Methods for building belief functions

Thierry Denœux

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

Fourth School on Belief Functions and their Applications
Xi'An, China, July 5, 2017

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

Definition (Least Commitment Principle)

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) 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

Example: LC mass function with given contour function

Problem statement

- Assume we ask an expert for the plausibility $\pi(\omega)$ of each $\omega \in \Omega$
- We get a function $\pi : \Omega \rightarrow [0, 1]$. 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 $\mathcal{M}(\pi)$?

LC mass function with given contour function

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

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

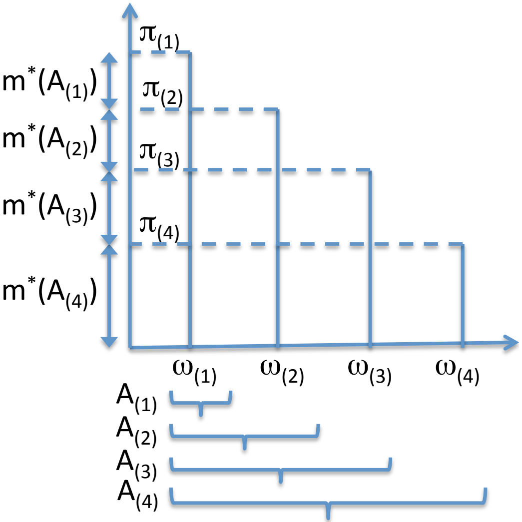
- Let Q^* be defined as $Q^*(\emptyset) = 1$ and

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

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

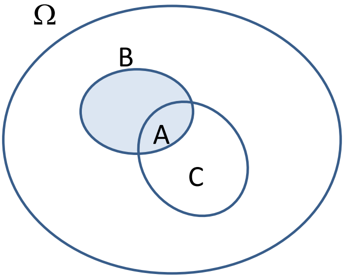
LC mass function with given contour function

Recovering the mass function



Outline

Deconditioning

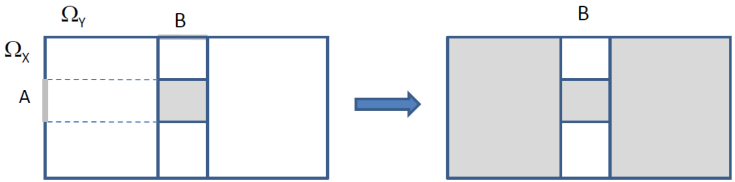


- Let m_0 be a mass function on Ω expressing our beliefs about X in a context where we know that $X \in B$
- We want to build a mass function m verifying the constraint $m(\cdot|B) = m_0$
- Any m built from m_0 by transferring each mass $m_0(A)$ to $A \cup C$ for some $C \subseteq \overline{B}$ satisfies the constraint
- s-least committed solution: transfer $m_0(A)$ to the largest such set, which is $A \cup \overline{B}$

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

Deconditioning

Conditional embedding



- More complex situation: two frames Ω_X and Ω_Y
- Let m_X^0 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 $\left(m_{XY} \oplus m_{Y[B]}\right)^{\downarrow X} = m_X^0$
- s-least committed solution: transfer $m_X^0(A)$ to $(A \times \Omega_Y) \cup (\Omega_X \times \overline{B})$
- Notation $m_{XY} = (m_X^0)_{\Uparrow XY}$ (conditional embedding)

Generalized Bayes Theorem

Problem statement

- Consider, for instance, a classification problem, where $X \in \Omega_X$ is a measurement vector and $Y \in \Omega_Y = \{y_1, \ldots, y_K\}$ is the class variable.
- Partial knowledge of X given each $Y = y_k$

$$m_X(\cdot|y_k), \quad k = 1, \ldots, K$$

- Prior knowledge about Y: m_Y^0 (may be vacuous)
- We observe $X \in A$
- Belief function on Y?

