NIVERSITE LARBI BEN M'HIDI OUM BOUAGHI
DEPARTEMENT MATHS ET INFORMATIQUE

TRAITEMENT STATISTIQUES DES DONNEES

COURS

MASTER 1 : INTELLIGENCE ARTIFICIELLE

DR BERKANE MOHAMED

Intitulé du Master :

Intelligence artificielle et sciences des données

Semestre : S1 Intitulé de l'UE : UEM1

Intitulé de la matière : Traitement statistique des données

Crédits : 04 Coefficients : 02

Mode d'enseignement : En présentiel

Objectifs de l'enseignement. La matière vise à donner aux étudiants les différentes techniques utilisées dans l'analyse, l'interprétation et la classification des données statistiques.

Connaissances préalables recommandées : Algorithmique, analyse et statistique.

Contenu de la matière

Chapitre 1. Introduction générale

Chapitre 2. Rappels d'algèbre linéaire et statistique descriptive

Chapitre 3. Analyses des caractères uni- et bivariées

Chapitre 4. Analyse en composantes principales (ACP)

Chapitre 5. Analyse des correspondances multiples (ACM)

Chapitre 6. Analyse factorielle discriminante (AFD)

Chapitre 7. Réseau de Kohonen

Mode d'évaluation : Continue 40% + Examen 60%

Références : - Bouroche, J.M et G. Saporta (2006), L'analyse des données, Presses Universitaires de France (PUF), Collection Que sais-je?, 9ème édition, 128p., 2006. - Saporta, G. (2006), Probabilités, analyse des données et statistiques, Technip, 2ème édition, 622p., 2006. - Denmat, A. et F. Héaulme (1999), Algèbre linéaire, Dunod, Collection Sciences Sup, Travaux Dirigés, DEUG MIAS/MASS, 1999.

Chapitre 1 :

Introduction Générale

-Statistique Descriptive -

Chapitre 1. Introduction générale

Le but du traitement statistiques des données ou bien analyse de données est de synthétiser, structurer l'information contenue dans des données multidimensionnelles. il s'agit d'un ensemble de méthodes statistiques, mathématiques ou informatiques qui permettent de transformer la donnée en information. L'analyse de données sert essentiellement à prédire et à comprendre.

1.1. La statistique

Définition : Le mot statistique désigne à la fois un ensemble de données, d'observations et l'activité qui consiste dans leur recueil, leur traitement et leur interprétation.

Deux classes de méthodes statistiques

- Statistique descriptive : Elle comprend la collecte des données, leur regroupement, leur représentation sous forme de tableaux et de graphiques, le calcul de totaux, de pourcentages, de moyennes et d'autres grandeurs caractéristiques. Autrement dit ; elle permet de dégager les caractéristiques essentielles du phénomène étudié et de suggérer des hypothèses pour une étude ultérieure plus poussée ; elle a pour but donc de résumer l'information contenue dans les données de façon synthétique et efficace par :
  - Représentations graphiques
  - Indicateurs de position, de dispersion et de relation
  - Régression linéaire ⇒ permet de dégager les caractéristiques essentielles du phénomène étudié et de suggérer des hypothèses pour une étude ultérieure plus poussée.

- Statistique inférentielle : Nécessite de définir des modèles probabilistes du phénomène aléatoire et savoir gérer les risques d'erreurs. Elle a pour but de faire des prévisions et de prendre des décisions au vu des observations par :
  - Estimation paramétrique
  - Intervalles de confiance,
  - Tests d'hypothèse.

Donc à partir d'observations faites sur un échantillon, de tirer des conclusions qui portent sur toute la population. La statistique inférentielle fait intervenir le calcul des probabilités.

1.2. Domaines d'application

- Economie, assurance, finance : études quantitatives de marchés, prévisions économétriques, analyse de la consommation des ménages, taxation des primes d'assurances et de franchises, gestion de portefeuille, évaluation d'actifs financiers, ...
- Biologie, médecine : essais thérapeutiques, épidémiologie, dynamique des populations, analyse du génome, ...
- Sciences de la terre : prévisions météorologiques, exploration pétrolière, ... Sciences humaines : enquêtes d'opinion, sondages, étude de population, ...
- Sciences de l'ingénieur : contrôle qualité, sûreté de fonctionnement, évaluation des performances, ...
- Sciences de l'information : traitement des images et des signaux, reconnaissance de forme et parole, machine learning, ...

1.3. Vocabulaire

Une population

Une population statistique est l'ensemble d'objets sur lequel on effectue des observations.

Les individus

sont les éléments de la population statistique étudiée. Pour chaque individu, on dispose d'une ou plusieurs observations.

Caractère statistique

ou variable statistique chaque individu d'une population est décrit par un ensemble de caractéristiques appelées variables ou caractères. On les note souvent par des lettres majuscules : X, Y.., donc c'est ce qui est observé ou mesuré sur les individus d'une population statistique.

Caractère quantitatif (variable quantitative)

C'est un caractère auquel on peut associer un nombre (poids, taille,). On distingue alors deux types de caractères quantitatifs :
• discret : c'est un caractère quantitatif qui ne prend qu'un nombre fini de valeurs
• continu : c'est un caractère quantitatif qui, théoriquement, peut prendre toutes les valeurs d'un intervalle de l'ensemble des nombres réels. Ses valeurs sont alors regroupées en classes.

Caractère qualitatif (variable qualitative)

C'est un caractère qu'on ne peut pas mesurer.

Effectif

On appelle effectif d'une valeur, le nombre de fois que cette valeur se répète dans l'échantillon.

1.4. Les matrices

Une matrice est un tableau de nombres. Le format de la matrice donne le nombre de lignes et le nombre de colonnes. Un terme (ou coefficient) de la matrice est indexé par deux indices donnant dans l'ordre le numéro de la ligne et le numéro de la colonne.

1.5. L'échantillonnage

Ensemble des méthodes permettant de sélection (de prélever) un échantillon de données au sein d'une population afin de juger cet ensemble en appliquant une méthode parmi de nombreuses mét hodes d'observations et de mesures.

1.6. Statistique univariée

La Statistique univariée (à 1 dimension) notée X est une application de l'ensemble de population vers l'ensemble des valeurs prise par le caractère.

