Statistical Inference for Incomplete Ranking Data: A Comparison of Two Likelihood-Based Estimators

Inés Couso and Eyke Hüllermeier

Abstract We consider the problem of statistical inference for ranking data, namely the problem of estimating a probability distribution on the permutation space. Since observed rankings could be incomplete in the sense of not comprising all choice alternatives, we propose to tackle the problem as one of learning from imprecise or coarse data. To this end, we associate an incomplete ranking with its set of consistent completions. We instantiate and compare two likelihood-based approaches that have been proposed in the literature for learning from set-valued data, the marginal and the so-called face-value likelihood. Concretely, we analyze a setting in which the underlying distribution is Plackett-Luce and observations are given in the form of pairwise comparisons.

1 Introduction

The study of rank data and related probabilistic models on the permutation space (symmetric group) has a long tradition in statistics, and corresponding methods for parameter estimation have been used in various fields of application, such as psychology and the social sciences [1]. More recently, applications in information retrieval (search engines) and machine learning (personalization, preference learning) have caused a renewed interest in the analysis of rankings and topics such as "learning-to-rank" [2]. Indeed, methods for learning and constructing preference models from explicit or implicit preference information and feedback are among the recent research trends in these disciplines [3].

In most applications, the rankings observed are *incomplete* or *partial* in the sense of including only a subset of the underlying choice alternatives, whereas no preferences are revealed about the remaining ones—pairwise comparisons can be seen as

I. Couso
University of Oviedo, Oviedo, Spain
e-mail: couso@uniovi.es

E. Hüllermeier (✉)
Paderborn University, Paderborn, Germany
e-mail: eyke@upb.de

S. Mostaghim et al. (eds.), *Frontiers in Computational Intelligence*,
Studies in Computational Intelligence 739,
https://doi.org/10.1007/978-3-319-67789-7_3

an important special case. In this paper, we therefore approach the problem of learning from ranking data from the point of view of statistical inference with *imprecise data*. The key idea is to consider an incomplete ranking as a *set-valued observation*, namely the set of all complete rankings consistent with the incomplete observation [4]. This approach is especially motivated by recent work on learning from imprecise, incomplete, or fuzzy data [5–9].

Thus, our paper can be seen as an application of general methods proposed in that field to the specific case of ranking data. This is arguably interesting for both sides, research on statistics with imprecise data and learning from ranking data: For the former, ranking data is an interesting test bed that may help understand, analyze, and compare methods for learning from imprecise data; for the latter, general approaches for learning from imprecise data may turn into new statistical methods for ranking.

In this paper, we compare two likelihood-based approaches for learning from imprecise data. More specifically, both approaches are used for inference about the so-called Plackett-Luce model, a parametric family of probability distributions on the permutation space.

2 Preliminaries and Notation

Let $\mathbb{S}_K$ denote the collection of rankings (permutations) over a set $U = \{a_1, \dots, a_K\}$ of K items a_k, $k \in [K] = \{1, \dots, K\}$. A complete ranking (a generic element of $\mathbb{S}_K$) is a bijection $\pi : [K] \longrightarrow [K]$, where $\pi(k)$ is the position of the kth item a_k in the ranking. We denote by π^{-1} the ordering associated with a ranking, i.e., $\pi^{-1}(j)$ is the index of the item on position j. We write rankings in brackets and orderings in parentheses; for example, $\pi = [2, 4, 3, 1]$ and $\pi^{-1} = (4, 1, 3, 2)$ both denote the ranking $a_4 \succ a_1 \succ a_3 \succ a_2$.

For a possibly incomplete ranking, which includes only some of the items, we use the symbol τ (instead of π). If the kth item does not occur in a ranking, then $\tau(k) = 0$ by definition; otherwise, $\tau(k)$ is the rank of the kth item. In the corresponding ordering, the missing items do simply not occur. For example, the ranking $a_4 \succ a_1 \succ a_2$ would be encoded as $\tau = [2, 3, 0, 1]$ and $\tau^{-1} = (4, 1, 2)$, respectively. We let $I(\tau) = \{k : \tau(k) > 0\} \subset [K]$ and denote the set of all rankings (complete or incomplete) by $\overline{\mathbb{S}}_K$.

An incomplete ranking τ can be associated with its set of consistent extensions $E(\tau) \subset \mathbb{S}_K$, where

$$E(\tau) = \Big\{ \pi \; : \; (\tau(i) - \tau(j))(\pi(i) - \pi(j)) \geq 0 \text{ for all } i, j \in I(\tau) \Big\}$$

An important special case is an incomplete ranking $\tau_{i,j}$ in the form of a pairwise comparison $a_i \succ a_j$ (i.e., $\tau_{i,j}(i) = 1$, $\tau_{i,j}(j) = 2$, $\tau_{i,j}(k) = 0$ otherwise), which is associated with the set of extensions

$$E(\tau_{i,j}) = E(a_i \succ a_j) = \{\pi \in \mathbb{S}_K \ : \ \pi(i) < \pi(j)\} \ .$$

Modeling an incomplete observation τ by the set of linear extensions $E(\tau)$ reflects the idea that τ has been produced from an underlying complete ranking π by some "coarsening" or "imprecisiation" process, which essentially consists of omitting some of the items from the ranking. $E(\tau)$ then corresponds to the set of all possible candidates π, i.e., all complete rankings that are compatible with the observation τ if nothing is known about the coarsening, except that it does not change the relative order of any items.

Sometimes, more knowledge about the coarsening is available, or reasonable assumptions can be made. For example, it might be known that τ is a top-t ranking, which means that it consists of the items that occupy the first t positions in π.

