UN CRITÈRE BASÉ SUR LA DISTANCE DE MAHALANOBIS POUR L'AFFECTATION D'OBJETS SUPPLÉMENTAIRES AUX CLASSES D'UNE CAH EUCLIDIENNE

Frédérik Cassor, Brigitte Le Roux

UN CRITÈRE BASÉ SUR LA DISTANCE DE MAHALANOBIS POUR L'AFFECTATION D'OBJETS SUPPLÉMENTAIRES AUX CLASSES D'UNE CAH EUCLIDIENNE

Frédérik Cassor[1] & Brigitte Le Roux[2]

[1] *CEVIPOF Sciences-po, 98 Rue de l'Université 75007 PARIS.*
frederik.cassor@sciencespo.fr
[2] *MAP5, Université Paris Descartes 45 rue des Saints Pères 75270 Paris Cedex 06.*
Brigitte.LeRoux@mi.parisdescartes.fr

Résumé. Dans une classification ascendante hiérarchique euclidienne (méthode de Ward), la méthode usuelle d'affectation d'un individu supplémentaire à une classe est basée sur la distance géométrique du point-individu au centre de la classe. Cette méthode présente l'inconvénient de ne pas tenir compte de ce que les classes différent, quant aux poids, à leurs formes et à leurs dispersions ; elle ne tient pas compte également des dichotomies successives de la hiérarchie de parties issue de la classification. C'est pourquoi nous proposons une nouvelle règle de classement adaptée à l'analyse géométrique des données qui tient compte de la forme de chacune des classes.

Partant d'un ensemble d'individus supplémentaires, nous proposons une stratégie d'affectation de ces individus aux classes issues d'une hiérarchie binaire de parties. L'idée est d'affecter les individus supplémentaires au niveau local d'un nœud à l'un de ses deux successeurs, jusqu'à parvenir à une classe de la partition étudiée. Nous définissons un critère qui repose sur le rapport des distances de Mahalanobis du point-individu au centre des deux classes constituant le nœud.

Nous présentons d'abord le principe de la méthode, puis nous l'appliquons à des enquêtes barométriques initiées par le CEVIPOF qui portent sur la confiance des électeurs français. Nous étudions l'évolution des classes d'individus entre 2009 et 2013. Pour cela, nous avons écrit un programme en langage R.

Mots-clés. Classification ascendante hiérarchique, classement, distance de Mahalanobis, données d'enquête, programme R.

Abstract. In a Euclidean hierarchical ascending clustering (HAC, Ward's method), the usual method for allocating a supplementary individual to a cluster is based on the geometric distance from the individual–point to the barycenter of the cluster. The main drawback of this method is that it does not take into consideration that clusters differ as regards weights, shapes and dispersions. Neither does it take into account successive dichotomies of the hierarchy of clustering. This is why we propose a new ranking rule adapted to geometric data analysis that takes the shape of clusters into account.

From a set of supplementary individuals, we propose a strategy for assigning these individuals to clusters stemming from a HAC. The idea is to assign supplementary individuals at the local level of a node to one of its two successors until a cluster of the partition under study is

reached. We define an allocation criterion based on the ratio of Mahalanobis distances from the individual–point to barycenters of the two clusters that make up the node.

We first introduce the principle of the allocation method, and we apply it to several barometric surveys carried out by the CEVIPOF on various components of trust among French voters. We study the evolution of clusters of individuals between 2009 and 2013. To do this, we have written a program in R language.

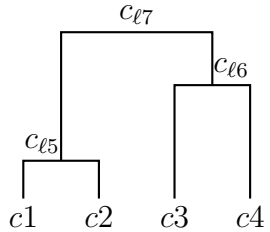
Keywords. Ascending Hierarchical Clustering, ranking, Mahalanobis' distance, survey data, R program.

1 Classement et hiérarchie

Classer un individu est une étape importante des méthodes de classification. Nous nous placerons ici dans le cas d'une classification ascendante hiérarchique de points d'un nuage euclidien.

Considérons un ensemble C de classes muni d'une *hiérarchie totale binaire* de parties dont les éléments terminaux sont les classes $c \in C$, que nous appellerons "classes primaires". Les n objets de l'ensemble I sont répartis dans les C classes primaires d'effectifs $(n_c)_{c \in C}$; les classes de la hiérarchie sont désignées par c_ℓ, avec ℓ allant de 1 à C pour les classes primaires, et de $C+1$ à $2C-1$ pour les classes associées aux nœuds de l'arbre.

