Annales Univ. Sci. Budapest., Sect. Comp. 36 (2012) 143–157

EM ALGORITHMS FOR GENERALIZED BRADLEY–TERRY MODELS

Villő Csiszár (Budapest, Hungary)

Communicated by Imre Kátai

(Received December 20, 2011; accepted January 12, 2012)

Abstract. The Bradley–Terry model and its generalizations are popular models for paired (or multiple) comparisons of individuals or teams. The literature describes several approaches to maximum likelihood estimation of the parameters. Here we propose the use of the EM scheme, we study its convergence properties, and compare it with previous algorithms.

1. Introduction

The Bradley–Terry model [2] is applicable to situations in which paired comparisons are made between individuals in a group. Suppose that there are m individuals, and there is a positive parameter λ_i attached to the ith individual, representing his overall ability ($i = 1, \dots, m$). The model then asserts that when comparing individuals i and j, the probability that i is the winner equals

$$P(\text{individual } i \text{ beats individual } j) = \frac{\lambda_i}{\lambda_i + \lambda_j}.$$

This model has widespread applications in areas such as statistics, sports, and machine learning.

Key words and phrases: Paired comparisons, ranking models, EM algorithm.
2010 Mathematics Subject Classification: 62F07, 62F10.
Supported by the European Union and the European Social Fund under the grant agreement no. TÁMOP 4.2.1./B-09/KMR-2010-0003. Also supported by OTKA grant K 76481.
https://doi.org/10.71352/ac.36.143

The Bradley–Terry model has been generalized in several different ways. One generalization, called the Plackett–Luce model [9], allows for the comparison and ordering of more than two individuals at a time. Also in this direction, Huang, Weng, and Lin [5] study a case when two *teams* are compared, where the team's overall ability depends on the abilities of its members. Agresti [1] introduced a model for paired comparisons when one of the contestants has a "home-field advantage". Rao and Kupper [10] modifies the Bradley–Terry model to allow for ties.

The maximum likelihood estimation of the parameters in these models has been an important issue from the beginning. Under mild conditions the existence of the ML estimator is guaranteed, and it can be found by iterative methods in each case. The Newton–Raphson method is applicable, of course. However, Hunter [6] proposes the use of MM algorithms, which are simpler, faster and more robust. Huang, Weng and Lin [5] describe a descent-direction type algorithm for their model. We remark that while maximum likelihood works well in many settings, Guiver and Snelson [4] have shown that, in the case of the Plackett–Luce model, it overfits when there is sparse data. They propose a Bayesian approach to overcome this difficulty.

The main contribution of our paper is that generalized Bradley–Terry models can be formulated using exponentially or geometrically distributed latent variables, and thus it is natural to consider the EM scheme for likelihood maximization.

The paper is structured as follows. In Section 2 we summarize the EM-scheme, and briefly review the work of Hunter on MM algorithms. In Sections 3 through 6 we derive EM algorithms for all the models mentioned above. Finally, Section 7 discusses the convergence properties of the algorithms.

2. EM and MM algorithms

Suppose that we have a parametric model for our observed data X; denote the vector of parameters by β. The likelihood function is denoted by $L(X, \beta)$; maximum likelihood estimation amounts to finding the parameter vector $\hat{\beta}$ where the likelihood function attains its global maximum over the parameter space.

The EM algorithm is an iterative procedure for finding $\hat{\beta}$ in incomplete-data settings. Suppose that the observed data X is "incomplete". By this we mean that there is a set of complete data Z, of which X is a function. Denote the complete data likelihood function by $L(Z, \beta)$. The EM algorithm then proceeds as follows:

Start with initial values $\beta^{(0)}$. In the $(s+1)$st iteration perform the following two steps:

1. Expectation step (E-step): Calculate the expected value of the complete data log-likelihood, given the incomplete data and parameter values $\beta^{(s)}$:

$$Q(\beta, \beta^{(s)}) = E(\log L(Z, \beta)|X, \beta^{(s)}).$$

2. Maximization step (M-step): Maximize $Q(\beta, \beta^{(s)})$ in β to obtain the values $\beta^{(s+1)}$.

The EM algorithm in its above generality was introduced by Dempster et al. [3]. They showed that the likelihood $L(X, \beta^{(s)})$ is increasing in s. However, in general it is not guaranteed that the sequence of likelihood iterates converges to the global maximum, and we can expect the convergence of the parameters $\beta^{(s)}$ even less. For more on EM algorithms we refer to the book by McLachlan and Krishnan [8].

The MM (minorization-maximization) algorithm is a more general scheme for maximizing the likelihood function. It also consists of two steps.

