On Bayesian inference for the Extended Plackett-Luce model

Stephen R. Johnson[*], Daniel A. Henderson[†], and Richard J. Boys[‡]

Abstract. The analysis of rank ordered data has a long history in the statistical literature across a diverse range of applications. In this paper we consider the Extended Plackett-Luce model that induces a flexible (discrete) distribution over permutations. The parameter space of this distribution is a combination of potentially high-dimensional discrete and continuous components and this presents challenges for parameter interpretability and also posterior computation. Particular emphasis is placed on the interpretation of the parameters in terms of observable quantities and we propose a general framework for preserving the mode of the prior predictive distribution. Posterior sampling is achieved using an effective simulation based approach that does not require imposing restrictions on the parameter space. Working in the Bayesian framework permits a natural representation of the posterior predictive distribution and we draw on this distribution to make probabilistic inferences and also to identify potential lack of model fit. The flexibility of the Extended Plackett-Luce model along with the effectiveness of the proposed sampling scheme are demonstrated using several simulation studies and real data examples.

Keywords: Markov chain Monte Carlo, MC^3, permutations, predictive inference, rank ordered data.

1 Introduction

Rank ordered data arise in many areas of application and a wide range of models have been proposed for their analysis; for an overview see Marden (1995) and Alvo and Yu (2014). In this paper we focus on the Extended Plackett-Luce (EPL) model proposed by Mollica and Tardella (2014); this model is a flexible generalisation of the popular Plackett-Luce model (Luce, 1959; Plackett, 1975) for permutations. In the Plackett-Luce model, entity $k \in \mathcal{K} = \{1, \ldots, K\}$ is assigned parameter $\lambda_k > 0$, and the probability of observing the ordering $\boldsymbol{x} = (x_1, x_2, \ldots, x_K)'$ (where x_j denotes the entity ranked in position j) given the entity parameters (referred to as "*support parameters*" in Mollica and Tardella (2014)) $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)'$ is

$$\Pr(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{\lambda}) = \prod_{j=1}^{K} \frac{\lambda_{x_j}}{\sum_{m=j}^{K} \lambda_{x_m}}. \tag{1}$$

[*]School of Mathematics, Statistics and Physics, Newcastle University, UK, stephen.johnson1@ncl.ac.uk

[†]School of Mathematics, Statistics and Physics, Newcastle University, UK, daniel.henderson@ncl.ac.uk

[‡]School of Mathematics, Statistics and Physics, Newcastle University, UK, richard.boys@ncl.ac.uk

We refer to (1) as the *standard* Plackett-Luce (SPL) probability. This probability is constructed via the so-called "*forward ranking process*" (Mollica and Tardella, 2014), that is, it is assumed that a rank ordering is formed by allocating entities from most to least preferred. This is a rather strong assumption. It is easy to imagine a scenario where an individual ranker might assign entities to positions/ranks in an alternative way. For example, it is quite plausible that rankers may find it easier to identify their most and least preferred entities first rather than those entities they place in the middle positions of their ranking (Mollica and Tardella, 2018). In such a scenario rankers might form their rank ordering by first assigning their most and then least preferred entities to a rank before filling out the middle positions through a process of elimination using the remaining (unallocated) entities, that is, they use a different *ranking process*. The Extended Plackett-Luce model relaxes the assumption of a fixed and known ranking process.

It is somewhat natural to recast the underlying ranking process in terms of a "*choice order*" where the choice order is the order in which rankers assign entities to positions/ranks. For example, suppose a ranker must provide a preference ordering of K entities; a choice order of $\boldsymbol{\sigma} = (1, K, 2, 3, \ldots, K-1)'$ corresponds to the ranking process where the most preferred entity is assigned first, followed by the least preferred entity and then the remaining entities are assigned in rank order from second down. Note that the choice order $\boldsymbol{\sigma}$ is simply a permutation of the ranks 1 to K and is referred to as the "*reference order*" in Mollica and Tardella (2014).

Whilst the EPL model is motivated in terms of a *choice order* as described above, this justification is not always appropriate. For example, the notion of a choice order clearly does not apply in the analysis of the Formula 1 data in Section 6, where the data are simply the finishing orders of the drivers in each race. We prefer to view the EPL model as a flexible probabilistic model for rank ordered data; ultimately all such probabilistic models induce a discrete distribution $P_{\boldsymbol{x}}$ over the set of all $K!$ permutations $\mathcal{S}_K$ and we wish this distribution to provide a flexible model for the observed data.

We adopt a Bayesian approach to inference which we find particularly appealing and natural as we focus heavily on predictive inference for observable quantities. We also make three main contributions, as outlined below. When the number of entities is not small, choosing a suitable prior distribution for $\boldsymbol{\sigma}$, the permutation of the ranks 1 to K, is a somewhat daunting task. We therefore propose to use the (standard) Plackett-Luce model to define the prior probability of each permutation, although we note that our inference framework is sufficiently general and does not rely on this choice. We also address the thorny issue of specifying informative prior beliefs about the entity parameters $\boldsymbol{\lambda}$ by proposing a class of priors that preserve the modal prior predictive rank ordering under different choice orders $\boldsymbol{\sigma}$. Constructing suitable posterior sampling schemes for the Extended Plackett-Luce model is challenging; the parameters $(\boldsymbol{\lambda}, \boldsymbol{\sigma}) \in \mathbb{R}^K_{>0} \times \mathcal{S}_K$ reside in a complicated mixed discrete-continuous space that must be effectively explored. This is made more challenging by the multi-modality of the marginal posterior distribution for $\boldsymbol{\sigma}$, with local modes separated by large distances within permutation space. We overcome the difficult sampling problem by appealing to Metropolis coupled Markov chain Monte Carlo (MC^3). To the best of our knowledge, the algorithm we

outline is the only one currently capable of sampling from the posterior distribution under the Extended Plackett-Luce model when the full parameter space is considered. Unfortunately the recent algorithm proposed by Mollica and Tardella (2020) relies on a rescaling of the entity parameters that does not preserve the target distribution; the solution given by Mollica and Tardella (2018) considers a restricted parameter space for $\boldsymbol{\sigma}$ and suffers from the same issue.

The remainder of the paper is structured as follows. In Section 2 we outline the Extended Plackett-Luce model and our associated notation, and in Section 2.2 we provide some guidance on interpreting the model parameters. In Section 3 we propose our Bayesian approach to inference. In particular we discuss suitable choices for the prior distribution and describe our simulation based scheme for posterior sampling. A simulation study illustrating the efficacy of the posterior sampling scheme and the performance of the EPL model over a range of values for the number of entities and number of observations is considered in Section 4; with further details also given in Section 2 of the supplementary material (Johnson et al., 2021). Section 5 outlines how we use the posterior predictive distribution for inference on observable quantities and for assessing the appropriateness of the model. Two real data analyses are considered in Section 6 to illustrate the use of the (unrestricted) EPL model. Section 7 offers some conclusions.

2   The Extended Plackett-Luce model

We now present the Extended Plackett-Luce model along with our associated notation and also discuss the interpretation of the model parameters in terms of the preferences of entities.

2.1   Model and notation

Recall that there are K entities to be ranked and that the collection of all entities is denoted $\mathcal{K} = \{1, \ldots, K\}$. The Extended Plackett-Luce model we consider is that proposed by Mollica and Tardella (2014) and, in its current form, is only appropriate for *complete* rank orderings in which all entities are included. Thus a typical observation is $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iK})'$ where x_{ij} denotes the entity ranked in position j in the ith rank ordering. The choice order is represented by $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_K)'$, where σ_j denotes the rank allocated at the jth stage. Conditional on $\boldsymbol{\sigma}$, each entity has a corresponding parameter $\lambda_k > 0$ for $k = 1, \ldots, K$; let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)'$. Crucially, the meaning and interpretation of $\boldsymbol{\lambda}$ depends on $\boldsymbol{\sigma}$ and this is addressed shortly.

The probability of a particular rank ordering under the Extended Plackett-Luce model (Mollica and Tardella, 2014) is defined as

$$\Pr(\boldsymbol{X}_i = \boldsymbol{x}_i | \boldsymbol{\lambda}, \boldsymbol{\sigma}) = \prod_{j=1}^{K} \frac{\lambda_{x_{i\sigma_j}}}{\sum_{m=j}^{K} \lambda_{x_{i\sigma_m}}}. \tag{2}$$

Therefore, the Extended Plackett-Luce probability (2) is simply the standard Plackett-Luce probability (1) evaluated with entity parameters $\boldsymbol{\lambda}$ and "permuted data" $\boldsymbol{x}_i^* =$

$\boldsymbol{x}_i \circ \boldsymbol{\sigma}$ where $\circ$ denotes composition of permutations which, in terms of vectors, implies that if $\boldsymbol{z} = \boldsymbol{x} \circ \boldsymbol{y}$ then $z_i = x_{y_i}$. Here x_{ij}^* denotes the entity chosen at the jth stage of the ith ranking process and therefore receiving rank σ_j.

Indeed, both the (forward ranking) standard Plackett-Luce model and (backward ranking) reverse Plackett-Luce model are special cases of (2) and are recovered when $\boldsymbol{\sigma} = (1, \ldots, K)' \equiv \mathcal{I}$, the identity permutation, and $\boldsymbol{\sigma} = (K, K-1, \ldots, 1)'$, the reverse of the identity permutation, respectively. We use the notation $\boldsymbol{X}_i | \boldsymbol{\lambda}, \boldsymbol{\sigma} \sim \text{EPL}(\boldsymbol{\lambda}, \boldsymbol{\sigma})$ to denote that the probability of rank ordering i is given by (2). Note that here and throughout we have adopted different notation from that in Mollica and Tardella (2014) however the essential components of the model remain unchanged.

It is clear that the EPL probability (2) is invariant to scalar multiplication of the entity parameters $\boldsymbol{\lambda}$, that is, $\Pr(\boldsymbol{X} | \boldsymbol{\lambda}, \boldsymbol{\sigma}) = \Pr(\boldsymbol{X} | c\boldsymbol{\lambda}, \boldsymbol{\sigma})$ for any positive constant $c > 0$. From a Bayesian perspective this type of identifiability issue is straightforwardly resolved by placing a non-uniform prior on $\boldsymbol{\lambda}$. In this case, although the likelihood function may be constant for multiple (scaled) $\boldsymbol{\lambda}$ values, the posterior distribution is not. That said, this issue can still lead to potential mixing problems for MCMC algorithms (Caron and Doucet, 2012) and thus an inefficient sampling scheme; this is revisited in Section 3.3.

2.2   Interpretation of the entity parameters $\boldsymbol{\lambda}$

A key aspect of analysing rank ordered data using Plackett-Luce type models is the interpretation of the entity parameters $\boldsymbol{\lambda}$. Moreover, it is essential to understand the interpretation of the λ parameters if one is to specify informative prior beliefs about the likely preferences of the entities.

