A Spatial Analysis of Professional Basketball:

Shot Selection of James Harden

Biancheng Wang[1]

1. Introduction

1.1 Background

Over the past ten years, basketball analytics becomes more and more popular and we can find some interesting paper published on the top statistics journal and machine learning conference. A Miller, L Bornn, R Adams and K Goldsberry (2014) use log-Gaussian Cox process to analysis the shot location of different NBA players and implements factorization to find the shooting type of players.

Daryl Morey may be the first one who applied statistics and data science methods to the professional basketball. He is the General Manager of the Houston Rockets of NBA, one of the best teams in the league in the recent years. Also, he is the co-founder of MIT Sloan Sports Analytics Conference. He has his own unique basketball philosophy, unlike the traditional one, which is called "Moreyball". This is a style of play that generates 3-pointers and layups, and uses the midrange as a last option. The basic idea behind "Moreyball" is very intuitive. He thinks that a smart basketball player should shoot at the location where he can get the largest expected scores. With the help of this idea, Houston Rockets has become more competitive in the league and has broken many NBA records, including most 3-pointers made in a regular season. In Houston Rockets, six-time NBA All-Star James Harden is definitely the leader of this team. He is also the best shooting guard of this era. Since he is the best iso player and the hardest player to guard in NBA, it might be important for us to find some interesting pattern in his shot selection.

[1] Department of Statistics, UCLA (email: wangbcbill@ucla.edu)

1.2 Motivation

There are several questions we want to answer in this paper. Which model can best capture his shot selection? Is there any shooting habit can be found based on our models? For the latter one, we mainly focus on the following two questions. Does the pattern of shot location accord with Morey's basketball philosophy? Is James Harden a "stubborn" player?

From the perspective of psychology, if one player make the shot at the certain position, he might be more likely to take more shots near that position. If a player has this kind of shooting habit, we regard him as a "stubborn" player. In this paper, we try to fit a Hawkes Model to check whether there exists self-exciting phenomenon by looking at the productivity estimated.

Basic basketball statistics have long been unfair to players who shoot a lot of 3-pointers. Players who take and make a large number of 3-point shots often don't have a very impressive field goal percentage, but that doesn't tell the whole story in terms of the points that player produces. Therefore, Effective Field Goal Percentage (eFG%) was created, which is defined as:

$$eFG\% = \frac{FGM + 0.5 \times 3PM}{FGA}$$

, where FGM is field goals made, $3PM$ is 3 point field goals made, and FGA is field goal attempts. This metric adjusts field goal percentage to account for the fact that three-point field goals count for three points while field goals only count for two points. With the help of eFG%, we can find out whether the shot selection of James Harden accords with the idea of "Moreyball" by checking the sign of coefficient of this covariate.

In the next section, we describe the dataset obtained and also some data pre-processing we did before analysis. Section 3 is dedicated to giving some basic analysis results. In Section 4, we fit Poisson Process and Hawkes Process to find the answer to the question we raised above. Finally, we conclude and give more discussion in Section 5.

2. Data

2.1 Data Description

The data we work with in this analysis comes from the NBA official website[2] and a sports data website[3] which collects data from ESPN's Shot Tracker. It contains shot locations, shot time and shot results of James Harden during 2017-2018 regular season. There are totally 1306 field goal attempts in 64 games. Also, we get the field goal percentage of James Harden in different ranges during 2016-2017 regular season.

2.2 Data Pre-processing

As we all know, the basketball court is a rectangle. It is 94 feet times 50 feet. Since the distance matters in some of our models and there is no reason to provide two dimensions with different weights, we choose a square part of full court and then normalize X and Y to $[0,1] \times [0,1]$, which is showed in the Figure 1. There is only one field goal attempt in the backcourt. We treat it as outlier and omit this observation. As shown in Figure 1, Kobe Bryant took more midrange shots than James Harden as expected.

As for time variable, we combine the 64 games together and transfer the time to $[0, 64 \times 48]$, where 48 is the total minutes per game. Actually, we ignore the time between different games. Then we can normalize it.

Besides, we calculate eFG% based on his field goal percentage in different ranges during the last season. From the Figure 2, we can find that James Harden performs better around the basket and outside the 3-point line.