Generalized Bayes Theorem

Solution

- Solution:

$$m_Y(\cdot|A) = \left(\bigoplus_{k=1}^{K} m_X(\cdot|y_k)_{\Uparrow XY} \oplus m_{X[A]} \oplus m_Y^0 \right)_{\downarrow Y}$$

- Expression

$$m_Y(\cdot|A) = \bigoplus_{k=1}^{K} \overline{\{y_k\}}^{Pl_X(A|y_k)} \oplus m_Y^0$$

where $\overline{\{y_k\}}^{Pl_X(A|\theta_k)}$ is the simple mass function that assigns the mass $1 - Pl_X(A|y_k)$ to $\overline{\{y_k\}}$ and $Pl_X(A|y_k)$ to Ω_Y

Generalized Bayes Theorem

Properties

- Property 1: Bayes' theorem is recovered as a special case when the conditional mass functions $m_X(\cdot|y_k)$ and m_Y^0 are Bayesian
- Property 2: If X_1 and X_2 are cognitively independent conditionally on Y, i.e.,

$$pl_{X_1 X_2}(A_1 \times A_2|y_k) = pl_{X_1}(A_1|y_k) \cdot pl_{X_2}(A_2|y_k)$$

for all $A_1 \subseteq \Omega_{X_1}$, $A_2 \subseteq \Omega_{X_2}$ and $y_k \in \Omega_Y$, then

$$m_Y(\cdot|X_1 \in A_1, X_2 \in A_2) = m_Y(\cdot|X_1 \in A_1) \oplus m_Y(\cdot|X_2 \in A_2)$$

Outline

Uncertainty measures

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

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

  with $f = Id$ (expected cardinality), $f(x) = -1/x$ (opposite of Yager's specificity), $f = \log_2$ (nonspecificity)
- 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 minimal 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 have a non linear optimization problem

Combination under unknown dependence (1/2)

- Consider two sources (S_1, P_1, Γ_1) and (S_2, P_2, Γ_2) generating 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

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

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

$$m(A) = \sum_{A_i \cap B_j = A} p^*_{ij},$$

for all $A \subseteq \Omega$, $A \neq \emptyset$, with $p^*_{ij} = p_{ij}/(1-\kappa)$, $\kappa =$ degree of conflict

Combination under unknown dependence (2/2)

- When the dependence between the two sources is unknown, the p_{ij}'s are unknown
- Maximizing the Shannon entropy yields Dempster's rule
- The least specific combined mass function can be found by solving the following linear optimization problem:

$$\max_{p^*_{ij}} \sum_{\{(i,j)|A_i \cap B_j \neq \emptyset\}} p^*_{ij} \log_2 |A_i \cap B_j|$$

under the constraints $\sum_{i,j} p^*_{ij} = 1$ and

$$\sum_i p^*_{ij} = q_j, \quad j = 1, \ldots, s$$

$$\sum_j p^*_{ij} = p_i, \quad i = 1, \ldots, r$$

$$p^*_{ij} = 0 \text{ for all } (i,j) \text{ s.t. } A_i \cap B_j = \emptyset$$

Outline

Most Commitment Principle

- Assume that the constraints imposed on a belief function by a certain problem are of the form

$$Bel(A) \leq f(A), \quad \forall A \subset \Omega,$$

  for some function f.
- The *pl*-least committed belief function verifying these constraints is vacuous: consequently, the LCP is ineffective in that case.
- Instead, it makes sense to select the most committed belief function verifying the constraints, if it exists.
- This principle can be called the Most Commitment Principle.
- Example: construction of a predictive belief function.

Motivation

- Let X be random variable (defined from a repeatable random experiment), with unknown probability $\mathbb{P}_X$.
- We have observed n independent replicates of X:

$$\boldsymbol{X} = (X_1, \ldots, X_n).$$

- Problem: quantify our beliefs regarding a future realization of X using a belief function $Bel(\cdot; \boldsymbol{X})$: predictive belief function.

Examples

1. Example 1:
   - We have drawn r black balls in n draws from an urn with replacement:
   - What is our belief that the next ball to be drawn from the urn will be black?
2. Example 2:
   - The lifetimes of 20 bearings have been observed:

     2398, 2812, 3113, 3212, 3523, 5236, 6215,
     6278, 7725, 8604, 9003, 9350, 9460, 11584,
     11825, 12628, 12888, 13431, 14266, 17809.
   - Let X be the lifetime of a bearing taken at random from the same population. Belief function on X?