1. Minorization step (M-step): Having the current iterate $\beta^{(s)}$, find a function $Q_s(\beta)$ minorizing the log-likelihood at $\beta^{(s)}$. By this we mean

$$Q_s(\beta) \leq \log L(X, \beta), \text{ with equality if } \beta = \beta^{(s)}.$$

2. Maximization step (M-step): Maximize $Q_s(\beta)$ in β to obtain the values $\beta^{(s+1)}$.

Again, the likelihood $L(X, \beta^{(s)})$ is increased in each step. In fact, the EM algorithm can be formulated as a special MM algorithm. For more on MM algorithms, see Hunter and Lange [7].

For most of the models considered in this paper Hunter [6] derives MM algorithms. He uses the simple inequality

$$-\ln x \geq 1 - \ln y - (x/y), \quad x, y > 0$$

to define minorizing functions $Q_s(\lambda)$. Moreover, in the minorizing functions, the components of λ are separated, thus maximization can be carried through explicitly.

3. The Plackett–Luce model

The Placket–Luce model can be viewed as a generalization of the Bradley–Terry model to multiple comparisons. In each comparison a subset $I \subset$

$\subset \{1, 2, \ldots, m\}$ of the individuals is picked, and ordered from best to worst. Thus the outcome is a permutation π of the elements in I, with $\pi(1)$ the overall winner, and $\pi(|I|)$ the overall loser. The model asserts that the probability of the ordering π equals

$$P(\pi) = \prod_{k=1}^{|I|} \frac{\lambda_{\pi(k)}}{\sum_{j=k}^{|I|} \lambda_{\pi(j)}}.$$

The next two lemmas describe two totally different ways how one can "realize" the Plackett–Luce model. One uses exponentially distributed random variables, the other uses draws from an urn. We omit the proofs, since they are straightforward.

Lemma 3.1. *The Plackett–Luce model is equivalent to the following. To each individual $i \in I$, independently of each other, attach an exponentially distributed random variable Z_i, with parameter λ_i. The order of the contestants is π if and only if $Z_{\pi(1)} < Z_{\pi(2)} < \ldots < Z_{\pi(|I|)}$.*

Lemma 3.2. *The Plackett–Luce model is equivalent to the following. Suppose that we have an urn with balls marked with numbers from* 1 *to m, and the proportion of balls with mark i is equal to λ_i (we may suppose that $\sum_{i=1}^{m} \lambda_i =$ $= 1$). We draw balls from the urn with replacement, until all numbers appear at least once. This defines an order of the individuals $1, \ldots, m$ (or of any subset I of the individuals), namely, the order in which they appeared.*

Now we can start to derive two different EM algorithms. First, let us use Lemma 3.1. Suppose that we have a sample of N comparisons, in the jth of which the set I_j of individuals is ordered, and their order is π_j. The task is to find the maximum likelihood estimator of the parameters $\lambda = (\lambda_1, \ldots, \lambda_m)$. We can treat the sample as incomplete data, whereas the complete data would be the values $Z = \{Z_{j,i} : j = 1, \ldots, N, \, i \in I_j\}$.

For the expectation step we need the complete data log-likelihood, which is

$$\log L(Z, \lambda) = \log \left(\prod_{j=1}^{N} \prod_{i \in I_j} \lambda_i e^{-\lambda_i Z_{j,i}} \right) = \sum_{j=1}^{N} \sum_{i \in I_j} (\log \lambda_i - \lambda_i Z_{j,i}).$$

Thus, denoting by N_i the number of comparisons in which individual i is a contestant,

$$E(\log L(Z, \lambda) | \pi_1, \ldots, \pi_N, \lambda^*) = \sum_{i=1}^{m} N_i \log \lambda_i - \sum_{i=1}^{m} \lambda_i \sum_{j : i \in I_j} E(Z_{j,i} | \pi_j, \lambda^*),$$

since $Z_{j,i}$ is independent of π_k, $k \neq j$. In order to complete the expectation step we need the following lemma (the proof is again omitted, since it is well-known).

Lemma 3.3. *Let Z_i be independent rv's, exponentially distributed with parameters λ_i, $i = 1, \ldots, n$. Then*

$$E(Z_i | Z_1 < Z_2 < \ldots < Z_n) = \sum_{k=1}^{i} \frac{1}{\sum_{r=k}^{n} \lambda_r}.$$

We need another notation: if $i \in I_j$, then let $\alpha_j(i)$ be such that $\pi_j(\alpha_j(i)) = i$ ($\alpha_j(i)$ is the rank of individual i in the jth comparison). Thus

$$E(\log L(Z, \lambda) | \pi_1, \ldots, \pi_N, \lambda^*) = \sum_{i=1}^{m} N_i \log \lambda_i - \sum_{i=1}^{m} \lambda_i \sum_{j: i \in I_j} \sum_{k=1}^{\alpha_j(i)} \frac{1}{\sum_{r=k}^{|I_j|} \lambda^*_{\pi_j(r)}}.$$

