Journal of Machine Learning Research 9 (2008) 2187-2216 Submitted 4/07; Revised 6/08; Published 10/08

Ranking Individuals by Group Comparisons

Tzu-Kuo Huang R93002@CSIE.NTU.EDU.TW
Chih-Jen Lin CJLIN@CSIE.NTU.EDU.TW
Department of Computer Science
National Taiwan University
Taipei 106, Taiwan

Ruby C. Weng CHWENG@NCCU.EDU.TW
Department of Statistics
National Chengchi University
Taipei 116, Taiwan

Editor: Greg Ridgeway

Abstract

This paper proposes new approaches to rank individuals from their group comparison results. Many real-world problems are of this type. For example, ranking players from team comparisons is important in some sports. In machine learning, a closely related application is classification using coding matrices. Group comparison results are usually in two types: binary indicator outcomes (wins/losses) or measured outcomes (scores). For each type of results, we propose new models for estimating individuals' abilities, and hence a ranking of individuals. The estimation is carried out by solving convex minimization problems, for which we develop easy and efficient solution procedures. Experiments on real bridge records and multi-class classification demonstrate the viability of the proposed models.

Keywords: ranking, group comparison, binary/scored outcomes, Bradley-Terry model, multi-class classification

1. Introduction

We address an interesting problem of estimating individuals' abilities from their group comparison results. This problem arises in some sports. One can evaluate a basketball player by his/her average points, but this criterion may be unfair as it ignores opponents' abilities. Comparison results in some sports, such as bridge, even do not reveal any direct information related to individuals' abilities. In a bridge match two partnerships form a team to compete with another two. The match record fairly reflects which two partnerships are better, but a partnership's raw score, depending on different boards, does not indicate its ability. Finding reasonable individual rankings using all group comparison records is thus a challenging task. Another application in machine learning/statistics is multi-class classification by coding matrices (Dietterich and Bakiri, 1995; Allwein et al., 2001). This technique decomposes a multi-class problem into several two-class problems, each of which is considered as the comparison between two disjoint subsets of class labels. The label with the greatest ability then serves as the prediction.

This line of research stems from the study of paired comparisons (David, 1988), in which one group/team consists of only one individual, and individuals' abilities are estimated from paired comparison results. Several models have been proposed, among which the most popular one is the Bradley-Terry model (Bradley and Terry, 1952): suppose there are k individuals whose abilities are indicated by a non-negative vector $\mathbf{p} = [p_1 \ p_2 \ \dots \ p_k]^T$. They proposed that

$$P(\text{individual } i \text{ beats } j) = \frac{p_i}{p_i + p_j}. \tag{1}$$

If comparisons are independent, then the maximum likelihood estimate of $\mathbf{p}$ is obtained by solving

$$\begin{aligned} \min_{\mathbf{p}} \quad & -\sum_{i \neq j} n_{ij} \log \frac{p_i}{p_i + p_j} \\ \text{subject to} \quad & \sum_{j=1}^{k} p_j = 1, \quad p_j \geq 0, j = 1, \dots, k, \end{aligned} \tag{2}$$

where n_{ij} is the number of times individual i beats j. The normalizing constraint in (2) is imposed because the objective function is scale-invariant. The solution to (2) can be found via a simple iterative procedure, which converges to the unique global minimum under mild conditions. Detailed discussions are in, for example, Hunter (2004).

Going from paired to group comparisons, we consider k individuals $\{1, \dots, k\}$ having m comparisons. The ith comparison setting involves a subset I_i, which is separated as two disjoint teams, I_i^+ and I_i^-. They have $n_i = n_i^+ + n_i^-$ comparisons, among which I_i^+ and I_i^- win n_i^+ and n_i^- times, respectively. Before seeking sophisticated models, an intuitive way to estimate the sth individual's ability is by the number of its winning comparisons normalized by the total number it involves:

$$\frac{\sum_{i:s \in I_i^+} n_i^+ + \sum_{i:s \in I_i^-} n_i^-}{\sum_{i:s \in I_i} n_i}. \tag{3}$$

In the case of paired comparisons, several authors (David, 1988; Hastie and Tibshirani, 1998) have shown that if

$$n_{si} > 0, \ n_{is} > 0, \ \text{and } n_{si} + n_{is} = \text{constant}, \forall s, i, \tag{4}$$

then the ranking by (3) is identical to that by the solution of (2). Note that under (4), the denominator of (3) is the same over all s, so the calculation is simplified to

$$\sum_{i:i \neq s} n_{si},$$

Although the above property may provide some support of (3), this approach has several problems. Firstly, (4) may not hold in most applications of paired comparisons. Secondly, (3) does not consider teammates' abilities, so strong players and weak ones receive the same credits. Because of these deficiencies, we use (3) as a baseline in experiments in Section 4 to demonstrate the need for more advanced methods. We refer to this approach as AVG.

As a direct extension of (1), Huang et al. (2006b) proposed a generalized Bradley-Terry model for group comparisons:

$$P(I_i^+ \text{ beats } I_i^-) = \frac{\sum_{j:j\in I_i^+} p_j}{\sum_{j:j\in I_i} p_j}, \tag{5}$$

which assumes that a team's ability is the sum of its members'. Under the assumption that comparisons are independent, individuals' abilities can be estimated by minimizing the negative log-likelihood of (5):

$$\begin{aligned} \min_{\mathbf{p}} \quad & -\sum_{i=1}^{m} \left(n_i^+ \log \frac{\sum_{j:j\in I_i^+} p_j}{\sum_{j:j\in I_i} p_j} + n_i^- \log \frac{\sum_{j:j\in I_i^-} p_j}{\sum_{j:j\in I_i} p_j} \right) \\ \text{subject to} \quad & \sum_{j=1}^{k} p_j = 1, \quad p_j \geq 0, j = 1, \ldots, k. \end{aligned} \tag{6}$$

Huang et al. (2006b) pointed out that (6) may not be a convex optimization problem, so global minima are not easy to obtain. Zadrozny (2002) was the first attempt to solve (6) by an iterative procedure, which, however, may fail to converge to a stationary point (Huang et al., 2006b). The algorithm of Huang et al. (2006b) converges to a stationary point under certain conditions. We refer to this approach as GBT.ML (Generalized Bradley-Terry Model using Maximum Likelihood).

Both models (1) and (6) consider comparisons' "binary" outcomes, that is, wins and losses. However, in many comparisons, results are also quantities reflecting opponents' performances/strengths, such as points in basketball or soccer games. Some work use these "measured" outcomes for paired comparisons; an example is Glickman (1993): instead of modeling the probability that one individual beats another, he considers the difference in two individuals' abilities as a random variable, whose realization is the difference in two scores. Individuals' abilities are then estimated via maximizing the likelihood.

In this paper we focus on the *batch* setting, under which individuals' abilities are not estimated until all comparisons are finished. This setting is suitable for annual sports events, such as the Bermuda Bowl for bridge considered in Section 4, where the goal is to rank participants according to their performances in the event. However, in some applications, competitions continue to take place without a clear end and a real-time ranking is required. An example is online gaming, where players make teams to compete against one another anytime they wish and expect a real-time update of their ranking right after a game is over. Several work deal with such an *online* scenario. For example, Herbrich and Graepel (2007) proposed the TrueSkill$^{\text{TM}}$ system, which generalizes the Elo system used in Chess (Elo, 1986). The system follows a Bayesian framework and obtains real-time rankings by an online learning scheme called *Gaussian density filtering* (Minka, 2001). Menke and Martinez (2007) re-parameterized the Bradley-Terry model (2) as a single-layer artificial neural network (ANN) and extended it for group competitions. Individuals' abilities are estimated by training the ANN with the delta rule, a typical online or incremental learning technique.

We managed to advance the state of the art in two directions. On the one hand, for comparisons with binary outcomes, we propose a new exponential model in Section 2. The

main advantage over Huang et al. (2006b) is that one can estimate individuals' abilities by minimizing unconstrained convex formulations. Hence global minima are easily obtained. On the other hand, we propose in Section 3 two models for comparisons with measured outcomes, which we call scored outcomes. The induced optimization problems are also unconstrained and convex; simple solution procedures are presented. This section may be the first study on finding individuals' abilities from *scored* group comparisons. Section 4 ranks partnerships in real bridge matches with the proposed approaches. Properties of different methods and their relations are studied in Section 5, which helps to explain experimental results. Section 6 demonstrates applications in multi-class classification. Section 7 concludes the work and discusses possible future directions.

Part of this work appears in a conference paper (Huang et al., 2006a).

2. Comparisons with Binary Outcomes

We denote individuals' abilities as a vector $\mathbf{v} \in R^k$, $-\infty < v_s < \infty$, $s = 1, \ldots, k$. Unlike $\mathbf{p}$ used in (5), $\mathbf{v}$ may have negative values. A team's ability is then defined as the sum of its members': for I_i^+ and I_i^-, their abilities are respectively

$$T_i^+ \equiv \sum_{s:s\in I_i^+} v_s \qquad \text{and} \qquad T_i^- \equiv \sum_{s:s\in I_i^-} v_s. \tag{7}$$

We consider teams' actual performances as random variables Y_i^+ and Y_i^-, $1 \le i \le m$ and define

$$P(I_i^+ \text{ beats } I_i^-) \equiv P(Y_i^+ - Y_i^- > 0). \tag{8}$$

The distribution of Y_i^+ and Y_i^- is generally unknown, but a reasonable choice should place the mode (the value at which the density function is maximized) around T_i^+ and T_i^-. To derive a computationally simple form for (8), we assume that Y_i^+ (and similarly Y_i^-) has a doubly-exponential extreme value distribution with

$$P(Y_i^+ \le y) = \exp(-e^{-(y-T_i^+)}), \tag{9}$$

whose mode is exactly T_i^+. Suppose Y_i^+ is independent of Y_i^-, from (8) and (9) we have

$$P(I_i^+ \text{ beats } I_i^-) = \frac{e^{T_i^+}}{e^{T_i^+} + e^{T_i^-}}. \tag{10}$$

The derivation is in Appendix A. One may assume other distributions (e.g., normal) in (9), but the resulting model is more complicated than (10). Such differences already occur for paired comparisons, where David (1988) gave some discussion. Thus (10) is our proposed model for binary outcomes.

For paired comparisons (i.e., each individual forms a team), (10) reduces to

$$P(\text{individual } i \text{ beats individual } j) = \frac{e^{v_i}}{e^{v_i} + e^{v_j}},$$

which is an equivalent re-parameterization (David, 1988; Hunter, 2004) of the Bradley-Terry model (1) by

$$p_i \equiv \frac{e^{v_i}}{\sum_{j=1}^k e^{v_j}}.$$

Therefore, our model (10) can also be considered as a generalized Bradley-Terry model. This re-parameterization however does not extend to the case of group comparisons, so (10) and (5) are different. Interestingly, (10) is a conditional exponential model or a *maximum entropy* model (Jaynes, 1957a,b), which is commonly used in the computational linguistic community (Berger et al., 1996). Thus we can use existing properties of this type of models.

Following the proposed model (10), we estimate $\mathbf{v}$ by using available comparison results. The following two sub-sections give two approaches: one minimizes a regularized least square formula, and the other minimizes the negative log-likelihood. Both are unconstrained convex optimization problems. Their differences are discussed in Section 5.