3. Basic Spatial Analysis

At the beginning of analysis, we implement kernel smoothing and calculate K, L, F, G, J functions. From the Kernel Smoothing plot (Figure 3), it seems that his shot location mainly

[2] http://stats.nba.com/
[3] http://nbasavant.com/

concentrates near the rim. We can also find some dark pixels above the break while both the color of pixels in the left corner and the right corner seem to be light. This is not a surprise because James Harden is the commander on the court and he rarely stands at the corner.

K, L function plotted in Figure 4 are far out of the confidence interval and J function plotted in Figure 5 is much smaller than 1, which can give us more evidence of clustering.

4. Modeling Analysis

In this part, we follow the logic flow in Section 1.2 by using both Poisson and Hawkes Process.

4.1 Poisson Process without covariates

We start from fitting an Inhomogeneous Poisson Process without covariates. Both the linear and quadratic intensity model fail as expected. Then we try to use cubic intensity as follows:

$$\lambda(x,y) = a_1 + a_2x + a_3x^2 + a_4x^3 + a_5y + a_6y^2 + a_7y^3 + a_8xy$$

However, Figure 5 shows that the fitted intensity still looks very different from the Kernel Smoothing plot (Figure 3). Therefore, based on what we see, maybe it is not a good idea to fit a Poisson Process without covariates.

4.2 Poisson Process with covariates

We add eFG% into our model as a covariate to improve the fit. Since this covariate has already contained some second-order information of x and y, we just use linear function of them. Therefore, the model of intensity can be written as follows:

$$\lambda(x,y) = a_1x + a_2y + a_3w(x,y) + a_4$$

, where $w(x,y)$ is $100 \times eFG\%$ at (x,y).

Parameters estimated are listed in Table 1. The positive coefficients of a_1 implies that James Harden seems more likely to shoot at the right part of the court. This might be reasonable because he is a left-hander. Also, we are interested in the sign of the coefficients of our covariate. The positive coefficient, which is a_3 here, seems to confirm that James Harden acts as Morey suggests.

After fitting this model, we calculate the fitted intensity and conduct super-thinning based on the fitted Poisson Process with covariates. As we can see from the Figure 6, this fitted intensity looks more similar to the Kernel Smoothing plot (Figure 3) than the one we get in Poisson Process without covariates. Figure 7 shows the super-thinned points, which are sparse in some area of the court. K function in Figure 8 becomes closer to the stationary Poisson one but both K and L function are still out of the confidence interval, suggesting that this model seems to fit better but still not good enough. Also, J function in Figure 9 still indicates clustering among the super-thinned points.

4.3 Hawkes Process with covariates

To find out whether James Harden is "stubborn" at shooting, we use a Hawkes Process with covariates to check the existence of triggering pattern. Since adding eFG% gives us better fit in the Poisson Process, we add this variable into our background rate in the Hawkes Process. Then we have the Hawkes Process with covariates:

$$\lambda(t,x,y) = \mu\, exp\big(a\, w(x,y)\big) + \kappa \sum_i g_t(t-t_i) g_{xy}(x-x_i, y-y_i)$$

In this function, $g_t(t) = \beta\, e^{-\beta t}$, $g_{xy}(x,y) = \alpha\, e^{-\alpha r^2}/\pi$ where $x^2 + y^2 = r^2$, $w(x,y)$ is eFG% at (x,y) and $T = 10^4$. From the results in Table 3, we can find that $\kappa = 2.814e-04$ which is close to 0. This means that only 0.3% of points are triggered directly by each point, implying that there might be little self-exciting pattern in James Harden's shot selection. Again, the coefficient of covariates remains positive.

From Figure 11, we can find that super-thinned points appear roughly homogeneous Poisson. Part of L function in Figure 12 locates within the confidence interval and J function in Figure 13 is larger than the one we get from super-thinned Poisson points but still smaller than 1. It seems that Hawkes Process fits better although not so excellently.

4.4 Further Analysis

One might think that the field goal missed may generate a shot far from this location and this may influence our test. Therefore, we try to fit Hawkes Process with covariates only for field goals made. However, we still get a κ which is close to 0. This may be a further evidence that James Harden is not "stubborn" at shooting.

