Assigning Objects to Classes of a Euclidean Ascending Hierarchical Clustering

Brigitte Le Roux, Frédérik Cassor

Assigning Objects to Classes of a Euclidean Ascending Hierarchical Clustering

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

[1] MAP5, Université Paris Descartes 45 Rue des Saints Pères 75270 Paris Cedex 06
[2] CEVIPOF Sciences-Po, 98 Rue de l'Université 75007 Paris

Abstract. In a Euclidean hierarchical ascending clustering (HAC, Ward's method), the usual method for allocating a supplementary object to a cluster is based on the geometric distance from the object-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 objects, we propose a strategy for assigning these objects to clusters stemming from a AHC. The idea is to assign supplementary objects 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 object–point to barycenters of the two clusters that make up the node.
We first introduce the principle of the method, and we apply it to a barometric survey carried out by the CEVIPOF on various components of trust among French citizens. We compare the evolution of clusters of individuals between 2009 and 2012 then 2013.

Keywords. Geometric Data Analysis, Correspondence Analysis and doubling coding, Ascending Hierarchical Clustering, Mahalanobis distance, survey data.

1 Assignment by Dichotomies to a System of Clusters

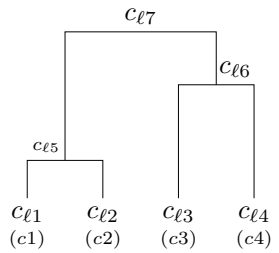
Let I be a set of n objects, and C a set[3] of C classes $c \in C$ defining a partition of I. We suppose that we have a nested system, providing a hierarchy of classes and more precisely a dichotomic hierarchical clustering represented by a binary hierarchical tree.

The end elements of the hierarchy are the classes $c \in C$, that we call "primary classes". The n objects of I are divided into the C primary classes, the absolute frequencies of classes are denoted $(n_c)_{c \in C}$ (with $n = \sum n_c$). The classes of the hierarchy are denoted c_ℓ with ℓ going from 1 to C for the primary classes, and

[3] As a general convention, we denote the cardinalities of finite sets like the sets themselves, except for set I.

from $C+1$ to $2C-1$ for the classes associated with the nodes of the hierarchical tree. The top node $c_{\ell=2C-1}$ is the set I of all objects.

As an example, take the following hierarchy with four terminal elements: $C = \{c1, c2, c3, c4\}$.



The assignment of a supplementary object will be made downwards as follows:

- we decide to which of the two immediate successors (here $c_{\ell 6}$ or $c_{\ell 5}$) of the top node ($c_{\ell 7}$) we assign the supplementary object;
- once the assignement to a node ℓ is made (say here $c_{\ell 6}$), we decide to which of the two immediate successors ($c_{\ell 3}$ or $c_{\ell 4}$) of this node ($c_{\ell 5}$) we assign the supplementary object (here to $c_{\ell 4}$);
- and so on until we reach a primary class.

2 Distance from Object to Class

In the sequel, we suppose that the objects are represented by points in a Euclidean space. We apply to the cloud of points a Euclidean clustering, that is, an ascending hierarchical clustering using the variance criterion (Ward's method) [LeRoux04].

We suppose that the space is referred to an orthonormal basis, for instance the one associated with the principal axes of the cloud. We denote $\mathbf{c}_\ell$ the column–of the coordinates of the mean point of the class c_ℓ and $\mathbf{V}_\ell$ its covariance matrix. If $\mathbf{y}$ denotes the column–vector of the coordinates of the supplementary object i_s, the index of proximity between object and class, denoted $\kappa_\ell(i_s)$, is equal to the Mahalanobis distance [6] from object–point i_s to the center of class c_ℓ, that is:

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

Comment

If, as an index of proximity, we take the geometric distance from the object-point to the center of the class, that is $\left((\mathbf{y} - \mathbf{c}_\ell)^\top (\mathbf{y} - \mathbf{c}_\ell)\right)^{1/2}$, we do not take into account the fact that the classes differ in weight, shape and dispersion. Now it seems natural that a point that is equidistant from the center of a highly concentrated class and from the one of a very dispersed class will be assigned to the latter. Hence it is preferable to choose as a distance from a point to a class the κ-norm, since it takes into account the shape of the class.