3 Probabilistic Models

Statistical inference requires a probabilistic model of the underlying data generating process, which, in our case, essentially comes down to specifying a probability distribution on the permutation space. One of the most well-known probability distributions of that kind is the Plackett-Luce (PL) model [1].

3.1 *The Plackett-Luce Model*

The PL model is parametrized by a vector $\theta = (\theta_1, \theta_2, \dots, \theta_K) \in \Theta = \mathbb{R}_+^K$. Each θ_i can be interpreted as the weight or "strength" of the option a_i. The probability assigned by the PL model to a ranking represented by a permutation $\pi \in \mathbb{S}_K$ is given by

$$\mathrm{pl}_\theta(\pi) \ = \ \prod_{i=1}^{K} \frac{\theta_{\pi^{-1}(i)}}{\theta_{\pi^{-1}(i)} + \theta_{\pi^{-1}(i+1)} + \cdots + \theta_{\pi^{-1}(K)}} \qquad (1)$$

Obviously, the PL model is invariant toward multiplication of θ with a constant $c > 0$, i.e., $\mathrm{pl}_\theta(\pi) = \mathrm{pl}_{c\theta}(\pi)$ for all $\pi \in \mathbb{S}_K$ and $c > 0$. Consequently, θ can be normalized without loss of generality (and the number of degrees of freedom is only $K - 1$ instead of K). Note that the most probable ranking, i.e., the mode of the PL distribution, is simply obtained by sorting the items in decreasing order of their weight:

$$\pi^* = \arg\max_{\pi \in \mathbb{S}_K} \mathrm{pl}_\theta(\pi) = \arg\operatorname*{sort}_{k \in [K]} \{\theta_1, \dots, \theta_K\} \, . \qquad (2)$$

As a convenient property of PL, let us mention that it allows for a very easy computation of marginals, because the marginal probability on a subset $U' = \{a_{i_1}, \dots, a_{i_J}\} \subset$

U of $J \leq K$ items is again a PL model parametrized by $(\theta_{i_1}, \ldots, \theta_{i_J})$. Thus, for every $\tau \in \overline{\mathbb{S}}_K$ with $I(\tau) = U'$,

$$\mathrm{pl}_\theta(\tau) = \sum_{\pi \in E(\tau)} \mathrm{pl}_\theta(\pi) = \prod_{j=1}^{J} \frac{\theta_{\tau^{-1}(j)}}{\theta_{\tau^{-1}(j)} + \theta_{\tau^{-1}(j+1)} + \ldots + \theta_{\tau^{-1}(J)}} \tag{3}$$

In particular, this yields pairwise probabilities

$$\mathrm{pl}_\theta(\tau_{i,j}) = \mathrm{pl}_\theta(a_i \succ a_j) = \frac{\theta_i}{\theta_i + \theta_j} \, .$$

This is the well-known Bradley-Terry model [1], a model for the pairwise comparison of alternatives. Obviously, the larger θ_i in comparison to θ_j, the higher the probability that a_i is chosen. The PL model can be seen as an extension of this principle to more than two items: the larger the parameter θ_i in (1) in comparison to the parameters $\theta_j, j \neq i$, the higher the probability that a_i occupies a top rank.

3.2 A Stochastic Model of Coarsening

While (1) defines a probability for every *complete* ranking π, and hence a distribution $p : \mathbb{S}_K \longrightarrow [0, 1]$, an extension of p from $\mathbb{S}_K$ to $\overline{\mathbb{S}}_K$ is in principle offered by (3). One should note, however, that marginalization in the traditional sense is different from coarsening. In fact, (3) assumes the subset of items U' to be fixed beforehand, prior to drawing a ranking at random. For example, focusing on two items a_i and a_j, one may ask for the probability that a_i will precede a_j in the next ranking drawn at random according to p.

Recalling our idea of a coarsening process, it is more natural to consider the data generating process as a two step procedure:

$$p_{\theta,\lambda}(\tau, \pi) = p_\theta(\pi) \cdot p_\lambda(\tau \,|\, \pi) \tag{4}$$

According to this model, a complete ranking π is generated first according to $p_\theta(\cdot)$, and this ranking is then turned into an incomplete ranking τ according to $p_\lambda(\cdot \,|\, \pi)$. Thus, the coarsening process is specified by a family of conditional probability distributions

$$\left\{ p_\lambda(\cdot \,|\, \pi) \; : \; \pi \in \mathbb{S}_K, \, \lambda \in \Lambda \right\} \, , \tag{5}$$

where λ collects all parameters of these distributions; $p_{\theta,\lambda}(\tau, \pi)$ is the probability of producing the data $(\tau, \pi) \in \overline{\mathbb{S}}_K \times \mathbb{S}_K$. Note, however, that π is actually not observed.

4 Statistical Inference

As for the statistical inference about the process (4), our main interest concerns the “precise part”, i.e., the parameter θ, whereas the coarsening is rather considered as a complication of the estimation. In other words, we are less interested in inference about λ or, stated differently, we are interested in λ only in so far as it helps to estimate θ. In this regard, it should also be noted that inference about λ will generally be difficult: Due to the sheer size of the family of distributions (5), λ could be very high-dimensional. Besides, concrete model assumptions about the coarsening process may not be obvious.

Therefore, what we are mainly aiming for is an estimation technique that is efficient in the sense of circumventing direct inference about λ, and at the same time robust in the sense that it yields reasonably good results for a wide range of coarsening procedures, i.e., under very weak assumptions about the coarsening (or perhaps no assumptions at all). As a first step toward this goal, we look at two estimation principles that have recently been proposed in the literature, both being based on the principle of likelihood maximization.

