Chapitre 6 : Classification non-supervisée

M1 du Master MMAS

James Ledoux

Dépt de mathématiques, Univ. Poitiers

12 juillet 2009

Classification automatique

Applications : Compression de données
Reconnaissance des formes, Traitement d'images
Biologie, ...

- **Segmentation d'images :** décomposer une image donnée en « segments », i.e. en régions contenant des pixels similaires

  Exemple : les segments peuvent être des régions de l'image décrivant un objet donné

  Les résultats de la segmentation sont utilisés pour aider à la détection de contour et la reconnaissance des formes.

- **En biologie :** Identification de classes de tumeurs à partir de données d'expression de gènes ; Classification de protéines : obtenir des groupes de protéines similaires, ...

Des méthodes de classification automatique

1. **Méthodes de ré-allocation**
   (a) Le nombre de classes est fixé :
   - agrégation autour des centres mobiles
   - nuées dynamiques
   - k-moyennes

   (b) Le nombre de classes n'est plus fixé :
   - ISODATA
   - Paramétrage de Wishart
2. **Classification hiérarchique**
   - méthodes ascendantes (CAH)

- **Données :**
  I un ensemble d'individus, d'objets, ...
- **Objectif :**
  Définir une typologie de l'ensemble des individus
- **Sorties :**
  une partition de l'ensemble I
  + définition d'éléments types représentatifs de chaque classe

Remarque : Les sorties d'un algorithme de classification automatique peuvent servir d'entrées pour une méthode de classement ou classification supervisée

- Recherche de la partition de I **optimale** / à la fonction objectif
  i.e. la **meilleure** partition
- L'ensemble I à classer est fini (n éléments)
  ➥ nombre fini de partitions possibles
- **Nombre de partitions $P_{n,k}$ en k classes pour n éléments**

$$P_{n,k} = P_{n-1,k-1} + kP_{n-1,k} \quad \text{(nombre de Stirling de 2ème espèce)}$$

et on peut montrer que

$$P_{n,k} \underset{n\to+\infty}{\sim} \frac{k^2}{k!}$$

Nombre total de partitions P_n (nombre de Bell) :

$$P_n = \sum_{k=0}^{n-1} \binom{k}{n-1} P_{n,k} = \frac{1}{e}\sum_{k=1}^{\infty} \frac{k^n}{k!}$$

Traitement de 1 million de partitions par seconde ➡ 126000 années pour P_{25}

Paramètres d'une classification automatique

- Une distance d (dans toute la suite) ou un indice de similarité s (ou de dissimilarité) :

$$s(i,j) = s(j,i), \quad s(i,j) \geq 0, \quad s(i,i) \geq s(i,j)$$

- La forme des éléments types recherchés
- La connaissance a priori ou non du nombre de classes ou regroupements
- Une configuration/partition initiale
- Une fonction objectif guidant la classification

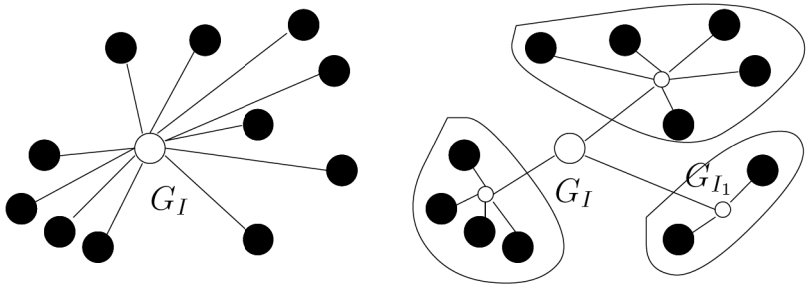
- $\mathcal{P}$ une partition de I en q classes $\{I_1, \ldots, I_q\}$ d'une famille de n points
- Chaque classe I_k :
  - poids $= m_k = \sum_{i\in I_k} p_i$
  - centre de gravité $G_{I_k} = \frac{1}{m_k}\sum_{i\in I_k} p_i\, x_i$

Rappel : le Th 2 (voir aussi Th 1) donnait une décomposition de la variance connue aussi comme le théorème de König-Huyghens

Proposition 9 (Théorème de Huyghens)

Pour un nuage $\mathcal{N}(I)$ de n points $x_1, \ldots, x_n$ de poids respectifs $p_1, \ldots, p_n$:

$$\underbrace{\textit{Inertie}(\mathcal{N}(I))}_{\textit{Inertie totale}} = \sum_i p_i\, d^2(x_i, G_I) = \underbrace{\sum_{k=1}^{q} m_k\, d^2(G_I, G_{I_k})}_{\textit{Inertie inter-classe}} + \underbrace{\sum_{k=1}^{q}\sum_{i\in I_k} p_i\, d^2(x_i, G_{I_k})}_{\textit{Inertie intra-classe}}$$



Inertie totale = Inertie inter-classe + Inertie intra-classe

FIGURE 30: Décomposition de l'inertie selon la relation de Huyghens

Notons que

Maximiser l'inertie inter-classe $\Longleftrightarrow$ Minimiser l'inertie intra-classe

Agrégation autour des centres mobiles

- Formalisme très limité
- Dans la pratique, très efficace
  - ➥ Partionnement de **grands** recueils de données
- Histoire : Forgy (1965), Mac Queen (1967) Ball et Hall (1967), ...
- Cas particulier de la technique dite des nuées dynamiques (Diday (1971))

- **Données :** tableau $n \times p$
  p variables $X_1, X_2, \ldots, X_p$ sont observées sur n individus
- L'espace $\mathbb{R}^p$ supportant les n points-individus est muni d'une distance d (euclidienne usuelle, distance du χ^2)
- Constitution d'au plus q classes

Etape 0 : sélection de q centres provisoires de classes

$$C_0 = \{C_1^0, \ldots, C_k^0, \ldots, C_q^0\}$$

➥ $\mathcal{P}_0 = \{I_1^0, \ldots, I_k^0, \ldots, I_q^0\}$ partition de I

avec | $i \in I_k^0$ s'il est plus proche de C_k^0 que de tous les autres centres |

