Combining statistical and subjective evidence: the belief function approach

Thierry Denœux[1]

[1]Université de Technologie de Compiègne
HEUDIASYC (UMR CNRS 6599)

Fifth Int. Conf. of the Thailand Econometric Society
Chiang Mai, January 12-13, 2012

Outline

Outline

Two kinds of evidence

- Inference and prediction procedures (as used in econometrics and other fields) can be seen as formal methods for computing statements about quantities of interests (e.g., confidence intervals, probability distributions, etc.), taking into account partial information (evidence).
- Two kinds of evidence:
  - Statistical data (observations that can be assumed to be drawn at random from a well-defined population);
  - Expert opinions (subjective statements summarizing observations and experiences gathered by individuals).
- In this talk, we are concerned with situations in which objective and subjective evidence coexist and have comparable importance in the reasoning and/or decision making process.

Example: 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.
- Traditionally, 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.
- A preliminary approach in the case of coastal infrastructure will be reported in this talk.

Outline

Bayesian framework

- The Bayesian framework has often been advocated as a suitable formalism for reasoning with objective and subjective evidence.
- However, this theory can be criticized as being too constrained to represent near-ignorance states of knowledge, as exemplified by the Von Mises Wine-Water paradox:

  There is a certain quantity of liquid. All that we know about the liquid is that it is composed entirely of wine and water, and the ratio of wine to water is between 1/3 and 3. What is the probability that the ratio of wine to water is less than or equal to 2?

Wine/water paradox

- Let θ denote the ratio of wine to water. All we know is that $\theta \in [1/3, 3]$. According to the Principle of Indifference (PI), $\theta \sim \mathcal{U}_{[1/3,3]}$. Consequently:

$$P(\theta \leq 2) = (2 - 1/3)/(3 - 1/3) = 5/8$$

- Now, let $\gamma = 1/\theta$ denote the ratio of water to wine. All we know is that $\gamma \in [1/3, 3]$. According to the PI, $\gamma \sim \mathcal{U}_{[1/3,3]}$. Consequently:

$$P(\theta \leq 2) = P(\gamma \geq 1/2) = (3 - 1/2)/(3 - 1/3) = 15/16$$

What's wrong?

- In the wine/water story, the information is provided as a set.
- A uniform distribution on a set A is not an adequate representation of the information $\theta \in A$, as it is not invariant with respect to non linear transformations: if $\boldsymbol{\theta}$ has uniform distribution on A, then $f(\boldsymbol{\theta})$ usually does not have a uniform distribution on $f(A)$.
- In contrast, set-membership methods such a interval analysis cannot cope with probabilistic information.
- We need a formalism combining set-theoretic (logical) and probabilistic approaches.
- The theory of belief functions is such a framework.

Outline

Historical perspective

- Also referred to as Dempster-Shafer (DS) or Evidence theory.
- Initially introduced by Dempster (1966, 1968) with the objective to reconcile Bayesian and fiducial inference.
- Shafer (1976) later formalized this approach as a general method for representing and combining any kind of evidence (such as testimonies).
- Although originating from Statistics, DS theory has been until now mainly popular in Artificial Intelligence (expert systems, robotics, machine learning, etc.).
- The theory of belief functions is a general framework allowing us to combine statistical data and expert judgements in a unified and principled way.

Main features

- The theory of belief function subsumes both the logical and probabilistic approaches to uncertainty: a belief function may be seen as
  - a non-additive measure or as
  - a generalized set.
- The belief function approach coincides with the Bayesian approach when all variables are described by probability distributions.
- Due to its greater expressive power, it allows for a more faithful representation of “weak” evidence when the available information cannot be described by a probability distribution without introducing unsupported assumptions.

Outline

Mass function

Definition

- Let θ be a variable taking values in a finite domain Θ, called the frame of discernment.
- Uncertain evidence about θ may be represented by a mass function m on Θ, defined as a function $2^\Theta \to [0,1]$, such that $m(\emptyset) = 0$ and

$$\sum_{A \subseteq \Theta} m(A) = 1.$$

- Any subset A of Θ 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).

