Advances in Data Analysis and Classification
Volume 3, Number 3, December 2009, pages 205 to 225

Aggregation invariance in general clustering approaches

François Bavaud

Received: 14.5.2009 Revised: 15.10.2009 Accepted: 26.102009 Published online: 14.11.2009

Abstract General clustering deals with weighted objects and fuzzy memberships. We investigate the group- or object-aggregation-invariance properties possessed by the relevant functionals (effective number of groups or objects, centroids, dispersion, mutual object-group information, etc.). The classical squared Euclidean case can be generalized to non-Euclidean distances, as well as to non-linear transformations of the memberships, yielding the c-means clustering algorithm as well as two presumably new procedures, the convex and pairwise convex clustering. Cluster stability and aggregation-invariance of the optimal memberships associated to the various clustering schemes are examined as well.

Keywords Aggregation invariance · c-means clustering · EM algorithm · Fuzzy clustering · Model-based clustering · Mutual information.

Subject classification AMS 34K20, 62P99, 65K10, 82B26, 94A17, 94D05.

1 Introduction

Clustering consists in optimally partitioning objects, whose data vectors or pairwise distances are known, into groups. The same holds for *general clustering*, dealing with *weighted objects* and *fuzzy membership matrices*.

This paper deals with the time-honored quest for optimally clustering a *configuration* (π, X), where X denotes the locations of the objects (matrix of features vectors) and π their weights (relative mass). Quite simply, each of the n objects is assigned to one of the m possible groups. This *crisp* clustering can be generalized to *fuzzy* clustering that is described by a matrix Z of degrees of membership.

François Bavaud
University of Lausanne, 1015 Lausanne, Switzerland
E-mail: francois.bavaud@unil.ch

Natural as is might be, the explicit consideration of weighted objects is not that frequent in the literature; by contrast, this paper emphasizes the weighted, non-uniform configurations, and addresses the question of their fuzzy clustering.

Quite pragmatically, practitioners apply various clustering algorithms, and might face solutions with *equivalent* fuzzy groups (Definition 2), as in Table 1. One would like to know if *aggregating* such equivalent groups is legitimate or not, and changes or not the value of the relevant quantities associated to the algorithms.

The same issues arise in the case of an *aggregation of objects*. Seeking to answer those questions calls for a deliberately *formal set-up* that starts by defining four kinds of group or object equivalences (section 2 and 3). The paper pursues by investigating the aggregation-invariant properties of a few relevant functionals (section 4), among which the group centroids, the group dispersions and the object-group mutual information are the most prominent.

Section 5 investigates investigates various clustering algorithms, which aim at determining the optimal clusterings as local minima of a few relevant functionals, whose convexity and local stability are further studied. In particular, the thermodynamic clustering, resulting from the minimization of the free energy, is shown to be equivalent to the weighted version of the model-based clustering, under broad conditions. Also, a new clustering scheme, called convex clustering, is introduced as an attempt to adapt the c-means clustering to the weighted, aggregation invariant situation.

A few theorems emphasize the main results; the proof of the most straightforward ones is left to the reader.

2 Formalism and notations

2.1 Weights, membership degrees, group distributions and quotients

A *crisp clustering* is a partition of n objects $i = 1, \ldots, n$ into m disjoint groups $g = 1, \ldots, m$. In full generality, a *fuzzy clustering* f is specified by an association matrix (f_{ig}) with non-negative entries normalized to $\sum_i \sum_g f_{ig} = 1$, whose components formally define a joint distribution $p(i, g)$ (for $i = 1, ..., n, g = 1, ..., m$). The quantity $\pi_i := \sum_g f_{ig} = p(i)$ is the relative *weight* of object i, $\rho_g := \sum_i f_{ig} = p(g)$ is the relative weight of group g and $z_{ig} := f_{ig}/\pi_i = p(g|i)$ is the *membership degree* of object i in group g, with $\sum_g z_{ig} = 1$ and $\sum_i \pi_i z_{ig} = \rho_g$ for all i and g. The $n \times m$ *membership matrix* $Z = (z_{ig})$ is crisp iff $z_{ig} = 0$ or 1 for all i, g.

A group g is specified by the objects it contains, namely by the distribution $\pi_i^g := f_{ig}/\rho_g = p(i|g)$, $i = 1, ..., n$, with $\sum_i \pi_i^g = 1$. The *quotient* q_{ig} is defined as the ratio of the joint probability $p(i, g)$ over its formal value under

independence $p(i)p(g)$, namely $q_{ig} := f_{ig}/(\pi_i\rho_g) = \pi_i^g/\pi_i = z_{ig}/\rho_g = \pi_i\rho_g$. By construction, $\sum_i \pi_i q_{ig} = 1$, $\sum_g \rho_g q_{ig} = 1$ and

$$f_{ig} = \pi_i z_{ig} = \rho_g \pi_i^g = \pi_i \rho_g q_{ig} \, .$$

2.2 Feature-based and dissimilarity-based clustering

The association matrix (f_{ig}) records the proportion of occurrences of the various objects in the various groups and exhibits the kind of formal row-column duality encountered in Factor Correspondence Analysis (FCA).

However, in Data Analysis, only the configuration (π, X), given by the weights π_i of the objects $i = 1, \ldots, n$ and their p-dimensional *feature* or *location vectors* $x_i = (x_{i1}, ..., x_{ip})' \in \mathbb{R}^p$, collected together in the location matrix $X := (x_1, \ldots x_n)'$, is known.

Feature-based clustering consists in determining an optimal association matrix f from the given (π, X). In *dissimilarity-driven* clustering, coordinates do not show up, but a $p \times p$ pairwise dissimilarity matrix $d = (d_{ij})$ among objects is known (see e.g. Runkler 2007 and references therein). One assumes (d_{ij}) to be *proper*, that is obeying, for all $i, j = 1, \ldots, n$: $d_{ij} \geq 0$, $d_{ij} = d_{ij}$ and $d_{ij} = 0 \Rightarrow [d_{ik} = d_{jk}$ for all $k]$ (and hence $d_{ii} = 0$).

In this paper, optimal clusterings f are defined as solutions to minimization problems of the form $\min_f \mathcal{A}[f]$, for a given (π, X) or (π, d), where $\mathcal{A}[f]$ is an *objective functional* (clustering criterion; see sections 4 and 5).

The distinction between feature- and dissimilarity-based clustering becomes irrelevant for the squared Euclidean dissimilarity noted by $D_{ij} := (x_i - x_j)'(x_i - x_j) = \|x_i - x_j\|^2$: by multidimensional scaling (see the appendix), the object locations can indeed be recovered (up to a translation and rotation) from a squared Euclidean dissimilarity matrix.

In feature-based clustering, each group g is furthermore characterized by a *prototypical location* or *centroid* $y_g \in \mathbb{R}^p$, defined as the location y minimizing a certain *dispersion functional* depending upon π^g and X: see sections 4.4, 4.6, 5.3, 5.4.1 and 5.6.

3 Definitions: aggregation and equivalences

Definition 1 (Object and group aggregation)
a) Two distinct objects i and j from $\{1, \ldots, n\}$ with given locations $x_i, x_j \in \mathbb{R}^p$ can be aggregated, resulting in an object $[i \cup j]$ that has by definition

i) weight $\pi_{[i\cup j]} = \pi_i + \pi_j$, membership degrees $z_{[i\cup j]g} = (\pi_i z_{ig} + \pi_j z_{jg})/(\pi_i + \pi_j)$, quotients $q_{[i\cup j]g} = (\pi_i q_{ig} + \pi_j q_{jg})/(\pi_i + \pi_j)$ and group distributions $\pi^g_{[i\cup j]} = \pi_i^g + \pi_j^g$ for $g = 1, \ldots, m$
ii) location $x_{[i\cup j]} = (\pi_i x_i + \pi_j x_j)/(\pi_i + \pi_j)$.

b) Two distinct groups g and h from $\{1, \ldots, m\}$ can be aggregated, creating a group $[g \cup h]$ that has weight $\rho_{[g \cup h]} = \rho_g + \rho_h$, membership $z_{i[g \cup h]} = z_{ig} + z_{ih}$, quotient $q_{i[g \cup h]} = (\rho_g q_{ig} + \rho_h q_{jg})/(\rho_g + \rho_h)$ and distribution $\pi_i^{[g \cup h]} = (\rho_g \pi_i^g + \rho_h \pi_i^h)/(\rho_g + \rho_h)$.