Principe dans le plan avec $q = 2$

FIGURE 31: Tirage des centres C_1^0 et C_2^0

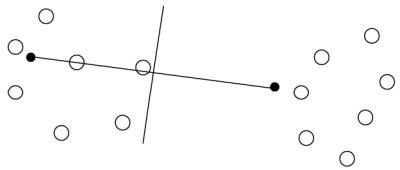


FIGURE 32: Nouveaux centres C_1^1 et C_2^1 et nouvelles classes I_1^1 et I_2^1

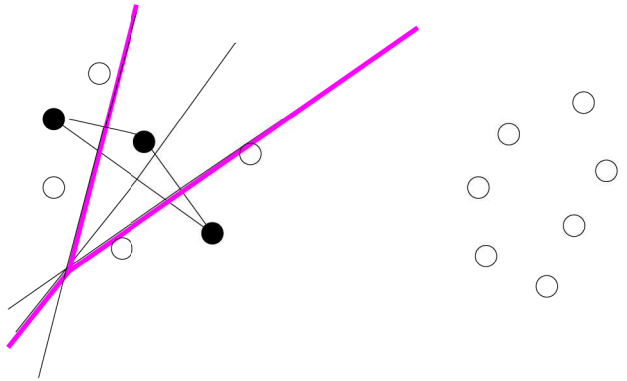


FIGURE 33: Les trois médiatrices du triangle définissent trois régions par : un point appartient à I_i^k s'il est plus proche du centre C_i^k que de tous les autres centres

Etape 1 :

- Les centres de gravité des classes de $\mathcal{P}_0$ déterminent q nouveaux centres

$$\{C_1^1, \ldots, C_k^1, \ldots, C_q^1\}$$
$$\hookrightarrow \mathcal{P}_1 = \{I_1^1, \ldots, I_k^1, \ldots, I_q^1\}$$

- Partition de I selon la même règle que pour $\mathcal{P}_0$

$$\vdots$$

Etape m :

- Les centres de gravité des classes de $\mathcal{P}_{m-1}$ déterminent q nouveaux centres

$$\{C_1^m, \ldots, C_k^m, \ldots, C_q^m\}$$
$$\hookrightarrow \mathcal{P}_m = \{I_1^m, \ldots, I_k^m, \ldots, I_q^m\}$$

- Partition de I

Proposition 10 (Convergence de l'algorithme)

Si on choisit le critère du maximum d'inter-interclasse, l'algorithme des centres mobiles s'arrête ou encore, est il est dit ***convergent****.*

Autrement dit, l'algorithme s'arrête lorsque :

- soit deux itérations successives donnent la même partition
- soit plus de gain / au critère d'inertie inter-classe maximale

- La preuve repose sur l'idée que chaque étape fait décroître strictement l'inertie intra-classe lorsqu'on remplace le centre de classe déterminant une nouvelle classe crée par son centre de gravité.
- La partition finale dépend du choix initial des centres : optimum local

Méthode de nuées dynamiques de Diday

Etape 1 :

Définir un **mode de représentation** d'un groupe d'individus i.e un élément caractéristique d'une classe

Ex : dans les centres mobiles, chaque classe est caractérisée par un point (le centre de gravité)

Quelques caractérisations d'une classe :

- un sous-ensemble de points (les plus centraux)
- un axe principal du nuage associé à la classe de points
- un plan factoriel
- une distribution de probabilité

Elément caractéristique appelé un **noyau**

Etape 2 :

Utiliser un algorithme de réallocation des points de l'ensemble I aux noyaux

Ex : chaque point est affecté au noyau dont il est le plus proche

Principe général

Alternance des deux étapes

- affectation $\rightarrow$ partition en k classes
- calcul des noyaux

Jusqu'à convergence pour la fonction objectif choisie

Problème : choix des noyaux initiaux

Technique des k-means (Mac Queen 1967)

- **Objectif :** partition de I en k (fixé) classes
- **Caractéristique :** stratégie mixte de ré-allocation et de calcul des centres
  1. Les k premiers individus de l'échantillon constituent les k centres initiaux
  2. Affectation des $(n-k)$ individus restants selon le centre le plus proche
     À chaque affectation d'un individu dans une classe selon la règle du plus proche centre
     recalcul du centre de gravité de la classe à laquelle l'individu vient d'être affecté
  3. Lorsque tous les individus sont affectés, la liste des centres constitue la liste des nouveaux centres provisoires qui sont utilisés pour une nouvelle itération avec la règle d'affectation au plus proche centre pour chaque individu

- **Économique** :
  choix des k premiers individus : configuration initiale « gratuite »
  $+$ nombre minimal d'itérations si on accepte la partition obtenue sans recherche de convergence
  $+$ « bonne » qualité de la partition obtenue
- **La partition dépend du choix initial** mais le faible coût de l'algorithme permet l'essai de plusieurs ensembles de centres initiaux
- La composition des classes issues de l'étape 2 est influencée par l'ordre d'apparition des individus dans l'échantillon

Méthode convergente des k-moyennes

1. Configuration initiale $=$ une **partition** de l'ensemble en k classes (Ex. issue des étapes 1,2 de la méthode originale des k-means)
2. Calcul de la distance de chaque individu à chacun des centres de gravité $G_1, \ldots, G_k$ initiaux
   Si le centre le plus proche d'un individu n'est pas celui de sa classe actuelle **alors** l'individu **change de classe** au bénéfice de celle dont le centre est le plus proche
   Les centres de gravité des deux classes affectées par ce changement sont recalculés
3. L'étape 2 est itérée jusqu'à convergence complète c'est à dire l'exploration complète de la population ne donne aucun changement de classe

Méthode convergente économique : nbre d'itérations souvent moins que le nombre de points. Cependant, Arthur et Vassilvitskii (06) ont montré récemment qu'il existait des jeux de données pour lesquels le nombre d'itérations des k-means pour converger était super-polynomial : $2^{\Omega(\sqrt{n})}$ où $\Omega(\sqrt{n})$ est une certaine fonction qui croît au moins linéairement avec $\sqrt{n}$.

- Dépendance aux conditions initiales : optima locaux
  - « Réponse » pratique : **formes fortes**
    - Effectuer plusieurs partitions avec plusieurs ensembles de centres initiaux
    - Regrouper les individus qui ont été affectés à une même classe dans chacune des partitions finales (partition produit)
      ➥ groupements stables ou formes fortes
  - Problème : faiblesse des effectifs des groupements stables
    $+$ aucune garantie théorique d'un gain

Exemple 15 (Exemple de calcul des formes fortes)

Première partition

113	38	35	40

Seconde partition

30
43
40

5	25	0
30	8	5
3	2	35

Partition produit

Ensemble de 113 individus partionnés en 3 classes

Méthode ISODATA (Ball et Hall, . . .)

