A robust Bayesian estimation approach for the imprecise Plackett-Luce model

Tathagata Basu, Sébastien Destercke, Benjamin Quost

A robust Bayesian estimation approach for the imprecise Plackett–Luce model*

Tathagata Basu[1] [0000−0002−6851−154X], Sébastien Destercke[1] [0000−0003−2026−468X], and Benjamin Quost[1] [0000−0002−0456−9953]

UMR CNRS 7253 Heudiasyc
Sorbonne Université, Université de Technologie de Compiègne
CS 60319 - 60203 Compiègne cedex, France
{name.surname}@hds.utc.fr

Abstract. Learning to rank has become an important part in the fields of machine learning and statistical learning. Rankings are indeed present in many applications, including cognitive psychology, recommender systems, sports tournament or automated algorithm choices. Rankings are however prone to subjectivity (when provided by users) and to incompleteness (when a contestant is missing, or users only report partial preferences). Robust or cautious approaches may overcome such issues. In this paper, we develop a Bayesian robust approach for a commonly used parametric model, the Plackett-Luce (PL) model. This allows us to obtain interval-valued parameter estimates for the strength parameter of the Plackett-Luce model. We illustrate our method with both synthetic and real data to show the usefulness of skeptic inference.

Keywords: preference learning, Plackett-Luce model, Bayesian analysis, imprecise probability.

1 Introduction

Dealing with preferences and rankings is an old topic in both statistics, AI and machine learning. They are present in cognitive psychology [12], recommender systems [7] or automated algorithm choices [11]. However, it always had an important role in sports or related events, where rankings or pairwise comparisons of participants are commonly used. One of the earliest works (dating back to the 30's) on pairwise comparisons [13] laid the foundations of the Bradley-Terry model [2]. When multiple comparisons are involved, this model naturally extends to the Plackett-Luce model, which is based on (and named after) the works of Plackett [10] and Luce [9].

Several estimation strategies were proposed for both models. In particular, Hunter [8] provided a class of minorisation-maximisation (or simply MM) algorithms, which are iterative optimisation methods to estimate the strength parameters of generalised Bradley-Terry models. Later on, several Bayesian methods

* This research was funded by the project PreServe (ANR Grant ANR-18-CE23-0008).

were introduced, such as the expectation propagation method for the Plackett-Luce model [6], a latent variable approach for the generalised Bradley-Terry models [3], etc. However, such strategies rely on a number of assumptions, and provide precise estimates whatever the amount and quality of available data.

In presence of limited information, however, it seems preferable to estimate a cautious ranking model, which would provide sets of possible rankings as outputs, or even abstain from making a prediction. Such a lack of information can occur in many different scenarios. For instance, we might have very little information before initialising a recommender system, or we may only observe partial rankings (eg, pairwise comparisons or top-k rankings). Moreover, rank data, especially those obtained by user feedback, can be subjective and uncertain, and should be treated accordingly. Therefore, it seems beneficial to carry on within a cautious framework such as described above.

Note that there exist a couple of existing such frameworks for the PL model. For instance, Cheng *et al* [4] propose to threshold pairwise ranking probabilities to obtain partial rankings, but are still based on a precise estimate and all the biases that can come with it. In Adam *et al* [1], an imprecise Plackett-Luce model was proposed for rank data; the α-cut of the contour likelihood was used to obtain robust estimates. This latter approach has the advantage that cautiousness directly depends on the lack of information; however, it has the drawbacks of likelihood-based approaches, notably a high sensitivity to the data at hand (especially for small sample sizes). In this paper, we consider using imprecise probabilities to develop a robust Bayesian PL model, where the imprecise posterior inference is the consequences of considering a set of possible priors.

The rest of the paper is organised as follows. In Section 2, we discuss our robust Bayesian approach for imprecise label ranking. Section 3 presents the proposed maximum a posteriori estimation strategy of the model. In Section 4, we illustrate our method on synthetic and real data, and we compare our strategy with other methods. Finally, we conclude this paper in Section 5.

2 Our robust PL model

This section introduces the model, and shows how we can perform robust Bayesian inference over such a model.

2.1 The Plackett-Luce model

Assume we want to learn a probabilistic model over rankings of p objects. A naive computation of the empirical frequencies of the rankings at hand is doomed to fail, since although the initial space is discrete, its size is factorial in p. The recourse to a parametric model seems therefore desirable. We consider the Plackett-Luce (PL for short) model, which relies on the Luce choice axiom for choosing an item from a subset of a set of items. The axiom states that the probability of selecting one object over another from an urn containing many objects is not affected by the presence or absence of other objects in the urn [9]. Therefore, this axiom allows us to define a probabilistic model over the total order of the labels.

