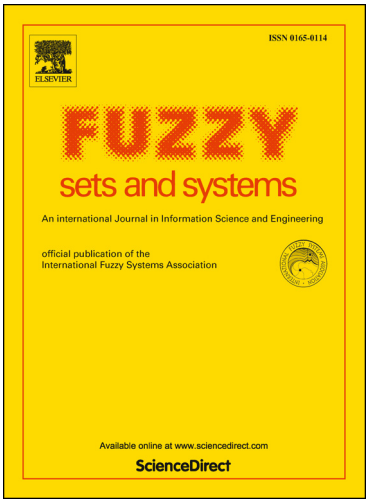
Inferring from an Imprecise Plackett–Luce model: application to label ranking

Loïc Adam, Arthur van Camp, Sébastien Destercke, Benjamin Quost

Journal Pre-proof



Inferring from an Imprecise Plackett–Luce model: application to label ranking

Loïc Adam, Arthur Van Camp, Sébastien Destercke and Benjamin Quost

PII: S0165-0114(24)00054-X
DOI: https://doi.org/10.1016/j.fss.2024.108908
Reference: FSS 108908

To appear in: *Fuzzy Sets and Systems*

Received date: 23 January 2023
Revised date: 22 December 2023
Accepted date: 7 February 2024

Please cite this article as: L. Adam, A. Van Camp, S. Destercke et al., Inferring from an Imprecise Plackett–Luce model: application to label ranking, *Fuzzy Sets and Systems*, 108908, doi: https://doi.org/10.1016/j.fss.2024.108908.

Inferring from an Imprecise Plackett–Luce model: application to label ranking

Loïc Adam[a], Arthur Van Camp[b], Sébastien Destercke[a], Benjamin Quost[a]

[a] *Université de Technologie de Compiègne, Heudiasyc, UMR 7253 CNRS, Compiègne, France*
[b] *University of Bristol, Department of Philosophy, Bristol, UK*

Abstract

The Plackett–Luce model is a popular parametric probabilistic model to define distributions between rankings of objects, modelling for instance observed preferences of users or ranked performances of algorithms. Since such observations may be scarce (users may provide partial preferences, or not all algorithms are run for a given experiment), it may be useful to consider the case where the parameters of the Plackett–Luce model are imprecisely known. In this paper, we first introduce the imprecise Plackett–Luce model, induced by a set of parameters (for instance, parameters with a high relative likelihood). Given a set of possible parameters for the model, we then provide an efficient algorithm to make cautious inferences, returning sets of possible optimal rankings (for instance in the form of partial orders). We illustrate the use of our imprecise model on label ranking, a specific kind of supervised learning.

Keywords: Preference learning, Cautious inference, Poor data, Imprecise Probability

1. Introduction

Learning and estimating probabilistic models over rankings of objects is an old problem, dating back to the 1920s [27]. In the last decades, this problem has known a revival, in particular due to a surge of interest from the machine learning community [14]. As the corresponding probabilities are defined over the space of permutations which grows exponentially with the number of objects, two classical approaches are either to split the initial problem in subcases (typically pairwise preferences [17]) or to use parametric models. In this latter case, two popular approaches consist either in associating a parametric random utility to each object and then considering the resulting distribution on rankings [4], or in directly defining a parametric distribution over the set of rankings [23].

There are multiple reasons to include cautiousness in both the estimation and inference steps of such models. The estimation may have to deal with scarce ranking information, such as in cold-start problems of recommender systems when predicting new user preferences [30], or with partial information, such as when one only observes top elements of a ranking or pairwise comparisons [21].

During the inference step, it may be useful to reinforce the reliability of the inferences made by outputting partial rankings as predictions, abstaining to predict when information is deemed unreliable. This could avoid recommending undesirable objects, or rejecting desirable ones, when only weak information is available, as well as allowing one to identify situations where obtaining more data or questioning the user may be instrumental.

When using precise probabilistic models, such abstentions are usually obtained by thresholding the estimated probabilities [9]. However, it can be argued that precise probabilities alone make it difficult to differentiate between ambiguous situations (*e.g.*, lots of observed preferences between two objects, half in favour of the first, half in favour of the second) and situations of lack of knowledge (*e.g*, no or very few observed preferences) [32]. This means in particular that approaches relying on precise probabilities may not be appropriate to deal with scarce data, due either to a lack of sensitivity or to being then strongly biased towards extreme values, thus lowering the interest of thresholding approaches. In contrast, relying on imprecise models to perform inferences makes it easier to reflect the lack of data by making the estimates more imprecise (and hence the predictions more partial) as data become scarcer. We will confirm this intuition in our experiments. Fundamental philosophical differences between precise and imprecise approaches to cautious inference lie behind this practical consideration: in the case of precise models, cautiousness is obtained through the decision/inference process, and is not reflected in the predictive model; whereas imprecise models encode a lack of knowledge in their structure during the estimation and learning steps, cautiousness merely being a consequence of the model encoding its limited state of knowledge. This argument, in addition to the aforementioned practical sensitivity to scarce data, supports the use of cautious approaches when handling scarce data.

It therefore makes sense to consider a theoretical framework that extends and enriches probabilities to better account for this distinction between ambiguity and ignorance. Imprecise probability theory [2], which models scarce knowledge by manipulating sets of probabilities, is an elegant mathematical framework that achieves this goal. However, to our knowledge, it has not yet been applied to the aforementioned approaches that are random utilities and parametric ranking models.

In this paper, we consider the latter, focusing more specifically on the well-known Plackett–Luce ranking model, which we present in Section 2. We focus on model inference in Section 3, showing that efficient methods can be developed to make cautious, guaranteed inferences based on sets of parameters. Section 4 then presents an application of the cautious Plackett–Luce model methods to label ranking, using relative likelihoods [5] to define the imprecise model via sets of parameters, similar to previous work [13]. Additionally, we provide in Appendix A some detailed proofs of two propositions introduced in Section 3, and in Appendix B we provide some complementary experimental results from Section 4. This work is an extension of a previously published work [1], and notably includes proofs as well as additional examples, a study of the case where parameters are interval-valued, and complementary experiments demonstrating

the usefulness of the proposed approach when compared to state-of-the-art thresholding approaches [9].

2. The imprecise Plackett–Luce model

We consider the problem of obtaining a probabilistic model over rankings of a finite set of objects or labels $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$. That is, we are interested in defining probabilities over *(strict) total orders on the labels—i.e.*, connective, transitive and irreflexive relations $\succ$ on Λ. We can (and will) identify any complete order $\succ$ over the labels—called *label ranking*—with its induced permutation $\tau : [1,n] \to [1,n]$ on indices $[1,n] := \{1, \ldots, n\}$, that is, the unique permutation of Λ such that

$$\lambda_{\tau(1)} \succ \lambda_{\tau(2)} \succ \cdots \succ \lambda_{\tau(n)}.$$

Because of this identification, in this paper, we will use the terms 'order on the labels', 'ranking' and 'permutation' interchangeably. We will denote the set which contains the $n!$ permutations on Λ by $\mathcal{L}$, a generic element of which will be denoted by τ.

We focus on one particular theoretical probability measure $P\colon 2^{\mathcal{L}} \to [0,1]$, namely the Plackett–Luce (PL) model [22, 26, 7, 15]. The PL model is parameterised by n parameters—called *strengths*—$v_1, \ldots, v_n$ in the set of (strictly) positive numbers $\mathbb{R}_{>0} := \{x \in \mathbb{R} : x > 0\}$.[1] We usually denote the *strength vector* $(v_1, \ldots, v_n)$ by v, which completely specifies the PL model. For any strength vector v, an arbitrary ranking τ in $\mathcal{L}$ is assigned probability:

$$\begin{aligned} P_v(\tau) &:= \prod_{k=1}^{n} \frac{v_{\tau(k)}}{\sum_{\ell=k}^{n} v_{\tau(\ell)}} \\ &= \frac{v_{\tau(1)}}{v_{\tau(1)} + \cdots + v_{\tau(n)}} \cdot \frac{v_{\tau(2)}}{v_{\tau(2)} + \cdots + v_{\tau(n)}} \cdots \frac{v_{\tau(n-1)}}{v_{\tau(n-1)} + v_{\tau(n)}}. \end{aligned} \tag{1}$$

The parameters $v_1, \ldots, v_n$ are defined up to a common positive multiplicative constant, so it is customary to assume that $\sum_{k=1}^{n} v_k = 1$. Therefore, the parameter $v = (v_1, \ldots, v_n)$ can be regarded as an element of the interior $\operatorname{int}(\Sigma) = \{(x_1, \ldots, x_n) \in \mathbb{R}^n_{>0} : \sum_{k=1}^{n} x_k = 1\}$ of the n-simplex $\Sigma := \{(x_1, \ldots, x_n) \in \mathbb{R}^n_{\geq 0} : \sum_{k=1}^{n} x_k = 1\}$.

The PL model has the following nice interpretation: the larger a weight v_i, the more a label λ_i tends to be preferred. This is reflected in the observation that the probability that label λ_i is the first ranked label is equal, for all $\tau \in \mathcal{L}$, to:

$$\sum_{\tau(1)=i} P_v(\tau) = v_i.$$

Given that λ_i is the first label, the probability that λ_j is the second label is equal to $v_j / \sum_{k=1, k\neq i}^{n} v_k$. This can be interpreted as the probability that λ_j is the first

[1] Next to $\mathbb{R}_{>0}$, we will also define the set of non-negative real numbers $\mathbb{R}_{\geq 0} := \{x \in \mathbb{R} : x \geq 0\}$.

amongst the remaining labels $\Lambda \setminus \{\lambda_i\}$. By recurrence, given that $\lambda_{\tau(1)}$ is the first label, $\lambda_{\tau(2)}$ the second, ..., $\lambda_{\tau(i-1)}$ the $i-1$-th one, the probability that $\lambda_{\tau(i)}$ is the i-th label is equal to $v_{\tau(i)}/\sum_{k=i}^{n} v_{\tau(k)}$, that is, the probability that $\lambda_{\tau(i)}$ is the first amongst the 'remaining' labels $\{\lambda_{\tau(i)}, \dots, \lambda_{\tau(n)}\}$.

For any PL model described by the strength vector v, finding the 'best' ranking—that is, the most probable (modal) ranking—is easy: it is sufficient to find the permutation τ that ranks the strengths in decreasing order. More specifically:

$$\tau \in \arg\max_{\tau' \in \mathcal{L}} P_v(\tau') \Leftrightarrow v_{\tau(1)} \geq v_{\tau(2)} \geq \cdots \geq v_{\tau(n-1)} \geq v_{\tau(n)}. \tag{2}$$

Example 1. *Consider the set $\Lambda = \{a, b, c\}$ of objects, together with the strengths $v_a = 0.3, v_b = 0.5, v_c = 0.2$. The most probable ranking is $b \succ a \succ c$ which has probability:*

$$P_v(b \succ a \succ c) = \frac{0.5}{0.5 + 0.3 + 0.2} \cdot \frac{0.3}{0.3 + 0.2} \cdot \frac{0.2}{0.2} = 0.3.$$

2.1. The imprecise Plackett–Luce model

We define an *imprecise* Plackett–Luce (IPL) model as the set of precise PL models obtained by letting the strengths vary over a subset $\Theta \subseteq \text{int}(\Sigma)$, rather than being precisely defined. It can be seen and interpreted as a robust, set-valued estimation of an unknown PL model, as Θ induces a corresponding set of precise PL models. We will assume that Θ is a subset of $\text{int}(\Sigma)$, rather than Σ, to ensure that all the strength values considered are positive, so that the PL model in Equation (1) is well-defined. A given ranking τ is now assigned several probabilities, each corresponding to one of the eligible precise PL models (or strength vectors). The *lower* and *upper probabilities* of a ranking τ are defined as:

$$\underline{P}_\Theta(\tau) \coloneqq \inf_{v \in \Theta} P_v(\tau) \quad \text{and} \quad \overline{P}_\Theta(\tau) \coloneqq \sup_{v \in \Theta} P_v(\tau) \quad \text{for all } \tau \text{ in } \mathcal{L},$$

and can be interpreted as bounds of a partially known PL model. A direct consequence is that the notion of 'best' or modal ranking is now ambiguous. Indeed, some ranking τ might maximise P_v for some strength vector v in Θ, while another ranking τ' maximises $P_u(\tau') > P_u(\tau)$ for another strength vector u in Θ, $u \neq v$. It results that classical decision rules and optimality conditions need to be redefined.

There are a number of imprecise-probabilistic optimality criteria. Since we are interested in returning cautious, set-valued predictions, we will consider here two of the most well-founded ones: *(Walley–Sen) maximality* [32, 28] and *E-admissibility* [20].

We call a ranking τ *maximal* if it is not dominated in the following order:

$$\tau_1 \succ_{\mathcal{M}} \tau_2 \Leftrightarrow (\forall v \in \Theta) P_v(\tau_1) > P_v(\tau_2). \tag{3}$$

This is indeed a 'robustification' of the precise rule, as $\tau_1 \succ_{\mathcal{M}} \tau_2$ only if $P_v(\tau_1) > P_v(\tau_2)$ is true for all possible models in Θ. If Θ contains more than one element,

then the ordering defined above can be a (strict) partial order—meaning that $\succ_{\mathcal{M}}$ is irreflexive, asymmetric and transitive—that might not be complete, and which might therefore admit more than one non-dominated element. The set of all maximal rankings—the rankings that are not dominated under $\succ_{\mathcal{M}}$, which we will denote further on by $\mathcal{M}_\Theta$—is therefore given by the set of rankings τ for which $\tau' \not\succ_{\mathcal{M}} \tau$ for all rankings τ':

$$\tau \in \mathcal{M}_\Theta \Leftrightarrow (\forall \tau' \in \mathcal{L}) \tau' \not\succ_{\mathcal{M}} \tau \tag{4}$$

$$\Leftrightarrow (\forall \tau' \in \mathcal{L})(\exists v \in \Theta) P_v(\tau) \geq P_v(\tau'). \tag{5}$$

A ranking τ is called *E-admissible* when there is a strength vector v for which it maximises P_v. In other words, the set of all E-admissible rankings, denoted further on by $\mathcal{E}_\Theta$, is given by the set of rankings τ for which:

$$(\exists v \in \Theta)(\forall \tau' \in \mathcal{L}) P_v(\tau) \geq P_v(\tau'). \tag{6}$$

Equivalently, the set of E-admissible rankings is given by:

$$\bigcup_{v \in \Theta} \arg\max_{\tau \in \mathcal{L}} P_v(\tau),$$

which corresponds to the union of all possible modal rankings. One can check the known fact [29] that any ranking that is E-admissible is also maximal, but not necessarily *vice versa*, by comparing Equations (4) and (6). The next example shows that, in our particular IPL setting, the two sets will not coincide in general.

Example 2. *Figure 1 displays the simplex representing the space of all possible parameters of a PL model for three objects, in barycentric coordinates. Each region is tagged by the corresponding optimal ranking, i.e., the most probable ranking whenever the strength vector lies in this region. This means that for a given set Θ of parameters, the set $\mathcal{E}_\Theta$ corresponds to the rankings whose region intersects with Θ. Any subset in this simplex can therefore be seen as a subset Θ introduced in this section.*

Now, consider the convex set Θ of parameters that is the interior of the convex hull of $v^1 = (1-\epsilon, 0, \epsilon)$ and $v^2 = (0, 0.5+\gamma, 0.5-\gamma)$ with $0.5 > \gamma > \epsilon > 0$, also represented in Figure 1 for the specific case $\epsilon = 0.25$ and $\gamma = 0.3$. That is, we look at all points $\alpha v^1 + (1-\alpha)v^2$, with $\alpha \in (0,1)$.

