A framework for learning fuzzy rule-based models with epistemic set-valued data and generalized loss functions

Luciano Sánchez[a,*], Inés Couso[b]

[a] *Universidad de Oviedo, Departamento de Informática, Campus de Viesques, 33071 Gijón, Asturias, Spain*
[b] *Universidad de Oviedo, Departamento de Estadística e I.O. y D.M., Campus de Viesques, 33071 Gijón, Asturias, Spain*

Abstract

A framework is proposed for learning fuzzy rule-based systems from low quality data where the differences between observed and true values may introduce systematic bias in the model. It is argued that there are problems where aggregating imprecise losses into numerical or fuzzy-valued risk functions discards useful information, thus generalizing the risk of a model to a vector of fuzzy losses is preferred. The principles governing a learner that is capable of optimizing these fuzzy multivariate risk functions are discussed. Illustrative use cases are worked to exemplify those situations where new framework could become the alternative of choice.

Keywords: Fuzzy Rule-based Models, Soft Computing, Imprecise Data.

1. Introduction

The term *uncomfortable science* was coined by Tukey [1] to describe cases where an inference must be drawn from a limited sample of data because the collection of further data is not feasible. Inducing knowledge from limited datasets poses a completely different set of challenges than big data analysis because, apart from computational efficiency considerations, small sets of data force the researcher to use the same information for both exploratory data analysis

*Corresponding author
Email address: luciano@uniovi.es (Luciano Sánchez)

(unveiling cause/effect relationships) and confirmatory data analysis (testing whether these relationships are supported by the data) [2]. Systematic bias is potentially introduced [3] which may invalidate the confirmatory analysis. Post-hoc theorizing from small datasets is still possible [4], but a great effort must be made to ensure that all the available data is exploited to the full. Discarding "less than perfect" data is not an option for small datasets.

In this paper, some techniques for exploiting "less than perfect" data are studied. Datasets comprising incomplete observations, intermediate between precise perceptions and missing data, will be considered. These will be referred to as "low quality data". An encompassing treatment of these data as epistemic fuzzy sets is adopted. In this last respect, the differences between "epistemic" and "ontic" fuzzy models will also be recalled in this section.

1.1. Low quality data

Generally speaking, "low quality data" comprises those cases where the inaccurate perception of the data can introduce systematic bias in the learning, as shown in the following prototypical examples:

- Discretized data ($a \leq x \leq b$, with x being the actual value and a, b perceived bounds) is very common because digital processing introduces this kind of uncertainty. For instance, if the true weight of an object is 7.3 and a digital scale displays "7" then our information about the weight is $6.5 \leq \text{weight} < 7.5$.

- Censored ($x \leq a$). For instance, in studies in mortality rate, the most that can be said about the death date of live individuals is that it is higher than the current date.

- Restrictions in multiple variables ($f(x_1, x_2, \ldots) = 0$). For instance, the marks of one student, given the class average score, or a high resolution image, given low resolution images [5].

- Tolerance intervals ($P(a \leq x \leq b) \geq 1 - \alpha$), that are slightly more informative than coarsely discretized data. For instance, a GPS sensor may

indicate us that our position is in a circle of radius 5 m., centered at a point in the map, and this information is guaranteed to be valid at least 95% of times. GPS sensors operate this way because the size of a radius that is true 100% of times would be too large for any practical purposes. Moreover, stacked tolerance intervals may be given for different degrees of confidence, i.e. the radius is lower than 5 m. more than 95% of times, lower than 10 m. 99% of times, etc. Under the level-cut representation of a fuzzy set, stacked tolerance intervals are fuzzy subsets of the observable space [6].

In all these cases, the key for avoiding the systematic bias is to distinguish the fact that an event is observed ("the value displayed at the scale is 7") from the fact that an event has happened ("the weight of the object is 7.3"). This problem has been realized early [7][8][9][10] but has not received a complete treatment until recently [11][12].

Observe that incomplete observations can be either crisp or fuzzy subsets of complete observations. However, not all of the interpretations of a fuzzy set are pertinent to this problem. This paper regards the "epistemic" or "disjunctive" approach, where sets are used to describe an incomplete knowledge about the vector of attributes and/or the response variable [13].

1.2. Context and aim of the study

In this paper, learning a model is understood as determining the model with the best empirical risk, given a sample or "training dataset" and a parametric family of models. We also consider that uncertain measurements are not different than partially missing values: the degree of knowledge about a value can be complete (precise data), null (missing data) or partial (imprecise data).

It is well known that missing data can be either removed or imputed while the data is preprocessed, and the same can be said about partially missing data, that can be either removed via instance or feature selection, or imputed (although the term "defuzzification" is preferred for fuzzy data). In turn, imputation can be single (the uncertain value is replaced by a precise measurement) or multiple (the

uncertain value is replaced by a set of precise measurements). Lastly, multiple imputation can be combined with also multiple models (a different model is learnt for each of the surrogate measurements) or a single model (whose risk is different for each of the surrogate measurements). The drawback of learning multiple models is that predictions become set-valued (one prediction for each model) but learning a single model is also troublesome because in this last case it is the empirical risk that becomes set-valued ("fuzzy fitness") and the learning algorithm must be designed in accordance.

This paper is about this last research line. We are interested in epistemic fuzzy models that operate with generalized definitions of loss and risk, thus a single model can be learnt without the need of preprocessing the data for removing the uncertainty. In this context, this paper aims to illustrate some cases where current techniques for learning fuzzy models from imprecise data still have room for improvement, in connection with recent advances in the use of generalized loss functions in machine learning [14].

The paper is organized as follows: in Section 2, a brief study about the state of the art in epistemic models for imprecise data is given, along with a discussion about the weakness of the current models. It is shown that none of the existing methods is better than the others, and also that not all the available information is being used in the learning, because models with different merits can be assigned the same risk. Because of this, new criteria for performing compared evaluations of fuzzy models are proposed in Section 3. These new criteria do not aggregate imprecise losses into fuzzy-valued risks, but the risk of a model is replaced by a vector of fuzzy losses. The principles governing a learner that would be capable of optimizing these fuzzy multivariate risk functions are discussed. In Section 4 three case studies are worked to illustrate those cases where the new framework is expected to have a competitive advantage. Section 5 concludes the paper. Lastly, the metaheuristic that was developed in order to preview the results of the new research line in the cases of study is described in the Appendix.

2. State of the art about epistemic models for imprecise data

When data is partially incomplete, machine learning is not straightforward. There are studies about the conditions for which the difference between true and observed data can be ignored. Others support learning a different model for each possible completion, in a process that shares certain points in common with multiple imputation. Finally, epistemic interval-valued or fuzzy representations of the incomplete data can be used, and this last representation can also be understood as an (implicit) set-valued imputation. The most relevant lines of research about these four subjects are reviewed in the present section, along with a critical view and future research paths.

2.1. Ignorability, MAR and CAR models and the AI&M algorithm

Initial studies about imprecise data focused in defining "ignorability" conditions [7], or conditions that must be assumed thus the learning algorithm can ignore the distinction between observed and true data. Examples of these conditions are the MAR or "missing at random" assumptions [7] or its counterpart CAR or "coarsening at random" [10], where the term "coarsening" is a particularization of what it is being called "low quality data" in this paper: "coarsening" refers to crisp subsets of the space in which the true data lie, and "low quality data" also includes fuzzy subsets. It has been shown that, if the data is CAR, probabilistic inferences can treat the observed data as it was grouped data, but large deviations may occur if the same inference is applied to non-CAR data [10]. Recent strategies for learning from imprecise data that is not CAR also exist, see for instance Jaeger's AI&M (Adjusting Imputation and Maximization) algorithm [15], however the relevant result from the point of view of fuzzy modeling is that, if MAR or CAR hypothesis would apply, then ordinary fuzzy models (i.e. fuzzy models for crisp data) can be applied to the perceived data as if it were the true data.

2.2. Conservative approach

The term "conservative" was coined by Jaeger [15] to refer to an interval-valued estimation of the model parameters obtained for each possible completion

of the data [16].This can be seen as an interval-valued imputation, where imprecise data are replaced by the tightest intervals that contain the true value of each variable. The interval-valued estimation of the model parameters is refined afterwards to a point estimate.

Many algorithms for regression with fuzzy data and "fuzzy coefficients" can be regarded as fuzzy counterparts of the conservative approach, see for instance the linear regression model in [17] or the neural network for fuzzy data in [18]. Moreover, fuzzy data can be understood as a possibilistic extension of the interval-valued imputation mentioned before.

As mentioned in the introduction, different foundations have been proposed for the fuzzy regression problem, and not all of them consist in a fuzzy-conservative inference. In short, those algorithms that match the "epistemic" view are related to the conservative approach, while "ontic" algorithms are not (see for instance [19]).

2.3. Data disambiguation

One of the latest results related to the learning of models from interval or fuzzy data has been proposed by Hüllermeier [20]. Data disambiguation is a novel process that shares certain points in common with multiple imputation. Imprecise data is disambiguated (the opposite operation of coarsening) but disambiguated values do not follow a separate statistical model for the uncertainty, but they depend on the same model that is being learnt. Under these assumptions, both the model and the best disambiguation of the data are learnt at the same time by means of the "minimin" strategy for model selection under uncertainty, and a loss function is produced that depends on the parametric model and also on the fuzzy memberships of the training data. It is found that, for certain common combinations of model and membership, the resulting loss functions mimic ϵ-insensitive losses for regression and hinge or exponential losses for classification.

2.4. Possibilistic risk functions

As mentioned before, the conservative approach consists in learning a parametric model for each possible completion of the data in the first place, then combining these parameters into a joint interval-valued set of parameters and lastly selecting the model given by the center-points of these intervals. The possibilistic loss function is akin to this view because all possible completions of the data are considered too, but it is different in that it is a single parametric model that is evaluated for all the completions, not a different model for every completion. This single model is assigned a different loss for each of the possible completions of every instance. In the interval case, these losses are aggregated first into an interval-valued loss for each instance, and interval losses are aggregated to obtain the risk function [21][22].

