Séance 3: Positionnement Multidimensionnel - Classification

Sébastien Gadat

Laboratoire de Statistique et Probabilités
UMR 5583 CNRS-UPS

www.lsp.ups-tlse.fr/gadat

Troisième partie III

Positionnement Multidimensionnel

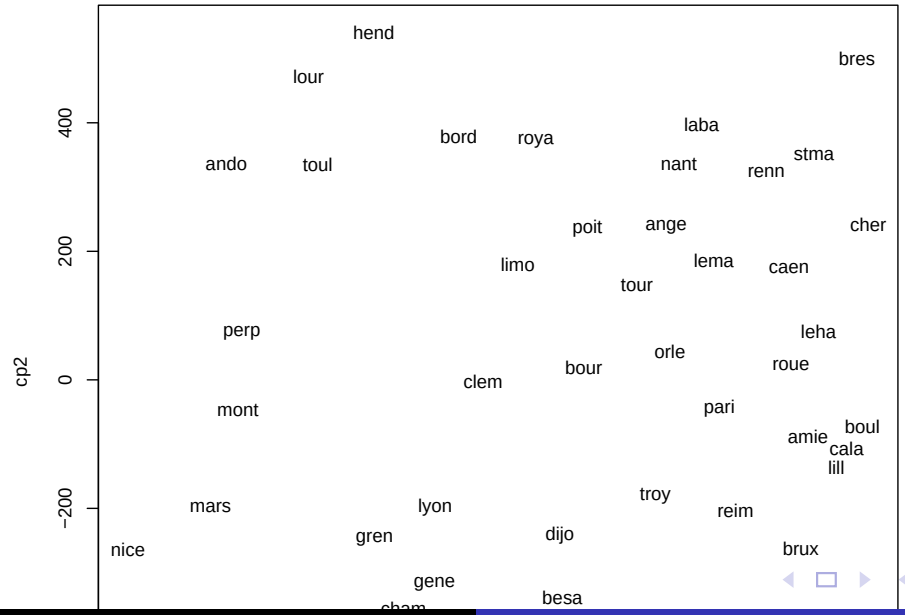
Introduction

Notations

- On suppose donnés n individus.
- p variables de valeurs inconnues pour chaque individu.
- Indice de dissimilarité/distance entre chacun des individus connus.
- ♣ Comment construire une représentation euclidienne dans un espace de taille réduite fidèle aux données ?

Exemples

- Données : tableau contenant les distances à parcourir par route entre différentes villes (en km).
- Les "coordonnées" des villes sont inconnues.



Exemples

- ♣ La matrice de distances est-elle euclidienne ?
- ♣ L'approximation euclidienne est-elle satisfaisante ?
- La MDS est une technique factorielle (nécessité de déterminer un nombre de dimension).
- Possibilité d'observer graphiquement les données à travers différentes optiques.

Définitions

Rappelons quelques propriétés et définitions élémentaires mais basiques à propos de la notion de distance.

- Une matrice $(n \times n)$ $\mathcal{D}$ est appelée matrice de distance si elle est symétrique et si :

$$d_j^j = 0 \text{ et } \forall (j,k), j \neq k, d_j^k \geq 0.$$

- Une matrice $(n \times n)$ $\mathcal{C}$ est appelée matrice de similarité si elle est symétrique et si

$$\forall (j,k), c_j^k \leq c_j^j.$$

- ♣ On peut transformer une matrice de similarité en $(c_j^k)_{j,k}$ en matrice de distance *via* :

$$d_j^k = (c_j^j + c_k^k - 2c_j^k)^{-1/2}$$

Définitions

Une matrice de distance est dite euclidienne s'il existe une configuration de vecteurs $\{x_1, \ldots, x_n\}$ vérifiant

$$(d_j^k)^2 =$$

On note A la matrice issue de $\mathcal{D}$ de terme général

$$(a_i^j) = -\frac{(d_j^k)^2}{2}$$

D désigne la matrice des poids des individus.

Définitions

Proposition :

- La matrice de projection D-orthogonale au vecteur $\mathbf{1}$ est donnée par
$$H = Id - \mathbf{1}\mathbf{1}'\mathbf{D}$$
- Une matrice de distance $\mathcal{D}$ est euclidienne si et seulement si $B = HAH'$ est symétrique définie positive. (B est la matrice obtenue par double centrage de A).
- Si la matrice de similarité C est positive, alors la matrice de distance $\mathcal{D}$ déduite de C est euclidienne.

