The Annals of Statistics
2004, Vol. 32, No. 1, 384–406

MM ALGORITHMS FOR GENERALIZED BRADLEY–TERRY MODELS

BY DAVID R. HUNTER

Pennsylvania State University

The Bradley–Terry model for paired comparisons is a simple and much-studied means to describe the probabilities of the possible outcomes when individuals are judged against one another in pairs. Among the many studies of the model in the past 75 years, numerous authors have generalized it in several directions, sometimes providing iterative algorithms for obtaining maximum likelihood estimates for the generalizations. Building on a theory of algorithms known by the initials MM, for minorization–maximization, this paper presents a powerful technique for producing iterative maximum likelihood estimation algorithms for a wide class of generalizations of the Bradley–Terry model. While algorithms for problems of this type have tended to be custom-built in the literature, the techniques in this paper enable their mass production. Simple conditions are stated that guarantee that each algorithm described will produce a sequence that converges to the unique maximum likelihood estimator. Several of the algorithms and convergence results herein are new.

1. Introduction. In a situation in which the individuals in a group are repeatedly compared with one another in pairs, Bradley and Terry (1952) suggested the model

$$P(\text{individual } i \text{ beats individual } j) = \frac{\gamma_i}{\gamma_i + \gamma_j}, \tag{1}$$

where γ_i is a positive-valued parameter associated with individual i, for each of the comparisons pitting individual i against individual j. As a concrete example, consider the individuals to be sports teams, where γ_i represents the overall skill of team i.

The Bradley–Terry model of (1) dates back to at least 1929 [Zermelo (1929)] and has applications in a broad range of problems. For instance, in any problem in which observed data may be represented in a directed graph with nonnegative-weighted edges and one wishes to attach an "influence" parameter to each node, a Bradley–Terry model may be useful. Note that the weights on the edges must be integers in this context; however, the algorithm of Section 2 could easily be applied to graphs with nonintegral weights. The sports scenario mentioned above gives one manifestation of such a graph—teams are nodes and the weight of edge (i, j) is the

Received January 2002; revised February 2003.
AMS 2000 subject classifications. 62F07, 65D15.
Key words and phrases. Bradley–Terry model, Luce's choice axiom, maximum likelihood estimation, MM algorithm, Newton–Raphson, Plackett–Luce model.

number of times i beats j—but there are many others. Examples in the literature range from the quantification of the influence of statistical journals [Stigler (1994)] to the transmission/disequilibrium test in genetics [Sham and Curtis (1995)]. There is also a vast literature on the properties and generalizations of the Bradley–Terry model; even as early as 1976, a published bibliography on the method of paired comparisons [Davidson and Farquhar (1976)] lists several hundred entries. Although it does not focus exclusively on the Bradley–Terry model, the book by David (1988) provides an in-depth examination of paired-comparison models. A more recent, albeit brief, history of the Bradley–Terry model is given by Simons and Yao (1999).

A simple iterative algorithm for finding maximum likelihood estimates in the Bradley–Terry model has been known for a long time [Zermelo (1929)]. Lange, Hunter and Yang (2000) demonstrated that this algorithm is a specific case of a general class of algorithms referred to here as MM algorithms. MM algorithms have been studied under various names for over 30 years, though the initials MM originate with a rejoinder of Hunter and Lange (2000). Surveys of some of this past work may be found in Heiser (1995) and Lange, Hunter and Yang (2000). Heiser (1995) uses the initials IM, for iterative majorization, to describe this class of algorithms; however, the initials MM better emphasize the close tie between MM algorithms and the best known special cases, EM (expectation–maximization) algorithms. For an explanation of why an EM algorithm is a special case of an MM algorithm—and, more precisely, why the E-step of EM is actually a minorization step—see, for example, Heiser (1995). The current paper demonstrates the potential power of the MM approach by showing how to extend the argument of Lange, Hunter and Yang (2000) to generalizations of the Bradley–Terry model. When numerical algorithms are published for these generalizations [Ford (1957), Rao and Kupper (1967) and Davidson (1970)], they tend to be ad hoc, designed specifically for the model at hand. Furthermore, convergence results are not always given. Here, we provide a sort of template for creating MM algorithms for these generalized Bradley–Terry models, showing how known results about MM algorithms can be applied to give sufficient conditions under which these algorithms can be guaranteed to converge to the maximum likelihood estimates.

We begin the development in Section 2 by describing the iterative algorithm for maximum likelihood estimation in the Bradley–Terry model and introducing several known generalizations of this model. Section 3 introduces MM algorithms and shows how to derive them for the models given in Section 2. Section 4 demonstrates that, under simple conditions, all of these algorithms are guaranteed to converge to the correct values regardless of the starting point. In Section 5, we discuss in greater depth a particular generalization of the Bradley–Terry model that allows for comparisons (rankings) involving more than two individuals. Finally, Section 6 gives a numerical example and provides some discussion about the estimation of standard errors.

2. Fitting and generalizing the model. Suppose we observe a number of pairings among m individuals or teams and we wish to estimate the parameters $\gamma_1, \ldots, \gamma_m$ using maximum likelihood estimation. If outcomes of different pairings are assumed to be independent, the log-likelihood based on the Bradley–Terry model (1) is

$$\ell(\boldsymbol{\gamma}) = \sum_{i=1}^{m} \sum_{j=1}^{m} \left[w_{ij} \ln \gamma_i - w_{ij} \ln(\gamma_i + \gamma_j) \right], \tag{2}$$

where w_{ij} denotes the number of times individual i has beaten individual j and we assume $w_{ii} = 0$ by convention. Since $\ell(\boldsymbol{\gamma}) = \ell(a\boldsymbol{\gamma})$ for $a > 0$, the parameter space should be regarded as the set of equivalence classes of $\mathbb{R}_+^m$, where two vectors are equivalent if one is a scalar multiple of the other. This is most easily accomplished by putting a constraint on the parameter space; to this end, we assume that $\sum_i \gamma_i = 1$.

As noted by Ford (1957), if it is possible to partition the set of individuals into two groups A and B such that there are never any intergroup comparisons, then there is no basis for rating any individual in A with respect to any individual in B. On the other hand, if all the intergroup comparisons are won by an individual from the same group, say group A, then if all parameters belonging to A are doubled and the resulting vector renormalized, the likelihood must increase; thus, the likelihood has no maximizer. The following assumption [Ford (1957)] eliminates these possibilities.

ASSUMPTION 1. In every possible partition of the individuals into two nonempty subsets, some individual in the second set beats some individual in the first set at least once.

Assumption 1 has a graph-theoretic interpretation: if the individuals are the nodes of a graph and the directed edge (i, j) denotes a win by i over j, then Assumption 1 is equivalent to the statement that there is a path from i to j for all nodes i and j. We will see later that Assumption 1 implies, among other things, that there exists a unique maximizer of the log-likelihood function (2).

We now describe an iterative algorithm to maximize $\ell(\boldsymbol{\gamma})$. Start with an initial parameter vector $\boldsymbol{\gamma}^{(1)}$. Dykstra (1956) considers several ways to select starting points; however, even though intelligent choice of a starting point can reduce the overall computational workload, in this context we assume that $\boldsymbol{\gamma}^{(1)}$ is chosen arbitrarily. For $k = 1, 2, \ldots$, let

$$\gamma_i^{(k+1)} = W_i \left[\sum_{j \neq i} \frac{N_{ij}}{\gamma_i^{(k)} + \gamma_j^{(k)}} \right]^{-1}, \tag{3}$$

where W_i denotes the number of wins by individual i and $N_{ij} = w_{ij} + w_{ji}$ is the number of pairings between i and j. If the resulting $\boldsymbol{\gamma}^{(k+1)}$ vector does

not satisfy the constraint $\sum_i \gamma_i^{(k+1)} = 1$, it should simply be renormalized. This renormalization step is to be understood as part of each algorithm described in this paper, though it is not mentioned henceforth because it does not essentially change the parameter vector.

Since (3) updates the parameter components one at a time, we may use the updates as soon as they are available; thus, (3) may be replaced by a cyclic version

$$\gamma_i^{(k+1)} = W_i \left[\sum_{j<i} \frac{N_{ij}}{\gamma_i^{(k)} + \gamma_j^{(k+1)}} + \sum_{j>i} \frac{N_{ij}}{\gamma_i^{(k)} + \gamma_j^{(k)}} \right]^{-1}. \tag{4}$$

Readers familiar with the literature on EM algorithms may notice the analogy with cyclic EM algorithms, also known as ECM algorithms [Meng and Rubin (1993)].

Under Assumption 1, both algorithm (3) and its cyclic version (4) produce a sequence $\boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)}, \ldots$ guaranteed to converge to the unique maximum likelihood estimator. In addition, the sequence $\ell(\boldsymbol{\gamma}^{(1)}), \ell(\boldsymbol{\gamma}^{(2)}), \ldots$ is monotone increasing. Rather than prove these facts directly as in Zermelo (1929), we adopt the approach of Lange, Hunter and Yang (2000), where algorithm (3) is shown to be a particular example from a class of algorithms we refer to here as MM algorithms. The monotonicity of the sequence $\{\ell(\boldsymbol{\gamma}^{(k)})\}$ is a characteristic property of all MM algorithms, and the guaranteed convergence follows from a theorem stated in Section 4. The cyclic algorithm (4) is technically also an MM algorithm, and, as such, it inherits these favorable convergence properties.

