Belief functions: basic theory and applications

Thierry Denœux[1]

[1] Université de Technologie de Compiègne, France
HEUDIASYC (UMR CNRS 7253)
https://www.hds.utc.fr/˜tdenoeux

ISIPTA 2013,
Compiègne, France,
July 1st, 2013

Foreword

- The theory of belief functions (BF) is not a theory of Imprecise Probability (IP)! In particular, it does not represent uncertainty using sets of probability measures.
- However, as IP theory, BF theory does extend Probability theory by allowing some imprecision (using a multi-valued mapping in the case of belief functions).
- These two theories can be seen as implementing two ways of mixing set-based representations of uncertainty and Probability theory:
  - By defining sets of probability measures (IP theory);
  - By assigning probability masses to sets (BF theory).

Theory of belief functions

History

- Also known as Dempster-Shafer (DS) theory or Evidence theory.
- Originates from the work of Dempster (1968) in the context of statistical inference.
- Formalized by Shafer (1976) as a theory of evidence.
- Popularized and developed by Smets in the 1980's and 1990's under the name Transferable Belief Model.
- Starting from the 1990's, growing number of applications in AI, information fusion, classification, reliability and risk analysis, etc.

Theory of belief functions

Important ideas

1. DS theory: a modeling language for representing elementary items of evidence and combining them, in order to form a representation of our beliefs about certain aspects of the world.
2. The theory of belief function subsumes both the set-based and probabilistic approaches to uncertainty:
   - A belief function may be viewed both as a generalized set and as a non additive measure.
   - Basic mechanisms for reasoning with belief functions extend both probabilistic operations (such as marginalization and conditioning) and set-theoretic operations (such as intersection and union).
3. DS reasoning produces the same results as probabilistic reasoning or interval analysis when provided with the same information. However, its greater expressive power allows us to handle more general forms of information.

Outline

Outline

1. Basic theory
   - Representation of evidence
   - Operations on Belief functions
   - Decision making
2. Applications
   - Classification
   - Preference aggregation
   - Object association
3. Statistical inference
   - Dempster's approach
   - Likelihood-based approach
   - Sea level rise example

Outline

Mass function
Definition

- Let ω be an unknown quantity with possible values in a finite domain Ω, called the frame of discernment.
- A piece of evidence about ω may be represented by a mass function m on Ω, defined as a function $2^\Omega \rightarrow [0,1]$, such that $m(\emptyset) = 0$ and
$$\sum_{A \subseteq \Omega} m(A) = 1.$$
- Any subset A of Ω such that $m(A) > 0$ is called a focal set of m. Special cases:
  - A logical mass function has only one focal set ($\sim$ set).
  - A Bayesian mass function has only focal sets of cardinality one ($\sim$ probability distribution).
- The vacuous mass function is defined by $m_\Omega(\omega) = 1$. It represents a completely uninformative piece of evidence.

Mass function
Example

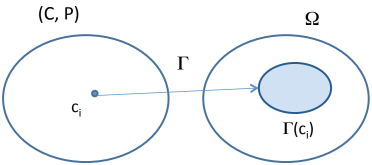
- A murder has been committed. There are three suspects: $\Omega = \{Peter, John, Mary\}$.
- A witness saw the murderer going away, but he is short-sighted and he only saw that it was a man. We know that the witness is drunk 20 % of the time.
- If the witness was not drunk, we know that $\omega \in \{Peter, John\}$. Otherwise, we only know that $\omega \in \Omega$. The first case holds with probability 0.8.
- Corresponding mass function:

$$m(\{Peter, John\}) = 0.8, \quad m(\Omega) = 0.2$$

Semantics

- What do these numbers mean?
- In the murder example, the evidence can be interpreted in two different ways and we can assign probabilities to the different interpretations:
  - With probability 0.8, we know that the murderer is either Peter or John;
  - With probability 0.2, we know nothing.
- A DS mass function encodes probability judgements about the reliability and meaning of a piece of evidence.
- It can be constructed by comparing our evidence to a situation where we receive a message that was encoded using a code selected at random with known probabilities.

Random code semantics



- A source holds some true information of the form $\omega \in A^*$ for some $A^* \subseteq \Omega$;
- It sends us this information as an encoded message using a code in $\mathcal{C} = \{c_1, \ldots, c_r\}$, selected at random according to some known probability measure on P;
- Decoding the message using code c produces a new message of the form "$\omega \in \Gamma(c) \subseteq \Omega$".

Then,

$$\forall A \subseteq \Omega, \quad m(A) = P(\{c \in \mathcal{C} | \Gamma(c) = A\})$$

is the chance that the original message was "$\omega \in A$", i.e., the probability of knowing only that $\omega \in A$.

Belief and plausibility functions

- For any $A \subseteq \Omega$, we can define:
  - The total degree of support (belief) in A as the probability that the evidence implies A:

$$Bel(A) = P(\{c \in \mathcal{C} | \Gamma(c) \subseteq A\}) = \sum_{B \subseteq A} m(B).$$

  - The plausibility of A as the probability that the evidence does not contradict A:

$$Pl(A) = P(\{c \in \mathcal{C} | \Gamma(c) \cap A \neq \emptyset\}) = 1 - Bel(\overline{A}).$$

- Uncertainty on the truth value of the proposition "$\omega \in A$" is represented by two numbers: $Bel(A)$ and $Pl(A)$, with $Bel(A) \leq Pl(A)$.

Characterization of belief functions

- Function $Bel : 2^\Omega \rightarrow [0,1]$ is a completely monotone capacity, i.e., it verifies $Bel(\emptyset) = 0$, $Bel(\Omega) = 1$ and

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

  for any $k \geq 2$ and for any family $A_1, \ldots, A_k$ in 2^Ω.
- Conversely, to any completely monotone capacity *Bel* corresponds a unique mass function m such that:

$$m(A) = \sum_{\emptyset \neq B \subseteq A} (-1)^{|A|-|B|} Bel(B), \quad \forall A \subseteq \Omega.$$

- m, *Bel* and *Pl* are thus three equivalent representations of a piece of evidence.

Special cases

- If all focal sets of m are singletons, then m is said to be Bayesian: it is equivalent to a probability distribution, and $Bel = Pl$ is a probability measure.
- If the focal sets of m are nested, then m is said to be consonant. Pl is a possibility measure, i.e.,

$$Pl(A \cup B) = \max(Pl(A), Pl(B)), \quad \forall A, B \subseteq \Omega,$$

and Bel is the dual necessity measure. The contour function $pl(\omega) = Pl(\{\omega\})$ is the possibility distribution.

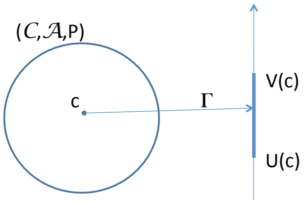
Extension to infinite frames

- In the finite case, we have seen that a belief function *Bel* can be seen as arising from an underlying probability space $(\mathcal{C}, \mathcal{A}, P)$ and a multi-valued mapping $\Gamma : \mathcal{C} \rightarrow 2^{\Omega}$.
- In the general case, given
  - a (finite or not) probability space $(\mathcal{C}, \mathcal{A}, P)$;
  - a (finite or not) measurable space $(\Omega, \mathcal{B})$ and
  - a multi-valued mapping $\Gamma : \mathcal{C} \rightarrow 2^{\Omega}$,

  we can always (under some measurability conditions) define a completely monotone capacity (i.e., belief function) *Bel* as:

$$Bel(A) = P(\{c \in \mathcal{C} | \Gamma(c) \subseteq A\}), \quad \forall A \in \mathcal{B}.$$

Random intervals ($\Omega = \mathbb{R}$)



- Let (U, V) be a two-dimensional random variable from $(\mathcal{C}, \mathcal{A}, P)$ to $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ such that $P(U \leq V) = 1$ and