1.6.1. La moyenne $\bar{x}$:

La moyenne est la mesure la plus commune de tendance centrale. Soit un échantillon de n valeurs observées x_1, x_2,..,x_i,..,x_n d'un caractère quantitatif X, on définit sa moyenne observée $\bar{x}$ comme la moyenne arithmétique des n valeurs :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Si les données observées sont regroupées en k classes d'effectifs fi (caractère continu ou discret), il faut les pondérer par les effectifs correspondants :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} f_i x_i \quad \text{avec } n = \sum_{i=1}^{k} f_i$$

Exemple :

On a relevé le prix de vente d'un stylo et le nombre de stylos vendus chez différents fournisseurs. Les résultats sont donnés dans le tableau suivant :

Prix de vente	15	16	17	18	19
Nombre de Stylos	97	34	43	20	6

Donner le prix de vente moyen p de ce stylo.

1.6.2. La médiane (Me) :

On appelle médiane d'une série statistique la valeur telle que l'effectif des valeurs inférieures à cette valeur est égal à l'effectif des valeurs supérieures. C'est donc la valeur qui partage l'échantillon en deux groupes de même taille.

1.6.3. Le mode (Mo)

Pour un caractère discret, le mode est la valeur du caractère la plus fréquente dans l'échantillon. Le mode correspond donc à la modalité dont l'effectif est le plus élevé. Il n'est pas toujours unique (un mode $\rightarrow$ série unimodale, deux modes $\rightarrow$ série bimodale, plusieurs modes $\rightarrow$ plurimodale). Pour un caractère continu, on appelle mode ou classe modale la classe correspondant au plus grand effectif.

Exemple :

1. Données groupées x_i 1 2 3 4 5 6 7 — un mode (série unimodale) : 4
   f_i 0 1 2 5 2 3 0
2. Donnéés brutes 1 2 2 3 3 3 3 4 4 4 4 5 6 — deux modes (série bimodale) : 3 et 4

1.6.4. Les quartiles

Les quartiles font un partage en quatre groupes égaux. Ils sont considérés comme un Indicateurs de position de tendance non centrale.

1. Le premier quartile Q1 (quartile inférieur) d'une série statistique est la valeur de la série qui correspond à la fréquence cumulée 0,25. C'est la médiane de la première série obtenue après avoir partagé la série initiale par sa médiane.
2. Le deuxième quartile Q2 est la médiane.
3. Le troisième quartile Q3 (quartile supérieur) d'une série statistique est la valeur de la série qui correspond à la fréquence cumulée 0,75. C'est la médiane de la deuxième série obtenue après avoir partagé la série initiale par sa médiane.

1.6.5. L'étendue

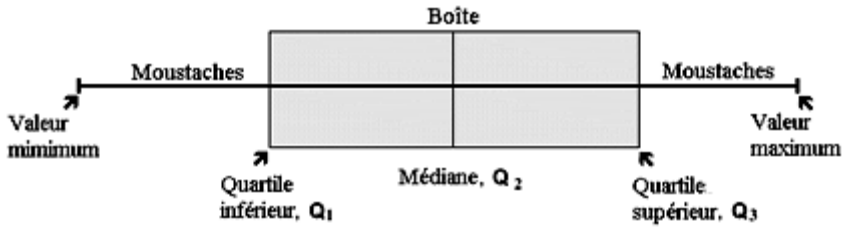
On appelle étendue d'une série statistique la différence entre la plus grande et la plus petite valeur observée.

1.6.6. L'écart interquartile

L'écart interquartile est une possibilité pour remédier au problème de la dépendance de l'étendue vis-à-vis des valeurs extrêmes. La différence **Q3 – Q1** est appelée écart interquartile. L'écart interquartile est l'étendue restante si on élimine les 25 % supérieurs et 25% inférieur de la distribution. C'est donc l'étendue des 50 % du milieu des observations.

1.6.7. Diagramme en boîte à moustaches (à pattes)

C'est un diagramme représenté par un rectangle délimité par le premier quartile et le troisième quartile. Pour l'obtenir, on trace un axe horizontal (ou vertical) sur lequel on place les valeurs de Q1 , Q3 et Q2 (la médiane) . L'un des côtés du rectangle a pour longueur l'écart interquartile, l'autre est quelconque. On complète ce diagramme en traçant deux traits horizontaux : l'un joignant Q1 au minimun de la série et l'autre joigant Q3 au maximun de la série. Ce sont les moustaches parfois appelées pattes.



- Les deux valeurs extrêmes représentent l'étendue
- Les quartiles Q_1 et Q_3 représentent l'écart interquartile $Q_3 - Q_1$
- La médiane.

1.6.8. Variance et écart-type

Pour mesurer le degré de dispersion des valeurs observées autour de la moyenne, on considère tous les écarts xi - $\overline{x}$, Au lieu de prendre la valeur absolue des écarts on peut aussi, pour éviter le signe négatif, élever les écarts au carré et en calculer ensuite la moyenne. Cette mesure nommée **variance** d'une série statistique (x_1 , x_2 , . . ., x_k) d'effectifs (n_1 , n_2 , . . ., n_k) est la valeur réelle notée *Var (X)* donné par

$$Var(X) = \frac{1}{N}\sum_{i=1}^{k} n_i(x_i - \overline{x})^2.$$

On appelle **écart type** de la distribution, la racine carrée positive de la variance, soit

$$\sigma_X = \sqrt{\frac{1}{N}\sum_{i=1}^{k} n_i(x_i - \overline{x})^2}$$

Plus l'écart type est petit, plus la distribution est rassemblée autour de la moyenne.

1.6.9. Covariance

La covariance est une mesure utilisée dans le cas de la statistique bivariée. Etant donnée une population de n individus, on souhaite étudier deux caractères différents X et Y et voir s'il existe un lien entre ces deux variables. Dans ce cas nous calculons la covariance qui n'est que la moyenne des produits des écarts pour chaque série d'observation.

$$cov(x,y) = \frac{1}{N}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

1.6.10. La corrélation

La corrélation est l'intensité de la relation existante entre deux séries de données. Le coefficient de corrélation mesure la dépendance linéaire entre les variables X et Y. c'est le rapport de la covariance sur le produit des écarts-types de deux variables X et Y.

$$r_{xy} = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

Remarque :

- -1 < r_{xy} < 1.
- Si r_{xy} est proche de 1 ou -1, les variables X et Y sont dits : fortement corrélées.

