On the median in imprecise ordinal problems

Sébastien Destercke

Noname manuscript No.
(will be inserted by the editor)

On the median in imprecise ordinal problems

Sébastien Destercke

the date of receipt and acceptance should be inserted later

Abstract When having to make a prediction under probabilistic uncertainty in ordinal problems, the median offers a number of interesting properties compared to other statistics such as the expected value. In particular, it does not depend on a particular metric defined over the elements, but still takes account of the ordinal nature of the data. It can also be shown to be the minimizer of the L_1 loss function. In this paper, we show that similar results can be obtained when the uncertainty is described not by a single probability distribution, but by a convex set of those. In particular, we relate the lower and upper medians to the L_1 loss function via the notion of lower and upper expectations (and extend these results to general quantiles). We also show that, using a different decision rule, the lower and upper median can be retrieved when assuming the cost to be strictly monotonic and symmetric, and nothing more. Finally, we run some tests to show the interest of using Median based predictions with convex sets of probabilities in ordinal regression problems.

keywords: Ordinal space, ordinal classification, ordinal regression, imprecise probabilities, median, sign-desirability

1 Introduction

There are many practical problems in which ordinal variables appear, as they are instrumental to model spaces where there exists a natural ordering of possible values, but where a numerical treatment of those values is unwarranted. For example, the rating of movies can be one of the following labels: `Very-Bad`, `Bad`, `Average`, `Good`, `Very-Good`, that are ordered from the worst situation to the best, but without claiming that a `Very-Bad` movie is five times worse than a `Very-Good`. Other examples include the selection of applicants for a position, the selection of papers for conferences (`Reject` statement is worse than `accept`, but they are not numerically related), judging the crisis level of some situations, the risk associated to particular loans or contracts for banks and insurances, ...

Université de Technologie de Compiegne U.M.R. C.N.R.S. 7253 Heudiasyc Centre de recherches de Royallieu F-60205 Compiegne Cedex FRANCE
Tel: +33 (0)3 44 23 79 85
Fax: +33 (0)3 44 23 44 77
E-mail: sebastien.destercke@hds.utc.fr, gen.yang@hds.utc.fr

Two of the main fields where such variables appear are the multi-criteria decision making and ordinal classification. In multi-criteria decision making, it is often easier to define an order between possible values of a criteria than to associate meaningful numbers (i.e., utility functions) to each of these values, unless such numbers (e.g., monetary units) already exist. Such ordinal variables can then be exploited by adequate methods [1,2] or by adapting existing numerical ones to an ordinal setting [3]. Ordinal classification [4,5] (or ordinal regression [6,7,8]), on the other hand, aims at learning from a set of examples a model that predict the output label of a new instance when the finite set of possible *labels* (elements, classes, ...) is naturally ordered.

Note that in these problems, ordinal variables should be treated differently than nominal ones (e.g., appearing in multi-class classification) and continuous or numerical ones, since in the former case there is no ordering between elements and in the latter there usually exists a metric on the outputs (value 5 is five times bigger than value 1). When our knowledge of the ordinal variable value is modelled by a probability, the median plays an important role as a predictive value. Indeed, in contrast to the notion of expected value, the median does not depend on a particular metric defined over the elements, therefore not facing the problem of defining such a metric, and in contrast with the notion of modal (most probable) value, it takes account of the ordinal nature of the labels.

Just as the modal value can be associated to the minimization of the $0/1$ loss function, and the expected value to the minimization of the L_2 loss function, the median is known to be the minimizer of the L_1 loss function [9]. This result, by connecting the median with usual loss minimization approaches, provides it with a sound interpretation in learning problems, and allows one to learn a model by directly minimizing the L_1 loss function rather than by estimating a probabilistic model.

All those results, however, rely on the fact that the estimated probabilistic model is reliable, in the sense that it is close to the theoretical unknown distribution. While this a reasonable assumption when sufficient, precise data are available, it may become unreasonable when data are scarce, missing or noisy. In such situations, imprecise probability approaches [10,11] may be instrumental. These approaches consider not a single but a (convex) set of probability models to describe uncertainty, with the idea that this set should converge towards a single estimate as more reliable data become available. Such imprecise probabilistic models can be used to produce cautious inferences in the form of imprecise predictions, making precise predictions only when the available information is sufficient to do so [12]. Imprecise probabilistic ideas have recently been applied to ordinal regression problems [13], yet those applications considered the usual 0/1 loss and its associated predictions which, as we shall see, may produce counter-intuitive results in an ordinal setting.

When applying imprecise probabilistic approaches to ordinal problems, a natural extension of the median is to look at lower and upper medians as potential predictors, yet we may wonder if they are still related to specific loss functions, and what is the nature of this relation. Exploring these questions is the main goal of the present paper, in which we show in Section 3 that lower and upper median intervals are natural predictions when considering the L_1 loss (Section 3.1). These intervals can also be retrieved by considering symmetric and strictly increasing losses, which are special cases of V-shaped costs [14,15], provided we use the notion of sign-preference (Section 3.2). We perform in Section 4 some experiments demonstrating the potential interest of imprecise median predictions. Finally, we show in Section 5 that some of our results can be extended to linear losses, by considering lower and upper quantiles.

2 Setting

In an ordinal problem such as ordinal classification, we are interested in making predictions and inferences on a space $\mathscr{Y} = \{y_1, \dots, y_m\}$ of ordered possible elements, that is $y_i \prec y_{i+1}$ for $i = 1, \dots, m-1$. This is different from usual classification problems, where the space $\mathscr{Y} = \{y_1, \dots, y_m\}$ is assumed to be unordered (i.e., without structure), and also from usual regression problems, as no metric is assumed between the elements of $\mathscr{Y}$, in the sense that y_5 is not "five times better" than y_1.

2.1 Ordinal problem under probabilistic uncertainty

When the uncertainty over $\mathscr{Y}$ is defined by a probability mass $p : \mathscr{Y} \to [0,1]$ with $p_j := p(y_j)$, classical predictions associated to p include the modal value

$$Mo_p = \arg\max_{y_i \in \mathscr{Y}} p_i, \tag{1}$$

the expected value

$$\mathbb{E}_p(f) = \sum_{y_i \in \mathscr{Y}} p_i f(y_i) \tag{2}$$

where $f : \mathscr{Y} \to \mathbb{R}$ is a real-valued function on $\mathscr{Y}$, typically $f(y_i) = i$, and finally the median value, defined as

$$Me_p = \{y_i \in \mathscr{Y} : P(\{y \geq y_i\}) \geq 0.5 \wedge P(\{y \leq y_i\}) \geq 0.5\}. \tag{3}$$

As said in the introduction, modal and expected values do not appear as the most natural choices in an ordinal setting.

The modal value ignores the ordinal nature of the problem, and has a particularly counter-intuitive behaviour in an ordinal setting: it is not monotonic w.r.t. stochastic dominance [14]. Recall that a probability p^2 stochastically dominates p^1 if $P^2(\{y \geq y_k\}) \geq P^1(\{y \geq y_k\})$ for all $k \in [1;m]$, where P^i denotes the probability measure induced by p^i. In such a case, it seems natural to require that $\hat{y}^2 \succ \hat{y}^1$, with $\hat{y}^i$ the prediction taken w.r..t. p^i.

Example 1 Consider the space $\mathscr{Y} = \{y_1, \dots, y_3\}$ and the probability masses (denoted in vectorial form)

$$p^1 = (0.34, 0.36, 0.3),$$
$$p^2 = (0.34, 0.33, 0.33).$$

p^2 stochastically dominates p^1, yet the modal value of p^1 and p^2 are respectively $\hat{y}^1 = y_2$ and $\hat{y}^2 = y_1$, hence $\hat{y}^2 \prec \hat{y}^1$, in contrast with what could be naturally expected in an ordinal context.

