Journal of Quantitative Analysis in Sports

Volume 6, *Issue* 3 2010 *Article* 4

An Improved LRMC Method for NCAA Basketball Prediction

Mark Brown, *City College, City University of New York*
Joel Sokol, *Georgia Institute of Technology*

Recommended Citation:
Brown, Mark and Sokol, Joel (2010) "An Improved LRMC Method for NCAA Basketball Prediction," *Journal of Quantitative Analysis in Sports*: Vol. 6: Iss. 3, Article 4.

DOI: 10.2202/1559-0410.1202

An Improved LRMC Method for NCAA Basketball Prediction

Mark Brown and Joel Sokol

Abstract

The LRMC method for predicting NCAA Tournament results from regular-season game outcomes is a two-part process consisting of a logistic regression model to estimate head-to-head differences in team strength, followed by a Markov chain model to combine those differences into an overall ranking. We consider replacing each of the two parts of LRMC with alternative models, empirical Bayes and ordinary least squares, that attempt to accomplish the same goal. Computational results show that replacing the logistic regression with either of two empirical Bayes models yields a statistically-significant improvement when the probabilities are jointly conditioned.

KEYWORDS: LRMC, logistic regression, Markov chain, empirical Bayes, college basketball, NCAA basketball

Author Notes: Brown is with the Department of Mathematics, City College, City University of New York. Sokol is with the Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology. Corresponding author: joel.sokol@isye.gatech.edu. Brown was supported by the National Security Agency, under grant H98230-06-01-0149. The authors would like to thank an anonymous referee for his helpful comments and suggestions.

1. Introduction

The annual National Collegiate Athletic Association (NCAA) Division I men's basketball tournament is the most-wagered-upon sporting event in the United States. It is a 6-round, 64-team single-elimination tournament played over three consecutive long weekends on neutral courts (so that no team plays on its own home court).

In most other sporting events in the United States, bets are generally placed on each individual game (or series of games between a pair of teams). In contrast, a popular way to bet on the NCAA Tournament is to predict the results of all 63 games before any of the games have been played. A 63-game prediction of this sort is often referred to as a *bracket*.

As one might expect, predicting the outcome of a 63-game tournament is difficult; to the best of our knowledge, nobody has ever made a perfect bracket prediction. However, there are a number of well-known rankings of college basketball teams that each implicitly make a prediction (assuming that the better team will beat the worse team in each tournament game). These rankings include national polls of sportswriters (conducted by the Associated Press) and coaches (conducted jointly by ESPN and *USA Today*), the Ratings Percentage Index (RPI) developed by the NCAA, and algorithm-based methods such as the Sagarin ratings (Sagarin, 2000-2009) and the Massey ratings (Massey, 2000).

In the last few years, a model called LRMC has been shown to be a more-accurate predictor of NCAA Tournament outcomes than the others, both game-by-game and in the more-important final rounds (Kvam and Sokol, 2006). LRMC has two pieces: a logistic regression (LR) model that estimates the probability that one team is better than another based on the outcome of a head-to-head game between the two, and a Markov chain (MC) model that combines all of these single-game estimates.

A reasonable question is whether either piece of LRMC can be improved by using a different (and specifically, more directly-focused) model for one or both parts of the process. In this paper, we present results for several related LRMC-type methods. To replace the logistic regression, we consider two separate empirical Bayes models, one that focuses specifically on predicting the outcome of individual games on neutral courts like the NCAA Tournament, and the other focuses on attempting to directly identify the best teams. Both methods have two variants: one uses each regular-season game as its own individual piece of information, and the other combines information from multiple regular-season games between the same two teams. We also consider an ordinary least squares model instead of the Markov chain in order to more directly combine the head-to-head estimates.

We note that recently, meta-methods such as those of West (2006a and 2006b) and Coleman and Lynch (2009) have been published. Both of these methods use rankings from other sources as input; for example, both use the RPI rankings, and West's method uses Sagarin's rankings as well. On the other hand, LRMC (and our improvements), Sagarin's and Massey's rankings, RPI, etc. are designed as standalone prediction systems based on simple data. Therefore, we usually do not make the apples-to-oranges comparison between LRMC et al. and meta-methods like those of West and Coleman and Lynch. However, in this paper we do descend into mixed-fruit comparisons a bit when we present our computational results.

The rest of this paper is organized as follows. In Section 2, we review the original LRMC model of Kvam and Sokol. In Section 3, we introduce our new LRMC-type approaches, and in Section 4 we provide computational results.

2. The LRMC Model

The LRMC model of Kvam and Sokol is a two-stage procedure. In the first stage, each game between two Division I opponents that season is evaluated to determine the probability, conditional on the margin of victory and the location of the game, that the winning team is better than the losing team. In the second stage, those single-game probabilities are used as transition probabilities in a Markov chain where each team is represented by a state. The teams are then ranked in descending order of their state's Markov chain steady-state probability. After describing each of these two stages more completely, we introduce in Section 3 alternative methods for each stage.

2.1. The Markov Chain

LRMC's Markov chain model follows that of Callaghan et al. (2004), but is generalized to allow for the outcome-specific probabilities generated in the first stage. Each state in the Markov chain corresponds to a Division I college basketball team, with one state per team.

For each game $g = (i,j)$ where i is the visiting (road) team and j is the home team, let $x(g)$ be the margin by which the home team's score exceeds the road team's score; of course, $x(g)$ will be negative when the road team has won the game. Scores at the end of regulation time (40 minutes) are used, so $x(g) = 0$ for all overtime games regardless of the eventual winner.

For any game where the home team scored x more points than the road team, let r_x be the probability that the home team is better than the road team. Then the transition probability between the road team's state to the home team's state is increased by r_x/N_{road} (where N_{road} is the number of games played by the

DOI: 10.2202/1559-0410.1202

road team) and the transition probability from the road team's state to itself is increased by $(1-r_x)/N_{road}$. Similarly, the transition probability between the home team's state to the road team's state is increased by $(1-r_x)/N_{home}$ and the transition probability from the home team's state to itself is increased by r_x/N_{home}.