- **Introduction d'un « calcul » du nombre de classes** :
  ➥ structure de la partition reflète au mieux la structure des données

  Appel à des démarches **heuristiques**
- **Caractéristique :** raffinement de la technique de ré-allocation dans la méthode des centres mobiles avec une gestion dynamique du nombre de classes

  Mais l'**objectif** reste une partition réalisant le

  minimum de la variance intra-classe

Etape 1 : Initialisation de 7 paramètres

Viter : nombre maximum d'étapes de type agrégation autour des centres mobiles

Nbiter : nombre maximum d'itérations dans la recherche de convergence de la méthode globale

Nbcl : paramètre permettant le contrôle du nombre de classes générées

Nbelt : effectif minimum de chaque classe

Paramètres de phases de fractionnement et d'agrégation :

Sfrac : valeur seuil pour la dispersion pour le fractionnement d'une classe

Sagreg : valeur seuil pour l'agrégation de deux classes

Nbagreg : nombre maximum de fusions admissibles lors de la phase de regroupement

+ Génération d'une séquence de noyaux (méthode qcq)

Etape 2

- Chacun des individus est affecté au noyau le plus proche
- Chaque noyau reste fixé tout au long du processus d'affectation
- Un nouvel ensemble de noyaux est constitué par calcul des centres de gravité des classes précédemment construites

Itération du cycle de réallocation / calcul des centres jusqu'à convergence (configuration stable) ou jusqu'à atteindre **Viter**

Etape 3

On néglige toute classe dont l'effectif est $<$ **Nbelt**

Les individus correspondants sont abandonnés pour le reste de l'analyse

Etape 4 : Cycle de regroupement ou de fractionnement des classes

- phase d'agrégation si le nbre courant de classes est ≥ 2 **Nbcl**
- phase de fractionnement si nbre courant de classes est $<$ **Nbcl**$/2$
- Dans tous les autres cas, on alterne les deux procédures de regroupement et de fractionnement

Etape 5

Les centres des classes ainsi établies sont calculés
➥ nouveau jeu de noyaux

Cycle de ré-allocation du type Etape 2 avec cette nouvelle famille de noyaux

Etape 6 : Itération des étapes 2,3,4,5 jusqu'à convergence (obtention d'une situation stable) ou jusqu'à atteindre **Nbiter**

Phase d'agrégation : Itération au plus **Nbagreg** fois de

1. Calcul de toutes les distances $d(G_i, G_j)$ avec G_i, G_j centres de gravité des classes courantes
2. pour i, j tel que $\min_{i,j} d(G_i, G_j) <$ **Sagreg**
   fusion des deux classes
   calcul du centre
   et Retour en 1

   Remarque : Si les centres sont suffisamment éloignés alors aucune fusion lors d'une itération

Phase de fractionnement

- Une classe est sujette à un fractionnement en deux classes si l'écart-type intra-classe pour une des variables de description est $>$ **Sfrac**
- Les éléments d'une telle classe sont affectés aux deux nouvelles classes de manière à fractionner la classe originelle dans la **direction de plus grande dispersion** : la valeur prise par la variable de plus fort écart-type (ou variable discriminante de la séparation) sur chaque individu est comparée à la moyenne de ces valeurs
  La moyenne constitue donc la frontière entre les deux sous-classes crées
- Calcul des centres des deux nouvelles classes
  + calcul de la distance $d_{1,2}$ séparant ces deux classes
- Fractionnement est accepté si $d_{1,2} \geq 1.1 *$ **Sagreg**

Remarques

- Algorithme optimise la variance intra-classe des partitions construites successivement
- L'étape 3 a pour effet, à chaque itération, d'éliminer les valeurs aberrantes
- Problème d'arrêt général de l'algorithme : aucune règle convaincante
  ➥ contrôle interactif du déroulement de l'algorithme
- Algorithme le plus coûteux des algorithmes de réallocation

Paramétrage de Wishart permet de surmonter le nombre fixé k de classes pour la méthode (convergente) des k-moyennes.

Classification hiérarchique (ascendante)

- **Principe** : création d'une séquence de partitions de I obtenues par regroupements successifs de classes
  i.e. chaque partition se déduit de la précédente par un regroupement de classes

  ➥ une hiérarchie de partitions de I
- **Représentation graphique de la séquence**

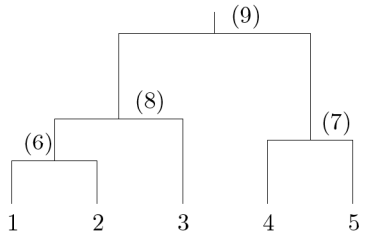


FIGURE 34: un arbre ou dendrogramme

Remarques :

- Le nombre de classes n'est plus fixé.
- Chaque coupure de l'arbre fournit une partition de I :

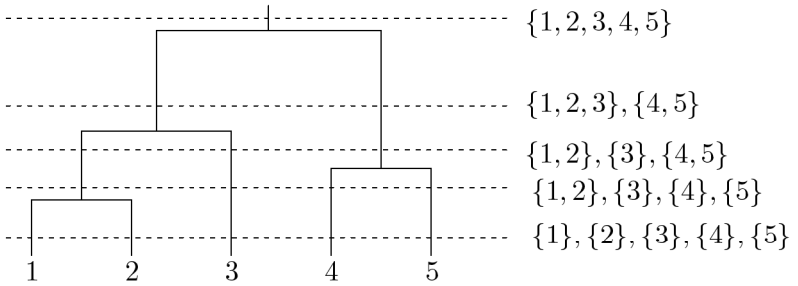


FIGURE 35: Liste des partitions

- Une famille $\mathcal{P}$ de parties de I forme une hiérarchie de I si

$$I \in \mathcal{P}, \qquad (i \in I \Longrightarrow \{i\} \in \mathcal{P})$$

$$\left.\begin{array}{l} P_1 \in \mathcal{P} \\ P_2 \in \mathcal{P} \end{array}\right| \Longrightarrow \left|\begin{array}{l} \text{soit } P_1 \cap P_2 = \emptyset \\ \text{soit } ((P_1 \subseteq P_2) \vee (P_2 \subseteq P_1)) \end{array}\right.$$

- Les individus sont les éléments terminaux de l'arbre (ou de la hiérarchie)
- À chaque étape, deux classes d'une partition sont regroupées pour donner une nouvelle classe.
  Dans la représentation sous forme d'arbre, le nœud matérialisant la nouvelle classe correspond au regroupement de deux nœuds appelés aîné et benjamin
- Une hiérarchie est dite indicée s'il existe une fonction $i : \mathcal{P} \to \mathbb{R}^+$ compatible avec la relation d'inclusion :

$$P_1 \subset P_2 \Longrightarrow i(P_1) < i(P_2)$$

- Les valeurs de la fonction indice sont souvent appelées les niveaux d'agrégation
  car $i(P)$ sera le niveau auquel on trouve agrégé pour la **première fois** tous les individus constituant la partie P
- **Problème :** la construction algorithmique d'une hiérarchie indicée ne doit pas engendrer **d'inversion** dans l'arbre :

x et y sont réunis avant z et w par l'algorithme
et $i(\{x, y\}) > i(\{z, w\})$!

