Methods for building belief functions

Thierry Denœux

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

Third School on belief functions and their applications,
Stella Plage, France, September 28, 2015

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
- We need methods for modeling evidence from
  - expert opinions or
  - statistical information
- In this lecture, we will review some general methods and give some practical examples

Outline

1. Least Commitment Principle
   - Example 1: LC mass function with given contour function
   - Deconditioning and the GBT
   - Uncertainty measures
2. Statistical estimation and prediction
   - Motivation
   - Likelihood-based belief function
   - Predictive belief function
3. Example: linear regression
   - Estimation and prediction
   - Example
   - Uncertain inputs

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

Outline

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

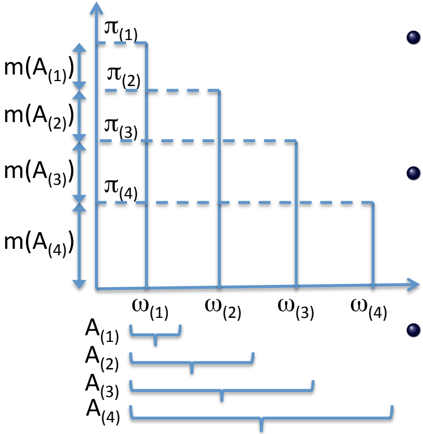
- Taking $\sqsubseteq_q$ as the informational ordering, the least committed element in $\mathcal{M}(\pi)$ is the consonant mass function whose contour function is π
- Its plausibility and commonality functions are defined as

$$Pl(A) = \max_{\omega \in A} \pi(\omega), \quad Q(A) = \min_{\omega \in A} \pi(\omega)$$

for all $A \subseteq \Omega$, $A \neq \emptyset$

LC mass function with given contour function

Recovering the mass function



- Let $1 = \pi_{(1)} \geq \pi_{(2)} \geq \ldots \geq \pi_{(K)}$ be the ordered values of π; $\omega_{(1)}, \ldots, \omega_{(K)}$ the elements of Ω in the corresponding order, and $A_{(k)} = \{\omega_{(1)}, \ldots, \omega_{(k)}\}$
- We have

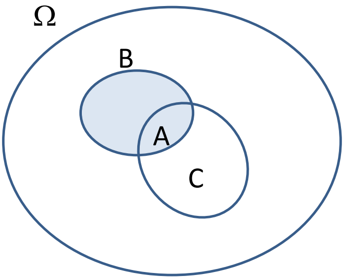
$$m(A_{(k)}) = \pi_{(k)} - \pi_{(k+1)}$$

for $k = 1, \ldots, K-1$ and $m(\Omega) = \pi_{(K)}$

- Source: $\Theta = [0, 1]$, P= Lebesgue measure, $\Gamma(\theta) = A_{(k)}$ if $\theta \in [\pi_{(k+1)}, \pi_{(k)}]$

Outline

Deconditioning



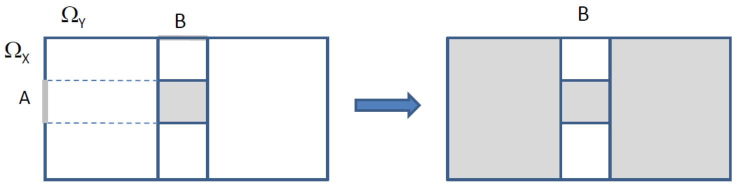
- Let m_0 be a mass function on Ω expressing our beliefs about X in a context where we know that $X \in B$
- We want to build a mass function m verifying the constraint $m(\cdot|B) = m_0$
- Any m built from m_0 by transferring each mass $m_0(A)$ to $A \cup C$ for some $C \subseteq \overline{B}$ satisfies 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_0^X be a mass function on Ω_X expressing our beliefs about X in a context where we know that $Y \in B$ for some $B \subseteq \Omega_Y$
- We want to find m^{XY} such that $\left(m^{XY} \oplus (m_B^Y)^{\uparrow XY}\right)^{\downarrow X} = m_0^X$
- s-least committed solution: transfer $m_0^X(A)$ to $(A \times \Omega_Y) \cup (\Omega_X \times \overline{B})$
- Notation $m^{XY} = (m_0^X)^{\Uparrow XY}$ (conditional embedding)

Generalized Bayes Theorem

Problem statement

- Two variables $X \in \Omega$ et $\theta \in \Theta = \{\theta_1, \ldots, \theta_K\}$
- Typically
  - X is observed (sensor measurement)
  - θ is not observed (class, unknown parameter)
- Partial knowledge of X given each $\theta = \theta_k$

$$m^{\Omega}(\cdot|\theta_k), \quad k = 1, \ldots, K$$

- Prior knowledge about θ: $m_0^{\Theta}(\Theta)$ (may be vacuous)
- We observe $X \in A$
- Belief function on Θ?