Overall, the transition probabilities between team states i and j are defined as

$$t_{ij}^R = \frac{1}{N_i}\left[\sum_{g=(i,j)} r_{x(g)} + \sum_{g=(j,i)}\left(1-r_{x(g)}\right)\right], \qquad \text{for all } j \neq i, \qquad (1)$$

$$t_{ii}^R = \frac{1}{N_i}\left[\sum_{g=(i,\cdot)}\left(1-r_{x(g)}\right) + \sum_{g=(\cdot,i)} r_{x(g)}\right]. \qquad (2)$$

Letting $T^R = [\,t_{ij}^R\,]$, the well-known Markov chain equations

$$\pi T^R = \pi, \qquad (3)$$

$$\textstyle\sum_i \pi_i = 1 \qquad (4)$$

can be solved to calculate the steady-state probability π_i for each team i.

2.2. Calibrating the Transition Probabilities

The difficult part of calibrating the LRMC model is determining values of r_x for each x. Recall that r_x is the probability that a team which has a margin of x points in a home game is better than the opponent it faced in that game. The difficulty is that the vast majority of college basketball games before the NCAA tournament take place on the home court of one of the two participants, which empirical evidence shows to be an advantage to the home team. Because NCAA tournament games all take place on neutral courts (i.e., courts that are not the home court of either participant), this home-court effect must be accounted for when regular-season scores are used to predict NCAA tournament outcomes.

Kvam and Sokol estimate r_x by a two-step process. First, they calculate s_x, the probability that a team which has a margin of x points in a home game would beat that same opponent at the opponent's home court. Then, they shift the curve of s_x to account for the difference between a game on the opponent's home court and a game at a neutral court.

To determine s_x, a logistic regression model is used. It takes all of the home-and-home matchups from four years of NCAA Division I basketball, and finds the best-fit estimate:

$$s_x = \frac{e^{0.0292x-0.6228}}{1+e^{0.0292x-0.6228}}. \quad (5)$$

Next, to determine r_x, the logistic regression curve is shifted to account for the difference between road and neutral-site games. Given two teams i and j, they assume that if i is d points stronger on a neutral court, then at home i has an advantage of $d+h$ points (where h is the size of the home-court advantage) and on the road i has an advantage of $d-h$ points. Thus, for the road matchup between i and an opponent to be even (50% chance of each team winning), d must equal h, and the home matchup would be a $2h$-point victory for team i.

Based on equation (5), $s_x = 0.5$ when $x = 0.6228/0.0292 \approx 21.33$. Thus, Kvam and Sokol define the home-court advantage to be $h = 21.33/2$, or about 10.67, and generate Markov chain transition probabilities using

$$r_x = \frac{e^{0.0292(x+10.67)-0.6228}}{1+e^{0.0292(x+10.67)-0.6228}}. \quad (6)$$

We note that this estimate of 10.67 points as the magnitude of home-court advantage is significantly different from the estimate used by Sagarin or by a simple least-squares fit where margin is estimated by the difference of the two teams' strength parameters plus the home-court advantage. Both of the other two methods estimate the home-court advantage to be close to 4 points.

3. Alternative Transition Probabilities

In this section, we present an alternative, empirical Bayes method for calculating transition probabilities. As we will show, the new method gives a similar fit to the logistic regression for s_x but estimates r_x differently. In Section 4, we show computational results that suggest this new method can be an improvement over r_x described in equation (6).

3.1. An Alternative Model for Road Success

We begin by assuming that the strengths (in points) of NCAA Division I college basketball teams are normally distributed with mean $\overline{M}$ and variance τ^2. Therefore, given any randomly-selected teams i and j, the inherent advantage of i over j is normally distributed with mean 0 and variance $2\tau^2$.

DOI: 10.2202/1559-0410.1202

In any matchup between two teams, there is also an inherent variation in score which we assume is independent of the teams in question. Breiter and Carlin (1997) estimated the Las Vegas betting lines and the ratings differences of Sagarin each differ from actual margins of victory by a mean of zero and a standard error of approximately $\sigma = 11$ points; nearly a decade later, Kvam and Sokol found similar estimates when they fit differences in teams' LRMC steady-state probabilities to point differences using 1999-2000 through 2004-2005 data.

Therefore, we estimate the margin (in points) x_{ij} between teams i (home) and j (road) to be a distributed according to $X_{ij} \sim N(m_{ij}+h,\sigma^2)$, where the m_{ij} are themselves distributed according to $M \sim N(0,2\tau^2)$.

Consequently, given an observation x between teams i and j, the conditional distribution of m_{ij} is

$$M \mid X=x \sim N\left(\frac{2\tau^2}{\sigma^2+2\tau^2}(x-h), \frac{2\sigma^2\tau^2}{\sigma^2+2\tau^2}\right) \qquad (7)$$

and the conditional distribution of the margin Y in a game between the same two teams at j's home court is

$$Y \mid X=x \sim N\left(\frac{2\tau^2}{\sigma^2+2\tau^2}(x-h)-h, \frac{2\sigma^2\tau^2}{\sigma^2+2\tau^2}+\sigma^2\right). \qquad (8)$$

The estimate of the probability that team i will beat team j on the road given an x-point margin at home is therefore

$$\Pr(Y>0 \mid X=x) = \Phi\left(\frac{x}{\sigma}\frac{2\tau^2}{\sqrt{(\sigma^2+2\tau^2)(\sigma^2+4\tau^2)}} - \frac{h}{\sigma}\sqrt{\frac{\sigma^2+4\tau^2}{\sigma^2+2\tau^2}}\right). \qquad (9)$$

In equation (9), all parameters can be estimated from other sources except for τ. Specifically, from Breiter and Carlin's work and Kvam and Sokol's work we can estimate $\sigma = 11$, and from Sagarin's work and simple least-squares fits we can estimate $h \approx 4$. (In Section 4, we discuss the robustness of our methods to variations in h; for the remainder of Section 3, we use the estimate h=4 in our numerical examples.) From the NCAA data and equation (5), both taken from Kvam and Sokol, we know that $\Pr(Y>0 \mid X=x) = \Phi(0) = 0.5$ when $x = 21.33$. In other words, the home team is equally likely to win or lose on the road against an opponent when they have beaten that team at home by about 21.33 points. Therefore, we can deduce from equation (9) that $\tau = 4.26$.

