Quantification of Prediction Uncertainty using Random Fuzzy Sets

Thierry Denœux

Université de technologie de Compiègne, France

Zhejiang University of Technology, Hangzhou, July 16, 2026

Need for uncertainty quantification

Prediction uncertainty

- Uncertainty quantification is a fundamental problem in statistics and AI.
- In prediction tasks such as, e.g., weather forecasting, we need to tell the user how confident they can be in the predictions.

Aleatory uncertainty

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

Epistemic uncertainty

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

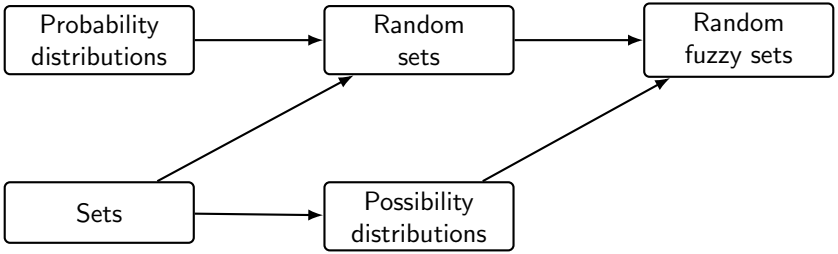
Why probability is not be enough

- Probability distributions are natural for representing random uncertainties.
- But epistemic uncertainties are very important in statistical and model-based prediction, because of
  - Limited sample size,
  - Poor data quality,
  - Model misspecification, etc.
- A single probability distribution may then give a false impression of precision.

Random fuzzy sets

Random fuzzy sets are designed to represent both aleatory and epistemic uncertainties simultaneously.

A hierarchy of uncertainty models



- Random sets extend probability and sets by allowing set-valued realisations.
- Possibility distributions encode graded compatibility.
- Random fuzzy sets combine both ideas.

Outline

1. Uncertainty representation frameworks
   - From mass functions to random sets
   - From random sets to random fuzzy sets
   - Gaussian random fuzzy models

2. Application to machine learning
   - Distance-based approach
   - Likelihood-based approach

Outline

Classical Dempster–Shafer theory

- Let Y be an unknown quantity with value in a finite frame Θ.
- Evidence about Y can be represented by a mass function assigning weights to subsets of Θ:

$$m(A) \geq 0, \qquad \sum_{A \subseteq \Theta} m(A) = 1.$$

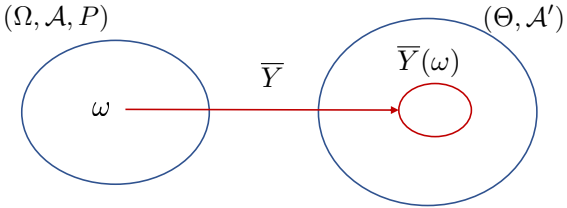
- $m(A)$ is the probability that the evidence only says that $Y \in A$.
- Belief and plausibility are lower and upper probabilities:

$$\mathrm{Bel}_m(B) = \sum_{A \subseteq B} m(A), \qquad \mathrm{Pl}_m(B) = \sum_{A \cap B \neq \varnothing} m(A) = 1 - \mathrm{Bel}_m(B^c).$$

- Interpretation:
  - $\mathrm{Bel}_m(B)$: degree of total support for B;
  - $\mathrm{Bel}_m(B)$: degree of compatibility of B with the evidence.

From finite frames to infinite spaces

Many unknowns take values in $\mathbb{R}^p$, not in a finite set.



- The mass function is replaced by the more general concept of random subset of Θ.
- This leads to the notion of a random set, defined by a probability space $(\Omega, \mathcal{A}, P)$ and a set-valued function

$$\overline{Y} : \Omega \to \mathcal{P}(\Theta).$$

Interpretation

- Here, Ω is a set of interpretations of the evidence.
- For each elementary interpretation ω, we observe a set-valued constraint:

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

Belief and plausibility for a random set

For $B \subseteq \Theta$,

$$\mathrm{Bel}_{\overline{Y}}(B) = P\{\overline{Y} \subseteq B\}, \qquad \mathrm{Pl}_{\overline{Y}}(B) = P\{\overline{Y} \cap B \neq \varnothing\} = 1 - \mathrm{Bel}_{\overline{Y}}(B^c).$$

Belief

Probability that the evidence supports B.

Plausibility

Probability that the evidence is compatible with B (does not support B^c).

Properties:

- $\mathrm{Bel}(\varnothing) = \mathrm{Pl}(\varnothing) = 0$
- $\mathrm{Bel}(\Theta) = \mathrm{Pl}(\Theta) = 1$
- $\forall B \subseteq \Theta$, $\mathrm{Bel}(B) \leq \mathrm{Pl}(B)$.
- If the focal sets are singletons, Bel = Pl is additive.

Random sets as evidence on infinite frames

Key equivalence

A random set is a probability distribution over crisp focal sets.

- This extends belief functions from finite frames to general spaces.
- But it still assumes that each focal element is crisp, i.e., the evidence imposes hard constraints on the unknown quantity.

What if the evidence is itself vague or graded, such as, e.g.,
"Y is close to 1.5", or "Y is small"?

