Cautious label ranking with label-wise decomposition

Sébastien Destercke, Marie-Hélène Masson, Michael Poss

▶ **To cite this version:**

Sébastien Destercke, Marie-Hélène Masson, Michael Poss. Cautious label ranking with label-wise decomposition. European Journal of Operational Research, 2015, 246 (3), pp.927-935. 10.1016/j.ejor.2015.05.005 . hal-01166139

HAL Id: hal-01166139
https://hal.science/hal-01166139v1

Submitted on 22 Jun 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Cautious label ranking with label-wise decomposition

Sébastien Destercke[1] and Marie-Hélène Masson[1,2] and Michael Poss[1]

1. UMR CNRS 7253 Heudiasyc UTC, Compiègne, France.
2. Université de Picardie Jules Verne, Amiens, France.

Abstract

In this paper, we are interested in the label ranking problem. We are more specifically interested in the recent trend consisting in predicting partial but more accurate (i.e., making less incorrect statements) orders rather than complete ones. To do so, we propose a ranking method based on label-wise decomposition. We estimate an imprecise probabilistic model over each label rank and we infer a partial order from it using optimisation techniques. This leads to new results concerning a particular bilinear assignment problem. Finally, we provide some experiments showing the feasibility of our method.

Keywords. Label ranking, label-wise decomposition, assignment problem, bilinear optimisation

1. Introduction

In recent years, machine learning problems with structured outputs received an increasing interest. These problems appear in a variety of fields, including biology [26], natural language treatment [4], image analysis [17], ...

In this paper, we deal with the problem of *label ranking*, where one has to learn a mapping from instances to rankings (complete orders) defined over a finite number of labels. Various ways have been proposed to solve the problem, most of them intending to reduce the initial complexity of the problem. Some approaches propose to fit a probabilistic ranking model with few parameters (Mallows [7], Placket-Luce [23]) using different approaches [24, 5]. Other approaches tend to decompose the problem. Ranking by pairwise comparison (RPC) [21] transforms the problem of label ranking into binary problems, combining the results to obtain the final ranking. Constraint classification and log-linear models [18, 14] learn, for each label, a (linear) utility function from which the ranking is deduced. More recently, some authors [8] have proposed to solve the problem by performing a label-wise (rather than a more classical pairwise) decomposition.

Email address:
sebastien.destercke@hds.utc.fr,mmasson@hds.utc.fr,michael.poss@hds.utc.fr
(Sébastien Destercke[1] and Marie-Hélène Masson[1,2] and Michael Poss[1])

$\mathbb{D}$

X_1	X_2	X_3	X_4	Y
107.1	25	*Blue*	60	$\lambda_1 \succ \lambda_3 \succ \lambda_2$
-50	10	*Red*	40	$\lambda_2 \succ \lambda_3 \succ \lambda_1$
200.6	30	*Blue*	58	$\lambda_2 \succ \lambda_1 \succ \lambda_3$
107.1	5	*Green*	33	$\lambda_1 \succ \lambda_2 \succ \lambda_3$
...	...	...	...	...

Figure 1: A label ranking data set $\mathbb{D}$

Additionally, some authors [10] have discussed the interest, in preference learning problems and in label ranking in particular, to predict partial orders rather than complete rankings. Such an approach can be seen as an extension of the reject option [2] or of the fact of making partial predictions [11]. Such cautious predictions can prevent harmful decisions based on incorrect predictions. In practice, current methods [10] exploit pairwise information and consist in thresholding a pairwise comparison matrix containing probabilistic estimates. It has been recently shown that such methods, when coupled with parametric probabilistic models (Placket-Luce and Mallows), are particularly interesting, as they are guaranteed to produce semi-orders, thus avoiding the presence of cycles in predicted relations.

In this paper, we retain the recent ideas of label-wise decomposition, and explore how partial predictions can be obtained from them. More precisely, we propose to learn for each label an imprecise probabilistic model of its rank, and use these models to infer a partial prediction, using robust optimisation techniques. Note that imprecise probabilistic approaches are well tailored to make partial predictions [11], as well as to deal with incomplete data [27].

Section 2 introduces the problem and our notations. Section 3 shows how rank can be predicted from imprecise probabilistic models. Section 4 presents the proposed inference method based on bilinear optimisation techniques. Finally, Section 5 shows some experiments on synthetic data sets.

2. Problem setting

The usual goal of classification problems is to associate an instance $\mathbf{x}$ coming from an instance space $\mathcal{X}$ to a single (preferred) label of the space $\Lambda = \{\lambda_1, \ldots, \lambda_k\}$ of possible classes. Label ranking problems correspond to the case where an instance $\mathbf{x}$ is no longer associated to a single label of Λ but to a total order over the labels, that is a connected, transitive, and asymmetric relation $\succ_{\mathbf{x}}$ over $\Lambda \times \Lambda$, that amounts to give a rank to each label $\lambda_1, \ldots, \lambda_k$. Hence, the space of prediction is now the whole set $\mathcal{L}(\Lambda)$ of complete rankings of Λ that contains $|\mathcal{L}(\Lambda)| = k!$ elements (i.e., the set of all permutations). Figure 1 illustrates a label ranking data set $\mathbb{D}$ with $k = 3$.

We can identify a ranking $\succ_{\mathbf{x}}$ with a permutation $\sigma_{\mathbf{x}}$ on $\{1, \ldots, k\}$ such that $\sigma_{\mathbf{x}}(i) < \sigma_{\mathbf{x}}(j)$ iff $\lambda_i \succ_{\mathbf{x}} \lambda_j$, as they are in one-to-one correspondence. $\sigma_{\mathbf{x}}(i)$ is

the *rank* of label λ_i in the ranking $\succ_{\mathbf{x}}$. In the label-wise decomposition method introduced later on, we will often refer to the assignment matrix $y_{\mathbf{x}}$ of ranking, which is a $k \times k$ Boolean matrix ($y_{\mathbf{x},ij} \in \{0,1\}$) of elements $y_{\mathbf{x},ij}$ such that

$$\begin{aligned} \sum_{i=1}^{k} y_{\mathbf{x},ij} &= 1, \quad j = 1, \ldots, k, \\ \sum_{j=1}^{k} y_{\mathbf{x},ij} &= 1, \quad i = 1, \ldots, k. \end{aligned} \tag{1}$$

The value $y_{\mathbf{x},ij} = 1$ means that label λ_i has rank j ($\sigma_{\mathbf{x}}(i) = j$), and constraints (1) ensure that each label has a different rank (hence defining a proper order). We denote by Y the set of all possible assignment matrices. Note that there is a one-to-one correspondence between assignment matrices, permutations and complete rankings. In the following, we will use these terms (rankings, permutations, assignment matrices) interchangeably.

Example 1. *Consider the set* $\Lambda = \{\lambda_1, \lambda_2, \lambda_3\}$ *and the observation* $\lambda_3 \succ \lambda_1 \succ \lambda_2$*, then we have*

$$\sigma_{\mathbf{x}}(1) = 2,\ \sigma_{\mathbf{x}}(2) = 3,\ \sigma_{\mathbf{x}}(3) = 1$$

and the associated assignment matrix y *is*

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

The task in label ranking is the same as the task in usual classification: to use the training instances $\mathbb{D} = \{(\mathbf{x}_i, y_i) | i = 1, \ldots, n\}$ to estimate the theoretical conditional probability measure $P_{\mathbf{x}} : 2^{\mathcal{L}(\Lambda)} \to [0,1]$ associated to an instance $\mathbf{x} \in \mathcal{X}$. However, in label ranking problems the size of $\mathcal{L}(\Lambda)$ quickly increases, even when k is small (for instance, $|\mathcal{L}(\Lambda)| \simeq 10^{12}$ for $k = 15$). This makes the estimation of $P_{\mathbf{x}}$ difficult and potentially quite inaccurate if only little data is available, hence an increased interest in providing accurate yet possibly partial predictions.

This rapid increase of $|\mathcal{L}(\Lambda)|$ also means that estimating directly $P_{\mathbf{x}}$ is in general not doable, except for very small problems. The most usual means to solve this issue is either to decompose the problem into many simpler ones or to assume that $P_{\mathbf{x}}$ follows some parametric law. In this paper, we shall focus on a label-wise decomposition of the problem. To simplify notations, we will drop the subscript $_{\mathbf{x}}$ in the following when there is no possible ambiguity.

