Separation and convexity properties of hierarchical and non hierarchical clustering

Patrice Bertrand[1]

[1]CEREMADE, Université Paris-Dauphine, Paris, France

Joint work with

Jean Diatta [2]

[2] LIM, Université de La Réunion, Saint-Denis, France

Plan

Plan

Plan

Plan

- *Multi-level clustering structures*
  - *Hierarchies*
    Johnson (1967), Benzécri (1973)
  - *Weak Hierarchies*

    Bandelt & Dress (1989, 1994), Diatta & Fichet (1994, 1998), Bertrand & Janowitz (2002)
  - *Pyramids (or pseudo-hierarchies)*

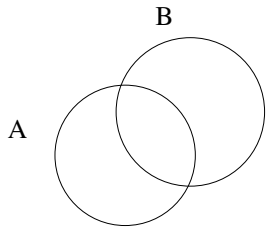
    Diday (1984, 1986), Fichet (1984, 1986)
  - *Paired hierarchies*

    Bertrand (2002, 2008), Bertrand & Brucker (2007)

Definitions

A pair $\{A, B\} \subseteq E$ (ground set) is said to be

- *hierarchical*: $A \cap B \in \{A, B, \emptyset\}$
  If $\{A, B\}$ is not hierarchical, then A and B *cross each other*



We use the following terminology for $\mathcal{F} \subseteq 2^E$:

- *set-system*: $\{\emptyset\} \notin \mathcal{F}$ and $E \in \mathcal{F}$
- *total*: for all $x \in E$, $\{x\} \subseteq \mathcal{F}$
- *closed*: $\mathcal{F}$ is closed under non empty intersections:

$$\forall \mathcal{G} \subseteq \mathcal{F}, \quad \bigcap \mathcal{G} \in \mathcal{F} \cup \{\emptyset\}$$

- *(strongly) hierarchical*: each pair $\{X, Y\} \subseteq \mathcal{F}$ is hierarchical

Weak hierarchies

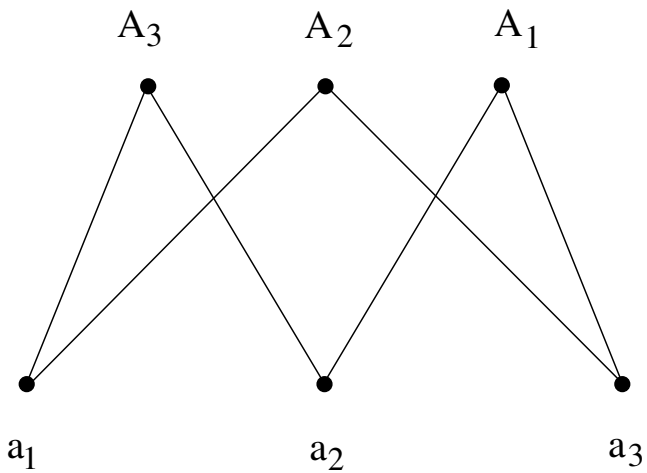
A collection $\mathcal{F} \subseteq 2^E$ is said to be *weakly hierarchical* if

$$\forall X, Y, Z \in \mathcal{F}, \quad X \cap Y \cap Z \in \{X \cap Y, Y \cap Z, X \cap Z\}$$

nsc

There are no $A_1, A_2, A_3 \in \mathcal{F}$ and $a_1, a_2, a_3 \in E$ s.t. $a_i \in A_j \iff i \neq j$

Forbidden configuration:

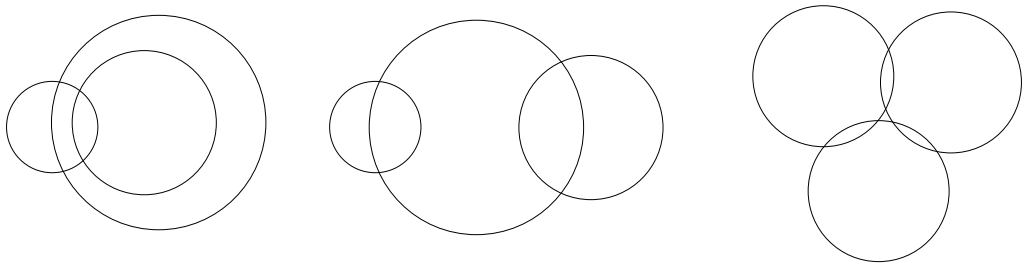


Paired hierarchies

A collection $\mathcal{F} \subseteq 2^E$ is called *paired hierarchical* if each $\mathcal{F}$-member crosses at most one $\mathcal{F}$-member

nsc

- $\forall X, Y, Z \in \mathcal{F}$, at least 2 of $\{X, Y\}, \{Y, Z\}, \{X, Z\}$ are hierarchical
- "X crosses Y" defines an equivalence relation whose class sizes are at most 2
- Forbidden configurations:

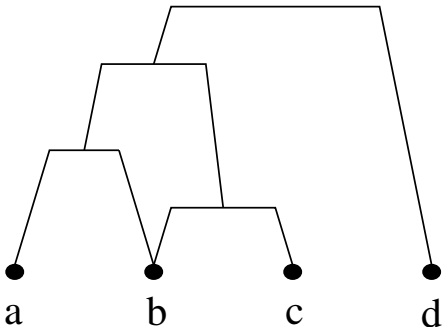


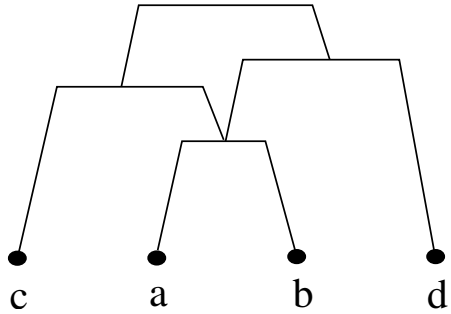
(a) (b) (c)

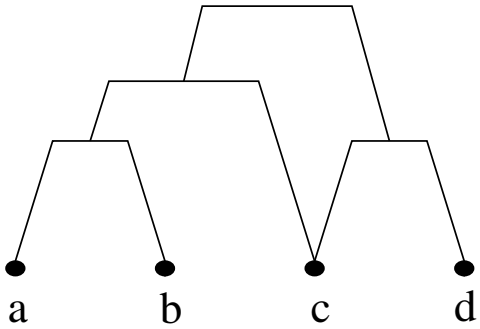
The term paired-hierarchy is used since $\{\bigcup \mathcal{G} \ : \ \mathcal{G} \text{ is a class}\}$ is a hierarchy

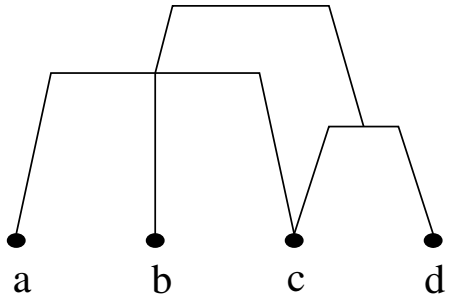
Examples and counter-examples

Paired-hierarchies

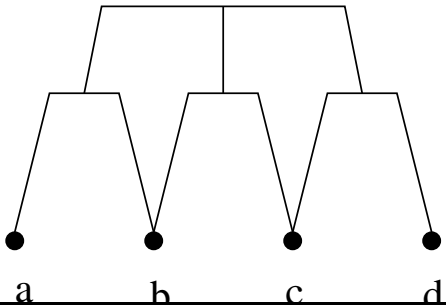


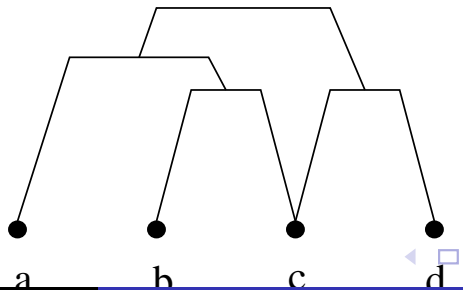






Weak-hierarchies





Correspondences between dissimilarities and multi-level clustering structures

d (dissimilarity on E) $\longleftrightarrow$ $(\mathcal{F}, f)$ ($\mathcal{F} \subseteq 2^E$ and $f : \mathcal{F} \mapsto \mathbb{R}^+$ being increasing)