Approach

- If we knew the conditional distribution $\mathbb{P}_X$, it would be natural to equate our degrees of belief $Bel_X(A|\boldsymbol{x})$ with degrees of chance $\mathbb{P}_X(A)$ for any event A, i.e., we would impose

$$Bel_X(\cdot|\boldsymbol{x}) = \mathbb{P}_X.$$

- In real situations, however, we only have limited information about $\mathbb{P}_X$ in the form of the observed data $\boldsymbol{x}$. Our predictive belief function should thus be less committed than $\mathbb{P}_X$, which can be expressed by the following inequalities

$$Bel_X(A|\boldsymbol{x}) \leq \mathbb{P}_X(A) \tag{1}$$

for all $A \subseteq \mathcal{X}$

Approach (continued)

- However, after observing $\boldsymbol{x}$, each probability $\mathbb{P}_X(A)$ can still be arbitrarily small.
- Consequently, the condition $Bel_X(\cdot|\boldsymbol{x}) \leq \mathbb{P}_X$ can only be guaranteed for the vacuous belief function, such that $Bel_X(A|\boldsymbol{x}) = 0$ for all $A \subset \mathcal{X}$.
- Solution: weaken condition (1) by imposing only that it hold for at least a proportion $1-\alpha \in (0,1)$ of the samples $\boldsymbol{x}$, under repeated sampling. We then have the following requirement,

$$\mathbb{P}_{\boldsymbol{X}}\left\{Bel_X(\cdot|\boldsymbol{X}) \leq \mathbb{P}_X\right\} \geq 1-\alpha, \tag{2}$$

for all $\theta \in \Theta$.

- A belief function verifying (2) is called a predictive belief function at confidence level $1-\alpha$. It is an approximate $1-\alpha$-level predictive belief function if Property (2) holds only in the limit as the sample size tends to infinity.

Meaning of Property (2)

$$\boldsymbol{x} = (x_1, \ldots, x_n) \rightarrow Bel(\cdot|\boldsymbol{x})$$
$$\boldsymbol{x}' = (x'_1, \ldots, x'_n) \rightarrow Bel(\cdot|\boldsymbol{x}')$$
$$\boldsymbol{x}'' = (x''_1, \ldots, x''_n) \rightarrow Bel(\cdot|\boldsymbol{x}'')$$
$$\vdots$$

- As the number of realizations of the random sample tends to ∞, the proportion of belief functions less committed than $\mathbb{P}_X$ should tend to $1-\alpha$.
- To achieve this property, we use
  - multinomial confidence regions in the discrete case
  - confidence bands in the continuous case

Outline

Multinomial Confidence Region

- Discrete random variable $X \in \mathcal{X} = \{\xi_1, \ldots, \xi_K\}$.
- Let $p_k = \mathbb{P}_X(\{\xi_k\})$ and $\boldsymbol{p} = (p_1, \ldots, p_K)$
- Let $\mathcal{R}(\boldsymbol{X}) \subseteq [0,1]^K$ be a random region of $[0,1]^K$. It is a confidence region for $\boldsymbol{p}$ at level $1-\alpha$ if

$$\mathbb{P}_{\boldsymbol{X}}\left\{\mathcal{R}(\boldsymbol{X}) \ni \boldsymbol{p}\right\} \geq 1-\alpha.$$

- $\mathcal{R}(\boldsymbol{X})$ is an asymptotic confidence region if the above inequality holds in the limit as $n \to \infty$.
- We consider a special kind of confidence regions: simultaneous confidence intervals:

$$\mathcal{R}(\boldsymbol{X}) = [P_1^-, P_1^+] \times \ldots \times [P_K^-, P_K^+]$$

Goodman's simultaneous confidence intervals

Goodman's simultaneous confidence intervals:

$$P_k^- = \frac{b + 2N_k - \sqrt{\Delta_k}}{2(n+b)},$$

$$P_k^+ = \frac{b + 2N_k + \sqrt{\Delta_k}}{2(n+b)},$$

with $N_k = \#\{i | X_i = \xi_k\}$, $b = \chi^2_{1;1-\alpha/K}$ and $\Delta_k = b\left(b + \frac{4N_k(n-N_k)}{n}\right)$.

