Methods for building belief functions

Thierry Denœux

Université de technologie de Compiègne, France
Institut Universitaire de France
https://www.hds.utc.fr/~tdenoeux

Fifth School on Belief Functions and their Applications
Sienna, Italy, October 28, 2019

Building belief functions

- The basic theory tells us how to reason and compute with belief functions, but it does not tell us where belief functions come from.
- To use DS theory in real applications, we need methods for modeling evidence from
  - Expert opinions or
  - Statistical information
- Two main strategies, often combined in applications:
  1. Decomposition: Start with elementary (often, simple) mass functions and transform/combine them using extension, marginalization and Dempster's rule (original DS approach).
  2. Global approach: Find the least (or the most) committed belief function compatible with given constraints.
- In this lecture, we will see several applications of these strategies.

Outline

Outline

Least Commitment Principle

Definition (Least Commitment Principle (LCP))

When several belief functions are compatible with a set of constraints, the least informative according to some informational ordering (if it exists) should be selected.

- General approach
  1. Express partial information (provided, e.g., by an expert or statistical data) as a set of constraints on an unknown mass function
  2. Find the least-committed mass function (according to some informational ordering), compatible with the constraints
- Examples of partial information
  1. Contour function
  2. Conditional mass function

Outline

Problem statement

- Assume an expert gives us the plausibility $\pi(\omega)$ of each $\omega \in \Omega$.
- We get a function $\pi : \Omega \to [0, 1]$. We assume that

$$\max_{\omega \in \Omega} \pi(\omega) = 1.$$

- Let $\mathcal{M}(\pi)$ be the set of mass functions m such that $pl = \pi$.
- What is the least committed mass function in $\mathcal{M}(\pi)$?
- A solution exists according to the q-ordering.

Solution

- Let $m \in \mathcal{M}(\pi)$ and Q its commonality function. We have

$$Q(\{\omega\}) = pl(\omega) = \pi(\omega), \quad \forall \omega \in \Omega$$

and

$$Q(A) \leq \min_{\omega \in A} Q(\{\omega\}) = \min_{\omega \in A} \pi(\omega), \quad \forall A \subseteq \Omega, A \neq \emptyset,$$

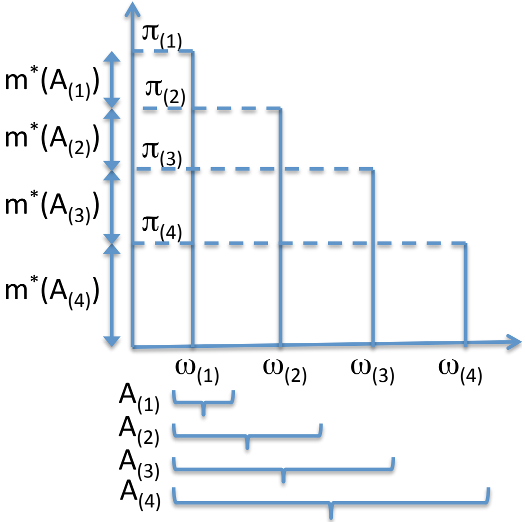
- Let Q^* be defined as $Q^*(\emptyset) = 1$ and

$$Q^*(A) = \min_{\omega \in A} \pi(\omega), \quad \forall A \subseteq \Omega, A \neq \emptyset.$$

Proposition

Q^* is the commonality function of a consonant mass function m^*, which is the q-least committed element in $\mathcal{M}(\pi)$.

Calculation of the mass function



Example

- Consider, for instance, the following contour distribution defined on the frame $\Omega = \{a, b, c, d\}$:

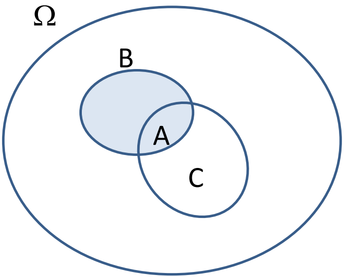
ω	a	b	c	d
$pl(\omega)$	0.3	0.5	1	0.7

- The corresponding mass function is

$$\begin{aligned} m(\{c\}) &= 1 - 0.7 = 0.3 \\ m(\{c, d\}) &= 0.7 - 0.5 = 0.2 \\ m(\{c, d, b\}) &= 0.5 - 0.3 = 0.2 \\ m(\{c, d, b, a\}) &= 0.3. \end{aligned}$$

Outline

Deconditioning



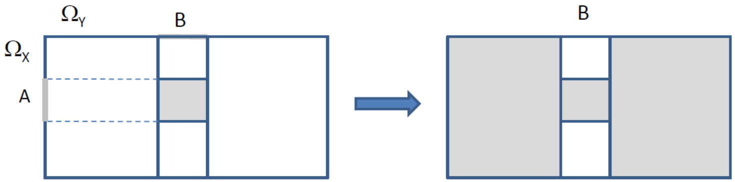
- Let m_0 be a mass function on Ω expressing our beliefs about X in a context where we know that $X \in B$.
- We want to build a mass function m verifying the constraint $m(\cdot \mid B) = m_0$.
- Any m built from m_0 by transferring each mass $m_0(A)$ to $A \cup C$ for some $C \subseteq \overline{B}$ satisfies the constraint.

Proposition

The s-least committed solution is obtained by transferring each mass $m_0(A)$ to the largest such set, which is $A \cup \overline{B}$:

$$m(D) = \begin{cases} m_0(A) & \text{if } D = A \cup \overline{B} \text{ for some } A \subseteq B \\ 0 & \text{otherwise.} \end{cases}$$

Conditional embedding



- More complex situation: two frames Ω_X and Ω_Y.
- Let m_0^X be a mass function on Ω_X expressing our beliefs about X in a context where we know that $Y \in B$ for some $B \subseteq \Omega_Y$.
- We want to find m^{XY} such that $\left(m^{XY} \oplus m_{[B]}^Y\right)^{\downarrow X} = m_0^X$.
- s-least committed solution: transfer $m_0^X(A)$ to $(A \times \Omega_Y) \cup (\Omega_X \times \overline{B})$.
- Notation $m^{XY} = (m_0^X)^{\Uparrow XY}$ (conditional embedding).

Discounting

Problem statement

- A source of information provides:
  - a value
  - a set of values
  - a probability distribution, etc.
- The information is:
  - not fully reliable or
  - not fully relevant.
- Examples:
  - Possibly faulty sensor
  - Measurement performed in unfavorable experimental conditions
  - Information is related to a situation or an object that only has some similarity with the situation or the object considered (case-based reasoning).

