Parameter Estimation in Large Dynamic Paired Comparison Experiments

Mark E. Glickman*
Boston University, USA

Summary

Paired comparison data in which the abilities or merits of the objects being compared may be changing over time can be modeled as a non-linear state-space model. When the population of objects being compared is large, likelihood-based analyses can be too computationally cumbersome to carry out on a regular basis. This presents a problem for rating populations of chess players and other large groups which often consist of tens of thousands of competitors. This problem is overcome through a computationally simple, non-iterative algorithm for fitting a particular dynamic paired comparison model. The algorithm, which improves over the commonly used algorithm of Elo by incorporating the variability in parameter estimates, can be performed on a regular basis even for large populations of competitors. The method is evaluated on simulated data, and is applied to ranking the best chess players of all time, and to ranking the top current tennis players.

Keywords: Approximate Bayesian estimation, Bradley-Terry model, Chess, Ranking, State-space models, Tennis.

1 Introduction

Paired comparison models address data that arise from situations in which two objects or treatments are compared at a time to determine a degree of preference. Examples include modeling choice behavior (preference of one soft drink to another, or the preference of the Democratic presidential candidate to the Republican candidate), competitive ability in sports (determining the strengths

*Address for correspondence: Department of Mathematics and Statistics, Boston University, 111 Cummington Street, Boston, MA 02215, USA. E-mail: mg@math.bu.edu

of teams in basketball or baseball), as well as many others. A review of some examples and issues involved in paired comparison modeling was given by David (1988) and Bradley (1984).

Recent work (Fahrmeir and Tutz, 1994; Glickman, 1993) extended the usual (static) paired comparison models by including parameters that are time-varying. These "dynamic" paired comparison models are appropriate, for example, for modeling paired comparison data arising from competitive sports where player or team abilities change over time. When the size of the population of competitors is reasonably small, the methodologies developed in these papers present no computational difficulties. However, these methods are inadequate for large populations of competitors because the computation becomes intractable. For example, with more than 30,000 chess players playing over 450,000 games each year in United States Chess Federation (USCF) chess tournaments, other less computationally intensive methods for fitting dynamic paired comparison models need to be considered.

This paper presents a non-iterative method for fitting dynamic paired comparison models. The method is especially useful when the population of objects or treatments to be compared is large, and where parameter estimates are desired on an ongoing basis. Measuring the abilities of chess players is the motivating example, though the method applies directly to other paired comparison settings where abilities or merits change over time. Section 2 introduces our dynamic paired comparison model. The non-iterative parameter updating algorithm is presented in Section 3. In our procedure, certain parameters need to be estimated before applying the updating algorithm, and the estimation of these model parameters is described in Section 4. The algorithm is then evaluated on simulated data in Section 5. Finally, the method is applied to two data sets: chess game outcomes among the best chess players of all time, and tennis match outcomes played among current tournament players.

2 A dynamic paired comparison model

The model we assume for competitor ability is closely related to the Bradley-Terry model for paired comparisons (Bradley and Terry, 1952). The Bradley-Terry model asserts that for two objects with merit parameters λ_1 and λ_2, object 1 is preferred to object 2 with probability $\lambda_1/(\lambda_1 + \lambda_2)$. For our specific problem, let θ_i and θ_j be the unknown (scalar) strengths for players i and j at a fixed point in time. Assume first that a game results in only two outcomes; a win or a loss. Let s_{ijk} be the k-th outcome of a game played between players i and j, where $s_{ijk} = 1$ when player i wins and $s_{ijk} = 0$ when player j wins. The model for game outcomes, not allowing for ties or partial preferences, is given by

$$\Pr(s_{ijk} = s) = \frac{(10^{(\theta_i - \theta_j)/400})^s}{1 + 10^{(\theta_i - \theta_j)/400}} \quad , \tag{1}$$

for $s = 0, 1$. This is simply a reparametrized version of the Bradley-Terry model. Likelihood-based inference for the Bradley-Terry model is straightforward from a set of paired comparison data. This particular reparametrization was chosen to produce parameter estimates that have an interpretation on the same scale as the USCF rating system, which corresponds to strength estimates roughly between 0 and 3000.

Several extensions to the Bradley-Terry model incorporating ties have been proposed, including those by Davidson (1970) and Rao and Kupper (1967), each of whom introduce a single parameter governing the frequency of ties. Joe (1990) found that Davidson's model is not well fitted by a particular chess data set. In our likelihood framework, instead of adopting an approach that models a tie as a possible outcome of a game, we act as if ties do not really occur, but treat a tie as half way between a win and a loss when constructing the likelihood. This approach avoids the complications of including extra (possibly time-varying) parameters in the model to account for the probability of a third outcome. More formally, we assume that the information about players' strengths resulting from a win followed by a loss is equivalent to the information resulting from two consecutive ties.

Thus, if p is the probability that the first player wins, so that the contribution to the likelihood of a win followed by a loss is $p(1-p)$, then the contribution to the likelihood of a single tie should be $\sqrt{p(1-p)}$. We therefore construct a likelihood using terms in model (1) where a tie corresponds to $s = 0.5$. Other extensions to the Bradley-Terry model, such as recognizing a home-field advantage for certain sports or the advantage of having the first move in chess, can be incorporated by the inclusion of an "order-effect" parameter, as in Davidson and Beaver (1977). We do not pursue this extension here.

Model (1) addresses measuring competitor ability when players' abilities remain fixed over time. This model can be extended to recognize that players' abilities can change over time. Glickman (1993) and Fahrmeir and Tutz (1994) explored an approach to modeling paired comparison data with time-varying abilities through the use of state-space models. To use this approach, we assume that groups of comparisons considered to be collected in a short time interval are assumed to fall in the same "rating period." For example, chess tournament games played during a two-month period could be considered to be part of the same rating period. Denoting the strength of player i during a rating period at time t_0 by $\theta_i^{(t_0)}$, and the strength of player i a rating period t units of time later by $\theta_i^{(t_0+t)}$, we adopt a model that assumes

$$\theta_i^{(t_0+t)} | \theta_i^{(t_0)}, \nu^2, t \sim \mathrm{N}(\theta_i^{(t_0)}, \nu^2 t) \quad , \tag{2}$$

where ν^2 is the increase in variance in competitors' strength per unit time. This model asserts that as time passes while a player is not competing, the description of a competitor's strength becomes more uncertain. Knowledge of a player's activities (e.g., preparation) between events could be incorporated into (2), but we assume such information is not generally available.

A likelihood-based analysis of paired comparison data using state-space models can follow either Fahrmeir and Tutz (1994), who used empirical Bayes methods, or Glickman (1993), who used Markov chain Monte Carlo simulation from the posterior distribution. In both approaches, a com-

plete analysis involves estimating all the parameters jointly. With a small population of competitors (e.g., teams in a league), this does not present a difficulty. However, because competition in many organizations involves populations of thousands of players (e.g., chess, interactive games on the internet), an exact likelihood-based analysis may not be computationally feasible.

