Adaptive paired comparison design

Mark E. Glickman*
Department of Mathematics and Statistics
Boston University

Shane T. Jensen
Department of Statistics
Harvard University

Abstract

An important aspect of paired comparison experiments is the decision of how to form pairs in advance of collecting data. A weakness of typical paired comparison experimental designs is the difficulty in incorporating prior information, which can be particularly relevant for the design of tournament schedules for players of games and sports. Pairing methods that make use of prior information are often ad hoc algorithms with little or no formal basis. The problem of pairing objects can be formalized as a Bayesian optimal design. Assuming a linear paired comparison model for outcomes, we develop a pairing method that maximizes the expected gain in Kullback-Leibler information from the prior to the posterior distribution. The optimal pairing is determined using a combinatorial optimization method commonly used in graph-theoretic contexts. We discuss the properties of our optimal pairing criterion, and demonstrate our method as an adaptive procedure for pairing objects multiple times. We compare the performance of our method on simulated data against random pairings, and against a system that is currently in use in tournament chess.

Keywords: Bayesian optimal design, Bradley-Terry model, competition, Swiss system, tournament schedule.

*Address for correspondence: Center for Health Quality, Outcomes & Economics Research, Edith Nourse Rogers Memorial Hospital (152), Bldg 70, 200 Springs Road, Bedford, MA 01730, USA. E-mail address: mg@bu.edu. Phone: (781) 687-2875, Fax: (781) 687-3106. The authors wish to thank Byron Ellis, Deian Palejev, and John Hartigan for their helpful comments.

1 Introduction

An important aspect of paired comparison experiments, in which objects are compared in pairs to evaluate their relative merits, is the choice of comparisons to be made. An overview of paired comparison design methods is discussed by David (1988, ch. 5) who outlines pre-1990 work on complete and incomplete designs, including both balanced and partially balanced designs. One feature of certain types of paired comparison experiments is that prior knowledge may be available about the merits of the objects. This knowledge may come, for example, from the outcomes of previous comparisons, or from prior beliefs about the objects' merits. A paired comparison situation where prior knowledge commonly exists is in the design of tournament or league schedules for games and sports competition. At the onset of a sports league, teams may be ranked through the results of previous competition or expert judgment which would lead to the choice of a league schedule. In tournaments for games such as chess and go, as well as many other games, game outcomes are used to construct player ratings which are then used as a basis to pair players. We describe in this paper a method for designing paired comparison experiments, particular to tournament scheduling, that incorporates prior information in a principled manner and which results in efficient inferences.

Most work on paired comparison designs that are used for tournament scheduling typically involve producing designs to infer the player/object with the highest merit. Two common designs are the knockout or elimination tournament, which has been studied by Hartigan (1968) and Hwang et al. (1991) among others, and the round-robin tournament, which has been examined by Daniels (1969). A more commonly implemented design is where each object is compared a number of times without being eliminated, and that the number of comparisons per object is typically far fewer than the number of objects. One such design is the "Swiss system," variants of which are in use in many gaming organizations.

A description of the Swiss system and its variants can be found in Kažić (1980, chapters 6–10). This approach is based on a reasonable though ad hoc idea that it is preferable to pair players who have similar cumulative results during a tournament. Among players with similar cumulative results, the Swiss system pairs players whose estimated a priori strengths are as different as possible, thereby avoiding having the best players compete early in a tournament.

Our approach, which can be viewed as an alternative to the Swiss system pairing method, relies on determining a set of pairings by maximizing the expected Kullback-Leibler distance between the prior and posterior densities for merit parameters. We formulate the problem as one of Bayesian optimal design in the sense of Lindley (1972, pp. 19–20), and describe how our method can be applied adaptively to multi-round tournaments. While the application of our method is general to paired comparison experiments where several sets of comparisons are to take place in sequence, we adopt a terminology that is specific to a chess tournament setting where the objects to be compared are players, and the merits are players' strengths. We describe the general framework for our approach, including the statistical model and optimality criterion, in Section 2. In Section 3, we examine how our framework can be applied to the Bradley-Terry model (Bradley and Terry, 1952) for paired comparisons. We demonstrate the application of our method to simulated data in Section 4 where we compare our approach to a variant of the Swiss system.

2 A method for optimal tournament design

Suppose that we want to make inferences about the playing strengths of N competitors, N assumed even, based on the results of games among the players. More generally, we may be interested in making inferences about the merits of N objects that are to be compared in

pairs. Rather than consider tournaments/designs where each player competes against every other (i.e., a round-robin tournament), or where only winners of games compete against other winners and losers are eliminated (i.e., a knockout or elimination tournament), we consider tournament designs where each player competes against r opponents (possibly multiple times against the same opponent), where r is typically much less than N. Each round consists of the N players distributed into $N/2$ pairs. The Swiss system is an example of such a design. Furthermore, rather than determining the r opponents per player in advance, we seek to use information from the game results after each round to pair players in the subsequent round.

The approach we develop is intended to be applied adaptively. After each round of a tournament, game outcome information is summarized as a prior distribution for the next round, and, incorporating player-pairing restrictions (such as players cannot compete against the same opponent more than once), the selection of the next set of pairings is treated no differently than pairing players at the first round of the tournament.

2.1 A utility function for pairings

In a given round of a tournament, suppose players i_k and j_k are to compete in game k, $k = 1, \ldots, N/2$. Let $Y_{i_k j_k}$ be 1 if player i_k defeats player j_k, and 0 if player j_k defeats i_k. For the development of our method, we assume no ties or other partial preferences. Then we assume that the probability of a game outcome is given by

$$p_k = \mathrm{P}(Y_{i_k j_k} = 1 \mid \theta_{i_k}, \theta_{j_k}) = F(\theta_{i_k} - \theta_{j_k}) \tag{1}$$

where F is a specified probability distribution function monotonically increasing in its argument. This model, the linear paired comparison model, assumes that win probabilities are functions of player strengths only through their difference. Two common special cases of this model include the Bradley-Terry model (Bradley and Terry, 1952) when F is a stan-

dard logistic distribution function, and the Thurstone-Mosteller model (Thurstone, 1927; Mosteller, 1951) when F is a standard normal distribution function. For the development of our optimality criterion, it is not necessary to commit to a particular choice of linear paired comparison model, though we do assume that the F takes support over the entire real line, that larger values of θ_i denote greater probability of winning, and that the values of the θ_i are unrestricted.