Choix du critère d'agrégation

Critères d'agrégation

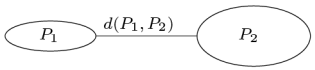
1. On suppose donné une distance d permettant de mesurer la distance entre individus
   - distance euclidienne usuelle
   - distance du chi-deux, ...
2. Définition, à partir de d, d'une distance entre un individu et un groupe d'individus
3. Définition d'une distance entre deux groupes ou classes

➥ Règles de calcul des distances entre classes disjointes

Quelques critères classiques d'agrégation :

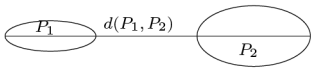
1. Distance du saut minimal (single linkage)

$$d_{\text{min}}((x, y); z) = \min(d(x, z); d(y, z))$$



2. Distance du saut maximal (diamètre)

$$d_{\text{max}}((x, y); z) = \max(d(x, z); d(y, z))$$



3. Distance moyenne

$$d_{\text{moy}}((x, y); z) = \frac{d(x, z) + d(y, z)}{2}$$

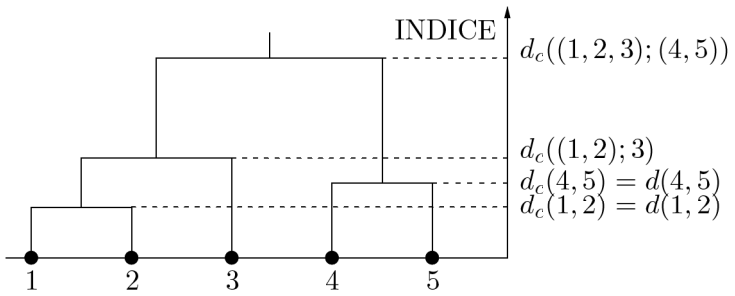
4. Distance de Ward

$$d_W((x, y), z) = \frac{(p_x + p_y)p_z}{(p_x + p_y) + p_z} d^2(G_{(x,y)}, z)$$

En général, on pose

$$i(\{x, y, z\}) = d_c((x, y); z)$$

Exemple 16



Etape 1 : $\mathcal{P}_n = \{\{i\}, \quad i \in I\}$

Etape 2 :

- Calcul des distances entre les n individus
- Recherche des deux éléments les plus proches
- Agrégation de ces deux éléments en un seul ($\mathcal{P}_{n-1}$)

Etape 3 : On a $n-1$ éléments à classer

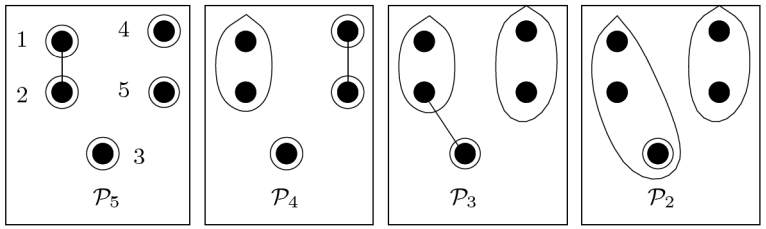
- Calcul des distances après agrégation (selon le critère d'agrégation choisi)
- Recherche des deux éléments les plus proches
- Agrégation de ces deux éléments en un seul ($\mathcal{P}_{n-2}$)

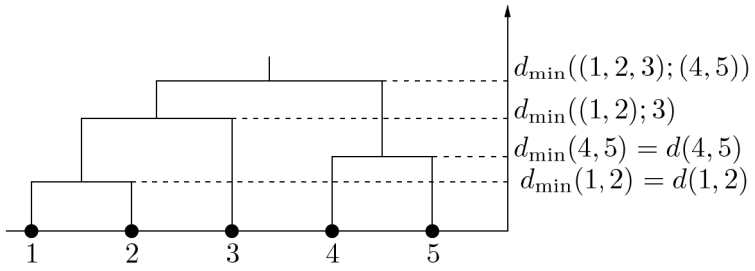
$$\mathcal{P}_{n-2} < \mathcal{P}_{n-1}$$
$$\vdots$$

Itération de l'étape 3 jusqu'à la partition $= I$

Remarque : Fonctionnement de l'algorithme ⇒ la distance n'intervient que par les inégalités qui existent entre elles

Exemple 17 (Construction du dendrogramme pour la distance $d_{\min}$)





- **Principe :** à chaque étape, déterminer une partition par agrégation de deux éléments qui optimise un critère lié à **l'inertie** (ou la variance)
- **Intérêt :** Mise en œuvre aisée lorsque la classification est effectuée après une analyse factorielle
- **Données :**
  - I ensemble des n individus
  - Chaque individu
    - un vecteur $x_i \in \mathbb{R}^p$
    - poids p_i
  - Poids total du nuage : $M = \sum_i p_i$
  - d distance entre les individus
  - $G_I = \frac{1}{M} \sum_i p_i \, x_i$ centre de gravité du nuage

Principe

- $\mathcal{P}_q$ une partition de I en q classes : $\{I_1, \ldots, I_q\}$
  Chaque classe I_k :
  - poids $-\ m_k\ -\ \sum_{i \in I_k} p_i$
  - centre de gravité $G_{I_k} = \frac{1}{m_k} \sum_{i \in I_k} p_i\ x_i$
- **Objectif :** créer des classes homogènes
  ➥ déterminer une partition réalisant le minimum l'inertie intra-classe