Finally, plugging back into equation (9) gives a simple estimate for the probability that team *i* will beat team *j* on the road, given that *i* beat *j* at home by *x* points:

$$\Pr(Y>0 \mid X=x) = \Phi(0.0189x - 0.4034). \qquad (10)$$

As we show in Table 1, our estimate in equation (10) is very similar to that derived by Kvam and Sokol using logistic regression.

x	$s_x = \frac{e^{0.0292x-0.6228}}{1+e^{0.0292x-0.6228}}$	$\Pr(Y>0 \mid X=x) = \Phi(0.0189x - 0.4034)$
-40	0.1430	0.1231
-35	0.1618	0.1435
-30	0.1826	0.1659
-25	0.2054	0.1905
-20	0.2303	0.2173
-15	0.2572	0.2461
-10	0.2860	0.2768
-5	0.3167	0.3093
0	0.3491	0.3433
5	0.3830	0.3787
10	0.4180	0.4151
15	0.4539	0.4523
20	0.4903	0.4899
25	0.5268	0.5275
30	0.5630	0.5650
35	0.5985	0.6018
40	0.6330	0.6378

Table 1. Comparison of s_x and $\Pr(Y>0 \mid X=x)$

3.2. Extension to Neutral-court Success

Given a model that gives both (i) accurate estimates of the probability that team *i* will beat team *j* on *j*'s home court given that *i* had a margin of *x* points over *j* on *i*'s home court, and (ii) an estimate of the magnitude of home-court advantage that is much more in line with other methods and common usage, we now use the model to predict success on a neutral court (including NCAA tournament games).

Kvam and Sokol estimate neutral-court success *r* relative to home-court success *s* as $r_x = s_{x+h} = \frac{e^{a(x+h)+b}}{1+e^{a(x+h)+b}}$. In this method, the effects of additional

DOI: 10.2202/1559-0410.1202

points scored and additional site advantage are exactly substitutable. For example, a win by $x+h$ points at a neutral court and a win by x points on a team's home court would lead to the same s and r. It is not necessary for this to be true, though; in fact, our method suggests that the two effects are not substitutable.

In our method, equation (7) gives the conditional distribution of the inherent difference between teams i and j given that team i beat team j by x points at home:

$$M \mid X=x \sim N\left(\frac{2\tau^2}{\sigma^2 + 2\tau^2}(x - h), \frac{2\sigma^2\tau^2}{\sigma^2 + 2\tau^2} \right) \quad . \tag{7}$$

Equation (8) is the conditional distribution of the margin of a matchup between the same two teams at j's home court:

$$Y \mid X=x \sim N\left(\frac{2\tau^2}{\sigma^2 + 2\tau^2}(x - h) - h, \frac{2\sigma^2\tau^2}{\sigma^2 + 2\tau^2} + \sigma^2 \right). \tag{8}$$

The only difference between equation (8) and the expression for the conditional distribution of the margin of a neutral-court matchup between i and j is the subtraction of h (for j's home court) in the mean. Removing that term gives

$$Z \mid X=x \sim N\left(\frac{2\tau^2}{\sigma^2 + 2\tau^2}(x - h), \frac{2\sigma^2\tau^2}{\sigma^2 + 2\tau^2} + \sigma^2 \right). \tag{11}$$

Note that the margin and the home-court advantage are not substitutable between equations (8) and (11).

From equation (11), we find the expression for the conditional probability of team i beating team j on a neutral court given i's x-point home margin over j. This probability $\Pr(Z>0 \mid X=x)$ is our equivalent of r_x:

$$\Pr(Z>0 \mid X=x) = \Phi\left(\frac{2\tau^2}{\sqrt{(\sigma^2 + 2\tau^2)(\sigma^2 + 4\tau^2)}} \left(\frac{x - h}{\sigma} \right) \right). \tag{12}$$

Alternatively, instead of the probability that team i would beat team j on a neutral court, we could also use to generate Markov chain transition probabilities $\Pr(M>0 \mid X=x)$, the probability that team i is simply better than team j, derived from equation (7):

$$\Pr(M>0 \mid X=x) = \Phi\left(\sqrt{\frac{2\tau^2}{\sigma^2+2\tau^2}}\left(\frac{x-h}{\sigma}\right)\right). \qquad (13)$$

$\Pr(Z>0 \mid X=x)$ and $\Pr(M>0 \mid X=x)$ were implicitly treated by Kvam and Sokol as if they were the same. However, our model shows how to treat the two differently; specifically, because of the extra variation inherent in the second game's results, $\Pr(Z>0 \mid X=x)$ will be less extreme than $\Pr(M>0 \mid X=x)$. This can be seen by looking at expressions (12) and (13); denoting a_i so that $\Pr(i>0 \mid X=x) = \Phi(a_i)$, $\frac{a_Z}{a_M} = \sqrt{\frac{2\tau^2}{\sigma^2+4\tau^2}} < 1$, so $\Pr(Z>0 \mid X=x)$ will tend more toward ½ than $\Pr(M>0 \mid X=x)$.