For the M step we need to maximize this expression in the parameter vector $(\lambda_1, \ldots, \lambda_m)$. The solution is explicit, since the parameters λ_i are separated. We obtain that

$$\lambda_i^{(s+1)} = N_i \left[\sum_{j: i \in I_j} \sum_{k=1}^{\alpha_j(i)} \frac{1}{\sum_{r=k}^{|I_j|} \lambda^{(s)}_{\pi_j(r)}} \right]^{-1}. \tag{3.1}$$

We notice that parameter update (3.1) is very similar to that obtained by Hunter via the MM algorithm. Denote by M_i the number of comparisons, in which individual i is ranked last. The difference between the two algorithms is that in Hunter's algorithm, when updating λ_i, one subtracts M_i from the numerator and $M_i / \lambda_i^{(s)}$ from the denominator of the right-hand side of expression (3.1).

Consider the special case, when only paired comparisons are made, that is, the Bradley–Terry model. Then the update equation simplifies to

$$\lambda_i^{(s+1)} = N_i \left[\sum_{k \neq i} \frac{N_{ik}}{\lambda_i^{(s)} + \lambda_k^{(s)}} + \frac{M_i}{\lambda_i^{(s)}} \right]^{-1}, \tag{3.2}$$

where N_{ik} is the number of comparisons between i and k.

We now turn to deriving an EM algorithm based on Lemma 3.2. The complete data then consists of the N drawing sequences $Z = \{Z_{j,k} : j = = 1, \ldots, N, k = 1, \ldots, R_j\}$, where R_j is the random number of draws necessary for all individuals to appear at least once.

For the expectation step we need the complete data log-likelihood, which is

$$\log L(Z, \lambda) = \log \left(\prod_{j=1}^{N} \prod_{k=1}^{R_j} \lambda_{Z_{j,k}} \right) = \sum_{j=1}^{N} \sum_{i=1}^{m} R_{j,i} \log \lambda_i,$$

where $R_{j,i}$ is the number of times i appears in the jth drawing sequence. Thus the conditional expectation is

$$E(\log L(Z,\lambda)|\pi_1,\dots,\pi_N,\lambda^*) = \sum_{i=1}^{m} \log\lambda_i \sum_{j=1}^{N} E(R_{j,i}|\pi_j,\lambda^*).$$

Once we calculate the conditional expectations of $R_{j,i}$, we can maximize the function subject to the condition that $\sum \lambda_i = 1$, which means that $\lambda_i^{(s+1)}$ will be proportional to $\sum_{j=1}^{N} E(R_{j,i}|\pi_j,\lambda^{(s)})$.

Lemma 3.4. *In the setting of Lemma* 3.2 *let* X_i *denote the number of times we draw an* i*-ball. Let* A *denote the event that the balls we are waiting for appear in the order* $1,2,\dots,k$*. Then*

$$E(X_t|A) = \lambda_t \sum_{j=1}^{k} \left(\sum_{i=j}^{k} \lambda_i\right)^{-1}, \text{ if } t > k,$$

and

$$E(X_t|A) = 1 + \lambda_t \sum_{j=t+1}^{k} \left(\sum_{i=j}^{k} \lambda_i\right)^{-1}, \text{ if } t \le k.$$

Proof. We will denote a geometrically distributed variable by $Y \sim \mathrm{Geo}(p)$, meaning $P(Y=k) = (1-p)^{k-1}p$. Given the event A, let $Y_{t,j}$ denote the number of times a t-ball is drawn strictly between the first occurrence of a $(j-1)$-ball and the first occurrence of a j-ball $(1 \le j \le k)$. Then $X_t = \sum_{j=1}^{k} Y_{t,j}$ if $t > k$, and $X_t = 1 + \sum_{j=t+1}^{k} Y_{t,j}$ if $t \le k$. It is easily shown that

$$Y_{t,j}|A \sim \mathrm{Geo}\left(\frac{\sum_{i=j}^{k}\lambda_i}{\lambda_t + \sum_{i=j}^{k}\lambda_i}\right) - 1,$$

if $t \notin \{j,\dots,k\}$, which is indeed the case in all terms occurring in X_t. The result follows. ■

Thus we arrived at the EM parameter-update

$$\lambda_i^{(s+1)} = N_i + \lambda_i^{(s)} \left[\sum_{j=1}^{N} \sum_{k=1}^{|I_j|} \frac{1}{\sum_{r=k}^{|I_j|} \lambda_{\pi_j(r)}^{(s)}} - \sum_{j:i\in I_j} \sum_{k=1}^{\alpha_j(i)} \frac{1}{\sum_{r=k}^{|I_j|} \lambda_{\pi_j(r)}^{(s)}} \right]. \tag{3.3}$$

We remark that it is not necessary to normalize the parameter vector in each step.