There are numerous generalizations of the basic Bradley–Terry model (1) in the literature. For instance, Agresti (1990) supposes that the individuals involved in any paired comparison are ordered and postulates that the probability of i beating j depends on which individual is listed first. If the individuals are sports teams, this assumption leads to the "home-field advantage" model

$$P(i \text{ beats } j) = \begin{cases} \theta\gamma_i/(\theta\gamma_i + \gamma_j), & \text{if } i \text{ is home,} \\ \gamma_i/(\gamma_i + \theta\gamma_j), & \text{if } j \text{ is home,} \end{cases} \tag{5}$$

where $\theta > 0$ measures the strength of the home-field advantage or disadvantage.

Extending the model in a different direction, suppose that ties are possible between teams. Rao and Kupper (1967) suppose that

$$\begin{aligned} P(i \text{ beats } j) &= \gamma_i/(\gamma_i + \theta\gamma_j), \\ P(j \text{ beats } i) &= \gamma_j/(\theta\gamma_i + \gamma_j), \\ P(i \text{ ties } j) &= (\theta^2 - 1)\gamma_i\gamma_j/\big[(\gamma_i + \theta\gamma_j)(\gamma_j + \theta\gamma_i)\big], \end{aligned} \tag{6}$$

calling $\theta > 1$ a "threshold" parameter. They justify this name by showing that model (6) can arise if each comparison is decided by a judge who estimates $\ln\gamma_i - \ln\gamma_j$ with error and declares a tie if this value is smaller than $\ln\theta$ in absolute value. Davidson (1970) gives a different adjustment to the Bradley–Terry model to account for ties, in which the probabilities are in the ratio

$$P(i \text{ beats } j) : P(j \text{ beats } i) : P(i \text{ ties } j) = \gamma_i : \gamma_j : \theta\sqrt{\gamma_i\gamma_j}. \tag{7}$$

The positive-valued parameter θ in model (7) is the constant of proportionality if the probability of a tie is proportional to the geometric mean of the probabilities of a win by either individual. Davidson (1970) calls $1/\theta$ an index of discrimination and points out that the use of the geometric mean is suggested by the fact that the merits of the individuals may be represented on a linear scale as $\ln \gamma_1, \ldots, \ln \gamma_m$.

The Bradley–Terry model has even been extended to allow for comparisons among more than two individuals at once. For instance, if individuals are compared in groups of three, where each comparison results in a ranking from best to worst, then Pendergrass and Bradley (1960) proposed the model

$$P(i \text{ best, } j \text{ in the middle and } k \text{ worst}) = \frac{\gamma_i \gamma_j}{(\gamma_i + \gamma_j + \gamma_k)(\gamma_j + \gamma_k)}. \tag{8}$$

We discuss model (8) and its generalization to comparisons of any number of individuals, termed the Plackett–Luce model by Marden (1995), in Section 5.

In the next section, we demonstrate how to obtain MM algorithms for fitting all of the models above using a method that is easily applicable to Bradley–Terry generalizations not discussed here.

3. Minorizing functions and MM algorithms. The strict concavity of the logarithm function implies for positive x and y that

$$-\ln x \geq 1 - \ln y - (x/y) \qquad \text{with equality if and only if } x = y. \tag{9}$$

Therefore, as shown in Lange, Hunter and Yang (2000), if we fix $\boldsymbol{\gamma}^{(k)}$ and define the function

$$Q_k(\boldsymbol{\gamma}) = \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij} \left[\ln \gamma_i - \frac{\gamma_i + \gamma_j}{\gamma_i^{(k)} + \gamma_j^{(k)}} - \ln\bigl(\gamma_i^{(k)} + \gamma_j^{(k)}\bigr) + 1 \right], \tag{10}$$

we may conclude that

$$Q_k(\boldsymbol{\gamma}) \leq \ell(\boldsymbol{\gamma}) \qquad \text{with equality if } \boldsymbol{\gamma} = \boldsymbol{\gamma}^{(k)}, \tag{11}$$

where $\ell(\boldsymbol{\gamma})$ is the log-likelihood of (2). A function $Q_k(\boldsymbol{\gamma})$ satisfying conditions (11) is said to *minorize* $\ell(\boldsymbol{\gamma})$ at the point $\boldsymbol{\gamma}^{(k)}$. It is easy to verify that, for any $Q_k(\boldsymbol{\gamma})$ satisfying the minorizing conditions (11),

$$Q_k(\boldsymbol{\gamma}) \geq Q_k\bigl(\boldsymbol{\gamma}^{(k)}\bigr) \quad \text{implies} \quad \ell(\boldsymbol{\gamma}) \geq \ell\bigl(\boldsymbol{\gamma}^{(k)}\bigr). \tag{12}$$

Property (12) suggests an iterative algorithm in which we let $\boldsymbol{\gamma}^{(k)}$ denote the value of the parameter vector before the kth iteration and define $\boldsymbol{\gamma}^{(k+1)}$ to be the maximizer of $Q_k(\boldsymbol{\gamma})$; thus, $\boldsymbol{\gamma}^{(k+1)}$ of (3) maximizes $Q_k(\boldsymbol{\gamma})$. Since this algorithm consists of alternately creating a minorizing function $Q_k(\boldsymbol{\gamma})$ and then maximizing it, Hunter and Lange (2000) call it an MM algorithm.

One feature of the function $Q_k(\boldsymbol{\gamma})$ defined in (10) that makes it easier to maximize than the original log-likelihood is the fact that it separates the

components of the parameter vector $\boldsymbol{\gamma}$. Thus, maximization of $Q_k(\boldsymbol{\gamma})$ is equivalent to maximization for each component $\boldsymbol{\gamma}_i$ separately. This separation of parameters is typical of many well-constructed minorizing functions in high-dimensional problems [Lange, Hunter and Yang (2000)].

The cyclic algorithm of (4) is itself an MM algorithm, since $\gamma_i^{(k+1)}$ is the maximizer of $Q_k(\gamma_1^{(k+1)}, \ldots, \gamma_{i-1}^{(k+1)}, \gamma_i, \gamma_{i+1}^{(k)}, \ldots, \gamma_m^{(k)})$, which minorizes $\ell(\boldsymbol{\gamma})$ at the point $\boldsymbol{\gamma} = (\gamma_1^{(k+1)}, \ldots, \gamma_{i-1}^{(k+1)}, \gamma_i^{(k)}, \ldots, \gamma_m^{(k)})$. Note that, in the cyclic case, there is some ambiguity about what should be considered one iteration of the algorithm; we discuss this further in Section 4.

In the home-field advantage model (5), we may use inequality (9) to construct a minorizing function for the log-likelihood function

$$\ell(\boldsymbol{\gamma}, \theta) = \sum_{i=1}^{m} \sum_{j=1}^{m} \left[a_{ij} \ln \frac{\theta \gamma_i}{\theta \gamma_i + \gamma_j} + b_{ij} \ln \frac{\gamma_j}{\theta \gamma_i + \gamma_j} \right], \tag{13}$$

where a_{ij} is 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. Letting $H = \sum_i \sum_j a_{ij}$ be the total number of home-field wins and W_i be the total number of wins by team i, we obtain

$$Q_k(\boldsymbol{\gamma}, \theta) = H \ln \theta + \sum_{i=1}^{m} W_i \ln \gamma_i - \sum_{i=1}^{m} \sum_{j=1}^{m} \left[\frac{(a_{ij} + b_{ij})(\theta \gamma_i + \gamma_j)}{\theta^{(k)} \gamma_i^{(k)} + \gamma_j^{(k)}} \right],$$

which minorizes $\ell(\boldsymbol{\gamma}, \theta)$ up to a constant; that is,

$$Q_k(\boldsymbol{\gamma}, \theta) + \left[\ell(\boldsymbol{\gamma}^{(k)}, \theta^{(k)}) - Q_k(\boldsymbol{\gamma}^{(k)}, \theta^{(k)}) \right] \leq \ell(\boldsymbol{\gamma}, \theta).$$

The presence of the product $\theta \gamma_i$ means that the parameters are not quite separated by the minorizing function, which makes direct maximization of the function slightly problematic. However, it is easy to maximize $Q_k(\boldsymbol{\gamma}, \theta^{(k)})$ as a function of $\boldsymbol{\gamma}$ and $Q_k(\boldsymbol{\gamma}^{(k+1)}, \theta)$ as a function of θ—therefore, we may construct a cyclic MM algorithm for this case.

