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

Abstract

This paper proposes new approaches to rank individuals from their group competition results. Many real-world problems are of this type. For example, ranking players from team games is important in some sports. We propose an exponential model to solve such problems. To estimate individual rankings through the proposed model we introduce two convex minimization formulas with easy and efficient solution procedures. Experiments on real bridge records and multi-class classification demonstrate the viability of the proposed model.

1. Introduction

We address an interesting problem of estimating individuals' abilities from their group competition 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. In some sports such as bridge, competition results 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 every partnership's raw score, depending on different boards, does not indicate a partnership's ability. Finding reasonable individual rankings using all group competition records is thus a challenging task. Another application in machine learning/statistics is multi-class probability estimates by error-correcting codes (Huang et al., 2005). Classification by error-correcting codes (Dietterich & Bakiri, 1995; Allwein et al., 2001) involves several two-class problems, each of which is considered as the competition between two disjoint subsets of class labels. Individuals' abilities are then an instance's probabilities in different classes.

Huang et al. (2005) propose a generalized Bradley-Terry model to solve this problem. They consider k individuals $\{1, \ldots, k\}$ having m competitions. The ith competition involves a subset I_i, which is separated to two disjoint teams, I_i^+ and I_i^-. They play $n_i = n_i^+ + n_i^-$ games, and we assume that I_i^+ and I_i^- win n_i^+ and n_i^- times, respectively. By representing the k individuals' abilities as a non-negative vector $\mathbf{p} \in R^k$, Huang et al. (2005) propose the following model:

$$P(I_i^+ \text{ beats } I_i^-) = \frac{\sum_{j:j\in I_i^+} p_j}{\sum_{j:j\in I_i} p_j}. \tag{1}$$

This model extends the Bradley-Terry model (1952) for pairwise comparison (i.e., games between any two individuals):

$$P(\text{individual } i \text{ beats individual } j) = \frac{p_i}{p_i + p_j}. \tag{2}$$

Huang et al. (2005) estimate $\mathbf{p}$ by minimizing the negative log-likelihood of (1):

$$\min_{\mathbf{p}} \; -\sum_{i=1}^{m} \left(n_i^+ \log \frac{\sum_{j:j\in I_i^+} p_j}{\sum_{j:j\in I_i} p_j} + n_i^- \log \frac{\sum_{j:j\in I_i^-} p_j}{\sum_{j:j\in I_i} p_j} \right)$$
$$\text{subject to} \quad \sum_{j=1}^{k} p_j = 1, 0 \le p_j, j = 1, \ldots, k. \tag{3}$$

They devise an iterative procedure to solve (3). However, since the negative log-likelihood may be non-convex, their procedure does not give a global minimum.

We propose a new exponential model in Section 2. The main advantage is that one can estimate individuals' abilities by minimizing unconstrained convex formulations, so global minima are easily obtained. Details are in Section 3. Section 4 presents a real application, ranking bridge partnerships from team matches,

Appearing in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006.

and shows that the proposed model gives better rankings than a naive approach. Section 5 applies the new model to classification by error-correcting codes. Results are competitive with those of Huang et al. (2005). Section 6 is conclusions.

2. A New Exponential Model

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 (1), $\mathbf{v}$ 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

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

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

$$P(I_i^+ \text{ beats } I_i^-) \equiv P(Y_i^+ - Y_i^- > 0). \tag{4}$$

The distribution of Y_i^+ and Y_i^- is generally unknown, but a reasonable choice should place the mode around T_i^+ and T_i^-. To derive a computationally simple form for (4), we assume that Y_i^+ (and similarly Y_i^-) has a doubly-exponential extreme-value distribution with

$$P(Y_i^+ \le y) = \exp(-e^{-(y-T_i^+)}), \tag{5}$$

whose mode is exactly T_i^+. Suppose Y_i^+ is independent of Y_i^-, from (4) and (5) we have

$$P(I_i^+ \text{ beats } I_i^-) = \frac{e^{T_i^+}}{e^{T_i^+} + e^{T_i^-}}. \tag{6}$$

We omit the basic but tedious derivation. One may assume other distributions, e.g., normal, in (5), but the resulting model may not be in a closed form. Such differences already occur for pairwise comparisons, of which David (1988) gave some discussion. Thus (6) is our proposed model in this paper.