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$ be the collection of N player strength parameters. Prior to competition, we assume that $\boldsymbol{\theta}$ has a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. Because the θ_i are unrestricted over the real numbers, using a multivariate normal distribution is a natural approach to describing prior beliefs about players' strengths. Before players compete, we would likely assume a multivariate normal prior density that factors into independent scalar densities for each player.

We adopt Lindley's (1956) framework for Bayesian optimal design by selecting the set of pairings that maximizes the expected Kullback-Leibler distance between the prior and posterior distributions for $\boldsymbol{\theta}$. As noted by Chaloner and Verdinelli (1995), using the Kullback-Leibler distance between prior and posterior densities as a utility function in a design context is appropriate when the goal is to make efficient inferences on the model parameters.

To be particular to our problem, let s be a specific set of $N/2$ pairings among the N players in the tournament. Henceforth, we refer to a set of pairings as a "design." Let $\mathcal{S}$ be the space of all designs for the N players, where the cardinality of $\mathcal{S}$ is at most $N!/2^{N/2}$ if there are no restrictions on pairings. For a specific design s, let $\mathcal{Y}_s$ be the collection of $2^{N/2}$ binary vectors $\boldsymbol{y}$ of potentially observable game outcomes. Define the utility of a design, U, to be the expected Kullback-Leibler information between the prior and posterior densities

of $\boldsymbol{\theta}$,

$$U(s) = \int \sum_{\boldsymbol{y} \in \mathcal{Y}_s} \ln \left\{ \frac{p(\boldsymbol{\theta} \mid \boldsymbol{y})}{p(\boldsymbol{\theta})} \right\} p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{2}$$

where $p(\boldsymbol{\theta})$ is the multivariate normal prior density for the strength parameters, $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ is the posterior density given game outcomes $\boldsymbol{y}$, and $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ is the joint distribution for the game outcomes $\boldsymbol{y}$ under design s. The goal of the optimal pairing problem is to find the design s^* such that $U(s^*) \geq U(s)$ for all $s \in \mathcal{S}$.

The expression for $U(s)$ in (2) can be rewritten in a form that allows for tractable computation. The summation over the $2^{N/2}$ terms can be reduced to a sum over only $N/2$ terms,

$$U(s) = \sum_{k=1}^{N/2} \left\{ \mathrm{E}(p_k \ln p_k) + \mathrm{E}(q_k \ln q_k) - \mathrm{E}(p_k) \ln \mathrm{E}(p_k) - \mathrm{E}(q_k) \ln \mathrm{E}(q_k) \right\} \tag{3}$$

where p_k is given in (1), with $q_k = 1 - p_k$, and $\mathrm{E}(\cdot)$ is taken with respect to the multivariate normal prior distribution for $(\theta_{i_k}, \theta_{j_k})$. The derivation of this reduction is shown in Appendix A. This important result implies that the utility of a design using the Kullback-Leibler divergence is the sum of game-wise contributions.

It is worth noting that, while the k-th summand of (3) can rarely, if ever, be computed analytically for most choices of linear paired comparison models, the means can be evaluated numerically. In fact, each integral, which is a mean over a bivariate normal density, can be transformed into an integral over a scalar normal variable. To see this, we first observe that all of the means in (3) are functions of the $\boldsymbol{\theta}$ only through the p_k, which involve only differences of two strength parameters. Letting $\Delta_{i_k j_k} = \theta_{i_k} - \theta_{j_k}$, so that $p_k = F(\Delta_{i_k j_k})$, the mean of any function of p_k can be taken with respect to the (scalar) normal prior distribution of $\Delta_{i_k j_k}$, which has mean

$$\mu_{i_k j_k} = \mathrm{E}(\Delta_{i_k j_k}) = \mathrm{E}(\theta_{i_k}) - \mathrm{E}(\theta_{j_k})$$

and variance

$$\sigma^2_{i_k j_k} = \mathrm{Var}(\Delta_{i_k j_k}) = \mathrm{Var}(\theta_{i_k}) + \mathrm{Var}(\theta_{j_k}) - 2\ \mathrm{Cov}(\theta_{i_k}, \theta_{j_k}),$$

both of which are trivially computed. Thus, for example,

$$\begin{aligned} \mathrm{E}(p_k \ln p_k) &= \int (p_k \ln p_k) p(\theta_{i_k}, \theta_{j_k}) d\theta_{i_k} d\theta_{j_k} \\ &= \int (p_k \ln p_k) p(\Delta_{i_k j_k}) d\Delta_{i_k j_k} \\ &= \int (F(\Delta_{i_k j_k}) \ln F(\Delta_{i_k j_k})) p(\Delta_{i_k j_k}) d\Delta_{i_k j_k} \end{aligned}$$

where $\Delta_{i_k j_k} \sim \mathrm{N}(\mu_{i_k j_k},\ \sigma^2_{i_k j_k})$.

Recognizing that the means in (3) are integrals over scalar normal densities, we can evaluate the means numerically using Gauss-Hermite quadrature (see, for example, Davis and Rabinowitz, 1975; Crouch and Spiegelman, 1990). This approach involves approximating an integral by a weighted sum at n grid points, where the n points are chosen to be a linear transformation of the zeroes of the n-th order Hermite polynomial. Fortran subroutines for determining the grid points and weights for Gauss-Hermite quadrature can be obtained from Press et al. (1997). Gauss-Hermite quadrature is particularly well-suited to our application because the densities over which the integrals are evaluated are exactly normal, and the functions multiplying the densities are, for conventional linear paired comparison models, well-behaved and slowly varying, and reasonably approximated by moderately high-order polynomials of $\Delta_{i_k j_k}$.