For the Rao–Kupper (1967) model of (6) that allows for ties, the likelihood is

$$\ell(\boldsymbol{\gamma}, \theta) = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \left\{ 2 w_{ij} \ln \left(\frac{\gamma_i}{\gamma_i + \theta \gamma_j} \right) + t_{ij} \ln \left(\frac{(\theta^2 - 1) \gamma_i \gamma_j}{(\theta \gamma_i + \gamma_j)(\gamma_i + \theta \gamma_j)} \right) \right\}, \tag{14}$$

where $t_{ij} = t_{ji}$ is the number of times that i and j have tied. Using inequality (9) as usual, we may construct the function

$$Q_k(\boldsymbol{\gamma}, \theta) = \sum_{i=1}^{m} \sum_{j=1}^{m} \left\{ (w_{ij} + t_{ij}) \left(\ln \gamma_i - \frac{\gamma_i + \theta \gamma_j}{\gamma_i^{(k)} + \theta^{(k)} \gamma_j^{(k)}} \right) + t_{ij} \ln(\theta^2 - 1) \right\},$$

which, up to a constant, minorizes $\ell(\boldsymbol{\gamma}, \theta)$ at $(\boldsymbol{\gamma}^{(k)}, \theta^{(k)})$. The parameters are not completely separated, but we may alternately maximize $Q_k(\boldsymbol{\gamma}, \theta^{(k)})$ as a function

of $\boldsymbol{\gamma}$ and $Q_k(\boldsymbol{\gamma}^{(k+1)}, \theta)$ as a function of θ to obtain a cyclic MM algorithm. Maximization of $Q_k(\boldsymbol{\gamma}, \theta^{(k)})$ with respect to $\boldsymbol{\gamma}$ gives

$$\gamma_i^{(k+1)} = \left[\sum_{j\neq i} s_{ij}\right]\left[\sum_{j\neq i}\left(\frac{s_{ij}}{\gamma_i^{(k)} + \theta^{(k)}\gamma_j^{(k)}} + \frac{\theta^{(k)} s_{ji}}{\theta^{(k)}\gamma_i^{(k)} + \gamma_j^{(k)}}\right)\right]^{-1}, \tag{15}$$

where $s_{ij} = w_{ij} + t_{ij}$ is the number of times individual i beat or tied individual j. Solving a quadratic equation to maximize $Q_k(\boldsymbol{\gamma}^{(k+1)}, \theta)$ with respect to θ gives

$$\theta^{(k+1)} = \frac{1}{2C_k} + \sqrt{1 + \frac{1}{4C_k^2}},$$

where

$$C_k = \frac{2}{T}\sum_{i=1}^{m}\sum_{j\neq i}\frac{\gamma_j^{(k+1)}(s_{ij})}{\gamma_i^{(k+1)} + \theta^{(k)}\gamma_j^{(k+1)}}$$

and T is the total number of ties observed among all of the comparisons. Equation (15) was suggested by Rao and Kupper (1967), though they did not explore the convergence properties of any algorithm derived from it. Equation (15) may also be modified to produce a cyclic update of $\boldsymbol{\gamma}$ in the same way that (4) is a modified version of (3).

In model (7), which also allows for ties, the log-likelihood

$$\ell(\boldsymbol{\gamma}, \theta) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\left[2w_{ij}\ln\frac{\gamma_i}{\gamma_i + \gamma_j + \theta\sqrt{\gamma_i\gamma_j}} + t_{ij}\ln\frac{\theta\sqrt{\gamma_i\gamma_j}}{\gamma_i + \gamma_j + \theta\sqrt{\gamma_i\gamma_j}}\right] \tag{16}$$

is minorized up to an irrelevant constant via inequality (9) by

$$Q_k^*(\boldsymbol{\gamma}, \theta) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\left[2w_{ij}\ln\gamma_i + t_{ij}\ln\bigl(\theta\sqrt{\gamma_i\gamma_j}\bigr) - \frac{(2w_{ij} + t_{ij})(\gamma_i + \gamma_j + \theta\sqrt{\gamma_i\gamma_j})}{\gamma_i^{(k)} + \gamma_j^{(k)} + \theta^{(k)}\sqrt{\gamma_i^{(k)}\gamma_j^{(k)}}}\right].$$

However, because of the second $\sqrt{\gamma_i\gamma_j}$ term above, direct maximization of $Q_k^*(\boldsymbol{\gamma}, \theta)$ is not convenient, even if θ is held fixed at $\theta^{(k)}$. Therefore, we employ a further well-known inequality, the arithmetic–geometric mean inequality, to create a minorizer of $Q_k^*(\boldsymbol{\gamma}, \theta)$.

In its most general form [Magnus and Neudecker (1988)], the arithmetic–geometric mean inequality states that $\prod_i x_i^{w_i} \leq \sum_i w_i x_i$ for $x_i \geq 0$, $w_i > 0$ and $\sum_i w_i = 1$, with equality if and only if all x_i are equal. With $w_1 = w_2 = 1/2$, we obtain

$$-\sqrt{\gamma_i\gamma_j} \geq -\frac{\gamma_i}{2}\sqrt{\frac{\gamma_j^{(k)}}{\gamma_i^{(k)}}} - \frac{\gamma_j}{2}\sqrt{\frac{\gamma_i^{(k)}}{\gamma_j^{(k)}}}, \tag{17}$$

with equality when $\boldsymbol{\gamma} = \boldsymbol{\gamma}^{(k)}$. Therefore, $Q_k^*(\boldsymbol{\gamma}, \theta)$ is minorized at $(\boldsymbol{\gamma}^{(k)}, \theta^{(k)})$ by

$$
\begin{aligned}
&Q_k(\boldsymbol{\gamma}, \theta) \\
&\quad = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \Bigg[2w_{ij} \ln \gamma_i + t_{ij} \ln\bigl(\theta \sqrt{\gamma_i \gamma_j}\bigr) - \frac{(2w_{ij} + t_{ij})(\gamma_i + \gamma_j)}{\gamma_i^{(k)} + \gamma_j^{(k)} + \theta^{(k)} \sqrt{\gamma_i^{(k)} \gamma_j^{(k)}}} \\
&\qquad - \frac{\theta(2w_{ij} + t_{ij})}{\gamma_i^{(k)} + \gamma_j^{(k)} + \theta^{(k)} \sqrt{\gamma_i^{(k)} \gamma_j^{(k)}}} \left(\frac{\gamma_i}{2} \sqrt{\frac{\gamma_j^{(k)}}{\gamma_i^{(k)}}} + \frac{\gamma_j}{2} \sqrt{\frac{\gamma_i^{(k)}}{\gamma_j^{(k)}}} \right) \Bigg].
\end{aligned}
$$

Minorization is a transitive relation: since $Q_k(\boldsymbol{\gamma}, \theta)$ minorizes $Q_k^*(\boldsymbol{\gamma}, \theta)$ at $(\boldsymbol{\gamma}^{(k)}, \theta^{(k)})$ and $Q_k^*(\boldsymbol{\gamma}, \theta)$ minorizes $\ell(\boldsymbol{\gamma}, \theta)$ at $(\boldsymbol{\gamma}^{(k)}, \theta^{(k)})$, we conclude that $Q_k(\boldsymbol{\gamma}, \theta)$ minorizes $\ell(\boldsymbol{\gamma}, \theta)$ at $(\boldsymbol{\gamma}^{(k)}, \theta^{(k)})$. The components of $\boldsymbol{\gamma}$ are now separated, and maximization of $Q_k(\boldsymbol{\gamma}, \theta^{(k)})$ with respect to $\boldsymbol{\gamma}$ is accomplished by

$$
\gamma_i^{(k+1)} = \frac{2W_i + T_i}{\sum_{j=1}^{m} g_{ij}(\boldsymbol{\gamma}^{(k)}, \theta^{(k)})}, \tag{18}
$$

where W_i is the total number of wins for individual i, T_i is the total number of ties for individual i and

$$
g_{ij}(\boldsymbol{\gamma}, \theta) = \frac{(w_{ij} + w_{ji} + t_{ij})(2 + \theta \sqrt{\gamma_j / \gamma_i})}{\gamma_i + \gamma_j + \theta \sqrt{\gamma_i \gamma_j}}.
$$

Naturally, the components of $\boldsymbol{\gamma}$ may be updated cyclically if the denominator of (18) is replaced by $\sum_{j<i} g_{ij}(\boldsymbol{\gamma}^{(k+1)}, \theta^{(k)}) + \sum_{j>i} g_{ij}(\boldsymbol{\gamma}^{(k)}, \theta^{(k)})$. Finally, we maximize $Q_k(\boldsymbol{\gamma}^{(k+1)}, \theta)$ as a function of θ by

$$
\theta^{(k+1)} = 4T \left[\sum_{i=1}^{m} \sum_{j=1}^{m} \frac{(2w_{ij} + t_{ij})(\gamma_i^{(k+1)} + \gamma_j^{(k+1)})}{\gamma_i^{(k+1)} + \gamma_j^{(k+1)} + \theta^{(k)} \sqrt{\gamma_i^{(k+1)} \gamma_j^{(k+1)}}} \right],
$$

where T is the total number of ties. Davidson (1970) used an argument nearly identical to that of Ford (1957) to prove that, under Assumption 1, the cyclic version of (18), along with a slightly different update to θ, is guaranteed to converge to the unique maximum likelihood estimator.