Besides, other parameter estimation methods have been tried. We get a larger κ by Stoyan's method but the log-likelihood is terrible. Thus, we think it might not be safe to accept this result.

5. Conclusion and Future Improvement

In conclusion, we can find many evidences of clustering among James Harden's shot locations. A Hawkes Process seems to fit better than a Poisson Process for our data and adding eFG% as a covariate can better capture the intensity. Based on the Poisson Process we fit, it seems that James Harden is more likely to shoot at the right side of court. Positive estimated coefficients of eFG% in both models seem to imply that his shot selections accord with "Moreyball". A small productivity might indicate that he is not a "stubborn" player in shooting. We think this is one of reasons why he is so hard to guard.

There is still something we can do in the future. First, a more accurate covariate can be used instead, e.g., eFG% by different zones. Second, we can try more complicated models, e.g., log-Gaussian Cox process, which has been used by A Miller, et al (2014). Last but not least, we can regard the coefficient of eFG% as a "Moreyball" index indicating how well his shot selection accord with Morey's basketball philosophy. Then we can also compare among different players using this index, which will be useful information for mangers of NBA teams.

6. Reference

[1] A Miller, L Bornn, R Adams, K Goldsberry (2014). Factorized point process intensities: A spatial analysis of professional basketball. International Conference on Machine Learning, 235-243.

7. Appendix

7.1 Figures

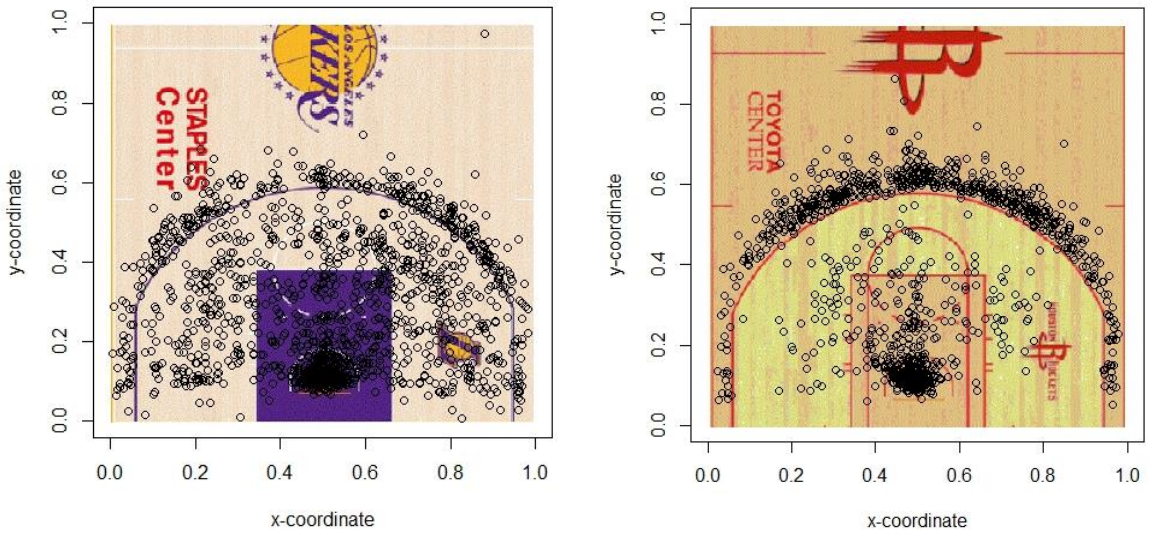

Figure 1: Shot Locations of Kobe Bryant and James Harden

(*Notes:* The left figure is shot locations of Kobe Bryant (2012-2013) and the right one belongs to James Harden (2017-2018).)

