Bayesian inference for Plackett-Luce ranking models

John Guiver JOGUIVER@MICROSOFT.COM
Edward Snelson ESNELSON@MICROSOFT.COM
Microsoft Research Limited, 7 J J Thomson Avenue, Cambridge CB3 0FB, UK

Abstract

This paper gives an efficient Bayesian method for inferring the parameters of a Plackett-Luce ranking model. Such models are parameterised distributions over rankings of a finite set of objects, and have typically been studied and applied within the psychometric, sociometric and econometric literature. The inference scheme is an application of Power EP (expectation propagation). The scheme is robust and can be readily applied to large scale data sets. The inference algorithm extends to variations of the basic Plackett-Luce model, including partial rankings. We show a number of advantages of the EP approach over the traditional maximum likelihood method. We apply the method to aggregate rankings of NASCAR racing drivers over the 2002 season, and also to rankings of movie genres.

1. Introduction

Problems involving ranked lists of items are widespread, and are amenable to the application of machine learning methods. An example is the subfield of "learning to rank" at the cross-over of machine learning and information retrieval (see e.g. Joachims et al., 2007). Another example is rank aggregation and meta-search (Dwork et al., 2001). The proper modelling of observations in the form of ranked items requires us to consider parameterised probability distributions over rankings. This has been an area of study in statistics for some time (see Marden, 1995 for a review), but much of this work has not made its way into the machine learning community. In this paper we study one particular ranking distribution, the Plackett-Luce, which has some very nice properties. Although parameter estimation in the Plackett-Luce can be achieved via maximum likelihood estimation (MLE) using MM methods (Hunter, 2004), we are unaware of an efficient Bayesian treatment. As we will show, MLE is problematic for sparse data due to overfitting, and it cannot even be found for some data samples that do occur in real situations. Sparse data within the context of ranking is a common scenario for some applications and is typified by having a small number of observations and a large number of items to rank (Dwork et al., 2001), or each individual observation may rank only a few of the total items. We therefore develop an efficient Bayesian approximate inference procedure for the model that avoids overfitting and provides proper uncertainty estimates on the parameters.

The Plackett-Luce distribution derives its name from independent work by Plackett (1975) and Luce (1959). The Luce Choice Axiom is a general axiom governing the choice probabilities of a population of 'choosers', choosing an item from a subset of a set of items. The axiom is best described by a simple illustration. Suppose that the set of items is $\{A, B, C, D\}$, and suppose that the corresponding probabilities of choosing from this set are (p_A, p_B, p_C, p_D). Now consider a subset $\{A, C\}$ with choice probabilities (q_A, q_C). Then Luce's choice axiom states that $q_A/q_C = p_A/p_C$. In other words, the choice probability ratio between two items is independent of any other items in the set.

Suppose we consider a set of items, and a set of choice probabilities that satisfy Luce's axiom, and consider picking one item at a time out of the set, according to the choice probabilities. Such samples give a total ordering of items, which can be considered as a sample from a distribution over all possible orderings. The form of such a distribution was first considered by Plackett (1975) in order to model probabilities in a K-horse race.

The Plackett-Luce model is applicable when each observation provides either a complete ranking of all items, or a partial ranking of only some of the items, or a ranking of the top few items (see section 3.5 for the

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009. Copyright 2009 by the author(s)/owner(s).

last two scenarios). The applications of the Plackett-Luce distribution and its extensions have been quite varied including horse-racing (Plackett, 1975), document ranking (Cao et al., 2007), assessing potential demand for electric cars (Beggs et al., 1981), modelling electorates (Gormley & Murphy, 2005), and modelling dietary preferences in cows (Nombekela et al., 1994).

Inferring the parameters of the Plackett-Luce distribution is typically done by maximum likelihood estimation (MLE). Hunter (2004) has described an efficient MLE method based on a minorise/maximise (MM) algorithm. In recent years, powerful new message-passing algorithms have been developed for doing approximate deterministic Bayesian inference on large belief networks. Such algorithms are typically both accurate, and highly scalable to large real-world problems. Minka (2005) has provided a unified view of these algorithms, and shown that they differ solely by the measure of information divergence that they minimise. We apply Power EP (Minka, 2004), an algorithm in this framework, to perform Bayesian inference for Plackett-Luce models.

In section 2, we take a more detailed look the Plackett-Luce distribution, motivating it with some alternative interpretations. In section 3, we describe the algorithm at a level of detail where it should be possible for the reader to implement the algorithm in code, giving derivations where needed. In section 4, we apply the algorithm to data generated from a known distribution, to an aggregation of 2002 NASCAR race results, and also to the ranking of genres in the MovieLens data set. Section 5 provides brief conclusions.