The fuzzy case consists in repeating the risk assignment for each level cut of the training data, giving rise to a fuzzy risk function [23]. The fuzzy risk consists in a conservative estimation of the true risk of the model, that is the result of applying the extension principle to the crisp risk function, given the fuzzy data. Standard learning algorithms cannot minimize fuzzy-valued functions and as such specialized algorithms were defined that depend on the definition of a partial order among fuzzy losses, see for instance [24] for the regression problem and [25] for classifiers.

2.5. Limitations of the current state of the art and future directions

There is not an approach that is better than the others. The best technique must be chosen for the case at hand. Not all datasets fulfill MAR and CAR assumptions that are also limited to coarse data (i.e. non-fuzzy) [26], but learning from CAR-compliant datasets require comparatively few computational resources. On the contrary, optimizing possibilistic risk functions is computationally very intensive. It has been argued against the conservative approach that an equal treatment for all instantiated models is not reasonable because it does not makes use of the model assumptions [20]. On the other side, using the same assumptions for imputing data and learning the model may also be

arguable [27], as exploratory analysis and confirmatory analysis would be using the same data.

The following example serves to show up some limitations of the current state of the art in learners for low quality data. In this case study, the risk of two models is evaluated over a dataset comprising coarse data. It will be shown that the two methods that fit this problem, which are "data disambiguation" and "possibilistic risk", lead to the wrong decision if the output contains stochastic noise.

Example 2.1. *The following dataset comprises three input-output pairs of the form* $y = 2x + \mathcal{U}(-1.5; 1.5)$*, where* x *is the input,* y *is the output and* $\mathcal{U}(-1.5; 1.5)$ *is a random value in the interval* $[-1.5, 1.5]$*. Suppose that the observed input at the second instance is an interval:*

instance	*true x*	*observed x*	y
1	$x_1 = 1$	$x_1 = 1$	$y_1 = 2.1$
2	$x_2 = 2$	$2 \leq x_2 \leq 3$	$y_2 = 4.5$
3	$x_3 = 3$	$x_3 = 3$	$y_3 = 4.8$

Let us evaluate the empirical risks of the true model ($y = 2.0x$) *and another arbitrary linear model* ($y = 1.6x$) *by means of their respective Mean Squared Errors (MSE), where*

$$MSE = \sum_{i=1}^{3} (model(x_i) - y_i)^2 : \tag{1}$$

1. *Model 1:* $y = 2.0x$*. MSE = (0.01 + 0.25 + 1.44)/3 = 0.56*
2. *Model 2:* $y = 1.6x$*. MSE = (0.25 + 1.69 + 0.00)/3 = 0.65*

If the true input data was observable, model #1 would be preferred. However, the perceived empirical risks are intervals:

1. *Model 1:* $y = 2.0x$. *MSE = (0.01 + [0.00,2.25] + 1.44)/3 = [0.48,1.23]*

2. *Model 2:* $y = 1.6x$. *MSE = (0.25 + [0.00,1.69] + 0.00)/3 = [0.08,0.65]*

Both intervals contain the true empirical risk, and are non-disjunct. In order to decide whether model 1 is preferred to model 2 or not, an interval ranking must be used. In this example, if the minimim criterion is used, model #2 is preferred because $0.08 < 0.48$. *If the minimax criterion is used, model #2 is preferred again because* $0.65 < 1.23$. *That is to say, the coarsening of the input can cause that the bounds of the perceived risk are lower at the wrong model.*

Lastly, if these models are used to disambiguate the imprecise perceptions of the input, the situation is the same:

instance	*true x*	*observed x*	*disambiguated x (Model #1)*	*disambiguated x (Model #2)*	*y*
1	$x_1 = 1$	$x_1 = 1$	$x_1 = 1$	$x_1 = 1$	$y_1 = 2.1$
2	$x_2 = 2$	$2 \leq x_2 \leq 3$	$x_2 = 2.25$	$x_2 = 2.81$	$y_2 = 4.5$
3	$x_3 = 3$	$x_3 = 3$	$x_3 = 3$	$x_3 = 3$	$y_3 = 4.8$

The empirical risks of the disambiguated models match those obtained when the minimin criterion:

1. *Model 1:* $y = 2.0x$. *MSE = (0.01 + 0.00 + 1.44)/3 = 0.48*

2. *Model 2:* $y = 1.6x$. *MSE = (0.25 + 0.00 + 0.00)/3 = 0.08*

With this example it was shown that those learning methods that propagate the uncertainty in the training data to the fitness function and then define an ordering among these generalized (interval or fuzzy) fitness values (see for instance [28]) heavily depend on the (also interval or fuzzy) ranking of choice. A ranking that is good for certain practical application may be wrong for a different application.

In a different subject, in this paper it is claimed that the expected risk (which is the average of the losses) might discard information that helps to make an informed decision. As mentioned in the introduction, this is not admissible: every bit of information must be kept in low quality datasets. In plain words, the average risk is not enough for learning models from low quality data, and we suggest that the losses at each training instance must be kept. Consider this second example:

Example 2.2. *The following dataset comprises three input-output pairs where the output variable is imprecisely observed (coarsely discretized data):*

instance	*observed x*	*true y*	*noisy y*	*observed y*
1	$x_1 = 1$	$y_1 = 2$	*2.1185*	$2.118 \leq y_1 \leq 2.119$
2	$x_2 = 2$	$y_2 = 4$	*4.7575*	$4.757 \leq y_2 \leq 4.758$
3	$x_3 = 3$	$y_3 = 6$	*6.9800*	$5.979 \leq y_3 \leq 7.981$

Let the risk of a model be defined by its Mean Squared Error (MSE), as before. The observed empirical risks of the true model $y = 2x$ and a second, arbitrary model $y = 2.653x$ are:

1. *Model 1: $y = 2x$. MSE = ([0.286,0.287] + [0.300,0.302] + [0.000,3.924])/3 = [0.196,1.504]*
2. *Model 2: $y = 2.653x$. MSE = ([0.014,0.014] + [0.573,0.575] + [0.000,3.924])/3 = [0.196,1.504]*

Observe that the observed empirical risks of the two models are the same, but the losses at each instance are different. Although the particular problem in this example is not of practical relevance, it illustrates the fact that choosing a model on the only basis of its imprecise risk may lead to ambiguities.

With this example it is shown that different models may share the same bounds of the empirical risk, being indistinguishable no matter the ranking.

Table 1: Main research lines in regard to learning models with low quality data

	Risk		
Imputation mechanism	Numerical	Interval/Fuzzy	Interval/Fuzzy Vector
Crisp	MAR and CAR models	N/A	N/A
Interval/Fuzzy	AI&M algorithm Conservative Data Disambiguation	Possibilistic risk	**Vector losses**

It is fast less common (although still possible) that different models share the same (interval or fuzzy) losses at each training instance. In this paper it is claimed that problems can be conceived where a good model cannot be learnt from imprecise data if the learner optimizes the empirical risk, but the same data leads to a valid model when the learning algorithm is based on a partial order defined over the (multivalued) empirical losses at each training instance. In other words, we suggested that models are not assigned (fuzzy) risks but *vectors of vague losses*. The new research line is positioned against the four mentioned approaches for learning models from low quality data in Table 1, where all methods are organized according to its implicit imputation mechanism (crisp, interval or fuzzy) and the nature of the risk function (also crisp, interval or fuzzy).

3. Learning models with low quality data: comparing vectors of fuzzy losses

The use of Vectors of Losses (VL) is less nonspecific than using generalized risks, but related machine learning algorithms are non-existent. VL algorithms would ultimately depend on an operator for making pairwise comparisons between models, that is introduced in this section.

The notation and other concepts that will be used in the forthcoming discussion are defined in Subsection 3.1. In Subsection 3.2 an algorithmic description is introduced in contraposition with the nearest paradigm, that is "possibilistic risk". This explanation is followed in Subsection 3.3 by a discussion about the

different criteria for sorting lists of vectors of fuzzy losses, where a stochastic order is studied that includes the same order used in "possibilistic risk" algorithms as a particular case. A brief discussion about the consequences of an incomplete knowledge about the function used for computing the losses is included in Section 3.4. Lastly, some issues related to the use of the proposed orders and their connection with multicriteria learning are discussed in Subsection 3.5.

3.1. Notation

In the following, ω is a training instance and $X(\omega)$ is the input vector, comprising crisp measurements of the features of ω. $\widetilde{X}(\omega)$ are the fuzzy observations of these features. $Y(\omega)$ is the output at ω, and $\widetilde{Y}(\omega)$ is the fuzzy observed output. θ is a parameter and Θ is the parametric space, $\theta \in \Theta$. A model is a function $f : \mathcal{X} \to \mathcal{Y}$ relating some response variable $Y : \Omega \to \mathcal{Y}$ to a collection of attributes $\mathbf{X} : \Omega \to \mathcal{X}$, both of them defined on the same population Ω. The output of the model for fuzzy data $f(\widetilde{X}(\omega))$ is computed with the help of the extension principle,

$$f(\widetilde{X}(\omega))(y) = \sup\{\widetilde{X}(x) : y = f(x)\}. \tag{2}$$

Fuzzy Rule Based Models (FRBMs) will be regarded as a particular type of parametric model $f_\theta : \mathcal{X} \to \mathcal{Y}$, where the membership functions of the linguistic variables and the rules in the Knowledge Base (KB) are represented by a list θ of numerical parameters. In other words, the parameter vector θ is a list of numbers that has two parts:

- A representation of the fuzzy partitions of input and output variables as a chain of real numbers.
- A list of integers comprising the antecedent-consequent pairs of the rules in the KB [29]. These rules are of either one of the two following types:
    1. "`if` $\mathbf{X}$ `is` $\widetilde{A}$ `then` Y is $\widetilde{C}$" where $\widetilde{A}$ and $\widetilde{C}$ are fuzzy sets (Mamdani rules, [30]).