From crisp focal sets to fuzzy focal sets

Random set

$$\overline{Y}(\omega) \subseteq \Theta$$

A crisp constraint:

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

Random fuzzy set

$$\widetilde{Y}(\omega) : \Theta \to [0,1]$$

A soft (graded, flexible) constraint:

$$Y = y \text{ is possible to degree } \widetilde{Y}(\omega)(y).$$

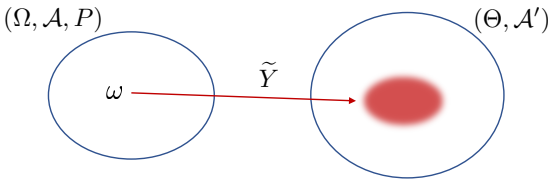
Possibility and necessity generated by a fuzzy set

- A fuzzy subset of Θ is a map $\pi : \Theta \to [0, 1]$. It is normal if $\sup_{\theta \in \Theta} \pi(\theta) = 1$.
- 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:

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

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

Definition of a random fuzzy set



Random fuzzy set

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

$$\widetilde{Y} : \Omega \to \mathcal{F}(\Theta),$$

where $\widetilde{Y}(\omega)$ is a fuzzy subset of Θ.

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

Belief and plausibility induced by a RFS

- For any $\omega \in \Omega$, $\widetilde{Y}(\omega)$ is a possibility distribution for X.
- For any $B \subseteq \Theta$, under measurability conditions, $\Pi_{\widetilde{Y}(\omega)}(B)$ and $\mathrm{N}_{\widetilde{Y}(\omega)}(B)$ are random variables.
- We define:

$$\mathrm{Pl}_{\widetilde{Y}}(B) = \mathbb{E}\left[\Pi_{\widetilde{Y}}(B)\right],$$

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

Interpretation

- Plausibility is the average possibility of B over all uncertain interpretations.
- Belief is the average necessity of B.

Contour function

The plausibility of a singleton gives the contour function:

$$\mathrm{pl}_{\widetilde{Y}}(y) = \mathrm{Pl}_{\widetilde{Y}}(\{y\}) = \mathbb{E}\left[\widetilde{Y}(y)\right].$$

- It is the average membership degree of x over the random fuzzy focal sets.
- It generalises both probability density-like information and possibility distributions.
- It is often easier to visualise than the full random fuzzy set.

One object, several limiting cases

Model	RFS representation	Main meaning
Random variable	random singleton	precise randomness
Random set	random crisp set	imprecise evidence
Possibility distribution	fixed fuzzy set	vague evidence
Random fuzzy set	random fuzzy set	variability + imprecision + vagueness

Unifying view

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

Combination of random fuzzy sets

- Let $\widetilde{Y}_1$ and $\widetilde{Y}_2$ be two independent random fuzzy sets on Θ.
- Their combination $\widetilde{Y}_1 \oplus \widetilde{Y}_2$ is obtained by intersecting their fuzzy focal sets:

$$\widetilde{Y}_1 \oplus \widetilde{Y}_2 : (\omega_1, \omega_2) \mapsto \widetilde{Y}_1(\omega_1) \cap \widetilde{Y}_2(\omega_2).$$

- Intersection is defined as the normalised product:

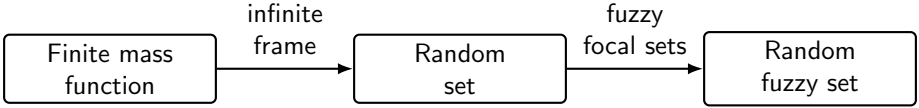
$$(\widetilde{Y}_1(\omega_1) \cap \widetilde{Y}_2(\omega_2))(x) = \frac{\widetilde{Y}_1(\omega_1)(x)\, \widetilde{Y}_2(\omega_2)(x)}{\sup_{u \in \Theta} \widetilde{Y}_1(\omega_1)(u)\, \widetilde{Y}_2(\omega_2)(u)}.$$

- The product probability measure $P_1 \otimes P_2$ is updated to account for the conflict.

Main idea

$\oplus$ combines two independent random fuzzy sets by intersecting their focal fuzzy sets, and renormalising.

Conceptual summary



- Random set: probability distribution over sets.
- Random fuzzy set: probability distribution over possibility distributions.
- Belief and plausibility remain lower and upper probabilities.

Need for practical continuous models

- General RFSs are very flexible but difficult to specify and propagate.
- For numerical uncertainty propagation, we need parametric models.
- Gaussian random fuzzy numbers and vectors provide such a tractable family.
- They are indexed by three parameters representing
  - location,
  - dispersion,
  - imprecision.

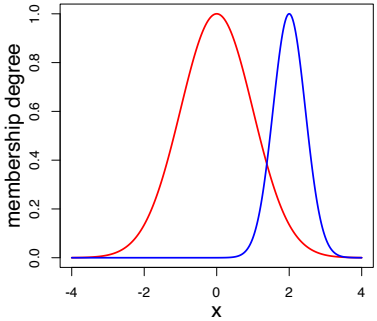
Gaussian fuzzy number

Definition

A Gaussian fuzzy number (GFN) is defined by

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

where $m \in \mathbb{R}$ is the mode and $h \geq 0$ is a precision parameter.