Example

- 220 psychiatric patients from some population, categorized as either neurotic, depressed, schizophrenic or having a personality disorder.
- Observed counts: $91, 49, 37, 43$.
- Goodman' confidence intervals at confidence level $1-\alpha=0.95$:

Diagnosis	n_k/n	P_k^-	P_k^+
Neurotic	0.41	0.33	0.50
Depressed	0.22	0.16	0.30
Schizophrenic	0.17	0.11	0.24
Personality disorder	0.20	0.14	0.27

From Confidence Regions to Lower Probabilities

- To each $\boldsymbol{p} = (p_1, \ldots, p_K)$ corresponds a probability measure $\mathbb{P}_X$.
- Consequently, $\mathcal{R}(\boldsymbol{X})$ may be seen as defining a family of probability measures, uniquely defined by the following lower probability measure:

$$P^-(A) = \min_{\boldsymbol{p} \in \mathcal{R}(\boldsymbol{X})} \sum_{\xi_k \in A} p_k = \max \left(\sum_{\xi_k \in A} P_k^-, 1 - \sum_{\xi_k \notin A} P_k^+ \right)$$

- P^- is verifies the following property,

$$\mathbb{P}_{\boldsymbol{X}} \left\{ P^- \leq \mathbb{P}_X \right\} \geq 1 - \alpha.$$

- P^- is 2-monotone, i.e., we have

$$P^-(A \cup B) \geq P^-(A) + P^-(B) - P^-(A \cap B), \quad \forall A, B \subseteq \mathcal{X}.$$

- However, it is not always completely monotone!

From Lower Probabilities to Belief Functions

Cases $K = 2$ and $K = 3$

- If $K = 2$ or $K = 3$, P^- is a belief function.
- Case $K = 2$:

$$m(\{\xi_1\}) = P_1^-, \quad m(\{\xi_2\}) = P_2^-, \quad m(\mathcal{X}) = 1 - P_1^- - P_2^-.$$

- Case $K = 3$:

$$\begin{aligned} m(\{\xi_k\}) &= P_k^-, \quad k = 1, 2, 3 \\ m(\{\xi_1, \xi_2\}) &= 1 - P_3^+ - P_1^- - P_2^- \\ m(\{\xi_1, \xi_3\}) &= 1 - P_2^+ - P_1^- - P_3^- \\ m(\{\xi_2, \xi_3\}) &= 1 - P_1^+ - P_2^- - P_3^- \\ m(\mathcal{X}) &= \sum_{k=1}^{3} (P_k^+ + P_k^-) - 2 \end{aligned}$$

From Lower Probabilities to Belief Functions

Case $K > 3$

- When $K > 3$, P^- is no longer guaranteed to be a belief function. We thus have to approximate P^- by a belief function.
- Let $\mathcal{B}(P^-)$ denote the set of belief functions *Bel* on $\mathcal{X}$ verifying $Bel \leq P^-$. We have, for any $Bel \in \mathcal{B}^{\mathcal{X}}(P^-)$:

$$\mathbb{P}(Bel \leq \mathbb{P}_X) \geq \mathbb{P}(P^- \leq \mathbb{P}_X) \geq 1 - \alpha.$$

- Most Commitment Principle: find a belief function $\mathcal{B}(P^-)$ as committed as possible, by maximizing a measure of specificity.

Optimization problem

- For instance, we can maximize criterion

$$J(m) = \sum_{A \subseteq \mathcal{X}} Bel(A) = 2^K \sum_{B \subseteq \mathcal{X}} 2^{-|B|} m(B).$$

subject to the constaints

$$\sum_{B \subseteq A} m(B) \leq P^-(A), \quad \forall A \subset \mathcal{X},$$

$$\sum_{A \subseteq \mathcal{X}} m(A) = 1,$$

$$m(A) \geq 0, \quad \forall A \subseteq \mathcal{X}.$$

- This is a linear optimization problem.

Example: Psychiatric Data

A	$P^{-}(A)$	$Bel^{*}(A)$	$m^{*}(A)$
$\{\xi_1\}$	0.33	0.33	0.33
$\{\xi_2\}$	0.16	0.14	0.14
$\{\xi_1, \xi_2\}$	0.50	0.50	0.021
$\{\xi_3\}$	0.11	0.097	0.097
$\{\xi_1, \xi_3\}$	0.45	0.45	0.020
$\{\xi_2, \xi_3\}$	0.28	0.28	0.036
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\{\xi_1, \xi_3, \xi_4\}$	0.70	0.66	0.038
$\{\xi_2, \xi_3, \xi_4\}$	0.50	0.48	0.019
$\mathcal{X}$	1	1	0