4. The Rao–Kupper threshold model

The generalization of the Bradley–Terry model due to Rao and Kupper allows a tie as the outcome of a comparison. In their model

$$P(i \text{ beats } j) = \lambda_i/(\lambda_i + \theta\lambda_j),$$

$$P(j \text{ beats } i) = \lambda_j/(\lambda_j + \theta\lambda_i),$$

$$P(i \text{ ties } j) = (\theta^2 - 1)\lambda_i\lambda_j/[(\lambda_i + \theta\lambda_j)(\lambda_j + \theta\lambda_i)].$$

They call $\theta > 1$ a threshold parameter. We can use similar latent variables as the ones in Lemma 3.1 for the Plackett–Luce model. The next lemma is straightforward again.

Lemma 4.1. *The Rao–Kupper threshold model is equivalent to the following. Suppose that we have to compare i with j. Attach to them independent and exponentially distributed random variables Z_i and Z_j, with parameters λ_i and λ_j, respectively. Then i wins if $Z_i < Z_j/\theta$, j wins if $Z_j < Z_i/\theta$, and otherwise a tie occurs.*

We now derive an EM algorithm assuming for simplicity that the parameter θ is known. In this case $Z = \{Z_{j,i} : j = 1, \ldots, N, \, i \in I_j\}$ is the complete data, whose log-likelihood is

$$\log L(Z, \lambda) = \log \left(\prod_{j=1}^{N} \prod_{i \in I_j} \lambda_i e^{-\lambda_i Z_{j,i}} \right) = \sum_{j=1}^{N} \sum_{i \in I_j} (\log \lambda_i - \lambda_i Z_{j,i}).$$

Thus, denoting by N_i the number of comparisons in which individual i is a contestant,

$$E(\log L(Z, \lambda)|\pi_1, \ldots, \pi_N, \lambda^*) = \sum_{i=1}^{m} N_i \log \lambda_i - \sum_{i=1}^{m} \lambda_i \sum_{j : i \in I_j} E(Z_{j,i}|\pi_j, \lambda^*),$$

since $Z_{j,i}$ is independent of π_k, $k \neq j$. We need the following lemma.

Lemma 4.2. *Let Z_i and Z_j be independent rv's, exponentially distributed with parameters λ_i and λ_j. Then*

$$E(Z_i|Z_i < Z_j/\theta) = \frac{1}{\lambda_i + \theta\lambda_j},$$

$$E(Z_i|Z_j < Z_i/\theta) = \frac{1}{\lambda_i} + \frac{\theta}{\lambda_j + \theta\lambda_i},$$

$$E(Z_i|Z_j/\theta < Z_i < Z_j\theta) = \frac{1}{\lambda_i + \theta\lambda_j} + \frac{\theta}{\theta\lambda_i + \lambda_j}.$$

Proof. The first two formulae are special cases of Lemma 3.3, since e.g. Z_j/θ is also exponentially distributed with parameter $\theta\lambda_j$. The third formula can be derived either by using the theorem of complete expectation, or by definition, using the conditional density function. ■

Returning now to the conditional expectation, we get

$$Q(\lambda,\lambda^*) = \sum_{i=1}^m N_i \log \lambda_i -$$

$$-\sum_{i=1}^m \lambda_i \sum_{k\neq i} \left((W_{ik}+T_{ik}) \frac{1}{\lambda_i^* + \theta\lambda_k^*} + (L_{ik}+T_{ik}) \frac{\theta}{\lambda_k^* + \theta\lambda_i^*} + L_{ik}\frac{1}{\lambda_i^*} \right),$$

where W_{ik}, L_{ik}, T_{ik} denote the number of times i won, lost, and tied with k, respectively. The coordinates of the parameter vector are again separated, the maximum is attained for

$$\lambda_i^{(s+1)} = N_i \left[\sum_{k\neq i} \left((W_{ik}+T_{ik}) \frac{1}{\lambda_i^{(s)} + \theta\lambda_k^{(s)}} + \right.\right.$$
$$\left.\left. +(L_{ik}+T_{ik}) \frac{\theta}{\lambda_k^{(s)} + \theta\lambda_i^{(s)}} + L_{ik}\frac{1}{\lambda_i^{(s)}} \right) \right]^{-1}. \tag{4.1}$$

Let $L_i = \sum_{k\neq i} L_{ik}$ be the total number of losses of i. Then again, we notice that our EM update is similar to the MM update of Hunter: his algorithm is obtained by subtracting L_i from our numerator and $L_i/\lambda_i^{(s)}$ from our denominator in (4.1).

