Efficient Bayesian Inference for Generalized Bradley-Terry Models

François Caron
INRIA Bordeaux - Sud-Ouest &
Institut de Mathématiques de Bordeaux, Université Bordeaux I
33405 Talence cedex, France
Francois.Caron@inria.fr

Arnaud Doucet
Departments of Computer Science and Statistics,
University of British Columbia,
Vancouver, BC, V6T 1Z2, Canada
Arnaud@stat.ubc.ca

Abstract

The Bradley-Terry model is a popular approach to describe probabilities of the possible outcomes when elements of a set are repeatedly compared with one another in pairs. It has found many applications including animal behaviour, chess ranking and multiclass classification. Numerous extensions of the basic model have also been proposed in the literature including models with ties, multiple comparisons, group comparisons and random graphs. From a computational point of view, Hunter (2004) has proposed efficient iterative MM (minorization-maximization) algorithms to perform maximum likelihood estimation for these generalized Bradley-Terry models whereas Bayesian inference is typically performed using MCMC (Markov chain Monte Carlo) algorithms based on tailored Metropolis-Hastings (M-H) proposals. We show here that these MM algorithms can be reinterpreted as special instances of Expectation-Maximization (EM) algorithms associated to suitable sets of latent variables and propose some original extensions. These latent variables allow us to derive simple Gibbs samplers for Bayesian inference. We demonstrate experimentally the efficiency of these algorithms on a variety of applications.

Keywords: Bradley–Terry model, data augmentation, EM algorithm, Gibbs sampler, maximum likelihood estimation, MCMC algorithms, MM algorithm, Plackett–Luce model.

1 Introduction

Consider a set of K elements. These elements are repeatedly compared with one another in pairs. For two elements i and j of this set, Bradley and Terry (1952) suggested the following model

$$\Pr(i \text{ beats } j) = \frac{\lambda_i}{\lambda_i + \lambda_j} \tag{1}$$

where $\lambda_l > 0$ is a parameter associated to element $l \in \{1, 2, \ldots, K\}$ that represents its skill rating and we denote $\lambda := \{\lambda_i\}_{i=1}^K$.

This model has found numerous applications. As mentioned in (Hunter, 2004), as early as 1976, a published bibliography on paired comparisons includes several hundred entries (Davidson and Farquhar, 1976). For example, it has been adopted by the World Chess Federation and the European Go Federation to rank players and it is a standard approach to build multiclass classifiers based on the output of binary classifiers (Hastie and Tibshirani, 1998). Various extensions have been proposed to handle home advantage (Agresti, 1990), draws (Rao and Kupper, 1967), multiple (Plackett, 1975; Luce, 1959) and team comparisons (Huang et al., 2006). In particular, the popular extension to multiple comparisons, named the Plackett-Luce model (Plackett, 1975; Luce, 1959), defines a prior distribution over permutations and has been used for ranking of multiple individuals and for choice models (Luce, 1977). The monographs of David (1988) and Diaconis (1988, Chap. 9) provide detailed discussions on the statistical foundations of these models.

For the basic Bradley-Terry model (1), it is possible to find the maximum likelihood (ML) estimate of the skill ratings λ using a simple iterative procedure (Zermelo, 1929; Hunter, 2004). Lange et al. (2000) established that this procedure is a specific case of the general class of algorithms referred to as MM algorithms. Generally speaking, MM algorithms use surrogate minimizing functions of the log-likelihood to define an iterative procedure converging to a local maximum. EM algorithms are thus just a special case of MM algorithms. An excellent survey of the MM approach and its applications can be found in (Lange et al., 2000). Hunter (2004) further derived MM algorithms for generalized Bradley-Terry models and established sufficient conditions under which these algorithms are guaranteed to converge towards the ML estimate.

Recently several authors have proposed to perform Bayesian inference for (generalized) Bradley-Terry models (Adams, 2005; Gormley and Murphy, 2009; Guiver and Snelson, 2009). The resulting posterior density is typically not tractable and needs to be approximated. An Expectation-Propagation method is developed in (Guiver and Snelson, 2009); this yields an approximation of the posterior which can be computed quickly and might be suitable for very large scale applications. However, it relies on a functional approximation of the posterior and the convergence properties of this algorithm are not well-understood. M-H algorithms have been proposed in (Adams, 2005; Gormley and Murphy, 2009). Gormley and Murphy (2009) suggested a carefully designed proposal distribution, though it can perform poorly in some scenarios as demonstrated in section 7.

Our contribution here is three-fold. First, we show that by introducing suitable sets of latent variables, the MM algorithms proposed by Hunter (2004) for the basic Bradley-Terry model and its generalizations to take into account home advantage, ties and multiple comparisons can be reinterpreted as standard EM algorithms. Hence, in cases where these EM algorithms converge slowly, all the acceleration techniques developed for EM algorithms can be directly used. We also believe that this non-trivial reinterpretation is potentially fruitful for statisticians who usually like thinking in terms of latent variables. Note that the latent variables introduced here differ from the ones introduced in the standard Thurstonian interpretation of the Bradley-Terry model (Diaconis, 1988, Chap. 9) and lead to more efficient algorithms as discussed in section 2. Second, using similar ideas, we propose original EM algorithms for some recent generalizations of the Bradley-Terry model including group comparisons and random graphs. Third, based on the sets of latent variables introduced to derive these EM algorithms, we propose Gibbs samplers to perform Bayesian inference in this important class of models. To the best of our knowledge, no Gibbs sampler has ever been proposed in this context. These algorithms have the great advantage of allowing us to bypass the design of proposal distributions for M-H updates and we demonstrate experimentally that they perform very well.

The rest of this paper is organized as follows. In section 2, we consider the basic Bradley-Terry model (1). Based on the introduction of a suitable set of latent variables, we present an EM reinterpretation of the MM algorithm presented by Hunter (2004) for Maximum a Posteriori (MAP) parameter estimation and an original data augmentation algorithm to sample from the posterior. In section 3, various standard extensions of the Bradley-Terry model allowing for home advantage, ties and competition between teams are described. EM algorithms and original Gibbs sampling schemes are proposed. The Plackett-Luce model (Plackett, 1975; Luce, 1959), a very popular generalization of the Bradley-Terry model for multiple comparisons, is presented in section 4. A discussion on identifiability issues and estimation of hyperparameters is given in section 5. Algorithms applicable to further extensions of the Bradley-Terry model to choice models, random graphs and classification are presented in section 6. In section 7, these algorithms are applied to the NASCAR 2002 dataset and to chess competition data.

2 Bradley-Terry model

Suppose we have observed a number of pairwise comparisons among K individuals. We denote by D the associated data. Let also w_{ij} denote the number of comparisons where i beats j, $w_i = \sum_{j=1,j\neq i}^{K} w_{ij}$ the total number of wins of element i and $n_{ij} = w_{ij} + w_{ji}$ the total number of comparisons between i and j. Based on the Bradley-Terry model (1), the log-likelihood function is given by

$$\begin{aligned}\ell(\lambda) &= \sum_{1\leq i\neq j\leq K} \left[w_{ij}\log\lambda_i - w_{ij}\log(\lambda_i+\lambda_j)\right]\\ &= \sum_{i=1}^{K} w_i \log\lambda_i - \sum_{1\leq i<j\leq K} n_{ij}\log(\lambda_i+\lambda_j)\end{aligned}$$

where the notation $1 \leq i \neq j \leq K$ is an abuse of notation to denote the set $\left\{(i,j)\in\{1,...,K\}^2\right.$ such that $\left. i\neq j\right\}$ and $1\leq i<j\leq K$ stands for $\left\{(i,j)\in\{1,...,K\}^2 \text{ such that } i<j\right\}$.

We seek to introduce latent variables which are such that the resulting complete log-likelihood admits a simple form. It is well-known that the Bradley-Terry model enjoys the following Thurstonian interpretation (Diaconis, 1988, Chap. 9): for each pair $1 \leq i < j \leq K$ and for each associated pair comparison $k = 1, \ldots, n_{ij}$, let $Y_{ki} \sim \mathcal{E}(\lambda_i)$ and $Y_{kj} \sim \mathcal{E}(\lambda_j)$ where $\mathcal{E}(\varsigma)$ is the exponential distribution of rate parameter ς then

$$\Pr(Y_{ki} < Y_{kj}) = \frac{\lambda_i}{\lambda_i + \lambda_j}.$$

These latent variables can be interpreted as arrival times and the individual with the lowest arrival time wins. These latent variables would allow us to define EM and data augmentation algorithms. However, instead of introducing these variables, we introduce for each pair i, j the latent random variable

$$Z_{ij} = \sum_{k=1}^{n_{ij}} \min(Y_{kj}, Y_{ki}).$$

It follows from the properties of the exponential distribution that $\min(Y_{ki}, Y_{kj}) \sim \mathcal{E}(\lambda_i + \lambda_j)$. Additionally, the sum of identically distributed exponential random variables is Gamma distributed with shape given by the number of variables and rate equal to the exponential rate so

$$Z_{ij} \sim \mathcal{G}(n_{ij}, \lambda_i + \lambda_j)$$

and the resulting complete log-likelihood remains simple. As Z_{ij} is a deterministic function of $\{Y_{ki}, Y_{kj}\}_{k=1,\ldots,n_{ij}}$, the fraction of missing information is reduced when the latent variables $\{Z_{ij}\}$ are introduced instead of $\{Y_{ki}, Y_{kj}\}$. This leads to faster rates of convergence for the resulting EM and data augmentation algorithms (Liu, 2001, Chap. 6).