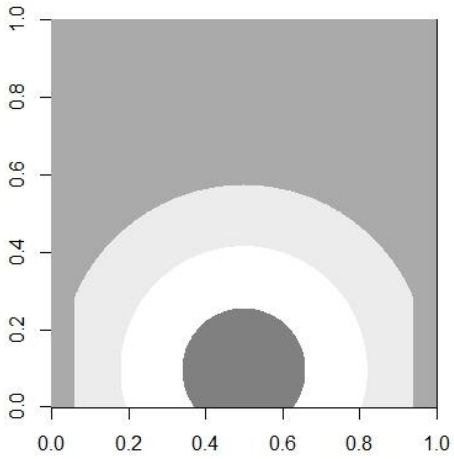

Figure 2: eFG% of James Harden in Different Ranges during 2016-2017 Season

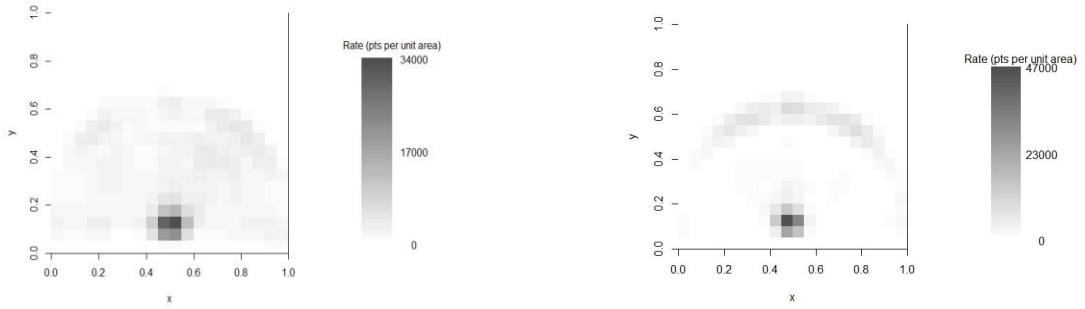

Figure 3: Kernel Smoothing of Shot Locations of Kobe Bryant and James Harden

(*Notes:* The left figure belongs to Kobe Bryant (2012-2013) and the right one belongs to James Harden (2017-2018).)

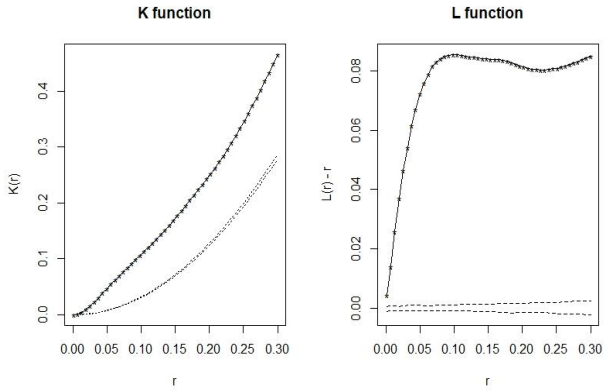

Figure 4: K, L Function of Original Data

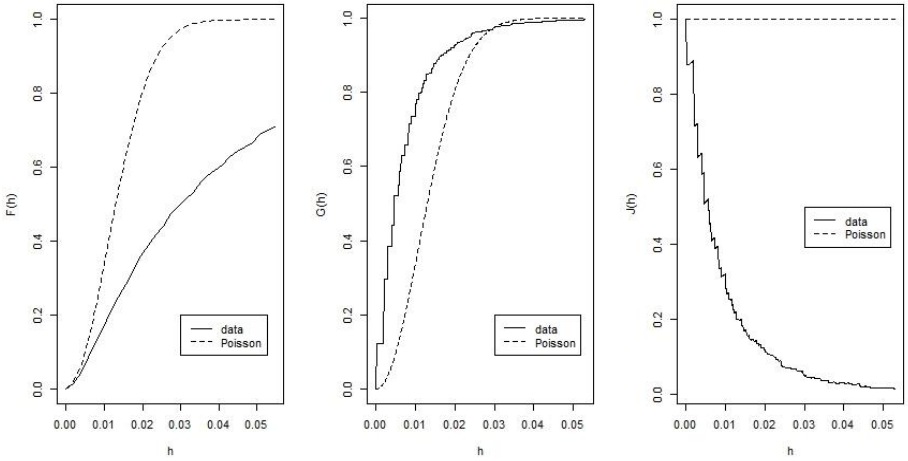

Figure 5: F, G, J Function of Original Data

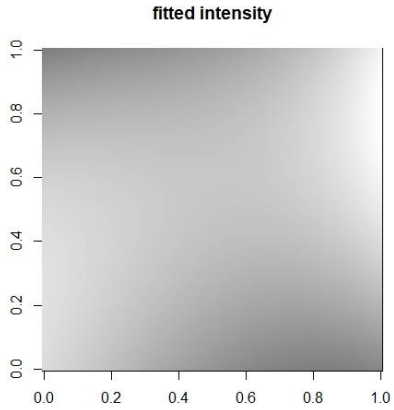

Figure 6: Fitted Intensity by Cubic Inhomogeneous Poisson Process without covariates