We have seen how to derive MM algorithms for some generalizations of the Bradley–Terry model using inequality (9) and in one case inequality (17). We may apply the same technique in the case of the triple-comparison model (8), but we postpone discussion of this case until Section 5. These MM algorithms, like all MM algorithms, are guaranteed to increase the value of the log-likelihood at each iteration. This monotonicity property alone is not enough to guarantee that the algorithms will eventually lead to the maximum likelihood estimators; the next section takes up the question of when such convergence must occur.

4. Convergence properties of MM algorithms. There is some ambiguity in deciding what it means for an algorithm to converge. Here, we say that an MM algorithm converges if $\boldsymbol{\gamma}^* = \lim_k \boldsymbol{\gamma}^{(k)}$ exists. This is a more stringent definition of convergence than is sometimes seen in the literature; for example, Hastie and Tibshirani (1998) merely note that $\lim_k \ell(\boldsymbol{\gamma}^{(k)})$ exists and is finite in declaring that algorithm (4) converges. We use the stronger definition here for two reasons. First, the final value of $\boldsymbol{\gamma}$ is usually more interesting than the final value of $\ell(\boldsymbol{\gamma})$; second, the existence of $\lim_k \ell(\boldsymbol{\gamma}^{(k)})$ is automatic for any MM algorithm, and this limit is finite as long as $\ell(\boldsymbol{\gamma})$ is bounded above. When $\boldsymbol{\gamma}^*$ does exist, we will also be interested in whether it maximizes $\ell(\boldsymbol{\gamma})$.

In general, it is not always possible to prove that the sequence of parameters defined by an MM algorithm converges at all, let alone to a global maximizer; McLachlan and Krishnan (1997) give examples of EM algorithms that converge to saddle points or fail to converge. Nonetheless, it is often possible to obtain convergence results in specific cases. For example, Ford (1957) showed that under Assumption 1 the algorithm of (4) converges to the unique maximum likelihood estimate, and Zermelo (1929) derived a similar result. This result may be obtained as a corollary of a more general theorem [Lange (1995)].

THEOREM 1 (Liapounov's theorem). *Suppose $M : \Omega \to \Omega$ is continuous and $\ell : \Omega \to \mathbb{R}$ is differentiable and for all $\boldsymbol{\gamma} \in \Omega$ we have $\ell[M(\boldsymbol{\gamma})] \geq \ell(\boldsymbol{\gamma})$, with equality only if $\boldsymbol{\gamma}$ is a stationary point of $\ell(\cdot)$ (i.e., the gradient is $\mathbf{0}$ at $\boldsymbol{\gamma}$). Then, for arbitrary $\boldsymbol{\gamma}^{(1)} \in \Omega$, any limit point of the sequence $\{\boldsymbol{\gamma}^{(k+1)} = M(\boldsymbol{\gamma}^{(k)})\}_{k \geq 1}$ is a stationary point of $\ell(\boldsymbol{\gamma})$.*

The proof is immediate. If $\gamma^* = \lim_n \boldsymbol{\gamma}^{(k_n)}$ for a subsequence $\boldsymbol{\gamma}^{(k_1)}, \boldsymbol{\gamma}^{(k_2)}, \ldots$, then the result is obtained by taking limits in

$$\ell\big(\boldsymbol{\gamma}^{(k_n)}\big) \leq \ell\big[M\big(\boldsymbol{\gamma}^{(k_n)}\big)\big] \leq \ell\big(\boldsymbol{\gamma}^{(k_{n+1})}\big).$$

For an MM algorithm, the map $M(\boldsymbol{\gamma})$ in the theorem is taken to be the map implicitly defined by one iteration of the algorithm, which guarantees $\ell[M(\boldsymbol{\gamma})] \geq \ell(\boldsymbol{\gamma})$. For each MM algorithm in this paper, the continuity of $M(\boldsymbol{\gamma})$ is clear; the fact that $\ell[M(\boldsymbol{\gamma})] = \ell(\boldsymbol{\gamma})$ implies that $\boldsymbol{\gamma}$ is a stationary point follows because the differentiable minorizing function is tangent to the log-likelihood at the current iterate.

In the case of a cyclic MM algorithm, $M(\boldsymbol{\gamma}^{(k)}) = \boldsymbol{\gamma}^{(k+1)}$ is actually the result of several MM iterations in succession, one for each subset of parameter components. Nevertheless, the continuity of M is clear, and the only way that $\ell[M(\boldsymbol{\gamma})] = \ell(\boldsymbol{\gamma})$ can occur is if each of the several MM iterations leaves $\boldsymbol{\gamma}$ unchanged, which means that $\boldsymbol{\gamma}$ is a stationary point of ℓ. Thus, Theorem 1 applies to cyclic MM algorithms as well as MM algorithms, and, in particular, cyclic MM algorithms share all of the convergence properties of MM algorithms detailed in this section.

The strategy for proving convergence of the Bradley–Terry MM algorithms is as follows. First, give a sufficient condition for upper compactness of the log-likelihood function [ℓ is defined to be upper compact if, for any constant c, the set $\{\boldsymbol{\gamma} \in \Omega : \ell(\boldsymbol{\gamma}) \geq c\}$ is a compact subset of the parameter space Ω]. Second, reparameterize the log-likelihood and give a sufficient condition for strict concavity of the reparameterized log-likelihood function. Since upper compactness implies the existence of at least one limit point and strict concavity implies the existence of at most one stationary point, namely the maximizer, we may conclude from Liapounov's theorem that the MM algorithm converges (irrespective of its starting point) to the unique maximum likelihood estimator. Note that, unlike some algorithms such as Newton–Raphson algorithms, an MM algorithm retains the same sequence of iterates after a reparameterization since reparameterization does not destroy the minorizing property or alter the maximum.

Almost all of the log-likelihood functions given in the previous section are upper compact if Assumption 1 is satisfied. The exception is the home-field advantage likelihood (13), for which we need a stronger assumption.

ASSUMPTION 2. In every possible partition of the teams into two nonempty subsets A and B, some team in A beats some team in B as home team, and some team in A beats some team in B as visiting team.

The following lemma gives sufficient, and in some cases necessary, conditions for the upper compactness of the likelihood functions seen in each of the models thus far.

LEMMA 1. *Let* $\Omega = \{\boldsymbol{\gamma} \in \mathbb{R}^m : \text{each } \gamma_i > 0, \sum_{i=1}^m \gamma_i = 1\}$. *The parameter space is assumed to be* Ω *for log-likelihoods* (2) *and* (8); $\Omega \times \{\theta \in \mathbb{R} : \theta > 0\}$ *for log-likelihoods* (16) *and* (13); *and* $\Omega \times \{\theta \in \mathbb{R} : \theta > 1\}$ *for log-likelihood* (14). *For the purpose of Assumption* 1, i *is said to beat* j *in a triple comparison if* i *is ranked higher than* j.

(a) *The log-likelihoods of* (2) *and* (8) *are upper compact if and only if Assumption* 1 *holds*.

(b) *Both of the log-likelihoods of* (14) *and* (16) *are upper compact if Assumption* 1 *holds and there is at least one tie*.

(c) *The log-likelihood of* (13) *is upper compact if Assumption* 2 *holds*.

The sufficient condition for upper compactness in the home-field advantage model, namely Assumption 2, may seem unusually strong, since it implies, for example, that each team must play at least four games, winning and losing both a home game and an away game. However, this is not an unrealistic assumption for many situations—consider, for example, the Major League Baseball schedule in the United States and Canada, in which no team ever wins or loses all of its home

games or all of its away games over the course of a season. In parts (b) and (c) of Lemma 1, it is not known whether the sufficient conditions are also necessary conditions.

As explained earlier, we now reparameterize the models and give conditions under which the log-likelihood functions are strictly concave. Let $\beta_i = \ln \gamma_i - \ln \gamma_1$ for $i = 1, \ldots, m$. The inverse function

$$\gamma_i = \frac{e^{\beta_i}}{\sum_{j=1}^m e^{\beta_j}}$$

establishes a one-to-one correspondence between $\{\boldsymbol{\gamma} \in \mathbb{R}_+^m : \sum_i \gamma_i = 1\}$ and $\{\boldsymbol{\beta} \in \mathbb{R}^m : \beta_1 = 0\}$. For models in which there is an additional parameter θ, let $\phi = \ln \theta$. Note that results (a) through (c) of Lemma 1 still hold after the reparameterization, since any sequence of parameter vectors that approaches the boundary of the original parameter space also approaches the boundary of the reparameterized space.

