Interval-valued regression and classification models in the framework of machine learning

Lev V. Utkin
Department of Computer Science
St.Petersburg State Forest Technical Academy
St.Petersburg, Russia
lev.utkin@mail.ru

Frank P.A. Coolen
Department of Mathematical Sciences
Durham University, Durham, UK
frank.coolen@durham.ac.uk

Abstract

We present a new approach for constructing regression and classification models for interval-valued data. The risk functional is considered under a set of probability distributions, resulting from the application of a chosen inferential method to the data, such that the bounding distributions of the set depend on the regression and classification parameter. Two extreme ('pessimistic' and 'optimistic') strategies of decision making are presented. The method is applicable with many inferential methods and risk functionals. The general theory is presented together with the specific optimisation problems for several scenarios, including the extension of the support vector machine method for interval-valued data.

Keywords. belief functions, classification, interval-valued observations, machine learning, p-box, regression, risk functional, support vector machines.

1 Introduction

A main goal of statistical machine learning is prediction of an unobserved output value y based on an observed input vector $\mathbf{x}$, which requires estimation of a predictor function f from training data consisting of pairs $(\mathbf{x}, y)$. Two major topics in statistics which fit into the statistical machine learning framework are regression analysis and classification. In regression analysis, one typically aims at estimation of a real-valued function based on a finite set of observations with random noise. In classification, the output variable is in one of a finite number of classes[1] and the main task is to classify the output y corresponding to each input $\mathbf{x}$ into one of the classes by means of a discriminant function. Many methods have been proposed for solving machine learning problems, but these are mostly based on rather restrictive assumptions, for example assuming the availability of a large amount of training data, known probability distribution for the random noise, or that all observations are point-valued ('precise'). Such assumptions are typically not fully satisfied in applications. For example, data often include interval-valued ('imprecise') observations, which may result from imperfection of measurement tools or imprecision of expert information if used as data. There may also be (partially) missing data, for example in classification problems the input vector ('pattern') $\mathbf{x}$ is often not fully observed. Many methods for dealing with such features use additional assumptions. In this paper, a general framework is presented that allows such important aspects to be incorporated in machine learning problems without additional assumptions, instead it uses the framework of imprecise probability [34] and it can be used for a wide variety of inferences, models and real-world situations.

Many methods have been presented for regression and classification with interval-valued data [11, 16, 23]. In some methods for machine learning, interval-valued observations are replaced by precise values based on some (often ad-hoc) additional assumptions, for example by taking middle points of the intervals [14]. Also, they may not be suitable if an observation is not restricted to an interval of finite length. This is an important restriction, as frequently it may only be known that an observation is larger (or smaller) than a specific value while the support of the corresponding random quantity is not finite. The method presented in this paper can deal with such information without additional assumptions and allows infinite support[2], including the use of $(-\infty, \infty)$ for missing elements of the input vector $\mathbf{x}$. Machine learning methods have been presented which use standard interval analysis and provide predictor functions with interval-valued parameter [2, 9, 26], and construction of second-order machine learning models for interval-valued patterns

[1] Often two classes, to which attention is restricted in this paper; generalization is possible but not addressed here.

[2] It should be noted that the support of elements of vector $\mathbf{x}$ can be arbitrary. Without loss of generality, we assume it to be $(-\infty, \infty)$.

was proposed in [4]. Although many methods have been presented for dealing with interval-valued data [23], these are mostly based on interval extension of the empirical risk functional [33] without benefiting from, or even considering, an imprecise probabilistic framework in direct relation to imprecise statistical data.

Pelckmans et al. [17] presented a detailed analysis of different methods and models for dealing with missing data in classification. Many methods do so by imputation of (partially) missing patterns, where missing (precise) values are replaced by some preferable values. Imputation using intervals, including the full support in case of missing elements of $\mathbf{x}$ has also been presented. De Cooman and Zaffalon [5] studied the classification problem with missing data in the framework of imprecise probability theory. An interesting approach for regression analysis with interval-valued and fuzzy data using belief functions and evidence theory has been proposed by Petit-Renaud and Denoeux [18]. One of the possible approaches to regression analysis is to consider a set of probability distributions for the random noise instead of a single distribution. This approach can be realized in the framework of imprecise probability theory [34] and has been developed by Walter et al. [36].

The novel approach for constructing a class of machine learning models and methods proposed in this paper uses risk functionals as in [18] and sets of probability distributions as in [36]. The starting point is a set of probability distributions related to the training data, which can just be a small amount of data or imprecise data, and this set can be generated by a variety of inferential methods and is assumed to be bounded by some lower and upper CDFs. Such sets of probability distributions are also called p-boxes [7]. In the regression and classification applications considered in this paper, these bounds for the set of probability distributions depend on the unknown parameter of the regression or discriminant function, because the sets of probability distributions considered are for the random residuals and as such they depend on the model parameter. It should be noted that the considered set of distributions is not the set of parametric distributions having the same parametric form as the bounding distributions, but it is the set of all possible distributions restricted by the lower and upper bounds. This is an important feature of the proposed approach in this paper.

Traditionally, machine learning methods have used a variety of simplifying assumptions in order to maintain acceptable computational effort required for implementation. The fact that the bounds for the set of probability distributions considered in the regression and classification problems depend on the model parameter makes it clear that any optimisation of risk functionals over the whole set of probability distributions is likely to require an enormous computational effort. It will be illustrated that, for a wide range of popular risk functions, computational is feasible due to new results for the optimisation. In addition to introduction of the general theory, the approach will be illustrated by presenting the resulting optimisation problem formulations for several combinations of loss functions and sets of probability distributions.

Generally, the parameter of a regression model is computed by minimising a risk functional defined by the combination of a certain loss function and a probability distribution for the random noise [10, 33]. When using a set of probability distributions instead of a precise distribution, we can choose a single distribution from this set which minimises or maximises the risk functional; the probability distribution maximising (minimising) the risk functional corresponds to the minimax (minimin) strategy. These cases can be called the ‘pessimistic’ and ‘optimistic’ decisions, respectively. The main problem in finding these two (‘extreme’ or ‘optimal’) precise distributions is that, like the bounds of the corresponding set of distributions, they depend on the unknown regression and classification model parameter which has to be computed. We will identify these optimal probability distributions as functions of the unknown parameter only, which enables us to substitute them into the expression for the risk functional and to compute the optimal model parameter by minimising the risk measure over the set of possible values for the parameter.

