Ranking rankings: an empirical comparison of the predictive power of sports ranking methods

D. Barrow, I. Drayer, P. Elliott, G. Gaut, and B. Osting

January 23, 2013

Abstract. In this paper, we empirically evaluate the predictive power of eight sports ranking methods. For each ranking method, we implement two versions, one using only win-loss data and one utilizing score-differential data. The methods are compared on 4 datasets: 32 National Basketball Association (NBA) seasons, 112 Major League Baseball (MLB) seasons, 22 NCAA Division 1-A Basketball (NCAAB) seasons, and 56 NCAA Division 1-A Football (NCAAF) seasons. For each season of each dataset, we apply 20-fold cross validation to determine the predictive accuracy of the ranking methods. The non-parametric Friedman hypothesis test is used to assess whether the predictive errors for the considered rankings over the seasons are statistically dissimilar. The post-hoc Nemenyi test is then employed to determine which ranking methods have significant differences in predictive power. For all datasets, the null hypothesis—that all ranking methods are equivalent—is rejected at the 99% confidence level. For NCAAF and NCAAB datasets, the Nemenyi test concludes that the implementations utilizing score-differential data are usually more predictive than those using only win-loss data. For the NCAAF dataset, the least squares and random walker methods have significantly better predictive accuracy at the 95% confidence level than the other methods considered.

Keywords. sports rankings, cross validation, hypothesis testing, Friedman test, and Nemenyi test

1 Introduction

In our meritocratic society, the concept of rank is paramount. Consumers seek the best product, search engines recommend the most relevant document, and sports fans demand to know the standing of their favorite sports team! The need for rankings in various contexts has led to the development of many different ranking algorithms. In this paper, we consider several existing ranking methods and, for several different sports leagues, statistically compare the

predictive power of the rankings. In what follows, we introduce the ranking methods considered, our comparison methodology, summarize our main results, and give an outline of the paper.

Ranking methods. There are a large number of methods used for sports ranking today. For example, as of October 31, 2012 the Massey Ratings website[1] compares 124 NCAA Division 1 FBS teams, using 101 different ranking methods. Admittedly the number of *official* ranking methods used in sports today is substantially smaller, however the recent survey of Stefani (2011) comparing the official rating systems of 159 international sports federations demonstrates that there are many in use. In this work, we compare 8 ranking methods, which we briefly summarize below. Detailed descriptions and references are given in §3. For each method, we consider two versions, labeled (α) and (β). The (α) version depends on win-loss information only and the (β) version depends on score differentials (magnitude or "margin" of victory).

WP **Winning percentage** is the simplest method for rating, evaluating each team based on the percentage of games won.

RPI The **Rating Percentage Index** is based on winning percentages; a team's RPI score is a weighted average of the winning percentages of teams, their opponents, and their opponent's opponents.

L2 The **least squares pairwise comparison method** seeks a rating such that the difference in ratings between two teams agrees with the game outcomes (pairwise comparison data). Depending on the way in which the pairwise comparison data is chosen, this method has many variations.

MP The **maximum posterior** rating is obtained by interpreting the $L2$ method in the Bayesian framework with a prior distribution on the rating.

K **Keener's** direct method for rating is based on the Perron-Frobenius eigenvector of a matrix describing the relative team strengths.

PR The **PageRank** rating is the Perron-Frobenius eigenvector of a stochastic matrix describing transition probabilities on a directed graph.

RW The **random walker** rating is the stationary state for a system of differential equations on a directed graph representing the game results.

E **Elo's** method updates the rating of a team after a game based on the difference between the observed game outcome and a prediction of that outcome based on past performance.

[1] http://www.masseyratings.com/cf/compare.htm

Comparison methodology. We consider 4 datasets: 32 National Basketball Association (NBA) seasons, 112 Major League Baseball (MLB) seasons, 22 NCAA Division 1 Basketball (NCAAB) seasons, and 56 NCAA Division 1-A Football (NCAAF) seasons. For each season of each dataset, we apply 20-fold cross validation to determine the predictive accuracy of the above listed ranking methods. The predictive accuracy is defined to be the proportion of games in the test data for which the ranking, obtained from the training data, correctly predicts the victor. The non-parametric Friedman test is then used to assess whether the predictive error for the considered rankings over the seasons are statistically dissimilar. To determine which ranking methods are significantly different from each other, the post-hoc Nemenyi test is used. Our comparison methodology, which follows Shaffer (1995) and Demšar (2006), is described in §4.

We exploit several properties of non-parametric hypothesis testing. First, non-parametric tests do not require assumptions on the probability distributions of the ratings being assessed. In fact, modeling these distributions for our datasets would be difficult; some seasons are simply more predictable than others and thus the predictive error of the ratings is season-dependent. Secondly, while parametric tests can be used to compare ratings (cardinal quantities), non-parametric tests are much more convenient when discussing rankings (ordinal quantities). Although these tests are well-developed tools for statistical inference and widely used in social sciences and for the design of experiments, we are unaware of their application to the comparison of sports ranking methods.

Summary of results. We find that the cross validation scores of the ranking methods vary among sports, seasons, and other ranking methods. This suggests that some sports are more predictable than others, that some seasons are more predictable than others, and finally that some ranking methods have greater predictive accuracy than others considered. Applying the Friedman test to compare the ranking methods, we find that for all datasets, the null hypothesis—that all rankings have equivalent predictive power—can be rejected at the 99% confidence level. That is, the predictive accuracy of the ranking methods differ. For the datasets considered, we make several observations on the predictive power of the rankings, including the following. For the NBA and MLB datasets, the Nemenyi test is unable to identify a ranking method which has significantly better predictive accuracy, although methods are identified which have significantly worse predictive accuracy. For NCAAF and NCAAB datasets, the Nemenyi test concludes that the implementations

utilizing score-differential data are usually more predictive than those using only win-loss data. For the NCAAF dataset, the least squares and random walker methods have significantly better predictive accuracy than the other methods considered.

Outline. In §2, we review previous work on comparing sports ranking methods. In §3, we discuss several commonly used ranking methods. In §4, we give a general discussion of hypothesis testing. In §5 we use cross-validation and hypothesis testing methods to analyze the predictive power of the ranking methods discussed in §3. We conclude in §6 with a discussion and further directions.

2 Previous Results

Ranking has an extensive history, with roots in debates between Condorcet and Borda over elections in the French Academy of Sciences in the late eighteenth century (David, 1963). With the recent swell of data, ranking methods are more prominent than ever with applications in web searches, e-commerce, and, of course, sports. A comprehensive survey of the literature is beyond the scope of this work; we reference here only a few recent contributions similar to our own.

Gill (2009) compares the the predictive performance of several least-squares type methods using leave-one-out cross validation. Using 1930-2007 NCAA football data, models are empirically identified which minimize prediction error. In particular, parameters for a pairwise comparison data model are found that reduce the effect of large victory margins. The study by Trono (2010) examines the efficacy of over twenty rating/ranking methods in predicting the winner and point spread of 1983-2003 NCAA Division 1-A football games. In particular, the methods are compared against the "Las Vegas betting line", which is a sports prediction used for betting purposes. The question of whether the rating/ranking methods with the highest predictive performance could be used profitably in sports betting is addressed. In both of these works, cross validation scores were used to evaluate the predictive performance; hypothesis testing is not used.

There are also several analytical comparisons of ranking methods. Chan (2011) analytically investigates the prediction accuracy of the Bradley-Terry and Thurstone-Mosteller models for three simple games, where the outcome is

probabilistically known. Tran (2011) constructs datasets for which the least-squares, PageRank, and tropical eigenvector methods produce arbitrarily different ranking orders.

Closely related to the predictive accuracy of a rating/ranking method is the sensitivity/stability of the method on the data (game results). Intuitively, one expects that a method which is robust to changes in the data, *i.e.*, less sensitive to anomalous games ("upsets"), will have higher predictive accuracy. In what follows, we give references to several works which further develop these ideas.

Using perturbative methods, Chartier, Kreutzer, Langville, and Pedings (2011b) analyze the sensitivity of three ranking methods: Massey, Colley, and PageRank methods. For "perfect seasons" (every team plays every other team exactly once and there are no upsets), the authors show that the Colley and Massey methods are less sensitive than the PageRank method. The study is illustrated with examples from sports data. In a closely related study, (Chartier, Kreutzer, Langville, and Pedings, 2011a) a method is introduced for weighting game results in sports rankings, allowing for a ranking to be more sensitive to some games than others. The method was used to produce successful predictions for the 2010 Division 1 NCAA Men's Basketball tournament ("March Madness").