For pairwise comparisons (i.e., each individual forms a team), (6) reduces to

$$P(\text{individual } i \text{ beats individual } j) = \frac{e^{v_i}}{e^{v_i} + e^{v_j}}, \tag{7}$$

which is an equivalent re-parameterization (David, 1988; Hunter, 2004) of the Bradley-Terry model (2) by

$$p_i \equiv \frac{e^{v_i}}{\sum_{j=1}^k e^{v_j}}. \tag{8}$$

Therefore, our model (6) can also be considered as a generalized Bradley-Terry model. The re-parameterization (8) does not extend to the case of group competitions, so (6) and (1) are different.

Interestingly, (6) is a conditional exponential model[1], which is commonly used in the computational linguistic community. Thus we can use existing properties of this type of models.

3. Estimations

Following the proposed model (6), we estimate $\mathbf{v}$ by using available competition results. This section proposes 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 4.2. We also discuss a naive approach by summing the number of games an individual wins.

3.1. Regularized Least Square (RLS)

Recall that n_i^+ and n_i^- are respectively the number of games teams I_i^+ and I_i^- win. From (6), we have

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

and furthermore

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

Thus one can solve

$$\min_{\mathbf{v}} \quad \sum_{i=1}^m ((T_i^+ - T_i^-) - \log(n_i^+/n_i^-))^2 \tag{9}$$

to estimate the vector $\mathbf{v}$ of individuals' abilities. To represent (9) in a simpler form, we define a vector $\mathbf{d} \in R^m$ with

$$d_i \equiv \log(n_i^+/n_i^-),$$

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

$$G_{ij} \equiv \begin{cases} 1 & \text{if individual } j \in I_i^+, \\ -1 & \text{if individual } j \in I_i^-, \\ 0 & \text{if individual } j \notin I_i. \end{cases} \tag{10}$$

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

$$\begin{bmatrix} 1 & 1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & -1 & -1 & 0 & 0 \\ -1 & -1 & 0 & 0 & 0 & 0 & -1 & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}, \tag{11}$$

[1] Tutorials are available at, for example, `http://www.cs.cmu.edu/~aberger/maxent.html`.

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 (9) as

$$\min_{\mathbf{v}} \quad (G\mathbf{v}-\mathbf{d})^T(G\mathbf{v}-\mathbf{d}), \tag{12}$$

which is equivalent to solving the following linear system:

$$G^TG\mathbf{v} = G^T\mathbf{d}. \tag{13}$$

The linear system (13) may have multiple solutions if G^TG is not invertible. To handle this situation, we add a regularization term $\mu\mathbf{v}^T\mathbf{v}$ to (12):

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

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

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

We refer to this approach as RLS (Regularized Least Square). We heuristically use $\mu = 0.001$ for experiments in this paper.

3.2. Maximum Likelihood (ML)

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

$$l(\mathbf{v}) \tag{14}$$
$$\equiv -\sum_{i=1}^{m}\left(n_i^+\log\frac{e^{T_i^+}}{e^{T_i^+}+e^{T_i^-}} + n_i^-\log\frac{e^{T_i^-}}{e^{T_i^+}+e^{T_i^-}}\right),$$

and we estimate $\mathbf{v}$ by

$$\arg\min l(\mathbf{v}). \tag{15}$$

It is well known that the log-likelihood of a conditional exponential model is concave. Thus $l(\mathbf{v})$ is convex, so one can easily find a global minimum, which satisfies the following optimality condition:

$$\frac{\partial l(\mathbf{v})}{\partial v_s} = -\Big(\sum_{i:s\in I_i^+} n_i^+ + \sum_{i:s\in I_i^-} n_i^-\Big) + \sum_{i:s\in I_i^+}\frac{n_i e^{T_i^+}}{e^{T_i^+}+e^{T_i^-}} + \sum_{i:s\in I_i^-}\frac{n_i e^{T_i^-}}{e^{T_i^+}+e^{T_i^-}} = 0, \quad s=1,\ldots,k. \tag{16}$$

This condition is reasonable as it implies that the total number of observed wins of individual s is the same as the expected number by the assumed model.