Discounting

Formalization

- A source $\mathcal{S}$ provides a mass function $m^{\Omega}_{\mathcal{S}}$.
- $\mathcal{S}$ may be reliable or not. Let $\mathcal{R} = \{R, NR\}$.
- Assumptions:
  - If $\mathcal{S}$ is reliable, we accept $m^{\Omega}_{\mathcal{S}}$ as a representation of our beliefs:

$$m^{\Omega}(\cdot \mid R) = m^{\Omega}_{\mathcal{S}}$$

  - If S is not reliable, we know nothing:

$$m^{\Omega}(\cdot \mid NR) = m^{\Omega}_{?}$$

  - The source has a probability α of not being reliable:

$$m^{\mathcal{R}}(\{NR\}) = \alpha, \quad m^{\mathcal{R}}(\{R\}) = 1 - \alpha$$

  (α is called the discount rate).

Discounting

Solution

- Solution:

$$ {}^{\alpha}m^{\Omega} = \left(m^{\mathcal{R}} \oplus m^{\Omega}(\cdot \mid R)^{\Uparrow\Omega\times\mathcal{R}}\right)^{\downarrow\Omega} = (1-\alpha)m^{\Omega}_{S} + \alpha m^{\Omega}_{\Omega}. $$

- ${}^{\alpha}m^{\Omega}$ can also be written as

$$ {}^{\alpha}m^{\Omega} = m^{\Omega}_{S} \bigcirc\!\!\!\!\cup\, m^{\Omega}_{0}, $$

with $m^{\Omega}_{0}(\Omega) = \alpha$ and $m^{\Omega}_{0}(\emptyset) = 1-\alpha$.

- Contour function:

$$ {}^{\alpha}pl(\omega) = (1-\alpha)pl(\omega) + \alpha, \quad \forall\omega \in \Omega. $$

- ${}^{\alpha}m^{\Omega}$ is a s-less committed than (a generalization of) m^{Ω}_{S}:

$$ {}^{\alpha}m^{\Omega} \sqsupseteq_{s} m^{\Omega}_{S}. $$

Generalization: Contextual Discounting

Formalization

- A more general model allowing us to take into account richer meta-information about the source.
- Let $\Theta = \{\theta_1, \ldots, \theta_L\}$ be a partition of Ω, representing different contexts.
- Let $m^{\mathcal{R}}(\cdot \mid \theta_k)$ denote the mass function on $\mathcal{R}$ quantifying our belief in the reliability of source $\mathcal{S}$, when we know that the actual value of X is in θ_k.
- We assume that:

$$m^{\mathcal{R}}(\{R\} \mid \theta_k) = 1 - \alpha_k, \quad m^{\mathcal{R}}(\{NR\} \mid \theta_k) = \alpha_k.$$

  for each $k \in \{1, \ldots, L\}$.
- Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_L)$.

Contextual Discounting

Example

- Let us consider a simplified aerial target recognition problem, in which we have three classes: airplane ($\omega_1 \equiv a$), helicopter ($\omega_2 \equiv h$) and rocket ($\omega_3 \equiv r$).
- Let $\Omega = \{a, h, r\}$.
- The sensor provides the following mass function: $m_S^\Omega(\{a\}) = 0.5$, $m_S^\Omega(\{r\}) = 0.5$.
- We assume that
  - The probability that the source is reliable when the target is an airplane is equal to $1 - \alpha_1 = 0.4$
  - The probability that the source is reliable when the target is either a helicopter, or a rocket is equal to $1 - \alpha_2 = 0.9$.
- We have $\Theta = \{\theta_1, \theta_2\}$, with $\theta_1 = \{a\}$, $\theta_2 = \{h, r\}$, and $\boldsymbol{\alpha} = (0.6, 0.1)$.

Contextual Discounting

Solution

- Solution:

$$
{}^{\alpha}m^{\Omega} = \left(\bigoplus_{k=1}^{L} m^{\mathcal{R}}(\cdot \mid \theta_k)^{\Uparrow \Omega \times \mathcal{R}} \oplus m^{\Omega}(\cdot \mid R)^{\Uparrow \Omega \times \mathcal{R}} \right)^{\downarrow \Omega}.
$$

- Result:

$$
{}^{\alpha}m^{\Omega} = m_S^{\Omega} \cup\!\!\!\!\bigcirc\, m_1^{\Omega} \cup\!\!\!\!\bigcirc \ldots \cup\!\!\!\!\bigcirc\, m_L^{\Omega}
$$

  with $m_k^{\Omega}(\theta_k) = \alpha_k$ and $m_k^{\Omega}(\emptyset) = 1 - \alpha_k$.

- Standard discounting is recovered as a special case when $\Theta = \{\Omega\}$.

Contextual Discounting

Example (continued)

- The discounted mass function can be obtained by combining disjunctively 3 mass functions:
  - $m_S^\Omega(\{a\}) = 0.5$, $m_S^\Omega(\{r\}) = 0.5$
  - $m_1^\Omega(\{a\}) = 0.6$, $m_1^\Omega(\emptyset) = 0.4$
  - $m_1^\Omega(\{h, r\}) = 0.1$, $m_1^\Omega(\emptyset) = 0.9$.
- Result:

A	$\{h\}$	$\{a\}$	$\{r\}$	$\{h, a\}$	$\{h, r\}$	$\{a, r\}$	Ω
$m_S^\Omega(A)$	0	0.5	0.5	0	0	0	0
${}^\alpha m^\Omega(A)$	0	0.45	0.18	0	0.02	0.27	0.08

Outline

Motivation

- In some cases, the least committed mass function compatible with some constraints does not exist, or cannot be found, for any informational ordering.
- An alternative approach is then to maximize a measure of uncertainty, i.e., find the most uncertain mass function satisfying some constraints.
- Many uncertainty measures have been proposed, some of which generalize the Shannon entropy. They can be classified in three categories:
  1. Measures of imprecision
  2. Measures of conflict
  3. Measures of total uncertainty

Measures of imprecision

- Idea: imprecision is higher when masses are assigned to larger focal sets:

$$\mathcal{I}(m) = \sum_{\emptyset \neq A \subseteq \Omega} m(A) f(|A|)$$

with $f = Id$ (expected cardinality), $f(x) = -1/x$ (opposite of Yager's specificity), $f = \log_2$.(nonspecificy)

- Nonspecificity $N(m)$ generalizes the Hartley function for set ($H(A) = \log_2(|A|)$) and was shown by Ramer (1987) to be the unique measure verifying some axiomatic requirements such as
  - Additivity for non-interactive mass functions: $N(m^{XY}) = N(m^X) + N(m^Y)$
  - Subadditivity for interactive mass functions: $N(m^{XY}) \leq N(m^X) + N(m^Y)$
  - ...
- Nonspecificity is equal to 0 for Bayesian mass function: we need to measure another dimension of uncertainty.