Burer (2012) uses perturbative methods to study the sensitivity of the Colley ranking method and proposes a variant of the method, motivated by recent results in robust optimization. The modification is intended to reduce the impact of "inconsequential games" on the rating/ranking. Using 2006-2011 regular season NCCA football data, the proposed ranking method is empirically shown to be more robust then the Colley method.

Osting, Brune, and Osher (2012a) uses the Fisher information of the least-squares ranking method to characterize the stability of the ranking with respect to changes in the dataset. Under certain assumptions, this measure of robustness has a nice graph theoretic interpretation in terms of the algebraic connectivity of the graph describing the dataset and is easily computable. A method is proposed for actively scheduling the games to optimally increase the informativeness of the dataset. This approach differs from previous approaches, attempting to improve rankings by modifying the dataset (schedule) rather than the ranking method.

3 Ranking methods

In this section, we describe the ranking methods compared in this paper. These methods will be compared in §5. As mentioned in the introduction, we are not attempting to compare all sports methods, but rather a representative subset. An introduction to ranking methods can be found in Langville and Meyer (2012). Before discussing individual ratings, we first give a few definitions.

Terminology and notation

Let V be a set of n teams to be rated, which we enumerate $V = \{i\}_{i=1}^{n}$. A *rating* $\phi\colon V \to \mathbb{R}$ assigns each team a quantitative "strength". A *ranking* is an ordering of the teams; a rank α team is "stronger" then $n-\alpha$ other teams. A ranking may be obtained from a rating on V simply by sorting. To be clear, a team is "good" if it has a large rating and a small ranking.

Consider a set of m games played among the teams in V. For each game played between teams i and j, we consider one of the following two datasets:

(α) The game result only, *i.e.*, win, loss, and tie information.
(β) The final game score.

The (β) dataset contains more information than the (α) dataset. Thus we expect that a ranking generated using the (β) dataset will have more predictive power than one generated using just the (α) dataset. For each of the ranking methods below, we consider two versions—one which uses the dataset labelled (α) and the other which uses the dataset labelled (β). In the Elo method (defined below), the order in which the games are played is also relevant.

Define the matrices $W, S \in \mathbb{R}^{n\times n}$, [2]

$$W_{ij} = \#\{\text{team } i \text{ beat team } j\} \; + \; \frac{1}{2}\,\#\{\text{ties between teams } i \text{ and } j\} \tag{1}$$

$$S_{ij} = \sum_{\substack{\text{games btwn}\\ \text{teams } i \text{ and } j}} \frac{\#\text{ points } i \text{ scored on } j}{\text{total points in game}}. \tag{2}$$

Here, W only depends on dataset (α) while S depends on dataset (β). Define the vectors $w, l, d \in \mathbb{R}^n$,

$$w = W\mathbf{1}, \qquad l = W^t\mathbf{1}, \quad \text{and} \quad d = (W + W^t)\mathbf{1}. \tag{3}$$

The vector element w_i (resp. l_i) is the number of games won (resp. lost) by team i plus one-half times the number of ties for team i. The vector element

[2] In equation (2), we take the fraction to be $\frac{1}{2}$ if the game results in a $0-0$ tie.

d_i is the number of games played by team i. Since each game results in a win, loss, or tie, we have $d = w + l$. Denote $D = \mathrm{diag}(d)$. We'll assume that each team has played in at least one game, implying D is invertible. We analogously define the vectors $s, t \in \mathbb{R}^n$,

$$s = S\mathbf{1} \quad \text{and} \quad t = S^t\mathbf{1}. \tag{4}$$

Note that $s + t = d$.

We define the vectors $y, z \in \mathbb{R}^m$ as follows. For each game g, define

$$y_g = \begin{cases} 0 & \text{game } g \text{ results in a tie} \\ \frac{\text{score of winning team} \ - \ \text{score of losing team}}{\text{total points in game } g} & \text{otherwise} \end{cases} \tag{5}$$

$$z_g = \mathrm{sgn}(y_g). \tag{6}$$

Here the function sgn is 1 if the argument is positive, -1 if the argument is negative, and 0 if the argument is zero. Note that since y has nonnegative entries, z does too. Define the matrix $B \in \mathbb{R}^{m \times n}$,

$$B_{gi} = \begin{cases} 1 & \text{game } g \text{ is between teams } i \text{ and } j,\ i \text{ beats } j \\ -1 & \text{game } g \text{ is between teams } i \text{ and } j,\ j \text{ beats } i \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Here, if a game g between teams i and j results in a tie, we assign a 1 and -1 to i and j arbitrarily. Note the following relationships:

$$B^tB = D - (W + W^t), \qquad B^ty = s - t, \quad \text{and} \quad B^tz = w - l. \tag{8}$$

Finally, define the matrix $\Delta = B^tB$.

Remark. Several of the quantities defined above can be interpreted on a graph (Foulds, 1992). Consider a directed multigraph where each vertex represents a team and each arc represents a game. If two teams play one another more than once, then more than one arc joins the two vertices. The arcs are oriented so that the head of the arc is the vertex representing the winning team. If there is a tie, we represent it one of two ways: (i) with two 'half' arcs with opposing orientations, or (ii) the orientation of arcs representing tied games are chosen arbitrarily. Note that in the case where there are no tied teams, the two graph representations are the same. If we represent ties using (i), the matrix W is the directed adjacency matrix for the directed multigraph and the vertex vectors w and l denote the in- and out-degrees respectively. If we represent ties using (ii), the matrix B is the arc-vertex incidence matrix for the directed multigraph. In both cases, the matrix Δ is the non-normalized graph Laplacian.

3.1 Winning percentage (WP)

The *winning percentage* of team i is defined $\phi_i^{\mathrm{WP}\alpha} = w_i/d_i$ where w and d are given in (3). Thus the vector of winning percentages $\phi^{\mathrm{WP}\alpha} \in \mathbb{R}^n$ can be computed

$$\phi^{\mathrm{WP}\alpha} = D^{-1}w. \tag{9}$$

This is perhaps the simplest rating method. A similar quantity using the dataset (β) which includes the margin of victory can be calculated

$$\phi^{\mathrm{WP}\beta} = D^{-1}s. \tag{10}$$

Here, s is defined in (4). The ratings (9) and (10) do not take into account the team's "strength of schedule". That is, a team is able to have a large winning percentage by playing poor teams. To compensate for this, the ratings percentage index was introduced.

3.2 Ratings percentage index (RPI)

The Ratings Percentage Index (RPI) is generated using the winning percentages of teams, their opponents, and their opponent's opponents (Pickle and Howard, 1981). Let $\phi_i^{\mathrm{WP}\alpha}$ be the winning percentage of team i as defined in (9). Let W be as defined in (1). Team i's average opponent's winning percentage can be computed

$$(\text{opponent's winning percentage})_i = (d_i)^{-1}\sum_j (W_{ij} + W_{ji})\phi_j^{\mathrm{WP}\alpha}.$$

Thus, $D^{-1}(W+W^t)\phi^{\mathrm{WP}\alpha}$ is the vector of average opponent's winning percentages and $\left(D^{-1}(W+W^t)\right)^2\phi^{\mathrm{WP}\alpha}$ is the vector of average opponent's opponent's winning percentages. Each team's RPI is then computed according to the following weighted average:

$$\phi^{\mathrm{RPI}\alpha} = \frac{1}{4}\phi^{\mathrm{WP}\alpha} + \frac{1}{2}D^{-1}(W+W^t)\phi^{\mathrm{WP}\alpha} + \frac{1}{4}\left(D^{-1}(W+W^t)\right)^2\phi^{\mathrm{WP}\alpha}. \tag{11}$$

Similarly, we define

$$\phi^{\mathrm{RPI}\beta} = \frac{1}{4}\phi^{\mathrm{WP}\beta} + \frac{1}{2}D^{-1}(W+W^t)\phi^{\mathrm{WP}\beta} + \frac{1}{4}\left(D^{-1}(W+W^t)\right)^2\phi^{\mathrm{WP}\beta}. \tag{12}$$

Remark. There is also a home-adjusted RPI ranking method which weights home and away games differently. We do not consider this method here.

3.3 Least-squares pairwise comparison method (L2)

Let B be defined as in (7). If $f \in \mathbb{R}^m$ has elements representing a game outcome, the pairwise comparison method for ranking is the solution to

$$\phi^{\mathrm{L2}} = \arg\min_{\phi} \ \|B\phi - f\|^2, \tag{13}$$

where $\|\cdot\|$ denotes the ℓ^2-norm. Choosing $f = y$ as defined in (5) and $f = z$ as defined in (6) gives two ranking methods, which we denote as $\phi^{\mathrm{L2}\beta}$ and $\phi^{\mathrm{L2}\alpha}$ respectively. Depending on the choice of f, the rating ϕ^{L2} is sometimes referred to as the Massey rating (Massey, 1997, Langville and Meyer, 2012), the Bradley-Terry rating (Bradley and Terry, 1952, David, 1963), Thurstone-Mosteller (David, 1963), HodgeRank (Jiang, Lim, Yao, and Ye, 2010, Xu, Yao, Jiang, Huang, Yan, and Lin, 2011), and least squares methods (Hirani, Kalyanaraman, and Watts, 2011, Osting, Darbon, and Osher, 2012b, Osting et al., 2012a, Chartier et al., 2011b).