5. The home-field advantage model

In the model suggested by Agresti pairs of individuals compete, but the probability that i beats j depends on whether i is at home or not. Explicitly, the winning probabilities are given by

$$P(i \text{ beats } j) = \begin{cases} \theta\lambda_i/(\theta\lambda_i + \lambda_j) & \text{if } i \text{ is at home,} \\ \lambda_i/(\lambda_i + \theta\lambda_j) & \text{if } j \text{ is at home,} \end{cases}$$

where $\theta > 0$ measures the strength of the home-field advantage or disadvantage. This model can be dealt with in the same manner as the Bradley–Terry model. In the following derivation we keep in mind Lemmas 3.1 and 3.3.

The complete data consist of the exponentially and independently distributed random variables $Z = \{H_j, F_j : j = 1, \ldots, N\}$, where H_j is the variable associated with the individual $h(j)$ at home and F_j with the individual $f(j)$ on foreign field in the jth game. Thus H_j has parameter $\theta\lambda_{h(j)}$ and F_j has parameter $\lambda_{f(j)}$. The complete data log-likelihood can be written as

$$\log L(Z, \lambda, \theta) = \log\left(\prod_{j=1}^{N} (\theta\lambda_{h(j)} e^{-\theta\lambda_{h(j)} H_j} \lambda_{f(j)} e^{-\lambda_{f(j)} F_j})\right) =$$

$$= N \log\theta + \sum_{i=1}^{m} N_i \log\lambda_i - \theta \sum_{j=1}^{N} \lambda_{h(j)} H_j - \sum_{j=1}^{N} \lambda_{f(j)} F_j.$$

Turning to the conditional expectation we get

$$E(\log L(Z, \lambda, \theta) | \pi_1, \ldots, \pi_N, \lambda^*, \theta^*) =$$

$$= N \log\theta + \sum_{i=1}^{m} N_i \log\lambda_i - \theta \sum_{j=1}^{N} \lambda_{h(j)} E(H_j | \pi_j, \lambda^*, \theta^*) - \sum_{j=1}^{N} \lambda_{f(j)} E(F_j | \pi_j, \lambda^*, \theta^*).$$

The conditional expectations of H_j and F_j are as follows:

$$E(H_j|\pi_j, \lambda^*, \theta^*) = \frac{1}{\theta^*\lambda^*_{h(j)} + \lambda^*_{f(j)}} + \frac{I(f(j) \text{ won})}{\theta^*\lambda^*_{h(j)}},$$

$$E(F_j|\pi_j, \lambda^*, \theta^*) = \frac{1}{\theta^*\lambda^*_{h(j)} + \lambda^*_{f(j)}} + \frac{I(h(j) \text{ won})}{\lambda^*_{f(j)}}.$$

Thus, the conditional expectation of the complete data log-likelihood equals

$$N \log\theta + \sum_{i=1}^{m} N_i \log\lambda_i - \sum_{i \neq k} N_{ik} \frac{\theta\lambda_i + \lambda_k}{\theta^*\lambda^*_i + \lambda^*_k} - \sum_{i=1}^{m} \left(HL_i \frac{\theta\lambda_i}{\theta^*\lambda^*_i} + FL_i \frac{\lambda_i}{\lambda^*_i}\right),$$

where N_i is the number of games played by i, N_{ik} is the number of games between i and k in which i is at home, HL_i is the number of games which i lost at home, and FL_i is the number of games which i lost on foreign ground. Since the parameters are not completely separated, it is not straightforward to maximize $Q((\lambda, \theta), (\lambda^{(s)}, \theta^{(s)}))$, but a cyclic algorithm is readily executable, by maximizing first in θ, then in λ. Writing out explicitly, the maximum of $Q((\lambda^{(s)}, \theta), (\lambda^{(s)}, \theta^{(s)}))$ is attained at

$$\theta^{(s+1)} = \frac{N}{\sum_{i \neq k} \frac{N_{ik}\lambda_i^{(s)}}{\theta^{(s)}\lambda_i^{(s)} + \lambda_k^{(s)}} + \frac{\sum_{i=1}^{m} HL_i}{\theta^{(s)}}}, \tag{5.1}$$

while the maximum of $Q((\lambda, \theta^{(s)}), (\lambda^{(s)}, \theta^{(s)}))$ is attained at

$$(5.2) \qquad \lambda_i^{(s+1)} = \frac{N_i}{\sum_{k \neq i} \left(\frac{N_{ik}\theta^{(s)}}{\theta^{(s)}\lambda_i^{(s)} + \lambda_k^{(s)}} + \frac{N_{ki}}{\theta^{(s)}\lambda_k^{(s)} + \lambda_i^{(s)}} \right) + \frac{HL_i + FL_i}{\lambda_i^{(s)}}}.$$

Again, we can notice the similarity with Hunter's MM algorithm, his algorithm follows by applying the following changes: in the formula (5.1) of $\theta^{(s+1)}$ subtract the second term from the denominator, and subtract $\sum_{i=1}^m HL_i$ from the numerator. In the formula (5.2) of $\lambda_i^{(s+1)}$ also subtract the second term from the denominator, and subtract $HL_i + FL_i$ from the numerator.

