Learning Stochastic Models for Basketball Substitutions from Play-by-Play Data

Harish S. Bhat, Li-Hsuan Huang, Sebastian Rodriguez

Applied Mathematics Unit
University of California, Merced
USA

Sept. 11, 2015

Basic Question

How should we model a basketball game between two teams?

- Suppose we seek a generative model that can be used to simulate.
- Of course we are interested in which team will win, but...
- We also want the model to generate a plausible game trajectory.

Problem

Substitutions and Defense

Traditional predictive models do not account for substitutions and focus mostly on offense.

What We Do

In our model, we...

1. Build dynamic, stochastic model of 5-man unit substitution.
2. Build model for average plus/minus rate of each 5-man unit.

Putting two elements together, we simulate separate game trajectory for home and visiting team. Whoever has higher final score wins.

Description of Data

Sample play-by-play data: Orl at Phl on 11/5/2014

Qtr	Time	Team	Event	Orl	Phi
1	10:46	Orl	Evan Fournier misses a 3-point jump shot from 26 feet out.	2	2
1	10:44	Orl	Nikola Vucevic with an offensive rebound.	2	2
1	10:41	Orl	Nikola Vucevic makes a putback layup from 1 foot out.	4	2
1	10:32	Phi	Brandon Davies makes a jump hook from 7 feet out. Tony Wroten with the assist.	4	4
1	10:17	Orl	Nikola Vucevic makes a hook shot from 1 foot out.	6	4
1	9:58	Phi	Nerlens Noel makes a jump shot from 17 feet out. Tony Wroten with the assist.	6	6
1	9:42	Orl	Channing Frye misses a 3-point jump shot from 25 feet out.	6	6
1	9:39	Orl	Tobias Harris with an offensive rebound.	6	6
1	9:29	Orl	Tony Wroten steals the ball from Channing Frye.	6	6
1	9:23	Phi	Tony Wroten makes a driving layup from 1 foot out.	6	8
1	9:17	Orl	Elfrid Payton makes a driving layup from 1 foot out.	8	8
1	9:04	Phi	Hollis Thompson makes a 3-point jump shot from 25 feet out. Luc Richard Mbah a Moute with the assist.	8	11
1	8:53	Orl	Nikola Vucevic misses a jump shot from 13 feet out.	8	11
1	8:51	Phi	Hollis Thompson with a defensive rebound.	8	11
1	8:39	Phi	Substitution: Henry Sims in for Nerlens Noel.	8	11
1	8:30	Phi	Henry Sims misses a jump shot from 20 feet out.	8	11
1	8:25	Orl	Magic with a defensive rebound.	8	11
1	8:23	Phi	Loose Ball foul committed by Henry Sims.	8	11
1	8:17	Orl	Henry Sims steals the ball from Elfrid Payton.	8	11
1	8:12	Phi	Elfrid Payton steals the ball from Tony Wroten.	8	11

Description of Data

Sources

- Grabbed play-by-play data for all 1230 regular-season NBA games from 2014-15. (Scraped from knbr.stats.com.)
- Also needed to verify lineup of players on court at beginning of each quarter. (Obtained from basketball-reference.com.)
- Parsed HTML data to produce one .csv file with 37203 rows, 20 columns.

Description of Data

After processing...

1 (Date)	2 (Home Team)	3 (Visiting Team)	4 (Home Player 1)	5 (Home Player 2)	6 (Home Player 3)	7 (Home Player 4)	8 (Home Player 5)	9 (Visiting Player 1)	10 (Visiting Player 2)	11 (Visiting Player 3)	12 (Visiting Player 4)	13 (Visiting Player 5)	14 (Seconds Played)	15 (Home Events)	16 (Visiting Events)	17 (Total Events)	18 (Home Score)	19 (Visiting Score)	20 (Δ_i)
20150127	Mia	Mil	478	479	480	487	481	57	426	425	431	427	350	13	21	34	15	17	-2
20150127	Mia	Mil	479	480	487	481	484	57	426	425	431	427	149	8	27	14	20	22	0
20150127	Mia	Mil	480	487	484	485	478	57	426	425	431	427	124	7	32	12	22	24	0
20150127	Mia	Mil	487	484	485	478	185	57	425	427	430	429	97	14	6	13	29	30	1
20150127	Mia	Mil	478	484	485	185	483	425	429	430	428	432	73	4	4	8	29	30	0

What is Δ_i?