Measures of conflict

- Idea: should be higher when masses are assigned to disjoint (or non nested) focal sets.
- Example: dissonance (Yager, 1983) is defined as

$$E(m) = -\sum_{A\subseteq\Omega} m(A)\log_2 Pl(A) = -\sum_{A\subseteq\Omega} m(A)\log_2(1-K(A))$$

  where $K(A)=\sum_{B\cap A=\emptyset} m(B)$ can be interpreted as measuring the degree to which the evidence conflicts with focal set A.
- Replacing $K(A)$ by

$$CON(A) = \sum_{\emptyset\neq B\subseteq\Omega} m(B)\frac{|A\setminus B|}{|A|},$$

  we get another conflict measure, called strife (Klir and Yuan, 1993).
- Both dissonance and strife generalize the Shannon entropy.

Measures of total uncertainty (1/2)

- Measure the degree of uncertainty of a belief function, taking into account the two dimensions of imprecision and conflict.
- Composite measures, e.g.,
  - $N(m) + S(m)$
  - Total uncertainty (Pal et al., 1993)

$$H(m) = -\sum_{\emptyset \neq A \subseteq \Omega} m(A) \log_2 \frac{|A|}{m(A)} = N(m) - \sum_{\emptyset \neq A \subseteq \Omega} m(A) \log_2 m(A)$$

- Agregate uncertainty

$$AU(m) = \max_{p \in \mathcal{P}(m)} \left(-\sum_{\omega \in \Omega} p(\omega) \log_2 p(\omega) \right)$$

where $\mathcal{P}(m)$ is the credal set of m.

Measures of total uncertainty (2/2)

- Other idea: transform m into a probability distribution and compute the corresponding Shannon entropy. Examples:
  1. Jousselme et al. (2006):

$$EP(m) = -\sum_{\omega \in \Omega} betp_m(\omega) \log_2 betp_m(\omega)$$

     where $betp_m$ the pignistic probability distribution is defined by

$$betp_m(\omega) = \sum_{A \subseteq \Omega : \omega \in A} \frac{m(A)}{|A|}$$

  2. Jirousek and Shenoy (2017)

$$H_{js}(m) = -\sum_{\omega \in \Omega} pl^*(\omega) \log_2 pl^*(\omega) + N(m)$$

     where $pl^*(\omega) = pl(\omega) / \sum_{\omega' \in \Omega} pl(\omega')$ is the normalized plausibility.
- Both measures extend the Hartley measure and the Shannon entropy.

Application of uncertainty measures

- Assume we are given (e.g., by an expert) some constraints that an unknown mass function m should satisfy, e.g., $Pl(A_i) = \alpha_i$, $Bel(A_i) \geq \beta_j$, etc.
- A minimally committed mass function can be found by maximizing some uncertainty measure $U(m)$, under the given constraints.
- With $U(m) = N(m)$ and linear constraints of the form $Bel(A_i) \geq \beta_j$, $Bel(A_i) \leq \beta_j$ or $Bel(A_i) = \beta_j$, we have a linear optimization problem, but the solution is generally not unique.
- With other measures and arbitrary constraints, we generally have to solve a non linear optimization problem.

Combination under unknown dependence (1/2)

- Consider two random sets (S_1, P_1, Γ_1) and (S_2, P_2, Γ_2) generating two mass functions m_1 and m_2.
- Let P_{12} on $S_1 \times S_2$ be a joint probability measure with marginals P_1 and P_2.
- Let $A_1, \ldots, A_r$ denote the focal sets of m_1, $B_1, \ldots, B_s$ the focal sets of m_2, $p_i = m_1(A_i)$, $q_j = m_2(B_j)$, and

$$p_{ij} = P_{12}\left(\{(s_1, s_2) \in S_1 \times S_2 \mid \Gamma_1(s_1) = A_i, \Gamma_2(s_2) = B_j\}\right).$$

- Assuming both sources to be reliable, the unnormalized combined mass function m has the following expression:

$$m(A) = \sum_{A_i \cap B_j = A} p_{ij}, \quad \forall A \subseteq \Omega.$$

- Independence assumption of Dempster's rule: $\forall (i,j)$, $p_{ij} = p_i p_j$.
- How to find the p_{ij}'s when the independence assumption is relaxed?

Combination under unknown dependence (2/2)

- Maximizing the Shannon entropy of the p_{ij}'s yields Dempster's rule.
- A least specific combined mass function (without normalization) can be found by solving the following linear optimization problem:

$$\max_{p_{ij}} \sum_{\{(i,j)|A_i \cap B_i \neq \emptyset\}} p_{ij} \log_2 |A_i \cap B_j|$$

under the constraints $\sum_{i,j} p_{ij} = 1$ and

$$\sum_i p_{ij} = p_j, \quad j = 1, \ldots, s$$

$$\sum_j p_{ij} = p_i, \quad i = 1, \ldots, r$$

- The mass function obtained as as solution of the above problem can be normalized.

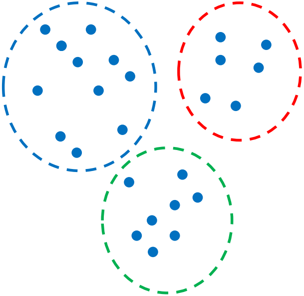
Outline

Decomposition approach

- In the original approach introduced by Dempster and Shafer, the available evidence is broken down into elementary items, each modeled by a mass function. The mass functions are then combined by Dempster's rule.
- Contrary to a common opinion, this approach can be applied even in situations where the frame of discernment is very large, provided
  - The combined mass functions have a simple form
  - We do not need to compute the full combined belief function, but only some partial information useful, e.g., for decision making.
- Two examples in which elementary mass functions are defined based on distances:
  1. Clustering
  2. Association

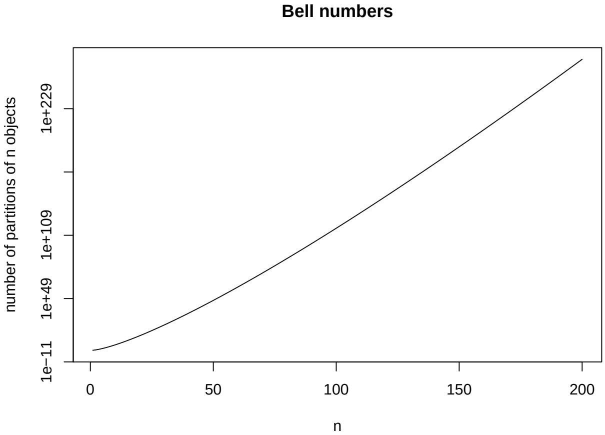
Outline

Clustering



- Finding a meaningful partition of a dataset.
- Assuming there is a true unknown partition, our frame of discernment should be the set $\mathcal{R}$ of all partitions of the set of n objects.
- But this set is huge!

Number of partitions of *n* objects



- Number of atoms in the universe $\approx 10^{80}$
- Can we implement evidential reasoning in such a large space?