This does not happen with the expected value, since it is known [16] that if p^2 stochastically dominates p^1, then for any function $f(y_1), \dots, f(y_m)$ increasing with the index i of y_i, we will have $\mathbb{E}_{p^1}(f) \leq \mathbb{E}_{p^2}(f)$. However, the expected value depends on a function f that, beyond being increasing with the rank i of labels y_i, can be defined arbitrarily and does not take account of the non-numerical nature of $\mathscr{Y}$. It can also result in a value that is different from all the values $f(y_1), \dots, f(y_m)$, therefore not corresponding to an element of $\mathscr{Y}$. It is also more sensible to the definition of $\mathscr{Y}$ to a certain extent, for instance if one splits a label into two sublabels.

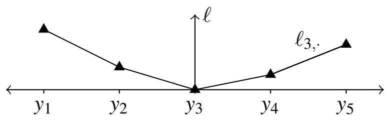

Fig. 1 A V-shaped loss function $\ell_{3,\cdot}$.

Example 2 Consider again the space $\mathscr{Y} = \{y_1, \dots, y_3\}$ and the probability mass

$$p = (0.2, 0.2, 0.6),$$

with $f(y_i) = i$, we obtain the expected value

$$\mathbb{E}_p(f) = 0.2 \cdot 1 + 0.2 \cdot 2 + 0.6 \cdot 3 = 2.4$$

that we would associate to y_2. Assume now that y_3 is split in two labels y_3, y_4 and that the new estimated distribution is

$$p = (0.2, 0.2, 0.3, 0.3).$$

this time, the expected value is

$$\mathbb{E}_p(f) = 0.2 \cdot 1 + 0.2 \cdot 2 + 0.3 \cdot 3 + 0.3 \cdot 4 = 2.7$$

which would no be associated to y_2 but to y_3. Yet, in both cases the median would be y_3.

Another usual means to derive a prediction $\hat{y}$ from a probability distribution consists of minimizing the expected value of some loss function $\ell : [1;m]^2 \to \mathbb{R}$ where $\ell_{i,j} := \ell(i,j)$ is the loss incurred by choosing y_i as a prediction when y_j is the true value. That is, to find

$$\hat{y} = \arg \min_{y_i \in \mathscr{Y}} \mathbb{E}(\ell_{i,\cdot}) \tag{4}$$

with $\mathbb{E}(\ell_{i,\cdot}) = \sum_{j \in 1,\dots,m} p_j \ell_{i,j}$. This prediction can then be considered as optimal w.r.t. to the loss ℓ. This is equivalent to compare every possible pair y_i, y_k, stating that y_i is preferred to y_k, denoted $y_i >_p y_k$, if

$$\mathbb{E}(\ell_{k,\cdot} - \ell_{i,\cdot}) > 0, \tag{5}$$

and then take the maximal element of the complete order $>_p$.

In ordinal problems, it is natural to ask the losses to follow a V-shaped form [15], that is $\ell_{i,j+1} \geq \ell_{i,j}$ if $j \geq i$, and $\ell_{i,j-1} \geq \ell_{i,j}$ if $j \leq i$. Such a generic V-shaped loss function is shown in Figure 1 for $i = 3$ and $m = 5$. This means that the loss $\ell_{i,\cdot}$ should not decrease as we consider elements further away from y_i. It is also natural to assume that $\ell_{i,i} = 0$ and $\ell_{i,j} > 0$ for $i \neq j$. We will call *strict* a V-shaped loss function where $\ell_{i,j+1} > \ell_{i,j}$ if $j \geq i$, and $\ell_{i,j-1} > \ell_{i,j}$ if $j \leq i$.

Both the mode, the expected value obtained for $f(y_i) = i$ and the median can be retrieved as predictions minimizing some particular losses, all being V-shaped. For instance, the mode is obtained by minimizing the 0/1 loss, that is $\ell_{i,j} = 1$ for all $i \neq j$, which is a very specific form of V-shaped loss where $\ell_{i,\cdot}$ is constant (hence neither increasing nor decreasing) on all elements but y_i.

The expected value is obtained by minimizing the L_2 loss defined as

$$\ell_{i,j} = (i - j)^2 \tag{6}$$

while the median Me_p of the distribution p corresponds to minimizing the L_1 loss

$$\ell_{i,j} = |i-j|, \tag{7}$$

which are both strict V-shaped loss functions, and are moreover symmetric. As recalled in the introduction, this connection between the Median and the L_1 loss provides an interesting justification for taking the Median as a predictive value, and allows to minimize this loss to directly get the median rather than first estimating p. In the next section, we will investigate to which extent the link between the median and the L_1 loss still holds when considering imprecisely defined probabilities. We will also propose a new connection between the median and more qualitative losses.

2.2 Imprecise probabilities and ordinal problems

In some cases, identifying the probability p accurately may be problematic, due to lack of information. This lack may be due to scarce, noisy or imprecise data, or to the use of expert opinions only providing partial information about p. In such situations, many authors [10, 17, 9] have argued that it is sensible to consider as estimate a convex set $\mathscr{P}$ of probabilities rather than a precise probability p. This is the approach we consider here. Recall that given a function $f : \mathscr{Y} \to \mathbb{R}$, the lower expectation $\underline{\mathbb{E}}(f)$ of f w.r.t. $\mathscr{P}$ is

$$\underline{\mathbb{E}}(f) = \inf_{p \in \mathscr{P}} \mathbb{E}(f) \tag{8}$$

and the upper expectation $\overline{\mathbb{E}}(f)$ is obtained by considering sup instead of inf in Equation (8). As $\mathscr{P}$ is most often described by linear constraints over $\mathscr{Y}$, solving Equation (8) can usually be done by using linear programming techniques. The lower expectation is translation invariant, i.e., $\underline{\mathbb{E}}(c+df) = c + d\underline{\mathbb{E}}(f)$ with c and d constants, and dual to the upper, i.e., $\underline{\mathbb{E}}(f) = -\overline{\mathbb{E}}(-f)$. The lower (upper) probability $\underline{P}(A)$ ($\overline{P}(A)$) of an event A corresponds to the lower (upper) expectation of the indicator function $\mathbf{1}_{(A)} : \mathscr{Y} \to \{0,1\}$.

Extending the decision rule and Equation (5) to the case where uncertainty is described by $\mathscr{P}$ can be done in several ways [12]. Here, we retain the notion of maximality, which states that y_i is preferred to y_k, denoted $y_i >_{\mathscr{P}} y_k$, if

$$\inf_{p \in \mathscr{P}} \mathbb{E}(\ell_{k,\cdot} - \ell_{i,\cdot}) = \underline{\mathbb{E}}(\ell_{k,\cdot} - \ell_{i,\cdot}) > 0, \tag{9}$$

that is if, considering all probabilities in $\mathscr{P}$, the lower bound of the expectation of $\ell_{k,\cdot} - \ell_{i,\cdot}$ is positive. The possibly imprecise decision $\hat{Y}$ then consists in taking the maximal elements of the partial order $>_{\mathscr{P}}$, that is

$$\hat{Y} = \{y_i \in \mathscr{Y} : \forall y_j, j \neq i, y_j \not>_{\mathscr{P}} y_i\}. \tag{10}$$

When $\mathscr{P} = \{p\}$, we retrieve the usual loss minimizer as a prediction. One question is then to know which kind of losses are adapted to the ordinal setting when predictions are obtained through Equations (9) (10). In addition to the problem illustrated in Example 1, 0/1 loss may provide, in the imprecise setting, predictions that contains "gaps", in the sense that $\hat{Y}$ may not be an *interval* $[y_i, y_j] = \{y_k : i \leq k \leq j\}$ with $i \leq j$, as illustrates the next example.

