Application of belief functions to statistical inference

Thierry Denœux

Université de Technologie de Compiègne
HEUDIASYC (UMR CNRS 6599)
http://www.hds.utc.fr/~tdenoeux

Beijing University of Technology
Beijing, China
July, 2016

Outline

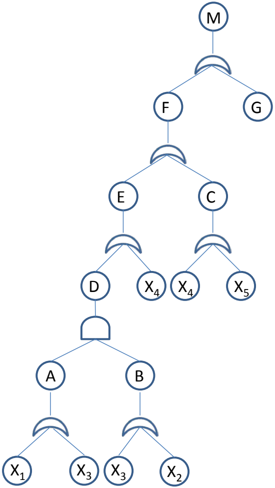
1. Complements on belief functions
   - Belief functions on product spaces
   - Belief functions on infinite spaces
2. Statistical estimation and prediction
   - Likelihood-based belief function
   - Predictive belief function
   - Some theoretical results
3. Applications
   - Linear regression
   - Innovation diffusion forecasting

Outline

1. Complements on belief functions
   - Belief functions on product spaces
   - Belief functions on infinite spaces
2. Statistical estimation and prediction
   - Likelihood-based belief function
   - Predictive belief function
   - Some theoretical results
3. Applications
   - Linear regression
   - Innovation diffusion forecasting

Belief functions on product spaces

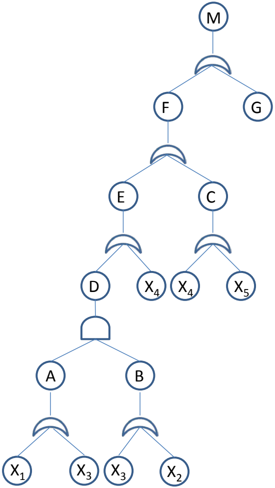
Motivation

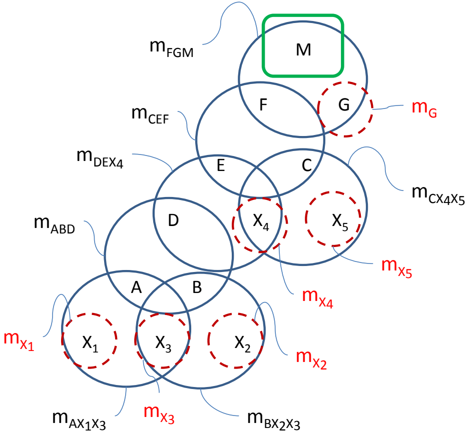


- In many applications, we need to express uncertain information about several variables taking values in different domains
- Example: fault tree (logical relations between Boolean variables and probabilistic or evidential information about elementary events)

Fault tree example

(Dempster & Kong, 1988)





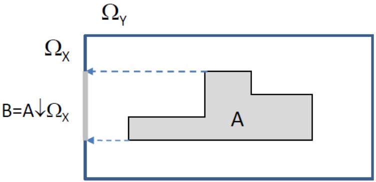
Hypergraph

Multidimensional belief functions

Marginalization, vacuous extension

- Let X and Y be two variables defined on frames Ω_X and Ω_Y
- Let $\Omega_{XY} = \Omega_X \times \Omega_Y$ be the product frame
- A mass function m^{XY} on Ω_{XY} can be seen as an generalized relation between variables X and Y
- Two basic operations on product frames
  1. Express a joint mass function m^{XY} in the coarser frame Ω_X or Ω_Y (marginalization)
  2. Express a marginal mass function m^X on Ω_X in the finer frame Ω_{XY} (vacuous extension)

Marginalization



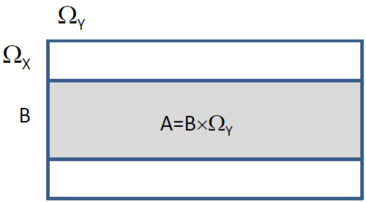
- Problem: express m^{XY} in Ω_X
- Solution: transfer each mass $m^{XY}(A)$ to the projection of A on Ω_X

- 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



- Problem: express m^X in Ω_{XY}
- Solution: transfer each mass $m^X(B)$ to the cylindrical extension of B: $B \times \Omega_Y$

- Vacuous extension:

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

Operations in product frames

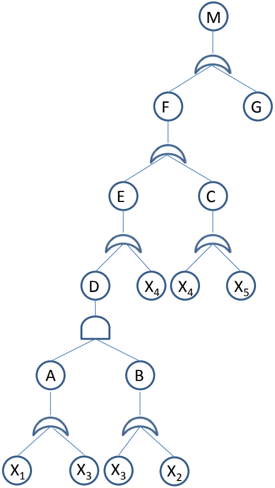
Application to approximate reasoning

- 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:

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

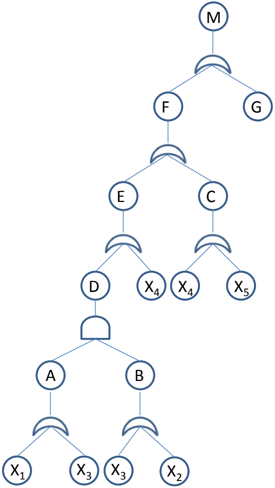
- Infeasible with many variables and large frames of discernment, but efficient algorithms exist to carry out the operations in frames of minimal dimensions

Fault tree example



Cause	$m(\{1\})$	$m(\{0\})$	$m(\{0,1\})$
X_1	0.05	0.90	0.05
X_2	0.05	0.90	0.05
X_3	0.005	0.99	0.005
X_4	0.01	0.985	0.005
X_5	0.002	0.995	0.003
G	0.001	0.99	0.009
M	0.02	0.951	0.029
F	0.019	0.961	0.02

Fault tree example (continued)



Cause	$m(\{1\})$	$m(\{0\})$	$m(\{0,1\})$
M	1	0	0
G	0.197	0.796	0.007
F	0.800	0.196	0.004
⋮	⋮	⋮	⋮
X_1	0.236	0.724	0.040
X_2	0.236	0.724	0.040
X_3	0.200	0.796	0.004
X_4	0.302	0.694	0.004
X_5	0.099	0.898	0.003

Outline

1. Complements on belief functions
   - Belief functions on product spaces
   - Belief functions on infinite spaces
2. Statistical estimation and prediction
   - Likelihood-based belief function
   - Predictive belief function
   - Some theoretical results
3. Applications
   - Linear regression
   - Innovation diffusion forecasting