3. Label-wise decomposition with probability sets

This section explains how the original label-ranking problem can be reduced to ordinal regression problems in a label-wise manner. Indeed, since in when an observation is precise (i.e., corresponds to a complete ranking over all labels),

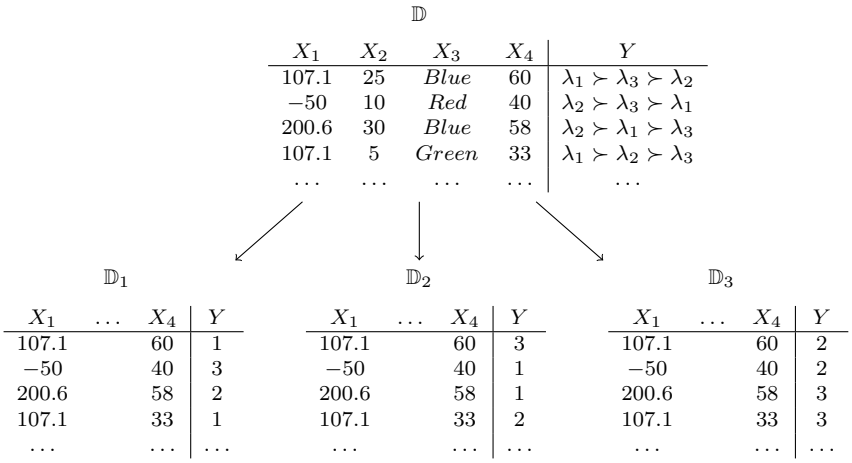

Figure 2: Label-wise decomposition of rankings

each label can be associated to a unique rank, a natural idea is to learn a probabilistic model $p_i : K \to [0, 1]$ with $K = \{1, 2, \ldots, k\}$ and where $p_{ij} = p_i(j)$ is interpreted as the probability $P(\sigma(i) = j)$ that label λ_i has rank j.

A first step is to decompose the original data set $\mathbb{D}$ into k data sets $\mathbb{D}_j = \{(\mathbf{x}_i, \sigma_{\mathbf{x}_i}(j))|i = 1, \ldots, n\}$, $j = 1, \ldots, k$. The decomposition procedure is illustrated by Figure 2. Estimating the probabilities p_{ij} for a given label λ_i then comes down to solving an ordinal regression problem. A natural way to estimate the expected cost c_{ij} of assigning label λ_i to rank j is then to consider a distance $D : K \times K \to \mathbb{R}$ between ranks and to compute

$$c_{ij} = \sum_{\ell=1}^{k} D(j, \ell) p_{i\ell}. \tag{2}$$

Common choices for the distances are the L_1 and L_2 norms, corresponding to

$$D_1(j, \ell) = |j - \ell| \tag{3}$$

and

$$D_2(j, \ell) = (j - \ell)^2. \tag{4}$$

These distances are often used in label-ranking, since D_1 is connected to Spearman's footrule and D_2 to Spearman's rank correlation (Spearman's ρ). In the sequel, we focus on the D_1 distance, yet all presented results extend to D_2 in a straightforward way. Note however that other distances or losses such as Kendall tau that are not label-wise decomposable cannot fit the current framework.

Precise estimates for p_i issued from the finite data set $\mathbb{D}_k$ may be unreliable, especially if these estimates rely on little, unreliable or incomplete data (e.g., if data are scarce around $\mathbf{x}$). Rather than relying on precise estimates in all cases, we propose to consider an imprecise probabilistic model, that is, to consider

for each label λ_i a polytope (a convex set) $\mathcal{P}_i$ of possible probabilities. We will denote by $\mathcal{P} = \times_{i=1}^{k} \mathcal{P}_i$ the Cartesian product of these polytopes. Under a distance D, each set $\mathcal{P}_i$ then induces through Equation (2) a corresponding set $\mathcal{C}_i$ of costs such that

$$\mathcal{C}_i = \{c_{ij} = \sum_{\ell=1}^{k} D(j,\ell) p_{i\ell} \mid p_i \in \mathcal{P}_i\}. \tag{5}$$

Note that, as $\mathcal{P}_i$ is convex and as c_{ij} is a linear function of p_{ik}, $\mathcal{C}_i$ is also a convex set. A common and simple way to define the set $\mathcal{P}_i$ is to provide bounds over the individual values p_{ij}, obtaining the set

$$\mathcal{P}_i = \{\underline{p}_{ij} \leq p_{ij} \leq \overline{p}_{ij}, \sum_{j \in K} p_{ij} = 1\}. \tag{6}$$

This model is commonly called probability intervals [12], yet other learning techniques may produce more complex sets, such as binary decomposition approaches [16] or the popular naive credal classifier [28].

Example 2. *The following bounds give an example of models $\mathcal{P}_i$ on the set $\Lambda = \{\lambda_1, \lambda_2, \lambda_3\}$*

	$\mathcal{P}_1$				$\mathcal{P}_2$				$\mathcal{P}_3$		
	1	2	3		1	2	3		1	2	3
$\overline{p}_{1j}$	0.3	0.5	0.2	$\overline{p}_{2j}$	0.6	0	0.7	$\overline{p}_{3j}$	0.7	0.6	0.4
$\underline{p}_{1j}$	0.3	0.5	0.2	$\underline{p}_{2j}$	0.3	0	0.4	$\underline{p}_{3j}$	0.3	0.2	0

These models indicate that we have precise information about λ_1 and believe its rank is 2, while knowledge about the two other labels are imprecise. We nevertheless know that λ_2 should not have rank 2, while there is much more imprecision about λ_3.

This approach requires to learn k different models, one for each label. This is to be compared to the RPC [21] approach, in which $k(k-1)2$ models (one for each pair of labels) have to be learnt. There is therefore a clear computational advantage for the current approach when k increases. It should also be noted that the two approaches rely on different models: while the label-wise decomposition uses learning methods issued from ordinal regression problems, the RPC approach usually uses learning methods issued from binary classification.

However, as for the RPC approach (and its cautious versions [10, 9, 15]), the label-wise decomposition requires to aggregate all decomposed models into a single (partial) prediction. Indeed, focusing only on decomposed models $\mathcal{P}_i$, nothing forbids to predict the same rank for multiple labels. In the next section, we focus on how to describe a set of potentially optimal solutions given the uncertain costs $\mathcal{C}_i$. To do this, we introduce a new way to handle the well-known assignment problem when costs are uncertain.

4. Cautious inference by optimisation

When costs c_{ij} are precisely valued, i.e., when $\mathcal{P}_i$ is reduced to a single probability p_i for any $i = 1, \ldots, k$, then finding an optimal ranking comes down to finding the assignment matrix y_{ij} that minimizes the overall cost c. One can easily see that this results in an assignment problem, which can be modelled with binary optimization variables y_{ij} equal to 1 iff label i is assigned to position j:

$$
\begin{aligned}
& \min && \sum_{i,j \in K} c_{ij} y_{ij} \\
(AP) \qquad & \text{s.t.} && \sum_{i \in K} y_{ij} = 1, \quad j \in K \\
& && \sum_{j \in K} y_{ij} = 1, \quad i \in K \\
& && y_{ij} \in \{0,1\}, \quad i,j \in K. \qquad (7)
\end{aligned}
$$

Solving the above optimisation problem can be done by using the well-known Hungarian algorithm [22] which has complexity $\mathcal{O}(k^3)$. However, in our case the costs are uncertain and belong to polytope $\mathcal{C} = \times_{i=1}^{k} \mathcal{C}_i$. A classical approach for such problems, such as the one used in Robust Optimisation [3], would seek *minmax* solutions of the problem [13], yet this would not give us a partial prediction reflecting our lack of information about the costs. Also, recent works suggest that minmax solutions are likely to be sub-optimal in Machine learning problems [19].

Let us denote by Y the set of binary matrices that are feasible for (AP) (the set of possible assignment matrices). Given two binary matrices y and y' feasible for (AP), we say that y dominates y', which is denoted $y \succ y'$, if $\sum_{i,j} c_{ij} y_{ij} < \sum_{i,j} c_{ij} y'_{ij}$ for all $c \in \mathcal{C}$. This way of defining dominance is strongly connected to the so-called maximality criterion used in imprecise probabilities [25]. Ideally, we would want to retrieve the full Pareto set of non-dominated solutions

$$\mathcal{Y} = \{y \in Y \mid \nexists y' \in Y \text{ with } y' \succ y\}$$

induced by this dominance criterion. Yet Y has the same size as $\mathcal{L}(\Lambda)$, and computing or even approximating $\mathcal{Y}$ with theoretical guarantee can be very difficult in practice. This is why we focus in what follows on finding outer approximations of $\mathcal{Y}$. We provide a means to infer a partial order on Λ whose linear extensions form a superset of $\mathcal{Y}$. The method comes down to assessing for a given pair of labels $\lambda_{i_1}, \lambda_{i_2}$ whether label λ_{i_1} is preferred to label λ_{i_2} for all elements of $\mathcal{Y}$. By doing this for every pair of labels, we then obtain a partial order. Note that as all elements of $\mathcal{Y}$ are proper rankings, there is no risk of inferring cyclical relations.

