Uncertainty Modeling using Random Fuzzy Sets: Theory and Application to Machine Learning

Thierry Denœux

Université de technologie de Compiègne, France
Institut universitaire de France

Distinguished Research Seminar Series

Department of Industrial and Systems Engineering
The Hong Kong Polytechnic University
16 November 2026

Uncertainty quantification in prediction

Prediction uncertainty

- Uncertainty quantification is a fundamental problem in statistics, artificial intelligence and engineering.
- When we predict, e.g., the survival time of a patient or the time to failure of a component, we need to tell the user how confident they can be in the prediction.

Aleatory uncertainty

- Inherent variability
- Cannot be reduced
- Usually represented by probability

Epistemic uncertainty

- Lack of knowledge
- Can be reduced by more information
- Imprecision or incompleteness of information

A motivating example

- A survival model is used to predict whether two cancer patients will survive more than five years ($T > 5$).

Patient A

Profile similar to that of many patients in the training data, about 50% of whom survived.

Patient B

Atypical profile: very few similar cases in the training data.

- A probabilistic model returns a single number $P(T > 5 \mid \mathbf{x}) \approx 0.5$ for each patient.
- A single number cannot distinguish
  - evidence of equal chances (many similar patients, about half of whom survived), and
  - lack of evidence (no similar patients).

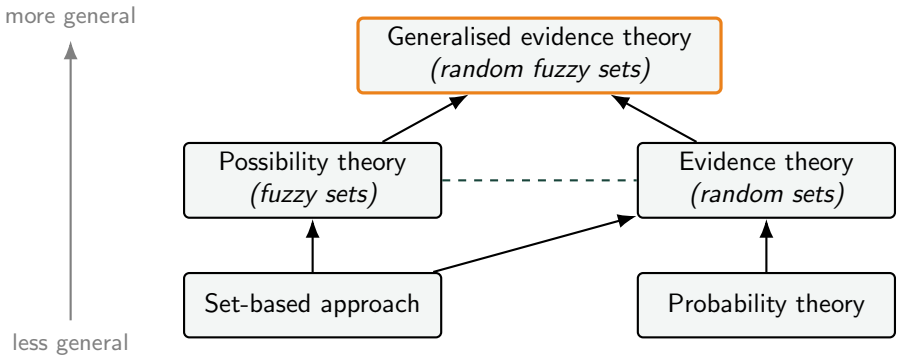
Degrees of support and plausibility

- Evidence theory assigns two numbers to each proposition A:
  - $\mathrm{Bel}(A)$: degree to which the evidence supports A;
  - $\mathrm{Pl}(A) = 1 - \mathrm{Bel}(A^c)$: degree to which A is compatible with the evidence (lack of support for the contrary).
- Example with $A = \{T > 5\}$ (illustrative values, obtained with the type of model presented in this talk):

	$\mathrm{Bel}(T > 5)$	$\mathrm{Pl}(T > 5)$	$\mathrm{Bel}(T \leq 5)$	$\mathrm{Pl}(T \leq 5)$
Patient A	0.48	0.55	0.45	0.52
Patient B	0.10	0.91	0.09	0.90

- Patient A: almost all the evidence bears on one proposition or the other ($0.48 + 0.45 = 0.93$): the situation is close to a probabilistic one.
- Patient B: the evidence supports neither proposition strongly (0.10 and 0.09), and both remain highly plausible (0.91 and 0.90).

A hierarchy of uncertainty theories



- Evidence (Dempster-Shafer) theory (Dempster, 1967; Shafer, 1976) combines sets and probabilities.
- Possibility theory (Zadeh, 1978; Dubois & Prade, 1988) is based on fuzzy sets (sets with unsharp boundaries).
- Random fuzzy sets encompass both theories in a single framework, suitable for uncertain reasoning with both discrete and continuous variables (Denœux, 2023, 2027).

Outline

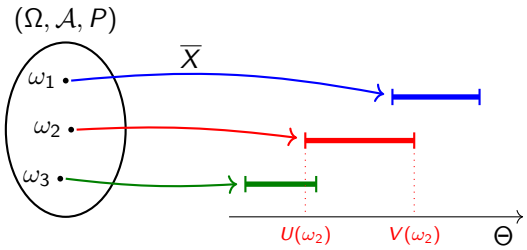
1. Evidence theory and random sets
2. From random sets to random fuzzy sets
   - General definitions
   - Gaussian random fuzzy numbers
3. Evidential machine learning
4. Application to survival analysis
   - Time-to-event prediction
   - Multimodal fusion
   - Competing risks

Outline

1. Evidence theory and random sets
2. From random sets to random fuzzy sets
3. Evidential machine learning
4. Application to survival analysis

Random sets

- Basic idea: represent evidence about an unknown quantity X with value in a set Θ by a random set. Example: the survival time T of a patient, with $\Theta =]0, +\infty[$.



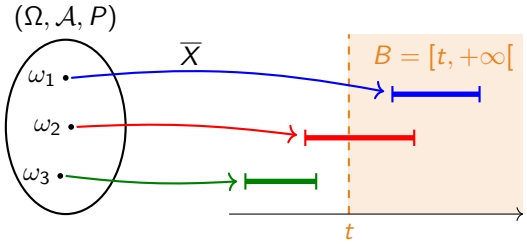
Definition

A random set is defined by

- a probability space $(\Omega, \mathcal{A}, P)$, where Ω is a set of interpretations of the evidence;
- a (measurable) set-valued mapping $\overline{X} : \Omega \to \mathcal{P}(\Theta)$.

- If interpretation ω holds, the evidence tells us that $X \in \overline{X}(\omega)$, and nothing more. The sets $\overline{X}(\omega)$ are the focal sets.
- Example: a random interval $\overline{X}(\omega) = [U(\omega), V(\omega)]$.
- When Θ is finite, a random set can be described by its mass function $m(A) = P(\{\omega : \overline{X}(\omega) = A\})$ (Dempster–Shafer theory).

Belief and plausibility



For any $B \subseteq \Theta$:

- degree of support (belief)

$$\mathrm{Bel}_{\overline{X}}(B) = P\left(\{\omega : \overline{X}(\omega) \subseteq B\}\right);$$

- degree of compatibility with the evidence (plausibility)

$$\mathrm{Pl}_{\overline{X}}(B) = P\left(\{\omega : \overline{X}(\omega) \cap B \neq \varnothing\}\right).$$

- In the picture: ω_1 implies $T \geq t$; ω_2 is compatible with $T \geq t$ without implying it; ω_3 contradicts $T \geq t$.
- For a random interval $[U, V]$: $\mathrm{Bel}_{\overline{X}}([t, +\infty[) = P(U \geq t)$ and $\mathrm{Pl}_{\overline{X}}([t, +\infty[) = P(V \geq t)$. As functions of t, these are the lower and upper survival functions behind the motivating example ($t = 5$).

Properties of belief and plausibility

- Duality: $\mathrm{Pl}_{\overline{X}}(B) = 1 - \mathrm{Bel}_{\overline{X}}(B^c)$, and $\mathrm{Bel}_{\overline{X}}(B) \leq \mathrm{Pl}_{\overline{X}}(B)$.
- Belief and plausibility are not additive: in general,

$$\mathrm{Bel}_{\overline{X}}(B) + \mathrm{Bel}_{\overline{X}}(B^c) \leq 1,$$

  i.e., the degrees of support for B and for its complement are only weakly linked.
- The difference

$$\mathrm{Pl}_{\overline{X}}(B) - \mathrm{Bel}_{\overline{X}}(B) = P\left(\{\omega : \overline{X}(\omega) \cap B \neq \varnothing \text{ and } \overline{X}(\omega) \cap B^c \neq \varnothing\}\right)$$

  is the probability that the evidence does not discriminate between B and B^c.