3 An approximate Bayesian updating algorithm

Suppressing the superscript t, let θ be the strength parameter of a player whose ability is to estimated. Prior to a rating period, we assume the prior distribution of a player's strength is

$$\theta|\mu,\sigma^2 \sim \mathrm{N}(\mu,\sigma^2) \tag{3}$$

with μ and σ^2 known. During a rating period, the player competes against m opponents, playing n_j games against opponent j, where $j = 1,\ldots,m$. We assume the distribution of the j-th opponent's strength, θ_j, is

$$\theta_j|\mu_j,\sigma_j^2 \sim \mathrm{N}(\mu_j,\sigma_j^2) \tag{4}$$

with known μ_j and σ_j^2. Let s_{jk} be the score of the k-th game against opponent j, with $s_{jk} = 1$ if the player wins game k, $s_{jk} = 0.5$ if the game results in a tie, and $s_{jk} = 0$ if the player loses. As before, we assume the likelihood will be a product of Bradley-Terry "probabilities"

$$p(s_{jk}|\theta,\theta_j) = \frac{(10^{(\theta-\theta_j)/400})^{s_{jk}}}{1+10^{(\theta-\theta_j)/400}} \quad . \tag{5}$$

The following sections develop a rating system based on closed-form approximations to the computations required to perform a fully Bayesian model fit of the state-space model described in Section 2. The rating algorithm is implemented as follows:

1. Collect game outcome data over a rating period.

2. At the end of the period, update players' rating distributions due to game outcomes from their pre-period (prior) rating distributions.

3. Subsequently update players' rating distributions due to the passage of time.

This is repeated for every rating period.

We proceed by describing prior distributions for players not having competed in tournaments, then the procedure for computing posterior distributions of strengths due to the passage of time, and finally the procedure for updating the strength distributions due to game results. The updating algorithm from game results is assessed in a simulation analysis. We also describe a smoothing procedure to make inferences on player strengths for early rating periods.

3.1 Prior distribution of strengths

Assume that, prior to competing, players' strengths are drawn independently from a normal distribution with mean 1500, and unknown variance. The initial variance, σ_0^2, is treated as a parameter to be inferred from data. Thus, the prior distribution assumed for any player with strength parameter θ before competing is

$$\begin{aligned} \theta|\sigma_0^2 &\sim \mathrm{N}(1500, \sigma_0^2) \\ \pi(\sigma_0^2) &\propto 1 \ . \end{aligned}$$

Incorporating useful sources of information (such as players' ages) could result in more informative prior distributions, though for the development of the algorithm such information is assumed to be unavailable.

3.2 Updating from the passage of time

In performing the updating computations, we act as if all games in a rating period are played at the beginning of the period. Over the duration of the rating period, we assume that knowledge of a player's strength becomes less certain, so that the parameter variance increases. At time t_0,

assume a player's strength is distributed as

$$\theta^{(t_0)}|\mu,\sigma^2 \sim \mathrm{N}(\mu,\sigma^2) \quad . \tag{6}$$

Integrating the distribution of $\theta^{(t_0+t)}|\theta^{(t_0)},\nu^2,t \sim \mathrm{N}(\theta^{(t_0)},\nu^2 t)$ with respect to the prior distribution (6) yields

$$\theta^{(t_0+t)}|\mu,\sigma^2,\nu^2,t \sim \mathrm{N}(\mu,\sigma^2+\nu^2 t) \quad , \tag{7}$$

where ν^2 is the increase in variance per unit time. In other words, an elapse of t units in time corresponds to an increase in $\nu^2 t$ in the variance. This model for the increase in variance preserves the additivity of variance with respect to time. The variance per unit time, ν^2, is a parameter that needs to be inferred in the model fitting process.

The choice of length of a rating period involves a variance/bias tradeoff. For short rating periods, little data may be available to estimate players' strengths, and the analytic approximations used in the algorithm in Section 3.3 may not be dependable. Conversely, if long rating periods are used, a player's ability may have changed substantially over a rating period, but this would not be detectable. The best compromise seems to be rating periods that are as short as possible, but where enough data is available to have some indication of players' strengths, perhaps at least 5–10 games per player on average.

3.3 Updating from game outcomes

At the beginning of a rating period, every competitor has a normal prior distribution of their playing strength. Rather than determine the posterior distribution of all strength parameters simultaneously making use of all paired comparison information, we carry out an approximate Bayesian analysis which leads to a tractable set of closed-form computations. The key idea is that the marginal posterior distribution of a player's strength is determined by integrating out the opponents' strength parameters over their *prior* distribution rather than over their posterior

distribution. The main disadvantage of this approach is that potentially important information is sacrificed, particularly the outcomes of opponents' games against other opponents. This is done in order to derive a set of closed-form computations. Thus inferences from our algorithm will not be as precise as a fully Bayesian analysis. This sacrifice of precision appears necessary if posterior updates are needed on an ongoing basis.

Letting $\boldsymbol{s}$ be the collection of game outcomes during a rating period, the marginal posterior distribution of θ can be approximated as the integral of the posterior distribution of θ conditional on opponents' strengths integrated over their prior distribution,

$$f(\theta|\boldsymbol{s}) \approx \int \cdots \int f(\theta|\theta_1, \ldots, \theta_m, \boldsymbol{s}) \varphi(\theta_1|\mu_1, \sigma_1^2) \cdots \varphi(\theta_m|\mu_m, \sigma_m^2) d\theta_1 \cdots d\theta_m \quad ,$$

where $\varphi(\cdot)$ is the normal density with the given mean and variance, and

$$f(\theta|\theta_1, \ldots, \theta_m, \boldsymbol{s}) \propto \varphi(\theta|\mu, \sigma^2) L(\theta, \theta_1, \ldots, \theta_m|\boldsymbol{s}) \quad .$$

Here, $L(\theta, \theta_1, \ldots, \theta_m|\boldsymbol{s})$ is the likelihood for all parameters. As terms in the likelihood that do not depend on θ (which correspond to games played between other players) may be treated as constant with respect to θ, the marginal posterior distribution of θ can be written as

$$\begin{aligned} f(\theta|\boldsymbol{s}) &\propto \int \cdots \int \varphi(\theta|\mu, \sigma^2) L(\theta, \theta_1, \ldots, \theta_m|\boldsymbol{s}) \varphi(\theta_1|\mu_1, \sigma_1^2) \cdots \varphi(\theta_m|\mu_m, \sigma_m^2) d\theta_1 \cdots d\theta_m \\ &= \varphi(\theta|\mu, \sigma^2) \int \cdots \int L(\theta, \theta_1, \ldots, \theta_m|\boldsymbol{s}) \varphi(\theta_1|\mu_1, \sigma_1^2) \cdots \varphi(\theta_m|\mu_m, \sigma_m^2) d\theta_1 \cdots d\theta_m \\ &\propto \varphi(\theta|\mu, \sigma^2) \prod_{j=1}^{m} \left(\int \prod_{k=1}^{n_j} \frac{(10^{(\theta-\theta_j)/400})^{s_{jk}}}{1 + 10^{(\theta-\theta_j)/400}} \varphi(\theta_j|\mu_j, \sigma_j^2) d\theta_j \right) \quad . \end{aligned} \tag{8}$$

This expression may be evaluated using numerical methods, such as Monte Carlo integration. Instead, we determine a set of closed-form computations that approximate the marginal posterior density in (8). The details of the derivation are in Appendix A.