For the Extended Plackett-Luce model, λ_k is proportional to the probability that entity k is selected at the first *stage* of the ranking process and therefore ranked in position σ_1 of the rank ordering $\boldsymbol{x}$. Then, conditional on an entity being assigned to position σ_1 in the rank ordering, the entity with the largest parameter of those remaining is that most likely to be assigned to position σ_2, and so on. For the standard Plackett-Luce model, arising from the forward ranking process with $\boldsymbol{\sigma} = \mathcal{I}$, we have that λ_k is proportional to the probability that entity k is assigned rank $\sigma_1 = 1$ (and is thus the most preferred entity), and so on. Therefore, for the standard Plackett-Luce model, entities with larger values are more likely to be given a higher rank. In other words, the λ parameters for the standard Plackett-Luce model correspond directly with *preferences* for entities. A consequence is that ordering the entities in terms of their values in $\boldsymbol{\lambda}$, from largest to smallest, will give the modal ordering $\hat{\boldsymbol{x}}$, that is, the permutation of the entities which yields the maximum Plackett-Luce probability (1), given $\boldsymbol{\lambda}$. Specifically, $\hat{\boldsymbol{x}} = \text{order}_{\downarrow}(\boldsymbol{\lambda})$, where $\text{order}_{\downarrow}(\cdot)$ denotes the ordering operation from largest to smallest. This makes specifying a prior distribution for $\boldsymbol{\lambda}$, when $\boldsymbol{\sigma} = \mathcal{I}$, relatively straightforward based on entity preferences. The interpretation of the λ parameters directly in terms of preferences can also be achieved in a straightforward manner with the backward ranking process ($\boldsymbol{\sigma} = (K, \ldots, 1)'$) of the reverse Plackett-Luce model. Apart from these special cases, however, the interpretation of the λ parameters in terms of preferences is not at all transparent for other choices of $\boldsymbol{\sigma}$. For example, suppose that $\lambda_i > \lambda_j$ and

$\boldsymbol{\sigma} = (2, 3, 1)'$. Here entity i is more likely to be ranked in second position than entity j. Further, if another entity $\ell \neq i, j$, is assigned to rank 2 then entity i is preferred for rank 3 (σ_2) over entity j.

Understanding the preference of the entities under the Extended Plackett-Luce model based on values of $\boldsymbol{\lambda}$ and $\boldsymbol{\sigma}$ can be made more straightforward if we first introduce the inverse of the choice order permutation $\boldsymbol{\sigma}^{-1}$. This is defined such that $\boldsymbol{\sigma} \circ \boldsymbol{\sigma}^{-1} = \boldsymbol{\sigma}^{-1} \circ \boldsymbol{\sigma} = \mathcal{I}$, the identity permutation. Here σ_j^{-1} denotes the stage of the ranking process at which rank j is assigned. We can then obtain directly the modal ordering of the entities under the EPL($\boldsymbol{\lambda}, \boldsymbol{\sigma}$) model, $\hat{\boldsymbol{x}}^{(\boldsymbol{\sigma}, \boldsymbol{\lambda})}$, and thus obtain a representation of the preference of the entities. Here $\hat{\boldsymbol{x}}^{(\boldsymbol{\sigma}, \boldsymbol{\lambda})}$ is obtained without enumerating any probabilities by permuting the entries in $\hat{\boldsymbol{x}}$ (the modal ordering under the standard Plackett-Luce model conditional on $\boldsymbol{\lambda}$), by $\boldsymbol{\sigma}^{-1}$, that is $\hat{\boldsymbol{x}}^{(\boldsymbol{\sigma}, \boldsymbol{\lambda})} = \hat{\boldsymbol{x}} \circ \boldsymbol{\sigma}^{-1}$. In other words, if $\hat{\boldsymbol{x}} = \mathrm{order}_{\downarrow}(\boldsymbol{\lambda})$ then $\hat{x}_j^{(\boldsymbol{\sigma}, \boldsymbol{\lambda})} = \hat{x}_{\sigma_j^{-1}}$, for $j = 1, \ldots, K$. Let $\hat{\boldsymbol{x}}^{-1}$ represent the ranks assigned to the entities under the standard Plackett-Luce model; this is obtained as the inverse permutation corresponding to $\hat{\boldsymbol{x}}$, that is, the permutation such that $\hat{\boldsymbol{x}}^{-1} \circ \hat{\boldsymbol{x}} = \mathcal{I}$. Now define $\boldsymbol{\eta}^{(\boldsymbol{\sigma})} = \hat{\boldsymbol{x}}^{(\boldsymbol{\sigma}, \boldsymbol{\lambda})} \circ \hat{\boldsymbol{x}}^{-1}$; this represents the permutation of the entities ranked under the EPL model at the stage corresponding to the rank assigned to entities 1 to K under the standard Plackett-Luce model. It follows that, if $\boldsymbol{\lambda}^{(\boldsymbol{\sigma})}$ has jth element $\lambda_j^{(\boldsymbol{\sigma})} = \boldsymbol{\lambda}_{\eta_j^{(\boldsymbol{\sigma})}}$, where $\eta_j^{(\boldsymbol{\sigma})}$ is the jth element of $\boldsymbol{\eta}^{(\boldsymbol{\sigma})}$, then $\hat{\boldsymbol{x}}^{(\boldsymbol{\sigma}, \boldsymbol{\lambda}^{\boldsymbol{\sigma}})} \equiv \hat{\boldsymbol{x}}$ for all $\boldsymbol{\sigma} \in \mathcal{S}_K$, and the modal preference ordering is preserved.

Some simplification is possible if we first order the entities in terms of preferences. Clearly, with $\hat{\boldsymbol{x}} = \mathcal{I}$ we have $\hat{\boldsymbol{x}}^{(\boldsymbol{\sigma}, \boldsymbol{\lambda})} = \boldsymbol{\sigma}^{-1}$ and so the modal ordering is given by the inverse choice order permutation. Moreover, if $\hat{\boldsymbol{x}} = \mathcal{I}$ then $\hat{\boldsymbol{x}}^{-1} = \mathcal{I}$ and so $\boldsymbol{\eta}^{(\boldsymbol{\sigma})} = \boldsymbol{\sigma}^{-1}$. It follows that choosing $\boldsymbol{\lambda}^{(\boldsymbol{\sigma})}$ such that its jth element is $\lambda_j^{(\boldsymbol{\sigma})} = \boldsymbol{\lambda}_{\sigma_j^{-1}}$, then $\hat{\boldsymbol{x}}^{(\boldsymbol{\sigma}, \boldsymbol{\lambda}^{\boldsymbol{\sigma}})} \equiv \mathcal{I}$ for all $\boldsymbol{\sigma} \in \mathcal{S}_K$. Therefore if the entities are labelled in preference order then permuting the λ parameters from the standard Plackett-Luce model by the inverse of the choice order permutation will preserve the modal permutation to be in the same preference order. This suggests a simple strategy for specifying prior distributions for the entity parameters which preserves modal preferences under different choice orders; we revisit this in Section 3.1.

3 Bayesian modelling

Suppose we have data consisting of n independent rank orderings, collectively denoted $\mathcal{D} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$. The likelihood of $\boldsymbol{\lambda}, \boldsymbol{\sigma}$ is

$$\pi(\mathcal{D}|\boldsymbol{\lambda}, \boldsymbol{\sigma}) = \prod_{i=1}^{n} \mathrm{Pr}(\boldsymbol{X}_i = \boldsymbol{x}_i | \boldsymbol{\lambda}, \boldsymbol{\sigma})$$

$$= \prod_{i=1}^{n} \prod_{j=1}^{K} \frac{\lambda_{x_{i\sigma_j}}}{\sum_{m=j}^{K} \lambda_{x_{i\sigma_m}}}. \tag{3}$$

We wish to make inferences about the unknown quantities in the model $\boldsymbol{\sigma}, \boldsymbol{\lambda}$ as well as future observable rank orderings $\tilde{\boldsymbol{x}}$. Specifically we adopt a Bayesian approach to

inference in which we quantify our uncertainty about the unknown quantities (before observing the data) through a suitable prior distribution.

3.1   Prior specification

We adopt a joint prior distribution for $\boldsymbol{\sigma}$ and $\boldsymbol{\lambda}$ of the form $\pi(\boldsymbol{\sigma}, \boldsymbol{\lambda}) = \pi(\boldsymbol{\lambda}|\boldsymbol{\sigma})\pi(\boldsymbol{\sigma})$ which explicitly emphasizes the dependence of $\boldsymbol{\lambda}$ on $\boldsymbol{\sigma}$.

Prior for σ

For the choice ordering $\boldsymbol{\sigma}$ we need to define a discrete distribution $P_{\boldsymbol{\sigma}}$ over the $K!$ elements of $\mathcal{S}_K$. If K is not small, perhaps larger than 4, then this could be a rather daunting task. Given the choice order parameter $\boldsymbol{\sigma}$ is a permutation, or equivalently a complete rank ordering, one flexible option is to use the Plackett-Luce model to define the prior probabilities for each choice order parameter. More specifically we let $\boldsymbol{\sigma}|\boldsymbol{q} \sim \mathrm{PL}(\boldsymbol{q})$ where $\boldsymbol{q} = (q_1, \ldots, q_K)' \in \mathbb{R}_{>0}^K$ are to be chosen *a priori* and

$$\mathrm{Pr}(\boldsymbol{\sigma}|\boldsymbol{q}) = \prod_{j=1}^{K} \frac{q_{\sigma_j}}{\sum_{m=j}^{K} q_{\sigma_m}}.$$

If desired, it is straightforward to assume each choice order is equally likely *a priori* by letting $q_k = q$ for $k = 1, \ldots, K$. Furthermore, the inference framework that follows is sufficiently general and does not rely on this prior choice. In particular, if we only wish to consider a subset of all the possible choice orderings $\mathcal{R} \subset \mathcal{S}_K$, as in Mollica and Tardella (2018), then this can be achieved by making an appropriate choice of prior probabilities for all $\boldsymbol{\sigma} \in \mathcal{R}$ and letting $\mathrm{Pr}(\boldsymbol{\sigma}) = 0$ for all $\boldsymbol{\sigma} \in \mathcal{S}_K \setminus \mathcal{R}$.

Prior for $\lambda|\sigma$

It is natural to wish to specify prior beliefs in terms of preferences for the entities. However, we have seen in Section 2.2 that the interpretation of the entity parameters $\boldsymbol{\lambda}$ in terms of preferences is dependent on the value of $\boldsymbol{\sigma}$. It follows that specifying an informative prior for the entity parameters is problematic unless the choice order $\boldsymbol{\sigma}$ is assumed to be known. We therefore consider separate prior distributions for $\boldsymbol{\lambda}$ conditional on the value of $\boldsymbol{\sigma}$. Since the entity parameters $\lambda_k > 0$ must be strictly positive, a suitable, relatively tractable, choice of conditional prior distribution is a gamma distribution with mean $a_k^{(\boldsymbol{\sigma})}/b_k^{(\boldsymbol{\sigma})}$, that is $\lambda_k|\boldsymbol{\sigma} \overset{indep}{\sim} \mathrm{Ga}(a_k^{(\boldsymbol{\sigma})}, b_k^{(\boldsymbol{\sigma})})$ for $k = 1, \ldots, K$ and $\boldsymbol{\sigma} \in \mathcal{S}_K$. Without loss of generality we set $b_k^{(\boldsymbol{\sigma})} = b = 1$, for all k and $\boldsymbol{\sigma}$ since b is not likelihood identifiable. Our proposed strategy for specifying the hyper-parameters $\boldsymbol{a}^{(\boldsymbol{\sigma})}$ is to first consider the prior distribution for $\boldsymbol{\lambda}$ under the standard Plackett-Luce model with $\boldsymbol{\sigma} = \mathcal{I}$. If we specify $\boldsymbol{a} = (a_1, \ldots, a_K)'$ then $\hat{\boldsymbol{x}}$, the modal preference ordering from the prior predictive distribution, is obtained by simply ordering the a_k values from largest to smallest and so $\hat{\boldsymbol{x}} = \mathrm{order}_\downarrow(\boldsymbol{a})$. Then, following the arguments of Section 2.2, we can

preserve our beliefs about the modal preference ordering for all values of $\boldsymbol{\sigma}$ by constructing $\boldsymbol{\eta}^{(\boldsymbol{\sigma})} = \hat{\boldsymbol{x}} \circ \boldsymbol{\sigma}^{-1} \circ \hat{\boldsymbol{x}}^{-1}$ and letting $a_k^{(\boldsymbol{\sigma})} = a_{\eta_k^{(\boldsymbol{\sigma})}}$ for $k = 1, \ldots, K$ and $\boldsymbol{\sigma} \in \mathcal{S}_K$. Again some simplification of notation is achievable if we first relabel the entities in terms of our *a priori* preferences. Specifically if we label the entities such that our most preferred entity is labelled 1, second most preferred labelled 2, and so on then $a_1 > a_2 > \cdots > a_K$ and $\hat{\boldsymbol{x}} = \mathcal{I}$. In this case $\boldsymbol{\eta}^{(\boldsymbol{\sigma})} = \boldsymbol{\sigma}^{-1}$ and so $a_k^{(\boldsymbol{\sigma})} = a_{\boldsymbol{\sigma}_k^{-1}}$ for $k = 1, \ldots, K$. Section 8 of the supplementary material contains two simple numerical examples to help illustrate the issues with prior specification and our proposed mode-preserving solution.