After reparameterization, the original Bradley–Terry model (1) becomes

$$\operatorname{logit}[P(i \text{ beats } j)] = \beta_i - \beta_j, \tag{19}$$

and the log-likelihood (2) becomes

$$\lambda(\boldsymbol{\beta}) = \sum_{i=1}^m \sum_{j=1}^m \bigl[w_{ij}\beta_i - w_{ij} \ln\bigl(e^{\beta_i} + e^{\beta_j}\bigr)\bigr]. \tag{20}$$

As (19) suggests, the Bradley–Terry model may be fitted using logistic regression. Agresti (1990) describes how to do this, noting that if a constant term is included in the model then it is the home-field advantage parameter $\phi = \log \theta$ of model (5) as long as the predictors are defined correctly. Logistic regression is not applicable to any of the other generalizations of the Bradley–Terry model discussed here.

The concavity of the log-likelihood (20) follows immediately because of the fact that the set of log-convex functions (i.e., functions whose logarithm is convex) is closed under addition. We can also prove concavity using Hölder's inequality, an approach with the additional benefit that it allows us to give sufficient conditions under which the concavity is strict. Taking logarithms in one form of Hölder's inequality [Magnus and Neudecker (1988)] shows that, for positive numbers $c_1, \ldots, c_N$ and $d_1, \ldots, d_N$ and $p \in (0, 1)$,

$$\ln \sum_{k=1}^N c_k^p d_k^{1-p} \le p \ln \sum_{k=1}^N c_k + (1-p) \ln \sum_{k=1}^N d_k, \tag{21}$$

with equality if and only if there exists some $\xi > 0$ such that $c_k = \xi d_k$ for all k. A log-likelihood λ is concave by definition if, for any parameter vectors $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $p \in (0, 1)$,

$$\lambda[p\boldsymbol{\alpha} + (1-p)\boldsymbol{\beta}] \ge p\lambda(\boldsymbol{\alpha}) + (1-p)\lambda(\boldsymbol{\beta}); \tag{22}$$

concavity is strict if $\boldsymbol{\alpha} \neq \boldsymbol{\beta}$ implies that the inequality in (22) is strict. Inequality (21) implies that

$$\begin{aligned} &-\ln\bigl[e^{p\alpha_i+(1-p)\beta_i} + e^{p\alpha_j+(1-p)\beta_j}\bigr] \\ &\qquad \geq -p\ln(e^{\alpha_i} + e^{\alpha_j}) - (1-p)\ln(e^{\beta_i} + e^{\beta_j}), \end{aligned} \tag{23}$$

so multiplying inequality (23) by w_{ij} and then summing over i and j demonstrates the concavity of the log-likelihood of (20).

The equality condition for Hölder's inequality (21) may be used to derive conditions for strict concavity of the reparameterized log-likelihood functions. These conditions are weaker than those of Lemma 1; for the most part, Assumption 3 is sufficient.

ASSUMPTION 3. In every possible partition of the individuals into two nonempty subsets, some individual in the second set is compared with some individual in the first set at least once.

The proof of the following lemma is given in the Appendix.

LEMMA 2. *For the reparameterization* $(\boldsymbol{\gamma}, \theta) \mapsto (\boldsymbol{\beta}, \phi)$ *in which* $\beta_i = \ln\gamma_i - \ln\gamma_1$ *and* $\phi = \ln\theta$, *let* $\Omega' = \{\boldsymbol{\beta} \in \mathbb{R}^m : \beta_1 = 0\}$.

(a) *The reparameterized versions of log-likelihoods* (2) *and* (8) *are strictly concave on* Ω' *if and only if Assumption* 3 *holds*.

(b) *The reparameterized version of* (14) *is strictly concave on* $\Omega' \times \mathbb{R}_+$ *and the reparameterized version of* (16) *is strictly concave on* $\Omega' \times \mathbb{R}$ *if and only if Assumption* 3 *holds and there is at least one tie*.

(c) *The reparameterized version of* (13) *is strictly concave on* $\Omega' \times \mathbb{R}$ *if Assumption* 3 *holds and there is a loop* $(i_0, i_1, \ldots, i_s = i_0)$ *such that* i_{j-1} *is home in at least one comparison against* i_j *for* $1 \leq j \leq s$.

Because the assumptions ensuring upper compactness given in Lemma 1 are stronger than those ensuring strict concavity, Liapounov's theorem (Theorem 1) implies that each of the MM algorithms, whether cyclic or not, is guaranteed to produce a sequence of parameter vectors converging to the maximum likelihood estimator under the assumptions of Lemma 1.

5. Multiple comparisons. Consider an extension of the Bradley–Terry model to comparisons involving $k \geq 3$ individuals, where the outcome of such a comparison is a ranking of the individuals from best to worst. This situation may arise, for example, when judges consider entries at a fair. Each judge might see only a few of the entries, then rank the entries seen. A thorough survey of models of this type is given by Marden (1995).

Suppose that there are m individuals, labeled 1 through m, in our population. For $A \subset \{1, \ldots, m\}$, say $A = \{1, \ldots, k\}$ with $k \leq m$, suppose that the individuals indexed by A are to be ranked. Let $\to$ denote the relation "is ranked higher than" and let $\mathcal{S}_k$ denote the group of permutations of k elements. Then, given A and some $\pi \in \mathcal{S}_k$, the probability we assign to the event $\pi(1) \to \cdots \to \pi(k)$ is

$$P_A[\pi(1) \to \cdots \to \pi(k)] = \prod_{i=1}^{k} \frac{\gamma_{\pi(i)}}{\gamma_{\pi(i)} + \cdots + \gamma_{\pi(k)}}. \tag{24}$$

This generalization of the Bradley–Terry model, termed the Plackett–Luce model by Marden (1995), was introduced by Plackett (1975). In the particular case of triple comparisons, model (24) reduces to the Pendergrass–Bradley (1960) model of (8). For any subset of A, say $\{1, 2\}$, we may interpret $P_A(1 \to 2)$ as

$$\sum_{\pi \in \mathcal{S}_k : \pi^{-1}(1) < \pi^{-1}(2)} P_A[\pi(1) \to \cdots \to \pi(k)],$$

the sum of the probabilities of all rankings of $\{1, \ldots, k\}$ such that $1 \to 2$. Ideally, the model should be *internally consistent* in the sense that the probability of a particular ranking does not depend on the subset from which the individuals are assumed to be drawn. In other words, if model (24) is internally consistent, then the subscript A in $P_A[\pi(1) \to \cdots \to \pi(k)]$ is unnecessary.

To prove internal consistency, let $A = \{1, \ldots, k\}$ as before and evaluate

$$\begin{aligned} &P_A(1 \to \cdots \to k-1) \\ &\quad = \gamma_1 \cdots \gamma_{k-1}\gamma_k \left[\frac{1}{(\gamma_1 + \cdots + \gamma_k) \cdots (\gamma_{k-1} + \gamma_k)\gamma_k} \right. \\ &\qquad \left. + \frac{1}{(\gamma_1 + \cdots + \gamma_k) \cdots (\gamma_k + \gamma_{k-1})\gamma_{k-1}} + \cdots \right], \end{aligned} \tag{25}$$

where the sum has k terms corresponding to the k distinct permutations in $\mathcal{S}_k$ that leave the order of $(1, \ldots, k-1)$ unchanged. Equation (25) simplifies to

$$\begin{aligned} &P_A(1 \to \cdots \to k-1) \\ &\quad = \frac{\gamma_1 \cdots \gamma_{k-1}}{(\gamma_1 + \cdots + \gamma_{k-1}) \cdots (\gamma_{k-2} + \gamma_{k-1})\gamma_{k-1}} \\ &\quad = P_{\{1,\ldots,k-1\}}(1 \to \cdots \to k-1). \end{aligned} \tag{26}$$

The subset $\{1, \ldots, k-1\}$ in (26) may be replaced by any subset of A with $k-1$ elements. Thus, using (26) repeatedly if necessary, for any $B = \{b_1, \ldots, b_l\} \subset A$,

$$P_A(b_1 \to \cdots \to b_l) = P_B(b_1 \to \cdots \to b_l). \tag{27}$$

Thus, the model is internally consistent so we may drop the subscripts A and B and simply write $P(b_1 \to \cdots \to b_l)$, or $P(\mathbf{b})$ for short. In particular, the number of individuals included in a given ranking need not be the same for all rankings. For example, a sport such as track and field is often contested at meets that can involve two or more teams. Results of an entire season of such meets could be combined using this model, resulting in an estimate of the relative strength of each team.