Under group aggregation $g, h \rightarrow [g \cup h]$, the $n \times m$ membership matrix Z transforms into a $n \times (m-1)$ matrix $\tilde{Z}$, whose columns other than g or h remain unchanged. Under object aggregation $i, j \rightarrow [i \cup j]$, Z updates into a $(n-1) \times m$ matrix $\tilde{Z}$, whose rows other than i or j remain unchanged.

Definition 2 (Equivalences)

- *Group equivalence* : two groups g and h are said to be *group-equivalent*, noted $g \overset{g}{\sim} h$ (tolerating some notational collision), if their distributions coincide, that is if $\pi_i^g = \pi_i^h$ or equivalently if $q_{ig} = q_{ih}$ for all objects $i = 1, \ldots, n$.
- *Centroid equivalence* : given their centroids y_g and y_h, two groups g and h are said to be *centroid-equivalent*, noted $g \overset{c}{\sim} h$, if $y_g = y_h$.
- *Membership equivalence* : two objects i and j are said to be *membership-equivalent*, noted $i \overset{m}{\sim} j$, if their membership degrees coincide, that is if $z_{ig} = z_{jg}$ or equivalently if $q_{ig} = q_{jg}$ for all groups $g = 1, \ldots, m$.
- *Dissimilarity equivalence* : given a proper dissimilarity d, two objects i and j are said to be *dissimilarity-equivalent*, noted $i \overset{d}{\sim} j$, if $d_{ij} = 0$.

3.1 An illustration

Table 1 displays a few rows (14 out of $n = 55$ Swiss towns) of a membership matrix $Z = (z_{ig})$ with $m = 10$ groups, resulting from thermodynamic clustering (see section 5.2) applied to a matrix of Euclidean "inter-town commuter distances" (Bavaud 2006).

Thun	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.96	0.00	0.01
Luzern	0.00	0.30	0.06	0.13	0.20	0.00	0.00	0.01	0.00	0.30
Einsiedeln	0.00	0.30	0.06	0.13	0.20	0.00	0.00	0.00	0.00	0.30
Lachen	0.00	0.30	0.06	0.13	0.20	0.00	0.00	0.00	0.00	0.30
Schwyz	0.00	0.30	0.07	0.13	0.20	0.00	0.00	0.00	0.00	0.30
Stans	0.00	0.30	0.06	0.13	0.20	0.00	0.00	0.00	0.00	0.30
Zug	0.00	0.30	0.06	0.13	0.20	0.00	0.00	0.01	0.00	0.30
Bulle	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.91	0.06	0.01
Fribourg	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.97	0.00	0.01
Grenchen	0.00	0.03	0.01	0.01	0.02	0.00	0.00	0.88	0.00	0.03
Olten-Zofingen	0.00	0.29	0.06	0.12	0.19	0.02	0.00	0.03	0.00	0.29
Solothurn	0.00	0.07	0.02	0.03	0.05	0.01	0.00	0.75	0.00	0.07
Basel	0.00	0.03	0.01	0.01	0.02	0.91	0.00	0.00	0.00	0.03
Schaffhausen	0.00	0.30	0.07	0.13	0.20	0.00	0.00	0.00	0.00	0.30

Table 1: *14 rows of a* 55×10 *membership matrix* $Z = (z_{ig})$*. Columns associated to groups* $g = 2, 3, 4, 5, 10$ *are (exactly or nearly) proportional (group equivalence). Rows associated to objects* $i = 2, 7$ *(and also* $i = 3, 4$ *and* $i = 5, 14$*) are identical (membership equivalence).*

Table 1 illustrates a familiar situation in fuzzy clustering, where many groups ($m = 10$) have been provided for the starting step of a general clustering algorithm, but few non-equivalent groups ($M = 5$) persist in the final solution. Intuitively, such equivalent groups could and should be aggregated, without changing the value of the relevant quantities of interest.

Similarly, one would expect that joining objects with identical membership degrees should change neither the clustering problem nor its solution:

Definition 3 (Aggregation invariance (AI): gAI, cAI, mAI, dAI) Let $\mathcal{A}[f]$ be an arbitrary functional depending upon the $n \times m$ association matrix $f = (f_{ig})$. Then

- $\mathcal{A}$ is *group-aggregation invariant* (gAI) if $\mathcal{A}$ is unchanged under the aggregation of any two group-equivalent groups g, h, i.e. with $g \overset{g}{\sim} h$
- $\mathcal{A}$ is *centroid-aggregation invariant* (cAI) if $\mathcal{A}$ is unchanged under the aggregation of any two centroid-equivalent groups g, h, i.e. with $g \overset{c}{\sim} h$
- $\mathcal{A}$ is *membership-aggregation invariant* (mAI) if $\mathcal{A}$ is unchanged under the aggregation of any two m-equivalent objects i, j, i.e. with $i \overset{m}{\sim} j$
- $\mathcal{A}$ is *dissimilarity-aggregation invariant* (dAI) if $\mathcal{A}$ is unchanged under the aggregation of any two d-equivalent objects i, j, i.e. with $i \overset{d}{\sim} j$.

4 Clustering functionals and their aggregation invariance

In sections 4 and 5 we define the main functionals involved in clustering problems and check their aggregation invariance properties. Strictly speaking, those functionals should we written as $\mathcal{A}[f]$, but since the weights and locations of objects (or their dissimilarities) are known and fixed, they will be written, with a slight notational abuse, as $\mathcal{A}[Z]$, in view of the identity $f_{ig} = \pi_i z_{ig}$.

4.1 Effective number of groups

Determining the "right" number of groups m is a pervasive issue in clustering. Actually, the number $m[Z]$ of rows of Z is not gAI : m decreases under the aggregation of groups that are equivalent in the sense of Definition 2. (*Note:* here and in the sequel, we use the term "decrease" and "increase" in the weak sense.)

However, the *effective number of groups* defined by $M[Z] := \text{rank}(Z) \leq \min(n, m)$ is gAI as well as mAI. Also and by construction, $M[Z]$ decreases under aggregation of groups or objects in general. Furthermore, $M[Z] = m[Z]$ for crisp clusterings involving non-empty groups.

4.2 Number of distinct objects

The number $\texttt{ND}[X] \leq n$ of objects with *distinct* feature vectors from $\{x_1, \ldots, x_n\}$ is gAI and dAI by construction. In general, $\texttt{ND}[X]$ can decrease but also increase under object aggregation: think of four points i, j, k, l, with $x_i = x_k \neq x_j = x_l$ (two distinct locations); under aggregation $[i \cup j]$, the location $x_{[i \cup j]}$ might well differ from x_k and x_l, thus increasing $\texttt{ND}[X]$ (three distinct locations).

4.3 Cohesiveness

Let $O = \{1, ..., n\}$ denote the set of objects and $G = \{1, ..., m\}$ the set of groups, with corresponding entropies and joint entropy:

$$H(O) = -\sum_i \pi_i \ln \pi_i \qquad H(G) := -\sum_g \rho_g \ln \rho_g \qquad H(O,G) = -\sum_{ij} f_{ig} \ln f_{ig}$$

The mutual object-group information

$$\begin{aligned} I[Z] :&= I(O:G) = H(O) + H(G) - H(O,G) \equiv H_O[Z] + H_G[Z] - H_{O,G}[Z] \\ &= \sum_{ig} f_{ig} \ln \frac{f_{ig}}{\pi_i \, \rho_g} = \sum_{ig} \pi_i \, z_{ig} \ln \frac{z_{ig}}{\rho_g} = \sum_{ig} \pi_i \, \rho_g \, q_{ig} \ln q_{ig} \end{aligned} \tag{1}$$

is a measure of object-group dependence or *cohesiveness*. $I[Z]$ is readily seen to be gAI and mA, and generally decreases under the aggregation of objects or groups - unless the latter ones are equivalent.