To summarize, for $1 \leq i < j \leq K$ such that $n_{ij} > 0$, we introduce the latent variables $Z = \{Z_{ij}\}$ which are such that

$$p\left(z \middle| D, \lambda\right) = \prod_{1 \leq i < j \leq K | n_{ij} > 0} \mathcal{G}(z_{ij}; n_{ij}, \lambda_i + \lambda_j) \tag{2}$$

The resulting complete log-likelihood is given by

$$
\begin{aligned}
\ell_c(\lambda) &= \sum_{i=1}^{K} w_i \log \lambda_i - \sum_{1\leq i<j\leq K} n_{ij} \log(\lambda_i + \lambda_j) \\
&\quad + \sum_{1\leq i<j\leq K} \left[n_{ij} \log(\lambda_i + \lambda_j) - (\lambda_i + \lambda_j) z_{ij} + (n_{ij} - 1) \log z_{ij} - \log \Gamma(n_{ij})\right] \\
&= \sum_{i=1}^{K} w_i \log \lambda_i - \sum_{1\leq i<j\leq K \mid n_{ij}>0} \left[(\lambda_i + \lambda_j) z_{ij} - (n_{ij} - 1) \log z_{ij} + \log \Gamma(n_{ij})\right] \qquad (3)
\end{aligned}
$$

where Γ is the Gamma function. If we assign additionally a prior to λ such that

$$
p\left(\lambda\right) = \prod_{i=1}^{K} \mathcal{G}(\lambda_i; a, b) \qquad (4)
$$

as in (Gormley and Murphy, 2009; Guiver and Snelson, 2009) then we can maximize the resulting log-posterior using the EM algorithm which proceeds as follows at iteration t:

$$
\lambda^{(t)} = \arg\max_{\lambda} Q(\lambda, \lambda^{(t-1)}), \qquad (5)
$$

where

$$
\begin{aligned}
Q(\lambda, \lambda^*) &= \mathbb{E}_{Z|D,\lambda^*}\left[\ell_c(\lambda)\right] + \log\ p\left(\lambda\right) \qquad (6) \\
&\equiv \sum_{i=1}^{K} \left[(a - 1 + w_i) \log \lambda_i - b\lambda_i\right] - \sum_{1\leq i<j\leq K} (\lambda_i + \lambda_j) \frac{n_{ij}}{\lambda_i^* + \lambda_j^*}
\end{aligned}
$$

with "$\equiv$" meaning "equal up to terms independent of the first argument of the Q function". Using (5), it follows that

$$
\lambda_i^{(t)} = \frac{a - 1 + w_i}{b + \sum_{j\neq i} \frac{n_{ij}}{\lambda_i^{(t-1)} + \lambda_j^{(t-1)}}}. \qquad (7)
$$

For $a = 1$ and $b = 0$, the MAP and ML estimates coincide. In this case (6) is exactly the minorizing function of the MM algorithm proposed in (Hunter, 2004, Eq. (10)) and thus the MM algorithm is given by (7).

Based on the same latent variables, we present a simple data augmentation algorithm for sampling from the posterior distribution $p\left(\lambda, z \middle| D\right)$. By construction, we can update Z conditional upon λ using (2) and the conditional $p\left(\lambda | z, D\right) \propto p\left(\lambda\right) \exp\left(\ell_c(\lambda)\right)$ can be expressed easily as the gamma prior is a conjugate prior for the complete data likelihood

$$\exp(\ell_c(\lambda)) \propto \prod_{i=1}^{K} \left[\lambda_i^{w_i} \exp\left\{ - \left(\sum_{i<j|n_{ij}>0} z_{ij} + \sum_{i>j|n_{ij}>0} z_{ji} \right) \lambda_i \right\} \right].$$

The data augmentation sampler at iteration t proceeds as follows:

- For $1 \leq i < j \leq K$ s.t. $n_{ij} > 0$, sample

$$Z_{ij}^{(t)} | D, \lambda^{(t-1)} \sim \mathcal{G}\left(n_{ij}, \lambda_i^{(t-1)} + \lambda_j^{(t-1)} \right). \tag{8}$$

- For $i = 1, \ldots, K$, sample

$$\lambda_i^{(t)} | D, Z^{(t)} \sim \mathcal{G}\left(a + w_i, b + \sum_{i<j|n_{ij}>0} Z_{ij}^{(t)} + \sum_{i>j|n_{ij}>0} Z_{ji}^{(t)} \right). \tag{9}$$

3 Generalized Bradley-Terry models

3.1 Home advantage

Consider now that the pairwise comparisons are modeled using the Bradley-Terry model with “home-field advantage” (Agresti, 1990) where

$$\Pr(i \text{ beats } j) = \begin{cases} \frac{\theta\lambda_i}{\theta\lambda_i + \lambda_j} & \text{if } i \text{ is home,} \\ \frac{\lambda_i}{\lambda_i + \theta\lambda_j} & \text{if } j \text{ is home.} \end{cases} \tag{10}$$

The parameter θ, $\theta > 0$, measures the strength of the home-field advantage ($\theta > 1$) or disadvantage ($\theta < 1$). Let a_{ij} be the number of times that i is at home and beats j and b_{ij} is the number of times that i is at home and loses to j.

The log-likelihood of the skill ratings λ and θ is given by

$$\ell(\lambda, \theta) = c \log \theta + \sum_{i=1}^{K} w_i \log \lambda_i - \sum_{1 \leq i \neq j \leq K} n_{ij} \log(\theta\lambda_i + \lambda_j)$$

where $n_{ij} = a_{ij} + b_{ij}$ is the number of times i plays at home against j, $c = \sum_{1\leq i\neq j\leq K} a_{ij}$ is the total number of home-field wins and w_i is the total number of wins of element i.

For $1 \leq i \neq j \leq K$ such that $n_{ij} > 0$, let us introduce the latent variables $Z = \{Z_{ij}\}$ which are such that

$$p\left(z\right| D, \lambda) = \prod_{1\leq i\neq j\leq K|n_{ij}>0} \mathcal{G}(z_{ij}; n_{ij}, \theta\lambda_i + \lambda_j). \tag{11}$$

The associated complete data log-likelihood is given by

$$\ell_c(\lambda, \theta) = c\log\theta + \sum_{i=1}^{K} w_i \log\lambda_i - \sum_{1\leq i\neq j\leq K|n_{ij}>0} \left[(\theta\lambda_i + \lambda_j)z_{ij} + \log\Gamma\left(n_{ij}\right) - \left(n_{ij} - 1\right)\log z_{ij}\right]. \tag{12}$$

Using independent priors for λ and θ, i.e. $p\left(\lambda, \theta\right) = p\left(\lambda\right) p\left(\theta\right)$, where $p\left(\lambda\right)$ is defined as (4) and

$$\theta \sim \mathcal{G}(a_\theta, b_\theta), \tag{13}$$

then we have

$$\begin{aligned} Q(\left(\lambda, \theta\right), \left(\lambda^*, \theta^*\right)) &= \mathbb{E}_{Z|D,\lambda^*,\theta^*}\left[\ell_c(\lambda, \theta)\right] + \log\ p\left(\lambda, \theta\right) \\ &\equiv \left(a_\theta - 1 + c\right)\log\theta - b_\theta\theta + \sum_{i=1}^{K}\left[\left(a - 1 + w_i\right)\log\lambda_i - b\lambda_i\right] \\ &- \sum_{1\leq i\neq j\leq K} n_{ij}\frac{\theta\lambda_i + \lambda_j}{\theta^*\lambda_i^* + \lambda_j^*}. \end{aligned}$$

We cannot maximize Q analytically w.r.t. (λ, θ). Instead we first maximize $Q(\left(\lambda, \theta^{(t-1)}\right), \left(\lambda^{(t-1)}, \theta^{(t-1)}\right))$ w.r.t. λ, then $Q((\lambda^{(t)}, \theta), (\lambda^{(t)}, \theta^{(t-1)}))$ w.r.t. to θ. The resulting algorithm is a special instance of the Expectation Conditional Maximization (ECM) algorithm (Meng and Rubin, 1993). We obtain

$$\lambda_i^{(t)} = \frac{a - 1 + w_i}{b + \sum_{1\leq i\neq j\leq K}\left\{\frac{\theta^{(t-1)}n_{ij}}{\theta^{(t-1)}\lambda_i^{(t-1)} + \lambda_j^{(t-1)}} + \frac{n_{ji}}{\theta^{(t-1)}\lambda_j^{(t-1)} + \lambda_i^{(t-1)}}\right\}} \text{ for } i = 1, \ldots, K, \tag{14}$$

$$\theta^{(t)} = \frac{a_\theta - 1 + c}{b_\theta + \sum_{1\leq i\neq j\leq K}\frac{n_{ij}\ \lambda_i^{(t)}}{\theta^{(t-1)}\lambda_i^{(t)} + \lambda_j^{(t)}}}. \tag{15}$$

For $a = a_\theta = 1$ and $b = b_\theta = 0$, i.e. if we use flat priors, this EM algorithm is similar to the MM algorithm proposed in (Hunter, 2004, pp. 389).