Example 3 Consider the space $\mathscr{Y} = \{y_1, \dots, y_3\}$ and the set $\mathscr{P}$ described by the following constraints

$$0.35 \leq p(y_1) \leq 0.45,$$

$$p(y_2) = 0.2,$$

$$0.35 \leq p(y_3) \leq 0.45.$$

In this case under the 0/1 loss ℓ we have that $\hat{Y} = \{y_1, y_3\}$, since for all probabilities within $\mathscr{P}$, we have $p(y_1)$ and $p(y_3)$ higher than $p(y_2)$, hence

$$\underline{\mathbb{E}}(\ell_{2,\cdot} - \ell_{1,\cdot}) = \inf_{p \in \mathscr{P}} p(y_1) - p(y_2) = 0.15$$

is higher than 0, meaning that y_1 is preferred to y_2 (by symmetry, the same happens for y_3).

Clearly, providing an interval $[y_i, y_j]$ as a prediction rather than an arbitrary set in an ordinal setting may be a desirable feature. For instance, it is enforced in the framework proposed by Del Coz *et al.* [18]. Also, since the median as a prediction has very good properties for ordinal problems when the uncertainty is described by p, it seems natural to consider its extension when uncertainty is described as a set $\mathscr{P}$. Given a set $\mathscr{P}$, the lower and upper medians [19] are defined as

$$\underline{Me}_{\mathscr{P}} = \inf_{p \in \mathscr{P}} \inf\{y_i \in \mathscr{Y} : P(\{y \geq y_i\}) \geq 0.5 \wedge P(\{y \leq y_i\}) \geq 0.5\}, \tag{11}$$

$$\overline{Me}_{\mathscr{P}} = \sup_{p \in \mathscr{P}} \sup\{y_i \in \mathscr{Y} : P(\{y \geq y_i\}) \geq 0.5 \wedge P(\{y \leq y_i\}) \geq 0.5\}, \tag{12}$$

or using the notation of Equation (3),

$$\underline{Me}_{\mathscr{P}} = \inf_{p \in \mathscr{P}} \inf Me_p \quad \text{and} \quad \overline{Me}_{\mathscr{P}} = \sup_{p \in \mathscr{P}} \sup Me_p.$$

In Example 3, the prediction using the lower and upper median would have been y_2, a quite different answer from the one obtained by using the 0/1 loss.

3 Retrieving median bounds with probability sets

The next sections explore how the prediction $[\underline{Me}_{\mathscr{P}}, \overline{Me}_{\mathscr{P}}]$ can be justified from a loss minimisation perspective. We will show that this can be done in at least two ways, one considering the L_1 loss within Equation (9), the other considering the notion of sign-preference [19] for strict V-shaped loss functions.

The connection with the L_1 loss shows that the results true in the precise setting remain true in the imprecise one (a feature that is not true for many concepts such as independence [20], conditioning [21], etc.), while the second connection links the median with a more qualitative view allowing some (numerical) imprecision in the loss definition.

To simplify notations, we will denote $\underline{Me}$ and $\overline{Me}$ the indices of respectively $\underline{Me}_{\mathscr{P}}$ and $\overline{Me}_{\mathscr{P}}$ in the rest of the paper, hence $y_{\underline{Me}} := \underline{Me}_{\mathscr{P}}$ and $y_{\overline{Me}} := \overline{Me}_{\mathscr{P}}$.

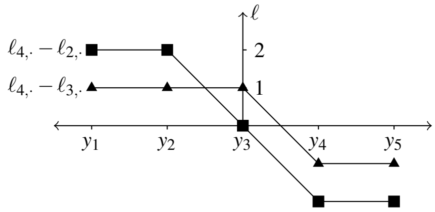

Fig. 2 Difference of L_1 loss function for $k=4$ and different values of $i<k$.

3.1 Imprecise median and the L_1 loss

Before showing that $[y_{\underline{Me}}, y_{\overline{Me}}]$ can be obtained by applying Equations (9) and (10) with the L_1 loss (7), let us first look at the form of this L_1 loss. Consider two labels y_i, y_k with $i<k$, then the difference

$$\ell_{k,j}-\ell_{i,j}=\begin{cases} k-i & \text{if } j<i \\ k+i-2j & \text{if } i\leq j\leq k \\ i-k & \text{if } k<j \end{cases} \tag{13}$$

is a constant for j outside the interval $[y_i, y_k]$, positive for $j<i$ and negative for $j>k$, and is decreasing between i and k. This is pictured for $m=5$ in Figure 2. Similarly, we have that

$$\ell_{i,j}-\ell_{k,j}=\begin{cases} i-k & \text{if } j<i \\ 2j-k-i & \text{if } i\leq j\leq k \\ k-i & \text{if } k<j \end{cases} \tag{14}$$

is a constant for j outside the interval $[y_i, y_k]$, negative for $j<i$ and positive for $j>k$, and is increasing between i and k. We can now demonstrate the following result

Proposition 1 *Under L_1 loss and given an uncertainty model $\mathscr{P}$, the prediction $\hat{Y}$ obtained by Equation* (10) *is*

$$\hat{Y}=[y_{\underline{Me}}, y_{\overline{Me}}] \tag{15}$$

Proof The proof will be done in two steps. First, we will show that any element outside $[y_{\underline{Me}}, y_{\overline{Me}}]$ is dominated (in the sense of Equation (9)) by an element within $[y_{\underline{Me}}, y_{\overline{Me}}]$, hence is not maximal. Second, we will prove that any two elements within $[y_{\underline{Me}}, y_{\overline{Me}}]$ are incomparable according to Equation (9)). Note that if $[y_{\underline{Me}}, y_{\overline{Me}}]$ reduces to one element, then the second step is not needed.

First part: consider an element y_i with $i<\underline{Me}$ and the element $y_{\underline{Me}}$. $y_i >_{\mathscr{P}} y_{\underline{Me}}$ if

$$\underline{\mathbb{E}}(\ell_{i,j}-\ell_{\underline{Me},j})$$

is positive. Now consider the following function $g:[1;m]\to\mathbb{R}$ such that

$$g(j)=\begin{cases} i-\underline{Me} & \text{if } j<\underline{Me} \\ \underline{Me}-i & \text{if } \underline{Me}\leq j. \end{cases}$$

g is such that $g < \ell_{i,\cdot} - \ell_{\underline{Me},\cdot}$, hence $\underline{\mathbb{E}}(g) < \underline{\mathbb{E}}(\ell_{i,\cdot} - \ell_{\underline{Me},\cdot})$. Now, if $\mathbf{1}_{(A)}$ stands for the indicator function of event A, we have that

$$\begin{aligned}
\underline{\mathbb{E}}(g) &= \underline{\mathbb{E}}(\mathbf{1}_{(j<\underline{Me})}\ (i-\underline{Me}) + \mathbf{1}_{(\underline{Me}\leq j)}\ (\underline{Me}-i)) \\
&\geq \underline{\mathbb{E}}(\mathbf{1}_{(j<\underline{Me})}\ (i-\underline{Me})) + \underline{\mathbb{E}}(\mathbf{1}_{(\underline{Me}\leq j)}\ (\underline{Me}-i)) \\
&= (i-\underline{Me})\overline{\mathbb{E}}(\mathbf{1}_{(j<\underline{Me})}\ ) + (\underline{Me}-i)\underline{\mathbb{E}}(\mathbf{1}_{(\underline{Me}\leq j)}\ ) \\
&= (\underline{Me}-i)(\underline{\mathbb{E}}(\mathbf{1}_{(\underline{Me}\leq j)}\ ) - \overline{\mathbb{E}}(\mathbf{1}_{(j<\underline{Me})}\ )) \\
&> 0.
\end{aligned}$$

The first inequality follows from $\underline{\mathbb{E}}(f+g) \geq \underline{\mathbb{E}}(f) + \underline{\mathbb{E}}(g)$, while the last one follows from the fact that by definition all probability measures within $\mathscr{P}$ are such that $P(y \geq y_{\underline{Me}}) \geq 0.5$ and $P(y < y_{\underline{Me}}) < 0.5$, hence $(\underline{\mathbb{E}}(\mathbf{1}_{(\underline{Me}\leq j)}\ ) - \overline{\mathbb{E}}(\mathbf{1}_{(j<\underline{Me})}\ ))$ is positive. The case y_i with $\overline{Me} < i$ can be proved using a similar reasoning. Hence the only possible optimal elements are those within the interval $[y_{\underline{Me}}, y_{\overline{Me}}]$.