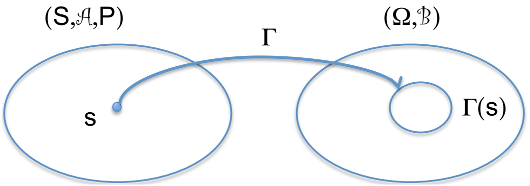
Belief function: general definition

- Let Ω be a set (finite or not) and $\mathcal{B}$ be an algebra of subsets of Ω
- A belief function (BF) on $\mathcal{B}$ is a mapping $Bel : \mathcal{B} \rightarrow [0,1]$ verifying $Bel(\emptyset) = 0$, $Bel(\Omega) = 1$ and the complete monotonicity property: for any $k \geq 2$ and any collection $B_1, \ldots, B_k$ of elements of $\mathcal{B}$,

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

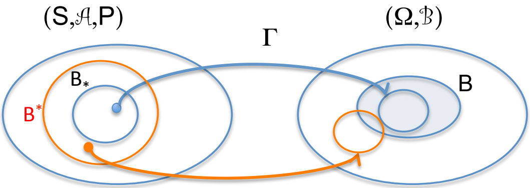
- A function $Pl : \mathcal{B} \rightarrow [0,1]$ is a plausibility function iff $B \rightarrow 1 - Pl(\overline{B})$ is a belief function

Source



- Let S be a state space, $\mathcal{A}$ an algebra of subsets of S, $\mathbb{P}$ a finitely additive probability on $(S, \mathcal{A})$
- Let Ω be a set and $\mathcal{B}$ an algebra of subsets of Ω
- Γ a multivalued mapping from S to $2^\Omega \setminus \{\emptyset\}$
- The four-tuple $(S, \mathcal{A}, \mathbb{P}, \Gamma)$ is called a source
- Under some conditions, it induces a belief function on $(\Omega, \mathcal{B})$

Strong measurability



- Lower and upper inverses: for all $B \in \mathcal{B}$,

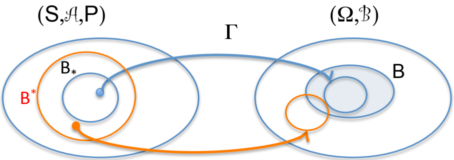
$$\Gamma_*(B) = B_* = \{s \in S | \Gamma(s) \neq \emptyset, \Gamma(s) \subseteq B\}$$

$$\Gamma^*(B) = B^* = \{s \in S | \Gamma(s) \cap B \neq \emptyset\}$$

- Γ is strongly measurable wrt $\mathcal{A}$ and $\mathcal{B}$ if, for all $B \in \mathcal{B}$, $B^* \in \mathcal{A}$
- $(\forall B \in \mathcal{B}, B^* \in \mathcal{A}) \Leftrightarrow (\forall B \in \mathcal{B}, B_* \in \mathcal{A})$

Belief function induced by a source

Lower and upper probabilities

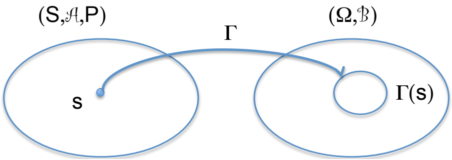


- Lower and upper probabilities:

$$\forall B \in \mathcal{B}, \quad \mathbb{P}_*(B) = \frac{\mathbb{P}(B_*)}{\mathbb{P}(\Omega^*)}, \quad \mathbb{P}^*(B) = \frac{\mathbb{P}(B^*)}{\mathbb{P}(\Omega^*)} = 1 - Bel(\overline{B})$$

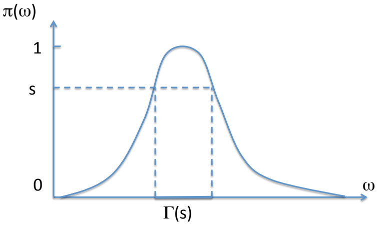
- $\mathbb{P}_*$ is a BF, and $\mathbb{P}^*$ is the dual plausibility function
- Conversely, for any belief function, there is a source that induces it (Shafer's thesis, 1973)

Interpretation



- Typically, Ω is the domain of an unknown quantity ω, and S is a set of interpretations of a given piece of evidence about ω
- If $s \in S$ holds, then the evidence tells us that $\omega \in \Gamma(s)$, and nothing more
- Then
  - $Bel(B)$ is the probability that the evidence supports B
  - $Pl(B)$ is the probability that the evidence is consistent with B

Consonant belief function



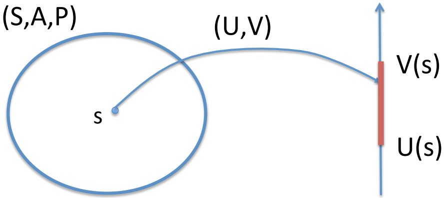
- Let π be a mapping from Ω to $S = [0, 1]$ s.t. $\sup \pi = 1$
- Let Γ be the multi-valued mapping from S to 2^Ω defined by

$$\forall s \in [0, 1], \quad \Gamma(s) = \{\omega \in \Omega | \pi(\omega) \geq s\}$$

- The source $(S, \mathcal{B}(S), \lambda, \Gamma)$ defines a consonant BF on Ω, such that $pl(\omega) = \pi(\omega)$ (contour function)
- The corresponding plausibility function is a possibility measure

$$\forall B \subseteq \Omega, \quad Pl(B) = \sup_{\omega \in B} pl(\omega)$$

Random closed interval

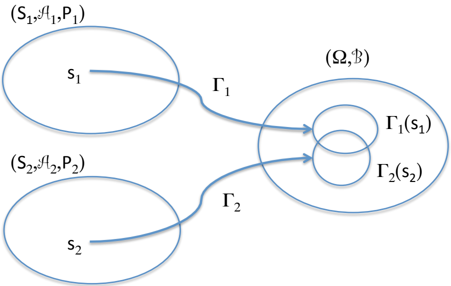


- Let (U, V) be a bi-dimensional random vector from a probability space $(S, \mathcal{A}, \mathbb{P})$ to $\mathbb{R}^2$ such that $U \leq V$ a.s.
- Multi-valued mapping:

$$\Gamma : s \rightarrow \Gamma(s) = [U(s), V(s)]$$

- The source $(S, \mathcal{A}, \mathbb{P}, \Gamma)$ is a random closed interval. It defines a BF on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$