Generalized Bayes Theorem

Solution

- Solution:

$$m^{\Theta}(\cdot|A) = \left(\bigoplus_{k=1}^{K} m^{\Omega}(\cdot|\theta_k)^{\Uparrow \Omega\times\Theta} \oplus m_A^{\Omega\uparrow\Omega\times\Theta} \oplus m_0^{\Theta\uparrow\Omega\times\Theta} \right)^{\downarrow\Theta}$$

- Expression

$$m^{\Theta}(\cdot|A) = \bigoplus_{k=1}^{K} \overline{\{\theta_k\}}^{pl^{\Omega}(A|\theta_k)} \oplus m_0^{\Theta}$$

where $\overline{\{\theta_k\}}^{pl^{\Omega}(A|\theta_k)}$ is the simple mass function that assigns the mass $1 - pl^{\Omega}(A|\theta_k)$ to $\overline{\{\theta_k\}}$ and $pl^{\Omega}(A|\theta_k)$ to Θ

Generalized Bayes Theorem

Properties

- Property 1: Bayes' theorem is recovered as a special case when the conditional mass functions $m^\Omega(\cdot|\theta_k)$ and m_0^Θ are Bayesian
- Property 2: If X and Y are cognitively independent conditionally on θ, i.e.,

$$pl^{XY}(A \times B|\theta_k) = pl^X(A|\theta_k) \cdot pl^Y(B|\theta_k)$$

for all $A \subseteq \Omega_X$, $B \subseteq \Omega_Y$ and $\theta_k \in \Theta$, then

$$m^\Theta(\cdot|X \in A, Y \in B) = m^\Theta(\cdot|X \in A) \oplus m^\Theta(\cdot|Y \in B)$$

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), $f = \log_2$ (nonspecificy)

- Nonspecificity $N(m)$ generalizes the Hartley function for set ($H(A) = \log_2(|A|)$) are was shown by Ramer (1987) to be the unique measure verifying some axiomatic requirements such as
  - Additivity for non-interactive mass functions: $N(m^{\Omega \times \Theta}) = N(m^{\Omega}) + N(m^{\Theta})$
  - Subadditivity for interactive mass functions: $N(m^{\Omega \times \Theta}) \leq N(m^{\Omega}) + N(m^{\Theta})$
  - ...
- 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
- Example: the pignistic probability distribution is defined by

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

  for all $\omega \in \Omega$
- Corresponding uncertainty measure

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

- As other total uncertainty measures, $EP(m)$ extends both 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

Estimation vs. prediction

- Consider an urn with an unknown proportion θ of black balls
- Assume that we have drawn n balls with replacement from the urn, y of which were black
- Problems
  1. What can we say about θ? (estimation)
  2. What can we say about the color Z of the next ball to be drawn from the urn? (prediction)
- Classical approaches
  - Frequentist: gives an answer that is correct most the time (over infinitely many replications of the random experiment)
  - Bayesian: assumes prior knowledge on θ and computes a posterior predictive probabilities $f(\theta|y)$ and $P(black|y)$

Criticism of the frequentist approach

- The frequentist approach makes a statement that is correct, say, for 95% of the samples
- However, 95% is not a correct measure of the confidence in the statement for a particular sample
- Example:
  - Let the prediction be $\{black, white\}$ with probability 0.95 and $\emptyset$ with probability 0.05 (irrespective of the data). This is a 95% prediction set.
  - This prediction is either know for sure to be true, or known for sure to be false.
- Also, the frequentist approach does not allow us to easily
  - Use additional information on θ, if it is available
  - Combine predictions from several sources/agents

Criticism of the Bayesian approach

- In the Bayesian approach, y, z and θ are seen as random variables
- Estimation: compute the posterior pdf of θ given y

$$f(\theta|y) \propto p(y|\theta)f(\theta)$$

  where $f(\theta)$ is the prior pdf on θ
- Prediction: compute the predictive posterior distribution

$$p(z|y) = \int p(z|\theta)f(\theta|y)d\theta$$

- We need the prior $f(\theta)$!

Main ideas

- None of the classical approaches to prediction (frequentist and Bayesian) is conceptually satisfactory
- Proposal of a new approach based on belief functions
- The new approach boils down to Bayesian prediction when a probabilistic prior is available, but it does not require the user to provide such a prior
- Application: linear Regression

Outline of the new approach (1/2)

- Let us come back to the urn example
- Let $Z \sim \mathcal{B}(\theta)$ be defined as

$$Z = \begin{cases} 1 & \text{if next ball is black} \\ 0 & \text{otherwise} \end{cases}$$

- We can write Z as a function of θ and a pivotal variable $W \sim \mathcal{U}([0,1])$,

$$Z = \begin{cases} 1 & \text{if } W \leq \theta \\ 0 & \text{otherwise} \end{cases}$$
$$= \varphi(\theta, W)$$