Whilst the framework above gives us a relatively straightforward way to encode information if we have a clear prior preference of the entities, additional complications arise in settings where preferences for some (or all) of the entities are exchangeable *a priori*. More specifically, if we wish to specify no (prior) preference between two distinct entities $j \neq k$ then, for the standard Plackett-Luce model, this can be achieved by setting $a_j = a_k$. Such choice gives rise to a multi-modal prior predictive distribution and so $\hat{\boldsymbol{x}} = \text{order}_\downarrow(\boldsymbol{a})$, and thus $\boldsymbol{\eta}^{(\boldsymbol{\sigma})} = \hat{\boldsymbol{x}} \circ \boldsymbol{\sigma}^{-1} \circ \hat{\boldsymbol{x}}^{-1}$, is not uniquely defined. Thankfully this issue can be straightforwardly resolved as part of our simulation-based inference approach. In particular if we let $\mathcal{X}$ denote the set of modes of the prior predictive distribution (under the SPL model) then we can simply sample (uniformly at random) $\hat{\boldsymbol{x}}$ from $\mathcal{X}$ at each iteration of our MCMC scheme. A formal discussion is provided in Section 4 of the supplementary material. Note that if we are unwilling to favour any particular preference ordering *a priori* then this can be achieved by letting $a_k = a$ (for all k) as this choice induces a uniform prior predictive distribution over all preference orders. Formally $\mathcal{X} = \mathcal{S}_K$, however given that $\boldsymbol{a}^{(\boldsymbol{\sigma})}$ is formed by permuting the elements of $\boldsymbol{a}$ it follows that $\boldsymbol{a}^{(\boldsymbol{\sigma})} = \boldsymbol{a}$ for all $\hat{\boldsymbol{x}}, \boldsymbol{\sigma} \in \mathcal{S}_K$ and so sampling $\hat{\boldsymbol{x}}$ from $\mathcal{X}$ is redundant in this special case.

3.2 Bayesian model

The complete Bayesian model is

$$\boldsymbol{X}_i | \boldsymbol{\lambda}, \boldsymbol{\sigma} \overset{indep}{\sim} \text{EPL}(\boldsymbol{\lambda}, \boldsymbol{\sigma}), \qquad i = 1, \ldots, n,$$

$$\lambda_k | \boldsymbol{\sigma}, \boldsymbol{a} \overset{indep}{\sim} \text{Ga}(a_k^{(\boldsymbol{\sigma})}, 1), \qquad k = 1, \ldots, K,$$

$$\boldsymbol{\sigma} | \boldsymbol{q} \sim \text{PL}(\boldsymbol{q}).$$

That is, we assume that our observations follow the distribution specified by the Extended Plackett-Luce model (2) and the prior distribution for $(\boldsymbol{\lambda}, \boldsymbol{\sigma})$ is as described in Section 3.1.

The full joint density of all stochastic quantities in the model (with dependence on fixed hyper-parameters suppressed) is

$$\pi(\boldsymbol{\sigma}, \boldsymbol{\lambda}, \mathcal{D}) = \pi(\mathcal{D} | \boldsymbol{\lambda}, \boldsymbol{\sigma}) \pi(\boldsymbol{\lambda} | \boldsymbol{\sigma}) \pi(\boldsymbol{\sigma}).$$

From which we quantify our beliefs about $\boldsymbol{\sigma}$ and $\boldsymbol{\lambda}$ through their joint posterior density

$$\pi(\boldsymbol{\sigma}, \boldsymbol{\lambda} | \mathcal{D}) \propto \pi(\mathcal{D} | \boldsymbol{\lambda}, \boldsymbol{\sigma}) \pi(\boldsymbol{\lambda} | \boldsymbol{\sigma}) \pi(\boldsymbol{\sigma}),$$

which is obtained via Bayes' Theorem. The posterior density $\pi(\boldsymbol{\sigma}, \boldsymbol{\lambda} | \mathcal{D})$ is not available in closed form and so we use simulation-based methods to sample from the posterior distribution as described in the next section.

3.3   Posterior sampling

Due to the complex nature of the posterior distribution we use Markov chain Monte Carlo (MCMC) methods in order to sample realisations from $\pi(\boldsymbol{\sigma}, \boldsymbol{\lambda} | \mathcal{D})$. The structure of the model lends itself naturally to consider sampling alternately from two blocks of full conditional distributions: $\pi(\boldsymbol{\sigma} | \boldsymbol{\lambda}, \mathcal{D})$ and $\pi(\boldsymbol{\lambda} | \boldsymbol{\sigma}, \mathcal{D})$.

Sampling the choice order parameter $\boldsymbol{\sigma}$ from $\pi(\boldsymbol{\sigma} | \boldsymbol{\lambda}, \mathcal{D})$

Given the choice order parameter $\boldsymbol{\sigma}$ is a member of $\mathcal{S}_K$ it is fairly straightforward to obtain its (discrete) full conditional distribution; specifically this is the discrete distribution with probabilities

$$\Pr(\boldsymbol{\sigma} = \boldsymbol{\sigma}_j | \boldsymbol{\lambda}, \mathcal{D}) \propto \pi(\mathcal{D} | \boldsymbol{\lambda}, \boldsymbol{\sigma} = \boldsymbol{\sigma}_j) \pi(\boldsymbol{\lambda} | \boldsymbol{\sigma} = \boldsymbol{\sigma}_j) \Pr(\boldsymbol{\sigma} = \boldsymbol{\sigma}_j)$$

for $j = 1, \ldots, K!$. Clearly sampling from this full conditional will require $K!$ evaluations of the EPL likelihood $\pi(\mathcal{D} | \boldsymbol{\lambda}, \boldsymbol{\sigma} = \boldsymbol{\sigma}_j)$ and so sampling from $\Pr(\boldsymbol{\sigma} = \boldsymbol{\sigma}_j | \boldsymbol{\lambda}, \mathcal{D})$ for $j = 1, \ldots, K!$ (a Gibbs update) is probably only plausible if K is sufficiently small; perhaps not much greater than 5. Of course, the probabilities $\Pr(\boldsymbol{\sigma} = \boldsymbol{\sigma}_i | \boldsymbol{\lambda}, \mathcal{D})$ and $\Pr(\boldsymbol{\sigma} = \boldsymbol{\sigma}_j | \boldsymbol{\lambda}, \mathcal{D})$ are conditionally independent for $i \neq j$ and so could be computed in parallel which may facilitate this approach for slightly larger values of K.

So as to free ourselves from the restriction to the case where K is small we instead consider a more general sampling strategy by constructing a Metropolis-Hastings proposal mechanism for updating $\boldsymbol{\sigma}$. Our investigation into the likelihood of the Extended Plackett-Luce model given different choice orders in Section 3 of the supplementary material revealed that $\pi(\mathcal{D} | \boldsymbol{\lambda}, \boldsymbol{\sigma})$ is likely to be multi-modal. Further, local modes can be separated by large distances within permutation space. In an attempt to effectively explore this large discrete space we consider 5 alternative proposal mechanisms; each of which occurs with probability p_ℓ for $\ell = 1, \ldots, 5$. The 5 mechanisms to construct the proposed permutation $\boldsymbol{\sigma}^\dagger$ are as follows.

1. The random swap: sample two positions $\phi_1, \phi_2 \in \{1, \ldots, K\}$ uniformly at random and let the proposed choice order $\boldsymbol{\sigma}^\dagger$ be the current choice order $\boldsymbol{\sigma}$ where the elements in positions ϕ_1 and ϕ_2 have been swapped.

2. The Poisson swap: sample $\phi_1 \in \{1, \ldots, K\}$ uniformly at random and let $\phi_2 = \phi_1 + m$ where $m = (-1)^\tau f$, $\tau \sim \text{Bern}(0.5)$ and $f \sim \text{Poisson}(t)$. Note that t is a tuning parameter and $\phi_2 \to \{(\phi_2 - 1) \bmod K\} + 1$ as appropriate. Again the proposed choice order $\boldsymbol{\sigma}^\dagger$ is formed by swapping the elements in positions ϕ_1 and ϕ_2 of the current choice order $\boldsymbol{\sigma}$.

3. The random insertion (Bezáková et al., 2006): sample two positions $\phi_1 \neq \phi_2 \in \{1, \ldots, K\}$ uniformly at random and let the proposed choice order $\boldsymbol{\sigma}^\dagger$ be formed by taking the value in position ϕ_1 and inserting it back into the permutation so that it is instead in position ϕ_2.

4. The prior proposal: here $\boldsymbol{\sigma}^\dagger$ is simply an independent draw from the prior distribution, that is, $\boldsymbol{\sigma}^\dagger | \boldsymbol{q} \sim \mathrm{PL}(\boldsymbol{q})$.

5. The reverse proposal: here $\boldsymbol{\sigma}^\dagger$ is defined to be the reverse ordering of the current permutation $\boldsymbol{\sigma}$, that is, $\boldsymbol{\sigma}^\dagger = \boldsymbol{\sigma}_{K:1} = (\sigma_K, \ldots, \sigma_1)'$.

Note that performing either of the swap or insertion moves (1–3) above may result in slow exploration of the set of all permutations as the proposal $\boldsymbol{\sigma}^\dagger$ may not differ much from the current value $\boldsymbol{\sigma}$. To alleviate this potential issue we propose to iteratively perform each of these moves S times, where S is to be chosen (and fixed) by the analyst. More formally (when using proposal mechanisms 1–3) we construct a sequence of intermediate proposals $\boldsymbol{\sigma}_s^\dagger$ from the "current" choice order $\boldsymbol{\sigma}_{s-1}^\dagger$. In particular we let $\boldsymbol{\sigma}_0^\dagger = \boldsymbol{\sigma}$ and generate $\boldsymbol{\sigma}_s^\dagger | \boldsymbol{\sigma}_{s-1}^\dagger$ for $s = 1, \ldots, S$; the proposed value for which we evaluate the acceptance probability is $\boldsymbol{\sigma}^\dagger = \boldsymbol{\sigma}_S^\dagger$. Further, for moves 1 and 2 it may seem inefficient to allow for the "null swap" $\phi_1 = \phi_2$, however this is done to avoid only proposing permutations with the same (or opposing) parity as the current value. Put another way, as $S \to \infty$ we would expect $\mathrm{Pr}(\boldsymbol{\sigma}^\dagger | \boldsymbol{\sigma}) > 0$ for all $\boldsymbol{\sigma}^\dagger, \boldsymbol{\sigma} \in \mathcal{S}_K$ and this only holds if we allow for the possibility that $\phi_1 = \phi_2$. Finally we note that each of these proposal mechanisms is "simple" in the respect that $\mathrm{Pr}(\boldsymbol{\sigma}^\dagger | \boldsymbol{\sigma}) = \mathrm{Pr}(\boldsymbol{\sigma} | \boldsymbol{\sigma}^\dagger)$ and so the proposal ratio cancels in each case. The full acceptance ratio is presented within the algorithm outline in Section 3.4.