The sets of probability distributions can be constructed from training data by a variety of statistical inference methods, including imprecise (‘generalized’) Bayesian inference models [19, 34, 35], nonparametric predictive inference [3] or belief functions [1, 6, 7, 13, 22]. The approach has recently been used in regression modelling with precise statistical data using Kolmogorov-Smirnov (KS) confidence bounds [30] and also includes imprecise Bayesian normal regression [28]. In this paper, there is special attention to the use of extended support vector machines (SVMs) [10, 33] to construct sets of probability distributions in case of interval-valued data, as SVMs are popular tools in machine learning. It will be interesting to implement the general approach presented here with a wide range of methods for constructing the sets of probability distributions and to compare the resulting inferences, for example also with regard to the effect of parameters such as the chosen confidence level if KS bounds are used; this is left as an important topic for future research.

2 Regression and classification in the machine learning framework

The standard learning problem can be formulated as follows [10, 33]. We select the best available function $f(\mathbf{x}, \alpha_{\text{opt}})$ from the set of functions $f(\mathbf{x}, \alpha)$ parameterized by parameter $\alpha \in \Lambda$ (this parameter is typically multi-dimensional), so the function $f(\mathbf{x}, \alpha_{\text{opt}})$ is considered to be the best approximation of the system response. The selection of the desired function is based on a (training) set of n observations $(\mathbf{x}_i, y_i)$, $i = 1, ..., n$, assumed to be independent (conditionally on the assumed model) and identically distributed with probability density function (PDF) $p(\mathbf{x}, y) = p(y \mid \mathbf{x})p(\mathbf{x})$ and CDF $F(\mathbf{x}, y)$. Here $\mathbf{x} \in \mathbb{R}^m$ is a multivariate input and y is a scalar output which takes values from $\mathbb{R}$ for the regression model and from the set $\{-1, 1\}$ for the classification model[3]. The regression and classification models can be regarded as special cases of the general learning problem, the method presented here is widely applicable.

The quality of an approximation $f(\mathbf{x}, \alpha)$ in a regression model is measured by the loss function $L(y, f(\mathbf{x}, \alpha))$ which typically depends on the difference $z = y - f(\mathbf{x}, \alpha)$. Therefore, we use the notation $L(z) = L(y, f(\mathbf{x}, \alpha))$. Common and convenient loss functions are the quadratic loss $L(z) = z^2$, the linear loss $L(z) = |z|$, and the so-called 'ε-insensitive' [33] and 'pinball' loss functions [12]. In classification models, commonly used loss functions are the indicator loss function $L(\mathbf{x}, y) = \mathbf{1}\{\text{sgn}(f(\mathbf{x}, \alpha)) \neq y\}$, the logistic loss, the hinge loss, the squared hinge loss and the least square loss functions [21]. All these loss functions can be implemented in the general approach presented in this paper.

The main goal of learning is to find the optimal parameter α_{opt} which minimises the following risk functional over the parametrized class of functions $f(\mathbf{x}, \alpha)$, $\alpha \in \Lambda$:

$$R(\alpha) = \int\int L(y - f(\mathbf{x}, \alpha))p(\mathbf{x}, y)\mathrm{d}\mathbf{x}\mathrm{d}y.$$

A commonly made assumption for regression models is that the random error (noise) Z, which takes the values $z = y - f(\mathbf{x}, \alpha)$, has mean zero and PDF $p(z \mid \alpha) = p(y \mid \mathbf{x})$, leading to

$$R(\alpha) = \int L(z \mid \alpha)p(z \mid \alpha)\mathrm{d}z.$$

If the joint density $p(\mathbf{x}, y)$ is unknown (or no specific form of it has been assumed), then the risk functional $R(\alpha)$ can be replaced by the *empirical risk functional*

$$R_{\text{emp}}(\alpha) = \frac{1}{n}\sum_{i=1}^{n} L(y - f(\mathbf{x}, \alpha)). \qquad (1)$$

If $p(\mathbf{x}, y)$ is known or of an assumed parametric form, then a common technique for computing α_{opt} is the maximum likelihood estimation method [33].

In this paper we assume that the function f is linear,

$$f(\mathbf{x}, \alpha) = \alpha_0 + \langle \alpha\varphi(\mathbf{x})\rangle$$

with $\langle\cdot\rangle$ the canonical dot product notation. In particular we consider the function with $\varphi_i(x_i) = x_i$, which corresponds to many popular models in learning. The use of more general functions f will be discussed elsewhere.

3 Regression with a set of distributions

Suppose that we do not know the precise CDF of Z, but we know that it belongs to a set $\mathcal{F}(\alpha)$ bounded by lower CDF $\underline{F}(z \mid \alpha)$ and upper CDF $\overline{F}(z \mid \alpha)$ which depend on the parameter α. As mentioned before, these bounds can result from the use of a wide range of inferential methods applied to the observations $(\mathbf{x}_i, y_i)$, $i = 1, ..., n$. It is important to emphasize that the dependence of the lower and upper CDFs on the parameter α is an important feature of the proposed approach. When we have a set of probability distributions instead of a single one, we can construct a corresponding set of regression models. For decision making, it is important to choose some of these models[4], we consider the use of the minimax ('pessimistic') and minimin ('optimistic') strategies to judge the quality of an estimator and hence of the corresponding regression model.

3.1 The minimax strategy

The minimax strategy can be motivated as follows. We do not know (or wish to assume) a precise CDF F and every CDF in $\mathcal{F}(\alpha)$ could be selected. Therefore, we should take the 'worst' distribution providing the largest value of the risk functional. The minimax criterion can be interpreted as an insurance against the worst case because it aims at minimising the expected loss in the least favorable case [20]. The upper risk functional $\overline{R}(\alpha)$ for α is defined as

$$\overline{R}(\alpha) = \max_{F(z \mid \alpha) \in \mathcal{F}(\alpha)} \int L(z \mid \alpha)\mathrm{d}F(z \mid \alpha). \qquad (2)$$

[3] Generally, y might take values in any finite set, the restriction to binary classification is due to space limitations.

[4] Alternative methods for dealing with the set of regression models can be of interest but are not investigated here.

It can be regarded as the upper expectation of the loss function. The optimal parameter α_{opt} is computed by minimising the upper risk functional over the set Λ.