We write

$$\boxed{\widetilde{y} = \mathrm{GFN}(m, h).}$$

Gaussian random fuzzy number

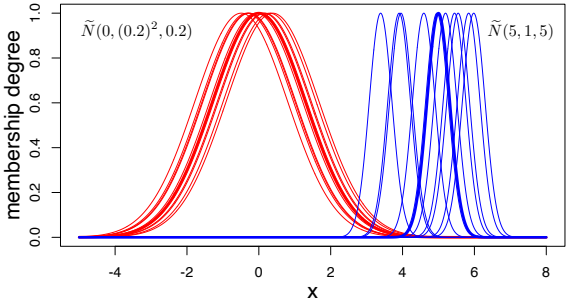
Definition

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

$$\widetilde{Y}(M) = \mathrm{GFN}(M, h),$$

with

$$M \sim \mathcal{N}(\mu, \sigma^2), \qquad h \geq 0.$$



We write

$$\widetilde{Y} \sim \widetilde{\mathcal{N}}(\mu, \sigma^2, h).$$

Interpreting the parameters

- μ is the central value.
- σ^2 measures how much the modes varies across interpretations.
- h measures the specificity of each focal fuzzy number.
- Three limiting cases:

Condition	Interpretation
$\sigma^2 = 0$	Gaussian fuzzy number
$h \to \infty$	Gaussian random variable
$h = 0$	vacuous information

Message

A GRFN continuously interpolates between probability, possibility and ignorance.

Gaussian fuzzy vector

Definition

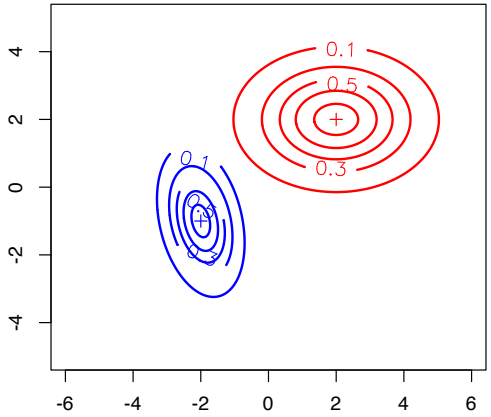
A Gaussian fuzzy vector (GFV) is defined by

$$\widetilde{\mathbf{y}}(\mathbf{y}) = \exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{m})^T\mathbf{H}(\mathbf{y}-\mathbf{m})\right),$$

where $\mathbf{m} \in \mathbb{R}^p$ is the mode and $\mathbf{H} \in \mathbb{R}^{p\times p}$ is a positive definite precision matrix.

We write

$$\boxed{\widetilde{\mathbf{y}} = \mathrm{GFV}(\mathbf{m}, \mathbf{H}).}$$



Gaussian random fuzzy vector

Definition

A Gaussian random fuzzy vector (GRFV) is a GFV whose mode is a normal random vector:

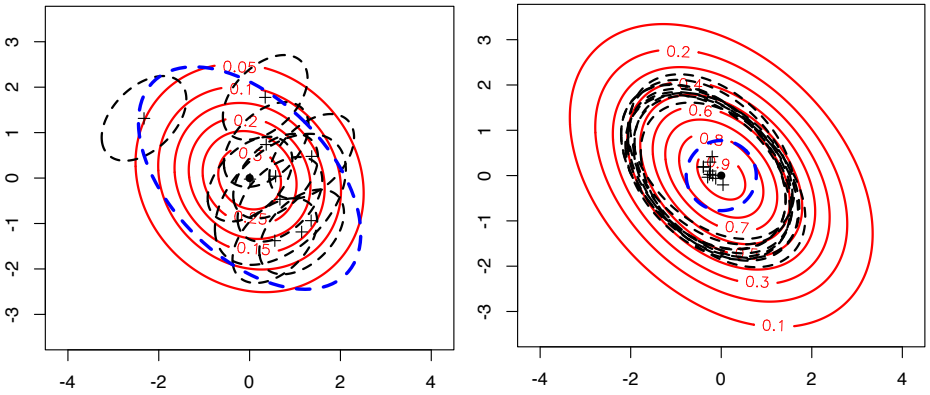
$$\widetilde{\mathbf{Y}}(\mathbf{M}) = \mathrm{GFV}(\mathbf{M}, \mathbf{H}), \qquad \mathbf{M} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \qquad \mathbf{H}, \boldsymbol{\Sigma} \succeq 0,$$

We write

$$\boxed{\widetilde{\mathbf{Y}} \sim \widetilde{\mathcal{N}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{H}).}$$

- $\boldsymbol{\Sigma}$ controls the dispersion of the random centres.
- $\mathbf{H}$ controls the shape and precision of each fuzzy focal ellipsoid.

Representation of GRFVs

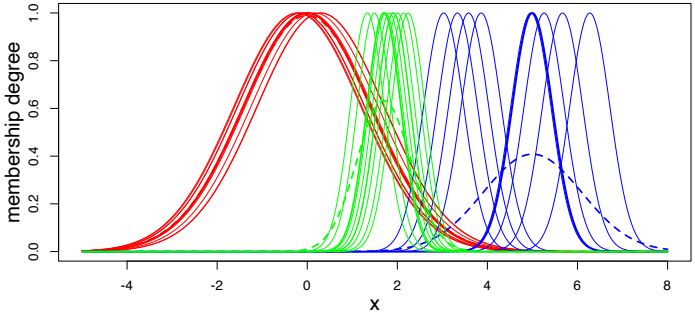


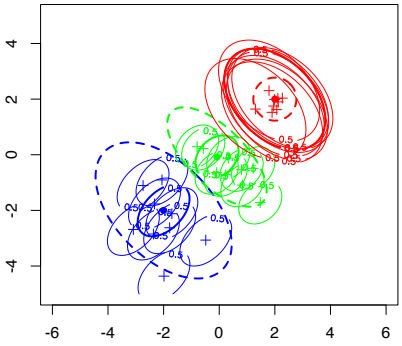
For each GRFV, we show 0.5-level curves of ten focal sets (black broken curves), a contour plot of the contour function (red solid curves), as well as a 95% confidence ellipse for the mode $\mathbf{M} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (blue broken curves).