Outline of the new approach (2/2)

- The equality

$$Z = \varphi(\theta, W)$$

allows us to separate the two sources of uncertainty on Z

1. uncertainty on W (random/aleatory uncertainty)
2. uncertainty on θ (epistemic uncertainty)

- Two-step method:

1. Represent uncertainty on θ using a likelihood-based belief function Bel_y^Θ constructed from the observed data y (estimation problem)
2. Combine Bel_y^Θ with the probability distribution of W to obtain a predictive belief function $Bel_y^{\mathbb{Z}}$

Outline

Parameter estimation

- Let $\boldsymbol{y} \in \mathbb{Y}$ denote the observed data and $f_{\boldsymbol{\theta}}(\boldsymbol{y})$ the probability mass or density function describing the data-generating mechanism, where $\boldsymbol{\theta} \in \Theta$ is an unknown parameter
- Having observed $\boldsymbol{y}$, how to quantify the uncertainty about Θ, without specifying a prior probability distribution?
- Likelihood-based solution (Shafer, 1976; Wasserman, 1990; Denœux, 2014)

Likelihood-based belief function

Requirements

Let $Bel_{\boldsymbol{y}}^{\Theta}$ be a belief function representing our knowledge about θ after observing $\boldsymbol{y}$. We impose the following requirements:

1. Likelihood principle: $Bel_{\boldsymbol{y}}^{\Theta}$ should be based only on the likelihood function

$$\theta \rightarrow L_{\boldsymbol{y}}(\theta) = f_\theta(\boldsymbol{y})$$

2. Compatibility with Bayesian inference: when a Bayesian prior P_0 is available, combining it with $Bel_{\boldsymbol{y}}^{\Theta}$ using Dempster's rule should yield the Bayesian posterior:

$$Bel_{\boldsymbol{y}}^{\Theta} \oplus P_0 = P(\cdot|\boldsymbol{y})$$

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

Likelihood-based belief function

Solution (Denœux, 2014)

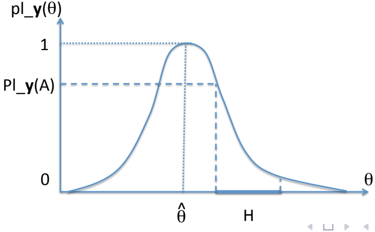
- $Bel_{\boldsymbol{y}}^{\Theta}$ is the consonant belief function induced by the relative likelihood function

$$pl_{\boldsymbol{y}}(\boldsymbol{\theta}) = \frac{L_{\boldsymbol{y}}(\boldsymbol{\theta})}{L_{\boldsymbol{y}}(\widehat{\boldsymbol{\theta}})}$$

where $\widehat{\boldsymbol{\theta}}$ is a MLE of $\boldsymbol{\theta}$, and it is assumed that $L_{\boldsymbol{y}}(\widehat{\boldsymbol{\theta}}) < +\infty$

- Corresponding plausibility function

$$Pl_{\boldsymbol{y}}^{\Theta}(H) = \sup_{\theta \in H} pl_{\boldsymbol{y}}(\theta), \quad \forall H \subseteq \Theta$$

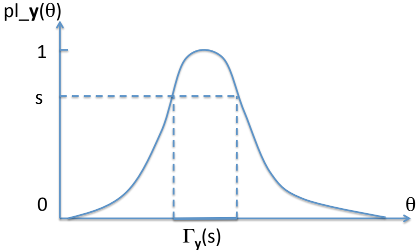


Source

- Corresponding random set:

$$\Gamma_{\boldsymbol{y}}(s) = \left\{ \theta \in \Theta \middle| \frac{L_{\boldsymbol{y}}(\theta)}{L_{\boldsymbol{y}}(\widehat{\theta})} \geq s \right\}$$

with s uniformly distributed in $[0, 1]$



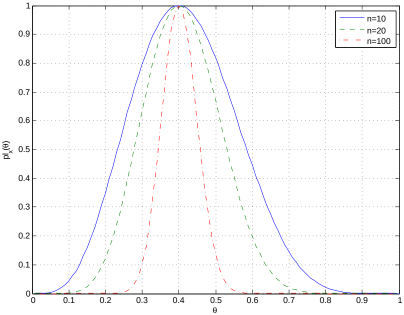
- If $\Theta \subseteq \mathbb{R}$ and if $L_{\boldsymbol{y}}(\theta)$ is unimodal and upper-semicontinuous, then $Bel_{\boldsymbol{y}}^{\Theta}$ corresponds to a random closed interval

Binomial example

In the urn model, $Y \sim \mathcal{B}(n, \theta)$ and

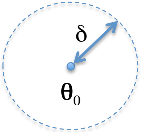
$$pl_y(\theta) = \frac{\theta^y (1-\theta)^{n-y}}{\widehat{\theta}^y (1-\widehat{\theta})^{n-y}} = \left(\frac{\theta}{\widehat{\theta}}\right)^{n\widehat{\theta}} \left(\frac{1-\theta}{1-\widehat{\theta}}\right)^{n(1-\widehat{\theta})}$$

for all $\theta \in \Theta = [0, 1]$, where $\widehat{\theta} = y/n$ is the MLE of θ.