Model

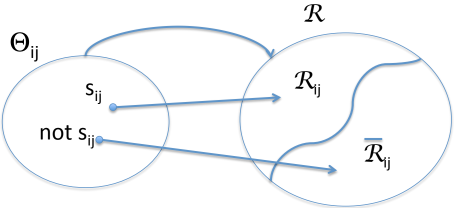
- Evidence: $n \times n$ matrix $D = (d_{ij})$ of dissimilarities between the n objects.
- For any $i < j$, let $\Theta_{ij} = \{s_{ij}, \neg s_{ij}\}$, where s_{ij} means "objects i and j belong to the same group" and $\neg s_{ij}$ is the negation of s_{ij}.
- Assumptions:
  1. Two objects have all the more chance to belong to the same group, that they are more similar. Each dissimilarity is a piece of evidence represented by the following mass function on Θ_{ij},

$$m_{ij}(\{s_{ij}\}) = \varphi(d_{ij}),$$
$$m_{ij}(\Theta_{ij}) = 1 - \varphi(d_{ij}),$$

     where φ is a non-increasing mapping from $[0, +\infty)$ to $[0, 1)$.
  2. The mass functions m_{ij} encode independent pieces of evidence (questionable, but acceptable as an approximation).
- How to combine these $n(n-1)/2$ mass functions to find the most plausible partition of the n objects?

Vacuous extension

- To be combined, the mass functions m_{ij} must be carried to the same frame, which will be the set $\mathcal{R}$ of all partitions of the dataset



- Let $\mathcal{R}_{ij}$ denote the set of partitions of the n objects such that objects o_i and o_j are in the same group ($r_{ij} = 1$).
- Each mass function m_{ij} can be vacuously extended to the $\mathcal{R}$ of all partitions:

$$
\begin{array}{ccc}
m_{ij}(\{s_{ij}\}) & \longrightarrow & \mathcal{R}_{ij} \\
m_{ij}(\Theta) & \longrightarrow & \mathcal{R}
\end{array}
$$

Combination

- The extended mass functions can then be combined by Dempster's rule.
- We will only combine the contour functions. The contour function of m_{ij} is

$$
\begin{aligned}
pl_{ij}(R) &= \begin{cases} m_{ij}(\mathcal{R}_{ij}) + m_{ij}(\mathcal{R}) & \text{if } R \in \mathcal{R}_{ij}, \\ m_{ij}(\mathcal{R}) & \text{otherwise,} \end{cases} \\
&= \begin{cases} 1 & \text{if } r_{ij} = 1, \\ 1 - \varphi(d_{ij}) & \text{otherwise,} \end{cases} \\
&= (1 - \varphi(d_{ij}))^{1 - r_{ij}}
\end{aligned}
$$

- Combined contour function:

$$
pl(R) \propto \prod_{i<j} (1 - \varphi(d_{ij}))^{1 - r_{ij}}
$$

for any $R \in \mathcal{R}$.

Decision

- The logarithm of the contour function can be written as

$$\ln pl(R) = -\sum_{i<j} r_{ij} \log(1 - \varphi(d_{ij})) + C$$

- Finding the most plausible partition is thus a binary linear programming problem. It can be solved exactly only for small n.
- However, the problem can be solved approximately using a heuristic greedy search procedure: the Ek-NNclus algorithm (Denoeux et al., 2015).

Outline

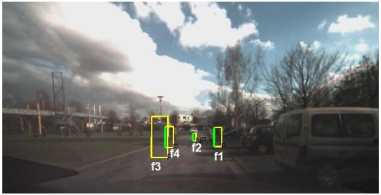
1. Least Commitment Principle
   - LC mass function with given contour function
   - Conditional embedding
   - Uncertainty measures

2. Combining elementary mass functions
   - Clustering
   - Object association

3. Predictive belief functions
   - Continuous belief functions
   - Application to prediction

Problem description

- Let $E = \{e_1, \ldots, e_n\}$ and $F = \{f_1, \ldots, f_p\}$ be two sets of objects perceived by two sensors, or by one sensor at two different times.
- Problem: given information about each object (position, velocity, class, etc.), find a matching between the two sets, in such a way that each object in one set is matched with at most one object in the other set.



Method of approach

1. For each pair of objects $(e_i, f_j) \in E \times F$, use sensor information to build a pairwise mass function m_{ij} on the frame $\Theta_{ij} = \{s_{ij}, \neg s_{ij}\}$, where
   - s_{ij} denotes the hypothesis that e_i and f_j are the same objects, and
   - $\neg s_{ij}$ is the negation of s_{ij}.
2. Vacuously extend the np mass functions m_{ij} in the frame $\mathcal{R}$ containing all admissible matching relations.
3. Combine the np extended mass functions $m_{ij}^{\uparrow \mathcal{R}}$ and find the matching relation with the highest plausibility.

Building the pairwise mass functions

Using position information

- Assume that each sensor provides an estimated position for each object. Let d_{ij} denote the distance between the estimated positions of e_i and f_j, computed using some distance measure.
- A small value of d_{ij} supports hypothesis s_{ij}, while a large value of d_{ij} supports hypothesis $\neg s_{ij}$. Depending on sensor reliability, a fraction of the unit mass should also be assigned to $\Theta_{ij} = \{s_{ij}, \neg s_{ij}\}$.
- This line of reasoning justifies a mass function $m_{ij}^{(p)}$ of the form:

$$m_{ij}^{(p)}(\{s_{ij}\}) = \alpha\varphi(d_{ij})$$
$$m_{ij}^{(p)}(\{\neg s_{ij}\}) = \alpha\left(1 - \varphi(d_{ij})\right)$$
$$m_{ij}^{(p)}(\Theta_{ij}) = 1 - \alpha,$$

where $\alpha \in [0, 1]$ is a degree of confidence in the sensor information and φ is a decreasing function taking values in $[0, 1]$.

Building the pairwise mass functions

Using velocity information

- Let us now assume that each sensor returns a velocity vector for each object. Let d'_{ij} denote the distance between the velocities of objects e_i and f_j.
- Here, a large value of d'_{ij} supports the hypothesis $\neg s_{ij}$, whereas a small value of d'_{ij} does not support specifically s_{ij} or $\neg s_{ij}$, as two distinct objects may have similar velocities.
- Consequently, the following form of the mass function $m_{ij}^{(v)}$ induce by d'_{ij} seems appropriate:

$$
\begin{aligned}
m_{ij}^{(v)}(\{\neg s_{ij}\}) &= \alpha' \left(1 - \psi(d'_{ij})\right) \\
m_{ij}^{(v)}(\Theta_{ij}) &= 1 - \alpha' \left(1 - \psi(d'_{ij})\right),
\end{aligned}
$$

where $\alpha' \in [0, 1]$ is a degree of confidence in the sensor information and ψ is a decreasing function taking values in $[0, 1]$.