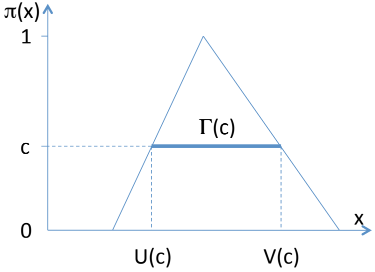
$$\Gamma(c) = [U(c), V(c)] \subseteq \mathbb{R}.$$

- This setting defines a random closed interval, which induces a belief function on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ defined by

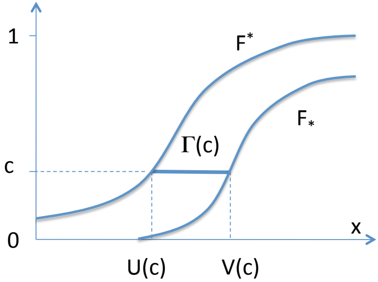
$$Bel(A) = P([U, V] \subseteq A), \quad \forall A \in \mathcal{B}(\mathbb{R}).$$

Examples

Consonant random interval



p-box



Outline

Basic operations on belief functions

1. Combining independent pieces of evidence (Dempster's rule);
2. Expressing evidence in a coarser frame (marginalization);
3. Expressing evidence in a finer frame (vacuous extension);

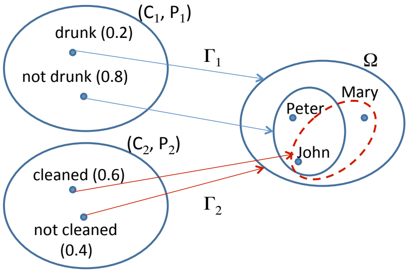
Combination of evidence

Murder example continued

- The first item of evidence gave us: $m_1(\{Peter, John\}) = 0.8$, $m_1(\Omega) = 0.2$.
- New piece of evidence: a blond hair has been found.
- There is a probability 0.6 that the room has been cleaned before the crime: $m_2(\{John, Mary\}) = 0.6$, $m_2(\Omega) = 0.4$.
- How to combine these two pieces of evidence?

Combination of evidence

Problem analysis



- If codes $c_1 \in \mathcal{C}_1$ and $c_2 \in \mathcal{C}_2$ were selected, then we know that $\omega \in \Gamma_1(c_1) \cap \Gamma_2(c_2)$.
- If the codes were selected independently, then the probability that the pair (c_1, c_2) was selected is $P_1(\{c_1\})P_2(\{c_2\})$.
- If $\Gamma_1(c_1) \cap \Gamma_2(c_2) = \emptyset$, we know that (c_1, c_2) could not have been selected.
- The joint probability distribution on $\mathcal{C}_1 \times \mathcal{C}_2$ must be conditioned to eliminate such pairs.

Dempster's rule
Definition

- Let m_1 and m_2 be two mass functions on the same frame Ω, induced by two independent pieces of evidence.
- Their combination using Dempster's rule is defined as:

$$(m_1 \oplus m_2)(A) = \frac{1}{1-K} \sum_{B \cap C = A} m_1(B) m_2(C), \quad \forall A \neq \emptyset,$$

where

$$K = \sum_{B \cap C = \emptyset} m_1(B) m_2(C)$$

is the degree of conflict between m_1 and m_2.
- $m_1 \oplus m_2$ exists iff $K < 1$.

Dempster's rule
Properties

- Commutativity, associativity. Neutral element: m_Ω.
- Generalization of intersection: if m_A and m_B are logical mass functions and $A \cap B \neq \emptyset$, then

$$m_A \oplus m_B = m_{A \cap B}$$

- Generalization of probabilistic conditioning: if m is a Bayesian mass function and m_A is a logical mass function, then $m \oplus m_A$ is a Bayesian mass function that corresponding to the conditioning of m by A.

Dempster's rule
Expression using commonalities

- Commonality function: let $Q : 2^\Omega \rightarrow [0,1]$ be defined as

$$Q(A) = \sum_{B \supseteq A} m(B), \quad \forall A \subseteq \Omega.$$

- Conversely,

$$m(A) = \sum_{B \supseteq A} (-1)^{|B \setminus A|} Q(B)$$

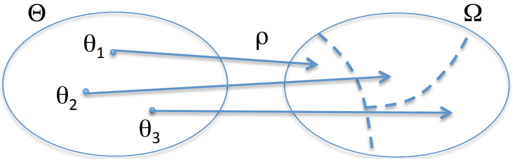
- Expression of $\oplus$ using commonalities:

$$(Q_1 \oplus Q_2)(A) = \frac{1}{1-K} Q_1(A) \cdot Q_2(A), \quad \forall A \subseteq \Omega, A \neq \emptyset.$$

$$(Q_1 \oplus Q_2)(\emptyset) = 1.$$

Refinement of a frame

- Assume we are interested in the nature of an object in a road scene. We could describe it, e.g., in the frame $\Theta = \{\text{vehicle}, \text{pedestrian}\}$, or in the finer frame $\Omega = \{\text{car}, \text{bicycle}, \text{motorcycle}, \text{pedestrian}\}$.
- A frame Ω is a refinement of a frame Θ (or, equivalently, Θ is a coarsening of Ω) if elements of Ω can be obtained by splitting some or all of the elements of Θ.
- Formally, Ω is a refinement of a frame Θ iff there is then a one-to-one mapping ρ between Θ and a partition of Ω:

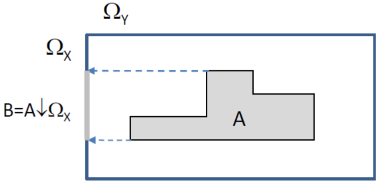


Compatible frames

- Two frames are said to be compatible if they have a common refinement.
- Example:
  - Let $\Omega_X = \{\text{red}, \text{blue}, \text{green}\}$ and $\Omega_Y = \{\text{small}, \text{medium}, \text{large}\}$ be the domains of attributes X and Y describing, respectively, the color and the size of an object.
  - Then Ω_X and Ω_Y have the common refinement $\Omega_X \times \Omega_Y$.

Marginalization

- Let Ω_X and Ω_Y be two compatible frames.
- Let m^{XY} be a mass function on $\Omega_X \times \Omega_Y$.
- It can be expressed in the coarser frame Ω_X by transferring each mass $m^{XY}(A)$ to the projection of A on Ω_X.



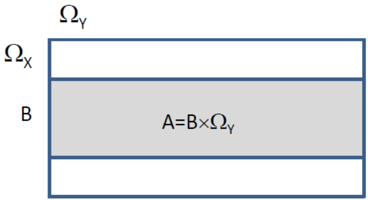
- Marginal mass function:

$$m^{XY\downarrow X}(B) = \sum_{\{A\subseteq\Omega_{XY}, A\downarrow\Omega_X=B\}} m^{XY}(A) \quad \forall B \subseteq \Omega_X.$$

- Generalizes both set projection and probabilistic marginalization.

Vacuous extension

- The "inverse" of marginalization.
- A mass function m^X on Ω_X can be expressed in $\Omega_X \times \Omega_Y$ by transferring each mass $m_X(B)$ to the cylindrical extension of B:



- This operation is called the vacuous extension of m_X in $\Omega_X \times \Omega_Y$. We have

$$m^{X\uparrow XY}(A) = \begin{cases} m^X(B) & \text{if } A = B \times \Omega_Y \\ 0 & \text{otherwise.} \end{cases}$$

Application to uncertain reasoning

- Assume that we have:
  - Partial knowledge of X formalized as a mass function m^X;
  - A joint mass function m^{XY} representing an uncertain relation between X and Y.
- What can we say about Y?
- Solution:

$$m^Y = \left(m^{X \uparrow XY} \oplus m^{XY} \right)^{\downarrow Y}.$$

- Infeasible with many variables and large frames of discernment, but efficient algorithms exist to carry out the operations in frames of minimal dimensions.