Mass function
Example

- A murder has been committed. There are three suspects: $\Theta = \{Peter, John, Mary\}$.
- 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 $\theta \in \{Peter, John\}$. Otherwise, we only know $\theta \in \Theta$. The first case holds with probability 0.8.
- Corresponding mass function:

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

Random code semantics

- Which meaning can exactly be assigned to such numbers $m(A)$?
- Possible interpretation: a piece of evidence can be represented by a mass function m with focal sets A_i, $i = 1, \ldots, r$ iff it induces the same state of knowledge about the variable of interest as would be induced by the following chance set-up:
  - We receive a coded message that was encoded using a code selected at random from a known set $\Omega = \{\omega_1, \ldots, \omega_r\}$;
  - We know the chance m_i of each code ω_i being selected.
  - Decoding the message using code ω_i produces a new message of the form "$\boldsymbol{\theta} \in A_i$".
- In this setting, $m(A_i) = m_i$ is the chance that the original message was "$\boldsymbol{\theta} \in A_i$", i.e., the probability of knowing only that $\boldsymbol{\theta} \in A_i$.

Link with the random set framework

- To each mass function m on Θ can thus be associated a triple (Ω, P, Γ), with $\Gamma : \Omega \to 2^{\Theta} \setminus \{\emptyset\}$.
- This formally defines a random set: mass functions and random sets are thus equivalent from a mathematical point of view.
- However, they have different interpretations:
  - Random set view: a random mechanism generates each set A with chance $m(A)$. Example: taking a handful of balls from an urn.
  - Belief function view: a given piece of evidence supports different hypotheses with different subjective probabilities. Example: taking a single ball from an urn and partially observing the result.

Belief function

Definition and interpretation

- The belief function induced by m is defined as

$$Bel(A) = \sum_{B \subseteq A} m(B), \quad \forall A \subseteq \Theta.$$

- $Bel(A)$ can be seen as the probability that the evidence implies that $\theta \in A$:

$$Bel(A) = P(\{\omega \in \Omega | \Gamma(\omega) \subseteq A\}.$$

- It can thus be interpreted as:
    - a total degree of support in A provided by the item of evidence;
    - a measure of our total belief committed to A after receiving that item of evidence.

Belief function
Characterization

- Function $Bel : 2^\Theta \to [0,1]$ is a completely monotone capacity: it verifies $Bel(\emptyset) = 0$, $Bel(\Theta) = 1$ and

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

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

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

- m and *Bel* are thus equivalent representations.

Plausibility function

- The plausibility function is defined by

$$Pl(A) = 1 - Bel(\overline{A}) = \sum_{B \cap A \neq \emptyset} m(B), \quad \forall A \subseteq \Theta$$

- Interpretation:
  - Degree to which the evidence is not contradictory with A:
  - Probability that A cannot be refuted by the available evidence.
- The function $pl : \Theta \rightarrow [0, 1]$ such that $pl(\theta) = Pl(\{\theta\})$ is called the contour function associated to m.

Belief and plausibility functions

Special cases

- If m is Bayesian, then $Bel = Pl$ is a probability measure.
- If the focal sets are nested, m is said to be consonant. Then, Pl is a possibility measure, i.e.,

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

and Pl is uniquely defined from the contour function pl as

$$Pl(A) = \max_{\theta \in A} pl(\theta), \quad \forall A \subseteq \Theta.$$

Dempster's rule
Principle

- Let m_1 and m_2 be two mass functions induced by triples $(\Omega_1, P_1, \Gamma_1)$ and $(\Omega_2, P_2, \Gamma_2)$ interpreted under the random code framework as before. Let us further assume that the codes are selected independently.
- The probability that any two codes $\omega_1 \in \Omega_1$ and $\omega_2 \in \Omega_2$ are both selected is $P_1(\{\omega_1\})P_2(\{\omega_2\})$, in which case we can conclude that $\boldsymbol{\theta} \in \Gamma_1(\omega_1) \cap \Gamma_2(\omega_2)$.
- If $\Gamma_1(\omega_1) \cap \Gamma_2(\omega_2) = \emptyset$, we know that the pair (ω_1, ω_2) could not have been selected: consequently, the joint probability distribution on $\Omega_1 \times \Omega_2$ must be conditioned, eliminating such pairs.