D'après le théorème de Huyghens (Proposition 9)

$$\text{minimiser l'inertie intra-classe : Intra } = \sum_{k=1}^{q} \sum_{i \in I_k} p_i\ d^2(x_i, G_{I_k})$$

$$\Longleftrightarrow \text{ maximiser l'inertie inter-classe : Inter} = \sum_{k=1}^{q} m_k\ d^2(G_I, G_{I_k})$$

Etape initiale : $\mathcal{P}_n = \{\{i\}, i \in I\}$

Intra $= 0$ et Inter $=$ Inertie totale

Etape finale : $\mathcal{P}_1 = I$

Intra $=$ inertie totale et Inter $= 0$

À chaque étape d'agrégation, nous allons avoir une perte d'inertie inter-classe. D'où déterminer une agrégation engendrant la perte minimale d'inertie inter-classes

Calcul de la perte d'inertie lors d'une agrégation :

- Agrégation des deux classes I_k et $I_{k'}$:

$$\left.\begin{matrix}(G_{I_k}, m_k)\\ (G_{I_{k'}}, m_{k'})\end{matrix}\right| \rightarrow G = \frac{m_k G_{I_k} + m_{k'} G_{I_{k'}}}{m_k + m_{k'}}, m_k + m_{k'}$$

- Inertie de $G_{I_k}, G_{I_{k'}}$ relativement à G_I

$$\begin{aligned} &= m_k\ d^2(G_{I_k}, G_I) + m_{k'}\ d^2(G_{I_{k'}}, G_I) \\ &= m_k\ d^2(G_{I_k}, G) + m_{k'}\ d^2(G_{I_{k'}}, G) \\ &\quad + \underbrace{\boxed{(m_k + m_{k'}) d^2(G, G_I)}}_{\text{Inertie de } G \text{ / } G_I} \end{aligned}$$

- Remplacement de $G_{I_k}, G_{I_{k'}}$ par le centre G génère la variation d'inertie inter-classe $\Delta_{kk'}$:

$$\begin{aligned} \Delta_{kk'} &= m_k d^2(G_{I_k}, G) + m_{k'} d^2(G_{I_{k'}}, G) \\ &= \frac{m_k m_{k'}}{m_k + m_{k'}} d^2(G_{I_k}, G_{I_{k'}}) \end{aligned}$$

À chaque étape, rechercher I_k et $I_{k'}$ réalisant

$$\min \Delta_{kk'}$$

En résumé :
le passage d'une partition à q classes $\mathcal{P}_q$ à une partition $\mathcal{P}_{q-1}$ à $(q-1)$ classes
$\Longrightarrow$perte d'inertie inter-classe $\Delta_q = \Delta_{kk'}$ où I_k et $I_{k'}$ sont les deux éléments agrégés
$\Longrightarrow$on peut indexer la hiérarchie grâce aux valeurs successives de Δ_q $(q = n, \ldots, 2)$

$$\text{et} \quad \boxed{\sum_{q=2}^{n} \Delta_q = \text{ Inertie totale}}$$

Critère d'agrégation de Ward :

$$d_W(I_k, I_{k'}) = \frac{m_k m_{k'}}{m_k + m_{k'}} d^2(G_{I_k}, G_{I_{k'}})$$

Remarques :

- Pour les calculs
  - soit on travaille sur les coordonnées des vecteurs : calcul du centre de gravité
  - soit on travaille sur les distances et la formule de mise à jour des distances entre éléments après une agrégation

$$d^2_W(I_{ag}, z) = \frac{1}{m_k + m_{k'}} \left(m_k d^2_W(I_k, z) + m_{k'} d^2_W(I_{k'}, z)) - \frac{m_k m_{k'}}{m_k + m_{k'}} d^2_W(I_k, I_{k'}) \right)$$

où I_{ag} est obtenu par agrégation de I_k et $I_{k'}$

- La valeur maximale du niveau d'agrégation ou de l'indice de la hiérarchie est égale à la valeur de l'inertie totale

- Construction d'un arbre hiérarchique : nombre d'opérations $O(n^3)$
- Agréger en une étape non plus deux éléments mais **plusieurs couples d'éléments** ↪ $O(n^2)$
- **Notion de voisins réciproques (Mac Quitty)**

  I_k et $I_{k'}$ sont voisins réciproques ($I_k \leftrightarrow I_{k'}$) :
  si I_k est le plus proche voisin de $I_{k'}$ ($I_k \rightarrow I_{k'}$)
  et si $I_{k'}$ est le plus proche voisin de I_k ($I_k \rightarrow I_{k'}$)

  Principe :
  - À chaque étape, création d'autant de nouveaux noeuds dans l'arbre qu'il y a de paires de voisins réciproques
  - À l'étape finale, tous les groupes sont combinés en un seul et l'arbre est complété

  ↪ Déterminer un algorithme efficace de recherche des voisins réciproques

Algorithme de recherche en chaîne

Étape 1 : on part d'un objet quelconque I_1 et on cherche son plus proche voisin, noté I_2, puis le plus proche voisin de I_2, noté I_3, ...

$$I_1 \rightarrow I_2 \rightarrow \cdots \rightarrow I_{k-2} \rightarrow I_{k-1} \rightarrow I_k \rightarrow \cdots$$

Arrêt lorsque deux éléments successifs sont voisins réciproques

$$\cdots \rightarrow I_{k-2} \rightarrow I_{k-1} \leftrightarrow I_k$$

Étape 2 : si $k = 2$ alors la chaîne commence par $I_1 \leftrightarrow I_2$
Sélection d'un nouvel élément comme origine de la chaîne

Étape 3 : si $k > 2$, on continue la recherche par extension de la chaîne à partir de l'élément I_{k-2}
L'algorithme se termine lorsque $n - 1$ noeuds ou agrégations ont été crées

- **Conditions d'applications de l'algorithme :**

  l'agrégation de I_{k-1} et I_k ne doit pas détruire la relation de plus proche voisin dans

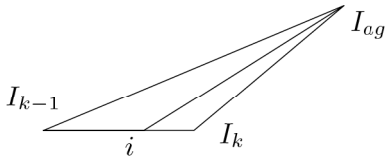
$$I_1 \rightarrow I_2 \rightarrow \cdots \rightarrow I_{k-2}$$

  Cette propriété est assurée si le critère d'agrégation utilisé ne crée pas **d'inversion** dans l'arbre i.e.

