Classification automatique

Plan du cours

I. Généralités

II. Méthodes hiérarchiques

III. Méthodes de partitionnement

IV. Méthodes mixtes

1 Généralités

Le but de la classification est de regrouper en classes des objets qui ont des caractéristiques communes à l'aide d'algorithmes formalisés, utilisant sur un critère bien défini. La *classification automatique* ou encore *classification non supervisée* permet de construire des groupes ou classes d'objets sans connaissances préalables sur les données à traiter. Il existe deux grandes familles de méthodes de classification : les méthodes hiérarchiques et les méthodes de partitionnement ou non hiérarchiques.

Réaliser une classification sur un ensemble donné E, nécessite de définir une distance (ou plus généralement une dissimilarité entre les éléments de cet ensemble E), mais aussi un *critère d'agrégation* pour regrouper deux parties disjointes de E.

Définition :
- On appelle *dissimilarité* sur E, toute application $\delta : E \times E \to \mathbb{R}_+$ vérifiant les propriétés suivantes:

1) $\forall i, j \in E, \delta(i,j) = \delta(j,i)$

2) $\forall i \in E, \delta(i,i) = 0$

- On appelle *similarité* sur E, toute application $s : E \times E \to \mathbb{R}_+$ vérifiant les propriétés suivantes:

1) $\forall i, j \in E, s(i,j) = s(j,i)$

2) $\forall i, j \in E, s(i,i) \geq s(i,j)$.

- $\mathcal{P}(E)$ étant l'ensemble des parties de E. On appelle *critère ou stratégie d'agrégation* sur E, toute application

$$\begin{array}{rl} d: & \mathcal{P}(E) \times \mathcal{P}(E) \to \mathbb{R}_+ \\ & (A,B) \mapsto d(A,B) \end{array}$$

avec $d(A,B)$ mesurant la dissimilarité entre A et B.

Pour une dissimilarité δ fixée, on peut définir plusieurs stratégies d'agrégation afin de regrouper deux parties disjointes A et B de E.

- ***Stratégie du minimum ou critère des voisins les plus proches (single linkage):***

$$d(A,B) = \min_{i \in A, j \in B} \delta(i,j).$$

Cette stratégie tend à constituer des groupes filiformes avec des effets de chaînage.

- ***Stratégie du maximum ou critère des voisins les plus éloignés (complete linkage):***

$$d(A,B) = \max_{i\in A, j\in B} \delta(i,j).$$

Cette stratégie tend à former des groupes compacts plus identifiables.

- ***Stratégie de la moyenne (average linkage):***

$$d(A,B) = moyenne_{i\in A, j\in B}\delta(i,j).$$

Cette stratégie conduit à des groupes de même variance.

- ***Stratégie de Ward:***
Les objets à classer forment un nuage de points $N = \{x_1, \cdots, x_n\}$ dans un espace euclidien, muni d'une distance d_e. Chaque point x_i est affecté d'un poids $p_i > 0$ de sorte que le poids associé à une partie $A \subset N$ est $p_A = \sum_{x_i\in A} p_i$ et que son centre de gravité est $g_A = \frac{1}{p_A}\sum_{x_i\in A} p_i x_i$. La stratégie de Ward est alors définie par :

$$d(A,B) = \frac{p_A p_B}{p_A + p_B} d_e^2(g_A, g_B).$$

- ***Stratégie du Barycentre:***

$$d(A,B) = d_e^2(g_A, g_B).$$

Le choix de la mesure de dissimilarité dépend du type de données.
Tableau de valeurs numériques: Les individus (ou objets) sont décrits par des variables quantitaives $Y_j, j = 1, \cdots, p$. y_{ij} représente la valeur de la variable Y_j pour l'individu i. Alors la dissimilarité entre deux individus i et i' peut être donnée par la distance euclidienne

$$\delta(i,i') = \sum_{j=1}^{p} (y_{ij} - y_{i'j})^2$$

ou la distance inverse des variances

$$\delta(i,i') = \sum_{j=1}^{p} \frac{1}{\sigma_j^2}(y_{ij} - y_{i'j})^2, \qquad \text{avec } \sigma_j^2 \text{ la variance de } Y_j.$$