Recherche d'une configuration de points

- Positionnement multidimensionnel : recherche d'une configuration de points dans un espace euclidien qui admette une matrice de distances :
  - égale à $\mathcal{D}$ si celle-ci satisfait la proposition précédente
  - meilleure approximation possible pour un rang de matrice donné (en général 2) de $\mathcal{D}$
- Il n'y a jamais unicité d'une telle représentation : si $(x_i)_i$ est une solution, alors

$$(z_i)_i = (Fx_i + b)_i$$

  est une solution lorsque F orthogonale et b quelconque.
- Une solution est donc définie à rotation et translation près.

Algorithme MDS

On se donne $\mathcal{D}$ matrice de distance et B la matrice centrée des lignes et colonnes, calculée comme précédemment.

- Si $\mathcal{D}$ est une matrice de distance euclidienne de points $\{x_1, \ldots, x_n\}$, alors B s'écrit en fait

$$b_{i,j} = \langle x_i - \bar{x}, x_j - \bar{x} \rangle$$

et B se met sous la forme $B = (HX)'(HX)$ et est appelée matrice des produits scalaires de la configuration centrée.

- Réciproquement, si $\mathbf{B}$ est positive de rang p, on écrit sa décomposition spectrale

$$\mathbf{B} = \mathbf{U}\mathbf{\Delta}\mathbf{U}'$$

Les vecteurs sont les lignes de la matrice centrée

Relation entre ACP et MDS

Supposons que Y soit un nuage de n points $(x_1, \dots x_n)$ décrits par p variables (connues). On définit $\mathcal{D}$ comme la matrice des distances entre les n individus :

$$d_i^j = \|x_i - x_j\|_M$$

La représentation graphique obtenue par MDS sur $\mathcal{D}$ est identique à la réalisation d'une ACP sur $(Y, M, 1/nId)$.

Quatrième partie IV

Classification

Introduction

Notations

- On suppose donnés n individus.
- Les données se présentent :
  - Sous la forme d'un tableau de distance
  - Les observations de p variables quantitatives sur les n individus
  - Un mélange de variables qualitatives et quantitatives
- Pour chacun des cas, on construit un tableau de distance entre individus
- Objectif : recherche d'une segmentation (ou partition) des individus

Introduction

Notations

- Problème : il y a trop de partitions possibles pour espérer explorer toutes les segmentations des individus
- Moyens : optimisation d'un critère pour fusionner deux groupes d'individus
- Le problème traité s'appelle *clustering*, c'est une technique d'apprentissage non supervisé

Construction d'un Critère

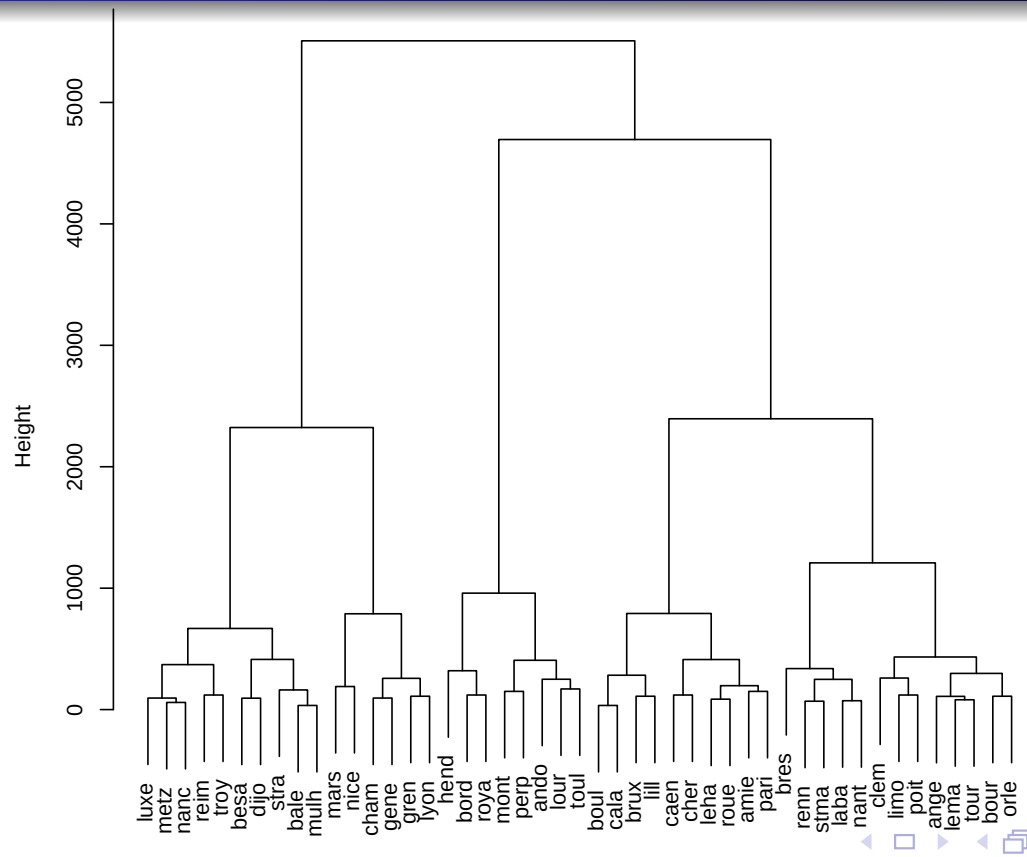
Toutes les techniques présentées seront des algorithmes itératifs convergeant vers une "bonne" partition.
Comment mesurer si une partition est "bonne" ?
Degrés de latitude :