Back to the motivating example

Patient B: $\mathrm{Bel}(T > 5) = 0.10$ and $\mathrm{Pl}(T > 5) = 0.91$. With probability 0.81, the evidence does not discriminate between the two propositions.

Special cases

Focal sets	Equivalent to
singletons	a random variable (Bel = Pl is a probability measure)
a single set A (constant random set)	the set-based statement "$X \in A$"
Θ only	total ignorance (vacuous belief function)

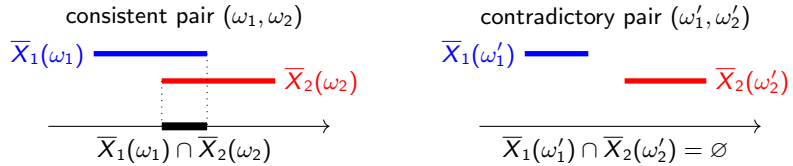
- Evidence theory thus generalises both set-based reasoning and probability theory.
- Under total ignorance, $\text{Bel}(B) = 0$ and $\text{Pl}(B) = 1$ for all $B \notin \{\varnothing, \Theta\}$; for instance, $\text{Bel}(T \geq t) = 0$ and $\text{Pl}(T \geq t) = 1$ for all $t > 0$.

Ignorance $\neq$ equiprobability

Evidence theory distinguishes lack of evidence (vacuous belief function) from evidence of equal chances (uniform probability distribution).

Combining independent evidence: Dempster's rule

- Two independent pieces of evidence are represented by random sets $\overline{X}_1 : \Omega_1 \to \mathcal{P}(\Theta)$ and $\overline{X}_2 : \Omega_2 \to \mathcal{P}(\Theta)$.
- If ω_1 and ω_2 both hold, the evidence tells us that $X \in \overline{X}_1(\omega_1) \cap \overline{X}_2(\omega_2)$.



- Noncontradictory pairs: $\Theta^* = \{(\omega_1, \omega_2) : \overline{X}_1(\omega_1) \cap \overline{X}_2(\omega_2) \neq \varnothing\}$; degree of conflict: $\kappa = 1 - (P_1 \otimes P_2)(\Theta^*)$.
- The orthogonal sum $\overline{X}_1 \oplus \overline{X}_2$ maps (ω_1, ω_2) to $\overline{X}_1(\omega_1) \cap \overline{X}_2(\omega_2)$, with probability measure $P_1 \otimes P_2$ conditioned on Θ^*.
- This is Dempster's rule; it is commutative and associative, and it generalises Bayesian conditioning.

Outline

1. Evidence theory and random sets
2. From random sets to random fuzzy sets
   - General definitions
   - Gaussian random fuzzy numbers
3. Evidential machine learning
4. Application to survival analysis

From crisp focal sets to fuzzy focal sets

Random set

$$\overline{X} : \omega \mapsto \overline{X}(\omega) \subseteq \Theta$$

$\overline{X}(\omega)$ expresses a hard constraint:

$$X \in \overline{X}(\omega).$$

Random fuzzy set

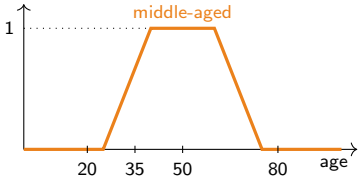
$$\widetilde{X} : \omega \mapsto \widetilde{X}(\omega) \text{ s.t. } \widetilde{X}(\omega) : \Theta \to [0,1]$$

$\widetilde{X}(\omega)$ expresses a soft (graded) constraint:

$$X = x \text{ is possible to degree } \widetilde{X}(\omega)(x).$$

Possibility and necessity generated by a fuzzy set

- A fuzzy subset of Θ is a map $\widetilde{\pi} : \Theta \to [0, 1]$. It is normal if $\sup_{\theta \in \Theta} \widetilde{\pi}(\theta) = 1$.
- A normal fuzzy set representing a soft constraint on an unknown X is called a possibility distribution.

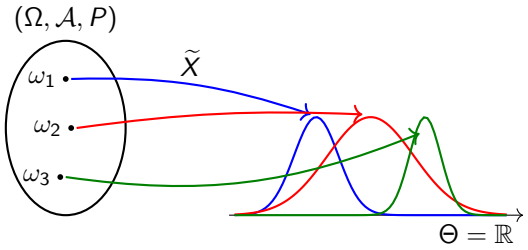


- A possibility distribution induces possibility and necessity measures:

$$\Pi_{\widetilde{\pi}}(B) = \sup_{x \in B} \widetilde{\pi}(x), \qquad \mathrm{N}_{\widetilde{\pi}}(B) = 1 - \Pi_{\widetilde{\pi}}(B^c).$$

  - $\Pi_{\widetilde{\pi}}(B)$ measures the extent to which B is compatible with the constraint.
  - $\mathrm{N}_{\widetilde{\pi}}(B)$ measures the extent to which B is implied by the constraint.
- A fuzzy focal set $\widetilde{X}(\omega)$ therefore induces possibility and necessity measures, conditional on ω.

Random fuzzy sets



Definition

A random fuzzy set (RFS) is defined by a probability space $(\Omega, \mathcal{A}, P)$ and a (measurable) fuzzy-valued mapping

$$\widetilde{X} : \Omega \to \mathcal{F}(\Theta).$$

If ω holds, X is constrained by the possibility distribution $\widetilde{X}(\omega)$.

- Random sets are recovered when all fuzzy focal sets are crisp.
- Random variables are recovered when all fuzzy focal sets are singletons.
- Possibility distributions are recovered when there is only one fuzzy focal set.

Belief and plausibility induced by an RFS

- For each $\omega \in \Omega$, $\widetilde{X}(\omega)$ induces possibility and necessity measures $\Pi_{\widetilde{X}(\omega)}$ and $\mathrm{N}_{\widetilde{X}(\omega)}$.
- Averaging over interpretations:

$$\mathrm{Pl}_{\widetilde{X}}(B) = \mathbb{E}\left[\Pi_{\widetilde{X}(\omega)}(B)\right], \qquad \mathrm{Bel}_{\widetilde{X}}(B) = \mathbb{E}\left[\mathrm{N}_{\widetilde{X}(\omega)}(B)\right] = 1 - \mathrm{Pl}_{\widetilde{X}}(B^c).$$

- $\mathrm{Bel}_{\widetilde{X}}$ is a belief function and $\mathrm{Pl}_{\widetilde{X}}$ is the dual plausibility function.
- The contour function is the plausibility of singletons:

$$\mathrm{pl}_{\widetilde{X}}(x) = \mathrm{Pl}_{\widetilde{X}}(\{x\}) = \mathbb{E}\left[\widetilde{X}(\omega)(x)\right].$$

Interpretation

- Plausibility is the expected possibility of B; belief is its expected necessity.
- An RFS is a probability distribution over possibility distributions.

One object, several limiting cases

Model	Fuzzy focal sets	Type of evidence
Random variable	random singletons	precise, random
Random set	random crisp sets	imprecise, random
Possibility distribution	one fixed fuzzy set	vague, nonrandom
Random fuzzy set	random fuzzy sets	random, imprecise and vague

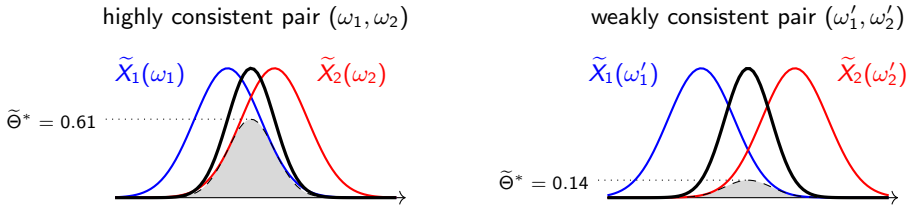
Unifying view

RFSs extend Dempster–Shafer theory in the same way as fuzzy sets extend crisp sets: focal sets become graded.