We first introduce some notations. Given $i_1, i_2 \in K$, let us define by

$$Y_{i_1 \succ i_2} = \{y \in Y \mid y_{i_1 j_1} = y_{i_2 j_2} = 1, j_1 < j_2\}$$

the set of all solutions where label λ_{i_1} is preferred to (has a lower rank/position than) label λ_{i_2}. Then label λ_{i_1} is preferred over label λ_{i_2}, which is denoted $\lambda_{i_1} \succ \lambda_{i_2}$, if $\mathcal{Y} \subseteq Y_{i_1 \succ i_2}$. This characterization is not practical, however, since we are unable to compute the full Pareto set $\mathcal{Y}$. Our objective below is to provide a *sufficient condition* to assert whether label λ_{i_1} is preferred over label λ_{i_2}. In other words, we compute a subset $\mathcal{I}'$ of the set $\mathcal{I}$ that contains the pairs of labels for which a preference can be established, i.e.

$$\mathcal{I} = \{(i_1, i_2) \in K \times K \mid \mathcal{Y} \subseteq Y_{i_1 \succ i_2}\}. \tag{8}$$

In this context, we introduce additional notation. For any $y \in Y_{i_1 \succ i_2}$, we define $y^{i_1 \| i_2}$ as the element in $Y_{i_2 \succ i_1}$ such that $y^{i_1 \| i_2}_{ij} = y_{ij}$ for $i \neq i_1, i_2$, and $y^{i_1 \| i_2}_{i_1 j} = y_{i_2 j}$ and $y^{i_1 \| i_2}_{i_2 j} = y_{i_1 j}$ for all $j \in K$. That is, $y^{i_1 \| i_2}$ corresponds to the ranking y where only the positions of the labels λ_{i_1} and λ_{i_2} have been swapped. The result below provides a sufficient condition for $\lambda_{i_1} \succ \lambda_{i_2}$ in the form of an optimization problem, whose objective function has the advantage to focus only on the ranks of labels $\lambda_{i_1}$and λ_{i_2}. The proof is presented in Appendix A.

Proposition 1. *Given $i_1, i_2 \in K$, a sufficient condition for $\lambda_{i_1} \succ \lambda_{i_2}$ is that the optimal solution cost of the optimization problem below is negative*

$$\begin{aligned} z(i_1, i_2) = \quad & \max \quad \sum_{j \in K} c_{i_1 j}(y_{i_1 j} - y_{i_2 j}) + c_{i_2 j}(y_{i_2 j} - y_{i_1 j}) \\ (WeakDom^{i_1}_{i_2}) \qquad & s.t. \quad c \in \mathcal{C} \\ & \quad\;\; y \in Y_{i_1 \succ i_2} \end{aligned} \tag{9}$$

Proposition 1 provides us with a first approach for finding relations between pairs of labels. However, the result suffers from two drawbacks. First, it amounts to solve a bilinear mixed-integer program, which is $\mathcal{NP}-hard$ to solve in general. Second, the sufficient condition is too restrictive. Indeed, when probabilities are precise (i.e., when $\mathcal{C}$ is reduced to a singleton), we would like that our approach returns the full set $\mathcal{I}$ defined in (8), determining a unique optimal solution (or a set of solutions with equal minimal costs, if the optimal solution is non-unique). Unfortunately, Proposition 1 fails to include all relations in $\mathcal{I}$, even when probabilities are precise and when there is only one optimal solution. Namely, let

$$\mathcal{I}^z = \{(i_1, i_2) \in K \times K \mid z(i_1, i_2) < 0\}$$

be the set returned by Proposition 1. The example below shows that Proposition 1 can lead to a set $\mathcal{I}^z$ strictly included in $\mathcal{I}$.

Example 3. *Consider again the space $\Lambda = \{\lambda_1, \lambda_2, \lambda_3\}$ with the following precise cost matrix c (possibly derived from precise probabilities)*

$$c = \begin{pmatrix} 2 & 3 & 4 \\ 4 & 2 & 3 \\ 3 & 4 & 2 \end{pmatrix}.$$

Solving the problem (AP) gives $y_{11} = 1$, $y_{22} = 1$ *and* $y_{33} = 1$, *hence the ranking* $\lambda_1 \succ \lambda_2 \succ \lambda_3$ *and the set* $\mathcal{I} = \{(1,2),(2,3),(1,3)\}$. *Now, consider the ranking* $\lambda_2 \succ \lambda_3 \succ \lambda_1$, *that is the matrix* y *with* $y_{13} = 1$, $y_{21} = 1$ *and* $y_{32} = 1$, *then the value*

$$\sum_{j \in K} c_{2j}(y_{2j} - y_{3j}) + c_{3j}(y_{3j} - y_{2j}) = (c_{21} - c_{31}) + (c_{32} - c_{22})$$
$$= (4-3) + (4-2) = 3$$

is positive, indicating that the maximum of $z(i_1, i_2)$ *for* $i_1 = 2$ *and* $i_2 = 3$ *is positive as well (as it is higher than* 3*). Hence, according to Proposition 1, we have* $(2,3) \notin \mathcal{I}^z$, *while* $(2,3) \in \mathcal{I}$.

The next proposition shows that in our setting, the general problem $(WeakDom_{i_2}^{i_1})$ of Proposition 1 can be solved efficiently. This result will be instrumental to solve a further optimisation problem resulting from a more stringent sufficient condition.

Proposition 2. *Assume that* $\mathcal{C} = \times_{i \in K} \mathcal{C}_i$ *and let* $i_1, i_2 \in K$ *be given. Then, the value of* $z(i_1, i_2)$ *is equal to the maximum of*

$$\max_{c \in \mathcal{C}} c_{i_1 j_1} + c_{i_2 j_2} - c_{i_1 j_2} - c_{i_2 j_1} \tag{10}$$

taken over all pairs $j_1 < j_2$ *such that* $(j_1, j_2) \in K \times K$

The proof can be found in Appendix B. Interestingly, this result showing that $(WeakDom_{i_2}^{i_1})$ can be solved in polynomial time relies on the fact that $\mathcal{C}$ is the Cartesian product of $\mathcal{C}_i$ for $i \in K$, and therefore can be used thanks to the label-wise decomposition we consider. More specifically, we show that $z(i_1, i_2)$ can be computed by solving $O(K^2)$ linear programs.

Example 4. *Let us pursue Example 3 considering* $i_1 = 2$ *and* $i_2 = 3$, *then we have:*

- *for* $j_1 = 1, j_2 = 2$, $c_{21} + c_{32} - c_{22} - c_{31} = 3$
- *for* $j_1 = 1, j_2 = 3$, $c_{21} + c_{33} - c_{23} - c_{31} = 0$
- *for* $j_1 = 2, j_2 = 3$, $c_{22} + c_{33} - c_{23} - c_{32} = -3$

and the maximal value is 3, *the one found in Example 3.*

We must now deal with the second issue, the one of conservatism. We can address this issue of Proposition 1 by restricting the number of elements considered in constraint (9) to a relevant subset of $Y_{i_1 \succ i_2}$. In this aim, we put a filter on any y that is considered in (9), by imposing that either y or $y^{i_1 \| i_2}$ be non-dominated by some $y^* \in \mathcal{Y}$. For instance, y^* can be the minmax solution, computable by dualizing the linear program [3], or more simply the solution of (AP) corresponding to sampled values of $\mathcal{P}_i$. We obtain the sufficient condition stated next, less conservative than Proposition 1, whose proof can be found in Appendix C.