Second part: consider two elements y_i, y_k with $\underline{Me} \leq i < k \leq \overline{Me}$ (remember that if $\underline{Me} = \overline{Me}$, the first part of the proof shows that this is the unique optimal element). Let us now consider the function f such that

$$f(j) = \begin{cases} i-k & \text{if } j < k \\ k-i & \text{if } k \leq j. \end{cases}$$

Clearly $f > \ell_{i,\cdot} - \ell_{k,\cdot}$, hence $\underline{\mathbb{E}}(f) > \underline{\mathbb{E}}(\ell_{i,\cdot} - \ell_{k,\cdot})$. We then have

$$\begin{aligned}
\underline{\mathbb{E}}(f) &= \underline{\mathbb{E}}(\mathbf{1}_{(j<k)}\ (i-k) + \mathbf{1}_{(k\leq j)}\ (k-i)) \\
&\leq \underline{\mathbb{E}}(\mathbf{1}_{(j<k)}\ (i-k)) + \overline{\mathbb{E}}(\mathbf{1}_{(k\leq j)}\ (k-i)) \\
&= (k-i)(\overline{\mathbb{E}}(\mathbf{1}_{(k\leq j)}\ ) - \underline{\mathbb{E}}(\mathbf{1}_{(j<k)}\ )) \\
&= (k-i)(1 - 2\underline{\mathbb{E}}(\mathbf{1}_{(j<k)}\ )) \\
&\leq 0
\end{aligned}$$

which shows that $y_k \not\succ_{\mathscr{P}} y_i$. The fourth equality follows from the fact that $\underline{\mathbb{E}}(A) = 1 - \overline{\mathbb{E}}(A^c)$ (where event A is equivalent to its indicator function), and since $\underline{Me} < k < \overline{Me}$, all probabilities within $\mathscr{P}$ are such that $P(y < y_k) > 0.5$. A similar reasoning can be followed to show $y_i \not\succ_{\mathscr{P}} y_k$.

Proposition 1 does show that the results holding for the L_1 loss with precise probabilities extend to convex sets of probabilities, in the sense that the obtained prediction is now the set of medians, that reduces to a unique element when the uncertainty model is a distribution p with a unique median. Also, using the L_1 loss avoids having "gaps" in the prediction, which is a reasonable requirement in an ordinal setting.

In practice, this also means that when we consider the L_1 loss function in ordinal problems, there is no need to compare every pair of possible elements when assessing 10, just to compute two boundary values. This may be especially interesting when m, the number of elements, is high.

Remark 1 The above results can be easily extended to the case where the space $\mathscr{Y}$ is the real line (following similar reasonings), in which case we can show that the prediction obtained with the L_1 loss is the interval bounded by the lower and upper Median. Similar results have been obtained for the L_2 loss and the interval bounded by lower and upper expectations [22].

3.2 Imprecise median and strict V-shaped symmetric losses

So far, we have assumed specific numeric form of the loss function, showing that using the L_1 loss leads to predict interval $[y_{\underline{Me}}, y_{\overline{Me}}]$ when using Equation (10). Yet using other V-shaped loss functions such as the L_2 loss will lead to other predictions [22].

Also, it may appear strange in some settings to consider that we can exactly specify a loss function while only providing imprecise estimates of probabilities. It would then be appealing to connect the median (or its interval-valued counter-part) with a looser or more qualitative definition of loss functions, that is without assuming a particular numerical form of this function.

With this idea in mind, we show that the notion of sign-preference, developed by Couso *et al.* [19] and that can be applied to any ordered values, can be used to justify producing $[y_{\underline{Me}}, y_{\overline{Me}}]$ as a prediction when

1. V-shaped loss function $\ell_{i,j}$ are strict and,
2. $\ell_{i,j} = c(|i-j|)$ is a function of $|i-j|$.

These two assumptions encompass all losses that are increasing and symmetric around label y_i, and therefore define a family $\mathscr{L}$ of loss functions that we will call strict V-shaped symmetric. Using only these two assumptions means that the difference $\ell_{k,j} - \ell_{i,j}$ is not numerically defined for two labels y_i, y_k with $i < k$, hence classical lower and upper expectations cannot be computed. However, we have that

$$\ell_{k,j} - \ell_{i,j} \text{ is } \begin{cases} > 0 & \text{if } |k-j| > |i-j| \\ = 0 & \text{if } |k-j| = |i-j| \\ < 0 & \text{if } |k-j| < |i-j| \end{cases} \tag{16}$$

This means that the notion of sign-preference can be applied to this kind of function. Recall that this notion states that y_i is sign-preferred to y_k, noted $y_i >_{SP} y_k$, if the following value

$$\underline{\mathbb{E}}(\mathbf{1}_{(\ell_{k,\cdot} - \ell_{i,\cdot} > 0)} - \mathbf{1}_{(\ell_{i,\cdot} - \ell_{k,\cdot} > 0)}) \tag{17}$$

is positive. Contrary to the expectation of $\ell_{k,\cdot} - \ell_{i,\cdot}$, Equation (17) can be evaluated under our assumptions (strict increase and symmetry), using Equation (16). A possibly imprecise prediction $\hat{Y}_{SP}$ can then be defined according to this new criterion, i.e.,

$$\hat{Y}_{SP} = \{y_i \in \mathscr{Y} : \forall y_j, j \neq i, y_j \not>_{SP} y_i\}, \tag{18}$$

Before demonstrating that $[y_{\underline{Me}}, y_{\overline{Me}}]$ is again the natural prediction in this framework, we first need an intermediate result.

Lemma 1 *Let $f : \mathscr{Y} \to \mathbb{R}$ and $g : \mathscr{Y} \to \mathbb{R}$ be two functions such that their sum $f + g = c$ is some constant $c \in \mathbb{R}$, then*

$$\underline{\mathbb{E}}(f-g) = \underline{\mathbb{E}}(f) - \overline{\mathbb{E}}(g)$$

Proof We have

$$\underline{\mathbb{E}}(f-g) = \underline{\mathbb{E}}(f+g-2g) = c - 2\overline{\mathbb{E}}(g)$$

by translation invariance and duality (with $\overline{\mathbb{E}}$) of $\underline{\mathbb{E}}$. Similarly, we have $\underline{\mathbb{E}}(f-g) = 2\underline{\mathbb{E}}(f) - c$. Finally, we have

$$2\underline{\mathbb{E}}(f-g) = 2\underline{\mathbb{E}}(f) - 2\overline{\mathbb{E}}(g)$$

by summing the two previous equalities.

Using Lemma 1, we can show the result

Proposition 2 *Under strict V-shaped symmetric losses $\mathscr{L}$ and given an uncertainty model $\mathscr{P}$, the prediction $\hat{Y}_{SP}$ obtained by Equation* (18) *is*

$$\hat{Y}_{SP} = [y_{\underline{Me}}, y_{\overline{Me}}] \tag{19}$$

Proof We will proceed in two steps similar to the ones of the proof of Proposition 1, and with a very similar reasoning.

