Classification ascendante hiérarchique, contrainte d'ordre : conditions d'applicabilité, interprétabilité des dendrogrammes

Nathanaël Randriamihamison [1,2] & Pierre Neuvial [2] & Nathalie Vialaneix [1]

[1] *INRA, UR 0875 MIAT, 31326 Castanet Tolosan cedex, France,*
{nathanael.randriamihamison,nathalie.vialaneix}@inra.fr
[2] *Institut de Mathématiques de Toulouse, Univ. Paul Sabatier, UMR 5219,*
pierre.neuvial@math.univ-toulouse.fr

Résumé. La classification ascendante hiérarchique (CAH) avec lien de Ward, ainsi que sa version sous contrainte d'ordre, sont couramment utilisées sur des données de différents types : distances, dissimilarités, noyaux, similarités. Nous précisons dans quel cas l'utilisation de ces méthodes est justifiée théoriquement. Nous étudions les conditions garantissant la cohérence entre les résultats de la CAH et leur représentation graphique classique sous forme de dendrogramme. Cette étude révèle une distinction importante entre cette propriété de cohérence et l'absence de croisement dans le dendrogramme.

Mots-clés. Classification ascendante hiérarchique, lien de Ward, classification ascendante hiérarchique sous contrainte, dendrogramme, croissance, inégalité ultramétrique

Abstract. Ward's Hierarchical Agglomerative Clustering (HAC) and its order-constrained version are widely used on various data types : distances, dissimilarities, kernels, or similarities. We clarify the theoretical justification of these methods in each case. We study the conditions that guarantee the consistency between the results of HAC and their graphical display as a dendrogram. This study reveals an important distinction between this consistency property and the absence of crossover within the dendrogram.

Keywords. Hierarchical agglomerative clustering, Ward's linkage, constrained hierarchical agglomerative clustering, dendrogram, monotonicity, ultrametric inequality

1 Classification Ascendante Hiérarchique

Cadre euclidien standard

Dans le cadre standard de la Classification Ascendante Hiérarchique (CAH), on suppose que l'ensemble des objets $\Omega = \{x_1, \ldots, x_n\}$ est un sous-ensemble de $\mathbb{R}^p$, et que la relation de proximité entre ces objets est encodée par la distance $D = (d_{ij})_{1 \leq i,j \leq n}$ induite par la norme euclidienne de $\mathbb{R}^p$ avec $d_{ij} = \|x_i - x_j\|_{\mathbb{R}^p}$.

L'algorithme de CAH part de la partition triviale en singletons $\mathcal{P}_1$, puis par fusions successives, se termine sur une autre partition triviale, $\mathcal{P}_n = \Omega$. L'étape t permettant de passer de la partition $\mathcal{P}_t$ à la partition $\mathcal{P}_{t+1}$ fusionne les deux classes de $\mathcal{P}_t$ les plus proches au sens d'une dissimilarité entre classes appelée *critère de lien* et notée δ.

Le *lien de Ward* [Ward, 1963] est le plus couramment utilisé en statistique car il a une interprétation simple. Pour un ensemble quelconque $G \subset \Omega$, l'inertie de G est définie par $I(G) = \sum_{x \in G} \|x - \bar{x}_G\|^2$, où $\bar{x}_G = |G|^{-1} \sum_{x \in G} x$ est le centre de gravité de G, et $|G|$ le cardinal de G. On appelle inertie intra-classes d'une partition $\mathcal{P} = (G_1, G_2, ..., G_K)$ la somme des inerties des classes : $I(\mathcal{P}) = \sum_{k=1}^{K} I(G_k)$. Le lien de Ward entre deux sous-ensembles disjoints G et G' est défini par : $\delta(G, G') = I(G \cup G') - I(G) - I(G')$ et correspond alors à la variation d'inertie intra-classes induite par cette fusion.