Proposition 3. *Let $y^* \in \mathcal{Y}$ be given and consider $i_1, i_2 \in K$. A sufficient condition for $\lambda_{i_1} \succ \lambda_{i_2}$ is that the optimal solution cost of the optimization problem below is negative*

$$
\begin{array}{lrll}
 & w(i_1,i_2) = & \max & \displaystyle\sum_{j \in K} c_{i_1 j}(y_{i_1 j} - y_{i_2 j}) + c_{i_2 j}(y_{i_2 j} - y_{i_1 j}) \\
(Dom_{i_2}^{i_1}) & & s.t. & c \in \mathcal{C} \\
 & & & y \in Y_{i_1 \succ i_2} \\
 & & & y \not\prec y^* \ \textbf{\textit{OR}}\ y^{i_1 \| i_2} \not\prec y^*
\end{array} \tag{11}
$$

One readily sees that $w(i_1, i_2) \leq z(i_1, i_2)$ for all $i_1, i_2 \in K$, as constraints of $(Dom_{i_2}^{i_1})$ defines a set of feasible solutions strictly included in the one described by the constraints of $(WeakDom_{i_2}^{i_1})$. Therefore, the sufficient condition from Proposition 3 is less conservative than the condition from Proposition 1. In particular, the result below shows that Proposition 3 never leads to the situation witnessed in Example 3. Its proof is provided in Appendix D.

Proposition 4. *Let $\mathcal{C}$ be a singleton and let $\mathcal{I}^w = \{(i_1, i_2) \in K \times K \mid w(i_1, i_2) < 0\}$. It holds that $\mathcal{I}^w = \mathcal{I}$.*

Proposition 3 indicates how to get an outer-approximation that is likely to not be too conservative, but does not solve the complexity issue. The next result shows that we can use Proposition 2 in a slightly modified way to integrate the constraint given by Equation (11), therefore keeping the complexity of computing $w(i_1, i_2)$ polynomial. Its proof can be found in Appendix E.

Proposition 5. *Assume that $\mathcal{C} = \times_{i \in K} \mathcal{C}_i$ and let $i_1, i_2 \in K$ be given. Then, the value of $w(i_1, i_2)$ is equal to the maximum of*

$$\max_{c \in \mathcal{C}} c_{i_1 j_1} + c_{i_2 j_2} - c_{i_1 j_2} - c_{i_2 j_1} \tag{12}$$

taken over all pairs $j_1 < j_2$ such that $v(j_1, j_2) \leq 0$ or $v(j_2, j_1) \leq 0$, where $v(j', j'')$ is the optimal solution cost of

$$
\begin{array}{lll}
\min & \displaystyle\sum_{i,j \in K} c^*_{ij} y_{ij} & \\
s.t. & y \in Y & \\
 & y_{i_1 j'} = y_{i_2 j''} = 1, &
\end{array}
$$

and $c^*_{ij} = \min\limits_{c_i \in \mathcal{C}_i} c_{ij} - c_{ij^*_i}$ for each $i, j \in K$, where j^*_i is such that $y^*_{ij^*_i} = 1$.

Compared to Proposition 2, the only added step is to check which pair (j_1, j_2) are such that $v(j_1, j_2) \leq 0$ or $v(j_2, j_1) \leq 0$, which can be done before solving (12). Let us denote by $S(y^*, i_1, i_2)$ the set that contains all pairs (j_1, j_2) such that $j_1 < j_2$ and $v(j_1, j_2) \leq 0$ or $v(j_2, j_1) \leq 0$ for the element y^*. Algorithm 1 summarizes the procedure to obtain a predicted partial ranking, given probability sets $\mathcal{P}_i$.

Example 5. *Let us pursue Example 4 considering* $i_1 = 2$, $i_2 = 3$, $j_1 = 1$ *and* $j_2 = 2$. *Let us now compute the values* $v(1,2)$ *and* $v(2,1)$. *The only possible* y^* *is* $y^*_{11} = 1, y^*_{22} = 1, y^*_{33} = 1$. *We then have the matrix*

$$c^* = \begin{pmatrix} 0 & 1 & 2 \\ 2 & 0 & 1 \\ 1 & 2 & 0 \end{pmatrix}.$$

We then have, for example, $v(1,2) = c^*_{13} + c^*_{21} + c^*_{32} = 6$ *which is indeed positive. Note that as matrix* c^* *is positive, the only case for which we have* $v(j_1, j_2) = 0$ *is when* $j_1 = 2$ *and* $j_2 = 3$, *which do correspond to the unique optimal solution.*

Algorithm 1: Algorithm to obtain Pareto set approximation $\mathcal{I}^w$ with one filter y^*

Input: Uncertainty models $\mathcal{P}_i$, element y^*
Output: $\mathcal{I}^w$
$\mathcal{I}^w = \emptyset$;
for *every* $(i_1, i_2) \in K^2$ **do**
  $w(i_1, i_2) = -\infty$;
  $\mathcal{S}(y^*, i_1, i_2) = \{(j_1, j_2) \in K^2 \mid j_1 < j_2\}$;
  for *each* $(j_1, j_2) \in K^2$ **do**
    Compute $v(j_1, j_2)$ and $v(j_2, j_1)$;
    if $v(j_1, j_2) > 0$ *and* $v(j_2, j_1) > 0$ **then**
    $\mathcal{S}(y^*, i_1, i_2) = \mathcal{S}(y^*, i_1, i_2) \setminus (j_1, j_2)$
  for *each* $(j_1, j_2) \in \mathcal{S}^*(i_1, i_2)$ **do**
    w_{cur}=Eq. (12) solution ;
    if $w(i_1, i_2) < w_{cur}$ **then** $w(i_1, i_2) = w_{cur}$
  if $w(i_1, i_2) \leq 0$ **then** $\mathcal{I}^w = \mathcal{I}^w \cup (i_1, i_2)$
return $\mathcal{I}^w$

Interestingly enough, we can notice that picking multiple elements $y^{*_1}, \ldots, y^{*_m}$ and adding the constraints

$$y \not\prec y^{*_j} \textbf{ OR } y^{i_1 \| i_2} \not\prec y^{*_j}$$

for $j \in \{1, \ldots, m\}$ to $(Dom^{i_1}_{i_2})$ simply comes down to considering the pairs (j_1, j_2) within the intersection $\cap_{j=1}^m S(y^{*_j}, i_1, i_2)$. We will denote by $\mathcal{S}^*(i_1, i_2) = \cap_{j=1}^m S(y^{*_j}, i_1, i_2)$ this intersection. Computing $\mathcal{S}^*(i_1, i_2)$ can then even be done iteratively, as the pairs eliminated for $S(y^{*_j}, i_1, i_2)$ do not have to be checked for $S(y^{*_{j+1}}, i_1, i_2), \ldots, S(y^{*_m}, i_1, i_2)$. The slightly more complex procedure to obtain $\mathcal{I}^w$ with multiple filtering solutions y^{*_j} is summarized in Algorithm 2.

Note that when the costs

$$c_{ij} = \sum_{\ell \in K} |\ell - j| p_{i\ell}, \tag{13}$$

Algorithm 2: Algorithm to obtain Pareto set approximation $\mathcal{I}^w$ with multiple filters

Input: Uncertainty models $\mathcal{P}_i$, elements $y^{*_1}, \ldots, y^{*_m}$
Output: $\mathcal{I}^w$
$\mathcal{I}^w = \emptyset$;
for *every* $(i_1, i_2) \in K^2$ **do**
  $w(i_1, i_2) = -\infty$;
  $\mathcal{S}^*(i_1, i_2) = \{(j_1, j_2) \in K^2 \mid j_1 < j_2\}$;
  for $\ell = 1, \ldots, m$ **do**
    for *each* $(j_1, j_2) \in \mathcal{S}^*(i_1, i_2)$ **do**
      $y^* = y^{*_\ell}$;
      Compute $v(j_1, j_2)$ and $v(j_2, j_1)$;
      if $v(j_1, j_2) > 0$ *and* $v(j_2, j_1) > 0$ **then**
      $\mathcal{S}^*(i_1, i_2) = \mathcal{S}^*(i_1, i_2) \setminus (j_1, j_2)$
  for *each* $(j_1, j_2) \in \mathcal{S}^*(i_1, i_2)$ **do**
    w_{cur}=Eq. (12) solution ;
    if $w(i_1, i_2) < w_{cur}$ **then** $w(i_1, i_2) = w_{cur}$
  if $w(i_1, i_2) \leq 0$ **then** $\mathcal{I}^w = \mathcal{I}^w \cup (i_1, i_2)$
return $\mathcal{I}^w$

are derived using D_1 distance, problem (12) can be rewritten as a linear program on variables p rather than c:

$$\begin{aligned} & \max_{p \in \mathcal{P}} \sum_{\ell \in K} |\ell - j_1| p_{i_1 \ell} + \sum_{\ell \in K} |\ell - j_2| p_{i_2 \ell} - \sum_{\ell \in K} |\ell - j_2| p_{i_1 \ell} - \sum_{\ell \in K} |\ell - j_1| p_{i_2 \ell} \\ = \quad & \max_{p \in \mathcal{P}} \sum_{\ell \in K} (|\ell - j_1| - |\ell - j_2|) p_{i_1 \ell} + (|\ell - j_2| - |\ell - j_1|) p_{i_2 \ell}. \end{aligned} \tag{14}$$