Outline

Decision making under uncertainty

- A decision problem can be formalized by defining:
  - A set Ω of states of the world ;
  - A set $\mathcal{X}$ of consequences;
  - a set $\mathcal{F}$ of acts, where an act is a function $f : \Omega \to \mathcal{X}$.
- Let $\succcurlyeq$ be a preference relation on $\mathcal{F}$, such that $f \succcurlyeq g$ means that f is at least as desirable as g.
- Savage (1954) has showed that $\succcurlyeq$ verifies some rationality requirements iff there exists a probability measure P on Ω and a utility function $u : \mathcal{X} \to \mathbb{R}$ such that

$$\forall f, g \in \mathcal{F}, \quad f \succcurlyeq g \Leftrightarrow \mathbb{E}_P(u \circ f) \geq \mathbb{E}_P(u \circ g).$$

  Furthermore, P is unique and u is unique up to a positive affine transformation.
- Does that mean that basing decisions on belief functions is irrational?

Savage's axioms

- Savage has proposed 7 axioms, 4 of which are considered as meaningful (the other three are technical).
- Let us examine the first two axioms.
- Axiom 1: $\succcurlyeq$ is a total preorder (complete, reflexive and transitive).
- Axiom 2 [Sure Thing Principle]. Given $f, h \in \mathcal{F}$ and $E \subseteq \Omega$, let fEh denote the act defined by

$$(fEh)(\omega) = \begin{cases} f(\omega) & \text{if } \omega \in E \\ h(\omega) & \text{if } \omega \notin E. \end{cases}$$

Then the Sure Thing Principle states that $\forall E, \forall f, g, h, h'$,

$$fEh \succcurlyeq gEh \Rightarrow fEh' \succcurlyeq gEh'.$$

- This axiom seems reasonable, but it is not verified empirically!

Ellsberg's paradox

- Suppose you have an urn containing 30 red balls and 60 balls, either black or yellow. Consider the following gambles:
  - f_1: You receive 100 euros if you draw a red ball;
  - f_2: You receive 100 euros if you draw a black ball.
  - f_3: You receive 100 euros if you draw a red or yellow ball;
  - f_4: You receive 100 euros if you draw a black or yellow ball.
- Most people strictly prefer f_1 to f_2, but they strictly prefer f_4 to f_3.

	R	B	Y
f_1	100	0	0
f_2	0	100	0
f_3	100	0	100
f_4	0	100	100

Now,

$$f_1 = f_1\{R,B\}0, \; f_2 = f_2\{R,B\}0$$

$$f_3 = f_1\{R,B\}100, \; f_4 = f_2\{R,B\}100.$$

- The Sure Thing Principle is violated!

Gilboa's theorem

- Gilboa (1987) proposed a modification of Savage's axioms with, in particular, a weaker form of Axiom 2.
- A preference relation $\succcurlyeq$ meets these weaker requirements iff there exists a (non necessarily additive) measure μ and a utility function $u : \mathcal{X} \rightarrow \mathbb{R}$ such that

$$\forall f, g \in \mathcal{F}, \quad f \succcurlyeq g \Leftrightarrow C_\mu(u \circ f) \geq C_\mu(u \circ g),$$

where C_μ is the Choquet integral, defined for $X : \Omega \rightarrow \mathbb{R}$ as

$$C_\mu(X) = \int_0^{+\infty} \mu(X > t)dt + \int_{-\infty}^{0} [\mu(X > t) - 1]dt.$$

- Given a belief function Bel on Ω and a utility function u, this theorem supports making decisions based on the Choquet integral of u with respect to Bel or Pl.

Lower and upper expected utilities

- For finite Ω, it can be shown that

$$C_{Bel}(u \circ f) = \sum_{B \subseteq \Omega} m(B) \min_{\omega \in B} u(f(\omega))$$

$$C_{Pl}(u \circ f) = \sum_{B \subseteq \Omega} m(B) \max_{\omega \in B} u(f(\omega)).$$

- Let $\mathcal{P}(Bel)$ be the set of probability measures P compatible with Bel, i.e., such that $Bel \leq P$. Then, it can be shown that

$$C_{Bel}(u \circ f) = \min_{P \in \mathcal{P}(Bel)} \mathbb{E}_P(u \circ f) = \underline{\mathbb{E}}(u \circ f)$$

$$C_{Pl}(u \circ f) = \max_{P \in \mathcal{P}(Bel)} \mathbb{E}_P(u \circ f) = \overline{\mathbb{E}}(u \circ f).$$

Decision making
Strategies

- For each act f we have two expected utilities $\underline{\mathbb{E}}(f)$ and $\overline{\mathbb{E}}(f)$. How to make a decision?
- Possible decision criteria:
  1. $f \succcurlyeq g$ iff $\underline{\mathbb{E}}(u \circ f) \geq \overline{\mathbb{E}}(u \circ g)$ (conservative strategy);
  2. $f \succcurlyeq g$ iff $\underline{\mathbb{E}}(u \circ f) \geq \underline{\mathbb{E}}(u \circ g)$ (pessimistic strategy);
  3. $f \succcurlyeq g$ iff $\overline{\mathbb{E}}(u \circ f) \geq \overline{\mathbb{E}}(u \circ g)$ (optimistic strategy);
  4. $f \succcurlyeq g$ iff
  $$\alpha\underline{\mathbb{E}}(u \circ f) + (1-\alpha)\overline{\mathbb{E}}(u \circ f) \geq \alpha\underline{\mathbb{E}}(u \circ g) + (1-\alpha)\overline{\mathbb{E}}(u \circ g)$$
  for some $\alpha \in [0,1]$ called a pessimism index (Hurwicz criterion).
- The conservative strategy yields only a partial preorder: f and g are not comparable if $\underline{\mathbb{E}}(u \circ f) < \overline{\mathbb{E}}(u \circ g)$ and $\underline{\mathbb{E}}(u \circ g) < \overline{\mathbb{E}}(u \circ f)$.

Ellsberg's paradox revisited

We have $m(\{R\}) = 1/3$, $m(\{B, Y\}) = 2/3$.

	R	B	Y	$\underline{\mathbb{E}}(u \circ f)$	$\overline{\mathbb{E}}(u \circ f)$
f_1	100	0	0	u(100)/3	u(100)/3
f_2	0	100	0	u(0)	u(200)/3
f_3	100	0	100	u(100)/3	u(100)
f_4	0	100	100	u(200)/3	u(200)/3

The observed behavior ($f_1 \succcurlyeq f_2$ and $f_4 \succcurlyeq f_3$) is explained by the pessimistic strategy.

Decision making
Special case

- Let $\Omega = \{\omega_1, \ldots, \omega_K\}$, $\mathcal{X} = \{correct, error\}$ and $\mathcal{F} = \{f_1, \ldots, f_K\}$, where
$$f_k(\omega_\ell) = \begin{cases} correct & \text{if } k = \ell \\ error & \text{if } k \neq \ell \end{cases}$$
and $u(correct) = 1$, $u(error) = 0$.
- Then $\underline{\mathbb{E}}(u \circ f_k) = Bel(\{\omega_k\})$ and $\overline{\mathbb{E}}(u \circ f_k) = pl(\omega_k)$.
- The optimistic (resp., pessimistic) strategy selects the hypothesis with the largest plausibility (resp., belief).
- Practical advantage of the maximum plausibility rule: if $m_{12} = m_1 \oplus m_2$, then
$$pl_{12}(\omega) \propto pl_1(\omega) pl_2(\omega), \forall \omega \in \Omega.$$
When combining several mass functions, we do not need to compute the complete mass function to make a decision.