2. Plackett-Luce models

A good source for the material in this section, and for rank distributions in general, is the book by Marden (1995). Consider an experiment where N judges are asked to rank K items, and assume no ties. The outcome of the experiment is a set of N rankings $\{y^{(n)} \equiv (y_1^{(n)}, \ldots, y_K^{(n)}) \mid n = 1, \ldots, N\}$ where a ranking is defined as a permutation of the K rank indices; in other words, judge n ranks item i in position $y_i^{(n)}$ (where highest rank is position 1). Each ranking has an associated ordering $\omega^{(n)} \equiv (\omega_1^{(n)}, \ldots, \omega_K^{(n)})$, where an ordering is defined as a permutation of the K item indices; in other words, judge n puts item $\omega_i^{(n)}$ in position i. Rankings and orderings are related by (dropping the judge index) $\omega_{y_i} = i$, $y_{\omega_i} = i$.

The Plackett-Luce (P-L) model is a distribution over rankings y which is best described in term of the associated ordering ω. It is parameterised by a vector $v = (v_1, \ldots, v_n)$ where $v_i \geq 0$ is associated with item index i:

$$PL(\omega \mid v) = \prod_{k=1,\ldots,K} f_k(v) \tag{1}$$

where

$$f_k(v) \equiv f_k(v_{\omega_k}, \ldots, v_{\omega_K}) \triangleq \frac{v_{\omega_k}}{v_{\omega_k} + \cdots + v_{\omega_K}} \tag{2}$$

2.1. Vase model interpretation

The vase model metaphor is due to Silverberg (1980). Consider a multi-stage experiment where at each stage we are drawing a ball from a vase of coloured balls. The number of balls of each colour are in proportion to the v_{ω_k}. A vase differs from an urn only in that it has an infinite number of balls, thus allowing non-rational proportions. At the first stage a ball ω_1 is drawn from the vase; the probability of this selection is $f_1(v)$. At the second stage, another ball is drawn — if it is the same colour as the first, then put it back, and keep on trying until a new colour ω_2 is selected; the probability of this second selection is $f_2(v)$. Continue through the stages until a ball of each colour has been selected. It is clear that equation 1 represents the probability of this sequence. The vase model interpretation also provides a starting point for extensions to the basic P-L model detailed by Marden (1995), for example, capturing the intuition that judges make more accurate judgements at the higher ranks.

2.2. Thurstonian interpretation

A Thurstonian model (Thurstone, 1927) assumes an unobserved random score variable x_i (typically independent) for each item. Drawing from the score distributions and sorting according to sampled score provides a sample ranking — so the distribution over scores induces a distribution over rankings. A key result, due to Yellott (Yellott, 1977), says that if the score variables are independent, and the score distributions are identical except for their means, then the score distributions give rise to a P-L model if and only if the scores are distributed according to a Gumbel distribution.

The CDF $\mathscr{G}(x \mid \mu, \beta)$ and PDF $g(x \mid \mu, \beta)$ of a Gumbel distribution are given by

$$\mathscr{G}(x \mid \mu, \beta) = e^{-z} \tag{3}$$

$$g(x \mid \mu, \beta) = \frac{z}{\beta} e^{-z} \tag{4}$$

where $z(x) = e^{-\frac{x-\mu}{\beta}}$. For a fixed β, $g(x \mid \mu, \beta)$ is an exponential family distribution with natural parameter $v = e^{\frac{\mu}{\beta}}$ which has a Gamma distribution conjugate prior. The use of the notation v for this natural parameter is deliberate — it turns out that $v_i = e^{\frac{\mu_i}{\beta}}$ is the P-L parameter for the i^{th} item in the ranking distribution induced by the Thurstonian model with score distributions $g(x_i \mid \mu_i, \beta)$. The TrueSkill rating system (Herbrich et al., 2007) is based on a Thurstonian model with a Gaussian score distributon. Although this model does not satisfy the Luce Choice Axiom, it has been applied in a large-scale commercial online rating system with much success.

2.3. Maximum likelihood estimation

The typical way to fit a P-L model is by maximum likelihood estimation (MLE) of the parameters v. Hunter (2004) describes a way to do this using a minorise/maximise (MM) algorithm (expectation maximisation (EM) is a special case of an MM algorithm), which is shown to be faster and more robust than the more standard Newton-Raphson method. Furthermore Hunter provides MATLAB code for this algorithm, along with an interesting example of learning a P-L to rank NASCAR drivers across the entire 2002 season of racing. We take up this example further in section 4.2, demonstrating that whilst MLE works well in some settings, it will overfit when there is sparse data. Furthermore, the MM algorithm requires a strong assumption (Assumption 1 of Hunter, 2004) to guarantee convergence: *in every possible partition of the individuals into two nonempty subsets, some individual in the second set ranks higher than some individual in the first set at least once.* As we shall see in the NASCAR data, this assumption is often not satisfied in real examples involving sparse data, and indeed the MM algorithm does not converge.