The updating algorithm approximates (8) by a normal density with mean and variance param-

eters μ' and σ'^2, respectively. The parameters are given by

$$\mu' = \mu + \frac{q}{1/\sigma^2 + 1/\delta^2} \sum_{j=1}^{m} \sum_{k=1}^{n_j} g(\sigma_j^2)(s_{jk} - \mathrm{E}(s|\mu, \mu_j, \sigma_j^2)) \tag{9}$$

$$\sigma'^2 = \left(\frac{1}{\sigma^2} + \frac{1}{\delta^2} \right)^{-1} \tag{10}$$

where

$$\begin{aligned} q &= \frac{\log 10}{400} = 0.0057565 \\ g(\sigma^2) &= \frac{1}{\sqrt{1 + 3q^2\sigma^2/\pi^2}} \\ \mathrm{E}(s|\mu, \mu_j, \sigma_j^2) &= \frac{1}{1 + 10^{-g(\sigma_j^2)(\mu - \mu_j)/400}} \\ \delta^2 &= \left(q^2 \sum_{j=1}^{m} n_j (g(\sigma_j^2))^2 \mathrm{E}(s|\mu, \mu_j, \sigma_j^2)(1 - \mathrm{E}(s|\mu, \mu_j, \sigma_j^2)) \right)^{-1} . \end{aligned}$$

These calculations are carried out in parallel for each player individually over the rating period to produce the approximating normal posterior distributions of each player's strength.

The posterior mean updating approximation in (9) has close connections to the chess rating system of Elo (1978). Elo's system was adopted by the USCF in the early 1960's, and subsequently by the World Chess Federation (FIDE) in 1970. Most national chess organizations, as well as national organizations for tournament table tennis and Scrabble, have adopted Elo's system with minor variations. If a player has an estimated strength $\tilde{\theta}$ at the onset of a rating period, then the Elo algorithm for updating a player's strength estimate from game outcomes is given by

$$\tilde{\theta}' = \tilde{\theta} + K \sum_{j=1}^{m} \sum_{k=1}^{n_j} (s_{jk} - \mathrm{We}(\tilde{\theta}, \tilde{\theta}_j)) \ , \tag{11}$$

where $\tilde{\theta}'$ is the player's posterior strength estimate, $\tilde{\theta}_j$ is the pre-period strength estimate of opponent j, K is a constant (e.g., $K = 32$ in the USCF system for amateur players), and

$$\mathrm{We}(\tilde{\theta}, \tilde{\theta}_j) = \frac{1}{1 + 10^{-(\tilde{\theta} - \tilde{\theta}_j)/400}} \tag{12}$$

is the approximate probability of defeating player j as a function of the estimates of strength.

The Elo updating algorithm is seen to be a special case of (9). If $\sigma_j^2 = 0$ for all opponents, implying that all opponents' mean strengths are known without error, then $g(\sigma_j^2) = 1$ for all j, and $\mathrm{E}(s|\mu, \mu_j, \sigma_j^2) = \mathrm{We}(\mu, \mu_j)$. Then, (9) reduces to the Elo updating formula with

$$K = \frac{q}{1/\delta^2 + 1/\sigma^2} \quad . \tag{13}$$

The computations in (9) and (10) may be preferable because they make use of the expected game outcome incorporating the uncertainty in the player's own strength and in the opponents' strengths, and the variability due to the game outcomes (represented by δ^2). The updating formula in (9) distinguishes the uncertainty in opponents' strengths by allowing $g(\cdot)$ to be less than 1; when an opponent's prior rating variance, σ_j^2, is large, then $g(\sigma_j^2)$ is substantially less than 1, and therefore the contribution of this opponent to the sum in (9) will be relatively small. Also, the fraction in (13) in effect weights the impact of game results relative to the precision of one's strength prior to competition. Thus, in contrast to the Elo updating formula, the value of K depends on the prior precision of one's strength. When one's prior strength is uncertain, game outcomes have a potentially substantial impact on one's strength distribution update, and this is reflected in (13) having a large value. When one's prior strength is precisely measured, then game outcomes should have little effect on one's strength update, and (13) is small.

3.4 Approximation accuracy

The accuracy of the approximation to the marginal posterior distribution (8) by a normal distribution with parameters given by (9) and (10) is assessed through simulation. We examine the accuracy of the updating algorithm when a player competes against 4, 10, 20 or 50 opponents in a rating period.

For a given number of opponents in a rating period, a player with prior strength distribution $\theta \sim \mathrm{N}(1500,\ 100^2)$ competes against opponents with normal prior distributions having means

	Average coverage fraction (95% intervals)	
Number of Opponents	Nominal 50% Posterior Interval	Nominal 95% Posterior Interval
4	0.4976 (0.448, 0.546)	0.9480 (0.929, 0.967)
10	0.4959 (0.449, 0.536)	0.9467 (0.924, 0.966)
20	0.4919 (0.437, 0.538)	0.9452 (0.922, 0.966)
50	0.4873 (0.433, 0.538)	0.9413 (0.911, 0.963)

Table 1: The averages and 95% central intervals of the proportion of weighted bootstrap simulated values that fall within nominal 50% and 95% posterior intervals constructed from the approximating normal distribution. The coverage fractions and intervals were computed based on 500 simulated data sets.

drawn from N(1500, 100^2) and standard deviations drawn from a scaled inverse-χ^2 distribution on 10 degrees of freedom with mean 50. Game outcomes were determined by simulating strengths from the prior distributions, and then simulating binary outcomes given the strengths. From a simulated collection of game outcomes and opponents' prior strength distributions, the computations in (9) and (10) were carried out. Additionally, 500 values of the marginal posterior distribution of θ given in (8) were simulated via Monte Carlo integration using the weighted bootstrap (Smith and Gelfand 1992). This was accomplished by generating 10000 random draws of θ from the player's prior distribution, determining the corresponding (approximate) marginal likelihoods evaluated at each of the draws (which involved averaging Bradley-Terry probabilities over 50 random draws from each opponent's prior distribution), and then drawing 500 values of θ without replacement from the original 10000 with probabilities proportional to the marginal likelihoods. The 500 values were summarized by means and standard deviations, and also by the proportion of values that lay within nominal 50% and 95% normal posterior intervals based on the approximating mean and standard deviation from (9) and (10). This entire process of calculating posterior summaries by the approximating method and the weighted bootstrap for each collection of simulated parameters and game outcomes was repeated 500 times for each fixed number of opponents.

The results of the analyses are displayed in Table 1. The entries of the table summarize the

results of the 500 simulated data sets per number of opponents. On average, for 4, 10, 20 or 50 opponents in one rating period, approximately 50% and 95% of weighted bootstrap draws fall within 50% and 95% nominal posterior intervals constructed from the approximating normal distribution using the updating formulas. This suggests that, on average, the approximating normal distribution produced by the updating algorithm closely approximates the distribution of values determined from the weighted bootstrap. Because the entries in the table are slightly less, on average, than the nominal coverage fractions, the updating algorithm produces mean and variance estimates that slightly overstate the precision of the marginal posterior distribution. One possible reason is that the updating algorithm produces posterior standard deviations that do not make full use of the data, resulting in variance estimates that are too small when a player has extreme results (relative to the prior distribution), and too large when a player's results are very consistent with the prior distribution. See Appendix A for computational details. This approximation becomes slightly amplified when more opponents are involved in the computations. It does appear, however, that the approximating algorithm produces estimates close enough to the Monte Carlo approach for practical purposes.

3.5 Parameter smoothing

The algorithm developed in Section 3.3 results in approximate posterior distributions of each player's strength parameter at the end of each rating period conditional only on preceding game outcomes. Thus, the algorithm makes efficient use of the data for estimating players' current strengths, but makes poor use of the data for estimating strengths of players during early rating periods. A smoothing procedure is now described to obtain approximate posterior distributions of strength parameters about earlier rating periods conditional on both previous and future data. This algorithm is the standard "backward filtering" of the Kalman filter to obtain smoothed estimates of past parameters.