Sampling the entity parameters $\boldsymbol{\lambda}$ from $\pi(\boldsymbol{\lambda} | \boldsymbol{\sigma}, \mathcal{D})$

Bayesian inference for variants of standard Plackett-Luce models typically proceeds by first introducing appropriate versions of the latent variables proposed by Caron and Doucet (2012), which in turn facilitate a Gibbs update for each of the entity parameters (assuming independent Gamma prior distributions are chosen). However we found that, when coupled with a Metropolis-Hastings update on $\boldsymbol{\sigma}$, this strategy does not work well for parameter inference under the Extended Plackett-Luce model (not reported here); this observation was also noted by Mollica and Tardella (2020). We therefore propose to use a Metropolis-Hastings step for sampling the entity parameters, specifically we use (independent) log normal random walks for each entity parameter in turn and so the proposed value is $\lambda_k^\dagger \overset{indep}{\sim} \mathrm{LN}(\log \lambda_k, \sigma_k^2)$ for $k = 1, \ldots, K$. We also implement an additional sampling step in the MCMC scheme, analogous to that in Caron and Doucet (2012), in order to mitigate the poor mixing that is caused by the invariance of the Extended Plackett-Luce likelihood to scalar multiplication of the $\boldsymbol{\lambda}$ parameters. Note that this step does not equate to an arbitrary rescaling of the $\boldsymbol{\lambda}$ parameters but rather corresponds to a draw of $\Lambda = \sum_k \lambda_k$ from its full conditional distribution. Full details are given in Section 3.4.

Metropolis coupled Markov chain Monte Carlo

Unfortunately the sampling strategy described above in the previous two subsections proves ineffective with the Markov chain suffering from poor mixing, particularly for $\boldsymbol{\sigma}$ where the chain is prone to becoming stuck in local modes (results not reported here). In an attempt to resolve these issues, and therefore aid the exploration of the posterior distribution we appeal to Metropolis coupled Markov chain Monte Carlo, or *parallel tempering.*

Metropolis coupled Markov chain Monte Carlo (Geyer, 1991), is a sampling technique that aims to improve the mixing of Markov chains in comparison to standard MCMC methods particularly when the target distribution is multi-modal (Gilks and Roberts, 1996; Brooks, 1998). The basic premise is to consider C chains evolving simultaneously, each of which targets a tempered posterior distribution $\tilde{\pi}_c(\theta_c|\mathcal{D}) \propto \pi(\mathcal{D}|\theta_c)^{1/T_c}\pi(\theta_c)$, where, for each chain c, $T_c \geq 1$ denotes the temperature and $\theta_c = \{\boldsymbol{\sigma}_c, \boldsymbol{\lambda}_c\}$ is the collection of unknown quantities. Note that the posterior of interest is recovered when $T_c = 1$. Further note that we have only considered a tempered likelihood component as we suggest that any prior beliefs should be consistent irrespective of the model choice. Now, as the posteriors $\tilde{\pi}_c(\theta_c|\mathcal{D})$ are conditionally independent given $\mathcal{D}$, we can consider them to be targeting the joint posterior

$$\pi(\theta_1, \ldots, \theta_C|\mathcal{D}) = \prod_{c=1}^{C} \tilde{\pi}_c(\theta_c|\mathcal{D}). \tag{4}$$

Suppose now we propose to swap θ_i and θ_j for some $i \neq j$ within a Markov chain targeting the joint posterior (4). If we let $\Theta = \{\theta_1, \ldots, \theta_C\}$ denote the current state and $\Theta^\dagger = \{\theta_1^\dagger, \ldots, \theta_C^\dagger\}$ the proposed state where $\theta_i^\dagger = \theta_j$, $\theta_j^\dagger = \theta_i$ and $\theta_\ell^\dagger = \theta_\ell$ for $\ell \neq i, j$, then, assuming a symmetric proposal mechanism, the acceptance probability of the state space swap is $\min(1, A)$ where

$$A = \frac{\pi(\mathcal{D}|\theta_j)^{1/T_i}\pi(\mathcal{D}|\theta_i)^{1/T_j}}{\pi(\mathcal{D}|\theta_i)^{1/T_i}\pi(\mathcal{D}|\theta_j)^{1/T_j}}.$$

Of course, if the proposal mechanism is not symmetric then the probability A must be multiplied by the proposal ratio $q(\Theta|\Theta^\dagger)/q(\Theta^\dagger|\Theta)$. Further, it is straightforward to generalise the above acceptance probability to allow the states of more than 2 chains to be swapped. However, this is typically avoided as such a proposal can result in poor acceptance rates. Our specific Metropolis coupled Markov chain Monte Carlo algorithm is outlined in the next section.

3.4 Outline of the posterior sampling algorithm

A parallel Metropolis coupled Markov chain Monte Carlo algorithm to sample from the joint posterior distribution of the parameters $\boldsymbol{\lambda}$ and the choice order parameter $\boldsymbol{\sigma}$ is as follows.

1. Tune:

   - choose the number of chains (C); let $T_1 = 1$ and choose $T_c > 1$ for $c = 2, \ldots, C$
   - choose appropriate values for the MH proposals outlined in Section 3.3

2. Initialise: take a prior draw or alternatively choose $\boldsymbol{\sigma}_c \in \mathcal{S}_K$ and $\boldsymbol{\lambda}_c \in \mathbb{R}_{>0}^K$ for $c = 1, \ldots, C$

3. For $c = 1, \ldots, C$ perform (in parallel) the following steps:

   - For $k = 1, \ldots, K$
     - draw $\lambda_{ck}^\dagger | \lambda_{ck} \sim \mathrm{LN}(\log \lambda_{ck}, \sigma_{\lambda_{ck}}^2)$
     - let $\lambda_{ck} \to \lambda_{ck}^\dagger$ with probability $\min(1, A)$ where

     $$A = \left\{ \frac{\pi(\mathcal{D} | \boldsymbol{\lambda}_{c,-k}, \lambda_{ck} = \lambda_{ck}^\dagger, \boldsymbol{\sigma}_c)}{\pi(\mathcal{D} | \boldsymbol{\lambda}_c, \boldsymbol{\sigma}_c)} \right\}^{1/T_c} \times \left(\frac{\lambda_{ck}^\dagger}{\lambda_{ck}} \right)^{a_k^{(\boldsymbol{\sigma}_c)}} e^{(\lambda_{ck} - \lambda_{ck}^\dagger)}$$

   - Sample ℓ with probabilities given by $\Pr(\ell = i) = p_{i,c}$ for $i = 1, \ldots, 5$
     - propose $\boldsymbol{\sigma}_c^\dagger$ using proposal mechanism ℓ
     - let $\boldsymbol{\sigma}_c \to \boldsymbol{\sigma}_c^\dagger$ with probability $\min(1, A)$ where

     $$A = \left\{ \frac{\pi(\mathcal{D} | \boldsymbol{\lambda}_c, \boldsymbol{\sigma}_c^\dagger)}{\pi(\mathcal{D} | \boldsymbol{\lambda}_c, \boldsymbol{\sigma}_c)} \right\}^{1/T_c} \times \frac{\pi(\boldsymbol{\lambda}_c | \boldsymbol{\sigma}_c^\dagger)}{\pi(\boldsymbol{\lambda}_c | \boldsymbol{\sigma}_c)} \frac{\Pr(\boldsymbol{\sigma}_c^\dagger)}{\Pr(\boldsymbol{\sigma}_c)}$$

   - Rescale
     - sample $\Lambda_c^\ddagger \sim \mathrm{Ga}\left(\sum_{k=1}^K a_k^{(\boldsymbol{\sigma}_c)}, 1 \right)$
     - calculate $\Sigma_c = \sum_{k=1}^K \lambda_{ck}$
     - let $\lambda_{ck} \to \lambda_{ck}\, \Lambda_c^\ddagger / \Sigma_c$ for $k = 1, \ldots, K$

4. Sample a pair of chain labels (i, j) where $1 \leq i \neq j \leq C$

   - let $(\boldsymbol{\lambda}_i, \boldsymbol{\sigma}_i) \to (\boldsymbol{\lambda}_j, \boldsymbol{\sigma}_j)$ and $(\boldsymbol{\lambda}_j, \boldsymbol{\sigma}_j) \to (\boldsymbol{\lambda}_i, \boldsymbol{\sigma}_i)$ with probability $\min(1, A)$ where
     $$A = \frac{\pi(\mathcal{D} | \boldsymbol{\lambda}_j, \boldsymbol{\sigma}_j)^{1/T_i} \pi(\mathcal{D} | \boldsymbol{\lambda}_i, \boldsymbol{\sigma}_i)^{1/T_j}}{\pi(\mathcal{D} | \boldsymbol{\lambda}_i, \boldsymbol{\sigma}_i)^{1/T_i} \pi(\mathcal{D} | \boldsymbol{\lambda}_j, \boldsymbol{\sigma}_j)^{1/T_j}}$$

5. Return to Step 3.

Tuning the MC^3 algorithm

The Metropolis coupled Markov chain Monte Carlo scheme targets the joint density (4) by simultaneously evolving C chains; each of which targets an alternative (tempered) density $\tilde{\pi}_c(\theta_c|\mathcal{D})$. Given the data $\mathcal{D}$, these chains are conditionally independent and should therefore be individually tuned to target their respective density in a typical fashion. Of course, it may not be possible to obtain near optimal acceptance rates within the posterior chain (and other chains with temperatures $\simeq 1$) however the analyst should aim to ensure reasonable acceptance rates; even if this results in small moves around the parameter space. Tuning the between chain proposal (Step 4 of the MC^3 algorithm) can be tricky in general. The strategy we suggest, also advocated by Wilkinson (2013), is that the temperatures are chosen such that they exhibit geometric spacing, that is, $T_{c+1}/T_c = r$ for some $r > 1$; this eliminates the burden of specifying $C - 1$ temperatures and instead only requires a choice of r. We also suggest only considering swaps between adjacent chains as intuitively the target densities are most similar when $|T_c - T_{c+1}|$ is small. It is generally accepted that between chain acceptance rates of around 20% to 60% provide reasonable mixing (with respect to the joint density of $\theta_1, \ldots, \theta_C$); see, for example, Geyer and Thompson (1995); Altekar et al. (2004). A suitable choice of the temperature ratio r can be guided via pilot runs of the MC^3 scheme and individual temperatures can also be adjusted as appropriate.