We mention briefly the connection between model (24) and Luce's choice axiom [Luce (1959)]. The axiom states that, for any model in which individual i has a positive probability of beating individual j when the two are compared as a pair for all $i \neq j$, we have

$$P_B(i \text{ wins}) = P_A(i \text{ wins})\, P_B(\text{anything from } A \text{ wins}) \qquad \text{for all } i \in A \subset B. \tag{28}$$

Luce (1959) showed that axiom (28) is equivalent to the statement

$$P_B(i \text{ wins}) = \frac{\gamma_i}{\sum_{j \in B} \gamma_j} \tag{29}$$

for positive-valued parameters γ_i. It is not hard to see that model (24) is equivalent to statement (29): Marden (1995) points out that model (24) arises from (29) if we envision the ranking process as first choosing a winner, then choosing a second-place finisher as the winner among those that remain and so on. The converse follows because, under (24),

$$\begin{aligned} P_A(i \text{ wins}) &= \sum_{\pi\,:\,\pi(1)=i} P_A[\pi(1) \to \cdots \to \pi(k)] \\ &= \sum_{\pi\,:\,\pi(1)=i} \frac{\gamma_i}{\gamma_1 + \cdots + \gamma_k} \prod_{j=2}^{k} \frac{\gamma_{\pi(j)}}{\gamma_{\pi(j)} + \gamma_{\pi(j+1)} + \cdots + \gamma_{\pi(k)}} \\ &= \frac{\gamma_i}{\gamma_1 + \cdots + \gamma_k}. \end{aligned}$$

Thus, model (24) is equivalent to Luce's choice axiom, which thus implies internal consistency in the sense defined above.

To fit model (24) using maximum likelihood, it is once again possible to construct a minorizing function using inequality (9). Suppose that the data consist of N rankings, where the jth ranking includes m_j individuals, $1 \leq j \leq N$. Let the ordered indices of the individuals in the jth ranking be denoted by $a(j,1), \ldots, a(j,m_j)$, so that $a(j,1) \to a(j,2) \to \cdots \to a(j,m_j)$ according to the jth ranking. Assuming independent rankings, the log-likelihood may be written as

$$\ell(\boldsymbol{\gamma}) = \sum_{j=1}^{N} \sum_{i=1}^{m_j - 1} \left[\ln \gamma_{a(j,i)} - \ln \sum_{s=i}^{m_j} \gamma_{a(j,s)} \right].$$

By inequality (9),

$$Q_k(\boldsymbol{\gamma}) = \sum_{j=1}^{N} \sum_{i=1}^{m_j - 1} \left[\ln \gamma_{a(j,i)} - \frac{\sum_{s=i}^{m_j} \gamma_{a(j,s)}}{\sum_{s=i}^{m_j} \gamma_{a(j,s)}^{(k)}} \right]$$

minorizes the log-likelihood $\ell(\boldsymbol{\gamma})$ at $\boldsymbol{\gamma}^{(k)}$, up to a constant. With the parameters now separated, maximization of $Q_k(\boldsymbol{\gamma})$ may be explicitly accomplished by

$$\gamma_t^{(k+1)} = \frac{w_t}{\sum_{j=1}^{N} \sum_{i=1}^{m_j - 1} \delta_{jit} [\sum_{s=i}^{m_j} \gamma_{a(j,s)}^{(k)}]^{-1}} \tag{30}$$

for $t = 1, \ldots, m$, where w_t is the number of rankings in which the tth individual is ranked higher than last and

$$\delta_{jit} = \begin{cases} 1, & \text{if } t \in \{a(j,i), \ldots, a(j,m_j)\}, \\ 0, & \text{otherwise.} \end{cases}$$

In other words, δ_{jit} is the indicator of the event that individual t receives a rank no better than i in the jth ranking. The MM algorithm of (30) generalizes (3); alternatively, the components of $\boldsymbol{\gamma}$ may be updated cyclically.

In this context, Assumption 1 makes sense if we interpret individual i beating individual j to mean that i is ranked higher than j in a ranking that includes both of them. As in Lemmas 1(a) and 2(a), Assumption 1 is necessary and sufficient for the upper compactness of the log-likelihood function, whereas Assumption 3 is necessary and sufficient for the strict concavity of the log-likelihood function under the reparameterization $\beta_i = \ln \gamma_i - \ln \gamma_1$. We conclude that the MM algorithms in this section are guaranteed to converge to the unique maximum likelihood estimator if Assumption 1 holds. The author does not know of another algorithm in the literature specifically for maximizing the Plackett–Luce likelihood; Plackett (1975) merely states that "the maximum of the likelihood can be determined only by numerical methods."

6. A numerical example. Competitive racing, whether it involves humans, animals or machines, provides ready examples in which subsets of a group of individuals are ranked. If we consider different races among some group of individuals to be independent, we may fit a Plackett–Luce model (24) to estimate the relative strengths of the individuals.

For example, consider the 36 automobile races for the 2002 NASCAR season in the United States. Each of these races involved 43 drivers, with some drivers participating in all 36 races and some participating in only one. Altogether, 87 different drivers participated in at least one race. However, Assumption 1 is violated due to the fact that four of the drivers placed last in each race they entered.

When these four are removed, we obtain a set of 83 drivers and 36 races, some of which involve 43 drivers and some of which involve only 42 drivers. As noted in the previous section, there is no problem in fitting a Plackett–Luce model when the comparisons involve different-sized subsets as in this case. Assumption 1 holds for the reduced field of 83 drivers.

For purposes of comparison, the Plackett–Luce model is fitted using two methods, the MM algorithm of (30) and a modified Newton–Raphson algorithm. The Newton–Raphson algorithm operates in the reparameterized parameter space $\{\boldsymbol{\beta} \in \mathbb{R}^{83} : \beta_1 = 0\}$, starting from the point $\boldsymbol{\beta} = \mathbf{0}$. Since the log-likelihood function is strictly concave on this space, as shown in the proof of Lemma 2(a), we might expect that an unmodified Newton–Raphson algorithm would be well behaved here; however, this is not the case. Therefore, a modified algorithm is used in which the proposed Newton–Raphson step is taken whenever that step results in an increase in the log-likelihood, and otherwise the MM step is taken. Since the MM algorithm is guaranteed to increase the log-likelihood at each iteration, it bails out the Newton–Raphson algorithm until the iterates are close enough to the maximizer that the Newton–Raphson step is effective.

The MM algorithm also starts from the point $\boldsymbol{\beta} = \mathbf{0}$, or, equivalently, $\boldsymbol{\gamma} = (\frac{1}{83}, \ldots, \frac{1}{83})$; as pointed out previously, the reparameterization does not change the MM algorithm. Although selection of the initial point affects the final iteration count of the algorithm, MM algorithms tend to get close to the answer quickly and then slow down; thus, the speed of convergence will likely be similar for all starting values far from the maximum. In fact, for numerous randomly chosen starting vectors in the NASCAR example, the algorithm always converged in either 25 or 26 iterations. In the implementation of the MM algorithm, the $\boldsymbol{\gamma}$ vector is not renormalized to satisfy $\sum_i \gamma_i = 1$ after each iteration, since doing so is unnecessary. Because the MATLAB code for the MM algorithm modifies the entire $\boldsymbol{\gamma}$ vector at once, the noncyclic version of the algorithm—that is, the algorithm of (30)—is used. Both the MM algorithm and the Newton–Raphson algorithm are terminated and convergence is declared when a simplistic stopping criterion is met, namely that the L_2-norm of the change in the value of the parameter vector is less than 10^{-9}. At convergence, the two algorithms produce MLE vectors that differ by less than 10^{-13} in each component.

MATLAB allows the counting of floating-point operations. Since MM algorithms tend to result in more but faster iterations than Newton–Raphson algorithms [Lange, Hunter and Yang (2000)] the counts of iterations until convergence are misleading. The number of floating-point operations is used as a measure of the overall computing effort required and serves as the basis of comparison here.

In the modified Newton–Raphson algorithm, even for an MM iteration, all of the work of determining the Newton–Raphson step is required. Therefore, the computational work reported in Table 1 for this algorithm is roughly the same as would be required for 10 Newton–Raphson iterations. It is instructive to note that the computational work required even to invert a single 82×82 matrix,

TABLE 1

Performance of an MM algorithm and a modified Newton–Raphson (MNR) algorithm in finding the maximum likelihood estimator for the Plackett–Luce model (24) fit to the 2002 NASCAR data. For the MNR algorithm, an MM iteration was taken whenever the Newton–Raphson step failed to increase the likelihood

	Number of iterations		Floating point operations
Type of algorithm	**MM**	**Newton–Raphson**	**required**
MM	26	—	0.22×10^6
Modified Newton–Raphson	4	6	14.45×10^6

which is done at each Newton–Raphson iteration, is roughly 1.13×10^6 floating-point operations according to MATLAB; this is itself over 5 times more than the MM algorithm requires to converge completely in 26 iterations. The number of iterations required by MM in this example is surprisingly small; as noted, for example, in Lange, Hunter and Yang (2000), an MM algorithm often requires several hundred iterations despite saving overall computational work as compared with other algorithms.