Most loss functions in regression models have one minimum at point 0. Utkin and Destercke [31, 32] have shown that the optimal CDF from the set $\mathcal{F}(\alpha)$ providing the upper bound for $R(\alpha)$ in case of such loss functions is of the form

$$F_U(z) = \begin{cases} \overline{F}(z), & z \leq \overline{F}^{-1}(\tau), \\ \tau, & \overline{F}^{-1}(\tau) < z < \underline{F}^{-1}(\tau), \\ \underline{F}(z), & z \geq \underline{F}^{-1}(\tau), \end{cases} \tag{3}$$

where τ is one of the roots of the equation

$$L\left(\overline{F}^{-1}(\tau)\right) = L\left(\underline{F}^{-1}(\tau)\right).$$

If the loss function is symmetric about 0, then τ can be derived from the equation $\underline{F}^{-1}(\tau) + \overline{F}^{-1}(\tau) = 0$. Using this optimal CDF, the upper risk functional $\overline{R}(\alpha)$ is

$$\overline{R}(\alpha) = \int_{-\infty}^{\overline{F}^{-1}(\tau)} L(z \mid \alpha) \mathrm{d}\overline{F}(z \mid \alpha) + \int_{\underline{F}^{-1}(\tau)}^{\infty} L(z \mid \alpha) \mathrm{d}\underline{F}(z \mid \alpha). \tag{4}$$

The optimal value of parameter α according to the minimax strategy can be derived by minimising $\overline{R}(\alpha)$ over $\alpha \in \Lambda$.

3.2 The minimin strategy

The minimin strategy can be interpreted as corresponding to an 'optimistic' decision, namely a CDF $F(z \mid \alpha) \in \mathcal{F}(\alpha)$ is used which provides the smallest value for the risk functional $R(\alpha)$ for arbitrary values of α. The corresponding lower risk functional for α is defined as

$$\underline{R}(\alpha) = \min_{F(z \mid \alpha) \in \mathcal{F}(\alpha)} \int L(z \mid \alpha) \mathrm{d}F(z \mid \alpha). \tag{5}$$

It can be regarded as the lower expectation of the loss function. The optimal parameter α_{opt} is computed by minimising the lower risk functional over the set Λ.

The optimal CDF from the set $\mathcal{F}(\alpha)$ providing the lower bound for the expectation is

$$F_L(z) = \begin{cases} \underline{F}(z), & z \leq 0, \\ \overline{F}(z), & z > 0. \end{cases} \tag{6}$$

Using this optimal CDF, which has a jump at point $z = 0$, the lower risk functional $\underline{R}(\alpha)$ is

$$\underline{R}(\alpha) = \int_{-\infty}^{0} L(z \mid \alpha) \mathrm{d}\underline{F}(z \mid \alpha) + \int_{0}^{\infty} L(z \mid \alpha) \mathrm{d}\overline{F}(z \mid \alpha). \tag{7}$$

The optimal value of parameter α according to the minimin strategy can be derived by minimising $\underline{R}(\alpha)$ over $\alpha \in \Lambda$.

4 Regression with interval-valued observations

Suppose that the training set consists of n independent observations $(\mathbf{x}_i, \mathcal{Y}_i)$, $i = 1, ..., n$, with intervals $\mathcal{Y}_i = [\underline{y}_i, \overline{y}_i]$ instead of point-valued observations[5]. This implies that the random noise Z takes values in intervals $\mathcal{Z}_i(\alpha)$ such that $y - f(\mathbf{x}_i, \alpha) \in \mathcal{Z}_i(\alpha)$ for all $y \in \mathcal{Y}_i$. The question that needs to be addressed is how to proceed with the interval-valued training set in the framework of predictive learning.

There are several ways in which one could deal with such an interval-valued data set. In this paper, we construct the lower and upper CDFs for a set of probability distributions corresponding to the available information through a chosen inferential method out of a wide range of possibilities, as discussed before. This set depends on the parameter α because the intervals $\mathcal{Z}_i(\alpha)$, $i = 1, ..., n$ are functions of α. With such intervals $\mathcal{Z}_i(\alpha)$, the same approach as proposed by Utkin and Coolen [30], who used p-boxes corresponding to Kolmogorov-Smirnov bounds, can be applied for parameter optimisation in the regression model under the minimax and minimin scenarios. Denoting the boundary points of intervals $\mathcal{Z}_i(\alpha)$ by $\underline{\mathcal{Z}}_i(\alpha) = \underline{y}_i - f(\mathbf{x}_i, \alpha)$ and $\overline{\mathcal{Z}}_i(\alpha) = \overline{y}_i - f(\mathbf{x}_i, \alpha)$, a p-box can be constructed from the observed intervals in the framework of Dempster-Shafer theory [6, 22]. If we assume for simplicity that every observation interval occurs only once, then

$$\underline{F}(z \mid \alpha) = \mathrm{Bel}((-\infty, z]) = n^{-1} \sum_{i: \overline{\mathcal{Z}}_i(\alpha) \leq z} 1,$$

$$\overline{F}(z \mid \alpha) = \mathrm{Pl}((-\infty, z]) = n^{-1} \sum_{i: \underline{\mathcal{Z}}_i(\alpha) \leq z} 1.$$

If some intervals occur more than once then the corresponding CDFs follow straightforwardly. These lower and upper CDFs, which depend on the parameter α, can be used for dealing with interval-valued y in regression, as is illustrated next for several scenarios.

[5] The method presented in this paper can also deal with interval-valued input variables $\mathbf{x}_i$. Due to space limitations, for regression the presentation is restricted to point-valued input variables, but for classification (Section 5) interval-valued input variables are used. Throughout, the intervals are not restricted, hence they can be any interval of possible values upto the whole of $(-\infty, \infty)$.

4.1 The minimax strategy

With the lower and upper CDFs corresponding to the interval-valued observations as discussed above, the upper risk functional in (4) is

$$\begin{aligned}\overline{R}(\alpha) = n^{-1} & \sum_{i:\underline{\mathcal{Z}}_i(\alpha)\leq \overline{F}^{-1}(\tau)} L(\underline{\mathcal{Z}}_i(\alpha)) \\ + n^{-1} & \sum_{i:\overline{\mathcal{Z}}_i(\alpha)\geq \underline{F}^{-1}(\tau)} L(\overline{\mathcal{Z}}_i(\alpha)).\end{aligned}$$

with τ such that $\underline{F}^{-1}(\tau) = -\overline{F}^{-1}(\tau)$. Note that this upper risk functional uses, for every α, only the boundary points $\underline{\mathcal{Z}}_i(\alpha)$ and $\overline{\mathcal{Z}}_i(\alpha)$ of the intervals $\mathcal{Z}_i(\alpha)$. This feature is important as it significantly simplifies computation of the optimal parameter α_{opt}.