Using the same latent variables, we can sample from the posterior distribution of (λ, θ, Z) using the Gibbs sampler which updates iteratively Z, λ and θ as follows at iteration t:

- For $1 \leq i \neq j \leq K$ s.t. $n_{ij} > 0$, sample

$$Z_{ij}^{(t)} | D, \lambda^{(t-1)}, \theta^{(t-1)} \sim \mathcal{G}\left(n_{ij}, \theta^{(t-1)}\lambda_i^{(t-1)} + \lambda_j^{(t-1)}\right). \tag{16}$$

- For $i = 1, \dots, K$, sample

$$\lambda_i^{(t)} | D, \theta^{(t-1)}, Z^{(t)} \sim \mathcal{G}\left(a + w_i, b + \theta^{(t-1)} \sum_{j \neq i | n_{ij} > 0} Z_{ij}^{(t)} + \sum_{j \neq i | n_{ij} > 0} Z_{ji}^{(t)}\right). \tag{17}$$

- Sample

$$\theta^{(t)} | D, \lambda^{(t)}, Z^{(t)} \sim \mathcal{G}\left(a_\theta + c, b_\theta + \sum_{i=1}^{K} \lambda_i^{(t)} \sum_{j \neq i | n_{ij} > 0} Z_{ij}^{(t)}\right). \tag{18}$$

3.2 Model with ties

If we now want to allow for ties in pairwise comparisons, we can use the following model proposed by Rao and Kupper (1967)

$$\Pr(i \text{ beats } j) = \frac{\lambda_i}{\lambda_i + \theta\lambda_j}, \tag{19}$$

$$\Pr(i \text{ ties } j) = \frac{(\theta^2 - 1)\lambda_i\lambda_j}{(\lambda_i + \theta\lambda_j)(\theta\lambda_i + \lambda_j)} \tag{20}$$

where $\theta > 1$. The log-likelihood function for (λ, θ) is given by

$$\begin{aligned}\ell(\lambda, \theta) &= \sum_{1 \leq i \neq j \leq K} \left[w_{ij} \log \frac{\lambda_i}{\lambda_i + \theta\lambda_j} + \frac{t_{ij}}{2} \log \frac{(\theta^2 - 1)\lambda_i\lambda_j}{(\theta\lambda_i + \lambda_j)(\lambda_i + \theta\lambda_j)} \right] \\ &= \sum_{1 \leq i \neq j \leq K} \left[s_{ij} \log \frac{\lambda_i}{\lambda_i + \theta\lambda_j} + \frac{t_{ij}}{2} \log(\theta^2 - 1) \right]\end{aligned}$$

where $t_{ij} = t_{ji}$ is the number of ties between i and j and $s_{ij} = w_{ij} + t_{ij}$.

For $1 \leq i \neq j \leq K$ such that $s_{ij} > 0$, let us introduce the latent variables $Z = \{Z_{ij}\}$

which are such that

$$p\left(z\right|D,\lambda)=\prod_{1\leq i\neq j\leq K|s_{ij}>0}\mathcal{G}(z_{ij};s_{ij},\lambda_i+\theta\lambda_j)$$

which yields the following complete log-likelihood

$$\ell_c(\lambda,\theta)=T\log(\theta^2-1)+\sum_{1\leq i\neq j\leq K|s_{ij}>0}\left[s_{ij}\log\lambda_i-(\lambda_i+\theta\lambda_j)z_{ij}+(s_{ij}-1)\log z_{ij}-\log\Gamma(s_{ij})\right] \tag{21}$$

where $T=\frac{1}{2}\sum_{1\leq i\neq j\leq K}t_{ij}$ is the total number of ties. Using independent priors for λ and θ, i.e. $p\left(\lambda,\theta\right)=p\left(\lambda\right)p\left(\theta\right)$, where $p\left(\lambda\right)$ is defined as (4) and

$$\overline{\theta}\sim\mathcal{G}(a_\theta,b_\theta), \tag{22}$$

where $\overline{\theta}=\theta-1$, then we obtain

$$\begin{aligned}Q(\left(\lambda,\theta\right),\left(\lambda^*,\theta^*\right))&=\mathbb{E}_{Z|D,\lambda^*,\theta^*}\left[\ell_c(\lambda,\theta)\right]+\log\ p\left(\lambda,\theta\right)\\&\equiv T\log(\theta^2-1)+(a_\theta-1)\log(\theta-1)-b_\theta\theta\\&+\sum_{1\leq i\neq j\leq K}s_{ij}\left(\log\lambda_i-\frac{\lambda_i+\theta\lambda_j}{\lambda_i^*+\theta^*\lambda_j^*}\right)+\sum_{i=1}^{K}\left[(a-1)\log\lambda_i-b\lambda_i\right]\end{aligned}$$

and we recover once again the minorizing function in (Hunter, 2004, pp. 389-390) for $a=1$ and $b=0$. Once more we cannot maximize $Q(\left(\lambda,\theta\right),\left(\lambda^*,\theta^*\right))$ analytically w.r.t (λ,θ) but we can use an ECM strategy and successively maximize Q w.r.t. λ conditional on $\theta^{(t-1)}$, then maximize w.r.t. θ conditional on $\lambda^{(t)}$. This yields the following procedure

- For $i=1,\ldots,K$, set

$$\lambda_i^{(t)}=\left(a-1+\sum_{j\neq i}s_{ij}\right)\left[b+\sum_{j\neq i}\frac{s_{ij}}{\lambda_i^{(t-1)}+\theta^{(t-1)}\lambda_j^{(t-1)}}+\frac{\theta^{(t-1)}s_{ji}}{\theta^{(t-1)}\lambda_i^{(t-1)}+\lambda_j^{(t-1)}}\right]^{-1}. \tag{23}$$

- Set

$$\theta^{(t)}=\frac{1}{2}\left(\frac{a_\theta-1+2T}{c^{(t)}+b_\theta}\right)\left(1+\sqrt{1+4\left(c^{(t)}+b_\theta\right)\frac{a_\theta-1+c^{(t)}+b_\theta}{(a_\theta-1+2T)^2}}\right) \tag{24}$$

where

$$c^{(t)} = \sum_{1\leq i\neq j\leq K} \frac{s_{ij}\lambda_j^{(t)}}{\lambda_i^{(t)} + \theta^{(t-1)}\lambda_j^{(t)}}. \tag{25}$$

Using the same latent variables, we can sample from the posterior distribution of (λ, θ, Z) using the following Gibbs sampler which updates iteratively Z, λ and θ as follows at iteration t:

- For $1 \leq i \neq j \leq K$ s.t. $s_{ij} > 0$, sample

$$Z_{ij}^{(t)}|D, \lambda^{(t-1)}, \theta^{(t-1)} \sim \mathcal{G}\left(s_{ij}, \lambda_i^{(t-1)} + \theta^{(t-1)}\lambda_j^{(t-1)}\right). \tag{26}$$

- For $i = 1, \ldots, K$, sample

$$\lambda_i^{(t)}|D, \theta^{(t-1)}, Z^{(t)} \sim \mathcal{G}\left(a + \sum_{j\neq i} s_{ij}, b + \sum_{j\neq i|s_{ij}>0} Z_{ij}^{(t)} + \theta^{(t-1)} \sum_{j\neq i|s_{ij}>0} Z_{ji}^{(t)}\right). \tag{27}$$

- Sample

$$\theta^{(t)}|D, \lambda^{(t)}, Z^{(t)} \sim p(\theta|D, \lambda^{(t)}, Z^{(t)}) \tag{28}$$

where

$$p(\theta|D, Z, \lambda) \propto (\theta^2 - 1)^T(\theta - 1)^{a_\theta - 1} \exp\left[-\left(b_\theta + \sum_{1\leq i\neq j\leq K|s_{ij}>0} Z_{ij}\right)\theta\right] 1_{\theta>1}. \tag{29}$$

It is possible to sample from (29) exactly. By performing a change of variable $\overline{\theta} = \theta - 1$, we obtain

$$p(\overline{\theta}|D, Z, \lambda) \propto (\overline{\theta}^2 + 2\overline{\theta})^T(\overline{\theta})^{a_\theta - 1} \exp\left[-\left(b_\theta + \sum_{1\leq i\neq j\leq K|s_{ij}>0} Z_{ij}\right)\overline{\theta}\right] \tag{30}$$

which is a mixture of Gamma distributions.