Extensions de la CAH

La CAH est couramment utilisée sur des données plus générales que celles décrites dans le cadre euclidien de la section précédente. On suppose désormais que les objets $(x_i)_i$ appartiennent à un espace arbitraire (non nécessairement euclidien).

Dissimilarités. Une dissimilarité $D = (d_{ij})_{1 \leq i,j \leq n}$ est une matrice symétrique à coefficients positifs et à diagonale nulle. Remarquons que dans le cas euclidien, l'inertie $I(G)$ s'exprime directement à partir des éléments de la matrice de distance D comme $I(G) = (2|G|)^{-1} \sum_{(x_i,x_j) \in G^2} d_{ij}^2$. On peut alors définir, par analogie au cas euclidien, une extension de l'inertie et du lien de Ward associée à D (voir par exemple [Székely and Rizzo, 2005] ou [Chavent et al., 2018]). Cependant, dans ce cas, on perd l'interprétabilité en termes de centre de gravité.

Noyaux. Dans un certain nombre de cas pratiques, les objets sont décrits par leurs ressemblances et non leurs dissemblances. C'est le cas, en particulier, lorsque les données sont décrites par un noyau [Schölkopf et al., 2004], c'est-à-dire, par une matrice symétrique définie positive $K = (k(x_i, x_j))_{1 \leq i,j \leq n}$. [Aronszajn, 1950] montre que le noyau peut être interprété comme une matrice de produit scalaire dans un certain espace de représentation $\mathcal{H}$. Ce cadre-ci est donc équivalent au cadre euclidien et on peut montrer que l'expression du lien de Ward s'écrit à partir des valeurs du noyau uniquement en utilisant l'*astuce noyau* [Dehman, 2015] :

$$\delta(G, G') = \frac{K(G)}{|G|} + \frac{K(G')}{|G'|} - \frac{K(G \cup G')}{|G \cup G'|}, \qquad \text{avec } K(G) = \sum_{x_i, x_j \in G^2} k(x_i, x_j). \tag{1}$$

Similarités. On peut interpréter la notion de similarité comme une généralisation de celle de noyau. Bien qu'il n'y ait pas de consensus sur la définition exacte d'une similarité, on appellera similarité une matrice $S = (s_{ij})_{1 \leq i,j \leq n}$ symétrique à diagonale positive. [Miyamoto et al., 2015] ont montré que l'approche « noyau » décrite ci-dessus pouvait être généralisée à toute matrice de similarité, par l'argument suivant : considérons la matrice $S + \lambda I_n$, c'est-à-dire la matrice S dont les élements diagonaux sont translatés de λ. D'une part, pour λ suffisamment grand, S_λ est définie positive, et peut donc s'interpréter comme un noyau. D'autre part, appliquer (formellement) l'algorithme de CAH à la matrice de similarité S_λ en utilisant l'équation (1) avec $K = S_\lambda$ aboutit exactement à la même suite de fusions, quelle que soit la valeur de λ. En effet, l'équation (1) implique : $\delta_{S_\lambda} = \delta_S + \lambda$.

Dans la suite, nous distinguerons simplement deux cas : le cas euclidien (qui contient d'après ce qui précède les cas « noyau » et « similarité quelconque »), et le cas non-euclidien, qui correspond aux dissimilarités quelconques.

CAH sous contrainte d'ordre

Dans un certain nombre de contextes applicatifs, il existe une information *a priori* sur les relations entre les objets. C'est le cas, en particulier, en statistique spatiale où les objets spatiaux sont en relation de voisinage, ou bien dans le contexte génomique, où les loci génomiques sont ordonnés le long d'une ligne (le chromosome). La classification ascendante hiérarchique sous contrainte de contiguïté [Ferligoj and Batagelj, 1982] restreint les fusions possibles aux objets dits contigus.