2.2 Determination of the optimal design

The utility function evaluated at a design is a sum of individual game contributions of the form

$$C_{ij} = \mathrm{E}(p_{ij} \ln p_{ij}) + \mathrm{E}(q_{ij} \ln q_{ij}) - \mathrm{E}(p_{ij}) \ln \mathrm{E}(p_{ij}) - \mathrm{E}(q_{ij}) \ln \mathrm{E}(q_{ij}), \tag{4}$$

where $p_{ij} = F(\theta_i - \theta_j)$ is the conditional probability that player i defeats player j given $\boldsymbol{\theta}$, $q_{ij} = 1 - p_{ij}$, and that $C_{ij} = C_{ji}$ due to symmetry. Also, $C_{ij} > 0$, which is a direct consequence of Jensen's inequality (briefly justified in Appendix B) assuming no degeneracy in the prior distribution of $\boldsymbol{\theta}$. Therefore, the problem of finding the optimal design reduces to evaluating all $\binom{N}{2}$ values of C_{ij}, and then determining the subset of $N/2$ of them corresponding to distinct pairs that produces the largest sum. Determining the optimizing set of pairs in our context is isomorphic to a well-known problem in graph theory called the "maximum-weight perfect matching" problem (see, for example, Lovász and Plummer, 1986). Viewing individual players as vertices in a graph, with incident edges corresponding to the game-wise contributions of (4), the maximum-weight perfect matching problem involves finding a subset of edges in the graph such that each vertex is met by only one edge (resulting in a "perfect matching"), and that the sum of the weights of the edges in the perfect matching is maximal. An efficient algorithm for determining the maximum-weight perfect matching was originally developed by Edmonds (1965), and improvements have more recently been worked out by Gabow and Tarjan (1991) and Cook and Rohe (1999) as well as others. An additional feature of the weighted perfect matching algorithms that makes it useful for optimal tournament pairing is that constraints on the inclusion of edges can be easily incorporated. For example, in the context of tournament design, if we want to optimize only over designs where players who have already competed do not compete again, then we apply the weighted matching algorithm to a graph with the corresponding edges removed.

When the number of players, N, is odd, a simple modification can be applied to use the

maximum-weight perfect matching algorithm. An N+1-st fictitious player is added to the set of N, with the requirement that $C_{i,N+1} = 0$ for all $i = 1, \ldots, N$. Setting $C_{i,N+1} = 0$ ensures that pairing player $N + 1$ with any other is less favorable than pairing two actual players. The perfect matching algorithm is now applied to the augmented set of $N + 1$ players, and the actual player who is paired with the fictitious player is left out of the pairing.

The adaptive procedure for pairing N players in round r can be summarized in the following steps.

1. Determine the multivariate normal distribution of players' strengths, $\boldsymbol{\theta}$, for round r.
   - If $r = 1$, use the pre-tournament multivariate normal prior distribution for $\boldsymbol{\theta}$.
   - If $r > 1$, use a multivariate normal density to approximate the actual posterior density computed from game results based on the first $r - 1$ rounds. The multivariate normal approximation depends on the choice of the specific linear paired comparison model.
2. Calculate all $\binom{N}{2}$ values of C_{ij} given in (4), using Gauss-Hermite quadrature of scalar normal variables to evaluate the means.
3. Apply the maximum-weight perfect matching algorithm to determine the optimal pairing.
   - If N is even, apply the perfect matching algorithm to the collection of the C_{ij}.
   - If N is odd, add an $N + 1$-st fictitious player with $C_{i,N+1} = 0$ for all i, and apply the perfect matching algorithm to the augmented set of $N + 1$ players, dropping the player that is paired with the fictitious player.

2.3 Ordering within pairs

In many games, and paired comparison settings in general, an asymmetry can exist within a pair that may be relevant for the outcome of a comparison. For example, in chess, one player is assigned to play white, and the other black, while in many professional sports, an asymmetry within a pair occurs through one team competing on its home field. In more general paired comparison situations, objects to be compared may be presented one after the other rather than simultaneously, so that the order in which the objects are presented may be relevant to the preference probability.

For optimal tournament design with the goal of inferring player strengths, it is not usually of interest to make inferences about order effects. Instead, the within-pair order is a nuisance feature of the design. Rather than guide the choice of within-pair order through modeling its effect in a probability model, we choose an approach that will minimize the confounding of order effects with player strength. This can be done simply by balancing order by player. In the context of chess, this involves assigning each player white and black the same number of times for tournaments with an even number of rounds, and a difference of one for tournaments with an odd number of rounds. David (1988, pg 143) argues that unless the order effects are expected to be large, balancing order in the design without estimating its effect should be sufficient.

To incorporate the balance of within-pair order into our optimal design approach, we include a set of pairing restrictions in the following manner. For odd-numbered rounds in a tournament, the algorithm proceeds without modification, and order is randomly assigned within each formed pair. For even-numbered rounds, a restriction is added to the maximum-weight perfect matching algorithm to ensure that players in the previous round who were randomly assigned "first" in the ordering will be paired with players who were assigned

"second" in the ordering. Once the optimal design is found under this restriction, the ordering within pairs is then reversed compared to the previous round. Thus, after even-numbered rounds, each player achieves exact balance in ordering.

3 Optimal design for the Bradley-Terry model

A commonly used linear paired comparison model is the Bradley-Terry model (Bradley and Terry, 1952). This model has had a long tradition of being used to measure competitor strength in tournaments. Recent work that incorporated the Bradley-Terry model for measuring competitor strength includes Bradley (1984), Joe (1990), and Glickman (1999). Letting θ_i and θ_j be the strength parameters of players i and j, and letting Y_{ij} be 1 if i defeats j, and 0 if j defeats i, the Bradley-Terry model asserts

$$\mathrm{P}(Y_{ij} = 1 \mid \theta_i, \theta_j) = \frac{\exp(\theta_i)}{\exp(\theta_i) + \exp(\theta_j)} = \frac{1}{1 + \exp(-(\theta_i - \theta_j))}. \tag{5}$$

The model in (5) is parameterized so that the merits take on an unrestricted range of real values.

