Clustering of relational data

Thierry Denœux

tdenoeux@utc.fr

University of Compiègne, France

Introduction

- Clustering methods
  - Finding groups in data
  - Generated structure: Hierarchy; hard, fuzzy, possibilistic partition
- Data type
  - Attribute data: objects described by attributes (features)
  - Proximity (Relational) data: pairwise dissimilarities between objects

Proximity Data

Let $\mathcal{P}$ be a collection of n objects $\{o_i\}_{i=1}^n$. The observations consist in pairwise dissimilarities between objects:

$$
\begin{array}{c|ccccc}
 & o_1 & \ldots & o_j & \ldots & o_n \\
\hline
o_1 & & & \vdots & & \\
\vdots & & & \vdots & & \\
o_i & & \ldots & d_{ij} & \ldots & \\
\vdots & & & \vdots & & \\
o_n & & & & &
\end{array}
$$

Origin of Proximity Data

- Distances computed from attribute data: allow to
  - handle heterogeneous data: quantitative, qualitative, structured, symbolic, etc.
  - incorporate prior knowledge in the distance function
- Intrinsically present in many domains: psychology, economics, biochemistry (structural comparison between protein sequences), web mining (clustering of web sites, etc.), etc.

Problem statement

- n objects described by dissimilarity matrix $D = (d_{ij})$.
- Assumption: each object belongs to one of c classes in $\Omega = \{\omega_1, \cdots, \omega_c\}$,
- Goal: express our beliefs regarding the class-membership of objects, in the form of belief functions on Ω.
- Resulting structure = Credal partition, generalizes hard, fuzzy and possibilistic partitions

Credal Partition

- Partial knowledge concerning class membership of o_i represented by a bba $m_i^\Omega : 2^\Omega \to [0,1]$.
- Credal partition: $M^\Omega \triangleq (m_1^\Omega, \ldots, m_n^\Omega)$
- Credal c-partition: each class is plausible for at least one object

$$\forall \omega \in \Omega, \exists i \in \{1, \ldots, n\}, pl_i(\{\omega\}) > 0$$

Credal Partition: example

A	$m_1(A)$	$m_2(A)$	$m_3(A)$	$m_4(A)$	$m_5(A)$
$\emptyset$	0	0	0	0	0
$\{\omega_1\}$	0	0	0	0.2	0
$\{\omega_2\}$	0	1	0	0.4	0
$\{\omega_1, \omega_2\}$	0.7	0	0	0	0
$\{\omega_3\}$	0	0	0.2	0.4	0
$\{\omega_1, \omega_3\}$	0	0	0.5	0	0
$\{\omega_2, \omega_3\}$	0	0	0	0	0
Ω	0.3	0	0.3	0	1

Special cases

- Each m_i^Ω is a *certain bba* $\rightarrow$ crisp partition of Ω.
- Each m_i^Ω is a *Bayesian bba* $\rightarrow$ fuzzy partition of Ω

$$u_{ik} = m_i^\Omega(\{\omega_k\}), \quad \forall i, k$$

- Each m_i^Ω is a *consonant bba* $\rightarrow$ possibilistic partition of Ω

$$u_{ik} = pl_i^\Omega(\{\omega_k\})$$

Learning a Credal Partition from data

- Problem: given a dissimilarity matrix $D = (d_{ij})$, how to build a "reasonable" credal partition ?
- Notion of cluster: objects within a cluster are assumed to be more similar among themselves than with objects from other clusters.
- Compatibility Principle: "The *more similar* two objects, the *more plausible* it is that they belong to the same class"

Formalization (1)

- Let S_{ij} be the event "objects i and j belong to the same class".

$$S_{ij} \triangleq \{(\omega_1, \omega_1), (\omega_2, \omega_2), \ldots, (\omega_c, \omega_c)\} \subset \Omega^2$$

- Computation of $pl^{\Omega^2}_{i \times j}(S_{ij})$ in the TBM: vacuously extend m_i and m_j to Ω^2, and combine using Dempster's rule:

$$\boxed{m^{\Omega^2}_{i \times j} = m_i^{\Omega \uparrow \Omega^2} \bigcirc\!\!\!\!\cap\, m_j^{\Omega \uparrow \Omega^2}}$$

Formalization(2)

$$
\begin{aligned}
\mathrm{pl}^{\Omega^2}_{i\times j}(S) &= \sum_{\{A\times B\subseteq\Omega^2 \mid (A\times B)\cap S\neq\emptyset\}} m_{i\times j}(A\times B)\\
&= \sum_{A\cap B\neq\emptyset} m_i(A)\cdot m_j(B)\\
&= 1-\sum_{A\cap B=\emptyset} m_i(A)\cdot m_j(B)\\
&= 1-K_{ij}
\end{aligned}
$$