In the following, the random variable X will denote the precise outcome of a single random experiment, i.e., a complete ranking π, whereas Y denotes the coarsening τ. We assume to be given an i.i.d. sample of size N and let $\tau = (\tau_1, \ldots, \tau_N) \in (\overline{\mathbb{S}}_K)^N$ denote a sequence of N independent incomplete observations of Y.

4.1 The Marginal Likelihood

The perhaps most natural approach is to consider the marginal likelihood function (also called “visible likelihood” in [10]), i.e., the probability of the observed data Y given the parameters θ and λ:

$$\begin{aligned} L_V(\theta, \lambda) = p(\boldsymbol{\tau} \mid \theta, \lambda) &= \prod_{i=1}^{N} p(Y = \tau_i \mid \theta, \lambda) \\ &= \prod_{i=1}^{N} \sum_{\pi \in \mathbb{S}_K} p_\theta(\pi) p_\lambda(\tau_i \mid \pi) \end{aligned} \tag{6}$$

The maximum likelihood estimate (MLE) would then be given by

$$(\theta^*, \lambda^*) = \underset{(\theta,\lambda)\in\Theta\times\Lambda}{\arg\max}\, L_V(\theta, \lambda)\,,$$

or, emphasizing inference about θ, by

$$\theta^* = \underset{\theta\in\Theta}{\arg\max}\, \max_{\lambda\in\Lambda} L_V(\theta, \lambda)\,.$$

As can be seen, this approach requires assumptions about the parametrization of the coarsening, i.e., the parameter space Λ. Of course, since both $\mathbb{S}_K$ and $\overline{\mathbb{S}}_K$ are finite, these assumptions can be "vacuous" in the sense of allowing all possible distributions. Thus, the family (5) would be specified in a tabular form by letting

$$p_\lambda(\tau \mid \pi) = \lambda_{\pi,\tau} \tag{7}$$

for all $\tau \in \overline{\mathbb{S}}_K$ and $\pi \in E(\tau)$ (recall that $p_\lambda(\tau \mid \pi) = 0$ for $\pi \notin E(\tau)$). In other words, Λ is given by the set of all these parametrizations under the constraint that

$$\sum_{\tau \in \overline{\mathbb{S}}_K} \lambda_{\pi,\tau} = \sum_{\tau \in E(\tau)} \lambda_{\pi,\tau} = 1$$

for all $\pi \in \mathbb{S}_K$. We denote this parametrization by Λ_{vac}.

4.2 The Face-Value Likelihood

The *face-value likelihood* is expressed as follows [11, 12]:

$$\begin{aligned} L_F(\theta, \lambda) &= \prod_{i=1}^{N} P\big(X \in E(\tau_i) \mid \theta, \lambda\big) \\ &= \prod_{i=1}^{N} \sum_{\pi \in E(\tau_i)} p_\theta(\pi) \end{aligned} \tag{8}$$

Note that the face-value likelihood does actually not depend on λ, which means that we could in principle write $L_F(\theta)$ instead of $L_F(\theta, \lambda)$. Indeed, this approach does not explicitly account for the coarsening process, or at least does not consider the coarsening as a stochastic process. The only way of incorporating knowledge about this process is to replace the set of linear extensions, $E(\tau_i)$, with a smaller set of complete rankings associated with an incomplete observation τ_i. This can be done if the coarsening is deterministic, like in the case of top-t selection.

5 Comparison of the Approaches

These two likelihood functions (6) and (8) coincide when the collection of possible values for Y forms a partition of the collection of permutations $\mathbb{S}_K$, since the events $Y = \tau_i$ and $X \in E(\tau_i)$ are then the same. But they do not coincide in the general case, where the event $Y = \tau_i$ implies but does not necessarily coincide with $X \in E(\tau_i)$.

In the following, we refer to the parameter estimation via maximization of (6) and (8) as MLM (marginal likelihood maximization) and FLM (face-value likelihood maximization), respectively.

5.1 Known Coarsening

A comparison between the marginal and face-value likelihood is arguable in the case where the coarsening is assumed to be known, because, as already said, the face-value likelihood is not able to exploit this knowledge (unless the coarsening is deterministic and forms a partition). Obviously, ignorance of the coarsening may lead to very poor estimates in general, as shown by the following example.

Let $K = 3$ and $U = \{a_1, a_2, a_3\}$. To simplify notation, we denote a ranking $a_i \succ a_j \succ a_k$ inducing $\pi^{-1} = (i, j, k)$ by $a_i a_j a_k$. We assume the PL model and suppose the coarsening to be specified by the following (deterministic) relation between complete rankings π and incomplete observations τ, which are all given in the form of pairwise comparisons:

	$a_1 \succ a_2$	$a_2 \succ a_1$	$a_1 \succ a_3$	$a_3 \succ a_1$	$a_2 \succ a_3$	$a_3 \succ a_2$
$a_1 a_2 a_3$	0	0	0	0	1	0
$a_1 a_3 a_2$	0	0	0	0	0	1
$a_2 a_1 a_3$	0	1	0	0	0	0
$a_2 a_3 a_1$	0	1	0	0	0	0
$a_3 a_1 a_2$	0	0	0	1	0	0
$a_3 a_2 a_1$	0	1	0	0	0	0

Denoting by $n_{i,j}$ the number of times $a_i \succ a_j$ has been observed, the face-value likelihood function reads as follows:

$$L_F(\boldsymbol{\tau}; \theta) = \prod_{i=1}^{3} \prod_{j \neq i} \left(\frac{\theta_i}{\theta_i + \theta_j} \right)^{n_{i,j}} .$$