Tableau de contingence: Les individus sont décrits par n modalités lignes et p modalités colonnes. k_{ij} est l'effectif conjoint des modalités ligne i et colonne j. La dissimilarité δ peut être choisie ici comme la distance du khi-deux entre deux profils lignes i et i':

$$\delta(i,i') = \sum_{j=1}^{p} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{ij}}{f_{\cdot j}} \right)^2,$$

où $f_{ij} = \frac{k_{ij}}{k}$, $f_{i\cdot} = \frac{k_{i\cdot}}{k}$, $f_{\cdot j} = \frac{k_{\cdot j}}{k}$.
Tableau de présence-absence: Les individus sont décrits par un tableau de variables binaires Y_j tel que l'élément $a_{ij} = 1$ si l'individu i présente la caractéristique associée à la variable Y_j, et $a_{ij} = 0$ sinon. On peut définir plusieurs mesures de dissimilarité en utilisant les nombres suivants :

a=Nombre de caractéristiques communes à i et i'
b=Nombre de caractéristiques possédées par i et non par i'
c=Nombre de caractéristiques possédées par i' et non par i
d=Nombre de caractéristiques possédées ni par i, ni par i'.

Par exemple,

$$\text{Jaccard : } \delta(i,i') = 1 - \frac{a}{a+b+c}.$$

$$\text{Dice : } \delta(i,i') = 1 - \frac{2a}{2a+b+c}.$$

2 Méthodes hiérarchiques : Classification Ascendante Hiérarchique (CAH)

Etant donné un ensemble $E = \{e_1, \cdots, e_n\}$ à partitionner, la méthode CAH repose sur un algorithme agglomératif qui calcule la dissimilarité entre deux parties disjointes à l'aide d'un critère d'agrégation donné, d. L'algorithme est le suivant:

- A l'étape initiale 0, chaque élément e_i forme une classe. La partition initiale est: $P_0 = \{\{e_1\}, \cdots, \{e_n\}\}$.
- A l'étape k, la partition P_k comprend $n - k$ classes. On passe de la partition P_k à la partition P_{k+1} en regroupant les deux classes A et B de la partition P_k les plus proches entre elles au sens du critère d'agrégation d, i.e. tel que

$$d(A, B) = \min\{d(C, D), C \in P_k, D \in P_k\}.$$

- A l'étape finale $n - 1$, il ne reste qu'une seule classe constituant la partition $P_{n-1} = \{E\}$.

Remarque. L'ensemble de ces partitions $P_0, P_1, \cdots, P_{n-1}$ forme une structure appelée hiérarchie de parties de E, i.e un sous-ensemble H de $\mathcal{P}$ tel que :

i) $E \in H$ et $\forall i, e_i \in H$, i.e H contient E et ses singletons.

ii) $\forall A, B \in H,\ A \cap B \in \{A, B, \phi\}$, i.e deux éléments de H sont soit emboîtés pour l'inclusion, soit d'intersection vide.

2.1 CAH sur tableau individus $\times$ variables

On considère un nuage de point-individus $N = \{x_1, \cdots, x_n\}$ dans un espace euclidien, muni d'une distance d_e. Chaque point x_i est affecté d'un poids $p_i > 0$. Le centre de gravité de ce nuage est noté $g = \sum_{i=1}^{n} p_i x_i$. L'inertie totale par rapport au centre de gravité g est donnée par :

$$I = \sum_{i=1}^{n} p_i d_e^2(x_i, g) = \sum_{i=1}^{n} p_i ||x_i - g||^2.$$

A chaque partie A de N, on associe son poids $p_A = \sum_{x_i \in A} p_i$ et son centre de gravité $g_A = \frac{1}{p_A} \sum_{x_i \in A} p_i x_i$.
Soit $P = \{C_1, \cdots, C_k\}$ une partition du nuage N en k classes. Notons:
$p_{C_1}, \cdots, p_{C_k}$: les poids respectifs des classes $C_1, \cdots, C_k$
$g_1, \cdots, g_k$: les centres de gravité respectifs des classes $C_1, \cdots, C_k$
$I_1, \cdots, I_k$: les inerties respectives des classes $C_1, \cdots, C_k$.
Pour tout $j = 1, \cdots, k$, l'inertie de la classe C_j est donnée par

$$I_j = \sum_{x_i \in C_j} p_i d_e^2(x_i, g_j).$$

On appelle inertie intra-classe, la quantité

$$I_W = \sum_{j=1}^{k} p_{C_j} I_j.$$

On appelle inertie inter-classe, la quantité

$$I_B = \sum_{j=1}^{k} p_{C_j} d_e^2(g_j, g).$$

D'après le théorème de Huygens, on a la décomposition suivante:

$$\begin{aligned} I &= I_W + I_B \\ Inertie\ totale &= Inertie\ intraclasse + Inertie\ interclasse \end{aligned}$$

