Statistical Inference from low-quality data using belief functions

Thierry Denœux[1]

[1] Université de Technologie de Compiègne
CNRS, UMR 7253 Heudiasyc
France

TES 2013
Chiang Mai, 10 January 2013

Motivation

Importance of data quality

- Current research in statistics and econometrics mainly focuses on the development of more complex models and inference procedures.
- However, data quality is recognized by applied statisticians as a key factor influencing the validity of the conclusions drawn from a statistical analysis:

"Issues of data quality and relevance, while underemphasized in the theoretical statistical and econometric literature, are certainly of great concern in much statistical work". (D. R. Cox, 2003)

Motivation

Need to combine data with expert judgement

- One of the reasons why data quality, in spite of its importance, has received relatively little attention in the statistical literature, may be that its evaluation often requires subjective judgements that do not easily fit with the standard frequentist or Bayesian frameworks.
- The Bayesian framework does allows for the introduction of personalistic prior information but...
- It does so in a very specific and questionable manner (by treating all unknown quantities as random variables), which raises a number of theoretical and practical issues.

Motivation

Alternative theories of uncertainty

- In the past thirty years, alternatives to the Bayesian framework for reasoning from weak information have emerged, including
  - Possibility Theory (Zadeh, 1978),
  - Imprecise Probabilities (Walley, 1991) and
  - The theory of Belief Functions (Dempster, 1967;Shafer, 1976).
- In this talk, we will focus on the Dempster-Shafer theory of Belief Functions and discuss its application to statistical inference, with emphasis on low-quality data.

Outline

Outline

1. Theory of belief functions
    - Representation of evidence
    - Combination of evidence
2. Application to statistical inference
    - Dempster's approach
    - Likelihood-based approach
3. Handling low-quality data
    - Partially relevant data
    - Uncertain data

Historical perspective

- Also known as Dempster-Shafer theory 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 framework for reasoning and decision-making under uncertainty.
- Many applications in Statistics, Artificial Intelligence, Information Fusion, etc.

Main features

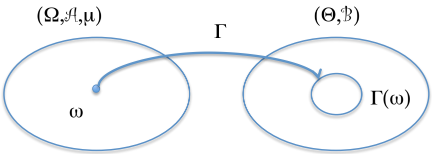
- The theory of belief function subsumes both the logical and probabilistic approaches to uncertainty: a belief function may be seen as
  - a non-additive measure or as
  - a generalized set.
- The belief function approach coincides with the Bayesian approach when all variables are described by probability 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

1. Theory of belief functions
   - Representation of evidence
   - Combination of evidence
2. Application to statistical inference
   - Dempster's approach
   - Likelihood-based approach
3. Handling low-quality data
   - Partially relevant data
   - Uncertain data

Belief and plausibility functions

Mathematical definitions



- Let $(\Omega, \mathcal{A}, \mu)$ be a probability space, $(\Theta, \mathcal{B})$ a measurable space, and Γ be a multivalued mapping from Ω to $\mathcal{B}$ defining a random set.
- Belief and plausibility functions on Θ are defined as follows (assuming certain measurability requirements):

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

$$Pl(A) = \mu(\{\omega \in \Omega | \Gamma(\omega) \cap A \neq \emptyset\} = 1 - Bel(\overline{A}),$$

for all $A \in \mathcal{B}$.

Belief and plausibility functions

Interpretation

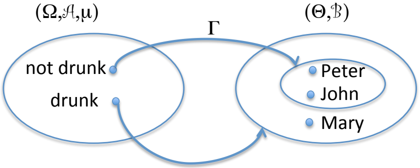
- Typically, Θ is the domain of an unknown quantity θ, and Ω is a set of interpretations of a given piece of evidence about θ, only one of which is true.
- If $\omega \in \Omega$ holds, then the evidence tells us that $\Gamma(\omega) \ni \theta$, and nothing more.
- Then
  - $Bel(A)$ is the probability that the evidence implies A;
  - $Pl(A)$ is the probability that the evidence is consistent with A.
- Obviously, $Bel(A) \leq Pl(A)$ for all A.
- *Bel* and *Pl* are non additive in general:

$$bel(A \cup B) \geq bel(A) + bel(B) - bel(A \cap B)$$

$$pl(A \cup B) \leq pl(A) + pl(B) - pl(A \cap B)$$

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.