Let now n_{ijk} denote the number of occurrences of the ranking $a_i a_j a_k$ in the sample. According to the above relation, we have the following:

$$\begin{aligned}
n_{1,2} &= 0 \\
n_{2,1} &= n_{213} + n_{231} + n_{321} \\
n_{1,2} &= 0 \\
n_{1,3} &= n_{312} \\
n_{2,3} &= n_{123} \\
n_{3,2} &= n_{132}
\end{aligned}$$

Therefore,

$$L_F(\theta) = \left(\frac{\theta_2}{\theta_1+\theta_2}\right)^{n_{213}+n_{231}+n_{321}} \times \left(\frac{\theta_3}{\theta_1+\theta_3}\right)^{n_{312}} \times \left(\frac{\theta_2}{\theta_2+\theta_3}\right)^{n_{123}} \times \left(\frac{\theta_3}{\theta_2+\theta_3}\right)^{n_{132}} .$$

For an arbitrary triplet $\theta = (\theta_1, \theta_2, \theta_2)$ with $\theta_1 + \theta_2 + \theta_3 = 1$, we observe that

$$L_F(\theta_1, \theta_2, \theta_3) \le L_F\left(0, \theta_2', \theta_3'\right) ,$$

where $\theta_2' = \frac{\theta_2}{\theta_2+\theta_3}$ and $\theta_3' = \frac{\theta_3}{\theta_2+\theta_3}$. In fact,

$$L_F(0, \theta_2', \theta_3') = (\theta_2')^{n_{123}} \cdot (\theta_3')^{n_{132}} ,$$

and therefore

$$L_F(\theta) = \left[\left(\frac{\theta_2}{\theta_1+\theta_2}\right)^{n_{213}+n_{231}+n_{321}} \cdot \left(\frac{\theta_3}{\theta_1+\theta_3}\right)^{n_{312}}\right] \times L_F(0, \theta_2', \theta_3') ,$$

which is clearly less than or equal to $L_F(0, \theta_2', \theta_3')$. Furthermore, according to Gibb's inequality, the above likelihood value, $L_F(0, \theta_2', \theta_3')$, is maximized for

$$(\hat{\theta}_2', \hat{\theta}_3') = \left(\frac{n_{123}}{n_{123}+n_{132}}, \frac{n_{132}}{n_{123}+n_{132}}\right) .$$

For instance, if we assume that the true distribution over $\mathbb{S}_3$ is PL with parameter $\theta = (\theta_1, \theta_2, \theta_3) = (0.99, 0.005, 0.005)$, then our estimation of θ based on the face-value likelihood function will be

$$\left(0, \frac{n_{123}}{n_{123}+n_{132}}, \frac{n_{132}}{n_{123}+n_{132}}\right) ,$$

which tends to $(0, 0.5, 0.5)$ as n tends to infinity.

5.2 Unknown Coarsening

The comparison between the two approaches appears to be more reasonable when the coarsening is assumed to be unknown. In that case, it might be fair to instantiate the marginal likelihood with the parametrization Λ_{vac}, because just like the face-value likelihood, it is then essentially ignorant about the coarsening. However, the estimation of the coarsening process under Λ_{vac} is in general not practicable, simply because the number of parameters (7) is too large: One parameter $\lambda_{\pi,\tau}$ for each $\tau \in \overline{\mathbb{S}}_K$ and $\pi \in E(\tau)$ makes about $2^K K!$ parameters in total. Besides, Λ_{vac} may cause problems of model identifiability. What we need, therefore, is a simplifying assumption on the coarsening.

5.2.1 Rank-Dependent Coarsening

The assumption we make here is a property we call *rank-dependent* coarsening. A coarsening procedure is rank-dependent if the incompletion is only acting on *ranks* (positions) but not on *items*. That is, the procedure randomly selects a subset of ranks and removes the items on these ranks, independently of the items themselves. In other words, an incomplete observation τ is obtained by projecting a complete ranking π on a random subset of positions $A \in 2^{[K]}$, i.e., the family (5) of distributions $p_\lambda(\cdot \mid \pi)$ is specified by a single measure on $2^{[K]}$. Or, stated more formally,

$$p_\lambda\big(\pi^{-1}(A) \mid \pi^{-1}\big) = p_\lambda\big(\sigma^{-1}(A) \mid \sigma^{-1}\big)$$

for all $\pi, \sigma \in \mathbb{S}^K$ and $A \subset [K]$, where $\pi^{-1}(A)$ denotes the projection of the ordering π^{-1} to the positions in A.

In the following, we make an even stronger assumption and assume observations in the form of (rank-dependent) pairwise comparisons. In this case, the coarsening is specified by probabilities

$$\Big\{ \lambda_{i,j} \mid 1 \le i < j \le K,\ \lambda_{i,j} \ge 0,\ \sum_{1 \le i < j \le K} \lambda_{i,j} = 1 \Big\},$$

where $\lambda_{i,j}$ denotes the probability that the ranks i and j are selected.

5.2.2 Likelihoods

Under the assumption of the PL model and rank-dependent pairwise comparisons as observations, the marginal likelihood for an observed set of pairwise comparisons $a_{i_n} \succ a_{j_n}$, $n \in [N]$, is given by

$$L_V(\theta, \lambda) = \prod_{n=1}^{N} \sum_{\pi \in \mathbb{S}_K, \pi(i_n) < \pi(j_n)} \lambda_{\pi(i_n), \pi(j_n)} \, \mathrm{pl}_\theta(\pi) \,. \tag{9}$$

The corresponding expression for the face-value likelihood is

$$L_F(\theta) = \prod_{n=1}^{N} \frac{\theta_{i_n}}{\theta_{i_n} + \theta_{j_n}} = \prod_{i \ne j} \left(\frac{\theta_i}{\theta_i + \theta_j} \right)^{n_{i,j}}, \tag{10}$$

where $n_{i,j}$ denotes the number of times $a_i \succ a_j$ has been observed. This is the Bradely-Terry-Luce (BTL) model, which has been studied quite extensively in the literature [13].