Change in point differential (plus/minus):

- Let us consider just one team, either home (H) or visiting (V).
- When a 5-man unit takes the court, we record the score $S_{i-1} = H_{i-1} - V_{i-1}$.
- When a substitution is made, the 5-man unit changes. We record the new score $S_i = H_i - V_i$ and then calculate the change $\Delta_i = S_i - S_{i-1}$.
- Δ_i is a simple way to account for defense.
- Note that we also record the time the 5-man unit played on the court during the period corresponding to Δ_i.

Continuous-time Markov chain (CTMC) model

We build one CTMC model for each team. Consider one team for now.

Simulation perspective:

- Each 5-man unit is a state. Let N = total number of units.
- CTMC is specified by an $N \times N$ transition rate matrix M.
- To simulate this team's trajectory in one game, starting in state i at time $t = 0$, loop as follows:
  1. For each $j \neq i$, sample exponential RV with parameter M_{ij}.
  2. Think of each exponential RV as an "alarm clock."
  3. Go to state corresponding to alarm clock that rings first. Advance t by time elapsed before alarm clock rings. Set i equal to the new state.
  4. Stop if the total elapsed time ≥ 48 minutes. Else go to step 1.

Continuous-time Markov chain (CTMC) model

We build one CTMC model for each team. Consider one team for now.

Inference:

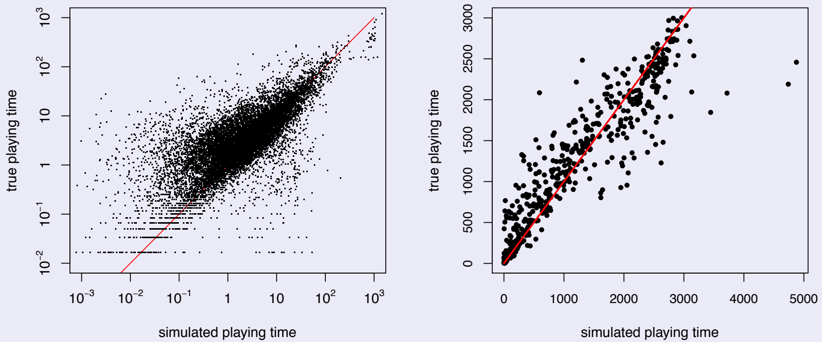
- Think of each game as a completely observed sample path of the CTMC.
- Then we have MLE (maximum likelihood estimator):

$$\widehat{M}^{j,k} = \frac{\#(j \to k)}{\alpha(j)}.$$

- $\#(j \to k)$ is the number of times we observe the transition from state j to state k.
- $\alpha(j)$ is the total time spent in state j.

True and simulated playing time, across all teams

5-man units (left) and individual players (right):



- Red line is $y = x$.
- Correlations are 0.834 (left) and 0.915 (right).
- Plenty of room for improvement!

Plus/minus rate model, i.e., what do we do with Δ_i?

Basic idea

- We can already use the CTMC to simulate dynamic presence of 5-man units on court.
- What we need: way to determine how much each 5-man unit contributes during its time on the court.
- We call this the "scoring rate" model, but it's actually an "average plus/minus rate" model.

Plus/minus rate model, i.e., what do we do with Δ_i?

Again, assume now we are working on a particular team's model.

Average vector $\vec{\beta}_0$

Let β_0^j be the j-th component of $\vec{\beta}_0$. For the j-th 5-man unit, set

$$\beta_0^j = \frac{\sum_{i \in S} \Delta_i}{\alpha(j)}$$

where S is the set of observations corresponding to the 5-man unit j.

Plus/minus rate model, i.e., what do we do with Δ_i?

Again, assume now we are working on a particular team's model.