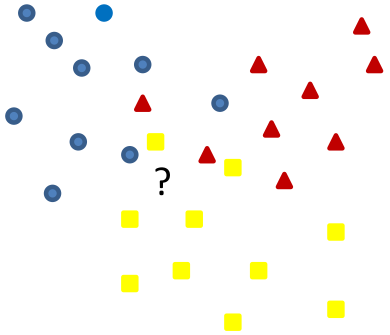
Outline

General Methodology

1. Define the frame of discernment Ω (may be a product space $\Omega = \Omega_1 \times \ldots \times \Omega_n$).
2. Break down the available evidence into independent pieces and model each one by a mass function m on Ω.
3. Combine the mass functions using Dempster's rule.
4. Marginalize the combined mass function on the frame of interest and, if necessary, find the elementary hypothesis with the largest plausibility.

Outline

Problem statement



- A population is assumed to be partitioned in c groups or classes.
- Let $\Omega = \{\omega_1, \ldots, \omega_c\}$ denote the set of classes.
- Each instance is described by
  - A feature vector $\mathbf{x} \in \mathbb{R}^p$;
  - A class label $y \in \Omega$.
- Problem: given a learning set $\mathcal{L} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, predict the class of a new instance described by $\mathbf{x}$.

What can we expect from belief functions?

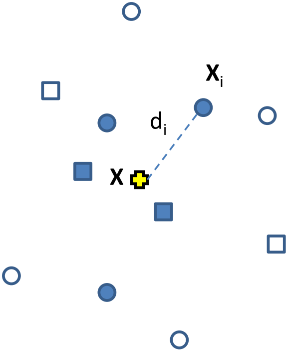
- Problems with "weak" information:
  - Non exhaustive learning sets;
  - Learning and test data drawn from different distributions;
  - Partially labeled data (imperfect class information for training data), etc.
- Information fusion: combination of classifiers trained using different data sets or different learning algorithms (ensemble methods).

Main belief function approaches

1. Approach 1: Convert the outputs from standard classifiers into belief functions and combine them using Dempster's rule or any other alternative rule (e.g., Quost al., *IJAR*, 2011);
2. Approach 2: Develop evidence-theoretic classifiers directly providing belief functions as outputs:
   - Generalized Bayes theorem, extends the Bayesian classifier when class densities and priors are ill-known (Appriou, 1991; Denœux and Smets, *IEEE SMC*, 2008);
   - Distance-based approach: evidential k-NN rule (Denœux, *IEEE SMC*, 1995), evidential neural network classifier (Denœux, *IEEE SMC*, 2000).

Evidential k-NN rule

Principle



- Let Ω be the set of classes.
- Let $\mathcal{N}_k(\mathbf{x}) \subset \mathcal{L}$ denote the set of the k nearest neighbors of $\mathbf{x}$ in $\mathcal{L}$, based on some distance measure.
- Each $\mathbf{x}_i \in \mathcal{N}_k(\mathbf{x})$ can be considered as a piece of evidence regarding the class of $\mathbf{x}$.
- The strength of this evidence decreases with the distance d_i between $\mathbf{x}$ and $\mathbf{x}_i$.

Evidential k-NN rule
Formalization

- The evidence of $(\mathbf{x}_i, y_i)$ with $\mathbf{x}_i \in \mathcal{N}_k(\mathbf{x})$ can be represented by a mass function m_i on Ω:

$$m_i(\{y_i\}) = \varphi(d_i)$$

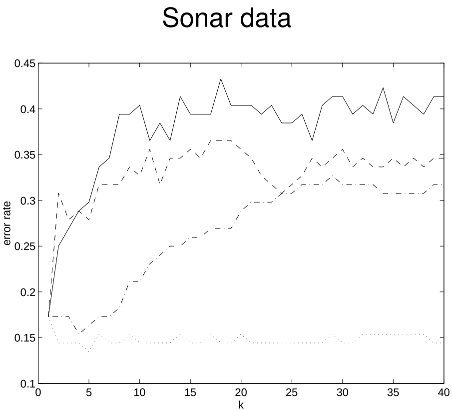
$$m_i(\Omega) = 1 - \varphi(d_i),$$

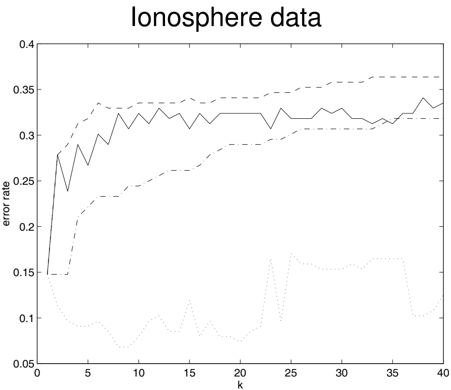
where φ is a decreasing function such that $\lim_{d \to +\infty} \varphi(d) = 0$.
- Pooling of evidence:

$$m = \bigoplus_{\mathbf{x}_i \in \mathcal{N}_k(\mathbf{x})} m_i.$$

- Function φ can be fixed heuristically or selected among a family $\{\varphi_\theta | \theta \in \Theta\}$ using, e.g., cross-validation.
- Decision: select the class with the highest plausibility.

Performance comparison (UCI database)





Test error rates as a function of *k* for the voting (-), evidential (:), fuzzy (–) and distance-weighted (-.) *k*-NN rules.

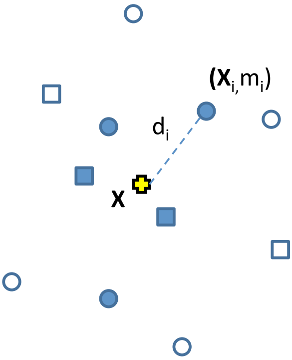
Partially supervised data

- In some applications, learning instances are labeled by experts or indirect methods (no ground truth). Class labels of learning data are then uncertain: partially supervised learning problem.
- Formalization of the learning set:

$$\mathcal{L} = \{(\mathbf{x}_i, m_i), i = 1, \ldots, n\}$$

  where
  - $\mathbf{x}_i$ is the attribute vector for instance i, and
  - m_i is a mass function representing uncertain expert knowledge about the class y_i of instance i.
- Special cases:
  - $m_i(\{\omega_k\}) = 1$ for all i: supervised learning;
  - $m_i(\Omega) = 1$ for all i: unsupervised learning;
- The evidential k-NN rule can easily be adapted to handle such uncertain learning data.

Evidential *k*-NN rule for partially supervised data



- Each mass function m_i is discounted with a rate depending on the distance d_i:

$$m'_i(A) = \varphi(d_i)\, m_i(A), \quad \forall A \subset \Omega.$$

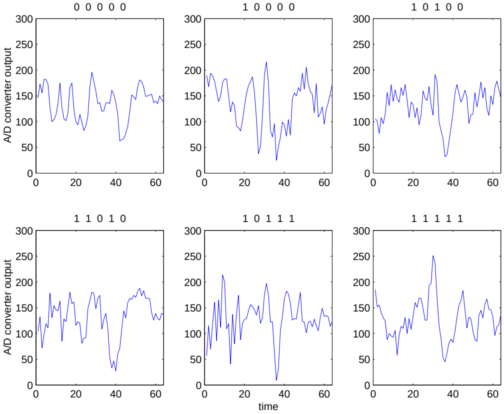
$$m'_i(\Omega) = 1 - \sum_{A \subset \Omega} m'_i(A).$$

- The k mass functions m'_i are combined using Dempster's rule:

$$m = \bigoplus_{\mathbf{x}_i \in \mathcal{N}_k(\mathbf{x})} m'_i.$$

Example: EEG data

EEG signals encoded as 64-D patterns, 50 % positive (K-complexes), 50 % negative (delta waves), 5 experts.