It is well known that the solution to the least squares problem (13) is given by $\phi^{\mathrm{L2}} = (B^tB)^\dagger B^t f$ where $C^\dagger$ denotes the Moore-Penrose pseudoinverse of the matrix C. Using (8), we obtain

$$\phi^{\mathrm{L2}\alpha} = \Delta^\dagger(w - l) \tag{14a}$$

$$\phi^{\mathrm{L2}\beta} = \Delta^\dagger(s - t). \tag{14b}$$

Remark. In Eq. (13), other ℓ^p norms have also been considered, see, *e.g.*, (Hochbaum, 2010, Osting et al., 2012b).

3.4 Maximum posterior estimate (MP)

It is useful to consider an alternative interpretation of (13). Suppose we model the probability of witnessing the pairwise comparisons f given a rating ϕ as

$$\pi(f|\phi) \propto \exp(-\|B\phi - f\|^2)$$

Then the maximum likelihood estimate $\arg\max_\phi \pi(f|\phi)$ is equivalent to (13). Additionally, assume a Bayesian prior on ϕ,

$$\pi_{\mathrm{pr}}(\phi) \propto \exp\left(-\gamma\|\phi\|^2\right).$$

This states that prior to any games, we assume that all teams have the same rating, 0. Then, using Bayes Law, the maximum posterior estimate is written

$$\phi^{\mathrm{MP}} = \arg\max_{\phi} \ \pi(f|\phi)\pi_{\mathrm{pr}}(\phi) = \arg\max_{\phi} \ \exp(-\|B\phi - f\|^2 - \gamma\|\phi\|^2).$$

The maximum is given by the solution to

$$(B^t B + \gamma \mathrm{Id})\phi^{\mathrm{MP}} = B^t f. \tag{15}$$

Choosing $f = y$ as defined in (5) and $f = z$ as defined in (6) gives two ranking methods, which we denote as $\phi^{MP\beta}$ and $\phi^{MP\alpha}$ respectively. Using (8), we obtain

$$\phi^{MP\alpha} = [\Delta + \gamma \mathrm{Id}]^{-1}(w - l) \tag{16a}$$

$$\phi^{MP\beta} = [\Delta + \gamma \mathrm{Id}]^{-1}(s - t). \tag{16b}$$

Note that the matrix $\Delta + \gamma \mathrm{Id}$ is positive definite for $\gamma > 0$ and thus invertible. This method is related to Tikhonov regularization in inverse problems and ridge regression in statistics. The ranking $\phi^{MP\alpha}$ with $\gamma = 2$ is related to the Colley method (Colley, 2002, Langville and Meyer, 2012). In Colley's method, f in (15) is taken to be a quantity similar to y, but non-normalized. In all ranking comparisons, we take $\gamma = 2$.

3.5 Keener's direct method (K)

Let $A \in \mathbb{R}^{n \times n}$ be a matrix where entry $A_{ij} \in [0, 1]$ describes the relative strength of team i over team j. Consider the normalized matrix, $D^{-1}A$. Keener's direct method interprets the Perron-Frobenius eigenvector[3] of $D^{-1}A$ as a rating (Keener, 1993).

It remains to describe the matrix A. We consider two constructions—using the datasets (α) and (β). Let $\phi^{\mathrm{K}\alpha}$ be the Perron-Frobenius eigenvector of $D^{-1}W$. Following Keener (1993), let $K \in \mathbb{R}^{n \times n}$ have matrix elements

$$K_{ij} = h\left(\frac{S_{ij} + 1}{S_{ij} + S_{ji} + 2}\right) \quad \text{where} \quad h(x) = \frac{1}{2} + \frac{1}{2}\mathrm{sgn}\left(x - \frac{1}{2}\right)\sqrt{|2x - 1|}. \tag{17}$$

Define $\phi^{\mathrm{K}\beta}$ to be the Perron-Frobenius eigenvector of $D^{-1}K$.

Note that the matrices $D^{-1}W$ and $D^{-1}K$ are not necessarily irreducible[4]. For example, $D^{-1}W$ is not irreducible if there is a winless team.

[3] Recall that for a matrix with non-negative entries, there exists a positive, real eigenvalue (called the Perron-Frobenius eigenvalue) such that any other eigenvalue is smaller in magnitude. The Perron-Frobenius eigenvalue is simple and the corresponding eigenvector (called the Perron-Frobenius eigenvector) has non-negative entries. See, for example, Horn and Johnson (1991).

[4] The matrices W and S are irreducible if the corresponding directed graph is strongly connected. The matrix W is irreducible if there is no partition of the teams $V = V_1 \sqcup V_2$ such that no team in V_1 has beat a team in V_2.

3.6 PageRank method (PR)

The PageRank method (Page, Brin, Motwani, and Winograd, 1999) considers the Perron-Frobenius eigenvector of the matrices

$$\Xi_p^\alpha := pW[\mathrm{diag}(l)]^\dagger + (1-p)\frac{1}{n}\mathbf{1}\mathbf{1}^t \quad \text{and} \quad \Xi_p^\beta := pS[\mathrm{diag}(t)]^\dagger + (1-p)\frac{1}{n}\mathbf{1}\mathbf{1}^t.$$

The matrices Ξ_p^α and Ξ_p^β are column stochastic matrices, *i.e.*, $\mathbf{1}^t\Xi_p = \mathbf{1}^t$. For $0 < p < 1$, Ξ_p^α and Ξ_p^β are irreducible so that the Perron-Frobenius theorem states that the Perron-Frobenius eigenvector has strict positive entries, is unique and simple, and the Perron-Frobenius eigenvalue is the largest magnitude eigenvalue. Since Ξ_p^α and Ξ_p^β are also stochastic, the Perron-Frobenius eigenvalue is equal to one. Thus the PageRank rankings, $\phi^{\mathrm{PR}\alpha}$ and $\phi^{\mathrm{PR}\beta}$ satisfy

$$\Xi_p^\alpha\ \phi^{\mathrm{PR}\alpha} = \phi^{\mathrm{PR}\alpha} \quad \text{and} \quad \Xi_p^\beta\ \phi^{\mathrm{PR}\beta} = \phi^{\mathrm{PR}\beta} \tag{18}$$

The convex combination parameter p has the interpretation that a random walker can randomly jump to any other node. In all ranking comparisons, we use the value $p = 0.95$.

3.7 Random walker method (RW)

The random walker method of Callaghan, Mucha, and Porter (2007) considers the system of differential equations

$$\dot{\phi} = \Gamma_p^\alpha\ \phi \quad \text{where} \quad \Gamma_p^\alpha := \left[p\left(W - \mathrm{diag}(l)\right) + (1-p)(W^t - \mathrm{diag}(w))\right]. \tag{19}$$

Here $p \in (\frac{1}{2}, 1)$ is taken to be a bias parameter. This system has the interpretation of a collection of "random walkers" on the directed graph representing the schedule which transition from node i to node j $(i \neq j)$ with probability $pW + (1-p)W^t$. Thus the walkers tend to move toward the teams which win games with probability p. When p is near $\frac{1}{2}$, the strength of schedule dominates and when p is near 1, the game results dominate. In all ranking comparisons, we use the value $p = 0.75$. The statement that the total number of walkers is conserved is written $\mathbf{1}^t\Gamma_p^\alpha = 0^t$. Similarly, we consider

$$\dot{\phi} = \Gamma_p^\beta\ \phi \quad \text{where} \quad \Gamma_p^\beta := \left[p\left(S - \mathrm{diag}(t)\right) + (1-p)\left(S^t - \mathrm{diag}(s)\right)\right]. \tag{20}$$

Note $\mathbf{1}^t\Gamma_p^\beta = 0^t$. The rankings, $\phi^{\mathrm{RW}\alpha}$ and $\phi^{\mathrm{RW}\beta}$ are defined to be the stationary state of the equations (19) and (20), *i.e.*, the solutions of

$$\Gamma_p^\alpha\ \phi^{\mathrm{RW}\alpha} = 0 \quad \text{and} \quad \Gamma_p^\beta\ \phi^{\mathrm{RW}\beta} = 0. \tag{21}$$