- Mesure d'éloignement des individus
- Critère d'homogénéité des classes à optimiser (souvent un critère de variance)
- Méthode de fusion (Classification ascendante ou nuées dynamiques)
- Nombre de classes
- Obtenir des classes "homogènes" (variance intra-classe faible)
- Obtenir des classes "bien" distinctes (variance inter-classe élevée)

Classification hiérarchique ascendante

- Regrouper 2 individus les plus proches
- Construction progressive d'un arbre du bas vers le haut
- Sélection de groupes à "fusionner" par le biais de calculs de sauts ou *linkage*
- Nombre de classes déterminé *a posteriori*

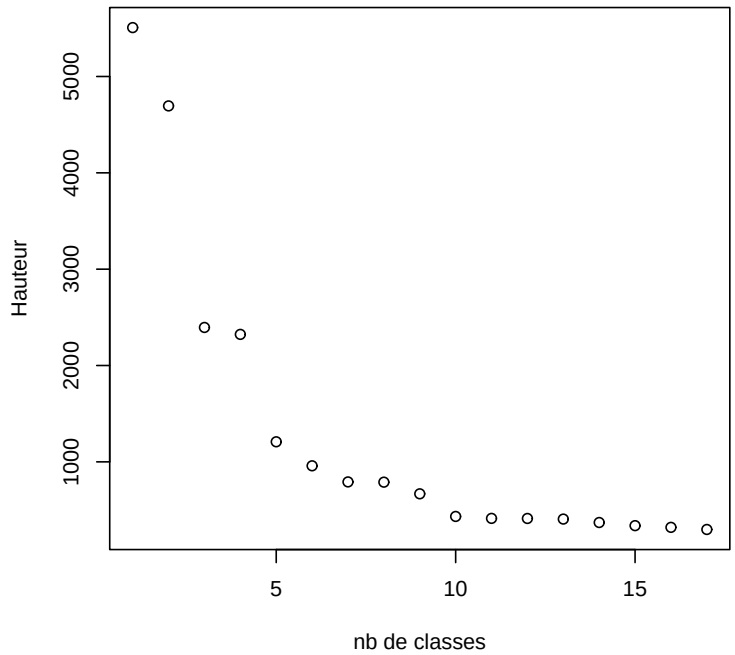
Classification Ascendante Hiérarchique sur les villes