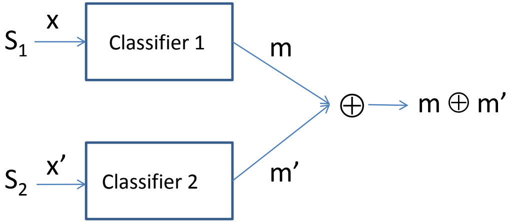


Results on EEG data
(Denoeux and Zouhal, 2001)

- $c = 2$ classes, $p = 64$
- For each learning instance $\mathbf{x}_i$, the expert opinions were modeled as a mass function m_i.
- $n = 200$ learning patterns, 300 test patterns

k	k-NN	w k-NN	Ev. k-NN (crisp labels)	Ev. k-NN (uncert. labels)
9	0.30	0.30	0.31	0.27
11	0.29	0.30	0.29	0.26
13	0.31	0.30	0.31	0.26

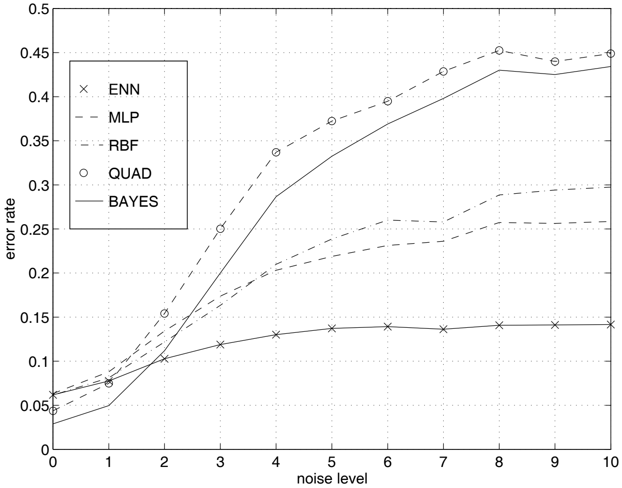
Data fusion example



- $c = 2$ classes
- Learning set ($n = 60$): $\mathbf{x} \in \mathbb{R}^5, \mathbf{x}' \in \mathbb{R}^3$, Gaussian distributions, conditionally independent
- Test set (real operating conditions): $\mathbf{x} \leftarrow \mathbf{x} + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$.

Results

Test error rates: $\mathbf{x} + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$



Summary on classification and ML

- The theory of belief functions has great potential to help solve complex machine learning (ML) problems, particularly those involving:
  - Weak information (partially labeled data, unreliable sensor data, etc.);
  - Multiple sources of information (classifier or clustering ensembles) (Quost et al., 2007; Masson and Denoeux, 2011).
- Other ML applications:
  - Regression (Petit-Renaud and Denoeux, 2004);
  - Multi-label classification (Denoeux et al. 2010);
  - Clustering (Denoeux and Masson, 2004; Masson and Denoeux 2008; Antoine et al., 2012).

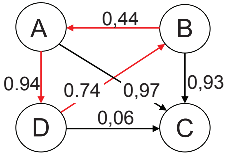
Outline

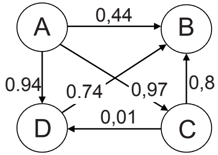
Problem

- We consider a set of alternatives $O = \{o_1, o_2, ..., o_n\}$ and an unknown linear order (transitive, antisymmetric and complete relation) on O.
- Typically, this linear order corresponds to preferences held by an agent or a group of agents, so that $o_i \succ o_j$ is interpreted as "alternative o_i is preferred to alternative o_j".
- A source of information (elicitation procedure, classifier) provides us with $n(n-1)/2$ paired comparisons, with some uncertainty.
- Problem: derive the most plausible linear order from this uncertain (and possibly conflicting) information.

Example (Tritchler & Lockwood, 1991)

- Four scenarios $O = \{A, B, C, D\}$ describing ethical dilemmas in health care.
- Two experts gave their preference for all six possible scenario pairs with confidence degrees.





- What can we say about the preferences of each expert?
- Assuming the existence of a unique consensus linear ordering L^* and seeing the expert assessments as sources of information, what can we say about L^*?

Pairwise mass functions

- The frame of discernment is the set $\mathcal{L}$ of linear orders over O.
- Comparing each pair of objects (o_i, o_j) yields a pairwise mass function $m^{\Theta_{ij}}$ on a coarsening $\Theta_{ij} = \{o_i \succ o_j, o_j \succ o_i\}$ with the following form:

$$
\begin{aligned}
m^{\Theta_{ij}}(o_i \succ o_j) &= \alpha_{ij},\\
m^{\Theta_{ij}}(o_j \succ o_i) &= \beta_{ij},\\
m^{\Theta_{ij}}(\Theta_{ij}) &= 1 - \alpha_{ij} - \beta_{ij}.
\end{aligned}
$$

- $m^{\Theta_{ij}}$ may come from a single expert (e.g., an evidential classifier) or from the combination of the evaluations of several experts.

Combined mass function

- Each of the $n(n-1)/2$ pairwise comparison yields a mass function $m^{\Theta_{ij}}$ on a coarsening Θ_{ij} of $\mathcal{L}$.
- Let $\mathcal{L}_{ij} = \{L \in \mathcal{L} | (o_i, o_j) \in L\}$. Vacuously extending $m^{\Theta_{ij}}$ in $\mathcal{L}$ yields

$$
\begin{aligned}
m^{\Theta_{ij}\uparrow\mathcal{L}}(\mathcal{L}_{ij}) &= \alpha_{ij}, \\
m^{\Theta_{ij}\uparrow\mathcal{L}}(\overline{\mathcal{L}_{ij}}) &= \beta_{ij}, \\
m^{\Theta_{ij}\uparrow\mathcal{L}}(\mathcal{L}) &= 1 - \alpha_{ij} - \beta_{ij}.
\end{aligned}
$$

- Combining the pairwise mass functions using Dempster's rule yields:

$$
m^{\mathcal{L}} = \bigoplus_{i<j} m^{\Theta_{ij}\uparrow\mathcal{L}}.
$$

Plausibility of a linear order

- We have $m^{\Theta_{ij}\uparrow\mathcal{L}}(\mathcal{L}_{ij}) = \alpha_{ij}$, $m^{\Theta_{ij}\uparrow\mathcal{L}}(\overline{\mathcal{L}_{ij}}) = \beta_{ij}$, $m^{\Theta_{ij}\uparrow\mathcal{L}}(\mathcal{L}) = 1 - \alpha_{ij} - \beta_{ij}$.
- Let pl_{ij} be the corresponding contour function:

$$pl_{ij}(L) = \begin{cases} 1 - \beta_{ij} & \text{if } (o_i, o_j) \in L, \\ 1 - \alpha_{ij} & \text{if } (o_i, o_j) \not\in L. \end{cases}$$

- After combining the $m^{\Theta_{ij}\uparrow\mathcal{L}}$ for all $i < j$ we get:

$$pl(L) = \frac{1}{1-K} \prod_{i<j} (1 - \beta_{ij})^{\ell_{ij}} (1 - \alpha_{ij})^{1-\ell_{ij}},$$

where $\ell_{ij} = 1$ if $(o_i, o_j) \in L$ and 0 otherwise.
- An algorithm for computing the degree of conflict K has been given by Tritchler & Lockwood (1991).