Asymptotic consistency

- $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ iid from $f_{\boldsymbol{\theta}}(y)$, $\boldsymbol{\theta}_0$ = true value
- Let $B_\delta(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta} \in \Theta | \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \delta\}$ be a ball centered on $\boldsymbol{\theta}_0$, with radius δ



- Under mild assumptions, for all $\delta > 0$,

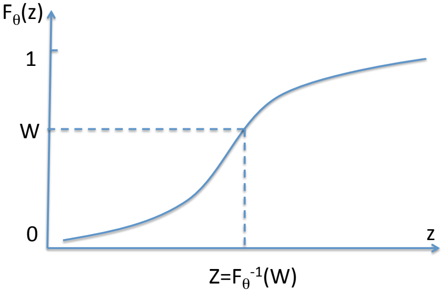
$$Bel_{\boldsymbol{Y}}^{\Theta}(B_\delta(\boldsymbol{\theta}_0)) \xrightarrow{a.s.} 1$$

Outline

Prediction problem

- Observed (past) data: $\boldsymbol{y}$ from $\boldsymbol{Y} \sim f_{\boldsymbol{\theta}}(\boldsymbol{y})$
- Future data: $Z|\boldsymbol{y} \sim F_{\boldsymbol{\theta},\boldsymbol{y}}(z)$ (real random variable)
- Problem: quantify the uncertainty of Z using a predictive belief function

φ-equation



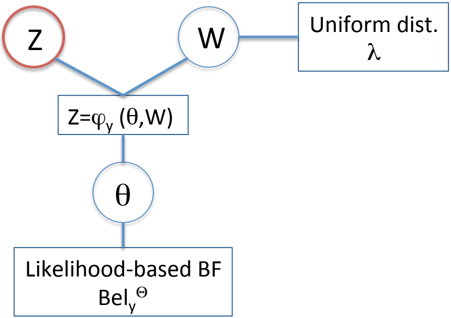
We can always write Z as a function of $\boldsymbol{\theta}$ and W as

$$Z = F_{\boldsymbol{\theta},\boldsymbol{y}}^{-1}(W) = \varphi_{\boldsymbol{y}}(\boldsymbol{\theta}, W)$$

where $W \sim \mathcal{U}([0,1])$ and $F_{\boldsymbol{\theta},\boldsymbol{y}}^{-1}$ is the generalized inverse of $F_{\boldsymbol{\theta},\boldsymbol{y}}$,

$$F_{\boldsymbol{\theta},\boldsymbol{y}}^{-1}(W) = \inf\{z | F_{\boldsymbol{\theta},\boldsymbol{y}}(z) \geq W\}$$

Main result

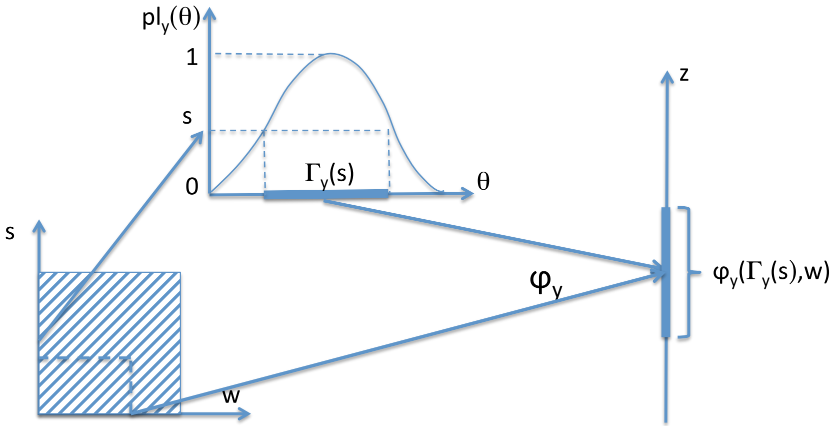


After combination by Dempster's rule and marginalization on $\mathbb{Z}$, we obtain the predictive BF on Z induced by the multi-valued mapping

$$(s, w) \to \varphi_{\boldsymbol{y}}(\Gamma_{\boldsymbol{y}}(s), w).$$

with (s, w) uniformly distributed in $[0, 1]^2$

Graphical representation



Practical computation

- Analytical expression when possible (simple cases), or
- Monte Carlo simulation:
  1. Draw N pairs (s_i, w_i) independently from a uniform distribution
  2. compute (or approximate) the focal sets $\varphi_{\boldsymbol{y}}(\Gamma_{\boldsymbol{y}}(s_i), w_i)$
- The predictive belief and plausibility of any subset $A \subseteq \mathbb{Z}$ are then estimated by