  > si I_{k-1} et I_k sont agrégés en I_{ag}
  > alors I_{ag} ne peut pas être plus près d'un élément que ne le sont les éléments I_{k-1} ou I_k

  ↪ **Condition de la médiane (Bruynoghe 78)**

- **Condition de la médiane**



$$\textbf{si } \boldsymbol{d_c(I_{k-1}, I_k)} < \inf\{d_c(I_{k-1}, I_{ag}); d_c(I_k, I_{ag})\}$$
$$\text{alors } \inf\{d_c(I_{k-1}, I_{ag}); d_c(I_k, I_{ag})\} < \boldsymbol{d_c(i, I_{ag})}$$

Cette propriété est satisfaite par les critères
- du saut minimal
- du saut maximal
- de la distance moyenne
- de Ward

Classification mixte

- **Méthode d'agrégation autour des centres mobiles**

  ➥ Partition d'ensembles volumineux de données à faible coût
  Mais partition dépend des premiers centres choisis et le nombre de classes est fixé a priori

- **Classification hiérarchique**

  ➥Stabilité des résultats + indication sur le nombre de classes à retenir
  mais mal adapté aux vastes recueils de données

- **Stratégie mixte de classification :**

combiner les deux techniques précédentes de classification

1. **Partitionnement initial de l'ensemble** à classer par la technique de l'agrégation autour des centres mobiles
   - On améliore la partition en calculant les groupements stables
     ➥ quelques centaines de groupes homogènes
2. **Agrégation hiérarchique des groupes issus de l'étape** 1
   - Elle permet de reconstituer des classes qui ont été fragmentées
   - Construction de l'arbre selon le critère de Ward qui prend en compte les poids au moment des choix des éléments à agréger
     ➥ dendrogramme → déterminer le nombre de classes finales
3. **Optimisation** (par la technique des centres mobiles) de la partition ou des partitions associées aux coupures choisies de l'arbre

- **Inspection visuelle de l'arbre :** coupure après les agrégations correspondantes à des valeurs peu élevées de l'indice
  c.a.d. après les agrégations des éléments les plus proches des uns des autres et avant celles regroupant les classes bien distinctes de la population

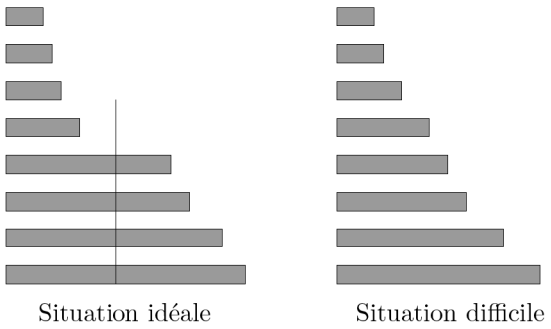


FIGURE 36: Histogramme des indices

- Application d'une procédure d'agrégation autour des centres mobiles à la partition obtenue en Phase 2 (ne fera qu'augmenter l'inertie inter-classe)
  - Les centres de classes sont les centres des classes obtenues par coupure de l'arbre
  - La première itération affecte les éléments au centre de gravité le plus proche
    ➥ Création de nouvelles classes
    Recalcul des centres
  - Les itérations suivantes consiste à réaffecter les éléments à leur centre le plus proche
    Arrêt dès que plus de réaffectation ou dès que l'inertie inter-classe cesse de croître significativement

- **Données :**
  n individus observés à travers les valeurs d'une famille de p variables (continues ou nominales)
- **Classification :**
  Regroupe en classes les individus se ressemblant le plus
- **Interprétation :**
  Comparer les valeurs des variables à l'intérieur d'une classe / aux valeurs pour la population totale
  ➥ Sélectionner les variables les plus caractéristiques de chaque classe

Automatisation de la description des classes
Notion de **valeurs-test**

Valeurs-test pour des variables continues

- X une variable continue
- $\overline{X}$ la moyenne de X sur la population totale
- $n(k)$ effectif de la classe k
- $\overline{X}(k)$ la moyenne de X évaluée sur la classe k
- $S^2(X)$ variance empirique de X
- $S_k^2(X)$ la variance de X évaluée sur la classe k

- Valeur-test pour la classe k

$$t_k(X) = \frac{\overline{X}(k) - \overline{X}}{S_k(X)} \quad \text{avec } S_k^2(X) = \frac{n - n(k)}{n-1} \frac{S^2(X)}{n(k)}$$

- Effectuer un classement des variables actives en fonction de la valeur absolue des valeurs-test
  ➥ Caractériser chaque classe en interprétant ces indicateurs comme une mesure de similarité entre variables et classes

Valeurs-test pour les variables nominales

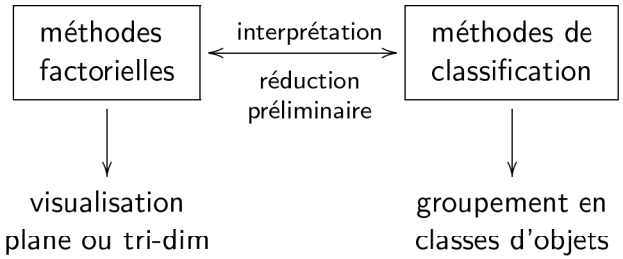
- $n_j(k) =$ effectif d'individus présentant la modalité j parmi les $n(k)$ individus de la classe
- $n_j =$ effectif des individus de la population complète présentant la modalité j
- $n =$ effectif total

- Pour évaluer le « degré de présence » de la modalité j d'une variable nominale dans une classe k d'individus, on compare

$$\frac{n_j(k)}{n(k)} \quad \text{et} \quad \frac{n_j}{n}$$

- Les différences les plus significatives permettent d'isoler les modalités les plus caractéristiques d'une classe d'individus

$$\text{Valeur test } = \frac{n_j(k) - n(k)\frac{n_j}{n}}{\sqrt{n(k)\frac{n-n(k)}{n-1}\frac{n_j}{n}(1-\frac{n_j}{n})}}$$



Insuffisances des méthodes factorielles

Représentations graphiques planes

1. Difficultés d'interprétation pour les plans factoriels $\neq$ (1,2)
2. Compression excessive et déformations
3. Manque de robustesse
4. Graphiques factoriels inextricables

Apports d'une classification

1-2 On complète l'analyse factorielle par une classification sur toute la population ou sur un sous-espace factoriel engendré par les premiers axes
- ➥ corrections de certaines déformations dues à l'opération de projection

Une classe peut être également typique d'un axe de rang élevé
- ➥ aide à l'interprétation

3. Pour la plupart des algorithmes de classification les parties basses de l'arbre sont robustes / points isolés

4. Pour une population de points-individu importante sur un plan factoriel
   - ➥ regroupement en classes homogènes
   - ➥les classes peuvent être utilisés pour aider l'interprétation des plans en identifiant des zones bien décrites

Mais :

- pour observer l'organisation spatiale des classes : positionnement des classes sur les axes factoriels
- mise en évidence de facteurs latents

Classification de données qualitatives

- Une solution consiste à effectuer une classification hiérarchique sur le tableau de toutes les coordonnées factorielles des n individus obtenu par une ACM

  On peut alors vérifier que cette procédure revient à utiliser un critère d'agrégation de Ward pour la distance du chi-deux entre individus, appliqué au tableau disjonctif complet des p variables qualitatives.