3.8 Elo's method (E)

Elo's method for rating is an iterative method where the ratings are updated after each game (Elo, 1961, Glickman, 1995). Let $\phi_i^{\mathrm{E}\alpha}$ be the rating of team i. For each game, the Elo method models the expected outcome as a logistic function applied to the difference in team ratings. That is, if $L(x) = \frac{1}{1+10^{-x/\xi}}$ is the logistic function with parameter $\xi > 0$, then the expected outcome of team i in a game against team j is

$$\mu_{ij} = L(\phi_i - \phi_j).$$

Note $0 < \mu_{ij} < 1$ and $\mu_{ij} + \mu_{ji} = 1$. For the (α) dataset, the Elo method takes the outcome for team i against team j to be

$$o_{ij}^{\alpha} = \begin{cases} 1 & \text{team } i \text{ beats team } j \\ \frac{1}{2} & \text{teams } i \text{ and } j \text{ tie} \\ 0 & \text{otherwise} \end{cases}$$

The ratings are then additively updated by a factor proportional to the difference between the observed and expected outcomes of the game, *i.e.*,

$$\begin{aligned} \phi_i^{\mathrm{E}\alpha} &\leftarrow \phi_i^{\mathrm{E}\alpha} + K(o_{ij}^{\alpha} - \mu_{ij}) \\ \phi_j^{\mathrm{E}\alpha} &\leftarrow \phi_j^{\mathrm{E}\alpha} + K(o_{ji}^{\alpha} - \mu_{ji}). \end{aligned}$$

For the (β) dataset, consider a game where the score between teams i and j is σ_i to σ_j. Then the outcome for the game is taken to be

$$o_{ij}^{\beta} = \frac{\sigma_i + 1}{\sigma_i + \sigma_j + 2}.$$

The Elo ratings for the (β) dataset are similarly updated according to

$$\begin{aligned} \phi_i^{\mathrm{E}\beta} &\leftarrow \phi_i^{\mathrm{E}\beta} + K(o_{ij}^{\beta} - \mu_{ij}) \\ \phi_j^{\mathrm{E}\beta} &\leftarrow \phi_j^{\mathrm{E}\beta} + K(o_{ji}^{\beta} - \mu_{ji}). \end{aligned}$$

In all ranking comparisons, we initialize all teams to have rating 1500 and use $K = 32$ and $\xi = 1000$. These parameter choices agree with those of Langville and Meyer (2012).

Remark. Other implementations of the Elo algorithm treat a teams rating as provisional until the team has played a fixed number, say 20, games. Some implementations reduce the parameter K as the number of games increases. TrueSkill is an adaption of the Elo method by Microsoft for multi-player online gaming where a team's performance is modeled as a normal distribution (Minka, Graepel, and Herbrich, 2007). After each game, both the mean and variance are updated based on the difference between the expected and observed game outcomes.

4 Comparing methods via hypothesis testing

In this section, we give a brief overview of hypothesis testing methods, which will be used to quantitatively compare the ranking methods described in §3. Our discussion closely follows Demšar (2006). See also Shaffer (1995).

We consider comparing M methods on D datasets. Let c_d^ℓ be a performance measure for the ℓ-th method on the d-th data set. (In §5, we'll take c_d^ℓ to be a cross validation score, $M = 16$, and D to be the number of seasons for a single sport.) Sorting the values of c_d^ℓ with respect to ℓ, we let r_d^ℓ be the rank of the ℓ-th method on the d-th dataset. A rank of r_d^ℓ indicates that method ℓ outperformed $M - r_d^\ell$ methods on the d-th dataset. In the event that $n > 1$ methods have the same cross validation score c_d^ℓ, we assign each of the tied methods the average ranking. For example, if in a comparison of 4 methods, the performance measures $c = [3.2, 2, 2, 1.5]$ are obtained for a given dataset, the corresponding rankings are $r = [1, 2.5, 2.5, 4]$.

Perhaps the simplest method for comparing ranking methods would be to simply compare the average rank of each algorithm over the datasets,

$$R_\ell = \frac{1}{D} \sum_{d=1}^{D} r_d^\ell. \tag{22}$$

However, this method of comparison is susceptible to variations in algorithm performance since excellent performance on a small subset of the dataset may compensate for a general lackluster performance. (Demšar, 2006).

The Friedman test takes as null hypothesis (H_0) that all ranking methods are equivalent. The alternative is that at least two of the methods differ in predictiveness. For each dataset, d, the sum of the ranks of the algorithms is given by $\sum_{\ell=1}^{M} r_d^\ell = \frac{M(M+1)}{2}$. The average of the ranks over all datasets is

then given by $\overline{R} = \frac{M+1}{2}$. Define the "sum of squares" quantities

$$SS_t = D\sum_{\ell=1}^{M}(R_\ell - \overline{R})^2 \;=\; D\left(\sum_{\ell=1}^{M} R_\ell^2\right) - D\frac{M(M+1)^2}{4}$$

$$SS_e = \frac{1}{D(M-1)}\sum_{\ell=1}^{M}\sum_{d=1}^{D}\left(r_d^\ell - \overline{R}\right)^2 = \frac{M(M+1)}{12}.$$

Here, SS_t is proportional to the variance of the average rank of the algorithms over the datasets and SS_e is proportional to the sample variance of the ranks. If H_0 holds, then we would expect that SS_t to be small compared to SS_e. Friedman considered the statistic

$$\chi_F^2 = \frac{SS_t}{SS_e} = \frac{12D}{M(M+1)}\left(\sum_{\ell=1}^{M} R_\ell^2\right) - 3D(M+1). \tag{23}$$

If M and D are large, then the null distribution of χ_F^2 can be approximated by the chi-square distribution with $M-1$ degrees of freedom. For smaller datasets, the null distribution can be computed (Demšar, 2006).

A less conservative statistic is given by

$$F_F = \frac{(D-1)\chi_F^2}{D(M-1) - \chi_F^2}. \tag{24}$$

The statistic F_F follows the F distribution with $M-1$ and $(M-1)(D-1)$ degrees of freedom. Thus, if F_F exceeds the critical value for the chosen significance level, α, then we can reject the null hypothesis, H_0. Critical values for the chi-squared and F cumulative distribution functions can be computed in Matlab using `chi2cdf` and `fcdf`.

If the Friedman test rejects the null hypothesis, then the post-hoc Nemenyi test can be applied. If the difference in average rank between two algorithms i and j exceeds a critical difference $\Delta_{\alpha,M,D}$, *i.e.*, $R_i - R_j > \Delta_{\alpha,M,D}$, then the performance of algorithm i is better than the performance of algorithm j with confidence α. The critical difference is given by

$$\Delta_{\alpha,M,D} = q_{\alpha,M}\sqrt{\frac{M(M+1)}{6D}} \tag{25}$$

where $q_{\alpha,M}$ is drawn from the studentized range distribution and depends on the significance level α as well as M, the number of methods compared. We obtained values of $q_{\alpha,M}$ from Miwa (2012).

sport	years	mean # teams	mean # games	mean alg. conn.	source
NBA	1980-2011	27	1093	52.5	NBA (2012)
MLB	1997-2012	29.7	1621	33.24	MLB (2012)
MLB-NL	1901-1996	9.4	730.1	163.9	MLB (2012)
MLB-AL	1901-1996	9.8	757.5	166.7	MLB (2012)
NCAAB	1991-2012	320.0	4,625	4.27	CBB (2012)
NCAAF	1956-2011	125.6	645.6	1.09	CFB (2012)

Table 1: Description of datasets analyzed. The algebraic connectivity, defined in (26), is a measure of the rankability of the dataset. See §5.1

5 Comparison of Ranking Methods

In this section, we describe the datasets, a cross-validation based measure of ranking error, and the results of hypothesis testing.

5.1 Data

We consider data from 4 sports leagues: 32 National Basketball Association (NBA) regular seasons, 112 Major League Baseball (MLB) regular seasons, 22 NCAA Basketball (NCAAB) Division 1 regular seasons, and 56 NCAA Football Division 1-A or FBS (NCAAF) seasons. In Table 1, we include a description of each dataset, including the mean number of games, mean number of teams, and algebraic connectivity of the directed graph representing the schedule, and source information. The algebraic connectivity of a directed graph with arc-vertex incidence matrix B is defined

$$\text{algebraic connectivity} = \min_{v^t\mathbf{1}=0} \frac{\|Bv\|^2}{\|v\|^2}. \tag{26}$$

Algebraic connectivity is a measure of the informativeness of the schedule represented by the graph (Osting et al., 2012a). Intuitively, a schedule with large algebraic connectivity has an associated ranking which is robust to anomalous games ("upsets").

In what follows, we comment on the general structure of each sports league.