Dempster's rule
Definition

- This line of reasoning yields the following combination rule, referred to as Dempster's rule:

$$(m_1 \oplus m_2)(A) = \frac{1}{1-\kappa} \sum_{B \cap C = A} m_1(B) m_2(C)$$

for all $A \subseteq \Theta$, $A \neq \emptyset$ and $(m_1 \oplus m_2)(\emptyset) = 0$, where

$$\kappa = \sum_{B \cap C = \emptyset} m_1(B) m_2(C) \quad (1)$$

is the degree of conflict between m_1 and m_2.
- Dempster's rule is commutative, associative, and it admits as neutral element the vacuous mass function defined as $m(\Theta) = 1$.

Outline

1. Motivation and background
   - Motivation
   - Theories of uncertainty
2. Theory of Belief functions
   - Belief functions on finite domains
   - Random closed intervals
   - Statistical inference using BFs
3. Sea level rise example
   - Problem statement and assumptions
   - Belief function model

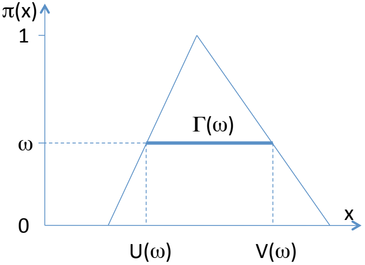
Random intervals
Definition

- The equivalence between belief functions and random sets allows us to define belief functions in infinite spaces. Here, we will restrict our discussion to random closed intervals on the real line.
- Let $(\Omega, \mathcal{A}, P)$ be a probability space and $(U, V) : \Omega \rightarrow \mathbb{R}^2$ a two-dimensional real random vector such that $P(\{\omega \in \Omega | U(\omega) \leq V(\omega)\}) = 1$.
- Let Γ be the multi-valued mapping that maps each $\omega \in \Omega$ to $[U(\omega), V(\omega)]$. This setting defines a random interval, as well as belief and plausibility functions defined by

$$
\begin{aligned}
Bel(A) &= P(\{\omega \in \Omega | [U(\omega), V(\omega)] \subseteq A\}) \\
Pl(A) &= P(\{\omega \in \Omega | [U(\omega), V(\omega)] \cap A \neq \emptyset\})
\end{aligned}
$$

for all $A \in \mathcal{B}(\mathbb{R})$.

Consonant random intervals



- Let $\pi : \mathbb{R} \to [0,1]$ be an upper semi-continuous function, $\Omega = [0,1]$, P the Lebesgues measure on $[0,1]$, and $\Gamma(\omega) = \{x \in \mathbb{R} | \pi(x) \geq \omega\}$.
- (Ω, P, Γ) defines a consonant random interval with contour function π and plausibility function

$$Pl(A) = \sup_{x \in A} \pi(x),$$

for all $A \in \mathcal{B}(\mathbb{R})$

Outline

Statistical inference problem

- Assume that we have observed a realization $\mathbf{x}$ of a random vector $\mathbf{X}$ with pdf $p(\mathbf{x};\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta$ is an unknown parameter.
- What does this item of evidence tell us about $\boldsymbol{\theta}$, and how to represent this information in the belief function framework?
- A solution was proposed by Shafer (1976) and justified axiomatically by Wasserman (1988).
- This solution depends only the likelihood function $L(\theta;\mathbf{x}) \propto p(\mathbf{x};\theta)$ and thus complies with the widely accepted likelihood principle.