Combination of Gaussian RFSs

Closeness property

GRFNs and GRFVs are closed under product-intersection.





Interpretability

- From $\widetilde{Y}(\mathbf{x}) \sim \widetilde{\mathcal{N}}\{\mu(\mathbf{x}), \sigma^2(\mathbf{x}), h(\mathbf{x})\}$, we can extract:
  - a point prediction $\mu(\mathbf{x})$;
  - a contour function $\mathrm{pl}_{\mathbf{x}}(y)$;
  - lower and upper CDFs;
  - belief intervals at prescribed confidence levels;
  - a measure of epistemic imprecision through $h(\mathbf{x})$.
- These summaries are interpretable by users who do not know the underlying theory.

Outline

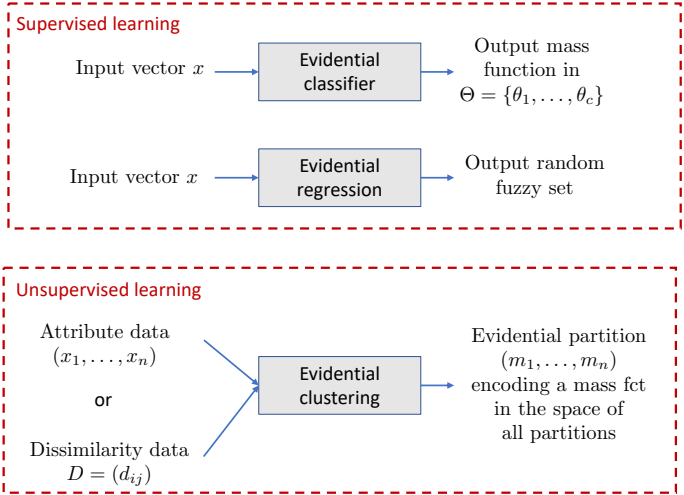
Machine learning: the setting

- Machine learning builds predictive models from data for classification, regression, decision support and knowledge discovery.
- In supervised learning, the learning set is

$$\mathcal{L} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}.$$

- The response may be
  - a class label: $Y \in \Theta = \{\theta_1, \ldots, \theta_c\}$;
  - a numerical value: $Y \in \Theta \subseteq \mathbb{R}$.
- Evidence theory enriches prediction by returning not only a label or value, but also a belief function describing predictive uncertainty.

Evidential machine learning



Distance-based evidential prediction

- The starting idea is simple:

> nearby training points or prototypes provide stronger evidence.

- A prototype close to the query input $\mathbf{x}$ gives an informative predictive statement.
- A prototype far from $\mathbf{x}$ gives weak evidence or even vacuous information.
- 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.

Prototype evidence as a GRFN

- The leaning vectors are summarised by K prototypes $\mathbf{w}_1, \ldots, \mathbf{w}_K$.
- Prototype k supplies
$$\widetilde{Y}_k(\mathbf{x}) \sim \widetilde{\mathcal{N}}\{\mu_k(\mathbf{x}), \sigma_k^2, s_k(\mathbf{x})h_k\}.$$
- A simple local mean is affine:
$$\mu_k(\mathbf{x}) = \beta_k^T \mathbf{x} + \beta_{k0}.$$
- The similarity term is
$$s_k(\mathbf{x}) = \exp\{-\gamma_k \|\mathbf{x} - \mathbf{w}_k\|^2\}.$$
- Close prototypes give high epistemic precision; distant prototypes give weak evidence.

Combining GRFN prototype outputs

A simplified product-intersection rule yields

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

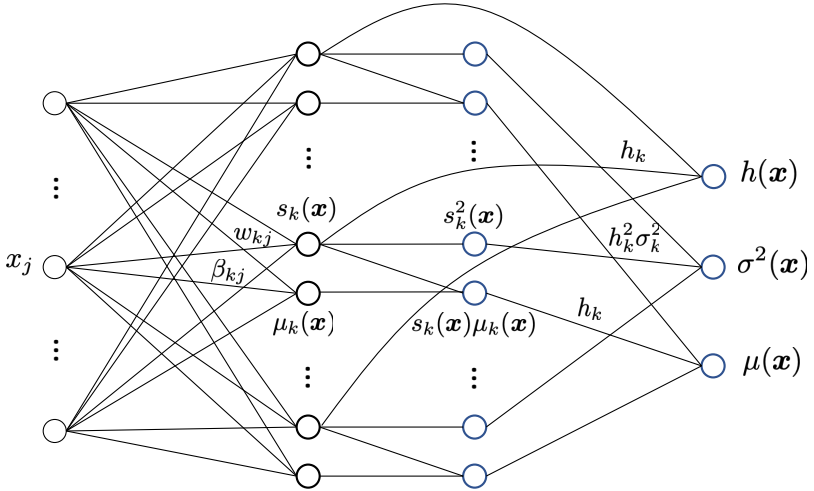
with

$$\mu(\mathbf{x}) = \frac{\sum_{k=1}^{K} s_k(\mathbf{x}) h_k \mu_k(\mathbf{x})}{\sum_{k=1}^{K} s_k(\mathbf{x}) h_k}, \quad h(\mathbf{x}) = \sum_{k=1}^{K} s_k(\mathbf{x}) h_k,$$

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

- $\sigma^2(\mathbf{x})$ captures aleatory uncertainty.
- $h(\mathbf{x})$ captures epistemic uncertainty.