The same can be said of the *chi-square* criterion that is a cohesiveness measure as well:

$$\chi^2[Z] := \sum_{ig} \frac{(f_{ig} - \pi_i \, \rho_g)^2}{\pi_i \, \rho_g} = \sum_{ig} \pi_i \, \rho_g \, (q_{ig} - 1)^2 \, .$$

The group entropy $H(G)$ is not gAI: it decreases under the aggregation of equivalent groups. Identity $I(O:G) = H(G)$ is readily seen to characterize crisp clusterings.

4.4 Dispersion and centroids (feature-based clustering)

In case the data have the form of an $n \times p$ feature matrix $X = (x_1, \ldots, x_n)' \in \mathbb{R}^{np}$, dissimilarities d_{ij} are typically computed by means of formulas such as $d_{ij} := d(x_i, x_j) = (\sum_{k=1}^p |x_{ik} - x_{jk}|^a)^b$ for a suitable choice of a and b. Let $y \in \mathbb{R}^p$ be an arbitrary location and consider the functional

$$\gamma_g[Z, y] := \sum_i \pi_i^g d(x_i, y) \, .$$

The *group centroid* y_g is defined as the location minimizing the latter criterion; the value of the minimum defines the *group dispersion* $\gamma_g[Z]$:

$$y_g[Z] := \arg\min_y \gamma_g[Z, y] \qquad \gamma_g[Z] := \min_y \gamma_g[Z, y] = \gamma_g[Z, y_g] \,. \tag{2}$$

Also, the *within-group dispersion* $\gamma_W[Z]$ and the *overall dispersion* $\gamma[Z]$ are defined as

$$\gamma_W[Z] := \sum_g \rho_g \gamma_g[Z] \qquad \gamma[Z] := \min_y \sum_{i=1}^n \pi_i d(x_i, y) =: \sum_{i=1}^n \pi_i d(x_i, y_0) \tag{3}$$

where $y_0 := \arg\min_y \sum_{i=1}^n \pi_i d(x_i, y)$ is the overall centroid of all n data points.

Theorem 1 (Aggregation invariance of dispersions and centroids)
a) *The group centroid $y_g[Z]$ is gAI, dAI and cAI.*
b) *The within-group dispersion $\gamma_W[Z]$ and the overall dispersion $\gamma[Z]$ are gAI, dAI and cAI.*
c) *The within-group dispersion $\gamma_W[Z]$ increases under group aggregation. In particular, $\gamma[Z] \geq \gamma_W[Z]$.*
d) *If $d(x, y)$ is of the form $h(x - y)$ with a convex and even function $h(\cdot)$, the group dispersion $\gamma_g[Z]$, the within-group dispersion $\gamma_W[Z]$ and the overall dispersion $\gamma[Z]$ decrease under object aggregation.*

Proof **a)** The first two assertions follow readily from definitions. To prove the cAI property, define $\lambda := \rho_g/(\rho_g + \rho_h) \in [0, 1]$. Then, assuming $y_g = y_h$,

$$\begin{aligned} \gamma_{[g\cup h]}[Z, y_{[g\cup h]}] &= \min_y \gamma_{[g\cup h]}[Z, y] \geq \lambda \min_y \gamma_g[Z, y] + (1-\lambda) \min_y \gamma_h[Z, y] \\ &= \lambda\gamma_g[Z, y_g] + (1-\lambda)\gamma_h[Z, y_g] = \gamma_{[g\cup h]}[Z, y_g] \,. \end{aligned} \tag{4}$$

b) Again, the first two assertions are readily proved; the cAI property follows from the identity $(\rho_g + \rho_h)\gamma_{[g\cup h]}[Z] = \rho_g y_g[Z] + \rho_h y_h[Z]$ found in (4).
c) Let $\gamma_W[\tilde{Z}]$ denote the within-group dispersion after groups g and h have been aggregated. Then $\gamma_W[\tilde{Z}] \geq \gamma_W[Z]$ iff $\gamma_{[g\cup h]}[Z] \geq \lambda\gamma_g[Z] + (1-\lambda)\gamma_h[Z]$ (first line in 4).
d) Let $\gamma[\tilde{Z}]$ denote the overall dispersion after objects i and j have been aggregated. Define $u(y) := \sum_{k\neq i,j} \pi_k d(x_k, y)$ and $x_{[i\cup j]} := (\pi_i x_i + \pi_j x_j)/(\pi_i + \pi_j)$. By convexity of h

$$\pi_i d(x_i, y) + \pi_j d(x_j, y) + u(y) \geq (\pi_i + \pi_j) d(x_{[i\cup j]}, y) + u(y) \,.$$

Applying $\min_y$ on both sides demonstrates $\gamma[Z] \geq \gamma[\tilde{Z}]$. The other claims are proved analogously. □

4.5 Pairwise dispersion (dissimilarity-driven clustering)

If the data consist of pairwise dissimilarities d_{ij} between objects, the dispersion in group g can be measured by the average *pair dissimilarity*:

$$\delta_g[Z] := \frac{1}{2}\sum_{ij} \pi_i^g \pi_j^g d_{ij} = \frac{1}{2}\sum_{ij} \pi_i \pi_j q_{ig} q_{jg} d_{ij}$$

which is dAI. The *within-group pair* dispersion

$$\delta_W[Z] := \sum_g \rho_g \delta_g[Z] = \frac{1}{2}\sum_{ij}\sum_g \rho_g \pi_i^g \pi_j^g d_{ij} \tag{5}$$

can be checked to be both gIA and dAI, as is the *overall dispersion*

$$\delta[Z] := \frac{1}{2}\sum_{ij}\sum_g \rho_g \pi_i \pi_j d_{ij} = \frac{1}{2}\sum_{ij} \pi_i \pi_j d_{ij} \; .$$

$\delta[Z]$ is trivially not affected by group aggregation, and one would expect $\delta_W[Z]$ to generally increase under group aggregation. This is, however, not the case, with the notable exception of squared Euclidean dissimilarities:

Theorem 2 (A new characterization of Euclidean distances) *The within-group pair dispersion $\delta_W[Z]$ increases under group aggregation for every Z iff d_{ij} is the squared Euclidean distance, noted $D_{ij} := \|x_i - x_j\|^2$.*

Proof Let $\delta_W[\tilde{Z}]$ denote the within-group pair dispersion after groups g and h have been aggregated. One readily finds $\delta_W[Z] - \delta_W[\tilde{Z}] = \frac{1}{2}\frac{\rho_g \rho_h}{\rho_g + \rho_h}\sum_{ij} s_i s_j d_{ij}$, where $s_i := \pi_i^g - \pi_i^h$. Hence $\delta_W[\tilde{Z}] \geq \delta_W[Z]$ for all Z iff $\sum_{ij} s_i s_j d_{ij} \leq 0$ for any s with $\sum_i s_i = 0$ (such a vector can indeed always be written as $s_i = a(\pi_i^g - \pi_i^h)$ where π_i^g and π_i^h are well-defined distributions and a is large enough). But the latter inequality, the *conditionally negative semi-definite condition*, is well-known to hold iff d_{ij} is a squared Euclidean distance D_{ij} (Schoenberg 1935; Blumenthal 1953). □

Theorem 3 (Huygens) *The within-group dispersion and pair dispersion coincide ($\gamma_W[Z] = \delta_W[Z]$) iff the given dissimilarity is a squared Euclidean distance ($d_{ij} = D_{ij}$).*

Proof Actually, Huygens' theorem traditionally covers the "if" part only; the "only if" part follows from Theorems (1) and (2). □

4.6 Gravity center and inertia (squared Euclidean dissimilarities)

Squared Euclidean dissimilarities, highlighted by Theorem 2, enjoy unique and well-known properties. First, the group dispersion $\gamma_g[Z]$ and the group pair dissimilarity $\delta_g[Z]$ coincide for all groups g and all clusterings Z, and define the *group inertia* $\Delta_g[Z]$ obeying *Huygens weak principle* (Bavaud 2002):

$$\Delta_g[Z] := \sum_i \pi_i^g D_i^g = \frac{1}{2}\sum_{ij} \pi_i^g \pi_j^g D_{ij} \qquad \text{where} \quad D_i^g := \|x_i - y_g\|^2 \,.$$

Also, the group centroid in (2) can be explicitly written as the *group gravity center*

$$y_g[Z] = \bar{x}^g := \sum_i \pi_i^g x_i = \sum_i \pi_i q_{ig} x_i \tag{6}$$

which is easily shown to be both mAI and dAI. The *overall gravity center*

$$y_0[Z] := \bar{x} := \sum_g \rho_g \bar{x}^g = \sum_{ig} \rho_g \pi_i q_{ig} x_i = \sum_i \pi_i x_i$$

does not depend on Z, and is thus completely aggregation-invariant, that is gAI, cAI, mAI and dAI.