From the picture, one can see that the set of E-admissible rankings is:

$$\mathcal{E}_\Theta = \{\lambda_1 > \lambda_2 > \lambda_3, \lambda_1 > \lambda_3 > \lambda_2, \lambda_2 > \lambda_1 > \lambda_3, \lambda_2 > \lambda_3 > \lambda_1\}$$

as the full line crosses only the regions corresponding to those four rankings.

Besides, it turns out that $\lambda_3 > \lambda_1 > \lambda_2 \in \mathcal{M}_\Theta$: it can be checked that for each τ, we may find a suitable set of parameters $v \in \Theta$ such that $P_v(\lambda_3 > \lambda_1 > \lambda_2) \geq P_v(\tau)$, meaning that $\tau \not\succ_{\mathcal{M}} (\lambda_3 > \lambda_1 > \lambda_2)$. For instance, for $\tau = \lambda_1 > \lambda_2 > \lambda_3$, we must find a strength vector $v \in \Theta$ such that:

$$v_3 \cdot \frac{v_1}{v_1 + v_2} > v_1 \cdot \frac{v_2}{v_3 + v_2};$$

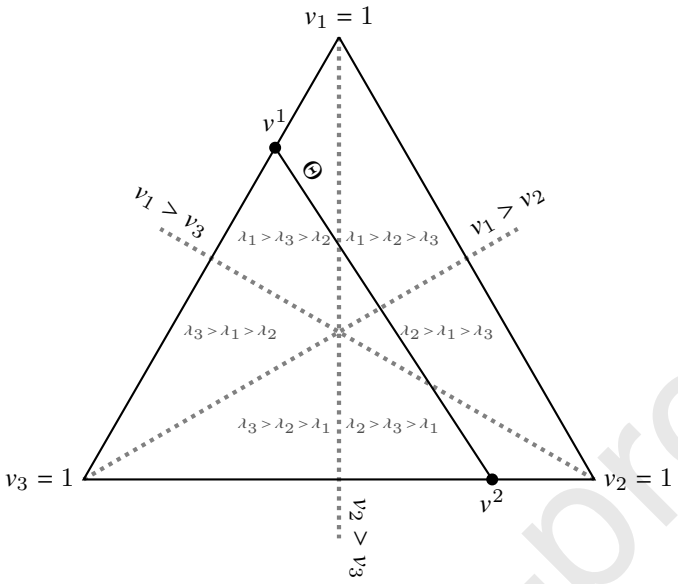


Figure 1: Simplex Σ with regions where rankings are optimal and E-admissible, and with parameter set Θ of Example 2

any vector with v_2 sufficiently close to 0 within Θ satisfies this inequality: one can therefore consider the point v^1. Other cases can be treated similarly, by picking adequate strength vectors within Θ.

3. Inference with IPL

We have seen in Section 2 that for a precise PL model, the 'best' (most probable) ranking can easily be found using Equation (2). Things become more complicated when the Plackett–Luce model becomes imprecise, since in this case, computing the set of all rankings satisfying Equation (4) to make robust and imprecise predictions generally requires comparing all pairs of possible answers.

This will be most of the time infeasible in practice, because the number of items to compare ($n!$) will quickly become huge as n grows: as a consequence, only problems with very few labels to rank will be tractable by sheer enumeration. Therefore, we need to find efficient ways to make predictions that remain coherent with imprecise probabilistic principles. Two different ways to do so is to consider approximate but guaranteed inferences in the general case, or to consider subcases (*i.e.*, domain restrictions) where making exact inferences become tractable.

In the following sections, we introduce two inference methods for the IPL model, one for each of these ideas. The first one, presented in Section 3.1, is an outer approximation to the set $\mathcal{M}_\Theta$ of (Walley–Sen) maximal rankings, and therefore also to the set $\mathcal{E}_\Theta$ of E-admissible ones. No further assumptions about

Θ need to be made. In Section 3.2, we introduce a second exact inference method where the set of strengths Θ has a specific form, namely that of *probability intervals* [12]. Such intervals can be obtained, *e.g.*, as lower/upper bounds resulting from projecting a generic set Θ on each strength value. We will introduce an efficient algorithm to compute the exact set $\mathcal{E}_\Theta$ of E-admissible rankings.

3.1. Outer approximation in the general case

We investigate here a criterion to decide whether a ranking is maximal. Rather than focusing on the whole ranking of objects, the idea in this section is to focus on individual pairs of objects: in this case, making inferences is easier and lead to outer approximations of $\mathcal{M}_\Theta$.

Inferring from Equation (3) and given two permutations τ and τ', we have:

$$\tau \succ_{\mathcal{M}} \tau' \Leftrightarrow (\forall v \in \Theta) \frac{P_v(\tau)}{P_v(\tau')} > 1. \tag{7}$$

Infer that in the expression for P_v in Equation (1), the numerator does not depend on τ, and hence we only have to deal with denominators in Equation (7).

Now, let us assume for a moment that the strengths are still precise, and consider τ and τ' such that $\tau(k) = \tau'(k)$ for all $k \in \{1,\dots,m\} \setminus \{i,j\}$ with $i \neq j$, and $\tau(j) = \tau'(i)$ and $\tau(i) = \tau'(j)$: the two rankings τ and τ' are equal, except for the positions i and j of two labels that are "swapped". We assume without loss of generality that $i < j$. This implies that $\sum_{\ell=k}^n v_{\tau(\ell)} = \sum_{\ell=k}^n v_{\tau'(\ell)}$ whenever k belongs to $\{1,\dots,n\} \setminus \{i+1,\dots,j\}$. Infer from Equation (7) that:

$$\begin{aligned}
\frac{P_v(\tau)}{P_v(\tau')} &= \prod_{k=1}^{n} \frac{\sum_{\ell=k}^n v_{\tau'(\ell)}}{\sum_{\ell=k}^n v_{\tau(\ell)}} \\
&= \underbrace{\prod_{k=1}^{i} \frac{\sum_{\ell=k}^n v_{\tau'(\ell)}}{\sum_{\ell=k}^n v_{\tau(\ell)}}}_{=1} \cdot \prod_{k=i+1}^{j} \frac{\sum_{\ell=k}^n v_{\tau'(\ell)}}{\sum_{\ell=k}^n v_{\tau(\ell)}} \cdot \underbrace{\prod_{k=j+1}^{n} \frac{\sum_{\ell=k}^n v_{\tau'(\ell)}}{\sum_{\ell=k}^n v_{\tau(\ell)}}}_{=1} \\
&= \prod_{k=i+1}^{j} \frac{v_{\tau'(j)} + \sum_{\ell=k,\ell\neq j}^n v_{\tau'(\ell)}}{v_{\tau(j)} + \sum_{\ell=k,\ell\neq j}^n v_{\tau(\ell)}} = \prod_{k=i+1}^{j} \frac{v_{\tau(i)} + \sum_{\ell=k,\ell\neq j}^n v_{\tau'(\ell)}}{v_{\tau(j)} + \sum_{\ell=k,\ell\neq j}^n v_{\tau(\ell)}}.
\end{aligned}$$

Consider for any k in $\{i+1,\dots,j\}$ the positive number $C_k \coloneqq \sum_{\ell=k,\ell\neq j}^n v_{\tau(\ell)} > 0$, then also $C_k = \sum_{\ell=k,\ell\neq j}^n v_{\tau'(\ell)}$ because $\tau'(\ell) = \tau(\ell)$ for any $\ell \neq i,j$, whence:

$$\frac{P_v(\tau)}{P_v(\tau')} = \prod_{k=i+1}^{j} \frac{v_{\tau(i)} + C_k}{v_{\tau(j)} + C_k}.$$

Since all the C_k are positive real numbers, this tells us that:

$$P_v(\tau) > P_v(\tau') \Leftrightarrow \frac{P_v(\tau)}{P_v(\tau')} > 1 \Leftrightarrow v_{\tau(i)} > v_{\tau(j)},$$

and therefore, for our specific rankings τ and τ':

$$\tau \succ_{\mathcal{M}} \tau' \Leftrightarrow (\forall v \in \Theta) v_{\tau(i)} > v_{\tau(j)}. \tag{8}$$

Determining whether the requirement in Equation (8) is fulfilled comes down to solving the optimisation problem

$$\inf_{v \in \Theta} (v_{\tau(i)} - v_{\tau(j)}) > 0. \tag{9}$$

This is simple in quite a number of cases: when Θ is a polytope defined by linear constraints, this can be done through standard linear programming; when Θ is a strict convex set and has an infinity of extreme points, one can resort to convex optimisation (*e.g.*, interior point methods) if needed. When Θ is characterised by a finite number of points (the extreme points of a polytope or points resulting from samplings), one can just apply the linear form (9) to every such point. Also, since (9) is linear, considering Θ or its convex hull would yield the same solution, thus making all previous approaches applicable to a set Θ of a general form.

Given an IPL model with strengths $\Theta \subseteq \text{int}(\Sigma)$, we can easily build a partial ordering outer-approximating $\mathcal{M}_\Theta$, in the sense that all rankings within $\mathcal{M}_\Theta$ are linear extensions of this partial order. Of course, this partial ordering may contain solutions that are not optimal under maximality, but we are sure that it will contain all optimal solutions, and it can be obtained easily. More formally, if we denote by $\lambda_k \succ_{\mathcal{P}} \lambda_\ell$ the fact that Equation (9) is satisfied, *i.e.*, $\inf_{v \in \Theta}(v_{\tau(k)} - v_{\tau(\ell)}) > 0$, then the set

$$\mathcal{P}_\Theta = \{\tau : \lambda_k \succ_{\mathcal{P}} \lambda_\ell \implies \tau(k) < \tau(\ell)\}$$

of permutations representable by the partial order $\succ_{\mathcal{P}}$ can be used as an outer approximation to the set of maximal linear orders, in the sense that $\mathcal{M}_\Theta \subseteq \mathcal{P}_\Theta$. The next example shows that this inclusion can be strict in some cases.

Example 3. *Let us consider the convex combination Θ between the two points $v^1 = (0.4, 0.2 - \epsilon, 0.4 + \epsilon)$ and $v^2 = (0.4, 0.4 + \epsilon, 0.2 - \epsilon)$, where $0 < \epsilon < 0.2$ (see Figure 2). One can check that $\mathcal{M}_\Theta = \{\lambda_1 > \lambda_2 > \lambda_3,\ \lambda_1 > \lambda_3 > \lambda_2,\ \lambda_2 > \lambda_1 > \lambda_3,\ \lambda_3 > \lambda_1 > \lambda_2\}$ by observing that $\mathcal{E}_\Theta$ is equal to this latter set and that we have, for instance, $\{\lambda_1 > \lambda_3 > \lambda_2\} \succ_{\mathcal{M}} \{\lambda_3 > \lambda_2 > \lambda_1\}$, as*

$$\begin{aligned}\inf_{v \in \Theta} \left(p_v(\{\lambda_1 > \lambda_3 > \lambda_2\}) - p_v(\{\lambda_3 > \lambda_2 > \lambda_1\})\right) &= \inf_{v \in \Theta} \left(v_1 \frac{v_3}{v_3 + v_2} - v_3 \frac{v_2}{v_1 + v_2}\right) \\ &= \inf_{v \in \Theta} \left(v_3 \frac{0.16 - 0.2 v_2}{0.6(0.4 + v_2)}\right)\end{aligned}$$

is positive, since both v_3 and $0.16 - 0.2v_2$ are always positive whatever the point chosen within Θ. However, one can easily check that $\inf_{v \in \Theta}(v_i - v_j) < 0$ for all pairs of i, j, therefore $\mathcal{P}_\Theta = \mathcal{L}$.

A simpler sufficient condition—which is not necessary—is that:

$$\underline{v}_{\tau(i)} \coloneqq \inf_{v \in \Theta} v_{\tau(i)} > \overline{v}_{\tau(j)} \coloneqq \sup_{v \in \Theta} v_{\tau(j)}.$$

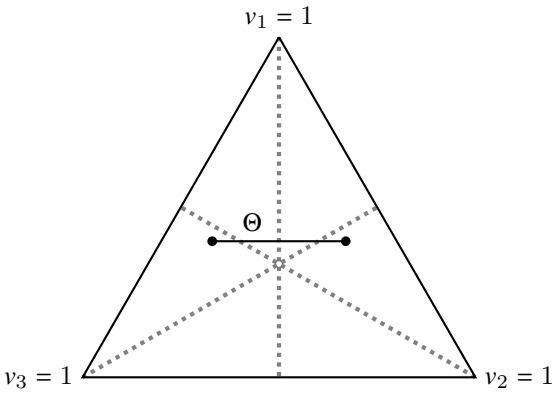


Figure 2: Parameter set Θ of Example 3

Indeed, this condition directly implies that if $\overline{v}_\ell < \underline{v}_k$ for two indices k and ℓ, then any ranking that prefers λ_ℓ over λ_k—in other words, any ranking τ for which $\tau(\ell) < \tau(k)$—will be dominated, according to Equation (3), by another ranking which only differs by the positions of λ_ℓ and λ_k. Among other things, this allows to conclude that if $\underline{v}_k > \overline{v}_\ell$, all maximally admissible rankings will be such that $\lambda_k \succ \lambda_\ell$. We can predict a partial ordering based on pairwise comparisons such that $\lambda_k \succ \lambda_\ell$ whenever $\underline{v}_k > \overline{v}_\ell$. This condition is weaker than $\inf_{v\in\Theta}(v_{\tau(k)} - v_{\tau(\ell)}) > 0$, because $\inf_{v\in\Theta} v_{\tau(k)} - \sup_{v\in\Theta} v_{\tau(\ell)} \geq \inf_{v\in\Theta}(v_{\tau(k)} - v_{\tau(\ell)})$. However, they are both equal when the set Θ is defined by intervals, a case that we explore in the next section and for which we give an efficient enumeration algorithm to get $\mathcal{E}_\Theta$.

3.2. Interval-valued case

In this section, we will make the simplifying assumption that the set of possible strengths is of the form:

$$\Theta = \left(\bigtimes_{k=1}^{n} [\underline{v}_k, \overline{v}_k]\right) \cap \mathrm{int}(\Sigma),$$

or in other words, that Θ is defined by the interval $[\underline{v}_k, \overline{v}_k] \subseteq (0,1)$ only, for each index k in $\{1,\ldots,n\}$. We believe such a restriction to be of particular practical interest, as it would be easy for a user to understand and interpret intervals of strength. Furthermore, we will see in this section that this restriction allows us to propose efficient inference algorithms.

We can interpret the possible strengths Θ as a subset of the simplex Σ, and therefore also as being equivalent to a set of probabilistic mass functions on $\{v_1,\ldots,v_n\}$. Since the possible strengths Θ are determined by the intervals $[\underline{v}_k, \overline{v}_k] \subseteq (0,1)$ for every index k in $\{1,\ldots,n\}$, it is formally equivalent to a so-called set of *probability intervals on singletons* [2, Section 4.4]. De Campos et

al. [12] showed that it is coherent—meaning the set Θ is non-empty, convex and tight (by which we mean that each pair of specified bounds $\underline{v}_k, \overline{v}_k$ is reachable by a point in Θ)—if and only if:

$$(\forall k \in \{1,\ldots,n\})\left(\underline{v}_k + \sum_{\substack{i=1\\ i\neq k}}^{n} \overline{v}_i \geq 1 \text{ and } \overline{v}_k + \sum_{\substack{i=1\\ i\neq k}}^{n} \underline{v}_i \leq 1\right). \tag{10}$$

We will assume in the following that Θ is a coherent set of possible strengths. It should however be noted that each point in the set Θ induces a corresponding probability over the space $\mathcal{L}$, in contrast with probability intervals that directly define a set of probabilities over the space Λ.

Remember that a given ranking τ is E-admissible if there is a parametrisation v in Θ such that τ maximises P_v. In this section, we are interested in the set of all E-admissible rankings $\bigcup_{v\in\Theta} \arg\max_{\tau\in\mathcal{L}} P_v(\tau)$.