When applying our optimal design approach, we use game outcomes from previous rounds to form a multivariate normal prior distribution for the current round. This can be accomplished using the Bayesian extension to the Bradley-Terry model proposed by Leonard (1973). His approach assumes a multivariate normal prior distribution for the strength parameters $\boldsymbol{\theta}$ in the Bradley-Terry model. Other Bayesian extensions to the Bradley-Terry model include those by Davidson and Solomon (1973) and Chen and Smith (1984). Leonard's approach involves approximating the posterior density, which is proportional to the product of a multivariate normal prior density and a product-binomial likelihood of logistic probabilities, by another multivariate normal density by retaining the first two moments of the

actual posterior density. In our setting, once game outcomes are observed, the approximating normal posterior distribution can be used as the prior distribution for the next round of the tournament in order to determine pairings in an adaptive fashion.

For the Bradley-Terry model, the values of the C_{ij} in (4) can be evaluated using Gauss-Hermite quadrature as a function of the prior mean and variance of $\Delta_{ij} = \theta_i - \theta_j$. We have found that using 30 grid points for Gauss-Hermite quadrature applied to functions of Bradley-Terry probabilities is sufficiently accurate for the computations of means in (4). The values of C_{ij} are summarized in a contour plot in Figure 1. The contour plot indicates the types of pairings that are considered preferable when forming pairs. Because the greatest values of C_{ij} occur in the upper left portion of the figure, the most preferred pairing occurs between two players that have a mean difference near zero, and a large standard deviation of a difference. This is intuitively reasonable because the greatest amount of information can be learned about two players' strengths when they are expected to be close in ability, and the amount of uncertainty is large. Figure 1 also reveals that, for constant standard deviation in strength difference, the value of C_{ij} decreases as the mean difference increases. Thus, pairings involving players with small mean differences are favored in the algorithm. Furthermore, for constant mean difference, the value of C_{ij} increases as the standard deviation of strength difference increases. Again, this makes sense intuitively because we should expect greater gain in information when the variance of the difference in strength is large.

One interesting feature of our optimal design algorithm is that it automatically lowers the utility of pairing players who have already competed against each other. When players i and j have competed in an earlier round, a positive correlation is induced on the distribution of (θ_i, θ_j) which is reflected in a positive covariance in the approximate normal posterior distribution. When the value of C_{ij} is computed for this player pair in a subsequent round, the positive covariance between θ_i and θ_j will reduce the variance of the difference, so that,

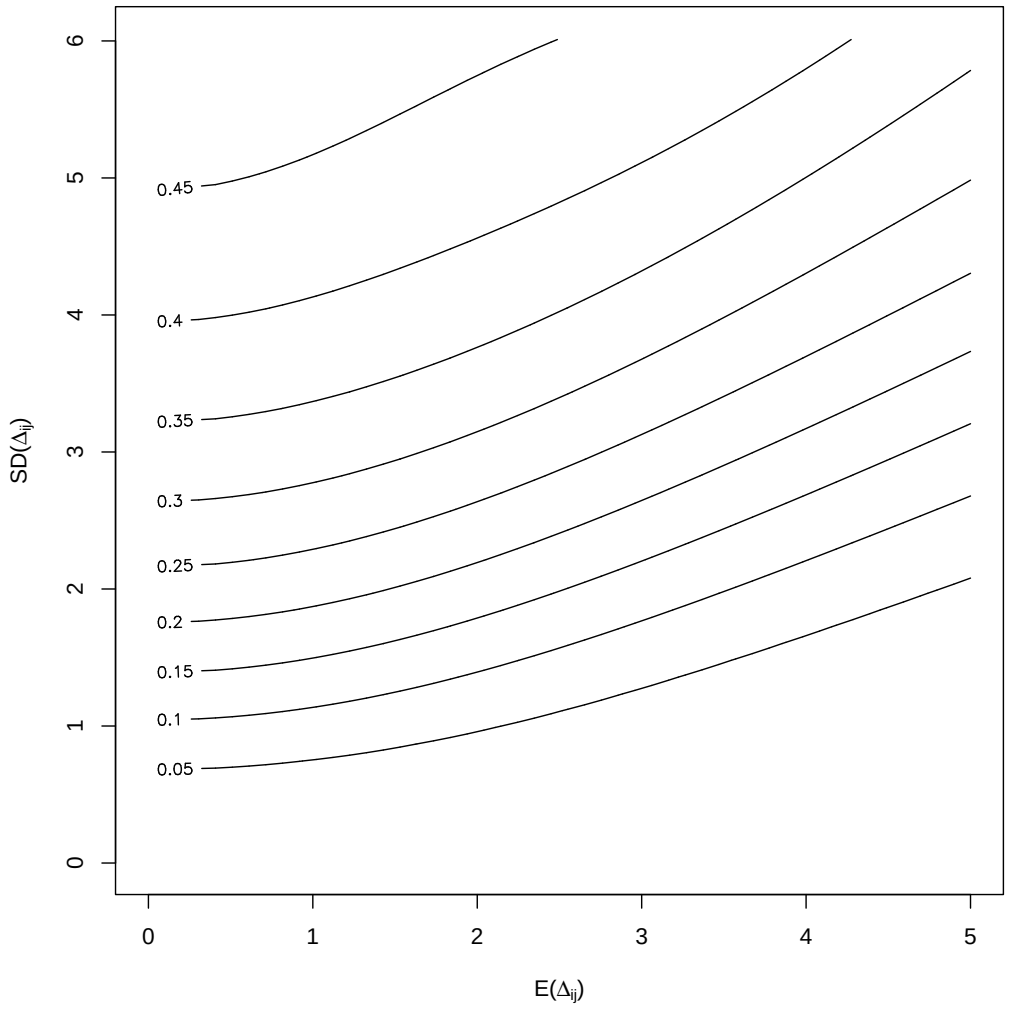

Figure 1: Contour plot of C_{ij} in (4) for the Bradley-Terry model as a function of the mean difference in strength parameters, and the standard deviation of the difference in strength parameters.

compared to pairs of players who have not yet competed, the value of C_{ij} will be small. This will lessen the likelihood that players i and j will compete for a second time. One situation, however, where players i and j may be paired a second time is if their mean difference is inferred to be small. This could occur, for example, if player i is assumed stronger than player j prior to competition, but player j defeats player i in a game played between them, at which point their strengths may be inferred to be nearly indistinguishable.