2.1 Regularized Least Square (Ext-B.RLS)

Recall that n_i^+ and n_i^- are respectively the number of comparisons teams I_i^+ and I_i^- win. From (10), we have

$$\frac{e^{T_i^+}}{e^{T_i^+} + e^{T_i^-}} \approx \frac{n_i^+}{n_i^+ + n_i^-},$$

and therefore

$$e^{T_i^+ - T_i^-} = \frac{e^{T_i^+}}{e^{T_i^-}} \approx \frac{n_i^+}{n_i^-}.$$

If $n_i^+ \neq 0$ and $n_i^- \neq 0$, one can solve

$$\min_{\mathbf{v}} \quad \sum_{i=1}^{m} \left((T_i^+ - T_i^-) - \log \frac{n_i^+}{n_i^-} \right)^2 \tag{11}$$

to estimate the vector $\mathbf{v}$ of individuals' abilities. In case of $n_i^+ = 0$ or $n_i^- = 0$, a simple solution is adding a small number to all n_i^+ and n_i^-. This technique is widely used in the computational linguistic community, known as the "add-one smoothing" for dealing with the zero-frequency problem. To represent (11) in a simpler form, we define a vector $\mathbf{d} \in R^m$ with

$$d_i \equiv \log \frac{n_i^+}{n_i^-},$$

and a "comparison setting matrix" $G \in R^{m \times k}$ with

$$G_{ij} \equiv \begin{cases} 1 & \text{if individual } j \in I_i^+, \\ -1 & \text{if individual } j \in I_i^-, \\ 0 & \text{if individual } j \notin I_i. \end{cases} \tag{12}$$

Take bridge in teams of four as an example. An individual stands for a partnership, so G's jth column records the jth partnership's team memberships in all m matches. Since a match is played by four partnerships from two teams, each row of G has two 1's, two -1's and $k-4$ 0's. Thus, G may look like

$$\begin{bmatrix} 1 & 1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & -1 & -1 & 0 & 0 \\ -1 & -1 & 0 & 0 & 0 & 0 & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}, \tag{13}$$

read as "The first match: the 1st, 2nd partnerships versus the 3rd, 4th; the second match: the 1st, 2nd versus the 5th, 6th; ...."

With the help of $\mathbf{d}$ and G, we rewrite (11) as

$$\min_{\mathbf{v}} \qquad (G\mathbf{v}-\mathbf{d})^T(G\mathbf{v}-\mathbf{d}), \tag{14}$$

which is equivalent to solving the following linear system:

$$G^TG\mathbf{v} = G^T\mathbf{d}. \tag{15}$$

If G^TG is not invertible, the linear system (15) may have multiple solutions, which lead to possibly multiple rankings. To see when G^TG is invertible, we prove the following result:

Theorem 1 *G^TG is invertible if and only if $rank(G) = k$.*

The proof is in Appendix B. This result shows that teams' members should change frequently across comparisons (as indicated by $rank(G) = k$) so that individuals' abilities are uniquely determined. To see how multiple rankings occur, consider an extreme case where several players always belong to the same team. Under the model (10), they can be merged as a single virtual player. After solving (14), their respective abilities can take any values but still remain optimal as long as the total ability is equal to the virtual player's. To handle such situations, we add a regularization term $\mu\mathbf{v}^T\mathbf{v}$ to (14):

$$\min_{\mathbf{v}} \qquad (G\mathbf{v}-\mathbf{d})^T(G\mathbf{v}-\mathbf{d}) + \mu\mathbf{v}^T\mathbf{v},$$

where μ is a small positive number. Then a unique solution exists:

$$\left(G^TG+\mu I\right)^{-1}G^T\mathbf{d}. \tag{16}$$

The rationale of the regularization is that individuals have equal abilities before having comparisons. We refer to this approach as Ext-B.RLS (Extreme value model for Binary outcomes using Regularized Least Square).

2.2 Maximum Likelihood (Ext-B.ML)

Under the assumption that comparisons are independent, the negative log-likelihood function is

$$l(\mathbf{v}) \equiv -\sum_{i=1}^{m}\left(n_i^+\log\frac{e^{T_i^+}}{e^{T_i^+}+e^{T_i^-}} + n_i^-\log\frac{e^{T_i^-}}{e^{T_i^+}+e^{T_i^-}}\right), \tag{17}$$

and we may estimate $\mathbf{v}$ by

$$\arg\min_{\mathbf{v}} l(\mathbf{v}).$$

It is well known that the log-likelihood of a conditional exponential model is concave, and hence $l(\mathbf{v})$ is convex. However, if $l(\mathbf{v})$ is not strictly convex, multiple global minima may result in multiple rankings. The following theorem gives the sufficient and necessary condition for strict convexity:

Theorem 2 *$l(\mathbf{v})$ is strictly convex if and only if $rank(G) = k$.*

The proof is in Appendix C. As discussed in Section 2.1, the condition may not hold, and a regularization term is usually added to ensure the uniqueness of the optimal solution. Here we consider a special one

$$\mu \sum_{s=1}^{k} (e^{v_s} + e^{-v_s}), \tag{18}$$

which is strictly convex and has unique minimum at $v_s = 0, s = 1, \ldots, k$. Later we will see that this function helps to derive a simple algorithm for maximizing the likelihood.

The modified negative log-likelihood is as the following:

$$l(\mathbf{v}) \equiv -\sum_{i=1}^{m} \left(n_i^+ \log \frac{e^{T_i^+}}{e^{T_i^+} + e^{T_i^-}} + n_i^- \log \frac{e^{T_i^-}}{e^{T_i^+} + e^{T_i^-}} \right) + \mu \sum_{s=1}^{k} (e^{v_s} + e^{-v_s}), \tag{19}$$

where μ is a small positive number. We estimate individuals' abilities by the unique global minimum

$$\arg\min_{\mathbf{v}} l(\mathbf{v}), \tag{20}$$

which satisfies the optimality condition:

$$\begin{aligned} \frac{\partial l(\mathbf{v})}{\partial v_s} &= -\Big(\sum_{i:s\in I_i^+} n_i^+ + \sum_{i:s\in I_i^-} n_i^- \Big) + \sum_{i:s\in I_i^+} \frac{n_i e^{T_i^+}}{e^{T_i^+} + e^{T_i^-}} + \sum_{i:s\in I_i^-} \frac{n_i e^{T_i^-}}{e^{T_i^+} + e^{T_i^-}} + \mu(e^{v_s} - e^{-v_s}) \\ &= 0, \quad s = 1, \ldots, k. \end{aligned}$$

Note that the strict convexity of (19) may not guarantee (20) to be attainable; we address this issue later in Theorem 3. Since μ is small,

$$\sum_{i:s\in I_i^+} n_i^+ + \sum_{i:s\in I_i^-} n_i^- \approx \sum_{i:s\in I_i^+} \frac{n_i e^{T_i^+}}{e^{T_i^+} + e^{T_i^-}} + \sum_{i:s\in I_i^-} \frac{n_i e^{T_i^-}}{e^{T_i^+} + e^{T_i^-}}, \tag{21}$$

which is a reasonable condition that the total number of observed wins of individual s is nearly the expected number by the assumed model. Meanwhile, the last term in $\partial l(\mathbf{v})/\partial v_s$ restricts the value of v_s from extremity, and thereby brings some robustness against huge n_i^+ or n_i^-.

Standard optimization methods (e.g., gradient or Newton's method) can be used to find a solution of (19). For conditional exponential models, an alternative technique to maximize the likelihood is the generalized iterative scaling by Darroch and Ratcliff (1972), which generates a sequence of iterations $\{\mathbf{v}^t\}_{t=0}^{\infty}$. The improved iterative scaling (Pietra et al., 1997) speeds up the convergence, but its update from $\mathbf{v}^t$ to $\mathbf{v}^{t+1}$ requires the solution of k one-variable minimization problems, which, however, usually do not have closed-form solutions. Goodman (2002) proposed the sequential conditional generalized iterative scaling, which changes only one variable at a time with a closed-form update rule. All the above techniques, however, need to be modified for solving (19) due to the regularization term (18). In the following we propose an iterative method that modifies one component of $\mathbf{v}$ at a time. Let $\boldsymbol{\delta} \equiv [0, \ldots, 0, \delta_s, 0, \ldots, 0]^T$ indicate the change of the sth component. Using the

inequality $\log x \le x - 1,\ \forall x > 0$,

$$\begin{aligned}
& l(\mathbf{v}+\boldsymbol{\delta}) - l(\mathbf{v}) \\
=& -\left(\sum_{i:s\in I_i^+} n_i^+ + \sum_{i:s\in I_i^-} n_i^-\right)\delta_s + \sum_{i:s\in I_i^+} n_i \log\left(\frac{e^{T_i^+ + \delta_s} + e^{T_i^-}}{e^{T_i^+} + e^{T_i^-}}\right) \\
& + \sum_{i:s\in I_i^-} n_i \log\left(\frac{e^{T_i^+} + e^{T_i^- + \delta_s}}{e^{T_i^+} + e^{T_i^-}}\right) + \mu e^{v_s}(e^{\delta_s} - 1) + \mu e^{-v_s}(e^{-\delta_s} - 1) \\
\le& -\left(\sum_{i:s\in I_i^+} n_i^+ + \sum_{i:s\in I_i^-} n_i^-\right)\delta_s + \left(\sum_{i:s\in I_i^+} \frac{n_i e^{T_i^+}}{e^{T_i^+} + e^{T_i^-}} + \sum_{i:s\in I_i^-} \frac{n_i e^{T_i^-}}{e^{T_i^+} + e^{T_i^-}}\right)(e^{\delta_s} - 1) \\
& + \mu e^{v_s}(e^{\delta_s} - 1) + \mu e^{-v_s}(e^{-\delta_s} - 1). \qquad (22)
\end{aligned}$$

If $\delta_s = 0$, (22) $= 0$. We then minimize (22) to obtain the largest reduction. It is easy to see that (22) is strictly convex. Taking the derivative with respect to δ_s to be zero, we find the root for a second-order polynomial of e^{δ_s}, so the update rule is:

$$v_s \leftarrow v_s + \log \frac{B_s + \sqrt{B_s^2 + 4\mu A_s e^{-v_s}}}{2A_s}, \qquad (23)$$

where

$$\begin{aligned}
A_s &\equiv \mu e^{v_s} + \sum_{i:s\in I_i^+} \frac{n_i e^{T_i^+}}{e^{T_i^+} + e^{T_i^-}} + \sum_{i:s\in I_i^-} \frac{n_i e^{T_i^-}}{e^{T_i^+} + e^{T_i^-}}, \qquad (24) \\
B_s &\equiv \sum_{i:s\in I_i^+} n_i^+ + \sum_{i:s\in I_i^-} n_i^-.
\end{aligned}$$

If using other regularization terms, minimizing (22) may not lead to a closed-form solution of δ_s. The algorithm is as the following:

Algorithm 1

1. Start with $\mathbf{v}^0$ and obtain $T_i^{0,+}, T_i^{0,-}, i = 1, \ldots, m$.
2. Repeat $(t = 0, 1, \ldots)$
   (a) Let $s = (t+1) \mod k$. Change the sth element of $\mathbf{v}^t$ by (23) to obtain $\mathbf{v}^{t+1}$.
   (b) Calculate $T_i^{t+1,+}, T_i^{t+1,-}, i = 1, \ldots, m$.

   until $\partial l(\mathbf{v}^t)/\partial v_j = 0,\ j = 1, \ldots, k$ are satisfied.