Define $D_i^0 := \|x_i - y_0\|^2$ and $D^{g0} := \|y_g - y_0\|^2$, as well as $\Delta_B[Z] := \sum_g \rho_g D^{g0}$. Then the *overall* or *total inertia* exactly decomposes as:

$$\Delta[Z] := \sum_i \pi_i D_i^0 = \sum_g \rho_g D^{g0} + \sum_g \rho_g \Delta_g[Z] = \Delta_B[Z] + \Delta_W[Z] \,.$$

The following results can be proven by repeated use of the Huygens principle, or by explicit manipulation of the coordinates in the Euclidean distance:

Theorem 4 (Behavior of the inertia under aggregation)
a) *Under object aggregation $Z \to \tilde{Z}$ with $i, j \to [i \cup j]$, the between inertia $\Delta_B[Z]$ remains constant; the total inertia and group inertias* decrease *as*

$$\Delta[\tilde{Z}] = \Delta[Z] - \frac{\pi_i \pi_j}{\pi_i + \pi_j} D_{ij} \qquad\qquad \Delta_g[\tilde{Z}] = \Delta_g[Z] - \frac{\pi_i^g \pi_j^g}{\pi_i^g + \pi_j^g} D_{ij}$$

b) *Under group aggregation $Z \to \tilde{Z}$ with $g, h \to [g \cup h]$, the total inertia $\Delta[\tilde{Z}]$ remains constant; the within inertia* increases*, and the between inertia* decreases *as*

$$\Delta_W[\tilde{Z}] = \Delta_W[Z] + \frac{\rho_g \rho_h}{\rho_g + \rho_h} D^{gh} \qquad\qquad \Delta_B[\tilde{Z}] = \Delta_B[Z] - \frac{\rho_g \rho_h}{\rho_g + \rho_h} D^{gh}$$

where $D^{gh} := \|y_g - y_h\|^2$.

5 Optimum clustering algorithms: the functional approach

Given n objects together with their weights and locations or dissimilarities, a fuzzy clustering is determined by a membership matrix Z (see section 2.1). We shall investigate the aggregation invariance properties of a few *functional clustering schemes* that are aiming at determining an *optimal clustering*, i.e., that minimizes a given *objective functional* $\mathcal{A}[Z]$ (clustering criterion). The group labels being arbitrary, we assume $\mathcal{A}[Z]$ to be *symmetric*, i.e., invariant w.r.t. permutations of the columns of Z.

A necessary condition for Z to be a local minimum is that the first derivatives of $\mathcal{A}[Z]$ with respect to all z_{ig} are vanishing where the constraints $z_{ig} \geq 0$ (for all i, g) and $\sum_g z_{ig} = 1$ (for all i) are to be taken into account by suitable Lagrange multipliers λ_i. Another necessary condition is the *local stability condition* requiring the non-negativity of the quadratic form $(\epsilon, \mathcal{F}[Z]\epsilon)$ where $\mathcal{F}[Z]$ is the matrix of second derivatives of $\mathcal{A}[Z]$ and ϵ any admissible variation satisfying the constraints.

5.1 Optimum clustering for convex functionals

Before examining specific clustering schemes, let us derive simple results based upon the convex or concave nature of clustering functionals. Consider two $n \times m$ membership matrices Z_a and Z_b as well as their *mixture* $\lambda Z_a + (1-\lambda)Z_b$ where $0 \leq \lambda \leq 1$. A functional $\mathcal{A}[Z]$ is said to be *convex* if

$$\mathcal{A}[\lambda Z_a + (1-\lambda)Z_b] \leq \lambda \mathcal{A}[Z_a] + (1-\lambda)\mathcal{A}[Z_b]$$

for all Z_a, Z_b and $0 \leq \lambda \leq 1$. Equivalently, convex functionals decrease under mixing (and concave functionals increase under mixing). When smooth enough, the functional is convex if its *Hessian* matrix with components

$$\mathcal{A}[Z]_{ig,jh} := \frac{\partial^2 \mathcal{A}[Z]}{\partial z_{ig} \partial z_{jh}}$$

is semi-positive definite. Straightforward differentiation shows this to be the case for the mutual information with

$$I[Z]_{ig,jh} = \delta_{gh}[\frac{\pi_i \delta_{ij}}{z_{ig}} - \frac{\pi_i \pi_j}{\rho_g}] = \delta_{gh} \frac{\rho_g}{z_{ig} z_{jg}} [\delta_{ij} \pi_i^g - \pi_i^g \pi_j^g] \tag{7}$$

By contrast, the Hessian of the within-group inertia turns out to be

$$\Delta_W[Z]_{ig,jh} = \delta_{gh} \frac{\pi_i \pi_j}{\rho_g} [D_{ij} - D_i^g - D_j^g] = -2\delta_{gh} \frac{\pi_i \pi_j}{\rho_g} B_{ij}^g \tag{8}$$

which is negative semi-definite (B_{ij}^g is a matrix of scalar products that is explained in the appendix). In summary, the mutual information is convex, while the inertia is concave.

Theorem 5 (Fuzziness and convexity)
A convex functional $\mathcal{A}[Z]$ takes its minimum value on fuzzy memberships and its maximum on crisp memberships. The opposite holds for a concave functional.

Proof The result essentially follows from the observation that any fuzzy membership can be written as a mixture of crisp memberships (partitions), and that the set of admissible membership matrices $\mathcal{Z} := \{Z \,|\, z_{ig} \geq 0\,, \sum_g z_{ig} = 1\}$ is convex. The situation where the minimizer Z_0 of the convex functional $\mathcal{A}[Z]$ lies on the boundary of $\mathcal{Z}$ (containing the partitions) cannot be ruled out at first thought, but the symmetry of $\mathcal{A}[Z]$ implies that membership matrices $\dot{Z}_0$, obtained from Z_0 by an arbitrary permutation of the columns, constitue also minimizers, whose mixing with Z_0 ultimately yields truly fuzzy membership matrices, lowering the value of the functional by convexity. □

Mixing is a binary operation, associating to two $n \times m$ membership matrices a membership matrix of the same type, and must be distinguished from the *aggregation* of groups, which is an unary operation yielding a $n \times (m-1)$ membership matrix $\tilde{Z}$ from an $n \times m$ membership matrix Z. The two concepts are, however, closely related:

Theorem 6 (Aggregation of groups and convexity)
Any convex gAI functional $\mathcal{A}[Z]$ decreases under group aggregation. Any concave gAI functional $\mathcal{A}[Z]$ increases under group aggregation.

Proof Consider two columns g and h of Z. Let $\dot{Z}$ be the membership obtained from Z by permuting columns g and h, and let $\tilde{Z}$ result from the aggregation of g and h. Abbreviate $\mathcal{A}[Z]$ as $a(z_g, z_h)$ and $\mathcal{A}[\tilde{Z}]$ as $a(z_{[g \cup h]})$, the other columns being fixed.Then

$$\begin{aligned}\mathcal{A}[\tilde{Z}] = a(z_{[g\cup h]}) \;=\; a(z_g + z_h) &\overset{(a)}{=} a(\frac{1}{2}(z_g + z_h), \frac{1}{2}(z_g + z_h)) \\ &\overset{(b)}{\leq} \frac{1}{2}a(z_g, z_h) + \frac{1}{2}a(z_h, z_g)) \overset{(c)}{=} a(z_g, z_h) = \mathcal{A}[Z]\end{aligned}$$

where (a) follows from gAI of $\mathcal{A}$, (b) from the convexity and (c) from the symmetry $\mathcal{A}[\dot{Z}] = \mathcal{A}[Z]$. □

The mutual information $I[Z]$ (1) is gAI and convex: it decreases under aggregation of groups, a well-known result in Information Theory (see e.g. Cover and Thomas 1991).