Prenons l'exemple suivant d'une hiérarchie avec 4 éléments terminaux : $C = \{c1, c2, c3, c4\}$.



Le classement d'un élément supplémentaire se fera par voie descendante en procédant comme suit :

- on décide d'abord auquel des deux successeurs (ici $c_{\ell 6}$ ou $c_{\ell 5}$) du sommet (nœud $c_{\ell 7}$, qui est l'ensemble I de tous les objets classés) affecter cet élément,
- une fois l'affectation faite à un nœud ℓ (disons ici $c_{\ell 6}$), on décide auquel des deux successeurs de ce nœud on doit affecter cet élément (disons ici $c_{\ell 4}$),
- et ainsi de suite jusqu'à parvenir à une classe primaire de la hiérarchie.

2 Distance d'un point à une classe

Pour mesurer la distance d'un point–individu à une classe, on utilise la κ–norme dont le carré est la distance de Mahalanobis associée à cette classe. Pour la classe c_ℓ et l'individu i, la

valeur de cette norme est notée $\kappa_\ell(i)$.

On se place dans une base orthonormée, typiquement la base principale du nuage des individus. On note $\mathbf{c}_\ell$ la colonne des coordonnées du point moyen de la classe c_ℓ, $\mathbf{V}_\ell$ la matrice de covariance de cette classe et $\mathbf{y}$ la colonne des coordonnées d'un individu i.

La distance de Mahalanobis du point–individu i à la classe c_ℓ (ou κ–norme) est telle que :

$$\kappa_\ell^2(i) = {}^t(\mathbf{y} - \mathbf{c}_\ell)\mathbf{V}_\ell^{-1}(\mathbf{y} - \mathbf{c}_\ell)$$

Si, comme indice de proximité entre un individu i et une classe, on prend la distance géométrique du point–individu i au centre de la classe, on ne tient aucun compte de ce que les classes différent quant au poids, à la forme et à la dispersion. Or il semble, par exemple, plus naturel qu'un point à égale distance du centre d'une classe très concentrée et du centre d'une classe très dispersée soit plutôt rattaché à cette dernière. Il est donc préférable de prendre comme distance entre un point et une classe, l'indice κ qui tient compte de la forme de la classe.

3 Critère d'affectation

Pour décider si un individu i doit être affecté à la classe c_ℓ ou à la classe $c_{\ell'}$, on comparera le rapport $\rho_{(\ell,\ell')}(i) = \kappa_\ell^2(i)/\kappa_{\ell'}^2(i)$ à un seuil $\alpha_{(\ell,\ell')}$:

$$i \text{ est affecté } c_\ell \text{ si } \rho_{(\ell,\ell')}(i) < \alpha_{(\ell,\ell')} \text{ et à } c_{\ell'} \text{ sinon}^{1}$$

Parmi les seuils α possibles on choisira le seuil $\widehat{\alpha}_{(\ell,\ell')}$ pour lequel le nombre d'erreurs obtenu en "ré-affectant" à ce seuil les individus ($i \in I$) qui ont constitué ces classes est minimum.

On note :

- $N_\ell(\alpha)$ le nombre d'individus de la classe c_ℓ qui sont mal classés au seuil α, c'est–à–dire pour lesquels $\rho_{(\ell,\ell')}(i) > \alpha_{(\ell,\ell')}$;
- $N_{\ell'}(\alpha)$ le nombre d'individus de la classe $c_{\ell'}$ mal classés au seuil α, c'est–à–dire pour lesquels $\rho_{(\ell,\ell')}(i) < \alpha_{(\ell,\ell')}$;
- $N_{(\ell,\ell')}(\alpha) = N_\ell(\alpha) + N_{\ell'}(\alpha)$ le nombre d'individus des classes c_ℓ et $c_{\ell'}$ mal classés au seuil α.

Le seuil $\widehat{\alpha}$ est la valeur α qui correspond au *minimum* de $N_{(\ell,\ell')}(\alpha)$.