Propriétés :

- Si le coefficient de corrélation est positif, les points du nuage sont alignés le long d'une droite croissante. Dans ce cas X et Y évoluent dans le même sens.
- Si le coefficient de corrélation est négatif, les points sont alignés le long d'une droite décroissante. Dans ce cas X et Y évoluent dans des sens opposés.
- Si le coefficient de corrélation est nul ou proche de zéro, il n'y a pas de dépendance linéaire.

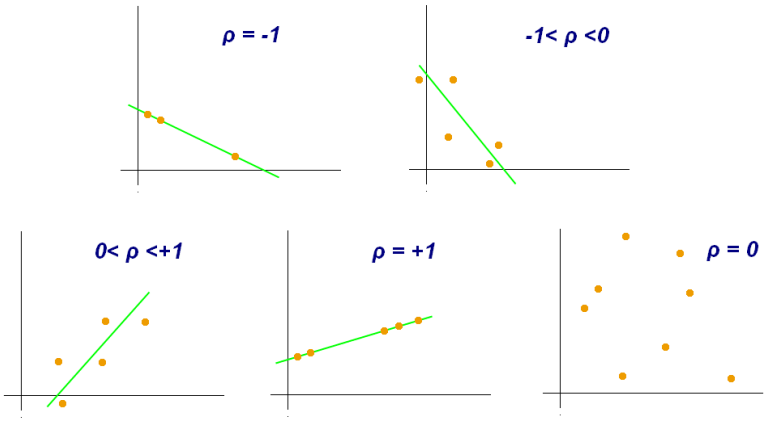


Figure : Diagrammes de dispersion avec différents cas de coefficient de corrélation.

1.6.11. Régression linéaire

La méthode des moindres carrés sert à ajuster les points, autrement dit, trouver la courbe qui représente mieux les données. Pour notre cas on se limite à une approximation linéaire. L'idée, donc, est de transformer un nuage de points en une droite la plus proche possible de chacun de ses points. On cherchera donc à minimiser les écarts entre les points et la droite. Cette méthode vise à représenter un nuage de points par une droite qui lie Y à X ayant une équation sous la forme :

$$Y = aX + b$$

Avec

$$a = \frac{Cov(x,y)}{Var(x)}$$

$$b = \bar{y} - a\bar{x}$$

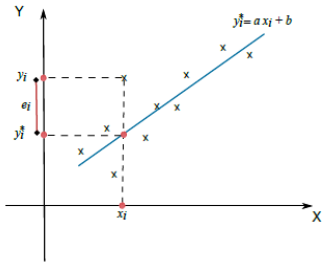


Figure : La droite la plus proche possible de chacun des points.

Exemple

L'étude de deux variables (X) le poids de 1000 grains en gramme et le rendement (Y) en quintaux par hectare chez le blé a donné les résultats suivants :

X (g)	43	46	48	50	52	55	56	58	60	62
Y (qx/ha)	30	33	35	36	37	39	39	42	43	45

1. Calculer la covariance Cov(xy) qui associe X et Y.
2. Calculer le coefficient de corrélation r_{xy} et quelle est votre conclusion ?
3. Déterminer l'équation de la régression qui lie X à Y.

Chapitre 2 : L'ACP

Analyse

En

Composantes Principales

Chapitre 2. Analyse en composantes principales

1. Introduction

L'Analyse en Composantes principales (ACP) est l'une des méthodes descriptives multidimensionnelles appelées méthodes factorielles. Dans la mesure où ce sont des méthodes descriptives, elles ne s'appuient pas sur un modèle probabiliste, mais elles dépendent d'un modèle géométrique. L'ACP propose, à partir d'un tableau rectangulaire de données comportant les valeurs de p variables quantitatives pour n individus, des représentations géométriques de ces individus et de ces variables. Ces données sont issues de l'observation d'une population. Les représentations des individus permettent de voir s'il existe une structure, non connue a priori, sur cet ensemble d'individus. De façon analogue, les représentations des variables permettent d'étudier les structures de liaisons linéaires sur l'ensemble des variables considérées. Ainsi, on cherchera si l'on peut distinguer des groupes dans l'ensemble des individus en regardant quelles sont les individus qui se ressemblent, celles qui se distinguent des autres, etc. Pour les variables, on cherchera quelles sont celles qui sont très corrélées entre elles, celles qui, au contraire ne sont pas corrélées aux autres, etc.

On cherche une représentation des n individus, dans un sous-espace F_k de R^p de dimension k (k petit 2, 3 … ; ex. un plan). Autrement dit, on cherche à définir k nouvelles variables combinaisons linéaires des p variables initiales qui feront perdre le moins d'information possible. Ces variables seront appelées « composantes principales »

Autrement dit, l'analyse en composantes principales (ACP) est une technique multivariée dite d'interdépendance. L'ACP, visent trois objectifs principaux :
1. Comprendre la structure d'un ensemble de variables afin de voir quelles variables sont associées.
2. Concevoir et raffiner des instruments de mesure comme les tests psychométriques qu'il est impossible de mesurer directement comme le degré de stress ou de bonheur d'une personne.
3. Condenser l'information contenue à l'intérieur d'un grand nombre de variables en un ensemble restreint de nouvelles dimensions tout en assurant une perte minimale d'informations.