The within-group inertia $\Delta_W[Z]$ (3) is concave and gAI: it takes on its minimum on crisp memberships (Theorem 6) and increases under aggregation of groups (Theorem 2).

The within-group pair dispersion $\delta_W[Z]$ (5) is gIA; since it does not increase under aggregation of groups in general (Theorem 2), it cannot be concave.

5.2 Thermodynamic clustering

5.2.1 Free energy

In statistical mechanics, equilibrium distributions emerge as a compromise between the opposite objectives of minimizing the *energy* (producing distributions concentrated in a ground state) and maximizing the *entropy* (producing uniform distributions). The conflict is arbitrated by the *temperature* $T \geq 0$, assessing the importance of the entropy term.

Transposed into the clustering context, those considerations lead to the functional scheme aiming at minimizing, w.r.t. Z, the gAI functional (see Rose et al. 1990; Rose 1998)

$$\mathcal{F}[Z] := I[Z] + \beta\, \gamma_W[Z] \tag{9}$$

which is proportional to the so-called *free energy* in thermodynamics. $I[Z]$ is the mutual information (1), $\gamma_W[Z]$ the within-group dispersion (3) and $\beta := 1/T$ is the inverse temperature. Minimizing $\gamma_W[Z]$ with a large enough initial number of clusters m tends to produce numerous crisp clusters tightly linked to objects, while minimizing $I[Z]$ favors a small number of fuzzy clusters scattered all over the objects. By contrast, both quantities (hence $\mathcal{F}[Z]$) decrease under object aggregation.

The free energy is made out of a mAI component and a dAI component, both gAI. Consequently, $\mathcal{F}[Z]$ is gAI only.

In the squared Euclidean case (see section 4.6), the minimizing conditions can be made explicit. Taking derivatives of (9) under the constraints $\sum_g z_{ig} = 1$ for all i yields, with Lagrange parameters λ_i,

$$\pi_i(\ln \frac{z_{ig}}{\rho_g} + \beta D_i^g) = \lambda_i \text{ for all } g, \qquad \text{i.e.} \qquad z_{ig} = \frac{\rho_g \exp(-\beta D_i^g)}{\sum_h \rho_h \exp(-\beta D_i^h)} \tag{10}$$

satisfying the constraints $z_{ig} \geq 0$ for all i, g. Let $Z(\beta)$ be the (supposed unique) membership matrix solution to (9), and define $R(\beta) := I[Z(\beta)]$ and $D(\beta) := \Delta_W[Z(\beta)]$. The parametric curve $(R(\beta), D(\beta))$ for $\beta = 1/T \geq 0$ is the *rate-distortion function* $R(D)$ of engineers and information theoreticians (see e.g. Cover and Thomas 1991 p. 362; Gray and Neuhoff 1998), which can also be written as:

$$R(D) := \min_{Z\ ;\ \Delta_W[Z] \leq D} I[Z]$$

$R(D)$ is decreasing, with $R(D) \equiv 0$ for $D \geq \Delta$ and $R(0) = H(O)$ (if all objects are distinct). In the usual information theoretical set-up, the centroids y_g (called *reproduction values* ; see e.g. Gray and Neuhoff 1998) are fixed, making $\Delta_W[Z]$ a *linear* function of the the memberships Z and thus ensuring the *convexity* of $R(D)$; the latter property cannot be guaranteed anymore in the present framework, due to the concavity of $\Delta_W[Z]$.

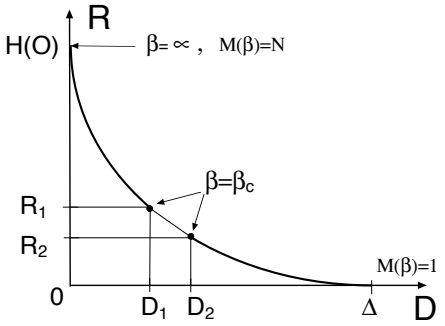

Fig. 1 Rate-distortion function $R(D)$, whose gap betrays a phase transition at critical inverse temperature β_c, equal to the slope $-(R_1 - R_2)/(D_2 - D_1)$. The limit $\beta = 1/T \to \infty$ creates as many groups as (distinct) objects, while β small enough yields a single group $M = 1$.

The existence of two minima $Z_1(\beta)$ and $Z_2(\beta)$ for (9) for some *critical inverse temperature* $\beta = \beta_c$, such that $R_1 > R_2$, $D_1 < D_2$ (and $R_1 + \beta_c D_1 = R_2 + \beta_c D_2$) characterizes a clustering *phase transition*, and leads to a gap in the rate distortion function (see Fig. 1 and section 5.2.2).

5.2.2 Phase transition

When the temperature T decreases, i.e. for an increasing $\beta = 1/T$, it may happen that formerly locally stable clusters cease to be so, and split into smaller clusters, thus increasing the effective number of groups. In the squared Euclidean case, this *phase transition* is controlled by λ_1^g, the projection of the group inertia Δ_g onto the first principal factor (see the appendix), as made explicit by the following Theorem:

Theorem 7 (Locally stable membership matrices) *A sufficient condition for local stability in thermodynamic clustering is* $\max_g \lambda_1^g \leq T/2$. *Beyond this limit, clusters may become unstable and fragment into smaller groups.*

Proof For a fixed temperature T, the membership matrix Z attains a local minimum iff it is stable with respect to perturbations of the form $z_{ig} \to z_{ig} + \epsilon_{ig}$ (with *admissibility conditions* $\sum_g \epsilon_{ig} = 0$ and $\epsilon_{ig} \geq -z_{ig}$), that is iff the Hessian of the free energy $\mathcal{F}[Z]$ satisfies (in view of (7) and (8))

$$\frac{1}{2} \sum_{ig,jh} \mathcal{F}_{ig,jh}[Z] \epsilon_{ig} \epsilon_{jh} = \frac{1}{2} \sum_{ijg} \left(\frac{\pi_i \delta_{ij}}{z_{ig}} - \frac{\pi_i \pi_j}{\rho_g} - 2\beta \frac{\pi_i \pi_j}{\rho_g} B_{ij}^g \right) \epsilon_{ig} \epsilon_{jg} \;\geq\; 0 \,.$$

Using the eigen-decomposition of Theorem 11, together with $\sum_{\alpha=1}^{n} u_{i\alpha}^g u_{j\alpha}^g = \delta_{ij}$ and $u_{in}^g u_{jn}^g = \sqrt{\pi_i^g \pi_j^g}$, the condition becomes

$$0 \leq \sum_{ijg} \frac{1}{\sqrt{\pi_i^g \pi_j^g}} \frac{\pi_i \pi_j}{\rho_g} (\delta_{ij} - \sqrt{\pi_i^g \pi_j^g} - 2\beta K_{ij}^g) \epsilon_{ig} \epsilon_{jg} = \sum_g \frac{1}{\rho_g} \sum_{\alpha=1}^{n-1} (1 - 2\beta \lambda_\alpha^g) (e_{\alpha g})^2 \quad (11)$$

where $e_{\alpha g} := \sum_i \frac{\pi_i}{\sqrt{\pi_i^g}} u_{i\alpha}^g \epsilon_{ig}$. Hence the condition $\max_{\alpha,g} \lambda_\alpha^g \leq 1/2\beta$, ensuring (11), is sufficient for local stability. Whether this condition is necessary or not depends on the proof of the existence of a matrix $(e_{\alpha g})$ associated to an admissible perturbation matrix (ϵ_{ig}), such that the inequality in (11) can be reverted whenever $\max_g \lambda_1^g > T/2$ - a question so far unsettled in full generality. □.

As a corollary, the 1-group solution is locally stable for $T > 2\Delta$.

Theorem 7 appears in the unweighted setting in Rose (1998) (see also Rose et al. (1990)), whose proof uses a perturbation of the gravity center y_g, instead of a perturbation of the membership matrix.

The variant consisting in minimizing the non-gAI functional $-H_{O,G}[Z] + \beta\Delta_W[Z]$, or equivalently $-H_{G|O}[Z]+\beta\Delta_W[Z]$ (first considered in Rose (1998)), yields to optimal solutions of the form $z_{ig} = \exp(-\beta D_i^g)/\sum_h \exp(-\beta D_i^h)$, assigning an a priori uniform weight on the groups.