ENNreg architecture



Learning

- The model is trained by minimising

$$C^{(R)}_{\lambda,\epsilon,\xi,\rho}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\lambda,\epsilon}(y_i, \widetilde{Y}(\mathbf{x}_i;\boldsymbol{\theta})) + \frac{\xi}{K}\sum_k h_k + \frac{\rho}{K}\sum_k \gamma_k,$$

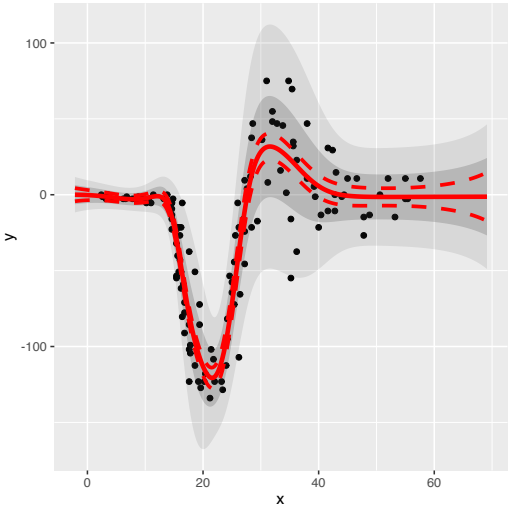
  where

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

  and $[y_i]_\epsilon = [y_i - \epsilon, y_i + \epsilon]$.
- Hyperparameters:
  - ϵ accounts for the finite precision of continuous observations (typically, $\epsilon = 0.01\widehat{\sigma}_Y$);
  - λ controls the trade-off between precise and cautious predictions (typically, $\lambda = 0.9$);
  - ξ and ρ regularise, respectively, the effective number of prototypes and the departure from a linear model. They are determined by cross-validation.

Example: motorcycle data

- Point predictions (solid red curve),
- lower and upper expectations (dashed red curves)
- belief intervals at levels 0.50 and 0.95 (shaded areas).



Comparison with classical methods (RMS)

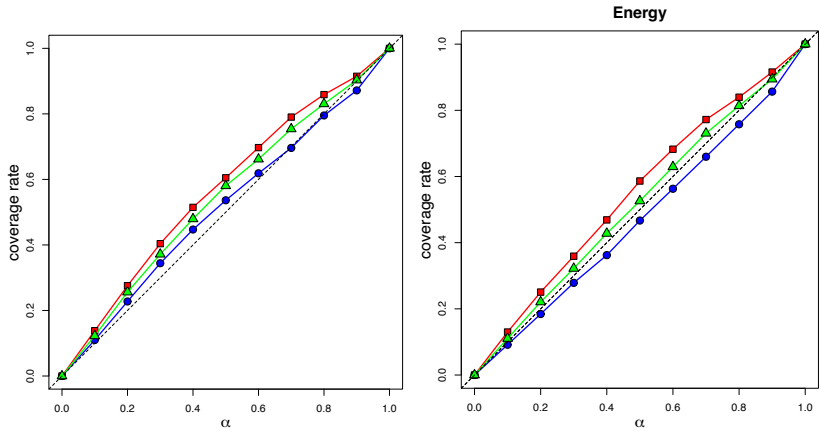
	ENNreg	RBF	RVM	SVM	GP	RF	MLP
Boston	**2.87 ± 0.14**	3.31 ± 0.19	3.42 ± 0.17	3.17 ± 0.15	3.70 ± 0.22	**3.11 ± 0.14**	**3.14 ± 0.14**
Energy	**1.06 ± 0.05**	2.06 ± 0.08	1.79 ± 0.05	1.39 ± 0.06	2.58 ± 0.07	1.75 ± 0.06	**0.95 ± 0.16**
Concr.	5.10 ± 0.12	6.30 ± 0.19	6.38 ± 0.16	5.62 ± 0.13	6.93 ± 0.13	**4.64 ± 0.12**	**4.82 ± 0.16**
Yacht	**0.44 ± 0.04**	2.00 ± 0.20	1.88 ± 0.20	1.93 ± 0.11	6.12 ± 0.31	0.96 ± 0.08	**0.50 ± 0.05**
Wine	0.63 ± 0.01	0.63 ± 0.01	0.80 ± 0.02	0.61 ± 0.01	0.61 ± 0.01	**0.56 ± 0.01**	0.77 ± 0.01
kin8nm	0.08 ± 0.00	0.11 ± 0.00	–	0.09 ± 0.00	0.08 ± 0.00	0.14 ± 0.00	**0.07 ± 0.00**
Crime	**0.14 ± 0.00**	**0.14 ± 0.00**	**0.14 ± 0.00**	**0.14 ± 0.00**	**0.14 ± 0.00**	**0.14 ± 0.00**	**0.14 ± 0.00**
Resid.	**0.11 ± 0.01**	0.16 ± 0.01	0.17 ± 0.01	0.15 ± 0.01	0.22 ± 0.01	0.16 ± 0.01	0.14 ± 0.01
Airfoil	**1.46 ± 0.03**	1.70 ± 0.04	2.58 ± 0.04	2.37 ± 0.04	2.49 ± 0.04	**1.44 ± 0.04**	1.53 ± 0.04
Bike	**6.59 ± 0.19**	**6.49 ± 0.15**	**6.64 ± 0.14**	7.11 ± 0.16	7.55 ± 0.14	6.86 ± 0.17	9.68 ± 0.20