Obviously, since the face-value likelihood is ignorant of the coarsening, we cannot expect the maximizer $\hat{\theta}$ of (10) to coincide with the true parameter θ. Interestingly, however, in our experimental studies so far, these parameters have always been asymptotically *comonotonic*, which is enough to recover the most probable ranking (2). That is, the face-value likelihood seems to yield reasonably strong estimates

$$\hat{\pi} = \underset{k \in [K]}{\arg\text{ sort}} \left\{ \hat{\theta}_1, \ldots, \hat{\theta}_K \right\} , \tag{11}$$

for sufficiently large samples, although the parameter $\hat{\theta}$ itself might be biased. The question whether or not this comonotonicity holds in general is still open. Yet, we could prove an affirmative answer at least under an additional assumption.

Theorem 1 *Suppose complete rankings to be generated by the PL model with parameters $\theta_1 > \theta_2 > \cdots > \theta_K$. Moreover, let the coarsening procedure be given by a rank-dependent selection of pairwise comparisons between items where λ satisfies the following condition:*

$$\lambda_{i,j} \geq \lambda_{i',j'}, \textit{ if } 1 \leq i \leq i' < j' \leq j \leq K.$$

Then, for an arbitrarily small $\epsilon > 0$ there exists $N_\epsilon \in \mathbb{N}$ such that, for every $N \geq N_\epsilon$ the maximizer $\hat{\theta}$ of (10) satisfies

$$\hat{\theta}_1 > \hat{\theta}_2 > \cdots > \hat{\theta}_K$$

with probability at least $1 - \epsilon$.

The proof of Theorem 1 can be derived from Lemmas 1 to 6.

Lemma 1 *Suppose complete rankings to be generated by the PL model with parameters $\theta_{i_1} > \theta_{i_2} > \cdots > \theta_{i_K}$. Given $i \neq j$, let $p_{i_k,i_l} = P(X \in E(a_{i_k} \succ a_{i_l})) = \frac{\theta_{i_k}}{\theta_{i_k}+\theta_{i_l}}$ denote the probability that a_{i_k} is preferred to a_{i_l}. Then, the following inequalities hold:*

$$p_{i_k,i_l} \geq p_{i_{k'},i_{l'}}, \ \forall\, k \leq k', l \geq l'.$$

Proof Straightforward, if we take into account that the mapping $f_c(x) = \frac{x}{x+c}$ is increasing on $\mathbb{R}_+$ for all $c > 0$ while $g_c : (x) = \frac{c}{c+x}$ is decreasing on $\mathbb{R}_+$.

Definition 1 Consider a complete ranking $\pi \in \mathbb{S}_K$, and let us consider two indices $i \neq j$. We define the (i,j)-*swap ranking*, $\pi_{i,j} : [K] \to [K]$, as follows: $\pi_{i,j}(k) = \pi(k)$, $\forall\, k \in [K] \setminus \{i,j\}$, $\pi_{i,j}(i) = \pi(j)$ and $\pi_{i,j}(j) = \pi(i)$.

Lemma 2 *Suppose complete rankings to be generated by the PL model* pl_θ*. Take* $i, j \in [K]$ *such that* $\theta_i > \theta_j$ *and* $\pi \in \mathbb{S}_K$*. Then:*

$$\pi(i) < \pi(j) \text{ if and only if } \mathrm{pl}_\theta(\pi) > \mathrm{pl}_\theta(\pi_{i,j}).$$

Proof Let us take an arbitrary ranking $\pi \in \mathbb{S}_k$ satisfying the restriction $\pi(i) < \pi(j)$ We can write:

$$\mathrm{pl}_\theta(\pi) = C_{i,j} \cdot \frac{\theta_{\pi^{-1}(\pi(i))}}{\sum_{s=\pi(i)}^{\pi(K)} \theta_{\pi^{-1}(s)}} \cdot \frac{\theta_{\pi^{-1}(\pi(j))}}{\sum_{s=\pi(j)}^{\pi(K)} \theta_{\pi^{-1}(s)}}$$

$$\mathrm{pl}_\theta(\pi_{i,j}) = C_{i,j} \cdot \frac{\theta_{\pi_{i,j}^{-1}(\pi_{i,j}(i))}}{\sum_{s=\pi_{i,j}(i)}^{\pi_{i,j}(K)} \theta_{\pi_{i,j}^{-1}(s)}} \cdot \frac{\theta_{\pi_{i,j}^{-1}(\pi_{i,j}(j))}}{\sum_{s=\pi_{i,j}(j)}^{\pi_{i,j}(K)} \theta_{\pi_{i,j}^{-1}(s)}},$$

where

$$C_{i,j} = \prod_{r \notin \{\pi(i), \pi(j)\}} \frac{\theta_{\pi^{-1}(r)}}{\theta_{\pi^{-1}(r)} + \theta_{\pi^{-1}(r+1)} + \ldots + \theta_{\pi^{-1}(K)}}$$
$$= \prod_{r \notin \{\pi_{i,j}(i), \pi_{i,j}(j)\}} \frac{\theta_{\pi_{i,j}^{-1}(r)}}{\theta_{\pi_{i,j}^{-1}(r)} + \theta_{\pi_{i,j}^{-1}(r+1)} + \ldots + \theta_{\pi_{i,j}^{-1}(K)}}.$$