Obtaining standard error estimates for parameters in a Bradley–Terry model is easy in principle; the inverse of the Hessian matrix of the log-likelihood (or, alternatively, the inverse of the Fisher information matrix) evaluated at the MLE gives an asymptotic approximation to the covariance matrix of the MLE. However, this is not ideal because part of the point of an MM algorithm is to avoid the inversion of large matrices; furthermore, computation of the Hessian or Fisher information matrix may be burdensome. For the NASCAR example, the inverse Hessian matrix is readily available because it is needed for the Newton–Raphson algorithm; several of the standard errors obtained from this matrix are listed in Table 2. Of note in this example is the fact that the standard errors, while larger for drivers who have not competed in many races, are not inversely proportional to the square root of the number of races. Furthermore, the rank order of the MLEs does not correspond completely with that of the average places. Both of these facts suggest that a great deal of the information about β_i is derived from the strengths of the other drivers in the race, which is one aspect of the Plackett–Luce model that might make it more appealing as a method for ranking the drivers than, say, average place.

Note that some care should be exercised in interpreting standard errors in this model. While standard asymptotic arguments enable one to build confidence intervals (for example) for parameters or contrasts, the standard errors should not be viewed as estimates of the true standard deviations of the MLE. The reason for this is that the MLE for any individual (on the β-scale) can be $\pm\infty$ with positive probability, since this happens whenever that individual wins or loses all contests.

TABLE 2
Top ten and bottom ten drivers according to average place, along with MLEs in $\boldsymbol{\beta}$-space and standard errors. The β-value of A. Cameron is constrained to be 0. Standard errors are obtained from the inverse Hessian matrix evaluated at the MLE

Driver	Races	Average place	$\hat{\beta}_i$	se($\hat{\beta}_i$)
P. Jones	1	4.00	4.15	1.57
S. Pruett	1	6.00	3.62	1.53
M. Martin	36	12.17	2.08	1.05
T. Stewart	36	12.61	1.83	1.05
R. Wallace	36	13.17	2.06	1.05
J. Johnson	36	13.50	1.94	1.05
S. Marlin	29	13.86	1.73	1.04
M. Bliss	1	14.00	2.23	1.47
J. Gordon	36	14.06	1.74	1.05
K. Busch	36	14.06	1.65	1.05
⋮				
C. Long	2	40.50	−0.32	1.30
C. Fittipaldi	1	41.00	−0.44	1.49
H. Fukuyama	2	41.00	−0.76	1.45
J. Small	1	41.00	−0.54	1.48
M. Shepherd	5	41.20	−0.45	1.16
K. Shelmerdine	2	41.50	−0.32	1.28
A. Cameron	1	42.00	0.00	0.00
D. Marcis	1	42.00	0.03	1.46
D. Trickle	3	42.00	−0.31	1.20
J. Varde	1	42.00	−0.15	1.48

7. Discussion. This paper does not claim to give a comprehensive treatment of all known generalizations of the Bradley–Terry model, nor is the main thrust of the paper to derive new algorithms and results, though some of the results here are new. The main thrust of the paper is to demonstrate that a single simple set of ideas may be applied to a broad range of generalizations of the Bradley–Terry model. Once certain fundamental properties of MM algorithms are established and a means for creating them for a particular class of log-likelihoods is in place, derivation of simple, reliable algorithms that are guaranteed to yield maximum likelihood estimates can be straightforward. Among the many distinct algorithms exhibited here, along with proofs that they converge to the maximum likelihood estimators, are several that have appeared earlier in the Bradley–Terry literature that were custom-built for particular log-likelihood functions. Here, they are mass-produced as specific examples of a general principle.

One of the well-known drawbacks (arguably the most important drawback) of EM algorithms in particular and MM algorithms in general is their slow

rate of convergence, particularly as the iterates close in on a stationary point. Computationally, MM algorithms tend to give fast, simple-to-code iterations, where each iteration moves in the right direction, but the tradeoff is they require many more iterations than an optimization method like Newton–Raphson. Newton–Raphson, on the other hand, despite converging in relatively few iterations, requires the computation and then the inversion of a square matrix at each iteration, operations that may be extremely time-consuming, particularly if the number of parameters is large. Furthermore, there is no guarantee that a Newton–Raphson step will increase the value of the objective function, so any well-designed Newton–Raphson algorithm must contain safeguards against erratic behavior. In the example of Section 6 in which the Plackett–Luce model is applied to NASCAR racing data, the MM algorithm fares extremely well as compared with the Newton–Raphson algorithms, even with respect to the number of iterations required for convergence.

Lange, Hunter and Yang (2000) discuss a method for accelerating an MM algorithm using a quasi-Newton approach. The idea is to construct a hybrid algorithm that attempts to retain the best features of both MM and Newton–Raphson. To do this, the algorithm begins as an MM algorithm, gradually building at each iteration an approximation to the inverse of the Hessian matrix that would be used in a Newton–Raphson algorithm. When the approximate inverse Hessian may be used productively in a Newton–Raphson step, it is used; until then, however, it is merely updated at each iteration using information in the MM algorithm iterations. Given the speed with which the unaltered MM algorithm converged in the example of Section 6, such an acceleration scheme was not needed—however, Lange, Hunter and Yang (2000) apply this quasi-Newton technique with dramatic effect in an example involving the original Bradley–Terry model (1).

Improved means for formulating standard error estimates via MM algorithms is an area where further research is called for. One possibility is to adapt known results from the EM algorithm literature, such as the SEM algorithm idea of Meng and Rubin (1991), to the more general MM case; however, the overhead for implementing SEM is great for problems with many parameters. Another possibility is to rely on an estimated inverse Hessian matrix obtained through a quasi-Newton acceleration of an MM algorithm, as described in Lange, Hunter and Yang (2000). Still another approach is a parametric bootstrap approach in which the model is fitted and then all comparisons are repeatedly resimulated according to the model. The problem with this approach is that, with positive probability, a simulated set of comparisons will violate Assumption 1. With many bootstrap samples being taken, even an event with a small probability can be quite likely to show up at least once. One possibility for avoiding this problem, and for making Assumptions 1–3 unnecessary, might be to give each individual a tiny fraction of a "win" against every other individual before fitting the model. The effect this correction might have on the estimates produced is as yet unknown.

APPENDIX

PROOF OF LEMMA 1. Part of the main idea of this proof may be found in the proof by Ford (1957) that algorithm (4) converges to the maximum likelihood estimator. Let $\Omega = \{\boldsymbol{\gamma} \in \mathbb{R}^m : \sum_i \gamma_i = 1\}$.

(a) Consider what happens to $\ell(\boldsymbol{\gamma})$ as $\boldsymbol{\gamma}$ approaches the boundary of Ω. If $\tilde{\boldsymbol{\gamma}}$ lies on the boundary of Ω, then $\tilde{\boldsymbol{\gamma}}_i = 0$ and $\tilde{\boldsymbol{\gamma}}_j > 0$ for some i and j. As noted earlier, if individuals are nodes of a directed graph in which edges represent wins, then Assumption 1 implies that a directed path exists from i to j. Therefore, there must be some individual a with $\tilde{\gamma}_a = 0$ who defeated some individual b with $\tilde{\gamma}_b > 0$, which means that, for $\boldsymbol{\gamma} \in \Omega$, taking limits in

$$\ell(\boldsymbol{\gamma}) \le \ln \gamma_a - \ln(\gamma_a + \gamma_b)$$

gives $\lim_{\boldsymbol{\gamma} \to \tilde{\boldsymbol{\gamma}}} \ell(\boldsymbol{\gamma}) = -\infty$. Thus, for any constant c, the set $\{\boldsymbol{\gamma} \in \Omega : \ell(\boldsymbol{\gamma}) \ge c\}$ is a closed and bounded, hence compact, set.

Conversely, suppose that the individuals may be partitioned into two groups A and B such that nobody from A ever beats anybody from B. The log-likelihood cannot decrease if every γ_i with $i \in B$ is doubled, then the resulting vector renormalized so that $\sum_i \gamma_i = 1$. In this way, $\boldsymbol{\gamma}$ may be driven to the boundary of Ω without decreasing the log-likelihood.

Note that these arguments apply equally well to the original Bradley–Terry model (1), the triple comparison model (8) or the Plackett–Luce model (24).

(b) As in part (a), suppose that $(\tilde{\boldsymbol{\gamma}}, \tilde{\theta})$ is on the boundary of the parameter space, which is $\Omega \times (1, \infty)$ for model (6) or $\Omega \times (0, \infty)$ for model (7). It suffices to show that

$$\lim_{(\boldsymbol{\gamma}, \theta) \to (\tilde{\boldsymbol{\gamma}}, \tilde{\theta})} \ell(\boldsymbol{\gamma}, \theta) = -\infty. \tag{31}$$

If $\tilde{\boldsymbol{\gamma}}$ is on the boundary of Ω, the proof of part (a), with minor modifications, proves that (31) holds regardless of the value of $\tilde{\theta}$. On the other hand, if $\tilde{\boldsymbol{\gamma}} \in \Omega$, then $T > 0$ implies (31) immediately if $\tilde{\theta} = 1$ in the case of model (14) or $\tilde{\theta} = 0$ in the case of model (16). Furthermore, since

$$\ell(\tilde{\boldsymbol{\gamma}}, \theta) \le \ln \tilde{\gamma}_i - \ln(\tilde{\gamma}_i + \theta \tilde{\gamma}_j)$$

in the case of (14) or

$$\ell(\tilde{\boldsymbol{\gamma}}, \theta) \le \ln \tilde{\gamma}_i - \ln\bigl(\tilde{\gamma}_i + \tilde{\gamma}_j + \theta \sqrt{\tilde{\gamma}_i \tilde{\gamma}_j}\bigr)$$

in the case of (16), we see that for $\tilde{\theta} = \infty$ we also obtain (31).