$\phi : (\mathcal{F}, f) \mapsto \phi(\mathcal{F}, f)$ with $\phi(\mathcal{F}, f)(x, y) = \min\{f(A) \ : \ a, b \in A, A \in \mathcal{F}\}$

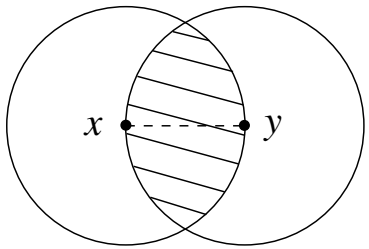
Conversely, each dissimilarity d is associated with:

- $\mathcal{D}^d(x, y)$: closed ball of center $x \in E$ and radius $r = d(x, y)$

$$\mathcal{D}^d(x, y) = \{z \in E : d(z, x) \leq d(x, y)\}$$

- $\mathcal{B}^d(x, y)$: 2-ball generated by $x, y \in E$, in the sense of d

$$\begin{aligned} \mathcal{B}^d(x, y) &= \mathcal{D}^d(x, y) \cap \mathcal{D}^d(y, x) \\ &= \{ z \in E : \max\{d(z, x), d(z, y)\} \leq d(x, y)\} \end{aligned}$$



Separation relation

A *ternary relation* designates any subset of E^3

A *(ternary) separation relation* is a ternary relation of the form:

- Given $\mathcal{F} \subseteq 2^E$, the ternary separation relation $s(\mathcal{F})$ is defined by $(x, y, z) \in s(\mathcal{F})$ if it exists a $\mathcal{F}$-member which contains x and y but not z.

  In what follows, we will write simply $xyz \in s(\mathcal{F})$ *in place of* $(x, y, z) \in s(\mathcal{F})$

Convexity

Abstract convexity (van de Vel, cf. early 1950s).

- A collection $\mathscr{C} \subseteq 2^E$ is called a *convexity* on E if $\emptyset, E \in \mathscr{C}$ and $\mathscr{C}$ is closed both under intersections and nested unions.

  $(E, \mathscr{C})$ is called a *convex structure* or a *convexity space*.

  Convex set: any member of $\mathscr{C}$
- $\forall A \subseteq E, \quad conv_{\mathscr{C}}(A) = \langle A \rangle_{\mathscr{C}} = \bigcap \{C : A \subseteq C \in \mathscr{C}\}$, is called the *(convex) hull* of A.
- Notations: $\langle a, b \rangle := \langle a, b \rangle_{\mathscr{C}}$
- *Segment joining a and b*: the 2-polytope $conv(\{a, b\})$
- $\langle , \rangle_{\mathscr{C}} : (a, b) \in E^2 \mapsto \langle a, b \rangle_{\mathscr{C}} \in 2^E$ is called the *segment operator* of the convexity $\mathscr{C}$.

$\mathcal{A}$rity

The arity of $\mathscr{C}$ is $\leq$ n if for all $C \in \mathscr{C}$ and $F \subseteq C$ with $\#F \leq n$, we have: $\langle F \rangle = conv(F) \subseteq C$

$\mathcal{R}$ank

$A \subseteq E$ is called *convexly independent* if $a \notin \langle A \setminus \{a\} \rangle$ for all $a \in A$
The *rank* of a convex structure $(E, \mathscr{C})$ is defined as the maximum size of a convexly independent set.