Dans la suite, on s'intéresse au cas particulier de la Classification Ascendante Hiérarchique sous Contrainte d'Ordre (CAHCO) : les objets initiaux sont reliés par une relation d'ordre total (temps, position génomique), et deux classes sont dites contiguës si et seulement si on peut en extraire un couple d'objets contigus. L'intérêt de cette contrainte est double : d'une part, elle permet de répercuter au mieux les relations existantes entre les objets au cours de la procédure, et ainsi de rendre les résultats plus interprétables. D'autre part, la complexité en temps de la CAH sans contrainte est cubique ($O(n^3)$) et celle de la CAHCO est seulement quadratique ($O(n^2)$) [Dehman, 2015].

2 Dendrogrammes

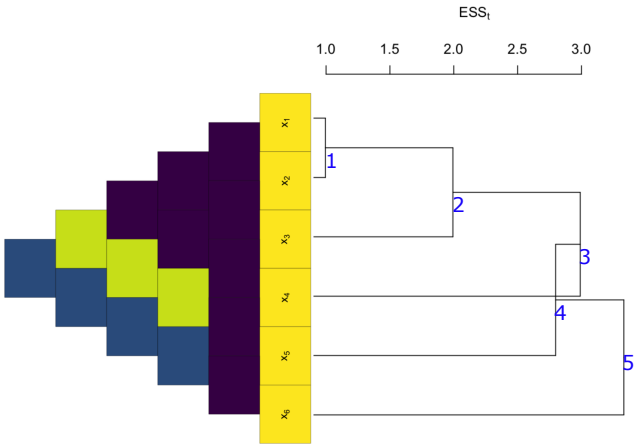
Les résultats d'une CAH sont fréquemment représentés à l'aide d'un *dendrogramme*, comme sur la figure 1. Un dendrogramme est un arbre binaire dont les nœuds sont les classes de la hiérarchie. Dans le cas particulier de la CAHCO, les feuilles, qui correspondent aux n objets à classer, sont représentées selon l'ordre naturel de ces objets. La hauteur du nœud correspondant à la t-ème fusion est notée h_t (les feuilles sont à la hauteur $h_0 = 0$). On utilise souvent comme hauteur la valeur du lien de Ward à l'étape t, notée m_t.

Croissance

Une propriété naturelle est que la suite de partitions induite par la CAH coïncide avec celle décrite par le dendrogramme, ce qui équivaut à la croissance de la suite $(h_t)_t$. Cependant, lorsque la hauteur est définie par le lien de Ward, la croissance n'est pas vérifiée pour la CAHCO [Grimm, 1987]. [Grimm, 1987] décrit des choix de hauteurs alternatifs au lien de Ward, dont les propriétés selon les données ou le type de CAH considérés sont précisées ci-dessous :

L'inertie intra-classes : la hauteur à l'étape t est définie par : $\mathrm{ESS}_t = \sum_{u=1}^{n-t} I(G_u^{t+1})$. Cette hauteur est croissante pour la CAH [Ward, 1963, Batagelj, 1981] et pour la CAHCO dans le cas euclidien [Grimm, 1987]. Mais nous démontrons qu'elle ne l'est pas

nécessairement dans le cas non euclidien. Ceci est illustré par la figure 1, qui donne un exemple simple où la CAHCO est appliquée à six objets, avec : $h_1 < h_2 < h_4 < h_3 < h_5$.



$$\begin{pmatrix} 0 & \sqrt{1.99} & \sqrt{1.99} & \sqrt{1.99} & 0.1 & 1 \\ \sqrt{1.99} & 0 & \sqrt{2} & \sqrt{1.99} & 0.1 & 1 \\ \sqrt{1.99} & \sqrt{2} & 0 & \sqrt{2} & 0.1 & 1 \\ \sqrt{1.99} & \sqrt{1.99} & \sqrt{2} & 0 & \sqrt{2} & 1 \\ 0.1 & 0.1 & 0.1 & \sqrt{2} & 0 & \sqrt{2} \\ 1 & 1 & 1 & 1 & \sqrt{2} & 0 \end{pmatrix}$$

FIGURE 1 – **Un croisement dû à la non-croissance de la hauteur définie par l'inertie intra-classes pour la CAHCO avec données non euclidiennes**. À gauche : représentation de la dissimilarité associée. Au centre : dendrogramme correspondant aux résultats de la CAHCO (l'ordre des fusions est noté en bleu). À droite : dissimilarité associée aux objets.