4 Comparison of design implementations

To examine the performance of our design approach, we evaluated our method along with two other pairing methods on simulated data. For our pairing approach, we considered a pairing scheme where no constraints were imposed in having players compete multiple times, and a second scheme where players who already have competed could not compete again. The first alternative to our method is pairing entirely at random. The second is a modified version of the Swiss system for tournament pairings. The modified Swiss system is due to Ólafsson (1990) who implements a pairing scheme using the weighted matching algorithm with a utility function that was optimized to mimic the Swiss system as closely as possible. To implement weighted maximum-weight perfect matching in our method and in that by Ólafsson, we used publicly available *C* code by Ed Rothberg (which can be obtained at `ftp://elib.zib-berlin.de` in the `/pub/mathprog/matching/weighted` directory) that implements Gabow's (1973) weighted matching algorithm.

Our simulations assumed a tournament of at most 16 rounds, consisting of 50 players with equally-spaced strength parameters from $\theta_1 = -2.5$ to $\theta_{50} = 2.5$. Because game outcomes depend on the pairing for any particular method, we carried out the following procedure for generating simulated game outcomes. First, we note that if X_i and X_j are independent

exponential random variables with means $\exp(\theta_i)$ and $\exp(\theta_j)$, respectively, then elementary probability yields $\Pr(X_i > X_j) = \exp(\theta_i)/(\exp(\theta_i) + \exp(\theta_j))$, the Bradley-Terry probability for player i defeating player j. This suggests that if we randomly generate values from independent exponential distributions with means $\exp(\theta_1), \ldots, \exp(\theta_{50})$, then player i is declared the winner of a game against player j with the correct probability if the simulated value X_i is greater than X_j; otherwise player j is declared the winner. This approach can be understood as each player displaying a "performance" for a game that varies from game to game, and the comparison of performances will yield the desired binary game outcome. An important feature of this approach is that generating game outcomes are separate and independent of the pairing method in the sense that players are assumed to produce game performances unrelated to the choice of their opponents. Thus, a single 16-round simulated tournament consists of generating, for each of the 50 players, 16 exponential random variables (one corresponding to each round of the tournament). The game outcomes, once a pairing for a round is determined, are the results of the comparisons of the simulated exponential random variables. The process of creating 16-round tournaments for 50 players was repeated 500 times to create 500 simulated tournaments on which to evaluate and compare the pairing methods.

Our simulation experiment consisted of varying three distinct factors for each pairing method. The first factor was the duration of the tournament. We examined the results after players completed simulated tournaments of 4 rounds, 8 rounds, and 16 rounds. The data from the 4 round and 8 round tournaments were simply the first 4 and first 8 rounds of the 16 round tournaments that were simulated. Secondly, we altered each of the pairing methods due to color/order constraints. Specifically, to incorporate color/order constraints within pairs, we altered our method in the manner described in Section 2.3. We carried out an identical procedure for incorporating order constraints for random pairings; players who were assigned "white" in an odd-numbered round could not compete against each other

in the subsequent round. For the Ólafsson approach, color balance is incorporated as a separate term in the utility function. In the version of Ólafsson pairing without color, we set this term to 0 in computing pairwise utilities. The third factor in our simulations was the assumed multivariate normal prior distribution for the strength parameters. We considered two different cases. The first case assumed an informative prior distribution with the means equal to the true values, a covariance matrix with variances equal to 0.3, and zero correlations. A prior distribution of this sort, particularly one with small variances, might be used in practice if a tournament involved players whose strengths were known to some precision in advance. The second case assumed a vague prior distribution with the mean vector set to a random permutation of 50 equally spaced values from -0.1 to 0.1, and a covariance matrix with zero correlations and variances equal to 4. This assumption would be appropriate for a tournament where no prior knowledge existed about players' strengths.

We evaluated the pairing methods by computing two statistics for each tournament, and reporting the mean and central 95% interval of each statistic across the 500 simulations. The first is the log of the determinant of the posterior variance (under the Bradley-Terry model) after the completion of a tournament. Larger values of this statistic indicate greater overall uncertainty in the player strengths. The second statistic is the sum of squared deviations between the ranks of the 50 players in order of the true strength parameters and the posterior means. A value of 0 indicates that the rank order from the posterior mean matches exactly the rank order of the true strengths. Values greater than 0 indicate a degree to which the relative ranking is not captured by the posterior mean of the strength parameters. It should be noted that our method is not optimized to produce rankings that are consistent with the truth in small numbers of rounds, mainly because our approach tends to pair players that are close in strength, so that the outcomes of games would tend to shuffle the means compared to the pairings in other methods. The results of the simulations when no color constraints are imposed are displayed in Tables 1 and 2. As we discuss below, the results for

the simulations with color constraints are similar. The statistic "log-variance" refers to the log of the determinant of the posterior variance, and "SSDR" refers to the sum of squared deviations in ranks. The column labeled "New" in the tables refers to our method with unrestricted pairings, and "New-X" refers to our method where players are not allowed to compete if they have played in an earlier round.

Table 1 shows the results for the simulations in which the prior distribution is informative about the strengths. In all cases, the posterior variance of strength parameters using our pairing method is substantially smaller than either random pairings, or Swiss system pairings. The restriction in our approach that players cannot compete more than once against the same opponent results in comparable variances to the unrestricted version. Random pairings appear to produce more variable inferences than either of our methods or the Swiss system. Furthermore, the magnitude of variance reduction in our method compared to the Swiss system improves as the number of rounds increases. While not displayed on the tables, these conclusions do not change whether color restrictions in the algorithms are incorporated.

The sum of squared difference in rankings reveal noteworthy features of the pairing methods when the prior distribution is informative. For 4-game tournaments, both the random and Swiss pairings outperform our approaches, as they both produce smaller sets of sums of squared differences in rank. The outperformance of the random and Swiss pairings decreases as the number of rounds increases. Our methods are slightly better on this statistic than the Swiss system for the 16-game tournaments, and about the same as the random pairings. Because our approach tends to pair players with similar inferred strength, players who are close in rank will be paired, even in the early rounds of a tournament. This is in contrast to the Swiss system where, in early rounds of a tournament, players further apart in strength are paired. The net effect of this difference is that, in our method, players close in strength will be further apart after they compete, and therefore early on the relative

rankings will be unreliable. In the Swiss system, two players will be paired initially who are already far apart in strength, and because the better player is likely to win, the posterior mean strengths will still be far apart and the relative rankings will be retained. Again, these conclusions persist in the analysis of simulations with color assignment restrictions.