2. "`if` $\mathbf{X}$ `is` $\widetilde{A}$ `then` $Y = M\mathbf{X} + B$", where $Y = M\mathbf{X} + B$ denotes an affine transformation of X (TSK models, [31][32]).

The purpose of this definition is to cast FRBMs as particular cases of parametric models thus learning FRBMs from data becomes the same statistical problem as estimating the parameter θ of a model f_θ from an imprecisely observed random sample. In this respect, learning a FRBM will not receive a different treatment in this paper than learning a linear regression model, a neural network or a decision tree: from an abstract point of view, learning any type of model from data consists in solving an optimization problem that aims at minimizing the loss at each instance, according to some loss function $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ that assigns a specific value to every pair $(Y(\omega), f(\mathbf{X}(\omega))$, composed by the outcome of the response variable and its estimate based on the collection of attributes, for every individual $\omega \in \Omega$. Common loss functions for regression models are the square and the absolute value of the difference $\Delta(y, \hat{y}) = (y - \hat{y})^2$ and $\Delta'(y, \hat{y}) = |y - \hat{y}|$, respectively, or the 0-1-valued loss $\Delta(y, \hat{y}) = 1_{\hat{y} \neq y}$ for classification models. The optimization algorithm, however, will be very much different if a regression model, a decision tree or a FRBM is being learnt. Lastly, in the following the acronym "PR" means 'Possibilistic Risk" and "VL" stands for "Vector Losses".

3.2. Algorithmic description of the computation of a vector of losses

The graph of the computations that are carried out when a vector of fuzzy losses is computed is depicted in the lower part of Figure 1, and it is compared to the relevant part of the algorithm of one of the PR models. Let us first restrict the explanation to crisp data and regression problems, and extend it to fuzzy data later.

Suppose that two models f_A and f_B are being compared. Given a crisp dataset, the output of either model is a vector of elements of the output space. The outputs of the models are lists of the same size as the training dataset. In the PR approach, the output of the model and the desired outputs are subtracted, and the differences squared and averaged: the model with a lower

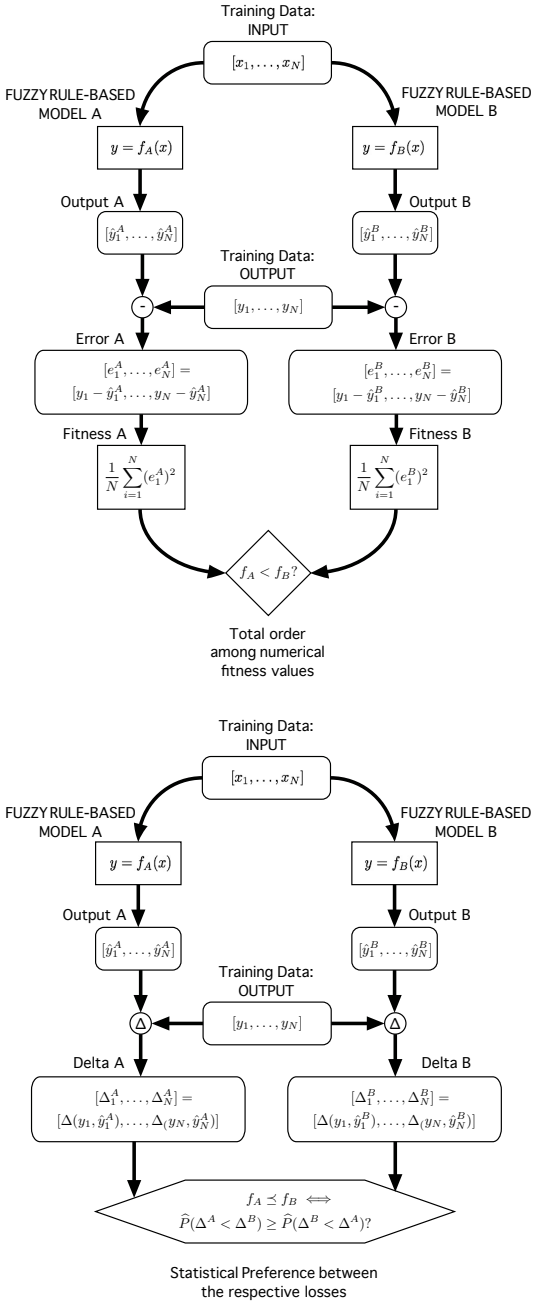

Figure 1: Graph of the computations involved in a comparison between two models in "possibilistic risk" (upper part) and "vector losses" (lower part) for crisp data and a numerical Δ function.

average is the best. In the VL approach, the output of the model and the desired output are composed by means of a function Δ that computes the loss at each instance. Δ is an increasing function of the differences between data and predictions, see Section 3.1. In this case, the results are not averaged: the whole list of losses is used when the models are compared. Specific methods (labelled "Statistical Preference" in the graph) are needed for doing this task, that will be discussed in Section 3.3.

Let us extend this definition to fuzzy data with the help of Figure 2, where imprecise variables have been colored in red. In this case, the training data is fuzzy thus the computation of the output of the models requires applying the extension principle, as seen in the preceding subsection. In PR models, this fuzzy output was subtracted from the observed data with a fuzzy arithmetic operator, resulting in a set-valued error and an also set-valued expected loss. The losses of the two models were compared by means of a fuzzy ranking [24]. In the new proposal, the Δ function is applied to the fuzzy output of the model and the list of fuzzy desired outputs. The composition of both lists produces a new list of fuzzy losses, and specific methods for comparing these lists are needed (those will be introduced in Section 3.3, as mentioned before).

Lastly, let us remark that the proposed method allows using a new configuration that, as far as we know, has not been explored in previous works: the function Δ, that grows with the difference between predictions and observed outputs, may be imprecise itself.

Let us explain this configuration with the help of an example. Consider a multiple-output model $f(x) = (f_1(x), f_2(x))$, where the loss function is a weighted combination of the squared losses of the outputs, with weights (w_1, w_2), i.e.

$$\Delta(y, f(x)) = w_1(y_1 - f_1(x))^2 + w_2(y_2 - f_2(x))^2. \tag{3}$$

The selection of the weights for these problems may be complex and there may happen that there is not information enough for choosing adequate weights. In this case, it is reasonable to think in using linguistic weights such as "LARGE"

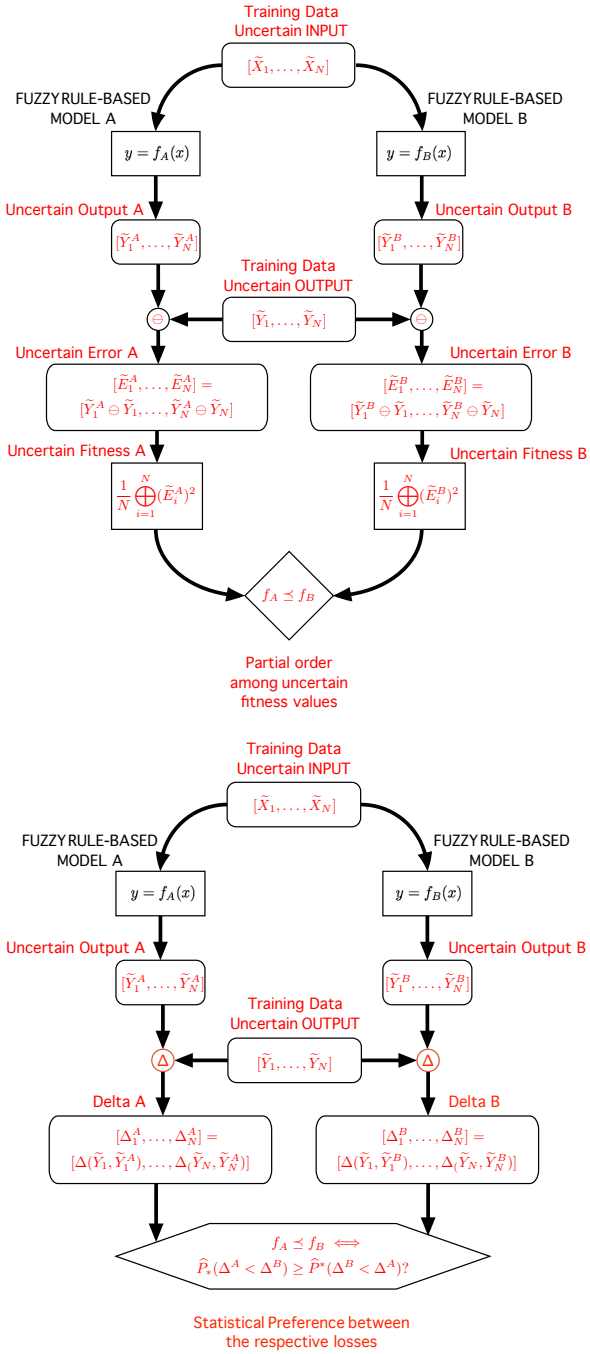

Figure 2: Graph of the computations involved in a comparison between two models in PR (upper part) and VL (lower part) for fuzzy data. Imprecise variables have been colored in red.