5.3 The weighted mixture model

Fuzzy membership matrices appear naturally in the maximum likelihood (ML) approach to the mixture model. Let us review this well-known topic (see e.g. Celeux and Govaert 1992; McLachlan and Krishnan 1997) in the weighted clustering setting.

Let $\{f(x|\theta)\}_{\theta\in\Theta}$ be a parametric family of probability density functions normalized by $\int_{\mathbb{R}^p} f(x|\theta)\, dx = 1$. Let $f_g(x) := f(x|\theta_g)$ denote the density associated to group g. Consider m groups with proportions $r = (r_1, \dots r_m)$ (with $r_g \geq 0$ and $\sum_{g=1}^m r_g = 1$) and parameters $\theta = (\theta_1, \dots \theta_m)$, forming the *mixture density* $\bar{f}(x|\theta, r) := \sum_g r_g\, f_g(x)$.

Consider N independent observations of a corresponding random vector $\tilde{X} \in \mathbb{R}^p$ that may take n distinct values $x_1, ..., x_n \in \mathbb{R}^p$ $(n \leq N)$. Let $N_i \geq 1$ be the number of observations located at x_i, for $i = 1, \dots n$, and define $\pi_i := N_i/N$. The probability density of the data (π, X) under the mixture model is $P(\pi, X, r, \theta) = \prod_{i=1}^n \bar{f}(x_i|\theta, r)^{N_i}$. Up to a factor N, the log-likelihood is

$$\mathtt{LL}(\theta, r|\pi, X) = \sum_i \pi_i \ln \bar{f}(x_i|\theta, r)\ .$$

Maximizing the log-likelihood w.r.t. (r, θ) for a fixed X amounts to determining the mixture of m distributions that optimally fits the object locations X, given their weights π.

If the objects were in addition *labelled*, that is if the counts N_{ig} = "number of times an object with location x_i is attributed to group g" were observable,

then associations $F = (f_{ig}) := (N_{ig}/N) = (\pi, Z)$ should also be taken into account and the relevant functional would be the *complete log-likelihood*

$$\mathtt{CLL}(r, \theta|\pi, Z, X) := \sum_{ig} \pi_i z_{ig} \ln(r_g f_g(x_i)) =$$
$$\sum_g \rho_g \ln r_g + \sum_g \rho_g \sum_i \pi_i^g \ln f_g(x_i) =: \phi(r, Z) + \psi(\theta, Z) .$$

Amazingly, determining the optimal mixture (r, θ) amounts to determining an optimal membership matrix Z:

Theorem 8 (Optimal mixture model and optimal membership) *Consider a a fixed configuration (π, X) and the clustering criterion*

$$\mathcal{M}[Z] := I[Z] + \eta_W[Z]$$

with

$$\eta_W[Z] := \sum_{g=1}^{m} \rho_g \eta_g[Z],\ \eta_g[Z] := \min_\theta \eta_g[Z, \theta],\ \eta_g[Z, \theta] := \sum_i \pi_i^g(-\ln f(x_i|\theta)) \tag{12}$$

Then

$$\max_{r,\theta} \mathtt{LL}(r, \theta) = -\min_Z \mathcal{M}[Z].$$

Proof Consider the alternative membership matrix $T(r, \theta)$ with components $t_{ig} := r_g f_g(x_i)/\bar{f}(x_i) \geq 0$ obeying $\sum_g t_{ig} = 1\ \forall i$, as well as the *cross-entropy*

$$\mathtt{CE}(Z, T) = \mathtt{CE}(Z, T(r, \theta)) := -\sum_{ig} \pi_i z_{ig} \ln t_{ig} .$$

Then

$$\mathtt{LL}(r, \theta) = \mathtt{CLL}(r, \theta|Z) + \mathtt{CE}(Z, T(r, \theta)) = \phi(r, Z) + \psi(\theta, Z) + \mathtt{CE}(Z, T(r, \theta)) \tag{13}$$

Consider two sets of parameters $(r^{(n)}, \theta^{(n)})$ and $(r^{(n+1)}, \theta^{(n+1)})$ together with a membership matrix $Z^{(n)} := T^{(n)} = T(r^{(n)}, \theta^{(n)})$. By (13)

$$\mathtt{LL}(r^{(n+1)}, \theta^{(n+1)}) - \mathtt{LL}(r^{(n)}, \theta^{(n)}) = [\mathtt{CLL}(r^{(n+1)}, \theta^{(n+1)}|T^{(n)}) - \mathtt{CLL}(r^{(n)}, \theta^{(n)}|T^{(n)})] + [\mathtt{CE}(T^{(n)}, T^{(n+1)}) - \mathtt{CE}(T^{(n)}, T^{(n)})] .$$

Choosing $(r^{(n+1)}, \theta^{(n+1)}) := \arg\max_{r,\theta} \mathtt{CLL}(r, \theta|T^{(n)})$ makes the first difference in the r.h.s. positive. The second difference in the r.h.s. is positive by virtue of inequality $\mathtt{CE}(Z, T) \geq \mathtt{CE}(Z, Z) = H_{G|O}[Z]$, valid for all membership

matrices Z and T. In summary, starting form any initial value $(r^{(0)}, \theta^{(0)})$ of the parameters, the "expectation-maximization" (EM) iteration scheme

$$z_{ig}^{(n)} := t_{ig}(r^{(n)}, \theta^{(n)}) = \frac{r_g^{(n)} f(x_i|\theta_g^{(n)})}{\bar{f}^{(n)}(x_i)} \qquad \text{E-step}$$

$$(r^{(n+1)}, \theta^{(n+1)}) := \arg\max_{r,\theta} \mathtt{CLL}(r, \theta|Z^{(n)}) \qquad \text{M-step}$$

increases the value of the log-likelihood, and hence converges towards a local maximum (r^*, θ^*). Let $Z^* := T(r^*, \theta^*)$. Then (13) yields $r_g^* = \rho_g^*$ and $\phi(r^*, Z^*) = \sum_g \rho_g^* \ln \rho_g^* = -H_G[Z^*]$, and finally, together with (12):

$$\begin{aligned} \mathtt{LL}(r^*, \theta^*) &= \phi(r^*, Z^*) + \psi(\theta^*, Z^*) + \mathtt{CE}(Z^*, Z^*) \\ &= -H_G[Z] - \eta_W(Z^*) + H_{G|O}[Z] = -I[Z^*] - \eta_W[Z^*]\,. \quad \square \end{aligned}$$

The aggregation invariance properties of the functional $\eta_g[Z, \theta]$ are entirely similar to those of the functional $\gamma_g[Z, y]$, as stated in Theorem 1. In particular, the properties a), b) and c) hold here as well (after the substitution $y \to \theta$). Also, $\eta_g[Z]$ decreases under object aggregation if $f(x|\theta)$ is log-concave in its first argument (see Theorem 1d) and its proof).

Model-based and thermodynamic clusterings coincide iff $\eta_g[Z, \theta] = \gamma_g[Z, y]$, that is iff $\theta \equiv y$ and $f_g(x) = f(x|y_g) = h(x - y_g)$ where $h(x)$ is symmetric and maximum at $x = 0$: in this class of distributions, called *location models*, the function $d(x, y) := -\ln(h(x - y))/h(0))$ constitutes a proper dissimilarity.

The class of *homogeneous location models* contains the distributions of the form $f_g(x) = \sigma^{-p} h((x - y_g)/\sigma)$, with common adjustable dispersion $\sigma > 0$. Application of the EM clustering algorithm collapses in the over-parameterized case where the number m of available groups equals (or exceeds) the number $\mathtt{ND}[X]$ of objects possessing distinct features (proof: attach each distinct object i to its "own private group" $g[i]$, hat is $\pi_i^g = \delta_{g,g[i]}$. Setting $y_{g[i]} = x_i$ yields $\min_Z \mathcal{M}[Z] \leq H_O[Z] - \ln f(0) + p \ln \sigma \to -\infty$ for $\sigma \to 0$). The same collapse occurs for *heterogeneous location models* of the form $f_g(x|\theta_g) = \sigma_g^{-p} f((x - y_g)/\sigma_g)$, as soon as $m \geq 2$.

