Exploration Statistique Multidimensionnelle
Chapitre 8 : Classification non supervisée

J-M Loubes

Laboratoire de Statistique et Probabilités

Institut de Mathématiques
www.lsp.ups-tlse.fr/Fp/Loubes

Objectifs

- Matrice $\mathbf{X}(n, p)$ des observations de p variables quantitatives et/ou qualitatives sur n individus
- Tableau de distances (ou dissemblance) des individus
- Recherche d'une typologie, segmentation ou partition des individus en classes par optimisation d'un critère
- Discrimination vs. classif. *classification vs. clustering*
- La complexité impose l'exécution d'un algorithme itératif

Choix de l'utilisateur

- Mesure d'éloignement ou distance entre individus
- Critère : trace de la matrice de variance intra
- Méthode : classif. hiérarchique ou par réallocation dynamique

Objectifs

- Matrice $\mathbf{X}(n, p)$ des observations de p variables quantitatives et/ou qualitatives sur n individus
- Tableau de distances (ou dissemblance) des individus
- Recherche d'une typologie, segmentation ou partition des individus en classes par optimisation d'un critère
- Discrimination vs. classif. *classification vs. clustering*
- La complexité impose l'exécution d'un algorithme itératif

Choix de l'utilisateur

- Mesure d'éloignement ou distance entre individus
- Critère : trace de la matrice de variance intra
- Méthode : classif. hiérarchique ou par réallocation dynamique

Indice de ressemblance ou similarité

- $\Omega = \{i = 1, \ldots, n\}$ ensemble des individus
- s définie de $\Omega \times \Omega$ dans $\mathbb{R}_+$ avec :

$$\begin{aligned} s(i,j) &= s(j,i), \forall (i,j) \in \Omega \times \Omega : \quad \text{symétrie} \\ s(i,i) &= S > 0, \forall i \in \Omega \ : \text{ ressemblance de } i \text{ avec lui-même} \\ s(i,j) &\leq S, \forall (i,j) \in \Omega \times \Omega : \text{ ressemblance majorée par } S \end{aligned}$$

- Indice de ressemblance normé s^* est défini à partir de s par :

$$s^*(i,j) = \frac{1}{S} s(i,j), \forall (i,j) \in \Omega \times \Omega$$

s^* est une application de $\Omega \times \Omega$ dans $[0,1]$

Indice de ressemblance ou similarité

- $\Omega = \{i = 1, \ldots, n\}$ ensemble des individus
- s définie de $\Omega \times \Omega$ dans $\mathbb{R}_+$ avec :

$$
\begin{aligned}
s(i,j) &= s(j,i), \forall (i,j) \in \Omega \times \Omega : \quad \text{symétrie} \\
s(i,i) &= S > 0, \forall i \in \Omega \ : \ \text{ressemblance de } i \text{ avec lui-même} \\
s(i,j) &\leq S, \forall (i,j) \in \Omega \times \Omega : \ \text{ressemblance majorée par } S
\end{aligned}
$$

- Indice de ressemblance normé s^* est défini à partir de s par :

$$s^*(i,j) = \frac{1}{S} s(i,j), \forall (i,j) \in \Omega \times \Omega$$

s^* est une application de $\Omega \times \Omega$ dans $[0,1]$

Indice de ressemblance ou similarité

- $\Omega = \{i = 1, \ldots, n\}$ ensemble des individus
- s définie de $\Omega \times \Omega$ dans $\mathbb{R}_+$ avec :

$$
\begin{array}{rcl}
s(i,j) & = & s(j,i), \forall (i,j) \in \Omega \times \Omega : \quad \text{symétrie} \\
s(i,i) & = & S > 0, \forall i \in \Omega \ : \ \text{ressemblance de } i \text{ avec lui-même} \\
s(i,j) & \leq & S, \forall (i,j) \in \Omega \times \Omega : \ \text{ressemblance majorée par } S
\end{array}
$$