3 Assignment Criterion

In order to decide if an individual i is assigned to class c_ℓ or to clas $c_{\ell'}$, we will compare the ratio $\rho_{(\ell,\ell')}(i) = \kappa^2_\ell(i)/\kappa^2_{\ell'}(i)$ to a threshold $\alpha_{(\ell,\ell')}$.

$$i \text{ is assigned to class } c_\ell \text{ if } \rho_{(\ell,\ell')}(i) < \alpha_{(\ell,\ell')} \text{ and to } c_{\ell'} \text{ if not}^4.$$

Among the possible thresholds, we will choose the one, denoted $\widehat{\alpha}_{(\ell,\ell')}$, for which, if we assign the n basic objects $i \in I$ according the preceding rule, the number of errors (misclassified objects) is minimum ([1, 3]).

We denote:

- $N_\ell(\alpha)$ the number of objects belonging to class c_ℓ that are misclassified at level α, that is, the number of $i \in c_\ell$ with $\rho_{(\ell,\ell')}(i) > \alpha_{(\ell,\ell')}$;
- $N_{\ell'}(\alpha)$ the number of objects belonging to class $c_{\ell'}$ that are misclassified at level α, that is, the number of $i \in c_{\ell'}$ with $\rho_{(\ell,\ell')}(i) < \alpha_{(\ell,\ell')}$;
- $N_{(\ell,\ell')}(\alpha) = N_\ell(\alpha) + N_{\ell'}(\alpha)$ the number of objects of the two classes c_ℓ and $c_{\ell'}$ misclassified at level α.

The threshold $\widehat{\alpha}_{(\ell,\ell')}$ is the value α corresponding to the *minimum* of $N_{(\ell,\ell')}(\alpha)$.

Calculation Algorithm. To calculate $\widehat{\alpha}_{(\ell,\ell')}$, the values $\rho_{(\ell,\ell')}(i)$ are ranked in ascending order, hence the sequence indexed by j (with $1 \leq j \leq n_{c_\ell} + n_{c_{\ell'}}$), with:

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

- If $\alpha < \rho_{(\ell,\ell')}(1)$, then all objects are assigned to class $c_{\ell'}$, hence there are n_{c_ℓ} errors.
- If $\rho_{(\ell,\ell')}(j) < \alpha < \rho_{(\ell,\ell')}(J+1)$, there is one less error if the object corresponding to j belongs to class c_ℓ and one additional if it belongs to class $c_{\ell'}$, and so on.

We denote $j_{\min}$ the rank corresponding to the minimum of $N_{(\ell,\ell')}(\alpha)$, *i.e.* the rank of the object for which the ratio ρ is taken as threshold corresponding to the minimum number of misclassified objects.

We can choose for $\widehat{\alpha}$ a value between $\rho_{(\ell,\ell')}(j_{\min})$ and $\rho_{(\ell,\ell')}(j_{\min+1})$, for instance: $\alpha = \big(\rho_{(\ell,\ell')}(j_{\min}) + \rho_{(\ell,\ell')}(j_{\min}+1)\big)/2$

4 Application to the Survey Data of Trust Barometer (CEVIPOF)

4.1 Data set

The data come from surveys initiated by CEVIPOF (Centre de Recherches Politiques de Sciences-Po Paris) that take account of several, and sometimes heterogeneous, components of trust. The aim of these surveys is to measure changes in trust between 2009 and 2012 then 2013 and 2014 in France.

[4] In case of equality, we can choose the more numerous class.

Six waves of on–line surveys has been conducted each year since 2009 by the research center CEVIPOF in partnership with the *Pierre Mendès France Institute* and the *Conseil Economique, Social et Environnemental*. The samples (about 1500 persons) are designed to be representative of the French registered electors, by using the quota method (gender, age, CSP) and categorization by type of agglomeration and size of the home town. The data are collected by "OpinionWay" using a CAWI (Computer Assisted Web Interview) system (See www.cevipof.com/fr/le-barometre-de-la-confiance-politique-du-cevipof/).