3. Bayesian Inference

This section makes heavy use of the ideas, notation, and algorithms in (Minka, 2005), and there is not the space to summarise those here. So although we give a complete description of our algorithm, a lot of background information from (Minka, 2005) is assumed.

Suppose that we have a set of observed full orderings $\Omega = \{\omega^{(n)}\}$. We would like to infer the parameters of a P-L model, placing proper priors on them. By Bayes' theorem, the posterior distribution over the parameters is proportional to:

$$p(v) \equiv p(v \mid \Omega) = \prod_{n=0,\ldots,N} \prod_{k=1,\ldots,K} f_k^{(n)}(v) \tag{5}$$

where $f_k^{(0)}$ is a prior, and the remaining $f_k^{(n)}$ are as in equation (2), but now indexed by datum also. As the v_i are positive values, it is natural to assign them Gamma distribution priors, and this is reinforced by the discussion in section 2.2. So we will assume that, for each k,

$$f_k^{(0)} = \mathrm{Gam}(v_k \mid \alpha_0, \beta_0) \tag{6}$$

In general we are interested in recovering the marginals of $p(v)$. We will be inferring a fully factorised approximation to $p(v)$, so the marginals will be a direct output of the inference algorithm. When the approximation is fully factorised, message-passing has a graphical interpretation as a factor graph, with messages passing between variables and factors.

3.1. Preliminaries

The message-passing algorithm described below will make use of both normalised and unnormalised versions of the Gamma distribution:

$$\mathrm{UGam}(x \mid \alpha, \beta) \triangleq x^{\alpha-1} e^{-\beta x} \tag{7}$$

$$\mathrm{Gam}(x \mid \alpha, \beta) \triangleq \frac{\beta^\alpha}{\Gamma(\alpha)} \mathrm{UGam}(x \mid \alpha, \beta) \tag{8}$$

where, for the normalised version, we require $\alpha > 0$ and $\beta > 0$. α is the *shape* parameter, and β is the *rate* parameter (i.e. $1/scale$). The UGam family is useful as it allows us to deal, in a consistent way, with improper distributions.

3.2. The factorisation

We will approximate $p(v)$ as a fully factorised product of Gamma distributions:

$$p(v) \approx q(v) = \prod_{i=1,K} q_i(v_i) = \prod_a \tilde{f}_a(v) \tag{9}$$

$a = (n, k)$ summarises the double index of datum and rank into a single index so as to keep the notation succinct and consistent with (Minka, 2005), and $q_i(v_i) = \mathrm{UGam}(v_i \mid \alpha_i, \beta_i)$. We follow the message-passing treatment in (Minka, 2005, section 4.1). The factors $\tilde{f}_a(v)$, which approximate the P-L factors $f_a(v)$, factorise fully into messages $m_{a\to i}$ from factor a to variable v_i:

$$\tilde{f}_a(v) = \prod_i m_{a\to i}(v_i) \tag{10}$$

where $m_{a\to i}(v_i) = \mathrm{UGam}(v_i \mid \alpha_{ai}, \beta_{ai})$. Collect all terms involving the same variable v_i to define messages from variable v_i to factor a

$$m_{i\to a}(v_i) = \prod_{b\neq a} m_{b\to i}(v_i) \tag{11}$$

The rationale of the message-passing algorithm is to improve the approximating factors $\tilde{f}_a(v)$ one at a time under the assumption that the approximation from the rest of the model is good — i.e. assuming that $q^{\backslash a}(v) = q(v)/\tilde{f}_a(v)$ is a good approximation of $p^{\backslash a}(v) = p(v)/f_a(v)$. Note that

$$q^{\backslash a}(v) = \prod_{b \neq a} \prod_i m_{b \to i}(v_i) = \prod_i m_{i \to a}(v_i) \qquad (12)$$

3.3. The message update

The key quantity that we need to calculate is:

$$q_i'(v_i) = \mathrm{proj}\big[m_{a \to i}(v_i)^{1-\alpha} m_{i \to a}(v_i) \quad \times \\ \int_{v \backslash v_i} \mathrm{d}v \, f_a(v)^\alpha \prod_{j \neq i} m_{a \to j}(v_j)^{1-\alpha} m_{j \to a}(v_j)\big] \qquad (13)$$

where proj denotes K-L projection, and where α is the α-divergence parameter which we can choose to make our problem tractable.[1] Define

$$m_{a \to j}(v_j)^{1-\alpha} m_{j \to a}(v_j) = \mathrm{UGam}(v_j \mid \gamma_{aj}, \delta_{aj}) \qquad (14)$$

The inference algorithm fails if $\mathrm{UGam}(v_j \mid \gamma_{aj}, \delta_{aj})$ becomes improper for any j — however, we have not seen this happen in practice. Individual messages, however, are allowed to and often will become improper. Assuming that $\mathrm{UGam}(v_j \mid \gamma_{aj}, \delta_{aj})$ is proper, we can equivalently replace it with its normalised version $g_{aj} \triangleq \mathrm{Gam}(v_j \mid \gamma_{aj}, \delta_{aj})$ — this simplifies the derivations.