Standard optimization methods (e.g., gradient or Newton's method) can be used to find a solution of (15). One may also use fixed-point type methods to minimize $l(\mathbf{v})$. A standard technique for conditional exponential models is improved iterative scaling (Pietra et al., 1997), which generates a sequence of iterations $\{\mathbf{v}^t\}_{t=0}^{\infty}$. The update from $\mathbf{v}^t$ to $\mathbf{v}^{t+1}$ requires the solution of k one-variable minimization problems. These k problems usually do not have closed-form solutions, and this situation happens for our problem (14). In the following we propose changing one component of $\mathbf{v}$ at a time. The resulting update rule is very simple. Let $\boldsymbol{\delta} \equiv [0,\ldots,0,\delta_s,0,\ldots,0]^T$ indicate the change of the sth component. We have

$$l(\mathbf{v}+\boldsymbol{\delta}) - l(\mathbf{v})$$
$$= -\Big(\sum_{i:s\in I_i^+} n_i^+ + \sum_{i:s\in I_i^-} n_i^-\Big)\delta_s + \tag{17}$$
$$\sum_{i:s\in I_i^+} n_i\log\left(\frac{e^{T_i^+}+\delta_s + e^{T_i^-}}{e^{T_i^+}+e^{T_i^-}}\right) + \sum_{i:s\in I_i^-} n_i\log\left(\frac{e^{T_i^+}+e^{T_i^-}+\delta_s}{e^{T_i^+}+e^{T_i^-}}\right)$$
$$\le -\Big(\sum_{i:s\in I_i^+} n_i^+ + \sum_{i:s\in I_i^-} n_i^-\Big)\delta_s + \tag{18}$$
$$\left(\sum_{i:s\in I_i^+}\frac{n_i e^{T_i^+}}{e^{T_i^+}+e^{T_i^-}} + \sum_{i:s\in I_i^-}\frac{n_i e^{T_i^-}}{e^{T_i^+}+e^{T_i^-}}\right)(e^{\delta_s}-1).$$

From (17), the inequality $x-1 \ge \log x$ yields (18). If $\delta_s = 0$, (18) $= 0$. We then minimize (18) to obtain the largest reduction. The solution has a simple closed form, which leads to the following update rule:

$$v_s \leftarrow v_s + \log\frac{\sum_{i:s\in I_i^+} n_i^+ + \sum_{i:s\in I_i^-} n_i^-}{\sum_{i:s\in I_i^+}\frac{n_i e^{T_i^+}}{e^{T_i^+}+e^{T_i^-}} + \sum_{i:s\in I_i^-}\frac{n_i e^{T_i^-}}{e^{T_i^+}+e^{T_i^-}}}. \tag{19}$$

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 (19) 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.

This algorithm is indeed the same as applying the sequential conditional generalized iterative scaling (Goodman, 2002) to (14). Since Goodman considers more complicated forms, here we give a derivation specific to our likelihood. The discussion also lets us know conditions for convergences:

Theorem 1 *If competition results satisfy*

$$\sum_{i:s\in I_i^+} n_i^+ + \sum_{s\in I_i^-} n_i^- > 0, \forall s, \tag{20}$$

then any limit point of the sequence $\{\mathbf{v}^t\}$ *generated by Algorithm 1 is a global minimum of* $l(\mathbf{v})$.

The proof is omitted. The condition (20) ensures both the numerator and the denominator of (19) are positive, so the update rule is well-defined. We refer to this approach as ML (Maximum Likelihood).

3.3. A Naive Approach (SUM)

We may estimate the sth individual's ability by summing the number of games it wins:

$$v_s \equiv \frac{\sum_{i:s\in I_i^+} n_i^+ + \sum_{i:s\in I_i^-} n_i^-}{\sum_{i:s\in I_i} 1}. \tag{21}$$

We refer to this method as SUM. This approach extends the following formula for pairwise comparisons:

$$v_s \equiv \frac{\sum_{i:i\neq s} n_{si}}{|\{i \mid n_{si}+n_{is}>0\}|},$$

where n_{si} is the number of games that individual s beats i. If

$$n_{si} > 0, n_{is} > 0, \text{ and } n_{si}+n_{is} = \text{constant}, \forall s, i, \tag{22}$$

then David (1988) show that the rankings by SUM and by ML are identical. Thus there is no need to maximize the likelihood. Practically (22) may not hold if individuals play different numbers of games. For group competitions, SUM and RLS/ML are quite different: SUM does not consider opponents' abilities, so its ranking is susceptible to individuals who played much fewer (or more) games but performed unusually well or poorly. Nor does it consider teammates' abilities, so strong players and weak ones receive the same credits. Ranking by SUM thus tends to be similar to that of teams. Because of the weak points mentioned above, SUM is used only as a baseline in experiments.