4.2 Structure of the trust space

In order to measure changes in trust [4], we take as a reference the 2009 survey (intermediate date between presidential elections in 2007 and 2012). In the analyses, the 1375 individuals of wave 1 (2009) are put as active elements, the ones of the three other other waves are put as supplementary.

In order to construct the trust space and to make a typology of the French registered electors according to trust [4], we retained five components of trust measured by 24 questions:

1. *Political trust*: trust in political roles (7 questions);
2. *Institutional trust*: trust in large public or private institutions (5 questions);
3. *Economic trust*: trust in organizations of the economic world (4 questions);
4. *Inter-individual trust*: trust in neighbors, people, foreigners ... (5 questions);
5. *Individual trust*: feeling of personal happiness, personal responsability, trust in one's own future (3 questions).

The questions are almost all in the same format: a four–level Likert scale (with levels: much trust, some trust, not much trust, no trust at all). For constructing the trust space, we use a procedure called doubling, that is, we attribute two scores by individual instead of a single score. We respectively coded the four levels (3,0), (2,1), (1,2) and (0,3) for four–level scales and by (1,0), (O,1) for the two levels of the two dichotomous questions. Then the table is doubled with for each individual a "trusted pole" and a "untrusted pole". We performed a correspondence analysis of the table with 2×24 columns and 1375 rows. In the correspondence analysis display of this table, we obtain two points for each question and one point for each individual). The line joining two poles of one question is going through the origin (as shown in figure 1 for the question about trust in banks).

Furthermore, the number of questions of trust components being different, we have weighted each question by the inverse of the number of questions of its component. Thus the components are about equivalent; the contributions to the cloud variance are respectively 22%, 19%, 21%, 23% and 15%. We will give the interpretation of the first two axes[5].

[5] For more details, see [3].

axe	variance	taux cumulé
1	0.0721	23.80
2	0.0310	34.04
3	0.0229	41.59
4	0.0210	48.52
5	0.0162	53.88
6	0.0125	58.02

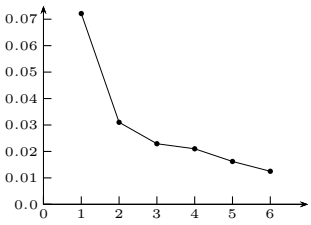


Table 1. Variances des axes ($> \overline{\lambda}$) et courbe de décroissance.