3.2.1. Checking E-admissibility

We will provide here an efficient way to check whether a given ranking τ is E-admissible. Our argument hinges on the observation, in Equation (2), that for any v in $\text{int}(\Sigma)$ the ranking τ maximises P_v if and only if the values in v are ranked (in decreasing order) according to the indices in τ. In other words, a ranking τ maximises P_v if and only if $v_{\tau(1)}$ is the highest strength, $v_{\tau(2)}$ is the second highest of the strengths, $v_{\tau(3)}$ is the third-highest rank, and so on.

Proposition 1. *Consider any parametrisation* $\Theta = \left(\bigtimes_{k=1}^{n}[\underline{v}_k, \overline{v}_k]\right) \cap \text{int}(\Sigma)$ *of an imprecise Plackett–Luce model, and any ranking* τ *in* $\mathcal{L}$*. Then* τ *is E-admissible—in other words,* $\tau \in \bigcup_{v\in\Theta} \arg\max_{\tau'\in\mathcal{L}} P_v(\tau')$*—if and only if there is a* k *in* $\{1,\ldots,n\}$ *such that:*

$$1 - \sum_{\ell=1}^{k-1} \min\{\overline{v}_{\tau(1)},\ldots,\overline{v}_{\tau(\ell)}\} - \sum_{\ell=k+1}^{n} \max\{\underline{v}_{\tau(\ell)},\ldots,\underline{v}_{\tau(n)}\} \in [\max\{\underline{v}_{\tau(k)},\ldots,\underline{v}_{\tau(n)}\}, \min\{\overline{v}_{\tau(1)},\ldots,\overline{v}_{\tau(k)}\}] \tag{11}$$

and

$$\underline{v}_{\tau(\ell)} \leq \min\{\overline{v}_{\tau(1)},\ldots,\overline{v}_{\tau(\ell)}\} \text{ for all } \ell \text{ in } \{1,\ldots,k-1\}, \text{ and } \overline{v}_{\tau(\ell)} \geq \max\{\underline{v}_{\tau(\ell)},\ldots,\underline{v}_{\tau(n)}\} \text{ for all } \ell \text{ in } \{k+1,\ldots,n\}. \tag{12}$$

The proof of Proposition 1 can be found in Appendix A. This proof shows that a possible solution of strength vectors being ordered as for τ is to let $v_{\tau(\ell)} := \min\{\overline{v}_{\tau(1)},\ldots,\overline{v}_{\tau(\ell)}\}$ for any ℓ in $\{1,\ldots,k-1\}$, $v_{\tau(\ell)} := \max\{\underline{v}_{\tau(\ell)},\ldots,\underline{v}_{\tau(n)}\}$ for any ℓ in $\{k+1,\ldots,n\}$, and also $v_{\tau(k)} := 1 - \sum_{\ell=1,\ell\neq k}^{n} v_{\tau(\ell)}$. Equation (12) actually ensures that for ℓ in $\{1,\ldots,k-1\}$ and ℓ in $\{k+1,\ldots,n\}$, such an assignment is within the intervals $[\underline{v}_{\tau(\ell)}, \overline{v}_{\tau(\ell)}]$, and Equation (11) ensures that

$v_{\tau(k)} \in [\underline{v}_{\tau(k)}, \overline{v}_{\tau(k)}]$, making sure that this assignment satisfies our interval constraints.

The condition in Proposition 1 has a polynomial complexity in the number n of labels. Indeed, we need to check n different values of k, and for each value k, we need by Equation (11) to calculate a sum of $n-1$ terms, and to check by Equation (12) $n-1$ inequalities, which yields a complexity of $n(2n-2)$. This can even be slightly reduced when some intervals in Equation (11) are empty, as for those values k where it happens, Equation (11) is trivially not satisfied, and we can avoid performing the summations and inequality checks.

3.2.2. Computing and enumerating all E-admissible rankings

Equation (11) offers a very quick way to check whether a given ranking is E-admissible, therefore allowing one to easily build an approximation of $\mathcal{E}_\Theta$ for instance through sampling. However, applying Equation (11) directly to obtain the exact $\mathcal{E}_\Theta$ is clearly not efficient enough. The main bottleneck is that it requires us to check E-admissibility for each individual ranking separately. Since there are $n!$ many such rankings, this quickly becomes intractable. In order to avoid this exponential blow-up, we will now develop an algorithm that is able to rule out the E-admissibility of many rankings at once, without having to explicitly check the E-admissibility of each of them individually.

Ruling out multiple rankings at once. The central idea of our algorithm is to use a search tree in order to navigate the set of all rankings $\mathcal{L}$, which makes it possible to determine whether a set of rankings is worth being further investigated. Each node in the tree corresponds to a sequence of labels at the beginning of a set of rankings; exploring further the branch consists in adding additional labels to the sequence (and thus restricting the corresponding set of rankings). If we are able to infer that there is no E-admissible ranking τ which contains a given sequence of labels, then we can completely ignore all rankings starting with this sequence. In Example 4 and Figure 4, we provide an example with $n = 4$ labels.

Consider any coherent parametrisation Θ determined by the probability intervals $[\underline{v}_k, \overline{v}_k]$ for all k in $\{1, \ldots, n\}$. Let $(\tau(1), \ldots, \tau(j)) = (k_1, \ldots, k_j)$ be an initial sequence of labels, with k_1, k_2, $\ldots$ k_j being distinct elements of $\{1, \ldots, n\}$. We want to infer whether there exists a ranking τ with the initial sequence $(k_1, \ldots, k_j)$ which is E-admissible with respect to Θ. To this end, let us introduce the following three equations:

$$\sum_{\ell=1}^{j} \min\{\overline{v}_{k_1}, \ldots, \overline{v}_{k_\ell}\} + \sum_{\substack{i=1 \\ i \notin \{k_1, \ldots, k_j\}}}^{n} \min\{\overline{v}_{k_1}, \ldots, \overline{v}_{k_j}, \overline{v}_i\} \geq 1; \qquad (A_j)$$

$$\overline{v}_{k_j} \geq \max\{\underline{v}_i : i \in \{1, \ldots, n\} \setminus \{k_1, \ldots, k_j\}\}; \qquad (B_j)$$

$$\max\{\underline{v}_i : i \in \{1,\dots,n\}\} + \max\{\underline{v}_i : i \in \{1,\dots,n\} \setminus \{k_1\}\} + \dots$$
$$+ \max\{\underline{v}_i : i \in \{1,\dots,n\} \setminus \{k_1,\dots,k_{j-1}\}\} + \sum_{\substack{i=1 \\ i \notin \{k_1,\dots,k_j\}}}^{n} \underline{v}_i \leq 1. \quad (C_j)$$

In the special case where $j = 1$—that is, we want to know whether a ranking starting with a single given element k_1 is E-admissible—the three Equations (A_j), (B_j) and (C_j) reduce to:

$$\sum_{i=1}^{n} \min\{\overline{v}_{k_1}, \overline{v}_i\} \geq 1; \quad (A_1)$$

$$\overline{v}_{k_1} \geq \max\{\underline{v}_i : i \in \{1,\dots,n\}\}; \quad (B_1)$$

$$\max\{\underline{v}_i : i \in \{1,\dots,n\}\} + \sum_{\substack{i=1 \\ i \neq k_1}}^{n} \underline{v}_i \leq 1. \quad (C_1)$$

Note that under the coherence requirement (10), Equation (C_1) is a direct consequence of Equation (B_1), but for $j \geq 2$ Equation (C_j) can no longer be deduced from the other equations.

Proposition 2. *Consider any coherent parametrisation Θ determined by a set of probability intervals $[\underline{v}_k, \overline{v}_k]$ for all k in $\{1,\dots,n\}$, and any initial segment $(\tau(1),\dots,\tau(m)) = (k_1,\dots,k_m)$ of length $m \in \{1,\dots,n-1\}$. Then, there exists an E-admissible ranking with initial segment $(k_1,\dots,k_m)$ if and only if the Equations (A_j), (B_j) and (C_j) are fulfilled for every j in $\{1,\dots,m\}$.*

The proof of Proposition 2 can be found in Appendix A. Let us now introduce an example illustrating Proposition 2, as well as the tree resulting from applying Algorithms 1 and 2 (introduced after this example), which simply check recursively whether Equations (A_j), (B_j) and (C_j) are fulfilled in a given branch in order to prolong it.

Example 4. *Let us consider a case where we have $n = 4$ labels, and where our set of possible parameters is given by the intervals $[\underline{v}_1, \overline{v}_1] = [3/8, 5/8]$, $[\underline{v}_2, \overline{v}_2] = [1/12, 1/12]$, $[\underline{v}_3, \overline{v}_3] = [1/30, 1/5]$ and $[\underline{v}_4, \overline{v}_4] = [1/8, 3/8]$ (which is easily verified to be coherent using Equation (10)). See Figure 3 for a visualisation of the intervals.*

A possible strength vector $v \in \Theta$, for which $\tau = (1,3,4,2)$ is the most probable ranking, is given by $(v_1, v_2, v_3, v_4) = (5/8, 1/12, 1/6, 1/8)$: we check easily that v belongs to Θ and that $v_{\tau(1)} = 5/8 \geq v_{\tau(2)} = 1/6 \geq v_{\tau(3)} = 1/8 \geq v_{\tau(4)} = 1/12$, so that Equation (2) guarantees that τ is indeed E-admissible. Another way to check it, as will be developed below in Algorithms 1-2, is to check that Equations (A_j), (B_j) and (C_j) are satisfied for the growing sequences $(1),(1,3),(1,3,4)$ and $(1,3,4,2)$. This is why the branch d_1, d_3, d_4, d_2 is fully developed to a depth of $n = 4$ in the tree represented by Figure 4. To give an overview, Table 1 displays strength vectors in Θ which yield as model rankings the different rankings τ corresponding to the leaves in Figure 4. This implies that all the rankings indicated in Figure 4

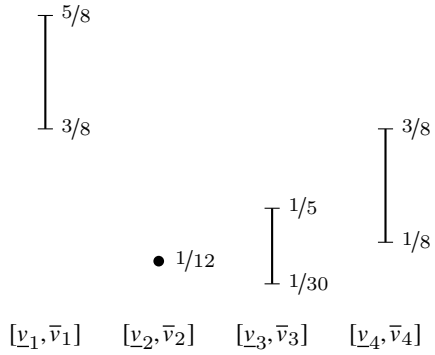


Figure 3: Probability intervals for Example 4

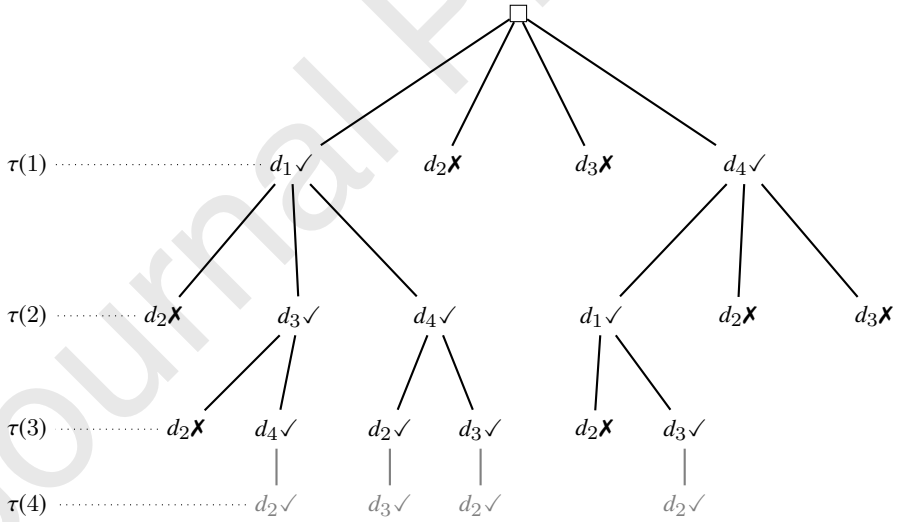


Figure 4: Search tree for $n = 4$, issued from Example 4

Table 1: Possible parameter values giving modal rankings of Example 4

$\nu = (\nu_1, \nu_2, \nu_3, \nu_4)$	$\tau = (\tau(1), \tau(2), \tau(3), \tau(4)) \in \arg\max_{\tau' \in \mathcal{L}} P_\nu(\tau')$
$(5/8, 1/12, 1/6, 1/8)$	$(1, 3, 4, 2)$
$(5/8, 1/12, 1/12, 5/24)$	$(1, 4, 2, 3)$
$(5/8, 1/12, 1/12, 5/24)$	$(1, 4, 3, 2)$
$(3/8, 1/12, 1/6, 3/8)$	$(4, 1, 3, 2)$

are indeed E-admissible with respect to Θ.

We can also show and check that every branch of the tree in Figure 4 that stops before reaching a depth of $n = 4$ *corresponds to a starting sequence whose completion cannot be an E-admissible ranking* τ. *Take for instance the sequence starting with* $(1, 2)$, *and assume* ex absurdo *that there would be such an E-admissible ranking* τ. *This would imply that there is a strength vector* ν *in* Θ *such that* $\nu_1 \geq \nu_2 \geq \max\{\nu_3, \nu_4\}$, *which by Equation* (2) *would imply that* $1/12 = \nu_2 \geq \nu_4 \geq \underline{\nu}_4 = 1/8$, *an impossibility.*

In practice, this impossibility can be checked by verifying that Equation (B_j) *is not satisfied for* $k_1 = 1, k_2 = 2$ *as indeed* $1/12 = \overline{\nu}_2 < \max\{\underline{\nu}_3, \underline{\nu}_4\} = 1/8$. *In essence, Equations* (A_j), (B_j) *and* (C_j) *allow one to check whether a given sequence* $\{k_1, \ldots, k_m\}$ *can or cannot be continued into an E-admissible ranking, and provides a set of mechanisms at the basis of the recursive algorithms 1 and 2 given below.* ◊

Algorithm. We propose an efficient algorithm based on Equations (A_j), (B_j) and (C_j) used in Proposition 2 to check whether there is an E-admissible ranking with a given initial segment. More precisely, the algorithm consists in using these equations recursively: to check whether there is an E-admissible ranking starting with $(k_1, \ldots, k_m)$ it suffices to check whether there is an E-admissible ranking starting with $(k_1, \ldots, k_{m-1})$ and whether the Equations (A_j), (B_j) and (C_j) hold for $j = m$.

Algorithms 1 and 2 provide pseudocodes describing a recursive method to find all E-admissible rankings given an interval-valued set Θ. Note that due to the pruning strategy, the algorithm is polynomial in the number of E-admissible options (hence finding one E-admissible option is fast), however this number may still be $|\Lambda|!$ in the worst case, and we may need to count that many rankings.

Algorithm 1 Find the E-admissible rankings opt_n

Require: probability intervals $[\underline{\nu}_k, \overline{\nu}_k]$ for k in $\{1, \ldots, n\}$
Ensure: $\{[\underline{\nu}_k, \overline{\nu}_k] : k \in \{1, \ldots, n\}\}$ coherent
  $\text{opt}_n \leftarrow \emptyset$
  for all $k_1 \in \{1, \ldots, n\}$ **do**
    `Recur`$(1, (k_1))$
  end for

Algorithm 2 `Recur`$(j,(k_1,\ldots,k_j))$

```
if j = n − 1 then
  append the unique k_n ∈ {1,...,n} \ {k_1,...,k_{n−1}} to the end of (k_1,...,k_{n−1})
  add (k_1,...,k_n) to opt_n                          ▷we found a solution.
else
  for all k_{j+1} ∈ {1,...,n} \ {k_1,...,k_j} do
    if Equations (A_{j+1}), (B_{j+1}) and (C_{j+1}) hold then
      append k_{j+1} to the end of (k_1,...,k_j)
      Recur(j + 1,(k_1,...,k_{j+1}))
    end if
  end for
end if
```

4. Application to label ranking

The previous sections have explored how cautious robust inference can be made when we only have imprecise knowledge about the parameters of a Plackett–Luce model. This section presents a possible use of our approach in a supervised machine learning problem, and discusses some possible ways to estimate the set of parameters from data.