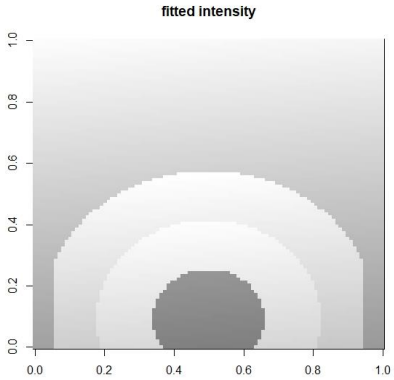

Figure 7: Fitted Intensity by Poisson Process with covariates

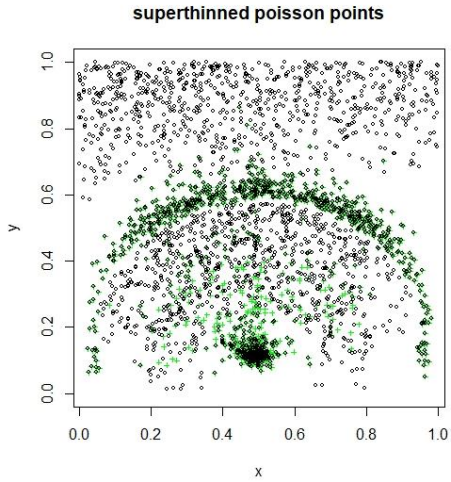

Figure 8: Super-thinned points by Poisson Process with covariates

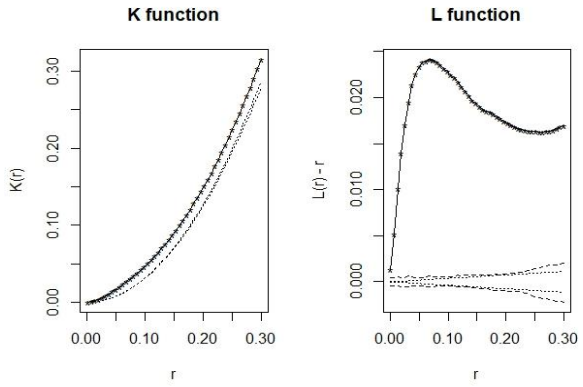

Figure 9: K, L Function of Super-thinned points by Poisson Process with covariates

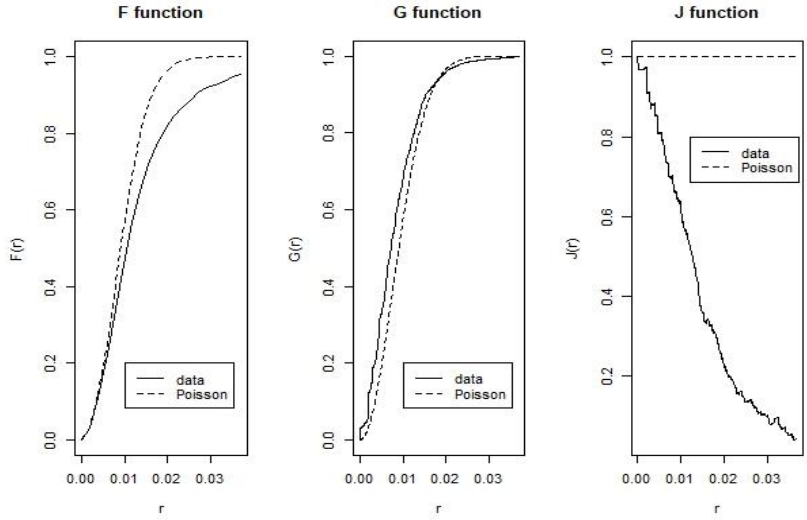

Figure 10: F, G, J Function of Super-thinned points by Poisson Process with covariates