First part: consider an element y_i with $i < \underline{Me}$. We have that $y_{\underline{Me}} >_{SP} y_i$, since

$$\begin{gathered}\underline{\mathbb{E}}(\mathbf{1}_{(\ell_{i,\cdot}-\ell_{\underline{Me},\cdot}>0)} - \mathbf{1}_{(\ell_{\underline{Me},\cdot}-\ell_{i,\cdot}>0)}) \geq \\ \underline{\mathbb{E}}(\mathbf{1}_{(\ell_{i,\cdot}-\ell_{\underline{Me},\cdot}>0)}) + \underline{\mathbb{E}}(-\mathbf{1}_{(\ell_{\underline{Me},\cdot}-\ell_{i,\cdot}>0)}) \geq \\ \underline{\mathbb{E}}(\mathbf{1}_{(\ell_{i,\cdot}-\ell_{\underline{Me},\cdot}>0)}) - \overline{\mathbb{E}}(-\mathbf{1}_{(\ell_{\underline{Me},\cdot}-\ell_{i,\cdot}>0)}) \geq \\ \underline{P}([y_{\underline{Me}}, y_m]) - \overline{P}([y_1, y_{\underline{Me}-1}]) > 0.\end{gathered}$$

The last inequality follows from the fact that $P([y_1, y_{\underline{Me}-1}]) < 0.5$ and $P([y_{\underline{Me}}, y_m]) \geq 0.5$ for any $p \in \mathscr{P}$. The third inequality is due to the fact that $\mathbf{1}_{(\ell_{\underline{Me},\cdot}-\ell_{i,\cdot}>0)} \subseteq [y_1, y_{\underline{Me}-1}]$ and that $[y_{\underline{Me}}, y_m] \subseteq \mathbf{1}_{(\ell_{i,\cdot}-\ell_{\underline{Me},\cdot}>0)}$. The case y_i with $i > \overline{Me}$ can be treated similarly.

Second part: consider two elements y_i, y_k with $\underline{Me} \leq i < k \leq \overline{Me}$ (again if $\underline{Me} = \overline{Me}$, the first part of the proof shows that this is the unique optimal element). We have that $y_k \not>_{SP} y_i$, since

$$\begin{gathered}\underline{\mathbb{E}}(\mathbf{1}_{(\ell_{i,\cdot}-\ell_{k,\cdot}>0)} - \mathbf{1}_{(\ell_{k,\cdot}-\ell_{i,\cdot}>0)}) \leq \\ \underline{\mathbb{E}}(\mathbf{1}_{(\ell_{i,\cdot}-\ell_{k,\cdot}>0)} - \mathbf{1}_{(\ell_{k,\cdot}-\ell_{i,\cdot}\geq 0)}) = \\ \underline{\mathbb{E}}(\mathbf{1}_{(\ell_{i,\cdot}-\ell_{k,\cdot}>0)}) - \overline{\mathbb{E}}(\mathbf{1}_{(\ell_{k,\cdot}-\ell_{i,\cdot}\geq 0)}) \leq 0\end{gathered}$$

where the second inequality follows from Lemma 1 and the last inequality follows by the fact that $[y_{\underline{Me}}, y_m] \subseteq \mathbf{1}_{(\ell_{i,\cdot}-\ell_{k,\cdot}>0)} \subseteq [y_{\overline{Me}}, y_m]$ (hence $\underline{P}(\{y_j : \ell_{j,\cdot} - \ell_{j,\cdot} > 0\}) \leq 0.5$) and that $[y_1, y_{\underline{Me}}] \subseteq \mathbf{1}_{(\ell_{k,\cdot}-\ell_{i,\cdot}\geq 0)} \subseteq [y_1, y_{\overline{Me}}]$ (hence $\underline{P}(\{y_j : \ell_{k,j} - \ell_{i,j} \geq 0\}) \geq 0.5$). This shows that $y_k \not>_{\mathscr{P}} y_i$.

This interesting property indicates that using the notion of sign-preference allows us to retrieve the median bounded interval under very mild assumptions. Actually, we can either see $\mathscr{L}$ as a qualitative definition of the loss (there is no need to specify numerical values, just the shape of the losses) or as considering a whole family of loss functions at once. In practice, this may be useful if an expert or a decision maker does not want to define a precise numerical loss function, but is ready to accept that the loss is symmetric and increasing as we get further away from the right label.

4 Experiments on ordinal classification

The goal of ordinal classification is to associate an instance $\mathbf{x}$ coming from an instance space $\mathscr{X}$ to a single label of the space $\mathscr{Y} = \{y_1, \ldots, y_m\}$ of possible classes. Ordinal classification differs from multi-class classification in that labels y_i are ordered, that is $y_i \prec y_{i+1}$ for $i = 1, \ldots, m-1$. An usual task is then to estimate the theoretical conditional probability measure

$P_{\mathbf{x}} : 2^{\mathscr{Y}} \to [0,1]$ associated to an instance $\mathbf{x}$ from a set of n training samples $(\mathbf{x}_i, \ell_{x_i}) \in \mathscr{X} \times \mathscr{Y}$, $i = 1, \ldots, n$, and to produce predictions from this conditional probability.

In this section, we will apply our previous results to this problem of ordinal classification, and will compare the case where the estimation is a precise model $P_{\mathbf{x}}$ to the case where it corresponds to a set $\mathscr{P}_{\mathbf{x}}$ of models.

4.1 Evaluation

Comparing classifiers that return cautious (partial) predictions in the form of multiple classes is a hard problem. Indeed, compared to the usual setting, measures of performance have to include the informativeness of the predictions in addition to the accuracy. Zaffalon et al. [23] discuss in details the case of comparing a cautious prediction with a precise one under a 0/1 loss assumption, using a betting interpretation.

However, under the mild assumption of Section 3.2, we do not even have access to numerical loss functions and it is not obvious how to properly define the numerical evaluation in this case. What we propose is a qualitative comparison of classifiers by declaring whether a prediction is better than another. Given two (cautious) classifications $\hat{Y}_1, \hat{Y}_2$ coming from two different classifiers C_1, C_2 and a ground truth y_k, we say that

- If $\min\{|j-k| : y_j \in \hat{Y}_1\} < \min\{|j-k| : y_j \in \hat{Y}_2\}$, C_1 wins. That is, among the classes predicted by C_1 is a better one (in the sense of V-shaped losses) than among the classes predicted by C_2 (and inversely for C_2 to win).
- If $\min\{|j-k| : y_j \in \hat{Y}_1\} = \min\{|j-k| : y_j \in \hat{Y}_2\}$, then C_1 wins if $|\hat{Y}_1| \leq |\hat{Y}_2|$, that is if C_1 is more informative than C_2 and their predictions are equally good.
- Else, the result is a tie.

Example 4 Assume that the observed ground truth for an instance $\mathbf{x}$ is y_2, with $\mathscr{Y} = \{y_1, \ldots, y_6\}$, and that classifier C_1 predicts $\hat{Y}_1 = \{y_3, y_4\}$. Then,

- if C_2 predicts $\hat{Y}_2 = \{y_2, y_3, y_4\}$, C_2 wins because
$$\min\{|j-2| : y_j \in \hat{Y}_1\} = 1 > 0 = \min\{|j-2| : y_j \in \hat{Y}_2\},$$
- if C_2 predicts $\hat{Y}_2 = \{y_3, y_4, y_5\}$, C_1 wins because
$$\min\{|j-2| : y_j \in \hat{Y}_1\} = 1 = \min\{|j-2| : y_j \in \hat{Y}_2\}$$
but $\hat{Y}_1$ is more precise.

Once the number of wins, ties, losses have been estimated over the test data set, we can apply a classical sign-test [24] to check whether the difference between win and loss is significant.

4.2 Method

The method we use is the extension of Frank and Hall [4] method to imprecise probabilities presented in details in [13]. Frank and Hall propose to estimate the $m-1$ probabilities $P_{\mathbf{x}}(A_k) := F(y_k)$ where $A_k = \{y_1, \ldots, y_k\}$, and the mapping $F : \mathscr{Y} \to [0,1]$ can be seen as discrete cumulative distribution. The probabilities $P_{\mathbf{x}}(\ell_{\mathbf{x}} = y_k)$ are then deduced through the formula $P_{\mathbf{x}}(y_k) = \max\{0, F_{\mathbf{x}}(y_k) - F_{\mathbf{x}}(y_{k-1})\}$.