Dempster's rule



- Let $(S_i, \mathcal{A}_i, \mathbb{P}_i, \Gamma_i)$, $i = 1, 2$ be two sources representing independent items of evidence, inducing BF Bel_1 and Bel_2
- The combined BF $Bel = Bel_1 \oplus Bel_2$ is induced by the source $(S_1 \times S_2, \mathcal{A}_1 \otimes \mathcal{A}_2, \mathbb{P}_1 \otimes \mathbb{P}_2, \Gamma_\cap)$ with

$$\Gamma_\cap(s_1, s_2) = \Gamma_1(s_1) \cap \Gamma_2(s_2)$$

Approximate computation

Monte Carlo simulation

Require: Desired number of focal sets N

$i \leftarrow 0$

while $i < N$ **do**

  Draw s_1 in S_1 from $\mathbb{P}_1$

  Draw s_2 in S_2 from $\mathbb{P}_2$

  $\Gamma_\cap(s_1, s_2) \leftarrow \Gamma_1(s_1) \cap \Gamma_2(s_2)$

  if $\Gamma_\cap(s_1, s_2) \neq \emptyset$ **then**

    $i \leftarrow i + 1$

    $B_i \leftarrow \Gamma_\cap(s_1, s_2)$

  end if

end while

$\widehat{Bel}(B) \leftarrow \frac{1}{N}\#\{i \in \{1, \ldots, N\} | B_i \subseteq B\}$

$\widehat{Pl}(B) \leftarrow \frac{1}{N}\#\{i \in \{1, \ldots, N\} | B_i \cap B \neq \emptyset\}$

Outline

1. Complements on belief functions
   - Belief functions on product spaces
   - Belief functions on infinite spaces

2. Statistical estimation and prediction
   - Likelihood-based belief function
   - Predictive belief function
   - Some theoretical results

3. Applications
   - Linear regression
   - Innovation diffusion forecasting

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 statistical inference (frequentist and Bayesian) is fully satisfactory, from a conceptual point of view
- Proposal of a new approach based on belief functions
- The new approach boils down to Bayesian inference 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 $W \sim \mathcal{U}([0, 1])$,

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

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 (estimation/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

1. Complements on belief functions
   - Belief functions on product spaces
   - Belief functions on infinite spaces

2. Statistical estimation and prediction
   - Likelihood-based belief function
   - Predictive belief function
   - Some theoretical results

3. Applications
   - Linear regression
   - Innovation diffusion forecasting

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

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

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)

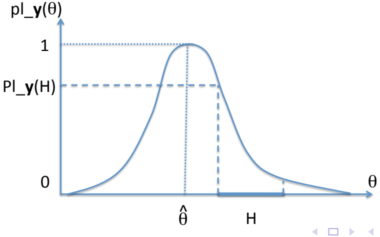
- $Bel_{\boldsymbol{y}}^{\Theta}$ is the consonant belief function induced by the relative likelihood function

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

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

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

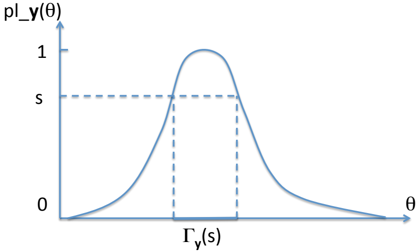


Source

- Corresponding random set:

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

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



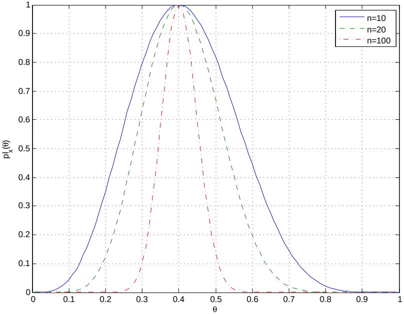
- If $\Theta \subseteq \mathbb{R}$ and if $L_{\boldsymbol{y}}(\theta)$ is unimodal and upper-semicontinuous, then $Bel_{\boldsymbol{y}}^{\Theta}$ corresponds to a random closed interval

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

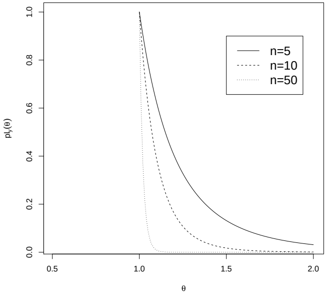


Uniform example

- Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be a realization from an iid random sample from $\mathcal{U}([0, \theta])$
- The likelihood function is

$$L_{\boldsymbol{y}}(\theta) = \theta^{-n} \mathbb{1}_{[y_{(n)}, +\infty)}(\theta)$$

- The predictive BF is induced by the random closed interval $[y_{(n)}, y_{(n)} S^{-1/n}]$, with $S \sim \mathcal{U}([0, 1])$



Profile likelihood

- Assume that $\theta = (\xi, \nu)$, where ξ is a parameter of interest and ν is a nuisance parameter
- Then, the marginal contour function for ξ is

$$pl_{\mathbf{y}}(\xi) = \sup_{\nu} pl_{\mathbf{y}}(\xi, \nu),$$

  which is the profile relative likelihood function
- The profiling method for eliminating nuisance parameter thus has a natural justification in our approach
- When the quantities $pl_{\mathbf{y}}(\xi)$ cannot be derived analytically, they have to be computed numerically using an iterative optimization algorithm

Relation with likelihood-based inference

- The approach to statistical inference outlined in the previous section is very close to the "likelihoodist" approach advocated by Birnbaum (1962), Barnard (1962), and Edwards (1992), among others
- The main difference resides in the interpretation of the likelihood function as defining a belief function
- This interpretation allows us to quantify the uncertainty in statements of the form $\theta \in H$, where H may contain multiple values. This is in contrast with the classical likelihood approach, in which only the likelihood of single hypotheses is defined
- The belief function interpretation provides an easy and natural way to combine statistical information with other information, such as expert judgements