Combination: the product-intersection rule

- Let $\widetilde{X}_1 : \Omega_1 \to \mathcal{F}(\Theta)$ and $\widetilde{X}_2 : \Omega_2 \to \mathcal{F}(\Theta)$ represent two independent pieces of evidence.
- If ω_1 and ω_2 both hold, X is constrained by the normalised product

$$\left(\widetilde{X}_1(\omega_1) \cap_p^* \widetilde{X}_2(\omega_2)\right)(\theta) = \frac{\widetilde{X}_1(\omega_1)(\theta)\, \widetilde{X}_2(\omega_2)(\theta)}{\widetilde{\Theta}^*(\omega_1, \omega_2)}.$$



Dashed (shaded): product $\widetilde{X}_1(\omega_1)\, \widetilde{X}_2(\omega_2)$, of height $\widetilde{\Theta}^*(\omega_1, \omega_2)$; thick black: normalised product $\widetilde{X}_1(\omega_1) \cap_p^* \widetilde{X}_2(\omega_2)$.

- The height of the product, $\widetilde{\Theta}^*(\omega_1, \omega_2) = \sup_{\theta \in \Theta} \widetilde{X}_1(\omega_1)(\theta)\, \widetilde{X}_2(\omega_2)(\theta)$, is the degree of consistency of the pair (ω_1, ω_2).

Product-intersection rule: conditioning on consistency

- As in Dempster's rule, $P_1 \otimes P_2$ is conditioned on the consistent pairs, which now form a fuzzy subset $\widetilde{\Theta}^*$ of $\Omega_1 \times \Omega_2$:

$$P_{12}(B) = \frac{\int_B \widetilde{\Theta}^* \, d(P_1 \otimes P_2)}{\int_{\Omega_1 \times \Omega_2} \widetilde{\Theta}^* \, d(P_1 \otimes P_2)}.$$

- Weakly consistent pairs thus receive a small probability. The degree of conflict is

$$\kappa = 1 - \int_{\Omega_1 \times \Omega_2} \widetilde{\Theta}^* \, d(P_1 \otimes P_2).$$

- The orthogonal sum $\widetilde{X}_1 \oplus \widetilde{X}_2$ maps (ω_1, ω_2) to $\widetilde{X}_1(\omega_1) \cap_p^* \widetilde{X}_2(\omega_2)$, with probability measure P_{12}.

Properties

Commutative and associative; generalises both Dempster's rule (crisp focal sets) and the normalised product of possibility distributions (one focal set).

Gaussian fuzzy numbers

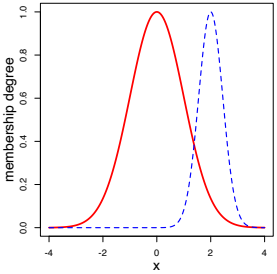
- General RFSs are very flexible, but we need parametric models to specify and propagate them in practice.

Gaussian fuzzy number (GFN)

$$\widetilde{F}(x) = \exp\left(-\frac{h}{2}(x-m)^2\right),$$

with mode $m \in \mathbb{R}$ and precision $h \geq 0$. We write $\widetilde{F} = \mathrm{GFN}(m, h)$.

$h = 0$: no information; $h \to +\infty$: singleton $\{m\}$.



- GFNs are closed under the normalised product intersection:

$$\mathrm{GFN}(m_1, h_1) \cap_p^* \mathrm{GFN}(m_2, h_2) = \mathrm{GFN}\left(\frac{h_1 m_1 + h_2 m_2}{h_1 + h_2}, h_1 + h_2\right).$$

Gaussian random fuzzy numbers

Definition

A Gaussian random fuzzy number (GRFN) is a GFN whose mode is a Gaussian random variable:

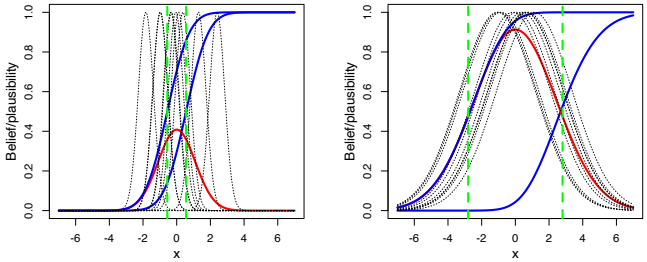
$$\widetilde{X} = \text{GFN}(M, h), \qquad M \sim \mathcal{N}(\mu, \sigma^2).$$

We write $\boxed{\widetilde{X} \sim \widetilde{\mathcal{N}}(\mu, \sigma^2, h)}$.

- Three parameters: a location μ, a variance σ^2 (aleatory uncertainty) and a precision h (epistemic uncertainty).
- A GRFN is the fuzzy analogue of the random interval $[M - \delta, M + \delta]$.

Condition	Special case
$h \to +\infty$	Gaussian random variable $\mathcal{N}(\mu, \sigma^2)$
$\sigma^2 = 0$	Gaussian possibility distribution $\text{GFN}(\mu, h)$
$h = 0$	vacuous belief function (total ignorance)

Visualising GRFNs



(a) (b)

- (a) $\widetilde{X}_1 \sim \widetilde{\mathcal{N}}(0, 1, 5)$; (b) $\widetilde{X}_2 \sim \widetilde{\mathcal{N}}(0, 1, 0.2)$.
- Ten focal fuzzy sets (dotted), contour function (red), lower and upper CDFs (blue), lower and upper expectations (green).
- The lower and upper CDFs move apart as the precision h decreases.

Closed-form expressions

Let $\widetilde{X} \sim \widetilde{\mathcal{N}}(\mu, \sigma^2, h)$ with $\sigma > 0$.

- Contour function:

$$\mathrm{pl}_{\widetilde{X}}(x) = \frac{1}{\sqrt{1 + h\sigma^2}} \exp\left(-\frac{h(x-\mu)^2}{2(1+h\sigma^2)}\right).$$

- Lower and upper CDFs:

$$\underline{F}_{\widetilde{X}}(x) = \mathrm{Bel}_{\widetilde{X}}(]-\infty, x]) = \Phi\left(\frac{x-\mu}{\sigma}\right) - \mathrm{pl}_{\widetilde{X}}(x)\,\Phi\left(\frac{x-\mu}{\sigma\sqrt{h\sigma^2+1}}\right),$$

$$\overline{F}_{\widetilde{X}}(x) = \mathrm{Pl}_{\widetilde{X}}(]-\infty, x]) = \underline{F}_{\widetilde{X}}(x) + \mathrm{pl}_{\widetilde{X}}(x).$$

- Closed-form expressions also exist for the belief and plausibility of any interval $[x, y]$, and for lower and upper expectations: $\underline{\mathbb{E}}(\widetilde{X}) = \mu - \sqrt{\pi/(2h)}$, $\overline{\mathbb{E}}(\widetilde{X}) = \mu + \sqrt{\pi/(2h)}$.

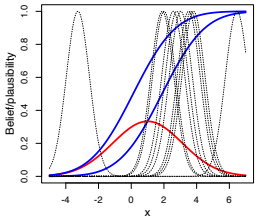
Message

GRFNs are as easy to handle as Gaussian random variables, with one extra parameter for epistemic uncertainty.