Comparison with SOTA methods (RMS & NLL)

	RMS				
	ENNreg	PBP	MC-dropout	Deep ens.	Deep ev. reg.
Boston	**2.87 ± 0.14**	**3.01 ± 0.18**	**2.97 ± 0.19**	**3.28 ± 1.00**	**3.06 ± 0.16**
Energy	**1.06 ± 0.05**	1.80 ± 0.05	1.66 ± 0.04	2.09 ± 0.29	2.06 ± 0.10
Concr.	**5.10 ± 0.12**	5.67 ± 0.09	**5.23 ± 0.12**	6.03 ± 0.58	5.85 ± 0.15
Yacht	**0.44 ± 0.04**	1.02 ± 0.05	1.11 ± 0.09	1.58 ± 0.48	1.57 ± 0.56
Wine	**0.63 ± 0.01**	**0.64 ± 0.01**	**0.62 ± 0.01**	**0.64 ± 0.04**	**0.61 ± 0.02**
kin8nm	**0.08 ± 0.00**	0.10 ± 0.00	0.10 ± 0.00	0.09 ± 0.00	0.09 ± 0.00

	NLL				
	ENNreg	PBP	MC-dropout	Deep ens.	Deep ev. reg.
Boston	2.53 ± 0.07	2.57 ± 0.09	**2.46 ± 0.06**	**2.41 ± 0.25**	**2.35 ± 0.06**
Energy	**1.14 ± 0.07**	2.04 ± 0.02	1.99 ± 0.02	**1.38 ± 0.22**	1.39 ± 0.06
Concr.	3.38 ± 0.13	3.16 ± 0.02	**3.04 ± 0.02**	**3.06 ± 0.18**	**3.01 ± 0.02**
Yacht	**0.13 ± 0.12**	1.63 ± 0.02	1.55 ± 0.03	1.18 ± 0.21	1.03 ±0.19
Wine	**0.94 ± 0.01**	0.97 ± 0.01	**0.93 ± 0.01**	**0.94 ± 0.12**	**0.89 ± 0.05**
kin8nm	**-1.19 ± 0.00**	-0.90 ± 0.01	-0.95 ± 0.01	**-1.20 ± 0.02**	-1.24 ± 0.01

Calibration plots



Probabilistic predictions (blue), raw evidential predictions (red) and adjusted evidential predictions (green).

A different starting point

- The distance-based approach is non-parametric.
- The likelihood-based approach starts from a statistical model, making it possible to derive a prediction equation

$$Y = \varphi(\mathbf{x}, \boldsymbol{\theta}, U),$$

  where U is a noise variable and $\boldsymbol{\theta}$ is unknown.
- The data inform us about $\boldsymbol{\theta}$ through the likelihood.
- Prediction is obtained by propagating parameter uncertainty and noise through the prediction equation.

Likelihood and relative likelihood

Let $\mathbf{Y}$ have density $f_{\boldsymbol{\theta}}$ and let $\mathbf{y}$ be observed.

Likelihood

$$L_{\mathbf{y}}(\boldsymbol{\theta}) = \alpha\, f_{\boldsymbol{\theta}}(\mathbf{y}),$$

where $\alpha > 0$ is arbitrary.

Relative likelihood

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

- It is normalised: $\sup_{\boldsymbol{\theta}} \widetilde{\pi}_{\mathbf{y}}(\boldsymbol{\theta}) = 1$.
- It ranks parameter values according to their compatibility with the data.

Possibilistic interpretation

- The relative likelihood is interpreted as a possibility distribution on the parameter space:

$$\widetilde{\pi}_{\mathbf{y}} : \Theta \to [0, 1].$$

- $\widetilde{\pi}_{\mathbf{y}}(\boldsymbol{\theta}) = 1$ for the most likely parameters.
- Small values mean that $\boldsymbol{\theta}$ is poorly supported by the data.
- This is a form of Bayesian inference without a probabilistic prior: if a probabilistic prior is added and combined by product-intersection, the Bayesian posterior is recovered.

Gaussian approximation

Let $\widehat{\boldsymbol{\theta}}$ be a maximiser of the log-likelihood. A second-order Taylor expansion gives

$$\log \widetilde{\pi}_{\mathbf{y}}(\boldsymbol{\theta}) \approx -\frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^T \mathbf{J}(\widehat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}),$$

where

$$\mathbf{J}(\widehat{\boldsymbol{\theta}}) = -\left.\frac{\partial^2 \log \widetilde{\pi}_{\mathbf{y}}}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$$

is the observed information matrix.