L'inertie de la fusion courante, $I_t = I(G_u^t \cup G_v^t)$, où G_u^t et G_v^t sont les deux classes fusionnées à l'étape t. Nous démontrons que la croissance de cette hauteur n'est jamais assurée (voir figure 2).

L'inertie moyenne de la fusion courante, est égale à $\overline{I}_t = \frac{I_t}{(|G_u^t|+|G_v^t|)}$. Pour cette hauteur, la croissance n'est jamais assurée dans le cas contraint [Grimm, 1987]. Nous démontrons qu'elle n'est pas non plus assurée dans le cas non contraint.

Le tableau 1 présente un récapitulatif des résultats de croissance dans ces différents contextes.

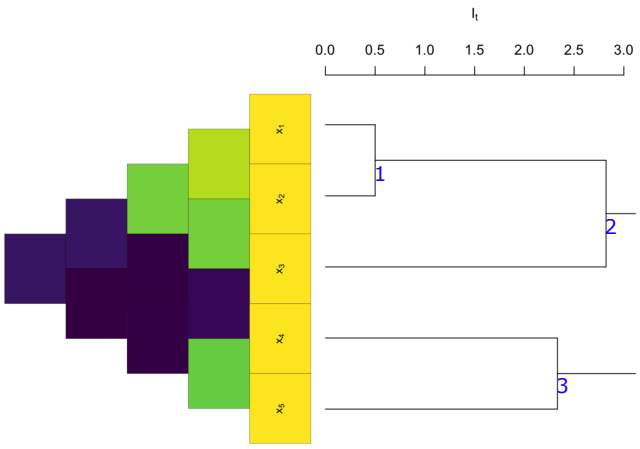
		m_t	ESS$_t$	I_t	$\overline{I}_t$
CAH	Euclidien	✓[Ward, 1963]	✓[Ward, 1963]	× [Fig. 2]	×
	Non euclidien	✓[Batagelj, 1981]	✓[Batagelj, 1981]	× [Fig. 2]	×
CAHCO	Euclidien	×[Grimm, 1987]	✓[Grimm, 1987]	× [Fig. 2]	×[Grimm, 1987]
	Non euclidien	×[Grimm, 1987]	× [Fig. 1]	× [Fig. 2]	×[Grimm, 1987]

TABLE 1 – Croissance des hauteurs pour la CAH (haut) et la CAHCO (bas).

Différence entre croissance et ultramétricité

Dans l'exemple de la figure 1, la non-croissance des hauteurs se manifeste par un *croisement* de branches dans le dendrogramme. Formellement, un croisement se produit lorsque la hauteur d'une classe produite par la fusion de G_u et G_v est plus petite que la hauteur de G_u ou de G_v. Cependant, croissance des hauteurs et absence de croisement ne sont pas équivalentes.

En effet, à partir du dendrogramme, on définit la *distance cophénétique* entre paires d'objets de $\{x_1, \ldots, x_n\}$ comme la hauteur du nœud du dendrogramme qui correspond à la première classe dans laquelle les deux objets considérés sont simultanément classés. Si on note $(h_{ij})_{i,j=1,\ldots,n}$ l'ensemble des distances cophénétiques entre paires $\{(x_i, x_j)\}_{ij}$, on appelle inégalité ultramétrique pour cette dissimilarité la propriété suivante : $\forall i, j, k \in \{1, ..., n\}$, $h_{ij} \leq \max\{h_{ik}, h_{kj}\}$. Lorsque cette propriété est vérifiée, la représentation graphique est assurée de ne comporter aucun croisement et réciproquement. Cependant, bien que croissance implique ultramétricité, la réciproque est, en général, fausse. La figure 2 montre un exemple où l'inégalité ultramétrique est respectée mais pas la croissance ($h_1 < h_3 < h_2$) : on y constate une *inversion de hauteur sans croisement*.