Whereas supervised classification consists in mapping instances $\mathbf{x}$ issued from an instance space $\mathcal{X}$ to single (preferred) labels of the space $\Lambda = \{\lambda_1,\ldots,\lambda_n\}$ of possible classes, we address here a more complex issue called label ranking, where we want to map any instance $\mathbf{x} \in \mathcal{X}$ to a *total order on the labels* $\succ_{\mathbf{x}}$ on Λ.

The task in label ranking is the same as in usual classification, *i.e.* using a set of training instances $(\mathbf{x}^i,\tau^i)$, $i \in \{1,\ldots,m\}$ to estimate the theoretical conditional probability measure $P_{\mathbf{x}}\colon 2^{\mathcal{L}} \to [0,1]$ associated to an instance $\mathbf{x} \in \mathcal{X}$. Ideally, observed outputs τ^i should be complete orders over Λ; this is however seldom the case. In order to prepare for this, we sometimes allow training instances τ^i to be incomplete (*i.e.*, partial orders over Λ).

In this case, we may apply the approach presented in Section 3.1 in order to infer an IPL model from such partial data. We will use the contour likelihood to get the parameter set corresponding to a specific instance $\mathbf{x}$, since efficient maximum likelihood estimation (MLE) methods can be used to infer a PL model. For justifications on the use of the contour likelihood to obtain sets of parameters as estimates, we refer for example to [31, 25, 6, 5].

4.1. Estimation method

We will now describe our estimation method in different steps, in order to obtain a parameter set Θ from observed data. Assume that we have observed a sample of K rankings $\mathcal{T} = \{\tau^1,\ldots,\tau^K\}$, with M_i the number of ranked labels in

τ^i. Given a strength vector v, the probability to observe $\mathcal{T}$ is:

$$P(\mathcal{T}|v) = \prod_{i=1}^{K} \prod_{m=1}^{M_i} \frac{v_{\tau^i(m)}}{\sum_{j=m}^{M_i} v_{\tau^i(j)}}. \tag{13}$$

4.1.1. Maximum likelihood estimation

Finding the Maximum Likelihood Estimation (MLE) of v comes down to maximizing Equation (13), or equivalently to doing the same with the log-likelihood:

$$Logl(v) = \sum_{i=1}^{K} \sum_{m=1}^{M_i} \left[\log(v_{\tau^i(m)}) - \log \sum_{j=m}^{M_i} v_{\tau^i(j)} \right]. \tag{14}$$

Unfortunately, no analytical solution to finding the MLE parameters of the PL model exists. Nevertheless, multiple efficient optimisation methods have been proposed in the literature. One of them, which we will use here, is the Minorisation-Maximisation (MM) algorithm by [18]. It is a generalisation of the Expectation–Maximisation (EM) algorithm. The MM algorithm is an iterative procedure which aims to maximise in each iteration a lower bound for the log-likelihood:

$$Q_k(v) = \sum_{i=1}^{K} \sum_{m=1}^{M_i} \left[\log(v_{\tau^i(m)}) - \frac{\log \sum_{j=m}^{M_i} v_{\tau^i(j)}}{\log \sum_{j=m}^{M_i} v^{(k)}_{\tau^i(j)}} \right], \tag{15}$$

where $v^{(k)}$ is the estimation of v during the k-th iteration. When the parameters are fixed, the maximisation of Q_k can be solved analytically and the algorithm provably converges to the MLE estimate v^* of v.

4.1.2. Set estimation via the contour likelihood

Given parameter values[2] $v \in \text{int}(\Sigma)$ and the likelihood function $l(v)$, the contour likelihood is:

$$l^*(v) = \frac{l(v)}{\max_{u \in \Sigma} l(u)} = \frac{l(v)}{l(v^*)}. \tag{16}$$

By construction, $l^*(v)$ take values in $]0,1]$. The closer l^* is to 1, the closer v is to a maximum of the likelihood function.

We can therefore naturally obtain imprecise estimates by considering the regions of the parameter space obtained by "cutting" the contour likelihood. Given β in $[0,1]$, the β-cut of the contour likelihood, written B^*_β, is defined by

$$B^*_\beta = \{v \in \Sigma : l^*(v) \geq \beta\}.$$

We stress here that the choice of β directly influences the precision (and thus the robustness) of the model: starting with $B^*_1 = v^*$, which generally leads to

[2] As before, we use the interior $\text{int}(\Sigma)$ of Σ to ensure that $\log \sum_{j=m}^{M_i} v_{\tau^i(j)}$ is well-defined.

a precise PL model, the IPL model then becomes less and less precise with decreasing β, possibly leading to partial (and even empty) predictions. The choice of β is thus directly linked to how imprecise we want our predictions to be. The interest of using β is that it allows us to control the precision/accuracy trade-off with a single parameter. Choosing the right value for this parameter therefore depends on how much precision an end-user or decision maker is willing to trade to obtain more robust/accurate predictions. As in other imprecise probabilistic classifiers [11], β can also be used as a way to "measure" how robust a given precise prediction is: if we need to decrease β a lot to make the maximum likelihood prediction imprecise, then this means the initial prediction was rather robust, else this may mean that the precise prediction relies on rather weak information.

4.1.3. Imprecise predictions

Once B^*_β is determined, for any test instance $\mathbf{x}$ to be processed, we can easily obtain an imprecise prediction $\hat{\tau}$ in the form of a partial ranking using the results of Section 3.1: we will retrieve $\hat{\tau}$ such that $\lambda_i \succ \lambda_j$ for all $v_k \in B^*_\beta$.

Example 5. *Let us assume that we want to determine the ranking τ of an instance $\mathbf{x}$ through a learning process, i.e. we predict the ranking of the instance $\mathbf{x}$ with the rankings of some other instances. To do so, we pick the five closest neighbours of $\mathbf{x}$ according to a distance (for example the Euclidean distance), as a classical scheme to get a local model estimation. Three of these neighbours have the associated ranking $(\lambda_2, \lambda_1, \lambda_3)$ and two have the associated ranking $(\lambda_1, \lambda_3, \lambda_2)$. Based on these neighbours, the ranking τ predicted by maximum likelihood is $(\lambda_1, \lambda_2, \lambda_3)$. Figure 5 displays the corresponding contour likelihood function, modelled using 20,000 randomly generated strengths v_k according to a Dirichlet distribution with $\alpha = 5v_{opt}$, with v_{opt} being the strength of the optimal Plackett–Luce model. Note that only v_1 and v_2 are represented in the Figure, since $v_3 = 1 - v_1 - v_2$, meaning we have only two degrees of freedom and that all strength vectors can be represented on a plane.*

The contour likelihood function takes values between 0 and 1, and its value decreases when the generated strengths v_k are far from the optimal strength v_{opt}. Moreover, it is possible to directly interpret the preferences between objects in Figure 5. Each median line corresponds to a situation where an object is equally preferred to another one. For example, $v_1 = v_2$ indicates that λ_1 and λ_2 are equally preferred. The intersection of the medians corresponds to the situation $v = [1/3, 1/3, 1/3]$, where all objects are equally preferred. In such a situation, all rankings are equally probable.

We can make an imprecise prediction on the ranking τ by "cutting" the contour likelihood function, ending up with a beta-cut B^*_β. In this example, we first take $\beta = 0.9$, giving a rather precise prediction, to the detriment of robustness. As in the precise case, we obtain $\tau = (\lambda_1, \lambda_2, \lambda_3)$, as observed from Figure 6: all the generated strengths v_k such that $L^*_k \geq 0.9$ stay in the same area delimited by the three median lines. The binary relations $\lambda_1 \succ \lambda_2$, $\lambda_2 \succ \lambda_3$ and

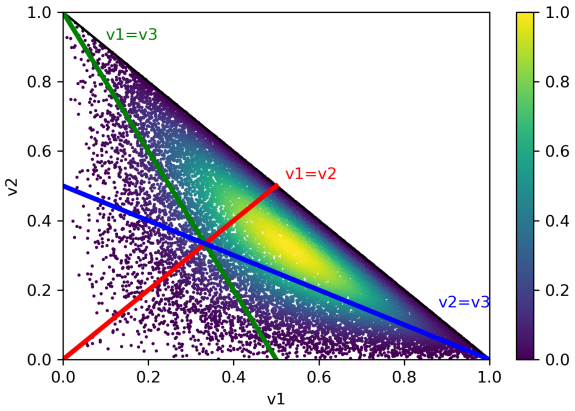


Figure 5: Full contour likelihood function

$\lambda_1 \succ \lambda_3$ *(that follows from the two previous ones) give the same final ranking as the precise approach.*

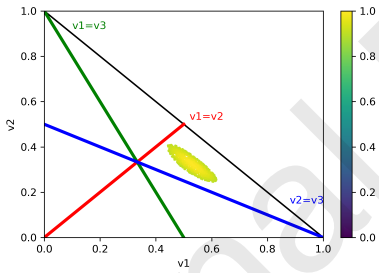


Figure 6: Beta-cut $B^*_{0.9}$

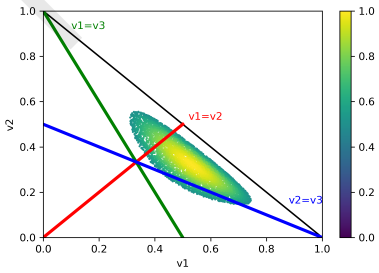


Figure 7: Beta-cut $B^*_{0.5}$

Using a new beta-cut B^*_β *with a coefficient* $\beta = 0.5$*, we obtain a different prediction. We observe from Figure 7 that a majority of generated strengths stay in the same delimited area, yet some generated strengths are outside this area, changing the predicted ranking or order: over the median* $v_1 = v_2$*, some generated strengths indicate that we could have* $\lambda_2 \succ \lambda_1$*; and under the median* $v_2 = v_3$*, some generated strengths indicate that we could have* $\lambda_3 \succ \lambda_2$*. In our approach, a binary relation* $\lambda_i \succ_{\mathrm{x}} \lambda_j$ *between two objects is kept if it is common to all generated strengths. In our case, this means that the prediction is reduced to* $\lambda_1 \succ \lambda_3$*, as* $v_1 = v_3$ *is the only median which does not intersect with* $B^*_{0.5}$*.*

4.2. Experimental setting

4.2.1. Likelihood approximation

In order to obtain the observations from which the contour likelihood is computed via (16), we consider here the method proposed by [8]. The approach is instance-based: for any $\mathbf{x} \in \mathcal{X}$, the predictions are made locally using its nearest neighbours. Let $\mathcal{N}_K(\mathbf{x})$ stand for the set of nearest neighbours of $\mathbf{x}$ in the training set, each neighbour $\mathbf{x}^i \in \mathcal{N}_K(\mathbf{x})$ being associated with a (possibly incomplete) ranking τ^i.

We model the contour likelihood by generating multiple strengths v according to a Dirichlet distribution with parameter $\beta = \gamma v^*$, where v^* is the MLE obtained with the best PL model (or equivalently, the best strength v) and $\gamma > 0$ is a coefficient which makes it possible to control the concentration of parameters generated around v^*.

4.2.2. Evaluation

When the observed and predicted rankings τ and $\hat{\tau}$ are complete, various accuracy measures [17] have been proposed to measure how close they are to each other (0/1 accuracy, Spearman's rank, etc.). Here, we retain Kendall's Tau:

$$A(\tau, \hat{\tau}) = \frac{C - D}{n(n-1)/2}, \tag{17}$$

where C and D are respectively the numbers of concordant and discordant pairs in τ and $\hat{\tau}$. In the case of imprecise predictions $\hat{\tau}$, the usual quality measures can be decomposed into two components [10]: correctness (CR), measuring the accuracy of the predicted comparisons, and completeness (CP):

$$CR(\tau, \hat{\tau}) = \frac{C - D}{C + D} \quad \text{and} \quad CP(\tau, \hat{\tau}) = \frac{C + D}{n(n-1)/2}, \tag{18}$$

where C and D are the same as in Equation (17). Should $\hat{\tau}$ be complete, $C + D = n(n-1)/2$, $CR(\tau, \hat{\tau}) = A(\tau, \hat{\tau})$ and $CP(\tau, \hat{\tau}) = 1$; while $CR(\tau, \hat{\tau}) = 1$ and $CP(\tau, \hat{\tau}) = 0$ if $\hat{\tau}$ is empty (since no comparison is done). Let us note a partial ranking has usually a higher correctness than its complete equivalent, suggesting that a partial ranking may be desirable if we want to avoid incorrectly ranked labels.

Example 6. *Let us suppose we want to estimate the ranking $\tau = (\lambda_2, \lambda_3, \lambda_1)$. We predict two rankings: a complete ranking $\hat{\tau}_1 = (\lambda_3, \lambda_2, \lambda_1)$ and a partial ranking $\hat{\tau}_2 = (\lambda_3, \lambda_1)$. We have $n(n-1)/2 = 3$ and the number of concordant and discordant pairs are $C_1 = 2$ and $D_1 = 1$ for $\hat{\tau}_1$, as we correctly predicted that $\lambda_3 \succ \lambda_1$ and $\lambda_2 \succ \lambda_1$, but also incorrectly predicted that $\lambda_3 \succ \lambda_2$; and $C_2 = 1$ and $D_2 = 0$ for $\hat{\tau}_2$, since $\lambda_3 \succ \lambda_1$ is correctly predicted, and we did not rank λ_2.*

We can now determine the correctness and completeness of each predicted ranking. We have $CR(\tau, \hat{\tau}_1) = 2-1/2+1 = 2/3$ and $CP(\tau, \hat{\tau}_1) = 2+1/3 = 1$, while $CR(\tau, \hat{\tau}_2) = 1-0/1+0 = 1$ and $CP(\tau, \hat{\tau}_2) = 1+0/3 = 1/3$: the ranking $\hat{\tau}_1$ is complete but partially incorrect, while the ranking $\hat{\tau}_2$ is fully correct (no label is incorrectly ranked) but does not rank all labels.

4.2.3. Thresholding

In the experiments, we compare our imprecise approach based on parameter sets to the abstention scheme proposed by [9]. Given a precise PL model with strength vector v, this latter approach uses the probability $P(\lambda_i \succ \lambda_j)$ of choosing the label λ_i over the label λ_j, given by:

$$P(\lambda_i \succ \lambda_j) = \frac{v_i}{v_i + v_j}, \tag{19}$$

indicating that $\lambda_i \succ \lambda_j$ only if $P(\lambda_i \succ \lambda_j) \geq \alpha$, with $\alpha \in [0.5, 1]$. For $\alpha = 0.5$, the prediction is simply the ordering induced by v, and for $\alpha = 1$, we retrieve the empty order. It has been proven in [9] that considering all values in-between provides a set of partial orders, *i.e.*, a set of partial predictions whose imprecision grows with α.

4.3. Experimental results

In the experiments[3], we use various datasets in order to compare our approach with that of [9]. They were adapted from classical datasets in [8], except for the SUSHI dataset, a standard in preference learning, in which the complete rankings over 10 types of sushi expressed by 5000 customers are recorded[4]. The datasets and their properties are quickly presented in Table 2, while more details on how these datasets were generated can be found in [8]. The number of attributes is only relevant to determine the closest neighbours of each instance, while the number of labels to rank is the heart of our problem: the more labels we have to rank, the more difficult the problem is, as we have to estimate the likelihood function in a higher-dimensional space. According to [8], the type of the dataset influences the difficulty of the prediction problem: in general, the correctness should be overall higher for datasets coming from classification problems. Nevertheless, we did not notice any additional difference on the ranking problem with our contour likelihood approach.

In order to limit the size of this section to a reasonable level, we only focus on a few datasets that are representative of all our experimental results, in the sense that results for other datasets follow the same trends. Experimental results on the other datasets can be found in Appendix B.