Approximate possibilistic information on $\boldsymbol{\theta}$

$$\widetilde{\pi}_{\mathbf{y}} \approx \mathrm{GFV}\left(\widehat{\boldsymbol{\theta}}, \mathbf{J}(\widehat{\boldsymbol{\theta}})\right).$$

Simple example: normal mean

- Suppose
$$Y_1, \ldots, Y_n \quad \text{i.i.d. from} \quad \mathcal{N}(\theta, \sigma^2), \quad \sigma^2 \text{ known.}$$
- The relative likelihood is
$$\widetilde{\pi}_{\mathbf{y}}(\theta) = \exp\left\{-\frac{n}{2\sigma^2}(\theta - \overline{y})^2\right\}.$$
- Hence the parameter uncertainty is represented exactly by a Gaussian possibility distribution centred at $\overline{y}$, with precision n/σ^2.
- As n increases, epistemic imprecision decreases.

Prediction equation

- For a new input $\mathbf{x}$, write the predictive model as

$$Y = \varphi(\mathbf{x}, \boldsymbol{\theta}, U),$$

  where U is random with known distribution.
- Parameter uncertainty is possibilistic:

$$\boldsymbol{\theta} \sim \widetilde{\pi}_{\mathbf{y}}.$$

- Noise uncertainty is probabilistic:

$$U \sim P_U.$$

- Combining them naturally yields a random fuzzy set for Y.

Constructing the predictive RFS

- For each simulated noise value u, propagate the parameter possibility distribution:

$$\widetilde{\pi}_{\mathbf{y},u}(y) = \sup_{\boldsymbol{\theta}:\, y=\varphi(\mathbf{x},\boldsymbol{\theta},u)} \widetilde{\pi}_{\mathbf{y}}(\boldsymbol{\theta}).$$

- Since U is random, the map

$$\widetilde{Y} : u \mapsto \widetilde{\pi}_{\mathbf{y},u}$$

  is a random fuzzy set.
- The resulting predictive RFS combines:
  - aleatory uncertainty from U;
  - epistemic uncertainty from $\widetilde{\pi}_{\mathbf{y}}$.

Linear-Gaussian case

- Suppose the prediction equation is linear:

$$Y = a(\mathbf{x}) + \mathbf{b}(\mathbf{x})^T \boldsymbol{\theta} + \sigma(\mathbf{x})U.$$

- If $U \sim N(0,1)$ and $\widetilde{\pi}_{\boldsymbol{\theta}}$ is Gaussian, the predictive RFS is a GRFN.
- This gives a closed-form approximation

$$\widetilde{Y}(\mathbf{x}) \sim \widetilde{\mathcal{N}}\left(\mu(\mathbf{x}), \sigma^2(\mathbf{x}), h(\mathbf{x})\right).$$

- The three parameters have direct predictive interpretations.
- These results hold approximately if the prediction equation is linearised about $\widehat{\boldsymbol{\theta}}$.

Exact and approximate workflows

- Exact propagation is possible in simple models.
- Monte Carlo propagation can be used when closed forms are unavailable.
- Gaussian approximation is often sufficient for large samples or weakly nonlinear models.
- A practical workflow is:
  1. fit the statistical model;
  2. compute or approximate the relative likelihood;
  3. propagate through the prediction equation;
  4. summarise the predictive RFS by contours, CDF bounds, and belief intervals.

Heteroscedastic regression model

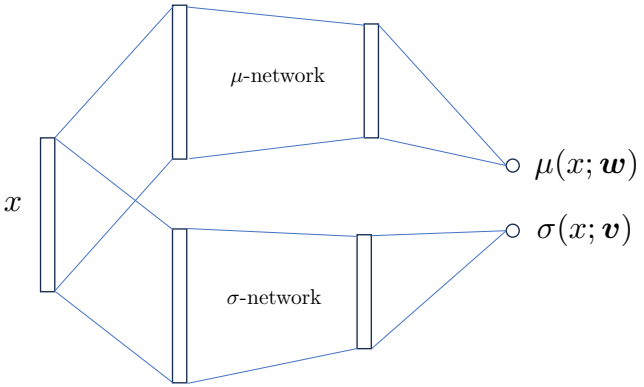
- In many prediction problems, the noise variance depends on the input.
- A neural network can output both

$$\mu(\mathbf{x};\mathbf{w}), \quad \sigma^2(\mathbf{x};\mathbf{v}).$$

- The prediction equation is

$$Y = \mu(\mathbf{x};\mathbf{w}) + \sigma(\mathbf{x};\mathbf{v})U, \quad U \sim \mathcal{N}(0,1).$$

- The likelihood is used to build a possibility distribution over all network parameters.



Parameter possibility distribution

- Let $\widehat{\boldsymbol{\theta}} = (\widehat{\mathbf{w}}^T, \widehat{\mathbf{v}}^T)^T$ be the maximum likelihood estimate.
- The relative likelihood is approximated by

$$\widetilde{\pi}_{\mathbf{y}}(\boldsymbol{\theta}) \approx \exp\{-\tfrac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^T \mathbf{J}(\widehat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})\}.$$

- The information matrix has a block structure separating mean and variance parameters.
- This approximation turns parameter uncertainty into a tractable Gaussian possibility distribution.