Let there be a total of p horses participating to n different races. Then, the PL model [10] can be defined in the following way:

$$P(X \mid \lambda) = \prod_{i=1}^{n} \prod_{j=1}^{p_i - 1} \frac{\lambda_{x_{ij}}}{\sum_{m=j}^{p_i} \lambda_{x_{im}}}. \tag{1}$$

where $p_i \leq p$ is the number of horses in the i-th race, $\lambda := (\lambda_1, \cdots, \lambda_p)$ is the vector of strength parameters and $X := [x_{ij}]$ is the $n \times p$ matrix containing the rankings (i.e., x_{ij} is the rank of the jth object/participant in the ith observed ranking, $i = 1, \ldots, n$, $j = 1, \ldots, p$). Sometimes, Eq. (1) is also called the Plackett-Luce distribution because of its probabilistic formulation.

Example 1. Consider the following ranking data from $n = 2$ races involving $p = 4$ different horses 'A', 'B', 'C' and 'D'. For instance, we have $x_{12} = 1$ and $x_{11} = 4$ (B and A have rank 1 and 4 in the first observation).

	1	2	3	4
1	B	D	C	A
2	B	A	C	–

Table 1. A toy example

Then, the PL model is given by

$$P(X \mid \lambda) = \left[\frac{\lambda_B}{\lambda_B + \lambda_D + \lambda_C + \lambda_A} \cdot \frac{\lambda_D}{\lambda_D + \lambda_C + \lambda_A} \cdot \frac{\lambda_C}{\lambda_C + \lambda_A} \right] \cdot \left[\frac{\lambda_B}{\lambda_B + \lambda_A + \lambda_C} \cdot \frac{\lambda_A}{\lambda_A + \lambda_C} \right]. \tag{2}$$

We then use this expression to estimate the strength parameters. For instance, we can empoly the MM algorithm by Hunter [8] to compute the maximum likelihood estimates.

We also have a Thurstonian interpretation for the PL model. A Thurstonian model considers a random score variable for each label. Drawing from the score distributions, and sorting based on these sampled scores, gives a sample ranking: that is, a distribution over the scores leads to a distribution over the rankings. Yellott [12] showed that a score distribution boils down to the PL model if and only if the scores follow a Gumbel distribution with fixed shape parameter.

2.2 Hierarchical model

For the robust Bayesian analysis, we adopt the latent variable model suggested by Caron and Doucet [3]. The latent variable Z_{ij} in the data augmentation process stands for the arrival time of the j-th item in the i-th race; it is assumed to be exponential:

$$Z_{ij} \overset{\text{ind}}{\sim} \text{Exp}\left(\sum_{m=j}^{p_i} \lambda_{x_{im}}\right).$$

Then, a natural choice for prior distribution over the strength parameters is the (conjugate) Gamma distribution, which ensures that $\lambda_k > 0$ for all $k = 1, \cdots, p$. This gives the following hierarchical model (assuming X are i.i.d. observations):

$$X \mid \lambda \sim \prod_{i=1}^{n} \prod_{j=1}^{p_i-1} \frac{\lambda_{x_i j}}{\sum_{m=j}^{p_i} \lambda_{x_{im}}}, \tag{3}$$

$$Z \mid X, \lambda \sim \prod_{i=1}^{n} \prod_{j=1}^{p_i-1} \left(\sum_{m=j}^{p_i} \lambda_{x_{im}}\right) \exp\left(-Z_{ij} \sum_{m=j}^{p_i} \lambda_{x_{im}}\right), \tag{4}$$

$$\lambda \sim \prod_{k=1}^{p} \frac{b_k^{a_k} \lambda_k^{a_k-1} e^{-b_k \lambda_k}}{\Gamma(a_k)}, \tag{5}$$

where $a_k > 0$ and $b_k > 0$.

In this paper, we will partially specify the prior parameters a_k by intervals $[\underline{a_k}, \overline{a_k}]$, in order to perform robust Bayesian analysis. This can also be seen as a scaled bound of the prior expectation of the k-th strength parameter. Therefore, for very limited information, we can simply consider a wider interval for a_k to perform a near-vacuous analysis.

3 Parameter Estimation

For parameter estimation, we need to investigate the posterior distributions of the strength parameters. The use of conjugate priors allows us to obtain analytic expressions of the full conditional as well as of the posteriors. However, the sensitivity analysis over a_k increases the computation cost. Therefore, we only compute the maximum a posteriori (MAP) estimates of the strength parameters instead of the full posterior analysis. We will see that this allows us to have efficient estimation procedures.