Algorithme de calcul du seuil. Pour calculer le seuil, on commence par ranger les valeurs $\rho_{(\ell,\ell')}(i)$ par ordre croissant, d'où la suite indexée par j (avec $1 \leq j \leq n_{c_\ell} + n_{c_{\ell'}}$) :

$$\rho_{(\ell,\ell')}(1) \leq \ldots \rho_{(\ell,\ell')}(j) \leq \ldots \leq \rho_{(\ell,\ell')}(n_{c_\ell} + n_{c_{\ell'}})$$

Si $\alpha < \rho_{(\ell,\ell')}(1)$, alors tous les individus sont affectés à la classe $c_{\ell'}$, et donc il y a n_{c_ℓ} erreurs.

Si $\rho_{(\ell,\ell')}(1) < \alpha < \rho_{(\ell,\ell')}(2)$, il y a une erreur de moins si l'individu correspondant à $j = 1$ appartient à la classe c_ℓ et une de plus s'il appartient à la classe $c_{\ell'}$, etc.

On note $j_{\min}$ le rang correspondant au minimum de $N_{(\ell,\ell')}(\alpha)$, *i.e.* le rang de l'individu dont le rapport ρ est pris comme seuil correspondant au nombre minimum de mal classés.

1. En cas d'égalité, on pourra le rattacher à la classe la plus nombreuse.

On prend un seuil compris entre $\rho_{(\ell,\ell')}(j_{\min})$ et $\rho_{(\ell,\ell')}(j_{\min+1})$, par exemple : $\alpha = \big(\rho_{(\ell,\ell')}(j_{\min}) + \rho_{(\ell,\ell')}(j_{\min}+1)\big)/2$

4 Application aux données du baromètre de la confiance (CEVIPOF)

4.1 Les données

Les données analysées ici proviennent d'enquêtes initiées par le CEVIPOF, qui prennent en compte les composantes diverses et parfois hétérogènes de la confiance.

> Une série de cinq enquêtes ont été effectuées, chaque année depuis Décembre 2009, par le CEVIPOF (centre de recherches politiques de Sciences-Po Paris) [2] en relation avec l'Institut Pierre Mendès France et le Conseil économique, social et environnemental. Les échantillons (d'environ 1500 personnes) sont représentatifs de la population française âgée de 18 ans et plus ; ils ont été constitués par la méthode des quotas, au regard des critères de sexe, d'âge, de catégorie socio–professionnelle, après stratification par région de résidence et de taille de commune. La méthode de recueil des données est l'enquête en ligne en utilisant le système CAWI (Computer Assisted Web Interview) ; la série d'enquêtes a été réalisée par Opinion Way.

L'objectif de cette étude est d'étudier l'évolution de la confiance entre 2009 et 2012 puis 2013.

4.2 Classification des individus

Pour comparer l'évolution de la confiance entre 2009 et 2012 puis 2013, nous avons pris comme période de référence 2009, date intermédiaire entre les élections présidentielles de 2007 et de 2012. Les *individus* actifs de l'analyse sont ceux de l'enquête 2009 (vague 1), ceux des enquêtes 2012 et 2013 ont été mis en éléments supplémentaires.

Pour construire l'espace de la confiance et établir une typologie des électeurs français au regard de la confiance, nous avons retenu cinq composantes de la confiance mesurées par 24 questions relevant de cinq thèmes : politique, institutionnel, économique, inter–individuel et individuel.

Nous avons effectué une classification euclidienne des individus de la vague 1 (2009). On voit (cf. figure 1) que les indices de niveau décroissent lentement à partir du 4^{e} : on interprète une partition en 4 classes. L'arbre hiérarchique supérieur aboutissant à une partition en 4 classes et les ellipses de concentration de ces classes dans le plan 1-2 de l'espace de la confiance (construit par une analyse des correspondances des données dédoublées) sont représentés sur la figure 2.

2. cf. www.cevipof.com/fr/le-barometre-de-la-confiance-politique-du-cevipof/.