4.3.1. Comparison

Here, we compare our approach based on the contour likelihood function with the abstention approach existing in the literature, using the instance-based algorithm. Nearest neighbours are identified based on the Euclidean distance. The optimal number of neighbours $K \in \{5, 10, 15, 20\}$ is determined via cross-validation. For each likelihood contour function, 200 points are generated according to a Dirichlet distribution with coefficient $\gamma \in \{1, 10\}$. A 10-Fold cross

[3] `https://github.com/LoicAdam/Imprecise_Plackett_Luce/`

[4] Available on `http://www.kamishima.net/sushi/`

Table 2: Datasets and their properties (the type refers to the original problem type: A for classification and B for regression

Dataset	Type	# instances	# attributes	# labels
Authorship	A	841	70	4
Bodyfat	B	252	7	7
Glass	A	214	9	6
Housing	B	506	6	6
Iris	A	150	4	3
Stock	B	950	5	5
Sushi	A	5000	11	10
Vehicle	A	846	18	4
Vowel	A	528	10	11
Wine	A	178	13	3
Wisconsin	B	194	16	16

validation is repeated 5 times for each setting. Moreover, a 95% confidence interval is provided, based on a Gaussian assumption. To compare both methods for different values of completeness, we used different thresholds and different values of β.

We further evaluate the robustness of the procedures. First, we delete some labels in each ranking, by choosing at random for each label whether it should be kept or not. We fix the probability of deleting a label to $p \in [0,1]$. In a second step, we swap neighbouring pairs of labels (we only consider neighbouring labels in a ranking to avoid unrealistic perturbations of the data). For example, $\lambda_{\tau(2)}$ can be swapped only with $\lambda_{\tau(1)}$ and $\lambda_{\tau(3)}$. Each neighbouring label pair is swapped with probability $p \in [0,1]$. Note that the order of the swaps is a random permutation, to allow for any label $\lambda_{\tau(i)}$ to be swapped with $\lambda_{\tau(j)}$, $\forall i, j \in \Lambda$.

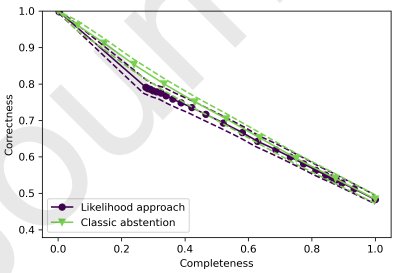


Figure 8: Comparison of methods on Wisconsin with no perturbations

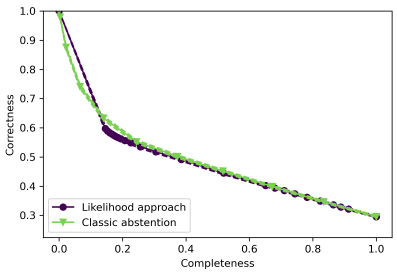


Figure 9: Comparison of methods on Sushi with no perturbations

As seen in Figures 8 and 9, the contour likelihood-based approach is on par with the method based on abstention, with no method giving a significantly

higher correctness for a given completeness value. This was the case with all datasets used in our experiments. As expected, when we have complete rankings, with $\beta = 1$ or $t = 0.5$ depending on the method, the correctness is rather low. Nevertheless, when abstention is allowed, correctness increases until it reaches one for a completeness of zero.

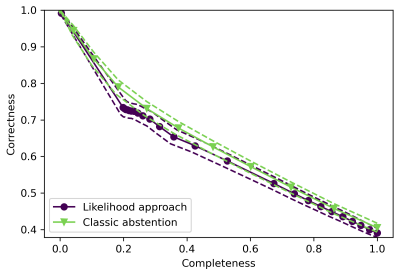


Figure 10: Comparison on Wisconsin with a missingness of 60 %

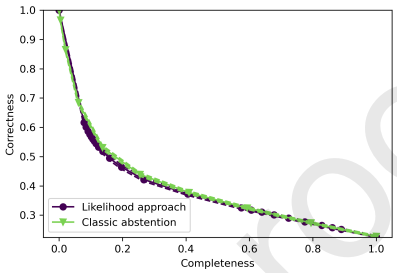


Figure 11: Comparison on Sushi with 60 % of swapped label pairs

Figures 10 and 11 show that the method is also on par even when the datasets are perturbed, meaning the correctness for a given completeness value is not higher for a given method, whether it be due to missing labels or swapped labels. It is also possible to notice that for a given completeness level, the correctness is lower than without noise. On average, the greater the perturbation is, the lower the average correctness is.

The results were similar on all datasets, with both methods being generally on par (see Appendix B.1). This indicates that a method based on the contour likelihood function can be used to make robust inferences for label ranking.

4.3.2. Influence of the amount of data

In this experiment, instead of adding perturbations to the training set, we reduce the training set size, in order to assess the influence of the amount of used data on the final result. Starting with a full training set, some points are randomly and progressively removed, until we obtain a training set containing only 10% of the original points. Moreover, in order to reflect the possible scarcity of data, we no longer systematically take K nearest neighbours to estimate the likelihood (as otherwise they would always rely on the same amount of data), but rather consider all neighbours within a given radius of the instance to classify. For this purpose, we compute the median M of all distances $d(\mathbf{x}^i, \mathbf{x}^j)$ between all pairs of training instances $(\mathbf{x}^i, \mathbf{x}^j)$. We then use M as a threshold in order to identify the training instances used to estimate the likelihood. If $\mathbf{x}$ is the instance for which we want to predict a ranking, we restrict the training set to $\mathcal{X}_t = \{\mathbf{x}^i : d(\mathbf{x}, \mathbf{x}^i) \leq M, i = 1, \ldots, m\}$.

The parameters for the likelihood contour function are the same as previously, and we still perform a 10-Fold cross validation repeated 5 times, with a confidence interval of 95%. A beta-cut of 10% is used in the likelihood approach. For the

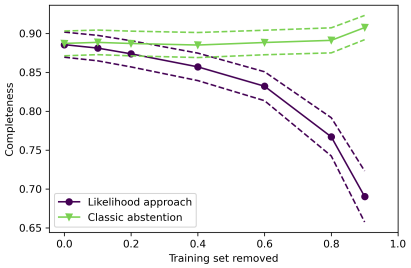


Figure 12: Completeness for Vehicle

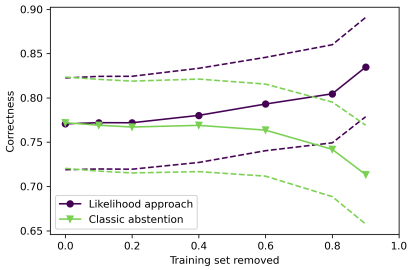


Figure 13: Correctness for Vehicle

abstention approach, a threshold $t \in [0.5, 0.6]$ is taken such that both methods have a similar starting point for completeness and correctness.

We observe in Figure 12 that completeness decreases when using the likelihood contour approach, while remaining at the same level with the abstention approach. This suggests that our approach tends to be more cautious when the available training data are scarcer. This property, *i.e.* the level of precision of the output reflects the amount of epistemic uncertainty, seems desirable. However, it should be noted that both methods have comparable accuracies in Figure 13, unless the training set becomes very small, indicating that in this case cautiousness may only be needed in situations of ambiguity.

One can check Appendix B.2 to see that the same behaviour is observed for all of our datasets: our approach is sensitive to the change in data quantity, while the thresholding approach is not. Even worse, as data become scarcer, the thresholding approach tends to provide more complete but also less accurate predictions. For instance, Figures 14 and 15 show that as completeness decreases, correctness notably increases. In other terms, for these data, abstaining is a better alternative than predicting when data are scarce. This behaviour obviously depends on the structure of the data: when many instances with clear natural groups are available, cautiousness is likely to have a marginal interest. However, with few training instances (*e.g* in the Iris data) or when groups are not well separated, our approach, being more cautious, clearly avoids making erroneous predictions for some instances.

Table 3, which summarises the results, confirms this observation. Usually, the two approaches start with the same completeness and correctness values.[5] Therefore, $\mathrm{Cp}_{\mathrm{Start}}$ ($\mathrm{Cr}_{\mathrm{Start}}$ respectively) is the average of the two starting completeness (correctness respectively) values. We can see that for the likelihood approach, completeness systematically decreases with data becoming scarcer, while correctness systematically increases. This is far from being true for the abstention approach, whose completeness can evolve in both ways (*e.g*, increases for

[5] We observe a maximal difference of 0.02 can exist, as it seems there are no explicit relation between β (beta-cut for likelihood approach) and t (threshold for abstention approach)

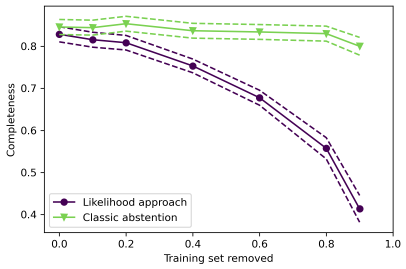


Figure 14: Completeness for Iris

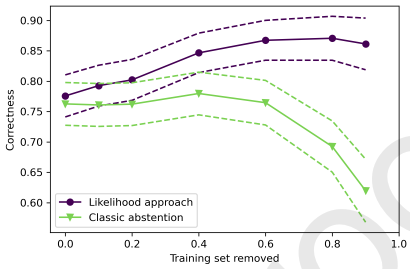


Figure 15: Correctness for Iris

Table 3: Influence of the amount of training data on completeness and correctness ($\beta = 0.1$). Here, $\boldsymbol{C}_{\text{Start}}$ stands for average values with no missing data, $\boldsymbol{C}_{\text{Lik}}$ and $\boldsymbol{C}_{\text{Abs}}$ for average values with the likelihood and abstention methods with 80% missing data. Bold letters indicate the best scores between the likelihood (Lik) and the abstention approach (Abs).

	Cp_{Start}	Cp_{Lik}	Cp_{Abs}	Cr_{Start}	Cr_{Lik}	Cr_{Abs}
Authorship	.955	**.912**	.953	.730	**.754**	.723
Bodyfat	.365	**.299**	.581	.284	**.206**	.135
Glass	.989	**.978**	.990	.706	**.718**	.713
Housing	.826	**.646**	.830	.537	**.621**	.532
Iris	.835	**.558**	.830	.770	**.871**	.692
Stock	.925	**.877**	.885	.569	**.580**	.542
Vehicle	.886	**.767**	.891	.771	**.805**	.742
Vowel	.883	**.741**	.877	.412	**.434**	.394
Wine	.696	**.553**	.770	.946	**.893**	.779
Wisconsin	.685	**.488**	.766	.552	**.476**	.380

Bodyfat, decreases for Stock), and whose correctness always decreases. Overall, this confirms that one of the interest of our approach, or of imprecise probabilistic estimation tools, lies in its sensitivity to the amount of available information, and the fact that this is reflected through the size of the set Θ of retained models.

5. Conclusions and perspectives

In this paper, we have addressed the problem of performing inference and making predictions with the well known Plackett–Luce model, a parametric ranking model. We have considered the case where the parameter vector is imprecise, in which case a set of Plackett–Luce models is valid. In this case, we have shown that imprecise predictions can be made in the form of sets of rankings. We have proposed two efficient inference methods: one allows for computing an outer approximation of the set of Walley-Sen maximal rankings and thus also of E-admissible rankings; another makes it possible to exactly compute the set of E-admissible rankings, if the parameters of the IPL model are each defined by lower and upper bounds. We have demonstrated the interest of our strategy for label ranking problems, showing that in presence of epistemic uncertainty, cautious inference—*i.e.* abstaining to make precise predictions when training data are scarce—is rewarding.

Possible future investigations may focus on improving the estimation strategy, for example by extending Bayesian approaches through the consideration of sets of priors [16]; or by developing a natively imprecise likelihood estimate, *e.g* by coupling recent estimation algorithms using stationary distribution of Markov chains [24] with recent works on imprecise Markov chains [19].

Additionally, since the Plackett–Luce is known to be strongly linked to particular random utility models [33, 3] (RUM), that models preferences between objects as real-valued random variables, it would be interesting to investigate what becomes of this relationship when making the RUM imprecise (in our case, considering Gumbel distributions with imprecise parameters).

Acknowledgements

Most of Arthur Van Camp's involvement in the research has been done as a postdoctoral researcher in the Heudiasyc laboratory of the Université de Technologie de Compiègne. He would like to thank the financial support of the national research agency (ANR), as he was funded by the ANR-18-CE23-0008 project PreServe.

References

[1] L. Adam, A. Van Camp, S. Destercke, and B. Quost. Inferring from an imprecise plackett–luce model: application to label ranking. In *International Conference on Scalable Uncertainty Management*, pages 98–112. Springer, 2020.

[2] T. Augustin, F. P. A. Coolen, G. de Cooman, and M. C. M. Troffaes, editors. *Introduction to Imprecise Probabilities*. John Wiley & Sons, 2014.

[3] H. Azari, D. Parks, and L. Xia. Random utility theory for social choice. In *Advances in Neural Information Processing Systems*, pages 126–134, 2012.

[4] G. Baltas and P. Doyle. Random utility models in marketing research: a survey. *Journal of Business Research*, 51(2):115–125, 2001.

[5] M. Cattaneo. *Statistical Decisions Based Directly on the Likelihood Function*. PhD thesis, ETH Zurich, 2007.

[6] Y. Y. Chen. Statistical inference based on the possibility and belief measures. *Transactions of the American Mathematical Society*, 347(5):1855–1863, 1995.

[7] 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*, 2010.

[8] W. Cheng, E. Hüllermeier, and K. J. Dembczynski. Label ranking methods based on the Plackett-Luce model. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 215–222, 2010.

[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] J. De Bock, C. P. De Campos, and A. Antonucci. Global sensitivity analysis for map inference in graphical models. *Advances in Neural Information Processing Systems*, 27, 2014.

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

[13] S. Destercke. A pairwise label ranking method with imprecise scores and partial predictions. In H. Blockeel, K. Kersting, S. Nijssen, and F. Železný, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 112–127, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[14] J. Fürnkranz and E. Hüllermeier, editors. *Preference Learning*. Springer Berlin Heidelberg, 2011.

[15] J. Gu and G. Yin. Fast algorithm for generalized multinomial models with ranking data. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2445–2453, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[16] J. Guiver and E. Snelson. Bayesian inference for Plackett-Luce ranking models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 377–384. ACM, 2009.

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

[18] D. R. Hunter et al. Mm algorithms for generalized bradley-terry models. *The annals of statistics*, 32(1):384–406, 2004.

[19] T. Krak, J. De Bock, and A. Siebes. Imprecise continuous-time Markov chains. *International Journal of Approximate Reasoning*, 88:452–528, 2017.

[20] I. Levi. *The Enterprise of Knowledge*. MIT Press, London, 1980.

[21] A. Liu, Z. Zhao, C. Liao, P. Lu, and L. Xia. Learning plackett-luce mixtures from partial preferences. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):4328–4335, Jul. 2019.

[22] R. D. Luce. *Individual Choice Behavior: A Theoretical analysis*. Wiley, New York, NY, USA, 1959.

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

[24] L. Maystre and M. Grossglauser. Fast and accurate inference of Plackett–Luce models. In *Advances in neural information processing systems*, pages 172–180, 2015.

[25] S. Moral and L. M. De Campos. Updating uncertain information. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 58–67. Springer, 1990.

[26] R. L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2):193–202, 1975.

[27] L. Thurstone. A law of comparative judgment. *Psychological Review*, 34:273–286, 1927.

[28] M. Troffaes. Generalising the conjunction rule for aggregating conflicting expert opinions. *I. J. of Intelligent Systems*, 21(3):361–380, March 2006.

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

[30] M. Volkovs, G. Yu, and T. Poutanen. Dropoutnet: Addressing cold start in recommender systems. *Advances in neural information processing systems*, 30, 2017.

[31] P. Walley. Belief Function Representations of Statistical Evidence. *The Annals of Statistics*, 15(4):1439 – 1465, 1987.