Monitoring convergence

The target (4) of our algorithm is a complicated mixed discrete-continuous distribution that means traditional convergence checks applied to each stochastic quantity are of little use. Indeed the unique samples of the discrete parameter ($\boldsymbol{\sigma}$) could be arbitrarily relabelled which may give a false impression of good mixing and convergence. Further, the natural dependence between the parameters ($\boldsymbol{\lambda}, \boldsymbol{\sigma}$) means that inspecting trace-plots of the continuous ($\boldsymbol{\lambda}$) components is not sensible as these parameters may make fairly large jumps around the parameter space when the choice order parameter ($\boldsymbol{\sigma}$) changes. These issues are akin to those that arise when fitting Bayesian mixture models and so following the approach used in that setting we choose to perform convergence checks with respect to the joint distribution. More specifically we apply the convergence diagnostic of Geweke (1992) to the (log) target distribution (4) as advocated by Cowles and Carlin (1996) and Brooks and Roberts (1998). Additionally we also perform this procedure on samples of the (log) posterior distribution of interest, that is, the marginal distribution $\pi(\boldsymbol{\lambda}, \boldsymbol{\sigma}|\mathcal{D})$ that corresponds to the chain with $T_c = 1$, and also the (log) observed data likelihood $\pi(\mathcal{D}|\boldsymbol{\lambda}, \boldsymbol{\sigma})$ evaluated at the posterior draws. Geweke-Brooks plots are also used to determine whether there is any evidence to suggest that a longer burn-in period may be required; these diagnostics are readily available within the R package *coda* (Plummer et al., 2006). To alleviate potential concerns about the sampling of the discrete choice order parameter ($\boldsymbol{\sigma}$) we also check that the marginal posterior distribution $\pi(\boldsymbol{\sigma}|\mathcal{D})$ is consistent under multiple (randomly initialised) runs of our algorithm.

4 Simulation study

To investigate the performance of the posterior sampling algorithm outlined in Section 3.4 we apply it on several synthetic datasets in which the values of the parameters used to simulate these data are known. We consider $K \in \{5, 10, 15, 20\}$ entities and generate 1000 rank orderings for each choice of K. Further we subset each of these datasets by taking the first $n \in \{20, 50, 200, 500, 1000\}$ orderings thus giving rise to 20 (nested) datasets. The parameter values $(\boldsymbol{\lambda}', \boldsymbol{\sigma}')$ from which these data are generated are drawn from the prior distribution outlined in Section 3.1 with $a_k = q_k = 1$ for $k = 1, \ldots, K$. That is, all choice orders and entity preferences (specified by the $(\boldsymbol{\lambda}', \boldsymbol{\sigma}')$ pair) are equally likely. The values of the parameters for each choice of K are given in Section 2 of the supplementary material. For each dataset, posterior samples were obtained via the algorithm outlined in Section 3.4. We choose to use $C = 5$ chains in each case, with both the temperatures and tuning parameters chosen appropriately. Convergence is assessed as outlined in Section 3.4 and the raw posterior draws are also thinned to obtain (approximately) 10K un-autocorrelated draws from the posterior distribution.

The aim of the simulation study is to investigate whether our algorithm is capable of generating posterior samples, for different (n, K) pairs. To this extent we would like to see that the values $(\boldsymbol{\lambda}', \boldsymbol{\sigma}')$ used to simulate these data look plausible under the respective posterior distributions. This can be judged by inspecting the marginal posterior distributions $\pi(\boldsymbol{\sigma}|\mathcal{D})$ and $\pi(\lambda_k|\mathcal{D})$. However, with space in mind, here we provide some key summaries and refer the interested reader to Section 2 of the supplementary material for a more detailed discussion. Table 1 shows the posterior probability $\Pr(\boldsymbol{\sigma}'|\mathcal{D})$ of the choice order parameter used to generate each respective dataset. Perhaps unsurprisingly we see that for each $K \in \{5, 10, 15, 20\}$ the posterior support for the choice order parameter used to generate the data increases with the number of

				n		
K		20	50	200	500	1000
5	$\Pr(\boldsymbol{\sigma}'|\mathcal{D})$	0.294^*	0.716^*	1.000^*	1.000^*	1.000^*
	$p_k \in$	$(0.36, 0.78)$	$(0.38, 0.73)$	$(0.36, 0.61)$	$(0.40, 0.60)$	$(0.54, 0.60)$
10	$\Pr(\boldsymbol{\sigma}'|\mathcal{D})$	0.156^*	0.604^*	1.000^*	1.000^*	1.000^*
	$p_k \in$	$(0.19, 0.78)$	$(0.11, 0.67)$	$(0.28, 0.64)$	$(0.30, 0.59)$	$(0.42, 0.59)$
15	$\Pr(\boldsymbol{\sigma}'|\mathcal{D})$	0.000	0.006	0.072	0.548^*	0.999^*
	$p_k \in$	—	$(0.40, 0.81)$	$(0.41, 0.73)$	$(0.44, 0.65)$	$(0.45, 0.64)$
20	$\Pr(\boldsymbol{\sigma}'|\mathcal{D})$	0.000	0.000	0.035	0.313^*	0.994^*
	$p_k \in$	—	—	$(0.30, 0.74)$	$(0.35, 0.83)$	$(0.38, 0.66)$

Table 1: Posterior probability $\Pr(\boldsymbol{\sigma}'|\mathcal{D})$ of the choice order used to generate each dataset along with range of tail area probability $p_k = \Pr(\lambda_k > \lambda'_k | \boldsymbol{\sigma} = \boldsymbol{\sigma}', \mathcal{D})$. *indicates that $\boldsymbol{\sigma}'$ is also the (posterior) modal observed choice order.
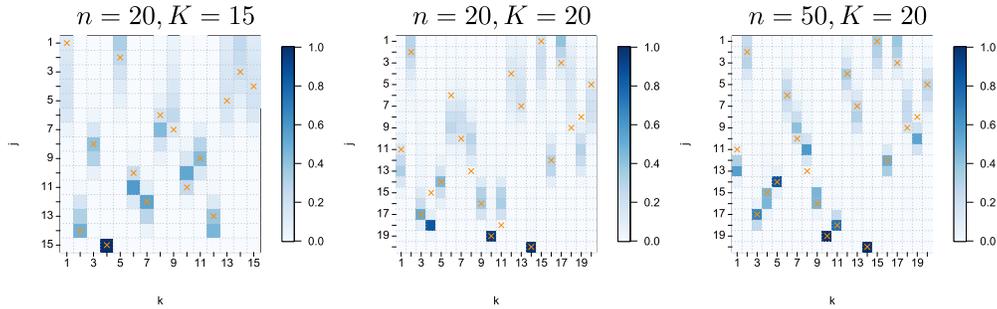
Figure 1: Synthetic data: heat maps of $\Pr(\sigma_j = k | \mathcal{D})$ for those analyses where $\boldsymbol{\sigma}'$ was not observed; the crosses highlight $\Pr(\sigma_j = \sigma_j' | \mathcal{D})$ in each case.

observations (rank orderings) considered. Interestingly we observe reasonable posterior support for $\boldsymbol{\sigma}'$ when only considering $n = 20$ preference orders of $K = 10$ entities. However for some of the analyses, those where n is relatively small in comparison to K, the choice order $\boldsymbol{\sigma}'$ is not observed in any of the 10K posterior draws. Further inspection reveals that the posterior draws of $\boldsymbol{\sigma}$ are reasonably consistent with the $\boldsymbol{\sigma}'$ used to generate the respective datasets; this can be seen by considering the marginal posterior distribution for each stage in the ranking process, that is, $\Pr(\sigma_j = k | \mathcal{D})$ for $j, k \in \{1, \ldots, K\}$. Figure 1 shows heat maps of $\Pr(\sigma_j = k | \mathcal{D})$ for those analyses where $\boldsymbol{\sigma}'$ was not observed; the crosses highlight $\Pr(\sigma_j = \sigma_j' | \mathcal{D})$ in each case. These figures reveal that, even with limited information, we are able to learn the final few entries in $\boldsymbol{\sigma}$ fairly well and much of the uncertainty resides within the earlier stages of the ranking process. Section 2.2 of the supplementary material presents the $\Pr(\sigma_j = k | \mathcal{D})$ from Figure 1 in tabular form along with the image plots for the remaining analyses.

For the Extended Plackett-Luce model we are trying to quantify our uncertainty not only about the choice order parameter but also about the entity parameters. As discussed in Section 2 the entity parameter values $\boldsymbol{\lambda}$ only have a meaningful interpretation for a given choice order parameter $\boldsymbol{\sigma}$. Section 2 of the supplementary material contains boxplots of the marginal posterior distributions $\pi(\lambda_k | \boldsymbol{\sigma} = \boldsymbol{\sigma}', \mathcal{D})$, that is, the marginal posterior distribution of λ_k where we only consider posterior draws with $\boldsymbol{\sigma} = \boldsymbol{\sigma}'$. To summarise the results we compute the tail area probabilities $p_k = \Pr(\lambda_k > \lambda_k' | \boldsymbol{\sigma} = \boldsymbol{\sigma}', \mathcal{D})$ of λ_k under its respective marginal distribution for $k = 1, \ldots, K$. The range of these values, that is, $(\min p_k, \max p_k)$ for each analysis is given in Table 1 and these indicate that there is reasonable posterior support for $\boldsymbol{\lambda}'$, even when n is small relative to K. In other words, the parameters $\boldsymbol{\lambda}'$ look plausible under the posterior distribution. Of course, we can not compute these quantities for those analyses where $\boldsymbol{\sigma}'$ is not observed. However, although prohibitive for $\boldsymbol{\lambda}$ inference, this does not prohibit inferences on observable quantities (rank orders) as this is achieved via the posterior predictive distribution; this is the topic of the next section.

5 Inference and model assessment via the posterior predictive distribution

In this section we consider methods for performing inference for the entities by appealing to the posterior predictive distribution which will also provide us with a mechanism for detecting lack of model fit. Obtaining the posterior predictive distribution, and, more generally, predictive quantities of interest, is computationally burdensome when the number of entities is not small and so we also outline Monte Carlo based approximations that can be used to facilitate this approach for larger values of K.

5.1 Inference for entity preferences

The Extended Plackett-Luce model we consider is only defined for complete rankings and so the posterior predictive distribution is a discrete distribution defined over all possible observations $\tilde{\boldsymbol{x}} \in \mathcal{S}_K$. These probabilities can be approximated by taking the sample mean of the EPL probability (2) over draws from the posterior distribution for $\boldsymbol{\lambda}$ and $\boldsymbol{\sigma}$, that is $\Pr(\boldsymbol{X} = \tilde{\boldsymbol{x}}|\mathcal{D}) \simeq N^{-1} \sum_{i=1}^{N} \Pr(\boldsymbol{X} = \tilde{\boldsymbol{x}}|\boldsymbol{\lambda}^{(i)}, \boldsymbol{\sigma}^{(i)})$ where $\boldsymbol{\lambda}^{(i)}, \boldsymbol{\sigma}^{(i)}$ for $i = 1, \ldots, N$ are sampled from the distribution with density $\pi(\boldsymbol{\lambda}, \boldsymbol{\sigma}|\mathcal{D})$. We can then use this posterior predictive distribution, for example, to obtain the marginal posterior predictive probability that entity k is ranked in position j, that is, $\Pr(\tilde{x}_j = k|\mathcal{D})$ for $j, k \in \{1, \ldots, K\}$. The posterior modal ordering $\hat{\boldsymbol{x}}$ is also straightforward to obtain and is simply that which has largest posterior predictive probability. However, when the number of entities is larger than say 9, this procedure involves enumerating the predictive probabilities for more than $\mathcal{O}(10^6)$ possible observations. Clearly this becomes computationally infeasible as the number of entities increases; particularly as computing the posterior predictive probability also involves taking the expectation over many thousands of posterior draws. When the number of entities renders full enumeration infeasible we suggest approximating the posterior predictive distribution via a Monte Carlo based approach as in Johnson et al. (2020). In particular we obtain a collection $\mathcal{P} = \{\tilde{\boldsymbol{x}}_\ell^{(m)} : m = 1, \ldots, M, \ \ell = 1, \ldots, L\}$ of draws from the posterior predictive distribution by sampling L rank orderings at each iteration of the M iterations of the posterior sampling scheme. We can then approximate $\Pr(\tilde{x}_j = k|\mathcal{D})$ by the empirical probability computed from the collection of rankings $\mathcal{P}$, that is $\Pr(\tilde{x}_j = k|\mathcal{D}) = \frac{1}{ML} \sum_{m=1}^{M} \sum_{\ell=1}^{L} \mathbb{I}(\tilde{x}_{\ell j}^{(m)} = k)$, where $\mathbb{I}(x)$ denotes an indicator function which returns 1 if x is true and 0 otherwise. If desired the mode of the posterior predictive distribution can be obtained via an efficient optimisation algorithm based on cyclic coordinate ascent (Johnson et al., 2020).