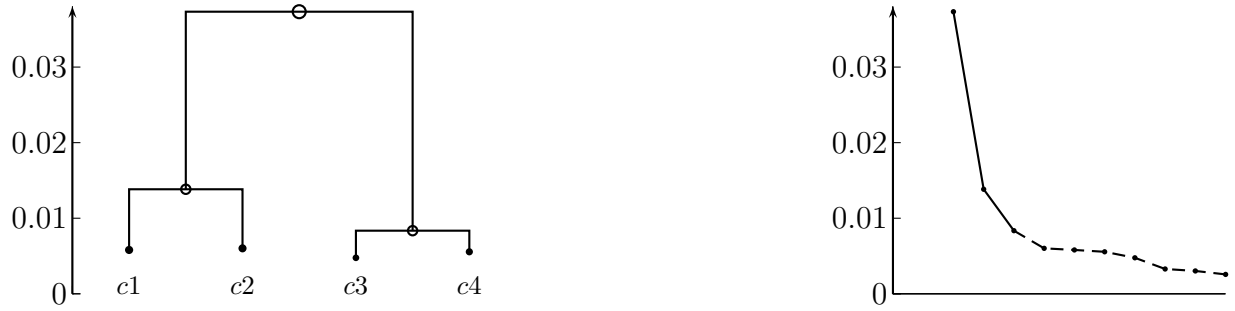


FIGURE 1 – Arbre hiérarchique supérieur et diagramme des indices de niveau.

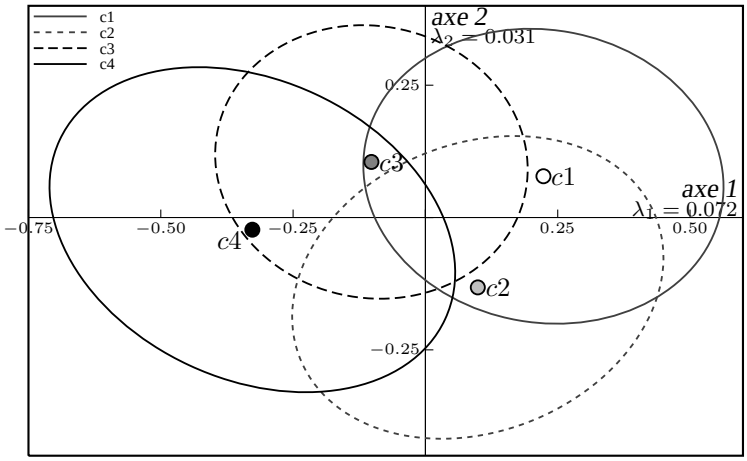


FIGURE 2 – Ellipses de concentration des 4 classes dans le plan 1-2.

La classe $c1$ ($n_{c1} = 402$) est celles des "hyperconfiants", la classe $c2$ ($n_{c2} = 396$) est celle des "confiants modérés", la classe $c3$ ($n_{c3} = 267$) est celle des "défiants modérés" et la classe $c4$ ($n_{c4} = 311$) est celle des "hyperdéfiants".

4.3 Classement des individus supplémentaires

Pour compléter cette étude, nous avons affecté les individus des vagues 2012 et 2013 aux classes définies par la CAH des individus de la vague 2009, en suivant la procédure de classement présentée précédemment. Les pourcentages d'individus dans chaque classe sont donnés dans le tableau suivant.

classes		Déc 2009	Déc 2012	Déc 2013
$c1$	hyperconfiants	29	29	28
$c2$	confiants modérés	29	20	17
$c3$	défiants modérés	19	31	34
$c4$	hyperdéfiants	23	20	21

Le niveau de confiance diminue : le classement des individus permet de préciser cette évolution : il existe un glissement important de la confiance modérée vers la défiance modérée, les classes extrêmes restant pratiquement stables.

Bibliographie

[1] Benzécri, J-P. (1977), Analyse discriminante et analyse factorielle, *Les cahiers de l'analyse des données*, 2 (4), 369-406.

[2] Fisher, R. A. (1936), The use of multiple measurements in taxonomic problems, *Annals of eugenics*, 7 (2), 179-188.

[3] Le Roux, B. (2014), *Analyse géométrique des données multidimensionnelles*, Dunod, Paris.

[4] Le Roux, B. et Perrineau, P. (2011), Les différents types d'électeurs au regard de différents types de confiance, *Les cahiers du CEVIPOF*, http ://www.cevipof.com/fr/les-publications/les-cahiers-du-cevipof/, 54, 5-35.