A Markov Model for Basketball

Kenny Shirley, Ph.D.
Applied Statistics Center
COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

THE IDEA

I model a basketball game as a sequence of transitions between discrete states. Specifically, the model is a Markov chain, which specifies that the probability distribution of the next state depends only on the present state. A good Markov model for basketball must, in my opinion, strike a compromise between being, on one hand, very detailed and complex, so as to capture all of the relevant (and sometimes rare) events that can occur during a game, and on the other hand, simple enough to fit and interpret, so that interesting strategic questions can be answered. A minimum requirement for the complexity of the Markov model is that the exact number of points scored by each team is determined by the transition count (i.e. the same transition cannot lead to different numbers of points scored at different times).

THE DEFINITION OF THE STATES

The states of the Markov chain are defined in terms of three factors:

1. Which team has possession (2): Home or Away
2. How that team gained possession (5): Inbound pass, Steal, Offensive Rebound, Defensive Rebound, Free Throws.
3. The number of points that were scored on the previous possession (4): 0, 1, 2, or 3.

The largest possible model would have 2 x 5 x 4 = 40 states, but since certain combinations of the 3 factors are impossible, the largest model (Figure 1a), has 30 states. Making certain assumptions about the course of action in a basketball game can further reduce the number of states. If one assumes, for example, that rare events like 4-point plays or loose ball fouls following missed free throws are impossible, then certain states can be eliminated without seriously affecting the usefulness of the model. The notation is relatively simple: Ai(2), for example, means that Team A gained possession via an inbound pass after 2 points were scored.

THE GOALS OF FITTING THE MODEL

If a Markov model fits the data well, then it can provide a very detailed "microsimulation" of a basketball game. Quantities of interest can be computed via simulation. Some examples of these might be (1) In-game win probabilities for a given team, (2) The expected number of points scored in a possession gained in different ways, such as offensive rebounds vs. defensive rebounds, and (3) The change in win probability as a function of the number of possessions in a game; i.e. how useful a strategy is it to "slow down the game?"

PREVIOUS WORK

Hal Stern (1994, JASA Vol. 89, p. 1128-1134) developed a Brownian motion model for the progress of sports scores that fits well for basketball, yielding good in-game win probability estimates. Two specific drawbacks to the Brownian motion model that Stern mentions are (1) The relative strengths of two teams playing are not included in the model, and (2) Which team has possession is not included in the model. The Markov model certainly incorporates the second piece of information, and can be fit to include the first as well. (Stern's model can be extended to incorporate team strengths as well – although it hasn't actually been done to my knowledge.)

PILOT STUDY

An 18-state model was fit using season-long summary data from the 2003-2004 NBA season. States that corresponded to rare events were eliminated (so as to reduce the model to 18 states), and the remaining transition probabilities were estimated using statistics like 2-pt FG %, 3-pt FG%, rebound % (off. and def.), steals and turnovers, for each team and its opponents. Win probabilities for each team, in a game vs. their average opponent, were estimated by simulating 1000 games per team. These win probabilities are very close to the actual winning percentages (Figure 2), suggesting that the Markov model does a good job of capturing the essence of play. That is, given estimates of transition probabilities and of the number of transitions in a game, the Markov model simulates realistic results.

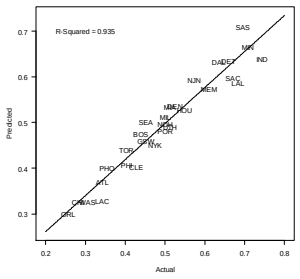

Figure 2: A plot of actual winning % vs. predicted winning %, where predictions were made using season-long summary statistics and simulations. On the scale of wins, the average error was about 3.3 wins out of 82 games.