Building the pairwise mass functions

Using class information

- Let us assume that the objects belong to classes. Let Ω be the set of possible classes, and let m_i and m_j denote mass functions representing evidence about the class membership of objects e_i and f_j.
- If e_i and f_j do not belong to the same class, they cannot be the same object. However, if e_i and f_j do belong to the same class, they may or may not be the same object.
- Using this line of reasoning, we can show that the mass function $m_{ij}^{(c)}$ on Θ_{ij} derived from m_i and m_j has the following expression:

$$m_{ij}^{(c)}(\{\neg s_{ij}\}) = \kappa_{ij}$$

$$m_{ij}^{(c)}(\Theta_{ij}) = 1 - \kappa_{ij},$$

where κ_{ij} is the degree of conflict between m_i and m_j.

Building the pairwise mass functions

Aggregation and vacuous extension

- For each object pair (e_i, f_j), a pairwise mass function $m^{\Theta_{ij}}$ representing all the available evidence about Θ_{ij} can finally be obtained as:

$$m_{ij} = m_{ij}^{(p)} \oplus m_{ij}^{(v)} \oplus m_{ij}^{(c)}.$$

- Let $\mathcal{R}$ be the set of all admissible matching relations, and let $\mathcal{R}_{ij} \subseteq \mathcal{R}$ be the subset of relations R such that $(e_i, f_j) \in R$.
- Vacuously extending m_{ij} in $\mathcal{R}$ yields the following mass function:

$$m_{ij}^{\uparrow\mathcal{R}}(\mathcal{R}_{ij}) = m_{ij}(\{s_{ij}\}) = \alpha_{ij}$$
$$m_{ij}^{\uparrow\mathcal{R}}(\overline{\mathcal{R}_{ij}}) = m_{ij}(\{\neg s_{ij}\}) = \beta_{ij}$$
$$m_{ij}^{\uparrow\mathcal{R}}(\mathcal{R}) = m_{ij}(\Theta_{ij}) = 1 - \alpha_{ij} - \beta_{ij}.$$

Combining pairwise mass functions

- Let pl_{ij} denote the contour function corresponding to $m_{ij}^{\uparrow\mathcal{R}}$. For all $R \in \mathcal{R}$,

$$pl_{ij}(R) = \begin{cases} 1-\beta_{ij} & \text{if } R \in \mathcal{R}_{ij}, \\ 1-\alpha_{ij} & \text{otherwise}, \end{cases}$$
$$= (1-\beta_{ij})^{r_{ij}}(1-\alpha_{ij})^{1-r_{ij}}$$

- Consequently, the contour function corresponding to the combined mass function

$$m^{\mathcal{R}} = \bigoplus_{i,j} m_{ij}^{\uparrow\mathcal{R}}$$

is

$$pl(R) \propto \prod_{i,j} (1-\beta_{ij})^{r_{ij}}(1-\alpha_{ij})^{1-r_{ij}}.$$

Finding the most plausible matching

- We have

$$\ln pl(R) = \sum_{i,j} [r_{ij} \ln(1-\beta_{ij}) + (1-r_{ij}) \ln(1-\alpha_{ij})] + C.$$

- The most plausible relation R^* can thus be found by solving the following binary linear optimization problem:

$$\max \sum_{i=1}^{n} \sum_{j=1}^{p} r_{ij} \ln \frac{1-\beta_{ij}}{1-\alpha_{ij}}$$

subject to $\sum_{j=1}^{p} r_{ij} \leq 1, \forall i$ and $\sum_{i=1}^{n} r_{ij} \leq 1, \forall j$.

- This problem can be shown to be equivalent to a linear assignment problem and can be solved in $o(\max(n,m)^3)$ time.

Outline

1. Least Commitment Principle
   - LC mass function with given contour function
   - Conditional embedding
   - Uncertainty measures
2. Combining elementary mass functions
   - Clustering
   - Object association
3. Predictive belief functions
   - Continuous belief functions
   - Application to prediction

Prediction vs. estimation

- Consider an urn with an unknown proportion θ of black balls
- Assume that we have drawn n balls with replacement from the urn, x of which were black
- Two categories of problems:

  Estimation: What can we say about θ?

  Prediction: What can we say about the color Y of the next ball to be drawn from the urn?

- Both kinds of problems have been addressed in the DS framework, starting from Dempster's original work.
- Problem addressed in this lecture:

How to quantify uncertainty in statistical prediction problems?

- We need to construct and manipulate continuous belief functions.

Outline

Belief function: general definition

Definition (Belief function)

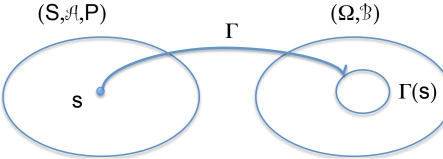
- *Let Ω be a set and $\mathcal{B}$ be an algebra of subsets of Ω (a nonempty family of subsets of Ω, closed under complementation and finite intersection).*
- *A mapping $Bel : \mathcal{B} \to [0,1]$ is a belief function (BF) iff $Bel(\emptyset) = 0$, $Bel(\Omega) = 1$ and Bel is completely monotone: for any $k \geq 2$ and any collection $B_1, \ldots, B_k$ of elements of $\mathcal{B}$,*

$$Bel\left(\bigcup_{i=1}^{k} B_i\right) \geq \sum_{\emptyset \neq I \subseteq \{1,\ldots,k\}} (-1)^{|I|+1} Bel\left(\bigcap_{i \in I} B_i\right)$$

Definition (Plausibility function)

Given a belief function $Bel : \mathcal{B} \to [0,1]$, the function $Pl : \mathcal{B} \to [0,1]$ such that $Pl(B) = 1 - Bel(\overline{B})$ is called its dual plausibility function.

Belief function induced by a random set



Consider a random set defined by a probability space $(S, \mathcal{A}, \mathbb{P})$, a set Ω equipped with an algebra $\mathcal{B}$ and a multi-valued mapping Γ from S to $2^{\Omega} \setminus \emptyset$.

Proposition

Under measurability conditions, the lower probability measure defined by

$$\mathbb{P}_*(B) = \mathbb{P}(\{s \in S \mid \Gamma(s) \subseteq B\}), \quad \forall B \in \mathcal{B}$$

is a belief function, and the upper probability measure

$$\mathbb{P}^*(B) = \mathbb{P}(\{s \in S \mid \Gamma(s) \cap B \neq \emptyset\}), \quad \forall B \in \mathcal{B}$$

is the corresponding dual plausibility function.