$\mathcal{I}$nterval operator

$I : E \times E \mapsto 2^E$ is called an *interval operator on E* if

$$\forall a, b \in E, \quad a, b \in I(a, b) = I(b, a).$$

$I(a, b)$: *interval between a and b*; (E, I): *interval space*.

Example: $\langle , \rangle_{\mathscr{C}} : (a, b) \in E^2 \mapsto \langle a, b \rangle_{\mathscr{C}}$ of any convexity $\mathscr{C}$

$\mathcal{N}$otations

$\mathcal{G}_I := \{C \subseteq E \mid \forall x, y \in C,\ I(x,y) \subseteq C\}$ is the convexity *induced* by I

$\mathcal{G}_{\langle\,,\,\rangle_{\mathscr{C}}}$:= interval convexity induced by the segment operator $\langle\,,\,\rangle_{\mathscr{C}}$

$\langle a, b\rangle_I$ segment between a and b in the sense of the convexity $\mathcal{G}_I$.

$\mathcal{P}$roperties (Calder (1971))

- $\forall a, b \in E, \quad I(a,b) \subseteq \langle a, b\rangle_I$
- A convexity is induced by an interval operator iff its arity is ≤ 2.
- The hull of a set A in an interval space is given by

$$\langle A \rangle = \bigcup_{k=0}^{\infty} A_k,$$

where $A_0 = A$ and for all $k \in \mathbb{N},\ A_{k+1} = \bigcup\{I(a,a') \mid a, a' \in A_k\}$.

Convexity induced by $\langle\ ,\ \rangle_{\mathscr{C}}$

Lemma 1

Let $\mathscr{C}$ be a convexity on E.

(i) $\langle\ ,\ \rangle_{\mathscr{C}}$ and $\langle\ ,\ \rangle_{\langle\ ,\ \rangle_{\mathscr{C}}}$ coincide.

(ii) We have:

$$\{\langle a,b\rangle_{\mathscr{C}} \mid a,b \in E\} \subseteq \mathscr{C} \subseteq \mathcal{G}_{\langle\ ,\ \rangle_{\mathscr{C}}},$$

where the two inclusions may be strict.

Remark

It is easily checked that: $xyz \in s(\mathscr{C}) \iff z \notin \langle x,y\rangle_{\mathscr{C}}$.

Interval operators and Cluster Analysis

- $\mathcal{B}^d$ and $\mathcal{D}^d$ are two interval operators defined on E

Lemma 2

For all dissimilarity d on E and all $x, y \in E$, there exist $u, v \in E$ such that:

$$\langle x\,,\, y\rangle_{\mathcal{B}^d} = \langle u\,,\, v\rangle_{\mathcal{B}^d} = \mathcal{B}^d(u, v).$$

Separation, Interval operators and Weak Hierarchies

- Bandelt and Dress (1994): A set-system $\mathscr{C}$ is weakly hierarchical iff for all x_1, x_2, x_3 distinct in E, $s(\mathscr{C})$ does not contains both

$$x_1x_2x_3,\ x_2x_3x_1 \text{ and } x_3x_1x_2$$

- Let I be an interval operator on E, and let

(w) No $x, y, z \in E$ exist s.t. $x \notin I(y, z)$, $y \notin I(x, z)$ and $z \notin I(x, y)$.

Proposition 3

Let I be an interval operator and let

(i) I satisfies *(W)*

(ii) $\langle , \rangle_I$ satisfies *(W)*

(iii) $\mathcal{G}_I$ is weakly hierarchical

(iv) $\mathcal{G}_I$ is of rank at most 2, i.e.
if $\emptyset \subsetneq A \subseteq E$, then $\langle A \rangle_I$ is of the form $\langle a, b \rangle_I$ for some $a, b \in A$.