Likelihood-based belief function

Case $\Theta = \{\theta_1, \theta_2\}$

- Assume $\Theta = \{\theta_1, \theta_2\}$ has only two points.
- Let $Bel(\cdot; \mathbf{x})$ be a belief function on Θ based on $\mathbf{x}$.
- It seems natural to impose the following requirements:
  1. If $L(\theta_1; \mathbf{x}) = L(\theta_2; \mathbf{x})$, then $Bel(\cdot; \mathbf{x})$ should be vacuous;
  2. $Bel(\{\theta\}; \mathbf{x})$ should be nondecreasing in $L(\theta; \mathbf{x})$;
  3. If $Bel = Bel(\cdot; \mathbf{x}) \oplus P_0$ and P_0 is a probability measure, then Bel should be equal to the Bayesian posterior,
- The solution is unique and it corresponds to a consonant belief function with following contour function:

$$pl(\theta; \mathbf{x}) = \frac{L(\theta; \mathbf{x})}{\max\left(L(\theta_1; \mathbf{x}), L(\theta_2; \mathbf{x})\right)}, \quad \forall \theta \in \Theta$$

Likelihood-based belief function

General case

- The above argument may be extended to the case where Θ is a complete, measurable space (Wasserman, 1988).
- In this case, the only belief function whose conditional on every two-point subset satisfies the previous three requirements and a certain continuity condition corresponds to the consonant belief function with contour function equal to the relative likelihood function:

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

Compatibility with Bayesian inference

- By construction, combining $Bel(\cdot; \mathbf{x})$ with a Bayesian prior P_0 on Θ using Dempster's rule yields a Bayesian belief function $Bel(\cdot; \mathbf{x}) \oplus P_0$ which is identical to the posterior probability obtained using Bayes' rule.
- Consequently, the proposed method of inference boils down to Bayesian inference when a Bayesian prior is available.
- However, prior knowledge need not be described by a probability measure. It can be any belief function, even vacuous.

Compatibility with 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

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

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

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

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

Outline

heudiasyc

Outline

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 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 $1 - 1/T$.
- 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 $z'_T = z_T + SLR$.

Outline

Statistical evidence on z_T

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

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

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

- The corresponding contour function is thus:

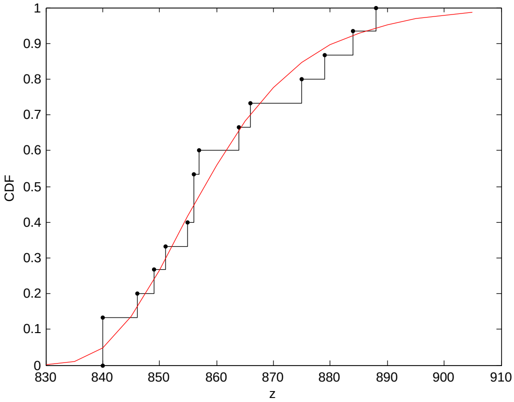
$$pl(z_T, \mu) = \frac{L(z_T, \mu)}{\sup_{z_T, \mu} L(z_T, \mu)}$$