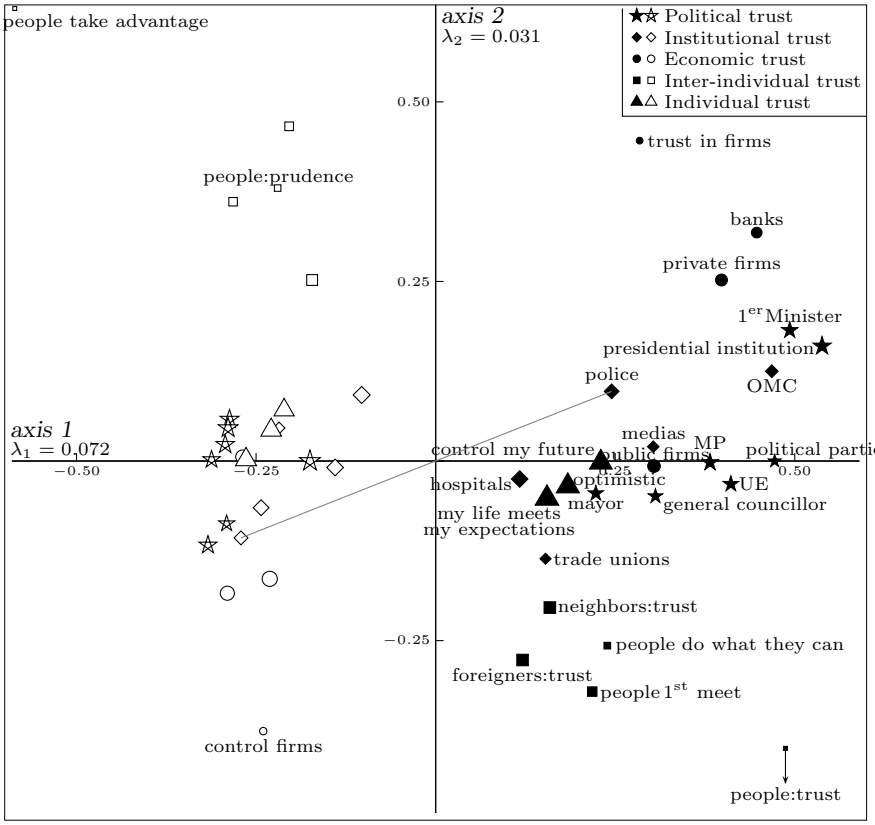


Fig. 1. Cloud of questions with their two poles: the trusted one with black markers and the distrusted one with white markers (the size markers are proportional to weights).

Interpretation of axes. The axis 1 is an axis of trust/distrust opposing the trusted poles of the questions (black markers) to the distrusted poles. The political, then the economical components of trust account for 34%+ 25% of the variance of axis 1.

For axis 2, the inter–individual component is predominant with a contribution of 63%, then the economic component contributes to 27%. It opposes, on

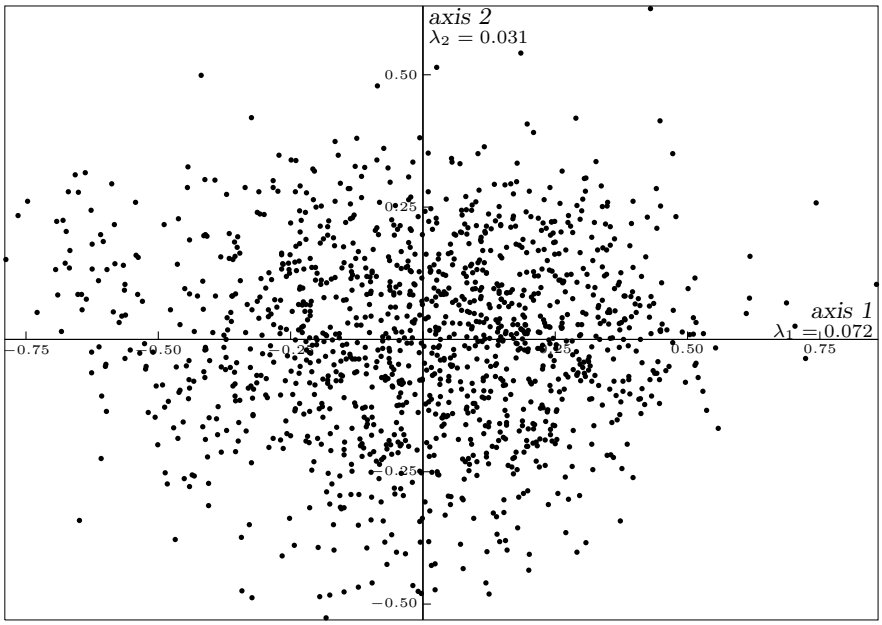


Fig. 2. Cloud of individual in plane 1-2 (the graphical scale is equal to 3/4 the one of figure 1).

the first hand (bottom in figure 1), an interpersonal trust (neighbors, foreigners, people meet for the fisrt time, ...) and a distrust for banks, firms, and on he other hand the opposite poles of these questions.

4.3 Clustering of individuals

On the cloud of individuals, we perform a Euclidean classification, precisely an AHC with variance criterion (Ward's method).

We can distinguish four groups of individuals as regards trust. The superior hierarchical tree associated with the partition in four clusters is given in Figure 3 and the concentration ellipses represented in the plane 1-2 of the trust space in Figure 4.

Interpretation of classes. Class $c1$ ($n_{c1} = 402$) is the one of the "hyper-trusters", class $c2$ ($n_{c2} = 396$) is the one of "moderate trusters", class $c3$ ($n_{c3} = 267$) the one of "moderate distrusters" and class $c4$ ($n_{c4} = 311$) is the one of "hyper-distrusters".