The same idea can be applied to sets of probabilities, in which case we estimate the bounds

$$\underline{P}_{\mathbf{x}}(A_k) := \underline{F}_{\mathbf{x}}(y_k) \text{ and } \overline{P}_{\mathbf{x}}(A_k) := \overline{F}_{\mathbf{x}}(y_k),$$

where $\underline{F}_{\mathbf{x}}, \overline{F}_{\mathbf{x}} : \mathscr{Y} \to [0,1]$ can be seen as lower and upper cumulative distributions defining a well-studied [25] probability set $\mathscr{P}_{\mathbf{x}}([\underline{F}, \overline{F}])$. Similarly to Frank and Hall, if estimated $\underline{F}, \overline{F}$ are non-increasing or do not satisfy inequality $\underline{F} \leq \overline{F}$, they can be corrected through Algorithm 1. Any base classifier returning probability bounds can be used to estimate $\underline{P}_{\mathbf{x}}(A_k), \overline{P}_{\mathbf{x}}(A_k)$.

Algorithm 1: Correction of estimates $\underline{F}, \overline{F}$ into proper estimates

Input: estimates $\underline{F}, \overline{F}$ obtained from data
Output: corrected estimates $\underline{F}, \overline{F}$

```
for k=1,...,m-1 do
    if F̄(y_k) > F̄(y_{k+1}) then F̄(y_{k+1}) ← F̄(y_k);
    if F̲(y_{m-k+1}) < F̲(y_{m-k}) then F̲(y_{m-k}) ← F̲(y_{m-k+1});
```

The lower expectation of any function f over $\mathscr{Y}$ can then be computed through the Choquet Integral: if we denote by $()$ a reordering of elements of $\mathscr{Y}$ such that $f(y_{(1)}) \leq \ldots \leq f(y_{(N)})$, this integral reads

$$\underline{\mathbb{E}}(f) = \sum_{i=1}^{N} (f(y_{(i)}) - f(y_{(i-1)}) \underline{P}(A_{(i)}) \tag{20}$$

with $f(x_{(0)}) = 0, A_{(i)} = \{x_{(i)}, \ldots, x_{(N)}\}$ and $\underline{P}(A_{(i)}) = \inf_{P \in \mathscr{P}_{\mathbf{x}}([\underline{F}, \overline{F}])} P(A_{(i)})$ is the lower probability of $A_{(i)}$. We refer to [25] or [13] for details about how these lower probabilities can be computed efficiently.

Example 5 Consider the case given by Table 1, where the considered function is the L_1 loss around y_2. The elements used in the computation of the Choquet integral (20) for this case are summarized in Table 1.

i	$y_{(i)}$	$f_{(i)}$	$A_{(i)}$	$\underline{P}_{\mathbf{x}}(A_{(i)})$
1	y_2	0	$\mathscr{Y}$	1
2	y_1	1	$\{y_1, y_3, y_4, y_5\}$	0.6
3	y_3	1	$\{y_3, y_4, y_5\}$	0.5
4	y_4	2	$\{y_4, y_5\}$	0.45
5	y_5	3	$\{y_5\}$	0.25

Table 1 Choquet integral components of Example 5

The lower probability of $\underline{\mathbb{E}}(f)$ of f is then

$$\underline{\mathbb{E}}(f) = (0-0) \cdot 1 + (1-0) \cdot 0.6 + (1-1) \cdot 0.5 + (2-1) \cdot 0.45 + (3-2)0.25 = 1.3$$

Name	#instances	#features	Name	#instances	#features
autoPrice	159	16	house 8L	22784	9
bank8FM	8192	9	house 16H	22784	17
bank32NH	8192	33	kinematics	8192	9
boston housing	506	14	puma8NH	8192	9
california housing	20640	9	puma32H	8192	33
cpu small	8192	13	stock	950	10
delta ailerons	7129	6	delta elevators	9517	7
friedman	40768	11			

Table 2 Data set details

As a base classifier to evaluate $\underline{P}_{\mathbf{x}}(A_k), \overline{P}_{\mathbf{x}}(A_k)$, we use the Naive Credal Classifier (NCC) [26] that extends the Naive Bayesian Classifier (NBC) by allowing conditional probabilities to become imprecise. This classifier relies on a positive real-valued hyper-parameter s, the imprecision of $\underline{P}_{\mathbf{x}}(A_k), \overline{P}_{\mathbf{x}}(A_k)$ increasing as the value s increases. For $s = 0$, we retrieve the classical NBC, which makes comparison between a precise method and its imprecise counter-part easy.

4.3 Results

In this section, our method is tested on 16 datasets of the UCI machine learning repository [27], whose details are given in Table 2. As there is a general lack of benchmark data sets for ordinal classification data, we used regression problems that we turned into ordinal classification by discretizing the output variable. The results reported in this section are obtained with a discretization into 7 classes of equal frequencies. We also performed experiments with 5 and 9 discretized classes, obtaining the same conclusions.

All experiments compare the results of the NBC with the results of the NCC used with $s = 2$, counting for each data set the number of win/loss/tie according to Section 4.1. We study the performances of each method when using a limited amount of data, respectively 100, 200 and 300 in three different experiments, as well as the performances when considering all data. The results of these experiments are reported in Table 3. For each experiment we performed a 10-fold cross validation. For experiments involving only a part of the data sets (either 100, 200 or 300 data), results are averages (rounded over the closest integer) over 10 repetitions in which data were randomly selected .

Table 3 clearly shows that while the imprecise approach is quite competitive and often wins when there are few data (10 victory and 1 loss when considering 100 samples, 8 victory and 1 loss with 200 samples), the results are much more balanced when the number of data increases (even with only 300 samples). When considering the whole data sets, it can be seen that most predictions are precise (given by the number of ties). It should also be noted that even in case of losses, the imprecise predictions still contain the precise predictions, but the added imprecision is not very useful (it may still warn the user that a second look at the prediction could be worthwhile, but do not improve much over the precise method). A tentative conclusion we may extract from these experiments is that using cautious predictions in an ordinal setting is mainly useful when the number of samples is quite low, i.e., when available information is very limited.

data sets	number of samples 100	200	300	all data
	W/L/T	W/L/T	W/L/T	W/L/T
autoPrice	20/43/37*			25/55/79*
bank8FM	**85/15/0****	**114/69/17***	128/114/58	93/104/7995
bank32NH	**81/19/0****	**171/29/0****	**264/34/2****	813/1063/6316*
Boston housing	42/29/29	61/68/71	54/76/170	54/90/362
California housing	**78/21/1****	**113/73/14****	134/117/49	128/195/20317
cpu small	**50/28/22***	61/60/79	54/107/139**	70/128/7994
delta ailerons	54/43/3	47/74/79	61/79/160	62/68/6999
friedman	**89/11/0****	**138/53/9****	**176/100/24****	209/320/40239
house 8L	**76/22/2****	59/79/62	60/103/137*	86/126/22572
house 16H	**64/24/12****	**100/39/61****	101/98/101	130/199/22455
kinematics	**87/13/0****	**138/55/7****	**178/104/18****	227/297/7668.
puma 8NH	**92/8/0****	**184/16/0****	**248/52/0****	233/358/7601
puma 32H	**87/13/0****	**179/21/0****	**245/55/0****	850/1032/6310*
stock	48/42/10	39/69/92*	42/93/165**	46/101/803
delta elevators	43/51/6	49/68/83	52/98/150*	77/108/9332

Table 3 Result of experiments. *: significant difference for significance level 0.05. **: significant difference for significance level 0.005. In bold are the case where the imprecise approach significantly win.