According to the relation between π and $\pi_{i,j}$, we can easily check the following equality:

$$\sum_{s=\pi(i)}^{\pi(K)} \theta_{\pi^{-1}(s)} = \sum_{s=\pi_{i,j}(j)}^{\pi_{i,j}(K)} \theta_{\pi_{i,j}^{-1}(s)}$$

(In fact, both θ_i and θ_j appear in both sums). Therefore, $\mathrm{pl}_\theta(\pi) > \mathrm{pl}_\theta(\pi_{i,j})$ if and only if $\sum_{s=\pi(j)}^{\pi(K)} \theta_{\pi^{-1}(s)} < \sum_{s=\pi_{i,j}(i)}^{\pi_{i,j}(K)} \theta_{\pi_{i,j}^{-1}(s)}$. Furthermore, we observe that:

$$\sum_{s=\pi(j)}^{\pi(K)} \theta_{\pi^{-1}(s)} - \sum_{s=\pi_{i,j}(i)}^{\pi_{i,j}(K)} \theta_{\pi_{i,j}^{-1}(s)} = \theta_j - \theta_i,$$

and therefore $\mathrm{pl}_\theta(\pi) > \mathrm{pl}_\theta(\pi_{i,j})$ if and only if $\theta_j < \theta_i$.

Lemma 3 *If* $a > a'$ *and* $b > b'$ *then* $ab + a'b' > ab' + a'b$*.*

Proof Straightforward.

Lemma 4 *Suppose that* $\lambda_{k,l} \geq \lambda_{k',l'}$ *for all* k, l, k', l' *such that* $k \leq k'$, $l \geq l'$, $k < l\, k' < l'$*. Suppose complete rankings to be generated according to a distribution p satisfying* $p(\pi) > p(\pi_{i,j})$ *for every* $\pi \in E(a_i \succ a_j)$*, for every pair* $i < j$*.*

Let $q_{i,j} = \sum_{\pi \in E(a_i > a_j)} p(\pi)\lambda_{\pi(i),\pi(j)}$, for all $i \neq j$ denote the probability of observing that a_i is preferred to a_j. Then

$$q_{i,j} > q_{i',j'} \text{ if } i \leq i' \text{ and } j \geq j'.$$

Proof We will divide the proof into three cases.

- Let us first prove that $q_{i,j} > q_{j,i}$, for all $i < j$. By definition, we have:

$$q_{i,j} = \sum_{\pi \in E(a_i > a_j)} p(\pi)\lambda_{\pi(i),\pi(j)} \text{ and } q_{j,i} = \sum_{\pi \in E(a_j > a_i)} p(\pi)\lambda_{\pi(j),\pi(i)}.$$

  Furthermore, $E(a_j \succ a_i) = \{\pi_{i,j} \,:\, \pi \in E(a_i \succ a_j)\}$ and thus we can alternatively write:

$$q_{j,i} = \sum_{\pi \in E(a_i > a_j)} p(\pi_{i,j})\lambda_{\pi_{i,j}(j),\pi_{i,j}(i)} = \sum_{\pi \in E(a_i > a_j)} p(\pi_{i,j})\lambda_{\pi(i),\pi(j)}.$$

  We easily deduce that $q_{i,j} > q_{j,i}$, $\forall\, i < j$ from the above hypotheses.
- Let us now prove that $q_{i,j} > q_{i+1,j}$, for all (i,j) with $i + 1 < j$. By definition we have:

$$q_{i,j} = \sum_{\pi \in E(a_i > a_j)} p(\pi)\lambda_{\pi(i),\pi(j)} = \sum_{\pi \in \mathbb{S}_k} p(\pi)\alpha_{\pi(i),\pi(j)},$$

  where $\alpha_{k,l} = \lambda_{k,l}$ for $k < l$ and $\alpha_{k,l} = 0$ otherwise. We can alternatively write:

$$q_{i,j} = \sum_{\pi \in E(a_i > a_{i+1})} p(\pi)\alpha_{\pi(i),\pi(j)} + \sum_{\pi \in E(a_{i+1} > a_i)} p(\pi)\alpha_{\pi(i),\pi(j)},$$

  or, equivalently:

$$q_{i,j} = \sum_{\pi \in E(a_i > a_{i+1})} [p(\pi)\alpha_{\pi(i),\pi(j)} + p(\pi_{i,i+1})\alpha_{\pi(i+1),\pi(j)}].$$

  Analogously, we can write:

$$q_{i+1,j} = \sum_{\pi \in E(a_i > a_{i+1})} [p(\pi)\alpha_{\pi(i+1),\pi(j)} + p(\pi_{i,i+1})\alpha_{\pi(i),\pi(j)}].$$

  Now, according to the hypotheses, for every $\pi \in E(a_i \succ a_{i+1})$, $p(\pi) > p(\pi_{i,i+1})$ and $\alpha_{\pi(i),\pi(j)} > \alpha_{\pi(i+1),\pi(j)}$. Then, we can easily deduce from Lemma 3 that $q_{i,j} > q_{i+1,j}$.
- It remains to prove that $q_{i,j} > q_{i,j-1}$ for all (i,j) with $i + 1 < j$. The proof is analogous to the previous one.

Lemma 5 *For every $i \neq j$ let $n_{i,j}$ the number of times $a_i \succ a_j$ is observed in a sample of size N. Given $\epsilon > 0$ there exists $N_\epsilon \in \mathbb{N}$ such that for every $N \geq N_\epsilon$, the following equalities hold with probability at least $1 - \epsilon$:*

$$n_{i,j} \geq n_{i',j'}, \ \forall\, 1 \leq i \leq i' < j' \leq j \leq K.$$

Proof This result is a direct consequence of the WLLN and Lemma 4.