Combination of GRFNs

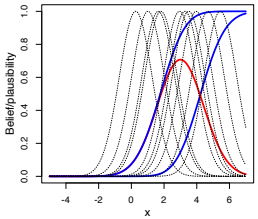
Closure property

If $\widetilde{X}_1 \sim \widetilde{\mathcal{N}}(\mu_1, \sigma_1^2, h_1)$ and $\widetilde{X}_2 \sim \widetilde{\mathcal{N}}(\mu_2, \sigma_2^2, h_2)$ are independent, then $\widetilde{X}_1 \oplus \widetilde{X}_2 \sim \widetilde{\mathcal{N}}(\mu_c, \sigma_c^2, h_1 + h_2)$, where μ_c and σ_c^2 have closed-form expressions (also for n GRFNs, and for dependent GRFNs).



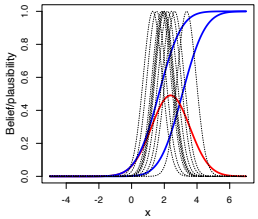
$\widetilde{\mathcal{N}}(1, 4, 2)$

$\oplus$



$\widetilde{\mathcal{N}}(3, 1, 1)$

$=$



$\widetilde{\mathcal{N}}(2.38, 1.05, 3)$

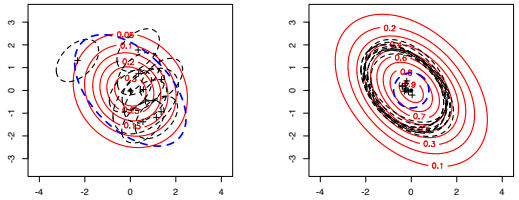
- Precisions add up: combining evidence reduces epistemic uncertainty.

A toolbox of tractable models

- Gaussian random fuzzy vectors (GRFVs):

$$\widetilde{\mathbf{X}} = \mathrm{GFV}(\mathbf{M}, \mathbf{H}), \quad \mathbf{M} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

  with $\mathrm{GFV}(\mathbf{m}, \mathbf{H})(\mathbf{x}) = \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T \mathbf{H}(\mathbf{x}-\mathbf{m})\right)$.
- Transformed GRFNs, e.g., lognormal random fuzzy numbers for positive quantities.
- Mixtures of GRFNs, for multimodal uncertainty.



Two GRFVs: focal sets (black), contour function (red), 95% ellipse of $\mathbf{M}$ (blue).

Key point

These families are closed under the product-intersection rule, which makes evidence-theoretic reasoning in continuous spaces computationally tractable.

Outline

1 Evidence theory and random sets

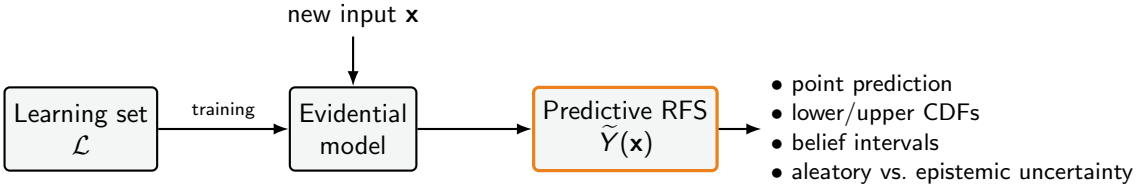
2 From random sets to random fuzzy sets

3 Evidential machine learning

4 Application to survival analysis

Supervised learning with uncertainty quantification

- In supervised learning, we learn a predictive model from a learning set $\mathcal{L} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^p$ is an input vector and $y_i \in \mathbb{R}$ a response.
- In evidential machine learning, the prediction for a new input $\mathbf{x}$ is a random fuzzy set $\widetilde{Y}(\mathbf{x})$, which represents both aleatory and epistemic uncertainty about the response Y.



- Two approaches: distance-based (this talk) and likelihood-based (briefly mentioned at the end).

Distance-based evidential prediction

- The starting idea is simple:

> nearby training points or prototypes provide stronger evidence.

- A prototype close to the input $\mathbf{x}$ provides strong evidence.
- A prototype far from $\mathbf{x}$ provides weak or even vacuous evidence.
- This produces a natural warning in regions with little or no training data.

Principle

Distances to prototypes induce pieces of evidence; these pieces are combined into a predictive belief function.

- History: evidential k-NN rule (1995) $\rightarrow$ evidential neural network classifier (2000) $\rightarrow$ ENNreg (2023) for regression.

ENNreg: prototype evidence as GRFNs

- The input space is summarised by K prototypes $\mathbf{w}_1, \ldots, \mathbf{w}_K$.
- Similarity between input $\mathbf{x}$ and prototype $\mathbf{w}_k$:

$$s_k(\mathbf{x}) = \exp\left(-\gamma_k \|\mathbf{x} - \mathbf{w}_k\|^2\right).$$

- Evidence supplied by prototype k:

$$\widetilde{Y}_k(\mathbf{x}) \sim \widetilde{\mathcal{N}}\left(\mu_k(\mathbf{x}), \sigma_k^2, s_k(\mathbf{x}) h_k\right), \quad \text{with} \quad \mu_k(\mathbf{x}) = \boldsymbol{\beta}_k^T \mathbf{x} + \beta_{k0}.$$

  - $\mu_k(\mathbf{x})$: local linear estimate of $\mathbb{E}(Y \mid \mathbf{x})$ in the neighbourhood of $\mathbf{w}_k$;
  - σ_k^2: local conditional variance;
  - $s_k(\mathbf{x}) h_k$: effective precision, which tends to 0 as $\|\mathbf{x} - \mathbf{w}_k\| \to \infty$.

Consequence

A prototype far from $\mathbf{x}$ supplies vacuous evidence: it no longer contributes to the prediction.

ENNreg: combining the prototype evidence

- The K GRFNs are combined by a simplified (unnormalised) product-intersection rule $\boxplus$:

$$\widetilde{Y}(\mathbf{x}) = \widetilde{Y}_1(\mathbf{x}) \boxplus \cdots \boxplus \widetilde{Y}_K(\mathbf{x}) \sim \widetilde{\mathcal{N}}\left(\mu(\mathbf{x}), \sigma^2(\mathbf{x}), h(\mathbf{x})\right),$$

with

$$\mu(\mathbf{x}) = \frac{\sum_{k=1}^K s_k(\mathbf{x}) h_k \mu_k(\mathbf{x})}{\sum_{k=1}^K s_k(\mathbf{x}) h_k}, \quad \sigma^2(\mathbf{x}) = \frac{\sum_{k=1}^K s_k(\mathbf{x})^2 h_k^2 \sigma_k^2}{\left(\sum_{k=1}^K s_k(\mathbf{x}) h_k\right)^2}, \quad h(\mathbf{x}) = \sum_{k=1}^K s_k(\mathbf{x}) h_k.$$

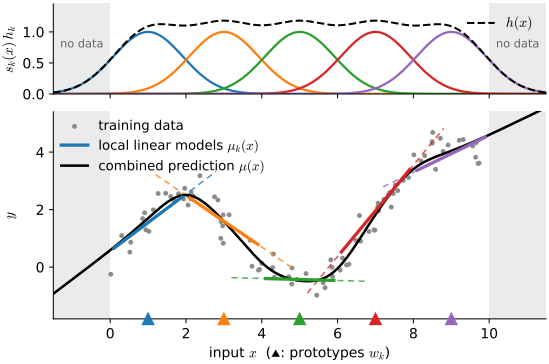
- $\boxplus$ neglects the conditioning step of $\oplus$; it is a good approximation when $\mathbf{x}$ is close to one prototype and far from the others.