Next we address the convergence issue. As $A_s > 0$, (23) is always well-defined. A formal proof of Algorithm 1's convergence is in the following theorem:

Theorem 3 *The modified negative log-likelihood $l(\mathbf{v})$ defined in* (19) *attains a unique global minimum, and the sequence $\{\mathbf{v}^t\}$ generated by Algorithm 1 converges to it.*

The proof is in Appendix D. In Huang et al. (2006b), some assumptions are needed to ensure their update rule to be well-defined as well as the convergence. In contrast, Algorithm 1 does not require any assumption since the regularization term provides very nice properties. We refer to the approach of minimizing (19) as Ext-B.ML (Extreme value distribution model for Binary outcomes using Maximum Likelihood).

3. Comparisons with Scored Outcomes

This section proposes estimating individuals' abilities based on measured outcomes, such as points in basketball or soccer games. We still use random variables Y_i^+ and Y_i^- for team performances, but give n_i^+ and n_i^- different meanings: they now denote scores of I_i^+ and I_i^-. Our idea is to view $n_i^+ - n_i^-$ as a realization of $Y_i^+ - Y_i^-$ and maximize the resulting likelihood. Note that we model *difference* in scores instead of the score itself. We propose two approaches in the following sub-sections. One assumes normal distributions for Y_i^+ and Y_i^-, while the other assumes the same extreme value distribution (9). Individuals' abilities are estimated by maximizing the likelihood of score differences. Properties of the two approaches are investigated in Section 5.

3.1 Normal Distribution Model (NM-S.ML)

As mentioned in Section 2, using normal distributions for comparisons with binary outcomes is computationally more difficult due to a complicated form of $P(I_i^+ \text{ beats } I_i^-)$. However, for scored paired comparisons, Glickman (1993) successfully applied normal distributions. He considers individuals' performances as normally distributed random variables

$$Y_i \sim N(v_i, \sigma^2),\ i = 1, \ldots, k,$$

and view the score difference of individuals i and j as a realization of $Y_i - Y_j$. By assuming Y_i and Y_j are independent for all individuals,

$$Y_i - Y_j \sim N(v_i - v_j, 2\sigma^2), \tag{25}$$

and individuals' abilities are estimated by maximizing the likelihood. We extend this approach to group comparisons. Recall that Y_i^+ and Y_i^- are random variables for two teams' performances. With the same assumption of independent normal distributions, we have

$$Y_i^+ \sim N(T_i^+, \sigma^2), \quad Y_i^- \sim N(T_i^-, \sigma^2).$$

and

$$Y_i^+ - Y_i^- \sim N(T_i^+ - T_i^-, 2\sigma^2).$$

Assuming comparisons are independent and defining a vector $\mathbf{b}$ with

$$b_i \equiv n_i^+ - n_i^-,$$

the negative log-likelihood then is

$$\begin{aligned} l(\mathbf{v}, \sigma) &= \log\sigma + \frac{1}{4\sigma^2} \sum_{i=1}^{m} \left(T_i^+ - T_i^- - (n_i^+ - n_i^-)\right)^2 \\ &= \log\sigma + \frac{(G\mathbf{v} - \mathbf{b})^T (G\mathbf{v} - \mathbf{b})}{4\sigma^2}, \end{aligned} \tag{26}$$

where G is the comparison setting matrix defined in (12). The maximum likelihood estimate of $\mathbf{v}$ is obtained by solving $\partial l(\mathbf{v},\sigma)/\partial v_s = 0\ \forall s$, which is the following linear system:

$$G^T G\mathbf{v} = G^T\mathbf{b}. \tag{27}$$

Similar to (14), (27) may have multiple solutions if $G^T G$ is not invertible. To overcome this problem, we add a regularization term and solve

$$\min_{\mathbf{v}} \quad l(\mathbf{v},\sigma) + \frac{\mu}{4\sigma^2}\mathbf{v}^T\mathbf{v}, \tag{28}$$

where μ is small positive number. The unique solution of (28) then is

$$\bar{\mathbf{v}} \equiv (G^T G + \mu I)^{-1} G^T\mathbf{b}. \tag{29}$$

In addition, we also obtain an estimate of the variance by solving

$$\frac{\partial\big(l(\mathbf{v},\sigma^2) + \frac{\mu}{4\sigma^2}\mathbf{v}^T\mathbf{v}\big)}{\partial\sigma} = 0,$$

which leads to

$$\bar{\sigma}^2 \equiv \frac{(G\bar{\mathbf{v}} - \mathbf{b})^T(G\bar{\mathbf{v}} - \mathbf{b}) + \mu\bar{\mathbf{v}}^T\bar{\mathbf{v}}}{2}.$$

We refer to this method as NM-S.ML (Normal distribution-based Model for Scored outcomes using Maximum Likelihood).

3.2 Extreme Value Distribution Model (Ext-S.ML)

Instead of the normal distribution in (25), we now consider that $Y_i^+ - Y_i^-$ is under the extreme value distribution for binary outcomes. Appendix A shows that

$$P(Y_i^+ - Y_i^- \le y) = \frac{e^{T_i^-}}{e^{T_i^+ - y} + e^{T_i^-}}, \tag{30}$$

and hence the density function is

$$f_{Y_i^+ - Y_i^-}(y) = \frac{e^{T_i^- + T_i^+ - y}}{(e^{T_i^+ - y} + e^{T_i^-})^2}.$$

The negative log-likelihood function is

$$-\sum_{i=1}^{m} \log \frac{e^{T_i^+ + T_i^- - (n_i^+ - n_i^-)}}{\big(e^{T_i^+ - (n_i^+ - n_i^-)} + e^{T_i^-}\big)^2}. \tag{31}$$

A similar proof to Theorem 2's shows that (31) is convex and shares the same condition for strict convexity in Section 2.2. Therefore, the problem of multiple solutions may also occur. We thus adopt the same regularization term as in Section 2.2 and solve

$$\min_{\mathbf{v}} \quad l(\mathbf{v}) \equiv -\sum_{i=1}^{m} \log \frac{e^{T_i^+ + T_i^- - (n_i^+ - n_i^-)}}{\big(e^{T_i^+ - (n_i^+ - n_i^-)} + e^{T_i^-}\big)^2} + \mu\sum_{s=1}^{k}(e^{v_s} + e^{-v_s}). \tag{32}$$

The unique global minimum satisfies for $s = 1, \ldots, k$,

$$\frac{\partial l(\mathbf{v})}{\partial v_s} = -m_s + 2\left(\sum_{i:s\in I_i^+} \frac{e^{T_i^+ + n_i^-}}{e^{T_i^+ + n_i^-} + e^{T_i^- + n_i^+}} + \sum_{i:s\in I_i^-} \frac{e^{T_i^- + n_i^+}}{e^{T_i^+ + n_i^-} + e^{T_i^- + n_i^+}} \right) + \mu(e^{v_s} - e^{-v_s})$$
$$= 0, \tag{33}$$

where

$$m_s \equiv \sum_{i:s\in I_i} 1.$$

From (30),

$$P(Y_i^+ - Y_i^- \geq T_i^+ - T_i^-) = \frac{1}{2},\ i = 1, \ldots, m. \tag{34}$$

Since μ is small, (33) and (34) imply that for $s = 1, \ldots, k$,

$$\sum_{i:s\in I_i^+} P(Y_i^+ - Y_i^- \geq n_i^+ - n_i^-) + \sum_{i:s\in I_i^-} P(Y_i^- - Y_i^+ \geq n_i^- - n_i^+)$$
$$\approx \frac{m}{2} = \sum_{i:s\in I_i^+} P(Y_i^+ - Y_i^- \geq T_i^+ - T_i^-) + \sum_{i:s\in I_i^-} P(Y_i^- - Y_i^+ \geq T_i^- - T_i^+).$$

As (21) in Section 2.2, the above condition also indicates that models should be consistent with observations. To solve (32), we use Algorithm 1 with a different update rule, which is in the form of (23) but with

$$\begin{aligned} A_s &\equiv \mu e^{v_s} + 2\left(\sum_{i:s\in I_i^+} \frac{e^{T_i^+ + n_i^-}}{e^{T_i^+ + n_i^-} + e^{T_i^- + n_i^+}} + \sum_{i:s\in I_i^-} \frac{e^{T_i^- + n_i^+}}{e^{T_i^+ + n_i^-} + e^{T_i^- + n_i^+}} \right), \\ B_s &\equiv m_s. \end{aligned}$$

The derivation is similar to (23)'s: let $\boldsymbol{\delta} \equiv [0, \ldots, 0, \delta_s, 0, \ldots, 0]^T$. Then

$$\begin{aligned} & l(\mathbf{v} + \boldsymbol{\delta}) - l(\mathbf{v}) \\ =& -m_s\delta_s + 2\left(\sum_{i:s\in I_i^+} \log \frac{e^{T_i^+ + n_i^- + \delta_s} + e^{T_i^- + n_i^+}}{e^{T_i^+ + n_i^-} + e^{T_i^- + n_i^+}} + \sum_{i:s\in I_i^-} \log \frac{e^{T_i^- + n_i^+ + \delta_s} + e^{T_i^+ + n_i^-}}{e^{T_i^+ + n_i^-} + e^{T_i^- + n_i^+}} \right) \\ & + \mu e^{v_s}(e^{\delta_s} - 1) + \mu e^{-v_s}(e^{-\delta_s} - 1) \\ \leq& -m_s\delta_s + 2\left(\sum_{i:s\in I_i^+} \frac{e^{T_i^+ + n_i^-}}{e^{T_i^+ + n_i^-} + e^{T_i^- + n_i^+}} + \sum_{i:s\in I_i^-} \frac{e^{T_i^- + n_i^+}}{e^{T_i^+ + n_i^-} + e^{T_i^- + n_i^+}} \right)(e^{\delta_s} - 1) \\ & + \mu e^{v_s}(e^{\delta_s} - 1) + \mu e^{-v_s}(e^{-\delta_s} - 1). \end{aligned} \tag{35}$$

Minimizing (35) leads to the update rule. Global convergence can be proved in a similar way to Theorem 3. We refer to this approach as Ext-S.ML (Extreme value distribution model for Scored outcomes using Maximum Likelihood).

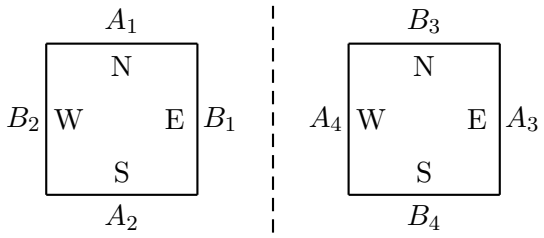


Figure 1: A typical bridge match setting. N, S, E and W stand for north, south, east, and west, respectively.

4. Ranking Partnerships from Real Bridge Records

This section presents a real application: ranking partnerships from match records of Bermuda Bowl 2005,[1] which is the most prestigious bridge event. In a match two partnerships (four players) from a team compete with two from another team. The rules require mutual understanding within a partnership, so partnerships are typically fixed while a team can send different partnerships for different matches. To rank partnerships using our model, an individual stands for a partnership, and every T_i^+ (or T_i^-) consists of two individuals. We caution the use of the term "team" here. Earlier we refer to each T_i^+ as a team and in bridge the two partnerships (or four players) of T_i^+ are really called a team. However, these four players are from a (super)-team (usually a country), which often has six members. We use "team" in both situations, which are easily distinguishable.

4.1 Experimental Settings

We discuss why a partnership's ability is not directly available from match results, and explain why our model is applicable here. Figure 1 illustrates the match setting. A_1, A_2, A_3, A_4 and B_1, B_2, B_3, B_4 are four players of Team A and Team B, sitting at two tables as depicted. A match consists of several boards, each of which is played at both tables. An important feature is that a board's four hands are at identical positions of two tables, but a team's two partnerships sit at complementary positions. In Figure 1, A_1 and A_2 sit at the north (N) and the south (S) sides of one table, so A_3 and A_4 must sit at the east (E) and the west (W) sides of the other table. This setting reduces the effect of uneven hands.