3.3 Group comparisons

Consider now that we have n pairwise comparisons between teams. For each comparison $i = 1, \ldots, n$, let $T_i^+ \subset \{1, \ldots, K\}$ be the winning team and $T_i^- \subset \{1, \ldots, K\}$ the losing team where $T_i^+ \cap T_i^- = \varnothing$ and $T_i = T_i^+ \cup T_i^-$. Recently Huang et al. (2006) have proposed the following model

$$\Pr(T_i^+ \text{ beats } T_i^-) = \frac{\sum_{j \in T_i^+} \lambda_j}{\sum_{j \in T_i} \lambda_j}. \tag{31}$$

The log-likelihood function for λ is thus given by

$$\ell(\lambda) = \sum_{i=1}^n \left[\log \left(\sum_{j \in T_i^+} \lambda_j \right) - \log \left(\sum_{j \in T_i} \lambda_j \right) \right].$$

For $i = 1, ..., n$ we introduce the latent variables $Z = \{Z_i\}$ and $C = \{C_i\}$ such that

$$p\left(z, c \mid D, \lambda\right) = p\left(z \mid D, \lambda\right) P\left(c \mid D, \lambda\right)$$

with

$$p\left(z \mid D, \lambda\right) = \prod_{i=1}^n \mathcal{E}\left(z_i; \sum_{j \in T_i} \lambda_j \right),$$
$$\Pr\left(c \mid D, \lambda\right) = \prod_{i=1}^n \frac{\lambda_{c_i}}{\sum_{j \in T_i^+} \lambda_j} \text{ with } c_i \in T_i^+$$

where $\mathcal{E}\left(x; \alpha\right)$ is the exponential density of argument x and rate parameter α. It follows that the complete log-likelihood is given by

$$\ell_c(\lambda) = \sum_{i=1}^n \left[\log \lambda_{c_i} - \left(\sum_{j \in T_i} \lambda_j \right) z_i \right]. \tag{32}$$

The Q function associated to the EM algorithm is given by

$$\begin{aligned} Q(\lambda, \lambda^*) &= \mathbb{E}_{Z,C|D,\lambda^*}\left[\ell_c(\lambda)\right] + \log\ p\left(\lambda\right) \\ &\equiv \sum_{i=1}^n \sum_{j \in T_i^+} \left[\log \lambda_j \frac{\lambda_j^*}{\sum_{k \in T_i^+} \lambda_k^*} - \frac{\sum_{j \in T_i} \lambda_j}{\sum_{j \in T_i} \lambda_j^*} \right] + \sum_{k=1}^K \left[(a-1) \log \lambda_k - b \lambda_k \right] \\ &\equiv \sum_{k=1}^K \left[\left(a - 1 + \lambda_k^* \sum_{i=1}^n \frac{\alpha_{ik}}{\sum_{j \in T_i^+} \lambda_j^*} \right) \log \lambda_k - \lambda_k \left(b + \sum_{i=1}^n \frac{\gamma_{ik}}{\sum_{j \in T_i} \lambda_j^*} \right) \right] \end{aligned}$$

where $\alpha_{ik} = 1$ if $k \in T_i^+$ and 0 otherwise and $\gamma_{ik} = 1$ if $k \in T_i$ and 0 otherwise. It follows that the EM update is given by

$$\lambda_k^{(t)} = \frac{a - 1 + \lambda_k^* \sum_{i=1}^n \frac{\alpha_{ik}}{\sum_{j \in T_i^+} \lambda_j^{(t-1)}}}{b + \sum_{i=1}^n \frac{\gamma_{ik}}{\sum_{j \in T_i} \lambda_j^{(t-1)}}}. \tag{33}$$

Using the same latent variables, we obtain a data augmentation sampler to sample from $p\left(\lambda, z, c \middle| D\right)$ by iteratively sampling (Z, C) and λ. This proceeds as follows at iteration t:

- For $i = 1, ..., n$, sample

$$\begin{aligned} & Z_i^{(t)} | D, \lambda^{(t-1)} \sim \mathcal{E}\left(\sum_{j \in T_i} \lambda_j^{(t-1)}\right), \\ & \Pr\left(C_i^{(t)} = k | D, \lambda^{(t-1)}\right) = \frac{\lambda_k^{(t-1)}}{\sum_{j \in T_i^+} \lambda_j^{(t-1)}},\ k \in T_i^+. \end{aligned} \tag{34}$$

- For $k = 1, \ldots, K$, sample

$$\lambda_k^{(t)} | D, Z^{(t)}, C^{(t)} \sim \mathcal{G}\left(a + \sum_{i=1}^n \delta_{k, C_i^{(t)}}, b + \sum_{i=1}^n \gamma_{ik} Z_i^{(t)}\right) \tag{35}$$

  where $\delta_{u,v} = 1$ if $u = v$ and 0 otherwise.

4 Multiple comparisons

We now consider the popular Plackett-Luce model (Luce, 1959; Plackett, 1975) which is an extension of the Bradley-Terry model to comparisons involving more than two elements. Assume that $p_i \leq K$ individuals are ranked for comparison i where $i = 1, ..., n$. We write $\rho_i = (\rho_{i1}, \ldots, \rho_{ip_i})$ where ρ_{i1} is the first individual, ρ_{i2}, the second, etc. The Plackett-Luce model assumes

$$\Pr(\rho_i | \lambda) = \prod_{j=1}^{p_i} \frac{\lambda_{\rho_{ij}}}{\sum_{k=j}^{p_i} \lambda_{\rho_{ik}}} = \prod_{j=1}^{p_i - 1} \frac{\lambda_{\rho_{ij}}}{\sum_{m=j}^{p_i} \lambda_{\rho_{im}}}. \tag{36}$$

For $i = 1, \ldots, n$ and $j = 1, \ldots, p_i - 1$, we introduce the following latent variables $Z = \{Z_{ij}\}$

$$p\left(z \middle| D, \lambda\right) = \prod_{i=1}^n \prod_{j=1}^{p_i - 1} \mathcal{E}(z_{ij}; \sum_{m=j}^{p_i} \lambda_{\rho_{im}}) \tag{37}$$

which leads to the complete log-likelihood

$$\ell_c(\lambda) = \sum_{i=1}^{n} \sum_{j=1}^{p_i - 1} \left[\log \lambda_{\rho_{ij}} - \left(\sum_{m=j}^{p_i} \lambda_{\rho_{im}} \right) z_{ij} \right]. \tag{38}$$

The Q function associated to the EM algorithm is given by

$$\begin{aligned} Q(\lambda, \lambda^*) &= \mathbb{E}_{Z|D,\lambda^*} \left[\ell_c(\lambda) \right] + \log \; p\left(\lambda\right) \\ &\equiv \sum_{i=1}^{n} \sum_{j=1}^{p_i - 1} \left[\log \lambda_{\rho_{ij}} - \frac{\sum_{m=j}^{p_i} \lambda_{\rho_{im}}}{\sum_{m=j}^{p_i} \lambda^*_{\rho_{im}}} \right] + \sum_{k=1}^{K} \; \left[(a-1) \log \lambda_k - b \lambda_k \right] \end{aligned}$$

which is once again equivalent to the majorizing function in (Hunter, 2004, pp. 398) for $a = 1$, $b = 0$. It follows that the EM algorithm is given at iteration t by

$$\lambda_k^{(t)} = (a - 1 + w_k) \left[b + \sum_{i=1}^{n} \left(\sum_{j=1}^{p_i - 1} \frac{\delta_{ijk}}{\sum_{m=j}^{p} \lambda_{\rho_{im}}^{(t-1)}} \right) \right]^{-1} \tag{39}$$

where w_k is the number of rankings where the k^{th} individual is not in the last ranking position and δ_{ijk} is defined as

$$\delta_{ijk} = \begin{cases} 1 & \text{if } k \in \{\rho_{ij}, \ldots, \rho_{ip_j}\} \\ 0 & \text{otherwise} \end{cases}$$

i.e. δ_{ijk} is the indicator of the event that individual k receives a rank no better than j in the i^{th} ranking.