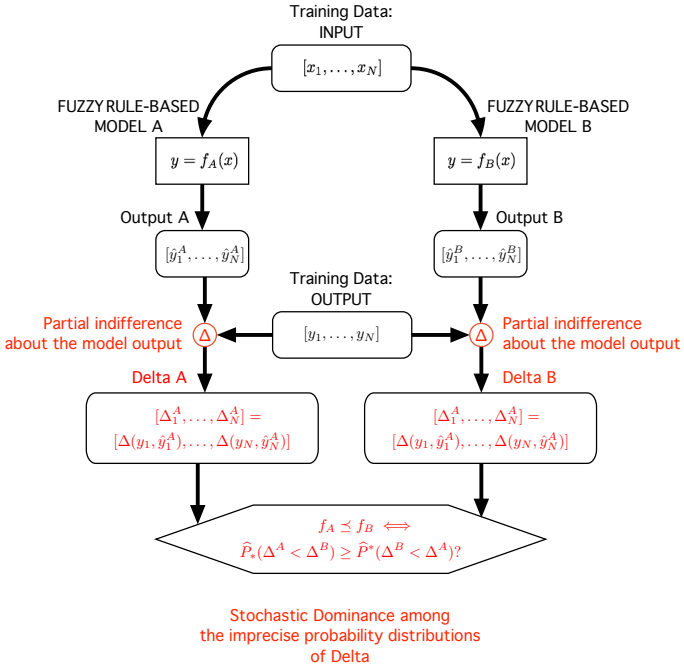

Figure 3: Graph of the computations involved in a comparison between two models for crisp data and a set-valued Δ function. Imprecise variables have been colored in red.

or "SMALL" associated to fuzzy membership functions named $\widetilde{W}_1$ and $\widetilde{W}_2$, in which case Δ becomes a fuzzy function $\widetilde{\Delta}$,

$$\widetilde{\Delta}_1(y, f(x))(z) =$$
$$\sup\{\min(\widetilde{W}_1(w_1), \widetilde{W}_2(w_2)) : z = w_1(y_1 - f_1(x))^2 + w_2(y_2 - f_2(x))^2\} \quad (4)$$

The graph of computations associated to the case where the data is crisp and the uncertainty is in the definition of the loss function is depicted in Figure 3. The uncertainty that is introduced at the definition of $\widetilde{\Delta}$ propagates to the vector of losses, and its treatment is the same as if the uncertainty were originated in the low quality of the data. In a loose sense, we may still regard this third case as "low quality data" because the weights of the cost function are input parameters to the learning problem. This case will be further discussed in Section 3.4 and a use case is provided in Section 4.3.

3.3. Sorting lists of vector losses

In order to compare fuzzy vector losses, a rank operator is needed in the first place to perform pairwise comparisons of the components of these vectors. There are many different operators in the literature: maximin [33][34], maximax [35], Teich's dominance [36], interval dominance [37], etc.; see [12] for a complete list and a study of the relations among them. A mechanism is needed to sort the lists of fuzzy losses of all the models being considered, and it is important to take into account that a total order is not possible. From a formal point of view, the vector of fuzzy losses associated to a given model can be identified with a fuzzy random variable representing its loss for each of the elements in the training dataset, and the purpose of the learning can also be described as determining the set of non dominated models for a suitable *stochastic ordering* defined among the losses. A complete discussion about these orders can be found in [14]. In this paper, the following three orderings are considered:

- Dominance in Expectation [38]
- Statistical Preference [39]
- Stochastical Dominance [40]

whose definitions follow. Given two vectors of losses Δ^A and Δ^B (recall that a vector of crisp losses can be regarded as a random variable, and a vector of fuzzy losses as a fuzzy random variable), the Dominance in Expectation is

$$f_A \preceq_{ED} f_B \iff E(\Delta^A) \leq E(\Delta^B), \tag{5}$$

where the model A is preferred to the model B if the expectation of the values $\Delta(y_i, \hat{y}_i^A)$ is lower than the expectation of the values $\Delta(y_i, \hat{y}_i^B)$. This ordering requires that the function Δ is numeric. Therefore, PR models are a particular case of VL models that combine a fuzzy ranking and a stochastic ordering of the kind "Dominance in Expectation".

The Statistical Preference is a second sorting criterion that has not a counterpart in previous works of modeling with low quality data. It is defined as

follows:

$$f_A \preceq_{SP} f_B \iff P(\Delta^A < \Delta^B) \geq P(\Delta^B < \Delta^A), \tag{6}$$

thus the model A is preferred to the model B if the number of training instances where $\Delta(y_i, \hat{y}_i^A) < \Delta(y_i, \hat{y}_i^B)$ is higher than the number of instances where $\Delta(y_i, \hat{y}_i^B) < \Delta(y_i, \hat{y}_i^A)$. This ordering does not require that Δ is numeric, provided that their values can be (partially) ordered.

Lastly, the First Stochastic Dominance, also without previous works in this field, is:

$$f_A \preceq_{SD} f_B \iff P(\Delta^A < c) \geq P(\Delta^B < c), \tag{7}$$

for a given constant $c \in \mathbb{R}$. The model A is preferred to the model B if the number of training instances where $\Delta(y_i, \hat{y}_i^A) < c$ is higher than the number of instances where $\Delta(y_i, \hat{y}_i) < c$. This ordering does not require that Δ is numeric, provided that their values can be compared to a number c.

3.4. Indifference, partial knowledge and non-numeric loss functions

Differences between observed and predicted values are assessed by means of the function Δ, as previously detailed. This function Δ may be, for instance, the square or the absolute value of the difference between predicted and observed values. In the particular case that $\Delta(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$, the best model for the Dominance in Expectation ordering is also optimal in the least squares sense. The First Stochastic Dominance is related to ϵ-insensitive learning, models for which the differences between predictions and observations are less than a given bound for a high percentage of the training instances [41][42][43].

Other cases exist where more complex definitions are appropriate. For instance, different weights can be applied to different training instances thus their relative importances are rebalanced. Also, small differences may be neglected, or very large differences may be clipped. One may even conceive non-numeric loss functions based on linguistic requisites such as "the weight of the second training instance must be higher than the height of the third training instance" or "if all differences are lower than 0.3, then a difference lower that 0.2 at the

second training instance is preferred to a difference lower than 0.1 at the third instance".

Lastly, it is remarked that each of these last sentences is compatible with multiple assignments of weights to the training instances, thus interval-valued or fuzzy weights can be used and that give rise to set or fuzzy-valued Δ functions, as already mentioned in Section 3.2 and Figure 3. Ambiguities in the definition of Δ are a secondary source of uncertainty that adds to the vagueness in the data. In the following example a degree of indifference between the risks of the different models is introduced, arising from a partial knowledge about the definition of the loss function:

Example 3.1. *A condition monitoring problem is considered where a model of the remaining useful life of an equipment is used to label it as "defective" or "normal". Defective equipment may be labelled as normal (false negatives, or type I error), or vice-versa (type II error). It is well known that passing a defective equipment as normal is much more expensive than coping with a false positive. Since models with a lower type I error are preferred, it was deemed reasonable to assign a different weight to each type of labelling:*

$$\Delta(Y(\omega), f_\theta(X(\omega))) = \begin{cases} w_{\mathrm{I}} & \text{if } Y(\omega) = \text{normal}, f_\theta(X(\omega)) = \text{defective} \\ w_{\mathrm{II}} & \text{if } Y(\omega) = \text{defective}, f_\theta(X(\omega)) = \text{normal} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

and compute the expected loss as follows:

$$\mathrm{risk}(\theta) = w_{\mathrm{I}} e_{\mathrm{I}}(\theta) + w_{\mathrm{II}} e_{\mathrm{II}}(\theta) \tag{9}$$

where $e_{\mathrm{I}}(\theta)$ and $e_{\mathrm{II}}(\theta)$ are the Type I and II errors of the model defined by the parameter θ. It is assumed that $w_{\mathrm{II}} = 1 - w_{\mathrm{I}}$. The knowledge about the values of

these weights is incomplete; w_{I} must be higher than w_{II}, but the precise values are not known.

Suppose that an interval ordering is defined that makes use of that information without assigning precise values to the weights. This is less arbitrary than any assignment of numerical values (unless the actual cost of a false negative can be established with precision). Notwithstanding, this also entails a secondary source of uncertainty. Consider the following three models:

	Type I error	*Type II error*
Model f_{θ_A}	*0.10*	*0.20*
Model f_{θ_B}	*0.20*	*0.10*
Model f_{θ_C}	*0.14*	*0.15*

If $w_{\mathrm{I}} > w_{\mathrm{II}}$, then Model f_{θ_A} is preferred to Model f_{θ_B} because $0.1w_{\mathrm{I}} + 0.2w_{\mathrm{II}} < 0.2w_{\mathrm{I}} + 0.1w_{\mathrm{II}}$, but Model f_{θ_A} is not comparable to Model f_{θ_C} because the risk of the model f_{θ_A} may be higher or lower than the risk of f_{θ_B}. This situation is intermediate between the multi-criteria case (where none of these models can be compared) and the scalar case (where all models are comparable, but the result of the comparison depends on the weights: Model f_{θ_A} is preferred to Model f_{θ_C} if $0 < w_{\mathrm{II}} < 4/9$, but Model f_{θ_C} is preferred to Model f_{θ_A} if $4/9 < w_{\mathrm{II}} < 1/2$).

Lastly, observe that the expected losses of the models become intervals even though the training data is crisp, because of the added imprecision in the preference operator. The risks of Models f_{θ_A}, f_{θ_B} and f_{θ_C} are

$$\mathrm{risk}(\theta_A) = 0.1w_{\mathrm{II}} + 0.1, \quad 0 < w_{\mathrm{II}} < 0.5 \tag{10}$$

$$\mathrm{risk}(\theta_B) = -0.1w_{\mathrm{II}} + 0.2, \quad 0 < w_{\mathrm{II}} < 0.5 \tag{11}$$

$$\mathrm{risk}_C(\theta_C) = 0.01w_{\mathrm{II}} + 0.14, \quad 0 < w_{\mathrm{II}} < 0.5 \tag{12}$$

thus the corresponding interval-valued risks are $\mathrm{risk}(\theta_A) = (0.1, 0.15)$, $\mathrm{risk}(\theta_B) =$

$(0.15, 0.2)$, $\text{risk}(\theta_C) = (0.145, 0.15)$.