2. Tableau de données

Les données sont les mesures effectuées sur n individus appelés aussi « unités » notés $\{u_1, u_2, .., u_i, ...u_n\}$. Les p variables quantitatives qui représentent ces mesures sont $\{v_1, v_2, ..., v_j, ...v_p\}$. Le tableau des données brutes à partir duquel on va faire l'analyse est noté X et a la forme suivante :

$$\mathbf{X} = \begin{array}{c} \\ u_1 \\ u_2 \\ . \\ u_i \\ . \\ u_n \end{array} \begin{array}{c} \begin{array}{cccccc} v_1 & v_2 & \dots & v_j & \dots & v_p \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2p} \\ . & . & \dots & . & \dots & . \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{ip} \\ . & . & \dots & . & \dots & . \\ x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix} \end{array}$$

Le principe est de résumer le tableau X en minimisant le nombre de variables

3. L'analyse en composante principale

Le principe pour déterminer la composante principale est de suivre les étapes suivantes :

1- Centrage et réduction des données
2- Déterminer la matrice de variance-covariance
3- Déterminer la matrice de corrélation
4- Déterminer les valeurs propres et les vecteurs propres sur la base de la matrice de corrélation entre les variables
5- Déterminer les axes factoriels → Sélectionner les composantes principales

3.1. Centrage et réduction des données

Soit : n individus caractérisés par p variables de nature différente ; pour homogénéisée les unités, les p variables seront centrées et réduites. Cela signifie que pour chaque variable, la moyenne sera nulle et la variance est égale à 1.

La matrice centrée et réduite est obtenue par la formule suivante :

$$x_{ij} = \frac{x_{ij} - \bar{x}}{\sigma_i}$$

Remarque : Centrer et réduire les données permet de donner le même poids à toutes les variables dans le calcul de la distance entre deux individus.

3.2. La matrice des variances covariances

La matrice des variances covariances permet de mesurer la liaison linéaire qui peut exister entre deux variables statistiques.

Exemple : Matrice à trois variables.

$$\begin{pmatrix} var(X1) & cov(X1,X2) & cov(X1,X3) \\ cov(X2,X1) & var(X2) & cov(X2,X3) \\ cov(X3,X1) & cov(X3,X2) & var(X3) \end{pmatrix}$$

Si Cov(X2,X1) =0 : alors les deux variables X1 et X2 sont indépendantes
Si Cov(X2,X1) <> 0 : alors les deux variables X1 et X2 sont dépendantes
La matrice de variances covariances est obtenue par la formule suivante :

$$V = \frac{1}{n} MC * MC^t$$

$$avec\ MC: la\ matrice\ centrée\ \ tel\ que\ x_{ij=}(x_{ij} - \bar{x})$$
$$MC^t: la\ matrice\ centrée\ transposée$$

3.3. Matrice des corrélations

Matrice des corrélations entre variables permet d'analyser les relations bilatérales entre les variables. Elle est obtenue par la formule suivante :

$$C = \frac{1}{n} CR^t * CR$$

$$avec\ CR: la\ matrice\ centrée\ \ réduie$$
$$CR^t: la\ matrice\ centrée\ réduite\ transposée$$

Exemple : Matrice à trois variables.

$$\begin{pmatrix} 1 & c(X1,X2) & c(X1,X3) \\ c(X2,X1) & 1 & c(X2,X3) \\ c(X3,X1) & c(X3,X2) & 1 \end{pmatrix}$$

3.4. Valeurs propres et vecteur propres

Soit $A(n,n) \in$ une matrice carrée de valeurs dans R :

- λ est dite **valeur propre** de la matrice A s'il existe un vecteur non nul $X \in$ K n tel que

$$AX = \lambda X$$

- Le vecteur *X* est alors appelé **vecteur propre** de A associé à la valeur propre λ.

Déterminer le valeurs propres et vecteurs propres revient à résoudre l'équation suivantes :

$$\det(A - \lambda I_n) = 0$$

Dans notre cas, A est la matrice de corrélation entre les variables, autrement dit l'équation s'écrit comme suit :

$$\det(C - \lambda I_n) = 0$$

Exemple : Matrice à trois variables.

$$det\left(\begin{pmatrix} 1 & c(X1,X2) & c(X1,X3) \\ c(X2,X1) & 1 & c(X2,X3) \\ c(X3,X1) & c(X3,X2) & 1 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right) = 0$$

$$\Rightarrow \left| \begin{pmatrix} 1 & c(X1,X2) & c(X1,X3) \\ c(X2,X1) & 1 & c(X2,X3) \\ c(X3,X1) & c(X3,X2) & 1 \end{pmatrix} - \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix} \right| = 0$$

$$\Rightarrow \begin{vmatrix} 1-\lambda & c(X1,X2) & c(X1,X3) \\ c(X2,X1) & 1-\lambda & c(X2,X3) \\ c(X3,X1) & c(X3,X2) & 1-\lambda \end{vmatrix} = 0$$

Exemple :

$$A = \begin{pmatrix} 1 & 3 & 3 \\ -2 & 11 & -2 \\ 8 & -7 & 6 \end{pmatrix}$$

3.5. Calcul d'inertie

Le pourcentage d'inertie exprimé par un axe factoriel permet d'évaluer la quantité d'information représentée par cet axe.

$$inertie\ d'un\ axe = \frac{val.\ propre\ associée}{\sum vals.\ propres}$$

4. Etude de cas

La problématique est la suivante : On dispose au préalable d'un ensemble de n visages de différentes personnes (parfois plusieurs pour une même personne). On cherche à déterminer si une nouvelle image que l'on reçoit appartient à une des personnes connues et, si c'est le cas, à laquelle.

Notons $\boldsymbol{V}=\boldsymbol{v}_i$, i = 1,...,n les visages. Chaque visage est représenté en dans son fichier, pixel par pixel où chaque pixel peut être exprimé par trois valeurs numérique représentant les trois couleurs de base : vert, bleu et rouge. On peut aussi utiliser des images noir et blanc, cad, en niveau de gris, donc chaque pixel est représenté par une seule valeur indiquant son intensité. Les différentes images sont supposées être de la même taille. Sous cette hypothèse, chaque image est alors transformée en vecteur. Chacun de ceux-ci est de même dimension N. On notera également $\boldsymbol{v}_i$ ces vecteurs

$$V = \begin{pmatrix} \boldsymbol{v}_1 \\ \vdots \\ \boldsymbol{v}_n \end{pmatrix} tel\ que:$$

$$\boldsymbol{v}_i = \begin{pmatrix} p_1 & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & p_m \end{pmatrix}$$