6. Paired team comparisons

Huang, Weng and Lin propose a generalization of all the models overviewed so far. Namely, in each comparison the participants are disjoint teams of individuals. If each individual has an ability parameter λ_i, as before, then a team's ability may be calculated as the sum of the abilities of its members.

Here we discuss the simplest case, the team version of the original Bradley–Terry model. Let the two teams be I and J, where $I, J \subset \{1, \ldots, m\}$, and $I \cap J = \emptyset$. Then

$$P(\text{team } I \text{ beats team } J) = \frac{\sum_{i \in I} \lambda_i}{\sum_{i \in I} \lambda_i + \sum_{j \in J} \lambda_j}.$$

We have again two equivalent "realizations" of the model, using latent variables. The straightforward proofs are once again left to the reader.

Lemma 6.1. *The team version of the Bradley–Terry model is equivalent to the following. To each individual, independently from each other, attach an exponentially distributed random variable Z_i, with parameter λ_i. Then team I beats team J if and only if $\min_{i \in I} Z_i < \min_{j \in J} Z_j$.*

Lemma 6.2. *The team version of the Bradley–Terry model is equivalent to the following. Suppose we have an urn with balls wearing numbers from 1 to m, with the proportion of balls with the number i written on them equalling λ_i (we may suppose that $\sum_{i=1}^m \lambda_i = 1$). We draw balls from the urn with replacement, until a ball with label in $I \cup J$ appears. If this ball's label is an element of I, then team I wins, otherwise team J is the winner.*

Let us start to derive an EM algorithm based on Lemma 6.1. Let there be N different team-formations $I_1, \ldots, I_N$, where I_j is separated into two disjoint teams. These two teams play N_j games against each other. For any player $\ell \in I_j$ denote by $W_j(\ell)$ ($L_j(\ell)$) the number of times ℓ's team wins (loses) in the jth team-formation. We will use the notation $q_j(\ell) = \sum_{i \in \ell\text{'s team}} \lambda_i$, $q_j = \sum_{i \in I_j} \lambda_i$, and $q_j(\ell^c) = q_j - q_j(\ell)$. Now define the complete data as

$$Z = \{Z_{j,k,\ell} : j = 1, \ldots, N, k = 1, \ldots, N_j, \ell \in I_j\},$$

where $Z_{j,k,\ell} \sim Exp(\lambda_\ell)$, and all these variables are independent.

Skipping some steps (very similar to the ones in the previous models), we arrive at the conditional expectation of the complete-data log-likelihood:

$$Q(\lambda, \lambda^*) = \sum_{\ell=1}^{m} \left(M_\ell \log \lambda_\ell - \lambda_\ell \sum_{j:\ell \in I_j} \sum_{k=1}^{N_j} E(Z_{j,k,\ell} | \pi_{j,k}, \lambda^*) \right),$$

where M_ℓ is the number of matches played by player ℓ. Again, we need a lemma to calculate the conditional expectation of $Z_{j,k,\ell}$.

Lemma 6.3. *Let $X_i \sim Exp(\lambda_i)$, $Y_j \sim Exp(\mu_j)$ be independent random variables, $i = 1, \ldots, n$, $j = 1, \ldots, m$. Let $X_1^{(n)} < \cdots < X_n^{(n)}$ denote the ordered random variables (similarly for the Y_j's). With the notation $\lambda = \sum \lambda_i$, $\mu = \sum \mu_j$,*

$$E(X_1 | \min\{X_1^{(n)}, Y_1^{(m)}\} = X_1^{(n)}) = \frac{1}{\lambda + \mu} + \frac{\lambda - \lambda_1}{\lambda} \frac{1}{\lambda_1},$$

and

$$E(X_1 | \min\{X_1^{(n)}, Y_1^{(m)}\} = Y_1^{(m)}) = \frac{1}{\lambda + \mu} + \frac{1}{\lambda_1}.$$