Ridge regression

For fixed λ, find $\vec{\beta}_1$ that minimizes

$$J_\lambda(\vec{\beta}_1) = \| \underbrace{(\vec{y} - X\vec{\beta}_0)}_{\vec{y}'} - X\vec{\beta}_1\|_2^2 + \frac{\lambda}{2}\|\vec{\beta}_1\|_2^2.$$

- X is an $82 \times N$ matrix, where X_{ij} is the number of seconds the 5-man unit j played in game i.
- $\vec{y}$ is 82×1 vector giving margin of victory or defeat in each game.
- Idea is to find $\beta = \vec{\beta}_0 + \vec{\beta}_1$ to minimize both $\|\vec{y} - X\vec{\beta}\|_2$ and $\|\vec{\beta}_1\|_2$.

Game simulation

Procedure for one game

- Run CTMC, proceeding from one 5-man unit to another.
- If unit j is on the floor for τ units of time, it contributes $\tau\beta^j$. Aggregating these contributions over a 48-minute game, we obtain the aggregate plus/minus score for one team.
- We do this for both teams; the team with larger score is declared the winner.
- Each time we simulate a game, we use 100 runs and majority vote to decide winner. Can also compute average margin of victory and probability of victory.

For a best-of-7 series

- Simulate game by game until one team accumulates 4 victories.
- Margin of victory is now in terms of # of games (max $=4$, min $=1$).

Test results

2015 NBA Playoffs

Winner	P. Margin	Prob.	T. Margin
GS	1.74	0.78	4
Hou	0.44	0.57	3
SA	0.42	0.54	LAC, 1
Por	0.29	0.56	Mem, 3
GS	0.32	0.53	2
Hou	0.01	0.53	1
GS	0.88	0.63	3
Atl	2.15	0.82	2
Cle	2.07	0.88	4
Chi	1.11	0.71	2
Tor	0.88	0.64	Was, 4
Atl	1.36	0.72	2
Cle	1.04	0.70	2
Cle	0.31	0.54	4
GS	0.16	0.51	2

- Close series: SA vs LAC and Hou vs LAC difficult to predict.
- Model does not account for other team, e.g., Memphis matched up very well against Portland, same with Houston against Dallas.
- Model does not account for injuries, fatigue. Assumes everyone is at regular-season health/fitness.

What-if scenarios

Eastern Conference Finals

- Atlanta's Kyle Korver was injured and did not play after first two games of the series against Cleveland.
- Our model predicts Cleveland should win this series with prob of 0.54 and margin of 0.31.
- We remove from Atlanta's CTMC any state that involves Kyle Korver and rerun simulation.
- Now Cleveland wins with prob of 0.79 and margin of 1.72, closer to reality.
- We suspect even better agreement will occur if we factor in effects of non-starter playing many minutes for Atlanta, poor matchup against Cleveland, etc.

Ongoing and Future Work

- If we keep CTMC, better ML methods to determine rate matrix M.
- Or, replace CTMC with semi-Markov model or other continuous-time stochastic model. Again, need better ML methods for inference.
- Account for fouls, specific matchups against other team, and time remaining in game.
- Compare against graphical model described by Oh, Keshri, and Iyengar (2015).

Small Advertisement

- UC Merced is the newest campus of the University of California system. Located in Merced, about 100 miles east from Silicon Valley, close to Yosemite National Park.
- The Applied Mathematics unit has an open Visiting Assistant Professor position, essentially a funded 3-year postdoc.
- Funding for PhD students also exist. PhD alumni from my group work as data scientists at Skytree, Microsoft, and startups.
- Please send email (`hbhat@ucmerced.edu`) if interested.

Training results

Confusion matrices

$$\begin{array}{c} \\ PredH \\ PredV \end{array} \begin{array}{c} \begin{array}{cc} TrueH & TrueV \end{array} \\ \begin{bmatrix} 506 & 202 \\ 200 & 322 \end{bmatrix} \end{array} \qquad \begin{array}{c} \\ H \\ V \end{array} \begin{array}{c} \begin{array}{cc} H & V \end{array} \\ \begin{bmatrix} 329 & 94 \\ 152 & 280 \end{bmatrix} \end{array} \qquad \begin{array}{c} \\ H \\ V \end{array} \begin{array}{c} \begin{array}{cc} H & V \end{array} \\ \begin{bmatrix} 220 & 54 \\ 90 & 186 \end{bmatrix} \end{array}$$

From left to right, we have

- all games, games where predicted margin ≥ 5, games where predicted margin ≥ 10
- accuracy increases: 0.67, 0.71, 0.73