5 From the median to other quantiles: linear losses

So far, we have considered symmetric loss functions. Yet there may be cases in ordinal classification where such symmetry is not desirable. For instance, when judging the seriousness of a given disease, it may be more damaging to underestimate its severity rather than overestimate it (or the other way around). The notion of *linear loss* [9] (a.k.a. pinball loss) is well adapted to the situation. A linear loss L_α is defined as

$$\ell_{i,j} = \begin{cases} \alpha(i-j) & \text{if } i > j \\ (1-\alpha)(j-i) & \text{if } i \leq j \end{cases} \tag{21}$$

with $0 < \alpha < 1$. The L_1 loss function is retrieved for $\alpha = 1/2$ (up to a constant), and $\alpha > 1/2$ means that predicting higher classes than the true one is more penalized than predicting lower ones. It has been shown [9] that the prediction optimizing the expected loss (21) using a single probability p is the $(1-\alpha)$ quantile $Q_p^{1-\alpha}$ defined as

$$Q_p^{1-\alpha} = \{y_i \in \mathscr{Y} : P(\{y \geq y_i\}) \geq 1-\alpha \wedge P(\{y \leq y_i\}) \geq \alpha\}. \tag{22}$$

When our uncertainty is given by a set $\mathscr{P}$ of potential probabilities, then this notion extends naturally to the one of lower and upper $(1-\alpha)$ quantiles

$$\underline{Q}_{\mathscr{P}}^{1-\alpha} = \inf_{p \in \mathscr{P}} \inf Q_p^{1-\alpha} \tag{23}$$

$$\overline{Q}_{\mathscr{P}}^{1-\alpha} = \sup_{p \in \mathscr{P}} \sup Q_p^{1-\alpha}. \tag{24}$$

Again, to simplify notation, we will denote $\underline{Q}^{1-\alpha}$ and $\overline{Q}^{1-\alpha}$ the indices of $\underline{Q}_{\mathscr{P}}^{1-\alpha}$ and $\overline{Q}_{\mathscr{P}}^{1-\alpha}$. An immediate question is then to know whether Proposition 1 extends to the case of L_α loss functions? Before showing that, first observe that for two labels y_i, y_k with $i < k$, then the difference

$$\ell_{k,j} - \ell_{i,j} = \begin{cases} \alpha(k-i) & \text{if } j < i \\ i - j + \alpha(k-i) & \text{if } i \leq j \leq k \\ (1-\alpha)(i-k) & \text{if } k < j \end{cases} \tag{25}$$

is a constant for j outside the interval $[y_i, y_k]$, positive for $j < i$ and negative for $j > k$, and is decreasing between i and k. This shape is similar to (13) for the L_1 loss. Similarly, the difference

$$\ell_{i,j} - \ell_{k,j} = \begin{cases} \alpha(i-k) & \text{if } j < i \\ j - i + \alpha(i-k) & \text{if } i \leq j \leq k \\ (1-\alpha)(k-i) & \text{if } k < j \end{cases} \tag{26}$$

behaves as the Equation (14) obtained for the L_1 loss.

Proposition 3 *Under L_α loss and given an uncertainty model $\mathscr{P}$, the prediction $\hat{Y}$ obtained by Equation* (10) *is*

$$\hat{Y} = [y_{\underline{Q}^{1-\alpha}}, y_{\overline{Q}^{1-\alpha}}] \tag{27}$$

Proof The proof will follow the same steps as the proof of Proposition 1, we will first show that elements outside $[y_{\underline{Q}^{1-\alpha}}, y_{\overline{Q}^{1-\alpha}}]$ are dominated by an element within it, and will then prove that any two elements within $[y_{\underline{Q}^{1-\alpha}}, y_{\overline{Q}^{1-\alpha}}]$ do not dominate each others.

First part: consider y_i with $i < \underline{Q}^{1-\alpha}$ and the element $y_{\underline{Q}^{1-\alpha}}$. Then the function

$$g(j) = \begin{cases} \alpha(i - \underline{Q}^{1-\alpha}) & \text{if } j < \underline{Q}^{1-\alpha} \\ (1-\alpha)(\underline{Q}^{1-\alpha} - i) & \text{if } \underline{Q}^{1-\alpha} \leq j. \end{cases}$$

is such that $g \leq \ell_{i,\cdot} - \ell_{\underline{Q}^{1-\alpha},\cdot}$. We also have that

$$\begin{aligned} \underline{\mathbb{E}}(g) &= \underline{\mathbb{E}}(\mathbf{1}_{(j<\underline{Q}^{1-\alpha})}\ \alpha(i - \underline{Q}^{1-\alpha}) + \mathbf{1}_{(\underline{Q}^{1-\alpha} \leq j)}\ (1-\alpha)(\underline{Q}^{1-\alpha} - i)) \\ &\geq \alpha(i - \underline{Q}^{1-\alpha})\overline{\mathbb{E}}(\mathbf{1}_{(j<\underline{Q}^{1-\alpha})}) + (1-\alpha)(\underline{Q}^{1-\alpha} - i)\underline{\mathbb{E}}(\mathbf{1}_{(\underline{Q}^{1-\alpha} \leq j)}) \\ &= (\underline{Q}^{1-\alpha} - i)((1-\alpha)\underline{\mathbb{E}}(\mathbf{1}_{(\underline{Q}^{1-\alpha} \leq j)}) - \alpha\overline{\mathbb{E}}(\mathbf{1}_{(j<\underline{Q}^{1-\alpha})})) > 0 \end{aligned}$$

where the last inequality follows from the fact that $\underline{\mathbb{E}}(\mathbf{1}_{(\underline{Q}^{1-\alpha} \leq j)}) \geq \alpha$ and $\overline{\mathbb{E}}(\mathbf{1}_{(j<\underline{Q}^{1-\alpha})}) < 1-\alpha$, hence the term

$$((1-\alpha)\underline{\mathbb{E}}(\mathbf{1}_{(\underline{Q}^{1-\alpha} \leq j)}) - \alpha\overline{\mathbb{E}}(\mathbf{1}_{(j<\underline{Q}^{1-\alpha})}))$$

is strictly positive.

Second part: consider two elements y_i, y_k with $\underline{Q}^{1-\alpha} \leq i < k \leq \overline{Q}^{1-\alpha}$ (again, this part is unnecessary if $\underline{Q}^{1-\alpha} = \overline{Q}^{1-\alpha}$). Let us now consider the function f such that

$$f(j) = \begin{cases} \alpha(i-k) & \text{if } j < k \\ (1-\alpha)(k-i) & \text{if } k \leq j. \end{cases}$$

Clearly $f > \ell_{i,\cdot} - \ell_{k,\cdot}$, hence $\underline{\mathbb{E}}(f) > \underline{\mathbb{E}}(\ell_{i,\cdot} - \ell_{k,\cdot})$. We then have

$$\begin{aligned}
\underline{\mathbb{E}}(f) &= \underline{\mathbb{E}}(\mathbf{1}_{(j<k)}\ \alpha(i-k) + \mathbf{1}_{(k\leq j)}\ (1-\alpha)(k-i)) \\
&\leq \underline{\mathbb{E}}(\mathbf{1}_{(j<k)}\ \alpha(i-k)) + \overline{\mathbb{E}}(\mathbf{1}_{(k\leq j)}\ (1-\alpha)(k-i)) \\
&= (k-i)((1-\alpha)\overline{\mathbb{E}}(\mathbf{1}_{(k\leq j)}\ ) - \alpha\underline{\mathbb{E}}(\mathbf{1}_{(j<k)}\ )) \\
&= (k-i)((1-\alpha)(1-\underline{\mathbb{E}}(\mathbf{1}_{(j<k)}\ ) - \alpha\underline{\mathbb{E}}(\mathbf{1}_{(j<k)}\ )) \\
&= (k-i)((1-\alpha) - \underline{\mathbb{E}}(\mathbf{1}_{(j<k)}\ )) \\
&\leq 0
\end{aligned}$$

which shows that $y_k \not\succ_{\mathscr{P}} y_i$. The last inequality follows from te fact that since $k \in]\underline{Q}^{1-\alpha}, \overline{Q}^{1-\alpha}[$, all probabilities within $\mathscr{P}$ are such that $P(j < y_k) > 1-\alpha$. A similar reasoning can be followed to show $y_i \not\succ_{\mathscr{P}} y_k$.