The upper risk functional for the minimax strategy with a fixed α can be written as the upper expectation corresponding to basic probability assignments [15, 24], giving

$$\overline{R}(\alpha) = n^{-1} \sum_{i=1}^{n} \max_{z\in[\underline{\mathcal{Z}}_i(\alpha),\overline{\mathcal{Z}}_i(\alpha)]} L(z).$$

We also concluded that this upper risk functional is achieved at boundary points of intervals $\mathcal{Z}_i$, with

$$\overline{R}(\alpha) = n^{-1} \sum_{i=1}^{n} \max\left\{L(\underline{\mathcal{Z}}_i(\alpha)), L(\overline{\mathcal{Z}}_i(\alpha))\right\}.$$

It should be pointed out that, if all observations are precise ('point-valued'), so $\underline{y}_i = \overline{y}_i = y_i$, this upper risk functional is equal to the standard empirical risk functional (1). We can now consider some of the most important loss function in regression, where the optimal parameter α_{opt} under minimax can be obtained by minimising $\overline{R}(\alpha)$ over all $\alpha \in \Lambda$.

4.1.1 Quadratic loss function

We consider the quadratic loss function $L(z) = z^2$, the most popular one in classical regression theory and applications. To minimise the corresponding upper risk functional we have to solve the optimisation problem functional:

$$\min_{\alpha} \left(\sum_{i=1}^{n} \max\left\{\underline{\mathcal{Z}}_i^{\,2}(\alpha), \overline{\mathcal{Z}}_i^{\,2}(\alpha)\right\} \right). \tag{8}$$

Introducing new optimisation variables G_i, $i = 1, ..., n$, such that $G_i^2 = \max\left\{\underline{\mathcal{Z}}_i^{\,2}(\alpha), \overline{\mathcal{Z}}_i^{\,2}(\alpha)\right\}$, problem (8) can be rewritten as

$$\min_{\alpha, G_i} \sum_{i=1}^{n} G_i^2, \tag{9}$$

subject to

$$\begin{aligned}& G_i \geq \underline{\mathcal{Z}}_i(\alpha), \quad G_i \geq \overline{\mathcal{Z}}_i(\alpha), \\ & G_i \geq -\underline{\mathcal{Z}}_i(\alpha), \ G_i \geq -\overline{\mathcal{Z}}_i(\alpha), \ i = 1, ..., n. \end{aligned} \tag{10}$$

The third and fourth constraints take into account the fact that residuals may be negative. If we assume that the function $f(\mathbf{x}, \alpha)$ is linear, i.e., $f(\mathbf{x}, \alpha) = \alpha_0 + \langle \alpha \mathbf{x} \rangle$, then the optimisation problem specified by (9) and (10) is a well-known quadratic programming problem with the optimisation variables α and G_i, $i = 1, ..., n$, which can be solved by means of standard methods.

4.1.2 Linear and pinball loss function

The pinball loss function with parameter $\tau \in [0, 1]$ is given by [12]

$$L_\tau(z) = \left\{ \begin{array}{cc} \tau z, & z > 0, \\ (\tau - 1)z, & z \leq 0. \end{array} \right.$$

The linear loss function is the special case of the pinball loss function with $\tau = 1$. We consider calculation of the optimal parameter of the regression model using the minimax criterion with the pinball loss function. We introduce new optimisation variables G_i, $i = 1, ..., n$, such that $G_i = \max\left\{L_\tau(\underline{\mathcal{Z}}_i(\alpha)), L_\tau(\overline{\mathcal{Z}}_i(\alpha))\right\}$. The condition $z \geq 0$ implies the condition $G_i \geq \tau \cdot z$. However, if $G_i \geq \tau \cdot z$ and $z \geq 0$, then $G \geq \tau \cdot z - z$. On the other hand, the condition $z < 0$ implies the condition $G_i \geq (\tau - 1) \cdot z = \tau \cdot z - z$. However, if $G_i \geq \tau \cdot z - z$ and $z < 0$, then $G_i \geq \tau \cdot z$. Finally, the condition $G_i \geq L_\tau(z)$ can be represented by means of two constraints $G_i \geq \tau \cdot z$ and $G_i \geq \tau \cdot z - z$, which simultaneously 'cover' all possible values of z. This implies that the optimisation problem for computing the optimal regression parameter can be written as

$$\min_{\alpha, G_i} \sum_{i=1}^{n} G_i, \tag{11}$$

subject to

$$\begin{aligned}& G_i \geq \tau \cdot \underline{\mathcal{Z}}_i(\alpha), \ G_i \geq \tau \cdot \overline{\mathcal{Z}}_i(\alpha), \\ & G_i \geq (\tau - 1) \cdot \underline{\mathcal{Z}}_i(\alpha), \\ & G_i \geq (\tau - 1) \cdot \overline{\mathcal{Z}}_i(\alpha), \ i = 1, ..., n. \end{aligned} \tag{12}$$

If we assume that the function $f(\mathbf{x}, \alpha)$ is linear, then this is a well-known linear programming problem.

4.2 SVM

Let us return to the case with the linear loss function and the minimax strategy, and compare the obtained optimisation problem with the popular SVM approach [10, 21, 33] which in regression is also called 'support vector regression'. The ε-insensitive loss function is applied in the corresponding regression models [33]. If all observations are point-valued, so $\underline{y}_i = \overline{y}_i = y_i$, then according to the standard SVR approach, parameter α is determined by the quadratic programming problem

$$\min_{\alpha} \left(\frac{1}{2} \langle \alpha, \alpha \rangle + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \right) \tag{13}$$

subject to

$$\begin{aligned} &\xi_i \geq 0,\ \xi_i + \varepsilon \geq (\langle \alpha \mathbf{x}_i \rangle + \alpha_0) - y_i, \\ &\xi_i^* \geq 0,\ \xi_i^* + \varepsilon \geq y_i - (\langle \alpha \mathbf{x}_i \rangle + \alpha_0), i = 1, .., n. \end{aligned} \tag{14}$$

Here C is a constant 'cost' parameter, ξ_i, ξ_i^*, $i = 1, ..., n$, are slack variables, and $\frac{1}{2} \langle \alpha, \alpha \rangle$ is the Tikhonov regularization term (the most popular penalty or smoothness term) [27] which enforces uniqueness by penalizing functions with wild oscillation and effectively restricting the space of admissible solutions [8]. The positive slack variables ξ_i, ξ_i^* represent the distance from y_i to the corresponding boundary values of the ε-tube.

