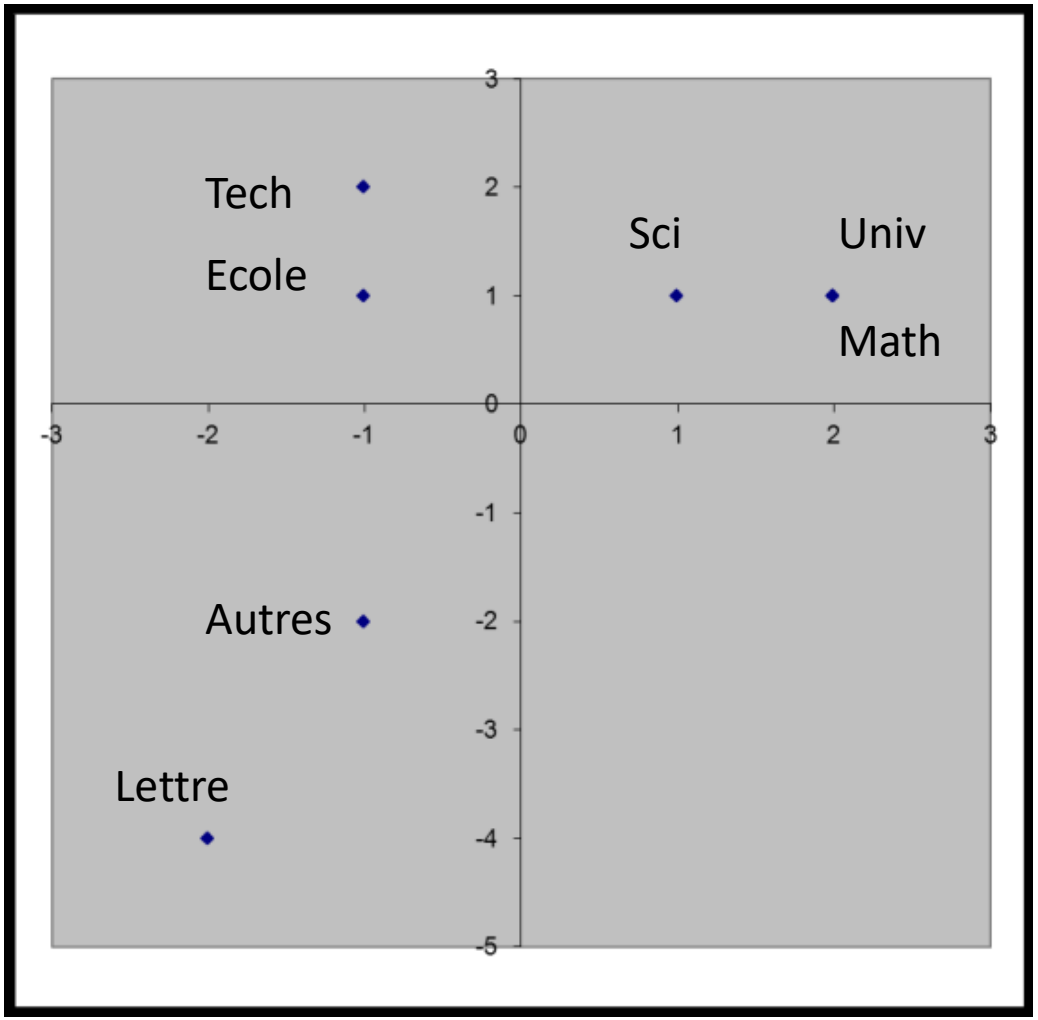
Suite

Sci	1
Math	2
Tech	-1
Lettre	-2

Univ	Ecole prepa	Autres
2	-1	-1

Graphique

Sci	1	1
Math	2	1
Tech	-1	2
Lettre	-2	-4
Univ	2	1
Ecole	-1	1
Autres	-1	-2



En effet $R = T_1 + T_2$… mais il existe aussi

$R = T'_1 + T'_2 = T''_1 + T''_2$ …

Quel est le critère (la métrique) qui permet de définir les meilleurs T_1 et T_2?

Pour une matrice de rang n, on cherche d'abord à trouver la meilleure T_1, puis la meilleure T_2 de telle manière à ce que le premier axe soit celui qui exprime le plus de sens..

$\chi^2(R) = \chi^2(T_1) + \chi^2(T_2)$

2491 = 1998 + 493

100% = 80.2% + 19.8%