On each board winning partnerships receive raw scores. Depending on the difference in two teams' total scores, the winning team gains International Match Points (IMPs). For example, Table 1 shows records of the first ten boards of the match between two Indian partnerships and two Portuguese partnerships. We can see that a larger difference in raw scores results in more IMPs for the winner. IMPs are then converted to Victory Points (VP) for the team ranking.[2] A quick look at Table 1 may motivate the following straightforward approach: a partnership's score in a match is the sum of raw scores over all boards, and its ability is the average over the matches it plays. However, this estimate is unfair due to raw

1. All match records are available at `http://www.worldbridge.org/tourn/Estoril.05/Estoril.htm`. The subset used here is available at `http://www.csie.ntu.edu.tw/~cjlin/papers/genBTexp/Data.zip`.
2. The IMP-to-VP conversion for Bermuda Bowl 2005 is on page 32, `http://www.worldbridge.org/departments/rules/GeneralConditionsOfContest2005.pdf`.

Board	Table I		Table II		IMPs	
	NS	EW	NS	EW	IN	PT
1		1510		1510		
2	100		650			11
3		630		630		
4		650		660		
5	690		690			
6	420			50	10	
7	140		600			10
8		420		100		8
9	460		400		2	
10		110		140	1	

Table 1: Records of the first ten boards between India (IN) and Portugal (PT). India: NS at Table I and EW at Table II. The four columns in the middle are boards' raw scores, and only winners get points. For example, in the second board IN's NS partnership won at Table I and got 100 points while PT's NS got 650 at Table II. Since PT got more points than IN, it obtained IMPs.

scores' dependency on boards and opponents. Summing a partnership's raw scores favors those who get better hands or play against weak opponents. Moreover, since boards are different across rounds and partnerships play in different rounds, the sum of raw scores can be more unfair. The above analysis indicates that a partnership's ability cannot be obtained directly from group comparison results. Hence the proposed models can be helpful.

We consider qualifying games: 22 teams from all over the world had a round robin tournament, which consisted of $\binom{22}{2} = 231$ matches and each team played 21. Most teams had six players in three fixed partnerships, and there were 69 partnerships in total. In order to obtain reasonable rankings, each partnership should play enough matches. The last column of Table 3 shows each partnership's number of matches. Most played 13 to 15 matches, which are close to the average (14=21×2/3) of a team with three fixed partnerships. Thus these match records are reasonable for further analysis.

To use our model, the comparison setting matrix G defined in (12) is of size 231×69; as shown in (13) each row records a match setting and has exactly two 1's (two partnerships from one team), two -1's (two partnerships from another team) and 65 0's (the remaining partnerships). The sum of two rival teams' scores (VPs) is generally 30, but occasionally between 25 to 30 as a team's maximal VP is 25. We use two rival teams' VPs as n_i^+ and n_i^-, respectively. Several matches have zero scores; we add one to all n_i^+ and n_i^- for Ext-B.RLS to avoid the numerical difficulties caused by $\log(n_i^+/n_i^-)$.

4.2 Evaluation and Results

In sport events, rankings serve two main purposes. On the one hand, they summarize the relative performances of players or teams based on outcomes in the event, so that people may easily distinguish outstanding ones from poor ones. On the other hand, rankings in

past events may indicate the outcomes of future events, and can therefore become a basis for designing future event schedules. Interestingly, we may connect these two purposes with two basic concepts in machine learning: minimizing the empirical error and minimizing the generalization error. For the first purpose, a good ranking must be consistent with available outcomes of the event, which relates to minimizing errors on training data, while for the second, a good ranking must predict well on the outcomes of future events, which is about minimizing errors on unseen data. We thus adopt these two principles to evaluate the proposed approaches, and in the context of bridge matches, they translate into the following evaluation criteria:

- **Empirical Performance**: How well do the estimated abilities and rankings fit the available match records?
- **Generalization Performance**: How well do the estimated abilities and rankings predict the outcomes of unseen matches?

Here we distinguish individuals' abilities from their ranking: Abilities give a ranking, but not vice versa. When we only have a ranking of individuals, groups' strengths are not directly available since the relation of individuals' ranks to those of groups is unclear. In contrast, if individuals' abilities are available, a group's ability can be the sum of its members'. We thus propose different error measures for abilities and rankings. Let $\{(I_1^+, I_1^-, n_1^+, n_1^-), \ldots, (I_m^+, I_m^-, n_m^+, n_m^-)\}$ be the group comparisons of interest and their outcomes. For a vector $\mathbf{v} \in R^k$ of individuals' abilities, we define the

- **Group Comparison Error**:

$$GCE(\mathbf{v}) \quad \equiv \quad \frac{\sum_{i=1}^{m} I\Big\{(n_i^+ - n_i^-)(T_i^+ - T_i^-) \leq 0\Big\}}{m},$$

where $I\{\cdot\}$ is the indicator function; T_i^+ and T_i^- are predicted group abilities of I_i^+ and I_i^-, as defined in (7). The GCE is essentially the proportion of wrongly predicted comparisons by the ability vector $\mathbf{v}$ to the m comparisons.

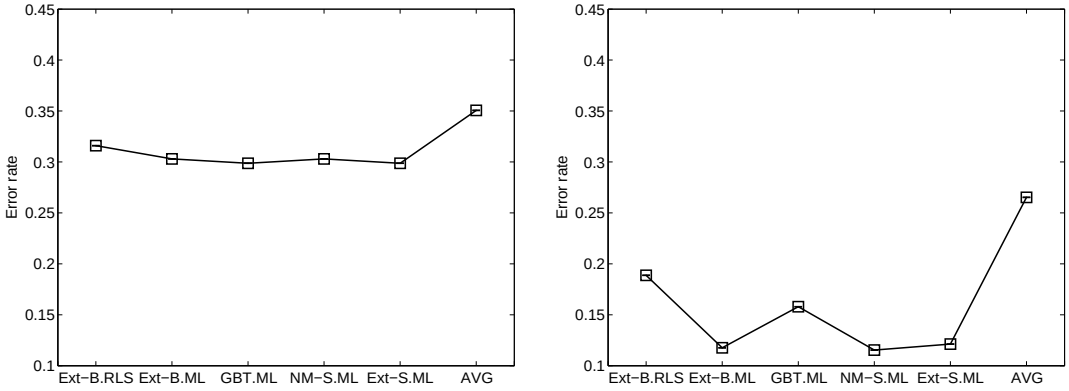
In the error measure for rankings, we use $\mathbf{r}$, a permutation of the k individuals, to denote a ranking, where r_s is the rank of individual s. Then we define the

- **Group Rank Error**:

$$GRE(\mathbf{r}) \quad \equiv \quad \frac{\sum_{i=1}^{m} \Big(I\Big\{n_i^+ > n_i^- \text{ and } U_i^+ > L_i^-\Big\} + I\Big\{n_i^+ < n_i^- \text{ and } L_i^+ < U_i^-\Big\}\Big)}{\sum_{i=1}^{m} \Big(I\Big\{U_i^+ > L_i^-\Big\} + I\Big\{L_i^+ < U_i^-\Big\}\Big)}, \tag{36}$$

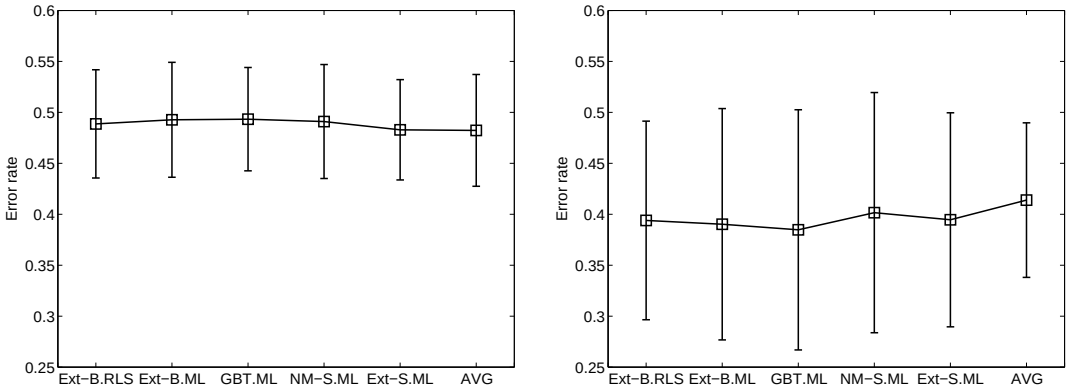
in which

$$U_i^+ \equiv \min_{j \in I_i^+} r_j, \qquad L_i^+ \equiv \max_{j \in I_i^+} r_j,$$
$$U_i^- \equiv \min_{j \in I_i^-} r_j, \qquad L_i^- \equiv \max_{j \in I_i^-} r_j.$$



(a) Group Comparison Error (b) Group Rank Error

Figure 2: Empirical performances of the six approaches.



(a) Group Comparison Error (b) Group Rank Error

Figure 3: Generalization performances of the six approaches, averaged over 50 random testing sets. Vertical bars indicate standard deviations.

Since a smaller rank indicates more strength, the U_i^+ and L_i^+ defined above represent the best and the weakest in I_i^+, respectively. The denominator in (36) is thus the number of comparisons where one group's members are all ranked higher (or lower) than the members of the competing group, and the numerator in (36) counts the number of wrong predictions, that is, comparisons in which members of the winning group are all ranked lower than those of the defeated group. In other words, GRE computes the error only on comparisons in which relative strengths of the participating groups can be clearly determined by their members' ranks, whereas GCE considers the error on all of the comparisons. From this point of view, GRE is a more conservative error measure.

Combining the two error measures with the two evaluation criteria, we conducted four sets of experiments: Empirical GCE, Empirical GRE, Generalization GCE, and Generalization GRE. We compared six approaches, including the newly proposed Ext-B.RLS, Ext-

Ext-B.RLS	Ext-B.ML	GBT.ML	NM-S.ML	Ext-S.ML	AVG
10/53	6/51	9/57	6/52	8/66	35/132

Table 2: Empirical Group Rank Errors in fraction.

B.ML, NM-S.ML, and Ext-S.ML; the generalized Bradley-Terry model GBT.ML (Huang et al., 2006b), and AVG, the simple approach (3) of summing individuals' scores, which serves as a baseline.[3] In the empirical part, we applied each approach on the entire 231 matches to estimate partnerships' abilities, and computed the two errors. Since the goal in the empirical part is to fit available records well, we set the regularization parameter μ for all approaches[4] except AVG to a small value 10^{-3}. In the generalization part, we randomly split the entire set as a training set of 162 matches and a testing set of 69 matches for 50 times. For each split, we searched for μ in $[2^5, 2^4, \ldots, 2^{-8}, 2^{-9}]$ by the Leave-One-Out (LOO) validation on the training set, estimated partnerships' abilities with the best μ, and then computed GCE and GRE on the testing set.