Let $t = 1, 2, \ldots, T$ sequentially index each of T equally spaced rating periods. Denote the collection of all available game outcomes in all rating periods before and including period t by $\boldsymbol{s}^{(t)}$. Carrying out the (forward) rating algorithm, we obtain approximate normal posterior distributions of $\theta_i^{(t)}|\boldsymbol{s}^{(t)} \sim \mathrm{N}(\mu_i^{(t)}, \sigma_i^{2\,(t)})$, for $t = 1, \ldots, T$ and for all players i, where $\theta_i^{(t)}$ denotes the strength parameter for player i during rating period t. The posterior distribution of $\theta_i^{(T-1)}|\boldsymbol{s}^{(T)}$ can be determined by noting that

$$f(\theta_i^{(T-1)}|\boldsymbol{s}^{(T)}) \propto \int f(\theta_i^{(T-1)}|\boldsymbol{s}^{(T-1)}) f(\theta_i^{(T)}|\theta_i^{(T-1)}, \nu^2) f(\theta_i^{(T)}|\boldsymbol{s}^{(T)}) d\theta_i^{(T)}. \tag{14}$$

Because all of the densities in the integrand are assumed normal, the posterior distribution of $\theta_i^{(T-1)}$ is also normal, that is, $\theta_i^{(T-1)}|\boldsymbol{s}^{(T)} \sim \mathrm{N}(M, V)$ where

$$\begin{aligned} V &= \left(\frac{1}{\sigma_i^{2\,(T-1)}} + \frac{1}{\nu^2 + \sigma_i^{2\,(T)}} \right)^{-1} \\ M &= V \left(\frac{\mu_i^{(T-1)}}{\sigma_i^{2\,(T-1)}} + \frac{\mu_i^{(T)}}{\nu^2 + \sigma_i^{2\,(T)}} \right). \end{aligned}$$

This procedure is then applied recursively to $t = T-2, T-3, \ldots, 1$ to obtain all the posterior distributions of strength parameters for player i conditional on all of the data. The same procedure is then applied to each player.

4 Model fitting

The algorithm in the preceding sections depends on the knowledge of σ_0^2, the variance describing the initial uncertainty in players' abilities, and ν^2, the variance increase per unit time. This section describes a procedure to estimate these parameters.

Let $t = 1, 2, \ldots, T$ index each of T rating periods, $\boldsymbol{s}^{(t)}$ represent the collection of all available game outcomes in all rating periods before and including period t, and $\theta_i^{(t)}$ and $\theta_j^{(t)}$ denote the strength parameters for players i and j during period t. For a given ordered pair of variances, (σ_0^2, ν^2), the rating algorithm provides the computations to obtain the approximate distributions of

$\theta_i^{(t)}|\boldsymbol{s}^{(t-1)} \sim \mathrm{N}(\mu_i, \sigma_i^2)$ and $\theta_j^{(t)}|\boldsymbol{s}^{(t-1)} \sim \mathrm{N}(\mu_j, \sigma_j^2)$. We may now define the "predictive discrepancy" for a game between i and j during period t as a measure of discrepancy between the predicted result for the game, given only information prior to period t, to the actual result of the game. In particular, let

$$d_{ij} = -s_{ij} \log p_{ij} - (1 - s_{ij}) \log(1 - p_{ij}) \tag{15}$$

be the discrepancy of a game played between i and j, where s_{ij} is the outcome of the game, and

$$p_{ij} = \frac{1}{1 + 10^{-g(\sigma_i^2 + \sigma_j^2)(\mu_i - \mu_j)/400}} \quad . \tag{16}$$

The expression in (16) is an approximation to the probability that i defeats j incorporating the variability of each player's strength estimate, and is derived analogously to the expected outcome integrated over an opponent's prior strength distribution. Thus the predictive discrepancy, d_{ij}, in (15) is the binomial log-likelihood for a game evaluated at a value that only depends on previous data.

The total discrepancy for the entire collection of game outcomes is computed as the sum of the d_{ij} over all games in the data set. This is accomplished by first summing the game discrepancies in the first rating period, then updating the strength distributions based on the game outcomes from the first period followed by updating the distributions due to the passage of time, then summing the game discrepancies in the second rating period given the updated strength distributions, and so on. Because the game discrepancies are cross-validatory, minimizing the total discrepancy is unlikely to result in an overfitted choice of variance parameters. Loss functions that involve measuring a discrepancy between predicted and actual outcomes, where the predicted outcome depends on all of the data, may result in excessively large estimates of ν^2 in order to allow a player's strengths to have unlimited variation from one rating period to the next.

A number of numerical methods for minimizing the total discrepancy as a function of σ^2 and ν^2 are feasible. However, because the total discrepancy is a complicated function of σ^2 and ν^2,

methods involving closed-form derivatives are not possible. To minimize the total discrepancy, the Nelder-Mead simplex algorithm (Nelder and Mead, 1965) was employed. This algorithm uses a direct search approach rather than numerically computing gradient information.

5 Evaluation of the estimation algorithm

To evaluate the variance estimation procedure, the algorithm was applied to simulated data under three different sets of assumed parameters and conditions:

Simulation 1. Ten players competing over 30 rating periods playing a total of 50 games per period, with $\sigma_0 = 200$, $\nu = 50$,

Simulation 2. Ten players competing over 120 rating periods playing a total of 50 games per period, with $\sigma_0 = 200$, $\nu = 50$, and

Simulation 3. Twenty players competing over 50 rating periods playing a total of 200 games per period, with $\sigma_0 = 200$, $\nu = 10$.

For each of the three simulation conditions, the fictitious competitors would compete amongst each other at random for the specified number of rating periods, over which their strengths would change by the addition of $\mathrm{N}(0, \nu^2)$ at the beginning of each successive rating period. This procedure was repeated 200 times for each of the three simulation conditions. From the results of each simulated data set, estimates of σ_0 and ν were obtained using the approach in Section 4, and then used to compute ratings from the final rating period.

A summary of the analyses is shown in Table 2. The estimates of σ_0 and ν were calculated using the simplex method, and the averages of these estimates over the 200 replications are shown in the third and fourth columns of Table 2. The estimation procedure seems to produce reasonably

	Parameters		Estimates		Coverage Fraction	
Simulation	σ_0	ν	$\hat{\sigma_0}$	$\hat{\nu}$	50%	95%
1	200	50	224.04	44.98	0.483	0.940
2	200	50	240.10	44.64	0.446	0.912
3	200	10	252.63	9.47	0.505	0.947

Table 2: Results of simulation analyses. Game outcomes were generated under differing model and data assumptions, and the table reports summaries over 200 replications. The first two columns show the assumed variances that generated the simulated data. The estimates of the two variance parameters, which are shown in the third and fourth columns, are the averages over the 200 replications. The fifth and sixth columns show the proportion of 50% and 95% central posterior intervals that contain players' true mean strengths for the last rating period.

close estimates to the underlying parameters. Sampling variability would explain some of the discrepancy. The slightly large estimates of σ_0 reflects that its posterior distribution is skewed. This can be seen in Figure 1, which shows the approximate joint posterior distribution of the two standard deviation parameters. These plots demonstrate that the posterior distribution of the standard deviation parameters are centered near the generating parameters. The large posterior variability of σ_0 reflects the small number of players in the simulations. Not surprisingly, the posterior variability of ν is smaller when the number of rating periods is larger.