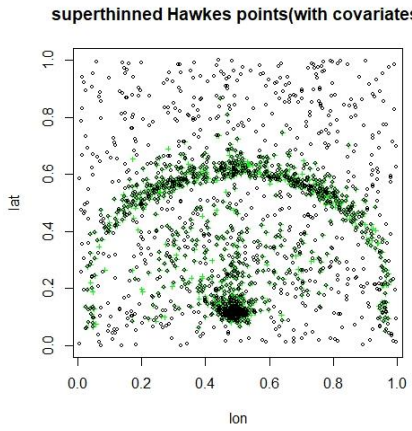

Figure 11: Super-thinned points by Hawkes Process with covariates

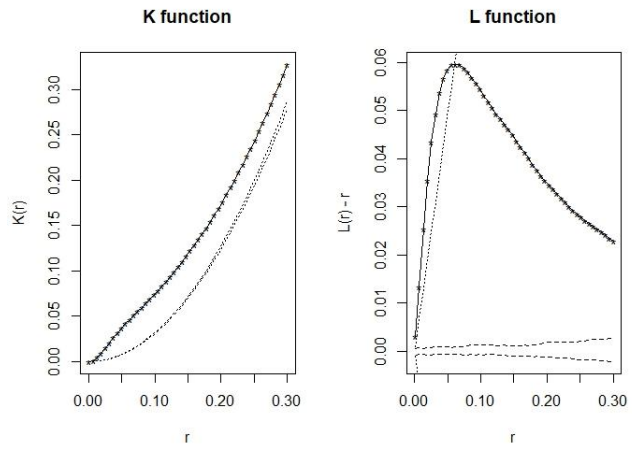

Figure 12: K, L Function of Super-thinned points by Hawkes Process with covariates

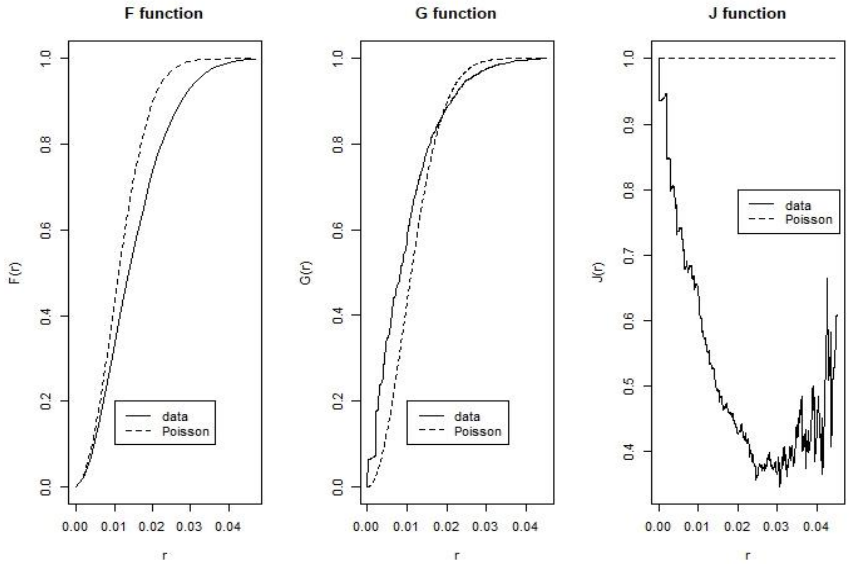

Figure 13: F, G, J Function of Super-thinned points by Hawkes Process with covariates

7.2 Tables

Table 1: Parameter Estimation for Cubic Inhomogeneous Poisson Process without covariates

Parameter	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
Estimate	1.884	27.889	27.273	-36.019	-9.848	11.673	19.797	-46.772

Table 2: Parameter Estimation for Poisson Process with covariates

Parameter	a_1	a_2	a_3	a_4
Estimate	**0.833**	-9.988	**0.507**	-15.348
SE	1.911	1.971	0.038	1.597

Table 3: Parameter Estimation for Hawkes Process with covariates

Parameter	μ	κ	α	β	a
Estimate	2.102e-05	**2.814e-04**	1.611	15.511	**17.025**
SE	1.213e-06	6.421e-06	19.803	26.438	0.113