NBA. For NBA, since 1968 each team has generally played 82 games during the regular season, 41 each home and away. Exceptions include the 1999

NBA lockout, where teams played only 50 games each. Although the teams are subdivided into conferences and divisions, the teams all play one another. Overtime prevents ties in the NBA. In the NBA, winning percentages are used to rank teams (and decide which teams will enter the postseason playoffs).

MLB. In MLB, the teams are subdivided into two leagues: the American League (AL) and the National League (NL). Prior to 1997, teams only played within their own league during the regular season. Consequently, for these years we consider the AL and NL separately. For the years considered, teams played on average 157 games. The variance in games is due to rain or player strikes. In MLB, rules allowing for "extra innings" reduce the number of ties, but they have occurred. In MLB, winning percentages are used to rank teams (and decide which teams will enter the postseason playoffs).

NCAAB. Our dataset includes only regular season games. For each season considered, we (recursively) prune the dataset by removing all teams which play less than 10 games. On average, we remove 65.4 teams and 94.0 games per year from the dataset. Overtime prevents ties in NCAAB. In NCAAB, the home-adjusted RPI method is used to rank teams (and decide which teams will enter the postseason "March madness" playoffs).

NCAAF. Our dataset includes both regular season and postseason "bowl" games. In NCAAF, Division 1-A teams mostly play ≈ 12 games per season against both Division 1-A and 1-AA teams. However, if one considers only games played between two Div. 1-A teams, then one finds that some teams play very few games. For each season considered, we (recursively) prune the dataset by removing all teams which play less than 4 games. On average, we remove 42.4 teams and 53.2 games per year from the dataset. On average, the remaining teams play 10 games each. The game to team ratio for the NCAAF is much lower than for NBA or MLB. In 1996, overtime was introduced to eliminate ties. The current ranking method used in NCAAF is to aggregate several rankings—some (proprietary) "computer-generated" methods and others based on expert opinion.

5.2 Ranking Error

To evaluate the prediction accuracy of ranking methods, we require a notion of error for a ranking. One can conceive of many possible ways of doing this. For example, if the goal of the ranking is to predict margin of victory in a

game, the error could be defined as the difference between the predicted and observed victory margins. Here, we simply try to predict which team wins each game. To evaluate this, we use cross validation. Each dataset is broken into two subsets: a training set and a test set. A ranking ϕ is generated by applying one of the methods described in §3 to the training set. We define the *prediction error*, $E^\ell(t)$, of a ranking, ℓ, for a given test set, t, to be the proportion of games in the test set such that the lower rated team beats the higher rated team, *i.e.*,

$$E^\ell(t) = \frac{\#\{\text{team } i \text{ beats } j \text{ in the dataset } t \text{ and } \phi_i^{(\ell)} \leq \phi_j^{(\ell)}\}}{\#\{\text{games in dataset } t\}} + \frac{\#\{\text{teams } i \text{ and } j \text{ tie and } \phi_i^{(\ell)} \neq \phi_j^{(\ell)}\}}{\#\{\text{games in dataset } t\}}. \tag{27}$$

The best performing ranking method is the one that has the lowest expected predictive error.

We use k-fold cross-validation where $k = 20$ to evaluate the prediction error for each method described in §3. More specifically, the data is partitioned into $k = 20$ disjoint subsets, and each subset is used once to test the accuracy of a ranking generated from the other $k - 1$ subsets. The average error across the k subsets is used as a cross-validation score. This process is repeated $\rho = 100$ times to average over the choice of partitions. We label the partitions P_j, for $j = 1, \ldots, \rho$. Thus, for each dataset d, taken to be one season, we define the cross validation score for ranking method ℓ,

$$c_d^\ell = \frac{1}{\rho} \sum_{j=1}^{\rho} \left(\frac{1}{k} \sum_{t \in \mathrm{P_j}} E^\ell(t) \right). \tag{28}$$

To generate a sensible ranking on each training set, we require that the partitions be chosen such that, for each training set, each team plays at least one game. We find that for $k = 20$, a randomly chosen partition will usually have this property, and we find that for each dataset considered, the property can be obtained by randomly selecting partitions and discarding partitions which do not satisfy this property. The same partitions are used to evaluate the cross validation score for each ranking method, ℓ.

In Fig. 1, for each dataset (sport), we plot c_d^ℓ, defined in (28), for each method ℓ as a function of season, d. In Fig. 2, for each dataset and each ranking method ℓ, we give a box and whisker plot for the distribution of cross validation scores, c_d^ℓ over the seasons d.

From Figs. 1 and 2, we observe that the cross validation scores are generally smallest for the MLB dataset and interpret this to mean that MLB games are generally less predictable than the other sports considered. From Fig. 1, we observe that the cross validation scores of all of the methods vary from season to season together, *i.e.*, have high correlation. We interpret this to mean that some seasons are more predictable than others. For the NCAAB dataset, the cross validation scores have much less inter-season variability than, for example, the NCAAF dataset.

From Figs. 1 and 2, we observe that the cross validation scores for some sports (NCAAB and NCAAF) have higher inter-method variability than others (MLB and NBA). The sports with high inter-method variability are precisely those with schedules represented by graphs with small algebraic connectivity as defined in (26), (see Table 1).

We observe that the PRβ ranking method has a considerably smaller cross validation score for the NBA dataset than the other methods. This is because score differentials in the NBA tend to be small compared to the score magnitudes and thus the matrix S, as defined in (2), is almost symmetric. Consequently, the matrix Ξ_p^β in (18) represents a walk on a directed graph where walkers are only weakly directed toward the winning team.

To measure the similarity of rankings generated by the considered ranking methods, we use the Kendall tau distance. Given two rankings, σ^1 and σ^2 for n teams, the *Kendall tau distance* is defined

$$d_K(\sigma^1,\sigma^2) := \frac{\#\{\text{teams } i \text{ and } j\colon (\sigma_i^1 > \sigma_j^1 \text{ and } \sigma_i^2 < \sigma_j^2) \text{ or } (\sigma_i^1 < \sigma_j^1 \text{ and } \sigma_i^2 > \sigma_j^2)\}}{\frac{1}{2}n(n-1)}.$$

Note that for all rankings σ^1 and σ^2, $0 \leq d_K(\sigma^1,\sigma^2) \leq 1$. Let $\pi^d(\ell)$ be the ranking generated by rating method ℓ on the dataset d. Define the matrix

$$\Phi^d_{\ell_1,\ell_2} = d_K\left(\pi^d(\ell_1), \pi^d(\ell_1)\right).$$

We then compute the average Frobenius norm of Φ^d over the seasons d,

$$\gamma = \frac{1}{D}\sum_{d=1}^{D} \|\Phi^d\|_F.$$

For each sport, we tabulate the value γ as follows:

	NBA	MLB	NL	AL	NCAAB	NCAAF
γ	1.02	1.94	1.44	1.40	2.23	2.70

Comparing NBA and NCAAF, it appears that a small variance in the cross validation scores might imply that the rankings themselves are similar. However, the cross validations scores across rankings for MLB has very small variance, but the rankings are fairly dissimilar compared to NBA. This is caused by the low cross validation scores for MLB for all rankings. Thus, similar cross validation scores across rankings for a given sport does not necessarily imply that the ranking methods produced similar rankings.

Finally, Fig. 1 illustrates the difficulty in describing the probability distribution of the cross validation score for each individual rating method. A parametric method for comparing cross validation scores would require many assumptions on these distributions and a method for taking into account inter-season variability; this is one of the primary motivations for employing non-parametric hypothesis testing methods.

5.3 Hypothesis testing

In this section, we apply the hypothesis testing method described in §4 to the datasets. The average rank of each algorithm over the seasons, given in (22), is computed using the cross validations scores, (28). For each dataset (sport), the average rank of each algorithm is indicated by the blue bar in Fig. 3. For each dataset, the Friedman statistics χ_F^2 and F_F, defined in Eqs. (23) and (24), exceed their respective critical values at the 99% c.l. (confidence level). Thus, in all cases we reject the null hypothesis—that each of the compared methods (described in §3) have equivalent predictive power.

Proceeding to the Nemenyi test, we compute the critical value given in (25) for each sport league. The red bar in Fig. 3 represents the Nemenyi critical value at the 95% c.l.. If the difference between the average ranks (blue bars) for two methods exceeds the distance indicated by the red bar, then the Nemenyi test concludes that the method with the smaller average ranking has statistically better predictive accuracy (at this confidence level). The Nemenyi critical values for comparing $M = 8$ and $M = 16$ methods at the 95% and 99% c.l.s are reported in Table 2.