- Une solution dans le cas de deux variables, est d'appliquer une classification suivant le critère de Ward pour la distance du chi-deux entre profils-lignes (ou profils-colonnes). Ceci revient à choisir d'agréger deux modalités faisant décroître le moins possible la valeur de l'inertie i.e. χ^2/n

  On peut également travailler sur l'ensemble de **toutes** les coordonnées factorielles du nuage des profils-lignes (ou du nuage des profils-colonnes)

Exemple 18 (Classification après ACP des données de denrées)

```
CLASSIFICATION HIERARCHIQUE (VOISINS RECIPROQUES)
SUR LES 8 PREMIERS AXES FACTORIELS
DESCRIPTION DES NOEUDS
 NUM. AINE  BENJ   EFF.    POIDS    INDICE   HISTOGRAMME DES INDICES DE NIVEAU
   9    4     5     2      2.00   0.06841   **
  10    2     1     2      2.00   0.23844   *****
  11    9     3     3      3.00   0.23893   *****
  12    7     6     2      2.00   0.45561   *********
  13   11    12     5      5.00   1.08829   **********************
  14    8    13     6      6.00   1.78439   ************************************
  15   14    10     8      8.00   4.12593   ********************************************************************************
SOMME DES INDICES DE NIVEAU =    8.00000
```

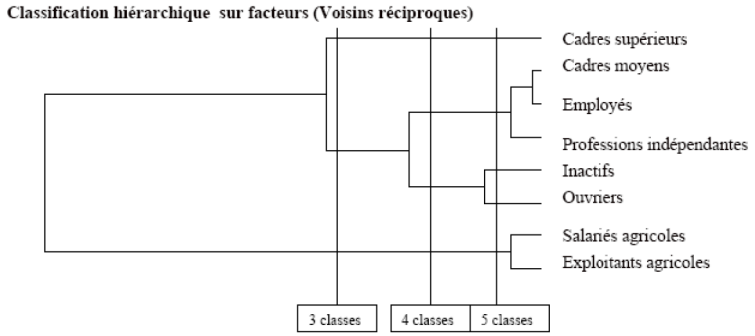
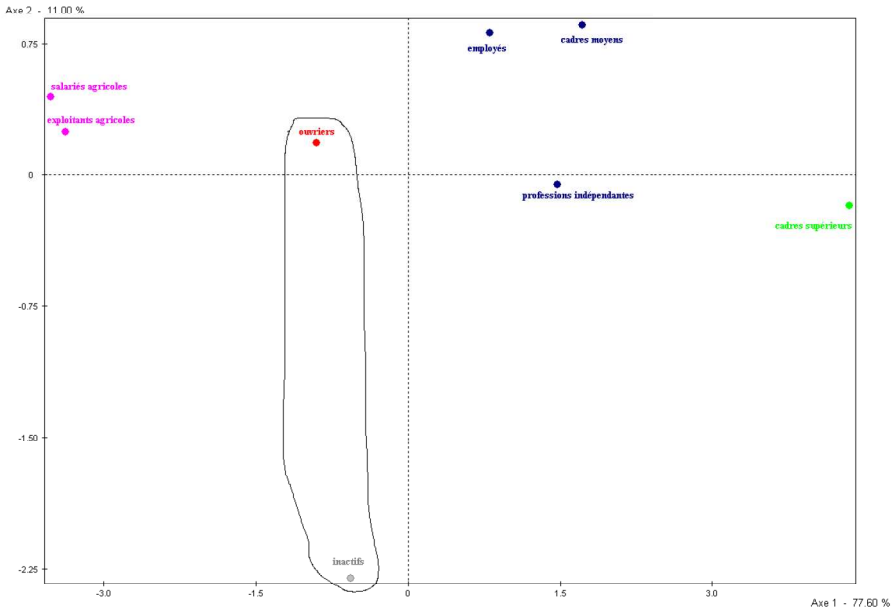


FIGURE 37: Agrégation avec le critère de Ward


```
CONSOLIDATION DE LA PARTITION AUTOUR DES 4 CENTRES DE CLASSES,
REALISEE PAR  5 ITERATIONS A CENTRES MOBILES
PROGRESSION DE L'INERTIE INTER-CLASSES
+-----------+------------+------------+------------+
| ITERATION |   I.TOTALE |    I.INTER |  QUOTIENT  |
+-----------+------------+------------+------------+
|     0     |    8.00000 |    6.99861 |   0.87483  |
|     1     |    8.00000 |    6.99861 |   0.87483  |
|     2     |    8.00000 |    6.99861 |   0.87483  |
+-----------+------------+------------+------------+
ARRET APRES L'ITERATION  2 L'ACCROISSEMENT DE L'INERTIE INTER-CLASSES
PAR RAPPORT A L'ITERATION PRECEDENTE N'EST QUE DE 0.000 %.

CONSOLIDATION DE LA PARTITION AUTOUR DES   5 CENTRES DE CLASSES,
REALISEE PAR  5 ITERATIONS A CENTRES MOBILES
PROGRESSION DE L'INERTIE INTER-CLASSES
+-----------+------------+------------+------------+
| ITERATION |   I.TOTALE |    I.INTER |  QUOTIENT  |
+-----------+------------+------------+------------+
|     0     |    8.00000 |    7.45422 |   0.93178  |
|     1     |    8.00000 |    7.45422 |   0.93178  |
|     2     |    8.00000 |    7.45422 |   0.93178  |
+-----------+------------+------------+------------+
ARRET APRES L'ITERATION  2 L'ACCROISSEMENT DE L'INERTIE INTER-CLASSES
PAR RAPPORT A L'ITERATION PRECEDENTE N'EST QUE DE 0.000 %.
```