To sample from $p\left(\lambda, z \middle| D\right)$, we can use the following data augmentation sampler. At iteration t, this proceeds as follows:

- For $i = 1, ..., n$, for $j = 1, \ldots, p_i - 1$, sample

$$Z_{ij}^{(t)} | D, \lambda^{(t-1)} \sim \mathcal{E}(\sum_{m=j}^{p_i} \lambda_{\rho_{im}}^{(t-1)}). \tag{40}$$

- For $k = 1, \ldots, K$, sample

$$\lambda_k^{(t)} | D, Z^{(t)} \sim \mathcal{G}(a + w_k, b + \sum_{i=1}^{n} \sum_{j=1}^{p_i - 1} \delta_{ijk} Z_{ij}^{(t)}). \tag{41}$$

Using the same augmentation (37) together with additional latent variables, EM and Gibbs samplers can be defined for further extensions of these models such as mixtures of Plackett-Luce models (Gormley and Murphy, 2008).

5 Discussion

5.1 Identifiability

Consider the basic Bradley-Terry model and its extensions to group comparisons and multiple comparisons. Let us define

$$\Lambda = \sum_{i=1}^{K} \lambda_i \text{ and } \pi_i = \frac{\lambda_i}{\Lambda}$$

and write $\pi := \{\pi_i\}_{i=1}^{K}$. The likelihood is invariant to a rescaling of the vector λ so the parameter Λ is not likelihood identifiable and

$$p(\pi, \Lambda | D) = p(\pi | D) p(\Lambda).$$

From (4), it follows that $\pi \sim \mathcal{D}(a, \ldots, a)$ where $\mathcal{D}$ is the Dirichlet distribution and $\Lambda \sim \mathcal{G}(Ka, b)$, hence

$$\Lambda^{MAP} = \frac{aK - 1}{b}.$$

To improve the mixing of the MCMC algorithms in this context, an additional sampling step can be added where we normalize the current parameter estimate $\lambda^{(t)}$ and then rescale them randomly using a prior draw for Λ.

- For $i = 1, \ldots, K$, set $\lambda_i^{*(t)} = \frac{\lambda_i^{(t)}}{\sum_{j=1}^{K} \lambda_j^{(t)}} \Lambda^{(t)}$ where $\Lambda^{(t)} \sim \mathcal{G}(Ka, b)$.

This step can drastically improve the mixing of the Markov chain. However, if we are only interested in the normalized values π of λ then this additional step is useless.

As an alternative, it is also possible to consider an EM algorithm for the basic Bradley-Terry model which does not require the introduction of a scale parameter. Assume $\pi \sim \mathcal{D}(a, \ldots, a)$. In order to construct a complete data likelihood for which the Dirichlet distribution is a conjugate prior, let us introduce latent variables M_{ij},

$C_{ij} = (C_{ij1}, \dots, C_{ijM_{ij}})$ for $1 \leq i \neq j \leq K$ such that $n_{ij} > 0$

$$\begin{aligned} M_{ij} &\sim NB(n_{ij}, \pi_i + \pi_j), \\ \Pr(C_{ijk} = l) &= \frac{\pi_l}{\sum_{n \neq i,j} \pi_n} \text{ for } k = 1, \dots M_{ij} \text{ and } l \neq i, j \end{aligned}$$

where $NB(r, p)$ is the negative binomial distribution. The complete log-likelihood is given by

$$\ell_c(\pi) = \sum_{i=1}^{K} \sum_{j=1, j \neq i}^{K} w_{ij} \log \pi_i + \sum_{i=1}^{K} \sum_{j=1, j \neq i}^{K} \sum_{k \neq i,j} \left[\log \binom{n_{ij} + m_{ij} - 1}{n_{ij} - 1} + r_{ijk} \log \pi_k \right] \quad (42)$$

where r_{ijk} is the number of $c_{ijl}, l = 1, \dots, m_{ij}$ that take value k. Omitting the terms independent of π, the Q function is given by

$$\begin{aligned} Q(\pi, \pi^*) &= \mathbb{E}_{M|D,\pi^*} \left[\mathbb{E}_{C|D,M,\pi^*} \left[\ell_c(\pi) \right] \right] + \log \ p(\pi) \\ &\equiv \mathbb{E}_{M|D,\pi^*} \left[\sum_{i=1}^{K} \sum_{j=1, j \neq i}^{K} w_{ij} \log \pi_i + \sum_{i=1}^{K} \sum_{j=1, j \neq i}^{K} \sum_{k \neq i,j} \log \binom{n_{ij} + M_{ij} - 1}{n_{ij} - 1} \right. \\ &\qquad \left. + M_{ij} \frac{\pi_k^*}{\sum_{l \neq i,j} \pi_l^*} \log \pi_k \right] + \log \ p(\pi) \\ &\equiv \sum_{i=1}^{K} \sum_{j=1, j \neq i}^{K} w_{ij} \log \pi_i + \sum_{i=1}^{K} \sum_{j=1, j \neq i}^{K} \sum_{k \neq i,j} n_{ij} \frac{\pi_k^*}{\pi_i^* + \pi_j^*} \log \pi_k + (a-1) \sum_{i=1}^{K} \log \pi_k + C \\ &\equiv \sum_{k=1}^{K} (w_k + \pi_k^* \sum_{i=1, i \neq k}^{K} \sum_{j=1, j \neq i,k}^{K} \frac{n_{ij}}{\pi_i^* + \pi_j^*} \log \pi_k) + (a-1) \sum_{i=1}^{K} \log \pi_k + C \end{aligned}$$

where C is a term independent of π. It follows that the EM update is given by

$$\pi_k^{(t)} \propto a - 1 + w_k + \pi_k^{(t-1)} \sum_{i=1, i \neq k}^{K} \sum_{j=1, j \neq i,k}^{K} \frac{n_{ij}}{\pi_i^{(t-1)} + \pi_j^{(t-1)}} \quad (43)$$

with $\sum_{k=1}^{K} \pi_k^{(t)} = 1$. Although the above EM algorithm does not rely on unidentifiable scaling parameters, it suffers from a slow convergence rate. When π_k takes small values, $\sum_{i \neq k} \sum_{j \neq k} \frac{n_{ij}}{\pi_i + \pi_j}$ is large and it slows down the convergence of the EM algorithm. The same augmentation can be used to define a Gibbs sampler, but the same slow mixing issues arise for the Markov chain.

5.2 Hyperparameter estimation

The prior (4) is specified by the hyperparameters a and b. However, the inverse scale parameter b is not likelihood identifiable so there is no point assigning a prior to it. However it might be interesting to set a prior $p(a)$ on a and estimate it from the data. Given λ, we have

$$p\left(a|\,\lambda\right) \propto p\left(a\right) \underbrace{\left(b^{K}\prod_{i=1}^{K}\lambda_{i}\right)^{a}}_{l_1(a)} \underbrace{\Gamma^{-K}(a)}_{l_2(a)}.$$

It is possible to sample from this density using auxiliary variables U_1, U_2 as described in (Damien et al., 1999). We introduce

$$p\left(a, u_1, u_2|\,\lambda\right) \propto p\left(a\right)\mathbb{I}\left\{u_1 < l_1\left(a\right)\right\}\mathbb{I}\left\{u_2 < l_2\left(a\right)\right\}.$$

A Gibbs sampler can now be implemented to sample from $p\left(a, u_1, u_2|\,\lambda\right)$. We can directly sample from the full conditionals of U_1 and U_2

$$U_1|\,\lambda \sim \mathcal{U}\left(0, l_1\left(a\right)\right), \quad U_2|\,\lambda \sim \mathcal{U}\left(0, l_2\left(a\right)\right)$$

where $\mathcal{U}\left(\alpha, \beta\right)$ is the uniform distribution on $\left(\alpha, \beta\right)$. The full conditional of a given u_1, u_2 is given by

$$p\left(a|\,\lambda, u_1, u_2\right) \propto p\left(a\right)\mathbb{I}_{A_1 \cap A_2}\left(a\right)$$

where

$$A_i = \left\{a : l_i\left(a\right) > u_i\right\}.$$

Alternatively we can update a using a M-H random walk on $\log(a)$. We can propose $a^{\star} = \exp(\sigma_a^2 z)a$ where $Z \sim \mathcal{N}(0,1)$ and $a^{\star}$ is accepted with probability

$$\min\left\{1, \frac{p(a^{\star})}{p(a)}\left(\frac{\Gamma(a)}{\Gamma(a^{\star})}\right)^{K}\left(b^{K}\prod_{i=1}^{K}\lambda_i\right)^{a^{\star}-a}\right\}.$$

6 Further extensions

6.1 Random graphs with a given degree sequence

A model closely related to Bradley-Terry has been proposed for undirected random graphs with K vertices (Holland and Leinhardt, 1981; Chatterjee et al., 2010; Park and Newman, 2004). In this model, graphs with the same degree sequence $(d_1, \dots, d_K)$, where d_i is the degree of node i, are supposed to be equiprobable. It can be formalized by saying that the degree sequence is a sufficient statistic for a probability distribution on graphs (Chatterjee et al., 2010).