Le processus d'agrégation entraîne une variation opposée des inerties intra-classe et inter-classe. En effet, le regroupement de deux classes disjointes A et B provoque une dimunition de l'inertie inter-classe et une augmentation de l'inertie intra-classe. Cette même variation d'inertie est donnée par:

$$\Delta = p_A d_e^2(g_A, g) + p_B d_e^2(g_B, g) - (p_A + p_B) d_e^2(g_{AB}, g),$$

où $g_{AB} = \frac{p_A g_A + p_B g_B}{p_A + p_B}$ est le centre de gravité de la classe AUB. On peut montrer que

$$\Delta = \frac{p_A p_B}{p_A + p_B} d_e^2(g_A, g_B).$$

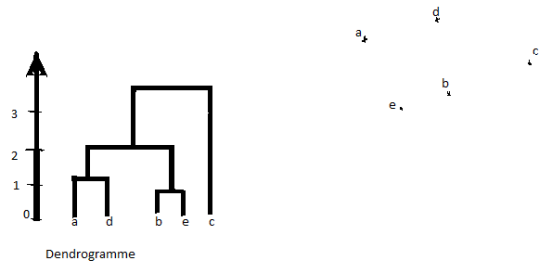
Pour classifier le nuage de points N, la stratégie de Ward consiste à prendre comme critère d'agrégation la variation d'inertie Δ, appelée *distance de Ward*. Dans la pratique, on agrégera les deux classes qui provoquent la plus petite diminution de l'inertie inter-classe, i.e. la plus petite augmentation de l'inertie intra-classe.

Qualité d'une partition.
L'algorithme de CAH conduit à une seule classe finale qui n'est pas l'objectif visé. Pour obtenir la meilleure partition (ou typologie des éléments à classer), on doit évaluer la qualité des partitions. En se basant sur la décomposition de l'inertie totale, la qualité d'une partition est définie par le rapport

$$Q = \frac{I_B}{I} = 1 - \frac{I_W}{I}.$$

A chaque étape de l'algorithme, on calculera la qualité de la partition obtenue et on choisira comme partition finale, celle obtenue à l'étape précédant une forte baisse de cette qualité, i.e. l'étape précédant la fusion de deux classes très éloignées. La meilleure partition est celle pour laquelle, il y a une bonne séparation des classes et une faible dispersion à l'intérieur des classes.

Arbre de classification.
Exemple : Soit $E = \{a, b, c, d, e\}$ un ensemble dont les éléments sont disposés dans le plan comme suit :
L'arbre de classification à gauche est obtenu en considérant la distance euclidienne qui sépare les points et le



critère d'agrégation du barycentre. On a au total 5 partitions :
$P_0 = \{\{a\}, \{b\}, \{c\}, \{d\}, \{e\}\}$

$P_1 = \{\{a\}, \{b, e\}, \{c\}, \{d\}, \}$
$P_2 = \{\{a, d\}, \{b, e\}, \{c\}\}$
$P_3 = \{\{a, d, b, e\}, \{c\}\}$
$P_4 = \{\{a, b, c, d, e\}$.
Remarque. Dans le dendrogramme, les éléments sont placés côte à côte en fonction de la distance qui les sépare et qui doit être minimale. La hauteur d'une branche est proportionnelle à la distance qui sépare les deux objets regroupés pour former cette branche.

Coupure de l'arbre.
L'arbre de classification (ou dendrogramme) indique l'ordre dans lequel les agrégations successives sont opérées, en précisant les valeurs (ou indices) correspondant aux différents niveaux d'agrégation. On choisit la partition finale en coupant l'arbre lorsqu'on observe un saut important sur ces indices ou niveaux d'agrégation. Les branches obtenues après coupure forment la partition finale ; chaque branche constituant une classe.