Proof. Starting with the first result the conditional probability density function of X_1 is

$$f(x_1 | \min\{X_1^{(n)}, Y_1^{(m)}\} = X_1^{(n)}) =$$

$$\frac{\lambda + \mu}{\lambda} \left[\lambda_1 e^{-\lambda_1 x_1} e^{-(\lambda+\mu-\lambda_1)x_1} + \sum_{i=2}^{n} \lambda_1 e^{-\lambda_1 x_1} \int_0^{x_1} \lambda_i e^{-\lambda_i y} e^{-(\lambda+\mu-\lambda_1-\lambda_i)y} dy \right] =$$

$$= \frac{\lambda + \mu}{\lambda} \left[\lambda_1 e^{-(\lambda+\mu)x_1} + \sum_{i=2}^{n} \frac{\lambda_1 \lambda_i}{\lambda + \mu - \lambda_1} e^{-\lambda_1 x_1} (1 - e^{-(\lambda+\mu-\lambda_1)x_1}) \right] =$$

$$= \frac{\lambda + \mu}{\lambda} \left[\lambda_1 e^{-(\lambda+\mu)x_1} + \frac{\lambda_1(\lambda - \lambda_1)}{\lambda + \mu - \lambda_1} (e^{-\lambda_1 x_1} - e^{-(\lambda+\mu)x_1}) \right].$$

Multiplying by x_1 and integrating between 0 and ∞, the result follows. The other formula can be derived similarly. ■

Substituting these results into $Q(\lambda,\lambda^*)$ and maximizing, we get the EM-update

(6.1)

$$\lambda_\ell^{(s+1)} = M_\ell \left\{ \sum_{j:\ell\in I_j} \frac{L_j(\ell)+W_j(\ell)}{q_j^{(s)}} + \left(L_j(\ell)+W_j(\ell) \frac{q_j^{(s)}(\ell)-\lambda_\ell^{(s)}}{q_j^{(s)}(\ell)} \right) \frac{1}{\lambda_\ell^{(s)}} \right\}^{-1}.$$

Let us turn to the EM algorithm based on the urn model of Lemma 6.2. The expectation of the complete data log-likelihood is similar to the one obtained for the Plackett–Luce model. We need to calculate the expected number of times a ball is drawn, given that one of the teams wins. This is solved by the next lemma.

Lemma 6.4. *In the setting of Lemma* 6.2*, denote by A the event that I wins, and let X_i be the random number of times an i-ball is drawn. Then*

$$E(X_i|A) = \begin{cases} \dfrac{\lambda_i}{\sum_{j\in I}\lambda_j} & \text{if} \quad i\in I \\ 0 & \text{if} \quad i \in J \\ \dfrac{\lambda_i}{\sum_{j\in I\cup J}\lambda_j} & \text{if} \quad i\notin I\cup J \end{cases}$$

Proof. The first two cases are trivial, while in the third case it is easy to show that $X_i|A \sim \mathrm{Geo}\left(\frac{\sum_{j\in I\cup J}\lambda_j}{\sum_{j\in I\cup J}\lambda_j+\lambda_i}\right)-1$. ■

A direct consequence of this lemma is the EM-update

$$\lambda_\ell^{(s+1)} = \left\{ \sum_{j:\ell\in I_j} \frac{W_j(\ell)}{q_j^{(s)}(\ell)} + \sum_{j:\ell\notin I_j} \frac{N_j}{q_j^{(s)}} \right\} \lambda_\ell^{(s)}. \tag{6.2}$$

It is interesting to note that this EM algorithm is similar to the one given by Huang, Weng and Lin. From their update formula we get ours by adding $\sum\limits_{j:\ell\notin I_j} \frac{N_j}{q_j^{(s)}}$ to both the numerator and the denominator of the multiplier of $\lambda_\ell^{(s)}$.

We remark that Huang, Weng and Lin studied generalizations of the Bradley–Terry team model. Namely, they considered the cases of multiple team comparisons, home-field advantage, and ties as well. These models could also be treated with our EM-approach, but we do not pursue these further.

7. Discussion of convergence

In this final section we briefly discuss convergence of our EM algorithms. Fortunately, all the work has been done by Hunter [6], who formulated conditions under which a unique ML estimate exists, and the MM algorithms converge to this ML estimate. The word-by-word copy of the conditions and the proof can be applied to our EM algorithms. The reason for this is that the convergence of the algorithm follows from the nice properties of the complete data likelihood function, the conditions on the iteration step are very mild and hold trivially in our case, too. The same holds for the work of Huang et al., thus the convergence results for their algorithm apply to our EM algorithms, too.