We choose $\alpha = -1$ as our α-divergence. This is needed in order to make the integral tractable, as the true factor $f_a(v)$ then gets inverted, leading to a sum of products of univariate integrals of Gamma distributions.

3.3.1. Projection for a P-L factor

Fixing attention on a specific factor $f_a(v)$ where $a = (n, k)$, with observed ordering ω, we have $f_a(v) = v_{\omega_k} / \sum_{l=k \ldots K} v_{\omega_l}$. So, with an α-divergence of -1, $f_a(v)^\alpha = \sum_{l=k \ldots K} (v_{\omega_l}/v_{\omega_k})$.

When evaluating $q_i'(v_i)$, if v_i does not appear in $f_a(v)$, then $m_{a \to i}(v) = 1$, so we can restrict the calculations to when $i = \omega_r$ for some r. Note, also, that we can ignore any factor in $\prod_{j \neq i} g_{aj}(v_j)$ for which $j \neq \omega_l$ for some l, because these integrate out to 1. We will consider the cases $r = k$ and $r \neq k$ separately.

Case 1 ($i = \omega_k$):

$$\begin{aligned} & g_{ai}(v_i) \int_{v \backslash v_i} f_a(v)^{-1} \prod_{j \neq i} g_{aj}(v_j) dv \\ &= g_{ai}(v_i) \sum_{l=k \ldots K} \int_{v \backslash v_i} (v_{\omega_l}/v_{\omega_k}) \prod_{j \neq i} g_{aj}(v_j) dv \\ &= g_{ai}(v_i) \Big(1 + \frac{1}{v_i} \sum_{l=k+1 \ldots K} \int v_{\omega_l} g_{a\omega_l}(v_{\omega_l}) dv_{\omega_l}\Big) \\ &= g_{ai}(v_i) \Big(1 + \frac{1}{v_i} \sum_{l=k+1 \ldots K} \frac{\gamma_{a\omega_l}}{\delta_{a\omega_l}}\Big) \\ &= \Big(\frac{\delta_{ai}}{\gamma_{ai} - 1} \sum_{l=k+1 \ldots K} \frac{\gamma_{a\omega_l}}{\delta_{a\omega_l}}\Big) \mathrm{Gam}(v_i \mid \gamma_{ai} - 1, \delta_{ai}) \\ &\qquad + \mathrm{Gam}(v_i \mid \gamma_{ai}, \delta_{ai}) \end{aligned} \qquad (15)$$

Case 2 ($i = \omega_r, r \neq k$):

$$\begin{aligned} & g_{ai}(v_i) \int_{v \backslash v_i} f_a(v)^{-1} \prod_{j \neq i} g_{aj}(v_j) dv \\ &= \Big(1 + \frac{\delta_{ai}}{\gamma_{ai} - 1} \sum_{l=k+1 \ldots K, l \neq r} \frac{\gamma_{a\omega_l}}{\delta_{a\omega_l}}\Big) \mathrm{Gam}(v_i \mid \gamma_{ai}, \delta_{ai}) \\ &\qquad + \Big(\frac{\delta_{ak}}{\gamma_{ak} - 1} \frac{\gamma_{ai}}{\delta_{ai}}\Big) \mathrm{Gam}(v_i \mid \gamma_{ai} + 1, \delta_{ai}) \end{aligned} \qquad (16)$$

Note that these both reduce to the general form

$$c \cdot \mathrm{Gam}(v_i \mid a, b) + d \cdot \mathrm{Gam}(v_i \mid a+1, b) \qquad (17)$$

The first two moments for an expression in the form of equation (17) are easily shown to be:

$$\begin{aligned} E(v_i) &= \frac{ca + d(a+1)}{b(c+d)} \\ E(v_i^2) &= \frac{ca(a+1) + d(a+1)(a+2)}{b^2(c+d)} \end{aligned} \qquad (18)$$

The unnormalised projection can then be calculated as

$$q_i'(v_i) = \mathrm{UGam}\Big(v_i \mid \frac{E(v_i)^2}{E(v_i^2) - E(v_i)^2}, \frac{E(v_i)}{E(v_i^2) - E(v_i)^2}\Big) \qquad (19)$$

3.3.2. Message update for a P-L factor

As a clarification to (Minka, 2005), and matching the original Power EP description in (Minka, 2004), the marginal updates and the message updates are:

$$q_i^{new}(v_i) = q_i(v_i)^2 / q_i'(v_i) \qquad (20)$$

$$m_{a \to i}^{new} = \frac{q_i^{new}(v_i)}{m_{i \to a}(v_i)} \qquad (21)$$

[1] For true K-L projection, we need to match the features of the Gamma distribution - namely $\ln(v_i)$ and v_i. However, we will approximate this by just matching the first two moments in order to avoid the non-linear iterative procedure required to retrieve gamma parameters from $E(\ln(v_i))$ and $E(v_i)$.