$$Bel(\{Peter, John\}) = \mu(\{nd\}) = 0.8,$$

$$Pl(\{Peter, John\}) = \mu(\{nd, d\}) = 1$$

Special cases

1. If there is some $A \subseteq \Theta$ such that $\Gamma(\omega) = A$ for all $\omega \in \Omega$, then the evidence tells us that $\theta \in A$ for sure, and nothing more. The corresponding belief function Bel^A is said to be logical. In particular, the vacuous belief function Bel^Θ encodes complete ignorance.
2. If all focal sets $\Gamma(\omega)$ are singletons, then $Bel = Pl$ is a probability measure.
3. If the focal sets $\Gamma(\omega)$ are nested, Bel is said to be consonant. We then have:
$$Pl(A) = \sup_{\theta \in A} pl(\theta)$$
where $pl : \theta \to Pl(\{\theta\})$ is the contour function of Bel.

Outline

1. Theory of belief functions
   - Representation of evidence
   - Combination of evidence
2. Application to statistical inference
   - Dempster's approach
   - Likelihood-based approach
3. Handling low-quality data
   - Partially relevant data
   - Uncertain data

Dempster's rule
Definition

- Let us assume that we have 2 pieces of evidence modeled by random sets $(\Omega_1, \mathcal{A}_1, \mu_1, \Gamma_1)$ and $(\Omega_2, \mathcal{A}_2, \mu_2, \Gamma_2)$ that induce belief functions Bel_1 and Bel_2 on Θ.
- We assume that these pieces of evidence are independent, in the sense that the probability measure on $\Omega_1 \times \Omega_2$ is the product measure $\mu_1 \otimes \mu_2$.
- The two pieces of evidence imply A iff $\Gamma_1(\omega_1) \cap \Gamma_2(\omega_2) \subseteq A$, assuming that $\Gamma_1(\omega_1) \cap \Gamma_2(\omega_2) \neq \emptyset$.
- We define a new combined belief function $Bel_{12} = Bel_1 \oplus Bel_2$ as follows: $Bel_{12}(A)$ is the probability that $\Gamma_1(\omega_1) \cap \Gamma_2(\omega_2) \subseteq A$, conditional on the event that these two random sets have nonempty intersection.

Dempster's rule

Properties

- Commutativity, associativity. Neutral element: Bel^{Θ}.
- Generalization of intersection: if Bel^A and Bel^B are logical belief functions and $A \cap B \neq \emptyset$, then

$$Bel^A \oplus Bel^B = Bel^{A \cap B}$$

- Generalization of probabilistic conditioning: if P is a probability measure and Bel^A is a logical function, then $P \oplus Bel^A$ is the conditional probability measure $P(\cdot|A)$.

Outline

The problem

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

Outline

Sampling model

- Let $X \in \mathbb{X}$ denote the observable data, $\theta \in \Theta$ the parameter of interest and $f(x;\theta)$ the probability mass or density function describing the data-generating mechanism.
- Let us represent such a sampling model by an equation

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

where $U \in \mathbb{U}$ is an unobserved auxiliary variable with known probability distribution μ independent of θ.
- This representation is quite natural in the context of data generation: for instance, to generate a continuous random variable X with cumulative distribution function (cdf) F_θ, one might draw U from $\mathcal{U}([0,1])$ and set $X = F_\theta^{-1}(U)$.

From *a*-equation to belief function

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

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

- The random set $(\mathbb{U}, \mathcal{B}(\mathbb{U}), \mu, \Gamma)$ induces a belief function $Bel_{\Theta \times \mathbb{X}}$ on $\mathbb{X} \times \Theta$.
- Conditioning $Bel_{\Theta \times \mathbb{X}}$ on θ yields the sampling distribution $f(\cdot; \theta)$ on $\mathbb{X}$;
- Conditioning it on $X = x$ gives a belief function $Bel_{\Theta}(\cdot; x)$ on Θ.