Relation with the likelihood-ratio test statistics

- We can also notice that $Pl_{\boldsymbol{y}}^{\Theta}(H)$ is identical to the likelihood ratio statistic for H
- From Wilk's theorem, we have asymptotically (under regularity conditions), when H holds,

$$-2 \ln Pl_{\boldsymbol{y}}(H) \sim \chi_r^2$$

  where r is the number of restrictions imposed by H
- Consequently, rejecting hypothesis H if its plausibility is smaller than $\exp(-\chi^2_{r;1-\alpha}/2)$ is a testing procedure with significance level approximately equal to α
- However, we consider these properties are incidental, as the approach outlined here is not based on frequentist inference

Outline

1 Complements on belief functions
- Belief functions on product spaces
- Belief functions on infinite spaces

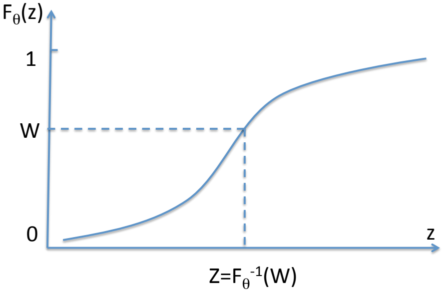
2 Statistical estimation and prediction
- Likelihood-based belief function
- Predictive belief function
- Some theoretical results

3 Applications
- Linear regression
- Innovation diffusion forecasting

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



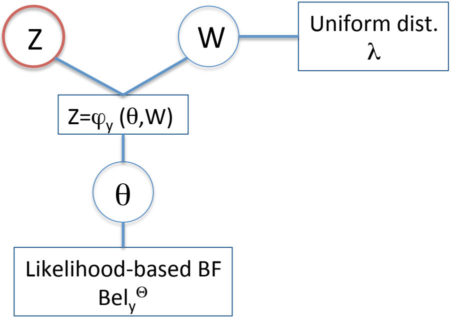
We can always write Z as a function of $\boldsymbol{\theta}$ and W as

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

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

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

Main result

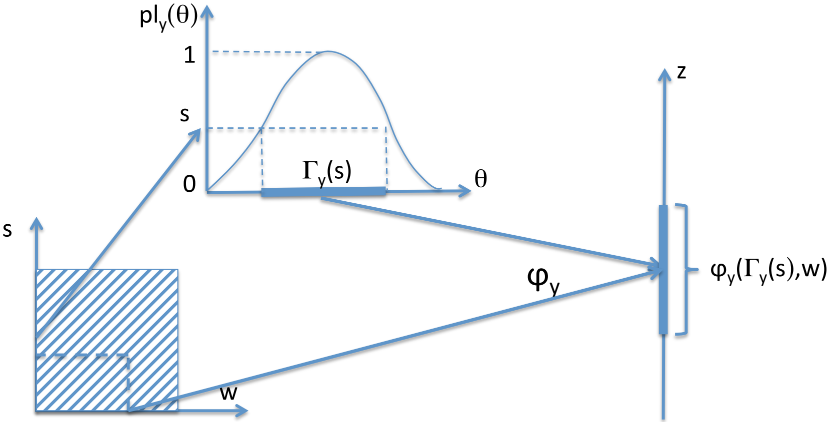


After combination by Dempster's rule and marginalization on $\mathbb{Z}$, we obtain the predictive BF on Z induced by the multi-valued mapping

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

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

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

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

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,

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

Example: the urn model

Analytical formula

We have

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

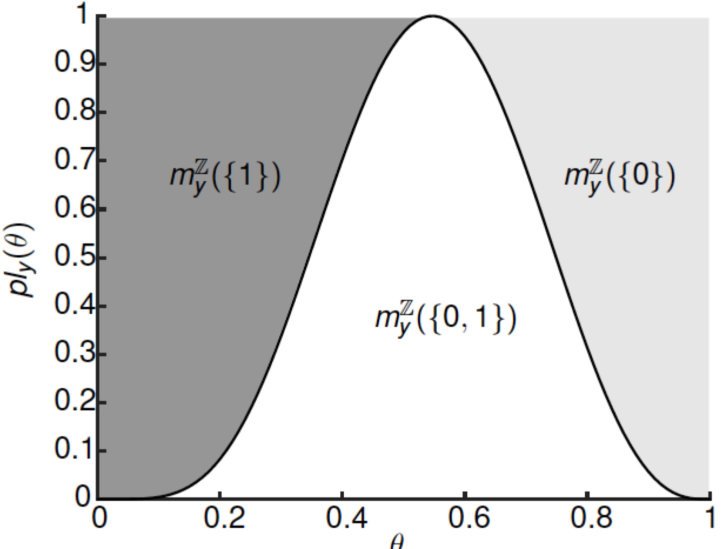
$$m_y^{\mathbb{Z}}(\{0\}) = \mathbb{P}\left(\varphi\left(\Gamma(s), W\right) = \{0\}\right) = 1 - \hat{\theta} - \frac{\underline{B}(1-\hat{\theta}; n-y+1, y+1)}{\hat{\theta}_j^y (1-\hat{\theta})^{n-y}}$$

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

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

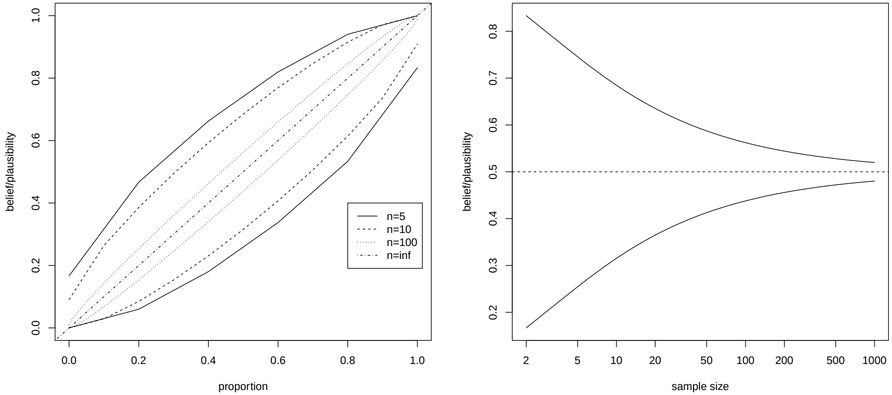
Example: the urn model

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,

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

Uniform example

- Assume that $Y_1, \ldots, Y_n, Z$ is iid from $\mathcal{U}([0,\theta])$
- Then $F_\theta(z) = z/\theta$ for all $0 \leq z \leq \theta$ and we can write $Z = \theta W$ with $W \sim \mathcal{U}([0,1])$
- We have seen that the belief function $Bel_{\boldsymbol{y}}^{\Theta}$ after observing $\boldsymbol{Y} = \boldsymbol{y}$ is induced by the random interval $[y_{(n)}, y_{(n)} S^{-1/n}]$
- Each focal set of $Bel_{\boldsymbol{y}}^{\mathbb{Z}}$ is an interval