  and the marginal contour function of z_T is

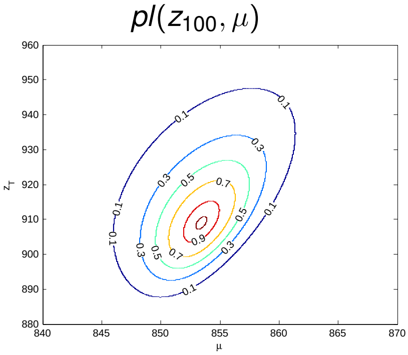
$$pl(z_T) = \sup_{\mu} pl(z_T, \mu).$$

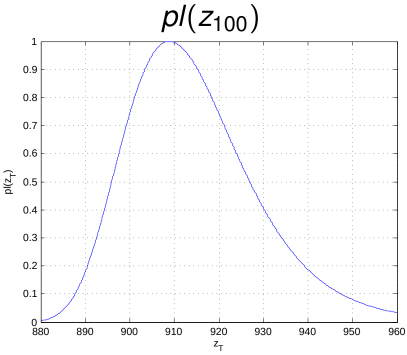
Data

15 years of sea level data at Le Havre, France



Results

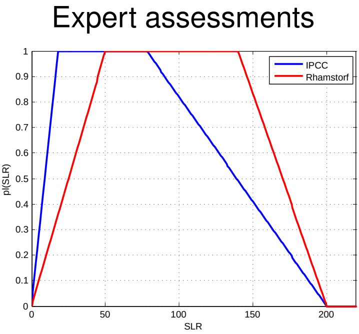

$pl(z_{100}, \mu)$
960
950
940
930
920
910
900
890
880
840
845
850
855
860
865
870
μ
z_T
0.1
0.3
0.5
0.7
0.9

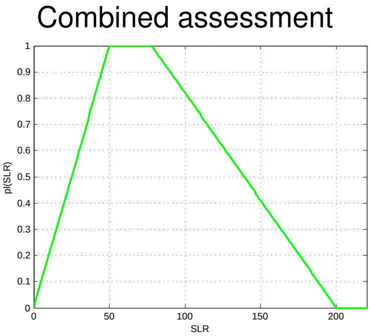

$pl(z_{100})$
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
880
890
900
910
920
930
940
950
960
z_T
$pl(z_T)$

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 opinions





Combination

Principle

- Let $[U_{z_T}, V_{z_T}]$ and $[U_{SLR}, V_{SLR}]$ be the independent random intervals corresponding to $pl(z_T)$ and $pl(SLR)$, respectively.
- The random interval for $z'_T = z_T + SLR$ is

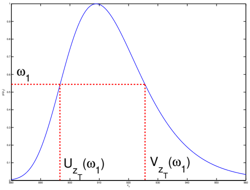
$$[U_{z_T}, V_{z_T}] + [U_{SLR}, V_{SLR}] = [U_{z_T} + U_{SLR}, V_{z_T} + V_{SLR}]$$

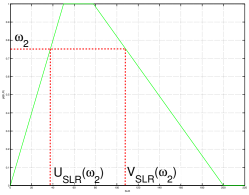
- The corresponding belief and plausibility functions are

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

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

Combination
Monte Carlo simulation



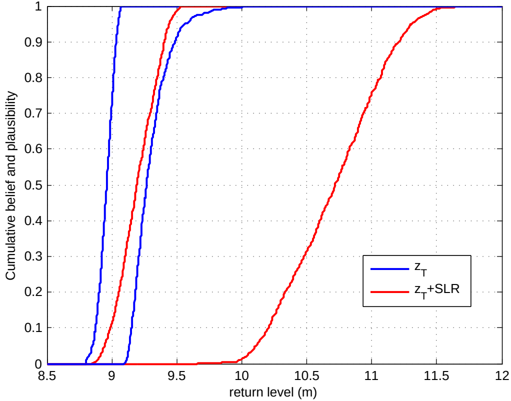


Algorithm to approximate $Pl(A)$:

$k = 0$
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
$\widehat{Pl}(A) = \frac{k}{N}$

heudiasyc

Result


Cumulative belief and plausibility
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
8.5
9
9.5
10
10.5
11
11.5
12
return level (m)
z_T
z_T+SLR

Summary and research challenges

- The theory of belief functions makes it possible to represent both expert opinions and statistical evidence in a unified framework.
- When applied to statistical inference, the belief function formalism is fully compatible with the likelihood and Bayesian methods, while allowing for intermediate situations in which only weak prior information is available.
- More work is needed on:
  - Elicitation of expert opinions (including the multiple experts case);
  - Comparison with alternative methods of statistical inference (e.g., Dempster);
  - Efficient combination and propagation procedures in infinite multidimensional spaces, etc.