In Fig. 3, we observe that the methods utilizing score-differential data (β) are usually more predictive than those utilizing win-loss data (α) only, although not always significantly more predictive. We also observe that there is no clear "most predictive" method for all sports considered. For the NBA dataset, we observe that the PRβ method is less predictive at the 99% c.l. than all but the PRα method. For each of the three MLB datasets, the ranking methods perform relatively similarly. For the MLB-NL dataset, the

PRβ method has less predictive accuracy at the 99% c.l. than 4 other methods. For the MLB-NL dataset, the Kβ method is less predictive at the 99% c.l. than 4 other methods. For the MLB-AL dataset, only the Eα method has worse predictive accuracy at the 99% c.l. than any of the other methods. For the NCAAB dataset, the RPIβ, L2β, MPβ, and RWβ methods each have higher predictive accuracy at the 99% c.l. than 8 other methods. In particular, the L2β and RWβ methods are more predictive at the 95% c.l. than the RPIα method. This is interesting because the official ranking method for NCAAB is home-adjusted RPI, a variant of the RPIα method. For the NCAAF dataset, the L2β and RWβ methods have higher predictive accuracy at the 99% c.l. than 10 and 9 other methods respectively.

Comparison of restricted methods. We repeat the Friedman and Nemenyi tests except only comparing ranking methods which utilize dataset (β) and plot the results in Fig. 4. For the NBA dataset, the PRβ method has lower predictive accuracy at the 95% c.l. than all other methods. For the MLB dataset, the Kβ method has the highest predictive accuracy and is more predictive at the 99% c.l. than the PRβ and Eβ ranking methods. For the MLB-NL dataset, the PRβ method has worse predictive accuracy at the 99% c.l. than all but the Eβ ranking method. For the MLB-AL dataset, only the WPβ ranking method is more predictive at the 95% c.l. than the PRβ and Eβ ranking methods. For the NCAAB dataset, the WPβ and Eβ methods are less predictive at the 95% c.l. than 5 other ranking methods. For the NCAAF dataset, the L2β method has the highest predictive accuracy and is more predictive at the 99% c.l. than all other β rankings considered, except RWβ. For the NCAAF dataset, the L2β and method RWβ methods are more predictive at the 95% c.l. than all other β rankings considered.

Cases where the Friedman test fails to reject H_0. By excluding some methods we found several examples where the Friedman test can fail to reject the null hypothesis, H_0, including the following. For the NBA dataset, if the PRβ, Eβ, and all α ranking methods are excluded, than the Friedman test does *not* reject H_0. For the MLB-AL dataset, if PRβ is excluded, the p-value for both the χ_F^2 and F_F statistics is 0.043. Thus, the Friedman test rejects H_0 at the 95% c.l., but accepts H_0 at the 99% c.l..

6 Discussion

In this paper, we empirically evaluated the predictive power of eight sports ranking methods, described in §3, and labelled WP, RPI, L2, MP, K, PR, RW, and E. For each ranking method, we implement two versions, one using only win-loss data (α) and one utilizing score-differential data (β). The methods were compared on data from four different sports leagues, NBA, MLB, NCAAF, and NCAAB. For each season of each dataset, we applied 20-fold cross validation to evaluate the predictive accuracy of the ranking methods. The non-parametric Friedman hypothesis test and post-hoc Nemenyi tests were used to assess whether the predictive error for the considered rankings over the seasons were statistically dissimilar and which ranking methods had significantly superior predictive accuracy. We found in all cases that the null hypothesis—that all ranking methods are equivalent—is rejected at the 99% confidence level. For NCAAF and NCAAB datasets, the Nemenyi test concludes that the implementations utilizing score-differential data (β) are usually more predictive than those using only win-loss data (α). For the NCAAF dataset, the least squares (L2β) and random walker (RWβ) methods have significantly better predictive accuracy than the other methods considered.

Our primary goal in this work was to demonstrate how hypothesis testing methods can be used for comparing sports ranking methods, not for finding the "best ranking method". We reiterate that we have only considered a small sample of ranking methods and have not incorporated all available information into the rankings, for example, home/away information, separate offensive/defensive ratings, *etc.*... Additionally, we have not "tuned" any method parameters for the individual datasets considered, although this would certainly improve the predictive accuracy of the methods. A natural future direction is to use the comparison methodology described here to compare a larger sample of rating methods or develop new rankings with improved predictive accuracy. It would also be interesting to consider the predictive power of a ranking obtained by rank aggregation (Dwork, Kumar, Naor, and Sivakumar, 2001a,b), especially as compared to the predictive power of its constituents.

Acknowledgements

This work was supported by the California Research Training Program in Computational and Applied Mathematics (NSF grant DMS-1045536). Braxton Osting is partially supported by ONR grant N000141010221, ONR grant

N000141210838, AFOSR MURI grant FA9550-10-1-0569, and a National Science Foundation (NSF) Postdoctoral Fellowship DMS-11-03959.

References

Bradley, R. A. and M. E. Terry (1952): "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, 39, 324–345.

Burer, S. (2012): "Robust rankings for college football," *JQAS*.

Callaghan, T., P. J. Mucha, and M. A. Porter (2007): "Random walker ranking for NCAA Division I-A football," .

CBB (2012): "http://www.sports-reference.com/cbb/," webpage accessed November 1, 2012.

CFB (2012): "http://www.sports-reference.com/cfb/," webpage accessed October 29, 2012.

Chan, V. (2011): "Prediction accuracy of linear models for paired comparisons in sports," *JQAS*, 7.

Chartier, T. P., E. Kreutzer, A. N. Langville, and K. E. Pedings (2011a): "Sensitivity and stability of ranking vectors," *SIAM J. Sci. Comput.*, 33, 1077–1102.

Chartier, T. P., E. Kreutzer, A. N. Langville, and K. E. Pedings (2011b): "Sports ranking with nonuniform weighting," *JQAS*, 7.

Colley, W. N. (2002): "Colley's bias-free college football ranking method: The Colley matrix explained," .

David, H. A. (1963): *The Method of Paired Comparisons*, Charles Griffin & Co.

Demšar, J. (2006): "Statistical comparisons of classifiers over multiple data sets," *JMLR*, 7, 1–30.

Dwork, C., R. Kumar, M. Naor, and D. Sivakumar (2001a): "Rank aggregation methods for the web," in *Proceedings of the 10th international converence on World Wide Web. ACM*, 613–622.

Dwork, C., R. Kumar, M. Naor, and D. Sivakumar (2001b): "Rank aggregation revisited," in *Proceedings International Conference World Wide Web (WWW10)*.

Elo, A. E. (1961): "The new U.S.C.F. rating system," *Chess Life*, 16, 160–161.

Foulds, L. R. (1992): *Graph Theory Applications*, Springer.

Gill, R. (2009): "Assessing methods for college football rankings," *JQAS*, 5.

Glickman, M. E. (1995): "A comprehensive guide to chess ratings," *American Chess Journal*, 3.

Hirani, A. N., K. Kalyanaraman, and S. Watts (2011): "Least squares ranking on graphs," arXiv:1011.1716v4.

Hochbaum, D. S. (2010): "The separation and separation-deviation methodology for group decision making and aggregate ranking," *TutORials in Operations Research*, 7, 116–141.

Horn, R. A. and C. R. Johnson (1991): *Matrix Analysis*, Cambridge University Press.

Jiang, X., L.-H. Lim, Y. Yao, and Y. Ye (2010): "Statistical ranking and combinatorial Hodge theory," *Math. Program. Ser. B*, 127, 203–244.

Keener, J. P. (1993): "The Perron-Frobenius theorem and the ranking of football teams," *SIAM Review*, 35, 80–93.

Langville, A. N. and C. D. Meyer (2012): *Who's #1?: The Science of Rating and Ranking*, Princeton University Press.

Massey, K. (1997): *Statistical models applied to the rating of sports teams*, Master's thesis, Bluefield College.

Minka, T., T. Graepel, and R. Herbrich (2007): "Trueskill™: A Bayesian skill rating system." *Advances in Neural Information Processing Systems*.

Miwa, T. (2012): "http://cse.niaes.affrc.go.jp/miwa/probcalc/s-range/," webpage accessed November 26, 2012.

MLB (2012): "http://www.baseball-reference.com/," webpage accessed October 29, 2012.

NBA (2012): "http://www.basketball-reference.com/," webpage accessed October 29, 2012.

Osting, B., C. Brune, and S. Osher (2012a): "Optimal data collection for improved rankings expose well-connected graphs," Submitted.

Osting, B., J. Darbon, and S. Osher (2012b): "Statistical ranking using the ℓ^1-norm on graphs," Submitted.

Page, L., S. Brin, R. Motwani, and T. Winograd (1999): "The PageRank citation ranking: bringing order to the web," Technical report, Stanford InfoLab Technical Report 1999-66.

Pickle, D. and B. Howard (1981): "Computer to aid in basketball championship selection." *NCAA News*, 4.