3.5. Stochastic orderings-based optimization

Given that there is only a partial order between the loss vectors, a set of models must be sought whose loss vectors are minimal with respect to the stochastic ordering being considered. Generally speaking, metaheuristics that can search in preference order rankings can also solve problems that depend on the Dominance in Expectation, as illustrated in the following example.

Example 3.2. *Suppose that a FRBM is defined by means of a parameter* $\theta \in \Theta$*, and its expected loss is an interval*

$$\text{risk}(\theta) \in [\text{risk}_*(\theta), \text{risk}^*(\theta)]. \tag{13}$$

Suppose also that the unknown function $\text{risk}(\theta)$ *has a minimum at* θ_0*. The tightest bounds of* θ_0 *that can be defined on the basis of the interval-valued risk, are (see Figure 4):*

$$\theta_0 \in \{\theta : \text{risk}_*(\theta) \leq \min_{\Theta}(\text{risk}^*(\theta))\} \tag{14}$$

and the knowledge about the value of the expected loss at θ_0 *is*

$$\text{risk}(\theta_0) \in [\min_{\Theta}(\text{risk}_*(\theta)), \min_{\Theta}(\text{risk}^*(\theta))]. \tag{15}$$

Consider the partial order that is defined in $\Theta \times \Theta$ *by the Fishburn interval dominance [37]:*

$$\begin{cases} \theta_1 \succeq \theta_2 & \text{if } \text{risk}^*(\theta_1) \leq \text{risk}_*(\theta_2) \\ \theta_2 \succeq \theta_1 & \text{if } \text{risk}^*(\theta_2) \leq \text{risk}_*(\theta_1) \\ \theta_1 \parallel \theta_2 & \text{otherwise} \end{cases} \tag{16}$$

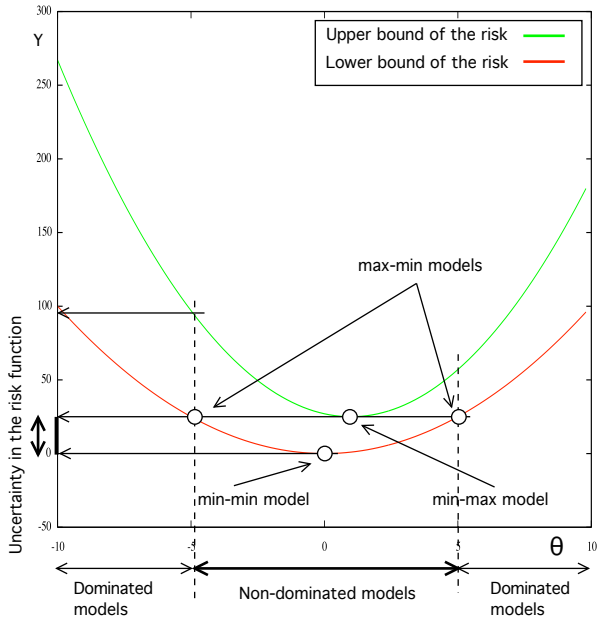

Figure 4: Optimization of an interval-valued loss function showing the uncertainty in the location of the minimum loss.

It is clear that

$$\{\theta : \nexists \theta' \in \Theta, \theta' \prec \theta\} = \{\theta : \text{risk}_*(\theta) \leq \min_{\Theta}(\text{risk}^*(\theta))\}. \tag{17}$$

In words, the most specific subset of the parameter space that contains the minimum θ_0 is also the set of the minimal elements for the preorder induced by the dominance relation between interval losses seen before.

Since the set of non-dominated solutions in Eq. 17 is akin to the "Pareto front" of multi-criteria problems, multi-objective optimization algorithms can be adapted to the search of the most specific subset of Θ that is known to contain θ_0 with the (imprecise) available information. The fuzzy case has a straightforward definition as a family of nested interval problems, whose respective solutions are the α-cuts of the fuzzy set defining the location of the optimal parameter.

Dominance operators between fuzzy-valued risk functions have been proposed elsewhere (see for instance [28] for a discussion about the subject) that allow solving all the nested problems at the same time.

However, the problem becomes much harder for orderings different than Dominance in Expectation. Transitivity cannot be taken for granted for all stochastic orderings (for instance, the Statistical Preference is not transitive) and hence elitist algorithms are not feasible. Operations such as finding a subset of non-dominated risks require specific methods [44], that are seldom studied in the context of numerical optimization [45]. The definition of a numerical algorithm that can efficiently search for the set of nondominated models for a given stochastic ordering is still an open problem.

4. Case studies

Three different situations are reviewed in this section:

1. A problem with partially missing data (nested tolerance intervals) where the difference between the true values and the modal points of the imputed fuzzy observations is heteroscedastic random noise (a mixture of distributions with unequal variances). This case illustrates certain limitations of current approaches (i.e. PR or VL+"dominance in expectation") that are deceived by outliers and may converge to the wrong model. On the contrary, VL+"statistical preference" and VL+"stochastic dominance" converge to the true model.

2. The second problem regards interval data that covers the true value of the independent variable with a total confidence. This is the paradigmatic case for PR algorithms, that achieve good results, on par with VL models. In this case, it is the properties of interval/fuzzy ranks that are discussed, showing that, for this particular problem, the minimin criteria proposed in [20] is the best choice for all stochastic orders.

3. Third and last, a crisp problem is presented where an uncertain loss function is used, with results that improve crisp loss functions. This is a new

family of problems that should be researched in future studies, and as such this case of study is presented without compared results.

The complexity of the case studies presented in this section prevents using hand calculations as done in Section 2. Therefore, a simple memetic algorithm that follows all the principles proposed in this paper has been developed specifically for solving the cases in this section. This algorithm, that could be considered as a baseline for future developments in this field, is described in Appendix A.

4.1. Fuzzy data with an epistemic interpretation

This problem comprises 50 pairs

$$(x_i, y_i) = (x_i, \sin \frac{x_i}{20}) \tag{18}$$

that are perceived through fuzzy sets with an epistemic interpretation, i.e. α-cuts of the fuzzy sets are confidence intervals of the true value of the variable at levels $1 - \alpha$. Triangular memberships (l, c, r) are used, where the support is the interval $[l, r]$ and the modal point of the set is c. The training dataset comprises 50 pairs

$$(x_i, \widetilde{Y}_i) = (x_i, (l_i, c_i, r_i)). \tag{19}$$

The modal points of the training data are

$$c_i = \begin{cases} y_i + \mathcal{U}(-0.1, 0.1) & \text{if } \mathcal{U}(0,1) < 0.95 \\ y_i + \mathcal{U}(-1, 1) & \text{else} \end{cases} \tag{20}$$

where $\mathcal{U}(a, b)$ stands for a random value with uniform distribution in $[a, b]$. Observe that 5% of the training data are outliers. The supports are

$$[l_i, r_i] = c_i \pm \mathcal{U}(0, 0.25). \tag{21}$$

FRBMs of the TSK type, comprising 10 rules each, were learned with the help of the algorithm described in the preceding section. A population of size 100 was used. The learning ends after 100 generations are completed. The mutation probability is of 0.05. A 5% of local optimization was added. Arithmetic

crossover and mutations were applied. The ranking among imprecise values was Teich's uniform dominance and the stochastic orderings were Dominance in Expectation, Statistical Preference and First Stochastical dominance with $c = 0.1$. Testing data is crisp and comprises the noiseless pairs (x_i, y_i). In addition to this, three more models were fitted to crisp data (the modal points of the training sample) with the same techniques.

The purpose of this experiment is to compare standard statistical models with PR and VL algorithms for learning FRBMs from low quality data. It is expected that optimizing a loss function underperforms statistical preference and stochastic dominance models because of the presence of outliers. It is also expected that fuzzy data is more informative than crisp data at this particular problem: lower test errors are expected if the fuzziness in the data is not discarded.

The results are summarized in Table 2 and displayed in Figure 5. As expected, Least Squares models overtrained because of the outliers. The fuzzy version improved the crisp model, and the VL models are noticeably more robust. The use of the Teich's uniform dominance in combination with the First Stochastic Dominance is intuitively the best setup, given that this ordering is resilient to those instances for which the residuals are higher than the specified bound.