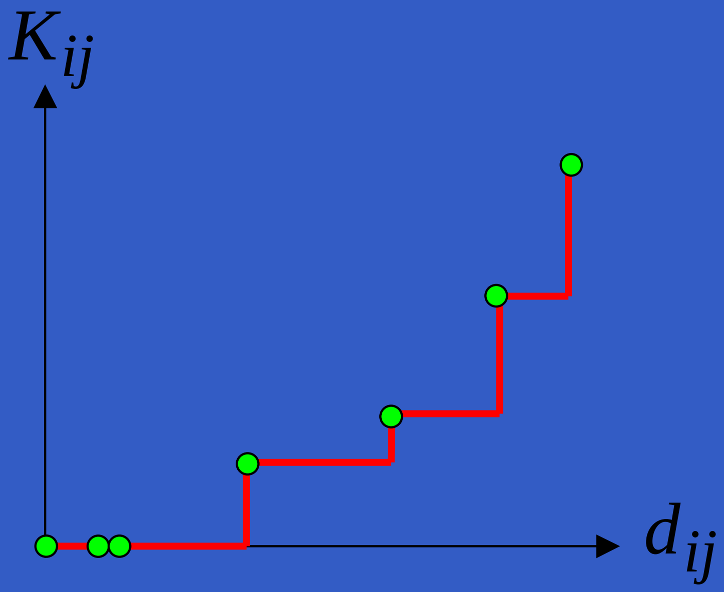
where K_{ij} = degree of conflict between m_i and m_j.

Compatibility criterion

Let $M^\Omega = (m_1^\Omega, \ldots, m_n^\Omega)$ be a credal c-partition of Ω. M is compatible with dissimilarity matrix $D = (d_{ij})$ iff:

For any 2 pairs of objects (o_i, o_j) and $(o_{i'}, o_{j'})$

$$d_{ij} > d_{i'j'} \Rightarrow K_{ij} \geq K_{i'j'}$$



The EVCLUS method

- Approach: minimize the discrepancy between the dissimilarities d_{ij} and the degrees of conflict K_{ij}, up to a monotonic transformation (similar to Muldimensional Scaling).
- Example of stress function (Sammon):

$$I(M,a,b) \triangleq \sum_{i<j} \frac{(aK_{ij}+b-d_{ij})^2}{d_{ij}}$$

- Minimization of I with respect to M and a, b by gradient descent.

Reducing the complexity

- Problem: large number of parameters ($n(2^c - 1)$ parameters for $n(n-1)/2$ dissimilarities).
- Solutions:
  - Reduce the focal elements to $\{\omega_i\}_{i=1}^c$, $\emptyset$, and Ω.
  - Add constraints to the problem: penalize "uninformative", "complex" credal partitions

$$I' = I + \lambda \sum_{i=1}^{n} H(m_i)$$

where H=generalized entropy measure.