Lemma 6 *Consider the mapping* $g : \mathscr{M}_K(\mathbb{R}_+) \times \mathscr{M}_K(\mathbb{R}_+) \to \mathbb{R}_+$ *defined over the collections of pairs of K-square matrices of positive numbers as follows:*

$$g(\mathbf{r}, \mathbf{s}) = g((r_{i,j})_{i,j \in [K]}, (s_{i,j})_{i,j \in [K]}) = \prod_{i \neq j} r_{i,j}^{s_{i,j}}.$$

Suppose that the matrix $\mathbf{s} = (s_{i,j})_{i,j \in [K]}$ *satisfies the following restriction:*

$$s_{i,j} \geq s_{i',j'}, \ \textit{if}\ i \leq i', j \geq j'.$$

Suppose that there exists $i^* \neq j^*$ *such that* $r_{i,j} \leq r_{i',i'}$ *for all* $(i,j), (i',j')$ *with:*

$$i \leq i', j \geq j', \ \{i,j\} \cap \{i^*, j^*\} \neq \emptyset \ \textit{and}\ \{i', j'\} \cap \{i^*, j^*\} \neq \emptyset.$$

Consider the matrix $\mathbf{r}' = (r'_{i,j})_{i \in K, j \in [K]}$ *where:* $r'_{i,j} = r_{\sigma(i),\sigma(j)}$, *where* $\sigma \in \mathbb{S}_K$ *swaps* i^* *and* j^*, *i.e.,* $\sigma(i^*) = j^*,\ \sigma(j^*) = i^*,\ \sigma(k) = k,\ \forall\, k \in [K] \setminus \{i^*, j^*\}$.

Then $g(\mathbf{r}', \mathbf{s}) \geq g(\mathbf{r}, \mathbf{s})$.

Proof It is easy to prove that, under the above conditions, the ratio $\frac{g(\mathbf{r}',\mathbf{s})}{g(\mathbf{r},\mathbf{s})}$ is greater than 1.

5.2.3 Experiments

In order to compare the two approaches experimentally, synthetic data was produced by fixing parameters θ and λ and drawing N samples at random according to (4). Then, estimations $\hat{\theta}$ and $\hat{\pi}$ were obtained for both likelihoods, i.e., by maximizing (9) and (10). As a baseline, we also included estimates of θ assuming the coarsening λ to be known; to this end, (9) is maximized as a function of θ only. The three approaches are called MLM, FLM, and TLM, respectively.

The quality of estimates is measured both for the parameters and the induced rankings (11), in terms of the Euclidean distance between θ and $\hat{\theta}$, and in terms of the Kendall distance (relative number of pairwise inversions between items) between π and $\hat{\pi}$. The expectations of the quality measures were approximated by averaging over 100 simulation runs.

Here, we present results for a series of experiments with parameters $K = 4$, $\theta = (0.4, 0.3, 0.2, 0.1)$, and different assumptions on the coarsening:

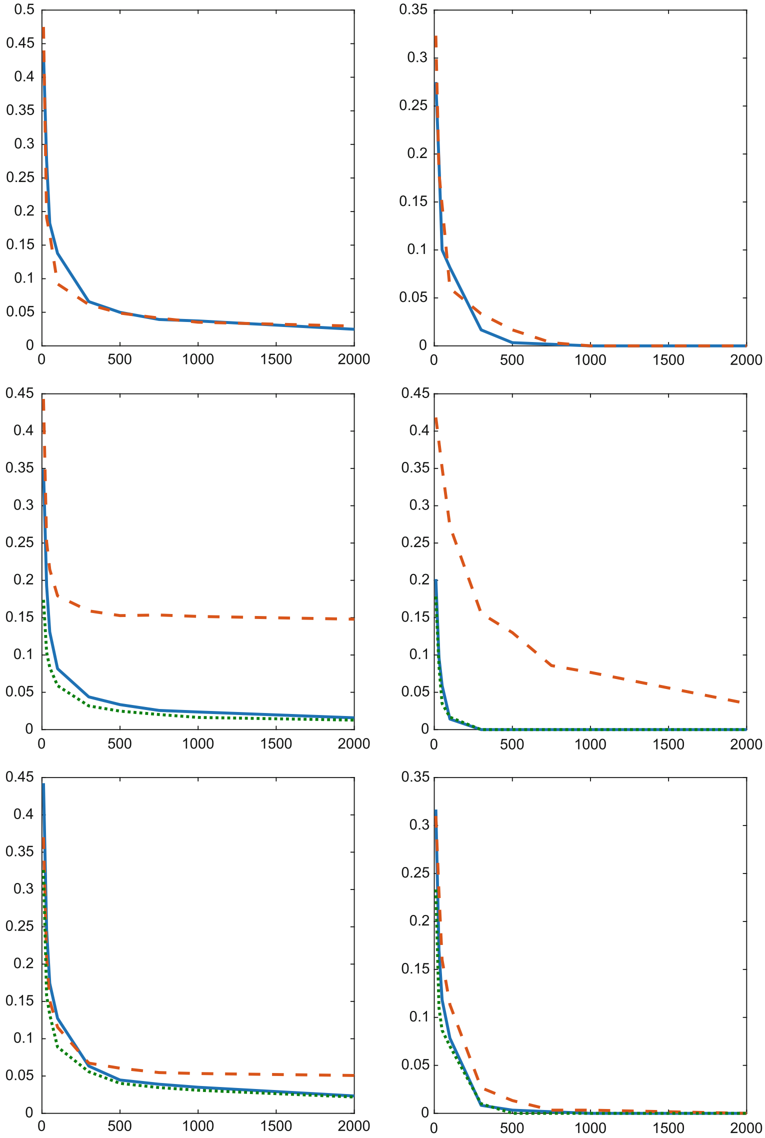

Fig. 1 Euclidean distance of parameter estimate $\hat{\theta}$ (left column) and Kendall distance of predicted ranking $\hat{\pi}$ (right column) for three experimental settings: uniform selection of pairwise comparisons (top), top-2 selection (middle), and rank-proportional selection (bottom). Curves are plotted in solid lines for MLM, dashed for FLM, and dotted for TLM