3.1 Maximum a posteriori estimation

We typically obtain MAP estimates by maximizing the posterior distribution $P(\lambda \mid X)$ with respect to λ, using an EM procedure. We first compute the complete log-likelihood which includes the latent variables in the likelihood function. From Eq. (1) and Eq. (4), we get

$$\ell(\lambda) \coloneqq \log P(X, Z|\lambda) = \sum_{i=1}^{n} \sum_{j=1}^{p_i-1} \left[\log(\lambda_{x_{ij}}) - Z_{ij} \sum_{m=j}^{p_i} \lambda_{x_{im}}\right]. \tag{6}$$

Computing the log-posterior from the log-prior defined above and this complete log-likelihood gives

$$\log(P(\lambda \mid X, Z)) = \ell(\lambda) + \log(P(\lambda)) + T \tag{7}$$

where T is an additive constant independent of λ — and therefore irrelevant for computing the MAP estimate. Therefore, for this latter purpose, following [3], we apply the EM algorithm [5] to $\ell(\lambda) + \log(P(\lambda))$.

E-step: In the E-step, we compute the expectation of $\ell(\lambda) + \log(P(\lambda))$ with respect to the latent variables Z, conditional on observing the data X and a current estimate λ^* for λ. This gives us a function of λ and λ^*:

$$\begin{aligned}
Q(\lambda, \lambda^*) &= \mathbb{E}_{Z|X,\lambda^*}\left[\ell(\lambda)\right] + \log(P(\lambda)), \\
&= \mathbb{E}_{Z|X,\lambda^*}\left[\sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\left[\log(\lambda_{x_ij}) - Z_{ij}\sum_{m=j}^{p_i}\lambda_{x_{im}}\right]\right] + \log(P(\lambda)), \\
&= \sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\left[\log(\lambda_{x_ij}) - \mathbb{E}_{Z|X,\lambda^*}(Z_{ij})\sum_{m=j}^{p_i}\lambda_{x_{im}}\right] + \log(P(\lambda));
\end{aligned}$$

since Z_{ij} follows an exponential distribution, we have

$$= \sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\left[\log(\lambda_{x_ij}) - \frac{\sum_{m=j}^{p_i}\lambda_{x_{im}}}{\sum_{m=j}^{p_i}\lambda^*_{x_{im}}}\right] + \log(P(\lambda));$$

now, from Eq. (5), we have

$$= \sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\left[\log(\lambda_{x_ij}) - \frac{\sum_{m=j}^{p_i}\lambda_{x_{im}}}{\sum_{m=j}^{p_i}\lambda^*_{x_{im}}}\right] + \sum_{k=1}^{p}\left[(a_k-1)\log(\lambda_k) - b_k\lambda_k\right] + C,$$

where C is additive constant independent of λ. Therefore,

$$Q(\lambda, \lambda^*) \equiv \sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\left[\log(\lambda_{x_ij}) - \frac{\sum_{m=j}^{p_i}\lambda_{x_{im}}}{\sum_{m=j}^{p_i}\lambda^*_{x_{im}}}\right] + \sum_{k=1}^{p}\left[(a_k-1)\log(\lambda_k) - b_k\lambda_k\right]. \tag{8}$$

M-step: In the M-step, we need to differentiate the function $Q(\lambda, \lambda^*)$ with respect to each λ_k to obtain the iterative formulation.

$$\begin{aligned}
\frac{\partial Q}{\partial \lambda_k} &= \frac{\partial}{\partial \lambda_k}\left[\sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\left[\log(\lambda_{x_ij}) - \frac{\sum_{m=j}^{p_i}\lambda_{x_{im}}}{\sum_{m=j}^{p_i}\lambda^*_{x_{im}}}\right] + \sum_{k=1}^{p}\left[(a_k-1)\log(\lambda_k) - b_k\lambda_k\right]\right], \\
&= \left[\sum_{i=1}^{n}\mathbb{I}_{k\in\{x_{i1},\cdots,x_{i(p_i-1)}\}} + a_k - 1\right]\frac{\partial\log(\lambda_k)}{\partial\lambda_k} - \left[\sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\frac{\mathbb{I}_{k\in\{x_{ij},\cdots,x_{ip_i}\}}}{\sum_{m=j}^{p_i}\lambda^*_{x_{im}}} + b_k\right]\frac{\partial\lambda_k}{\partial\lambda_k},
\end{aligned}$$

$$= \frac{1}{\lambda_k}\left[\sum_{i=1}^{n} \mathbb{I}_{k\in\{x_{i1},\cdots,x_{i(p_i-1)}\}} + a_k - 1\right] - \left[\sum_{i=1}^{n}\sum_{j=1}^{p_i-1} \frac{\mathbb{I}_{k\in\{x_{ij},\cdots,x_{ip_i}\}}}{\sum_{m=j}^{p_i}\lambda^*_{x_{im}}} + b_k\right].$$