Case of ordered data

- Assume $\mathcal{X}$ is ordered: $\xi_1 < \ldots < \xi_K$.
- The focal sets of $Bel(\cdot|\boldsymbol{x})$ can be constrained to be intervals $A_{k,r} = \{\xi_k, \ldots, \xi_r\}$.
- Under this additional constraint, an analytical solution to the previous optimization problem can be found:

$$m^*(A_{k,k}) = P_k^-,$$

$$m^*(A_{k,k+1}) = P^-(A_{k,k+1}) - P^-(A_{k+1,k+1}) - P^-(A_{k,k}),$$

$$m^*(A_{k,r}) = P^-(A_{k,r}) - P^-(A_{k+1,r}) - P^-(A_{k,r-1}) + P^-(A_{k+1,r-1})$$

for $r > k+1$, and $m^*(B) = 0$, for all $B \notin \mathcal{I}$.

Example: rain data

- January precipitation in Arizona (in inches), recorded during the period 1895-2004.

class ξ_k	n_k	n_k/n	p_k^-	p_k^+
< 0.75	48	0.44	0.32	0.56
$[0.75, 1.25)$	17	0.15	0.085	0.27
$[1.25, 1.75)$	19	0.17	0.098	0.29
$[1.75, 2.25)$	11	0.10	0.047	0.20
$[2.25, 2.75)$	6	0.055	0.020	0.14
≥ 2.75	9	0.082	0.035	0.18

- Degree of belief that the precipitation in Arizona next January will exceed, say, 2.25 inches?

Rain data: Result

$m(A_{k,r})$	1	2	3	4	5	6
1	0.32	0	0	0.13	0.11	0
2	-	0.085	0	0	0.012	0.14
3	-	-	0.098	0	0	0
4	-	-	-	0.047	0	0
5	-	-	-	-	0.020	0
6	-	-	-	-	-	0.035

- We get $Bel(X \geq 2.25) = Bel^*(\{\xi_5, \xi_6\}) = 0.055$ and $Pl(X \geq 2.25) = 0.317$.
- In 95 % of cases, the intervals $[Bel^*(A), Pl^*(A)]$ computed using this method simultaneously contain $\mathbb{P}_X(A)$ for all $A \subseteq \mathcal{X}$.

Outline

Continuous case

- If X is absolutely continuous, $\Omega = \mathbb{R}$
- A solution can be obtained using a confidence band on the cumulative distribution function F_X of X.
- Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be an iid sample from X with cdf F_X.
- A pair of functions $(\underline{F}(\cdot; \boldsymbol{X}), \overline{F}(\cdot; \boldsymbol{X}))$ computed from $\boldsymbol{X}$ and such that $\underline{F}(\cdot; \boldsymbol{X}) \leq \overline{F}(\cdot; \boldsymbol{X})$ is a confidence band at level $\alpha \in (0, 1)$ if

$$P\left\{\underline{F}(x; \boldsymbol{X}) \leq F_X(x) \leq \overline{F}(x; \boldsymbol{X}), \ \forall x \in \mathbb{R}\right\} = 1 - \alpha,$$

Kolmogorov Confidence band

- A non parametric confidence band can be computed using the Kolmogorov statistic:

$$D_n = \sup_x |S_n(x; \boldsymbol{X}) - F_X(x)|,$$

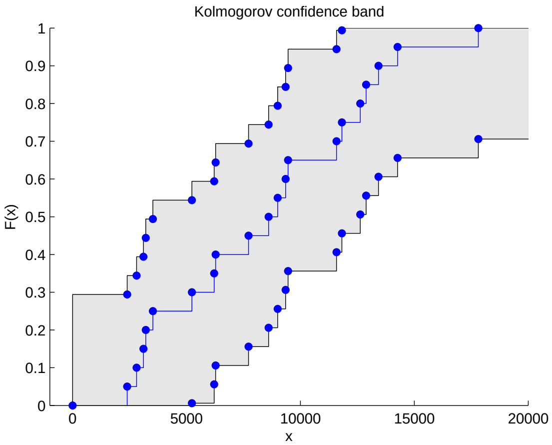
where $S_n(\cdot; \boldsymbol{X})$ is the sample cdf.