4. Experiments: Ranking Partnerships from Real Bridge Records

This section presents a real application: ranking partnerships from match records of Bermuda Bowl 2005[2], 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.

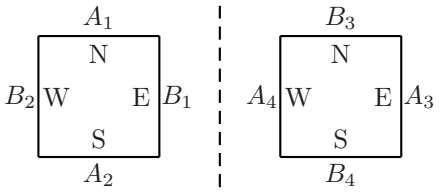

Figure 1. A typical match setting. N, S, E and W stand for north, south, east and west, respectively.

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 eight boards of the match between from India and Portugal in Bermuda Bowl 2005. 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[3]. A quick look at Table 1 may motivate the following straightforward approach: A partnership's score in a match is the sum

[2]All match records are available at `http://www.worldbridge.org/tourn/Estoril.05/Estoril.htm`.

[3]The IMP-to-VP conversion for Bermuda Bowl 2005 is on page 32, `http://www.worldbridge.org/departments/rules/GeneralConditionsOfContest2005.pdf`.

Table 1. Records of the first eight 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.

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

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 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 competition results. Hence the proposed model 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. Table 2 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 game setting matrix G defined in (10) is of size 231×69; as shown in (11) each row records a match's 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. Occasionally it is between 25 to 30 as a team's maximal score is 25. We normalize two VPs by their sum as n_i^+ and n_i^-, respectively.

4.2. Results and Analysis

Table 2 lists partnership rankings by four approaches, RLS, ML, Huang et al., 2005 (HNG) and SUM. Before investigating which one is better, we check the differences between the four approaches. Table 3(a) presents correlation coefficients by Kendall's tau, a standard way to find correlation between various rankings. Clearly RLS/ML/HNG behave similarly, but SUM is very different. We further measure the distance between the ranking by one approach and those by the other three:

$$d(\text{rank by method 1}, \text{ranks by other methods}) \equiv \begin{cases} \min(\text{others' ranks}) - \text{rank1} & \text{if rank1 the smallest,} \\ \text{rank1} - \max(\text{others' ranks}) & \text{if rank1 the largest,} \\ 0 & \text{otherwise.} \end{cases}$$

For example, from Table 2 the 2nd partnership of U.S.A.1 (US1) ranks 67/66/66/29 by RLS/ML/HNG/SUM. Then

$$d(29, \{67, 66, 66\}) = \min(67, 66, 66) - 29 = 37.$$

Checking all 69 partnerships' ranks gives

$$|d(\text{RLS, \{ML, HNG and SUM\}}) \geq 20| = 2, \quad (23)$$
$$|d(\text{ML, \{RLS, HNG and SUM\}}) \geq 20| = 0, \quad (24)$$
$$|d(\text{HNG, \{RLS, ML and SUM\}}) \geq 20| = 0, \quad (25)$$
$$|d(\text{SUM, \{RLS, HNG and ML\}}) \geq 20| = 10. \quad (26)$$

In Table 2 we respectively underline and boldface partnerships satisfying (23) and (26). From (26), SUM produces a very different ranking from those by the other three, an observation consistent with the correlation matrix in Table 3(a). In addition, SUM's ranking is closer to the team ranking (by total VPs). Partnerships satisfying (26) have higher ranks than those by RLS/ML/HNG when the team ranks are high, but have the opposite when the team ranks are low. This observation indicates that SUM may fail to identify weak (strong) partnerships from strong (weak) teams. From (24) and (25), ML's and HNG's rankings are always close to at least one other ranking. In fact, they give very similar rankings, as indicated by a high correlation coefficient of 0.87. However, compared with ML, the ranking by HNG is more correlated to that by SUM. Next we use match records to evaluate these approaches.