CAH sur les variables

On peut aussi réaliser une CAH sur les variables en les regroupant de sorte que les variables appartenant à une même classe soient fortement corrélées et les variables appartenant à des classes différentes soient faiblement corrélées. Cela permet de réduire le nombre de variables de l'étude en identifiant les éventuelles redondances. La proximité entre deux variables X_i et X_j est mesurée par la valeur absolue de leur coefficient de corrélation linéaire, $|r(X_i, X_j)|$. Donc une mesure de dissimilarité entre deux variables X_i et X_j est donnée par $1 - |r(X_i, X_j)|$. Pour regrouper deux ensembles de variables A et B, on peut utiliser la stratégie suivante :

$$d(A, B) = \max_{X_i \in A, X_j \in B} (1 - |r(X_i, X_j)|).$$

Cette stratégie d'agrégation correspond au critère des voisins les plus éloignés et permet d'obtenir des classes homogènes avec des variables suffisamment corrélées.

2.2 CAH sur tableau de dissimilarité

On considère un tableau de dissimilarités ou de similarités entre n objets. L'algorithme de CAH calcule la dissimilarité entre deux sous-ensembles disjoints de l'ensemble des n objets en utilisant une stratégie d'agrégation (voir section 1) suivant l'objectif visé. Lorsqu'on veut obtenir une typologie des n objets avec des classes bien identifiées, on utilisera le critère d'agrégation des voisins les plus éloignés. Par contre, lorsque l'on veut construire une hierarchie ou **arbre de longueur minimale**, on utilise le critère des voisins les plus proches. En effet, le critère des voisins les plus proches pourrait entraîner des effets de chaîne, ce qui implique que deux objets très éloignés, mais reliés par une série de points proches, peuvent se retrouver dans une même classe.
L'arbre de longueur minimale peut être utilisé pour déterminer le chemin le plus court si les n objets constituent des localités à relier. Il est aussi utilisé pour déceler de façon rapide la structure de corrélation d'une matrice contenant un grand nombre de variables.

3 Méthodes de partitionnement

On dispose d'un ensemble d'individus décrits par des variables numériques. On se propose de construire une partition de cet ensemble en k classes. Contrairement aux méthodes hiérarchiques, ici le nombre de classes k est fixé du début à la fin de l'algorithme. Il existe plusieurs techniques de partitionnement d'un ensemble en k classes disjointes.

3.1 Méthode des centes mobiles (Forgy, 1965)

En choisissant initialement k points comme centres parmi l'ensemble des individus, l'algorithme des centres mobiles permet de construire une suite de partitions en k classes, $(\{C_1^m, \cdots, C_k^m\}, m \geq 1)$ en faisant décroitre l'inertie intra-classe. La qualité d'une partition étant donnée par $Q = 1 - I_W/I$. Plus l'inertie intra-classe est faible , meilleure est la partition.

Algorithme de la méthode des centres mobiles:

Soit $N = \{x_1, \cdots, x_n\}$ un nuage d'individus à classer.

Etape 0: On choisit au hasard k centres parmi les n points du nuage, notés $c_1, \cdots, c_k$. On regroupe autour de chaque centre c_i les points x_j qui lui sont plus proches pour obtenir une première partition

$$P_0 = \{C_1^0, \cdots, C_k^0\}.$$

On détermine ensuite les centres de gravité de ces classes $\{C_1^0, \cdots, C_k^0\}$ que l'on note $\{g_1^0, \cdots, g_k^0\}$.

Etape 1: On considère les centres de gravité $\{g_1^0, \cdots, g_k^0\}$ comme les nouveaux centres de classe et on regroupe autour de ces centres de gravité les points x_j les plus proches pour former une nouvelle partition en k classes

$$P_1 = \{C_1^1, \cdots, C_k^1\}.$$

On détermine ensuite les centres de gravité $\{g_1^1, \cdots, g_k^1\}$ de ces nouvelles classes $\{C_1^1, \cdots, C_k^1\}$ qui deviennent les nouveaux centres de classe pour l'étape suivante.

Etape m: On regroupe autour des centres de gravité $\{g_1^{m-1}, \cdots, g_k^{m-1}\}$ obtenues à l'étape précédente les points x_j qui leur sont plus proches pour former la partition

$$P_m = \{C_1^m, \cdots, C_k^m\}.$$

On détermine les centres de gravité de ces nouvelles classes, notés $\{g_1^m, \cdots, g_k^m\}$; puis on itère le processus.

Etape finale: L'algorithme s'arrête lorsque deux itérations successives conduisent à une faible décroissance de l'inertie intra-classe ou à sa stabilisation.