Après transformation des matrices visages sous forme de vecteur unidimensionnelle, on obtient :

$$v_i = (p_1, \cdots \quad p_j \cdots \quad p_m)$$

$$par\ conséquent: V = \begin{pmatrix} p_{11} \cdots p_{1m} \\ \vdots \ \ddots \ \vdots \\ p_{n1} \cdots p_{nm} \end{pmatrix}$$

Les moyennes des visages

$$\overline{v_i} = \frac{1}{n} \sum_{j=1}^{m} p_{ij}$$

Le vecteur des moyennes de tous les visages **v** est le suivant :

$$\overline{\boldsymbol{v}} = \begin{pmatrix} \overline{\boldsymbol{v}}_1 \\ \vdots \\ \overline{\boldsymbol{v}}_n \end{pmatrix}$$

La matrice centrée est comme suit :

$$MC = V - \overline{\boldsymbol{v}}$$

A partir de la matrice *MC,* on calcule la matrice centrée réduite et par conséquent, on calcule matrice de corrélation. Par la suite on calcule les valeurs propres et les vecteurs propres qui sont de taille N avec N inférieur ou égale à n. chaque vecteur propre correspond à une image. Ces images (ainsi que ces vecteurs) sont appelées les visages propres. L'espace des vecteurs propres est appelé espace des visages.

Dans la pratique, on n'utilise qu'un nombre réduit de vecteur propres, on sélectionne les vecteurs ayant des inerties importantes. Donc, chaque visage peut être projeté dans ce nouvel espace avec une perte minimale.

Chapitre 3 : L'ACM

Analyse

En

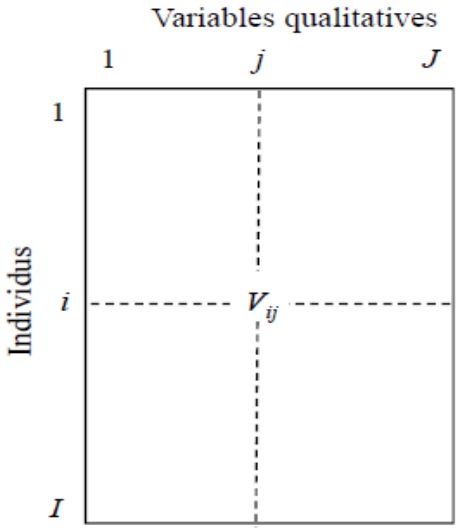
Composantes Multiples

Chapitre 3. Analyse en composantes multiples

1. Introduction

On est supposé être devant un tableau de variables qualitatives décrit par des individus. Sur ce tableau, on veut savoir les différents regroupements suivant certains aspects. Par exemple résultat d'une enquête où $\boldsymbol{I}$ personnes sont interrogées sur $\boldsymbol{J}$ questions à choix multiples. On cherche à regrouper les individus qui se ressembles en se posant deux questions : combien de groupe former, où mettre les coupures.

L'ACM s'intéresse à des tableaux de données rectangulaires qualitatives où les individus sont en lignes et les variables en colonnes. Chaque individu est décrit par les numéros des catégories où il est classé pour les p variables. Les données brutes se présentent sous forme d'un tableau à n lignes et p colonnes.



Etude des individus Un individu = une ligne du TDC = ensemble de ses modalités Ressemblance des individus Variabilité des individus Principales dimensions de la variabilité des individus (en relation avec les modalités) 2 Etude des variables Liaisons entre variables qualitatives (en relation avec les modalités) Visualisation d'ensemble des associations entre modalités Variable synthétique (Indicateur quantitatif fondé sur des variables qualitatives)

On notera :

- I : nombre d'individus
- J : Nombre de variables
- K_j le nombre de modalités de la variable j
- v_{ij} la modalité prise par l'individu i sur la variable j.
- $K = K_1 + \ldots + K_j$ le nombre total de modalités.

2. Tableau disjonctif complet

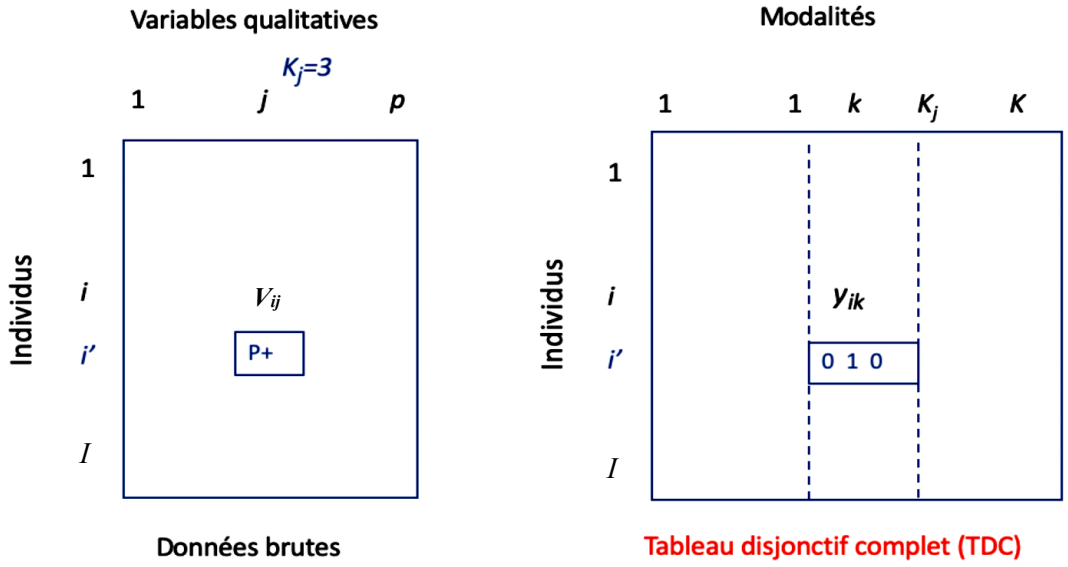
La forme mathématique utile pour les calculs est alors le tableau disjonctif des modalités des J variables obtenues en juxtaposant les J tableaux des modalités de chaque variable. Le résultat de cette opération est un tableau dit « tableau disjonctif complet ».
Exemple :

$I=5$; $J=3$; $K_1=3$; $K_2=2$; $K_3=3$

	codage réduit	codage disjonctif
1	1 2 3	1 0 0 | 0 1 | 0 0 1
2	2 1 1	0 1 0 | 1 0 | 1 0 0
3	2 2 2	0 1 0 | 0 1 | 0 1 0
4	3 2 1	0 0 1 | 0 1 | 1 0 0
5	3 1 2	0 0 1 | 1 0 | 0 1 0

$$X = (X_1 \mid X_2 \mid X_3)$$

La forme générale du tableau disjonctif complet est comme suit :



Remarques :

- La somme des éléments de chaque ligne est égale à p : nombre de variables
- La somme des éléments d'une colonne donne l'effectif marginal de la catégorie correspondante.