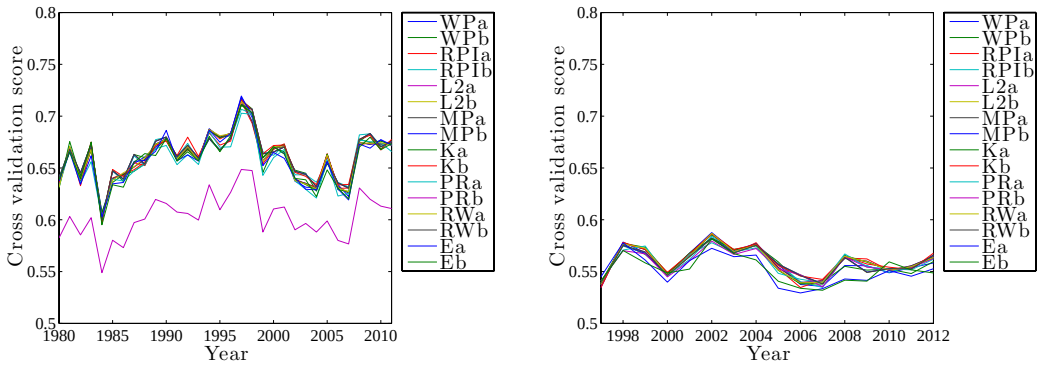
Shaffer, J. P. (1995): "Multiple hypothesis testing," *Annu. Rev. Psychol.*, 46, 561–584.

Stefani, R. (2011): "The methodology of officially recognized international sports rating systems," *JQAS*, 7.

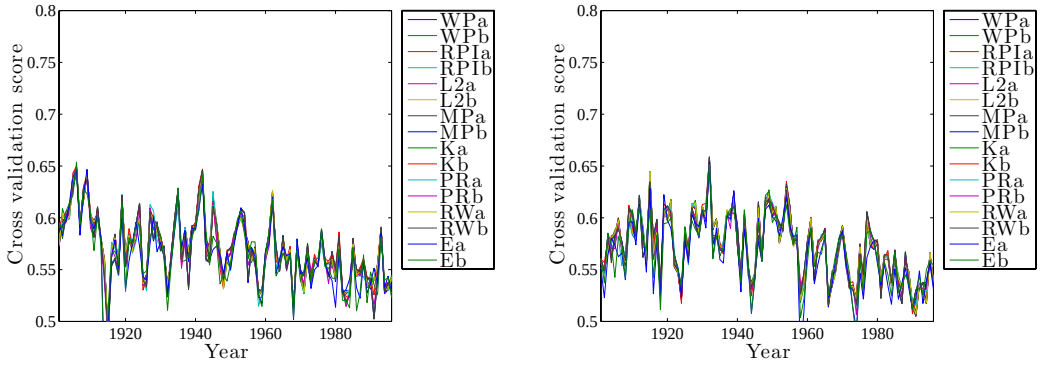
Tran, N. M. (2011): "Pairwise ranking: choice of method can produce arbitrarily different rank order," arXiv:1103.1110v1.

Trono, J. A. (2010): "Rating/ranking systems, post-season bowl games, and "the spread"," *JQAS*, 6.

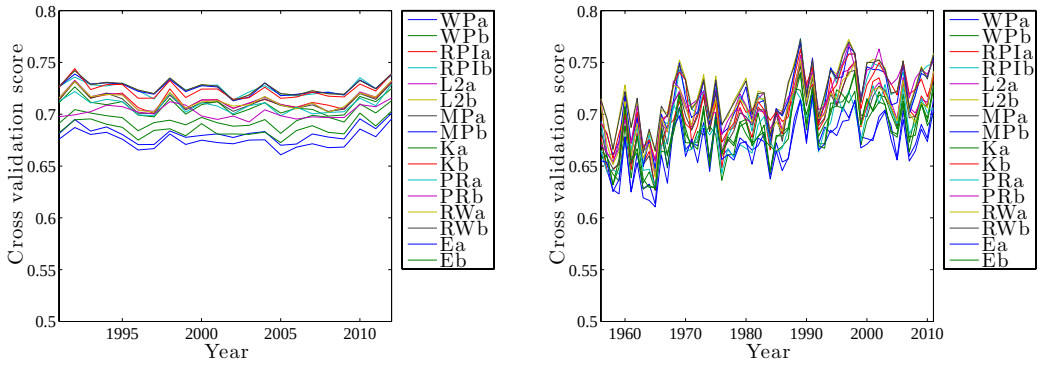
Xu, Q., Y. Yao, T. Jiang, Q. Huang, B. Yan, and W. Lin (2011): "Random partial paired comparison for subjective video quality assessment via HodgeRank," in *ACM Multimedia*.



(a) NBA (b) MLB

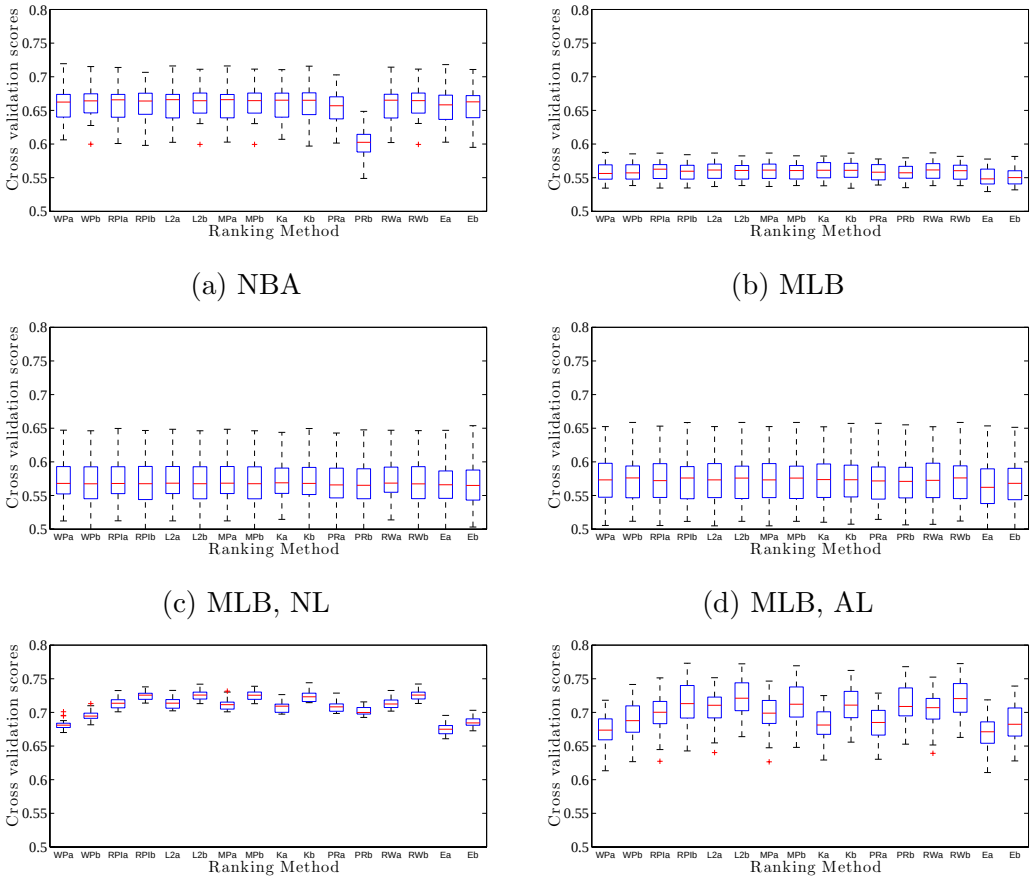


(c) MLB, NL (d) MLB, AL



(e) NCAAB (f) NCAAF

Figure 1: For each dataset (sport), we plot the cross validation scores, c_d^ℓ, as defined in (28), for each method, ℓ, as a function of season, d. The dataset for each sport is described in Table 1. A large cross validation score indicates that the method has good predictive accuracy. See §5.2.