3.4. Summary of the algorithm

1. Initialise $m_{a\to i}(v_i)$ for all a,i to be uniform, except when $a = (0, k)$, corresponding to the constant prior messages. We set each of these to a broad prior of $\mathrm{UGam}(v_i \mid 3.0, 2.0)$.
2. Repeat until all $m_{a\to i}(v_i)$ converge:
   (a) Pick a factor $a = (n, k)$.
   (b) Compute the messages into the factor using equation (11).
   (c) Compute the projections $q'_i(v_i)$ using equation (19) via equations. (15), (16), (17), (18).
   (d) Update the factor's outgoing messages using equations (20) and (21).

Note that marginals can be recovered at any time by $q_i(v_i) = \prod_a m_{a\to i}$. As there is a degree of freedom in the v_i, the *rate* parameter of the marginals can be collectively scaled so that, for example, the means of the v_i's sum to a specified value; this is useful, for example if you are trying to identify known parameters as we do in section 4.1. Finally, there isn't the space to show the evidence calculation here, but it can be easily derived from the converged unnormalised messages as shown in (Minka, 2005, section 4.4). This computation of the evidence is a further advantage of the fully Bayesian approach as it allows us to build mixture models with different numbers of mixture components and evaluate their Bayes factors (model selection).

3.5. Incomplete rankings

One of the nice properties of the P-L distribution is that it is internally consistent: the probability of a particular ordering does not depend on the subset from which the individuals are assumed to be drawn (see Hunter, 2004 for an outline of a proof, and relation to the Luce Choice Axiom). Suppose we have two sets of items A and B where $B \subset A$. This means that the probability of a particular ordering of the items in B, marginalizing over all possible unknown positions of the items left over in A, is exactly the same as the P-L probability of simply ordering those items in B completely independently from A. The consequence of internal consistency is that each datum can be an incomplete ordering of the total set of items, and yet they can still be combined together consistently, with each datum's likelihood being a simple product of the factors $f_k^{(n)}$ of the items that are ranked in that particular datum. This is extremely useful in practice, as in many applications an individual "judge" may only rank some of the items. An example is the NASCAR data of section 4.2 where a different, but overlapping, set of drivers compete in each race. In terms of our inference algorithm, the incomplete ranking case simply decreases the number of factors that have to be included in the message-passing graph.

Another variation is where top-S rankings have been given. An example might be where users are asked to rank their top-10 movies, or in meta-search where each search engine reports its top-10 (or top-100 etc) documents for a given query. Again this situation can be handled consistently, and in this case the factors $f_k^{(n)}$ for which $k > S$ are removed from the likelihood (1). This is equivalent to marginalizing over all the unknown positions of the other items, but assuming that they are ranked somewhere below the top-S items.

4. Examples

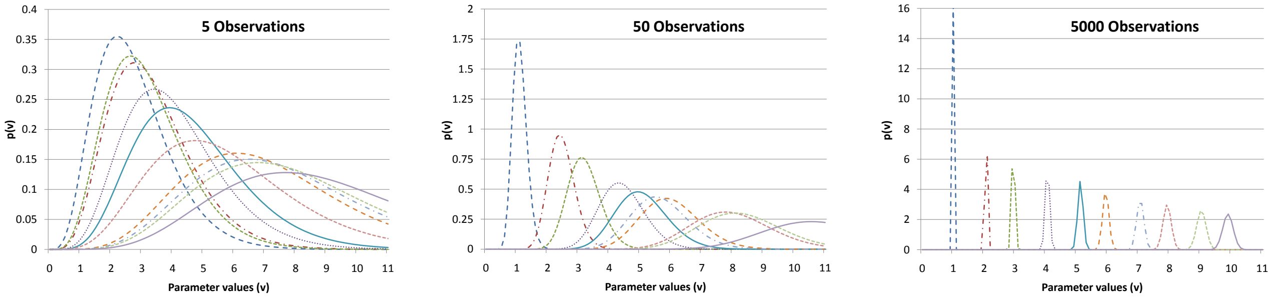
4.1. Inferring known parameters

To verify that the algorithm is doing the right thing, we can generate data from a P-L distribution with known parameters, and then try to infer the parameters. Figure 1 shows the inferred parameters from a P-L model with parameter vector $v = (1.0, 2.0, \ldots, 10.0)$. The marginals in 1a are inferred from 5 observations, those in 1b from 50, and those in 1c from 5000 observations. As expected, the spread of the marginals decreases as the data increases, and the true parameter values are reasonably represented by the marginals.

It is interesting to observe that estimates become less certain for larger parameters. This is perhaps to be expected, as the ratio v_{10}/v_9 in this example is much smaller than the ratio of $v_2/v1$, so the top rank choices are less clear-cut decisions than the bottom ones.