NB: L'algorithme des centres mobiles est basé sur le résultat suivant:

Théorème 1 *L'inertie intra-classe I_W diminue aucours de l'algorithme des centres mobiles.*

Preuve: On utilise le théorème de Huygens qui, pour une classe donnée C_i de centre de gravité g_i et d'effectif n_i, s'écrit comme suit:

$$\forall a \in \mathbb{R}^p, \quad \sum_{x_j \in C_i} d^2(x_j, a) = \sum_{x_j \in C_i} d^2(x_j, g_i) + n_i d^2(a, g_i) \tag{1}$$

qui veut dire que le centre de gravité g_i d'une classe est le point le plus proche, au sens des moindres carrés, de l'ensemble des points de cette classe.

A l'étape m, on a k classes $\{C_1^m, \cdots, C_k^m\}$ de centres de gravité respectifs $\{g_1^m, \cdots, g_k^m\}$. Donc l'inertie intra-classe à l'étape m est

$$I_W^m = \sum_{i=1}^{k} \sum_{x_j \in C_i^m} d^2(x_j, g_i^m).$$

Soit g_i^{m+1} le centre de gravité de la classe C_i^{m+1} à l'étape $m+1$, alors en combinant (1) et le fait que si un point quitte une classe sa distance avec le centre de gravité de sa nouvelle classe est plus petite que celle avec le centre de gravité de son ancienne classe, on obtient :

$$\begin{aligned} I_W^m = \sum_{i=1}^{k} \sum_{x_j \in C_i^m} d^2(x_j, g_i^m) &\geq \sum_{i=1}^{k} \sum_{x_j \in C_i^{m+1}} d^2(x_j, g_i^m) \\ &\geq \sum_{i=1}^{k} \sum_{x_j \in C_i^{m+1}} d^2(x_j, g_i^{m+1}) \\ &\geq I_W^{m+1}. \qquad \text{cqfd!} \end{aligned}$$

3.2 Méthode des nuées dynamiques (Diday, 1971)

Elle généralise la méthode des centres mobiles et, est bien adaptée aux gros volumes de données. Au lieu des centres ponctuels, cette méthode choisit au hasard des groupes d'individus représentatifs des classes appelés noyaux. Le principe de l'algorithme est le même que celui des centres mobiles. On peut cependant rechercher des **"formes fortes"** c'est à dire des groupes d'éléments qui appartiennent à une même classe dans toutes les partitions finales. Par exemple, si on a deux patrtitions finales, les formes fortes sont les classes d'effectif non nul dans la partition produit.

Remarque. La détermination des formes fortes ne fournit pas généralement une partition intéressante, mais elle peut suggérer le nombre de classes k à retenir, qui peut être considéré comme le nombre de formes fortes d'effectif important.

3.3 Méthode des $k - means$

C'est une amélioration de la méthode des centres mobiles. L'algorithme des $k - means$ choisit au hasard des centres ponctuels comme celui des centres mobiles, mais la règle de calcul des nouveaux centres n'est pas la même. A chaque réaffectation d'un individu, on modifie la position du centre correspondant plutôt que d'attendre la réaffectation de tous les individus de la classe. Ainsi, en une seule itération, cette méthode peut donner une partition de bonne qualité.

4 Méthodes mixtes

Lorsque les données à classer sont très nombreuses, les méthodes hiérarchiques ne sont pas directement applicables. La méthode de partitionnement des centres mobiles est adaptée à ces données volumineuses, mais présente l'inconvénient de produire une partition finale qui dépend du choix initial des centres de classe et, de fixer le nombre de classes *a priori*. Il convient alors de combiner ces deux types de méthodes pour aboutir à une partition finale satisfaisante. L'algorithme de la classification mixte procède en trois étapes :

- Etape 1: On utilise par exemple la méthode des centres mobiles pour obtenir q (quelques dizaines ou centaines) formes fortes.
- Etape 2: On applique la CAH en considérant comme partition initiale, celle constituée par les q formes fortes obtenues à l'étape 1. on obtient alors une partition en k classes, avec $k < q$.
- Etape 3: On consolide la partition en k classes, obtenue à l'étape 2, en appliquant de nouveau la méthode des centres mobiles pour aboutir à une partition finale en k classes de meilleure qualité.