- The probability distribution of D_n can be computed exactly. Let $d_{n,\alpha}$ by the α-critical value of D_n, i.e., $\mathbb{P}(D_n \geq d_{n,\alpha}) = \alpha$.
- The two step functions

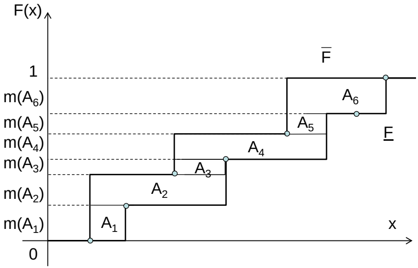
$$\begin{aligned} \underline{F}(x; \boldsymbol{X}) &= \max(0, S_n(x; \boldsymbol{X}) - d_{n,\alpha}), \\ \overline{F}(x; \boldsymbol{X}) &= \min(1, S_n(x; \boldsymbol{X}) + d_{n,\alpha}) \end{aligned}$$

form a confidence band at level $1 - \alpha$.

Bearings data ($1 - \alpha = 0.95$)

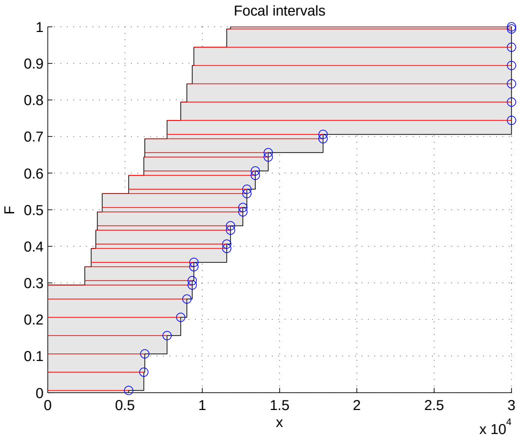


p-boxes and belief functions

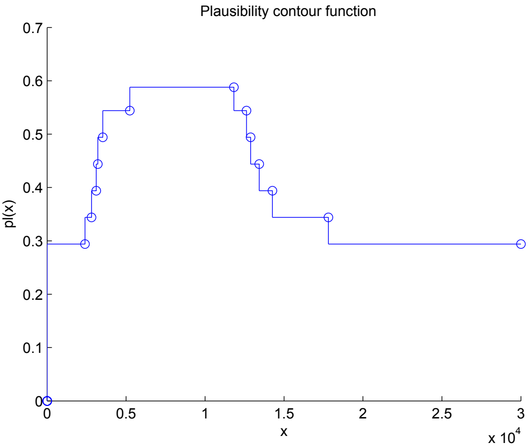


- A Kolmogorov confidence band defines a p-box (a set of probability measures with cdf constrained by 2 step functions).
- A p-box is equivalent to a discrete random interval.
- The belief function constructed from a Kolmogorov confidence band at level $1 - \alpha$ is a predictive belief function at level $1 - \alpha$.

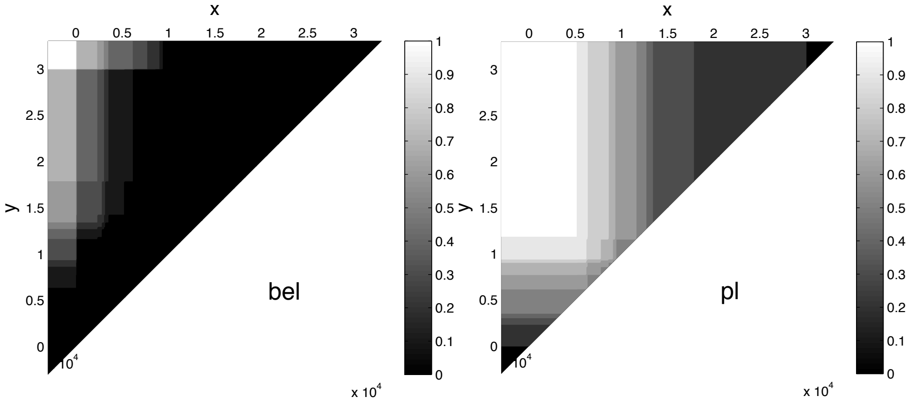
Bearings data: Construction of a mass function from a p-box



Bearings data: Contour function



Bearings data: Belief and plausibility functions



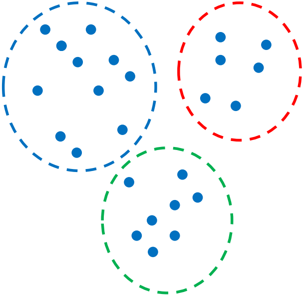
Outline

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:
  1. Clustering
  2. Association