4.2. Ranking NASCAR racing drivers

Hunter (2004) performs a case-study of fitting a P-L model to the NASCAR 2002 season car racing results. In this section we also study this data because it serves as a comparison and highlights a number of advantages of our fully Bayesian approach to the MM MLE method. The 2002 US NASCAR season consisted of 36 races in which a total of 87 different drivers competed. However, any one race involved only 43 drivers. This ranged from some drivers competing in all the races, and some only in one race in the season. This is therefore a good example of the incomplete rankings case discussed in section 3.5. As discussed in section 2.3, Hunter's MM algorithm requires quite a strong assumption for convergence. In many cases, and indeed in this case, this assumption will not be satisfied. In the NASCAR data 4 drivers placed last in every race they entered, thus violating this assumption. There-



(a) 5 observations

(b) 50 observations

(c) 5000 observations

Figure 1. Marginal P-L parameter distributions inferred from data generated from $PL(\omega \mid (1.0, 2.0, \ldots, 10.0))$

fore Hunter had to simply remove these drivers from the model. In contrast, our Bayesian method can be applied to all the data with no problems due to the proper priors that are placed on the P-L parameters. However, for the purposes of making a direct comparison with their work, we follow this and remove these drivers so as to be using exactly the same data set, with a total of 83 drivers.

Table 1 shows the top and bottom 10 drivers as ordered by average place, as well as their rank assigned by both maximum likelihood and Bayesian EP inference. For maximum likelihood the ordering is done by MLE P-L parameter, and for EP the ordering is done by mean P-L parameter. There are some clear differences between the two methods. The MLE method places Jones and Pruett in first and second place respectively — this certainly ties in with their very high (numerically low) average place. However, they only actually raced in one race each compared with some drivers who raced the whole season of 36 races. This is an example of the MLE algorithm overfitting — one race is not enough evidence on which to judge the skill of these drivers, and yet MLE places them right at the top. In contrast the EP inference places these drivers mid-way down the ranking, and also their P-L parameters have high uncertainty compared with other drivers. With further evidence, it is possible that these drivers would rise up the ranking. The EP method ranks Mark Martin in first place, followed by Rusty Wallace: drivers who have raced all 36 races. Similarly, toward the bottom of the table EP method puts Morgan Shepherd at the very bottom instead of some of the other drivers with similar average place but who raced in only one or two races. Morgan Shepherd has raced in 5 races, and so enough evidence has accumulated that he consistently does poorly. Notice that even when the number of races raced is the same (e.g. Martin, Stewart, Wallace, Johnson raced 36 races), neither MLE P-L or EP P-L are equivalent to simply ordering by average place — the P-L model takes into account exactly who is racing in each race: it is better to have won in a race full of good drivers rather than a race of poor drivers.

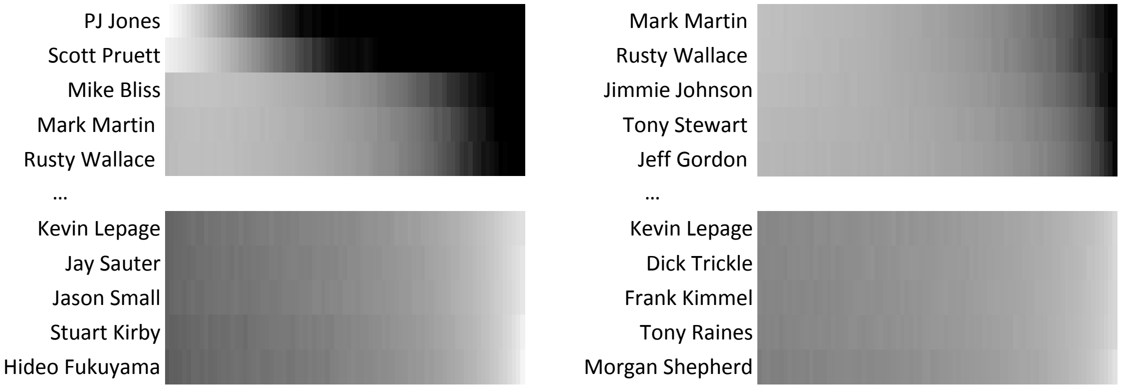
Figure 2 is an alternative way of viewing the inferences about selected NASCAR drivers — the top and bottom 5 drivers as ordered by MLE (2a) and by EP (2b). Instead of showing the inferred P-L parameters, which are a little hard to interpret in themselves, we show the inferred rank marginal distributions implied by the inferred P-L parameters for each driver. This is grey-scale visualisation of the probability that each driver will come at a certain place in a race involving all 83 drivers. As we see the MLE plot is dominated by the over-fitting to the two drivers P J Jones and Scott Pruett, who both have highly skewed distributions toward the top ranks. In contrast the EP plot shows much broader and more reasonable rank marginal distributions, reflecting the fact that even for the best drivers there is high uncertainty in any given race about where they will place.