$$\widehat{Bel}^{\mathbb{Z}}_{\boldsymbol{y}}(A) = \frac{1}{N} \#\{i \in \{1, \ldots, N\} | \varphi_{\boldsymbol{y}}(\Gamma_{\boldsymbol{y}}(s_i), w_i) \subseteq A\}$$

$$\widehat{Pl}^{\mathbb{Z}}_{\boldsymbol{y}}(A) = \frac{1}{N} \#\{i \in \{1, \ldots, N\} | \varphi_{\boldsymbol{y}}(\Gamma_{\boldsymbol{y}}(s_i), w_i) \cap A \neq \emptyset\}$$

Example: the urn model

- Here, $Y \sim \mathcal{B}(n, \theta)$. The likelihood-based BF is induced by a random interval

$$\Gamma(s) = \{\theta : pl_y(\theta) \geq s\} = [\underline{\theta}(s), \overline{\theta}(s)]$$

- We have

$$Z = \varphi(\theta, W) = \begin{cases} 1 & \text{if } W \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

- Consequently,

$$\varphi(\Gamma(s), W) = \varphi\left(\left[\underline{\theta}(s), \overline{\theta}(s)\right], W\right) = \begin{cases} \{1\} & \text{if } W \leq \underline{\theta}(s) \\ \{0\} & \text{if } \overline{\theta}(s) < W \\ \{0, 1\} & \text{otherwise} \end{cases}$$

Example: the urn model

Analytical formula

We have

$$m_y^{\mathbb{Z}}(\{1\}) = \mathbb{P}\left(\varphi\left(\Gamma(s), W\right) = \{1\}\right) = \hat{\theta} - \frac{\underline{B}(\hat{\theta}; y+1, n-y+1)}{\hat{\theta}^y (1-\hat{\theta})^{n-y}}$$

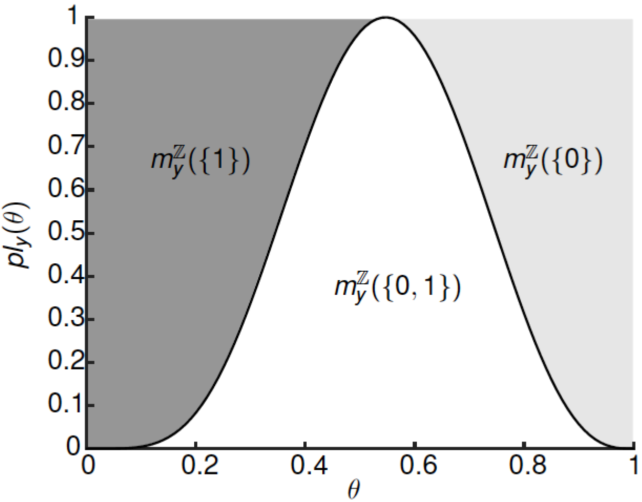
$$m_y^{\mathbb{Z}}(\{0\}) = \mathbb{P}\left(\varphi\left(\Gamma(s), W\right) = \{0\}\right) = 1 - \hat{\theta} - \frac{\underline{B}(1-\hat{\theta}; n-y+1, y+1)}{\hat{\theta}_j^y (1-\hat{\theta})^{n-y}}$$

$$m_y^{\mathbb{Z}}(\{0,1\}) = 1 - m_y^{\mathbb{Z}}(\{0\}) - m_y^{\mathbb{Z}}(\{1\})$$

where $\underline{B}(z; a, b) = \int_0^z t^{a-1}(1-t)^{b-1} dt$ is the incomplete beta function

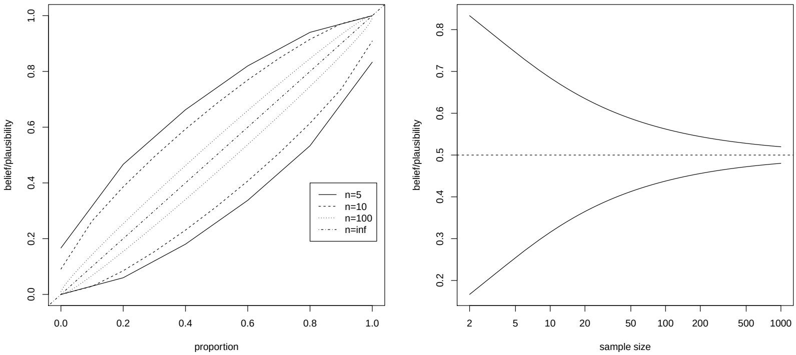
Example: the urn model

Geometric representation



Example: the urn model

Belief/plausibility intervals



Consistency

- Here, it is easy to show that

$$m_y^{\mathbb{Z}}(\{1\}) \xrightarrow{P} \theta_0 \quad \text{and} \quad m_y^{\mathbb{Z}}(\{0\}) \xrightarrow{P} 1 - \theta_0$$

  as $n \to \infty$, i.e., the predictive belief function converges to the true distribution of Z
- When the predictive belief function is induced by a random interval $[\underline{Z}, \overline{Z}]$, we can show that, under mild conditions,

$$\underline{Z} \xrightarrow{d} Z \quad \text{and} \quad \overline{Z} \xrightarrow{d} Z$$