4.2. Interval-valued data with minimin and minimax ranks

The second problem comprises 50 pairs of the same function seen in the preceding section that are perceived through intervals that contain the true value with total confidence. The training set is

$$(x_i, \overline{Y}_i) = (x_i, [l_i, r_i]). \tag{22}$$

where

$$[l_i, r_i] = \begin{cases} [y_i, y_i + 0.1] & \text{if } \mathcal{U}(0,1) < 0.5 \\ [y_i - 0.1, y_i] & \text{else} \end{cases} \tag{23}$$

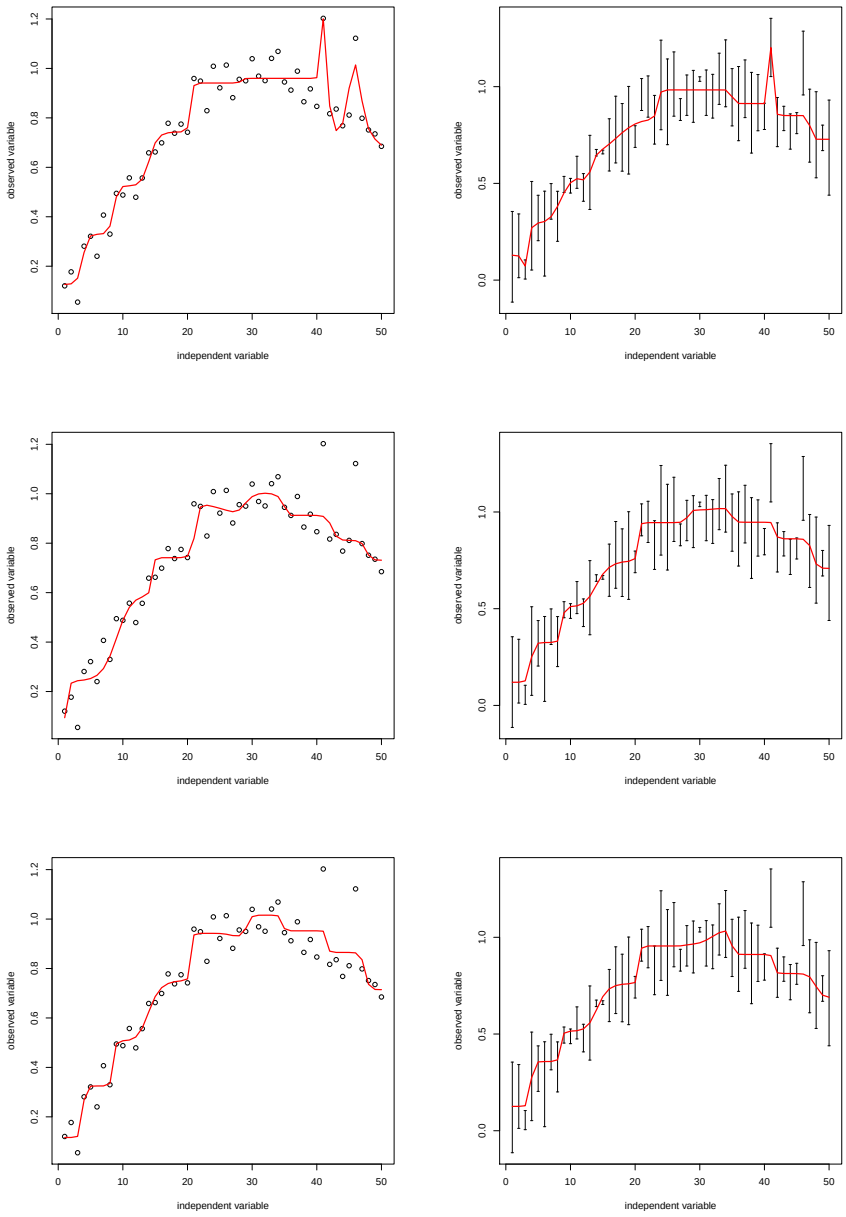

Figure 5: From upper to bottom, left to right: Crisp Least Squares vs Domination in Expectation, crisp Statistical Preference vs Fuzzy Statistical Preference, Crisp Stochastical Dominance vs Fuzzy Stochastical Dominance. Least Squares model overtrained and tried to learn the outliers, while stochastic ordering-based models were more robust.

Table 2: Crisp Least Squares compared to PR and VL algorithms for learning FRBMs from fuzzy data

Crisp	MSE Train	MSE Test
Nonlinear Least Squares	0.00311	0.00612
LQD PR	MSE Train	MSE Test
Fuzzy Least Squares	(0.00174, 0.00392, 0.03601)	0.00409
LQD VL	MSE Train	MSE Test
Crisp Statistical Preference	0.00710	0.00257
Crisp Stochastical Dominance	0.00484	0.00258
Fuzzy Statistical Preference	(0.00212, 0.00285, 0.03710)	0.00228
Fuzzy Stochastical Dominance	(0.00212, 0.00343, 0.03613)	0.00216

The purpose of this second problem is to compare the different order rankings among imprecise data. Minimin and minimax ranks were selected, and the results are summarized in Figure 6 and Table 3. Observe that the upper bounds at 50% of the intervals are the true values of the training data, and the remaining 50% are the lower bounds of the corresponding intervals. Minimax-based ranks favor the models that pass near the majority of the center-points of the intervals, while minimin models are the smoothest curves that are contained in the same intervals. The last alternative is the best solution for this particular case.

This case study shows that the selection of a particular combination of stochastic ordering and rank depends on the properties of the problem. Minimax or Uniform ranks, combined with statistical preference or stochastic dominance, are intuitively preferred for noisy data, while least squares and minimin ranks are suitable for noiseless interval or fuzzy data.

4.3. Crisp data with uncertainty in the loss function

A novel type of problems is introduced in this last case study. The data is crisp but there is an incomplete knowledge about the loss function. It will be shown that the use of an imprecise loss function can be actually beneficial: the outcome of the learning is a model that cannot be found by methods that depend on crisp loss functions.

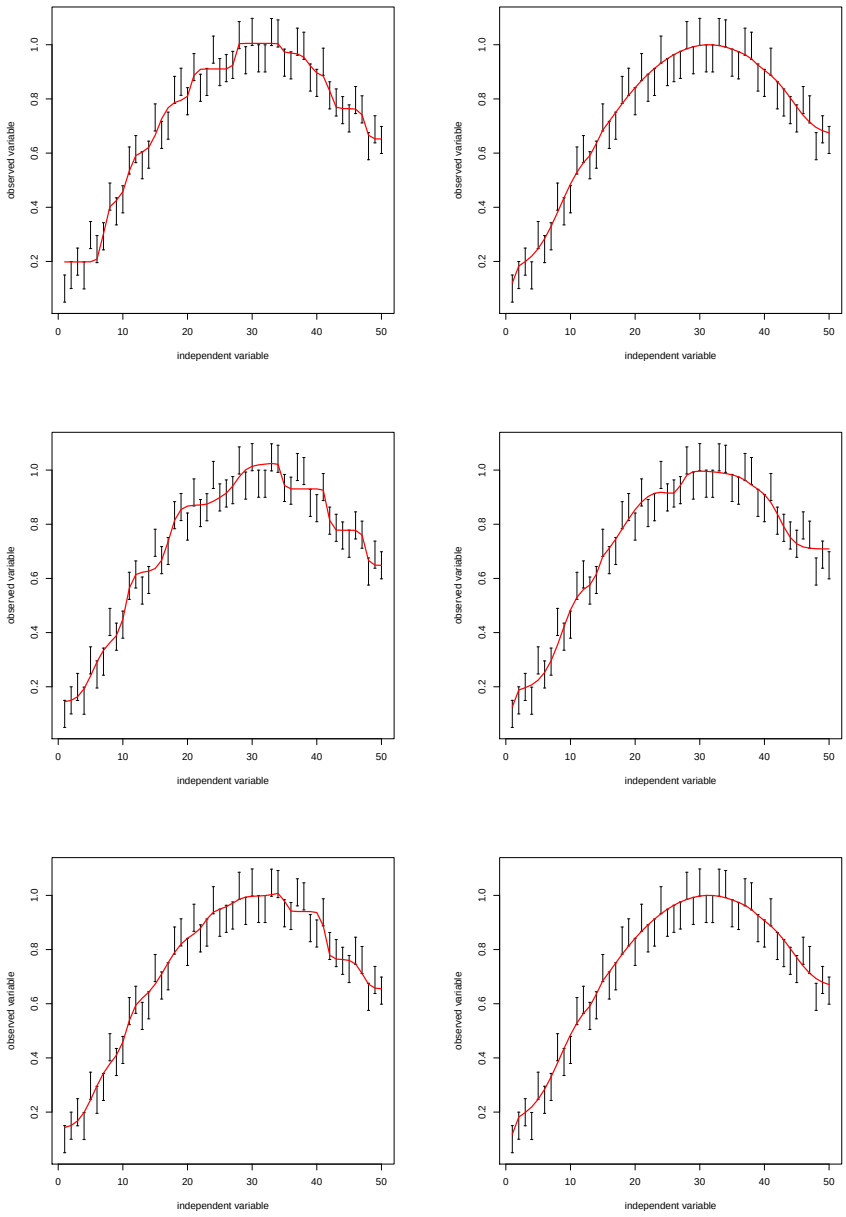

Figure 6: From upper to bottom, left to right: Minimax vs Minimin for Domination in Expectation, Statistical Preference and Stochastical Dominance. The minimin rank is better suited for the case where the true values of the observed variable are known to be covered by the interval output.

Table 3: Compared ranks for imprecise data: Minimax vs. Minimin

Minimax	MSE Train	MSE Test
Fuzzy Least Squares	(0.00060, 0.00189, 0.00783)	0.00117
Fuzzy Statistical Preference	(0.00066, 0.00196, 0.00792)	0.00122
Fuzzy Stochastical Dominance	(0.00103, 0.00211, 0.00792)	0.00071
Minimin	**MSE Train**	**MSE Test**
Fuzzy Least Squares	(0.00100, 0.00229, 0.00942)	0.00047
Fuzzy Statistical Preference	(0.00057, 0.00225, 0.00897)	0.00114
Fuzzy Stochastical Dominance	(0.00100, 0.00228, 0.00943)	0.00045

The description of this case is as follows: data of the torque and energetic efficiency of a motor vs. angular speed is used. The purpose of the model is to predict both the torque and the efficiency of the motor with a FRBM comprising three rules of the following form:

```
If angular speed is LOW then Torque = LT and Efficiency = LE
If angular speed is MEDIUM then Torque = MT and Efficiency = ME
If angular speed is HIGH then Torque = HT and Efficiency = HE
```

The linguistic terms have Gaussian memberships. The size of the rule base prevents that a perfect matching is achieved for both variables at the same time. An accurate prediction of the efficiency requires that the linguistic term "MEDIUM" is positioned at a place where the approximation of the torque is suboptimal, and vice-versa.