Computing the MAP estimate requires to set this derivative to zero, which gives

$$\lambda_k = \frac{a_k + w_k - 1}{b_k + \sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\frac{\delta_{ijk}}{\sum_{m=j}^{p_i}\lambda^*_{x_{im}}}}, \tag{9}$$

where $w_k = \sum_{i=1}^{n}\mathbb{I}_{k\in\{x_{i1},\cdots,x_{i(p_i-1)}\}}$ and $\delta_{ijk} = \mathbb{I}_{k\in\{x_{ij},\cdots,x_{ip_i}\}}$. Now, to show that Eq. (9) gives a maximum for $Q(\lambda,\lambda^*)$, we investigate the Hessian matrix

$$H_Q = \begin{bmatrix} -\frac{a_1+w_1-1}{\lambda_1^2} & 0 & \cdots & 0 \\ 0 & -\frac{a_2+w_2-1}{\lambda_2^2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & -\frac{a_p+w_p-1}{\lambda_p^2} \end{bmatrix}. \tag{10}$$

Assuming that a horse does not come last in every race ($w_k \geq 1$, for all k), H_Q is negative definite for all values of λ. Therefore, the Q function attains its maximum when Eq. (9) is satisfied. This gives us the following iterative formula to obtain an approximate solution for λ:

$$\lambda_k^{(t)} = \frac{a_k + w_k - 1}{b_k + \sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\frac{\delta_{ijk}}{\sum_{m=j}^{p_i}\lambda^{(t-1)}_{x_{im}}}}. \tag{11}$$

3.2 Imprecise MAP estimation

Remark that after Eq. (11), we have

$$\lambda_k^{(t)} \propto a_k \quad \text{and} \quad \lambda_k^{(t)} \propto \sum_{m=j}^{p_i}\lambda^{(t-1)}_{x_{im}};$$

Therefore, bounds on the parameter value $\lambda_k^{(t)}$ estimated in the t-th iteration can be computed as

$$\underline{\lambda}_k^{(t)} = \frac{\underline{a_k} + w_k - 1}{b_k + \sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\frac{\delta_{ijk}}{\sum_{m=j}^{p_i}\underline{\lambda}^{(t-1)}_{x_{im}}}}, \quad \overline{\lambda}_k^{(t)} = \frac{\overline{a_k} + w_k - 1}{b_k + \sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\frac{\delta_{ijk}}{\sum_{m=j}^{p_i}\overline{\lambda}^{(t-1)}_{x_{im}}}}. \tag{12}$$

To show that these bounds reflect epistemic uncertainty (i.e., the larger the dataset, the closer the bounds are to each other), we make the following simple but necessary assumption.

Assumption 1 (Boundedness of λ_k) *There exists a global bound $M < \infty$ such that $0 < \underline{\lambda}_k^{(t)} \leq \overline{\lambda}_k^{(t)} \leq M$ for each parameter k and all iteration t.*

Note that this assumption is not very limiting, as in practice we will start from bounded initial estimates, and (12) will not diverge. Given this assumption is satisfied, we can show that our estimates become more and more precise as data are gathered.

Theorem 1 (Convergence of imprecision). *Let $\Delta_k^{(t)}(n) := \overline{\lambda}_k^{(t)} - \underline{\lambda}_k^{(t)}$ denote the imprecision in the k-th strength parameter estimated at the t-th iteration. Let $p \geq \max_i\{p_i\}$, and let $\lambda^{(0)} = (1/p, \cdots, 1/p)$ be the initial guess of the strength parameter vector. Then, as $n \to \infty$, we have $\Delta_k^{(t)}(n) \to 0$ for each parameter k and all iteration t.*

Proof. Since we obtain the lower and upper estimates by an iterative algorithm, we will prove the theorem by induction.

For $t = 1$ From Eq. (12), we have

$$\begin{aligned}\Delta_k^{(1)}(n) &= \frac{\overline{a_k} + w_k - 1}{b_k + \sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\frac{\delta_{ijk}}{\sum_{m=j}^{p_i} 1/p}} - \frac{\underline{a_k} + w_k - 1}{b_k + \sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\frac{\delta_{ijk}}{\sum_{m=j}^{p_i} 1/p}} \\ &= \frac{\overline{a}_k - \underline{a}_k}{b_k + p\sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\frac{\delta_{ijk}}{p_i+1-j}}\end{aligned}$$

Note that by construction, we have $w_k = \sum_{i=1}^{n}\mathbb{I}_{k\in\{x_{i1},\cdots,x_{i(p_i-1)}\}}$ and $\delta_{ijk} = \mathbb{I}_{k\in\{x_{ij},\cdots,x_{ip_i}\}}$. Then there exists a sequence of numbers $\{k_s\}^{\infty}$, and a sequence $\{j_{k_s}|j_{k_s} < p_{k_s}, \forall k_s \in \{k_s\}^{\infty}\}^{\infty}$, such that

$$\Delta_k^{(1)}(n) = \frac{\overline{a}_k - \underline{a}_k}{b_k + \sum_{i\in\{k_s\}^{\infty}}\sum_{j=1}^{j_i}\frac{p}{p_i+1-j}}.$$

As $n \to \infty$, we have $\sum_{i\in\{k_s\}^{\infty}}\sum_{j=1}^{j_i}\frac{p}{p_i+1-j} \to \infty$, and therefore $\Delta_k^{(1)}(n) \to 0$.