- Indice de ressemblance normé s^* est défini à partir de s par :

$$s^*(i,j) = \frac{1}{S} s(i,j), \forall (i,j) \in \Omega \times \Omega$$

s^* est une application de $\Omega \times \Omega$ dans $[0,1]$

Indice de dissemblance ou dissimilarité

- d de $\Omega \times \Omega$ dans $\mathbb{R}_+$ avec :

$$\begin{aligned} d(i,j) &= d(j,i), \forall (i,j) \in \Omega \times \Omega : \quad \text{symétrie} ; \\ d(i,i) &= 0, \forall i \in \Omega \ : \end{aligned}$$

- Si s est une similarité, d est une dissimilarité

$$d(i,j) = S - s(i,j), \forall (i,j) \in \Omega \times \Omega$$

- Un indice de dissemblance normé est défini par :

$$d^*(i,j) = \frac{1}{D} d(i,j), \forall (i,j) \in \Omega \times \Omega$$

avec $d^* = 1 - s^*$ et $s^* = 1 - d^*$

Indice de dissemblance ou dissimilarité

- d de $\Omega \times \Omega$ dans $\mathbb{R}_+$ avec :

$$\begin{aligned} d(i,j) &= d(j,i), \forall (i,j) \in \Omega \times \Omega : \quad \text{symétrie} ; \\ d(i,i) &= 0, \forall i \in \Omega \ : \end{aligned}$$

- Si s est une similarité, d est une dissimilarité

$$d(i,j) = S - s(i,j), \forall (i,j) \in \Omega \times \Omega$$

- Un indice de dissemblance normé est défini par :

$$d^*(i,j) = \frac{1}{D} d(i,j), \forall (i,j) \in \Omega \times \Omega$$

avec $d^* = 1 - s^*$ et $s^* = 1 - d^*$

Indice de dissemblance ou dissimilarité

- d de $\Omega \times \Omega$ dans $\mathbb{R}_+$ avec :

$$\begin{aligned} d(i,j) &= d(j,i), \forall (i,j) \in \Omega \times \Omega : \quad \text{symétrie} ; \\ d(i,i) &= 0, \forall i \in \Omega \ : \end{aligned}$$

- Si s est une similarité, d est une dissimilarité

$$d(i,j) = S - s(i,j), \forall (i,j) \in \Omega \times \Omega$$

- Un indice de dissemblance normé est défini par :

$$d^*(i,j) = \frac{1}{D} d(i,j), \forall (i,j) \in \Omega \times \Omega$$

avec $d^* = 1 - s^*$ et $s^* = 1 - d^*$

Indice de distance

- $d(i,j) = 0 \Rightarrow i = j$
- Pour éviter des incohérences entre dissemblances

Distance

- $d(i,j) = d(j,i), \forall (i,j) \in \Omega \times \Omega$;
- $d(i,i) = 0 \Leftrightarrow i = j$;
- $d(i,j) \leq d(i,k) + d(j,k); \forall (i,j,k) \in \Omega^3$.

Si Ω est fini, la distance peut être normée

Distance euclidienne

- Si Ω est muni d'un produit scalaire :
  $d(i,j) = [\langle i-j, i-j \rangle]^{1/2} = \|i-j\|$

Indice de distance

- $d(i,j) = 0 \Rightarrow i = j$
- Pour éviter des incohérences entre dissemblances

Distance

- $d(i,j) = d(j,i), \forall (i,j) \in \Omega \times \Omega$;
- $d(i,i) = 0 \Leftrightarrow i = j$;
- $d(i,j) \leq d(i,k) + d(j,k); \forall (i,j,k) \in \Omega^3$.

Si Ω est fini, la distance peut être normée

Distance euclidienne

- Si Ω est muni d'un produit scalaire :
  $d(i,j) = [\langle i-j, i-j \rangle]^{1/2} = \|i-j\|$

Indice de distance

- $d(i,j) = 0 \Rightarrow i = j$
- Pour éviter des incohérences entre dissemblances

Distance

- $d(i,j) = d(j,i), \forall (i,j) \in \Omega \times \Omega$;
- $d(i,i) = 0 \Leftrightarrow i = j$;
- $d(i,j) \leq d(i,k) + d(j,k); \forall (i,j,k) \in \Omega^3$.