Under the nonlinear least squares setup, the loss of the FRBM depends on the accuracy of the prediction of two different output variables, weighted as follows:

$$\begin{aligned} \text{loss} = & \frac{1}{N}\sum_{i=1}^{N} w_1 \cdot (\text{ModelTorque}_i - \text{ObservedTorque}_i)^2 + \\ & w_2 \cdot (\text{ModelEffic}_i - \text{ObservedEffic}_i)^2 \end{aligned} \tag{24}$$

However, the selection of the weights w_1 and w_2 is not immediate. In the upper

part of Figure 7 different weights have been assigned to each variable. In the left part, $w_1 = 0.1$ and $w_2 = 0.9$. In the middle part, $w_1 = w_2 = 0.5$. In the right part, $w_1 = 0.9$ and $w_2 = 0.1$. Observe that only one of the curves is accurately modeled at each case; the approximation error is always suboptimal for the other curve. The case where both weights are equal is biased towards a better accuracy of the efficiency. This is because the values of this variable are also higher and contribute more to the average loss.

Now suppose that it is requested that $w_2 > w_1$, i.e. it is stated that a good fitting of the efficiency is preferred to the torque, but precise values of w_1 and w_2 are not given. Without loss of generality, it is assumed that $w_1 + w_2 = 1$ and the following interval-valued loss defined:

$$\begin{aligned}\overline{\text{loss}} = \{ & \frac{1}{N} \sum_{i=1}^{N} w \cdot (\text{ModelTorque}_i - \text{ObservedTorque}_i)^2 + \\ & (1-w) \cdot (\text{ModelEffic}_i - \text{ObservedEffic}_i)^2 : w \in [0, 0.5]\}\end{aligned} \tag{25}$$

In the lower, left part of the same figure this loss function has been optimized (dominance in expectation). The lower, right part is the opposite case

$$\begin{aligned}\overline{\text{loss}'} = \{ & \frac{1}{N} \sum_{i=1}^{N} w \cdot (\text{ModelTorque}_i - \text{ObservedTorque}_i)^2 + \\ & (1-w) \cdot (\text{ModelEffic}_i - \text{ObservedEffic}_i)^2 : w \in [0.5, 1]\}\end{aligned} \tag{26}$$

where the accuracy of the torque is preferred to that of the efficiency. Observe that in both cases the balance between the two outputs of the model was better than any of the weighted least squares models, and a rule base is obtained that was not found by none of the crisp weights that were essayed.

5. Concluding remarks and future work

Learning fuzzy rules from low quality data poses two fundamental challenges: (a) exploiting the available information and (b) avoiding systematic bias. On the one hand, in order to make the most of the available information, incomplete and/or imprecise training data cannot be removed, but this "less than perfect"

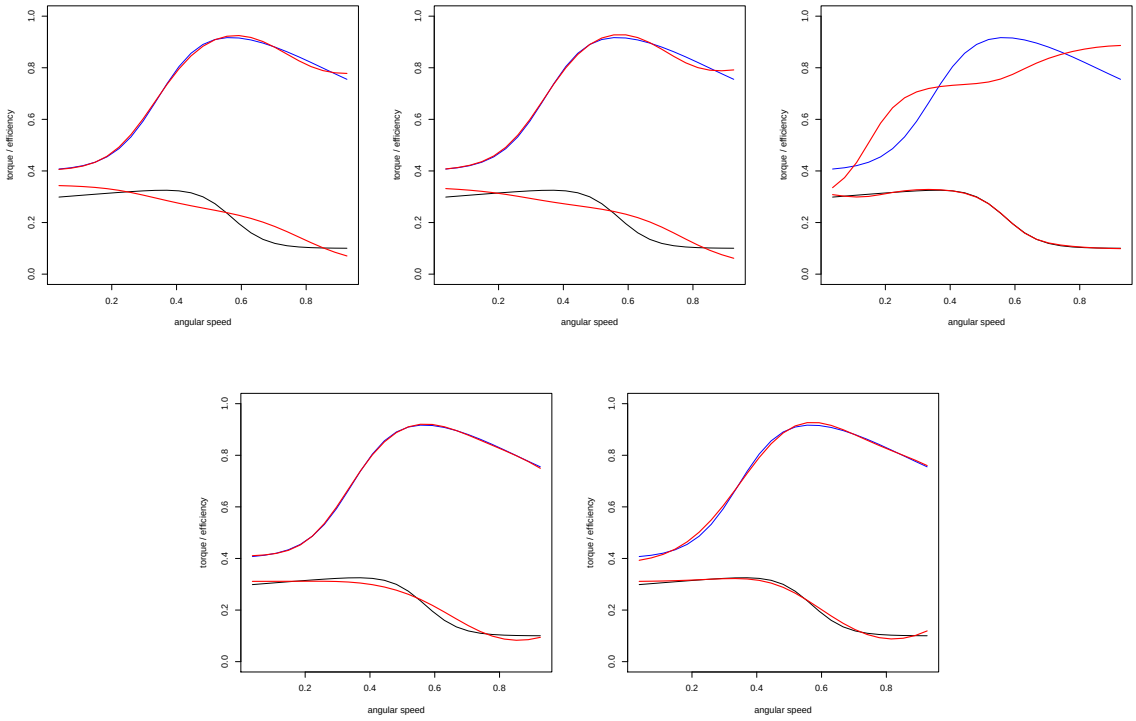

Figure 7: Crisp data and incomplete knowledge about the loss function. A model comprising rules of the form "If input is VALUE then output1 = VALUE and output2 = VALUE" is optimized on the basis of a risk function which is, in turn, a weighted average of the MSEs the two outputs. It is hard to find a set of crisp weights that balance the accuracies of the two outputs of the model, but imprecise weights exist that level these errors. Upper part: crisp loss function with $w_1 = 0.1$, $w_2 = 0.9$ (left), $w_1 = w_2 = 0.5$ (center) and $w_1 = 0.9$, $w_2 = 0.1$ (right). Lower part: interval-valued loss function with $w_2 > w_1$ (left) and $w_1 > w_2$ (right).

data is kept in the training dataset. On the other hand, avoiding systematic bias is also difficult because, as shown in this paper, it is possible that the ranks among the perceived risks of a set of models are different than the ranks among the actual risks, deceiving the learning algorithm. We have also shown that the perceived risks of different models may be identical.

Because of these reasons, in this paper it is proposed that empirical risks are not the best option when learning fuzzy rule-based models from low quality data, but a multivariate fitness function comprising a vector of fuzzy losses (VL) is preferred.

Three representative case studies have been provided where the new framework is the best alternative. In the first case, the training data was coarsely discretized and stochastic noise was added. This case was solved with crisp, possibilistic risk (PR, also known as "fuzzy fitness") and lastly VL. We concluded that VL was the most appropriate strategy. In the second case, stochastic noise was not added. We concluded that problems with pure epistemic noise are well suited for PR (which is a particular case of VL), but certain stochastic orders have interesting properties. In the third and last case, which was a multi-criteria problem, it was not the data that was imprecise, but the loss function. The loss function was given a (fuzzy rule-based) linguistic definition. It was concluded that the VL algorithm finds models that could not be found if the uncertainty in the loss function was removed.

The technologies introduced in this study are at the stage of initial development. New lines of research should be launched to study the theoretical properties of the models that arise from the different combinations of stochastic orders and interval or fuzzy orderings. Also, new metaheuristics for finding the set of minimal elements of a partial order on the basis of a preference operator can be designed, paying special interest on their scalability, including the parallelization of the code for its use in computing clusters or other specialized equipment.

Acknowledgements

This work has been partially supported by "Ministerio de Economía y Competitividad" from Spain under grants TIN2014-56967-R, TIN2017-84804-R and by the Regional Ministry of the Principality of Asturias under grant FC-15-GRUPIN14-073.

References

[1] J. W. Tukey, Unsolved Problems of Experimental Statistics, Journal of the American Statistical Association 49 (268) (1954) 706–731.

[2] J. W. Tukey, We Need Both Exploratory and Confirmatory, The American Statistician 34 (1) (1980) 23–25.

[3] W. G. Cochran, Errors of Measurement in Statistics, Technometrics 10 (4) (1968) 637.

[4] P. Diaconis, Theories of Data Analysis: From Magical Thinking Through Classical Statistics, in: Exploring Data Tables, Trends, and Shapes, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2006, pp. 1–36.

[5] F. Graba, F. Comby, O. Strauss, Non-additive imprecise image super-resolution in a semi-blind context, IEEE Transactions on Image Processing 26 (3) (2017) 1379–1392.

[6] I. Couso, S. Montes, P. Gil, The necessity of the strong α-cuts of a fuzzy set, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 9 (02) (2001) 249–262.

[7] D. B. Rubin, Inference and missing data, Biometrika 63 (3) (1976) 581–592.

[8] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society. Series B (Methodological) 39 (1) (1977) 1–38.

[9] R. J. Mislevy, R. J. A. Little, D. B. Rubin, Statistical Analysis with Missing Data, Journal of Educational Statistics 16 (2) (1991) 150.

[10] D. F. Heitjan, D. B. Rubin, Ignorability and Coarse Data, The Annals of Statistics 19 (4) (1991) 2244–2253.

[11] L. Sánchez, I. Couso, Guest editorial: special issue on "Knowledge extraction from low quality data: theoretical, methodological and practical issues", Soft Computing - A Fusion of Foundations, Methodologies and Applications 16 (5) (2012) 739–740.

[12] I. Couso, L. Sánchez, Guest editorial: Special issue "Harnessing the information contained in low-quality data sources", International Journal of Approximate Reasoning 55 (7) (2014) 1485–1486.

[13] I. Couso, D. Dubois, Statistical reasoning with set-valued information: Ontic vs. epistemic views, International Journal of Approximate Reasoning 55 (7) (2014) 1502–1518.