For $t = 2$ As previously, we have

$$\begin{aligned}\Delta_k^{(2)}(n) &= \frac{\overline{a_k} + w_k - 1}{b_k + \sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\frac{\delta_{ijk}}{\sum_{m=j}^{p_i}\overline{\lambda}_{x_{im}}^{(1)}}} - \frac{\underline{a_k} + w_k - 1}{b_k + \sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\frac{\delta_{ijk}}{\sum_{m=j}^{p_i}\underline{\lambda}_{x_{im}}^{(1)}}} \\ &= \frac{\overline{a_k} + w_k - 1}{b_k + \sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\frac{\delta_{ijk}}{\sum_{m=j}^{p_i}\overline{\lambda}_{x_{im}}^{(1)}}} - \frac{\underline{a_k} + w_k - 1}{b_k + \sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\frac{\delta_{ijk}}{\sum_{m=j}^{p_i}\left(\overline{\lambda}_{x_{im}}^{(1)} - \Delta_{x_{im}}^{(1)}(n)\right)}} \\ &= \delta_k^{(2)}(n) + S_k^{(2)}(n),\end{aligned}$$

where

$$\delta_k^{(2)}(n) = \left[\frac{\overline{a_k}}{b_k + \sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\frac{\delta_{ijk}}{\sum_{m=j}^{p_i}\overline{\lambda}_{x_{im}}^{(1)}}} - \frac{\underline{a_k}}{b_k + \sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\frac{\delta_{ijk}}{\sum_{m=j}^{p_i}\left(\overline{\lambda}_{x_{im}}^{(1)} - \Delta_{x_{im}}^{(1)}(n)\right)}}\right] \tag{13}$$

$$S_k^{(2)}(n) = \left[\frac{w_k - 1}{b_k + \sum_{i=1}^{n} \sum_{j=1}^{p_i - 1} \frac{\delta_{ijk}}{\sum_{m=j}^{p_i} \overline{\lambda}_{x_{im}}^{(1)}}} - \frac{w_k - 1}{b_k + \sum_{i=1}^{n} \sum_{j=1}^{p_i - 1} \frac{\delta_{ijk}}{\sum_{m=j}^{p_i} \left(\overline{\lambda}_{x_{im}}^{(1)} - \Delta_{x_{im}}^{(1)}(n) \right)}} \right]. \tag{14}$$

To show $\lim_{n\to\infty} \Delta_k^{(2)}(n) \to \infty$, we need that both $\delta_k^{(2)}(n) \to 0$ and $S_k^{(2)}(n) \to 0$ as $n \to \infty$. Due to space limitations, we will only provide details for the former. Now, from Eq. (13), we have

$$\lim_{n\to\infty} \delta_k^{(2)}(n) = \lim_{n\to\infty} \left[\frac{\overline{a_k}}{b_k + \sum_{i=1}^{n} \sum_{j=1}^{p_i - 1} \frac{\delta_{ijk}}{\sum_{m=j}^{p_i} \overline{\lambda}_{x_{im}}^{(1)}}} \right.$$
$$\left. - \frac{\underline{a_k}}{b_k + \sum_{i=1}^{n} \sum_{j=1}^{p_i - 1} \frac{\delta_{ijk}}{\sum_{m=j}^{p_i} \left(\overline{\lambda}_{x_{im}}^{(1)} - \lim_{n\to\infty} \Delta_{x_{im}}^{(1)}(n) \right)}} \right].$$

Since $\Delta_k^{(1)}(n) \to 0$ as $n \to \infty$, we have

$$\lim_{n\to\infty} \delta_k^{(2)}(n) = \lim_{n\to\infty} \left[\frac{\overline{a_k}}{b_k + \sum_{i=1}^{n} \sum_{j=1}^{p_i - 1} \frac{\delta_{ijk}}{\sum_{m=j}^{p_i} \overline{\lambda}_{x_{im}}^{(1)}}} - \frac{\underline{a_k}}{b_k + \sum_{i=1}^{n} \sum_{j=1}^{p_i - 1} \frac{\delta_{ijk}}{\sum_{m=j}^{p_i} \overline{\lambda}_{x_{im}}^{(1)}}} \right]$$
$$= \lim_{n\to\infty} \left[\frac{\overline{a_k} - \underline{a_k}}{b_k + \sum_{i=1}^{n} \sum_{j=1}^{p_i - 1} \frac{\delta_{ijk}}{\sum_{m=j}^{p_i} \overline{\lambda}_{x_{im}}^{(1)}}} \right].$$