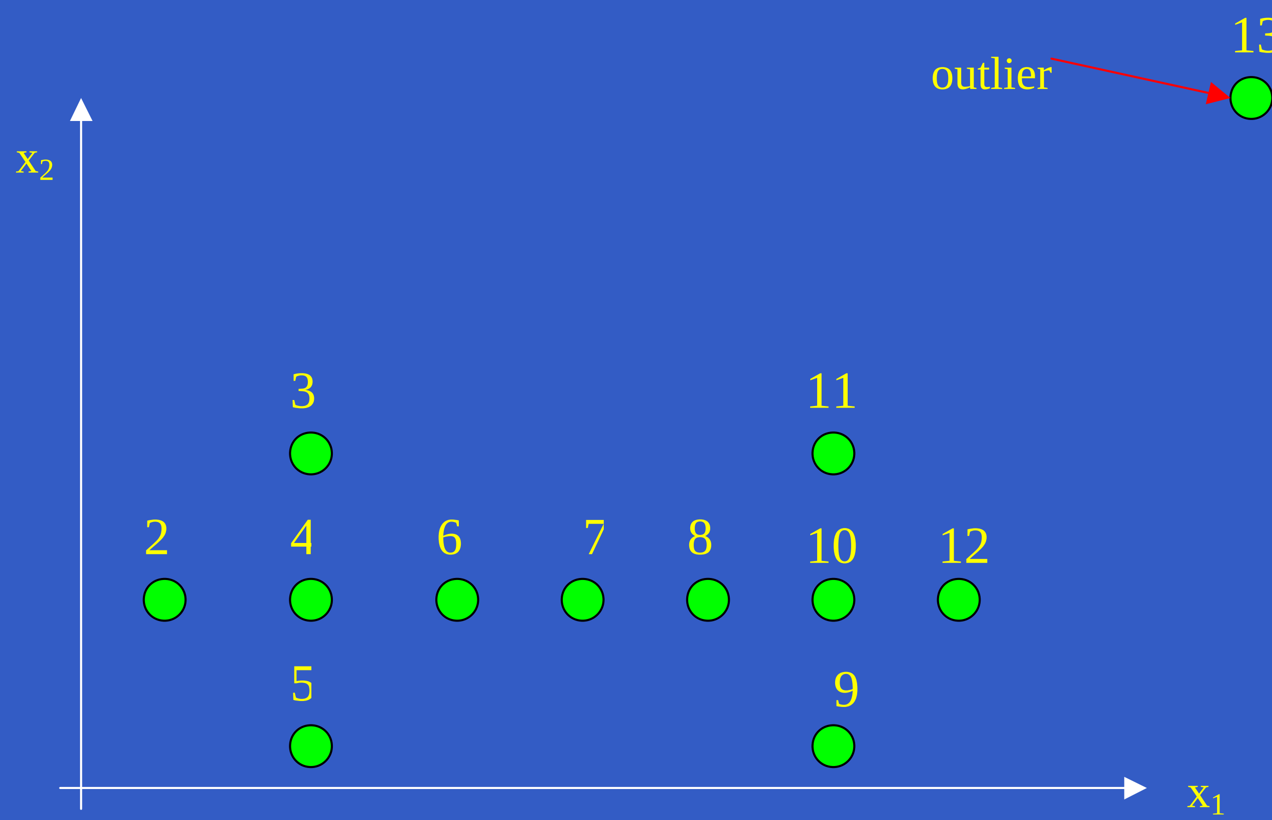
Entropy measure

Possible choice for the entropy function (Pal and Bezdek):

$$H(m_i) = \sum_{A \in \mathcal{F}(m_i) \setminus \{\emptyset\}} m_i(A) \log_2 \left(\frac{|A|}{m_i(A)} \right) + m_i(\emptyset) \log_2 \left(\frac{|\Omega|}{m_i(\emptyset)} \right)$$

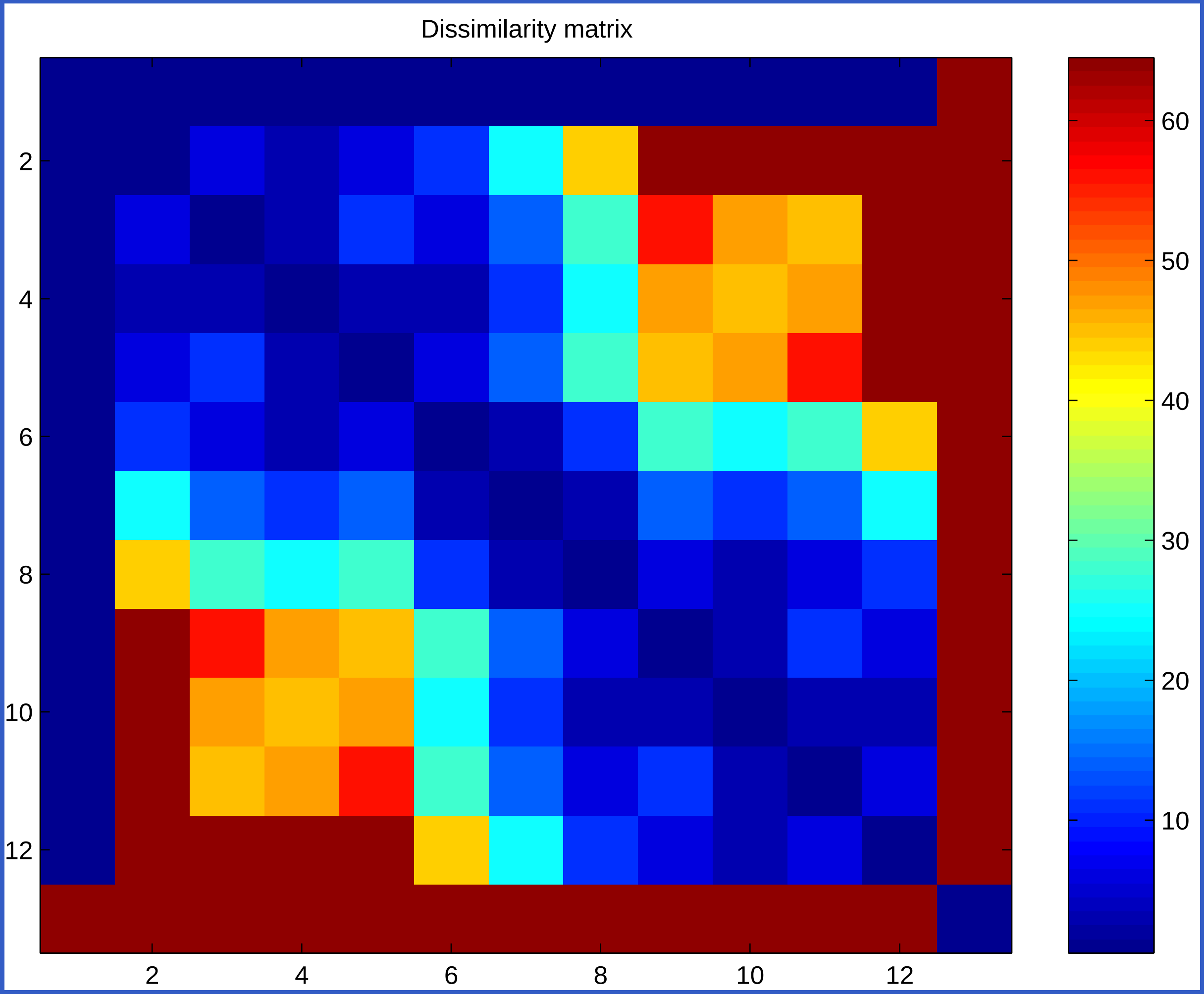
H favors the allocation of the mass to a small number of focal elements with low cardinality.