- The consistency remains to be proved in the general case

Outline

Model

We consider the following standard regression model

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

- $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is the vector of n observations of the dependent variable
- X is the fixed design matrix of size $n \times (p+1)$
- $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)' \sim \mathcal{N}(\mathbf{0}, I_n)$ is the vector of errors
- The vector of coefficients is $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma)'$.

Likelihood-based belief function

- The likelihood function for this model is

$$L_{\boldsymbol{y}}(\boldsymbol{\theta}) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y} - X\beta)'(\boldsymbol{y} - X\beta)\right]$$

- The contour function can thus be readily calculated as

$$pl_{\boldsymbol{y}}(\boldsymbol{\theta}) = \frac{L_{\boldsymbol{y}}(\boldsymbol{\theta})}{L_{\boldsymbol{y}}(\widehat{\boldsymbol{\theta}})}$$

  with $\widehat{\boldsymbol{\theta}} = (\widehat{\beta}', \widehat{\sigma})'$, where
  - $\widehat{\beta} = (X'X)^{-1}X'\boldsymbol{y}$ is the ordinary least squares estimate of β
  - $\widehat{\sigma}$ is the standard deviation of residuals

Plausibility of linear hypotheses

- Assertions (hypotheses) H of the form $A\beta = \boldsymbol{q}$, where A is a $r \times (p+1)$ constant matrix and $\boldsymbol{q}$ is a constant vector of length r, for some $r \leq p+1$
- Special cases: $\{\beta_j = 0\}$, $\{\beta_j = 0, \forall j \in \{1, \ldots, p\}\}$, or $\{\beta_j = \beta_k\}$, etc.
- The plausibility of H is

$$Pl_{\boldsymbol{y}}^{\Theta}(H) = \sup_{A\beta = \boldsymbol{q}} pl_{\boldsymbol{y}}(\boldsymbol{\theta}) = \frac{L_{\boldsymbol{y}}(\widehat{\boldsymbol{\theta}}_*)}{L_{\boldsymbol{y}}(\widehat{\boldsymbol{\theta}})}$$

where $\widehat{\boldsymbol{\theta}}_* = (\widehat{\boldsymbol{\beta}}_*', \widehat{\sigma}_*)'$ (restricted LS estimates) with

$$\widehat{\boldsymbol{\beta}}_* = \widehat{\boldsymbol{\beta}} - (X'X)^{-1} A' [A(X'X)^{-1} A']^{-1} (A\widehat{\boldsymbol{\beta}} - \boldsymbol{q})$$

$$\widehat{\sigma}_* = \sqrt{(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_*)'(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_*)/n}$$

Linear model: prediction

- Let z be a not-yet observed value of the dependent variable for a vector $\boldsymbol{x}_0$ of covariates:

$$z = \boldsymbol{x}_0' \beta + \epsilon_0,$$

with $\epsilon_0 \sim \mathcal{N}(0, \sigma^2)$

- We can write, equivalently,

$$z = \boldsymbol{x}_0' \beta + \sigma \Phi^{-1}(w) = \varphi_{\boldsymbol{x}_0, \boldsymbol{y}}(\boldsymbol{\theta}, w),$$

where w has a standard uniform distribution

- The predictive belief function on z can then be approximated using Monte Carlo simulation

Linear model: prediction

- Let z be a not-yet observed value of the dependent variable for a vector $\boldsymbol{x}_0$ of covariates:

$$z = \boldsymbol{x}_0' \beta + \epsilon_0,$$

with $\epsilon_0 \sim \mathcal{N}(0, \sigma^2)$

- We can write, equivalently,

$$z = \boldsymbol{x}_0' \beta + \sigma \Phi^{-1}(w) = \varphi_{\boldsymbol{x}_0, \boldsymbol{y}}(\boldsymbol{\theta}, w),$$

where w has a standard uniform distribution

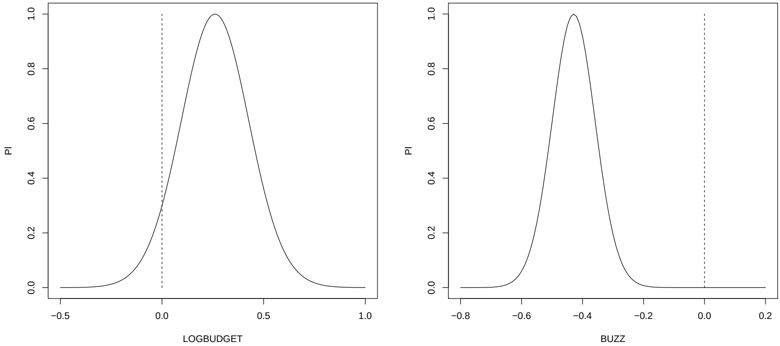
- The predictive belief function on z can then be approximated using Monte Carlo simulation