4.3. Ranking movie genres

The MovieLens data set was collected and is owned by the GroupLens Research Project at the University of Minnesota. The data set consists of 100,000 ratings (1–5) from 943 users on 1682 movies. This data is interesting in that it (a) provides simple demographic information for each user, and (b) provides information about each film as a list of genre vectors — a film can have more than one genre — for example *Romantic Comedy*. We obtained ranking data by creating, for each user, an average rating of each genre across all films seen by the particular user. Each user rated at least 20 films so they each see many genres, but there is no guarantee that a user will see all types of genre. This means the genre rankings are partial lists and the absence of a given genre from an observation is not an indication that a user is giving it a

Table 1. Posterior P-L rankings for top and bottom ten 2002 NASCAR drivers, as given by average place. The parameter estimates v have been normalised to sum to 1 for both MLE and EP so that they are comparable (for EP their means sum to 1). The EP SDev v column shows the standard deviation of the posterior gamma distribution over v.

Driver	Races	Av. place	MLE Rank	MLE v	EP Rank	EP Mean v	EP SDev v
PJ Jones	1	4.00	1	0.1864	18	0.0159	0.0079
Scott Pruett	1	6.00	2	0.1096	19	0.0156	0.0078
Mark Martin	36	12.17	4	0.0235	1	0.0278	0.0047
Tony Stewart	36	12.61	7	0.0184	4	0.0229	0.0040
Rusty Wallace	36	13.17	5	0.0230	2	0.0275	0.0046
Jimmie Johnson	36	13.50	6	0.0205	3	0.0250	0.0042
Sterling Marlin	29	13.86	9	0.0167	6	0.0207	0.0040
Mike Bliss	1	14.00	3	0.0274	23	0.0146	0.0073
Jeff Gordon	36	14.06	8	0.0168	5	0.0213	0.0036
Kurt Busch	36	14.06	12	0.0153	8	0.0198	0.0034
⋮							
Carl Long	2	40.50	75	0.0021	73	0.0062	0.0029
Christian Fittipaldi	1	41.00	77	0.0019	68	0.0075	0.0039
Hideo Fukuyama	2	41.00	83	0.0014	77	0.0054	0.0028
Jason Small	1	41.00	81	0.0017	71	0.0067	0.0036
Morgan Shepherd	5	41.20	78	0.0019	83	0.0041	0.0016
Kirk Shelmerdine	2	41.50	76	0.0021	75	0.0059	0.0028
Austin Cameron	1	42.00	68	0.0029	62	0.0083	0.0043
Dave Marcis	1	42.00	67	0.0030	61	0.0083	0.0043
Dick Trickle	3	42.00	74	0.0022	80	0.0050	0.0022
Joe Varde	1	42.00	71	0.0025	66	0.0078	0.0041



(a) Maximum likelihood

(b) Bayesian EP inference

Figure 2. Marginal posterior rank distributions for top and bottom 5 drivers as ordered by (a) MLE or (b) EP. White indicates high probability and rankings are from left (1st place) to right (83rd place).

Table 2. Normalised P-L parameters for ranking MovieLens genres, with no. of data points in each category in parentheses.

All (943)			Age 25-29 (175)			Age 55-59 (32)		
Genre	Mean	SDev	Genre	Mean	SDev	Genre	Mean	SDev
War	0.0968	0.0036	Film-Noir	0.0920	0.0101	War	0.0873	0.0165
Drama	0.0902	0.0032	Drama	0.0911	0.0075	Thriller	0.0805	0.0147
Film-Noir	0.0828	0.0039	Documentary	0.0867	0.0117	Drama	0.0741	0.0137
Romance	0.0709	0.0026	War	0.0820	0.0070	Film-Noir	0.0681	0.0153
Crime	0.0619	0.0023	Romance	0.0730	0.0060	Mystery	0.0676	0.0131
Mystery	0.0607	0.0023	Crime	0.0570	0.0050	Crime	0.0655	0.0124
Thriller	0.0563	0.0020	Sci-Fi	0.0533	0.0045	Adventure	0.0607	0.0119
Sci-Fi	0.0545	0.0020	Animation	0.0513	0.0049	Western	0.0603	0.0149
Documentary	0.0538	0.0034	Thriller	0.0501	0.0041	Action	0.0595	0.0112
Action	0.0514	0.0018	Mystery	0.0487	0.0043	Romance	0.0569	0.0104
Western	0.0511	0.0027	Action	0.0479	0.0039	Sci-Fi	0.0535	0.0113
Adventure	0.0489	0.0018	Western	0.0461	0.0053	Documentary	0.0459	0.0139
Animation	0.0478	0.0022	Comedy	0.0450	0.0037	Comedy	0.0450	0.0083
Comedy	0.0428	0.0015	Adventure	0.0446	0.0038	Animation	0.0418	0.0119
Musical	0.0397	0.0017	Musical	0.0411	0.0039	Fantasy	0.0418	0.0148
Children's	0.0348	0.0014	Children's	0.0386	0.0036	Musical	0.0365	0.0081
Horror	0.0313	0.0013	Horror	0.0285	0.0027	Horror	0.0278	0.0065
Fantasy	0.0244	0.0013	Fantasy	0.0229	0.0026	Children's	0.0272	0.0064