Theorem 7.1. *Under the assumptions of Hunter* [6]*, the EM algorithms defined by the iteration steps* (3.1), (3.3), (4.1), *and* (5.1)–(5.2) *converge to the unique maximum likelihood estimate of the parameter vector. Moreover, if the conditions of Huang et al.* [5]*, Theorem* 4*, are satisfied, then the EM algorithms defined by the iteration steps* (6.1) *and* (6.2) *converge to the unique maximum likelihood estimate of the parameter vector.*

For convenience, we sketch the proof for the original Bradley–Terry model, the simplest case. Let us denote the parameter space by $\Lambda = \{\lambda \in \mathbb{R}^m : : \sum_{i=1}^m \lambda_i = 1,\ \lambda_i > 0\ \forall i\}$. The first step is to show that the log-likelihood function $\ell(\lambda) = \log L(X, \lambda)$ is upper compact on Λ if and only if in every possible partition of the individuals into two nonempty subsets one individual in the second set beats some individual in the first set at least once. Since in all EM algorithms $\ell(\lambda^{(s+1)}) \ge \ell(\lambda^{(s)})$, this assumption implies the existence of at least one limit point of the sequence of iterates. Next, by the following theorem any limit point is a stationary point of $\ell(\lambda)$, that is the gradient is 0 at λ.

Theorem 7.2. (Lyapunov's theorem.) *Suppose $M : \Lambda \to \Lambda$ is continuous and $\ell : \Lambda \to \mathbb{R}$ is differentiable and for all $\lambda \in \Lambda$ we have $\ell(M(\lambda)) \ge \ell(\lambda)$, with equality only if λ is a stationary point of $\ell(\cdot)$. Then, for arbitrary $\lambda^{(1)} \in \Lambda$, any limit point of the sequence $\{\lambda^{(s+1)} = M(\lambda^{(s)})\}_{s\ge 1}$ is a stationary point of $\ell(\lambda)$.*

We can apply Lyapunov's theorem with the map M defined implicitly by the iteration step of the algorithm, which is clearly continuous. Moreover, we noted before that the EM algorithm is a special case of MM algorithms: the function $Q(\lambda, \lambda^{(s)})$ calculated in the E-step is in fact a minorizing function of $\ell(\lambda)$ at $\lambda^{(s)}$, up to an additive constant. Notice that if the minorizing function is differentiable (as is the case now), it must be tangent to the log-likelihood at the current iterate. Therefore $\ell(\lambda^{(s+1)}) = \ell(\lambda^{(s)})$ implies that $\lambda^{(s)}$ is a stationary point of $\ell(\lambda)$. In order to finish the proof, one reparametrizes the log-likelihood so that it becomes strictly concave (note that MM algorithms do not change under reparametrization of the log-likelihood). For the Bradley–Terry model the reparametrization $\beta_i = \log \lambda_i - \log \lambda_1$ is suitable. Finally, strict concavity implies the existence of at most one stationary point, namely the maximizer, hence the sequence of iterates must converge to this unique maximum likelihood estimator.

The similarity of the MM algorithms in the literature and our EM algorithms seems to be a "coincidence", although one wonders if there is any deeper reason for it. Another issue, which we do not pursue here further is the speed of convergence. Preliminary simulation studies suggest that the EM algorithms are generally slower than the MM ones. It is well known that EM algorithms are in general quite slow, but there are several methods to speed them up. We leave it for future work to explore the applicability of these methods to our EM algorithms.

References

[1] **Agresti, A.,** *Categorical Data Analysis*, Wiley, New York, 1990.

[2] **Bradley, R.A. and M. Terry,** The rank analysis of incomplete block designs: I. the method of paired comparisons, *Biometrika*, **39** (1952), 324–345.

[3] **Dempster, A.P., N. Laird, and D.B. Rubin,** Maximum likelihood from incomplete data via the EM algorithm (with discussion), *J. Roy. Statist. Soc. Ser. B*, **39** (1977), 1–38.

[4] **Guiver, J. and E. Snelson,** Bayesian inference for Plackett-Luce ranking models, *Proceedings of the 26th International Conference on Machine Learning* (2009).

[5] **Huang, T.-K., R.C. Weng, and C.-J. Lin,** Generalized Bradley–Terry models and multi-class probability estimates, *J. Mach. Learn. Res.*, **4** (2006), 85–115.

[6] **Hunter, D.R.,** MM algorithms for generalized Bradley–Terry models, *Ann. Statist.*, **32** (2004), 384–406.

[7] **Hunter, D.R. and K. Lange,** A tutorial on MM algorithms, *The American Statistician*, **58** (2004), 30–37.

[8] **McLachlan, G.J. and T. Krishnan,** *The EM Algorithm and Extensions*, John Wiley&Sons, New York 1997.

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

[10] **Rao, P. V. and L. L. Kupper,** Ties in paired-comparison experiments: A generalization of the Bradley–Terry model, *J. Amer. Statist. Assoc.*, **62** (1967), 194–204.

V. Csiszár
Department of Probability Theory and Statistics
Faculty of Science
Eötvös Loránd University
H-1117 Budapest, Pázmány P. sétány 1/C
Hungary
villo@cs.elte.hu