The constraints (12) and (14) coincide if the variables G_i coincide with the slack variables ξ_i, ξ_i^* and $\underline{y}_i = \overline{y}_i$, $\varepsilon = 0$, $\tau = 1$. Consequently, the proposed approach for constructing the regression model with interval-valued data, supplemented by the regularization term and the constant 'cost' parameter C, can be regarded as an extension of the SVM approach to the case of interval-valued data, i.e. we have the same objective function and the following constraints in terms of SVR for every $i = 1, ..., n$:

$$\begin{aligned} \xi_i \geq 0, &\quad \xi_i + \varepsilon \geq (\langle \alpha \mathbf{x}_i \rangle + \alpha_0) - \underline{y}_i, \\ \xi_i \geq 0, &\quad \xi_i + \varepsilon \geq (\langle \alpha \mathbf{x}_i \rangle + \alpha_0) - \overline{y}_i, \\ \xi_i^* \geq 0, &\quad \xi_i^* + \varepsilon \geq \underline{y}_i - (\langle \alpha \mathbf{x}_i \rangle + \alpha_0), \\ \xi_i^* \geq 0, &\quad \xi_i^* + \varepsilon \geq \overline{y}_i - (\langle \alpha \mathbf{x}_i \rangle + \alpha_0). \end{aligned}$$

Now the slack variables ξ_i, ξ_i^* are additionally constrained and represent the largest distance from $\underline{y}_i$ and $\overline{y}_i$ to the corresponding boundary values of the ε-tube, respectively. This implies that the minimax strategy searches for the largest residuals (or 'margins' in terms of classification) from all residuals in every interval $\mathcal{Z}_i$, $i = 1, ..., n$. The corresponding dual optimisation problem is

$$\begin{aligned} \max \Bigg(& -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (Q_i - T_i)(Q_j - T_j) \langle \mathbf{x}_i \mathbf{x}_j \rangle \\ & - \varepsilon \sum_{i=1}^{n} (Q_i + T_i) - \sum_{i=1}^{n} \underline{y}_i (Q_i - T_i) \\ & + \sum_{i=1}^{n} (\overline{y}_i - \underline{y}_i)(\varphi_i^* - \varphi_i) \Bigg), \end{aligned}$$

subject to

$$\sum_{i=1}^{n} (Q_i - T_i) = 0,\ 0 \leq Q_i \leq C,\ 0 \leq T_i \leq C.$$

Here $\psi_i, \psi_i^*, \varphi_i, \varphi_i^*$ are Lagrange multipliers and $Q_i = \psi_i + \varphi_i$, $T_i = \psi_i^* + \varphi_i^*$.

It can be seen from this dual optimisation problem that in the regression model we use a point in every observation interval which is a linear combination of its bounds $\underline{y}_i$ and $\overline{y}_i$ with coefficients determined by the values of the Lagrange multipliers. If $\overline{y}_i = \underline{y}_i$ we get the dual optimisation problem of the standard SVM method with variables Q_i and T_i.

If the quadratic loss function is used instead of the ε-insensitive loss function, then the proposed regression model (optimisation problem (9)-(10)) is the 'least squares SVM' approach [25] which is solved through a system of linear equations.

4.3 The minimin strategy

Using the lower and upper CDFs corresponding to the interval-valued observatons, as discussed at the start of this section, we can rewrite the lower risk functional (7) as

$$\begin{aligned} \underline{R}(\alpha) = n^{-1} & \sum_{i:\overline{\mathcal{Z}}_i(\alpha) \leq 0} L(\overline{\mathcal{Z}}_i(\alpha)) \\ & + n^{-1} \sum_{i:\underline{\mathcal{Z}}_i(\alpha)) \geq 0} L(\underline{\mathcal{Z}}_i(\alpha)). \end{aligned}$$

As in the case of the upper risk function, this lower risk functional is, for every α, defined only by the boundary points $\underline{\mathcal{Z}}_i(\alpha)$ and $\overline{\mathcal{Z}}_i(\alpha)$ of the intervals $\mathcal{Z}_i(\alpha)$. However, not all observation intervals contribute to the lower risk functional because the optimal CDF has a jump at point 0.

The lower risk functional for the minimin strategy with a fixed α can be written as the lower expectation corresponding to basic probability assignments [15, 24], giving

$$\underline{R}(\alpha) = n^{-1} \sum_{i=1}^{n} \min_{z \in [\underline{\mathcal{Z}}_i(\alpha), \overline{\mathcal{Z}}_i(\alpha)]} L(z). \tag{15}$$

It follows that the risk measure is 0 if there exist one or more values of α such that $0 \in [\underline{\mathcal{Z}}_i(\alpha), \overline{\mathcal{Z}}_i(\alpha)]$ for every $i = 1, ..., n$. If this is the case for multiple vectors α, one can consider to have found several 'perfect fits' to the available data, which either could be considered all together (this would be in line with some fundamental ideas behind imprecise probability) or which could be compared by a secondary criterion (the same comment applies generally if there are multiple optimal vectors α). This is an interesting topic for future research, for now let us assume that a unique best estimate of α can be obtained and that the corresponding lower risk functional is positive (so there is no 'perfect fit'). A term in the objective function is non-zero if one of the following two conditions holds

$$\overline{\mathcal{Z}}_i(\alpha) < 0, \quad \underline{\mathcal{Z}}_i(\alpha) > 0.$$

Let us consider the pinball loss function for this situation. Introducing new optimisation variables H_i, $i = 1, ..., n$, it is easy to prove that the optimisation problem can be written as

$$\min_{\alpha, H_i} \sum_{i=1}^{n} H_i,$$

subject to

$$\begin{aligned} & H_i \geq \tau \underline{\mathcal{Z}}_i(\alpha), \ H_i \geq (\tau - 1)\overline{\mathcal{Z}}_i(\alpha), \\ & H_i \geq 0, \ i = 1, ..., n. \end{aligned}$$

The quadratic loss function leads to the similar problem with H_i replaced by G_i^2, minimisation over G_i, and $\tau = 1$. These are well-known optimisation problems that can be solved efficiently by standard methods.