Table 2 shows the difference between the old r_x and the new $\Pr(Z>0 \mid X=x)$ and $\Pr(M>0 \mid X=x)$.

x	$r_x = \frac{e^{0.0292(x+10.67)-0.6228}}{1+e^{0.0292(x+10.67)-0.6228}}$	$\Pr(Z>0 \mid X=x) = \Phi(0.0189x-0.0756)$	$\Pr(M>0 \mid X=x) = \Phi(0.0437x-0.1747)$
-40	0.1855	0.2027	0.0273
-35	0.2086	0.2304	0.0443
-30	0.2338	0.2601	0.0688
-25	0.2609	0.2917	0.1027
-20	0.2900	0.3250	0.1473
-15	0.3210	0.3597	0.2034
-10	0.3536	0.3956	0.2705
-5	0.3876	0.4324	0.3472
0	0.4228	0.4699	0.4307
5	0.4588	0.5075	0.5174
10	0.4952	0.5452	0.6033
15	0.5316	0.5824	0.6845
20	0.5678	0.6189	0.7576
25	0.6032	0.6543	0.8204
30	0.6376	0.6885	0.8719
35	0.6706	0.7211	0.9121
40	0.7020	0.7520	0.9420

Table 2. Comparison of r_x, $\Pr(Z>0 \mid X=x)$, and $\Pr(M>0 \mid X=x)$

DOI: 10.2202/1559-0410.1202

3.3. Multi-game Joint Conditioning

During the course of an NCAA basketball season, some pairs of teams (usually in the same conference) play each other more than once. Within a conference, all or most pairs of teams might play each other twice, once on each team's home court; the original LRMC model took advantage of this fact when fitting the logistic regression curve for s_x. In some cases, pairs of teams even play each other three times, once on each team's home court during the regular season, and then on a neutral court (or one team's home court) as part of the conference tournament. It is even possible for teams to play each other four times before the NCAA tournament, if they also meet in one of the preseason invitational tournaments.

In the LRMC model, every time a pair of teams plays each other it is treated as a brand new event. The probability r_x that one team would beat another is found for each outcome independently; that is, given two margins x_1 and x_2 between a pair of teams, r_{x_1} is determined independently of x_2 and r_{x_2} is determined independently of x_1. In the empirical Bayes model of Sections 3.1 and 3.2, we determine $\Pr(Z>0 \mid X=x)$ and $\Pr(M>0 \mid X=x)$ similarly: $\Pr(Z>0 \mid X_1=x_1)$ and $\Pr(M>0 \mid X_1=x_1)$ are determined independently of x_2, and $\Pr(Z>0 \mid X_2=x_2)$ and $\Pr(M>0 \mid X_2=x_2)$ are determined independently of x_1.

However, this need not be the case. Instead, we can condition jointly on all outcomes between a pair of teams. If we denote the number of games teams two teams have played against each other to be G, then we can define $Z \mid X_1=x_1, X_2=x_2, \ldots, X_G=x_G$ and $M \mid X_1=x_1, X_2=x_2, \ldots, X_G=x_G$. Specifically,

$$M \mid X_1=x_1, X_2=x_2, \ldots, X_G=x_G \sim N\left(\frac{2\tau^2}{\sigma^2 + 2\tau^2 G} \sum_{g=1}^{G} \left(x_g - h_g\right), \frac{2\sigma^2\tau^2}{\sigma^2 + 2\tau^2 G} \right) \qquad (14)$$

and

$$Z \mid X_1=x_1, X_2=x_2, \ldots, X_G=x_G \sim N\left(\frac{2\tau^2}{\sigma^2 + 2\tau^2 G} \sum_{g=1}^{G} \left(x_g - h_g\right), \frac{2\sigma^2\tau^2}{\sigma^2 + 2\tau^2 G} + \sigma^2 \right), \qquad (15)$$

where h_g is the appropriate home-court adjustment for game g (either h for a home game, $-h$ for a road game, or 0 for a neutral-court game).

So, the jointly-conditioned probabilities that we can use for Markov chain transitions become

$$\Pr_{JOINT}(M>0 \mid X_1=x_1, X_2=x_2, \ldots, X_G=x_G) = \Phi\left(\sqrt{\frac{2\tau^2}{\sigma^2+2\tau^2 G}}\sum_{g=1}^{G}\left(\frac{x_g-h_g}{\sigma}\right)\right) \quad (16)$$

and

$$\Pr_{JOINT}(Z>0 \mid X_1=x_1, X_2=x_2, \ldots, X_G=x_G) = \Phi\left(\frac{2\tau^2}{\sqrt{\left(\sigma^2+2\tau^2 G\right)\left(\sigma^2+2\tau^2(G+1)\right)}}\sum_{g=1}^{G}\left(\frac{x_g-h_g}{\sigma}\right)\right). \quad (17)$$

In the original LRMC method, the estimates from each individual game between two opponents are implicitly averaged to get an overall estimate for the probability that one of those two teams is better than the other head-to-head. Essentially, if we were to adopt this method for our empirical Bayes probabilites, it would estimate

$$\Pr_{AVG}(M>0 \mid X_1=x_1, X_2=x_2, \ldots, X_G=x_G) = \frac{1}{G}\sum_{g=1}^{G}\Phi\left(\sqrt{\frac{2\tau^2}{\sigma^2+2\tau^2}}\left(\frac{x_g-h_g}{\sigma}\right)\right) \quad (18)$$

in place of equation (16).

Table 3 demonstrates the difference between using single-game estimates and the jointly-conditioned estimates using the Duke/North Carolina (UNC) rivalry as an example.

3.4. Creating Transition Probabilities

We use the empirical Bayes probabilities we generate in the same way as the original LRMC method, as transition probabilities in a Markov chain.

It is straightforward to incorporate our probabilities when each game is treated as a separate event. Recall that the original LRMC method creates a Markov chain transition matrix $T^R = [\, t_{ij}^R \,]$, where

$$t_{ij}^R = \frac{1}{N_i}\left[\sum_{g=(i,j)} r_{x(g)} + \sum_{g=(j,i)}\left(1-r_{x(g)}\right)\right], \qquad \text{for all } j \neq i, \quad (1)$$

$$t_{ii}^R = \frac{1}{N_i}\left[\sum_{g=(i,\cdot)}\left(1-r_{x(g)}\right) + \sum_{g=(\cdot,i)} r_{x(g)}\right]. \quad (2)$$