Possibility measures

- If, for any $(s, s') \in S^2$, $\Gamma(s) \subseteq \Gamma(s')$ or $\Gamma(s') \subseteq \Gamma(s)$, the BF *Bel* is said to be consonant.
- The plausibility distribution is then a possibility measure: it verifies

$$Pl(A \cup B) = \max\left(Pl(A), Pl(B)\right), \quad \forall (A, B) \in \mathcal{B}^2,$$

and

$$Pl(A) = \sup_{\omega \in A} pl(\omega),$$

where the mapping $pl : \omega \to Pl(\{\omega\})$ (called the contour function of *Bel*) is the corresponding possibility distribution.

Monte Carlo approximation

- Except in very simple cases, it is usually impossible to derive exact expressions for
$$Bel(B) = \mathbb{P}(\{s \in S \mid \Gamma(s) \subseteq B\})$$
and
$$Pl(B) = \mathbb{P}(\{s \in S \mid \Gamma(s) \cap B \neq \emptyset\})$$
for a given $B \in \mathcal{B}$.
- We can approximate these quantities by drawing N elements $s_1, \ldots, s_N$ of S randomly from $\mathbb{P}$. By the law of large numbers,
$$\widehat{Bel}(B) = \frac{1}{N} \sum_{i=1}^{N} I(\Gamma(s_i) \subseteq B) \xrightarrow{a.s.} Bel(B)$$
and
$$\widehat{Pl}(B) = \frac{1}{N} \sum_{i=1}^{N} I(\Gamma(s_i) \cap B \neq \emptyset) \xrightarrow{a.s.} Pl(B).$$
as $N \longrightarrow \infty$.

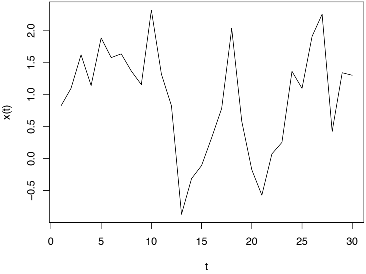
Outline

Example of a prediction problem

As an example of a statistical prediction problem, consider an AR(1) model

$$X_t = \rho X_{t-1} + \epsilon_t, \quad t = 1, 2, \ldots,$$

where $\rho \in (-1, 1)$ is a parameter and $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$.



Problem: having observed $\boldsymbol{x}_{1:T} = (x_1, \ldots, x_T)$, predict the next h future values $\boldsymbol{Y} = (X_{T+1}, \ldots, X_{T+h})$.

General approach

Approach: express $\boldsymbol{Y}$ as a function of the parameter $\boldsymbol{\theta} = (\rho, \sigma)$ and a random vector with known distribution.

- For instance, assuming $h = 2$, we can write

$$
\begin{aligned}
X_{T+1} &= \rho x_T + \epsilon_{T+1} = \rho x_T + \sigma \Phi^{-1}(U_1) \\
X_{T+2} &= \rho X_{T+1} + \epsilon_{T+2} \\
&= \rho^2 x_T + \rho\sigma\Phi^{-1}(U_1) + \sigma\Phi^{-1}(U_2)
\end{aligned}
$$

  with $U_1, U_2 \sim \text{Unif}(0,1)$, so we have

$$
\boxed{\boldsymbol{Y} = (X_{T+1}, X_{T+2}) = \varphi(\boldsymbol{\theta}, \boldsymbol{U})}
$$

  where $\boldsymbol{U} = (U_1, U_2) \sim \text{Unif}([0,1]^2)$.

Random vs. epistemic uncertainty

The "φ-equation"

$$\boldsymbol{Y} = \varphi(\boldsymbol{\theta}, \boldsymbol{U})$$

allows us to separate the two sources of uncertainty on $\boldsymbol{Y}$:

1. Uncertainty on $\boldsymbol{U}$ (random/aleatory uncertainty)
2. Uncertainty on $\boldsymbol{\theta}$ (epistemic uncertainty)

Two-step method:

1. Represent uncertainty on $\boldsymbol{\theta}$ using an estimative belief function $Bel_{\boldsymbol{\theta}}$ constructed from the observed data
2. Combine $Bel_{\boldsymbol{\theta}}$ with the probability distribution of $\boldsymbol{U}$ to obtain a predictive belief function $Bel_{\boldsymbol{Y}}$

Properties of the predictive BF

The properties of the predictive BF $Bel_{\boldsymbol{Y}}$ depends on the BF $Bel_{\boldsymbol{\theta}}$:

- If $Bel_{\boldsymbol{\theta}}(\{\theta_0\}) = 1$, where θ_0 is the true value of $\boldsymbol{\theta}$, then $Bel_{\boldsymbol{Y}}$ is the true probability distribution of $\boldsymbol{Y}$ given $\boldsymbol{x}_T$.
- If $Bel_{\boldsymbol{\theta}}(\{\widehat{\boldsymbol{\theta}}\}) = 1$, where $\widehat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, then $Bel_{\boldsymbol{Y}}$ is the plug-in estimate of the true probability distribution of $\boldsymbol{Y}$ given $\boldsymbol{x}_T$.
- If $Bel_{\boldsymbol{\theta}}(A) = I(R_{1-\alpha} \subseteq A)$, where $R_{1-\alpha}$ is a $1-\alpha$ confidence region on $\boldsymbol{\theta}$, then $Bel_{\boldsymbol{Y}}$ has a frequentist property: it is dominated by the true conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{x}_T$ with probability $1-\alpha$.
- If $Bel_{\boldsymbol{\theta}}$ is the likelihood-based BF, then $Bel_{\boldsymbol{Y}}$ generalizes the Bayesian posterior probability distribution of $\boldsymbol{Y}$.

Here, I focus on the last method as an illustration.

Likelihood-based belief function

Definition (Likelihood-based belief function)

The likelihood-based belief function is the consonant BF with contour function (possibility distribution)

$$pl(\theta) = \frac{p(\mathbf{x}_{1:T}; \theta)}{p(\mathbf{x}_{1:T}; \widehat{\theta})} = \frac{L(\theta; \mathbf{x}_{1:T})}{L(\widehat{\theta}; \mathbf{x}_{1:T})},$$

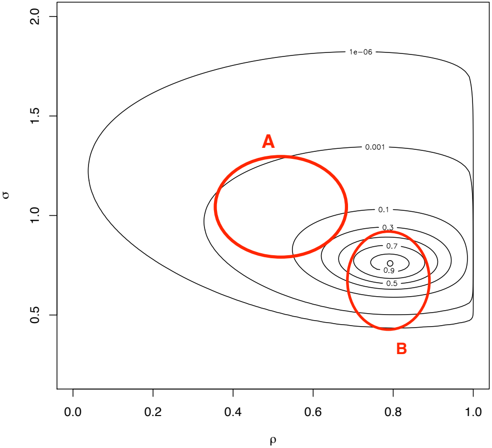
where L denotes the likelihood function and $\widehat{\theta}$ the MLE of θ. It represents the information provided by the data about θ.