4.4 SVM

We consider the case with the linear loss function under the minimin strategy and derive the optimisation problem in the SVM framework. By using the standard Tikhonov regularization term, we can formulate the following convex optimisation problem

$$\min_{\alpha} \left(\frac{1}{2} \langle \alpha, \alpha \rangle + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \right),$$

subject to

$$\begin{aligned} & \xi_i \geq 0, \ \xi_i + \varepsilon \geq (\langle \alpha \mathbf{x}_i \rangle + \alpha_0) - \overline{y}_i, \ i = 1, ..., n, \\ & \xi_i^* \geq 0, \ \xi_i^* + \varepsilon \geq \underline{y}_i - (\langle \alpha \mathbf{x}_i \rangle + \alpha_0), \ i = 1, ..., n. \end{aligned}$$

This is a quadratic programming problem, with slack variables ξ_i, ξ_i^* representing the distance from $\overline{y}_i$ and $\underline{y}_i$ to the corresponding lower and upper boundary values of the ε-tube, respectively. The minimin strategy searches for the smallest residuals in each interval $\mathcal{Z}_i$, $i = 1, ..., n$, under condition that there are positive residuals. As in Subsection 4.2, the corresponding dual optimisation problem provides further insights into the optimal solution, a detailed analysis will be presented elsewhere.

5 Classification with interval-valued observations

We consider classification problems where the system output y is restricted to two values, the proposed method can be generalized to more possible values. The input variables (patterns) $\mathbf{x}$ may be interval-valued. Suppose that we have a training set $(\mathcal{X}_i, y_i)$, $i = 1, ..., n$. Here $\mathcal{X}_i \subset \mathbb{R}^m$ is the Cartesian product of m intervals $[\underline{x}_k^{(i)}, \overline{x}_k^{(i)}]$, $k = 1, ..., m$, which again are not restricted so could even include intervals $(-\infty, \infty)$, and $y_i \in \{-1, 1\}$. Let the $n_{-1} = r$ observations $\mathcal{X}_i$ with $i = 1, \ldots, r$ correspond to the class (with) $y = -1$ and the $n_{+1} = n - r$ observations $\mathcal{X}_i$ with $i = r + 1, \ldots, n$ correspond to the class $y = 1$.

The risk functional can be written as $R(\alpha) = R_{-1}(\alpha) + R_{+1}(\alpha)$, with

$$\begin{aligned} R_y(\alpha) = & \int_{\mathbb{R}^n} L(\mathbf{x}, y) \mathrm{d}F(\mathbf{x}, y) \\ & = \pi_y \int_{\mathbb{R}^n} L(\mathbf{x}, y) \mathrm{d}F(\mathbf{x} \mid y), \end{aligned}$$

where $\pi_y = p(y)$ is a prior probability[6] for class y. Suppose that the CDFs $F(\mathbf{x} \mid y)$ are unknown. As discussed before, a wide range of inferential methods can be chosen to, in combination with the dataset containing interval-valued observations, produce a set of CDFs $F(\mathbf{x} \mid y)$. One additional obstacle due to the interval-valued input variables is that $\mathbf{x}$ is a vector so now p-boxes of multivariate distributions must be constructed. We propose that this problem can be resolved as follows. Note that interval-valued data $\mathbf{x}$ lead to an interval-valued discriminant function $f(\mathbf{x}, \alpha)$ whose parameter α is unknown and has to be determined. Therefore, in contrast to many alternative approaches in classification, we propose to consider the CDF $F(f \mid y)$ instead of the multivariate CDF $F(\mathbf{x} \mid y)$. This is briefly discussed further below, detailed explanation and illustrations will be presented elsewhere. With this change, the risk functional becomes

$$R_y(\alpha) = \pi_y \int_{\mathbb{R}} L(f \mid y) \mathrm{d}F(f \mid y).$$

[6] Choice of prior probabilities is not addressed here. However, it is worth noting that generalization to allow imprecise prior probabilities is possible.

But we allowed explicitly the use of a set of CDFs, so now consider the set $\mathcal{F}(y)$ of probability distributions produced by lower CDF $\underline{F}(f \mid y)$ and upper CDF $\overline{F}(f \mid y)$ CDFs , i.e.

$$\mathcal{F}(y) = \{F(f) \mid \forall f \in \mathbb{R}, \underline{F}(f|y) \leq F(f) \leq \overline{F}(f|y)\}.$$

It is important to emphasize that, although we have not explicitly included α in the notation for these distribution sets, $\mathcal{F}(y)$ depends on the parameter α because f is a function of α and the lower and upper CDFs depend on α. We introduce notation

$$f_L = \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x},\alpha), \quad f_U = \max_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x},\alpha).$$

If the function f is linear, then the lower and upper bounds for the discriminant functions are determined only by the bounds of pattern intervals, i.e.

$$\begin{aligned} f_L &= \min_{x_k\in\{\underline{x}_k,\overline{x}_k\},\ k=1,\dots,m} f(\mathbf{x},\alpha), \\ f_U &= \max_{x_k\in\{\underline{x}_k,\overline{x}_k\},\ k=1,\dots,m} f(\mathbf{x},\alpha). \end{aligned}$$

This property is also valid for arbitrary monotone discriminant functions. For every interval-valued observation $(\mathcal{X}_i, y_i)$, we have the interval $\mathbf{f}_i = [f_{L,i}, f_{U,i}]$ of values of the discriminant function. These intervals depend on the parameter α, so the bounds $f_{L,i}$ and $f_{U,i}$ cannot be computed explicitly, but inference is again possible in many important scenarios through specification of the optimisation problems involved, and the use of standard algorithms to solve such problems. We illustrate this next for the minimax strategy, methods for the minimim strategy can be developed similarly and will be presented elsewhere.