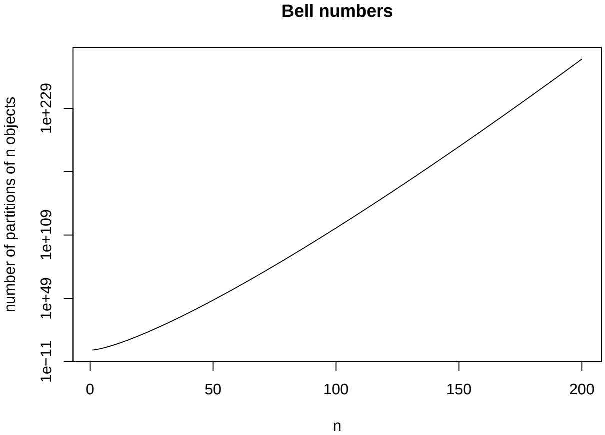
Outline

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

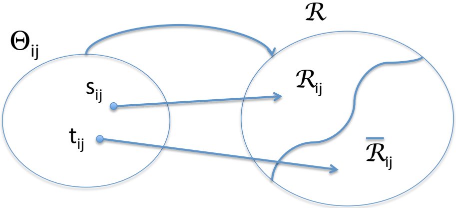
- Evidence: $n \times n$ matrix $D = (d_{ij})$ of dissimilarities between the n objects.
- For any $i < j$, let $\Theta_{ij} = \{s_{ij}, t_{ij}\}$, where s_{ij} means "objects i and j belong to the same class" and t_{ij} means "objects i and j do not belong to the same group".
- Assumptions:
  1. Two 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 (not true, but maybe 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

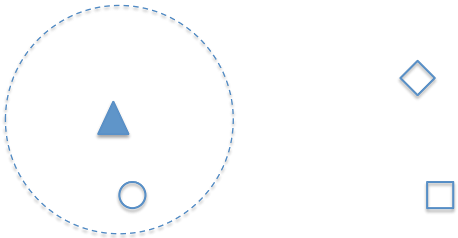
$$\log pl(R) = -\sum_{i<j} r_{ij} \log(1 - \varphi(d_{ij})) + C$$

- 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.
- This is a decision-directed clustering procedure, using the evidential k-nearest neighbor (Ek-NN) rule as a base classifier.

Example
Toy dataset

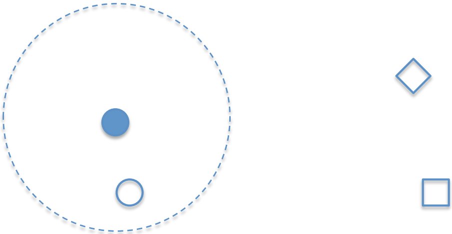
Example

Iteration 1



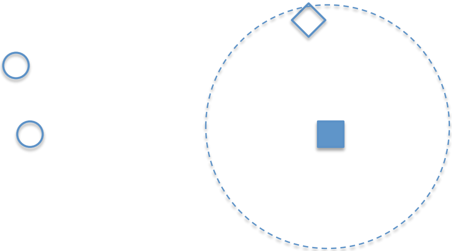
Example

Iteration 1 (continued)



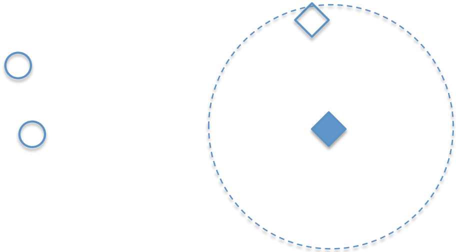
Example

Iteration 2



Example

Iteration 2 (continued)