Exemple 19 (Classification hiérarchique sur facteurs AFC des données couleurs cheveux/yeux)

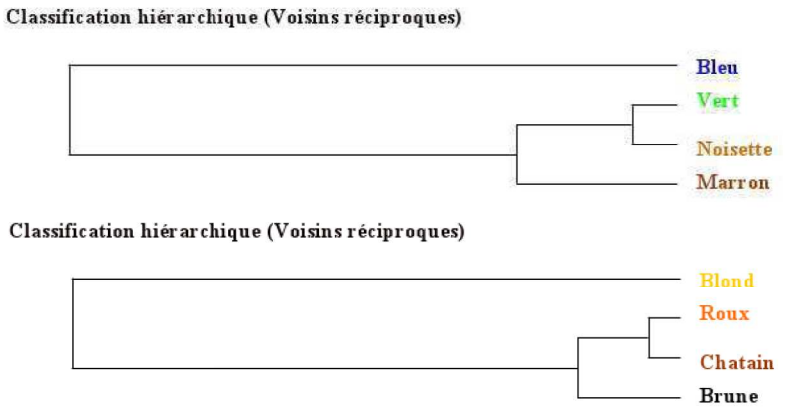


FIGURE 38: Agrégation selon le critère de Ward

Exemple 20 (Classification sur facteurs d'une ACM des races canines)

CLASSIFICATION HIERARCHIQUE (VOISINS RECIPROQUES) SUR LES 10 PREMIERS AXES FACTORIELS
DESCRIPTION DES NOEUDS

NUM.	AINE	BENJ	EFF.	POIDS	INDICE	HISTOGRAMME DES INDICES DE NIVEAU
28	3	1	2	2.00	0.00000	*
29	19	11	2	2.00	0.00000	*
30	26	5	2	2.00	0.00000	*
31	22	8	2	2.00	0.00000	*
32	23	12	2	2.00	0.01236	***
33	10	25	2	2.00	0.01236	***
34	20	16	2	2.00	0.01236	***
35	27	24	2	2.00	0.01236	***
36	29	4	3	3.00	0.01648	****
37	21	13	2	2.00	0.01759	****
38	35	6	3	3.00	0.02925	******
39	18	15	2	2.00	0.02972	******
40	36	14	4	4.00	0.03033	******
41	32	28	4	4.00	0.03091	******
42	17	7	2	2.00	0.03121	******
43	31	2	3	3.00	0.03297	*******
44	34	33	4	4.00	0.04708	*********
45	43	30	5	5.00	0.04826	*********
46	38	37	5	5.00	0.04943	**********
47	39	44	6	6.00	0.05659	**********
48	40	9	5	5.00	0.06920	**********
49	45	42	7	7.00	0.07758	**********
50	47	41	10	10.00	0.10537	**********
51	46	50	15	15.00	0.23052	**********
52	49	48	12	12.00	0.27610	**********
53	52	51	27	27.00	0.43861	**********

SOMME DES INDICES DE NIVEAU = 1.66667

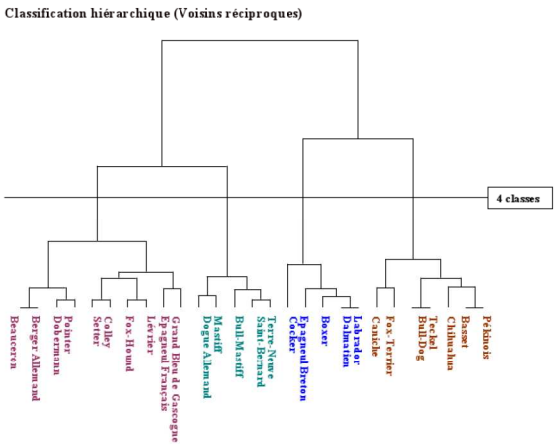
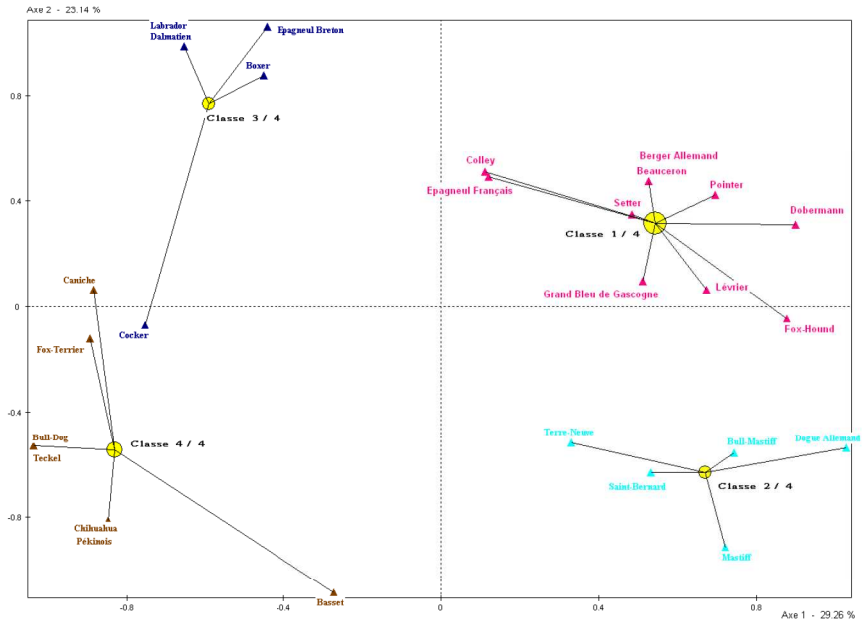


FIGURE 39: Agrégation suivant le critère de Ward

CONSOLIDATION DE LA PARTITION AUTOUR DES 4 CENTRES DE CLASSES,
REALISEE PAR 10 ITERATIONS A CENTRES MOBILES
PROGRESSION DE L'INERTIE INTER-CLASSES

ITERATION	I.TOTALE	I.INTER	QUOTIENT
0	1.66667	0.94523	0.56714
1	1.66667	0.94523	0.56714
2	1.66667	0.94523	0.56714

ARRET APRES L'ITERATION 2 L'ACCROISSEMENT DE
L'INERTIE INTER-CLASSES PAR RAPPORT A L'ITERATION PRECEDENTE
N'EST QUE DE 0.000 %.



SAS

Les procédures SAS/STAT **FASTCLUS**, **CLUSTER**, **VARCLUS**, **TREE**