Outline

Example: movie Box office data

- Dataset about 62 movies released in 2009 (from Greene, 2012)
- Dependent variable: logarithm of Box Office receipts
- 11 covariates:
  - 3 dummy variables (G, PG, PG13) to encode the MPAA (Motion Picture Association of America) rating, logarithm of budget (LOGBUDGET), star power (STARPOWR),
  - a dummy variable to indicate if the movie is a sequel (SEQUEL),
  - four dummy variables to describe the genre (ACTION, COMEDY, ANIMATED, HORROR)
  - one variable to represent internet buzz (BUZZ)

Some marginal contour functions

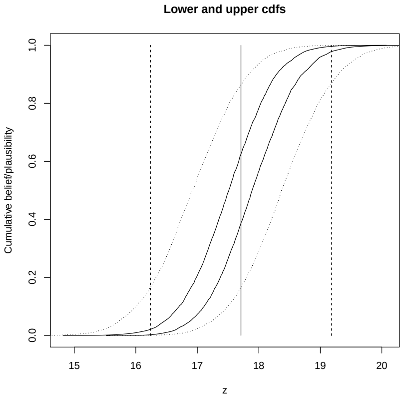


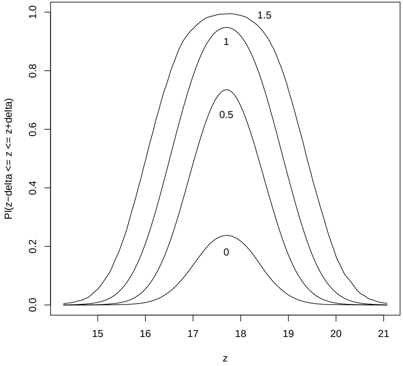
Regression coefficients

	Estimate	Std. Error	t-value	p-value	$Pl(\beta_j = 0)$
(Intercept)	15.400	0.643	23.960	$<$ 2e-16	1.0e-34
G	0.384	0.553	0.695	0.49	0.74
PG	0.534	0.300	1.780	0.081	0.15
PG13	0.215	0.219	0.983	0.33	0.55
LOGBUDGET	0.261	0.185	1.408	0.17	0.30
STARPOWR	4.32e-3	0.0128	0.337	0.74	0.93
SEQUEL	0.275	0.273	1.007	0.32	0.54
ACTION	-0.869	0.293	-2.964	4.7e-3	6.6e-3
COMEDY	-0.0162	0.256	-0.063	0.95	0.99
ANIMATED	-0.833	0.430	-1.937	0.058	0.11
HORROR	0.375	0.371	1.009	0.32	0.54
BUZZ	0.429	0.0784	5.473	1.4e-06	4.8e-07

Movie example

BO success of an action sequel film rated PG13 by MPAA, with LOGBUDGET=5.30, STARPOWER=23.62 and BUZZ= 2.81?





Outline

Ex ante forecasting

Problem and classical approach

- Consider the situation where some explanatory variables are unknown at the time of the forecast and have to be estimated or predicted
- Classical approach: assume that $\boldsymbol{x}_0$ has been estimated with some variance, which has to be taken into account in the calculation of the forecast variance
- According to Green (Econometric Analysis, 7th edition, 2012)
  - "*This vastly complicates the computation. Many authors view it as simply intractable*"
  - "*analytical results for the correct forecast variance remain to be derived except for simple special cases*"

Ex ante forecasting

Belief function approach

- In contrast, this problem can be handled very naturally in our approach by modeling partial knowledge of $\boldsymbol{x}_0$ by a belief function $Bel^{\mathbb{X}}$ in the sample space $\mathbb{X}$ of $\boldsymbol{x}_0$
- We then have

$$Bel_{\boldsymbol{y}}^{\mathbb{Z}} = \left(Bel_{\boldsymbol{y}}^{\Theta} \oplus Bel_{\boldsymbol{y}}^{\mathbb{Z}\times\Theta} \oplus Bel^{\mathbb{X}}\right)^{\downarrow \mathbb{Z}}$$