Alternatively, Eq. (14) can be rewritten

$$\begin{aligned} & \max_{p_{i_1} \in \mathcal{P}_{i_1}} \sum_{\ell \in K} (|\ell - j_1| - |\ell - j_2|) p_{i_1 \ell} + \max_{p_{i_2} \in \mathcal{P}_{i_2}} (|\ell - j_2| - |\ell - j_1|) p_{i_2 \ell} = \\ & - \min_{p_{i_1} \in \mathcal{P}_{i_1}} \sum_{\ell \in K} (|\ell - j_2| - |\ell - j_1|) p_{i_1 \ell} - \min_{p_{i_2} \in \mathcal{P}_{i_2}} (|\ell - j_1| - |\ell - j_2|) p_{i_2 \ell} = \\ & - \underline{\mathbb{E}}_{i_1}(D_1(j_2, \cdot) - D_1(j_1, \cdot)) - \underline{\mathbb{E}}_{i_2}(D_1(j_1, \cdot) - D_1(j_2, \cdot)) \end{aligned}$$

where $\underline{\mathbb{E}}_{i_1}$ (resp. $\underline{\mathbb{E}}_{i_2}$) is a lower expectation under $\mathcal{P}_{i_1}$ (resp. $\mathcal{P}_{i_2}$). Said in other words, $\lambda_{i_1} \succ \lambda_{i_2}$ if

$$\underline{\mathbb{E}}_{i_1}(D_1(j_2, \cdot) - D_1(j_1, \cdot)) + \underline{\mathbb{E}}_{i_2}(D_1(j_1, \cdot) - D_1(j_2, \cdot)) \tag{15}$$

is positive for every $(j_1, j_2) \in \mathcal{S}^*(i_1, i_2)$. Expression (15) has a nice interpretation, as it says that $\lambda_{i_1} \succ \lambda_{i_2}$ if the expected cost of swapping i_1 from rank j_2 to j_1 $(D_1(j_2, \cdot) - D_1(j_1, \cdot))$ and swapping i_2 from rank j_1 to j_2 is positive.

This, again, emphasizes the strong links this approach has with the notion of maximality [25].

Furthermore, when probability sets $\mathcal{P}_i$ correspond to probability intervals (see Equation (6)) and when the costs are derived using the D_1 distance (Equation (3)), we can solve problem (14) by using two sorting algorithms. We denote $a_k = |k-j_1|-|k-j_2|$ and $b_k = |k-j_2|-|k-j_1|$ for each $k \in K$, reorder variables $p_{i_1 k}$ (resp. $p_{i_2 k}$) according to the decreasing values of a_k (resp. b_k), and define $a_0 = 1$ and $b_0 = 1$. Then, the solution of problem (14) is obtained by fixing $p_{i_1 1}$ to $a^* = \min(\overline{p}_{i_1 1}, a_0)$ (resp. $p_{i_2 1}$ to $b^* = \min(\overline{p}_{i_2 1}, b_0)$), subtracting a^* from a_0 (resp. b^* from b_0), and repeating the operation with the second element of the list until a_0 and b_0 are equal to 0. This corresponds to applying the so-called Choquet integral to obtain lower expectations.

5. Experiments

This section shows results of some experiments done on several synthetic data sets. In particular, our goal was to compare our proposed approach with the construction of an inner approximation of the Pareto front, obtained by sampling precise distributions within sets $\mathcal{P}_i$.

5.1. Data sets and classifiers

The data sets we used in the experiment are presented in [21]. There are two types of data sets [20]:

- synthetic data sets consisting in multiclass data sets issued from the UCI machine learning repository [1] that have been transformed into label ranking data by a procedure proposed in [21]. To obtain this transformation, a naive Bayes classifier is first trained on the entire data set. Then, for each instance, the labels are ordered according to the predicted class probabilities;
- real-world data sets issued from the bioinformatics fields where the outputs are qualitative profiles of yeast gene expressions. More details can be found in [21].

All these data sets are described in Table 1.

As explained in Section 2, the problem of label ranking is decomposed in a label-wise manner into k ordinal regression problems. Logistic regression is used as a base learner. For each classifier, and thus for each label λ_i, B bootstrap replicates of the learning set are used to provide B precise estimates of $\mathbf{p}_i$. Polytope $\mathcal{P}_i$ is obtained as a set of simultaneous confidence intervals $[\underline{p}_{ij}; \overline{p}_{ij}]$ $j = 1, k$, determined so as to contain the α % most central values of $\mathbf{p}_i$, where α is a given level of confidence.

The results of our method are compared to those obtained with a Monte-Carlo (MC) approach. This one consists in computing several solutions (linear orders) of (AP) using precise costs computed from randomly sampled distributions within $\mathcal{P}_i$. Note that this procedure provides a set of complete orders that form an inner approximation of $\mathcal{Y}$ (a subset of $\mathcal{Y}$).

Table 1: Data sets description for label ranking

data set	#features	#labels	#instances
iris	4	3	150
glass	9	6	214
vehicle	18	4	846
segment	18	7	2310
authorship	70	4	841
dtt	24	4	2465
cold	24	4	2465

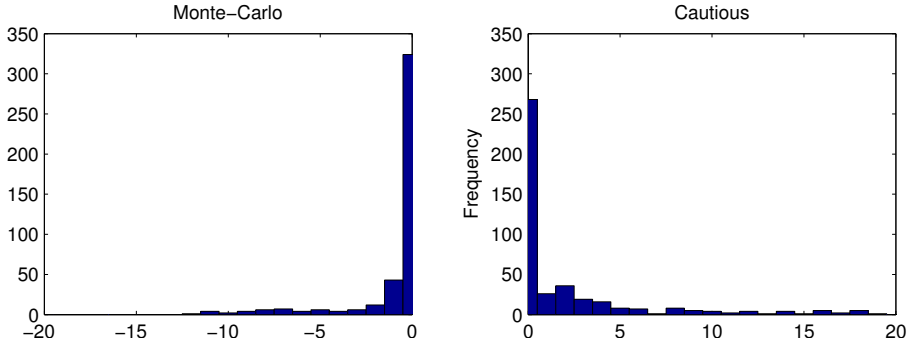

Figure 3: Distribution of differences of cardinalities between $\mathcal{Y}$ and the approximations (left:MC approach; right: cautious approach).

5.2. Approximation quality

In a first experiment, we compare the approximation quality of the two methods using the Vehicle data set. The number of labels in this case is limited (four labels) so that the enumeration of all possible rankings remains possible (24 possible rankings). The exact Pareto set $\mathcal{Y}$ of non dominated solutions can thus be determined and compared with the approximations given by both methods. For this experiment, half of the data set has been used for training, the other part being kept for the test. Parameter α was set to 95%. The number of bootstrap replicates was fixed to 100 and we used 100 random samplings in the MC approach. The distributions of the deviations between the cardinality of the exact set $\mathcal{Y}$ and the cardinality of the approximate ones are shown in Figure 3 for both methods.

It can be seen that the approximation quality is generally good for both methods. The difference of cardinalities for the MC approach seems to be slightly smaller than for the cautious approach, but the MC approach, even with a high number of random samplings, provides no guarantees that the result will be exact (in fact, there are a few experiments in which the MC approach underestimate significantly the size of $\mathcal{Y}$). On the contrary, the cautious approach guarantees that the result is a superset of $\mathcal{Y}$. It should also be noted that these good results are obtained using Algorithm 1 and only one filtering

element y^* obtained by random sampling. From this, it appears that adding further filtering elements is unnecessary in most cases, yet adding some further filtering elements when producing very partial rankings may help to improve the approximation quality.