Le tableau disjonctif complet (TDC) est prétraité avant d'être analysé en ACM.

TDC

	1 …	k	… K
1			
⋮		⋮	
i	…	y_{ik}	…
		⋮	
I			
n			
moy	…	$p_k = \frac{n_k}{I}$	…

TDC centrée et "réduit"

	1 …	k	… K
1			
⋮		⋮	
i	…	$x_{ik} = \frac{y_{ik}}{p_k} - 1$	…
		⋮	
I			
n			
moy	…	0	…

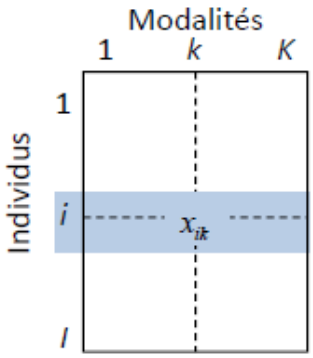
La moyenne	Valeur centrée	Valeur centrée réduite
$p_k = \frac{n_k}{I}$ tel que $n_k = \sum_{i=1}^{n} y_{ik}$	$x_{ik,centrée} = (y_{ik} - p_k)$	$x_{ik} = \frac{(y_{ik} - p_k)}{p_k}$

Le tableau disjonctif complet est :

- **Centré** : soustraction de la moyenne $p_k = n_k/n$
- **Réduit** : division par la fréquence p_k.

3. Nuage des individus

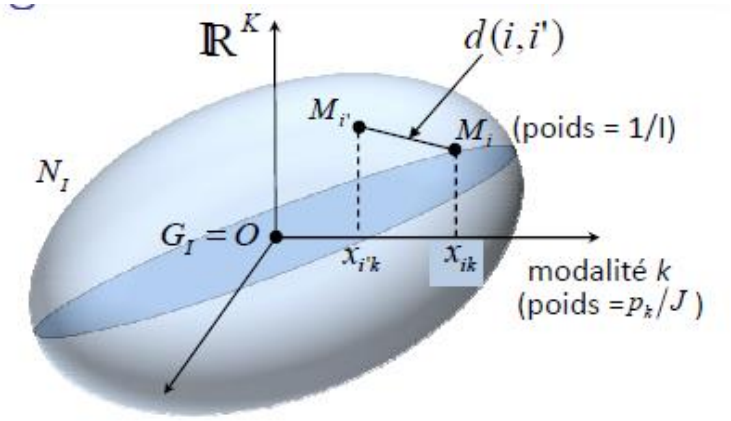
La proximité entre deux individus se mesure avec une distance Euclidienne pondérée.



$$d^2_{i,i'} = \sum_{k=1}^{K} \frac{p_k}{J} (x_{ik} - x_{i'k})^2 = \sum_{k=1}^{K} \frac{p_k}{J} \left(\frac{y_{ik}}{p_k} - \frac{y_{i'k}}{p_k} \right)^2 = \frac{1}{J} \sum_{k=1}^{K} \frac{1}{p_k} (y_{ik} - y_{i'k})^2$$

$$d^2(i, i') = \frac{1}{p} \sum_{k=1}^{K} \frac{1}{p_k} (y_{ik} - y_{i'k})^2$$

- 2 individus prennent les mêmes modalités : distance = 0
- 2 individus ont en commun beaucoup de modalités : distance petite
- 2 individus dont l'un des 2 possède une modalité rare : distance grande
- 2 individus ont en commun une modalité rare : distance petite



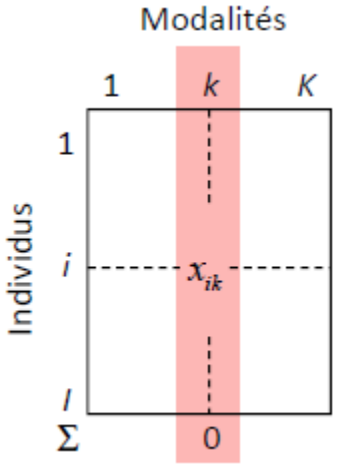
$$d(i, G_I)^2 = \sum_{k=1}^{K} \frac{p_k}{J} (x_{ik})^2 = \sum_{k=1}^{K} \frac{p_k}{J} \left(\frac{y_{ik}}{p_k} - 1 \right)^2$$

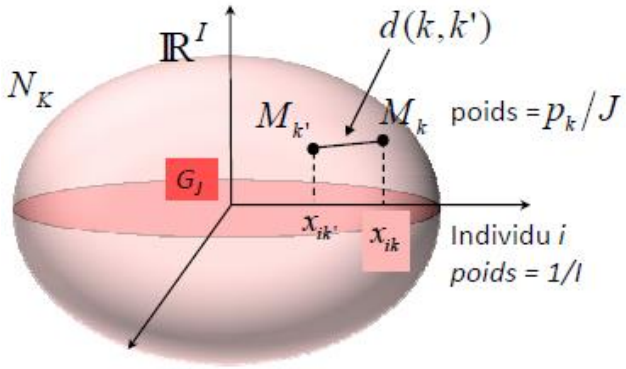
$$d(i, nG_I)^2 = \frac{1}{J} \sum_{k=1}^{K} \frac{y_{ik}}{p_k} - 1$$

L'inertie du nuage est :

$$\text{Inertie}(N_I) = \sum_{i=1}^{I} \underbrace{\frac{1}{I} d^2(i, O)}_{\text{inertie de } i} = \sum_{i=1}^{I} \left(\frac{1}{IJ} \sum_{k=1}^{K} \frac{y_{ik}}{p_k} - \frac{1}{I} \right) = \frac{K}{J} - 1$$