Interpretation of the output

- $\mu(\mathbf{x})$: point prediction;
- $\sigma^2(\mathbf{x})$: aleatory uncertainty (heteroscedastic);
- $h(\mathbf{x})$: epistemic uncertainty (small $h(\mathbf{x})$ far from the data).

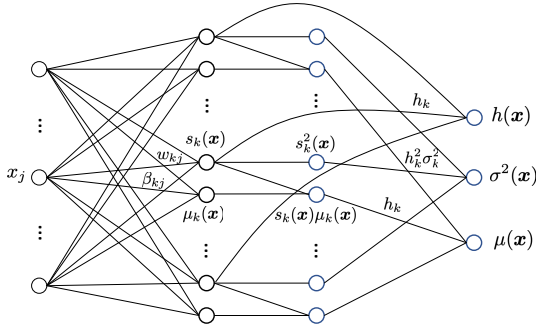
ENNreg: a one-dimensional illustration

- Each prototype w_k carries a local linear model $\mu_k(x)$ (coloured); the weights $s_k(x)h_k$ blend them into a smooth nonlinear prediction $\mu(x)$ (black).
- Outside the training data, $\mu(x)$ is extrapolated linearly, but $h(x) \to 0$: the prediction is supported by little evidence (high epistemic uncertainty).



ENNreg architecture

- Input layer, two hidden layers with $2K$ units each, three output units.
- Complexity of a forward pass: $O(pK)$, as for an RBF network.
- Generalises both normalised RBF networks ($\boldsymbol{\beta}_k = \mathbf{0}$) and linear regression ($\gamma_k = 0$).



Learning

- The model is trained by minimising

$$C_{\lambda,\epsilon,\xi,\rho}(\vartheta) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\lambda,\epsilon}\left(y_i, \widetilde{Y}(\mathbf{x}_i;\vartheta)\right) + \frac{\xi}{K}\sum_{k=1}^{K} h_k + \frac{\rho}{K}\sum_{k=1}^{K}\gamma_k,$$

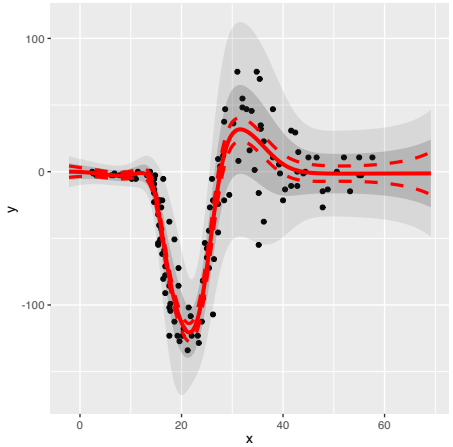
  with the generalised negative log-likelihood

$$\ell_{\lambda,\epsilon}\left(y_i, \widetilde{Y}(\mathbf{x}_i)\right) = -\lambda \ln \mathrm{Bel}_{\widetilde{Y}(\mathbf{x}_i)}([y_i]_\epsilon) - (1-\lambda)\ln \mathrm{Pl}_{\widetilde{Y}(\mathbf{x}_i)}([y_i]_\epsilon),$$

  where $[y_i]_\epsilon = [y_i - \epsilon, y_i + \epsilon]$ accounts for the finite precision of observations.
- Hyperparameters:
  - $\lambda \in [0,1]$ controls the trade-off between precise ($\lambda \to 1$) and cautious ($\lambda \to 0$) predictions;
  - ξ and ρ regularise, respectively, the effective number of prototypes and the departure from a linear model (tuned by cross-validation).
- As $h \to +\infty$ and $\epsilon \to 0$, the loss reduces to the usual negative log-likelihood (up to a constant).

Example: motorcycle data

- Head acceleration in a simulated motorcycle accident, as a function of time.
- ENNreg with $K = 30$ prototypes.
- Point predictions (solid red curve),
- lower and upper expectations (dashed red curves),
- belief intervals at levels 0.50 and 0.95 (shaded areas).



Calibration of predictive belief functions

- A predictive RFS $\widetilde{Y}(\mathbf{x})$ is calibrated if, for every event A and almost every $\mathbf{x}$,

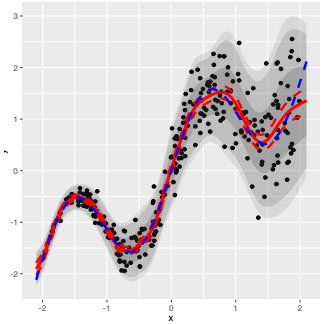
$$\text{Bel}_{\widetilde{Y}(\mathbf{x})}(A) \leq P_{Y|\mathbf{x}}(A) \leq \text{Pl}_{\widetilde{Y}(\mathbf{x})}(A),$$

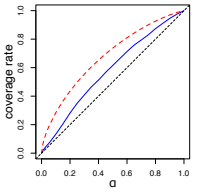
  where $P_{Y|\mathbf{x}}$ is the true conditional distribution of Y.
- An α-level belief interval $B_\alpha(\mathbf{x})$ verifies $\text{Bel}_{\widetilde{Y}(\mathbf{x})}(B_\alpha(\mathbf{x})) = \alpha$.
- Calibration implies

$$\alpha \leq \mathbb{E}_{\mathbf{X}}\left[P_{Y|\mathbf{X}}(B_\alpha(\mathbf{X}))\right] \leq \mathbb{E}_{\mathbf{X}}\left[\text{Pl}_{\widetilde{Y}(\mathbf{X})}(B_\alpha(\mathbf{X}))\right].$$

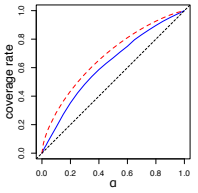
- This can be checked on test data by plotting, as functions of α,
  - the coverage rate $\widehat{p}(\alpha)$ of belief intervals (should lie above the diagonal),
  - the average plausibility $\overline{\text{Pl}}(\alpha)$ of belief intervals (should lie above the coverage curve).

Example: simulated heteroscedastic data

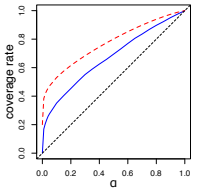




(a)



(b)



(c)

- Left: true regression function (blue dashed), predictions $\mu(x)$ (red), lower and upper expectations (red dashed), belief intervals at levels 0.5, 0.9 and 0.99 (grey); $n = 300$, $K = 30$.
- Right: calibration plots computed with (a) lower bounds, (b) upper bounds and (c) two-sided belief intervals: coverage rates (solid blue) and average plausibilities (dashed red).
- Predictions are cautious, and the necessary condition for calibration is satisfied.

Outline

1 Evidence theory and random sets

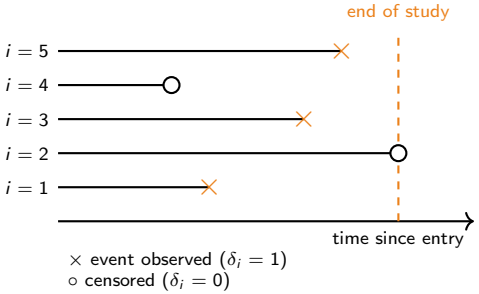
2 From random sets to random fuzzy sets

3 Evidential machine learning

4 Application to survival analysis
- Time-to-event prediction
- Multimodal fusion
- Competing risks

Time-to-event data

- $T > 0$: time to an event of interest (death, relapse, failure of a component, etc.).
- Data are right-censored: for some individuals, we only know that $T_i > t_i$.
- Observations: $(\mathbf{x}_i, t_i, \delta_i)$, with $\delta_i = 1$ if the event was observed, $\delta_i = 0$ otherwise.
- Target: the survival function $S(t \mid \mathbf{x}) = P(T > t \mid \mathbf{x})$.