Finding the most plausible linear order

- We have

$$\ln pl(L) = \sum_{i<j} \ell_{ij} \ln \left(\frac{1-\beta_{ij}}{1-\alpha_{ij}} \right) + c$$

- $pl(L)$ can thus be maximized by solving the following binary integer programming problem:

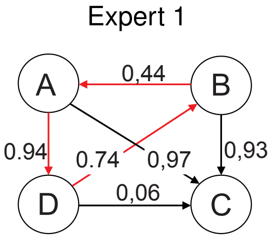
$$\max_{\ell_{ij}\in\{0,1\}} \sum_{i<j} \ell_{ij} \ln \left(\frac{1-\beta_{ij}}{1-\alpha_{ij}} \right),$$

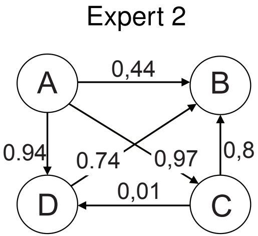
subject to:

$$\begin{cases} \ell_{ij} + \ell_{jk} - 1 \leq \ell_{ik}, & \forall i<j<k, \quad (1) \\ \ell_{ik} \leq l_{ij} + \ell_{jk}, & \forall i<j<k. \quad (2) \end{cases}$$

- Constraint (1) ensures that $\ell_{ij}=1$ and $\ell_{jk}=1 \Rightarrow \ell_{jk}=1$, and (2) ensures that $\ell_{ij}=0$ and $\ell_{jk}=0 \Rightarrow \ell_{ik}=0$.

Example





$$L_1^* = A \succ D \succ B \succ C$$

$$pl(L_1^*) = 0.807$$

$$L_2^* = A \succ C \succ D \succ B$$

$$pl(L_2^*) = 1$$

Example: combination of expert evaluations

- Dempster's rule of combination:

(o_i, o_j)	$o_i \succ o_j$	$o_j \succ o_i$	Θ_{ij}
(A,B)	0.3056	0.3056	0.3889
(A,C)	0.9991	0	0.0009
(A,D)	0.9964	0	0.0036
(B,C)	0.7266	0.2187	0.0547
(B,D)	0	0.9324	0.0676
(C,D)	0.0594	0.0094	0.9312

- $L^* = A \succ D \succ B \succ C$ and $pl(L^*) = 0.8893$.
- We get the same linear order as the one given by Expert 1.

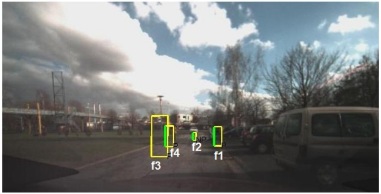
Summary on preference aggregation

- The framework of belief functions allows us to model uncertainty in paired comparisons.
- The most plausible linear order can be computed efficiently using a binary linear programming approach.
- The approach has been applied to label ranking, in which the task is to learn a "ranker" that maps p-dimensional feature vectors x describing an agent to a linear order over a finite set of alternatives, describing the agent's preferences (Denœux and Masson, BELIEF 2012).
- The method can easily be extended to the elicitation of preference relations with indifference and/or incomparability between alternatives (Denœux and Masson. *Annals of Operations Research* 195(1):135-161, 2012).

Outline

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 a sensor at two different times.
- Problem: given information 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 $m^{\Theta_{ij}}$ on the frame $\Theta_{ij} = \{h_{ij}, \overline{h}_{ij}\}$, where
   - h_{ij} denotes the hypothesis that e_i and f_j are the same objects, and
   - $\overline{h}_{ij}$ is the hypothesis that e_i and f_j are different objects.
2. Vacuously extend the np mass functions $m^{\Theta_{ij}}$ in the frame $\mathcal{R}$ containing all admissible matching relations.
3. Combine the np extended mass functions $m^{\Theta_{ij}\uparrow\mathcal{R}}$ 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 h_{ij}, while a large value of d_{ij} supports hypothesis $\overline{h}_{ij}$. Depending on sensor reliability, a fraction of the unit mass should also be assigned to $\Theta_{ij} = \{h_{ij}, \overline{h}_{ij}\}$.
- This line of reasoning justifies a mass function $m_p^{\Theta_{ij}}$ of the form:

$$
\begin{aligned}
m_p^{\Theta_{ij}}(\{h_{ij}\}) &= \alpha\varphi(d_{ij}) \\
m_p^{\Theta_{ij}}(\{\overline{h}_{ij}\}) &= \alpha\left(1 - \varphi(d_{ij})\right) \\
m_p^{\Theta_{ij}}(\Theta_{ij}) &= \alpha,
\end{aligned}
$$

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 $\overline{h}_{ij}$, whereas a small value of d'_{ij} does not support specifically h_{ij} or $\overline{h}_{ij}$, as two distinct objects may have similar velocities.
- Consequently, the following form of the mass function $m_v^{\Theta_{ij}}$ induce by d'_{ij} seems appropriate:

$$m_v^{\Theta_{ij}}(\{\overline{h}_{ij}\}) = \alpha' \left(1 - \psi(d'_{ij})\right) \quad (1a)$$

$$m_v^{\Theta_{ij}}(\Theta_{ij}) = 1 - \alpha' \left(1 - \psi(d'_{ij})\right), \quad (1b)$$

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 $m_c^{\Theta_{ij}}$ on Θ_{ij} derived from m_i and m_j has the following expression:

$$
\begin{aligned}
m_c^{\Theta_{ij}}(\{\overline{h}_{ij}\}) &= \kappa_{ij} \\
m_c^{\Theta_{ij}}(\Theta_{ij}) &= 1 - \kappa_{ij},
\end{aligned}
$$

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:
$$m^{\Theta_{ij}} = m_p^{\Theta_{ij}} \oplus m_v^{\Theta_{ij}} \oplus m_c^{\Theta_{ij}}.$$
- 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 $m^{\Theta_{ij}}$ in $\mathcal{R}$ yields the following mass function:

$$\begin{aligned}
m^{\Theta_{ij}\uparrow\mathcal{R}}(\mathcal{R}_{ij}) &= m^{\Theta_{ij}}(\{h_{ij}\}) = \alpha_{ij} \\
m^{\Theta_{ij}\uparrow\mathcal{R}}(\overline{\mathcal{R}_{ij}}) &= m^{\Theta_{ij}}(\{\overline{h}_{ij}\}) = \beta_{ij} \\
m^{\Theta_{ij}\uparrow\mathcal{R}}(\mathcal{R}) &= m^{\Theta_{ij}}(\Theta_{ij}) = 1 - \alpha_{ij} - \beta_{ij}.
\end{aligned}$$

Combining pairwise mass functions

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

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

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

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

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 using, e.g., the Hungarian algorithm in $O(\max(n,p)^3)$ time.

Conclusion

- In this problem as well as in the previous one, the frame of discernment can be huge (e.g., $n!$ in the preference aggregation problem).
- Yet, the belief function approach is manageable because:
  - The elementary pieces of evidence that are combined have a simple form (this is almost always the case);
  - We are only interested in the most plausible alternative: hence, we do not have to compute the full combined belief function.
- Other problems with very large frame for which belief functions have been successfully applied:
  - Clustering: Ω is the space of all partitions (Masson and Denoeux, 2011) ;
  - Multi-label classification: Ω is the powerset of the set of classes (Denoeux et al., 2010).

Outline

The problem

- We consider a statistical model $\{f(x,\theta), x \in \mathbb{X}, \theta \in \Theta\}$, where $\mathbb{X}$ is the sample space and Θ the parameter space.
- Having observed x, how to quantify the uncertainty about Θ, without specifying a prior probability distribution?
- Two main approaches using belief functions:
  1. Dempster's approach based on an auxiliary variable with a pivotal probability distribution (Dempster, 1967);
  2. Likelihood-based approach (Shafer, 1976, Wasserman, 1990).

Outline

Sampling model

- Suppose that the sampling model $X \sim f(x; \theta)$ can be represented by an "a-equation" of the form

$$X = a(\theta, U),$$

where $U \in \mathbb{U}$ is an (unobserved) auxiliary variable with known probability distribution μ independent of θ.

- This representation is quite natural in the context of data generation.
- For instance, to generate a continuous random variable X with cumulative distribution function (cdf) F_θ, one might draw U from $\mathcal{U}([0, 1])$ and set $X = F_\theta^{-1}(U)$.