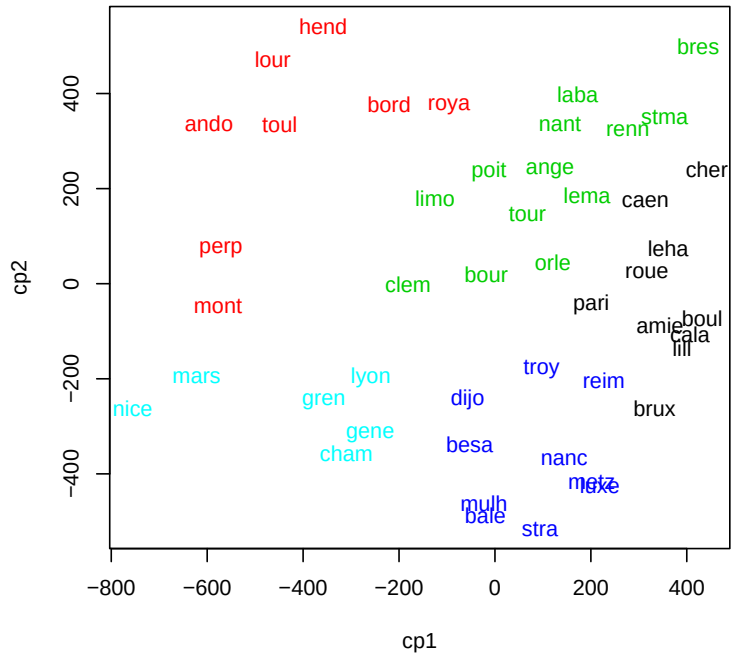
Classification par ré-allocation dynamique

- On fixe le nombre de classes k *a priori*
- Tirage aléatoire (uniforme) de k centres de classe
- Partition des individus en regroupant vers le centre le plus proche possible
- Mise à jour des centres comme barycentre des groupes obtenus
- Itération du processus jusqu'à stabilisation des partitions

CAH sur l'exemple des villes



CAH sur l'exemple des villes



Stabilité des classifications

Choix du nombre de classe : k est le plus petit entier correspondant à une baisse non significative de la variance inter-classes : ici $k = 5$.
CAH : instabilité importante par rapport aux modifications des distances entre individus initiaux
Ré-allocation dynamique : plus grande stabilité

Mesures d'éloignement entre individus

L'indice de ressemblance s satisfait :

- $\forall (i,j) \in \Omega \qquad s(i,j) = s(j,i)$
- $\forall i \in \Omega \qquad s(i,i) = S > 0$
- $\forall (i,j) \in \Omega \qquad s(i,j) < S$

L'indice de dissemblance d satisfait :

- $\forall (i,j) \in \Omega \qquad d(i,j) = d(j,i)$
- $\forall i \in \Omega \qquad d(i,i) = 0$

On passe d'une notion à l'autre en utilisant :

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

Par ailleurs, on peut normer l'indice d par

$$d^\star(i,j) = \frac{1}{-}d(i,j)$$

Mesures d'éloignement entre individus

Un indice de distance est un indice de dissemblance satisfaisant en plus :

$$\forall (i,j) \in \Omega \qquad d(i,j) = 0 \Longrightarrow i = j$$

Une distance est un indice de distance vérifiant en plus l'inégalité triangulaire :

$$\forall (i,j,k) \in \Omega \qquad d(i,j) \leq d(i,k) + d(k,j)$$

- Métrique Euclidienne $M = Id$, distance euclidienne :

$$d(x,y) = \|x - y\|_M$$

- Métrique réduite matrice diagonale M : inverse des écarts types
- Métrique de Mahalanobis : M inverse de la matrice de variance

Mesures d'éloignement entre groupes

Pour effectuer la CAH, il est nécessaire de pouvoir calculer des distances entre groupes. Si A et B désignent ces 2 groupes, on peut opter pour plusieurs stratégies :

- ♣ $d(A,B)=$ Saut minimum, single linkage
- ♣ $d(A,B)=$ Saut maximum, complete linkage
- ♣ $d(A,B)=$ Saut moyen, average linkage
- ♣ $d(A,B)=$ Barycentres, centroïds
- ♣ $d(A,B)=$ Saut de Ward

Le saut de Ward est la stratégie la plus courante : dans le cas Euclidien, ce critère correspond à chaque itération à minimiser la décroissance de la variance interclasse.

Propriétés des Centres mobiles

On initialise les k centres sur k individus parmi les n choisis au hasard.
Propriété La variance inter-classes augmente à chaque itération.
L'algorithme converge vers un optimum local de la variance inter/intra classes.
Plusieurs aménagement ont été proposés :

- k-means : les barycentres des classes sont mis à jour à chaque allocation d'un individu dans une classe
- on remplace le noyau barycentrique par un noyau représentatif de la classe

Classification par Centres mobiles