- In the first experiment, we set $\lambda_{1,2} = \cdots = \lambda_{3,4} = 1/6$. Thus, pairwise comparisons are selected uniformly at random. In this case, the face-value likelihood coincides with the likelihood of θ assuming the coarsening to be known, so this setting is clearly in favor of FLM (which, as already said, also coincides with TLM). Indeed, as can be seen in Fig. 1 (top), FLM yields very accurate estimates that improve with an increasing sample size. Nevertheless, MLM is not much worse and performs more or less on a par.
- In the second experiment, $\lambda_{1,2} = 1$ and $\lambda_{1,3} = \lambda_{1,4} = \cdots = \lambda_{3,4} = 0$. This corresponds to the top-2 setting, in which always the two items on the top of the ranking are observed. As expected, FLM now performs worse than MLM. As can be seen in Fig. 1 (middle, left), the parameter estimates of FLM are biased. Nevertheless, the estimation $\hat{\pi}$ is still decent (Fig. 1, middle, right) and continues to improve with increasing sample size.
- In the last experiment, items are selected with a probability inversely proportional their ranks: $\lambda_{i,j} \propto (8 - i - j)$. Thus, pairs on better ranks are selected with a higher probability than pairs on lower ranks. The results are shown in Fig. 1 (bottom). As can be seen, FLM is again biased and performs worse than MLM. However, the bias and the difference in performance are much smaller than in the top-2 scenario. This is hardly surprising, given that the coarsening λ in this experiment is less extreme than in the top-2 case. Instead, it is closer to the uniform coarsening of the first experiment, for which, as already said, FLM is the right likelihood.

6 Conclusion

This paper is meant as a first step toward learning from incomplete ranking data based on methods for learning from imprecise (set-valued) data. Needless to say, the scope of the paper is very limited, both in terms of the methods considered (inference based on the marginal and the face-value likelihood) and the setting analyzed (observation of pairwise comparisons based on the PL model with rank-dependent coarsening)—generalizations in both directions shall be considered in future work. Nevertheless, our results clearly reveal some important points:

- The arguably “correct” way of tackling the problem is complete inference about (θ, λ), i.e., about the complete data generating process, as done by MLM. While this approach will guarantee theoretically optimal results, it will not be practicable in general, unless the number of items is small or the parametrization of the coarsening process is simplified by very restrictive assumptions.
- Simplified estimation techniques such as FLM, which make incorrect assumptions about the coarsening or even ignore this process altogether, will generally lead to biased results.
- Yet, in the context of ranking data, one has to distinguish between the estimation of the parameter θ, i.e., the identification of the model, and the prediction of a related ranking π (typically the most probable ranking given θ, i.e., the mode of

the distribution). Indeed, the main interest often concerns π, while θ only serves an auxiliary purpose. As shown by the case of FLM, a biased estimation of θ does not exclude an accurate prediction of π, at least under certain assumptions on the coarsening process.

These observations suggest a natural direction for future work, namely the search for methods that achieve a reasonable compromise in the sense of being practicable and robust at the same time, where we consider a method robust if it guarantees a strong performance over a broad range of relevant coarsening procedures. Such methods should improve on techniques that ignore the coarsening, albeit at an acceptable increase in complexity.

References

1. Marden JI (1995) Analyzing and modeling rank data. Chapman and Hall, London, New York
2. Liu TY (2011) Learning to rank for information retrieval. Springer
3. Fürnkranz J, Hüllermeier E (2010) Preference learning. Springer
4. Ahmadi Fahandar M, Hüllermeier E, Couso I (2017) Statistical inference for incomplete ranking data: the case of rank-dependent coarsening. In: Proceedings ICML–2017, 34th international conference on machine learning, Sydney, Australia
5. Denoeux T (2011) Maximum likelihood estimation from fuzzy data using the EM algorithm. Fuzzy Sets Syst 183(1):72–91
6. Denoeux T (2013) Maximum likelihood estimation from uncertain data in the belief function framework. IEEE Trans Knowl Data Eng 25(1):119–130
7. Hüllermeier E (2014) Learning from imprecise and fuzzy observations: data disambiguation through generalized loss minimization. Int J Approx Reason 55(7):1519–1534
8. Plass J, Cattaneo M, Schollmeyer G, Augustin T (2016) Testing of coarsening mechanism: coarsening at random versus subgroup independence. In: Proceedings of SMPS 2016, 8th international conference on soft methods in probability and statistics. Springer, pp 415–422
9. Viertl R (2011) Statistical methods for fuzzy data. Wiley
10. Couso I, Dubois D. A general framework for maximizing likelihood under incomplete data (Submitted for publication)
11. Dawid AP, Dickey JM (1977) Likelihood and bayesian inference from selectively reported data. J Am Stat Assoc 72:845–850
12. Jaeger M (2005) Ignorability for categorical data. Ann Stat 33(4):1964–1981
13. Bradley RA, Terry ME (1952) The rank analysis of incomplete block designs I. The method of paired comparisons. Biometrika 39:324–345