Example: Bernoulli sample

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

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

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

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

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

Discussion

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

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

- Drawbacks:
  - It often leads to cumbersome or even intractable calculations except for very simple models, which imposes the use of Monte-Carlo simulations.
  - More fundamentally, the analysis depends on the a-equation $X = a(\theta, U)$ and the auxiliary variable U, which are not unique for a given statistical model $\{f(\cdot;\theta), \theta \in \Theta\}$: the method fails to obey the likelihood principle.

Outline

Likelihood-based belief function

Requirements

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

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

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

Likelihood-based belief function

Solution

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

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

- The corresponding plausibility function is:

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

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

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

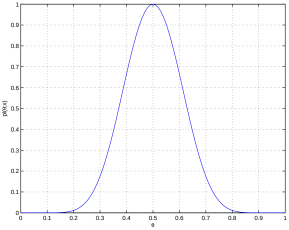
Example: Bernoulli sample

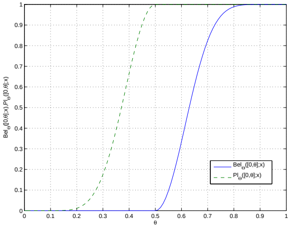
- Let $X = (X_1, \ldots, X_n)$ consist of independent Bernoulli observations and $\theta \in \Theta = [0, 1]$ is the probability of success.
- We get

$$pl(\theta; x) = \frac{\theta^y (1-\theta)^{n-y}}{\hat{\theta}^y (1-\hat{\theta})^{n-y}},$$

where $y = \sum_{i=1}^n x_i$ and $\hat{\theta}$ is the MLE.

- Example for $n = 20$ and $y = 10$:





Discussion

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

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

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

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

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

Outline

1. Theory of belief functions
   - Representation of evidence
   - Combination of evidence
2. Application to statistical inference
   - Dempster's approach
   - Likelihood-based approach
3. Handling low-quality data
   - Partially relevant data
   - Uncertain data

Motivation

- Classical statistical procedures address idealized situations where the data are precisely observed and can be considered as being drawn from a well defined population described by some parameter of interest θ.
- There are situations, however, where this simple model does not apply. Two such situations will be considered:
  - Some of data may collected from a population that is only known to "resemble" the population of interest (because, e.g., there were collected at different times or places) $\rightarrow$ partially relevant data.
  - Sometimes, the data are only imperfectly observed and we may only have a set of possible values, a value with a confidence degree or a combination of both $\rightarrow$ uncertain data.

Outline

Problem statement

- Assume that we are interested in a parameter $\theta \in \Theta$ related to a certain population and we observe a random variable X with probability density or mass function $f(x; \theta')$, where $\theta' \in \Theta$ is a parameter believed to be "close" to θ.
- For instance, θ might be the death rate in some hospital, and X the number of deaths in a neighboring hospital.
- Having observed $X = x$, our belief about θ' is represented by the contour function

$$pl'(\theta'; x) = \frac{L(\theta'; x)}{\sup_{\theta'} L(\theta'; x)}.$$

- What does x tell us about θ?

Solution

- Assume that the statement "θ' is close to θ" can be formalized as $d(\theta, \theta') \leq \delta$, where d is a distance measure defined on Θ and δ is a known constant.
- This piece of information can be modeled by a logical belief function with focal set $S_\delta = \{(\theta, \theta') | d(\theta, \theta') \leq \delta\} \subset \Theta^2$.
- Combining it with $pl'(\theta'; x)$ using Dempster's rule yields a consonant belief function on $\Theta \times \Theta'$, with contour function

$$pl(\theta, \theta'; x) = pl'(\theta'; x) \mathbb{1}_{S_\delta}(\theta, \theta').$$

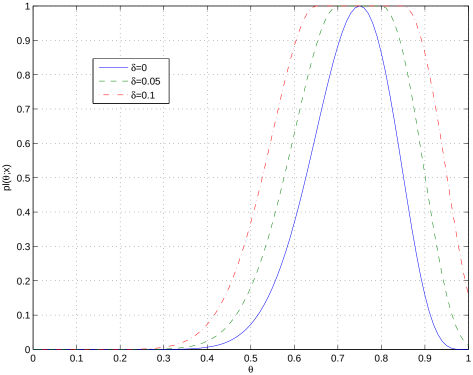
- Marginalizing out θ' yields:

$$pl(\theta; x) = \sup_{\theta'} pl(\theta, \theta'; x) = \sup_{\theta' \in B_\delta(\theta)} pl'(\theta'; x),$$

where $B_\delta(\theta) = \{\theta' \in \Theta | d(\theta, \theta') \leq \delta\}$.