4. Nuage des modalités





La distance entre deux modalités est :

$$d^2(k,k') = \sum_{i=1}^{I} \left(\frac{y_{ik}}{p_k} - \frac{y_{ik'}}{p_{k'}} \right)^2 = \frac{p_k + p_{k'} - 2p_{kk'}}{p_k p_{k'}}$$

La variance d'une modalité est donnée par la formule suivante :

$$Var(k) = d^2(k,O) = \sum_{i=1}^{I} \frac{1}{I} x_{ik}^2 = \sum_{i=1}^{I} \left(\frac{y_{ik}}{p_k} - 1 \right)^2 = \frac{1}{p_k} - 1$$

L'inertie d'une modalité est comme suit :

$$Inertie(k) = \frac{p_k}{J} d^2(k,O) = \frac{1 - p_k}{J}$$

Exemple :

	p_k	1/2	1/5	1/10	1/101
	$d(k,O)$	1	2	3	10
(si $J = 10$)	$Inertie(k)$	0.05	0.08	0.09	0.099

L'inertie d'une variable s'écrit :

$$Inertie(j) = \frac{1}{J} \sum_{k=1}^{K_j} (1 - p_k) = \frac{K_j - 1}{J}$$

On déduit l'inertie totale :

$$Inertie\ totale = \sum_{j=1}^{J} \frac{K_j - 1}{J} = \frac{K}{J} - 1$$

Chapitre 4 :

Les Réseaux

de

Kohonen

Chapitre 4. Les réseaux de Kohonen

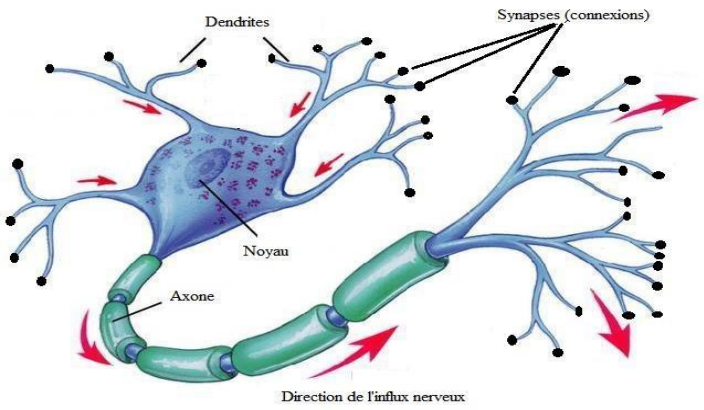
1. Introduction

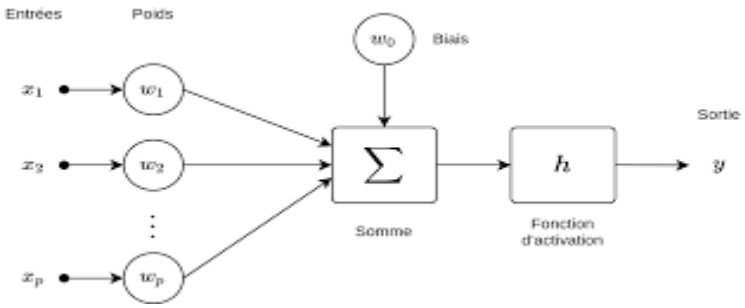
L'organisation et le fonctionnement des neurones biologiques a suscité la création de réseaux de neurones dits « artificiels » ou « formels » (RNF) composés d'un ensemble de neurones reliés entre eux et agissant selon une dynamique qui est différente en fonction du type de réseau. On peut par conséquent établir une correspondance entre un réseau de neurones biologiques et un réseau de neurones artificiels en termes de dynamique et de topologie.

2. Concept de base

2.1. Le neurone formel

Un neurone formel est un processeur élémentaire ayant pour entrée un certain nombre de variables en provenance de neurones antécédents, simulant les dendrites, et pour sortie une réponse par analogie à l'axone alimentant d'autres neurones. Cette sortie est calculée selon une fonction dite de *transfert* (ou fonction d'activation) sur la base des entrées pondérées par des poids.





2.2. Topologie

Les connexions entres neurones décrivent la topologie du réseau. La complexité du réseau biologique ainsi que le nombre important des neurones rendent impossible la transposition

directe en réseau de neurones formels de même taille. Il s'agit donc soit de simuler une partie de la topologie du réseau biologique comme pour les cartes de Kohonen soit de proposer des topologies totalement différentes de la réalité comme c'est le cas des réseaux de neurones cellulaire est une structure bidimensionnelle ou multidimensionnelle composée de processeurs analogiques non linéaires identiques appelés cellules ayant une interaction locale influencée par un ensemble d'opérateurs.

2.3. Apprentissage

L'apprentissage est un processus qui permet à un système de mémoriser une connaissance. Pour un réseau de neurones formels, l'apprentissage consiste à fournir au réseau un ensemble d'informations (ensemble d'apprentissage) en entrées. Le réseau adapte son comportement progressivement en fonction des nouvelles informations présentées suivant un algorithme d'apprentissage pour converger vers un état final. On distingue trois types d'apprentissage.

a. Apprentissage supervisé

Un apprentissage est dit *supervisé* lorsqu'on force le réseau à converger vers un état cible précis en lui présentant des exemples corrects en entrée. Le réseau doit parvenir à une bonne représentation mathématique afin de généraliser ces exemples pour mieux approcher les nouvelles situations.

b. Apprentissage non supervisé

Un apprentissage est dit *non supervisé* lorsqu'on présente des entrées au réseau afin qu'il converge vers un état final quelconque. Cet état est déduit suivant la conception du réseau et des exemples présentés sur ses entrées. Donc nous n'avons, dans ce cas aucune information externe permettant d'influencer cette convergence. Cette forme d'apprentissage est basée sur la notion d'apprentissage compétitif.