Let $r = (r_1, r_2)$ be the ranks of two partnerships. We define an order relationship between two groups $r = (r_1, r_2)$ and $\bar{r} = (\bar{r}_1, \bar{r}_2)$:

$$r \text{ better than } \bar{r} \text{ if } \max(r_1, r_2) < \min(\bar{r}_1, \bar{r}_2). \quad (27)$$

That is, if the weakest partnership from r is better than the strongest one from $\bar{r}$, then the group r should

Table 2. Partnerships' rankings. A partnership corresponds to the same position in columns. For example, The second partnership of Italy (IT) is ranked 18th, 17th, 14th and 4th by RLS, ML, HNG and SUM, respectively, and it plays 14 matches. Rankings satisfying (23) and (26) are boldfaced and underlined, respectively. Teams are ordered by team ranks; abbreviations of teams follow `http://www.paladinosoftware.com/Generic/countries.htm`.

Team	Partnership rankings												#match		
	RLS			ML			HNG			SUM					
IT	21	18	13	8	17	18	7	14	16	5	4	12	15	14	13
US2	63	**67**	1	52	**66**	1	43	**66**	1	47	**29**	2	8	17	17
US1	9	36	**41**	10	19	**37**	10	15	**38**	23	6	**10**	18	10	14
SE	2	55	37	2	25	53	2	12	64	1	19	39	14	13	15
IN	14	40	**42**	9	32	**41**	9	30	**42**	20	14	**15**	15	14	13
AR	33	26	32	26	21	29	25	23	34	16	17	28	15	14	13
EG	47	30	27	43	27	13	52	22	13	38	22	3	14	20	7
	46			**57**			**50**			**8**			1		
BR	19	4	**66**	31	7	**61**	24	5	**65**	25	11	**32**	11	18	13
JP	8	62	39	3	67	39	3	68	29	7	44	49	14	14	14
NL	6	60	11	30	44	28	28	47	31	37	34	21	15	15	12
CN	61	49	5	46	45	6	45	46	6	30	52	9	13	14	15
ZA	48	34	20	49	24	15	51	27	20	48	35	27	15	13	14
RU	38	28	31	35	16	47	36	18	49	40	24	53	14	14	14
PT	15	25	54	34	11	55	26	11	61	50	26	46	14	14	14
AU	44	51	**16**	42	51	**20**	40	53	**21**	42	51	**41**	16	11	15
NZ	68	23	3	68	48	5	67	39	4	64	36	13	9	16	17
UK	**10**	24	59	**12**	36	64	**17**	33	63	**45**	18	54	17	12	13
CA	17	35	56	14	38	58	19	35	62	33	43	60	14	16	12
TW	45	57	43	60	65	56	56	60	55	57	56	66	2	12	1
	7	**29**	58	**4**	**23**	54	**8**	**37**	54	**31**	**63**	61	4	7	16
PL	**12**	52	53	**22**	50	59	**32**	48	59	**58**	55	62	15	15	12
GP	50	22	69	40	33	69	44	41	69	65	59	69	14	14	14
JO	64	65		62	63		57	58		67	68		21	21	

be superior to $\bar{r}$. Then for each match, we define two kinds of events:

$$\begin{cases} \text{Violation:} & r \text{ better than } \bar{r} \text{ but } \bar{r} \text{ beats } r, \\ \text{Hit:} & r \text{ better than } \bar{r} \text{ and } r \text{ beats } \bar{r}. \end{cases}$$

A good ranking should produce many hits while causing few violations, so we use the following evaluation criterion:

$$\frac{\text{Number of violations}}{\text{Number of hits}}, \tag{28}$$

whose value should be minimized. Table 3(b) shows the ratio (28) of each of the four rankings and numbers of violations co-occurred for any two approaches. There are some interesting observations:

1. SUM produces the largest number of hits, but also the largest number of violations. Overall its ratio (28) is the largest: $32/96 = 0.33$. The other three methods achieve better balance between violations and hits: ML performs slightly better than HNG ($6/45 = 0.13 < 9/48 = 0.19$), while RLS is worse ($12/45 = 0.27$).

Table 3. Properties of rankings by the four approaches. For violations and hits, the diagonal shows #violations/#hits of each approach, and the off-diagonals show #violations co-occurred for any two approaches.

(a) Correlation coefficients

	RLS	ML	HNG	SUM
RLS	1.0	0.71	0.68	0.39
ML	0.71	1.0	0.87	0.50
HNG	0.68	0.87	1.0	0.53
SUM	0.39	0.50	0.53	1.0

(b) Violations and hits

	RLS	ML	HNG	SUM
RLS	12/45	3	3	4
ML	3	6/45	6	5
HNG	3	6	9/48	8
SUM	4	5	8	32/96

2. Results are highly related to the findings in (23)-(26). Among 12 violations of RLS, the two partnerships from (23) involve in four. For SUM, nearly all violations are related to the 10 partnerships from (26).