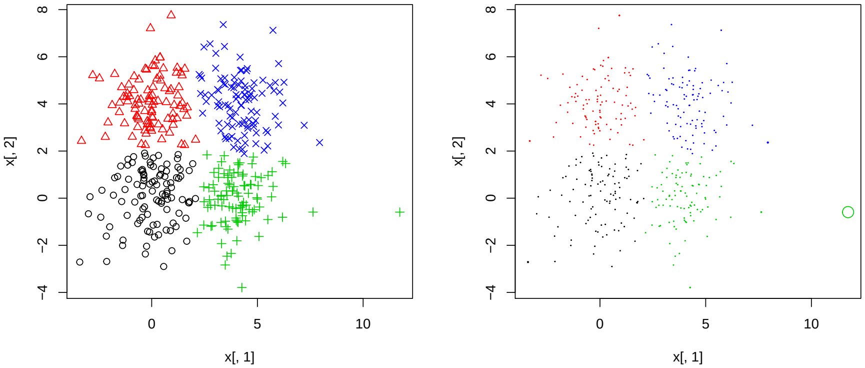
Example

Result

Ek-NNclus

- Starting from a random initial partition, classify each object in turn, using the Ek-NN rule.
- The algorithm converges to a local maximum of the contour function $pl(R)$ if $k = n - 1$.
- With $k < n - 1$, the algorithm converges to a local maximum of an objective function that approximates $pl(R)$.

Example



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.
- Problem: 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.

Formalization

- Let R_{ij} be a binary variable equal to 1 if e_i and f_j are the same object, 0 otherwise.
- We know the distances d_{ij} between the positions of each objects e_i and f_j.
- Each distance d_{ij} that induces a mass function m_{ij} on Θ_{ij}, for instance,

$$\begin{aligned} m_{ij}(\{1\}) &= \rho\varphi(d_{ij}) = \alpha_{ij} \\ m_{ij}(\{0\}) &= \rho\,(1 - \varphi(d_{ij})) = \beta_{ij} \\ m_{ij}(\Theta_{ij}) &= 1 - \rho = 1 - \alpha_{ij} - \beta_{ij}, \end{aligned}$$

  where $\rho \in [0, 1]$ is a degree of confidence in the sensor information and φ is a decreasing function taking values in $[0, 1]$.
- As before these *np* mass functions can be carried to the same frame and combine by Dempster's rule.

Vacuous extension

- Let $\mathcal{R}$ be the set of matching relations between sets E and F (each object in E can be matched to at most one object in F, and conversely).
- Let $\mathcal{R}_{ij}$ be the set of matching relations where object e_i is matched to object f_j.
- As before, each m_{ij} is vacuously extended to $\mathcal{R}$,

$$\begin{array}{ccc} m_{ij}(\{1\}) & \longrightarrow & \mathcal{R}_{ij} \\ m_{ij}(\{0\}) & \longrightarrow & \overline{\mathcal{R}_{ij}} \\ m_{ij}(\Theta) & \longrightarrow & \mathcal{R} \end{array}$$

Combination

- The contour function of m_{ij} is

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

- The combined contour function is thus

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

and its logarithm is

$$
\begin{aligned}
\ln pl(R) &= \sum_{i,j}[R_{ij}\ln(1-\beta_{ij}) + (1-R_{ij})\ln(1-\alpha_{ij})] + C \\
&= \sum_{i,j} w_{ij}R_{ij} + C
\end{aligned}
$$

Decision

- To find the matching relation R with greatest plausibility, we need to solve the following linear optimization problem,

$$\max \sum_{i,j} w_{ij} R_{ij} + C$$

subject to

$$\begin{array}{ll} \sum_{j=1}^{p} R_{ij} \leq 1 & \forall i \in \{1,\ldots,n\} \\ \sum_{i=1}^{n} R_{ij} \leq 1 & \forall j \in \{1,\ldots,p\} \\ R_{ij} \in \{0,1\} & \forall i \in \{1,\ldots,n\}, \forall j \in \{1,\ldots,p\}, \end{array}$$

- This is a linear assignment problem, which can be solved in $o(\max(n,m)^3)$ time.

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
  - Predictive belief function
  - Likelihood-based 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. Denoeux 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. Denoeux.
Constructing Belief Functions from Sample Data Using Multinomial Confidence Regions.
International Journal of Approximate Reasoning, 42(3):228–252, 2006.

A. Aregui and T. Denoeux.
Constructing Predictive Belief Functions from Continuous Sample Data Using Confidence Bands.
In G. De Cooman and J. Vejnarova and M. Zaffalon (Eds), *Proceedings of the Fifth International Symposium on Imprecise Probability: Theories and Applications (ISIPTA '07)*, pages 11-20, Prague, Czech Republic, July 2007.

References II

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

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

T. Denoeux, 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. Denoeux, 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.