The same problem in reliability engineering

Time to failure of components or systems; units still working at the end of a test or field study are censored.

Classical and machine-learning approaches

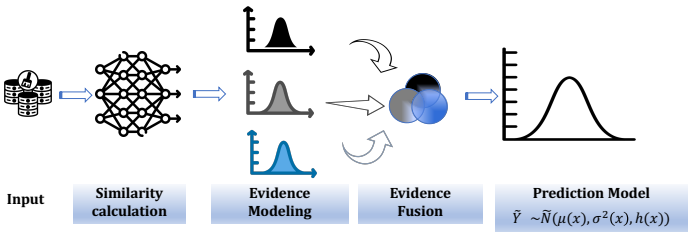
- Cox proportional hazards model:

$$\lambda(t \mid \mathbf{x}) = \lambda_0(t) \exp(\boldsymbol{\beta}^T \mathbf{x}).$$

  Simple and widely used, but the proportional hazards assumption is often unrealistic, and the model is linear in $\mathbf{x}$.

- Machine-learning approaches:
  - random survival forests (RSF);
  - neural extensions of the Cox model: DeepSurv, Cox-Time (non-proportional hazards);
  - discrete-time models: DeepHit (requires discretising time).
- Limitations:
  - strong distributional assumptions or loss of information due to time discretisation;
  - predictions are probability distributions: epistemic uncertainty (small samples, heavy censoring, atypical patients) is not represented, or only through computationally intensive ensembles or Monte Carlo dropout.

Evidential time-to-event prediction



- ENNreg is applied to the log-time $Y = \ln T$:

$$\widetilde{Y}(\mathbf{x}) \sim \widetilde{\mathcal{N}}\left(\mu(\mathbf{x}), \sigma^2(\mathbf{x}), h(\mathbf{x})\right) \iff \widetilde{T}(\mathbf{x}) = \exp \circ \widetilde{Y}(\mathbf{x})$$ is a lognormal random fuzzy number.

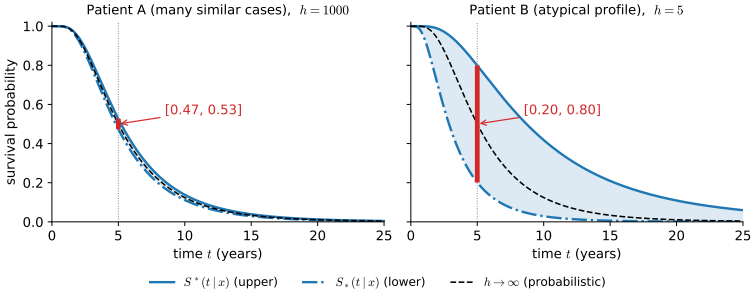
- $\exp(\mu(\mathbf{x}))$ is the most plausible event time; $\sigma^2(\mathbf{x})$ reflects aleatory uncertainty; $h(\mathbf{x})$ reflects epistemic uncertainty.
- No proportional hazards assumption, and no time discretisation.

Lower and upper survival functions

- The predictive RFS yields lower and upper survival functions:

$$S_*(t \mid \mathbf{x}) = \mathrm{Bel}_{\widetilde{T}(\mathbf{x})}([t, +\infty[) = 1 - \overline{F}_{\widetilde{Y}(\mathbf{x})}(\ln t),$$
$$S^*(t \mid \mathbf{x}) = \mathrm{Pl}_{\widetilde{T}(\mathbf{x})}([t, +\infty[) = 1 - \underline{F}_{\widetilde{Y}(\mathbf{x})}(\ln t),$$

with $S^*(t \mid \mathbf{x}) - S_*(t \mid \mathbf{x}) = \mathrm{pl}_{\widetilde{Y}(\mathbf{x})}(\ln t)$.



Back to the motivating example: same μ and σ, different precisions h (illustrative parameter values).

Learning from censored data

- A censored observation is an imprecise observation: we only know that $T_i \geq t_i$, i.e., $Y_i \in [y_i, +\infty[$ with $y_i = \ln t_i$.
- This is naturally handled by the generalised negative log-likelihood:

$$\ell_i = -\lambda \ln \mathrm{Bel}_{\widetilde{Y}(\mathbf{x}_i)}(A_i) - (1-\lambda) \ln \mathrm{Pl}_{\widetilde{Y}(\mathbf{x}_i)}(A_i),$$

with

$$A_i = \begin{cases} [y_i - \epsilon, y_i + \epsilon] & \text{if } \delta_i = 1 \text{ (event observed)}, \\ [y_i, +\infty[& \text{if } \delta_i = 0 \text{ (censored)}. \end{cases}$$

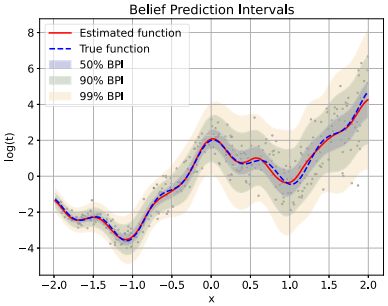
- The network is trained by minimising the regularised average loss, using standard gradient-based optimisation (e.g., Adam) with early stopping.

Remark

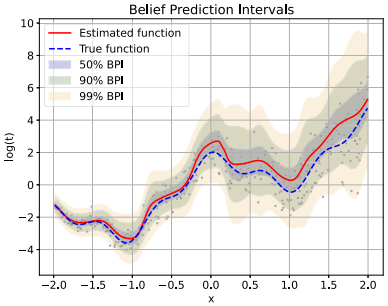
Evidence theory provides a unified treatment of precise and set-valued (censored) observations.

Simulated data: effect of censoring

- One input $X \sim \text{Unif}([-2, 2])$; ln T is a nonlinear function of X plus heteroscedastic Gaussian noise; 1600 training and 400 test instances.



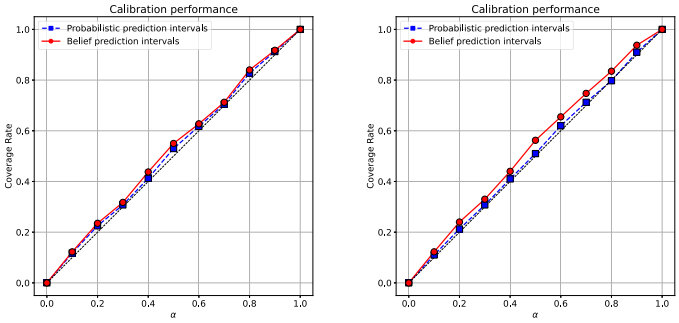
(a) No censoring



(b) 50% censoring

- With 50% censoring, predictions become slightly biased, but belief intervals widen and still contain most of the data.

Simulated data: calibration



(a) No censoring

(b) 50% censoring

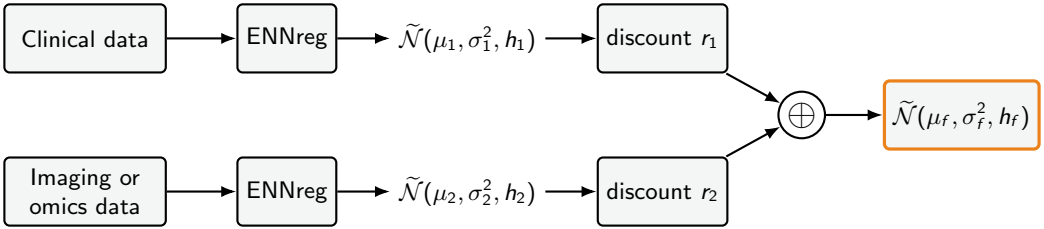
- Coverage rates of belief prediction intervals (red) and probabilistic prediction intervals (blue) as functions of α.
- Belief intervals remain calibrated with 50% censoring; they become more conservative, owing to higher epistemic uncertainty (lower $h(\mathbf{x})$).