[32] P. Walley. *Statistical reasoning with imprecise Probabilities*. Chapman and Hall, New York, 1991.

[33] J. I. Yellott Jr. The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 15(2):109–144, 1977.

Appendix A. Proofs

Proposition 1. *Consider any parametrisation* $\Theta = \left(\bigtimes_{k=1}^{n}[\underline{v}_k, \overline{v}_k]\right) \cap \operatorname{int}(\Sigma)$ *of an imprecise Plackett–Luce model, and any ranking* τ *in* $\mathcal{L}$*. Then* τ *is E-admissible—in other words,* $\tau \in \bigcup_{v\in\Theta} \arg\max_{\tau'\in\mathcal{L}} P_v(\tau')$*—if and only if there is a* k *in* $\{1,\dots,n\}$ *such that:*

$$1 - \sum_{\ell=1}^{k-1} \min\{\overline{v}_{\tau(1)},\dots,\overline{v}_{\tau(\ell)}\} - \sum_{\ell=k+1}^{n} \max\{\underline{v}_{\tau(\ell)},\dots,\underline{v}_{\tau(n)}\} \in [\max\{\underline{v}_{\tau(k)},\dots,\underline{v}_{\tau(n)}\}, \min\{\overline{v}_{\tau(1)},\dots,\overline{v}_{\tau(k)}\}] \quad (11)$$

and

$$\underline{v}_{\tau(\ell)} \leq \min\{\overline{v}_{\tau(1)},\dots,\overline{v}_{\tau(\ell)}\} \text{ for all } \ell \text{ in } \{1,\dots,k-1\}, \text{ and } \overline{v}_{\tau(\ell)} \geq \max\{\underline{v}_{\tau(\ell)},\dots,\underline{v}_{\tau(n)}\} \text{ for all } \ell \text{ in } \{k+1,\dots,n\}. \quad (12)$$

Proof 1 (of Proposition 1, recalled above). *For sufficiency, assume that there is a* k *in* $\{1,\dots,n\}$ *such that Equations* (11) *and* (12) *hold. Then*

1. *By letting* $v_{\tau(\ell)} \coloneqq \min\{\overline{v}_{\tau(1)},\dots,\overline{v}_{\tau(\ell)}\}$ *for any* ℓ *in* $\{1,\dots,k-1\}$*,* $v_{\tau(\ell)} \coloneqq \max\{\underline{v}_{\tau(\ell)},\dots,\underline{v}_{\tau(n)}\}$ *for any* ℓ *in* $\{k+1,\dots,n\}$*, and also* $v_{\tau(k)} \coloneqq 1 - \sum_{\ell=1,\ell\neq k}^{n} v_{\tau(\ell)}$*, then by definition* $\sum_{\ell=1}^{n} v_{\tau(\ell)} = 1$*, so the elements in* v *sum up to* 1*.*

2. *Furthermore, for all* ℓ *in* $\{1,\dots,k-1\}$*, we see that* $v_{\tau(\ell)} \leq \overline{v}_{\tau(\ell)}$ *by definition, and for all* ℓ *in* $\{k+1,\dots,n\}$*, we similarly find* $v_{\tau(\ell)} \geq \underline{v}_{\tau(\ell)}$*. Equation* (12) *tells us in addition that* $v_{\tau(\ell)} \geq \underline{v}_{\tau(\ell)}$ *for all* ℓ *in* $\{1,\dots,k-1\}$*, and* $v_{\tau(\ell)} \leq \overline{v}_{\tau(\ell)}$ *for* ℓ *in* $\{k+1,\dots,n\}$*, whence* $v_{\tau(\ell)} \in [\underline{v}_{\tau(\ell)}, \overline{v}_{\tau(\ell)}] \subseteq (0,1)$ *for all* ℓ *in* $\{1,\dots,n\} \setminus \{k\}$*.*

3. *Since* $v_{\tau(k)} = 1 - \sum_{\ell=1}^{k-1} \min\{\overline{v}_{\tau(1)}, \ldots, \overline{v}_{\tau(\ell)}\} - \sum_{\ell=k+1}^{n} \max\{\underline{v}_{\tau(\ell)}, \ldots, \underline{v}_{\tau(n)}\}$, *it follows from Equation* (11) *that the strength* $v_{\tau(k)}$ *belongs to*

$$[\max\{\underline{v}_{\tau(k)}, \ldots, \underline{v}_{\tau(n)}\}, \min\{\overline{v}_{\tau(1)}, \ldots, \overline{v}_{\tau(k)}\}],$$

which is equal to

$$[\max\{v_{\tau(k+1)}, \underline{v}_{\tau(k)}\}, \min\{v_{\tau(k-1)}, \overline{v}_{\tau(k)}\}] \subseteq [\underline{v}_{\tau(k)}, \overline{v}_{\tau(k)}] \subseteq (0, 1).$$

Therefore v *belongs to* $\mathrm{int}(\Sigma)$.

We will show that the values in v *are ranked according to* τ, *because then Equation* (2) *guarantees that* τ *is E-admissible. To this end, let us first remark that* $v_{\tau(1)} \geq v_{\tau(2)} \geq \cdots \geq v_{\tau(k-1)}$ *because their defining minima are taken over increasingly bigger supersets, and similarly that* $v_{\tau(n)} \leq v_{\tau(n-1)} \leq \cdots \leq v_{\tau(k+1)}$ *because their defining maxima are taken over increasingly bigger supersets. Since we have already inferred that* $v_{\tau(k)}$ *belongs to*

$$\begin{aligned}[\max\{\underline{v}_{\tau(k)}, \ldots, \underline{v}_{\tau(n)}\}, \min\{\overline{v}_{\tau(1)}, \ldots, \overline{v}_{\tau(k)}\}] \\ = [\max\{v_{\tau(k+1)}, \underline{v}_{\tau(k)}\}, \min\{v_{\tau(k-1)}, \overline{v}_{\tau(k)}\}] \subseteq [v_{\tau(k+1)}, v_{\tau(k-1)}],\end{aligned}$$

we infer that $v_{\tau(k+1)} \leq v_{\tau(k)} \leq v_{\tau(k-1)}$, *whence indeed* $v_{\tau(1)} \geq v_{\tau(2)} \geq \cdots \geq v_{\tau(n-1)} \geq v_{\tau(n)}$.

For necessity, assume that τ *is E-admissible, so that there is a parametrisation* v *in* Θ *such that* $v_{\tau(1)} \geq v_{\tau(2)} \geq \cdots \geq v_{\tau(n-1)} \geq v_{\tau(n)}$. *We let* $\alpha := \min\{\overline{v}_{\tau(1)} - v_{\tau(1)}, v_{\tau(n)} - \underline{v}_{\tau(n)}\} \in \mathbb{R}_{\geq 0}$, *and replace* $v_{\tau(1)}$ *with* $v^{\alpha}_{\tau(1)} := v_{\tau(1)} + \alpha$, *and similarly,* $v_{\tau(n)}$ *with* $v^{\alpha}_{\tau(n)} := v_{\tau(n)} - \alpha$. *Note that this replacement does not alter the order:* $v^{\alpha}_{\tau(1)} \geq v_{\tau(2)} \geq \cdots \geq v_{\tau(n-1)} \geq v^{\alpha}_{\tau(n)}$, *and furthermore, it still sums to* 1*:* $v^{\alpha}_{\tau(1)} + v_{\tau(2)} + \cdots + v_{\tau(n-1)} + v^{\alpha}_{\tau(n)} = v_{\tau(1)} + \alpha + v_{\tau(2)} + \cdots + v_{\tau(n-1)} + v_{\tau(n)} - \alpha = 1$. *We also infer that* $v^{\alpha}_{\tau(1)} \leq \overline{v}_{\tau(1)}$ *and* $v^{\alpha}_{\tau(n)} \geq \underline{v}_{\tau(n)}$, *with one the inequalities being an equality, guaranteeing that the new parametrisation also belongs to* Θ. *All this means that* τ *maximises the probability under the new parametrisation as well, so we may assume without loss of generality that* $v_{\tau(1)} = \overline{v}_{\tau(1)}$ *or* $v_{\tau(n)} = \underline{v}_{\tau(n)}$. *In other words, we may assume that* $v_{\tau(1)}$ *or* $v_{\tau(n)}$ *is* extreme, *which means in this case being equal to* $\overline{v}_{\tau(1)}$ *or* $\underline{v}_{\tau(n)}$ *respectively.*

Now there are two cases: either (i) $v_{\tau(1)}$ *is extreme, i.e.* $v_{\tau(1)} = \overline{v}_{\tau(1)}$, *or (ii)* $v_{\tau(n)}$ *is extreme, i.e.* $v_{\tau(n)} = \underline{v}_{\tau(n)}$. *If (i), we let* $\beta := \min\{\min\{\overline{v}_{\tau(1)}, \overline{v}_{\tau(2)}\} - v_{\tau(2)}, v_{\tau(n)} - \underline{v}_{\tau(n)}\} \in \mathbb{R}_{\geq 0}$ *and replace* $v_{\tau(2)}$ *with* $v^{\beta}_{\tau(2)} := v_{\tau(2)} + \beta$, *and similarly,* $v_{\tau(n)}$ *with* $v^{\beta}_{\tau(n)} := v_{\tau(n)} - \beta$. *Then, again, this replacement does not alter the order, and sums to* 1. *We also infer that* $v^{\beta}_{\tau(2)} \leq \min\{\overline{v}_{\tau(1)}, \overline{v}_{\tau(2)}\}$ *and* $v^{\beta}_{\tau(n)} \geq \underline{v}_{\tau(n)}$, *with one the inequalities being an equality, guaranteeing that the new parametrisation also belongs to* Θ. *So we have found yet another parametrisation for which* τ *maximises the associated probability. We therefore may assume without loss of generality that* $v_{\tau(2)}$ *is extreme—equal to* $\min\{\overline{v}_{\tau(1)}, \overline{v}_{\tau(2)}\}$*—or* $v_{\tau(n)}$ *is extreme—equal to* $\underline{v}_{\tau(n)}$. *If (ii), a similar reasoning as above leads us to conclude that* $v_{\tau(1)}$

is extreme—equal to $\overline{v}_{\tau(1)}$*—or* $v_{\tau(n-1)}$ *is extreme—equal to* $\max\{\underline{v}_{\tau(n-1)}, \underline{v}_{\tau(n)}\}$*. In any case, we infer that the first* i *and the last* $j := 2 - i$ *(with* i *in* $\{0,1,2\}$*) of* $v_{\tau(1)}$*,* $v_{\tau(2)}$*, …,* $v_{\tau(n)}$ *are extreme.*

We repeat this process iteratively, each time considering the smallest index $i+1$ *such that* $v_{\tau(i+1)}$ *is non-extreme, and the biggest index* $j-1$ *such that* $v_{\tau(j-1)}$ *is non-extreme:*

$$\begin{aligned}
&v_{\tau(1)} = \overline{v}_{\tau(1)},\ v_{\tau(2)} = \min\{\overline{v}_{\tau(1)}, \overline{v}_{\tau(2)}\},\ \ldots,\ v_{\tau(i)} = \min\{\overline{v}_{\tau(1)}, \ldots, \overline{v}_{\tau(i)}\},\\
&v_{\tau(n)} = \underline{v}_{\tau(n)},\ v_{\tau(n-1)} = \max\{\underline{v}_{\tau(n)}, \underline{v}_{\tau(n-1)}\},\ \ldots,\\
&v_{\tau(j)} = \max\{\underline{v}_{\tau(n)}, \underline{v}_{\tau(n-1)}, \ldots, \underline{v}_{\tau(j)}\}.
\end{aligned}$$

If $i + 1 < j - 1$*, then, using a similar reasoning as above, without loss of generality we may replace* $v_{\tau(i+1)}$ *or* $v_{\tau(j-1)}$ *with its extreme variant—meaning that* $v_{\tau(i+1)} = \min\{\overline{v}_{\tau(1)}, \ldots, \overline{v}_{\tau(i+1)}\}$ *or* $v_{\tau(j-1)} = \max\{\underline{v}_{\tau(n)}, \underline{v}_{\tau(n-1)}, \ldots, \underline{v}_{\tau(j-1)}\}$*. We therefore may assume that* $i+1 = j-1 =: k$*. Clearly,* $v_{\tau(k)} \in [\underline{v}_{\tau(k)}, \overline{v}_{\tau(k)}]$*, but since* v *is ordered according to* τ*, we furthermore infer that* $v_{\tau(k+1)} \leq v_{\tau(k)} \leq v_{\tau(k-1)}$*, whence* $v_{\tau(k)}$ *belongs to*

$$\begin{aligned}
&[\max\{\underline{v}_{\tau(k)}, v_{\tau(k+1)}\}, \min\{\overline{v}_{\tau(k)}, v_{\tau(k-1)}\}]\\
&\qquad = [\max\{\underline{v}_{\tau(k)}, \ldots, \underline{v}_{\tau(n)}\}, \min\{\overline{v}_{\tau(1)}, \ldots, \overline{v}_{\tau(k)}\}].
\end{aligned}$$

On the other hand, since $v_{\tau(1)}$*,* $v_{\tau(2)}$*, …,* $v_{\tau(n)}$ *sum up to* 1*, we have that*

$$v_{\tau(k)} = 1 - \sum_{\ell=1, \ell\neq k}^{n} v_{\tau(\ell)} = 1 - \sum_{\ell=1}^{k-1} \min\{\overline{v}_{\tau(1)}, \ldots, \overline{v}_{\tau(\ell)}\} - \sum_{\ell=k+1}^{n} \max\{\underline{v}_{\tau(\ell)}, \ldots, \underline{v}_{\tau(n)}\},$$

whence Equation (11) *indeed is satisfied. Moreover, Equation* (12) *is satisfied since, for every* ℓ *in* $\{1, \ldots, k-1\}$*, the parameter* $v_{\tau(\ell)}$ *belongs to* $[\underline{v}_{\tau(\ell)}, \overline{v}_{\tau(\ell)}]$ *whence in particular* $\min\{\overline{v}_{\tau(1)}, \ldots, \overline{v}_{\tau(\ell)}\} = v_{\tau(\ell)} \geq \underline{v}_{\tau(\ell)}$*, and for every* ℓ *in* $\{k+1, \ldots, n\}$*, the parameter* $v_{\tau(\ell)}$ *belongs to* $[\underline{v}_{\tau(\ell)}, \overline{v}_{\tau(\ell)}]$ *whence in particular* $\max\{\underline{v}_{\tau(1)}, \ldots, \underline{v}_{\tau(\ell)}\} = v_{\tau(\ell)} \leq \overline{v}_{\tau(\ell)}$*.* □

Proposition 2. *Consider any coherent parametrisation* Θ *determined by a set of probability intervals* $[\underline{v}_k, \overline{v}_k]$ *for all* k *in* $\{1, \ldots, n\}$*, and any initial segment* $(\tau(1), \ldots, \tau(m)) = (k_1, \ldots, k_m)$ *of length* $m \in \{1, \ldots, n-1\}$*. Then, there exists an E-admissible ranking with initial segment* $(k_1, \ldots, k_m)$ *if and only if the Equations* (A_j)*,* (B_j) *and* (C_j) *are fulfilled for every* j *in* $\{1, \ldots, m\}$*.*

Proof 2 (of Proposition 2, recalled above). *Note first that due to Equation* (2)*, a ranking* τ *which admits* $(k_1, \ldots, k_m)$ *as initial sequence (i.e., such that* $(\tau(1), \ldots, \tau(m)) = (k_1, \ldots, k_m)$*) is E-admissible if and only if for some strength vector* v *in* Θ*,*

$$v_{k_1} \geq v_{k_2} \geq \cdots \geq v_{k_m} \geq \max\{v_i : i \in \{1, \ldots, n\} \setminus \{k_1, \ldots, k_m\}\}. \tag{A.1}$$