5.4 Convex clustering

Before examining the weighted c-means algorithm (section 5.6), we introduce a presumably new family of *convex dispersion* functionals whose minimization defines a *convex clustering* procedure. Those functionals depend upon a smooth non-negative strictly convex, strictly increasing function $c(q)$ with derivative $c'(q)$. We shall make use of the function $r(\cdot)$ obeying $r(c'(q)) = q$ (inverse of $c'(q)$) which exists and is unique by construction.

5.4.1 Aggregation properties and optimal memberships

Let $D(x,y) = \|x-y\|^2$ denote the squared Euclidean distance and consider a fixed configuration (π, X). Define the *convex dispersion* $\kappa_g[Z]$ for group g as

$$\kappa_g[Z] := \min_y \kappa_g[Z,y] =: \kappa_g[Z,y_g] \qquad \kappa_g[Z,y] := \sum_i \pi_i c(q_{ig}) D(x_i,y)$$

$$\text{where} \qquad y_g[Z] = \sum_i \hat{\pi}_i^g x_i \qquad \text{with} \qquad \hat{\pi}_i^g := \frac{\pi_i c(q_{ig})}{\sum_j \pi_j c(q_{jg})} \,. \quad (14)$$

Theorem 9 (Aggregation-invariance of convex functionals)
a) The centroid $y_g[Z]$ is mAI but not dAI.
b) The convex dispersion $\kappa_g[Z]$ is neither dAI nor mAI. It decreases under aggregation of objects.
c) The within-group convex dispersion $\kappa_W[Z] := \sum_g \rho_g \kappa_g[Z]$ *is gAI.*

Proof a) Let $i \overset{m}{\sim} j$. Define $\tau_g[Z] := \sum_i \pi_i \, c(q_{ig})$. Aggregating i and j leaves $y_g[Z]$ unchanged in view of the identities

$$\frac{\hat{\pi}^g_{[i\cup j]}}{\pi^g_{[i\cup j]}} = \frac{\hat{\pi}^g_i}{\pi^g_i} = \frac{\hat{\pi}^g_j}{\pi^g_j} = \frac{c(q_{[i\cup j]g})}{\tau_g[Z]} \,.$$

On the other hand, $\hat{\pi}^g_{[i\cup j]} \neq \hat{\pi}^g_i + \hat{\pi}^g_j$ in general, thus proving the second assertion.

b) Consider a group g and two objects i,j with quotients q_{ig}, q_{jg} and locations x_i, x_j, the other objects being fixed. Then

$$\kappa_g(q_{ig}, q_{jg}, x_i, x_j) \overset{(a)}{\geq} \kappa_g(q_{ig}, q_{jg}, x_{[i\cup j]}, x_{[i\cup j]}) \overset{(b)}{\geq} \kappa_g(q_{[i\cup j]g}, q_{[i\cup j]g}, x_{[i\cup j]}, x_{[i\cup j]})$$

where (a) follows from the convexity of $D(x,y)$, and (b) follows from the convexity of $c(q)$.

c) Immediate from the definition. □

Theorem 10 (Optimal convex membership)
The membership matrix Z minimizing $\kappa_W[Z]$ fulfills

$$c'(q_{ig}) D(x_i, y_g) = b_i + a_g[Z] - \kappa_g[Z]$$

i.e.,

$$z_{ig} = \rho_g \, r\Big(\frac{b_i + a_g[Z] - \kappa_g[Z]}{D_{ig}}\Big) \quad (15)$$

where $a_g[Z] := \sum_j \pi_i q_{ig} c'(q_{jg}) D(x_j, y_g)$ and b_i is determined by the normalization condition $\sum_g z_{ig} = \sum_g \rho_g q_{ig} = 1$.

Proof Differentiating κ_W under the above constraint, and taking into account

$$z_{ig} = \rho_g q_{ig} \quad \Rightarrow \quad \frac{\partial}{\partial z_{ig}} = \sum_k \frac{\partial q_{kg}}{\partial z_{ig}} \frac{\partial}{\partial q_{kg}} = \frac{1}{\rho_g}\left(\frac{\partial}{\partial q_{ig}} - \pi_i \sum_k q_{kg} \frac{\partial}{\partial q_{kg}}\right)$$

as well as $\sum_j \pi_j c(q_{ig})(x_j - y_g)' \partial y_g / \partial q_{ig} = 0$ (see 14) yields

$$\lambda_i = \frac{\partial \kappa_W}{\partial z_{ig}} = \pi_i \left(\kappa_g[Z] + c'(q_{ig}) D(x_i, y_g) - a_g[Z]\right) .$$

Setting $b_i := \lambda_i / \pi_i$ achieves the proof. The constraint $z_{ig} \geq 0$ is satisfied due to the positivity of $r(\cdot)$. □

5.5 Examples

The case $c(q) = q^2/2$ entails $r(q) = q$, $a_g[Z] = 2\kappa_g[Z]$, and $b_i = (1 - \sum_g \frac{\rho_g \kappa_g[Z]}{D_{ig}}) / \sum_h \frac{\rho_h}{D_{ih}}$. The design of an iterative scheme intended to converge to (15) is obvious; we will not tackle this question, nor study the stability of the solution in the present paper.

The choice $c(q) = q$ yields $\kappa_W[Z] = \Delta_W[Z]$, the within-group dispersion, whose minimum it attained for crisp memberships (section 5.1). The function $r()$ is not defined, and Theorem 10 does not apply. The same holds for $c(q) \equiv 1$, yielding $\kappa_W[Z] = \Delta[Z]$, the total-group dispersion, which does not depend upon the membership.

5.6 Weighted c-means

We now turn to the question of generalizing the *c-means clustering* to the weighted case. The latter actually obtains by defining

$$\mathcal{J}_g[Z, y] := \sum_i \pi_i c(z_{ig}) D(x_i, y) \quad \text{and} \quad \mathcal{J}_g[Z] := \min_y \mathcal{J}_g[Z, y] = \mathcal{J}_g[Z, y_g]$$

$$\text{with solution} \qquad y_g = \sum_i \check{\pi}_i^g x_i \qquad\qquad \text{with} \qquad \check{\pi}_i^g = \frac{\pi_i c(z_{ig})}{\sum_j \pi_j c(z_{jg})} .$$

Minimizing further $\mathcal{J}_W[Z] := \sum_g \mathcal{J}_g[Z]$ with respect to Z yields the solution (compare with (15)):

$$z_{ig} = r\left(\frac{b_i}{D_{ig}}\right)$$

where b_i is fixed such as $\sum_g z_{ig} = 1$. In particular, the algebraic case $c(q) = Cq^B$ with $C > 0$ and $B > 1$ yields the celebrated "fuzzy c-means clustering" (see e.g. Bezdek 1981 or Miyamoto et al. 2008)

$$z_{ig} = \frac{D_{ig}^{-\frac{1}{B-1}}}{\sum_h D_{ih}^{-\frac{1}{B-1}}} = \frac{1}{\sum_h \left(\frac{D_{ig}}{D_{ih}}\right)^{\frac{1}{B-1}}}$$

while the exponential case $c(q) = C\exp(Aq)$ with $C, A > 0$ gives

$$z_{ig} = \frac{1}{m} + \frac{1}{Am}\sum_{h=1}^{m} \ln \frac{D_{ih}}{D_{ig}} \quad . \tag{16}$$

As before, the centroids y_g are mAI but not dAI. But $\mathcal{J}_W[Z]$ is not gAI, contrarily to $\kappa_W[Z]$. In other words, aggregating two equivalent groups generally modifies the value of $\mathcal{J}[Z]$, which is certainly posing a difficulty of interpretation, irrespectively of the algorithmic merits of the c-means.

5.7 Convex pairwise clustering

In the spirit of sections 4.5 and 5.4, define the average *convex pairwise dissimilarity*

$$\nu_g[Z] := \frac{1}{2}\sum_{ij} \pi_i \pi_j c(q_{ig}) c(q_{jg}) d_{ij} \qquad \text{and} \qquad \nu_W[Z] := \sum_g \rho_g \nu_g[Z] \,.$$

Minimizing $\nu_W[Z]$ under the constraint $\sum_g z_{ig} = 1$ yields (see section 5.4)

$$\lambda_i = \frac{\partial \nu_W[Z]}{\partial z_{ig}} = \pi_i(\nu_g[Z] + c'(q_{ig})E_{ig}[Z] - a_g[Z])$$

$$\text{that is} \qquad z_{ig} = \rho_g \, r(\frac{b_i + a_g[Z] - \nu_g[Z]}{E_{ig}[Z]})$$

with $E_{ig}[Z] := \sum_j \pi_j c(q_{jg}) d_{ij}$ and $a_g[Z] := \sum_k \pi_k q_{kg} c'(q_{kg}) E_{kg}[Z]$.