Year-by-year Duke/UNC Single-game outcomes				Estimated probability that Duke is better than UNC head-to-head	
Year	*Home team*	*Game outcome*	Pr(M>0 | X=x) for Duke	$\Pr_{AVG}(M>0 \mid X_1=x_1, X_2=x_2, \ldots, X_G=x_G)$	$\Pr_{JOINT}(Z>0 \mid X_1=x_1, X_2=x_2, \ldots, X_G=x_G)$
2000	UNC	Duke by 4	64%	65%	76%
	Duke	Duke by 14	67%		
2001	Duke	UNC by 2	40%	68%	92%
	UNC	Duke by 14	78%		
	Neutral	Duke by 26	87%		
2002	UNC	Duke by 29	93%	82%	99%
	Duke	Duke by 25	82%		
	Neutral	Duke by 12	70%		
2003	Duke	Duke by 9	59%	60%	74%
	UNC	UNC by 3	52%		
	Neutral	Duke by 12	70%		
2004	UNC	Duke by 2	60%	56%	61%
	Duke	Duke by 5	52%		
2005	Duke	Duke by 1	45%	49%	48%
	UNC	UNC by 2	53%		
2006	UNC	Duke by 4	64%	48%	45%
	Duke	UNC by 7	32%		
2007	Duke	UNC by 6	33%	33%	22%
	UNC	UNC by 14	33%		
2008	UNC	Duke by 11	74%	52%	55%
	Duke	UNC by 8	30%		

Table 3. Effect of jointly conditioning on all head-to-head outcomes

We can use our probability estimates conditioned on the results of each game g, $z_g = \Pr(Z>0 \mid X=x)$ and $m_g = \Pr(M>0 \mid X=x)$, in a similar way. Specifically, we can replace the logistic regression estimates $r_{x(g)}$ either with z_g to create a transition matrix T^Z defined by

$$t_{ij}^Z = \frac{1}{N_i}\left[\sum_{g=(i,j)} z_g + \sum_{g=(j,i)}\left(1-z_g\right)\right], \qquad \text{for all } j \neq i, \tag{19}$$

$$t_{ii}^Z = \frac{1}{N_i}\left[\sum_{g=(i,\cdot)}\left(1-z_g\right) + \sum_{g=(\cdot,i)} z_g\right], \tag{20}$$

or with m_g to create a transition matrix T^M defined by

$$t_{ij}^M = \frac{1}{N_i}\left[\sum_{g=(i,j)} m_g + \sum_{g=(j,i)}\left(1-m_g\right)\right], \qquad \text{for all } j \neq i, \tag{21}$$

$$t_{ii}^M = \frac{1}{N_i}\left[\sum_{g=(i,\cdot)}\left(1-m_g\right) + \sum_{g=(\cdot,i)} m_g\right]. \tag{22}$$

In each of these cases, every game is given equal weight. So, in the case where two teams playing $G > 1$ games against each other, this is equivalent to the transitions between those teams being $\Pr_{\text{AVG}}(M>0 \mid X_1=x_1, X_2=x_2, \ldots, X_G=x_G)$ weighted G times.

On the other hand, when using jointly-conditioned probabilities it is less obvious how they should be weighted. One option is to take the jointly-conditioned probabilities and weight them G times. We denote such a transition matrix as J_G, so that J_G^Z and J_G^M are the transition matrices obtained by weighting G times the jointly-conditioned probabilities of a neutral-court victory and of being the better team. A second option is to weight each jointly-conditioned probability just once; we denote these transition matrices as J_1^Z and J_1^M.

In a sense, J_G is an over-weighting because the variance of the G-weighted probabilities is greater than the variance of G independent games. For example, the variance of $M \mid X_1=x_1, X_2=x_2, \ldots, X_G=x_G$ is $\frac{2\sigma^2\tau^2}{\sigma^2+2\tau^2 G}$, so after weighting G times the overall variance will be $\frac{2\sigma^2\tau^2 G^2}{\sigma^2+2\tau^2 G}$, which is larger than the variance of G independent single-game estimates, $\frac{2\sigma^2\tau^2 G}{\sigma^2+2\tau^2} = \frac{2\sigma^2\tau^2 G^2}{\sigma^2 G+2\tau^2 G}$. In other words, G single-game estimates give more information than an estimate that is jointly-conditioned on G game outcomes and then weighted G times.

On the other hand, J_1 is an under-weighting, because the variance of a unit weighting of a G-game jointly-conditioned estimate, $\frac{2\sigma^2\tau^2}{\sigma^2+2\tau^2 G}$, is smaller than

DOI: 10.2202/1559-0410.1202

the variance of G independent single-game estimates, $\frac{2\sigma^2\tau^2 G}{\sigma^2+2\tau^2} = \frac{2\sigma^2\tau^2}{\sigma^2/G+2\tau^2/G}$.

Determining the "right" way of weighting games is unclear. As a heuristic, in addition to J_G and J_1, we test a "balanced" weighting J_B, where each jointly-conditioned estimate is weighted so that its overall variance will be exactly equal to the variance of G independent single-game estimates. For estimates of M, the balance factor is

$$B_M(G) = \sqrt{\frac{G(\sigma^2+2\tau^2 G)}{\sigma^2+2\tau^2}}, \tag{23}$$

and for estimates of Z, the balance factor is

$$B_Z(G) = \sqrt{\frac{G(\sigma^2+4\tau^2)(\sigma^2+2\tau^2 G)}{(\sigma^2+2\tau^2)(\sigma^2+2\tau^2(G+1))}}. \tag{24}$$

Although we do not have mathematical justification for these balance factors being the correct ones to use, we show in Section 4 that empirically they outperform J_G and J_1.

3.5. Combining Head-to-Head Estimates

Given a set of head-to-head estimates of relative team quality, the LRMC method uses a Markov chain to combine these often-contradictory estimates into one overall ranking. The Markov chain is an indirect way of considering each of the head-to-head estimates together; a more direct method is to use a least-squares regression. Given any vector of estimated head-to-head strength differences d between teams, the regression fits team strength estimates α for each team. Specifically, given a matrix B such that the row for a comparison between teams i and j has a "1" in column i, a "−1" in column j, and zeros everywhere else, the model estimates α as

$$\hat{\alpha} = \left(B^{\mathrm{T}}B\right)^{-1}B^{\mathrm{T}}d\,. \tag{25}$$

4. Computational Results

In order to compare these various NCAA tournament prediction methods with those of the original LRMC, we compare their performance in ten years of tournament games, from the 1999-2000 season through the 2008-2009 season. With 63 tournament games played each year, this gives a set of 630 games in our test set. All outcome data were downloaded from Yahoo (1999-2009).