5.1 The minimax strategy

According to the minimax strategy, we select a probability distribution from the set $\mathcal{F}(-1)$ and a probability distribution from the set $\mathcal{F}(+1)$ such that the risk measures $R_{-1}(\alpha)$ and $R_{+1}(\alpha)$ achieve their maxima for every fixed α. It must be emphasized that the 'optimal' probability distributions may be different for different values of parameter α, which implies that the corresponding 'optimal' probability distributions depend on α. Since the sets $\mathcal{F}(-1)$ and $\mathcal{F}(1)$ are obtained independently for $y = -1$ and $y = 1$, the upper risk functional with respect to the minimax strategy is of the form

$$\overline{R}(\alpha) = \max_{F(f|-1)\in\mathcal{F}(-1)} R_{-1}(\alpha) + \max_{F(f|1)\in\mathcal{F}(1)} R_{+1}(\alpha).$$

For many popular loss functions in such classification the loss function $L(f,-1)$ is increasing. If this is the case, then the upper bound for $R_{-1}(\alpha)$ is achieved at the distribution $\underline{F}(f,-1)$, hence

$$\overline{R}_{-1}(\alpha) = \int_{\mathbb{R}} L(f,-1)\mathrm{d}\underline{F}(f,-1).$$

In this case the function $L(f,1)$ is decreasing, so

$$\overline{R}_{+1}(\alpha) = \int_{\mathbb{R}} L(f,1)\mathrm{d}\overline{F}(f,1).$$

The upper expectation $\overline{R}_{-1}(\alpha)$ corresponding to given basic probability assignments $m(\mathbf{f}_i) = r^{-1}$ for intervals $\mathbf{f}_i$, $i = 1,...,r$, can be derived for fixed α by [15, 24]

$$\begin{aligned} \overline{R}_{-1}(\alpha) &= r^{-1}\sum_{i=1}^{r} \max_{f\in[f_{L,i}(\alpha),f_{U,i}(\alpha)]} L(f,-1) \\ &= r^{-1}\sum_{i=1}^{r} L(f_{U,i}(\alpha),-1). \end{aligned}$$

And similarly, the corresponding upper expectation $\overline{R}_{+1}(\alpha)$ is

$$\overline{R}_{+1}(\alpha) = (n-r)^{-1}\sum_{i=r+1}^{n} L(f_{L,i}(\alpha),1).$$

Finally, we minimise $\overline{R}(\alpha)$ to compute α_{opt}, with

$$\begin{aligned} \overline{R}(\alpha) &= \frac{\pi_-}{r}\sum_{i=1}^{r} L(f_{U,i}(\alpha),-1) \\ &\quad + \frac{\pi_+}{n-r}\sum_{i=r+1}^{n} L(f_{L,i}(\alpha),1). \end{aligned}$$

Further steps towards the solution of the problem depend on the chosen loss function, we briefly consider one important special case. For the hinge loss function $L(\mathbf{x},y) = \max(1-yf,0)$,

$$\begin{aligned} \overline{R}(\alpha) &= \frac{\pi_-}{r}\sum_{i=1}^{r} \max\left(0, 1 + f_{U,i}(\alpha)\right) \\ &\quad + \frac{\pi_+}{n-r}\sum_{i=r+1}^{n} \max\left(0, 1 - f_{L,i}(\alpha)\right). \end{aligned}$$

After simple modifications, we get the linear problem

$$\min_{\alpha}\left(\frac{\pi_-}{r}\sum_{i=1}^{r} G_i + \frac{\pi_+}{n-r}\sum_{i=r+1}^{n} G_i\right) \qquad (16)$$

subject to

$$\begin{aligned} &G_i \geq 1 - y_i\left(\langle\alpha\mathbf{x}_i\rangle + \alpha_0\right),\ \forall x_k^{(i)} \in \{\underline{x}_k^{(i)}, \overline{x}_k^{(i)}\}, \\ &G_i \geq 0,\ i = 1,...,n. \qquad (17) \end{aligned}$$

By adding the standard Tikhonov regularization term to the objective function, we get the SVM classifier with cost parameters $C_- = \pi_-/r$ and $C_+ = \pi_+/(n - r)$. We introduce notation

$$Q_i = \sum_{k \in J_i} \psi_{ik},\ T_j(i) = \sum_{k \in J_i(j)} \psi_{ik} x_j^{(i,k)}$$

where the set $J_i(j)$ is a 'projection' of the set of indices on the j-th element of the vector $\mathbf{x}_i$. Then the dual optimisation problem is

$$\max \left(\sum_{i=1}^{n} Q_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \left(\sum_{v=1}^{m} T_v(i) T_v(j) \right) \right),$$

subject to

$$\sum_{i=1}^{n} y_i Q_i = 0,\ 0 \leq Q_i \leq C_-,\ i = 1, ..., r,$$
$$0 \leq Q_i \leq C_+,\ i = r+1, ..., n.$$

This is the SVM classification approach with interval-valued data under the minimax strategy. Space restrictions prevent further details, illustration or discussion of this result and related results for different loss functions and for the minimin strategy. However, it is clear that the general approach presented in this paper leads to a wide variety of attractive methods for machine learning, with relatively straightforward inclusion of interval-valued observations.

6 Concluding remarks

In this paper, a new class of imprecise regression and classification models has been proposed which are capable to deal with interval-valued data as frequently occur in practice. The class has been illustrated for several important specific cases, and it has been shown that the resulting inference problems can be formulated as standard optimisation problems, so the method can be implemented using readily available software. This new method has several important features. First, it has a clear explanation and justification in the decision making framework. Secondly, it allows a wide variety of inferential methods for constructing the p-boxes. For example, imprecise ('generalized') Bayesian inference models [19] can be used and these provide an exciting opportunity for developing learning models for a wide range of different applications. Thirdly, the method can deal with (partly) missing data as the intervals for observations are not restricted, which is important as complete data sets are the exception in practice. Finally[7], resulting statistical inferences are similar some well-known robust statistics methods, for which the current approach provides formal justifications and interpretations in a decision theoretic framework. Detailed study of these aspects, and development of further models and corresponding inferences, is ongoing. The main disadvantage of the proposed approach is that it is often not straightforward how the bounding CDFs can be explicitly defined as functions of the regression or classification parameter, which may add to computational complexity but the results show that the approach can be developed to allow real-world applications. A main strength of the proposed method is the link with the popular SVM approach. A key feature of SVMs is the use of kernels which are functions that transform the input data to a high-dimensional space where the learning problem is solved. Such kernel functions can be linear or nonlinear, which will allow us to significantly extend the class of regression or discriminant functions that can be used. Our approach directly showed how the regular SVM approach can be generalized for dealing with interval-valued observations.