Si Ω est fini, la distance peut être normée

Distance euclidienne

- Si Ω est muni d'un produit scalaire :
  $d(i,j) = [\langle i-j, i-j \rangle]^{1/2} = \|i-j\|$

Données quantitatives

- p variables toutes quantitatives,
- matrice de produit scalaire sur l'espace $\mathbb{R}^P$; $\mathbf{M} = \mathbf{I}_p$
- réduire les variables de variances hétérogènes : $\mathbf{M} = \Sigma^{-1}$
- Mahalanobis pour atténuer la structure de corrélation.

Données quantitatives

- p variables toutes quantitatives,
- matrice de produit scalaire sur l'espace $\mathbb{R}^P$; $\mathbf{M} = \mathbf{I}_p$
- réduire les variables de variances hétérogènes : $\mathbf{M} = \Sigma^{-1}$
- Mahalanobis pour atténuer la structure de corrélation.

Données quantitatives

- p variables toutes quantitatives,
- matrice de produit scalaire sur l'espace $\mathbb{R}^P$; $\mathbf{M} = \mathbf{I}_p$
- réduire les variables de variances hétérogènes : $\mathbf{M} = \Sigma^{-1}$
- Mahalanobis pour atténuer la structure de corrélation.

Données quantitatives

- p variables toutes quantitatives,
- matrice de produit scalaire sur l'espace $\mathbb{R}^P$; $\mathbf{M} = \mathbf{I}_p$
- réduire les variables de variances hétérogènes : $\mathbf{M} = \Sigma^{-1}$
- Mahalanobis pour atténuer la structure de corrélation.

Variables binaires

- Notations
  - a_{ij} = nombre de caractères communs à i et j sur les p considérés
  - b_{ij} = nombre de caractères possédés par i mais pas par j
  - c_{ij} = nombre de caractères possédés par j mais pas par i
  - d_{ij} = nombre de caractères que ne possèdent ni i ni j
  - avec bien sûr, $a_{ij} + b_{ij} + b_{ij} + d_{ij} = p$
- Indices :
  - Concordance : $\frac{a_{ij}+d_{ij}}{p}$
  - Jaccard : $\frac{a_{ij}}{a_{ij}+b_{ij}+c_{ij}}$
  - Dice : $\frac{2a_{ij}}{2a_{ij}+b_{ij}+c_{ij}}$

Variables qualitatives

Distance euclidienne du χ^2 entre profils-lignes du tableau disjonctif complet

$$d^2_{\chi^2} = \frac{n}{p} \sum_{j=1} p \sum_{\ell=1}^{m_j} \delta_{ik}^{j\ell} \frac{1}{n_\ell^j}$$

- m_j : nombre de modalités de la variable qualitative Y^j
- n_ℓ^j : effectif de la ℓième modalité de Y^j
- $\delta_{ik}^{j\ell} = 1$ si les individus i et k présentent une discordance pour la ℓième modalité de la variables Y^j et 0 sinon

Mélange quantitatif, qualitatif

- Rendre tout qualitatif : métrique du χ^2
- Rendre tout quantitatif à l'aide d'une AFCM

Résumé

¡2¿

- soit un tableau de mesures $n \times p$ associé à un produit scalaire $p \times p$ (en général $\mathbf{I}_p$)
- soit un tableau $n \times n$ de dissemblances ou distances entre individus
- Attention, si n grand, problèmes de stockage d'un tableau $n \times n$

Mélange quantitatif, qualitatif

- Rendre tout qualitatif : métrique du χ^2
- Rendre tout quantitatif à l'aide d'une AFCM

Résumé

¡2¿

- soit un tableau de mesures $n \times p$ associé à un produit scalaire $p \times p$ (en général $\mathbf{I}_p$)
- soit un tableau $n \times n$ de dissemblances ou distances entre individus
- Attention, si n grand, problèmes de stockage d'un tableau $n \times n$

Mélange quantitatif, qualitatif

- Rendre tout qualitatif : métrique du χ^2
- Rendre tout quantitatif à l'aide d'une AFCM