The last two columns of Table 2 show how well the estimated strengths from the last rating period correspond with the true parameter values. For each set of simulations, normal posterior distributions at the end of the last rating period are obtained for all players using the estimated values of σ_0 and ν. Also, the true mean strengths during the last rating period were recorded. After centering the estimated and true mean strengths to 1500 for each of the 200 replications, we determined the proportion of the players across the 200 replications that have 50% and 95% central posterior intervals containing their true mean strength at the last rating period. These appear as the last two column in Table 2. The strengths were centered because drift in average strength cannot be detected from paired comparison data; if all players increased in strength by the same amount from one rating period to the next, it would be impossible to determine this from the data. In all cases, the actual proportion of coverage is close to the nominal level of coverage. This is

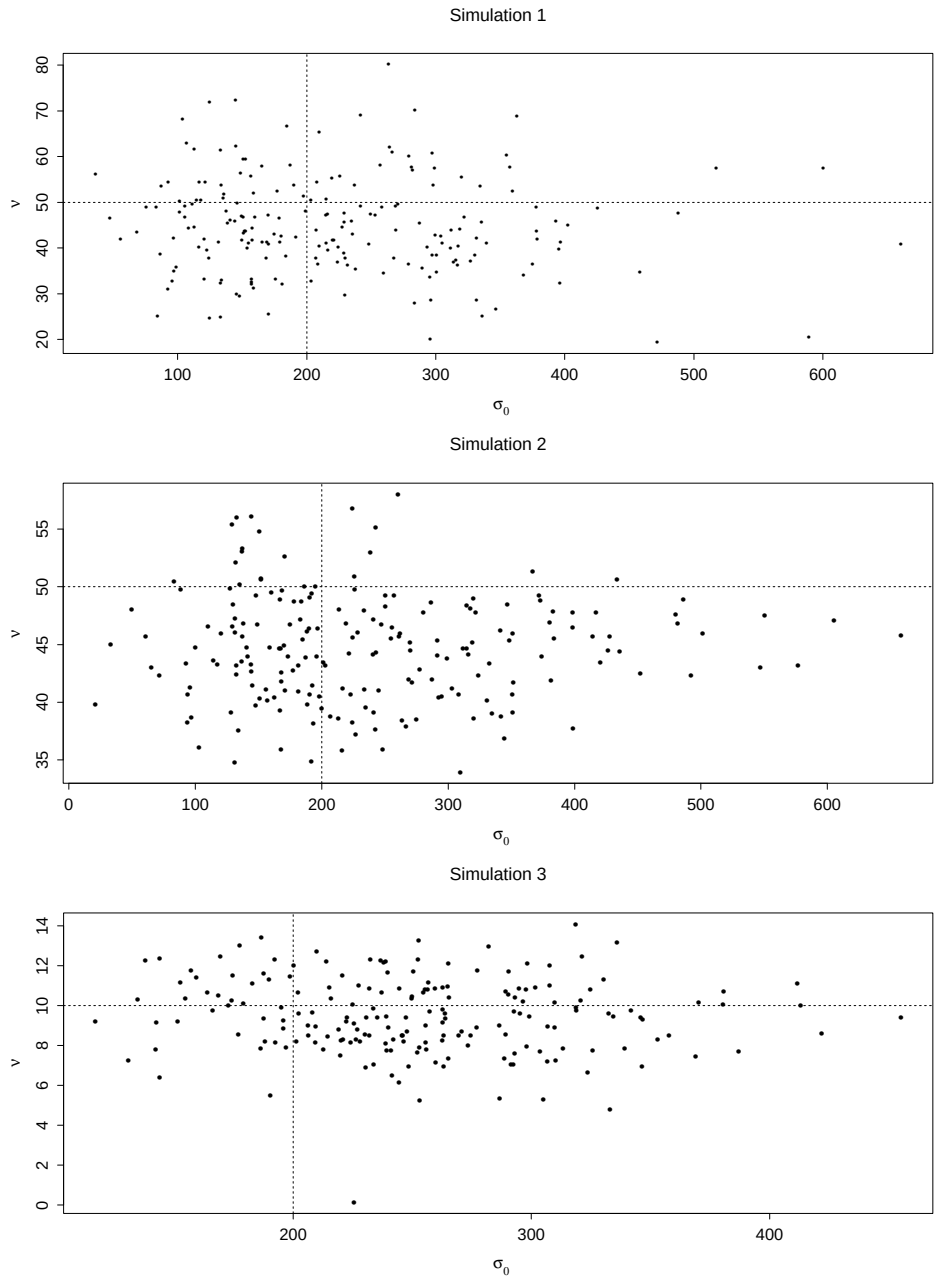

Figure 1: Joint posterior distribution of (σ_0, ν) from each of the three simulations. The vertical and horizontal dotted lines indicate the values of the generating parameters.

particularly true for the third simulation which had parameters that were more precisely estimated because the size of the simulated data set per replication was large. For the first two simulations, the actual fraction of coverage is slightly less than the reported level of coverage, indicating that a player's approximate normal posterior distribution from the updating algorithm may be overstating the precision under similar conditions. However, the coverage is still fairly close to nominal, and with a large set of data, coverage does not appear to be a problem.

6 Examples

We apply the methodology of Section 3 to the analysis of two data sets. The first data set consists of all known game results from 1857 to 1991 among 88 of the world's all-time best chess players. The second data set involves the outcomes of tennis matches played among 1190 competitors from 1986 through 1995.

6.1 The best chess players of all time

Professor Nathan Divinsky has compiled a data set consisting of all known tournament or match game results among 88 of the top chess players of all time. The data set contains 15664 game outcomes played among 1367 pairs of players. Not all $\binom{88}{2} = 3828$ pairs of players competed against each other because some players lives, much less their playing careers, did not overlap. Several analyses based on smaller versions of this data set have been published, including those by Elo (1978), Keene and Divinsky (1989), Joe (1990) and Henery (1992). Elo fitted paired comparison models separately in each year, and then smoothed the estimates. Keene and Divinsky fitted a Bradley-Terry model to the data, acting as if all the games were played in one large tournament. Joe divided players' careers into "peak" and "off-peak" periods, and fitted an extension of the Bradley-Terry model explicitly accounting for ties. Henery fitted a model that accounts for differing

frequencies of ties, but not the possibility of changes in abilities over time.

For our analysis, a single rating period is a year, so the number of rating periods is 135. In some years no game outcomes were observed (for example 1859, 1874, 1875), but this is handled by applying the computations in Section 3.2 in succession. The prior variance for a player was constrained to be σ^2 for the year of first competition. The total discrepancy was minimized for $\sigma = 38.19$ and $\nu = 18.87$. Using these values, estimates of all players' strengths for every year were computed using the rating algorithm followed by the backward smoothing procedure.

In Table 3, the top 20 players are ranked according to their peak posterior mean strength, along with the posterior standard deviation of the strength and the year in which the player attained highest strength (the peak year). The top eight players on the list are former and current world champions. The top eight players identified by Joe (1990) matches the top eight in Table 3, except Steinitz is replaced by Paul Morphy. The data set only consists of 25 games against two opponents played by Morphy between 1857–1858, so there is very little information in our analysis that would indicate such profound playing strength (Morphy is ranked 27 on our list). The ninth on the list, Akiba Rubinstein, is the highest ranked player who never held the world championship. Due to possible increase in overall playing strength (which could result from an increased understanding of chess over time), the posterior mean strengths of players at different points in time cannot be directly compared. Instead, one should interpret Table 3 as indicating the level of strength relative to current competitors. Thus, according to the data analysis, Lasker dominated his opponents around 1916 more so than Fischer did in 1972.