Then

$$(\mathrm{i}) \Rightarrow (\mathrm{ii}) \Leftrightarrow (\mathrm{iii}) \Leftrightarrow (\mathrm{iv})$$

Corollary 4

If the interval operator I satisfies (w), then $\mathcal{G}_I = \{\langle a, b\rangle_I \mid a, b \in E\}$

Definition 5 (k-ball)

Let $A \subseteq E$ with $\#A = k > 2$, and denote

$$\mathcal{B}_A^d = \{x \in E \mid \forall a \in A,\ d(a, x) \leq \mathrm{diam}_d\, A\}.$$

Proposition 6

If $(\mathscr{C}, f)$ is an indexed closed weak-hierarchical set system s.t. $f^{-1}(0) = \{X \in \mathscr{C} \mid f(X) = 0\}$ is a partition of E, then

$$\mathcal{B}_A^{\phi((\mathscr{C},f))} = \langle A\rangle_{\mathscr{C}\cup\{\emptyset\}},$$

for all nonempty subset A of E.

Notation 7

$\mathcal{B}^{(\mathscr{C},f)} := \mathcal{B}^{\phi((\mathscr{C},f))}$

Corollary 8

Let $\mathscr{C}$ be a set-system on E. The following are equivalent:

(i) $\mathscr{C}$ is closed and weakly hierarchical

(ii) $\mathscr{C} \cup \{\emptyset\} = \mathcal{G}_I$ for some interval operator I satisfying (w)

(iii) $\mathscr{C} \cup \{\emptyset\} = \mathcal{G}_{\mathcal{B}^{(\mathscr{C},f)}}$ for some index f on $\mathscr{C}$ satisfying $\bigcup f^{-1}(0) = E$

(iv) $\mathscr{C} \cup \{\emptyset\} = \mathcal{G}_{\mathcal{B}^{(\mathscr{C},f)}}$ for all index f on $\mathscr{C}$ satisfying $\bigcup f^{-1}(0) = E$

Criterion to recognize whether a set system is weakly hierarchical: define f by $f(A) := \mid A \mid \ -1$

Separation relation, Interval operators, Hierarchies

Proposition 9

For all collection $\mathscr{C}$ of subsets of E, the following are equivalent:

(i) $\mathscr{C}$ is hierarchical;

(ii) For all $x, y, z \in E$, $xyz \in s(\mathscr{C}) \Rightarrow \big(yzx \notin s(\mathscr{C})$ and $zxy \notin s(\mathscr{C})\big)$.

Definition 10

Let I be an interval operator on E, we denote:

(H) For all $x, y, z \in E$, either $I(x, y) \subseteq I(x, z)$ or $I(x, z) \subseteq I(x, y)$.

Remark 11

Clearly, (H) $\Rightarrow$ (W) and $\mathcal{D}^d$ satisfies (H)

Proposition 12

Let I be any interval operator on E, and denote:

(i) I satisfies the condition (H)

(ii) The segment operator $\langle\, , \,\rangle_I$ satisfies the condition (H)

(iii) $\mathcal{G}_I$ is a hierarchical set system

Then

$$(\mathrm{i}) \Rightarrow (\mathrm{ii}) \Leftrightarrow (\mathrm{iii})$$

Notation 13

$$\begin{aligned} \mathcal{D}_A^d &:= \{x \in E \mid \exists\, a \in A \text{ s.t. } d(a,x) \leq \operatorname{diam}_d A\} \\ \mathcal{D}^{(\mathscr{C},f)}(x,y) &:= \mathcal{D}^{\phi((\mathscr{C},f))}(x,y) \end{aligned}$$

for any $\emptyset \subsetneq A \subseteq E$ and any $x, y \in E$.