We ran a large number of tests, and found that for all of our new methods, and for a wide range of values of home court advantage h, the Markov chain outperformed the least squares method for combining head-to-head estimates. In fact, for almost all methods, the h that performed worst with the Markov chain was better than the best h for least squares. Moreover, the least squares methods were all worse than the original LRMC. Therefore, in the next section we discuss only the Markov chain results.

4.1. Home Court Effects

First, we consider the "correct" value of h. We tested our new LRMC methods for all integer values of h on the interval [–11,11]. Figure 1 shows the average number of games correctly predicted by our method (out of 630) for each h, with the original LRMC method shown as a reference. (Because the original LRMC does not use the parameter h but instead deduces its own h-value from past data, it appears as a straight line in Figure 1.) We believe that the results contain some noise (hence the local non-monotonicity in the results) and that results are not independent, so we first examine the average performance of the eight new methods rather than looking at each individual method.

Figure 1 shows a clear performance peak when the home court advantage is between 1 and 4, with slightly better performance at h=3 and h=4. This seems to suggest that the choice of h is not so important, as long as it is in the interval [1,4]. Figure 2 shows a finer-grained breakdown at intervals of 0.1 between h=0 and h=5. There is a slightly better performance from h=2 to h=4, with a very slight maximum around h=3.5.

We note that, as the amount of noise in Figures 1 and especially 2 suggests, the individual methods exhibit slightly-different qualitative behavior from each other. Among the Z-based methods (based on the probability of winning at a neutral site), the jointly conditioned, heuristically balanced method J_B^Z performs best with a peak at h=3.5, while the jointly conditioned, unit weighted method J_1^Z peaks at h=2.5. Among M-based methods (based on the probability of being a better team), the jointly conditioned, heuristically balanced method J_B^M again performs best, with a peak from h=3.5 to h=4.0, while the jointly conditioned,

DOI: 10.2202/1559-0410.1202

unit weighted method J_1^M peaks from h = 0.0 to h=0.2. (We suspect this peak is due to random noise.)

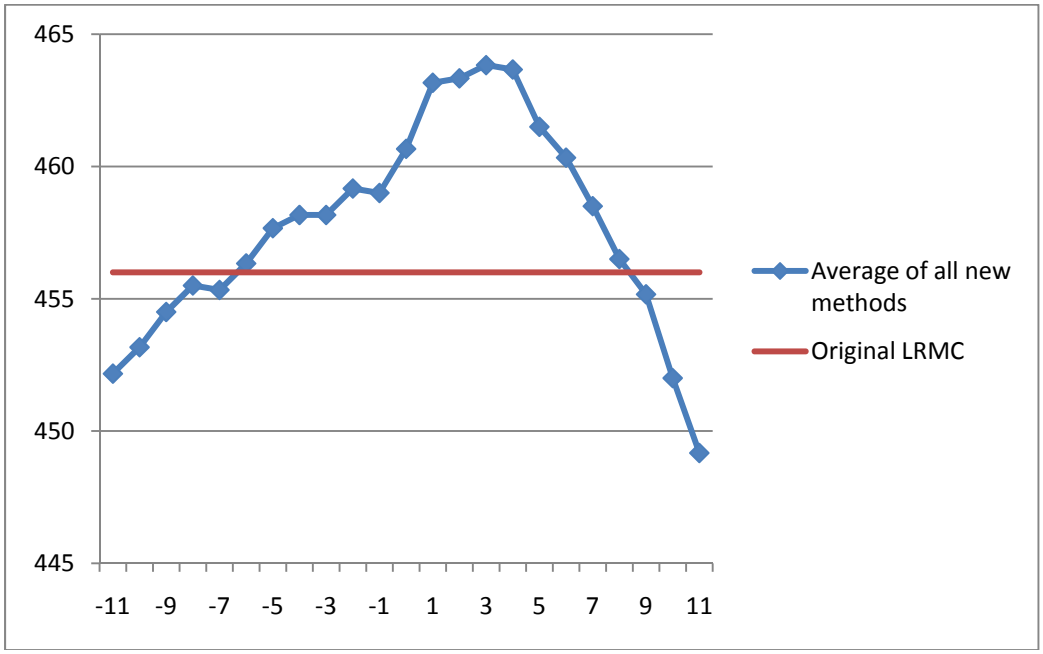

Figure 1. Average performance of new LRMC methods as a function of h, as compared to original LRMC. Performance is measured as number of correct predictions out of 630 games.

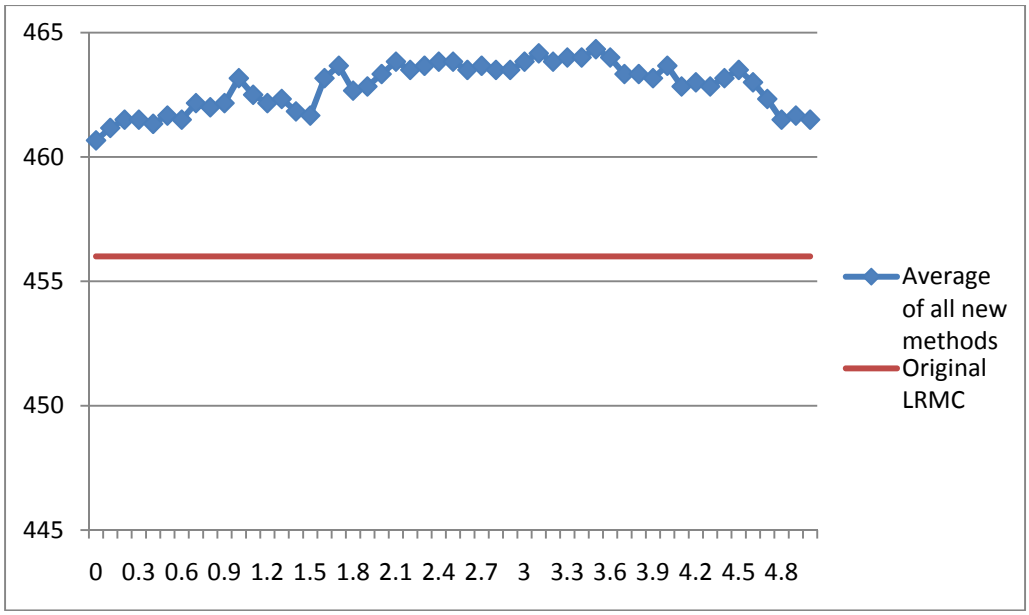

Figure 2. Average performance of new LRMC methods as a function of h, as compared to original LRMC. Performance is measured as number of correct predictions out of 630 games.