Results are in Figures 2 and 3 for empirical and generalization performances, respectively. In the empirical part, the four proposed approaches and GBT.ML perform obviously better than AVG, and the improvement in GRE is very significant. In particular, Ext-B.ML, NM-S.ML, and Ext-S.ML cause very small GREs, to the order of 10^{-1}. These results show that the proposed approaches are effective in fitting the available bridge match records. However, in the generalization part, all of the approaches result in poor GCEs, nearly as large as a random predictor does, and the proposed approaches did not improve over AVG. For GREs, values are smaller, but the improvements over AVG are rather marginal. In the following we give some accounts of the poor generalization performances. As mentioned in Section 4.1, each match setting can be viewed as a vector in $\{1, 0, -1\}^{69}$, in which only two dimensions have 1's, and another two have -1's. Moreover, we are using records in the qualifying stage, a round-robin tournament in which every two teams (countries) played exactly one match. Consequently, when a match is removed from the training set, the four competing partnerships of that match have no chance to meet directly during the training stage. Indirect comparisons may only be marginally useful in predicting those partnerships' competition outcomes due to the lack of transitivity. In conclusion, the outcome of a match in this bridge data set may not be well indicated by outcomes of the other matches, and therefore all of the approaches failed to generalize well.

To further study the rankings by the six approaches, we show in Table 2 their empirical GREs. Since GRE only looks at the subset of matches in which group members' ranks clearly decide groups' relative strengths, the size of this subset, that is, the denominator in GRE, may also be a performance indicator of each approach. We thus present GREs in fraction. It is clear that the ranking by AVG is able to determine the outcomes of more matches, but at the same time causes more errors. Similar results are also found in the generalization experiments. We may therefore say that the proposed approaches and

3. AVG gives individuals' abilities. We then use the same summation assumption to obtain groups' abilities for computing GCEs.
4. In order to ensure the convergence of their algorithm, Huang et al. (2006b) added to the objective function (5) what they called a "barrier term," which is also controlled by a small positive number μ (See Eq. (14) in Huang et al. 2006b). Here we simply refer to it as a regularization parameter.

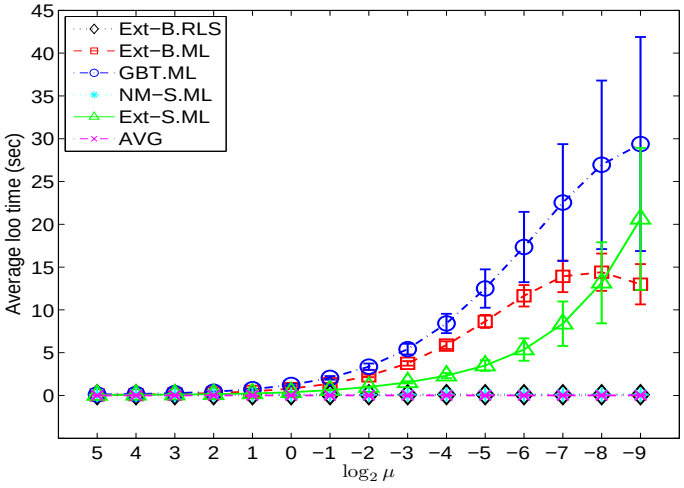


Figure 4: Average LOO time (sec) over 50 training/testing splits. Vertical bars indicate standard deviations.

GBT.ML lead to rankings with more "precision," in the sense that they may not be able to decide groups' relative performances in the majority of comparisons, but once they do, their decisions are accurate.

In addition to the efficacy of the six approaches, we also reported their efficiency. Figure 4 shows the average LOO time over the 50 training/testing splits under different values of μ. We obtained these timing results on an Intel® Core™2 Quad CPU (2.66GHz) machine with 8G main memory; the linear systems of Ext-B.RLS and NM-S.ML were solved by Gaussian Elimination. AVG,[5] Ext-B.RLS, and NM-S.ML finished LOO almost instantly under all values of μ, while Ext-B.ML, GBT.ML, and Ext-S.ML, the three approaches using iterative algorithms, took more time as μ decreased. However, for large-scale problems with a huge k or m, traditional linear system solvers may encounter memory or computational difficulties, and the efficiency of the proposed approaches requires a more thorough study.

Finally, we list the top ten partnerships ranked by Ext-B.ML in Appendix F. Most of them are famous bridge players.

5. Properties of Different Approaches

Although we distinguish binary comparisons from scored ones, they are similar in some situations. On the one hand, if two teams had a series of comparisons, the number of victories can be viewed as a team's score in a super-game. On the other hand, scores in a game might be the sum of binary outcomes; for example, scores in soccer games are total numbers of successful shots. It is therefore interesting to study the properties of different methods and their relation. Table 3 lists partnership rankings obtained by applying the six approaches to the entire set of match records. We first investigate the

5. Apparently there is no need to run LOO for AVG, which is independent of μ; we do it here only for timing comparisons.

Team (ordered by team rankings)	Partnership rankings																		#match		
	Ext-B.RLS			Ext-B.ML			GBT.ML			NM-S.ML			Ext-S.ML			AVG					
Italy (IT)	14	18	11	7	14	21	4	12	19	6	18	22	7	4	40	5	4	11	15	14	13
U.S.A.2 (US2)	57	**67**	1	53	**65**	1	39	**67**	1	53	**65**	1	54	50	1	42	**25**	2	8	17	17
U.S.A.1 (US1)	8	27	**37**	11	17	**38**	11	13	**38**	11	14	**38**	35	9	16	23	6	**10**	18	10	14
Sweden (SE)	2	43	50	2	23	55	2	10	65	2	23	56	5	8	47	1	14	38	14	13	15
India (IN)	10	35	**39**	9	29	**41**	9	28	**40**	9	28	**41**	12	28	37	19	12	**15**	15	14	13
Argentina (AR)	29	25	28	27	20	30	25	23	34	26	19	30	41	10	52	16	18	26	15	14	13
Egypt (EG)	47	23	24	51	18	22	51	22	15	51	17	21	51	18	17	37	20	3	14	20	7
	49			**52**			**50**			**52**			44			**8**			1		
Brazil (BR)	31	4	**66**	28	8	**59**	24	8	**63**	29	7	**58**	26	57	11	28	13	**31**	11	18	13
Japan (JP)	5	**65**	38	3	**67**	39	3	**68**	27	3	**68**	40	3	68	15	7	**44**	46	14	14	14
Netherlands (NL)	16	52	17	32	43	31	30	45	33	32	43	31	30	34	49	36	32	24	15	15	12
China (CN)	51	48	7	45	44	6	47	46	7	45	44	8	43	31	21	30	52	9	13	14	15
South Africa (ZA)	45	30	20	49	26	15	52	26	20	50	24	15	55	29	13	49	35	27	15	13	14
Russia (RU)	34	21	42	35	16	46	36	16	49	36	16	47	48	6	23	39	21	53	14	14	14
Portugal (PT)	22	12	58	34	10	56	29	14	60	37	10	55	33	22	56	50	29	47	14	14	14
Australia (AU)	40	55	19	42	50	19	43	53	21	42	49	20	20	45	32	43	51	40	16	11	15
New Zealand (NZ)	68	41	3	68	48	5	66	42	5	66	48	5	66	58	2	64	41	17	9	16	17
England (UK)	**9**	33	61	**12**	36	64	**17**	32	64	**13**	35	63	36	25	64	**48**	22	55	17	12	13
Canada (CA)	13	36	56	13	40	58	18	35	62	12	39	57	19	24	67	34	45	62	14	16	12
Chinese Taipei (TW)	53	62	46	63	66	57	56	61	55	64	67	59	59	65	60	57	56	66	2	12	1
	6	**26**	59	**4**	**25**	54	**6**	**37**	54	**4**	**25**	54	14	27	53	**33**	**63**	61	4	7	16
Poland (PL)	**15**	54	60	**24**	47	60	**31**	48	59	**27**	46	60	39	38	61	**58**	54	60	12	15	15
Guadeloupe (GP)	**44**	32	69	**37**	33	69	**44**	41	69	**34**	33	69	42	46	69	**65**	59	69	14	14	14
Jordan (JO)	63	64		62	61		57	58		61	62		62	63		67	68		21	21	

Table 3: Partnerships' rankings. A partnership corresponds to the same position in columns. For example, Italy's second partnership is ranked 18th, 14th, 12th, 18th, 4th and 4th by Ext-B.RLS, Ext-B.ML, GBT.ML, NM-S.ML, Ext-S.ML and AVG, respectively, and it plays 14 matches. Rankings satisfying (38) and (39) are underlined and boldfaced, respectively.

similarity between these rankings by Kendall's tau, a standard correlation coefficient that quantifies the consistency between two rankings. We computed Kendall's tau for every pair of the six rankings and present them in Table 4, which indicates roughly three groups: Ext-B.RLS, Ext-B.ML, GBT.ML and NM-S.ML give similar rankings; the one by Ext-S.ML is quite different, while AVG seems to be uncorrelated with the others. We then measure the distance between two groups of rankings $g1$ and $g2$: For each partnership,

$$d(\text{ranks by } g1, \text{ranks by } g2) \equiv \begin{cases} \min(\text{ranks by } g2) - \max(\text{ranks by } g1) & \text{if ranks by } g1 \text{ are all smaller,} \\ \min(\text{ranks by } g1) - \max(\text{ranks by } g2) & \text{if ranks by } g2 \text{ are all smaller,} \\ 0 & \text{otherwise.} \end{cases} \tag{37}$$

	Ext-B.RLS	Ext-B.ML	GBT.ML	NM-S.ML	Ext-S.ML	AVG
Ext-B.RLS	1.00	0.84	0.79	0.82	0.50	0.44
Ext-B.ML	0.84	1.00	0.87	0.97	0.61	0.49
GBT.ML	0.79	0.87	1.00	0.86	0.62	0.53
NM-S.ML	0.82	0.97	0.86	1.00	0.60	0.49
Ext-S.ML	0.50	0.61	0.62	0.60	1.00	0.50
AVG	0.44	0.49	0.53	0.49	0.50	1.00

Table 4: Kendall's tau (correlation coefficients).

For example, from Table 3 the second partnership of U.S.A.2 is ranked 67th/65th/67th/65th by Ext-B.RLS/Ext-B.ML/GBT.ML/NM-S.ML and 25th by AVG. Therefore,

$$d(\{67, 65, 67, 65\}, 25) = \min(67, 65, 67, 65) - 25 = 40.$$

Checking all 69 partnerships' ranks gives

$$|d(\{\text{Ext-B.RLS,Ext-B.ML,GBT.ML,NM-S.ML}\},\ \text{Ext-S.ML}) \geq 20| = 6, \tag{38}$$

$$|d(\{\text{Ext-B.RLS,Ext-B.ML,GBT.ML,NM-S.ML}\},\ \text{AVG}) \geq 20| = 11. \tag{39}$$

In Table 3 we respectively underline and boldface partnerships satisfying (38) and (39). The eleven ranks satisfying (39) shows that AVG's ranking is closer to the team ranking:[6] Partnerships satisfying (39) have higher ranks than those by the others when the team ranks are high, but have the opposite when the team ranks are low. This observation indicates that AVG may fail to identify weak (strong) individuals from strong (weak) groups.

The above results suggest that approaches based on different types of comparisons may produce similar rankings, such as Ext-B.ML and NM-S.ML, while those based on the same type of outcomes may lead to diverse results, such as NM-S.ML and Ext-S.ML. Therefore, in the next two subsections we study their formulations and obtain the following results:

- When all n_i's are equal, that is, the number of games or the total score in every group comparison is the same, and estimated group abilities are approximately even, Ext-B.ML and NM-S.ML give similar rankings.
- When all n_i's are equal, Ext-B.RLS is more sensitive than Ext-B.ML and NM-S.ML to extreme outcomes ($n_i^+ \approx 0$ or $n_i^+ \approx n_i$).
- For the two scored-outcome approaches, extreme outcomes have a greater impact on NM-S.ML than on Ext-S.ML.