5.3. Performances on label ranking problems

In this section, we illustrate the behaviour of the MC and the cautious approaches using the five data sets described in Table 1. The result of each method being a partial order, we use two specific measures proposed in [10] to evaluate the performances of the methods. The first one, *correctness*, quantifies how the predicted (partial) ranking matches the true ranking, whereas the second one is intended to measure the degree of *completeness* of the relation. A good method predicting partial orders should see its correctness increase as the completeness decrease, and there is usually a trade-off to be found between the two criteria. They can be formally defined as follows. Let $\mathbb{D}_j = \{(\mathbf{x}_i, \sigma_{\mathbf{x}_i}(j))|i = 1, \ldots, n\}$, $j = 1, \ldots, k$ denote the test sets used in the experiments. Let $\mathcal{I}_{\mathbf{x}_i}$ denote the approximation of $\mathcal{I}$ for instance $\mathbf{x}_i$ found either by Algorithm 1 or by a MC approach. Then, for each $\mathbf{x}_i$, a pair of labels $\{\lambda_k, \lambda_l\}$ is said to be concordant if:

$$(k, l) \in \mathcal{I}_{\mathbf{x}_i} \text{ and } \sigma_{\mathbf{x}_i}(k) < \sigma_{\mathbf{x}_i}(l) \text{ or } (l, k) \in \mathcal{I}_{\mathbf{x}_i} \text{ and } \sigma_{\mathbf{x}_i}(l) < \sigma_{\mathbf{x}_i}(k).$$

It is said to be discordant if:

$$(k, l) \in \mathcal{I}_{\mathbf{x}_i} \text{ and } \sigma_{\mathbf{x}_i}(l) < \sigma_{\mathbf{x}_i}(k) \text{ or } (l, k) \in \mathcal{I}_{\mathbf{x}_i} \text{ and } \sigma_{\mathbf{x}_i}(k) < \sigma_{\mathbf{x}_i}(l).$$

Let c_i and d_i denote the number of concordant and discordant pairs of labels for instance $\mathbf{x}_i$, respectively. The correctness measure for the whole test set is defined as:

$$Correctness = \frac{1}{n} \sum_{i=1}^{n} \frac{c_i - d_i}{c_i + d_i},$$

whereas the completeness is defined by:

$$Completeness = \frac{1}{n} \sum_{i=1}^{n} \frac{c_i + d_i}{n(n-1)/2}.$$

Note that the correctness measure can be considered as a generalization for partial rankings of the Kendall tau, classically used for evaluating the correlation of complete rankings.

In the experiments, we used the following settings to obtain the probability sets $\mathcal{P}_i$: we produced 100 Bootstrap samples from the initial training data sets, resulting in 100 estimations $\mathbf{p}_i$ of the rank probabilities for each label λ_i. To get confidence intervals of increasing size (and thus sets $\mathcal{P}_i$ of increasing sizes), we varied the value from 0 to 0.95. Note that, by convention, when α is set to zero, only the most central value of the 100 $\mathbf{p}_i$ is kept, so that precise costs are used and the completeness is equal to one.

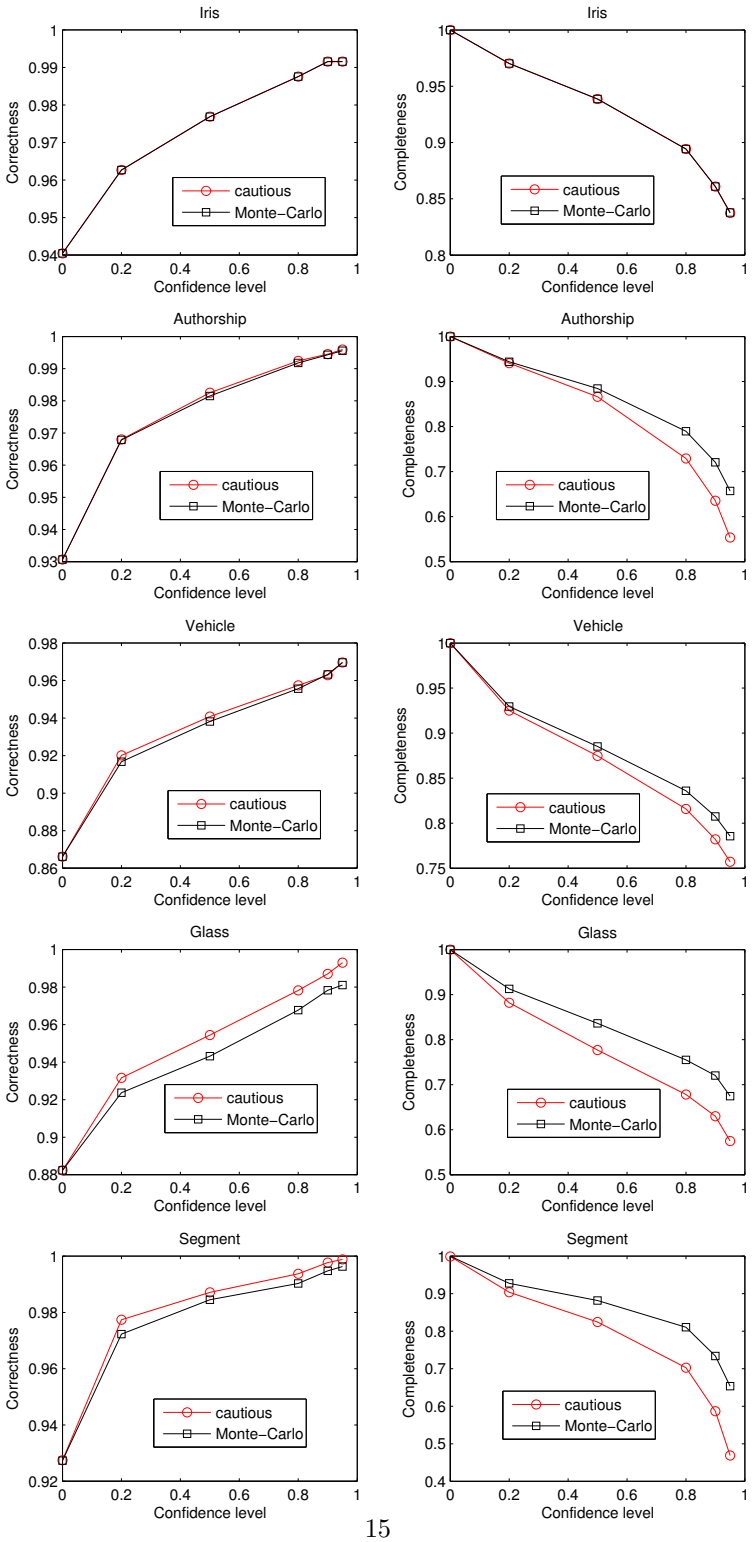

Figure 4: Experimental results on synthetic data sets.

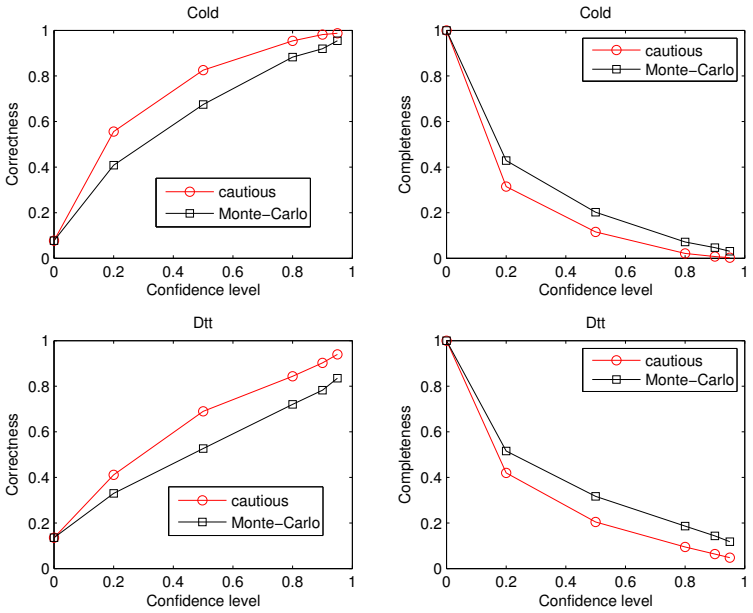

Figure 5: Experimental results on real-world data sets.

The results of the experiments are displayed in Figures 4 and 5, and are mean values computed over five repetitions of a 5-fold cross-validation procedure. In the Monte Carlo approach, the final partial order from which correctness and completeness are computed is obtained by keeping the pairwise preference relations common to all orders resulting from the sampling. From these experiments, the following conclusions may be drawn:

1. As expected, for both methods, an increase of α leads to a greater abstention but to a significant improvement of the correctness. The abstention is thus done on poorly reliable predictions. This shows the interest of the use of imprecise probabilities when solving a label ranking problem through a label-wise approach.
2. The performances of the two methods are very similar when the number of labels is small (iris, authorship, vehicle) but show some qualitative differences when the number of classes, and thus the number of possible rankings, grows. The cautious approach allows to reach a better correctness while keeping the completeness at a reasonable level. Beyond the theoretical guarantees it gives of being cautious, this confirms the interest of our approach.