From a-equation to belief function

- The equation $X = a(\theta, U)$ defines a multi-valued mapping

$$\Gamma : U \rightarrow \Gamma(U) = \{(X, \theta) \in \mathbb{X} \times \Theta | X = a(\theta, U)\}.$$

- Under measurability conditions, the probability space $(\mathbb{U}, \mathcal{B}(\mathbb{U}), \mu)$ and the multi-valued mapping Γ induce a belief function $Bel_{\Theta \times \mathbb{X}}$ on $\mathbb{X} \times \Theta$.
- Conditioning $Bel_{\Theta \times \mathbb{X}}$ on θ yields the sampling distribution $f(\cdot; \theta)$ on $\mathbb{X}$;
- Conditioning it on $X = x$ gives a belief function $Bel_{\Theta}(\cdot; x)$ on Θ.

Example: Bernoulli sample

- Let $X = (X_1, \ldots, X_n)$ consist of independent Bernoulli observations and $\theta \in \Theta = [0, 1]$ is the probability of success.
- Sampling model:

$$X_i = \begin{cases} 1 & \text{if } U_i \leq \theta \\ 0 & \text{otherwise,} \end{cases}$$

  where $U = (U_1, \ldots, U_n)$ has pivotal measure $\mu = \mathcal{U}([0, 1]^n)$.
- Having observed the number of successes $y = \sum_{i=1}^{n} x_i$, the belief function $Bel_\Theta(\cdot; x)$ is induced by a random closed interval

$$[U_{(y)}, U_{(y+1)}],$$

  where $U_{(i)}$ denotes the i-th order statistics from $U_1, \ldots, U_n$.
- Quantities like $Bel_\Theta([a, b]; x)$ or $Pl_\Theta([a, b]; x)$ are readily calculated.

Discussion

- Dempster's model has several nice features:
  - It allows us to quantify the uncertainty on Θ after observing the data, without having to specify a prior distribution on Θ;
  - When a Bayesian prior P_0 is available, combining it with $Bel_\Theta(\cdot, x)$ using Dempster's rule yields the Bayesian posterior:

$$Bel_\Theta(\cdot, x) \oplus P_0 = P(\cdot|x).$$

- Drawbacks:
  - It often leads to cumbersome or even intractable calculations except for very simple models, which imposes the use of Monte-Carlo simulations.
  - More fundamentally, the analysis depends on the a-equation $X = a(\theta, U)$ and the auxiliary variable U, which are not unique for a given statistical model $\{f(\cdot; \theta), \theta \in \Theta\}$. As U is not observed, how can we argue for an a-equation or another?

Outline

1. Basic theory
   - Representation of evidence
   - Operations on Belief functions
   - Decision making
2. Applications
   - Classification
   - Preference aggregation
   - Object association
3. Statistical inference
   - Dempster's approach
   - Likelihood-based approach
   - Sea level rise example

Likelihood-based belief function

Requirements

1. Likelihood principle: $Bel_\Theta(\cdot; x)$ should be based only on the likelihood function $L(\theta; x) = f(x; \theta)$.
2. Compatibility with Bayesian inference: when a Bayesian prior P_0 is available, combining it with $Bel_\Theta(\cdot, x)$ using Dempster's rule should yield the Bayesian posterior:

$$Bel_\Theta(\cdot, x) \oplus P_0 = P(\cdot|x).$$

3. Principle of minimal commitment: among all the belief functions satisfying the previous two requirements, $Bel_\Theta(\cdot, x)$ should be the least committed (least informative).

Likelihood-based belief function

Solution

- $Bel_\Theta(\cdot; x)$ is the consonant belief function with contour function equal to the normalized likelihood:

$$pl(\theta; x) = \frac{L(\theta; x)}{\sup_{\theta' \in \Theta} L(\theta'; x)},$$

- The corresponding plausibility function is:

$$Pl_\Theta(A; x) = \sup_{\theta \in A} pl(\theta; x) = \frac{\sup_{\theta \in A} L(\theta; x)}{\sup_{\theta \in \Theta} L(\theta; x)}, \quad \forall A \subseteq \Theta.$$

- Corresponding random set: $(\Omega, \mathcal{B}(\Omega), \mu, \Gamma_x)$ with $\Omega = [0, 1]$, $\mu = \mathcal{U}([0, 1])$ and

$$\Gamma_x(\omega) = \{\theta \in \Theta | pl(\theta; x) \geq \omega\}.$$

Discussion

- The likelihood-based method is much simpler to implement than Dempster's method, even for complex models.
- By construction, it boils down to Bayesian inference when a Bayesian prior is available.
- It is compatible with usual likelihood-based inference:
    - Assume that $\theta = (\theta_1, \theta_2) \in \Theta_1 \times \Theta_2$ and θ_2 is a nuisance parameter. The marginal contour function on Θ_1

$$pl(\theta_1; x) = \sup_{\theta_2 \in \Theta_2} pl(\theta_1, \theta_2; x) = \frac{\sup_{\theta_2 \in \Theta_2} L(\theta_1, \theta_2; x)}{\sup_{(\theta_1, \theta_2) \in \Theta} L(\theta_1, \theta_2; x)}$$

    is the relative profile likelihood function.
    - Let $H_0 \subset \Theta$ be a composite hypothesis. Its plausibility

$$Pl(H_0; x) = \frac{\sup_{\theta \in H_0} L(\theta; x)}{\sup_{\theta \in \Theta} L(\theta; x)}.$$

    is the usual likelihood ratio statistics $\Lambda(x)$.

Outline

Climate change

- Climate change is expected to have enormous economic impact, including threats to infrastructure assets through
  - damage or destruction from extreme events;
  - coastal flooding and inundation from sea level rise, etc.
- Adaptation of infrastructure to climate change is a major issue for the next century.
- Engineering design processes and standards are based on analysis of historical climate data (using, e.g. Extreme Value Theory), with the assumption of a stable climate.
- Procedures need to be updated to include expert assessments of changes in climate conditions in the 21th century.

Adaptation of flood defense structures

- Commonly, flood defenses in coastal areas are designed to withstand at least 100 years return period events.
- However, due to climate change, they will be subject during their life time to higher loads than the design estimations.
- The main impact is related to the increase of the mean sea level, which affects the frequency and intensity of surges.
- For adaptation purposes, statistics of extreme sea levels derived from historical data should be combined with projections of the future sea level rise (SLR).

Assumptions

- The annual maximum sea level Z at a given location is often assumed to have a Gumbel distribution

$$P(Z \leq z) = \exp\left[-\exp\left(-\frac{z-\mu}{\sigma}\right)\right]$$

with mode μ and scale parameter σ.

- Current design procedures are based on the return level z_T associated to a return period T, defined as the quantile at level $1 - 1/T$. Here,

$$z_T = \mu - \sigma \log\left[-\log\left(1 - \frac{1}{T}\right)\right]$$

- Because of climate change, it is assumed that the distribution of annual maximum sea level at the end of the century will be shifted to the right, with shift equal to the SLR :

$$z'_T = z_T + SLR.$$

Approach

1. Represent the evidence on z_T by a likelihood-based belief function using past sea level measurements;
2. Represent the evidence on SLR by a belief function describing expert opinions;
3. Combine these two items of evidence to get a belief function on $z'_T = z_T + SLR$.

Statistical evidence on z_T

- Let $z_1, \ldots, z_n$ be n i.i.d. observations of Z. The likelihood function is:

$$L(z_T, \mu; z_1, \ldots, z_n) = \prod_{i=1}^{n} f(z_i; z_T, \mu),$$

where the pdf of Z has been reparametrized as a function of z_T and μ.