5.1 Comparing Binary- and Scored-outcome Approaches

Experimental results in Section 4 show that the binary-outcome approach Ext-B.ML and the scored-outcome approach NM-S.ML give very similar rankings. By analyzing their optimization problems, we find that

6. Recall that in the beginning of Section 4, we mentioned that all teams, after the qualifying stage was over, were ranked according to their total VPs gained in the tournament.

Claim 1 *If all n_i's are equal and the optimal* $\mathbf{v}$ *for Ext-B.ML satisfies*

$$T_i^+ \approx T_i^- \quad \forall i,$$

then Ext-B.ML and NM-S.ML give very close rankings.

The proof is in Appendix E. For the bridge data used in Section 4, n_i's are two rival teams' total VPs and are mostly 30; the average $|T_i^+ - T_i^-|$ from the optimal $\mathbf{v}$ for Ext-B.ML is 0.3983.

However, in applications where n_i's are unequal, these two approaches may give different results. Clearly, they use different approximations:

$$\frac{e^{T_i^+}}{e^{T_i^-}} \approx \frac{n_i^+}{n_i^-} \qquad \text{and} \qquad T_i^+ - T_i^- \approx n_i^+ - n_i^-. \tag{40}$$

One considers the ratio, which is independent from the values of n_i's, but the other considers the difference, whose value scales with those of n_i's. Therefore, the estimate by NM-S.ML may be more biased than Ext-B.ML to fit comparison outcomes with large n_i.

Another issue is the small but perceivable dissimilarity of the ranking by Ext-B.RLS from those by Ext-B.ML and NM-S.ML, as revealed in the empirical GREs in Table 2 and the Kendall's tau in Table 4. Investigating them more carefully, we find that

$$|d(\text{Ext-B.RLS}, \{\text{Ext-B.ML, NM-S.ML}\}) \geq 10| = 8, \tag{41}$$

where the distance is defined in (37). Interestingly, five of these eight partnerships played matches where weak teams beat strong teams by an extreme amount, such as Netherlands beating U.S.A.2 by 25:0, and Ext-B.RLS ranks them higher than Ext-B.ML and NM-S.ML do. This result suggests that Ext-B.RLS is vulnerable to even only few extreme outcomes so as to change the overall ranking. We verify this property by comparing the estimates by Ext-B.RLS and NM-S.ML. Suppose $n_i = n \ \forall i$ (which is the case here), and then according to (16), the ability estimate of individual s by Ext-B.RLS is

$$v_s = \sum_{i=1}^{m} A_{si}\big(\log n_i^+ - \log n_i^-\big) = \sum_{i=1}^{m} A_{si}\big(\log n_i^+ - \log(n - n_i^+)\big),$$

where $A = \big(G^T G + \mu I\big)^{-1} G^T$. To check the sensitivity of v_s with respect to the change of n_i^+, we calculate

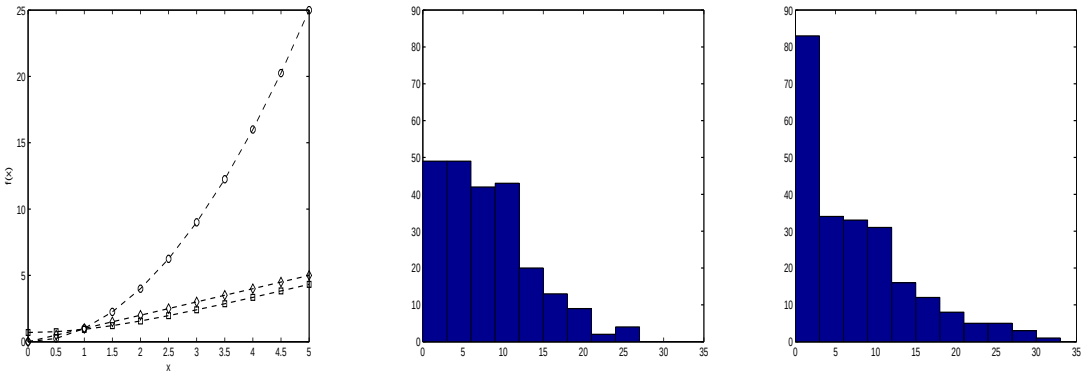
$$\frac{\partial v_s}{\partial n_i^+} = A_{si}\big(\frac{1}{n_i^+} + \frac{1}{n - n_i^+}\big) = \frac{n A_{si}}{n_i^+(n - n_i^+)}.$$

Clearly, the estimate v_s is more sensitive to extreme values of n_i^+, that is, $n_i^+ \approx 0$ or $n_i^+ \approx n$. However, for NM-S.ML we have

$$v_s = \sum_{i=1}^{m} A_{si}(n_i^+ - n_i^-) = \sum_{i=1}^{m} A_{si}(2n_i^+ - n)$$

and

$$\frac{\partial v_s}{\partial n_i^+} = 2A_{si}.$$



(a) Loss function curves. Circles: x^2. Diamonds: x. Squares: $\log(1+\cosh(x))$.

(b) Error histogram for NM-S.ML

(c) Error histogram for Ext-S.ML

Figure 5: Error function curves and histograms. The x-axis of histograms is $|T_i^+ - T_i^- - (n_i^+ - n_i^-)|$.

That is, different values of n_i^+ have equal impact on the estimate by NM-S.ML.

In conclusion, when n_i remains a constant and the estimates by Ext-B.ML have $T_i^+ \approx T_i^- \ \forall i$, NM-S.ML and Ext-B.ML give similar estimates, which are less sensitive than that by Ext-B.RLS to extreme outcomes. When n_i's are unequal, the discussion in (40) indicates that NM-S.ML is more affected than Ext-B.ML.

5.2 Comparing the Two Scored-outcome Approaches

As shown in (38), the ranking by Ext-S.ML is rather diverse from those by Ext-B.RLS, Ext-B.ML, and NM-S.ML. We explore this issue by first re-writing the objective functions of NM-S.ML and Ext-S.ML respectively as

$$\min_{\mathbf{v}} \quad \sum_{i=1}^{m} \left(T_i^+ - T_i^- - (n_i^+ - n_i^-) \right)^2 + \mu \sum_{s=1}^{k} v_s^2$$

and

$$\min_{\mathbf{v}} \quad \sum_{i=1}^{m} \log \left(1 + \cosh \left(T_i^+ - T_i^- - (n_i^+ - n_i^-) \right) \right) + \mu \sum_{s=1}^{k} (e^{v_s} + e^{-v_s}),$$

where cosh is the *hyperbolic cosine* function. Although these two formulations are derived to maximize the likelihood, they can be viewed as minimizing estimation errors

$$\left| T_i^+ - T_i^- - (n_i^+ - n_i^-) \right|$$

with two different loss functions. As μ is small, we ignore the effect of the regularization term. It is easy to show that as $\left|T_i^+ - T_i^- - (n_i^+ - n_i^-)\right| \to \infty$,

$$\frac{\log\left(1 + \cosh\left(T_i^+ - T_i^- - (n_i^+ - n_i^-)\right)\right)}{\left|T_i^+ - T_i^- - (n_i^+ - n_i^-)\right|} \to 1.$$

To show the behaviors of the three functions: x^2, x and $\log(1 + \cosh(x))$, we plot their curves in Figure 5(a). One can see that $\log(1 + \cosh(x))$ increases almost linearly with x. In the machine learning community, it is well known that quadratic loss functions may lead to a very different estimation from linear ones. The reason is that quadratic loss functions penalize large errors more severely than linear ones do; estimations are thus dominated by even only few extreme observations, and as a side effect, may cause quite a few moderate errors. In contrast, estimations under linear loss functions may allow several large errors in order to make most errors small. Figures 5(b) and 5(c) are histograms of $\left|T_i^+ - T_i^- - (n_i^+ - n_i^-)\right|$ for NM-S.ML and Ext-S.ML, respectively; we see clearly the aforementioned two error patterns: Compared with NM-S.ML, Ext-S.ML has a lot more errors in the first bin and also some in the last two. In addition, we find that the empirical GRE of Ext-S.ML in Section 4.2 is highly related to its error pattern: Among the 24 correct rank predictions[7] produced by Ext-S.ML but not by NM-S.ML, twelve have $\left|T_i^+ - T_i^- - (n_i^+ - n_i^-)\right|$ smaller than 3 (the first bin of histograms); NM-S.ML has no $\left|T_i^+ - T_i^- - (n_i^+ - n_i^-)\right|$ larger than 27 (the last two bins of histograms) while Ext-S.ML has four, among which the partnerships satisfying (38) participate in three. Interestingly, the two types of loss functions seem to reflect two different ranking criteria: one focuses more on performances against extreme opponents, so wins over strong opponents and losses to weak opponents greatly influence the ranking; the other is less sensitive to extreme outcomes and treat comparisons more evenly. Consequently, deciding which loss function, and hence which approach to use may eventually be contingent on game-specific factors and subjective preferences.

6. Multi-class Classification

Multi-class classification using coding matrices (Dietterich and Bakiri, 1995; Allwein et al., 2001) is a general scheme to decompose a problem into several two-class problems. The widely-used methods "one-against-one" and "one-against-the rest" are special cases of this framework. The decomposition is usually specified by a coding matrix $G \in \{1, 0, -1\}^{m \times k}$, where k is the number of classes and m is the number of two-class problems. Each row of G describes how the k classes are separated to two groups: those with 1 are in one group while those with -1 are in the other; those with 0 are not used in this two-class problem. The coding matrix in Table 5 illustrates four common types of codes: one-against-one, one-against-all, dense, and sparse; their definitions are given by Items 1 to 4 on Page 2209. At the training stage, m binary classifiers are trained for the m two-class problems. For an unlabeled instance, its label is predicted by combining results of the m binary classifiers.

There are several schemes for deciding the final prediction. Dietterich and Bakiri (1995) proposed choosing the class whose column in G has the smallest distance to the m binary decisions on the instance. This method can correct errors made by some decision rules,

7. Correct rank predictions are the denominator of GRE minus its numerator.

One-against-one	$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 \end{bmatrix}$
One-against-the rest	$\begin{bmatrix} -1 & -1 & -1 & 1 & -1 & -1 & -1 & -1 \end{bmatrix}$
Dense	$\begin{bmatrix} 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \end{bmatrix}$
Sparse	$\begin{bmatrix} 1 & -1 & 0 & 0 & 1 & 0 & 0 & -1 \end{bmatrix}$
$\vdots$	$\begin{bmatrix} \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$

Table 5: A coding matrix ($k = 8$). The four rows illustrate four types of codes.

and thus is called *error-correcting output codes* (ECOC). Allwein et al. (2001) proposed a more general framework, the *loss-based decoding*, which exploits not only binary decisions, but also *decision values* of binary classifiers. In particular, they adopted the *exponential loss-based decoding* (EXPLOSS): let $\hat{f}_i$ be the decision function of the ith binary classifier, and $\hat{f}_i(\mathbf{x}) > 0$ (< 0) specifies that an instance $\mathbf{x}$ to be in classes of I_i^+ (I_i^-). Then,

$$\text{predicted label } \equiv \arg\min_s \Big(\sum_{i=1}^{m} e^{-G_{is}\hat{f}_i} \Big).$$