Résumé

¡2¿

- soit un tableau de mesures $n \times p$ associé à un produit scalaire $p \times p$ (en général $\mathbf{I}_p$)
- soit un tableau $n \times n$ de dissemblances ou distances entre individus
- Attention, si n grand, problèmes de stockage d'un tableau $n \times n$

CAH : Objectif

- Agglomération itérative de 2 éléments de la partition
- Construction d'un dendrogramme ou arbre binaire
- Problème : définir $d(A, B)$ A et B deux groupes ou éléments d'une partition à partir de
  - w_A et w_B leurs pondérations
  - $d_{i,j}$ la dissemblance ou distance entre deux individus

CAH : Objectif

- Agglomération itérative de 2 éléments de la partition
- Construction d'un dendrogramme ou arbre binaire
- Problème : définir $d(A, B)$ A et B deux groupes ou éléments d'une partition à partir de
  - w_A et w_B leurs pondérations
  - $d_{i,j}$ la dissemblance ou distance entre deux individus

CAH : Objectif

- Agglomération itérative de 2 éléments de la partition
- Construction d'un dendrogramme ou arbre binaire
- Problème : définir $d(A, B)$ A et B deux groupes ou éléments d'une partition à partir de
  - w_A et w_B leurs pondérations
  - $d_{i,j}$ la dissemblance ou distance entre deux individus

CAH : Objectif

- Agglomération itérative de 2 éléments de la partition
- Construction d'un dendrogramme ou arbre binaire
- Problème : définir $d(A, B)$ A et B deux groupes ou éléments d'une partition à partir de
  - w_A et w_B leurs pondérations
  - $d_{i,j}$ la dissemblance ou distance entre deux individus

CAH : Objectif

- Agglomération itérative de 2 éléments de la partition
- Construction d'un dendrogramme ou arbre binaire
- Problème : définir $d(A, B)$ A et B deux groupes ou éléments d'une partition à partir de
  - w_A et w_B leurs pondérations
  - $d_{i,j}$ la dissemblance ou distance entre deux individus

Cas d'une dissemblance

- $d(A,B) = \min_{i\in A, j\in B}(d_{ij})$ saut minimum, single linkage
- $d(A,B) = \sup_{i\in A, j\in B}(d_{ij})$ saut maximum ou diamètre, complete linkage
- $d(A,B) = \frac{1}{\text{card}(A)\text{card}(B)} \sum_{i\in A, j\in B} d_{ij}$ saut moyen, group average linkage