From the boundedness assumption on λ_k, we have

$$\lim_{n\to\infty} \delta_k^{(2)}(n) \le \lim_{n\to\infty} \left[\frac{\overline{a_k} - \underline{a_k}}{b_k + \sum_{i=1}^{n} \sum_{j=1}^{p_i - 1} \frac{\delta_{ijk}}{\sum_{m=j}^{p_i} M}} \right]$$

which becomes, after applying previous arguments,

$$\le \lim_{n\to\infty} \left[\frac{\overline{a_k} - \underline{a_k}}{b_k + \frac{1}{Mp} \sum_{i\in\{k_s\}^\infty} \sum_{j=1}^{j_i} \frac{p}{p_i + 1 - j}} \right].$$

Since Mp is finite and the term $\sum_{i\in\{k_s\}^\infty} \sum_{j=1}^{j_i} \frac{p}{p_i+1-j}$ diverges to ∞ as $n \to \infty$, we therefore have

$$\lim_{n\to\infty} \delta_k^{(2)}(n) = 0. \tag{15}$$

Coupled with the same result for $S_k^{(2)}(n)$, we get $\Delta_k^{(2)}(n) \to 0$ as $n \to \infty$.

Now, by mathematical induction, we can assume that $\lim_{n\to\infty} \Delta_k^{(t)}(n) \to 0$ for $t = 1, 2, \cdots, r$. Using the same reasoning as above (which is not detailed here, due to the lack of place), we can show that this holds for $t = r + 1$ as well, which proves our theorem.

4 Illustration

For the illustration of our method, we use both synthetic and real-life datasets. We show that as we accumulate more data or ignore sparse information, our method gives us a more precise answer. To do so, we first define completeness as

$$\text{Completeness} = \frac{\text{nb observed comparisons}}{\text{nb possible comparisons}}. \tag{16}$$

Here, for two strength parameters λ_i and λ_j, we consider that $\lambda_i > \lambda_j$ if

$$\inf_{a_1,\cdots,a_P} \{\lambda_i(a_1,\cdots,a_P) - \lambda_j(a_1,\cdots,a_P)\} > 0; \tag{17}$$

note that if $\lambda_i \not> \lambda_j$ and $\lambda_j \not> \lambda_i$, we call the objects (drivers) incomparable.

4.1 Synthetic Dataset

For the synthetic dataset, we set $\lambda_k = (81-k)$ for $1 \leq k \leq 80$. We then use these fixed values of λ to generate 10000 observations of rankings. We also add some noise in the data by adding 100 observations obtained by using a different set of λ_k such that $\lambda_k = (21-k)$ for $1 \leq k \leq 20$ and $\lambda_k \sim \mathrm{U}(0,1)$ for $21 \leq k \leq 80$.

To perform our analyses, we consider two different settings. In one case, we consider the completeness of our estimate for different top k-rankings and in the other case, we consider the completeness against the total number of observations.

In the Fig. 1, we show these analyses of completeness averaged over 10 replications. We start our analyses with 20 observations and keep adding 5 observations in each step. We notice that the estimates become more complete as we increase the number of observations. It tends to be monotone except for few cases. This happens as in the new observations, a parameter may perform unexpectedly better/worse and force the estimates to be more imprecise. We also check this for 4 different values of top k-ranks. In one case, we consider all the rankings and the completeness is equal to 1 even after repeating for 10 times. But, this value changes for other cases. It is usually lower for smaller values of k. However, we notice that for $k = 20$, this completeness increases faster than the others and at a certain point the completeness is better than the completeness for $k = 40$.

4.2 NASCAR Dataset

We use the NASCAR 2002 dataset to illustrate our method with a real dataset. This dataset contains the results of 2002 season where a total of 36 races took place. In each race, 43 drivers participate and across the season 87 different drivers participate. This dataset was studied by Hunter [8] for illustration. Note that the data are pre-processed so as to eliminate 4 drivers, who always come last.

We aim at using these data in order to infer a ranking model of the drivers. More precisely, we aim at showing that as we infer from more complete data (see below for a definition of completeness), or as we ignore sparse information, our method results in a more precise ranking model.