Finally, let us note that the closeness of the results again indicate that using Algorithm 1 and only one filtering element y^* was sufficient for the considered data sets. We can conjecture that adding further filtering elements will become more interesting as the number of labels grows, yet it should be noted that in

the label ranking setting, the number of labels is typically limited (compared to, e.g., instance ranking or object ranking problems, where considering millions of items to rank is not unusual).

6. Conclusion

Label ranking is a hard learning problem with structured outputs where making cautious (partial) and reliable predictions may be preferable to making complete yet partially unreliable ones. To reduce the complexity of the problem, we consider a label-wise decomposition, while we consider using sets of probabilities to make cautious predictions about the label ranks.

This leads us to consider an assignment problem with imprecise costs from which we want to extract a Pareto front (i.e., the set of undominated rankings). Such a problem is in general $\mathcal{NP} - hard$ to solve, however for the specific case of label-wise decomposition, we provide a new (to our knowledge), elegant and efficient (polynomial) solution to the problem, that provides an outer approximation of the Pareto front.

Our experiments on benchmark data sets show that the approach is sound, and does not provide overly conservative predictions, something we may fear with outer-approximation. In fact, most of the time it will provide exact solutions. Experiments also seems to indicate that the interest of the approach increases with the number of labels to rank. In future experiments, we also plan to integrate missing data, since though label-wise decomposition tends to perform better than pairwise decomposition when no data are missing, it is also more sensible to missing data than pairwise approaches [6]. Yet there are different ways into which missing data can be considered within imprecise probabilistic inferences [29], hence considering them is beyond the scope of this paper.

As a perspective to this work, we plan to work on developing efficient methods to find other outer-approximations. As an example, we may search to tell which rank each label can take within $\mathcal{Y}$. Such an approximation would be able to express some situations that pairwise comparisons cannot, for example it is possible to perfectly represent the Pareto set $\mathcal{Y} = \{\lambda_1 \succ \lambda_2 \succ \lambda_3, \lambda_3 \succ \lambda_2 \succ \lambda_1\}$ by giving the possible rank of each label, while it is impossible to exactly represent it with pairwise comparisons. Note that the converse is also true, as there are Pareto sets that can be exactly represented by pairwise comparisons, but cannot be represented exactly by giving the set of ranks a label can assume. Also, finding efficiently which rank a label can assume in $\mathcal{Y}$ is a trickier problem, as it becomes difficult to exploit the ordinal structure of ranks.

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).

Appendix A. Proof of Proposition 1

Proof. We must show that $z(i_1, i_2) < 0$ implies that $\mathcal{Y} \subseteq Y_{i_1 \succ i_2}$. Notice that since $Y = Y_{i_1 \succ i_2} \cup Y_{i_2 \succ i_1}$, any $y \in \mathcal{Y}$ belongs either to $Y_{i_1 \succ i_2}$ or to $Y_{i_2 \succ i_1}$. We show next that $z(i_1, i_2) < 0$ implies that $Y_{i_2 \succ i_1} \cap \mathcal{Y} = \emptyset$, which proves the result.

Let y be any element in $Y_{i_2 \succ i_1}$ and let us compare the cost of y with the cost of $y^{i_1 \| i_2}$ (notice that $y^{i_1 \| i_2} \in Y_{i_1 \succ i_2}$) for any $c \in \mathcal{C}$:

$$\sum_{i,j \in K} c_{ij} y_{ij}^{i_1 \| i_2} - \sum_{i,j \in K} c_{ij} y_{ij} = \sum_{j \in K} \Big((c_{i_1 j} y_{i_1 j}^{i_1 \| i_2} + c_{i_2 j} y_{i_2 j}^{i_1 \| i_2}) - (c_{i_1 j} y_{i_1 j} + c_{i_2 j} y_{i_2 j}) \Big) \tag{A.1}$$

$$= \sum_{j \in K} c_{i_1 j} (y_{i_1 j}^{i_1 \| i_2} - y_{i_2 j}^{i_1 \| i_2}) + c_{i_2 j} (y_{i_2 j}^{i_1 \| i_2} - y_{i_1 j}^{i_1 \| i_2}) \tag{A.2}$$

$$\leq z(i_1, i_2), \tag{A.3}$$

where (A.1) holds because y and $y^{i_1 \| i_2}$ are equal for all rows different from i_1 and i_2, (A.2) follows from $y_{i_1} = y_{i_2}^{i_1 \| i_2}$ and $y_{i_2} = y_{i_1}^{i_1 \| i_2}$, and (A.3) follows from the definition of $z(i_1, i_2)$. Hence, $z(i_1, i_2) < 0$ implies that $\sum_{i,j} c_{ij} y_{ij}^{i_1 \| i_2} < \sum_{i,j} c_{ij} y_{ij}$ for all $c \in \mathcal{C}$, and therefore, $y^{i_1 \| i_2} \succ y$. Thus, $y \notin \mathcal{Y}$. □

Appendix B. Proof of Proposition 2

Proof. In view of the objective function of $(WeakDom_{i_2}^{i_1})$, the only relevant components of y are $y_{i_1 j}$ and $y_{i_2 j}$ for all $j \in K$. Projecting set $Y_{i_1 \succ i_2}$ over these components yields the finite set below

$$\sum_{m=1}^{\ell} y_{i_1 m} \geq \sum_{m=1}^{\ell} y_{i_2 m}, \qquad \ell \in K \tag{B.1}$$

$$\sum_{j \in K} y_{ij} = 1, \qquad i \in \{i_1, i_2\} \tag{B.2}$$

$$y_{ij} \in \{0, 1\}, \qquad j \in K, i \in \{i_1, i_2\}. \tag{B.3}$$

Constraint (B.1) ensures that the rank given to i_1 is lower than the rank given to i_2. Hence, we can compute $z(i_1, i_2)$ by enumerating the $K(K-1)/2$ solutions to (B.1)–(B.3). □

Appendix C. Proof of Proposition 3

Proof. The proof is similar to the proof of Proposition 1, with y any element in $Y_{i_2 \succ i_1}$ such that both y and $y^{i_1 \| i_2}$ are not dominated by y^*. Indeed, if $y \prec y^*$ and $y^{i_1 \| i_2} \prec y^*$, it means that neither y nor $y^{i_1 \| i_2}$ are in $\mathcal{Y}$, therefore we do not need to consider them in the previous proof. □

It is tempting to replace constraint (11) by the stronger constraint

$$y \not\prec y^*,$$

yet in this case we could have disregarded a matrix $y' \in Y_{i_1 \succ i_2}$ such that $y'^{i_1 \| i_2} \in \mathcal{Y}$, and hence, we could have concluded falsely that $\lambda_{i_1} \succ \lambda_{i_2}$. Indeed, that $y \not\prec y^*$ tells us nothing about whether $y^{i_1 \| i_2} \in \mathcal{Y}$. For this reason, constraint (11) mentions that we must also check matrices y that *are* dominated by y^* whenever $y^{i_1 \| i_2}$ is not dominated by y^*.

Appendix D. Proof of Proposition 4

Proof. The inclusion $\mathcal{I}^w \subseteq \mathcal{I}$ follows from Proposition 3. To prove the reverse inclusion, consider any pair of labels $(i_1, i_2) \in \mathcal{I}$, and suppose that $(i_1, i_2) \notin \mathcal{I}^w$. Hence, there exists $y \in Y_{i_1 \succ i_2}$ such that

$$w(i_1, i_2) = \sum_{j \in K} c_{i_1 j}(y_{i_1 j} - y_{i_2 j}) + c_{i_2 j}(y_{i_2 j} - y_{i_1 j}) \geq 0. \tag{D.1}$$

Two cases may occur, depending on whether $y^{i_1 \| i_2} \not\prec y^*$ or $y \not\prec y^*$.

- If $y^{i_1 \| i_2} \not\prec y^*$, then $y^{i_1 \| i_2} \in \mathcal{Y}$ because when $\mathcal{C}$ is a singleton, $\succ$ is a complete pre-order, and any matrix is in $\mathcal{Y}$ as soon as it is not dominated by a single element of $\mathcal{Y}$. This leads to a contradiction because $y^{i_1 \| i_2} \in Y_{i_2 \succ i_1}$, and thus, $\mathcal{Y} \nsubseteq Y_{i_1 \succ i_2}$.
- If $y \not\prec y^*$, then constraint (D.1) implies that $y^{i_1 \| i_2} \not\prec y$, and thus, $y^{i_1 \| i_2} \in \mathcal{Y}$, which again leads to a contradiction.