(c) Note that, for $\tilde{\boldsymbol{\gamma}} \in \Omega$, if $\tilde{\theta} = 0$ then Assumption 2 implies (31) because there must be at least one home win; similarly, at least one home loss implies (31) if $\tilde{\theta} = \infty$ and $\tilde{\boldsymbol{\gamma}} \in \Omega$. On the other hand, if $\tilde{\boldsymbol{\gamma}}$ is on the boundary of Ω, then by the

argument of part (a), there exist a and b such that $\tilde{\gamma}_a = 0$, $\tilde{\gamma}_b > 0$ and a defeated b with a at home. Therefore,

$$\ell(\boldsymbol{\gamma}, \theta) \le \ln(\theta\gamma_a) - \ln(\theta\gamma_a + \gamma_b)$$

implies (31) as long as $\tilde{\theta} < \infty$. Similarly, since there also exist a' and b' such that $\tilde{\gamma}_{a'} = 0$, $\tilde{\gamma}_{b'} > 0$ and a' defeated b' with b' at home,

$$\ell(\boldsymbol{\gamma}, \theta) \le \ln(\gamma_{a'}) - \ln(\gamma_{a'} + \theta\gamma_{b'})$$

implies (31) even if $\tilde{\theta} = \infty$. $\square$

PROOF OF LEMMA 2. Let $\Omega' = \{\boldsymbol{\beta} \in \mathbb{R}^m : \beta_1 = 0\}$. Recall the definition of strict concavity given after inequality (22).

(a) In the case of the Bradley–Terry model (1), take $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \Omega'$ and $p \in (0, 1)$. By the equality condition for Hölder's inequality (21), equality in (22) implies, in particular, that $\beta_i - \alpha_i = \beta_j - \alpha_j$ for all i and j for which $\max\{w_{ij}, w_{ji}\} > 0$. Thus, Assumption 3 plus the fact that $\beta_1 = \alpha_1 = 0$ means that $\boldsymbol{\alpha} = \boldsymbol{\beta}$ and so the concavity is strict.

Conversely, if Assumption 3 is violated, then the individuals may be partitioned into two groups, A and B, such that no intergroup comparisons take place. One of the groups, say A, does not contain individual 1, so there are no constraints on the parameters in group A. Thus, adding the same constant to each of the β_i for i in group A does not change the likelihood, and the concavity fails to be strict.

By a nearly identical argument, the same results hold for the triple-comparison model (8) and the Plackett–Luce model (24).

(b) Let $\lambda(\boldsymbol{\beta}, \phi)$ denote the log-likelihood function, either (14) or (16) as the case may be, after the reparameterization. In the case of the Rao–Kupper model (6),

$$\lambda[p(\boldsymbol{\alpha}, \phi_1) + (1-p)(\boldsymbol{\beta}, \phi_2)] = p\lambda(\boldsymbol{\alpha}, \phi_1) + (1-p)\lambda(\boldsymbol{\beta}, \phi_2) \tag{32}$$

implies by the equality condition for Hölder's inequality that, whenever $\max\{w_{ij}, t_{ij}\} > 0$, $\beta_i - \alpha_i = \beta_j - \alpha_j + \phi_2 - \phi_1$. However, when $T > 0$ the log-likelihood includes the strictly concave term $\ln[\exp(2\phi) - 1]$, so (32) implies $\phi_2 = \phi_1$. Thus, Assumption 3 along with $\beta_1 = \alpha_1 = 0$ implies that $(\boldsymbol{\alpha}, \phi_1) = (\boldsymbol{\beta}, \phi_2)$ and the concavity is strict.

In the case of the Davidson model (7), (32) implies by the Hölder equality condition that, whenever $\max\{w_{ij}, w_{ji}, t_{ij}\} > 0$, we obtain

$$\beta_i - \alpha_i = \beta_j - \alpha_j = \phi_2 - \phi_1 + \frac{\beta_i + \beta_j - \alpha_i - \alpha_j}{2}. \tag{33}$$

But (33) implies both $\beta_i - \alpha_i = \beta_j - \alpha_j$ and $\phi_1 = \phi_2$, so Assumption 3 implies $(\boldsymbol{\alpha}, \phi_1) = (\boldsymbol{\beta}, \phi_2)$.

For the converse, the argument for both models is the same as in part (a).

(c) With $\lambda(\boldsymbol{\beta}, \phi)$ denoting the log-likelihood function for the reparameterized home-field advantage model, (32) implies by the Hölder equality condition that $\phi_1 - \phi_2 + \beta_i - \alpha_i = \beta_j - \alpha_j$ whenever i is home in a comparison against j. Therefore, the existence of a loop as described in the lemma implies that $s(\phi_1 - \phi_2) + \beta_{i_1} - \alpha_{i_1} = \beta_{i_1} - \alpha_{i_1}$ for some positive integer s, which means that $\phi_1 = \phi_2$. Therefore, Assumption 3 implies $\boldsymbol{\beta} = \boldsymbol{\alpha}$ as in part (a). □

REFERENCES

AGRESTI, A. (1990). *Categorical Data Analysis*. Wiley, New York.

BRADLEY, R. A. and TERRY, M. E. (1952). Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika* **39** 324–345.

DAVID, H. A. (1988). *The Method of Paired Comparisons*, 2nd ed. Oxford Univ. Press.

DAVIDSON, R. R. (1970). On extending the Bradley–Terry model to accommodate ties in paired comparison experiments. *J. Amer. Statist. Assoc.* **65** 317–328.

DAVIDSON, R. R. and FARQUHAR, P. H. (1976). A bibliography on the method of paired comparisons. *Biometrics* **32** 241–252.

DYKSTRA, O., JR. (1956). A note on the rank of incomplete block designs—applications beyond the scope of existing tables. *Biometrics* **12** 301–306.

FORD, L. R., JR. (1957). Solution of a ranking problem from binary comparisons. *Amer. Math. Monthly* **64** 28–33.

HASTIE, T. and TIBSHIRANI, R. (1998). Classification by pairwise coupling. *Ann. Statist.* **26** 451–471.

HEISER, W. J. (1995). Convergent computation by iterative majorization. In *Recent Advances in Descriptive Multivariate Analysis* (W. J. Krzanowski, ed.) 157–189. Oxford Univ. Press.

HUNTER, D. R. and LANGE, K. (2000). Rejoinder to discussion of "Optimization transfer algorithms using surrogate objective functions." *J. Comput. Graph. Statist.* **9** 52–59.

LANGE, K. (1995). A gradient algorithm locally equivalent to the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **57** 425–437.

LANGE, K., HUNTER, D. R. and YANG, I. (2000). Optimization transfer using surrogate objective functions (with discussion). *J. Comput. Graph. Statist.* **9** 1–59.

LUCE, R. D. (1959). *Individual Choice Behavior*. Wiley, New York.

MAGNUS, J. R. and NEUDECKER, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley, New York.

MARDEN, J. I. (1995). *Analyzing and Modeling Rank Data*. Chapman and Hall, London.

MCLACHLAN, G. J. and KRISHNAN, T. (1997). *The EM Algorithm and Extensions*. Wiley, New York.

MENG, X.-L. and RUBIN, D. B. (1991). Using EM to obtain asymptotic variance–covariance matrices: The SEM algorithm. *J. Amer. Statist. Assoc.* **86** 899–909.

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

PENDERGRASS, R. N. and BRADLEY, R. A. (1960). Ranking in triple comparisons. In *Contributions to Probability and Statistics* (O. Olkin et al., eds.) 331–351. Stanford Univ. Press.

PLACKETT, R. L. (1975). The analysis of permutations. *Appl. Statist.* **24** 193–202.

RAO, P. V. and KUPPER, L. L. (1967). Ties in paired-comparison experiments: A generalization of the Bradley–Terry model. *J. Amer. Statist. Assoc.* **62** 194–204. [Corrigendum *J. Amer. Statist. Assoc.* **63** 1550–1551.]

SHAM, P. C. and CURTIS, D. (1995). An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Ann. Human Genetics* **59** 323–336.

SIMONS, G. and YAO, Y.-C. (1999). Asymptotics when the number of parameters tends to infinity in the Bradley–Terry model for paired comparisons. *Ann. Statist.* **27** 1041–1060.

STIGLER, S. M. (1994). Citation patterns in the journals of statistics and probability. *Statist. Sci.* **9** 94–108.

ZERMELO, E. (1929). Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Math. Z.* **29** 436–460.

DEPARTMENT OF STATISTICS
PENNSYLVANIA STATE UNIVERSITY
UNIVERSITY PARK, PENNSYLVANIA 16802
USA