Figure 2 shows the strengths of the top eight players plotted over time. Typically, at the beginning of their careers, they are moderate in strength, with a gradual climb toward their peak strength, and then a gradual decline. The most obvious exception is Fischer, who gave up competitive chess in 1972 after winning the world championship. The plot also suggests that Kasparov

Rank	Player	Posterior Mean Strength in Peak Year	Posterior Standard Deviation	Peak Year
1.	EMANUEL LASKER	1693	29	1916
2.	JOSE CAPABLANCA	1680	28	1921
3.	ROBERT FISCHER	1656	38	1972
4.	ALEXANDER ALEKHINE	1647	24	1930
5.	GARRY KASPAROV	1643	32	1991
6.	MIKHAIL BOTVINNIK	1623	27	1947
7.	ANATOLY KARPOV	1609	20	1984
8.	WILHELM STEINITZ	1608	29	1876
9.	AKIBA RUBINSTEIN	1584	24	1912
10.	MAX EUWE	1579	23	1935
11.	BORIS SPASSKY	1578	21	1968
12.	SIEGBERT TARRASCH	1576	25	1905
13.	VIKTOR KORCHNOI	1573	21	1978
14.	GEZA MAROCZY	1572	25	1908
15.	DAVID BRONSTEIN	1571	23	1953
16.	VASSILY IVANCHUK	1570	32	1991
17.	SAMUEL RESHEVSKY	1569	24	1952
18.	VASSILY SMYSLOV	1567	22	1954
19.	ARON NIMZOVITCH	1565	26	1931
20.	TIGRAN PETROSIAN	1564	22	1963

Table 3: Twenty of the best chess players of all time ranked according to their posterior mean strength from the fitted model for the year in which the player attained highest strength, the peak year.

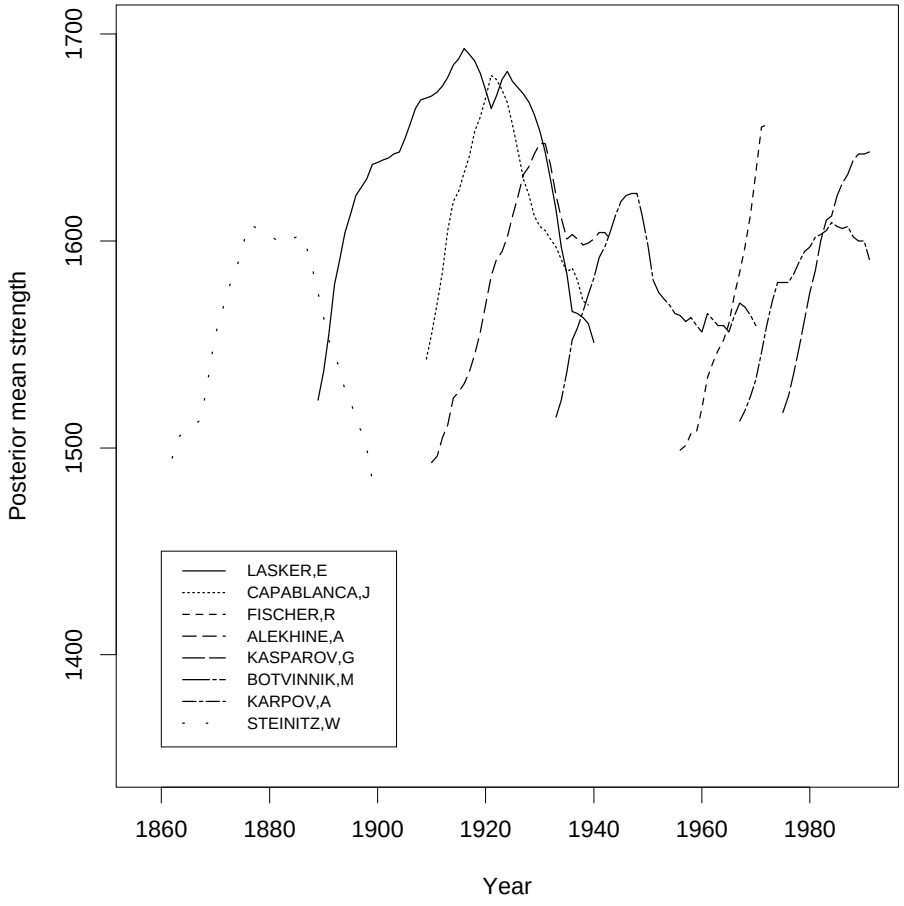

Figure 2: Smoothed posterior mean strengths of eight players over time.

may have still been on the rise in 1991, and that Karpov has been gradually declining in strength since his peak in the mid-1980s. Both Alekhine and Botvinnik had careers that ended without a steep decline.

6.2 Ranking current tennis players

The methodology of Section 3 was applied to rank the world's current top tennis players. The data set consisted of all 33359 matches played among 1190 male participants from 1986 to the end of 1995 in events on the ATP Tour. The ATP Tour is generally regarded as the main international organizer of men's professional tennis. Games were excluded from the analysis if they were played in events that did not award ATP "points," for example the Davis Cup and some ATP Tour championships. Rating periods were designated to be two months long, resulting in a total of 60 rating periods. As in the analysis of chess game outcomes, the prior variance for a player was assumed to be σ^2 during the rating period in which he first competed.

Minimizing the total discrepancy for these data yielded values of $\sigma = 113.65$ and $\nu = 22.35$. By contrast to the chess outcomes analysis, the spread of players' initial strengths is much larger, indicating a wider variation of playing strengths among the tennis competitors. Because the chess players were selected to be (roughly) the 88 best in the world over all time, whereas the 1190 tennis players were likely to be the best only in a 10 year period, the greater variability in the tennis players' strengths is not surprising. The variance increase per year was much smaller for the chess player data than for tennis, suggesting more stability in chess playing strength than in tennis. This seems intuitively sensible because the factors that result in a tennis player performing well (e.g., physical condition, responsiveness, lack of injuries) may be more variable over time than factors that result in a chess player performing well (judgment, insight, ability to calculate variations).

Table 4 displays the posterior means and standard deviations of the strengths for the top 20

Rank	Player	Posterior Mean Strength	Posterior Standard Deviation	ATP Tour Rank End of 1995
1.	ANDRE AGASSI	1992	53	2
2.	PETE SAMPRAS	1987	51	1
3.	THOMAS MUSTER	1892	46	3
4.	MICHAEL CHANG	1885	50	5
5.	BORIS BECKER	1860	51	4
6.	JIM COURIER	1841	48	8
7.	MICHAEL STICH	1804	52	12
8.	YEVGENY KAFELNIKOV	1790	46	6
9.	THOMAS ENQVIST	1780	46	7
10.	WAYNE FERREIRA	1776	47	9
11.	TODD MARTIN	1767	50	18
12.	MAGNUS LARSSON	1764	57	17
13.	SERGI BRUGUERA	1764	52	13
14.	GORAN IVANISEVIC	1757	51	10
15.	STEFAN EDBERG	1752	53	23
16.	RICHARD KRAJICEK	1747	50	11
17.	MARC ROSSET	1720	47	15
18.	ARNAUD BOETSCH	1704	43	14
19.	ANDREI MEDVEDEV	1702	51	16
20.	MALIVAI WASHINGTON	1695	48	26

Table 4: Twenty of the current ATP Tour tennis participants ranked according to their posterior mean strengths at the end of 1995 from the fitted model.

players ranked according to their posterior means. Players were included in the list if they had played within four rating periods (eight months) of the end of 1995. The posterior means for these top players have greater spread than the top chess players, though because the standard deviations are also larger it is more difficult to assert differences in playing strength. Agassi and Sampras have posterior means about 100 higher than all other competitors. Posterior probabilities of game outcomes can be computed using formula (16). According to the parameter estimates, Sampras, ranked second, would defeat Muster, ranked third, with a posterior probability of 0.63.