- We then have

$$Pl_\theta(A) = \sup_{\theta \in A} pl(\theta) \quad \text{for all } A \subseteq \Theta.$$

- Combining Bel_θ with a Bayesian prior on θ then yields the Bayesian posterior.
- Justified by axiomatic arguments (Denœux, 2014).

Example



$$Pl_\theta(A) = 0.3$$

$$Bel_\theta(A) = 1 - Pl_\theta(\overline{A}) = 0$$

$$Pl_\theta(B) = 1$$

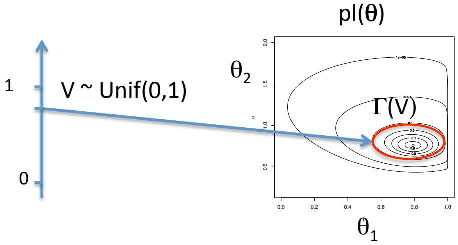
$$Bel_\theta(B) = 1 - Pl_\theta(\overline{B}) = 0.3$$

Random set representation of Bel_θ

- We can show that the likelihood-based BF Bel_θ on θ is induced by the random set

$$\Gamma(V) = \{\theta \in \mathbf{\Theta} \mid pl(\theta) \geq V\}$$

with $V \sim \mathrm{Unif}(0,1)$.

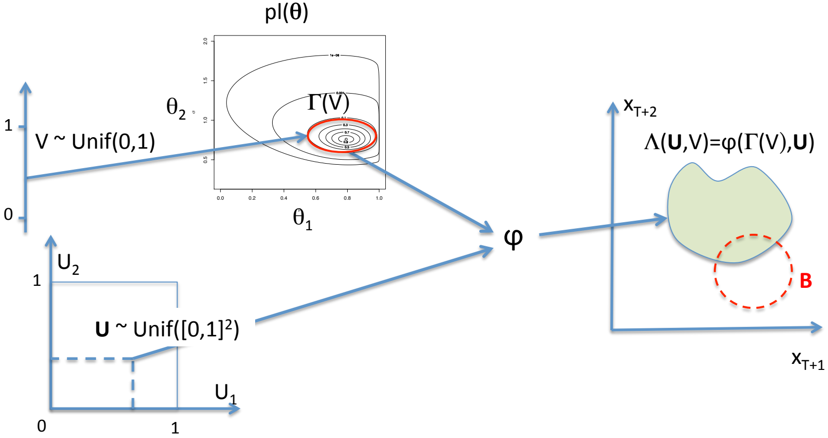


- We have

$$Bel_\theta(A) = \mathbb{P}(\Gamma(V) \subseteq A) \quad \text{and} \quad Pl_\theta(A) = \mathbb{P}(\Gamma(V) \cap A \neq \emptyset)$$

Predictive belief function

A predictive BF $Bel_{\boldsymbol{Y}}$ on $\boldsymbol{Y}$ is obtained by propagating Bel_{θ} together with the probability distribution of $\boldsymbol{U}$ through the φ-equation $\boldsymbol{Y} = \varphi(\theta, \boldsymbol{U})$:



The mapping $\Lambda : (\boldsymbol{U}, V) \to \varphi(\Gamma(V), \boldsymbol{U})$ defines the predictive BF $Bel_{\boldsymbol{Y}}$ on $\boldsymbol{Y}$.

Practical computation of $Bel_{\boldsymbol{Y}}$

The belief and plausibility degrees of events $B \subseteq \mathbb{R}^h$, defined as

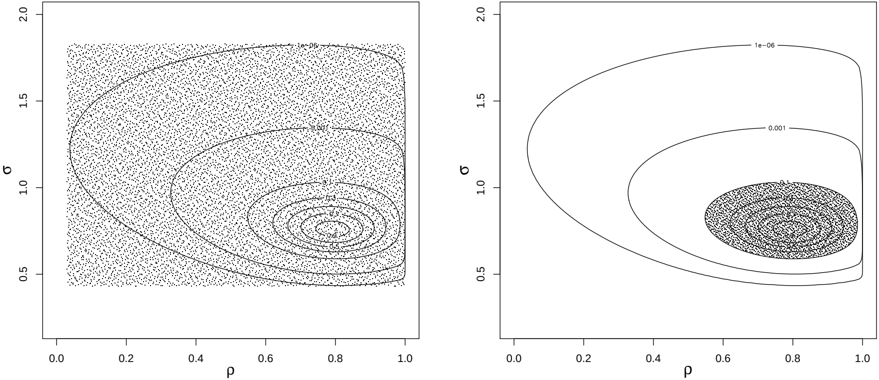
$$Bel_{\boldsymbol{Y}}(B) = \mathbb{P}\left(\varphi(\Gamma(V), \boldsymbol{U}) \subseteq B\right), \quad \text{and}$$

$$Pl_{\boldsymbol{Y}}(B) = \mathbb{P}\left(\varphi(\Gamma(V), \boldsymbol{U}) \cap B \neq \emptyset\right)$$

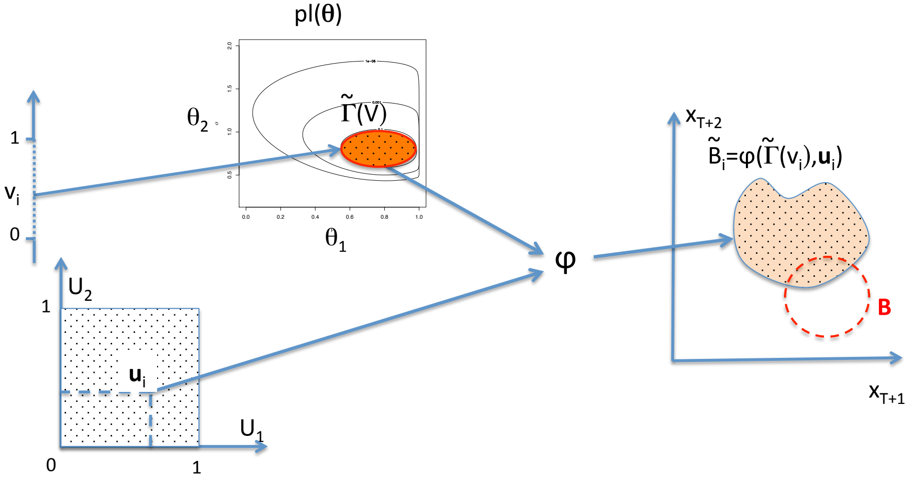
can be approximated by combining Monte Carlo (MC) simulation and set representation techniques:

- We start by approximating the parameter space $\boldsymbol{\Theta}$ by a finite set of points $\widetilde{\boldsymbol{\Theta}} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\} \subset \boldsymbol{\Theta}$.
- For each $v \in [0, 1]$, the set $\Gamma(v)$ is approximated by the finite set $\widetilde{\Gamma}(v) = \{\boldsymbol{\theta} \in \widetilde{\boldsymbol{\Theta}} \mid pl(\boldsymbol{\theta}) > v\}$.
- The distributions of $\boldsymbol{U}$ and V are approximated by MC simulation

Point cloud representation



Point cloud propagation

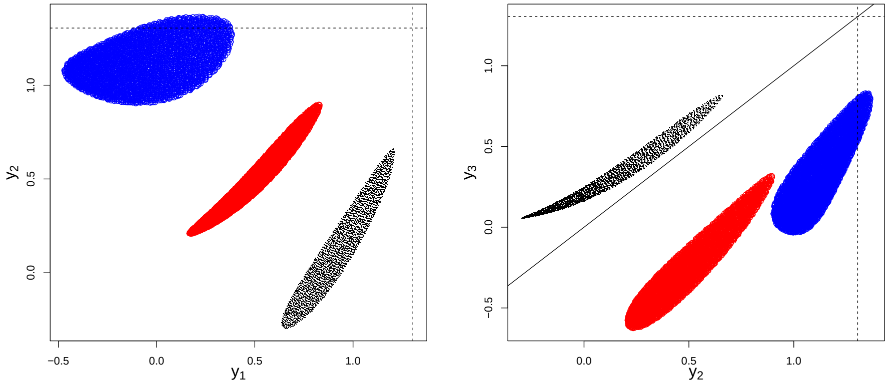


Point cloud propagation algorithm

Require: Point cloud $\widetilde{\boldsymbol{\Theta}} := \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\} \subset \boldsymbol{\Theta}$
Require: Desired number of focal sets N
for $i = 1$ to N **do**
Draw independently v_i from Unif([0, 1]) and $\boldsymbol{u}_i$ from Unif($[0, 1]^2$)
Find $\widetilde{\Gamma}(v_i) := \{\boldsymbol{\theta} \in \widetilde{\boldsymbol{\Theta}} \mid pl(\boldsymbol{\theta}) > v_i\}$
Compute $\widetilde{B}_i := \varphi(\widetilde{\Gamma}(v_i), \boldsymbol{u}_i)$
end for
$\widehat{Bel}(B) := \frac{1}{N}\sum_{i=1}^{N} I(\widetilde{B}_i \subseteq B)$
$\widehat{Pl}(B) := \frac{1}{N}\sum_{i=1}^{N} I(\widetilde{B}_i \cap B \neq \emptyset)$

Example

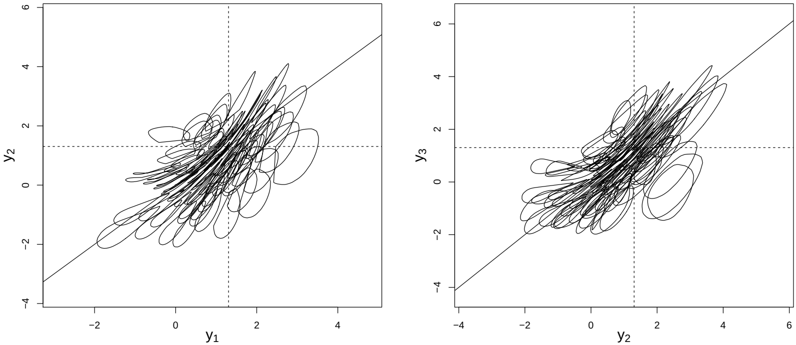
Approximated focal sets ($h = 3$)



2-D projections of focal sets $\widetilde{B}_i = \varphi(\widetilde{\Gamma}(v_i), \boldsymbol{u}_i)$ for $v_i = 0.1$ and three different values of $\boldsymbol{u}_i$

Example

100 focal sets $\widetilde{\Lambda}(s_i, \boldsymbol{u}_i)$ ($h = 3$)



Convex hulls of the two-dimensional projections of 100 focal sets $\widetilde{B}_i = \varphi(\widetilde{\Gamma}(v_i), \boldsymbol{u}_i)$ on the planes spanned by (Y_1, Y_2) (left) and (Y_2, Y_3) (right)

Example

Belief and plausibility of some events

Event	*Bel*	*Pl*	True proba.
$(X_{T+1} > X_{T+2} > X_{T+3})$	0.26	0.38	0.31
$(X_{T+1} < 0)$ & $(X_{T+2} < 0)$ & $(X_{T+3} < 0)$	0.024	0.073	0.086
$(X_{T+1} > 0)$ & $(X_{T+2} > 0)$ & $(X_{T+3} > 0)$	0.51	0.75	0.50

Summary

- Developing practical applications using the Dempster-Shafer framework requires modeling expert knowledge and statistical information using belief functions:
- Systematic and principled methods now exist:
  - Least-commitment principle
  - GBT
  - Likelihood-based belief function
  - Predictive belief functions
  - etc.
- Specific methods will be studied in following lectures (correction mechanisms, classification, clustering, etc.).
- More research on expert knowledge elicitation and statistical inference is needed.

References I

cf. `https://www.hds.utc.fr/~tdenoeux`

T. Denœux and P. Smets.
Classification using Belief Functions: the Relationship between the Case-based and Model-based Approaches
IEEE Transactions on Systems, Man and Cybernetics B, 36(6):1395–1406, 2006.

T. Denœux, O. Kanjanatarakul and S. Sriboonchitta.
EK-NNclus: a clustering procedure based on the evidential K-nearest neighbor rule.
Knowledge-Based Systems, 88:57–69, 2015.

T. Denœux, N. El Zoghby, V. Cherfaoui and A. Jouglet.
Optimal object association in the Dempster-Shafer framework.
IEEE Transactions on Cybernetics, 44(22):2521–2531, 2014.

References II

cf. https://www.hds.utc.fr/~tdenoeux

T. Denœux.
Likelihood-based belief function: justification and some extensions to low-quality data.
International Journal of Approximate Reasoning, 55(7):1535–1547, 2014.

O. Kanjanatarakul, T. Denœux and S. Sriboonchitta.
Prediction of future observations using belief functions: a likelihood-based approach.
International Journal of Approximate Reasoning, 72:71–94, 2016.

T. Denœux and S. Li.
Frequency-Calibrated Belief Functions: Review and New Insights.
International Journal of Approximate Reasoning, 92:232–254, 2018.

References III

cf. https://www.hds.utc.fr/~tdenoeux

T. Denœux and O. Kanjanatarakul.
Multistep Prediction using Point-Cloud Approximation of Continuous Belief Functions.
2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2019), New Orleans, USA, June 23-26, 2019.