- The corresponding contour function is thus:

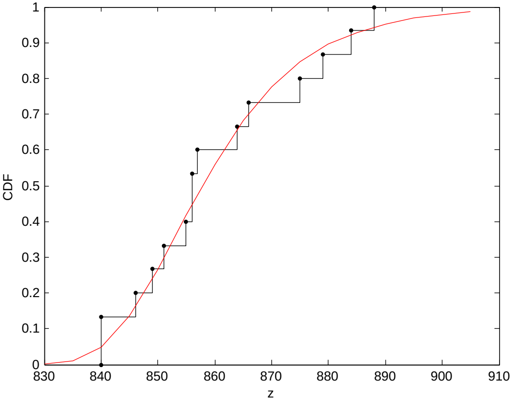
$$pl(z_T, \mu; z_1, \ldots, z_n) = \frac{L(z_T, \mu; z_1, \ldots, z_n)}{\sup_{z_T, \mu} L(z_T, \mu; z_1, \ldots, z_n)}$$

and the marginal contour function of z_T is

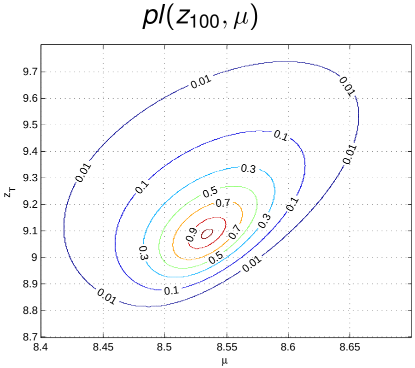
$$pl(z_T; z_1, \ldots, z_n) = \sup_{\mu} pl(z_T, \mu; z_1, \ldots, z_n).$$

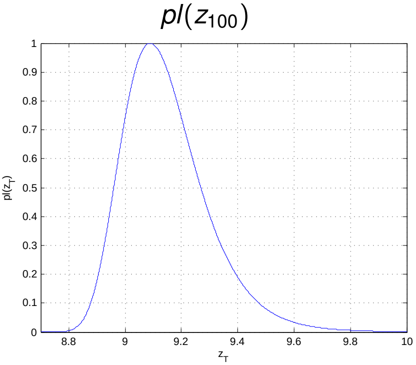
Data

15 years of sea level data at Le Havre, France



Results


$pl(z_{100}, \mu)$
0.01
0.1
0.3
0.5
0.7
0.9
z_T
μ
8.4
8.45
8.5
8.55
8.6
8.65
8.7
8.8
8.9
9
9.1
9.2
9.3
9.4
9.5
9.6
9.7


$pl(z_{100})$
$pl(z_T)$
z_T
8.8
9
9.2
9.4
9.6
9.8
10
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

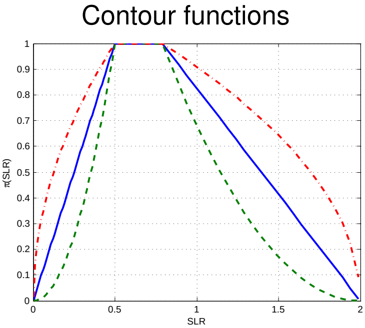
Expert evidence on SLR

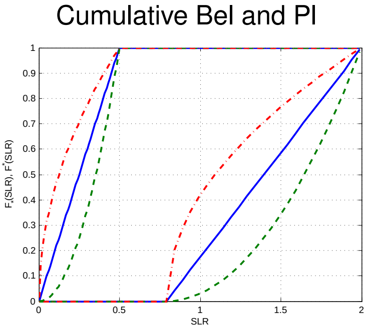
- Future SLR projections provided by the IPCC last Assessment Report (2007) give [0.18 m, 0.79 m] as a likely range of values for SLR over the 1990-2095 period. However, it is indicated that higher values cannot be excluded.
- Other recent SLR assessments based on semi-empirical models have been undertaken. For example, based on a simple statistical model, Rahmstorf (2007) suggests [0.5m, 1.4 m] as a likely range.
- Recent studies indicate that the threshold of 2 m cannot be exceeded by the end of this century due to physical constraints.

Representation of expert evidence

- The interval $[0.5, 0.79] = [0.18, 0.79] \cap [0.5, 1.4]$ seems to be fully supported by the available evidence, as it is considered highly plausible by all three sources, while values outside the interval $[0, 2]$ are considered as impossible.
- Three representations:
  - Consonant random intervals with core $[0.5, 0.79]$, support $[0, 2]$ and different contour functions π;
  - p-boxes with same cumulative belief and plausibility functions as above;
  - Random sets $[U, V]$ with independent U and V and same cumulative belief and plausibility functions as above.

Representation of expert opinions





Combination

Principle

- Let $[U_{z_T}, V_{z_T}]$ and $[U_{SLR}, V_{SLR}]$ be the independent random intervals representing evidence on z_T and SLR, respectively.
- The random interval for $z'_T = z_T + SLR$ is

$$[U_{z_T}, V_{z_T}] + [U_{SLR}, V_{SLR}] = [U_{z_T} + U_{SLR}, V_{z_T} + V_{SLR}]$$

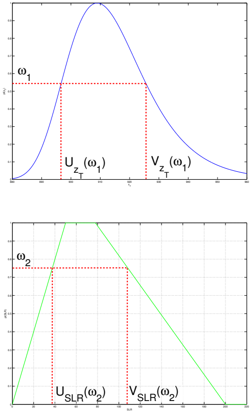
- The corresponding belief and plausibility functions are

$$\begin{aligned} Bel(A) &= P([U_{z_T} + U_{SLR}, V_{z_T} + V_{SLR}] \subseteq A) \\ Pl(A) &= P([U_{z_T} + U_{SLR}, V_{z_T} + V_{SLR}] \cap A \neq \emptyset) \end{aligned}$$

  for all $A \in \mathcal{B}(\mathbb{R})$.
- $Bel(A)$ and $Pl(A)$ can be estimated by Monte Carlo simulation.

Combination

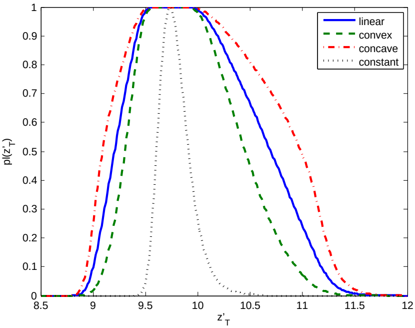
Monte Carlo simulation

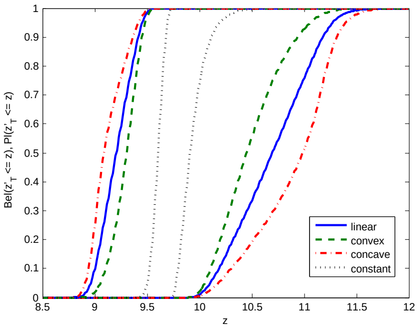


Algorithm to approximate $Pl(A)$:

$k = 0$
for $i = 1 : N$ **do**
  Pick $\omega_1 \sim U(0,1)$, $\omega_2 \sim U(0,1)$
  $I = [U_{z_T}(\omega_1) + U_{SLR}(\omega_2), V_{z_T}(\omega_1) + V_{SLR}(\omega_2)]$
  if $I \cap A \neq \emptyset$ **then**
    $k = k + 1$
  end if
end for
$\widehat{Pl}(A) = \frac{k}{N}$

Result


pl(z'_T)
z'_T
linear
convex
concave
constant


Bel(z'_T <= z), Pl(z'_T <= z)
z
linear
convex
concave
constant

Summary

- The theory of belief functions is a modeling language for representing elementary items of evidence and combining them, in order to form a representation of our beliefs about certain aspects of the world.
- This theory is relatively simple to implement and has been successfully used in a wide range of applications, such as classification and sensor fusion.
- Evidential reasoning can be implemented even in very large spaces, because
  - Elementary items of evidence induce simple belief functions, which can be combined very efficiently;
  - The most plausible hypothesis can be found without computing the whole combined belief function.
- Statistical evidence may be represented by likelihood-based belief functions, generalizing both likelihood-based and Bayesian inference.

Papers and Matlab software available at:

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

THANK YOU!