There are interesting possibilities for combining corresponding 'minimin' and 'minimax' strategies. For example, the method for cautious decision making proposed by Utkin and Augustin [29], which uses the extreme points of a set of probability distributions produced by imprecise data, can be applied. In our approach, the values of the extreme points are determined from the optimal CDFs (3) and (6) for the minimax and minimin strategies, respectively. Detailed analysis of this cautious strategy and the possibility to arrive at set-based predictions and related final decisions on the basis of our model outputs, are interesting topics for future research, together with dealing with imprecise input variables for the object to predict, imprecision in the dependent variables and of course comparison with more established methods.

Acknowledgements

We thank two referees for detailed comments and suggestions which all have improved this paper and will guide future research.

References

[1] A. Arequi, T. Denoeux. Constructing predictive belief functions from continuous sample data using confidence bands. *ISIPTA'07*[8], pp 11-20, 2007.

[2] C. Angulo, D. Anguita, L. Gonzalez-Abril, J.A. Ortega. Support vector machines for interval discriminant analysis. *Neurocomputing*, 71:1220–1229, 2008.

[7] This was discussed by Utkin and Coolen [30] for p-boxes based on Kolmogorov-Smirnov bounds

[8] Proceedings ISIPTA conferences available from *www.sipta.org*

[3] T. Augustin, F.P.A. Coolen. Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference*, 124:251–272, 2004.

[4] E. Carrizosa, J. Gordillo, F. Plastria. Classification problems with imprecise data through separating hyperplanes. Technical Report MOSI/33, Vrije Universiteit Brussel, 2007.

[5] G. de Cooman, M. Zaffalon. Updating beliefs with incomplete observations. *Artificial Intelligence*, 159:75–125, 2004.

[6] A.P. Dempster. Upper and lower probabilities induced by a multi-valued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

[7] S. Destercke, D. Dubois, E. Chojnacki. Unifying practical uncertainty representations - i: Generalized p-boxes. *International Journal of Approximate Reasoning*, 49:649–663, 2008.

[8] T. Evgeniou, T. Poggio, M. Pontil, A. Verri. Regularization and statistical learning theory for data analysis. *Computational Statistics & Data Analysis*, 38:421–432, 2002.

[9] P.Y. Hao. Interval regression analysis using support vector networks. *Fuzzy Sets and Systems*, 60:2466–2485, 2009.

[10] T. Hastie, R. Tibshirani, J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York, 2001.

[11] H. Ishibuchi, H. Tanaka, N. Fukuoka. Discriminant analysis of multi-dimensional interval data and its application to chemical sensing. *International Journal of General Systems*, 16:311–329, 1990.

[12] R. Koenker, G. Bassett. Regression quantiles. *Econometrica*, 46:33–50, 1978.

[13] E. Kriegler, H. Held. Utilizing belief functions for the estimation of future climate change. *International Journal of Approximate Reasoning*, 39:185–209, 2005.

[14] E.A. Lima Neto, F.A.T. de Carvalho. Centre and range method to fitting a linear regression model on symbolic interval data. *Computational Statistics and Data Analysis*, 52:1500–1515, 2008.

[15] H.T. Nguyen, E.A. Walker. On decision making using belief functions. In: R.Y. Yager, M. Fedrizzi, J. Kacprzyk (Eds), *Advances in the Dempster-Shafer Theory of Evidence*. Wiley, New York, pp 311–330, 1994.

[16] P. Nivlet, F. Fournier, J.J. Royer. Interval discriminant analysis: An efficient method to integrate errors in supervised pattern recognition. *ISIPTA'01*, pp 284-292, 2001.

[17] K. Pelckmans, J. De Brabanter, J.A.K. Suykens, B. De Moor. Handling missing values in support vector machine classifiers. *Neural Networks*, 18:684 – 692, 2005.

[18] S. Petit-Renaud, T. Denoeux. Nonparametric regression analysis of uncertain and imprecise data using belief functions. *International Journal of Approximate Reasoning*, 35:1–28, 2004.

[19] E. Quaeghebeur, G. de Cooman. Imprecise probability models for inference in exponential families. *ISIPTA'05*, pp 287–296, 2005.

[20] C.P. Robert. *The Bayesian Choice*. Springer, New York, 1994.

[21] B. Scholkopf, A.J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, 2002.

[22] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[23] A. Silva, P. Brito. Linear discriminant analysis for interval data. *Computational Statistics*, 21:289–308, 2006.

[24] T.M. Strat. Decision analysis using belief functions. *International Journal of Approximate Reasoning*, 4:391–418, 1990.

[25] J.A.K. Suykens, J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9:293–300, 1999.

[26] H. Tanaka, H. Lee. Interval regression analysis by quadratic programming approach. *IEEE Transactions on Fuzzy Systems*, 6:473–481, 1998.

[27] A.N. Tikhonov, V.Y. Arsenin. *Solution of Ill-Posed Problems*. W.H. Winston, Washington DC, 1977.

[28] L.V. Utkin. Regression analysis using the imprecise Bayesian normal model. *International Journal of Data Analysis Techniques and Strategies*, 2:356–372, 2010.

[29] L.V. Utkin, T. Augustin. Efficient algorithms for decision making under partial prior information and general ambiguity attitudes. *ISIPTA'05*, pp 349–358, 2005.

[30] L.V. Utkin, F.P.A. Coolen. On reliability growth models using Kolmogorov-Smirnov bounds. *International Journal of Performability Engineering*, 7:5–19, 2011.

[31] L.V. Utkin, S. Destercke. Computing expectations with p-boxes: two views of the same problem. *ISIPTA'07*, pp 435–444, 2007.

[32] L.V. Utkin, S. Destercke. Computing expectations with continuous p-boxes: Univariate case. *International Journal of Approximate Reasoning*, 50:778 – 798, 2009.

[33] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[34] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[35] P. Walley. Inferences from multinomial data: Learning about a bag of marbles. (With discussion) *Journal of the Royal Statistical Society, Series B*, 58:3–57, 1996.

[36] G. Walter, T. Augustin, A. Peters. Linear regression analysis under sets of conjugate priors. *ISIPTA'07*, pp 445–455, 2007.