low ranking. We then built a P-L model using these observations. The advantage of using user rankings rather than ratings is that it removes user bias on the ratings scale, and indeed ordering the genres by mean rating gives significantly different results. Note that we are not ranking genre popularity here — instead we are ranking how well a particular genre was received, although there is likely to be genre-dependent bias in movie selection. So, for example, the algorithm put the *War* genre at the top of the ranking; although war movies were not the most watched type of movie, when watched, they were ranked highly. Table 2 shows the means of the posterior parameter estimates and the corresponding rankings for the whole user population; these are compared with the parameter estimates/rankings for the sub-populations of age 25–29 users and age 55–59 users. Not only are the rankings different, with the younger category preferring *Film-Noir* to the older category's *War* films, but also the uncertainties are higher for the older category due to there only being 32 age 55–59 data points. The division of the users into different categories hints at a straightforward extension of the basic P-L model — a mixture of P-L distributions. An advantage of the EP Bayesian inference is that model evidence can be used to determine the optimum number of components in a mixture. The resulting mixture model can then be used as the basis for a recommender system. We leave this extension for future work.

5. Conclusions

We have described a message-passing algorithm for inferring parameters of a P-L ranking distribution. We have shown that this can accurately learn parameters and their uncertainties from data generated from a known P-L model. We have shown the scalability of the algorithm by running it on real-world data sets, and demonstrated significant advantages over the maximum likelihood approach, especially the avoidance of over-fitting to sparse data.

Future work involves extending the algorithm to learn mixtures of these models. A Bayesian treatment of mixtures should yield insights into clusters of users in e.g. movie rating data such as MovieLens. The Thurstonian interpretation of these models provides insights to how we might build more complex models where the P-L parameters are outputs of other feature-based models, thus extending the range of applications. For example, in a "learning to rank" application we could build a feature-based regression model to link query-document features to P-L ranking parameters. The EP method described is also straightforwardly applied to the extensions of the basic P-L model briefly discussed in section 2.1. These "multi-stage" models have many more parameters, and therefore are likely to benefit even more from a Bayesian treatment.

References

Beggs, S., Cardell, S., & Hausman, J. (1981). Assessing the potential demand for electric cars. *Journ. Econometrics*, *17*, 1–19.

Cao, Z., Liu, T.-Y., Tsai, M.-F., & Li, H. (2007). *Learning to rank: from pairwise approach to listwise approach* (Technical Report). Microsoft Research.

Dwork, C., Kumar, R., Naor, M., & Sivakumar, D. (2001). Rank aggregation methods for the web. *World Wide Web (WWW)* (pp. 613–622).

Gormley, I., & Murphy, T. (2005). *Exploring Irish election data: A mixture modelling approach* (Technical Report). Trinity College Dublin, Dept. Stat.

Herbrich, R., Minka, T., & Graepel, T. (2007). TrueSkill(TM): A Bayesian skill rating system. In *Adv. in Neur. Inf. Proc. Sys. (NIPS) 19*, 569–576.

Hunter, D. R. (2004). MM algorithms for generalized Bradley-Terry models. *Ann. of Stats.*, *32*, 384–406.

Joachims, T., Li, H., Liu, T.-Y., & Zhai, C. (2007). Learning to rank for information retrieval. *Spec. Int. Grp. Info. Retr. (SIGIR) Forum*, *41*, 58–62.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis.* Wiley.

Marden, J. (1995). *Analyzing and modeling rank data.* Chapman and Hall.

Minka, T. (2004). *Power EP* (Technical Report). Microsoft Research.

Minka, T. (2005). *Divergence measures and message passing* (Technical Report). Microsoft Research.

Nombekela, S., Murphy, M., Gonyou, J., & Marden, J. (1994). Dietary preferences in early lactation cows as affected by primary tastes and some common feed flavors. *Journal of Dairy Science*, *77*, 2393–2399.

Plackett, R. (1975). The analysis of permutations. *Applied Stat.*, *24*, 193–202.

Silverberg, A. (1980). *Statistical models for q-permutations.* Doctoral dissertation, Princeton Univ., Dept. Stat.

Thurstone, L. (1927). A law of comparative judgement. *Psychological Reviews*, *34*, 273–286.

Yellott, J. (1977). The relationship between Luce's choice axiom, Thurstone's theory of comparative judgement, and the double exponential distribution. *Journ. Math. Psych.*, *15*, 109–144.