6 Conclusion

Clustering problems are a nice instance of the interplay between the *geometric content* and the *probabilistic content*, typical of Data Analysis. This paper has underlined four classes of functionals involved in general clustering, namely:

- *centroids* y_g, representing the central, typical location of a group g, or the optimal model parameter θ_g associated with group g.
- *dispersions* γ_g, Δ_g or η_g measuring the average dissimilarity between the locations of the objects in the group and the centroid (or measuring the average dissimilarity δ_g between the pairs of objects of the group g).
- *entropies* H_O, H_G, $H_{O,G}$ or $I_{O:G}$, respectively denoting the object uncertainty, the group uncertainty, the object-group association uncertainty and the object-group dependence. While the previous classes are of geometric nature, entropy functionals express the probabilistic aspects of the clustering problem, and measure which groups are made out of which objects, and conversely.
- *"mixed functionals"* such as κ_g, $\mathcal{J}_g$ and ν_g, of twofold nature, interweaving geometric and probabilistic content.

Discussing the aggregation-invariance properties of algorithms and functionals necessitates a general clustering approach: aggregating objects requires the consideration of weighted objects, and aggregating groups requires the the consideration of fuzzy memberships. This main theme was developped throughout this paper, whose main results are summarized in eleven theorems, among which Theorems 1,2,4,5,6,9 and 10 are presumably new, as is the proof of Theorem 7.

Among open issues, let us mention the question

- of the consequences of aggregating groups which are only nearly equivalent
- of the stability of the convex clustering and other variants
- of the convexity of the associated "mixed functionals", under various families of functions $c(q)$
- of the numerical testing of alternative weighted c-means procedures, such as the "exponential c-means" of equation (16), as well as of the convex clustering procedure of section 5.4, or its pairwise version of section 5.7.

The last issue is of practical rather than formal nature, and should be given the highest priority if one wants to determine whether or not convex clustering can be recommanded to practionners. Although convex clustering possesses better aggregation-invariance properties than the c-means algorithm, which is defective in that respect (section 5.6), the question of its numerical behavior remains so far open.

Acknowledgments: Particularly detailed remarks and helpful suggestions from the Editor as well as from three anonymous Reviewers have led to substantial improvements in the manuscript and are most gratefully acknowledged.

References

- Bavaud F (2002) Quotient dissimilarities, Euclidean embeddability, and Huygens' weak principle. In: K. Jaguja, A.Sokolowski, H.-H. Bock (eds) Classification, Clustering and Data Analysis. Springer, New York, pp. 194-202
- Bavaud F (2006) Spectral clustering and multidimensional scaling: a unified view. In: V. Batagelj, H.-H. Bock H-H, A. Ferligoj, A. Ziberna (eds) Data Science and Classification. Springer, New York, pp 131-139
- Bezdek D (1981). Pattern Recognition. Plenum Press, New York
- Blumenthal L M (1953) Theory and applications of distance geometry. University Press, Oxford.
- Celeux G, Govaert G (1992) A classification EM algorithm and two stochastic versions. Computational Statistics and Data Analysis 14: 315-332
- Cover T M, Thomas J A (1991). Elements of Information Theory. Wiley, New York
- Gray R M, Neuhoff D L (1998) Quantization. IEEE Transactions on Information Theory 44: 2325 - 2383
- Miyamoto S, Ichihashi H, Honda K (2008) Algorithms for Fuzzy Clustering: Methods in c-Means Clustering with Applications, Springer, New York
- McLachlan G J, Krishnan T (1997) The EM algorithm and extensions. John Wiley, New York
- Rose K (1998) Deterministic Annealing for clustering, compression, classification, regression, and related optimization problems. Proceedings of the IEEE, 86: 2210-2239

- Rose K, Gurewitz E, Fox G C (1990) Statistical mechanics and phase transitions in clustering. Phys. Rev. Lett. 65: 945-948
- Runkler T A (2007) Relational fuzzy clustering. In: J. Valente de Oliveira, W. Pedrycz (eds) Advances in fuzzy clustering and its applications. John Wiley, Chichester, pp 31-52
- Schoenberg I J (1935) Remarks to Maurice Fréchet's article "Sur la définition axiomatique d'une classe d'espaces vectoriels distancés applicables vectoriellement sur l'espace de Hilbert". Annals of Mathematics 36: 724–732

Appendix: weighted, group-specific multidimensional scaling

Theorem 7 above makes essential use of the quantity $\lambda_1^g = \max_\alpha \lambda_\alpha^g$, controlling the local stability of the thermodynamic clustering. Here $\lambda_\alpha^g[Z]$, measuring the *proportion of the group inertia* $\Delta_g[Z]$ expressed in dimension α, denotes the α-th eigenvalue associated to the multidimensional scaling problem (MDS) in its weighted, group-specific version (that is attached to group g), where $\alpha = 1, \ldots, n$, and $\lambda_\alpha^g = 0$ if $\alpha \geq p$, where p is the dimensionality of the object features.

For definition sake, and given the scarcity of expositions of weighted MDS in the scientific literature, we present the main results of weighted, group-specific MDS in the following Theorem, whose proof follows the usual (unweighted) steps (see Bavaud 2006).

Theorem 11 (weighted, group-specific MDS)
a) *Consider a fixed group group g of n objects with data points x_i, weights π_i^g, and gravity center y_g (6). Define $D_{ij} := \|x_i - x_j\|^2$ and $D_i^g := \|x_i - y_g\|^2$. The* $n \times n$ matrix of scalar products *with components $B_{ij}^g := \frac{1}{2}(D_i^g + D_j^g - D_{ij}) = (x_i - y_g)'(x_j - y_g)$ is* positive semi-definite*, and so is the matrix K^g with elements $K_{ij}^g := \sqrt{\pi_i^g \pi_j^g}\, B_{ij}^g$. Let $K^g = U^g \Lambda^g (U^g)'$ be its decomposition, where $U^g = (u_{i\alpha}^g)$ is orthogonal and $\Lambda^g = diag(\lambda_\alpha^g)$ is diagonal with decreasingly ordered eigenvalues $\lambda_1^g \geq \lambda_2^g \geq \ldots \geq 0$, where $\lambda_n^g = 0$ is associated to the eigenvector $u_{in}^g = \sqrt{\pi_i^g}$. Then*

$$D_{ij} = \sum_{\alpha \geq 1} (x_{i\alpha}^g - x_{j\alpha}^g)^2 \qquad \textit{where} \quad x_{i\alpha}^g := \frac{\sqrt{\lambda_\alpha^g}}{\sqrt{\pi_i^g}}\, u_{i\alpha}^g \qquad i, j, \alpha = 1, \ldots, n\,.$$

$$\textit{Also,} \qquad \Delta_g[Z] = \sum_{i=1}^n \pi_i B_{ii}^g = \sum_{i=1}^n K_{ii}^g = \sum_{\alpha=1}^n \lambda_\alpha^g = \sum_{\alpha=1}^{n-1} \lambda_\alpha^g\,.$$

b) *Replacing the distribution π_i^g by the overall weights $\pi_i^0 := \pi_i$ and repeating the above construction yields the usual weighted MDS, with spectral decomposition $K^0 = U^0 \Lambda^0 (U^0)'$ providing coordinates $x_{i\alpha}^0 = \sqrt{\lambda_\alpha^0}\, u_{i\alpha}^0 / \sqrt{\pi_i}$ and eigenvalues λ_α^0 such that*

$$\Delta[Z] = \sum_i \pi_i B_{ii}^0 = \sum_i K_{ii}^0 = \sum_\alpha \lambda_\alpha^0 \geq \sum_{\alpha, g} \rho_g\, \lambda_\alpha^g = \Delta_W\,. \qquad \square$$