It is also worth mentioning that the SSDR statistic increases with the number of rounds regardless of the pairing method. Because the means are ordered correctly at the start, and that game outcomes tend to shift the posterior distribution of the strength parameters, the posterior means are continually shuffled somewhat by the different pairing schemes and the randomness of the game results.

When a vague prior distribution on the strength parameters is assumed, with a random order of the means, our methods again produce consistently smaller posterior variance measures than random pairings or Swiss pairings. The results are shown in Table 2. The variance reduction is more pronounced as the number of rounds increases. The SSDR criterion appears worse for our method for 4-round and 8-round tournaments than the Swiss pairings, but better in 16-round tournaments. Our approach is consistently better on the SSDR criterion than random pairings. All of these conclusions persist when color allocation restrictions are included. The magnitude of the SSDR statistics in Table 2 compared to those in Table 1 indicate that even with as many as 16 games per player, the inferred ranks are still substantially different from the true ranks, regardless of the pairing method used.

In our pairing method in which players are allowed to compete against the same opponent multiple times, we find that such occurrences are relatively rare. Our simulations reveal that, for 4-round tournaments, players never compete more than once when the prior distribution is vague, and only about 1% of the games when the prior distribution is informative. In 8-round tournaments, less than 1% of the games in simulated tournaments involve repeat

opponents when the prior distribution is vague, and about 5% of the games when the prior distribution is informative. The greatest fraction of repeat opponents occurs for 16-round tournaments, in which 5.5% of the games involve opponents who play more than once when the prior distribution is vague, and about 38% of the games when the prior distribution is informative.

	Informative prior distribution			
	Random	Swiss	New	New-X
log-variance	–67.71	–67.98	–72.66	–72.66
4 games	(–68.40, –66.99)	(–68.36, –67.53)	(–72.72, –72.60)	(–72.72, –72.59)
log-variance	–74.24	–76.80	–82.49	–82.34
8 games	(–75.12, –73.34)	(–77.21, –76.30)	(–82.58, –82.37)	(–82.51, –82.13)
log-variance	–85.15	–91.24	–97.39	–95.54
16 games	(–86.23, –83.98)	(–91.71, –90.74)	(–97.53, –97.22)	(–96.01, –95.06)
SSDR	157.6	155.3	232.9	232.7
4 games	(90.0, 258.1)	(79.0, 244.0)	(147.0, 336.0)	(146.0, 336.0)
SSDR	243.3	265.7	295.0	291.9
8 games	(144.0, 366.0)	(153.0, 402.0)	(177.0, 440.0)	(175.9, 428.0)
SSDR	312.8	346.6	311.6	316.3
16 games	(189.0, 481.2)	(204.0, 514.0)	(178.0, 476.2)	(196.0, 466.1)

Table 1: Results of 500 simulations in which the prior mean of the merits were equal to the true means, the prior variances were 0.3, and the prior correlations were zero. The mean and 95% intervals of the two summary statistics across the 500 simulations are displayed.

5 Discussion

The adaptive paired comparison approach developed here has several appealing features. First, the pairing method relies on maximizing the expected Kullback-Leibler information from the prior distribution to the posterior distribution, which is a sensible utility in that pairings will produce efficient and informative inferences. Secondly, the form of the utility allows for a convenient reduction in computation so that the utility for a collection of games

	Non-informative prior distribution			
	Random	Swiss	New	New-X
log-variance	13.13	8.69	8.28	8.28
4 games	(9.34, 16.96)	(7.28, 10.19)	(7.35, 9.07)	(7.35, 9.07)
log-variance	–13.06	–21.84	–23.38	–23 41
8 games	(–17.17, –8.21)	(–23.21, –20.54)	(–24.32, –22.40)	(–24.48, –22.31)
log-variance	–42.63	–54.96	–57.73	–57.35
16 games	(–46.80, –38.24)	(–56.94, –52.78)	(–58.69, –56.78)	(–58.87, –55.29)
SSDR	5557	3714	5201	5201
4 games	(3304, 8247)	(1992, 5592)	(3322, 7726)	(3322, 7726)
SSDR	3128	2206	2612	2607
8 games	(1905, 4565)	(1314, 3417)	(1548, 4018)	(1506, 4023)
SSDR	1674	1316	1231	1265
16 games	(1009, 2453)	(775, 2002)	(783, 1864)	(784, 1929)

Table 2: Results of 500 simulations in which the prior mean of the merits was a random permutation of the equally-spaced values from -0.1 to 0.1, the prior variances were 4, and the prior correlations were zero. The mean and 95% intervals of the two summary statistics across the 500 simulations are displayed.

is simply the sum of the game-wise contributions. Finally, the combinatorial optimization of selecting pairs whose sum of game-wise utilities is maximal is isomorphic to a standard problem in graph theory, the solution to which is implementable in a straightforward manner. The method is adaptive so that the pairing scheme can be applied recursively after each round of a tournament, or, more generally, after each set of pairings in a paired comparison experiment.

Our design approach can be applied to paired comparison problems that involve ties or partial preferences following an idea in Glickman (1999). The design assumes that no partial preferences are observable, but when a tie or partial preference is actually observed, its contribution to the likelihood is counted as a fraction of a win and a corresponding fraction of a loss. For example, suppose p is the probability one player defeats another. If a tie results, then the contribution to the likelihood for this outcome is $p^{0.5}(1-p)^{0.5}$. For our

method, the multivariate normal prior distribution for a new round of a tournament may be computed based on having observed ties and partial preferences in such a manner.