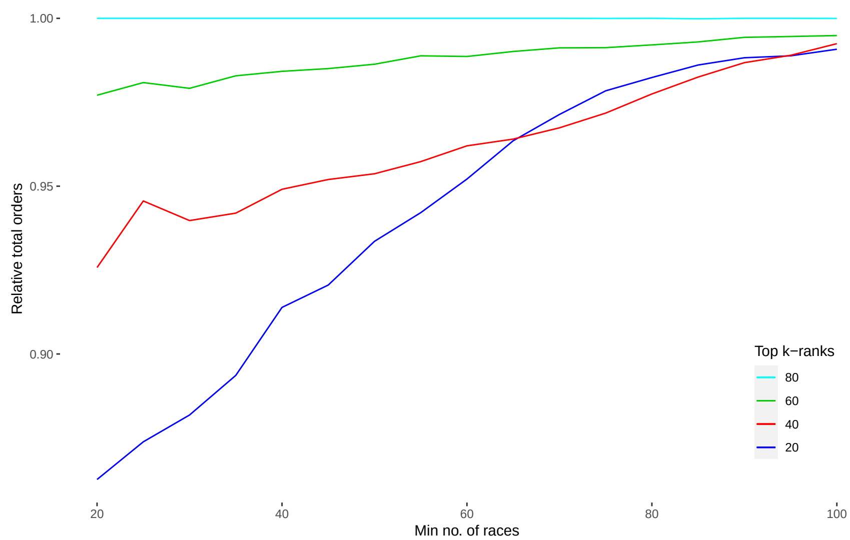

Fig. 1. Completeness of the estimated rankings with respect to different size of observations and different top k-rankings.

Cautious ranking inference First, we simply estimate the rankings of the drivers. For this purpose, we consider the whole dataset with 83 drivers. We set $a_1 = a_2 = \cdots = a_P = a$ for near vacuous analysis and consider $a \in (1/83, 82/83)$. We set $b_1 = b_2 = \cdots = b_P = 1$ and perform our robust Bayesian analysis over a to obtain the bounds for the strength parameters. We display these strength parameters (in the increasing order) of their lower bounds in Fig. 2.

Note that $\lambda_i > \lambda_j$ (or ith driver is better than the jth) in the sense of Equation (17) whenever $\underline{\lambda}_i > \underline{\lambda}_j$ and $\overline{\lambda}_i > \overline{\lambda}_j$, due to their dependencies to a_i values. We can see that a majority of drivers can be ordered; some of them are however incomparable, such as for instance driver 58, whose 4th place in one race led our robust approach to be cautious about its potential strength.

Sensitivity to completeness of training data In a second step, we analyse what becomes of our predictions if we consider more complete rankings in order to infer our model. Since all races contain different drivers, we do this in two ways.

First, we consider those drivers that have participated in at least r races. Obviously, for $r = 1$, we consider all drivers; by letting r increase, we eliminate some of them: then, we get more complete information about the remaining ones. In addition, we consider only the top-k information corresponding to race outcomes, thereby truncating the observed rankings.

Fig. 3 summarises this analysis, showing that globally, the completeness of the inferred rankings increases as we get more information. We can see that as we increase r (required number of races for a driver to be included in the data), the comparisons become more complete. This is also the case when we use the

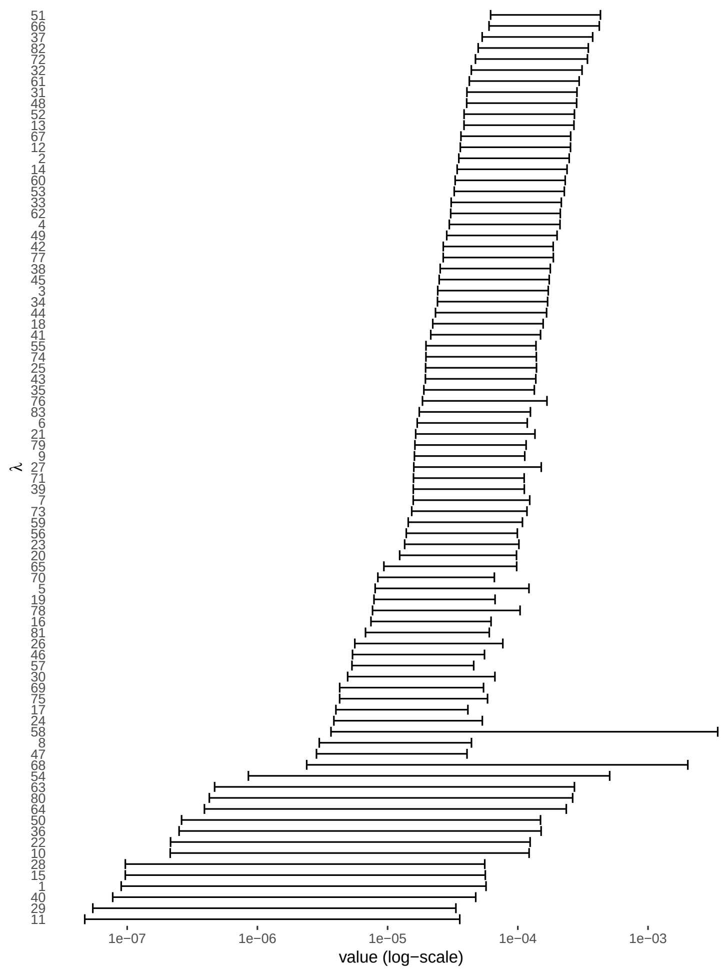

Fig. 2. MAP estimate of λ with 83 drivers

complete racing outcomes instead of the top k-ranks. In this latter case, since some drivers may perform well in some races and arrive below the kth position in others, this selection is likely to change the bounds of the estimated strength parameters, thus resulting in incomplete comparisons. We also notice that for smaller values of r, completeness is better for top 15-rankings than for top 20-rankings. This may seem contradictory, but this behaviour is actually plausible, since for those particular cases, the refined race results are more consistent for the first 15 ranks, and therefore the strength parameter bounds are tighter.

