# Theory of Belief Functions: Application to machine learning and statistical inference

## Lecture 3: Multinomial predictive belief function

Thierry Denœux

Summer 2023

# The problem

- Research on Belief Functions (Dempster-Shafer theory) $\rightarrow$ developing new tools for manipulating belief functions:
  - Combination rules,
  - Propagation in evidential networks,
  - General Bayesian Theorem, ...
- Where do the belief functions come from?
  - Expert opinions: belief function elicitation (see paper by Ben Yaghlane *et al.* in this conference);
  - Data: the topic of this talk.

# Introductory example

- Consider an urn with white ($\xi_1$), red ($\xi_2$) and black ($\xi_3$) balls in proportions $p_1$, $p_2$ and $p_3$.
- Let $X \in \mathcal{X} = \{\xi_1, \xi_2, \xi_3\}$ be the color of a ball that will be drawn from the urn: belief on $X$?
- Two cases:
  1. We know the proportions $p_k$: then $Bel^{\mathcal{X}}(\{\xi_k\}) = p_k$ ( Hacking's Principle);
  2. We have observed the result of $n$ drawings from the urn with replacement, e.g. 5 white balls, 3 red balls and 2 black balls.
- How to build a belief function from data in the 2nd case ?

# Formalization

- Discrete variable $X \in \mathcal{X} = \{\xi_1, \ldots, \xi_K\}$ defined as the result of a random experiment.
- $X$ is characterized by an unknown frequency (probability) distribution $\mathbb{P}_X$.
- $\mathbb{P}_X(A)$: limit frequency of the event $A \subseteq \mathcal{X}$ in an infinite sequence of trials.
- We have observed a realization $\mathbf{x}_n$ of an iid random sample $X_n = (X_1, \ldots, X_n)$ with parent distribution $\mathbb{P}_X$.
- Problem: build a belief function $Bel^{\mathcal{X}}[\mathbf{x}_n]$ with well-defined properties with respect to the unknown frequency distribution $\mathbb{P}_X \rightarrow$ predictive belief function.

# Previous work

- Dempster (1966) provided a solution for the case $K = 2$:

$$m^{\mathcal{X}}(\{\xi_1\}) = \frac{N_1}{n+1}, \quad m^{\mathcal{X}}(\{\xi_2\}) = \frac{N_2}{n+1},$$

$$m^{\mathcal{X}}(\mathcal{X}) = \frac{1}{n+1},$$

  with $N_k = \#\{i | X_i = \xi_k\}$, $N_1 + N_2 = n$.
- Same result obtained by Smets (1994) in the TBM framework.
- Both approaches become intractable when $K > 2$.

# Previous work (cont.)

- Same problem tackled by Walley (1996) in the imprecise probability framework $\rightarrow$ Imprecise Dirichlet Model (IDM).
- The obtained lower probability measure happens to be a belief function, with mass function:

$$
\begin{aligned}
m^{\mathcal{X}}(\{\xi_k\}|\mathbf{N}, s) &= \frac{N_k}{n+s}, \quad k = 1, \ldots, K, \\
m^{\mathcal{X}}(\mathcal{X}|\mathbf{N}, s) &= \frac{s}{s+n},
\end{aligned}
$$

  with $N_k = \#\{i | X_i = \xi_k\}$ and $s > 0$.
- Well justified in the IP framework, not in the BF framework.

# New approach

- Let $Bel^{\mathcal{X}}[\mathbf{x}_n]$ be the BF on $X$ after observing a realization $\mathbf{x}_n$ of random sample $\mathbf{X}_n = (X_1, \ldots, X_n)$.
- Which properties should $Bel^{\mathcal{X}}[\mathbf{x}_n]$ verify with respect to $\mathbb{P}_X$ ?
- Hacking's principle (1965): if $\mathbb{P}_X$ is know, then $Bel^{\mathcal{X}}[\mathbf{x}_n] = \mathbb{P}_X$.
- Weak version:

$$\forall A \subseteq \mathcal{X}, \quad Bel^{\mathcal{X}}[\mathbf{X}_n](A) \xrightarrow{P} \mathbb{P}_X(A), \text{ as } n \to \infty.$$

(Requirement $R_1$)

# New approach (continued)

- Least Commitment Principle: for fixed $n$, $Bel^{\mathcal{X}}[\mathbf{x}_n]$ should be less informative that $\mathbb{P}_X$:

$$Bel^{\mathcal{X}}[\mathbf{x}_n](A) \leq \mathbb{P}_X(A), \quad \forall A \subseteq \mathcal{X}.$$

- This condition is too restrictive (it leads to the vacuous BF).
- Weaker condition (Requirement $R_2$):

$$\mathbb{P}(Bel^{\mathcal{X}}[\mathbf{X}_n] \leq \mathbb{P}_X) \geq 1 - \alpha,$$

for some $\alpha \in (0, 1)$.