The simulations in Section 4 demonstrate that our approach produces inferences that are more precise than other methods in terms of total variance of the posterior distribution, regardless of using a vague prior distribution or an informative prior distribution. However, because players that are paired tend to be similar in strength using our approach, the gain in precision is related to the movement in the means after a round of games is played. Because most tournaments, even for large numbers of players, typically schedule at most six to eight rounds, this is a potential concern of the application of our method. Unlike the Swiss system, which tends to pair players of similar strength (by virtue of having similar cumulative game results) in later rounds of a tournament, our approach pairs players of similar strength as soon as possible, assuming little to no correlation in the strength parameters. Even though our pairing scheme produces inferences that are generally less variable than the other methods for the types of simulation sets we analyzed, our pairing approach must be applied for a substantial number of rounds if the prior variances of the strength parameters are large. If the number of comparisons is not as limited, as is the case in conventional tournaments, then the efficiency gain using our approach per game makes it an attractive alternative to other pairing methods.

Extensions to our approach are possible, including ones that address the issue of the inferiority of our method in producing reliable rankings for a small number of rounds compared to the Swiss system. For example, rather than assuming a prior distribution with uncorrelated strength parameters, one could assume a correlation structure that where correlations are large for players whose prior means are close, and small for players whose prior means are far apart. The justification for assuming such a correlation structure is that it mimics the effect of players already competing several rounds using our pairing method. This might

be desirable if it is known that players are going to be competing in a small number of rounds because assuming correlated parameters acts as if players close in ability have already competed (which they would have in our method) and players different in ability have not competed. Computationally, this correlation structure forces players further apart in strength to be paired because the variance of the difference, accounting for the prior correlation, will be larger than for players with mean strengths that are close. The net result is a pairing that is closer to the Swiss system for initial rounds. Such extensions to our approach highlight the flexibility of our general framework for paired comparison design.

A Simplification of utility

Let

$$U(s) = \int \sum_{\boldsymbol{y} \in \mathcal{Y}_s} \ln \left\{ \frac{p(\boldsymbol{\theta} \mid \boldsymbol{y})}{p(\boldsymbol{\theta})} \right\} p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

as in (2). Noting that $p(\boldsymbol{\theta} \mid \boldsymbol{y})/p(\boldsymbol{\theta}) = p(\boldsymbol{y} \mid \boldsymbol{\theta})/p(\boldsymbol{y})$, where $p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$, we obtain

$$\begin{aligned} U(s) &= \int \left[\sum_{\boldsymbol{y} \in \mathcal{Y}_s} \ln \left\{ \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})}{p(\boldsymbol{y})} \right\} p(\boldsymbol{y} \mid \boldsymbol{\theta}) \right] p(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \int \left(\sum_{\boldsymbol{y} \in \mathcal{Y}_s} p(\boldsymbol{y}|\boldsymbol{\theta}) \ln p(\boldsymbol{y}|\boldsymbol{\theta}) \right) p(\boldsymbol{\theta}) d\boldsymbol{\theta} - \sum_{\boldsymbol{y} \in \mathcal{Y}_s} p(\boldsymbol{y}) \ln p(\boldsymbol{y}). \end{aligned} \tag{6}$$

The sum inside the integrand of the first term in (6) can be simplified by collapsing terms through a marginalization strategy, as follows.

$$\begin{aligned} \sum_{\boldsymbol{y} \in \mathcal{Y}_s} p(\boldsymbol{y}|\boldsymbol{\theta}) \ln p(\boldsymbol{y}|\boldsymbol{\theta}) &= \sum_{y_1=0}^{1} \sum_{y_2=0}^{1} \cdots \sum_{y_{N/2}=0}^{1} \left(\prod_{k=1}^{N/2} p(y_k|\boldsymbol{\theta}) \sum_{k=1}^{N/2} \ln p(y_k|\boldsymbol{\theta}) \right) \\ &= \sum_{k=1}^{N/2} \left\{ \sum_{y_1=0}^{1} \sum_{y_2=0}^{1} \cdots \sum_{y_{N/2}=0}^{1} \ln p(y_k|\boldsymbol{\theta}) \prod_{m=1}^{N/2} p(y_m|\boldsymbol{\theta}) \right\} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=1}^{N/2} \left\{ \sum_{y_k=0}^{1} p(y_k|\boldsymbol{\theta}) \ln p(y_k|\boldsymbol{\theta}) \sum_{y_1=0}^{1} \cdots \sum_{y_{k-1}=0}^{1} \sum_{y_{k+1}=0}^{1} \cdots \sum_{y_{N/2}=0}^{1} \prod_{m \neq k} p(y_m|\boldsymbol{\theta}) \right\} \\
&= \sum_{k=1}^{N/2} \sum_{y_k=0}^{1} p(y_k|\boldsymbol{\theta}) \ln p(y_k|\boldsymbol{\theta}) \\
&= \sum_{k=1}^{N/2} \sum_{y_k=0}^{1} p(y_k|\theta_{i_k}, \theta_{j_k}) \ln p(y_k|\theta_{i_k}, \theta_{j_k}) \qquad (7)
\end{aligned}
$$

The subtracted term in (6) can be simplified by recognizing, first, that $p(\boldsymbol{y}|s)$ factors into independent densities, and then applying the same marginalization strategy as above. To show the factorization,

$$
\begin{aligned}
p(\boldsymbol{y}) &= \int p(\boldsymbol{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int \left\{ \prod_{k=1}^{N/2} p(y_k|\theta_{i_k}, \theta_{j_k}) \right\} p(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \prod_{k=1}^{N/2} \left(\int p(y_k|\theta_{i_k}, \theta_{j_k}) p(\theta_{i_k}, \theta_{j_k}) d\theta_{i_k} d\theta_{j_k} \right) = \prod_{k=1}^{N/2} p(y_k).
\end{aligned}
$$

The simplification of the subtracted term in (6) proceeds analogously to (7) in the following manner.

