A Markov Model for Basketball

Kenny Shirley, Ph.D.
Applied Statistics Center

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, in my opinion, strikes 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 sharing 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 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, Four Threes
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 1c) has 30 states. Making various assumptions about the course of action in a basketball game can further reduce the number of states. For instance, for rare events like 6- or 8-point plays or four-all, following missed free throws are impossible, then certain states can be eliminated without seriously affecting the usefulness of the model. The transition is relatively simple, although, for mathematicians, the model 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 in different ways, such as offensive rebounds vs. defensive rebounds, (3) The change in win probability as a function of the number of possessions in a game, i.e. how useful is my strategy in "showing 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 fit 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-point FG %, 3-point FG %, rebound % (off, 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 win-loss 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.

HOW DO WE COLLECT DATA?
The sequence of plays below occurred during the 1st quarter of the 2007 Harlem vs. 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, whereas this event is represented by a transition in the Markov model, from state Bi(3) to state Bi(0). The number of transitions in a game, the Markov model simulates realistic results.

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 balanced.

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.

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 higher the number of transitions, 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 “showing down” or “gossiping 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 not shown, 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: 46 and 99. Not surprisingly, 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, and for a very small number of possessions left, and a large lead for the home team, there is a great difference between win probabilities for the situations (46) and (99). This confirms that the Brownian motion model misses an important piece of information near the end of a game, because it doesn’t account for which team currently has possession.

THE NEXT STEP
If we model rows of the transition matrix using multivariate logit models, then we can incorporate effects for the individual teams into the transition probabilities, building on the “baseline” transition matrix, and fit a unique transition matrix for every transition of each team. This model would most likely incorporate 4 x 10 x 5 x 2 = 400 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!