1. *For necessity, assume that there is an E-admissible ranking τ whose initial sequence is $(k_1,\dots,k_m)$. This implies, for any ℓ in $\{1,\dots,m\}$, that*

$$\begin{aligned} v_{k_1} \geq v_{k_2} \geq \cdots \geq v_{k_\ell} &\geq \max\{v_i : i \in \{1,\dots,n\} \setminus \{k_1,\dots,k_m\}\} \\ &\geq \max\{v_i : i \in \{1,\dots,n\} \setminus \{k_1,\dots,k_\ell\}\}. \end{aligned} \tag{A.2}$$

We will prove that then the Equations (A_j), (B_j) and (C_j) are fulfilled for every j in $\{1,\dots,m\}$. To this end, consider any such j. Use Equation (A.2) *with $\ell = j$ to infer that indeed Equation (B_j) is fulfilled. Equation* (A.2) *implies also that $v_{k_\ell} = \min\{v_{k_1},\dots,v_{k_\ell}\}$ for every ℓ in $\{1,\dots,j\}$, and $v_i = \min\{v_{k_1},\dots,v_{k_j},v_i\}$ for all i in $\{1,\dots,n\} \setminus \{k_1,\dots,k_j\}$. Use the fact that v sums to* 1 *to infer that*

$$\sum_{\ell=1}^{j} \min\{v_{k_1},\dots,v_{k_\ell}\} + \sum_{\substack{i=1 \\ i\notin\{k_1,\dots,k_j\}}}^{n} \min\{v_{k_1},\dots,v_{k_j},v_i\} = 1,$$

and hence we infer immediately that indeed Equation (A_j) is fulfilled. Finally, to show that also Equation (C_j) is fulfilled, infer from Equation (A.2) *that*

$$v_{k_\ell} \geq \max\{\underline{v}_i : i \in \{1,\dots,n\} \setminus \{k_1,\dots,k_\ell\}\}$$

for every ℓ in $\{1,\dots,j\}$. Use again the fact that v sums to 1 *to infer that indeed Equation (C_j) is fulfilled.*

2. *For sufficiency, let us define two vectors u and w that satisfy the condition in Equation* (A.1)*, as we will see below. Let*

$$u_{k_j} := \min\{\overline{v}_{k_1},\dots,\overline{v}_{k_j}\} \text{ and } w_{k_j} := \max\{\underline{v}_i : i \in \{1,\dots,n\} \setminus \{k_1,\dots,k_{j-1}\}\}$$

for all j in $\{1,\dots,m\}$, and

$$u_i := \min\{\overline{v}_{k_1},\dots,\overline{v}_{k_m},\overline{v}_i\} \quad \text{and} \quad w_i := \underline{v}_i$$

for all i in $\{1,\dots,n\} \setminus \{k_1,\dots,k_m\}$. Then by definition

$$u_{k_1} \geq u_{k_2} \geq \cdots \geq u_{k_m} \geq \max\{u_i : i \in \{1,\dots,n\} \setminus \{k_1,\dots,k_m\}\} \tag{A.3}$$

and

$$w_{k_1} \geq w_{k_2} \geq \cdots \geq w_{k_m} \geq \max\{w_i : i \in \{1,\dots,n\} \setminus \{k_1,\dots,k_m\}\}, \tag{A.4}$$

so we see that both vectors u and w respect the order described in Equation (A.1)*. These vectors are however not guaranteed to be strength vectors: they do not necessarily sum to* 1*, and hence, do not necessarily belong to Θ. They nevertheless exhibit useful properties: we will show (i) that u is an upper probability, and w a lower probability—which means that the former sums to a value that is at least* 1*, and the latter to a value that is at most* 1*; and (ii) that u_i and w_i belong to $[\underline{v}_i,\overline{v}_i]$ for all i in $\{1,\dots,n\}$. By taking a suitable convex combination of them, we eventually show that we will end up with a coherent strength vector v that belongs to Θ, and that satisfies the inequalities in Equation* (A.1)*.*

i *To show that* u *and* w *are an upper and a lower probability, respectively, use Equation* (A_j) *with* $j = m$ *to infer that* $\sum_{i=1}^{n} u_i \geq 1$*, and use Equation* (C_j) *with* $j = m$ *to infer that* $\sum_{i=1}^{n} w_i \leq 1$*.*

ii *We show that* u_i *and* w_i *belong to* $[\underline{v}_i, \overline{v}_i]$ *for every* i *in* $\{1,\ldots,n\}$ *by proving that* $\underline{v}_i \leq w_i \leq u_i \leq \overline{v}_i$ *for every* i *in* $\{1,\ldots,n\}$*, which implies the former. By their definitions, we immediately have that* $\underline{v}_i \leq w_i$ *and* $u_i \leq \overline{v}_i$ *for every* i *in* $\{1,\ldots,n\}$*, so it remains to show that* $w_i \leq u_i$ *for every* i *in* $\{1,\ldots,n\}$*. To this end, consider first any* j *in* $\{1,\ldots,m\}$*. Infer from Equations* (B_1) *and* (A.4) *that*

$$\overline{v}_{k_1} \geq \max\{\underline{v}_i : i \in \{1,\ldots,n\}\} \geq w_{k_1} \geq w_{k_2} \geq \cdots \geq w_{k_m}.$$

Similarly, infer from Equations (B_j) *with* $j = 2$ *and* (A.4) *that*

$$\overline{v}_{k_2} \geq \max\{\underline{v}_i : i \in \{1,\ldots,n\} \setminus \{k_1,k_2\}\} \geq w_{k_2} \geq w_{k_3} \geq \cdots \geq w_{k_m},$$

which, together with similar applications of Equations (B_j) *for* j *in* $\{3,\ldots,j\}$ *and* (A.4)*, leads to the desired inequality*

$$u_{k_j} = \min\{\overline{v}_{k_1}, \overline{v}_{k_2}, \ldots, \overline{v}_{k_j}\} \geq w_{k_j}.$$

Since the choice of j *in* $\{1,\ldots,m\}$ *was arbitrary, we have shown that* $w_{k_j} \leq u_{k_j}$ *for every* j *in* $\{1,\ldots,m\}$*. Consider now any* i *in* $\{1,\ldots,n\} \setminus \{k_1,\ldots,k_m\}$*. Use Equation* (B_j) *to infer that, for every* j *in* $\{1,\ldots,m\}$*,*

$$\overline{v}_{k_j} \geq \max\{\underline{v}_\ell : \ell \in \{1,\ldots,n\} \setminus \{k_1,\ldots,k_j\}\} \geq \underline{v}_i,$$

whence

$$\min\{\overline{v}_{k_1}, \ldots, \overline{v}_{k_m}\} \geq \underline{v}_i.$$

Since also $\overline{v}_i \geq \underline{v}_i$*, we infer that indeed*

$$u_i = \min\{\overline{v}_{k_1}, \ldots, \overline{v}_{k_m}, \overline{v}_i\} \geq \underline{v}_i = w_i.$$

This shows that $\underline{v}_i \leq w_i \leq u_i \leq \overline{v}_i$ *for every* i *in* $\{1,\ldots,n\}$*.*

In order to use our vectors u *and* w *for our goal, let* $\alpha := \sum_{i=1}^{n} u_i$ *and* $\beta := \sum_{i=1}^{n} w_i$*. We have already inferred above that* $\alpha \geq 1$ *and* $\beta \leq 1$*. If* $\alpha = 1$ *or* $\beta = 1$ *we are done, because then* u *or* w *belong to* Θ*, so one of them is a* strength vector *for which we already know that it satisfies the order of Equation* (A.1) *which implies that there is an E-admissible ranking that starts with* $(k_1,\ldots,k_m)$*. Assume therefore that* $\beta < 1 < \alpha$*, so that* $\alpha - \beta > 0$*,* $\frac{\alpha-1}{\alpha-\beta} \in (0,1)$*,* $\frac{1-\beta}{\alpha-\beta} \in (0,1)$ *and* $\frac{\alpha-1}{\alpha-\beta} + \frac{1-\beta}{\alpha-\beta} = 1$*. Let the vector* v *be defined as*

$$v_i := \frac{1-\beta}{\alpha-\beta} u_i + \frac{\alpha-1}{\alpha-\beta} w_i \quad \text{for all } i \text{ in } \{1,\ldots,n\},$$

so v is a convex combination of u and w, and it therefore too satisfies (the order described in) Equation (A.1)*, and $\underline{v}_i \leq w_i \leq v_i \leq u_i \leq \overline{v}_i$ for every i in $\{1,\ldots,n\}$. Also,*

$$\sum_{i=1}^{n} v_i = \frac{1-\beta}{\alpha-\beta}\sum_{i=1}^{n} u_i + \frac{\alpha-1}{\alpha-\beta}\sum_{i=1}^{n} w_i = \frac{1-\beta}{\alpha-\beta}\alpha + \frac{\alpha-1}{\alpha-\beta}\beta = \frac{\alpha-\beta}{\alpha-\beta} = 1,$$

so v belongs to Θ. This means that v is a strength vector of our model that satisfies the desired ordering from Equation (A.1)*, which implies that there is indeed an E-admissible ranking that starts with $(k_1,\ldots,k_m)$.* □

Appendix B. Additional experimental results

In this appendix, we introduce the experimental results on the different datasets that we didn't show in Subsection 4.3, as the results are pretty similar between each dataset. This appendix is divided in two subsections: in a first subsection, we compare our approach based on the contour likelihood function with the state-of-the-art abstention approach when 60 % of the labels are missing, or when 60 % of the labels are swapped, as presented in Paragraph 4.3.1. In a second subsection, we compare both approaches when the amount of user data in the training set is reduced, as presented in Paragraph 4.3.2. To evaluate both approaches, we use correctness and completeness as presented in Paragraph 4.2.2.

Appendix B.1. Missing and swapped labels

In this subsection, we want to see how robust both methods are when the training dataset is perturbed either due to missing labels or swapped labels on the datasets we didn't show before: Authorship, Bodyfat, Glass, Housing, Iris, Stock, Vehicle, Vowel, and Wine. For each dataset, we first provide a comparison of both methods when there are no perturbations on the dataset. Then, we provide on the left a comparison when 60 % of labels are missing, and on the right a comparison when 60 % of labels are swapped.

In general, both approaches have similar results, especially when labels are swapped. We provide for each dataset additional comments if needed.

Authorship. We notice on Figures B.16 and B.17 that our likelihood-based approach provides a higher correctness than the classic abstention approach when the completeness is around 0.85. However, our approach has difficulties reaching very low completeness values, even with β values close to 0.

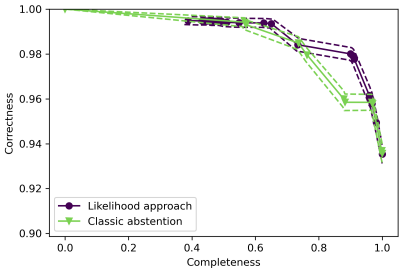


Figure B.16: Comparison of methods on Authorship with no perturbations

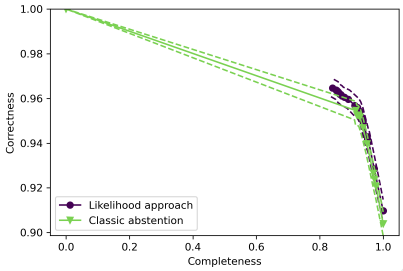


Figure B.17: Comparison on Authorship with a missingness of 60 %

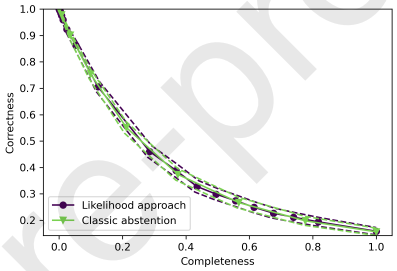


Figure B.18: Comparison on Authorship with 60 % of swapped label pairs

Bodyfat. Both methods perform very similarly on this dataset, and we have no difficulties obtaining different completeness values. Perturbing the dataset does indeed diminish the correctness for a given completeness value.

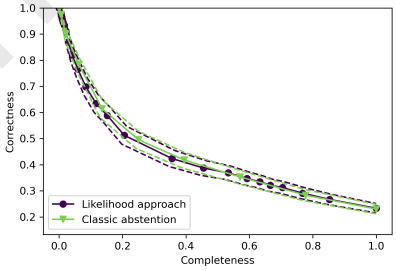


Figure B.19: Comparison of methods on Bodyfat with no perturbations

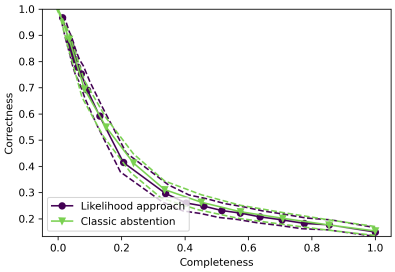


Figure B.20: Comparison on Bodyfat with a missingness of 60 %

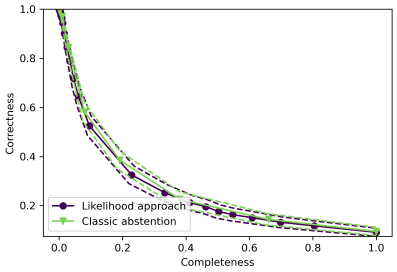


Figure B.21: Comparison on Bodyfat with 60 % of swapped label pairs

Glass. Similarly to Authorship, as we can see on Figure B.23, our likelihood-based approach provides a higher correctness than the classic abstention approach, but this time for low completeness values, while having difficulties to reach the lowest correctness values.

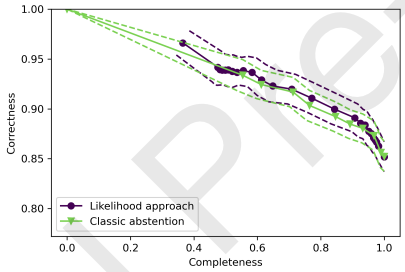


Figure B.22: Comparison of methods on Glass with no perturbations

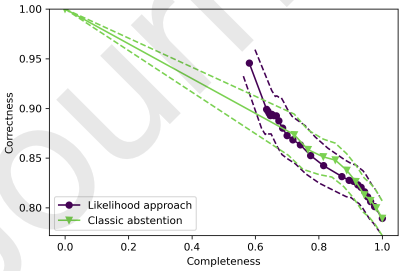


Figure B.23: Comparison on Glass with a missingness of 60 %

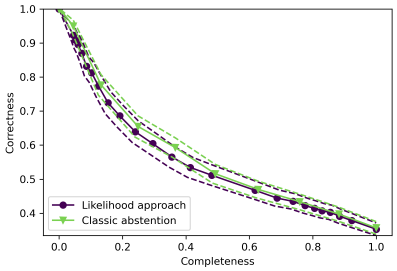


Figure B.24: Comparison on Glass with 60 % of swapped label pairs

Housing. Similarly to Bodyfat, both approaches are similar, but this time we are unable to reach a completeness of less than 0.4.

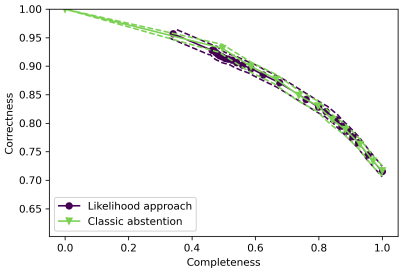


Figure B.25: Comparison of methods on Housing with no perturbations

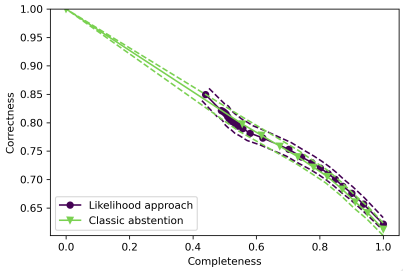


Figure B.26: Comparison on Housing with a missingness of 60 %

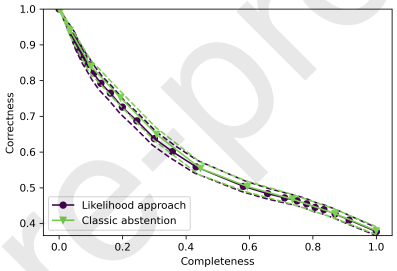


Figure B.27: Comparison on Housing with 60 % of swapped label pairs

Iris. We have the same type of behaviour as Authorship and Glass, with a higher correctness for some values of the completeness with our approach, as seen on Figure B.28, and a difficulty to reach low completeness values, as seen on Figure B.29. Let us note that, despite having a very high correctness on the standard dataset and the dataset with missing labels, the increase of the correctness is very different when the labels are swapped, as seen on Figure B.30, and is actually very similar to the increase of the correctness on the other datasets when labels are swapped.