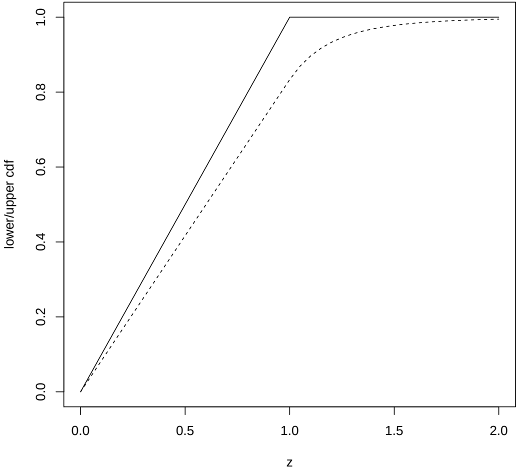
$$\varphi(\Gamma_{\boldsymbol{y}}(s), w) = [y_{(n)} w, y_{(n)} s^{-1/n} w]$$

- The predictive belief function $Bel_{\boldsymbol{y}}^{\mathbb{Z}}$ is induced by the random interval

$$[\widehat{Z}_{\boldsymbol{y}*}, \widehat{Z}_{\boldsymbol{y}}^*] = [y_{(n)} W, y_{(n)} S^{-1/n} W]$$

Uniform example

Lower and upper cdfs



Uniform example

Consistency

- From the consistency of the MLE, $Y_{(n)}$ converges in probability to θ_0, so

$$\widehat{Z}_{\boldsymbol{Y}*} = Y_{(n)} W \xrightarrow{d} \theta_0 W = Z$$

- We have $\mathbb{E}(S^{-1/n}) = n/(n-1)$, and

$$\text{Var}(S^{-1/n}) = \frac{n}{(n-2)(n-1)^2}$$

- Consequently, $\mathbb{E}(S^{-1/n}) \to 1$ and $\text{Var}(S^{-1/n}) \to 0$, so $S^{-1/n} \xrightarrow{P} 1$
- Hence,

$$\widehat{Z}^*_{\boldsymbol{Y}} = Y_{(n)} S^{-1/n} W \xrightarrow{d} \theta_0 W = Z$$

Outline

1. Complements on belief functions
   - Belief functions on product spaces
   - Belief functions on infinite spaces
2. Statistical estimation and prediction
   - Likelihood-based belief function
   - Predictive belief function
   - Some theoretical results
3. Applications
   - Linear regression
   - Innovation diffusion forecasting

Consistency of the likelihood-based belief function

- Assume that the observed data $\boldsymbol{y} = (y_1, \ldots, y_n)$ is a realization of an iid sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ from $Y \sim f_{\boldsymbol{\theta}}(y)$
- From Fraser (1968):

Theorem

If $\mathbb{E}_{\boldsymbol{\theta}_0}[\log f_{\boldsymbol{\theta}}(Y)]$ exists, is finite for all $\boldsymbol{\theta}$, and has a unique maximum at $\boldsymbol{\theta}_0$, then, for any $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, $pl_n(\boldsymbol{\theta}) \rightarrow 0$ almost surely under the law determined by $\boldsymbol{\theta}_0$

Consistency of the likelihood-based belief function (continued)

- The property $pl_n(\theta_0) \to 1$ a.s. does not hold in general (under regularity assumptions, $-2 \log pl_n(\theta_0)$ converges in distribution to χ^2_1)
- But we have the following theorem:

Theorem

Under some assumptions (Fraser, 1968), for any neighborhood N of θ_0, $Bel_n^{\Theta}(N) \to 1$ and $Pl_n^{\Theta}(N) \to 1$ almost surely under the law determined by θ_0

Consistency of the predictive belief function

- Assume that
  - The observed data $\boldsymbol{y} = (y_1, \ldots, y_n)$ is a realization of an iid sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$
  - The likelihood function $L_n(\boldsymbol{\theta})$ is unimodal and upper-semicontinuous, so that its level sets $\Gamma_n(s)$ are closed and connected, and that function $\varphi(\boldsymbol{\theta}, w)$ is continuous
- Under these conditions, the random set $\varphi(\Gamma_n(S), W)$ is a closed random interval $[\widehat{Z}_{*n}, \widehat{Z}^*_n]$
- Then:

Theorem

Assume that the conditions of the previous theorem hold, and that the predictive belief function $Bel_n^{\mathbb{Z}}$ is induced by a random closed interval $[\widehat{Z}_{*n}, \widehat{Z}^*_n]$. Then $\widehat{Z}_{*n}$ and $\widehat{Z}^*_n$ both converge in distribution to Z when n tends to infinity.

Outline

1. Complements on belief functions
   - Belief functions on product spaces
   - Belief functions on infinite spaces
2. Statistical estimation and prediction
   - Likelihood-based belief function
   - Predictive belief function
   - Some theoretical results
3. Applications
   - Linear regression
   - Innovation diffusion forecasting

Outline

1. Complements on belief functions
   - Belief functions on product spaces
   - Belief functions on infinite spaces
2. Statistical estimation and prediction
   - Likelihood-based belief function
   - Predictive belief function
   - Some theoretical results
3. Applications
   - Linear regression
   - Innovation diffusion forecasting

Model

We consider the following standard regression model

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

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 $\boldsymbol{\theta} = (\beta', \sigma)'$

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 $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}', \widehat{\sigma})'$, where
  - $\widehat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y}$ 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