[14] I. Couso, L. Sánchez, Machine learning models, epistemic set-valued data and generalized loss functions: An encompassing approach, Information Sciences 358-359 (2016) 129–150.

[15] M. Jaeger, The AI&M Procedure for Learning from Incomplete Data, in: 22nd Conference on Uncertainty in Artificial Intelligence, 2006, pp. 225–232.

[16] M. Ramoni, P. Sebastiani, Robust Learning with Missing Data, Machine Learning 45 (2) (2001) 147–170.

[17] H. Ishibuchi, H. Tanaka, Fuzzy regression analysis using neural networks, Fuzzy Sets and Systems 50 (3) (1992) 257–265.

[18] H. Ishibuchi, M. Nii, Fuzzy regression using asymmetric fuzzy coefficients and fuzzified neural networks, Fuzzy Sets and Systems 119 (2) (2001) 273–290.

[19] P. Diamond, Fuzzy least squares, Information Sciences 46 (3) (1988) 141–157.

[20] E. Hüllermeier, Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization., Int. J. Approx. Reasoning 55 (7) (2014) 1519–1534.

[21] P. Zywica, A. Wójtowicz, A. Stachowiak, K. Dyczkowski, Improving medical decisions under incomplete data using interval-valued fuzzy aggregation., in: 16th World Congress of the International Fuzzy Systems Association (IFSA-EUSFLAT), Atlantis Press, 2015, pp. 577–584.

[22] A. Wójtowicz, P. Żywica, A. Stachowiak, K. Dyczkowski, Solving the problem of incomplete data in medical diagnosis via interval modeling, Applied Soft Computing 47 (2016) 424–437.

[23] L. Sánchez, I. Couso, Advocating the use of imprecisely observed data in genetic fuzzy systems, IEEE Transactions on Fuzzy Systems 15 (4) (2007) 551–562.

[24] L. Sánchez, J. Otero, I. Couso, Obtaining linguistic fuzzy rule-based regression models from imprecise data with multiobjective genetic algorithms, Soft Computing 13 (5) (2009) 467–479.

[25] A. Palacios, L. Sánchez, I. Couso, Extending a simple genetic cooperative-competitive learning fuzzy classifier to low quality datasets, Evolutionary Intelligence 2 (1-2) (2009) 73–84.

[26] J. Plass, M. E. Cattaneo, G. Schollmeyer, T. Augustin, Testing of coarsening mechanisms: Coarsening at random versus subgroup independence, in: Soft Methods for Data Science, Springer, 2017, pp. 415–422.

[27] L. Sánchez, Comments on "Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization" by Eyke Hüllermeier., Int. J. Approx. Reasoning 55 (7) (2014) 1583–1587.

[28] L. Sánchez, I. Couso, J. Casillas, Genetic learning of fuzzy rules based on low quality data, Fuzzy Sets and Systems 160 (17) (2009) 2524–2552.

[29] H. Ishibuchi, T. Nakashima, M. Nii, Classification and Modeling with Linguistic Information Granules, Springer (2005).

[30] E. H. Mamdani, S. Assilian, An experiment in linguistic synthesis with a fuzzy logic controller, International Journal of Man-Machine Studies 7 (1) (1975) 1–13.

[31] T. Takagi, M. Sugeno, Fuzzy identification of systems and its applications to modeling and control, Systems, Man and Cybernetics, IEEE Transactions on SMC-15 (1) (1985) 116–132.

[32] M. Sugeno, G. T. Kang, Structure identification of fuzzy model, Fuzzy Sets and Systems 28 (1) (1988) 15–33.

[33] M. Sniedovich, Wald's maximin model: a treasure in disguise!, The Journal of Risk Finance 9 (3) (2008) 287–291.

[34] A. Wald, Statistical Decision Functions Which Minimize the Maximum Risk, The Annals of Mathematics 46 (2) (1945) 265.

[35] J. K. Satia, R. E. Lave Jr., Markovian Decision Processes with Uncertain Transition Probabilities, Operations Research 21 (3) (1973) 728–740.

[36] J. Teich, Pareto-Front Exploration with Uncertain Objectives, in: Advances in Artificial Intelligence, Springer Berlin Heidelberg, Berlin, Heidelberg, 2001, pp. 314–328.

[37] R. R. Laxton, P. C. Fishburn, Interval Orders and Interval Graphs: A Study of Partially Ordered Sets., Journal of the Royal Statistical Society. Series A (General) 148 (4) (1985) 393.

[38] L. J. Savage, The Foundations of Statistics, Courier Corporation, 2012.

[39] H. A. David, The method of paired comparisons, A Charles Griffin Book, 1988.

[40] J. Hadar, W. Russell, Rules for Ordering Uncertain Prospects, American Economic Review (1969) 25–34.

[41] L. Sánchez, Interval-valued GA-P algorithms, Evolutionary Computation, IEEE Transactions on 4 (1) (2000) 64–72.

[42] L. Sánchez, I. Couso, Fuzzy random variables-based modeling with GA-P algorithms, in: Information, uncertainty and fusion, Springer, 2000, pp. 245–256.

[43] M. Serrurier, H. Prade, A general framework for imprecise regression, in: 2007 IEEE International Fuzzy Systems Conference, IEEE, 2007, pp. 1–6.

[44] S. Chen, T. Joachims, Modeling intransitivity in matchup and comparison data, in: Proceedings of the ninth acm international conference on web search and data mining, ACM, 2016, pp. 227–236.

[45] L. B. Said, S. Bechikh, K. Ghédira, The r-dominance: a new dominance relation for interactive evolutionary multicriteria decision making, IEEE Transactions on Evolutionary Computation 14 (5) (2010) 801–818.

[46] J. Nelder, R. Mead, A simplex method for function minimization, The computer journal 7 (4) (1965) 308–313.

[47] F. di Pierro, S.-T. Khu, D. A. Savi, An Investigation on Preference Order Ranking Scheme for Multiobjective Evolutionary Optimization, IEEE Transactions on Evolutionary Computation 11 (1) (2016) 17–45.

Appendix A. Description of the memetic algorithm designed for the use cases

The memetic algorithm applied in the use cases is described here. Its pseudocode is outlined in Figures A.8 and A.9. The main idea of this algorithm

is to use the domination count as a primary rank, and then use a secondary measure to guide the search and preserve diversity. The domination count is the number of individuals in the current population that are preferred to the individual being assessed, according to the chosen combination of fuzzy ranking and stochastic ordering. The population is ranked first according to the domination count of each individual, and individuals with the same domination count are sorted by their average distance to the other individuals in the population, promoting the presence of models in the less dense areas of the genotype space. It is suggested that a low selective pressure is applied thus the domination count is not subjected to large variations at each generation. (i.e. at each generation the best M individuals are copied to the intermediate population; it is suggested that M is a significant fraction of the total size of the population). Genetic operators are applied to individuals selected through a double binary tournament. Finally, in this baseline algorithm the local optimization is applied to a constant fraction of the best individuals of each generation, but more complex strategies should be developed in future works. It is remarked that specific crossover and mutation operators could be adapted to the parametric expression of the fuzzy model but the definition of these operators is out of the scope of this study.

The local optimizer is derived from the Nelder and Mead's algorithm [46]. The numerical-valued objective function used by this greedy algorithm is replaced by the rank in the simplex. The rank of an individual in the simplex is obtained by sorting the simplex by the domination count (that is computed with respect to the whole population of the genetic algoritm). Ties in the domination count are broken with the mean squared sum of the modal points of the values of the Δ function. Observe that this is not the same rank used to sort the genetic population, where the second criterion was intended to promote the diversity. In the local optimizer, this criterion is a secondary preference in the same sense as [47].

```
for gen in (1,NITER):
    compute the domination count of the population θ_1, θ_2, ..., θ_N
    sort θ_1, θ_2, ..., θ_N by domination count (first) and crowding (second)
    insert M individuals from the population into the intermediate population
    while size of intermediate population < N:
       select two individuals of the population
       perform crossover
       if rnd() < p perform mutation
       insert the offspring into the intermediate population
    compute the domination count of the intermediate population θ'_1, θ'_2, ..., θ'_N
    sort θ'_1, θ'_2, ..., θ'_N by domination count (first) and crowding (second)
    for i in (1,opt):
       apply local optimization to θ'_i
    population = intermediate population
return θ_i whose domination count = 0
```

Figure A.8: Pseudocode of a memetic algorithm used for searching the minimal elements of a stochastic ordering

```
Compute N + 1 models θ_1, θ_2, ..., θ_{n+1} defining the simplex
repeat
    compute the domination count of θ_1, θ_2, ..., θ_{n+1}
    sort θ_1, θ_2, ..., θ_{n+1} by domination count (first) and MSE (second)
    θ_cen = (1/n) Σ_{i=1}^{n} θ_i
    θ_ref = θ_cen + α(θ_cen − θ_{n+1})
    C_1 = Δ_ref ≺ Δ_n
    C_2 = Δ_ref ≺ Δ_1
    if C_1 and not C_2:
       θ_{n+1} = θ_ref
    else:
       if C_1:
          θ_exp = θ_ref + γ(θ_ref − θ_cen)
          if Δ_exp ≺ Δ_ref:
             θ_{n+1} = θ_exp
          else:
             θ_{n+1} = θ_ref
       else:
          θ_con = θ_cen + ρ(θ_{n+1} − θ_cen)
          if Δ_con ≺ Δ_{n+1}:
             θ_{n+1} = θ_con
          else:
             for i in (2,n+1):
                θ_i = θ_1 + σ(θ_i − θ_1)
until iteration limit or the simplex is small enough
return θ_1
```

Figure A.9: Pseudocode of the local optimizer (derived from Nelder and Mead's simplex) used for searching the minimal elements of a stochastic ordering