$$
\begin{aligned}
\sum_{\boldsymbol{y} \in \mathcal{Y}_s} p(\boldsymbol{y}) \ln p(\boldsymbol{y}) &= \sum_{y_1=0}^{1} \sum_{y_2=0}^{1} \cdots \sum_{y_{N/2}=0}^{1} \left(\prod_{k=1}^{N/2} p(y_k) \sum_{k=1}^{N/2} \ln p(y_k) \right) \\
&= \sum_{k=1}^{N/2} \left\{ \sum_{y_1=0}^{1} \sum_{y_2=0}^{1} \cdots \sum_{y_{N/2}=0}^{1} \ln p(y_k) \prod_{m=1}^{N/2} p(y_m) \right\} \\
&= \sum_{k=1}^{N/2} \left\{ \sum_{y_k=0}^{1} p(y_k) \ln p(y_k) \sum_{y_1=0}^{1} \cdots \sum_{y_{k-1}=0}^{1} \sum_{y_{k+1}=0}^{1} \cdots \sum_{y_{N/2}=0}^{1} \prod_{m \neq k} p(y_m) \right\} \\
&= \sum_{k=1}^{N/2} \sum_{y_k=0}^{1} p(y_k) \ln p(y_k) \qquad (8)
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
U(s) &= \sum_{k=1}^{N/2} \sum_{y_k=0}^{1} \left\{ \int p(y_k|\theta_{i_k}, \theta_{j_k}) \ln p(y_k|\theta_{i_k}, \theta_{j_k}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} - p(y_k) \ln p(y_k) \right\} \\
&= \sum_{k=1}^{N/2} \sum_{y_k=0}^{1} \left\{ \int p(y_k|\theta_{i_k}, \theta_{j_k}) \ln p(y_k|\theta_{i_k}, \theta_{j_k}) p(\theta_{i_k}, \theta_{j_k}) d\theta_{i_k} d\theta_{j_k} - p(y_k) \ln p(y_k) \right\}
\end{aligned}
$$

$$= \sum_{k=1}^{N/2} \left\{ \mathrm{E}(p_k \ln p_k) + \mathrm{E}(q_k \ln q_k) - \mathrm{E}(p_k) \ln \mathrm{E}(p_k) - \mathrm{E}(q_k) \ln \mathrm{E}(q_k) \right\}$$

where $p_k = 1 - q_k = \mathrm{P}(Y_k = 1 \mid \theta_{i_k}, \theta_{j_k})$, and $\mathrm{E}(\cdot)$ is taken with respect to the multivariate normal prior distribution for $(\theta_{i_k}, \theta_{j_k})$. Thus the sum over $2^{N/2}$ terms can be collapsed into a sum over only $N/2$ terms.

B Positivity of game-wise utility contribution

Let

$$f(p) = p \ln p$$

Noting that $f(p)$ is strictly convex over $0 < p < 1$, Jensen's inequality (which itself, therefore, is strict for this function) yields

$$\mathrm{E}(f(X)) > f(\mathrm{E}(X)) \Longleftrightarrow \mathrm{E}(p \ln p) > \mathrm{E}(p) \ln \mathrm{E}(p).$$

Applying this result separately to p_{ij} and q_{ij} in (4) proves that $C_{ij} > 0$.

Bibliography

Bradley RA (1984) "Paired comparisons: some basic procedures and examples." Handbook of Statistics, vol 4, ed. PR Krishnaiah and PK Sen. Amsterdam: North-Holland, 299–326.

Bradley RA and Terry ME (1952) "The rank analysis of incomplete block designs. 1. The method of paired comparisons." *Biometrika*, **39**, 324–45.

Chaloner KC and Verdinelli I (1995) "Bayesian experimental design: A review." *Statistical Science*, **10**, 273–304.

Chen C and Smith TM (1984) "A Bayes-type estimator for the Bradley-Terry model for paired comparisons." *Journal of Statistical Planning and Inference*, **10**, 9–14.

Cook WJ and Rohe A (1999) "Computing minimum-weight perfect matchings." *INFORMS Journal on Computing*, **11**, 138–148.

Crouch EAC and Spiegelman D (1990) "The evaluation of integrals of the form $\int f(t)\exp(-t^2)dt$: application to logistic normal models." *Journal of the American Statistical Association*, **85**, 464–9.

Daniels HE (1969) "Round-robin tournament scores." *Biometrika*, **56**, 295–299.

David HA (1988) The method of paired comparisons (2nd ed). New York: Oxford University Press.

Davidson RR and Solomon DL (1973) "A Bayesian approach to paired comparison experimentation." *Biometrika*, **60**, 477–487.

Davis PJ and Rabinowitz P (1975) Methods of numerical integration. New York: Dover.

Edmonds J (1965) "Paths, trees and flowers." *Canadian Journal of Mathematics*, **17**, 449–467.

Gabow HN (1973) "Implementation of algorithms for maximum matching on nonbipartite graphs." Ph.D. Thesis, Stanford University.

Gabow HN and Tarjan RE (1991) "Faster scaling algorithms for general graph matching problems." *Journal of the ACM*, **38**, 815–853.

Glickman ME (1999) "Parameter estimation in large dynamic paired comparison experiments." *Applied Statistics*, **48**, 377–394.

Hartigan JA (1968) "Inference from a knockout tournament." *Annals of Mathematical Statistics*, **39**, 583–92.

Hwang FK, Zong-zhen L and Yao YC (1991) "Knockout tournaments with diluted Bradley-Terry preference schemes." *Journal of Statistical Planning and Inference*, **28**, 99–106.

Joe H (1990) "Extended use of paired comparison models, with application to chess rankings." *Applied Statistics*, **48**, 631–644.

Kažić BM (1980) The chess competitor's handbook. New York: Arco.

Leonard T (1973) "An alternative Bayesian approach to the Bradley-Terry model for paired comparisons." *Biometrics*, **33**, 121–132.

Lindley DV (1956) "On the measure of information provided by an experiment." *Annals of Statistics*, **27**, 986–1005.

Lindley DV (1972) Bayesian statistics – a review. Philadelphia: SIAM.

Lovász L and Plummer MD (1986) Matching theory. Budapest: Akadémia i Kiadoó.

Mosteller F (1951) "Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations." *Psychometrika*, **16**, 3–9.

Ólafsson S (1990) "Weighted matching in chess tournaments." *Journal of the Operational Research Society*, **41**, 17–24.

Press WH, Teukolsky SA, Vetterling WT and Flannery BP (1997) Numerical recipes in Fortran 77: The art of scientific computing (2nd ed). New York: Cambridge University Press.

Thurstone L (1927) "A law of comparative judgment." *Psychological Review*, **34**, 273–286.