In this model, we have $r_{ij} = 1$ if there is an edge between i and j for $1 \leq i < j \leq K$ and it is assumed that

$$\Pr(r_{ij} = 1) = \frac{\lambda_i \lambda_j}{1 + \lambda_i \lambda_j}$$

where $\lambda_k > 0$ for $k \in \{1, \dots, K\}$. Given the observations $D = \{r_{ij}\}_{1 \leq i < j \leq K}$, the log-likelihood function for λ is given by

$$\ell(\lambda) = \sum_{1 \leq i < j \leq K} \left[r_{ij} \log\left(\lambda_i \lambda_j\right) - \log\left(1 + \lambda_i \lambda_j\right) \right].$$

We introduce the following latent variables $Z = \{Z_{ij}\}_{1 \leq i < j \leq K}$ such that

$$p\left(z \middle| D, \lambda\right) = \prod_{1 \leq i < j \leq K} \mathcal{E}(z_{ij}; \lambda_i + \frac{1}{\lambda_j}).$$

The complete log-likelihood is given by

$$\ell_c(\lambda) = \sum_{1 \leq i < j \leq K} \left[r_{ij} \log \lambda_i - (1 - r_{ij}) \log \lambda_j - (\lambda_i + \frac{1}{\lambda_j}) z_{ij} \right] \tag{44}$$

The Q function associated to the EM algorithm is given by

$$\begin{aligned} &Q(\lambda, \lambda^*) = \mathbb{E}_{Z|D,\lambda^*}\left[\ell_c(\lambda)\right] + \log\ p\left(\lambda\right) \\ &\equiv \sum_{i=1}^{K} \left\{ \log \lambda_i \left[(a-1) + \sum_{j>i} r_{ij} - \sum_{j<i} (1 - r_{ij}) \right] - \lambda_i \left(b + \sum_{j>i} \frac{1}{\lambda_i^* + \frac{1}{\lambda_j^*}} \right) - \frac{1}{\lambda_i} \sum_{j<i} \frac{1}{\lambda_j^* + \frac{1}{\lambda_i^*}} \right\}. \end{aligned}$$

Solving $\partial Q(\lambda, \lambda^*)/\partial \lambda_i = 0$ requires solving a quadratic equation. For sake of brevity, we do not present these details here.

Once again, we can define a data augmentation sampler to sample from $p\left(\lambda, z\middle| D\right)$ by iteratively sampling Z and λ. This proceeds as follows at iteration t:

- For $1 \leq i < j \leq K$, sample

$$Z_{ij}^{(t)}\Big| D, \lambda^{(t-1)} \sim \mathcal{E}(\lambda_i^{(t-1)} + \frac{1}{\lambda_j^{(t-1)}}). \tag{45}$$

- For $i = 1, ..., K$, sample

$$\lambda_i^{(t)}|D, Z^{(t)} \sim \mathcal{GIG}\left(2\left(\sum_{j>i} Z_{ij}^{(t)} + b\right), 2\sum_{j<i} Z_{ij}^{(t)}, a + \sum_{j>i} r_{ij} - \sum_{j<i}(1 - r_{ij})\right). \tag{46}$$

Here $\mathcal{GIG}\left(\alpha, \beta, \gamma\right)$ denotes the generalized inverse Gaussian distribution (see e.g. (Barndorff-Nielsen and Shephard, 2001)) whose density for an argument x is proportional to

$$x^{\gamma-1}\exp\left\{-\left(\alpha x + \beta/x\right)/2\right\}.$$

Algorithms to sample exactly from this distribution are available.

6.2 Choice models

Other extensions of the Bradley-Terry model are the choice models introduced by Restle (1961) and Tversky (1972a,b) in psychology; see also (Wickelmaier and Schmid, 2004; Görür et al., 2006). In these models, we are given a set of n elements. To each element i is associated a set of K features represented by a binary vector $f_i \in \{0, 1\}^K$. The probability that element i is chosen over element j is given by

$$\pi_{ij} = \frac{\sum_{k=1}^K \lambda_k f_{ik}(1 - f_{jk})}{\sum_{k=1}^K \lambda_k f_{ik}(1 - f_{jk}) + \sum_{k=1}^K \lambda_k f_{jk}(1 - f_{ik})}$$

where $\lambda_k > 0$ is a weight representing the importance of feature k. The term $\sum_{k=1}^K \lambda_k f_{ik}(1-f_{jk})$ corresponds to the sum of the weights of features possessed by object i but not object j. EM and Gibbs algorithms can be derived by following the same construction as with group comparisons.

6.3 Categorical data

Let consider the following original model for categorical data analysis

$$\begin{aligned}\Pr(Y_i = k) &= \frac{\sum_{j=1}^{p} \exp(X_{ij})\lambda_{kj}}{\sum_{l=1}^{K}\sum_{j=1}^{p} \exp(X_{ij})\lambda_{lj}} \\ &= \frac{\exp(X_i)^T \lambda_k}{\sum_{l=1}^{K} \exp(X_i)^T \lambda_l} \end{aligned} \tag{47}$$

where $X_i \in \mathbb{R}^p$ is a vector of covariates and $\lambda_k \in \mathbb{R}_+^p$ for $k = 1, \dots, K$. This model could be used as an alternative to the multinomial logit model (Agresti, 1990). By introducing latent variables $Z_i \sim \mathcal{E}\left(\sum_{l=1}^{K} \exp(X_i)^T \lambda_l\right)$, we can define EM and Gibbs algorithms resp. to maximize the posterior distribution of the parameters λ_k and sample from it when the prior is given by (4).

7 Experimental results

In all the above models, the parameter b is just a scaling parameter on λ_k. As the likelihood is invariant to a rescaling of the λ_k, this parameter does not have any influence on inference. Hence to ensure that the MAP estimate satisfies $\sum_{k=1}^{K} \widehat{\lambda}_k = 1$, we set $b = Ka - 1$ henceforth as explained in section 5. We demonstrate our algorithms on one synthetic and two real-world data sets.

7.1 Synthetic Data

We first study the Plackett-Luce model, comparing experimentally the mixing properties of the Gibbs sampler relative to a slightly modified version of the M-H algorithm proposed by Gormley and Murphy (2009). In this latter paper, the authors propose to update the skill parameters simultaneously using the following proposal distribution[1] at iteration t

$$\text{for } k = 1, \dots, K,\ \lambda_k^\star \sim \mathcal{G}\left(a + w_k, b + \sum_{i=1}^{n}\left(\sum_{j=1}^{p_i - 1} \frac{\delta_{ijk}}{\sum_{m=j}^{p} \lambda_{\rho_{im}}^{(t-1)}}\right)\right)$$

We simulated 500 dataset of n rankings of $K = 4$ individuals, for various values of n with $a = 5$. For each dataset, 10,000 iterations of the Gibbs sampler presented in section 4 were

[1] The authors actually use a normal approximation of the gamma distribution, and work with normalized data. To obtain similar algorithms, we consider unnormalized data.

run. The sample lag-1 autocorrelation was then computed for the four skill parameters. For a given sample size n, the mean value over skill parameters and simulated data is reported on Figure 1 together with 90% confidence bounds. The algorithm of Gormley and Murphy (2009) performs reasonably well when the sample size is large, which is the case for the voting data they considered, but poorly for small sample sizes.

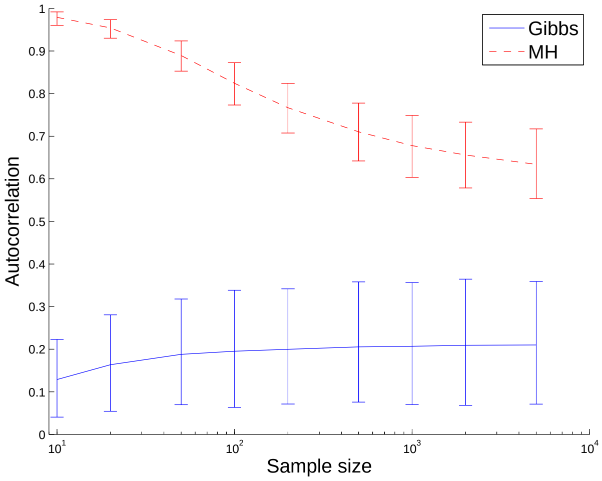


Figure 1: Sample lag-1 autocorrelation as a function of the sample size n for the Gibbs sampler and a modified version of the M-H algorithm of Gormley and Murphy (2009).