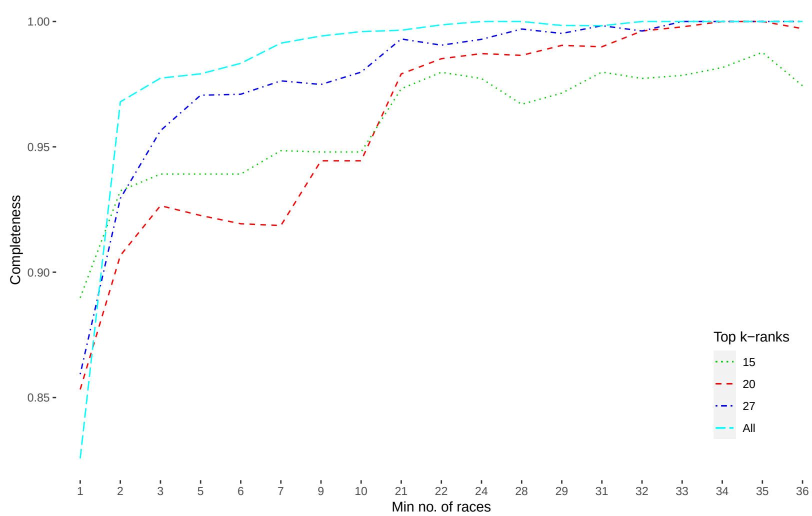

Fig. 3. Completeness of the estimated ranking for different top k-ranks with respect to minimum number of races

5 Conclusion

In this article, we propose a robust Bayesian approach for the estimation of a Plackett-Luce model, together with a robust estimation of the strength parameters, in order to perform cautious ranking inference. Such a cautious model is particularly useful for ranking problems where only limited information is available. We show that our estimation approach is consistent in that it results in tighter strength estimates as the training sample size grows.

We use a synthetic dataset to show the completeness of our imprecise estimation. We also apply our approach to the well-known NASCAR dataset containing race information. The experiments confirm that our robust Bayesian approach tends to be more complete when we have more information, satisfying our theoretical result on the imprecision of the posterior estimates.

We plan to extend our approach in several ways: studying the effect of adding prior information; using expected estimates rather than MAP ones; extending our model to mixture of PL, for instance to perform preference clustering.

Bibliography

[1] Adam, L., Van Camp, A., Destercke, S., Quost, B.: Inferring from an Imprecise Plackett–Luce Model: Application to Label Ranking. In: Springer (ed.) 14th International Conference on Scalable Uncertainty Management (SUM 2020). Lecture Notes in Computer science, vol. 12322, pp. 98–112. Bolzano, Italy (Sep 2020)

[2] Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika **39**(3/4), 324–345 (1952)

[3] Caron, F., Doucet, A.: Efficient bayesian inference for generalized bradley—terry models. Journal of Computational and Graphical Statistics **21**(1), 174–196 (2012)

[4] Cheng, W., Hüllermeier, E., Waegeman, W., Welker, V.: Label ranking with partial abstention based on thresholded probabilistic models. Advances in neural information processing systems **25** (2012)

[5] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological) **39**(1), 1–38 (1977)

[6] Guiver, J., Snelson, E.: Bayesian inference for plackett-luce ranking models. In: Proceedings of the 26th Annual International Conference on Machine Learning. p. 377–384. ICML '09, Association for Computing Machinery, New York, NY, USA (2009). https://doi.org/10.1145/1553374.1553423

[7] He, J., Li, X., Liao, L.: Category-aware next point-of-interest recommendation via listwise bayesian personalized ranking. In: IJCAI. vol. 17, pp. 1837–1843 (2017)

[8] Hunter, D.R.: MM algorithms for generalized Bradley-Terry models. The Annals of Statistics **32**(1), 384 – 406 (2004). https://doi.org/10.1214/aos/1079120141

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

[10] Plackett, R.L.: The analysis of permutations. Journal of the Royal Statistical Society: Series C (Applied Statistics) **24**(2), 193–202 (1975). https://doi.org/https://doi.org/10.2307/2346567

[11] Tornede, A., Wever, M., Hüllermeier, E.: Algorithm selection as recommendation: From collaborative filtering to dyad ranking. In: CI Workshop, Dortmund (2019)

[12] Yellott, J.I.: 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). https://doi.org/https://doi.org/10.1016/0022-2496(77)90026-8

[13] Zermelo, E.F.F.: Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. Mathematische Zeitschrift **29**, 436–460 (1929)