□

Appendix E. Proof of Proposition 5

Proof. If we forget about constraint (11) that filters dominated solutions, then we get back ($WeakDom_{i_2}^{i_1}$) and can use the proof of Appendix B. However, we do not want to compute $z(i_1, i_2)$, but $w(i_1, i_2)$.

In fact not all solutions to (B.1)–(B.3) must be considered in the computation of $w(i_1, i_2)$. We need to check which of these enumerated solutions satisfy (11). From the definition of dominance, we must check that there exists $c \in \mathcal{C}$ such that $c^T(y - y^*) \leq 0$ or such that $c^T(y^{i_1 \| i_2} - y^*) \leq 0$, where c^T denotes the transpose of c. This amounts to show that $v(j_1, j_2) \leq 0$ or $v(j_2, j_1) \leq 0$ where $v(j', j'')$ is the optimal solution cost of the following optimization problem:

$$(K1) \qquad \begin{aligned} \min \quad & \sum_{i,j \in K} c_{ij}(y_{ij} - y^*_{ij}) \\ \text{s.t.} \quad & c \in \mathcal{C} \\ & y \in Y \\ & y_{i_1 j'} = y_{i_2 j''} = 1. \end{aligned}$$

We show next that the optimal solution cost of $(K1)$ is equal to the optimal solution cost of

$$(K2) \qquad \begin{array}{ll} \min & \sum_{i,j \in K} c^*_{ij} y_{ij} \\ \text{s.t.} & y \in Y \\ & y_{i_1 j'} = y_{i_2 j''}. \end{array}$$

Remark first that the set of binary matrices y feasible for $(K1)$ and $(K2)$ are equal. Let $\overline{y}$ be any binary vector feasible for $(K1)$, and for each $i \in K$, let $\overline{j}_i \in K$ be such that $\overline{y}_{i\overline{j}_i} = 1$ and $\overline{y}_{ij} = 0$ for $j \neq \overline{j}_i$. The optimal value of c associated to vector $\overline{y}$ is computed as

$$\begin{aligned} \min_{c \in \mathcal{C}} \sum_{i,j \in K} c_{ij}(\overline{y}_{ij} - y^*_{ij}) &= \min_{c \in \mathcal{C}} \sum_{i \in K} (c_{i\overline{j}_i} - c_{ij^*_i}) \\ &= \sum_{i \in K} \min_{c_i \in \mathcal{C}_i} (c_{i\overline{j}_i} - c_{ij^*_i}) \qquad \text{(E.1)} \\ &= \sum_{i \in K} c^*_{i\overline{j}_i}, \qquad \text{(E.2)} \end{aligned}$$

where (E.1) follows from the assumption $\mathcal{C} = \times_{i \in K} \mathcal{C}_i$ and (E.2) follows from the definition of c^*. Hence, choosing $y = \overline{y}$ in $(K2)$ provides a feasible solution for $(K2)$ with the same value as the solution of $(K1)$. Since the argument is valid for all y feasible for $(K1)$ and $(K2)$, the optimal solution costs of both problems are equal.

Hence, checking whether $(j_1, j_2) \in S(y^*, i_1, i_2)$ can be done by solving problem $(K2)$ where costs c^* have been computed in a pre-processing phase. □

References

[1] K. Bache and M. Lichman. UCI machine learning repository, 2013.

[2] P. Bartlett and M. Wegkamp. Classification with a reject option using a hinge loss. *The Journal of Machine Learning Research*, 9:1823–1840, 2008.

[3] A. Ben-Tal, L. E. Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton University Press, 2009.

[4] A. Bordes, X. Glorot, J. Weston, and Y. Bengio. Joint learning of words and meaning representations for open-text semantic parsing. *Journal of Machine Learning Research - Proceedings Track*, 22:127–135, 2012.

[5] W. Cheng, K. Dembczynski, and E. Hüllermeier. Label ranking methods based on the plackett-luce model. In *Proceedings of the 27th Annual International Conference on Machine Learning - ICML*, pages 215–222, 2010.

[6] W. Cheng, S. Henzgen, and E. Hüllermeier. Labelwise versus Pairwise Decomposition in Label Ranking. In *Workshop on Knowledge Discovery, Data Mining and Machine Learning*, 2013.

[7] W. Cheng, J. Hühn, and E. Hüllermeier. Decision tree and instance-based learning for label ranking. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML' 09*, 2009.

[8] W. Cheng and E. Hüllermeier. A Nearest Neighbor Approach to Label Ranking based on Generalized Labelwise Loss Minimization. In *M-PREF'13, 7th Multidisciplinary Workshop on Preference Handling*, Beijing, 2013.

[9] W. Cheng, E. Hüllermeier, W. Waegeman, and V. Welker. Label ranking with partial abstention based on thresholded probabilistic models. In *Advances in Neural Information Processing Systems 25 (NIPS-12)*, pages 2510–2518, 2012.

[10] W. Cheng, M. Rademaker, B. De Baets, and E. Hüllermeier. Predicting partial orders: ranking with abstention. *Machine Learning and Knowledge Discovery in Databases*, pages 215–230, 2010.

[11] G. Corani, A. Antonucci, and M. Zaffalon. Bayesian networks with imprecise probabilities: Theory and application to classification. *Data Mining: Foundations and Intelligent Paradigms*, pages 49–93, 2012.

[12] L. de Campos, J. Huete, and S. Moral. Probability intervals: a tool for uncertain reasoning. *Int. J. of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2:167–196, 1994.

[13] V. G. Deineko and G. J. Woeginger. On the robust assignment problem under a fixed number of cost scenarios. *Operations Research Letters*, 34(2):175 – 179, 2006.

[14] O. Dekel, C. D. Manning, and Y. Singer. Log-linear models for label ranking. In *Advances in Neural Information Processing Systems 16 (NIPS-03)*, 2003.

[15] S. Destercke. A pairwise label ranking method with imprecise scores and partial predictions. In *ECML/PKDD (2)*, pages 112–127, 2013.

[16] S. Destercke and B. Quost. Correcting binary imprecise classifiers: Local vs global approach. In *Scalable Uncertainty Management - 6th International Conference, Proceedings*, pages 299–310, 2012.

[17] X. Geng. Multilabel ranking with inconsistent rankers. In *Proceedings of Conference on Computer Vision and Pattern Recognition 2014*, 2014.

[18] S. Har-peled, D. Roth, and D. Zimak. Constraint classification : A new approach to multiclass classification and ranking. In *Advances in Neural Information Processing Systems 15 (NIPS-02)*, pages 785–792, 2002.

[19] E. Hüllermeier. Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *International Journal of Approximate Reasoning*, 55(7):1519 – 1534, 2014. Special issue: Harnessing the information contained in low-quality data sources.

[20] E. Hüllermeier and W. Cheng. Label Ranking Datasets. `http://www.uni-marburg.de/fb12/kebi/research/repository/labelrankingdata`, 2009. [Online; accessed 13-November-2014].

[21] E. Hüllermeier, J. Furnkranz, W. Cheng, and K. Brinker. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172:1897–1916, 2008.

[22] H. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quaterly*, 2(1-2):83–97, 1955.

[23] J. Marden. *Analyzing and modeling rank data*, volume 64. Chapman & Hall/CRC, 1996.

[24] M. Meila and H. Chen. Dirichlet process mixtures of generalized mallows models. In *Proceedings of the 26th Conference in Uncertainty in Artificial Intelligence*, pages 358–367, 2010.

[25] M. Troffaes. Decision making under uncertainty using imprecise probabilities. *Int. J. of Approximate Reasoning*, 45:17–29, 2007.

[26] N. Weskamp, E. Hullermeier, D. Kuhn, and G. Klebe. Multiple graph alignment for the structural analysis of protein active sites. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 4(2):310–320, 2007.

[27] M. Zaffalon. Exact credal treatment of missing data. *J. of Statistical Planning and Inference*, 105(1):105–122, 2002.

[28] M. Zaffalon. The naive credal classifier. *J. Probabilistic Planning and Inference*, 105:105–122, 2002.

[29] M. Zaffalon and E. Miranda. Conservative inference rule for uncertain reasoning under incompleteness. *Journal of Artificial Intelligence Research*, 34(2):757, 2009.