7.2 Nascar 2002 dataset

NASCAR is the primary sanctioning body for stock car auto racing in the United States. Each race involves 43 drivers. During the 2002 season, 87 different drivers participated in 36 races. Some drivers participated in all of the races while others participated in only one. We propose to apply the Plackett-Luce model with gamma prior on the parameters. The NASCAR dataset[2] has been studied by Hunter (2004), who noted that the MLE cannot be found for the original data set, as the likelihood has no maximizer. Assumption 1 in (Hunter, 2004) is not met as four drivers placed last in each race they entered, and therefore had to be removed. This does not need to be done if we follow a Bayesian approach.

We first focus here on predicting the outcome of the next race based on the previous ones, starting from race 5, evaluating the performances with the test log-likelihood (TLL);

[2] The data can be downloaded from http://www.stat.psu.edu/ dhunter/code/btmatlab/

i.e. we compute for $i = 5, \ldots, 35$

$$TLL(i) = \log \Pr(\rho_{i+1} | \widehat{\lambda}^{(i)})$$

where ρ_i is the ranking of race i and $\widehat{\lambda}^{(i)}$ is the MAP/posterior mean estimate obtained with races 1 to i.

The mean value and 90% confidence bounds of TLL are represented in Figure 2 w.r.t. the value of a for the EM algorithm. The EM algorithm was initialized using $(\lambda_1^{(0)}, \ldots, \lambda_{87}^{(0)}) = (\frac{1}{87}, \ldots, \frac{1}{87})$. The Gibbs sampler was initialized at the same value and the parameter a was assigned a flat improper prior, initialized at the value 1 and sampled as described in section 5.2. We ran 50,000 iterations with 2,000 burn-in. As detailed in Section 5, only the normalized weights π_i are likelihood identifiable. The mean test log-likelihood obtained with the posterior mean estimate of the normalized weights is also represented in Figure 2.

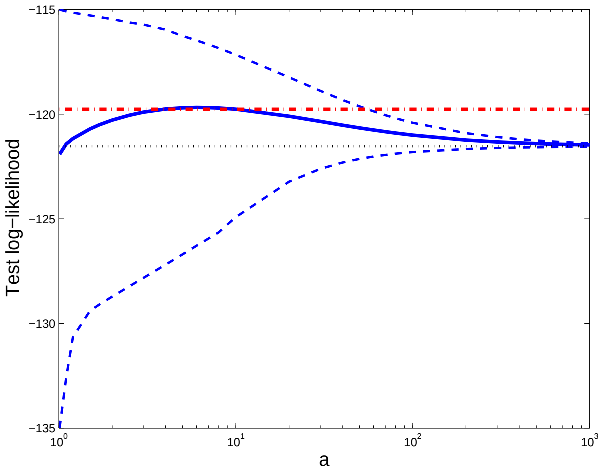


Figure 2: Test log-likelihood on the Nascar 2002 dataset. From race 5 to 35, we compute the log-likelihood of the next race based on the MAP/posterior mean estimates. Mean (solid blue line) and 90% interval (dashed blue line) of the test log-likelihood with MAP estimates is represented w.r.t. to the parameter a. The black dotted line represents the test log-likelihood obtained with a uniform prior. The dash-dot red line represents the mean test log-likelihood using the posterior mean estimate of the normalized weights, with a flat prior on a.

Skill ratings are usually represented on the real line, and we use the following one-to-one mapping $\beta_i = \log \pi_i - \log 1/87$. The marginal posterior densities of the reparameterized skill ratings for the first four drivers according to their average place are reported in Figure 3. The Bayesian approach can effectively capture the uncertainty in the skill ratings of the drivers. ML and posterior mean estimates together with standard deviations of the marginal posteriors are reported in Table 1 for the first ten and last ten drivers according to average place. As explained above, MLE cannot be performed with the full set of drivers, and skills estimates are not available for the four drivers who always finished at the last place.

Table 1: Top ten and bottom ten drivers according to average place, along with ML and posterior mean estimates of the skill parameters in β space. Marginal posterior standard deviations are also provided.

Driver	Races	Average place	ML estimate	Posterior Mean	Posterior Std
P. Jones	1	4.00	2.79	0.14	0.53
S. Pruett	1	6.00	2.25	0.12	0.53
M. Martin	36	12.17	0.71	0.85	0.17
T. Stewart	36	12.61	0.47	0.66	0.17
R. Wallace	36	13.17	0.70	0.84	0.17
J. Johnson	36	13.50	0.58	0.74	0.17
S. Marlin	29	13.86	0.37	0.55	0.19
M. Bliss	1	14.00	0.87	0.05	0.53
J. Gordon	36	14.06	0.38	0.58	0.17
K. Busch	36	14.06	0.29	0.51	0.17
...					
M. Shepherd	5	41.20	-1.81	-1.11	0.41
K. Shelmerdine	2	41.50	-1.68	-0.81	0.50
A. Cameron	1	42.00	-1.36	-0.50	0.55
D. Marcis	1	42.00	-1.34	-0.49	0.54
D. Trickle	3	42.00	-1.67	-0.94	0.45
J. Varde	1	42.00	-1.51	-0.55	0.55
A. Hillenburg	2	43.00	—	-1.46	0.69
G. Bradberry	1	43.00	—	-1.09	0.69
J. Hedlesky	1	43.00	—	-1.02	0.68
R. Renfrow	1	43.00	—	-1.04	0.70

7.3 Chess data

Rating the skills of chess players is of major practical interest. It allows organizers of a tournament to avoid having strong players playing against each other at early stages, or to restrict the tournament to players with skills above a given threshold. The international chess federation adopted the so-called “Elo” system which is based on the Bradley-Terry model (Elo, 1978). For historical considerations about the chess rating system, the reader

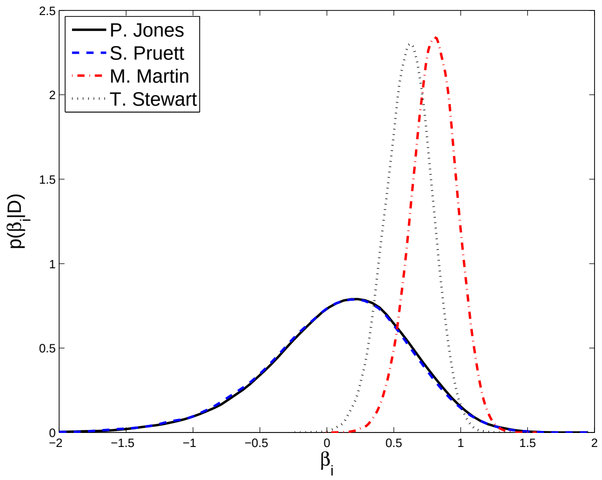


Figure 3: Marginal posterior distribution for the modified skill ratings β_i of the first 4 drivers according to their average place. P. Jones and S. Pruett only participated in 1 race each, while M. Martin and T. Stewart participated in 36 races.

should refer to Glickman (1995).

We consider here game-by-game chess results over 100 months, consisting of 65,053 matches between 8631 players[3]. The outcome E_i of each game i is either win (+1), tie (+0.5) or loss (0). We estimate the parameters of the Bradley-Terry model with ties presented in section 3.2 on the first 95 months and then compute the expected outcome $\widehat{E}_i \in [0,1]$ of the games of the last 5 months. The hyperparameters for the tie parameter θ are set to $a_\theta = 1$, $b_\theta = 0$. Given the large sample size, it is not possible to sample from Eq. (29) as the number of elements in the mixture is very large. We therefore use a M-H step with a normal random walk proposal of standard deviation 0.1. The root mean squared error, defined by

$$\sqrt{\frac{1}{n_t}\sum_{i=1}^{n_t}\left(E_i - \widehat{E}_i\right)^2}$$

where $n_t = 11182$ is the number of games in the last 5 months, is reported for predictions based on MAP estimates and full Bayesian prediction based on the Gibbs sampler outcomes, for different values of the hyperparameter a. We also report the results of the Gibbs sampler outcome when a is assigned a flat prior. EM and Gibbs samplers were initialized at $(\lambda_1^{(0)}, \dots, \lambda_{8631}^{(0)}) = (\frac{1}{8631}, \dots, \frac{1}{8631})$ and $\theta^{(0)} = 1.5$. The Gibbs samplers were run with 10,000 iterations and 1,000 burn-in iterations. The results are reported in Figure 4. The results demonstrate the benefit of penalizing the skill rating parameters

[3] Chess data can be downloaded from http://kaggle.com/chess

and the improvement brought up by a full Bayesian analysis. In Figure 5 we also report the autocorrelation function associated to the parameter θ and the skill parameters with largest mean values. The Markov chain displays good mixing properties.