Cas d'une distance euclidienne

g_A et g_B : barycentres des classes

$$d(A,B) = d(g_A, g_B) \quad \text{(distance des barycentres, centroïd)}$$

$$d(A,B) = \frac{w_A w_B}{w_A + w_B} d(g_A, g_B) \quad \text{(saut de Ward)}$$

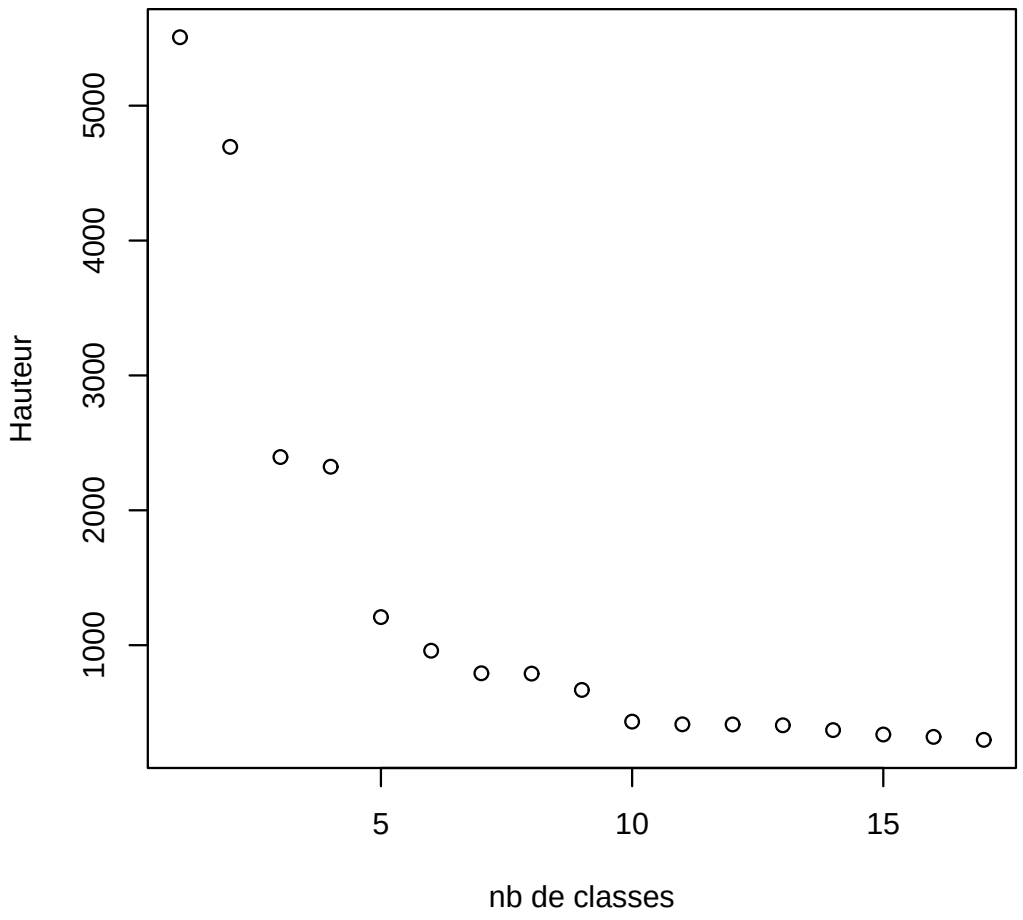
Saut de Ward et maximisation de la variance inter

Algorithme de classification ascendante hiérarchique

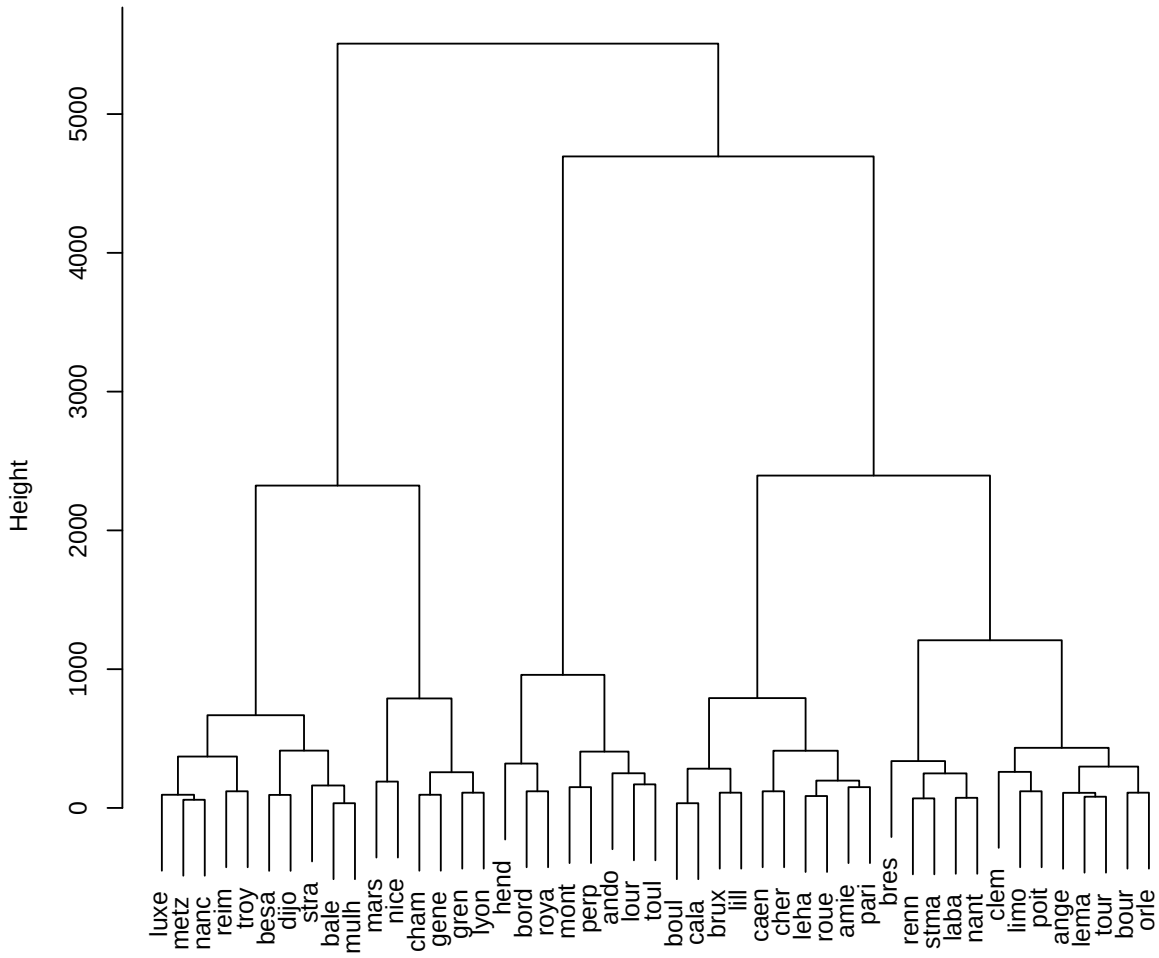
- Initialisation : singletons, calcul des distances
- Itérer jusqu'à agrégation en une seule classe :
  1. regrouper les deux classes les plus proches au sens de la "distance" entre groupes choisie
  2. mise à jour du tableau de distance en remplaçant les deux classes par la nouvelle et en calculant sa "distance" avec les autres classes