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

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

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

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

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

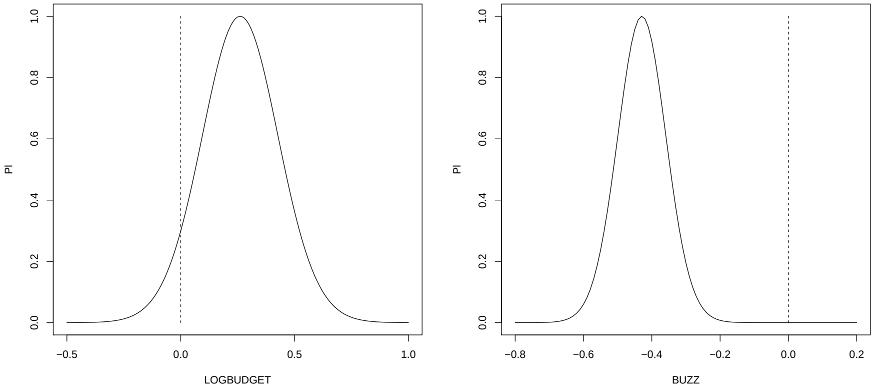
Linear model: prediction

- Let z be a not-yet observed value of the dependent variable for a vector $\boldsymbol{x}_0$ of covariates:
$$z = \boldsymbol{x}_0'\boldsymbol{\beta} + \epsilon_0,$$
with $\epsilon_0 \sim \mathcal{N}(0, \sigma^2)$
- We can write, equivalently,
$$z = \boldsymbol{x}_0'\boldsymbol{\beta} + \sigma\Phi^{-1}(w) = \varphi_{\boldsymbol{x}_0,\boldsymbol{y}}(\boldsymbol{\theta}, w),$$
where w has a standard uniform distribution
- The predictive belief function on z can then be approximated using Monte Carlo simulation

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

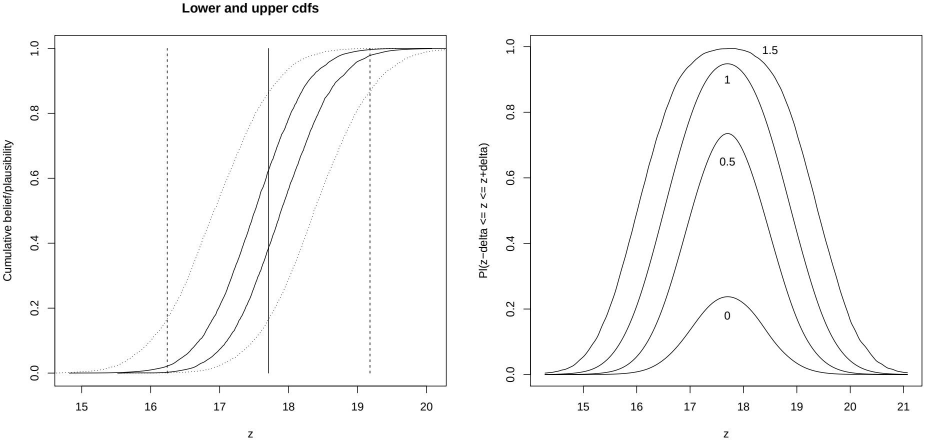


Regression coefficients

	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?



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
$$(\omega, s, w) \to \varphi_y(\Lambda(\omega), \Gamma_{\boldsymbol{y}}(s), w)$$
- $Bel_{\boldsymbol{y}}^{\mathbb{Z}}$ 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 $\boldsymbol{x}_0 \in \Lambda(\omega)$

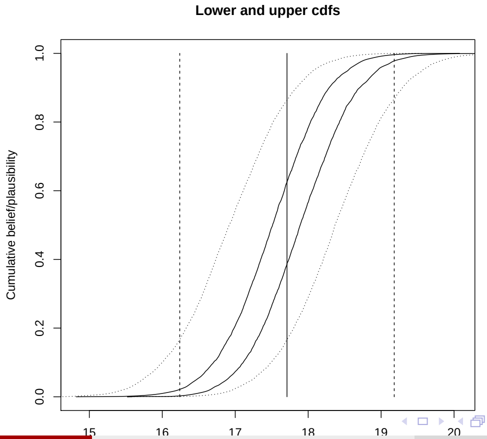
$B_i \leftarrow [z_{*i}, z_i^*]$

end for

Movie example

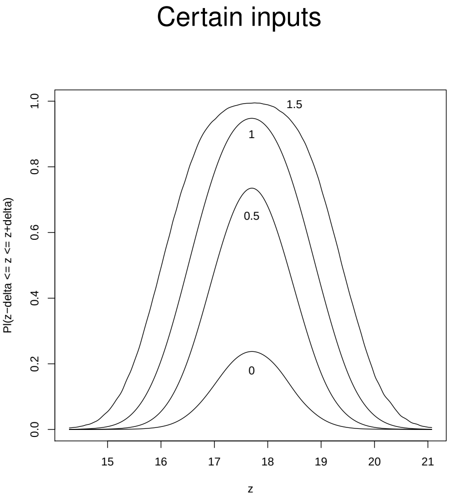
Lower and upper cdfs

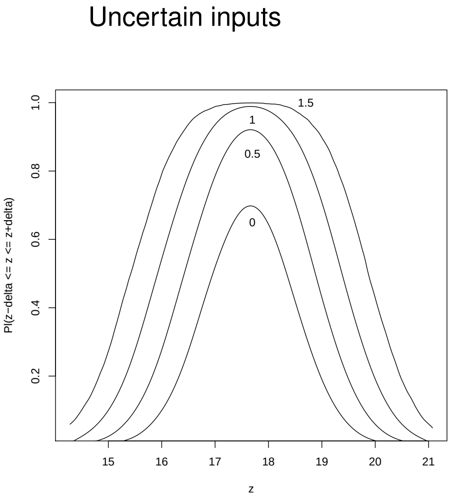
BO success of an action sequel film rated PG13 by MPAA, with LOGBUDGET=5.30, STARPOWER=23.62 and BUZZ= (0,2.81,5) (triangular possibility distribution)?