4.1 Description statistique des classes

Après avoir regroupé les individus en classes, on peut vouloir caractériser ces classes à l'aide des variables.

4.1.1 Caractérisation par des variables illustratives

Les variables illustratives sont des variables qui n'ont pas contribué à la construction des classes, mais que l'on peut utiliser *a posteriori* pour identifier et caractériser les classes obtenues.

a. Cas d'une variable nominale:

Soit N une variable aléatoire représentant le nombre d'individus de la classe C_k possédant la modalité "j" de la variable nominale considérée. Sous l'hypothèse H_0 selon laquelle les n_k individus de la classe C_k sont tirés au hasard et sans remise dans l'ensemble des individus constituant l'échantillon global, N suit une loi

hypergéométrique de moyenne $E(N) = n_k \frac{n_j}{n}$ et de variance

$$s_k^2(N) = \frac{n - n_k}{n - 1} \frac{s^2(N)}{n_k},$$

où $s^2(N) = \frac{n_j}{n}(1 - \frac{n_j}{n})$ est la variance de N dans l'échantillon global de taille n, n_k est l'effectif de la classe C_k, n_j est le nombre total d'individus possédant la modalité "j", et n_{kj} le nombre d'individus de la classe C_k possédant la modalité "j".
Pour n, n_k, n_{kj} assez grands, la variable

$$t_k(N) = \frac{N - E(N)}{s_k(N)} \sim N(0, 1).$$

$t_k(N)$ est appelée valeur-test. Elle mesure l'écart entre la proportion d'individus possédant la modalité "j" dans la classe C_k et la proportion d'individus possédant la modalité "j" dans l'échantillon global en termes d'écart type d'une loi normale. On pourra calculer la probabilité $p_k(j) = \mathbb{P}(|N(0, 1)| > t_k(n_{kj}))$ ou comparer la valeur absolue de $t_k(n_{kj})$ à 2 (propriété de la loi normale). Plus la probabilité $p_k(j)$ est faible, plus la modalité "j" est caractéristique de la classe C_k, et donc $|t_k(n_{kj})| > 2$.
b. Cas d'une variable continue:
Sous la même hypothèse H_0 que les n_k individus de la classe C_k sont tirés au hasard et sans remise, la variable $\bar{X}_k$ représentant la moyenne de la variable considérée X dans la classe C_k a pour espérance $\bar{X}$ et pour variance

$$s_k^2(X) = \frac{n - n_k}{n - 1} \frac{s^2(X)}{n_k}, \qquad s^2(X) \text{ est la variance globale de } X.$$

La valeur-test est ici

$$t_k(X) = \frac{\bar{X}_k - \bar{X}}{s_k(X)} \sim N(0, 1).$$

Plus cette valeur-test est grande, plus la moyenne de X dans la classe C_k est différente de la moyenne globale, et donc plus cette variable est carartéristique de la classe C_k.
Remarque. On peut ranger les variables par valeurs-test décroissantes et ne retenir que les variables les plus significatives ; ce qui permet de caractériser rapidement les classes.

4.1.2 Caractérisation par des variables actives

Comme les variables actives ne sont pas indépendantes des classes, car ayant participé à leur construction, on ne peut baser un test statistique rigoureux sur ces valeurs-test. Néanmoins, on peut les utiliser pour classer les variables actives en vue de caractériser les classes. Les valeurs absolues de ces valeurs-test constituent alors de simples mesures de similarité entre variables et entre classes.

4.1.3 Complémentarité entre Analyse factorielle et Classification

Pour décrire un ensemble de données volumineux, il est important de faire une utilisation conjointe de l'analyse factorielle et de la classification. Cette démarche peut se résumer en les étapes suivantes:

1) Analyse factorielle: résume les données sur un plan factoriel avec une bonne description ;

2) Classification à partir des premiers facteurs: élimine les fluctuations aléatoires dues à la variance des autres facteurs, et permet d'avoir des classes plus homogènes ;

3) Description des classes : On calcule les valeurs-test pour les variables afin de comparer les moyennes (ou fréquences) des classes aux moyennes (ou fréquences) gobales ;

4) Positionnement des classes dans le plan factoriel: Les centres de gravité des classes sont projetés au sein des variables ou modalités actives dans le premier plan factoriel. La position de chaque individu repéré par le numéro de sa classe permet de représenter la densité et la dispersion des classes.