Nombre de classes : Rupture dans la décroissance du R^2 partiel (Ward)

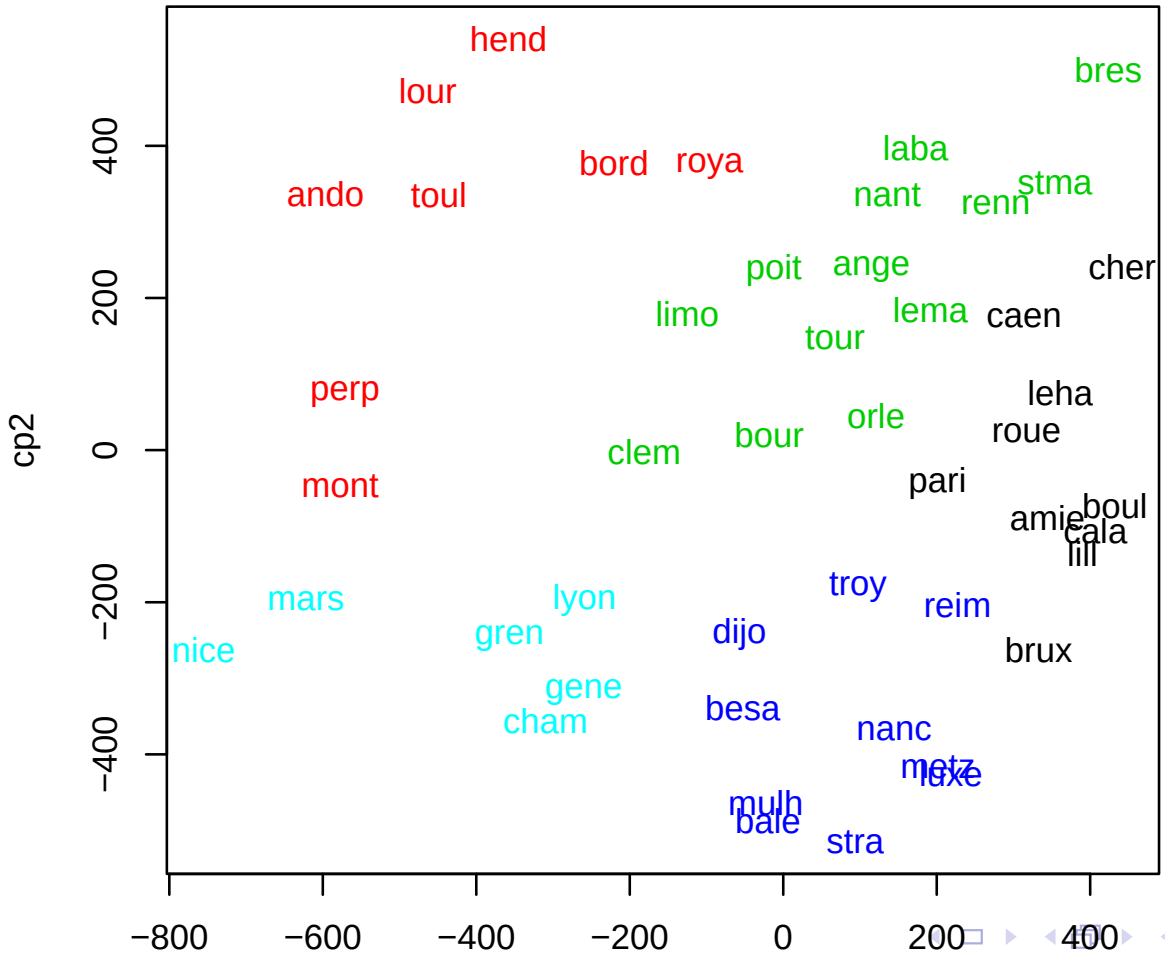
Villes : Décroissance de la variance inter classes