4.4 Assignment of supplementary individuals

To complement this study, we have assigned, using the procedure described above, the individuals of the other waves to the classes defined by the AHC

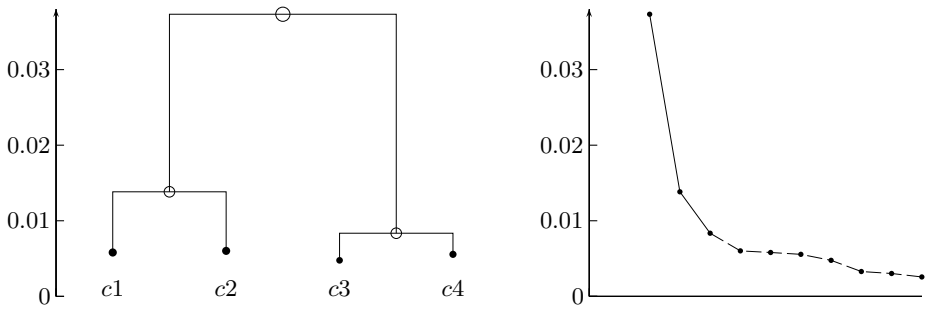


Fig. 3. Superior hierarchical tree and diagram of level indexes.

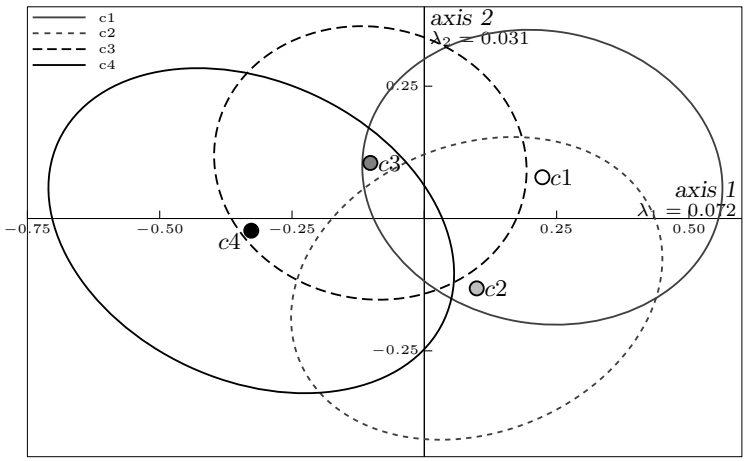


Fig. 4. Concentration ellipses of the 4 classes in plane 1-2.

of individuals of wave 1 (2009). The percentages of individuals in each class for waves 1 (2009), 4 (2012) and 5 (1013) are given in table 2.

Comment: The level of trust diminishes. The method used here enables us to specify the evolution: there exists an important shift of moderate trust to moderate distrust, and the evolution in the extreme classes is weak.

5 Conclusion

As we have seen, this method is of particular interest for the study of data tables indexed by time, that is for longitudinal studies. It can also be used in the case of a cloud of individuals equipped with structuring factors, for instance for comparing, as Pierre Bourdieu says, “positions in the field and position taking”.

classes	Dec 2009	Dec 2012	Dec 2013
$c1$ hyper–trusters	29	29	28
$c2$ moderate trusters	29	20	17
$c3$ moderate distrusters	19	31	34
$c4$ hyper–distrusters	23	20	21

Table 2. Percentages of individuals in each class for the 3 waves.

A calculation algorithm program was written in R, and it is invoked from SPAD software[6].

References

1. Benzécri, J.-P. (1977), Analyse discriminante et analyse factorielle, *Les cahiers de l'Analyse des Données*, 2 (4), 369-406.
2. Le Roux, B. (2004), *Geometric Data Analysis: from Correspondence Analysis to Structured Data Analysis*, Kluwer, Dordrecht.
3. Le Roux, B. (2014), *Analyse géométrique des données multidimensionnelles*, Dunod, Paris.
4. Le Roux, B. and 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.
5. Mahalanobis, P.C. (1936), On the generalized distance in statistics *Proceedings of the National Institute of Sciences (Calcutta)*, 2, 49-55.
6. Murtagh, F. (2005), *Correspondence Analysis and Data coding with Java and R*, Chapman and Hall, London.

[6] The analyses of the CEVIPOF barometer data were performed using SPAD (distributed by Coheris SPAD©).