Real-world datasets: discrimination

n (censoring rate)	METABRIC 1,904 (42%)	GBSG 2,232 (43%)	SUPPORT 8,873 (32%)	MIMIC-IV 53,512 (89%)
Cox	0.643 ± 0.010	0.662 ± 0.006	0.568 ± 0.002	0.779 ± 0.005
RSF	0.626 ± 0.008	0.663 ± 0.006	$\underline{0.624 \pm 0.002}$	0.814 ± 0.004
SurvivalSVM	0.649 ± 0.011	0.666 ± 0.006	0.572 ± 0.001	0.791 ± 0.005
DeepSurv	0.642 ± 0.037	0.669 ± 0.043	0.602 ± 0.004	0.834 ± 0.004
Cox-CC	0.643 ± 0.006	$\underline{0.673 \pm 0.003}$	0.603 ± 0.002	0.837 ± 0.003
Cox-Time	0.661 ± 0.005	0.672 ± 0.005	0.610 ± 0.002	$\underline{0.841 \pm 0.003}$
DeepHit	$\underline{0.669 \pm 0.010}$	0.667 ± 0.005	$\mathbf{0.626 \pm 0.002}$	0.823 ± 0.004
ENNreg	$\mathbf{0.672 \pm 0.007}$	$\mathbf{0.679 \pm 0.005}$	$\mathbf{0.626 \pm 0.004}$	$\mathbf{0.845 \pm 0.003}$

- Time-dependent concordance index, showing how well a model ranks patients by risk (higher is better), mean $\pm$ standard error over five random splits; best results in bold, second best underlined.
- In terms of calibration (integrated Brier score), ENNreg is best on GBSG and second best on METABRIC and MIMIC-IV (see backup slides).

Multimodal survival analysis: EsurvFusion



- One ENNreg module per modality, each producing a GRFN for the log-survival time.
- Possibilistic discounting with a learned reliability coefficient $r_i \in [0, 1]$:

$$\widetilde{\mathcal{N}}(\mu_i, \sigma_i^2, h_i) \longrightarrow \widetilde{\mathcal{N}}(\mu_i, \sigma_i^2, r_i h_i).$$

- Combination by product-intersection: precision $h_f = \sum_i r_i h_i$, and mean equal to the reliability- and precision-weighted average $\mu_f = \sum_i r_i h_i \mu_i / \sum_i r_i h_i$.
- HECKTOR 2022 data (head-and-neck cancer, PET/CT radiomics + clinical data, 81% censoring): C-index 0.703 vs. 0.686 for DeepSurv with late fusion; learned reliabilities $r_{\text{clin}} \approx 0.86$, $r_{\text{rad}} \approx 0.97$.

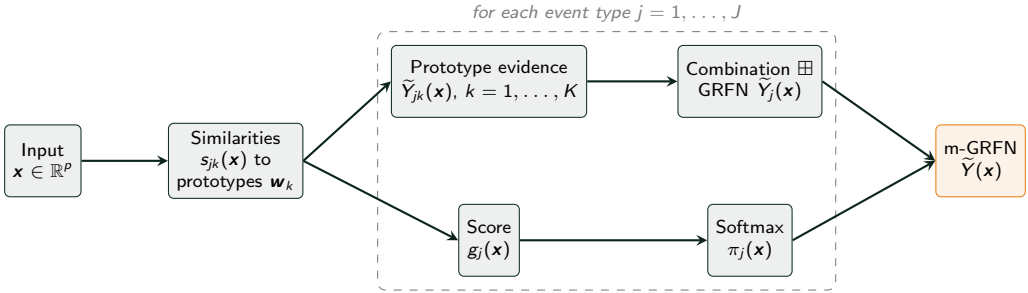
Competing risks

- Individuals may experience one of J mutually exclusive event types: e.g., death from cancer vs. death from other causes; relapse vs. graft-versus-host disease vs. infection.
- Data: $\{(\boldsymbol{x}_i, t_i, \delta_i)\}_{i=1}^N$, with covariates $\boldsymbol{x}_i \in \mathbb{R}^p$, observed time t_i and indicator $\delta_i \in \{0, 1, \ldots, J\}$, where 0 denotes censoring and J is the number of event types.
- Let T be the time to the first event and $D \in \{1, \ldots, J\}$ its type ($\delta_i = D_i$ if uncensored).
- Quantity of interest: the cumulative incidence function (CIF) of each event type,

$$F_j(t \mid \mathbf{x}) = P(T \leq t, D = j \mid \mathbf{x}) = \underbrace{P(D = j \mid \boldsymbol{x})}_{\pi_j(\boldsymbol{x})} \underbrace{P(T \leq t \mid D = j, \boldsymbol{x})}_{\text{described by a GRFN}}, \quad j = 1, \ldots, J.$$

- Existing approaches: cause-specific Cox model, Fine–Gray subdistribution hazards model, DeepHit, CR-SAVAE (variational autoencoder).
- These approaches generally do not quantify the epistemic uncertainty of event-specific predictions.

CR-ENNreg architecture



- One GRFN per prototype and event type, combined into an event-specific GRFN $\widetilde{Y}_j(\boldsymbol{x})$ describing $y = \ln T$ given $D = j$.
- The same similarities feed a softmax layer giving the event probabilities $\pi_j(\boldsymbol{x})$.
- Output: $\widetilde{Y}(\boldsymbol{x}) \sim \sum_{j=1}^{J} \pi_j(\boldsymbol{x}) \widetilde{\mathcal{N}}\big(\mu_j(\boldsymbol{x}), \sigma_j^2(\boldsymbol{x}), h_j(\boldsymbol{x})\big)$, an m-GRFN whose components are indexed by the event types.

CR-ENNreg: a mixture of GRFNs

- Probability of each event type, computed from the similarities:

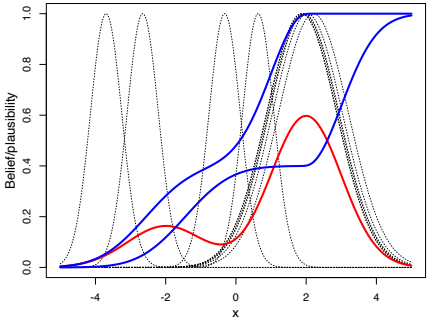
$$\pi_j(\mathbf{x}) = P(\delta = j \mid \mathbf{x}) = \frac{\exp(g_j(\mathbf{x}))}{\sum_{l=1}^{J} \exp(g_l(\mathbf{x}))},$$

with $g_j(\mathbf{x}) = \alpha_j + \sum_{k=1}^{K} \omega_{jk} s_{jk}(\mathbf{x})$.

- The predictive model is a mixture GRFN (m-GRFN):

$$\widetilde{Y}(\mathbf{x}) \sim \sum_{j=1}^{J} \pi_j(\mathbf{x}) \widetilde{\mathcal{N}}\left(\mu_j(\mathbf{x}), \sigma_j^2(\mathbf{x}), h_j(\mathbf{x})\right),$$

with $\mathrm{Bel}_{\widetilde{Y}(\mathbf{x})} = \sum_j \pi_j(\mathbf{x}) \mathrm{Bel}_{\widetilde{Y}_j(\mathbf{x})}$ and $\mathrm{Pl}_{\widetilde{Y}(\mathbf{x})} = \sum_j \pi_j(\mathbf{x}) \mathrm{Pl}_{\widetilde{Y}_j(\mathbf{x})}$.