6 Conclusions and perspectives

In this paper, we have studied the problem of making prediction in an ordinal setting when the uncertainty about the labels is described by a set of probabilities (rather than a single one), and when the associated prediction is set-valued. In such a setting, considering the set of predictions induced by the 0/1 loss appears somewhat unnatural, in particular because the resulting prediction set can contain gaps, as shows Example 3.

Another solution is to consider V-shaped loss, and in particular losses depending on the absolute value of rank differences. Considering such losses, we have shown that:

- when considering the L_1 loss, the predicted set is the set bounded by the lower and upper medians, thus generalizing results obtained in the precise case. Among other things, this means that rather than making pairwise comparisons to produce the final prediction when using the L_1 loss, one can just compute two values;
- the set bounded by the lower and upper medians could also be justified as a prediction when considering qualitative (i.e., not numerically defined) symmetric V-shaped losses, using the notion of sign-desirability;
- when considering linear losses L_α, the predicted set is the set bounded by $(1-\alpha)$ quantiles, again generalizing results obtained in the precise case.

The second result in particular seems quite interesting, as it indicates that sign-desirability is, in some situations, a suitable tool to study families of costs defined solely by inequalities.

We have applied our findings to the problem of ordinal regression or classification, and have proposed a way to compare precise and imprecise predictions in such settings (since numerical comparisons are not possible with symmetric V-shaped losses without specific numerical forms). Results indicate that providing cautious predictions in the form of Median bounds is mainly interesting when only few learning data are available.

This work focused on the median and its relation to loss functions within an imprecise setting, with an application to ordinal regression. However, this study suggests different interesting avenues of research:

- Investigating the necessary and sufficient conditions to impose to loss functions in order to retrieve the median interval, under different decision rules (e.g., classical maximality as in Section 3.1 or sign-preference as in Section 3.2). Results from Section 5 suggest

that symmetry and strict convexity are necessary conditions. Also, in the case of maximality, results from [22] linking expected value and L_2 loss, and Example 3 indicate that retrieving the median interval with other symmetric losses than the L_1 loss may be difficult;
- More generally, it would be interesting to know what are the conditions to impose on losses for the prediction to be a closed interval, as this means that one only needs to compute the bounds of such intervals. Again, this paper and [22] suggest that strict convexity of the loss function is a necessary condition;
- Study to which extent the presented results can be used in other applications involving ordinal variables, in particular the field of multi-criteria decision making, in which some methods are closely related to ordinal regression problems [8]. Another potential field of application is the one of robust statistics and quantile regression [28,29], which would nevertheless require to properly extend the results of this paper to a continuous setting.

Acknowledgements

This work was carried out in the framework of the Labex MS2T, which was funded by the French Government, through the program Investments for the future managed by the National Agency for Research (Reference ANR-11-IDEX-0004-02)

References

1. C. Zopounidis and M. Doumpos, "Multicriteria classification and sorting methods: A literature review," *European Journal of Operational Research*, vol. 138, no. 2, pp. 229–246, 2002.
2. M. Grabisch and C. Labreuche, "A decade of application of the choquet and sugeno integrals in multi-criteria decision aid," *Annals of Operations Research*, vol. 175, no. 1, pp. 247–286, 2010.
3. B. Ahn and S. Choi, "Aggregation of ordinal data using ordered weighted averaging operator weights," *Annals of Operations Research*, vol. 201, no. 1, pp. 1–16, 2012.
4. E. Frank and M. Hall, "A simple approach to ordinal classification," in *Proceedings of the 12th European Conference on Machine Learning*. Springer-Verlag, 2001, pp. 145–156.
5. S. Lievens, B. De Baets, and K. Cao-Van, "A probabilistic framework for the design of instance-based supervised ranking algorithms in an ordinal setting," *Annals of Operations Research*, vol. 163, no. 1, pp. 115–142, 2008.
6. W. Chu and S. S. Keerthi, "Support vector ordinal regression," *Neural computation*, vol. 19, no. 3, pp. 792–815, 2007.
7. K. Uematsu and Y. Lee, "Statistical optimality in multipartite ranking and ordinal regression," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 37, no. 5, pp. 1080–1094, May 2015.
8. S. Angilella, M. Bottero, S. Corrente, V. Ferretti, S. Greco, and I. M. Lami, "Non additive robust ordinal regression for urban and territorial planning: an application for siting an urban waste landfill," *Annals of Operations Research*, pp. 1–30, 2013.
9. J. O. Berger, *Statistical decision theory and Bayesian analysis*. Springer, 1985.
10. P. Walley, *Statistical reasoning with imprecise Probabilities*. New York: Chapman and Hall, 1991.
11. G. Corani, A. Antonucci, and M. Zaffalon, "Bayesian networks with imprecise probabilities: Theory and application to classification," *Data Mining: Foundations and Intelligent Paradigms*, pp. 49–93, 2012.
12. M. Troffaes, "Decision making under uncertainty using imprecise probabilities," *Int. J. of Approximate Reasoning*, vol. 45, pp. 17–29, 2007.
13. S. Destercke and G. Yang, "Cautious ordinal classification by binary decomposition," in *Machine Learning and Knowledge Discovery in Databases*. Springer, 2014, pp. 323–337.
14. W. Kotlowski and R. Slowinski, "On nonparametric ordinal classification with monotonicity constraints," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 11, pp. 2576–2589, 2013.
15. L. Li and H.-t. Lin, "Ordinal regression by extended binary classification," in *Advances in Neural Information Processing Systems*, 2006, pp. 865–872.
16. H. Levy, "Stochastic dominance and expected utility: survey and analysis," *Management Science*, vol. 38, no. 4, pp. 555–593, 1992.

17. I. Levi, *The Enterprise of Knowledge*. London: MIT Press, 1980.
18. J. José del Coz and A. Bahamonde, "Learning nondeterministic classifiers," *The Journal of Machine Learning Research*, vol. 10, pp. 2273–2293, 2009.
19. I. Couso and L. Sánchez, "The behavioral meaning of the median," in *Combining Soft Computing and Statistical Methods in Data Analysis*. Springer, 2010, pp. 115–122.
20. I. Couso, S. Moral, and P. Walley, "A survey of concepts of independence for imprecise probabilities," *Risk Decision and Policy*, vol. 5, no. 02, pp. 165–181, 2000.
21. D. Dubois and H. Prade, "Focusing vs. belief revision: A fundamental distinction when dealing with generic knowledge," in *Qualitative and quantitative practical reasoning*. Springer, 1997, pp. 96–107.
22. A. Benavoli and M. Zaffalon, "Density-ratio robustness in dynamic state estimation," *Mechanical Systems and Signal Processing*, vol. 37, no. 1-2, pp. 54–75, 2013.
23. M. Zaffalon, G. Corani, and D. Mauá, "Evaluating credal classifiers by utility-discounted predictive accuracy," *International Journal of Approximate Reasoning*, 2012.
24. J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research,*, vol. 9, no. 7, pp. 1–30, 2006.
25. S. Destercke, D. Dubois, and E. Chojnacki, "Unifying practical uncertainty representations - i: Generalized p-boxes," *Int. J. Approx. Reasoning*, vol. 49, no. 3, pp. 649–663, 2008.
26. M. Zaffalon, "The naive credal classifier," *J. Probabilistic Planning and Inference*, vol. 105, pp. 105–122, 2002.
27. A. Frank and A. Asuncion, "UCI machine learning repository," 2010. [Online]. Available: http://archive.ics.uci.edu/ml
28. M. E. Cattaneo and A. Wiencierz, "Likelihood-based imprecise regression," *International Journal of Approximate Reasoning*, vol. 53, no. 8, pp. 1137–1154, 2012.
29. R. Koenker, *Quantile regression*. Cambridge university press, 2005, no. 38.