5.2 Model assessment via posterior predictive checks

In the Bayesian framework assessment of model fit to the data can be provided by comparing observed quantities with potential future observations through the posterior predictive distribution; the basic idea being that the observed data $\mathcal{D}$ should appear to be a plausible realisation from the posterior predictive distribution. This approach

to Bayesian goodness of fit dates back at least to Guttman (1967) and is described in detail in Gelman et al. (2014), for example. Several methods for assessing goodness of fit for models of rank ordered data were proposed in Cohen and Mallows (1983) and more recently similar methods have been developed in a Bayesian framework by, amongst others, Yao and Böckenholt (1999), Mollica and Tardella (2017) and Johnson et al. (2020). In the illustrative examples on real data in Section 6 we propose a range of diagnostics tailored to the specific examples. For example, one generic method for diagnosing lack of model fit is to monitor the (absolute value of the) discrepancy between the marginal posterior predictive probabilities of entities taking particular ranks with the corresponding empirical probabilities computed from the observed data. That is, we consider $d_{jk} = |\Pr(\tilde{x}_j = k|\mathcal{D}) - \Pr(x_j = k)|$ where $\Pr(x_j = k) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(x_{ij} = k)$ denotes the empirical probabilities computed from those $x \in \mathcal{D}$ and the posterior predictive probabilities $\Pr(\tilde{x}_j = k|\mathcal{D})$ are computed as described in Section 5.1. These discrepancies d_{jk} for $j, k \in \{1, \ldots, K\}$ can then be depicted as a heat map where large values could indicate potential lack of model fit. By focusing on the marginal probabilities $\Pr(x_j = k)$ we obtain a broad-scale "first-order" check on the model, but, as described in Cohen and Mallows (1983), we could also look at finer-scale features such as pairwise comparisons, triples and so on. Of course, if the full posterior predictive distribution over all $K!$ possible observations is available (that is, if K is small) then we could also check that the observed data look plausible under this distribution.

6   Illustrative examples

We now summarise analyses of two real datasets which together highlight how valuable insights can be obtained by considering the Extended Plackett-Luce model as opposed to simpler alternatives. With no direct competitor for Bayesian analyses of the Extended Plackett-Luce model we compare our conclusions to those obtained under standard and reverse Plackett-Luce analyses. For the standard and reverse Plackett-Luce analyses posterior samples are obtained using the (partially collapsed) Gibbs sampling scheme of Caron and Doucet (2012) as this algorithm has been shown to be very efficient for these models. Of course, our algorithm is also capable of targeting these much simpler posterior distributions. This could be achieved by considering a single chain ($C = 1$) and repeatedly performing the MH update of λ followed by the rescaling step with $\sigma = \mathcal{I}$ or $\sigma = (K, \ldots, 1)'$ fixed throughout.

6.1   Song data

For our first example we consider a dataset with a long standing in the literature that was first presented in Critchlow et al. (1991). The original dataset was formed by asking ninety-eight students to rank $K = 5$ words, (1) *score*, (2) *instrument*, (3) *solo*, (4) *benediction* and (5) *suit*, according to the association with the target word "*song*". However, the available data given in Critchlow et al. (1991) is in grouped format and the ranking of 15 students are unknown and hence discarded. The resulting dataset therefore comprises $n = 83$ rank orderings and is reproduced in the supplementary material.

Posterior samples are obtained via the algorithm outlined in Section 3.4 where the prior specification is as in Section 3.1 with $q_k = 1$ and $a_k = 1$ (for $k = 1, \ldots, K$) and so all choice and preference orderings are equally likely *a priori*. The following results are based on a typical run of our (appropriately tuned) MC^3 scheme initialised from the prior. We performed 1M iterations, with an additional 10K discarded as burn-in, which were thinned by 100 to obtain 10K (almost) un-autocorrelated realisations from the posterior distribution. The algorithm runs fairly quickly, with C code on five threads of an Intel Core i7-4790S CPU (3.20GHz clock speed) taking around 9 minutes. The results of the convergence diagnostics described in Section 3.4, along with the tuning parameter values, are given in Section 5 of the supplementary material.

Investigation of the posterior distribution reveals there is no support for the standard (or reverse) Plackett-Luce model(s) with $\Pr(\boldsymbol{\sigma} = (3, 2, 1, 4, 5)'|\mathcal{D}) = 0.9983$, $\Pr(\boldsymbol{\sigma} = (5, 4, 1, 2, 3)'|\mathcal{D}) = 0.0015$ and the remaining posterior mass (0.0002) assigned to $\boldsymbol{\sigma} = (2, 3, 1, 4, 5)'$. It is interesting to see that, although it receives relatively little posterior support, the 2nd most likely choice order parameter value is that given by reversing the elements of the posterior modal value. It is also worth noting that the posterior modal choice order ($\boldsymbol{\sigma} = (3, 2, 1, 4, 5)'$) is not contained within the restricted set considered by Mollica and Tardella (2018).

Inference for the Extended Plackett-Luce model is substantially more challenging than for the simpler standard/reverse Plackett-Luce models and so it is important to question whether this additional complexity allows us to better describe the data. Put another way, does the EPL model give rise to improved model fit? To this extent we investigate the discrepancies $d_{jk} = |\Pr(\tilde{x}_j = k|\mathcal{D}) - \Pr(x_j = k)|$. For comparative purposes we also compute the discrepancies obtained from under both standard and reverse Plackett-Luce analyses of these data; Figure 2 shows these values as a heat map for $j, k \in \{1, \ldots, K\}$. Visual inspection clearly suggests that the observed data look more plausible under the EPL model when compared to the simpler models. In particular, there are rather large discrepancies between the predictive and empirical probabilities that entity $k = 1$ (Score) is ranked in position $j = 3$ under the SPL (0.34) and RPL (0.40) analyses. The superior fit of the EPL model is further supported by
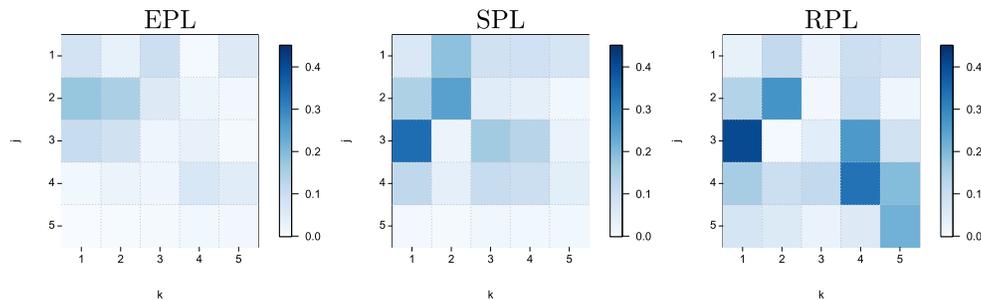

Figure 2: Song data: heat maps showing $d_{jk} = |\Pr(\tilde{x}_j = k|\mathcal{D}) - \Pr(x_j = k)|$ for the extended, standard and reverse Plackett-Luce analyses.
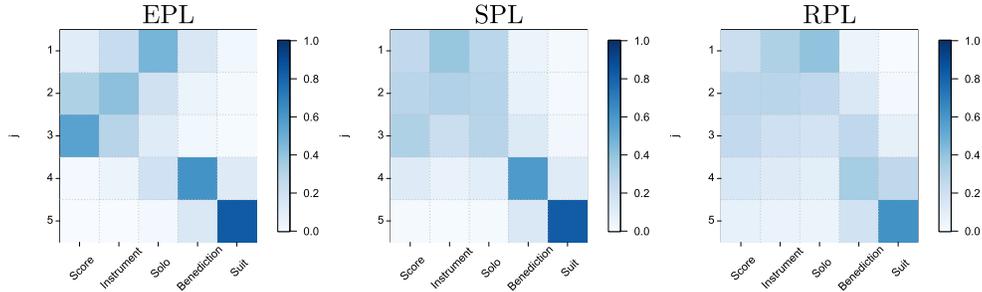
Figure 3: Song data: heat maps showing $\Pr(\tilde{x}_j = k|\mathcal{D})$ for the extended, standard and reverse Plackett-Luce analyses.

the Watanabe-Akaike information criterion (Watanabe, 2010) with values of 464.12, 546.84 and 654.21 for the extended, standard and reverse PL models respectively. Note that these values are on the deviance scale and the effective number of parameters is computed as in Gelman et al. (2014). Given we have a mixed discrete-continuous parameter space we suggest that the WAIC is preferable to other popular information criteria such as AIC, DIC or BIC due to their reliance on point estimates and/or the assumption of a (multivariate) Gaussian posterior which may not be appropriate.

Turning now to inference for observable quantities we again appeal to the posterior predictive distribution. More specifically we can now use the (predictive) probabilities $\Pr(\tilde{x}_j = k|\mathcal{D})$ to deduce the likely positions of entities within rankings. Figure 3 shows these probabilities as a heat map for $j, k \in \{1, \ldots, K\}$. Focusing on the Extended Plackett-Luce analysis, it is fairly clear that "Suit" (5) is the least preferred entity and "Benediction" (4) is the 4th most preferred, with relatively little (predictive) support for any other entities in these positions. There is perhaps more uncertainty on those entities that are ranked within positions $j = 1, 2, 3$, although the figure would suggest that the preference of the entities is (Solo, Instrument, Score, Benediction, Suit). Indeed this is the modal predictive ranking and has predictive probability 0.232. Interestingly the same modal (predictive) ranking is obtained under the RPL analysis, although it only receives predictive probability 0.07, whereas for the SPL analysis the modal (predictive) ranking is (Instrument, Solo, Score, Benediction, Suit) and occurs with probability 0.122. Figure 3 also shows there to be much more uncertainty, particularly for the top 3 entities, under the SPL and RPL analyses; this is perhaps more naturally seen through Figure 4 that shows the (full) posterior predictive distribution for each possible future observation $\tilde{\boldsymbol{x}} \in \mathcal{S}_K$. The crosses ($\times$) show the empirical probabilities of the observed data (those $\tilde{\boldsymbol{x}} \in \mathcal{D}$). Figure 4 illustrates that the observed modal ranking is (Solo, Instrument, Score, Benediction, Suit).

6.2 Formula 1 data

We now analyse a dataset containing the finishing orders of drivers within the 2018 Formula 1 (F1) season and so we have $n = 21$ rank orderings of the $K = 20$ drivers.
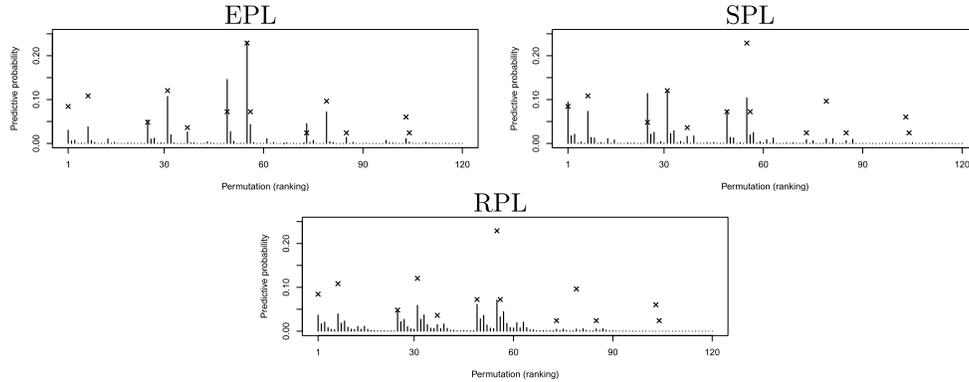
Figure 4: Song data: Full posterior predictive distribution for each of the $5! = 120$ possible observations $\tilde{\boldsymbol{x}} \in \mathcal{S}_K$ under the EPL (left), SPL (right) and RPL (bottom) analyses. Crosses ($\times$) show the empirical probabilities of the observed data (those $\tilde{\boldsymbol{x}} \in \mathcal{D}$).