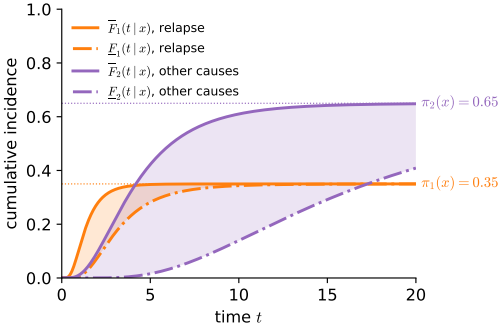
m-GRFN $0.4\,\widetilde{\mathcal{N}}(-2, 1, 5) + 0.6\,\widetilde{\mathcal{N}}(2, 0.01, 1)$: focal sets, contour function (red), lower/upper CDFs (blue).

CR-ENNreg: lower and upper CIFs

- Lower and upper CIFs:

$$\underline{F}_j(t \mid \mathbf{x}) = \pi_j(\mathbf{x})\, \mathrm{Bel}_{\widetilde{Y}_j(\mathbf{x})}\left(]-\infty, \ln t]\right),$$
$$\overline{F}_j(t \mid \mathbf{x}) = \pi_j(\mathbf{x})\, \mathrm{Pl}_{\widetilde{Y}_j(\mathbf{x})}\left(]-\infty, \ln t]\right).$$

- Training: same loss as before, with
  - $\pi_j(\mathbf{x}_i)\, \mathrm{Bel}_{\widetilde{Y}_j(\mathbf{x}_i)}([y_i]_\epsilon)$ and $\pi_j(\mathbf{x}_i)\, \mathrm{Pl}_{\widetilde{Y}_j(\mathbf{x}_i)}([y_i]_\epsilon)$ if $\delta_i = j$;
  - $\mathrm{Bel}_{\widetilde{Y}(\mathbf{x}_i)}([y_i, +\infty[)$ and $\mathrm{Pl}_{\widetilde{Y}(\mathbf{x}_i)}([y_i, +\infty[)$ if $\delta_i = 0$.



Illustrative parameter values.

CR-ENNreg: results

	MELANOMA ($n = 205,\ J = 2$)		MGUS2 ($n = 1{,}338,\ J = 2$)		EBMT ($n = 8{,}966,\ J = 6$)	
	C-index ↑	IBS ↓	C-index ↑	IBS ↓	C-index ↑	IBS ↓
cs-Cox	0.594	0.126	0.656	0.316	0.584	0.071
Fine–Gray	0.596	0.124	0.647	0.334	0.588	0.070
DeepHit	0.680	0.126	0.686	**0.141**	0.560	**0.069**
CR-SAVAE	0.671	0.167	0.682	0.168	0.558	0.079
CR-ENNreg	**0.705**	**0.116**	**0.687**	0.214	**0.602**	0.071

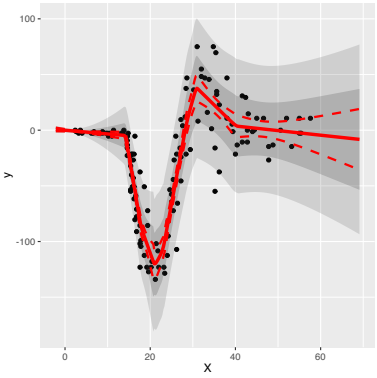
- Averages over event types; mean over 10 random splits.
- CR-ENNreg achieves the best average C-index on all three datasets, with the largest gain on EBMT (six competing events).
- Calibration is competitive on MELANOMA and EBMT but not uniformly better: improving calibration is ongoing work.

Beyond prototypes: the likelihood-based approach

- Evidential likelihood-based inference: the relative likelihood

$$\widetilde{\pi}_{\mathbf{y}}(\boldsymbol{\theta}) = \frac{L_{\mathbf{y}}(\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta}'} L_{\mathbf{y}}(\boldsymbol{\theta}')}$$

  is interpreted as a possibility distribution on the parameter. Adding a probabilistic prior by product-intersection gives back Bayesian inference.
- Prediction: propagating $\widetilde{\pi}_{\mathbf{y}}$ and the random noise U through a prediction equation $Y = \varphi(\mathbf{x}, \boldsymbol{\theta}, U)$ yields a predictive RFS.
- Neural networks: Gaussian approximation and linearisation give a GRFN $\widetilde{\mathcal{N}}\left(\mu(\mathbf{x}), \sigma^2(\mathbf{x}), h(\mathbf{x})\right)$, with $h(\mathbf{x})$ derived from the observed information matrix.



Heteroscedastic neural network on the motorcycle data: point predictions, lower/upper expectations and belief intervals at levels 0.50 and 0.95.

Take-home messages

- Random fuzzy sets provide a unified language for evidence that is both uncertain and fuzzy, encompassing evidence theory and possibility theory.
- Gaussian random fuzzy numbers (and their multivariate, transformed and mixture extensions) make this framework tractable in continuous spaces, with closed-form combination.
- In machine learning, predictions take the form of RFSs, which separate aleatory (σ^2) and epistemic (h) uncertainty and yield lower/upper CDFs and belief intervals.
- In survival analysis, this approach
  - handles censored data as imprecise observations,
  - avoids proportional hazards and time discretisation,
  - provides lower and upper survival functions (or CIFs) with calibrated belief intervals,
  - extends to multimodal fusion (discounting + combination) and competing risks (mixtures).

Main references (1/2): theory

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



T. Denœux

Modern Evidence Theory: A Unified Approach to Uncertainty in AI, Engineering and Statistics.

Springer Nature, to appear.



T. Denœux

Reasoning with fuzzy and uncertain evidence using epistemic random fuzzy sets: general framework and practical models.

Fuzzy Sets and Systems, 453:1–36, 2023.



T. Denœux

Parametric families of continuous belief functions based on generalized Gaussian random fuzzy numbers.

Fuzzy Sets and Systems, 471:108679, 2023.



T. Denœux

Combination of dependent and partially reliable Gaussian random fuzzy numbers.

Information Sciences, 681:121208, 2024.

Main references (2/2): machine learning and survival analysis



T. Denœux

Quantifying prediction uncertainty in regression using random fuzzy sets: the ENNreg model.

IEEE Transactions on Fuzzy Systems, 31(10):3690–3699, 2023.



T. Denœux

Uncertainty quantification in regression neural networks using evidential likelihood-based inference.

International Journal of Approximate Reasoning, 182:109423, 2025.



L. Huang, Y. Xing, S. Mishra, T. Denœux and M. Feng

Evidential time-to-event prediction with calibrated uncertainty quantification.

International Journal of Approximate Reasoning, 181:109403, 2025.



L. Huang, Y. Xing, Q. Lin, J. Duan, S. Ruan and M. Feng

EsurvFusion: An evidential multimodal survival fusion model based on epistemic random fuzzy sets.

IEEE Transactions on Fuzzy Systems, 34(1):76–88, 2026.



C. Mi, S. Ruan, L. Huang, P. Vera and T. Denœux

CR-ENNreg: An evidential survival model for competing risks analysis.

In *Belief Functions: Theory and Applications (BELIEF 2026)*, LNAI 17186, pp. 202–209, Springer, 2027.