Linearisation for prediction

- For a fixed input $\mathbf{x}$, linearise both networks around $\widehat{\boldsymbol{\theta}}$:

$$\mu(\mathbf{x};\mathbf{w}) \approx \mu(\mathbf{x},\widehat{\mathbf{w}}) + \mathbf{g}_\mu(\mathbf{x})^T(\mathbf{w}-\widehat{\mathbf{w}}),$$

$$\sigma(\mathbf{x};\mathbf{v}) \approx \sigma(\mathbf{x},\widehat{\mathbf{v}}) + \mathbf{g}_\sigma(\mathbf{x})^T(\mathbf{v}-\widehat{\mathbf{v}}).$$

- This yields an approximate predictive equation

$$Y \approx \begin{pmatrix} \mathbf{g}_\mu(\mathbf{x}) \\ U\mathbf{g}_\sigma(\mathbf{x}) \end{pmatrix}^T \boldsymbol{\theta} + \mu(\mathbf{x};\widehat{\mathbf{w}}) - \mathbf{g}_\mu(\mathbf{x})^T\widehat{\mathbf{w}} + U\left(\sigma(\mathbf{x};\widehat{\mathbf{v}}) - \mathbf{g}_\sigma(\mathbf{x})^T\widehat{\mathbf{v}}\right), \quad U \sim \mathcal{N}(0,1)$$

- This is linear in the full parameter vector $\boldsymbol{\theta}$. Hence Gaussian possibilistic propagation can be applied.
- The resulting RFS can be further approximated by a fixed-precision GRFN.

Interpretation of the predictive GRFN

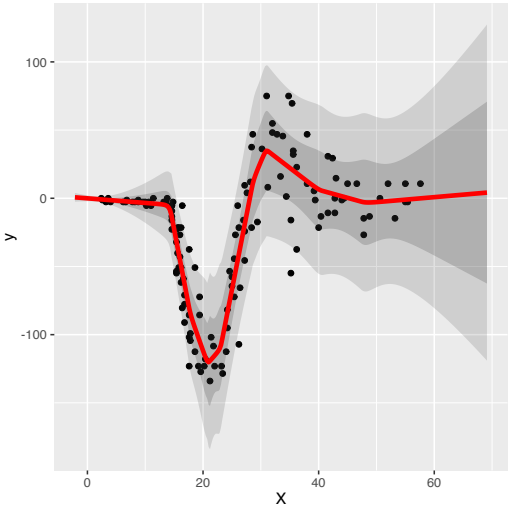
Final predictive approximation

$$\widetilde{Y}(\mathbf{x}) \overset{\text{app.}}{\sim} \widetilde{\mathcal{N}}\left(\mu(\mathbf{x};\widehat{\mathbf{w}}), \sigma^2(\mathbf{x};\widehat{\mathbf{v}}), h(\mathbf{x},\widehat{\boldsymbol{\theta}})\right).$$

Component	Interpretation
$\mu(\mathbf{x};\widehat{\mathbf{w}})$	point prediction / expected response
$\sigma^2(\mathbf{x};\widehat{\mathbf{v}})$	aleatory variability of Y given x
$h(\mathbf{x},\widehat{\boldsymbol{\theta}})$	epistemic precision due to finite data and parameter uncertainty

Example: motorcycle data

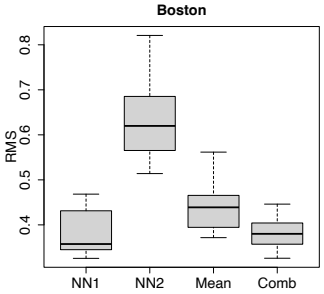
- Point predictions (solid red curve),
- belief intervals at levels 0.50 and 0.95 (shaded areas)
- μ-network: 20 hidden units; σ-network: 10 hidden units.

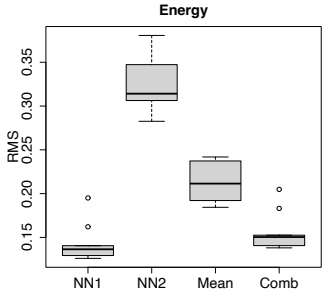


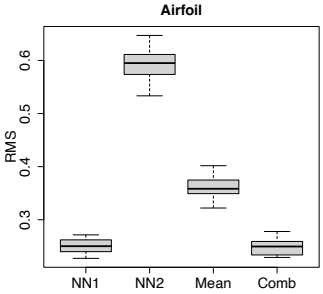
Experiment

- The reliability of predictions depends on the size of the training set.
- To simulate another fusion scenario involving the combination of predictions with different reliabilities, we combined the outputs of two neural networks, trained with 60% and 10% of the data.
- The process was repeated ten times with different random training/validation/test partitions.

Results







Box plots of root-mean-squared errors for the five datasets and the following methods: neural network trained on 60% of the data (NN1), neural network trained on 10% of the data (NN2), average prediction (Mean) and combination by the product-intersection rule (Comb).

Distance-based vs. likelihood-based approaches

Distance-based	Likelihood-based
Evidence induced by distance to prototypes	Evidence induced by relative likelihood of parameters
Good for flexible local models and deep features	Good when a prediction equation is available
Uncertainty learned from geometry and loss	Epistemic uncertainty tied to likelihood theory
Often easy to visualise	Often easier to justify statistically

Take-home messages

- Random fuzzy sets provide a language for predictive uncertainty with both aleatory and epistemic components.
- Belief and plausibility can be introduced operationally as lower and upper support for events.
- Distance-based methods build predictive RFSs from local evidence supplied by neighbours or learned prototypes.
- Likelihood-based methods build predictive RFSs by interpreting relative likelihood as a possibility distribution and propagating it through the prediction equation.
- For users, the practical outputs are point predictions, contours, lower/upper CDFs, belief intervals, and extrapolation warnings.

References

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

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
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