(a) NBA

(b) MLB

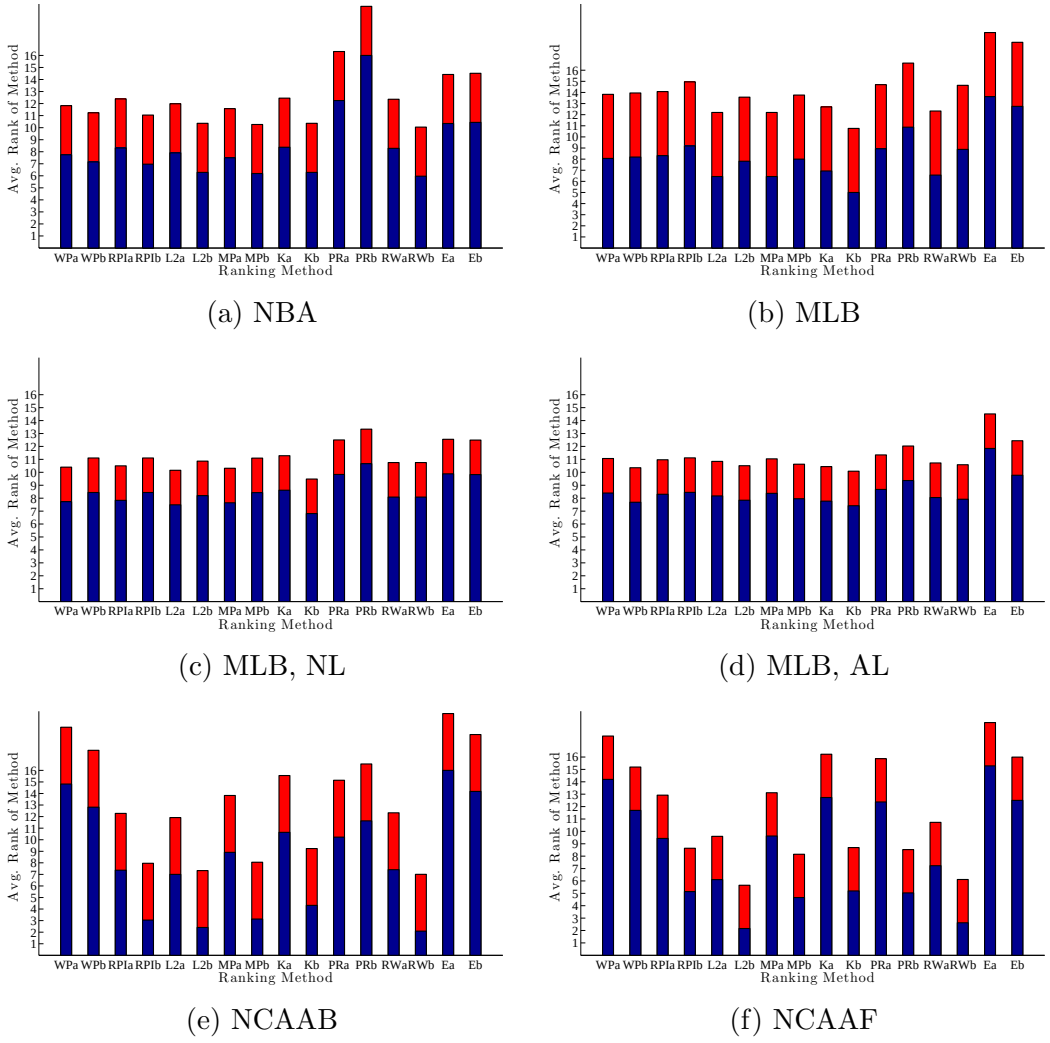
(c) MLB, NL

(d) MLB, AL

(e) NCAAB

(f) NCAAF

Figure 2: For each dataset (sport) and each ranking method ℓ, we give a box and whisker plot for the distribution of cross validation scores c_d^ℓ, defined in (28), over the seasons d. The median is indicated by the red line, the first and third quartile are indicated by the blue box, and the 'whiskers' extend to the most extreme data points not considered outliers. The outliers are plotted individually by red (+) markers. See §5.2.

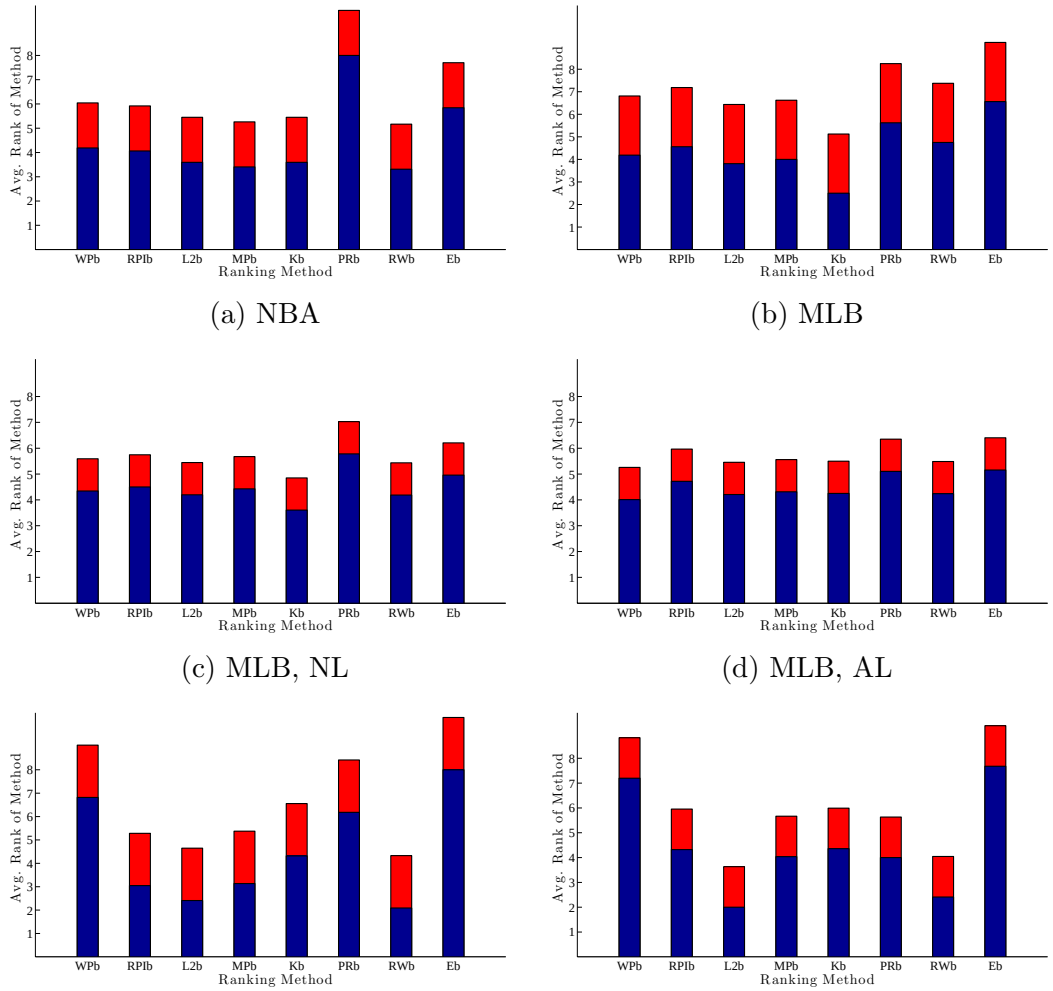


(a) NBA (b) MLB

(c) MLB, NL (d) MLB, AL

(e) NCAAB (f) NCAAF

Figure 3: For each dataset, the blue bars indicate the average rank of each algorithm over the seasons, as defined in (22). Lower average rank indicates that the method has better predictive accuracy. The red bars indicate the Nemenyi critical distance at the $\alpha = 0.05$ significance level, defined in (25). If the difference between the average ranks for two methods exceeds the distance indicated by the red bar, then Nemenyi test concludes that the method with the smaller average ranking has statistically better predictive accuracy. See §5.3.



(a) NBA (b) MLB

(c) MLB, NL (d) MLB, AL

(e) NCAAB (f) NCAAF

Figure 4: This figure is similar to Fig. 3, except only rankings depending on the score differential data, labelled (β), are compared. See §5.3.

$M = 8$

	NBA $D = 32$	MLB $D = 16$	NL $D = 96$	AL $D = 96$	NCAAB $D = 22$	NCAAF $D = 56$
$\Delta_{\alpha,M,D}$, $\alpha = 95\%$	1.86	2.62	1.07	1.07	2.24	1.40
$\Delta_{\alpha,M,D}$, $\alpha = 99\%$	2.16	3.05	1.25	1.25	2.60	1.63

$M = 16$

	NBA $D = 32$	MLB $D = 16$	NL $D = 96$	AL $D = 96$	NCAAB $D = 22$	NCAAF $D = 56$
$\Delta_{\alpha,M,D}$, $\alpha = 95\%$	4.08	5.77	2.35	2.35	4.92	3.08
$\Delta_{\alpha,M,D}$, $\alpha = 99\%$	4.62	6.54	2.67	2.67	5.58	3.49

Table 2: Nemenyi test critical values, $\Delta_{\alpha,M,D}$, for comparing M ranking methods among D seasons at the α confidence level. See §5.3 and Fig. 3.