It will be interesting to see whether we are able to gain more valuable insights using the EPL model when compared to the simpler variants. In particular whether we are able to gain any information about the choice order parameter $\boldsymbol{\sigma}$ in this setting as K is fairly large, relative to n. The rank orderings considered here were collected from www.espn.co.uk and are reproduced in Section 7 of the supplementary material.

Numerous variants of the Plackett-Luce model have previously been developed for the analysis of F1 finishing orders; see Henderson and Kirrane (2018) and the discussion therein. In general, models derived from the reverse Plackett-Luce (RPL) model appear to perform better than the standard Plackett-Luce model in the sense that they give rise to better model fit. We choose to incorporate this prior information by letting $\boldsymbol{q} = (1, \ldots, K)'$ and so *a priori* the modal choice ordering is $\hat{\boldsymbol{\sigma}} = (K, \ldots, 1)'$, that is, the choice ordering corresponding to the reverse Plackett-Luce model. Figure 5 (right) shows the (log) prior probabilities $\Pr(\sigma_j = k)$ for $j, k \in \{1, \ldots, K\}$. Regarding the drivers, although their individual ability no doubt plays a part, ultimately Formula 1 is a team sport and the performance of a particular driver within a race is often limited by the quality of the car their respective team is able to produce. To this extent we suppose that teams with larger budgets, and hence more capacity to invest in research and development, are more likely to produce a superior car and thus improved race performance. We therefore let the a_k value be equal to the team budget for each of the K drivers and then rescale the a_k values such that the smallest value is 1. The implication is that, under the standard Plackett-Luce model, an individual who drives for any particular team is twice as likely to win a race when compared to another driver who is part of a team with half of the budget. The budgets were obtained from www.racefans.net and the a_k values for each driver are given in Section 7 of the supplementary material. Because this prior specification does not lead to a unique prior predictive mode under the SPL model, we use a slight modification of our mode preserving procedure which preserves all modes when averaged over the prior for $\boldsymbol{\sigma}$; details of this modification are
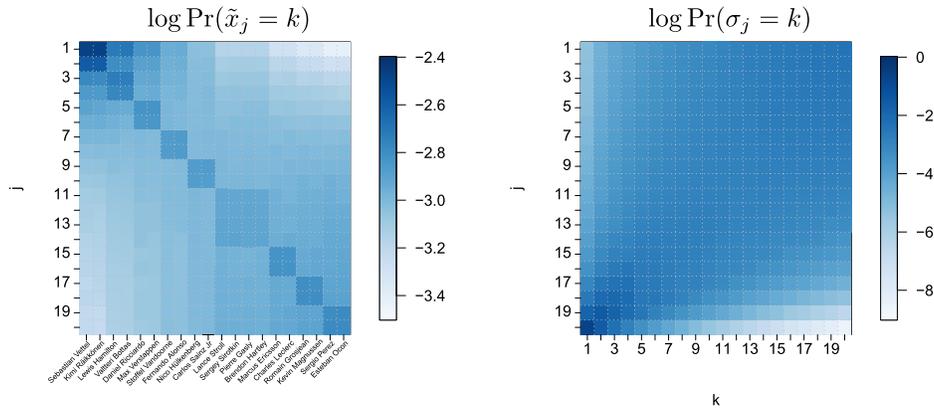
Figure 5: F1 data: heat maps showing the prior predictive distribution (left) and the prior distribution of the choice order parameter (right).

provided in the supplementary material. Of course, for the EPL model there is also additional uncertainty on the choice order parameter and the resulting prior predictive distribution is shown in Figure 5 (left). Here the prior (predictive) probabilities $\Pr(\tilde{x}_j = k)$ are obtained via a Monte Carlo based approach akin to that outlined in Section 5 where the $(\boldsymbol{\lambda}, \boldsymbol{\sigma})$ draws are obtained from the prior as opposed to the posterior. Section 6 of the supplementary material provides posterior summaries from alternative analyses with $\boldsymbol{a} = \mathbf{1}$ (a uniform prior predictive distribution) and $\boldsymbol{q} = \mathbf{1}$ (a uniform prior distribution for $\boldsymbol{\sigma}$). These analyses reveal that the posterior distribution is not particularly sensitive to the choices of prior distribution considered here.

The following results are based on a typical run of our (appropriately tuned) MC^3 scheme initialised from the prior. We allow a burn-in period of 100K iterations after which we perform a further 2M iterations, with these thinned by 200 to obtain 10K (almost) un-autocorrelated realisations from the posterior distribution. This analysis takes around an hour using C code on five threads of an Intel Core i7-4790S CPU (3.20GHz clock speed). Section 6 of the supplementary material contains the results of the convergence diagnostics and the tuning parameter values used for this analysis.

First, to address the model selection issue, the estimated WAIC values for the extended, standard and reverse PL models are (on the deviance scale) 1501.58, 1672.49 and 1507.32 respectively. These values suggest the marginal superiority of the EPL model over the RPL model, with the SPL model a distant third.

Focusing now on the EPL analysis, investigation of the posterior distribution reveals that there is a large amount of uncertainty on the choice order parameter $\boldsymbol{\sigma}$ and also potential bi-modality within certain ranking stages. That said, further inspection of the marginal posterior distributions given by $\Pr(\sigma_j = k | \mathcal{D})$ reveals that there is a surprisingly small amount of uncertainty on the ranks allocated in the 13th–20th stages; see Figure 6 (right). Further within these positions $(\sigma_{13}, \ldots, \sigma_{20})$ the ranks allocated

are consistent with the choice order parameter corresponding to the reverse Plackett-Luce model which suggests why previous authors may have found the RPL model to be preferable to the SPL model for modelling F1 results. We also note that, although there is some uncertainty within certain ranking stages, it is clear that these data are rather informative about the likely values of $\boldsymbol{\sigma}$ in the sense that the posterior distribution is not that similar to the prior; see Figure 5 (right).

Again we can turn to the posterior predictive distribution to deduce the likely positions of entities within rankings. Note that here complete enumeration of the posterior predictive probabilities for each $\tilde{\boldsymbol{x}} \in \mathcal{S}_K$ is computationally infeasible as $K!$ is of $\mathcal{O}(10^{18})$. Figure 6 (left) therefore shows the (log) probabilities $\Pr(\tilde{x}_j = k|\mathcal{D})$ as a heat map for $j, k \in \{1, \ldots, K\}$; the corresponding figures for RPL and SPL analyses are provided within the supplementary material. Note that the predictive probabilities $\Pr(\tilde{x}_j = k|\mathcal{D})$ are computed based on synthetic data simulated from the predictive distribution as discussed in Section 5 with $L = 10$. This figure shows that, perhaps unsurprisingly, there is a fairly large amount of uncertainty about the finishing positions of some of the drivers. That said, this does not mean that useful inferences cannot be made based on these predictive probabilities. For example, the figure would suggest that Stoffel Vandoorne (column 7) is most likely to place somewhere between 11th–15th (probability 0.46) whereas the probability of this driver finishing within the top-10 positions is only 0.21. In contrast, Fernando Alonso (column 8) is much more likely to obtain a top-10 finish (probability 0.43) and we would only expect Alonso to finish within positions 11 to 15 in around 25% of the races. This is particularly interesting given that both of these individuals drove for the same team (McLaren Renault) within the 2018 season and hence were exchangeable *a priori* under our model. It is also clear that the first 6 drivers are those most likely to perform well during a race. However, a particularly interesting aspect is the bi-modality of the distribution for some of these drivers. Take,
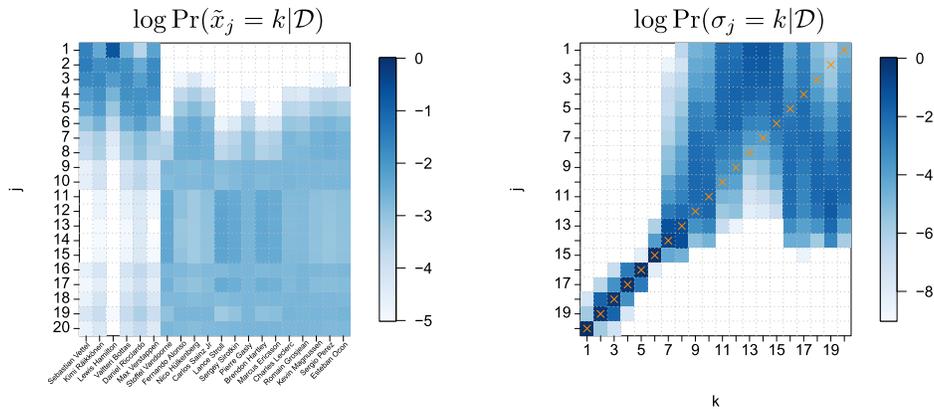

Figure 6: F1 data: heat maps showing the posterior predictive distribution (left) and the marginal prior distribution of the choice order parameter (right). Crosses ($\times$) highlight the probabilities corresponding to $\boldsymbol{\sigma} = (K, \ldots, 1)'$, the reverse Plackett-Luce model.

for example, Sebastian Vettel (column 1); we can see that although most of the probability is assigned to this driver finishing within the top-10 positions, there is also an increased chance of finishing within the bottom 5 when compared to the middle (11–15) positions. Although perhaps counter-intuitive this result is perhaps not as surprising as it first seems. Indeed, although the drivers in the top cars are likely to do well, there is also a chance they could suffer from a mechanical issue, or perhaps be involved in an accident, and hence obtain a surprisingly low finishing position within any given race. We also note that this particular aspect is not captured by either the standard or reverse Plackett-Luce analyses. The corresponding heat maps can be found in Section 6 of the supplementary material and these show that the distribution of the finishing position of each driver, that is, the distribution within each column, is uni-modal and this perhaps explains why the Extended Plackett-Luce model is favoured here.

We can also look at other quantities of interest, for example, the number of times we would expect each of the top 6 drivers to win a race, feature on the podium (top 3), and also obtain a points (top 10) finish based on the predictive probabilities. More specifically Table 2 shows $n \times \sum_{k=1}^{p} \Pr(\tilde{x}_j = k|\mathcal{D})$ for $p = 1, 3, 10$ along with the observed number of times computed from those $x \in \mathcal{D}$. It is interesting to see that both the extended and reverse Plackett-Luce models are able to capture these aspects of the distribution fairly well. Further the expected number of points (top 10) finishes under the standard Plackett-Luce model is also fairly consistent with the observed data. However, the shortcomings of the more simple standard Plackett-Luce model become clear if we instead consider the expected number of wins and podiums. For example, we observed that Hamilton won 11 races and the SPL model suggests that he would be expected to win around 5 races within an F1 season whereas the EPL model suggests 10 wins which is much more consistent with the observed data. Again, additional insight into the question of model fit can be obtained via heat maps showing the discrepancies $d_{jk} = |\Pr(\tilde{x}_j = k|\mathcal{D}) - \Pr(x_j = k)|$ for $j, k \in \{1, \ldots, K\}$; these are provided in Section 6 of the supplementary material. The total discrepancies for each model, that is $\sum_{jk} d_{jk}$, are 13.35, 17.33 and 14.14 for the extended, standard and reverse PL models respectively. These values support the conclusion from the comparison of the WAIC values; the EPL model provides the best fit out of the three models.