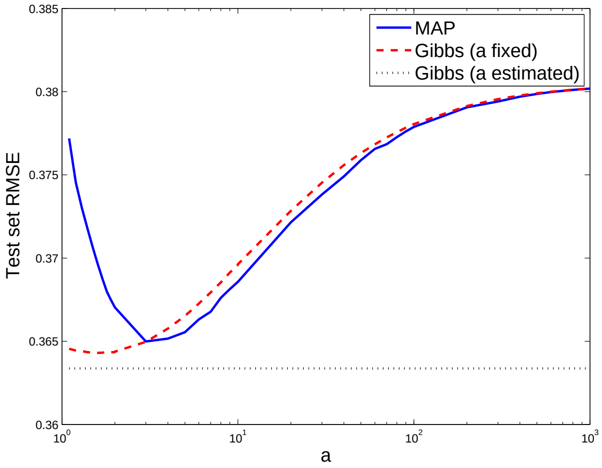
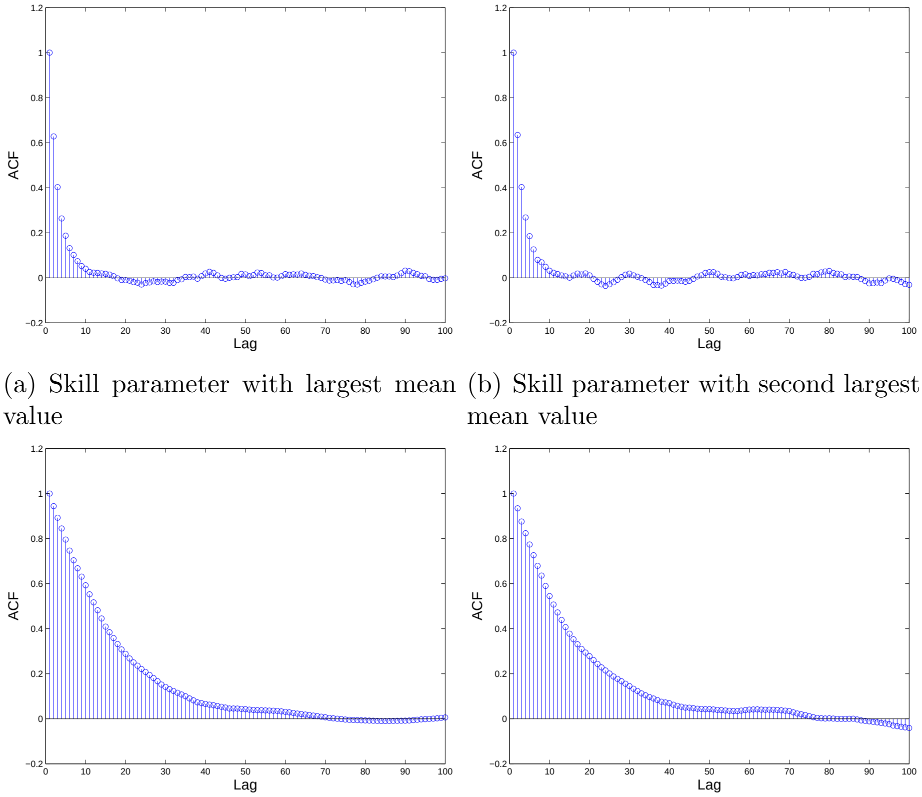


Figure 4: Test root mean square error on the chess dataset for different values of the parameter a. Based on an history of 95 months, we predict the outcome of the games of the last 5 months.

8 Conclusion

The Bradley-Terry model and its generalizations arise in numerous applications. We have shown here that most of the MM algorithms proposed in Hunter (2004) can be reinterpreted as special cases of EM algorithms. Additionally we have proposed original EM algorithms for some recent generalizations of the Bradley-Terry models. Finally we have shown how the latent variables introduced to derive these EM algorithms lead straightforwardly to Gibbs sampling algorithms. These elegant MCMC algorithms mix experimentally well and outperform a recently proposed M-H algorithm.

Acknowledgment. The authors are grateful to Persi Diaconis for helpful discussions and pointers to references on the Plackett-Luce and random graph models and to Luke Bornn for helpful comments.



(a) Skill parameter with largest mean value

(b) Skill parameter with second largest mean value

(c) Tie parameter θ

(d) Hyperparameter a

Figure 5: Autocorrelation functions for (a-b) the two skill parameters with largest mean values, (c) the parameter θ and for (d) the hyperparameter a. The fast decrease indicates that the Markov chain mixes well. The parameters θ and a are updated with a M-H step which explains their relatively low mixing.

9 Supplementary material

Matlab files for Bayesian inference with generalized Bradley-Terry models: Zip file containing Matlab files to apply EM algorithms and Gibbs samplers for the following models: Bradley-Terry model (with or without ties, with or without home advantage), Plackett-Luce model. A README file describes the different files in the archive. (BayesBT.zip, zip file)

References

Adams, E. (2005). Bayesian analysis of linear dominance hierarchies. *Animal Behaviour*, 69:1191–1201.

Agresti, A. (1990). *Categorical Data Analysis*. Wiley.

Barndorff-Nielsen, O. and Shephard, N. (2001). Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society B*, 63:167–241.

Bradley, R. and Terry, M. (1952). Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika*, 39:324–345.

Chatterjee, S., Diaconis, P., and Sly, A. (2010). Random graphs with a given degree sequence. Technical report, Stanford University.

Damien, P., Wakefield, J., and Walker, S. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society B*, 61:331–344.

David, H. (1988). *The method of paired comparisons.* Oxford University Press.

Davidson, R. and Farquhar, P. (1976). A bibliography on the method of paired comparisons. *Biometrics*, 32:241–252.

Diaconis, P. (1988). *Group representations in probability and statistics, IMS Lecture Notes*, volume 11. Institute of Mathematical Statistics.

Elo, A. (1978). *The rating of Chess Players, Past and present.* Arco Pub.

Glickman, M. (1995). A comprehensive guide to chess ratings. Technical report, Department of Statistics, Boston University.

Gormley, I. and Murphy, T. (2008). Exploring voting blocs with the Irish electorate: a mixture modeling approach. *Journal of the American Statistical Association*, 103:1014–1027.

Gormley, I. and Murphy, T. (2009). A grade of membership model for rank data. *Bayesian Analysis*, 4:265–296.

Görür, D., Jäkel, F., and Rasmussen, C. (2006). A choice model with infinitely many latent features. In *International Conference on Machine Learning.*

Guiver, J. and Snelson, E. (2009). Bayesian inference for Plackett-Luce ranking models. In *International Conference on Machine Learning.*

Hastie, T. and Tibshirani, R. (1998). Classification by pairwise coupling. *Annals of Statistics*, 26:451–471.

Holland, P. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76:33–65.

Huang, T.-K., Weng, R., and Lin, C.-J. (2006). Generalized Bradley-Terry models and multi-class probability estimates. *Journal of Machine Learning Research*, 7:85–115.

Hunter, D. (2004). MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, 32:384–406.

Lange, K., Hunter, D., and Yang, I. (2000). Optimization transfer using surrogate objective functions (with discussion). *Journal of Computational and Graphical Statistics*, 9:1–59.

Liu, J. (2001). *Monte Carlo Methods for Scientific Computing.* Springer-Verlag: New York.

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

Luce, R. (1977). The choice axiom after twenty years. *Journal of Mathematical Psychology*, 15:215–233.

Meng, X.-L. and Rubin, D. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80:267–278.

Park, J. and Newman, M. (2004). The statistical mechanics of networks. *Physical Review E*, 70:066117.

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

Rao, P. and Kupper, L. (1967). Ties in paired-comparison experiments: A generalization of the Bradley-Terry model. *Journal of the American Statistical Association*, 62:194–204.

Restle, F. (1961). *Psychology of judgement and choice*. New-York: Wiley.

Tversky, A. (1972a). Choice by elimination. *Journal of Mathematical Psychology*, 9:341–367.

Tversky, A. (1972b). Elimination by aspects: a theory of choice. *Psychological Review*, 79:281–299.

Wickelmaier, F. and Schmid, C. (2004). A Matlab function to estimate choice model parameters from paired-comparison data. *Behavior Research Methods, Instruments and Computers*, 36:29–40.

Zermelo, E. (1929). Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. *Math. Z.*, 29:436–460.