4.2. Comparison to Original LRMC

Our initial goal in this research was to determine whether any of our ideas would result in an improvement over the original LRMC method. Figures 1 and 2 would seem to imply that the answer is yes, since (for reasonable choices of h) even the average of the eight new methods predicts more games correctly than the original LRMC. However, it is not enough to simply look at the number of games whose winners are correctly predicted by each method. This is because there are many games in which any reasonable prediction method would pick the same winner; for example, in the tournament's first round there are four games that match up one of the four participants the selection committee ranked as best against one of the four participants the selection committee ranked as worst. None of the four worst teams has ever won one of these matchups, so every prediction method should make the same (obvious) prediction for these games.

Therefore, to make comparisons between methods we look only at games in which the two methods being compared disagreed in their prediction; we refer to these games as the *disagreement set* between two methods. We then use a one-tailed McNemar's test to measure the significance level of one method's superiority over another.

Table 4 shows these head-to-head results using h=3.5. We test each of the three singly-conditioned methods (using r_x, Pr(Z>0 | X=x), and Pr(M>0 | X=x)) to derive transition probabilities T^R, T^Z, and T^M, as well as six jointly-conditioned methods (using transition matrices J_G^Z, J_B^Z, J_1^Z, J_G^M, J_B^M, and J_1^M). For each, we display test results based on both a Markov chain (MC) and a least-squares regression (LS).

Each cell of the table shows the comparison between the tested algorithm and the original LRMC method. The cell contains the number of games in the disagreement set for which the test method was correct, followed by the number of games in the disagreement set for which the original method was correct. The second line of each cell is the McNemar's test p-value for the test method's superiority. For example, the original LRMC method and our singly-conditioned empirical Bayes estimate of Pr(Z>0 | X=x) (the probability of winning at a neutral court) with a Markov chain disagreed on the predicted winner of 47 out of the 630 tournament games. The original LRMC method was correct 20 times and our new method was correct 27 times, and McNemar's test returns a p-value of 0.19 indicating that the difference between the two methods is not statistically significant.

DOI: 10.2202/1559-0410.1202

	Singly-conditioned probabilities			Jointly-conditioned probabilities					
	Original LR	Pr(win @ neutral)	Pr(better team)	Pr(win @ neutral)			Pr(better team)		
	T^R	T^Z	T^M	J_G^Z	J_B^Z	J_1^Z	J_G^M	J_B^M	J_1^M
Original MC	---	27-20 .19	23-19 .32	24-19 .27	25-13 .04	21-14 .16	22-14 .12	20-10 .05	18-10 .09
Least-squares	63-79 .92	58-64 .74	63-78 .91	78-155 1.00	80-156 1.00	84-158 1.00	78-140 1.00	78-140 1.00	80-140 1.00

Table 4. Differences between the original LRMC and the eight new algorithms tested, when h=3.5. Each cell shows the results in the disagreement set: correct predictions by the test algorithm, correct predictions by the original method, and McNemar's test p-value for the test method's superiority.

	Singly-conditioned probabilities			Jointly-conditioned probabilities					
	Original LR	Pr(win @ neutral)	Pr(better team)	Pr(win @ neutral)			Pr(better team)		
	T^R	T^Z	T^M	J_G^Z	J_B^Z	J_1^Z	J_G^M	J_B^M	J_1^M
Original MC	---	26-18 .15	23-20 .38	25-20 .28	25-15 .08	22-13 .09	22-17 .26	20-10 .05	18-9 .06
Least-squares	63-79 .92	58-63 .71	63-78 .91	79-155 1.00	80-152 1.00	83-154 1.00	78-140 1.00	78-138 1.00	82-138 1.00

Table 5. Differences between the original LRMC and the eight new algorithms tested, where h varies from season to season based on a least squares estimate. Each cell shows the results in the disagreement set: correct predictions by the test algorithm, correct predictions by the original method, and McNemar's test p-value for the test method's superiority.

From Table 4, several results are clear. First, as we noted earlier, the use of the Markov chain is clearly superior to replacing it with a least-squares estimate, a conclusion which gets even more pronounced for jointly-conditioned head-to-head probability estimates. Second, both of the empirical Bayes estimates for head-to-head probabilities improve over the original logistic regression. Finally, when jointly conditioning, using the heuristic balance factor is important; it provides the final step that allows our new methods to claim statistically significant improvements over the original LRMC method. Both jointly conditioned, heuristically balanced methods are significantly better than original LRMC at the 0.05 level.

One additional observation we can take away from Table 4 is that basing the empirical Bayes step on the probability of being a better team, rather than the probability of winning at a neutral site, leads both to fewer disagreements with the original LRMC method, and generally to better head-to-head results against original LRMC.

4.3. Varying Home Court Effect By Season

In our experiments so far, we have assumed that the effect h of playing on a home court is constant from season to season. However, that might not be true. For example, Sagarin calculates a home court value each season and finds it to be slightly different one season to the next. Rather than use home court values from Sagarin, we prefer that our methods be "stand-alone", so we computed simple least-squares estimates for the home court value for each season. We then tested our new LRMC methods using those estimates. (We also did test our methods using values from Sagarin, and found no statistically significant differences between using those home court values and the least squares values.)

The results in Table 5 are very similar to those in Table 4, suggesting that our methods are somewhat robust to possible changes in home court advantage from season to season. Table 6 shows the least squares home court values for each of the ten seasons we tested. (We note that it is still unclear how much of the slight changes observed in Table 6 is real, and how much is simply noise.)