Example

Binomial distribution with $n = 20$, $x = 15$ and $\delta \in \{0, 0.05, 0.1\}$



Outline

Problem statement

- We consider the situation where the data x have been generated by a random process but have been imperfectly observed.
- Our partial knowledge of x can then be described by a belief function $Bel_{\mathbb{X}}$ on the sample space $\mathbb{X}$.
- For instance, we may know that $x \in B$ with some confidence degree α, which is described by the following belief function

$$Bel_{\mathbb{X}}(A) = \begin{cases} \alpha & \text{if } B \subseteq A \subset \mathbb{X}, \\ 1 & \text{if } A = \mathbb{X}, \\ 0 & \text{otherwise.} \end{cases}$$

- Problem: how to quantify the uncertainty on Θ given $Bel_{\mathbb{X}}$?

Solution

- Assume $\mathbb{X}$ is finite and induced by a random set $(\Omega, 2^\Omega, P_\Omega, \Gamma)$, where Ω is a finite space.
- We define

$$L(\theta; Bel_{\mathbb{X}}) = \sum_{\omega \in \Omega} p(\omega) P_{\mathbb{X}}(X \in \Gamma(\omega)) = \sum_{x \in \mathbb{X}} f(x;\theta) pl(x)$$

and

$$pl(\theta; Bel_{\mathbb{X}}) = \frac{L(\theta; Bel_{\mathbb{X}})}{\sup_{\theta \in \Theta} L(\theta; Bel_{\mathbb{X}})}.$$

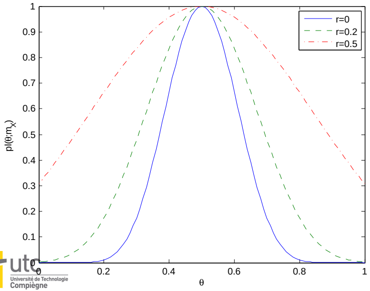
- When $\mathbf{X} = (X_1, \ldots, X_n)$, where each X_i is a random variable taking values in $\mathbb{X}_i$, then under independence assumptions:

$$pl(\theta; Bel_{\mathbb{X}}) \propto \prod_{i=1}^{n} pl(\theta; Bel_{\mathbb{X}_i})$$

Example

- Assume that $\mathbf{X} = (X_1, \ldots, X_n)$ is iid with $X_i \sim \mathcal{B}(\theta)$.
- We have

$$pl(\theta; Bel_{\mathbb{X}}) \propto \prod_{i=1}^{n} [\theta pl_{\mathbb{X}_i}(1) + (1-\theta) pl_{\mathbb{X}_i}(0)]$$



- $n = 20$,
- $pl_i(1) = 1$, $pl_i(0) = r$ for $i = 1, \ldots, 10$,
- $pl_i(1) = r$, $pl_i(0) = 1$ for $i = 11, \ldots, 20$.

Conclusion

- The theory of belief function is a general framework for uncertain reasoning, which encompasses probabilistic reasoning as a special case.
- When applied to statistical inference, this formalism allows us to quantify uncertainty on the parameter without having to specify a prior probability distribution. When such a probabilistic prior is available, the resulting belief function is identical to the Bayesian posterior.
- The belief function approach to statistical inference is general enough to model real-world problems that are not easily handled by classical statistical methods, such as situations of low-quality (partially relevant or imperfectly observed) data.
- The method has been applied to various statistical models such as linear regression (Denoeux, 2011), Gaussian mixtures (Côme, 2009), independent factor analysis (Cherfi et al, 2012) and HMMs (Ramasso and Denoeux, 2012).

Papers and Matlab software available at:

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

THANK YOU!