Villes : Exemple d'un dendogramme



Villes : Représentation des classes avec MDS



Principe des centres mobiles

- réallocation dynamique des individus à des classes
- Le nombre de classes k est fixé *a priori*

Algorithme de Forgy

- Initialisation Tirer au hasard ou sélectionner, k points dans l'espace des individus, en général k individus de l'ensemble, appelés centres ou noyaux
- Itérer jusqu'à stagnation du critère de variance inter-classe
  1. Allouer chaque individu à l'un des noyaux, c'est-à-dire à une classe
  2. Calculer le centre de gravité de chaque classe, il devient le nouveau noyau

Attention : optimum local

Principe des centres mobiles

- réallocation dynamique des individus à des classes
- Le nombre de classes k est fixé *a priori*

Algorithme de Forgy

- Initialisation Tirer au hasard ou sélectionner, k points dans l'espace des individus, en général k individus de l'ensemble, appelés centres ou noyaux
- Itérer jusqu'à stagnation du critère de variance inter-classe
  1. Allouer chaque individu à l'un des noyaux, c'est-à-dire à une classe
  2. Calculer le centre de gravité de chaque classe, il devient le nouveau noyau

Attention : optimum local

Variantes

- Algorithme $k$means : les noyaux des classes, ici les barycentres, sont recalculés à chaque allocation d'un point à une classe ; algorithme plus efficace
- Nuées dynamiques : Un centre de classes est un noyau d'éléments représentatifs d'une classe
- Partitionning Around Medoïds