Finally, whilst the number of wins, podium and points finishes is clearly of interest, ultimately the F1 Drivers' Championship is determined by the number of points each

	Observed			EPL			SPL			RPL		
Driver Name	W	Pod	Pts	W	Pod	Pts	W	Pod	Pts	W	Pod	Pts
Lewis Hamilton	11	17	20	10.31	16.47	20.21	4.74	11.89	20.52	10.45	16.55	20.33
Sebastian Vettel	5	12	20	4.27	12.68	19.46	3.06	8.68	19.46	4.31	12.72	19.66
Max Verstappen	2	11	17	2.20	9.61	18.74	1.41	4.43	15.20	2.15	9.59	19.00
Daniel Ricciardo	2	2	13	0.61	4.83	16.90	0.83	2.67	11.21	0.55	4.76	17.37
Kimi Räikkönen	1	12	17	1.92	9.07	18.58	1.47	4.58	15.50	1.79	8.84	18.82
Valtteri Bottas	0	8	19	1.64	8.41	18.37	2.34	6.95	18.42	1.71	8.68	18.77

Table 2: F1 data: Observed number of wins (W), podiums (Pod) and points (Pts) finishes and also the expected numbers under the predictive distributions for the EPL, SPL and RPL analyses for the top six drivers in the 2018 season.

Driver Name	Points	Expected points			Pr (observed position)		
		EPL	SPL	RPL	EPL	SPL	RPL
Lewis Hamilton	408	395.68	310.20	399.05	0.94	0.85	0.95
Sebastian Vettel	320	308.63	251.29	311.24	0.69	0.62	0.73
Kimi Räikkönen	251	250.45	158.02	246.54	0.28	0.11	0.24
Max Verstappen	249	259.74	152.79	259.31	0.29	0.32	0.28
Valtteri Bottas	247	240.91	216.02	245.02	0.37	0.04	0.33
Daniel Ricciardo	170	185.34	101.34	183.98	0.81	0.17	0.83
Nico Hülkenberg	69	65.12	59.29	62.70	0.40	0.03	0.37
Sergio Perez	62	52.22	93.70	52.47	0.16	0.12	0.18
Kevin Magnussen	56	46.63	70.26	45.13	0.13	0.07	0.13
Carlos Sainz Jr	53	48.40	100.28	58.17	0.14	0.08	0.11
Fernando Alonso	50	51.22	59.34	43.03	0.11	0.07	0.13
Esteban Ocon	49	45.22	67.19	40.47	0.12	0.08	0.13
Charles Leclerc	39	37.33	61.63	35.58	0.14	0.08	0.13
Romain Grosjean	37	34.99	65.78	30.87	0.14	0.08	0.14
Pierre Gasly	29	27.34	62.07	31.17	0.16	0.08	0.13
Stoffel Vandoorne	12	15.01	78.41	23.63	0.14	0.05	0.16
Marcus Ericsson	9	16.94	63.00	19.44	0.16	0.08	0.17
Lance Stroll	6	11.48	57.91	13.40	0.18	0.09	0.21
Brendon Hartley	4	14.19	45.42	12.61	0.19	0.15	0.24
Sergey Sirotkin	1	14.15	47.02	7.16	0.19	0.19	0.50

Table 3: F1 dataset: 2018 Drivers' Championship final standings, expected points and posterior predictive probabilities of finishing in the observed final position under the EPL, SPL and RPL analyses.

driver is able to accrue over an F1 season. Given the amount of points awarded is simply a deterministic function of the finishing orders we can straightforwardly compute the expected number of points each driver obtains by simulating synthetic F1 seasons $\tilde{\mathcal{D}} = \{\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_n\}$ from the posterior predictive distribution. Table 3 shows the observed number of points in addition to the expected number under each model considered. We can also use $\tilde{\mathcal{D}}$ to compute the (predictive) probability that each driver finishes within their respective (observed) position; these values are given in Table 3 (right).

7   Conclusion

We have considered the problem of implementing a Bayesian analysis of rank ordered data using the Extended Plackett-Luce model. In particular we have considered carefully the problem of prior specification, proposing a Plackett-Luce model as the prior for the choice order parameter $\boldsymbol{\sigma}$ and proposing a prior distribution on the entity parameters that preserves the modal ordering under the prior predictive distribution. We have also tackled the challenging issue of posterior sampling of a potentially highly multi-modal posterior distribution with both discrete and continuous components via a Metropolis coupled Markov chain Monte Carlo scheme. This has enabled efficient

posterior sampling which potentially facilitates further analyses based on the Extended Plackett-Luce model and further extensions of the model. Finally, we have focused on predictive inference for observable quantities that naturally facilitates the assessment of model adequacy.

Reproducibility

With reproducibility in mind, the code to run the algorithm outlined in Section 3.4 can be found at the GitHub repository https://github.com/srjresearch/ExtendedPL. In addition, C code for performing a standard Plackett-Luce analysis along with each of the synthetic and real datasets presented are also provided.

Supplementary Material

Supplementary material of "On Bayesian inference for the Extended Plackett-Luce model" (DOI: 10.1214/21-BA1258SUPP; .pdf). The supplementary material contains details of the data generating mechanism, further discussion of our mode preserving prior specification and additional simulation and real data results.

References

Altekar, G., Dwarkadas, S., Huelsenbeck, J. P., and Ronquist, F. (2004). "Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference." *Bioinformatics*, 20(3): 407–415. 476

Alvo, M. and Yu, P. (2014). *Statistical Methods for Ranking Data*. Springer, New York. MR3308929. doi: https://doi.org/10.1007/978-1-4939-1471-5. 465

Bezáková, I., Kalai, A., and Santhanam, R. (2006). "Graph model selection using maximum likelihood." In *Proceedings of the 23rd International Conference on Machine Learning*, 105–112. ACM. 473

Brooks, S. P. (1998). "Markov chain Monte Carlo method and its application." *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(1): 69–100. 474

Brooks, S. P. and Roberts, G. O. (1998). "Convergence assessment techniques for Markov chain Monte Carlo." *Statistics and Computing*, 8(4): 319–335. 476

Caron, F. and Doucet, A. (2012). "Efficient Bayesian inference for generalized Bradley–Terry models." *Journal of Computational and Graphical Statistics*, 21(1): 174–196. MR2913362. doi: https://doi.org/10.1080/10618600.2012.638220. 468, 473, 480

Cohen, A. and Mallows, C. L. (1983). "Assessing goodness of fit of ranking models to data." *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(4): 361–374. 480

Cowles, M. K. and Carlin, B. P. (1996). "Markov chain Monte Carlo convergence diagnostics: a comparative review." *Journal of the American Statistical Association*, 91(434): 883–904. MR1395755. doi: https://doi.org/10.2307/2291683. 476

Critchlow, D. E., Fligner, M. A., and Verducci, J. S. (1991). "Probability models on rankings." *Journal of Mathematical Psychology*, 35(3): 294–318. MR1128236. doi: https://doi.org/10.1016/0022-2496(91)90050-4. 480

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*. CRC Press, 3rd edition. MR3235677. 480, 482

Geweke, J. (1992). "Evaluating the accuracy of sampling-based approaches to calculating posterior moments." In Bernardo, J. M., Berger, J., Dawid, A. P., and Smith, J. F. M. (eds.), *Bayesian Statistics 4*, 169–193. Oxford University Press. MR1380276. 476

Geyer, C. J. (1991). "Markov chain Monte Carlo maximum likelihood." In *Proceedings of the 23rd Symposium on the Interface*, Computing Science and Statistics, 156–163. Interface Foundation of North America. 474

Geyer, C. J. and Thompson, E. A. (1995). "Annealing Markov chain Monte Carlo with applications to ancestral inference." *Journal of the American Statistical Association*, 90(431): 909–920. MR1341319. doi: https://doi.org/10.2307/1390763. 476

Gilks, W. R. and Roberts, G. O. (1996). "Strategies for improving MCMC." In Gilks, W., Richardson, S., and Spiegelhalter, D. (eds.), *Markov chain Monte Carlo in Practice*, Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis. MR1397966. doi: https://doi.org/10.1007/978-1-4899-4485-6. 474

Guttman, I. (1967). "The use of the concept of a future observation in goodness-of-fit problems." *Journal of the Royal Statistical Society, Series B*, 29(1): 83–100. MR0216699. 480

Henderson, D. A. and Kirrane, L. J. (2018). "A comparison of truncated and time-weighted Plackett–Luce models for probabilistic forecasting of Formula One results." *Bayesian Analysis*, 13(2): 335–358. MR3780426. doi: https://doi.org/10.1214/17-BA1048. 483

Johnson, S. R., Henderson, D. A., and Boys, R. J. (2020). "Revealing subgroup structure in ranked data using a Bayesian WAND." *Journal of the American Statistical Association*. To appear. MR4189765. doi: https://doi.org/10.1080/01621459.2019.1665528. 479, 480

Johnson, S. R., Henderson, D. A., and Boys, R. J. (2021). "Supplementary Material of "On Bayesian inference for the Extended Plackett-Luce model"." *Bayesian Analysis*. doi: https://doi.org/10.1214/21-BA1258SUPP. 467

Luce, R. (1959). *Individual Choice Behavior: A Theoretical Analysis*. Wiley. MR0108411. 465

Marden, J. I. (1995). *Analyzing and Modeling Rank Data*. London: Chapman and Hall. MR1346107. 465

Mollica, C. and Tardella, L. (2014). "Epitope profiling via mixture modelling of ranked data." *Statistics in Medicine*, 33(21): 3738–3758. MR3260657. doi: https://doi.org/10.1002/sim.6224.   465, 466, 467, 468

Mollica, C. and Tardella, L. (2017). "Bayesian Plackett–Luce mixture models for partially ranked data." *Psychometrika*, 82(2): 442–458. MR3665302. doi: https://doi.org/10.1007/s11336-016-9530-0.   480

Mollica, C. and Tardella, L. (2018). "Constrained Extended Plackett-Luce model for the analysis of preference rankings." In *Book of Short Papers – SIS2018: 49th Scientific meeting of the Italian Statistical Society*, 480–486.   466, 467, 470, 481

Mollica, C. and Tardella, L. (2020). "Bayesian analysis of ranking data with the Extended Plackett–Luce model." *Statistical Methods & Applications*, 1–20.   467, 473

Plackett, R. L. (1975). "The analysis of permutations." *Applied Statistics*, 24: 193–202. MR0391338. doi: https://doi.org/10.2307/2346567.   465

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). "CODA: Convergence diagnosis and output analysis for MCMC." *R News*, 6(1): 7–11. URL https://journal.r-project.org/archive/   476

Watanabe, S. (2010). "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory." *Journal of Machine Learning Research*, 11(Dec): 3571–3594. MR2756194.   482

Wilkinson, D. (2013). "Parallel tempering and Metropolis coupled MCMC." https://darrenjw.wordpress.com/tag/mc3/. [Online; accessed 13-Feb-2020].   476

Yao, G. and Böckenholt, U. (1999). "Bayesian estimation of Thurstonian ranking models based on the Gibbs sampler." *British Journal of Mathematical and Statistical Psychology*, 52(1): 79–92. MR2694087.   480