The last column of the table shows the official ATP Tour rankings of the players at the end of 1995, based on players accruing points depending on their successes in their best 14 tournaments over the previous 12 months. The players are then ranked according to the sum of their points.

While this system does not use a probabilistic model to measure players' strengths, it does produce a ranking list that conforms to generally held perceptions. The top 18 players in the ATP Tour rankings appear in Table 4. The other two players are Stefan Edberg, who is ranked 15 on our list, and Malivai Washington, who is ranked 20. Edberg appears high on our list because he was inferred to be much stronger in the early 1990s, and that the model fit purports that he did not decline in ability as much as the ATP ranking system indicates. Washington, interestingly, had the newsworthy result of competing in the finals at Wimbledon in 1996, so that he may have been stronger, as our estimate shows, than indicated by the ATP rank.

7 Discussion

The problem of developing a rating system for large paired comparison experiments with time-varying abilities involves a tradeoff between making full use of outcome information and keeping a system simple enough to be used on a regular basis. The likelihood-based state-space approach provides inferences that are consistent with the model assumptions, but the computational complexity can be enormous. The algorithm developed in this paper attempts to combine desirable features of each approach to produce a system that is both usable and accurate for time-dependent paired comparison situations.

While computationally straightforward, the algorithm ignores certain features of the model that a likelihood-based analysis would recognize. For example, information about the results of opponents' games is sacrificed to aid the computational ease of the algorithm. More precision could be gained in a likelihood-based analysis because all players' results would have an impact on inferences. A related issue is that posterior correlations of players' strengths are not accounted for in our algorithm. If a player competes against one particular opponent more frequently than against others, then a strong correlation may be induced. Then both strength distributions should

be affected when either player competes. On the other hand, the benefit in excluding correlations is that the number of model parameters is greatly reduced. This is quite desirable when the number of players in the population is large. In any event, the simulations in this paper seem to show the approximating algorithm performs reasonably well, and the application to the chess and tennis game data produce player rankings that match external criteria.

Acknowledgements

The author is grateful to Miki Singh and Harry Joe for providing the tennis data and chess data, respectively, and to Christopher Avery for his help preparing the tennis data.

A Derivation of closed-form computations

We derive a closed-form normal approximation to the approximate posterior distribution of a player's strength given in equation (8). The derivation of the formulas given in (9) and (10) can be summarized in three steps:

- Approximate the likelihood, marginalized over the opponents' prior strength distribution, by a product of logistic cumulative distribution functions.
- Approximate the resulting expression by a normal distribution.
- Construct a linear approximation using a Taylor series expansion around the prior mean to determine the mean and variance of the approximating normal posterior distribution.

The likelihood, integrated over the distribution of the opponents' prior strength distribution, is given by

$$L(\theta|\boldsymbol{s}) \quad = \quad \prod_{j=1}^{m} \left(\int \prod_{k=1}^{n_j} \frac{(10^{(\theta-\theta_j)/400})^{s_{jk}}}{1+10^{(\theta-\theta_j)/400}} \varphi(\theta_j|\mu_j, \sigma_j^2) d\theta_j \right)$$

$$\approx \prod_{j=1}^{m}\prod_{k=1}^{n_j}\int \frac{(10^{(\theta-\theta_j)/400})^{s_{jk}}}{1+10^{(\theta-\theta_j)/400}}\varphi(\theta_j|\mu_j,\sigma_j^2)\,d\theta_j \quad . \tag{17}$$

Under the actual model, an opponent plays at a fixed strength (which can be viewed as a single value of θ_j drawn from the prior distribution) for all games in a rating period. This last approximation is justified by allowing the possibility that an opponent is able to display different strengths for different games, the strengths being drawn independently from the opponent's prior distribution.

An integral in the above product could be approximated easily using numerical methods. Instead, we approximate the integrals in (17), which are logistic cumulative distribution functions (cdfs) integrated over normal densities, by rescaled logistic cdfs. This is accomplished by first approximating each logistic cdf in an integrand by a normal cdf with the same mean and variance so that the integral can be evaluated in closed form to a normal cdf. The resulting normal cdf is then converted back to a logistic cdf with the same mean and variance. This yields the approximation

$$\int \frac{(10^{(\theta-\theta_j)/400})^{s_{jk}}}{1+10^{(\theta-\theta_j)/400}}\varphi(\theta_j|\mu_j,\sigma_j^2)\,d\theta_j \approx \frac{(10^{g(\sigma_j^2)(\theta-\mu_j)/400})^{s_{jk}}}{1+10^{g(\sigma_j^2)(\theta-\mu_j)/400}} \quad , \tag{18}$$

where

$$g(\sigma^2) = \frac{1}{\sqrt{1+3q^2\sigma^2/\pi^2}} \tag{19}$$

and

$$q = \frac{\log 10}{400} \quad .$$

Approximating integrals of this type, which commonly arise in logistic regression models with random effects, has been addressed similarly in Aitchinson and Begg (1976), Lauder (1978), and Boys and Dunsmore (1987). The (approximate) marginal likelihood, therefore, may now be written as

$$L(\theta|\boldsymbol{s}) \quad = \quad \prod_{j=1}^{m}\prod_{k=1}^{n_j}\frac{(10^{g(\sigma_j^2)(\theta-\mu_j)/400})^{s_{jk}}}{1+10^{g(\sigma_j^2)(\theta-\mu_j)/400}} \quad .$$

We now obtain a normal approximation to this marginal likelihood. To do so, we first find an expression for the mode by setting the derivative of the log-marginal likelihood equal to 0. Note

that

$$\log L(\theta|\boldsymbol{s}) = \sum_{j=1}^{m}\sum_{k=1}^{n_j}(qg(\sigma_j^2)s_{jk}(\theta-\mu_j) - \log(1+10^{g(\sigma_j^2)(\theta-\mu_j)/400}))$$

so that

$$\frac{\partial \log L(\theta|\boldsymbol{s})}{\partial\theta} = q\sum_{j=1}^{m}\sum_{k=1}^{n_j} g(\sigma_j^2)(s_{jk} - \frac{1}{1+10^{-g(\sigma_j^2)(\theta-\mu_j)/400}}) \quad .$$

Let

$$\mathrm{E}(s|\theta,\mu_j,\sigma_j^2) = \frac{1}{1+10^{-g(\sigma_j^2)(\theta-\mu_j)/400}} \quad , \tag{20}$$

which is the approximate expected outcome of a game against opponent j, incorporating the uncertainty contained in the prior distribution of this opponent's strength. At the mode of the marginal likelihood, $\hat{\theta}$, we therefore have

$$\sum_{j=1}^{m}\sum_{k=1}^{n_j} g(\sigma_j^2)\left(s_{jk} - \mathrm{E}(s|\hat{\theta},\mu_j,\sigma_j^2)\right) = 0 \quad . \tag{21}$$