Movie example

Pl-plots





Outline

1 Complements on belief functions
- Belief functions on product spaces
- Belief functions on infinite spaces

2 Statistical estimation and prediction
- Likelihood-based belief function
- Predictive belief function
- Some theoretical results

3 Applications
- Linear regression
- Innovation diffusion forecasting

Innovation diffusion

- Forecasting the diffusion of an innovation has been a topic of considerable interest in marketing research
- Typically, when a new product is launched, sale forecasts have to be based on little data and uncertainty has to be quantified to avoid making wrong business decisions based on unreliable forecasts
- Our approach uses the Bass model (Bass, 1969) for innovation diffusion together with past sales data to quantify the uncertainty on future sales using the formalism of belief functions

Bass model

- Fundamental assumption (Bass, 1969): for eventual adopters, the probability $f(t)$ of purchase at time t, given that no purchase has yet been made, is an affine function of the number of previous buyers

$$\frac{f(t)}{1-F(t)} = p + qF(t)$$

  where p is a coefficient of innovation, q is a coefficient of imitation and $F(t) = \int_0^t f(u)du$.

- Solving this differential equation, the probability that an individual taken at random from the population will buy the product before time t is

$$\Phi_\theta(t) = cF(t) = \frac{c(1-\exp[-(p+q)t])}{1+(p/q)\exp[-(p+q)t]}$$

  where c is the probability of eventually adopting the product and $\theta = (p, q, c)$

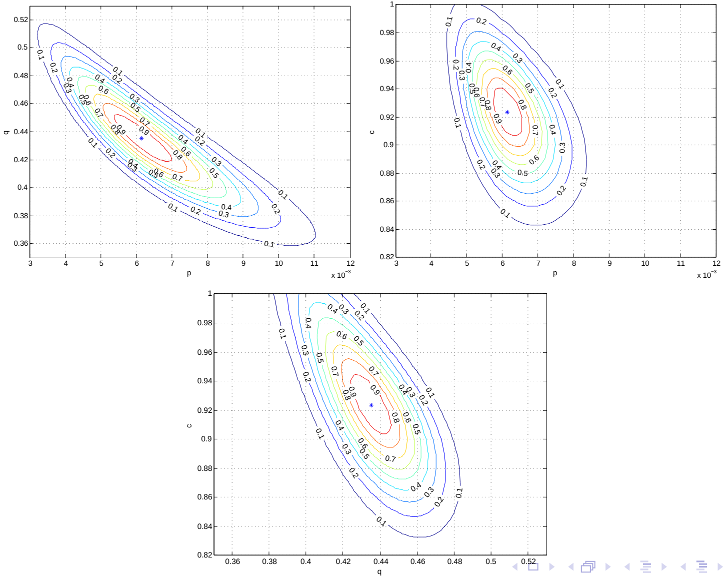
Parameter estimation

- Data: $y_1, \ldots, y_{T-1}$, where y_i = observed number of adopters in time interval $[t_{i-1}, t_i)$
- The number of individuals in the sample of size M who did not adopt the product at time t_{T-1} is $y_T = M - \sum_{i=1}^{T-1} y_i$
- The probability of adopting the innovation between times t_{i-1} and t_i is $p_i = \Phi_\theta(t_i) - \Phi_\theta(t_{i-1})$ for $1 \leq i \leq T-1$, and the probability of not adopting the innovation before t_{T-1} is $p_T = 1 - \Phi_\theta(t_{T-1})$
- Consequently, $\boldsymbol{x} = (x_1, \ldots, x_T)$ is a realization of $\boldsymbol{X} \sim \mathcal{M}(M, p_1, \ldots, p_T)$ and the likelihood function is

$$L_{\boldsymbol{y}}(\theta) \propto \prod_{i=1}^{T} p_i^{y_i} = \left(\prod_{i=1}^{T-1} [\Phi_\theta(t_i) - \Phi_\theta(t_{i-1})]^{y_i} \right) [1 - \Phi_\theta(t_{T-1})]^{y_T}$$

- The belief function on θ is defined by $pl_{\boldsymbol{y}}(\theta) = L_{\boldsymbol{y}}(\theta) / L_{\boldsymbol{y}}(\widehat{\theta})$

Results



Sales forecasting

- Let us assume we are at time t_{T-1} and we wish to forecast the number Z of sales between times τ_1 and τ_2, with $t_{T-1} \leq \tau_1 < \tau_2$
- Z has a binomial distribution $\mathcal{B}(Q, \pi_\theta)$, where
  - Q is the number of potential adopters at time $T-1$
  - π_θ is the probability of purchase for an individual in $[\tau_1, \tau_2]$, given that no purchase has been made before t_{T-1}

$$\pi_\theta = \frac{\Phi_\theta(\tau_2) - \Phi_\theta(\tau_1)}{1 - \Phi_\theta(t_{T-1})}$$

- Z can be written as $Z = \varphi(\theta, \boldsymbol{W}) = \sum_{i=1}^{Q} \mathbb{1}_{[0,\pi_\theta]}(W_i)$ where

$$\mathbb{1}_{[0,\pi_\theta]}(W_i) = \begin{cases} 1 & \text{if } W_i \leq \pi_\theta \\ 0 & \text{otherwise} \end{cases}$$

and $\boldsymbol{W} = (W_1, \ldots, W_Q)$ has a uniform distribution in $[0,1]^Q$.

Predictive belief function

Multi-valued mapping

- The predictive belief function on Z is induced by the multi-valued mapping $(s, \boldsymbol{w}) \to \varphi(\Gamma_{\boldsymbol{y}}(s), \boldsymbol{w})$ with

$$\Gamma_{\boldsymbol{y}}(s) = \{\theta \in \Theta : pl_{\boldsymbol{y}}(\theta) \geq s\}$$

- When θ varies in $\Gamma_{\boldsymbol{y}}(s)$, the range of π_θ is $[\underline{\pi}_\theta(s), \overline{\pi}_\theta(s)]$, with

$$\underline{\pi}_\theta(s) = \min_{\{\theta | pl_{\boldsymbol{y}}(\theta) \geq s\}} \pi_\theta, \quad \overline{\pi}_\theta(s) = \max_{\{\theta | pl_{\boldsymbol{y}}(\theta) \geq s\}} \pi_\theta$$

- We have

$$\varphi(\Gamma_{\boldsymbol{y}}(s), \boldsymbol{w}) = [\underline{Z}(s, \boldsymbol{w}), \overline{Z}(s, \boldsymbol{w})],$$

  where $\underline{Z}(s, \boldsymbol{w})$ and $\overline{Z}(s, \boldsymbol{w})$ are, respectively, the number of w_i's that are less than $\underline{\pi}_\theta(s)$ and $\overline{\pi}_\theta(s)$