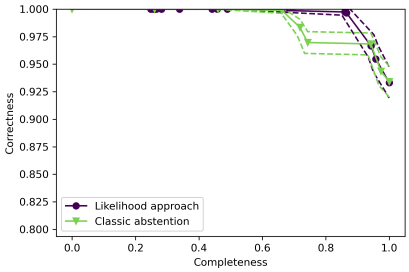


Figure B.28: Comparison of methods on Iris with no perturbations

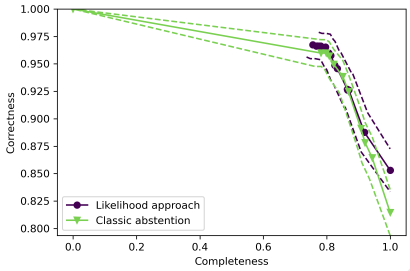


Figure B.29: Comparison on Iris with a missingness of 60 %

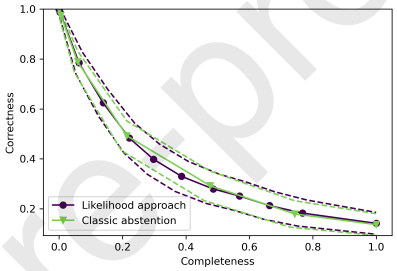


Figure B.30: Comparison on Iris with 60 % of swapped label pairs

Stock. Similarly to Bodyfat or Housing, both approaches are similar, but reaching low values of completeness is even more difficult.

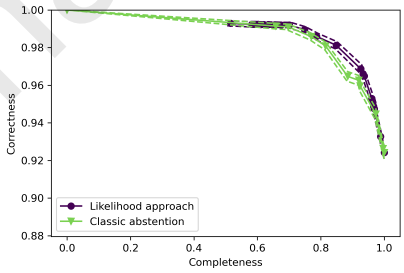


Figure B.31: Comparison of methods on Stock with no perturbations

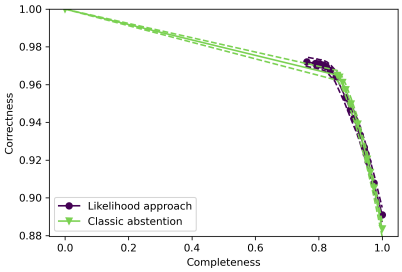


Figure B.32: Comparison on Stock with a missingness of 60 %

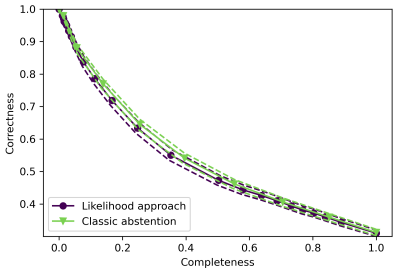


Figure B.33: Comparison on Stock with 60 % of swapped label pairs

Vehicle. Similarly to Bodyfat, Housing or Stock, both approaches are similar, with a difficulty to reach low values of completeness.

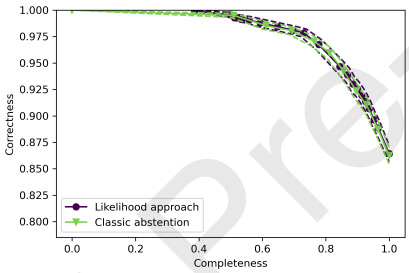


Figure B.34: Comparison of methods on Vehicle with no perturbations

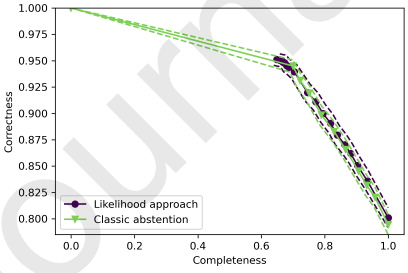


Figure B.35: Comparison on Vehicle with a missingness of 60 %

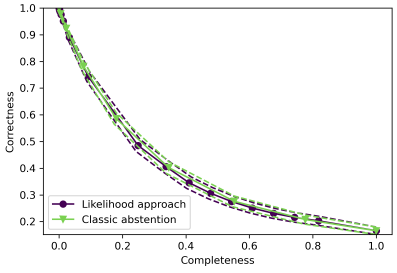


Figure B.36: Comparison on Vehicle with 60 % of swapped label pairs

Vowel. This dataset is different from the others, as our method this time actually gives a slightly lower correctness than the classic abstention approach for given completeness values, like Wisconsin dataset on Figures 8 and 10. This

is especially visible on Figures B.38 and B.39. This might be because both Vowel and Wisconsin datasets have the most labels to rank (11 and 16 respectively), and we may reach the curse of dimensionality, as we need to sample weights v on a 10 and 15 dimensional space respectively.

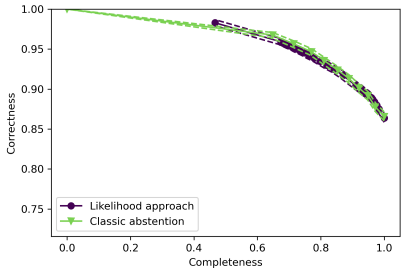


Figure B.37: Comparison of methods on Vowel with no perturbations

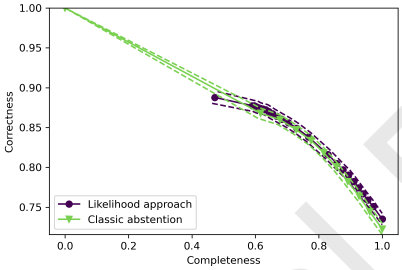


Figure B.38: Comparison on Vowel with a missingness of 60 %

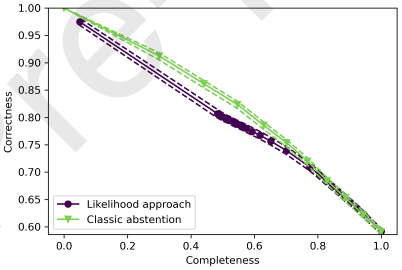


Figure B.39: Comparison on Vowel with 60 % of swapped label pairs

Wine. We have the same type of behaviour as Authorship, Glass and Iris, with a higher correctness for some values of the completeness with our approach, as seen on Figure B.41, and a difficulty to reach low completeness values, as seen on the same figure. This is one of the easiest dataset to predict on (with Iris), and we reach very high correctness values very easily, even with very high completeness values (meaning we have full rankings). Nevertheless, we have the same behaviour for swapped levels as Iris, as the increase of the correctness is very different.

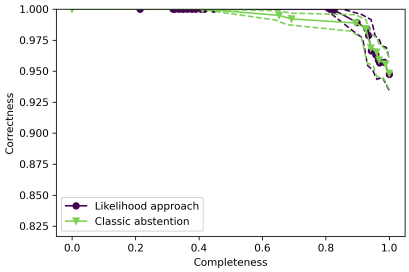


Figure B.40: Comparison of methods on Wine with no perturbations

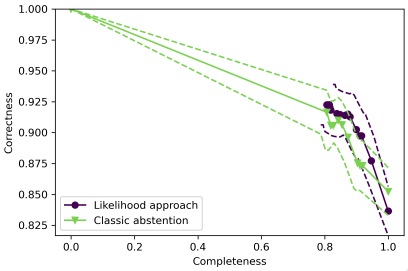


Figure B.41: Comparison on Wine with a missingness of 60 %

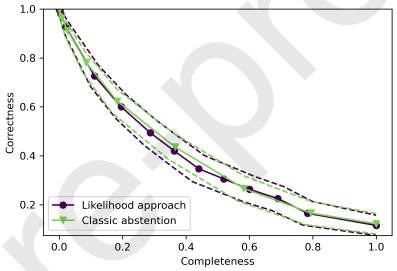


Figure B.42: Comparison on Wine with 60 % of swapped label pairs

Appendix B.2. Change in the amount of data available

In the subsection, we want to see how both methods behave when the training dataset is reduced, on the 8 datasets we didn't show before: Authorship, Bodyfat, Glass, Housing, Stock, Vowel, Wine and Wisconsin. For each dataset, we compare the completeness and the correctness between both methods.

Compared to the previous subsection, we will not provide individual comments for each dataset, as the results are very similar: the completeness of the predictions with our likelihood-based approach decreases as the training set diminishes in size, while the completeness of the predictions with the classic abstention approach does not change, or increases after a certain point. On the correctness, it is always higher for our approach, but the difference between both approaches is not always significative on some datasets.

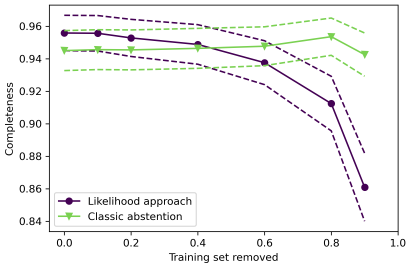


Figure B.43: Completeness for Authorship

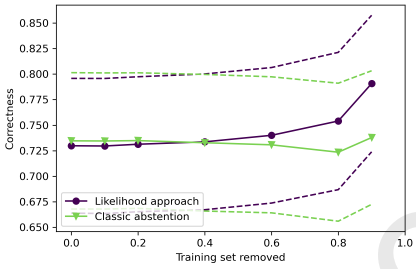


Figure B.44: Correctness for Authorship

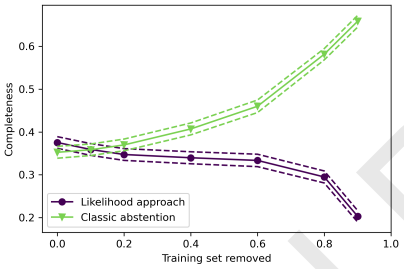


Figure B.45: Completeness for Bodyfat

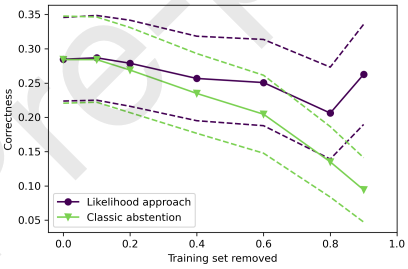


Figure B.46: Correctness for Bodyfat

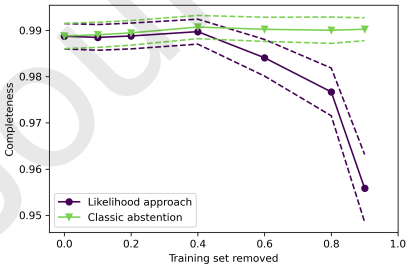


Figure B.47: Completeness for Glass

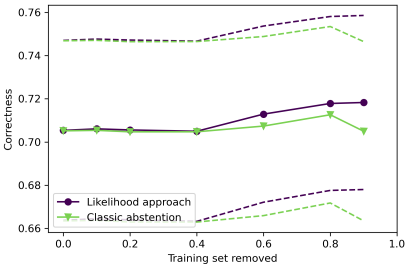


Figure B.48: Correctness for Glass

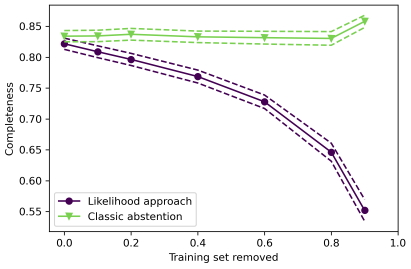


Figure B.49: Completeness for Housing

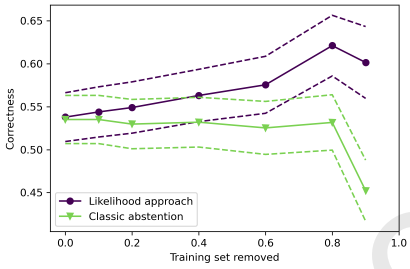


Figure B.50: Correctness for Housing

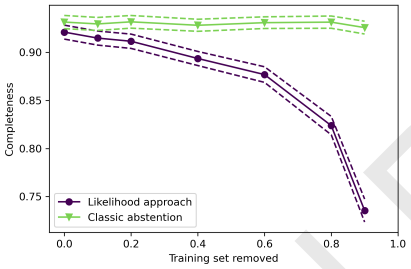


Figure B.51: Completeness for Stock

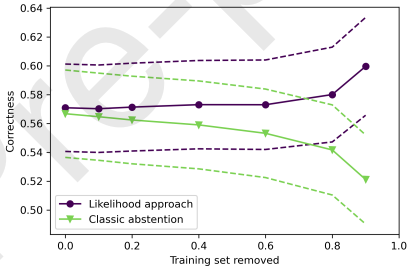


Figure B.52: Correctness for Stock

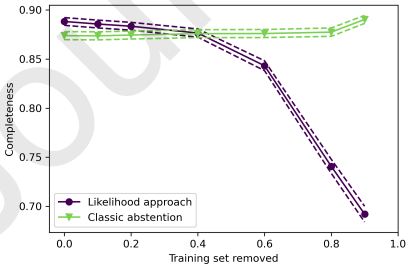


Figure B.53: Completeness for Vowel

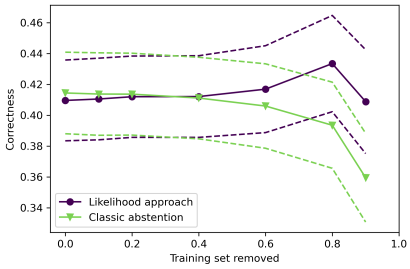


Figure B.54: Correctness for Vowel

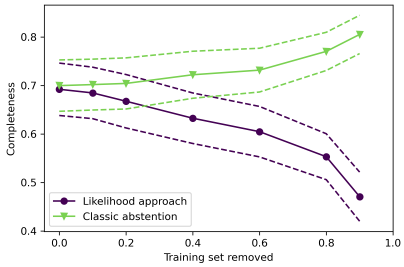


Figure B.55: Completeness for Wine

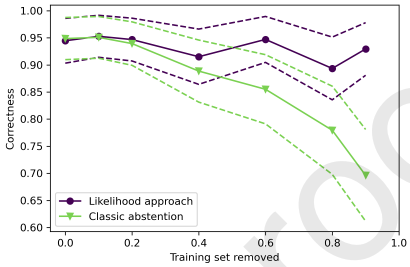


Figure B.56: Correctness for Wine

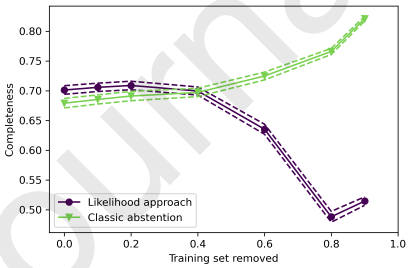


Figure B.57: Completeness for Wisconsin

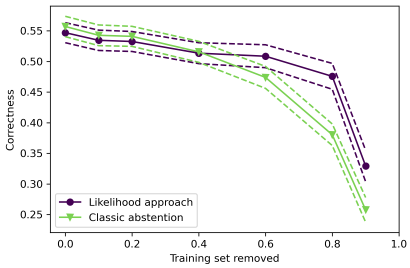


Figure B.58: Correctness for Wisconsin

Declaration of interests

☐ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☒ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Sebastien destercke reports financial support was provided by French National Research Agency.