3. In few cases weak partnerships beat strong ones. For example, two partnerships from Egypt (EG, team rank 6) beat two from Italy (IT, team rank 1). All four approaches however rank the Italian partnerships higher than the Egyptian ones: RLS (18 13 vs. 47 30), ML (17 18 vs. 27 43), HNG (14 16 vs. 22 52), and SUM (4 12 vs. 22 38). Such ranks are reasonable as the Italian partnerships win many other matches.

We find that the six violations of ML are either exceptional games where weaker ones win or rankings with a small amount of violation (i.e., $\max(r_1, r_2) - \min(\bar{r}_1, \bar{r}_2)$ in (27) is small). Such violations are thus not very serious. In contrast, RLS and SUM have additional violations related to partnerships identified in (23)-(26), whose ranks are likely to be wrong. For HNG, its nine violations are the six of ML plus three additional ones. Since ML and HNG give quite similar results, we investigate more carefully the distance between their rankings and find that

$$|d(\text{ML, HNG}) \geq 10| = 5. \tag{29}$$

Similar to the second observation discussed earlier, Table 4 shows that three of the five partnerships in (29) involve in two of the three additional violations. Moreover, each partnership's rank by ML is higher than that by HNG if the team rank is low (PL and TW), but lower if the team rank is high (JP). Therefore, ML may be better than HNG in identifying weak (strong) partnerships from strong (weak) teams. From all aspects discussed so far, ML is the best and HNG is almost as good. Table 5 lists the top ten partnerships by ML, which are also the top ten by HNG with only minor re-ordering. Some are famous players.

In addition to violations, another evaluation measure is the mean-squared error (MSE):

$$\frac{1}{m}\sum_{i=1}^{m}\left(\tilde{P}(I_i^+ \text{ beats } I_i^-) - \frac{n_i^+}{n_i}\right)^2,$$

Table 4. The three matches where HNG causes violations but ML does not. We show the ranks of playing partnerships by ML and HNG, boldfacing those satisfying (29).

Method	Partnership ranks				VPs	
	NL		BR			
ML	30	44	7	31	20	10
HNG	28	47	5	24		
	AR		TW			
ML	21	26	**23**	54	14	16
HNG	23	25	**37**	54		
	JP		PL			
ML	3	**39**	**22**	50	6	24
HNG	3	**29**	**32**	48		

where $\tilde{P}(I_i^+ \text{ beats } I_i^-)$ is the estimate. The MSEs for RLS, ML and HNG are 0.0365, 0.0283 and 0.0284, respectively. We did not calculate MSE for SUM because it does not assume a predictive model. Again we see that ML and HNG behave similarly, while RLS performs worse. One may wonder if experiments under a standard training/testing setting should be conducted. However, in contrast to classification or regression where generalization ability is emphasized, the goal of ranking is to explain available outcomes as well as possible, so evaluating rankings on unseen match results would be unreasonable.

We then explore why RLS is slightly worse than ML. The two partnerships satisfying (23) respectively have scores 25:0 and 1:25 in two matches. Note that the highest VP one can obtain is 25. Among eight 25:0/1 and 0/1:25 matches in all games, two occur for two partnerships of New Zealand, which rank (3rd, 23rd) by RLS. However, ML and HNG give $(5, 48)$ and $(4, 39)$, respectively. Thus extreme scores seem to more significantly affect RLS. We explain this phenomenon by checking the optimization formulas of RLS and ML. If the scores are 25:0, $n_i^+/n_i^- = \infty$ causes problems in (9), so we set $n_i^- = 0.001$. However, $\log n_i^+/n_i^-$ is still large, so RLS essentially hopes

$$(T_i^+ - T_i^- - \text{a large value})^2 \tag{30}$$

is small. For (14), if $n_i^- = 0$, it intends to have small

$$\left| \frac{e^{T_i^+}}{e^{T_i^+} + e^{T_i^-}} - 1 \right| = \left| \frac{1}{1 + e^{T_i^- - T_i^+}} - 1 \right|. \tag{31}$$

One needs $T_i^+ \gg T_i^-$ so that (30) is reasonably small, but a small (31) does not require very large $T_i^+ - T_i^-$. In other words, extreme scores cause large terms in (9), but only moderate ones in (14).