- Assume that the belief function $Bel^{\mathbb{X}}$ is induced by a source $(\Omega, \mathcal{A}, \mathbb{P}^{\Omega}, \Lambda)$, where Λ is a multi-valued mapping from Ω to $2^{\mathbb{X}}$
- The predictive belief function $Bel_{\boldsymbol{y}}^{\mathbb{Z}}$ is then induced by the multi-valued mapping

$$(\omega, s, w) \to \varphi_{\boldsymbol{y}}(\Lambda(\omega), \Gamma_{\boldsymbol{y}}(s), w)$$

- $Bel_{\boldsymbol{y}}^{\mathbb{Z}}$ can be approximated by Monte Carlo simulation

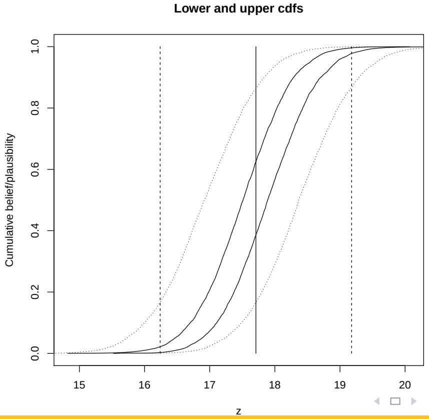
Monte Carlo algorithm

Require: Desired number of focal sets N
  for $i = 1$ **to** N **do**
    Draw (s_i, w_i) uniformly in $[0, 1]^2$
    Draw ω from $\mathbb{P}^\Omega$
    Search for $z_{*i} = \min_{\boldsymbol{\theta}} \varphi_{\boldsymbol{y}}(\boldsymbol{x}_0, \boldsymbol{\theta}, w_i)$ such that $pl_{\boldsymbol{y}}(\boldsymbol{\theta}) \geq s_i$ and $\boldsymbol{x}_0 \in \Lambda(\omega)$
    Search for $z_i^* = \max_{\boldsymbol{\theta}} \varphi_{\boldsymbol{y}}(\boldsymbol{x}_0, \boldsymbol{\theta}, w_i)$ such that $pl_{\boldsymbol{y}}(\boldsymbol{\theta}) \geq s_i$ and $\boldsymbol{x}_0 \in \Lambda(\omega)$
    $B_i \leftarrow [z_{*i}, z_i^*]$
  end for

Movie example

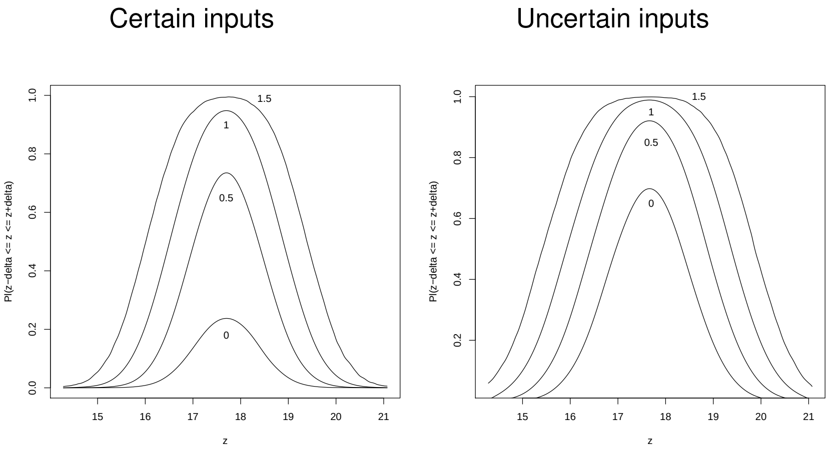
Lower and upper cdfs

BO success of an action sequel film rated PG13 by MPAA, with LOGBUDGET=5.30, STARPOWER=23.62 and BUZZ= (0,2.81,5) (triangular possibility distribution)?



Movie example

Pl-plots



Summary

- Developing pratical applications using the Dempster-Shafer framework requires modeling expert knowledge and statistical information using belief functions
- Systematic and principled methods now exist
    - Least-commitment principle
    - GBT
    - Likelihood-based belief functions
    - Predictive belief function
    - 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`

Ph. Smets.
Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem.
International Journal of Approximate Reasoning, 9:1-35, 1993.

G. J. Klir and M. J. Wierman.
Uncertainty-based information. Elements of generalized information theory.
Springer-Verlag, Studies in Fuzziness and Soft Computing, Vol. 15, 1999.

T. Denoeux.
Maximum likelihood estimation from Uncertain Data in the Belief Function Framework.
IEEE Transactions on Knowledge and Data Engineering, 25(1):119-130, 2013.

References II

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

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

O. Kanjanatarakul, S. Sriboonchitta and T. Denœux
Forecasting using belief functions. An application to marketing econometrics.
International Journal of Approximate Reasoning, 55(5):1113–1128, 2014.