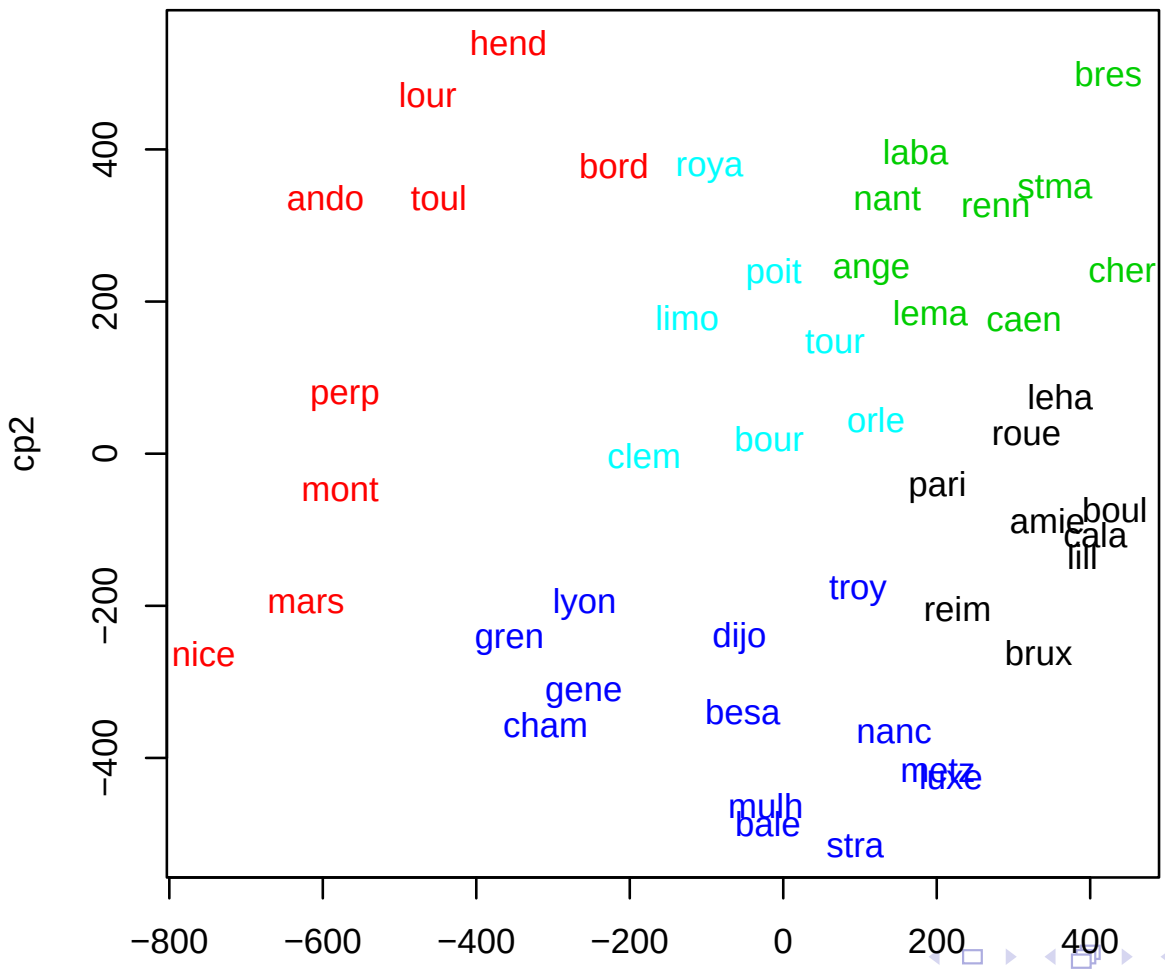
Variantes

- Algorithme $k$means : les noyaux des classes, ici les barycentres, sont recalculés à chaque allocation d'un point à une classe ; algorithme plus efficace
- Nuées dynamiques : Un centre de classes est un noyau d'éléments représentatifs d'une classe
- Partitionning Around Medoïds

Variantes

- Algorithme $k$means : les noyaux des classes, ici les barycentres, sont recalculés à chaque allocation d'un point à une classe ; algorithme plus efficace
- Nuées dynamiques : Un centre de classes est un noyau d'éléments représentatifs d'une classe
- Partitionning Around Medoïds

Villes : Classes d'un PAM avec MDS



Classification de grands tableaux

1. Réallocation dynamique avec un grand nombre de classes $(n/10)$
2. Classification hiérarchique des barycentres
3. détermination d'un nombre "optimal" k de classes
4. Réallocation dynamique de l'ensemble avec k classes en choisissant pour noyaux les barycentres des classes de l'étape précédente

Conclusion

- Résultat : une variable qualitative T dont les modalités précisent la classe retenue pour chaque individu
- Problèmes : interprétation des classes
- Surtout dans le cas de variables qualitatives