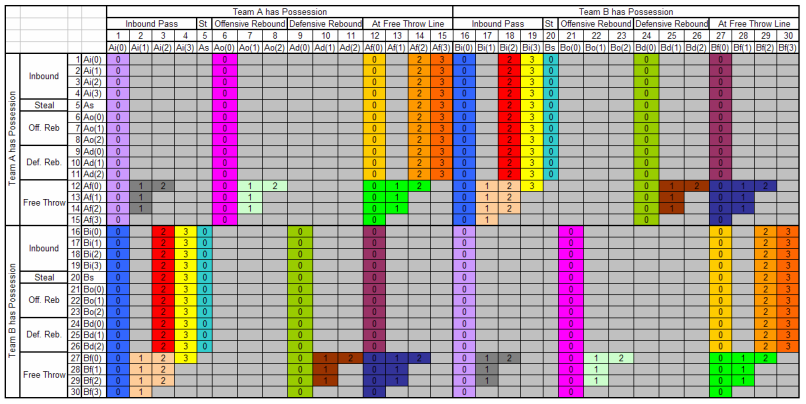

Figure 1a: Points Scored for each Transition, where Team A is the home team; gray boxes are transitions with zero probability.

Figure 1b: Color key for Transitions

THE FIT OF THE MODEL

The model was fit to a very small sample of 18 quarters (4.5 games) of NBA basketball from the 2006-2007 season. There were 1162 transitions recorded in this sample, yielding an estimate of about 260 transitions per game. Figure 3 contains Bayesian estimates of the transition probabilities, in which states that shared the same method of gaining possession of the ball were constrained to have the same transition probabilities. With more data, this constraint can be lifted, but with such a small sample, I think the benefits outweigh the costs.

The expected number of points scored from each state is calculated and displayed in the rightmost column of Figure 3. For the home team (Team A), offensive rebounds are the best way to gain possession, followed by steals, defensive rebounds, and finally the inbound pass (free throws aren't as interesting to analyze here). For the away team (Team B), surprisingly, defensive rebounds produce the most points on average. I strongly suspect this is an artifact of the small sample and that with more data, the expected points vector for the away team would look much like that for the home team, except slightly less.

To estimate the total number of points scored by each team in a full game using the Markov model, I just calculate the stationary distribution of the transition matrix, which yields the long-run probabilities of being in each state, and multiply this by the expected points vector, and then multiply the product by the estimated number of transitions in a game, which is about 260. This yielded an expected score of 96.8 – 91.4 in favor of the home team, which is roughly consistent with other estimates of total points and home court advantage.

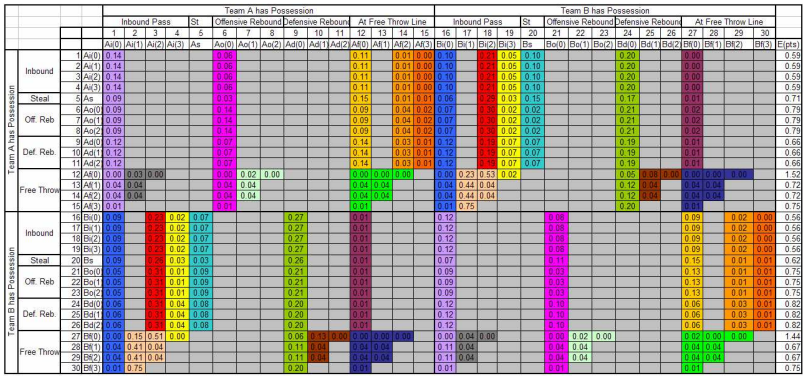

Figure 3: Estimated transition probabilities for 30-state model without incorporating team-specific variables. A Bayesian model with a flat Dirichlet prior was used to estimate each row of the transition matrix as a multinomial distribution. Also, data was pooled for estimating transition probabilities for rows in which a team acquired possession the same way. The last column contains the expected number of points scored in the next transition, given the current state.

WIN PROBABILITY AS A FUNCTION OF THE NUMBER OF TRANSITIONS AND WHICH TEAM HAS POSSESSION

How does the number of transitions in the game affect the probability of the home team winning? Since the home team is the favorite, the more transitions that occur, the higher should be the probability the home team wins. Interestingly, Figure 4a (below, left) shows that this probability is almost constant for the entire range of realistic numbers of transitions – the (average) home team always has about a 61-65% chance of winning! This result seems to suggest that the randomness inherent in each possession swamps the difference in win probabilities for a wide range of transition counts - a somewhat surprising find that needs closer inspection. If the model proves to be a good fit, then this result means that there is no use in "slowing down" or "speeding up" the game in order to gain a strategic advantage – just make more shots!