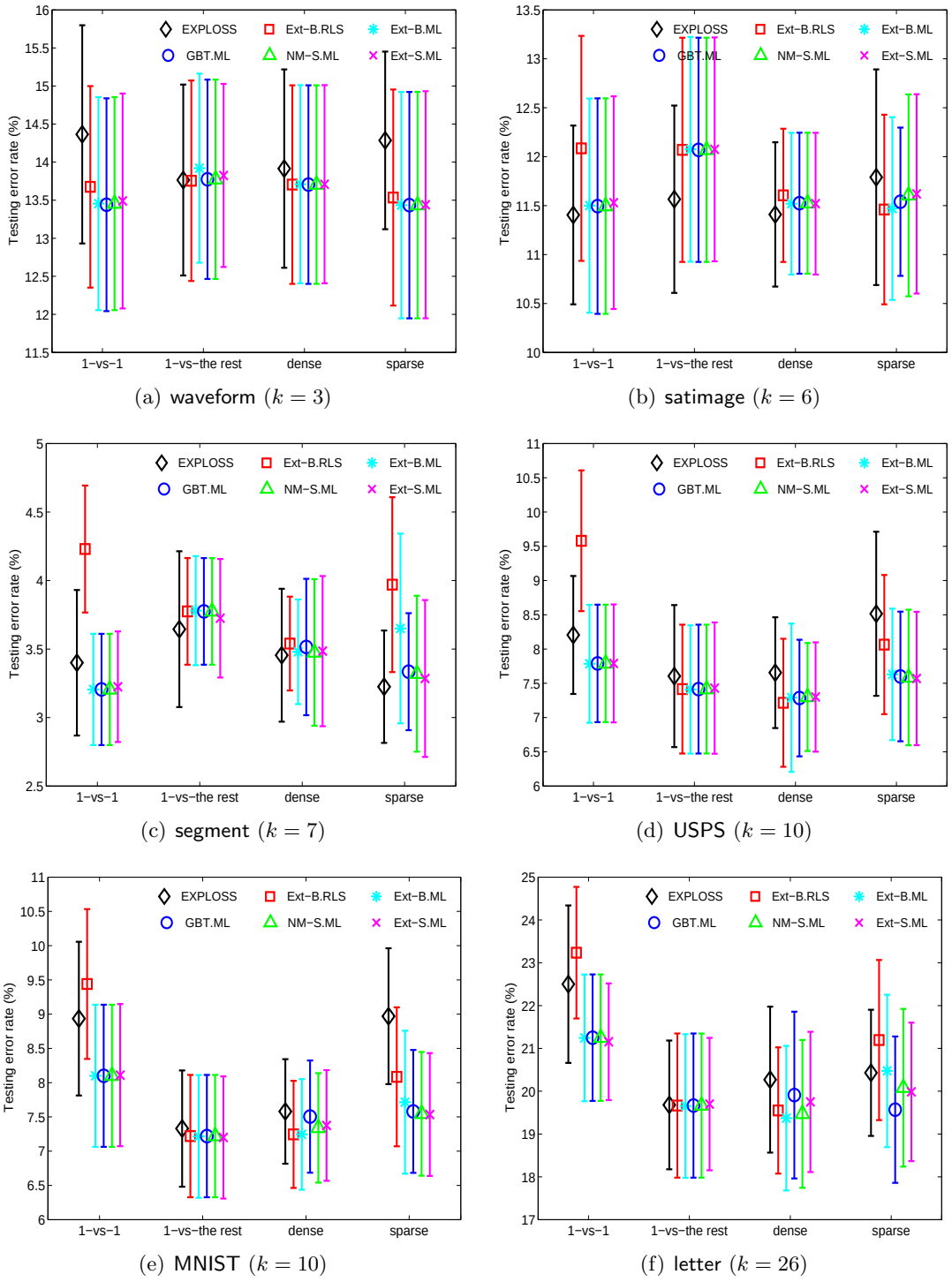
If $G_{is} = 1$ and $\hat{f}_i(\mathbf{x})$ says $s \in I_i^+$, then $e^{-G_{is}\hat{f}_i}$ gives a small loss. By using decision values, the loss-based decoding incorporates the confidence of each binary prediction in making the final decision.

Table 5 is in the same format as our "comparison setting matrix" G defined in (12) and (13). Huang et al. (2006b) (GBT.ML) thus consider classes as individuals and two-class problems as group comparisons; the 1's and -1's in the ith row of G correspond to I_i^+ and I_i^-, respectively. The group competition results n_i^+ and n_i^- are assumed to be available from two-class classifiers. For an unlabeled instance, classes are ranked according to their estimated "abilities" and the highest one (with the largest ability) serves as the prediction. All of our newly proposed models can be applied in the same way, but there are two minor issues. Firstly, all of our proposed methods except Ext-B.RLS assume that group comparisons are independent. This property does not hold for multi-class classification since two-class classifiers involving the same classes share training data. Huang et al. (2006b) pointed out that GBT.ML can be interpreted as minimizing the Kullback-Leibler distance between the model and the observations. It is easy to see that their argument also applies to Ext-B.ML but not to NM-S.ML nor Ext-S.ML. Secondly, the n_i^+ and n_i^- given by two-class classifiers are real values, for which the binary-outcome approaches, according to their definition, may not be suitable. Despite of these minor issues, as we will show, our proposed methods perform quite well in practice.

We compare our methods with EXPLOSS and GBT.ML on six real data sets: waveform, satimage, segment, USPS, MNIST, and letter; numbers of classes range from 3 to 26. The settings of experiments are the same as those in Huang et al. (2006b). We use the 20 subsets of 800 training and 1,000 testing instances[8] and consider the same four types of coding matrices:

1. One-against-one: $|I_i^+| = |I_i^-| = 1,\ i = 1, \ldots, k(k-1)/2$.

8. Available at `http://www.csie.ntu.edu.tw/~cjlin/papers/svmprob/data`.



(a) waveform ($k = 3$) (b) satimage ($k = 6$)

(c) segment ($k = 7$) (d) USPS ($k = 10$)

(e) MNIST ($k = 10$) (f) letter ($k = 26$)

Figure 6: Testing error rates on the 800-training-1000-testing data sets by six approaches under four codes: one-against-one (1-vs-1), one-against-the rest (1-vs-the rest), dense, and sparse. Vertical bars indicate standard deviations.

2. One-against-all: $|I_i^+| = 1,\ |I_i^-| = k-1,\ i = 1, \ldots, k.$

3. Dense: $|I_i^+| = |I_i^-| = k/2,\ \forall i;\ m = [10 \log_2 k].$

4. Sparse: $E(|I_i^+|) = E(|I_i^-|) = k/4,\ \forall i;\ m = [15 \log_2 k].$

$[x]$ rounds a real number x to its nearest integer. We choose support vector machines (SVM) (Boser et al., 1992) with the RBF (Radial Basis Function) kernel $e^{-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2}$ as the two-class classifier, where $\mathbf{x}_i$ and $\mathbf{x}_j$ are two training instances. An improved version (Lin et al., 2007) of Platt (2000) generates n_i^+ and $n_i^- = 1 - n_i^+$ from SVM decision values. We implement our methods by modifying LIBSVM (Chang and Lin, 2001). For all of the 20 subsets, we select SVM parameters by cross validation before testing. Figures 6(a)-6(f) report the average testing error rates and standard deviations of the six methods: EXPLOSS, Ext-B.RLS, Ext-B.ML, GBT.ML, NM-S.ML and Ext-S.ML. Each figure summarizes the results on one data set by six groups of colored error bars, which represent the error rates of the six methods under the four types of codes. We can see that EXPLOSS (black diamond) and Ext-B.RLS (red square) perform worse than the others under the one-against-one and the sparse codes as k becomes large, while GBT.ML, Ext-B.ML, NM-S.ML and Ext-S.ML are almost equally good. Regarding the performances of the four types of codes, one-against-one and sparse are less effective for large values of k, an observation consistent with the results in (Huang et al., 2006b). Recall that in Section 4.2 Ext-S.ML behaves differently from the others, but here its predictions are similar to those of NM-S.ML and Ext-B.ML. The reason is that the n_i^+ and n_i^- produced by (Lin et al., 2007) are probabilities satisfying $n_i^+ + n_i^- = 1$, so values of $|T_i^+ - T_i^- - (n_i^+ - n_i^-)|$ are mostly small and the difference between quadratic and linear loss functions is negligible. Results here suggest that the proposed methods are useful for multi-class classification with coding matrices.

7. Conclusions

We propose new and useful methods to rank individuals from group comparisons. For comparisons with binary outcomes, earlier work solves non-convex problems, but here convex formulations with easy solution procedures are developed. For scored outcomes, our formulations are probably the first for this type of problems. Experiments show that the proposed approaches give reasonable partnership rankings from bridge records and perform effectively in multi-class classification. We give theoretical accounts for behaviors of proposed approaches, which demonstrate how different models reflect diverse ranking criteria. We also develop techniques to evaluate different rankings, which may be used in other ranking tasks.

Appendix A. Derivation of (10) from (8)

$$P(Y_i^+ - Y_i^- > 0) \equiv \int_{-\infty}^{\infty} \int_{y^-}^{\infty} de^{-e^{-(y^+ - T_i^+)}} de^{-e^{-(y^- - T_i^-)}}. \tag{42}$$

Let

$$x^+ \equiv e^{-(y^+ - T_i^+)} \text{ and } x^- \equiv e^{-(y^- - T_i^-)}.$$

Consequently,

$$de^{-e^{-(y^+-T_i^+)}} = -e^{-x^+}dx^+ \text{ and } de^{-e^{-(y^--T_i^-)}} = -e^{-x^-}dx^-.$$

Then,

$$\begin{aligned}(42) &= \int_0^\infty -e^{-x^-}\int_0^{x^- e^{T_i^+ - T_i^-}} -e^{-x^+}dx^+dx^- \\ &= \frac{e^{T_i^+}}{e^{T_i^+}+e^{T_i^-}}.\end{aligned}$$

Appendix B. Proof of Theorem 1

If $rank(G) < k$, G^TG is obviously not invertible; if $rank(G) = k$, the Singular Value Decomposition of G can be written as

$$G = U\Lambda V^T,$$

where $U \in R^{m\times k}$ and $V \in R^{k\times k}$ are orthonormal and $\Lambda \in R^{k\times k}$ is diagonal with

$$\Lambda_{ii} \neq 0,\ i = 1, \ldots, k.$$

Therefore,

$$G^TG = V\Lambda U^T U\Lambda V^T = V\Lambda^2 V^T$$

is invertible.

Appendix C. Proof of Theorem 2

We first rewrite $l(\mathbf{v})$ as

$$l(\mathbf{v}) = -\sum_{i=1}^m (n_i^+ T_i^+ + n_i^- T_i^-) + \sum_{i=1}^m n_i \log(e^{T_i^+} + e^{T_i^-}).$$

The first summation is obviously convex. For the second summation, by using Hölder's inequality we have

$$\begin{aligned}&\sum_{i=1}^m n_i \log\left(e^{\lambda T_i^+ + (1-\lambda)\tilde{T}_i^+} + e^{\lambda T_i^- + (1-\lambda)\tilde{T}_i^-}\right) \\ =\ &\sum_{i=1}^m n_i \log\left((e^{T_i^+})^\lambda (e^{\tilde{T}_i^+})^{1-\lambda} + (e^{T_i^-})^\lambda (e^{\tilde{T}_i^-})^{1-\lambda}\right) \\ \leq\ &\sum_{i=1}^m n_i \log\left(e^{T_i^+} + e^{T_i^-}\right)^\lambda \left(e^{\tilde{T}_i^+} + e^{\tilde{T}_i^-}\right)^{1-\lambda} \\ =\ &\sum_{i=1}^m n_i\lambda \log\left(e^{T_i^+} + e^{T_i^-}\right) + \\ &\sum_{i=1}^m n_i(1-\lambda) \log\left(e^{\tilde{T}_i^+} + e^{\tilde{T}_i^-}\right) \qquad (43)\end{aligned}$$

for any $\mathbf{v}, \tilde{\mathbf{v}}$ and $\lambda \in (0,1)$, and the equality holds if and only if

$$T_i^+ - T_i^- = \tilde{T}_i^+ - \tilde{T}_i^- \ \forall i,$$

which can be re-written as

$$G(\mathbf{v} - \tilde{\mathbf{v}}) = \mathbf{0}. \tag{44}$$

If $rank(G) = k$, then (44) holds if and only if $\mathbf{v} = \tilde{\mathbf{v}}$, so $l(\mathbf{v})$ is strictly convex. If $l(\mathbf{v})$ is strictly convex, then the equality in (43) holds if and only if $\mathbf{v} = \tilde{\mathbf{v}}$, so

$$G(\mathbf{v} - \tilde{\mathbf{v}}) = \mathbf{0} \Leftrightarrow \mathbf{v} = \tilde{\mathbf{v}},$$

which implies $rank(G) = k$.