Butterfly example

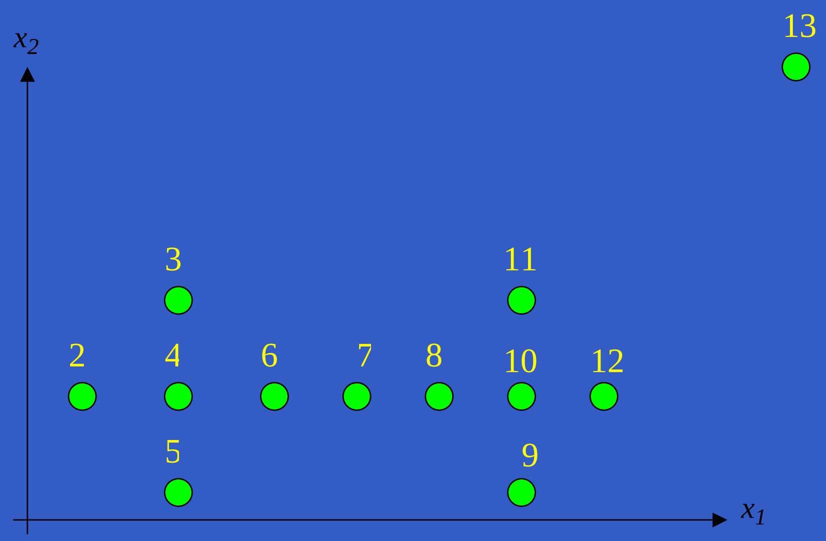


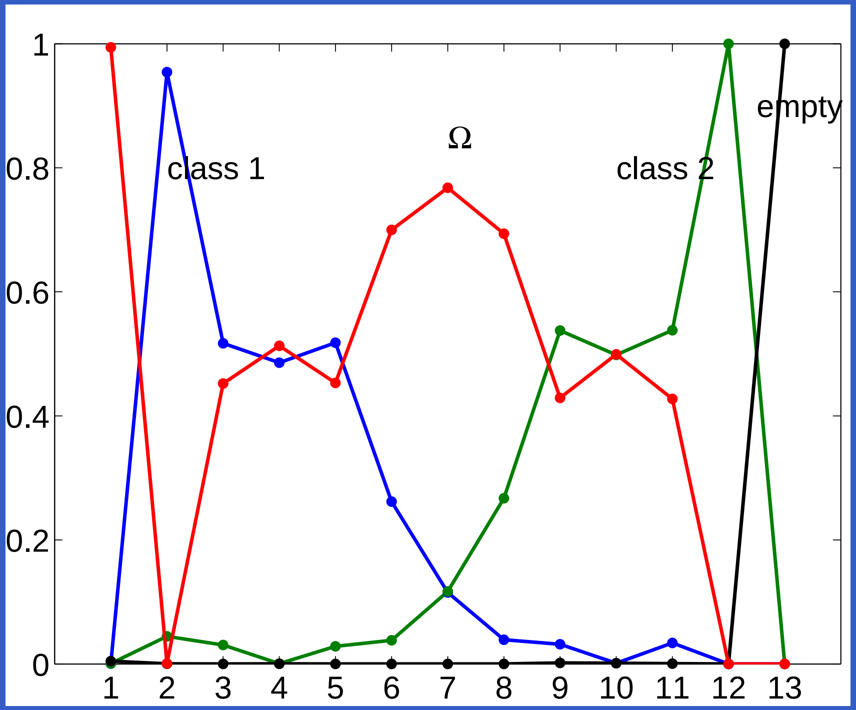
+ object #1 similar to all other objects ("inlier")

