NBA Fantasy Basketball Prediction

Xingchi Dai, Andy Shi, Hyungmok Son, Hidenori Tanaka

Institute for Applied Computational Science, Harvard John A. Paulson School of Engineering Applied Sciences

Abstract

Fantasy sports have become so popular in the United States, as many companies offer money fantasy sports players for picking the best line-up of players to compete in the game. In this project, we used what we have learned from AM207 to find the best line-up constrained by factors such as salary, and number of players. We also came up a model to measure NBA players' contributions to their teams and assess how players will perform on a game-to-game basis.

Modeling Approach

We decided to use the offensive rate and the defensive rate to measure one team. The i^{th} team's ratings could be calculated by the following formulas:

$$\begin{aligned} O_i &= \beta_{i,0} + \beta_{i,1}x_{i,1} + \beta_{i,2}x_{i,2} + \ldots + \beta_{i,J}x_{i,J} \\ D_i &= \gamma_{i,0} + \gamma_{i,1}y_{i,1} + \gamma_{i,2}y_{i,2} + \ldots + \gamma_{i,J}y_{i,J} \end{aligned} \tag{1}$$

where D and O are defensive and offensive ratings. The j^{th} player's x and y shall be calculated using the following formulas:

$$\begin{aligned} y_j &= \text{blocks} + \text{steals} + \text{defensive rebounds} \\ x_j &= \text{points per game} + \text{assists per game} \end{aligned} \tag{2}$$

The model is trained on the score differences from NBA games in 2016. The score difference follows a normal distribution, justifying some model choices.

$$\mathcal{N}(O_{guest} - D_{host} - (O_{host} - D_{guest}), \sigma)$$

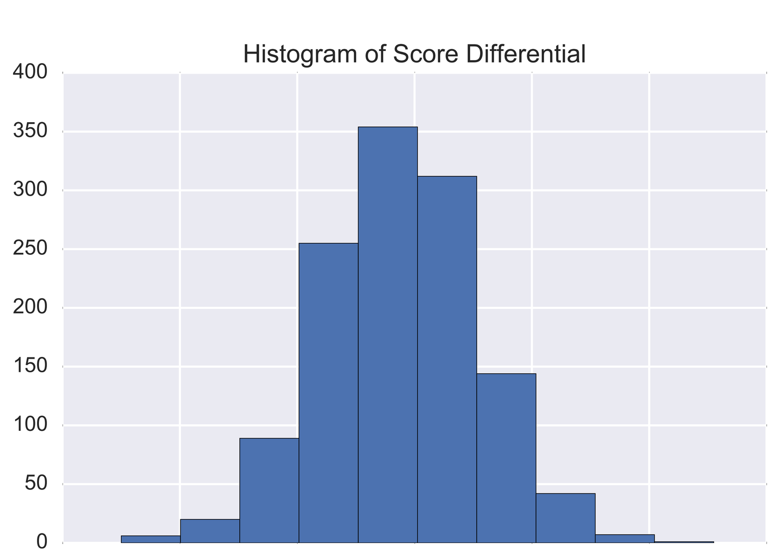

Figure 1: Distribution of Score Differences.

Lineup Selection

Players were evaluated on points per game, and our team was constrained with a salary cap of $55 million. Simulated annealing outperforms the greedy method by achieving a value of 236 vs. 176.

Naive Method	Simulated Annealing
Stephen Curry	Stephen Curry
James Harden	Jordan Clarkson
Jordan Hamilton	C.J. McCollum
Michael Carter-Williams	Michael Beasley
Harrison Barnes	DeMarcus Cousins
C.J. Miles	Hassan Whiteside
Omri Casspi	Damian Lillard
Ish Smith	Dahntay Jones
J.R. Smith	Giannis Antetokounmpo
Joe Johnson	Evan Fournier
Isaiah Thomas	Isaiah Thomas
J.J. O'Brien	Anthony Davis

Table 1: Lineups generated with the naive method (greedy algorithm) and simulated annealing.

Estimating Player Importance

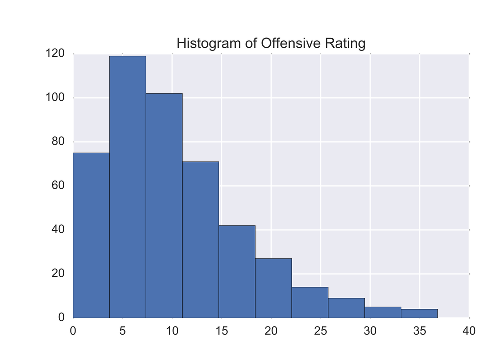

Figure 2: Histogram of offensive rating

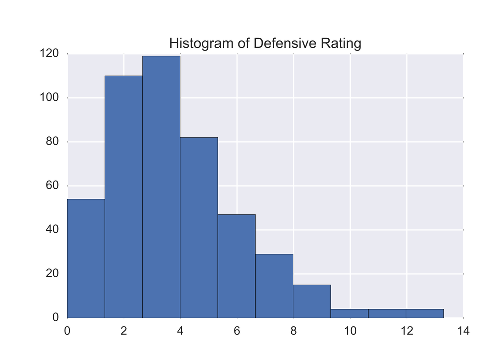

Figure 3: Histogram of defensive rating.