- For fixed s, $\underline{Z}(s, \boldsymbol{W}) \sim \mathcal{B}(Q, \underline{\pi}_\theta(s))$ and $\overline{Z}(s, \boldsymbol{W}) \sim \mathcal{B}(Q, \overline{\pi}_\theta(s))$

Predictive belief function

Calculation

- The belief and plausibilities that Z will be less than z are

$$Bel_{\mathbf{y}}^{\mathbb{Z}}([0,z]) = \int_0^1 F_{Q,\underline{\pi}_\theta(s)}(z)ds$$

$$Pl_{\mathbf{y}}^{\mathbb{Z}}([0,z]) = \int_0^1 F_{Q,\overline{\pi}_\theta(s)}(z)ds$$

where $F_{Q,p}$ denotes the cdf of the binomial distribution $\mathcal{B}(Q,p)$

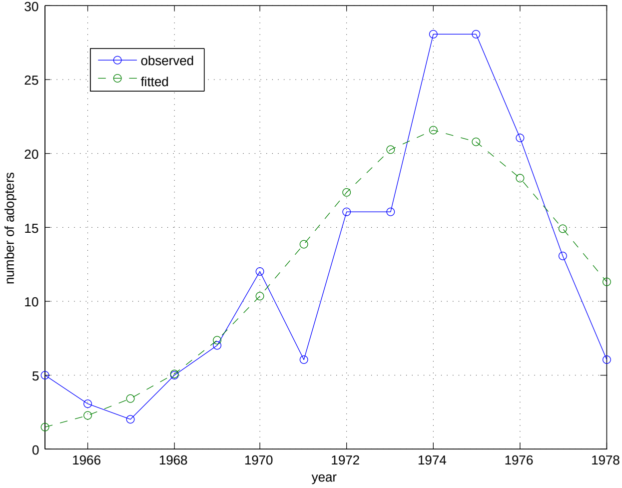
- The contour function of Z is

$$pl_{\mathbf{y}}(z) = \int_0^1 \left(F_{Q,\underline{\pi}_\theta(s)}(z) - F_{Q,\overline{\pi}_\theta(s)}(z-1)\right) ds$$

- Theses integrals can be approximated by Monte-Carlo simulation

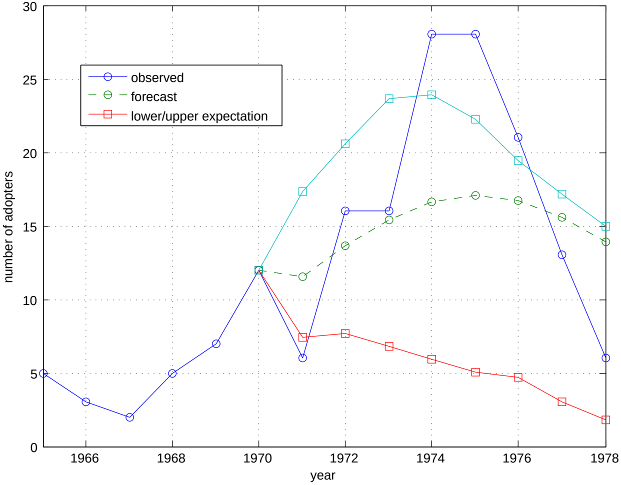
Ultrasound data

Data collected from 209 hospitals through the U.S.A. (Schmittlein and Mahajan, 1982) about adoption of an ultrasound equipment



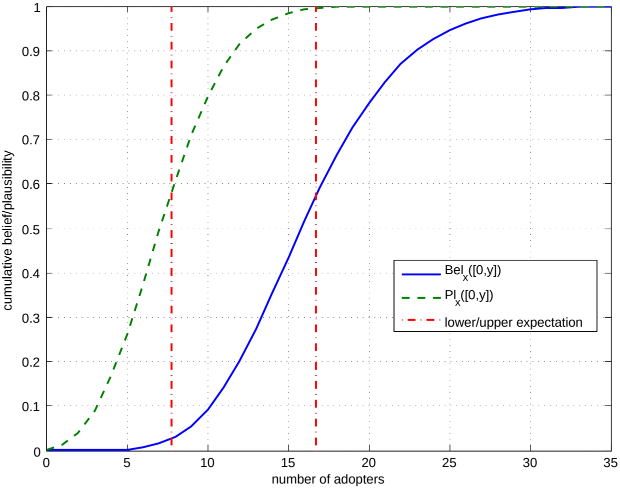
Forecasting

Predictions made in 1970 for the number of adopters in the period 1971-1978, with their lower and upper expectations



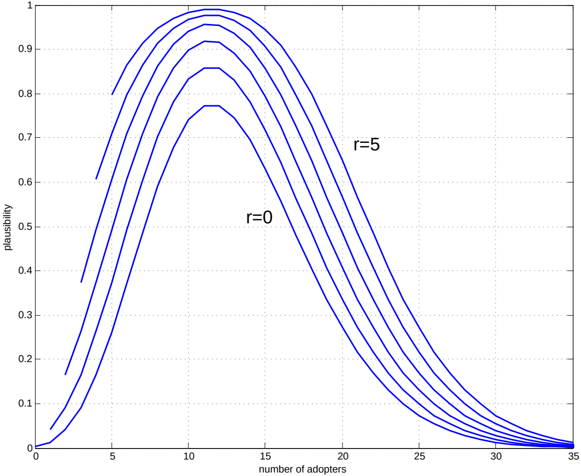
Cumulative belief and plausibility functions

Lower and upper cumulative distribution functions for the number of adopters in 1971, forecasted in 1970



Pl-plot

Plausibilities $Pl_{\boldsymbol{y}}^{\mathbb{Y}}([z-r, z+r])$ as functions of z, from $r=0$ (lower curve) to $r=5$ (upper curve), for the number of adopters in 1971, forecasted in 1970:



Conclusions

- Uncertainty quantification is an important component of any forecasting methodology. The approach introduced in this lecture allows us to represent forecast uncertainty in the belief function framework, based on past data and a statistical model
- The proposed method is conceptually simple and computationally tractable
- The belief function formalism makes it possible to combine information from several sources (such as expert opinions and statistical data)
- The Bayesian predictive probability distribution is recovered when a prior on θ is available
- The consistency of the method has been established under some conditions

References I

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

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