Appendix D. Proof of Theorem 3

It is easy to verify that the level sets of $l(\mathbf{v})$ are bounded. Since $l(\mathbf{v})$ is strictly convex, it then attains a unique global minimum. To prove the convergence of Algorithm 1, we first show that if $\partial l(\mathbf{v})/\partial v_s \neq 0$, then minimizing (22) leads to

$$l(\mathbf{v} + \boldsymbol{\delta}) < l(\mathbf{v}). \tag{45}$$

From (23), if the optimal δ_s for (22) is zero, then

$$\frac{B_s + \sqrt{B_s^2 + 4\mu A_s e^{-v_s}}}{2A_s} = 1,$$

which implies

$$4A_s(\mu e^{-v_s} - A_s + B_s) = -4A_s \frac{\partial l(\mathbf{v})}{\partial v_s} = 0. \tag{46}$$

Since $A_s \neq 0$ throughout iterations, (46) implies $\partial l(\mathbf{v})/\partial v_s = 0$. Thus if $\partial l(\mathbf{v})/\partial v_s \neq 0$, the optimal $\delta_s \neq 0$. With (22) $= 0$ if $\delta_s = 0$, (45) follows.

Next we show that the sequence $\{\mathbf{v}^t\}$ generated by Algorithm 1 is bounded. If not, there must exist j such that $|v_j^t| \to \infty$. Then

$$\begin{aligned} l(\mathbf{v}^t) &\geq \mu \sum_{s=1}^{k} (e^{v_s^t} + e^{-v_s^t}) \\ &= \mu \sum_{s=1}^{k} (e^{|v_s^t|} + e^{-|v_s^t|}) \\ &\geq \mu e^{|v_j^t|} + e^{-|v_j^t|} \\ &\to \infty, \end{aligned}$$

which contradicts the fact that

$$l(\mathbf{v}^0) > l(\mathbf{v}^t) \ \forall t.$$

Since $\{\mathbf{v}^t\}$ is bounded, it has limit points. For any limit point $\mathbf{v}^*$, there is an infinite set $\bar{N}$ such that

$$\lim_{t \in \bar{N}, t \to \infty} \mathbf{v}^t = \mathbf{v}^*.$$

Since $\mathbf{v}$ is finite dimensional, there is one component s updated in an infinite set $N \subset \bar{N}$:

$$(t \bmod k) + 1 = s \text{ for } t \in N.$$

Because $l(\mathbf{v})$ is convex, to prove that $\mathbf{v}^*$ is a global minimum, it suffices to show that

$$\frac{\partial l(\mathbf{v}^*)}{\partial v_s} = 0 \text{ for } s = 1, \ldots, k. \tag{47}$$

Suppose the contrary is true, then among $s, s+1, \ldots, k, 1, \ldots, s-1$, there is $\bar{s}$ such that

$$\frac{\partial l(\mathbf{v}^*)}{\partial v_s} = \cdots = \frac{\partial l(\mathbf{v}^*)}{\partial v_{\bar{s}-1}} = 0, \ \frac{\partial l(\mathbf{v}^*)}{\partial v_{\bar{s}}} \neq 0. \tag{48}$$

From (45), updating $v_{\bar{s}}^*$ by (23) yields $\mathbf{v}^{*+1} \neq \mathbf{v}^*$ and

$$l(\mathbf{v}^{*+1}) < l(\mathbf{v}^*).$$

We have that $\partial l(\mathbf{v}^*)/\partial v_s = 0$ implies

$$\frac{B_s + \sqrt{B_s^2 + 4\mu A_s^* e^{-v_s^*}}}{2A_s^*} = 1,$$

where A_s^* is defined according to (24) and B_s is a constant independent of $\mathbf{v}$. Therefore,

$$\begin{aligned} \lim_{\substack{t \in N, \\ t \to \infty}} v_s^{t+1} &= \lim_{\substack{t \in N, \\ t \to \infty}} \left(v_s^t + \log \frac{B_s + \sqrt{B_s^2 + 4\mu A_s^t e^{-v_s^t}}}{2A_s^t} \right) \\ &= v_s^* + \log \frac{B_s + \sqrt{B_s^2 + 4\mu A_s^* e^{-v_s^*}}}{2A_s^*} \\ &= v_s^*, \end{aligned} \tag{49}$$

and

$$\lim_{t \in N, t \to \infty} \mathbf{v}^{t+1} = \lim_{t \in N, t \to \infty} \mathbf{v}^t = \mathbf{v}^*. \tag{50}$$

Let $\bar{t}$ be the iteration corresponding to $\bar{s}$. Using (48), a similar derivation to (49) and (50) shows that

$$\lim_{\substack{t \in N, \\ t \to \infty}} \mathbf{v}^{t+1} = \cdots = \lim_{\substack{t \in N, \\ t \to \infty}} \mathbf{v}^{\bar{t}} = \mathbf{v}^* \text{ and } \lim_{\substack{t \in N, \\ t \to \infty}} \mathbf{v}^{\bar{t}+1} = \mathbf{v}^{*+1};$$

consequently,

$$\lim_{t \in N, t \to \infty} l(\mathbf{v}^{\bar{t}+1}) = l(\mathbf{v}^{*+1}) < l(\mathbf{v}^*),$$

which contradicts the fact that

$$l(\mathbf{v}^*) \leq \cdots \leq l(\mathbf{v}^{t+1}) \leq l(\mathbf{v}^t).$$

Thus (47) holds for all limit points. Since $l(\mathbf{v})$ is strictly convex, every limit point is the unique global minimum. Moreover, the sequence $\{\mathbf{v}^t\}$ is bounded, so it globally converges to the global minimum.

Appendix E. Proof of Claim 1

From (29) it is clear that the ranking by NM-S.ML is invariant to the scale of n_i; we thus assume

$$n_i^+ + n_i^- = 2, \forall i.$$

Then (26) can be rewritten as

$$\min_{\mathbf{v}} \sum_{i=1}^{m} \left((T_i^+ - T_i^-)^2 - (4n_i^+ - 4)(T_i^+ - T_i^-) \right).$$

For Ext-B.ML, as μ is small and can be ignored, we consider the objective function in (17), which can be re-written as

$$\sum_{i=1}^{m} -n_i^+(T_i^+ - T_i^-) + n_i \log(e^{T_i^+ - T_i^-} + 1) \tag{51}$$

$$= \sum_{i=1}^{m} -n_i^+(T_i^+ - T_i^-) + 2\Big(\log 2 + \frac{1}{2}(T_i^+ - T_i^-) + \frac{1}{8}(T_i^+ - T_i^-)^2 + O\big((T_i^+ - T_i^-)^3\big) \Big) \tag{52}$$

$$\approx \frac{1}{8} \sum_{i=1}^{m} \left((T_i^+ - T_i^-)^2 - (4n_i^+ - 4)(T_i^+ - T_i^-) \right).$$

From (51) to (52) we use the Taylor expansion of the function $\log(e^x + 1)$ at $x = 0$ and the assumption that $T_i^+ \approx T_i^- \ \forall i$. Therefore, the rankings by NM-S.ML and Ext-B.ML are similar.

Appendix F. Top 10 Partnerships by Ext-B.ML

Team	Players	
U.S.A.2	Eric Greco	Geoff Hampson
Sweden	Peter Bertheau	Fredrik Nystrom
Japan	Yoshiyuki Nakamura	Yasuhiro Shimizu
Chinese Taipei	Chih-Kuo Shen	Jui-Yiu Shih
New Zealand	Tom Jacob	Malcolm Mayer
China	Zhong Fu	Jie Zhao
Italy	Norberto Bocchi	Giorgio Duboin
Brazil	Gabriel Chagas	Miguel Villas-boas
India	Subhash Gupta	Rajeshwar Tewari
Portugal	Jorge Castanheira	Sofia Pessoa

References

Erin L. Allwein, Robert E. Schapire, and Yoram Singer. Reducing multiclass to binary: a unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1: 113–141, 2001. ISSN 1533-7928.

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

Bernhard E. Boser, Isabelle Guyon, and Vladimir Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152. ACM Press, 1992.

Ralph A. Bradley and Milton E. Terry. The rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324–345, 1952.

Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

John N. Darroch and Douglas Ratcliff. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43(5):1470–1480, 1972.

Herbert A. David. *The method of paired comparisons*. Oxford University Press, second edition, 1988.

Thomas G. Dietterich and Ghulum Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995.

Arpad E. Elo. *The Rating of Chessplayers, Past and Present*. Arco Pub., New York, 2nd edition, 1986.

Mark E. Glickman. *Paired comparison models with time-varying parameters*. PhD thesis, Department of Statistics, Harvard University, 1993.

Joshua Goodman. Sequential conditional generalized iterative scaling. In *ACL*, pages 9–16, 2002.

Trevor Hastie and Robert Tibshirani. Classification by pairwise coupling. *The Annals of Statistics*, 26(1):451–471, 1998.

Ralf Herbrich and Thore Graepel. TrueSkill$^{\text{TM}}$: A Bayesian skill rating system. In *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, 2007.

Tzu-Kuo Huang, Chih-Jen Lin, and Ruby C. Weng. Ranking individuals by group comparisons. In *Proceedings of the Twenty Third International Conference on Machine Learning (ICML)*, 2006a.

Tzu-Kuo Huang, Ruby C. Weng, and Chih-Jen Lin. Generalized Bradley-Terry models and multi-class probability estimates. *Journal of Machine Learning Research*, 7:85–115, 2006b. URL `http://www.csie.ntu.edu.tw/~cjlin/papers/generalBT.pdf`.

David R. Hunter. MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, 32:386–408, 2004.

Edwin T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4): 620–630, 1957a.

Edwin T. Jaynes. Information theory and statistical mechanics ii. *Physical Review*, 108(2): 171–190, 1957b.

Hsuan-Tien Lin, Chih-Jen Lin, and Ruby C. Weng. A note on Platt's probabilistic outputs for support vector machines. *Machine Learning*, 68:267–276, 2007. URL `http://www.csie.ntu.edu.tw/~cjlin/papers/plattprob.pdf`.

Joshua E. Menke and Tony R. Martinez. A Bradley-Terry artificial neural network model for individual ratings in group competitions. *Neural Computing and Applications*, 2007. To appear.

Thomas Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, MIT, 2001.

Stephen Della Pietra, Vincent Della Pietra, and John Lafferty. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):380–393, 1997.

John Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, Cambridge, MA, 2000. MIT Press.

Bianca Zadrozny. Reducing multiclass to binary by coupling probability estimates. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 1041–1048. MIT Press, Cambridge, MA, 2002.