Player	Team	Salary	PPG	beta
Stephen Curry	GSW	11.4	30.1	0.112917
Kawhi Leonard	SAS	16.5	21.2	0.10513
LaMarcus Aldridge	SAS	19.5	18	0.086141
Tony Parker	SAS	13.4	11.9	0.075976
Klay Thompson	GSW	15.5	22.1	0.074252
Russell Westbrook	OKC	16.7	23.5	0.073085
Kevin Durant	OKC	20.2	28.2	0.071576
Draymond Green	GSW	14.3	14	0.065665
LeBron James	CLE	23.0	25.3	0.059836
Manu Ginobili	SAS	2.8	9.6	0.056099
Patrick Mills	SAS	3.6	8.5	0.049918
Tim Duncan	SAS	5.0	8.6	0.049911
Kyrie Irving	CLE	14.8	19.6	0.045292

Table 2: Player importances, as estimated using L-BFGS optimization.

Prediction of Player's Condition

- **Short term condition**: Used 1st and 2nd order Markov model to predict player's score in the next game given results from previous games.
- We divided player's condition into 4 states and successfully predicted condition in the next game with around 50% accuracy in both models.
- **Medium term condition**: Predicted trend for next 10 games with an ARMA model.

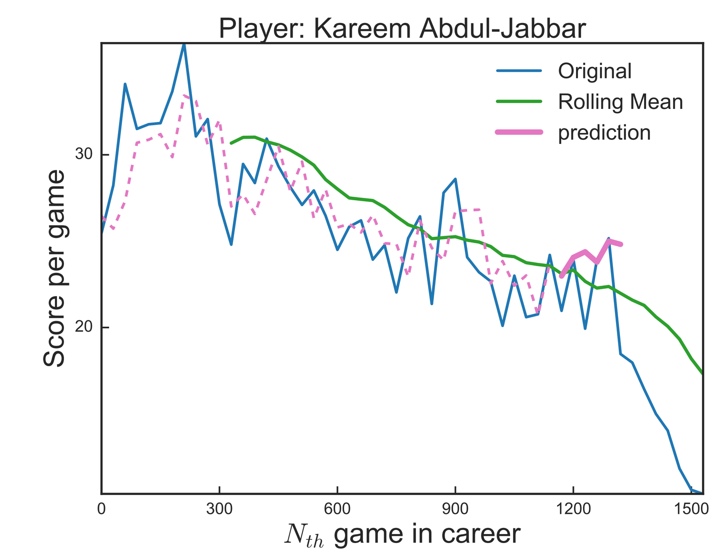

Figure 4: Points/game in Kareem Abdul-Jabbar's entire career

.

Simulated Annealing

- Simulated annealing (SA) converged faster as the temperature changed over the time.
- Got different results as SA might still get stuck in local optima. Fixed this by using random restarts.

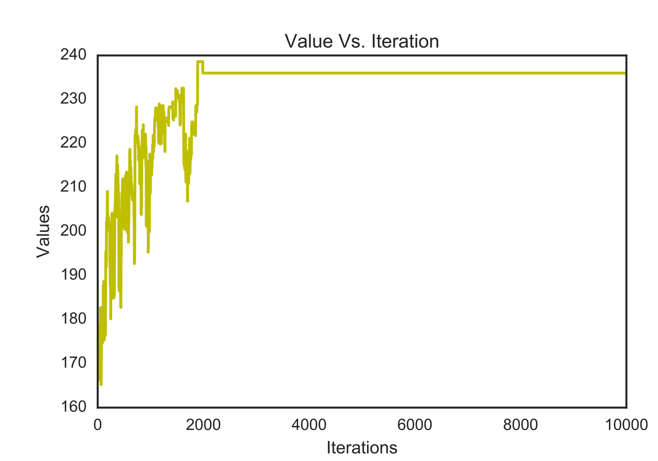

Figure 5: Value vs. iteration of simulated annealing. This plot shows the eventual convergence of SA.

L-BFGS and SGD

- We directly optimized an ℓ_2 regularized version of our likelihood using L-BFGS and SGD.
- SGD did not converge because we used one data point at a time and the gradient was too noisy.
- L-BFGS did converge.

Model Validation

- Using L-BFGS model fitting, achieved 70% accuracy on winner prediction for last 20% of games, when training on the first 80%.

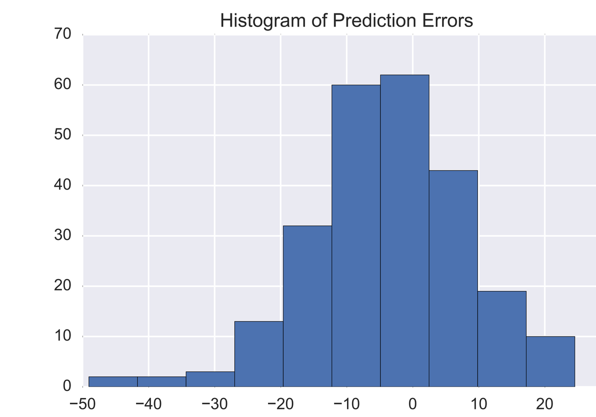

Figure 6: Histogram of prediction errors for the last 20% of the games in the 2015-2016 season

.

Sampling Methods

- Used PyMC (implementation of Metropolis-Hastings) to obtain posterior samples from β and γ parameters.
- Used independent normal priors for each parameter.
- PyMC doesn't seem to converge.
- Another approach used elliptical slice sampling (ESS). ESS is another way to efficiently sample from parameters with Gaussian priors. Also did not converge.

Discussion and Next Steps

- Pooling: Currently, we assume each player is independent. However, players in the same team should be correlated since they have the same coach and team organization, but it's hard to quantify this correlation.
- Sampling Performance: Some of our methods still haven't converged yet. We need to get more samples, and improve our sampling performance.

References

- Adams, R.P., Dahl, G.E. and Murray, I., (2010). Incorporating side information in probabilistic matrix factorization with gaussian processes. arXiv preprint arXiv:1003.4944.
- Murray, I., Adams, R. P., and MacKay, D. J. (2009). Elliptical slice sampling. arXiv preprint arXiv:1001.0175.