Proposition 14

If d is an ultrametric, then $\boxed{\mathcal{D}_A^d = \mathcal{B}_A^d}$

Corollary 15

If $(\mathscr{C}, f)$ is an indexed hierarchy on E such that $f^{-1}(0) = \{X \in \mathscr{C} \mid f(X) = 0\}$ is a partition of E, then

$$\forall A \in 2^E \setminus \{\emptyset\}, \quad \mathcal{D}_A^{(\mathscr{C},f)} = \langle A \rangle_{\mathscr{C} \cup \{\emptyset\}}.$$

Corollary 16

Let $\mathscr{C}$ be a set-system on E. The following are equivalent:

(i) $\mathscr{C}$ is hierarchical

(ii) $\mathscr{C} \cup \{\emptyset\} = \mathcal{G}_I$ for some interval operator I satisfying (H)

(iii) $\mathscr{C} \cup \{\emptyset\} = \mathcal{G}_{\mathcal{D}(\mathscr{C},f)}$ for some index f on $\mathscr{C}$ s. t. $f^{-1}(0) = \bigcup E$

(iv) $\mathscr{C} \cup \{\emptyset\} = \mathcal{G}_{\mathcal{D}(\mathscr{C},f)}$ for all index f on $\mathscr{C}$ s.t. $f^{-1}(0) = \bigcup E$

Criterion to recognize whether a set system is hierarchical: define f by $f(A) := |A| - 1$

Interval operators, separation relation and paired hierarchies

Definition 17 (property $\mathrm{P}[\mathscr{C}]$)

For all $A \in \mathcal{F}$, all distinct elements x, y of A and all $u, v \notin A$,

$$\left(uyx \in s(\mathscr{C}) \Rightarrow vxy \notin s(\mathscr{C})\right) \text{ and } \left(uyv \in s(\mathscr{C}) \Rightarrow vyx \notin s(\mathscr{C})\right)$$

Proposition 18

A set-system $\mathscr{C}$ is hierarchical iff $\mathrm{P}[\mathscr{C}]$ is true

Definition 19

Given an interval operator I on E, let (P) be defined as:

(P) No $x, y, u, v \in E$, with $x \neq y$, satisfy both $u, v \notin I(x, y)$ and either

$$\big(x \notin I(y, u) \text{ and } y \notin I(x, v)\big) \text{ or } \big(v \notin I(y, u) \text{ and } x \notin I(y, v)\big).$$

Proposition 20

Let $\mathscr{C}$ be a closed set-system on E. The following assertions are equivalent:

(i) $\mathscr{C}$ is hierarchical

(ii) $\mathscr{C} \cup \{\emptyset\} = \mathcal{G}_I$ for some interval operator I satisfying (P)

Application to Cluster Analysis

Algorithm of construction of the interval convexity $\mathcal{G}_I$

Denote I an interval operator on E and $E := \{e_1, \ldots, e_n\}$

Denote $\mathcal{G}_I^k$ the convexity induced on $E_k := \{e_1, \ldots, e_k\}$ $(k \leq n)$ by the restriction of I to E_k

1. Put $\mathcal{G}_I^0 = \{\varnothing\}$
2. Assume $\mathcal{G}_I^k$ is known for $k < n$. For each $C \in \mathcal{G}_I^k$,
   - **2a** mark C iff $e_{k+1} \notin I(e, e')$ for all $e, e' \in C$
   - **2b** mark $C \cup \{e_{k+1}\}$ iff $I(e, e_{k+1}) \subseteq C \cup \{e_{k+1}\}$ for all $e \in C \cup \{e_{k+1}\}$
   - **2c** add to $\mathcal{G}_I^{k+1}$ any subset that was marked either by **2a** or by **2b**
3. If $k < n$, increment the value of k and repeat step 2

Some Discussion

1. Characterizations of clustering structures in terms of ternary separation relations and (abstract) convexity
2. $\mathcal{D}^d$ satisfies (H) and $\mathcal{B}^d$ satisfies (W): is there some similar map φ_d^p, that can be derived from any dissimilarity d, satisfying (P)?

   Rmk: $\mathcal{B}^d(a,b) \subseteq \varphi_d^p(a,b) \subseteq \mathcal{D}^d(a,b)$ *must hold, for all* $a, b \in E$
3. Separation properties $\overset{?}{\longleftrightarrow}$ convexity properties