The second derivative of the log-marginal likelihood, evaluated at $\hat{\theta}$, is given by

$$\begin{aligned}\left.\frac{\partial^2 \log L(\theta|\boldsymbol{s})}{\partial\theta^2}\right|_{\theta=\hat{\theta}} &= -q^2\sum_{j=1}^{m}\sum_{k=1}^{n_j}(g(\sigma_j^2))^2\mathrm{E}(s|\hat{\theta},\mu_j,\sigma_j^2)(1-\mathrm{E}(s|\hat{\theta},\mu_j,\sigma_j^2)) \\ &= -q^2\sum_{j=1}^{m} n_j(g(\sigma_j^2))^2\mathrm{E}(s|\hat{\theta},\mu_j,\sigma_j^2)(1-\mathrm{E}(s|\hat{\theta},\mu_j,\sigma_j^2)) \quad . \end{aligned} \tag{22}$$

The marginal likelihood can be approximated by a normal density with a mean equal to $\hat{\theta}$ and a variance which is the negative reciprocal of the expression in (22). Let δ^2 be the variance associated with the normal approximation to the marginal likelihood. We can effectively approximate δ^2 by substituting μ, the prior mean, for $\hat{\theta}$,

$$\delta^2 \approx \left(q^2\sum_{j=1}^{m} n_j(g(\sigma_j^2))^2\mathrm{E}(s|\mu,\mu_j,\sigma_j^2)(1-\mathrm{E}(s|\mu,\mu_j,\sigma_j^2))\right)^{-1} \quad , \tag{23}$$

because the binomial-type variance in (22) is roughly constant over a wide range of values of θ.

The approximate marginal posterior distribution for θ can now be expressed as proportional to a product of two normal densities,

$$f(\theta|\boldsymbol{s}) \propto \varphi(\theta|\mu,\sigma^2)\varphi(\theta|\hat{\theta},\delta^2) \quad . \tag{24}$$

The posterior mean, μ', and posterior variance, σ'^2, can therefore be expressed as

$$\begin{aligned} \sigma'^2 &= \left(\frac{1}{\sigma^2} + \frac{1}{\delta^2}\right)^{-1} && (25) \\ \mu' &= \sigma'^2 \left(\frac{\mu}{\sigma^2} + \frac{\hat{\theta}}{\delta^2}\right) \\ &= \mu + \frac{1/\delta^2}{1/\sigma^2 + 1/\delta^2}(\hat{\theta} - \mu) \quad . && (26) \end{aligned}$$

Rather than calculate $\hat{\theta}$ directly using an iterative numerical procedure, we use a linear closed-form approximation to $(\hat{\theta} - \mu)$. Define

$$h(\theta) = \sum_{j=1}^{m} \sum_{k=1}^{n_j} \frac{g(\sigma_j^2)}{1 + 10^{-g(\sigma_j^2)(\theta - \mu_j)/400}} \quad . \tag{27}$$

From (21), we now have that

$$h(\hat{\theta}) = \sum_{j=1}^{m} \sum_{k=1}^{n_j} g(\sigma_j^2) s_{jk} \quad . \tag{28}$$

A Taylor series expansion of $h(\theta)$ around μ yields

$$h(\hat{\theta}) \approx h(\mu) + (\hat{\theta} - \mu) h'(\mu) \tag{29}$$

where

$$h'(\mu) = q \sum_{j=1}^{m} \sum_{k=1}^{n_j} (g(\sigma_j^2))^2 \mathrm{E}(s|\mu, \mu_j, \sigma_j^2)(1 - \mathrm{E}(s|\mu, \mu_j, \sigma_j^2)) \quad . \tag{30}$$

Substituting the linear approximation implied by (29) for $(\hat{\theta} - \mu)$, equation (26) can be rewritten as

$$\begin{aligned} \mu' &\approx \mu + \frac{1/\delta^2}{1/\sigma^2 + 1/\delta^2} \left(\frac{h(\hat{\theta}) - h(\mu)}{h'(\mu)}\right) \\ &= \mu + \frac{q}{1/\sigma^2 + 1/\delta^2} \Big(h(\hat{\theta}) - h(\mu)\Big) && (31) \\ &= \mu + \frac{q}{1/\sigma^2 + 1/\delta^2} \sum_{j=1}^{m} \sum_{k=1}^{n_j} g(\sigma_j^2)(s_{jk} - \mathrm{E}(s|\mu, \mu_j, \sigma_j^2)) \quad . && (32) \end{aligned}$$

The equality in (31) follows as $1/(\delta^2 h'(\mu)) = q$. The last equality in (32) is justified by expanding the expressions for $h(\hat{\theta})$ and $h(\mu)$. The posterior mean and variance, therefore, can be computed using closed form equations given in (25) and (32).

References

Aitchison, J., and Begg, C. B. (1976) Statistical diagnosis when basic cases are not classified with certainty. *Biometrika*, **63**, 1–12.

Boys, R. J., and Dunsmore, I. R. (1987) Diagnostic and sampling models in screening. *Biometrika*, **74**, 365–374.

Bradley R. A. (1984) Paired comparisons: Some basic procedures and examples. In *Handbook of Statistics* 4, eds. P. R. Kirshnajah and P. K. Sen, Amsterdam: Elsevier, pp. 299–326.

Bradley R. A., and Terry, M. E. (1952) The Rank Analysis of Incomplete Block Designs. 1. The Method of Paired Comparisons. *Biometrika*, **39**, 324–45.

David, H. A. (1988) *The method of paired comparisons* (2nd ed.), London: Chapman and Hall.

Davidson, R. R. (1970) On Extending the Bradley-Terry Model to Accommodate Ties in Paired Comparison Experiments. *J. Am. Statist. Ass.*, **65**, 317–28.

Davidson, R. R. and Beaver, R. J. (1977) On Extending the Bradley-Terry Model to Incorporate Within-Pair Order Effects. *Biometrics*, **33**, 693–702.

Elo, A. E. (1978) *The rating of chess players past and present*, New York: Arco Publishing.

Fahrmeir, L. and Tutz, G. (1994) Dynamic stochastic models for time-dependent ordered paired comparison systems. *J. Am. Statist. Ass.*, **89**, 1438–1449.

Glickman, M. E. (1993) *Paired comparison models with time-varying parameters.* PhD Dissertation,

Harvard University Department of Statistics, Cambridge, USA.

Henery, R. J. (1992) An extension to the Thurstone-Mosteller model for chess. *The Statistician*, **41**, 559–567.

Joe, H. (1990) Extended Use of Paired Comparison Models, with Application to Chess Rankings. *Appl. Statist.*, **39**, 85–93.

Keene, R. and Divinsky, N. (1989) *Warriors of the mind: A quest for the supreme genius of the chess board.*" Brighton (England): Hardinge Simpole.

Lauder, I. J. (1978) Computational problems in predictive diagnosis. Proceedings of COMPSTAT, Third Symposium on Computational Statistics, Leiden, Netherlands, 186–192.

Nelder, J. A. and Mead, R. (1965) A simplex method for function minimization. *The Computer Journal*, **7**, 308–313.

Rao, P. V. and Kupper, L. L. (1967) Ties in Paired-Comparison Experiments: A Generalization of the Bradley-Terry Model. *J. Am. Statist. Ass.*, **62**, 194–204.

Smith, A. F. M. and Gelfand, A. E. (1992) Bayesian statistics without tears: A sampling-resampling perspective, *Am. Statist.*, **46**, 84–88.