$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ 0 & -1 \\ 10 & 0 \\ 10 & 2.16 \end{pmatrix}$$

FIGURE 2 – **Inversion de la hauteur définie par l'inertie de la classe fusionnée pour la CAH (ou la CAHCO : sur cet exemple, les deux donnent des résultats identiques) avec données euclidiennes**. À gauche : Dissimilarité. Au centre : Dendrogramme (l'ordre des fusions est noté en bleu). À droite : coordonnées des objets.

Parmi les hauteurs que nous avons étudiées, les seules où ultramétricité implique croissance dans le contexte de la CAH sont celles où la hauteur est définie par le lien de Ward ou par ESS_t. Ces deux hauteurs sont certainement à privilégier. Pour ces deux cas, on peut bien conclure qu'un dendrogramme sans croisement garantit une suite des hauteurs croissante, et donc cohérence entre les résultats de la CAH et le dendrogramme.

3 Conclusion

La CAH ainsi que sa version sous contrainte d'ordre peuvent s'appliquer de façon justifiée dans un cadre plus vaste que le cadre euclidien. En revanche, certaines propriétés, comme la croissance des hauteurs de fusion, peuvent ne pas être garanties en fonction du contexte. Par ailleurs, la cohérence entre la procédure de classification hiérarchique et le dendrogramme associé n'est pas toujours équivalente à l'absence de croisement dans la représentation. Cette distinction vient de la différence entre la notion d'ultramétricité pour la distance cophénétique induite par la CAH, et la croissance des hauteurs. Il convient d'y être attentif pour interpréter correctement les résultats de la classification.

Remerciements

Les auteurs remercient Marie Chavent pour des discussions stimulantes autour de ce travail. Celui-ci a été effectué dans le cadre du projet SCALES, financé par la Mission pour les Initiatives Transverses et Interdisciplinaires du CNRS. La thèse de N.R. est financée par le programme INRA/Inria.

Références

[Aronszajn, 1950] Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3) :337–337.

[Batagelj, 1981] Batagelj, V. (1981). Note on ultrametric hierarchical clustering algorithms. *Psychometrika*, 46(3) :351–352.

[Chavent et al., 2018] Chavent, M., Kuentz-Simonet, V., Labenne, A., and Saracco, J. (2018). ClustGeo2 : an R package for hierarchical clustering with spatial constraints. *Computational Statistics*, 33(4) :1799–1822.

[Dehman, 2015] Dehman, A. (2015). *Spatial Clustering of Linkage Disequilibrium Blocks for Genome-Wide Association Studies*. PhD thesis, Université Paris Saclay.

[Ferligoj and Batagelj, 1982] Ferligoj, A. and Batagelj, V. (1982). Clustering with relational constraint. *Psychometrika*, 47(4) :413–426.

[Grimm, 1987] Grimm, E. C. (1987). CONISS : a FORTRAN 77 program for stratigraphically constrained analysis by the method of incremental sum of squares. *Computers & Geosciences*, 13(1) :13–35.

[Miyamoto et al., 2015] Miyamoto, S., Abe, R., Endo, Y., and Takeshita, J.-I. (2015). Ward method of hierarchical clustering for non-Euclidean similarity measures. In *Proceedings of the VIIth International Conference of Soft Computing and Pattern Recognition (SoCPaR 2015)*, Fukuoka, Japan. IEEE.

[Schölkopf et al., 2004] Schölkopf, B., Tsuda, K., and Vert, J.-P. (2004). *Kernel Methods in Computational Biology*. MIT PR.

[Székely and Rizzo, 2005] Székely, G. J. and Rizzo, M. L. (2005). Hierarchical clustering via joint between-within distances : extending Ward's minimum variance method. *Journal of Classification*, 22(2) :151–183.

[Ward, 1963] Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301) :236–244.