In addition to individual-game results, we studied the quality of predictions of the overall tournament winner. There was almost no difference between our new methods and the original LRMC method, or between pairs of our new methods; almost all predicted four winners out of ten correctly (in 2001, 2005, 2008, and 2009), ranked eight of the ten tournament winners in the top 5 (missing only in 2003 and 2006), and ranked four of the last five tournament winners first or second overall (all except 2006).

DOI: 10.2202/1559-0410.1202

Season	Home court estimate	
	Sagarin	Least squares
1999-2000	4.23	3.92
2000-2001	4.39	4.16
2001-2002	4.31	4.50
2002-2003	4.11	4.14
2003-2004	4.02	3.99
2004-2005	4.10	4.02
2005-2006	4.04	3.63
2006-2007	4.20	3.86
2007-2008	4.13	3.94
2008-2009	3.81	3.38

Table 6. Estimates of home court advantage

4.4. Year-to-Year Performance

Coleman and Lynch compared the original LRMC method's performance to RPI, Sagarin, and the methods of West and Coleman and Lynch in predicting the 2009 NCAA Tournament. LRMC actually was slightly worse than the other methods that year, predicting 45 games correctly compared to 46-48 for the other methods. Coleman and Lynch also note that for a 10-year period, the RPI correctly predicted 69.6% of NCAA Tournament games, Sagarin correctly predicted 70.8%, and the two models in West correctly predicted 73.2% and 73.8%.

In Table 7, we extend the one-year experiment of Coleman and Lynch to include ten years of data (nine of which are overlapping). We do not have year-by-year data for West or Coleman and Lynch, but we can compare with the overall accuracy that Coleman and Lynch report for their models. We note that the original LRMC results reported here are slightly different from those reported by Kvam and Sokol because the Yahoo data set has become much cleaner in the interim.

As Table 7 shows, the best predictor can vary from year to year, as does the absolute and relative performance of each method. However, the overall fraction of games predicted correctly by RPI and Sagarin is (as expected) about the same in our 10-year sample and that of Coleman and Lynch, so it is not unreasonable to assume that their models' performance (73.2% and 73.8%) would be about the same on this data set. The performance of our best improvement, the jointly conditioned, heuristically balanced Bayesian LRMC, is only slightly better, at 74.0% and 74.3%. We suspect that this difference is not statistically significant, though without knowing the disagreement sets, it is impossible to test.

	Popular rankings		Singly-conditioned probabilities			Jointly-conditioned probabilities					
			Orig-inal LR	Pr(win at neut-ral)	Pr(better team)	Pr(win @ neutral)			Pr(better team)		
	RPI	Saga-rin	T^R	T^Z	T^M	J_G^Z	J_B^Z	J_1^Z	J_G^M	J_B^M	J_1^M
2000	41	44	44	46	46	45	48	48	46	45	45
2001	43	43	45	45	45	49	46	45	49	48	47
2002	41	45	48	45	45	46	47	44	47	47	46
2003	44	42	44	43	43	44	44	42	44	43	43
2004	46	45	41	47	45	43	44	44	43	44	43
2005	46	45	47	47	47	45	45	45	46	45	45
2006	40	42	43	44	44	43	45	46	43	46	47
2007	46	51	48	49	49	50	51	52	50	50	50
2008	46	46	51	54	51	52	52	51	52	53	53
2009	46	48	45	44	45	44	46	46	44	45	45
Total	439	451	456	463	460	461	468	463	464	466	464
% of 630	69.7	71.6	72.4	73.5	73.0	73.2	74.3	73.5	73.7	74.0	73.7

Table 7. Year-by-year results for LRMC methods (using h=3.5) and popular rankings.

We note in passing that Las Vegas betting favorites have won about 73.6% of games over these same two years; that number hides a significant difference between 2000-2004 (70.9%) and 2005-2009 (76.4%). [Note that these values could be off in the last digit, because for games that are close to 50/50 picks, some Las Vegas betting houses could have a team as a slight favorite while others might call the game a pick-'em (no favorite, and thus omitted from our sample); in addition, the betting lines change slightly over time in response to the number of bets placed on each team.]

While we don't know what caused the difference between 2000-2004 and 2005-2009, we do believe that a static system such as ours and those of West, Coleman and Lynch, Sagarin, etc. that predicts winners before any games are played is unlikely to be able to outperform the dynamic Las Vegas lines that take updated information into account. Therefore, it seems reasonable that 76.4% might be an approximate upper bound for the performance of any static models.

DOI: 10.2202/1559-0410.1202

References

Breiter, D. and B. Carlin (1997). How to Play Office Pools If You Must. *Chance* **10** (1), pp. 5-11.

Callaghan, T., P.J. Mucha, and M.A. Porter (2004). The Bowl Championship Series: A Mathematical Review. *Notices of the American Mathematical Society* **51**, pp. 887-893.

Coleman, B.J. and A.K. Lynch (2009). NCAA Tournament Games: The Real Nitty-Gritty. *Journal of Quantitative Analysis in Sports* **5** (3), Article 8.

Kvam, P. and J.S. Sokol (2006). A Logistic Regression/Markov Chain Model for NCAA Basketball. *Naval Research Logistics* **53**, pp. 778-803.

Massey, K. (2000) Description of the Massey rating system. http://www.masseyratings.com/theory/massey.htm

Sagarin, J. (2000-2009). Jeff Sagarin Computer Rankings. Updated weekly; end-of-season rankings archived. http://www.usatoday.com/sports/sagarin.htm

West, B.T. (2006a). A Simple and Flexible Rating Method for Predicting Success in the NCAA Basketball Tournament. *Journal of Quantitative Analysis in Sports* **2** (3), Article 3.

West, B.T. (2006b). A Simple and Flexible Rating Method for Predicting Success in the NCAA Basketball Tournament: Updated Results from 2007. *Journal of Quantitative Analysis in Sports* **4** (2), Article 8.

Yahoo (1999-2009) Yahoo! Sports NCAA Men's Basketball Scores & Schedule. (1999-2009). Updated daily. http://sports.yahoo.com/ncaab/scoreboard