# Meaning of Requirement $R_2$

$$
\begin{aligned}
\mathbf{x}_n = (x_1, \ldots, x_n) &\rightarrow Bel^{\mathcal{X}}[\mathbf{x}_n] \\
\mathbf{x}'_n = (x'_1, \ldots, x'_n) &\rightarrow Bel^{\mathcal{X}}[\mathbf{x}'_n] \\
\mathbf{x}''_n = (x''_1, \ldots, x''_n) &\rightarrow Bel^{\mathcal{X}}[\mathbf{x}''_n] \\
&\vdots
\end{aligned}
$$

- As the number of realizations of the random sample tends to $\infty$, the proportion of belief functions less committed than $\mathbb{P}_X$ should tend to $1 - \alpha$.
- To achieve this property: use of a multinomial confidence region.

# Multinomial Confidence Region

- Let $N_k = \#\{i | X_i = \xi_k\}$. Vector $\mathbf{N} = (N_1, \ldots, N_K)$ has a multinomial distribution $\mathcal{M}(n, p_1, \ldots, p_K)$, with $p_k = \mathbb{P}_X(\{\xi_k\})$.
- Let $\mathcal{S}(\mathbf{N}) \subseteq [0,1]^K$ a random region of $[0,1]^K$. It is a confidence region for $\mathbf{p}$ at level $1-\alpha$ if
$$\mathbb{P}(\mathcal{S}(\mathbf{N}) \ni \mathbf{p}) \geq 1-\alpha.$$
- $\mathcal{S}(\mathbf{N})$ is an asymptotic confidence region if the above inequality holds in the limit as $n \to \infty$.
- Simultaneous confidence intervals: $\mathcal{S}(\mathbf{N}) = [P_1^-, P_1^+] \times \ldots \times [P_K^-, P_K^+]$

# Multinomial Confidence Region (cont.)

- Goodman's simultaneous confidence intervals:

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

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

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

# Example

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

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

# From ConfidenceRegions to Lower Probabilities

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

$$P^-(A) = \max\left(\sum_{\xi_k \in A} P_k^-, 1 - \sum_{\xi_k \notin A} P_k^+\right)$$

- $P^-$ satisfies requirements $R_1$ and $R_2$:
  - $P^-(A) \xrightarrow{P} \mathbb{P}_X(A)$ as $n \to \infty$, for all $A \subseteq \mathcal{X}$,
  - $\mathbb{P}(P^- \leq \mathbb{P}_X) \geq 1 - \alpha$.

# From lower probabilities to belief functions

## Case $K \leq 3$

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

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

- With Goodman intervals:

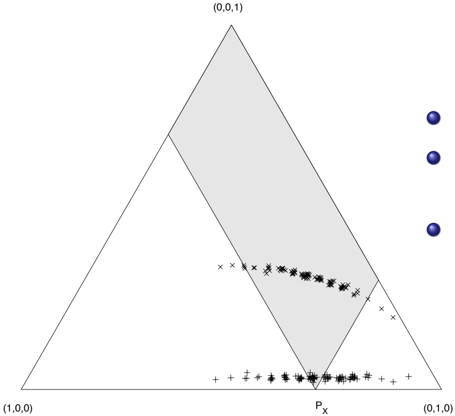
$$m(\{\xi_1\}) \approx \widehat{p} - u_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

$$m(\{\xi_2\}) \approx 1 - \widehat{p} - u_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

$$m(\mathcal{X}) \approx 2u_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}},$$

where $\widehat{p} = N_1/n$.

# Simulation example




- $K = 2$, $p_1 = \mathbb{P}_X(\{\xi_1\}) = 0.3$.
- 100 realizations of a random sample of size $n = 30$.
- 100 predictive belief functions at level $1 - \alpha = 0.95$.

# From lower probabilities to belief functions

## Case $K > 3$

- If $K > 3$, $P^-$ is not a belief function in general. We can find the most committed belief function satisfying $Bel \leq P^-$ by solving the following linear optimization problem:

$$\max_m J(m) = \sum_{A \subseteq \mathcal{X}} Bel(A) = \sum_{A \subseteq \mathcal{X}} \sum_{B \subseteq A} m(B)$$

subject to the constraints:

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

$$\sum_{A \subseteq \mathcal{X}} m(A) = 1, \quad m(A) \geq 0, \quad \forall A \subseteq \mathcal{X}$$

- The solution satisfies requirements $R_1$ and $R_2$: it is a predictive belief function (at confidence level $1 - \alpha$).

# Example: Psychiatric Data

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

# Case of ordered data

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

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

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

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

# Example: rain data

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

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

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

# Rain data: Result

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

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

# Conclusions

- A "frequentist" approach, based on multinomial confidence regions, for building a belief function quantifying the uncertainty about a discrete random variable $X$ with unknown probability distribution, based on observed data.
- Two "reasonable" properties of the solution with respect to the true frequency distribution $\mathbb{P}_X$:
  - it is less committed than $\mathbb{P}_X$ with some user-defined probability, and
  - it converges towards $\mathbb{P}_X$ in probability as the size of the sample tends to infinity.
- Another approach based on the likelihood function will be described later.