Butterfly example: dissimilarity matrix



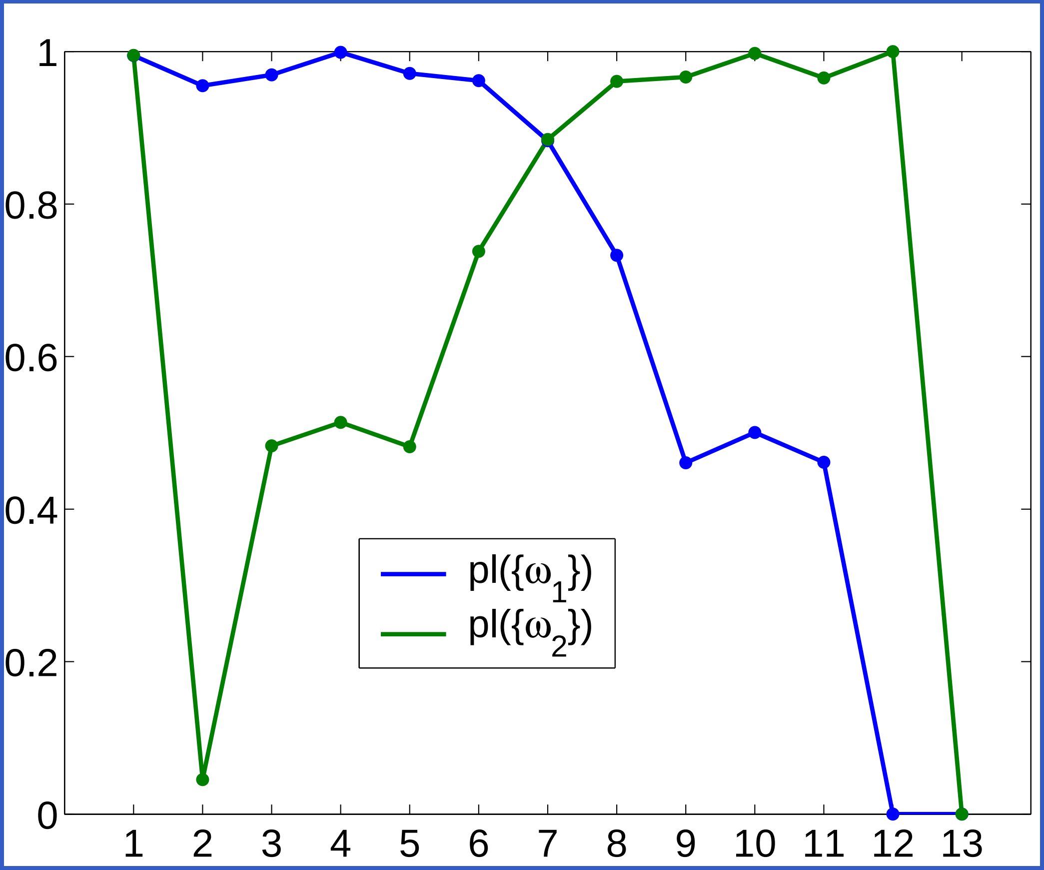
Butterfly example: Credal partition



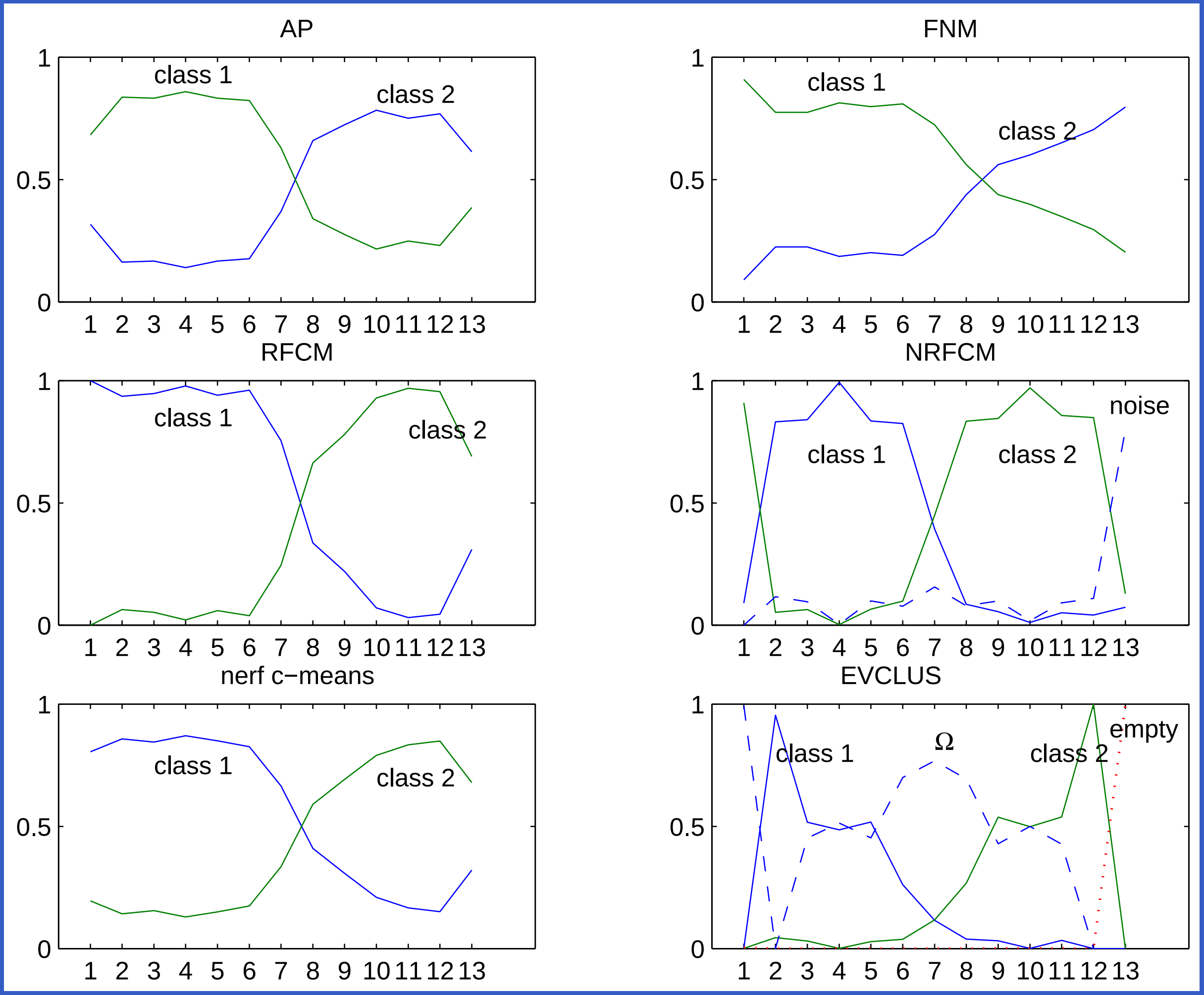


Butterfly example: Plausibilities

Plausibilities of ω_1 and ω_2



Butterfly example: comparison



Experiments with real data

- Cat cortex data: 65 objects (cortical areas), ordinal dissimilarities (connection strenghs expressed on an ordinal scale), “true” partition in 4 clusters (4 functional regions of the cortex)
- Protein data set: 213 proteins, dissimilarities derived from structural comparison, ‘true” partition in 4 clusters (4 classes of globins)
- sensory data: 13 objects, subjective assessments of dissimilarity by several experts, fusion of credal partitions.

Conclusion

- EVCLUS: a new clustering method for relational data, based on belief functions.
- The concept of credal partition extends those of hard or fuzzy partitions, greater flexibility
- Advantages of the method
  - Detection and representation of atypical observations (in/out-liers),
  - Robustness to non metric data
  - Fusion of credal partitions, and combination with prior knowledge.