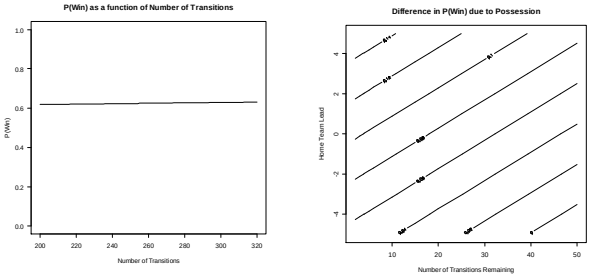

Figure 4b (above, right) attempts to answer the question of how important it is to know which team currently has possession in order to estimate the win probability of the home team, as a function of the number of transitions left in the game (which is never known, but possibly can be estimated), and the home team's lead. The figure shows the difference in win probabilities for the home team for two starting points: Ai(0) and Bi(0). Not surprisingly, as the number of transitions remaining increases, the current possession of the ball is less important – in the lower right hand corner of the plot, the difference is about zero. But as the number of transitions left decreases, the current possession of the ball becomes more important, until for a very small number of possessions left, and a larger lead for the home team, there is a great difference between win probabilities for the situations Ai(0) and Bi(0). This confirms that the Brownian motion model misses some potentially important information near the end of a game, because it doesn't account for which team currently has possession.

HOW DO WE COLLECT DATA?

The sequence of plays below occurred during the 1st quarter of the 4/6/07 Cleveland @ Washington game. CLE, the visitor, is coded as Team B in the Markov model. Notice the highlighted event: The standard play-by-play failed to record a WAS deflection following a CLE steal sometime between 2:55 and 2:39, whereas this event is represented by a transition in the Markov model, from state Bs to state Bi(0).

ESPN play-by-play				Markov Model	
Time	Event (CLE)	Score	Event (WAS)	State	Event
3:24		14-19	Etan Thomas makes 2-foot hook shot (Antonio Daniels assists)	Bi(2)	CLE inbound (WAS made 2pt FG)
3:08	Etan Thomas blocks LeBron James's 6-foot jumper	14-19			
3:05		14-19	Antonio Daniels defensive rebound	As	WAS steal (block)
3:02	Anderson Varejao personal foul (Antawn Jamison draws the foul)	14-19		Ai(0)	WAS inbound (CLE non-shooting foul)
3:02	Eric Snow enters the game for Sasha Pavlovic	14-19			
3:02	Donyell Marshall enters the game for Drew Gooden	14-19			
2:55		14-19	Antawn Jamison bad pass (Donyell Marshall steals)	Bs	CLE steal
(??:??)				Bi(0)	CLE inbound (WAS deflection)
2:39	Larry Hughes misses 17-foot jumper	14-19			
2:36		14-19	Jarvis Hayes defensive rebound	Ad(0)	WAS def. rebound (CLE miss 2pt FG)
2:29		14-21	DeShawn Stevenson makes 16-foot jumper (Antawn Jamison assists)	Bi(2)	CLE inbound (WAS made 2pt FG)
2:10	LeBron James makes two point shot (Eric Snow assists)	16-21			
2:10		16-21	Etan Thomas shooting foul (LeBron James draws the foul)	Bf(2)	CLE FT (CLE made 2pt FG, WAS shooting foul)
2:10		16-21	Roger Mason enters the game for DeShawn Stevenson		
2:10	LeBron James makes free throw 1 of 1	17-21		Ai(1)	WAS inbound (CLE made FT)

WHERE DO WE GO FROM HERE?

If we model rows of the transition matrix using multinomial logit models, then we can incorporate effects for the individual teams into the transition probabilities, resulting in one "baseline" transition matrix, and then a unique transition matrix for every matchup between teams. This model would most likely incorporate 4 x 10 x 3 x 2 =240 parameters. In a basketball season, there are about 320,000 transitions total. With about half a season of data, I think fitting the large model with team strength parameters would be possible. For the time being, we need more data! THANKS!

Email: ks2632@columbia.edu or kshirley@wharton.upenn.edu