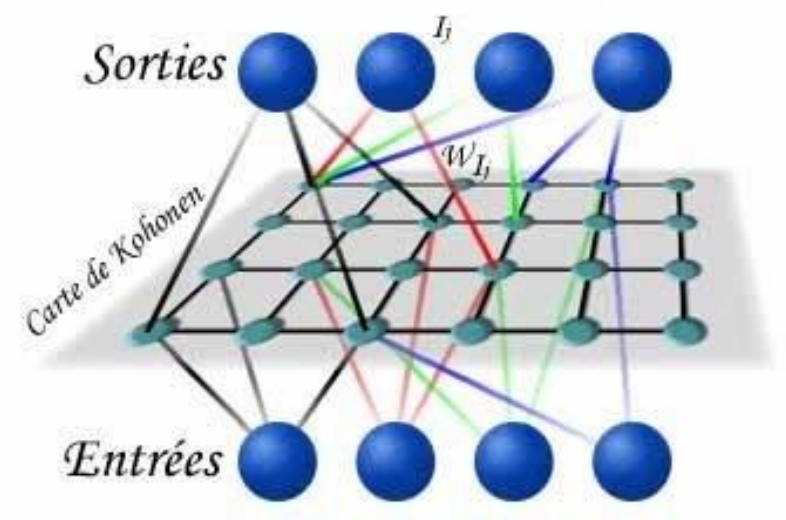
c. Apprentissage par renforcement

L'apprentissage par *renforcement* permet de contourner certaines des limitations de l'apprentissage supervisé. Il consiste en un apprentissage supervisé combiné à un indice de satisfaction, scalaire réel. Lorsqu'une décision prise par le réseau engendre un indice de satisfaction positif, alors la tendance du réseau à prendre cette décision doit être renforcée. Dans le cas contraire, la tendance à prendre cette décision doit être réduite.

3. Théorie de Kohonen

On comprendra par auto-organisation, un procédé capable, bien que non supervisé, de trouver la solution à un problème d'optimisation au sens d'un certain critère. Sur cette base, Kohonen a proposé un réseau de neurones formel connu sous le nom de « cartes auto-organisatrices de Kohonen » (ou Self Organizing Maps en anglais, SOM). Le but d'une carte auto-organisatrice

est de représenter un ensemble de données où chaque neurone se spécialise pour représenter un sous-ensemble bien particulier des données selon les points communs qui les rassemblent. Par conséquent, on obtient une classification en dimension multiple des données. Autrement dit, on peut simuler le fonctionnement d'une carte auto-organisatrice comme une carte topographique pour laquelle l'intensité des liaisons entre les neurones et les stimuli (élément de l'ensemble de données) donneraient l'information de relief. L'intérêt est de savoir, pour un ensemble de stimuli, quels neurones seront les vainqueurs. Suite à l'analyse de ces réponses, nous pouvons déduire un lien entre chaque stimulus et un neurone de la carte. Dans les versions actuelles, le vecteur poids d'un neurone de la carte de Kohonen permet de stocker un motif semblable à une catégorie d'entrées caractérisée par ce neurone après apprentissage.



3.1. Topologie

La carte de Kohonen est en général à deux dimensions (Figure 2.5). Chaque neurone de la carte est relié à tous les neurones de la couche d'entrée. Il existe plusieurs topologies de carte de Kohonen ayant des propriétés différentes : unidimensionnelle(linéaire) ou à deux dimensions (rectangulaire, carré, hexagonale), monocouche ou multicouches. Etant donné que chaque neurone est caractérisé par ses poids et par sa position topologique, cette structure représente un facteur important dans le déroulement de l'algorithme d'apprentissage.

3.2. Le réseau de Kohonen et l'apprentissage

Les cartes auto-organisatrices sont basées sur un apprentissage non supervisé dont le principe clé repose sur l'apprentissage compétitif. Nous exposons dans le paragraphe suivant le principe de ce type d'apprentissage. Contrairement aux autres types d'apprentissage où, généralement, tous les neurones peuvent apprendre simultanément et de la même manière, l'apprentissage compétitif, consiste à créer une compétition ou une concurrence entre les différents neurones de la carte à chaque présentation d'un stimulus afin de déterminer lequel, parmi tous les

neurones, sera le plus actif à un instant donné. Par conséquent, l'apprentissage compétitif produit un *vainqueur* et, parfois, un ensemble de neurones voisins du neurone vainqueur. Seul ce vainqueur et, potentiellement, son voisinage bénéficient d'une modification de leur poids (valeurs portées par ces connexions avec les neurones d'entrée). Donc une petite communauté parmi tous les neurones de la carte sera concernée par l'arrivée d'un stimulus ; l'apprentissage est donc local. L'interaction d'un neurone avec un stimulus est déterminée par une *distance* caractérisant la similarité entre le neurone et ce stimulus. Il y a donc une différenciation des neurones qui se spécialisent pour traiter certains types de stimuli. Ainsi, les neurones individuels peuvent apprendre à se spécialiser sur des sous-ensembles de stimuli similaires pour devenir des détecteurs de caractéristiques.

3.3. L'algorithme d'apprentissage

Liste des variables :

A : Ensemble d'apprentissage
E : Couche d'entrée de taille « K »
C : Carte de Kohonen à deux dimensions de taille Mx N
W : matrice poids des connexions reliant la carte de Kohonen et la couche d'entrée E

Phase 1 : Créer l'ensemble d'apprentissage **A**

Phase 2 : Initialiser la matrice poids **W** (initialisation aléatoire)

Phase 3 : **1/** Affecter un élément de **A** à **E** (choix aléatoire)

2/ Calculer la similarité entre tous les neurones de la carte c_{ij} $(i=1..M \text{ et } j=1..N)$ et E
ex : distance euclidienne

$$v_{ij} = \sum_{k=1}^{K} (w_{ij}^{k} - e^{k})^2$$

3/ Modifier les poids du neurone vainqueur v^* ayant h,l comme indice
ainsi que les poids des neurones voisins comme suit :

$$w_{hl} = w_{hl} + f(w_{hl}, E)$$
$$w_{hl,voisins} = w_{hl,voisins} + g(w_{hl,voisins}, E)$$
$$Remarque : \ f \ et \ g : deux \ fonctions \ décroissante$$

Phase 4 : Aller à « phase 3 »