Discussion so far globally compares the three approaches. Next we investigate two teams' rankings in detail. Table 6 lists match records of U.S.A.2 and Poland. For U.S.A.2, the 2nd partnership (called P2) ranks 29th by SUM but RLS, ML and HNG give 67, 66 and 66, respectively. In addition, RLS and ML have that P2 is similar to P1, but for SUM, P2 is better. When (P2,P3) are together, they win 7 matches but lose 5. For (P1,P3), they significantly win 3, but lose 1. Thus P2 is not better than P1, but SUM fails to capture such relationships. Moreover, SUM does not consider opponents' abilities, so P2's 25:3 match against Jordan (team rank 22, the last) and P1's 25:5 match against India (team rank 5) respectively give them similar credits. For the Polish team, all approaches rank P2 and P3 to be around 55. SUM considers P1 to be similar as well (58th), but RLS, ML and HNG give P1 a much higher rank (12, 22 and 32). To find which one is more reasonable, we list opponents' team ranks:

P1,P2: big wins over $9, 11, 17$; big losses to $6, 8$
P2,P3: big wins over $19, 21, 22$; big losses to $2, 14, 18$
P1,P3: small wins over $1, 4$; big losses to $10, 13$.

Clearly, results of (P1,P2) and (P2,P3) imply that P1 is better than P3. The reason is that (P1,P2) wins over/loses to stronger teams. Similarly, comparing (P2,P3) and (P1,P3) shows that P1 is better than P2. This example shows that the proposed approach nicely captures indirect relationships. Earlier we stressed the difference between team and partnership rankings, so one may doubt the use of opponents' team ranks above. However, a match involves two but not one partnership of a team. As most teams have only three partnerships, team ranks should reasonably indicate the ability of two participating ones.

Table 5. Top 10 partnerships by the approach 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
Brazil	Gabriel Chagas	Miguel Villas-boas
Italy	Norberto Bocchi	Giorgio Duboin
India	Subhash Gupta	Rajeshwar Tewari
U.S.A.1	Jeff Meckstroth	Eric Rodwell

While ML seems to be the best for this data, it is not perfect. We suspect that it overestimates a Chinese Taipei partnership as the 4th. This team (six players) has six different partnerships, more than any team else. As some play very few matches, without enough records, the obtained ranks are less reliable. Earlier we criticized that SUM is vulnerable if some partnerships play very few matches. This is observed in its rank for an Egyptian partnership which plays only one

Table 6. Match records of U.S.A.2 and Poland. A star indicates playing in a match, and we boldface the score of the winning team. The last column shows rival teams and their team rankings. For U.S.A.2, rankings by RLS/ML/HNG/SUM are P1: 63/52/43/47, P2: 67/66/66/29 and P3: 1/1/1/2. For Poland, rankings are P1: 12/22/32/58, P2: 52/50/48/55 and P3: 53/59/59/62.

(a) U.S.A.2.

P1	P2	P3	Score		vs.
⋆	⋆		0	**25**	NL(10)
⋆	⋆		12	**18**	ZA(12)
⋆	⋆		11	**19**	UK(17)
⋆	⋆		14	**16**	TW(19)
⋆		⋆	**25**	5	IN(5)
⋆		⋆	**22**	8	NZ(16)
⋆		⋆	**22**	8	PL(20)
⋆		⋆	12	**18**	US1(3)
	⋆	⋆	**22**	8	SE(4)
	⋆	⋆	**19**	11	EG(7)
	⋆	⋆	**19**	11	BR(8)
	⋆	⋆	**24**	6	RU(13)
	⋆	⋆	**20**	10	PT(14)
	⋆	⋆	**20**	10	CA(18)
	⋆	⋆	**25**	3	JO(22)
	⋆	⋆	15	15	AU(15)
	⋆	⋆	12	**18**	IT(1)
	⋆	⋆	13	**17**	AR(6)
	⋆	⋆	14	**16**	JP(9)
	⋆	⋆	13	**17**	CN(11)
	⋆	⋆	14	**16**	GP(21)

(b) Poland

P1	P2	P3	Score		vs.
⋆	⋆		**24**	6	JP(9)
⋆	⋆		**19**	11	CN(11)
⋆	⋆		**22**	8	UK(17)
⋆	⋆		15	15	ZA(12)
⋆	⋆		9	**21**	US1(3)
⋆	⋆		14	**16**	IN(5)
⋆	⋆		4	**25**	AR(6)
⋆	⋆		5	**25**	BR(8)
⋆	⋆		11	**19**	AU(15)
⋆		⋆	**16**	14	IT(1)
⋆		⋆	**16**	14	SE(4)
⋆		⋆	11	**19**	EG(7)
⋆		⋆	9	**21**	NL(10)
⋆		⋆	7	**23**	RU(13)
⋆		⋆	11.5	**17.5**	NZ(16)
	⋆	⋆	**25**	5	TW(19)
	⋆	⋆	**19**	11	GP(21)
	⋆	⋆	**20**	10	JO(22)
	⋆	⋆	8	**22**	US2(2)
	⋆	⋆	4	**25**	PT(14)
	⋆	⋆	3	**25**	CA(18)

match. While RLS, ML and HNG give more reasonable ranks for this partnership, unfortunately they are not so successful on the Chinese Taipei partnership.

5. Multi-class Classification

In Huang et al. (2005), the main application is multi-class probability estimates under error-correcting codes. Since their formulation (3) incorporates normalizing and positive constraints, class probability estimates are immediately available from optimal parameters; the proposed model, however, cannot directly deliver such probability estimates. Nevertheless, it can still be used for classification by predicting the label with the largest v_s. We conduct experiments on the seven real problems used in Huang et al. (2005). They prepared 20 subsets of 800 training and 1,000 testing instances[4] and considered four error-correcting codes. Due to space limitation, we present only results on the "sparse" and "dense" codes. Average test error rates by RLS, ML and HNG are in Table 7. For the dense code, RLS and ML are marginally better than HNG, but for the sparse code, RLS performs worse

[4] `http://www.csie.ntu.edu.tw/~cjlin/papers/svmprob/data`

Table 7. Average test error rates (in percentage). We boldface the lowest ones for each error-correcting code.

		Dense			Sparse		
Problem	#class	RLS	ML	HNG	RLS	ML	HNG
dna	3	6.35	**6.34**	6.39	6.88	6.29	**6.24**
waveform	3	**13.71**	**13.71**	13.92	13.54	**13.45**	14.27
satimage	6	11.61	11.52	**11.41**	**11.46**	11.58	11.79
segment	7	3.54	3.46	**3.45**	3.97	3.54	**3.23**
USPS	10	**7.22**	7.29	7.66	8.06	**7.68**	8.52
MNIST	10	**7.25**	**7.25**	7.58	8.09	**7.74**	8.97
letter	26	19.55	**19.37**	20.27	21.20	20.47	**20.43**

while ML and HNG are almost equally good. Results indicate that the proposed model is also useful for multi-class classification by error-correcting codes.

6. Conclusions

We propose a new and useful method to rank individuals from group comparisons. Contrary to early work, which solves non-convex problems, here convex formulations with easy solution procedures are developed. Experiments show that the proposed approach gives reasonable partnership rankings from bridge records. We also develop techniques to evaluate different rankings, which may be used in other ranking tasks.

Acknowledgements This work was partially supported by National Science Council, Taiwan.

References

Allwein, E. L., Schapire, R. E., & Singer, Y. (2001). Reducing multiclass to binary: a unifying approach for margin classifiers. *JMLR*, *1*, 113–141.

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

David, H. A. (1988). *The method of paired comparisons*. Oxford University Press. Second edition.

Dietterich, T. G., & Bakiri, G. (1995). Solving multiclass learning problems via error-correcting output codes. *J. Artificial Intelligence Res.*, *2*, 263–286.

Goodman, J. (2002). Sequential conditional generalized iterative scaling. *ACL* (pp. 9–16).

Huang, T.-K., Weng, R. C., & Lin, C.-J. (2005). A generalized Bradley-Terry model: From group competition to individual skill. In *NIPS 17*.

Hunter, D. R. (2004). MM algorithms for generalized Bradley-Terry models. *Ann. Statist.*, *32*, 386–408.

Pietra, S. D., Pietra, V. D., & Lafferty, J. (1997). Inducing features of random fields. *IEEE PAMI*, *19*, 380–393.