Data Science in Action: An Application to NBA Basketball Analytics

Bradley Lecture II

Dipak K. Dey

University of Connecticut

Dipak.dey@uconn.edu

NBA Basketball Analytics

- Data Analytics has become a necessary part of sports activities. It enterprises as an alternative, we can say in each component, Data Analytics has become an integral part, and the NBA is no exception.
- NBA stands for the National Basketball Association, which is an expert basketball league in North America. It includes 30 teams, 29 from the USA and 1 from Canada, and is extensively considered to be the world's premier men's expert basketball league. The NBA season runs from October to June, with the playoffs taking place with inside the spring. The league was founded in 1946 and has since become a worldwide phenomenon, with a big fan base and numerous international players.

Player Performance Analysis

- NBA teams use Data Analytics to advantage a competitive edge by analyzing player performance, scouting opponents, and enhancing game strategy. We will discuss how NBA teams use Data Analytics to enhance their performance.
- For example, one team used Data Analytics to identify that their players were no longer performing properly in the third quarter. By analyzing participant data, they decided that the players were experiencing fatigue within side the third quarter. To fight this, the crew implemented a brand new training regime to enhance participant endurance, resulting in a significant development in overall performance in the course of the third quarter.

Scouting Opponents

- NBA teams also use Data Analytics to scout opponents by analyzing player statistics and game footage. They use this information to develop game strategies and identify weaknesses in their opponents. By analyzing their opponent’s tendencies, teams can develop defensive strategies to shut down their opponents' strengths and exploit their weaknesses.
- NBA teams use Data Analytics to research shot selection and determine which shots are simplest for their players. Teams can use records to determine which regions of the court their players shoot from with the very best accuracy, which sorts of shots are a maximum success for each player, and which shots are best in opposition to specific opponents.

Performance Analysis

- NBA teams use Data Analytics to investigate players overall performance and identify potential acquisitions. Teams use data to decide which players will be the best match for their crew, primarily based totally on their playing style, strengths, and weaknesses.

Injury Prevention

- NBA teams use Data Analytics to monitor player health and prevent injuries. Teams collect data from wearable technology and sensors on the court to monitor player movement and identify to potentially injury risks.

NBA Basketball Analytics

- NBA teams use Data Analytics to benefit the competitive side by analyzing participants' overall performance, scouting opponents, and enhancing game strategy. By using the information to optimize participants' overall performance, develop game strategies, and perceive opponent weaknesses, teams can enhance their overall performance and increase their probability of winning. As Data Analytics continues to evolve, we can count on seeing even extra advanced data analysis strategies being used within the NBA and different expert sports activities leagues.

Objectives of Our Research

- Analyzing players' "hotspots", i.e., locations where they make the most shooting attempts, is an indispensable part of basketball data analytics. Identifying such hotspots, as well as which players tend to have similar hotspot locations, provides valuable information for coaches as well as for teams who are aiming at making transactions and looking for players of a specific type.
- One preliminary tool for representing shot locations is the shot chart, which is rather rough as there is no clear-cut way of defining "similarity", which calls for the need for more rigorous statistical modeling.

Related Works

- Various tools have been proposed to model point patterns. Among them, spatial point processes is a family of models that assume event locations are random, and realized from an underlying process, which has an intensity surface.
- Spatial point processes have a wide range of variants, the most prominent of which being the Poisson process the Gibbs process and the log-Gaussian Cox process.
- Reich et al. (2006) developed a multinomial logit model that incorporates spatially varying coefficients, which were assumed to follow a heterogeneous Poisson process. Miller et al. (2014) discussed creating low-dimensional representation of players' shooting habits using several different spatial point processes. These works, however, focus mainly on characterizing the shooting behavior of individual players. Which players are similar to each other, however, remains un-answered by these works.

Different Modeling Strategies

- Using LGCP (Log Gaussian Cox process) to obtain the underlying intensity, and then defined a similarity measure on the intensities of different players, which was later used in a hierarchical model that employed mixture of finite mixtures (MFM; Miller and Harrison, 2018) to perform clustering.
- A Bayesian nonparametric matrix clustering approach to analyze the latent heterogeneity structure of estimated intensity surfaces. Note that in all the previous works, the intensity function always played a certain role, which adds another layer of modeling between the shots and the grouping structure.

Motivating Data

- Our data consists of both made and missed field goal attempt locations from the offensive half court of games in the 2017/2018 NBA regular season.
- The data is available at http://nbasavant.com/index.php, and also on GitHub (https://github.com/ys-xue/MFM-ZIP-Basketball-Supplemental).
- We focus on players that have made more than 400 field goal attempts. Also, players who just started their careers in the 17/18 season, are not considered. A total of 191 players who meet the two criteria above are included in our analysis.
- We model a player's shooting location choices and outcomes as a spatial point pattern.

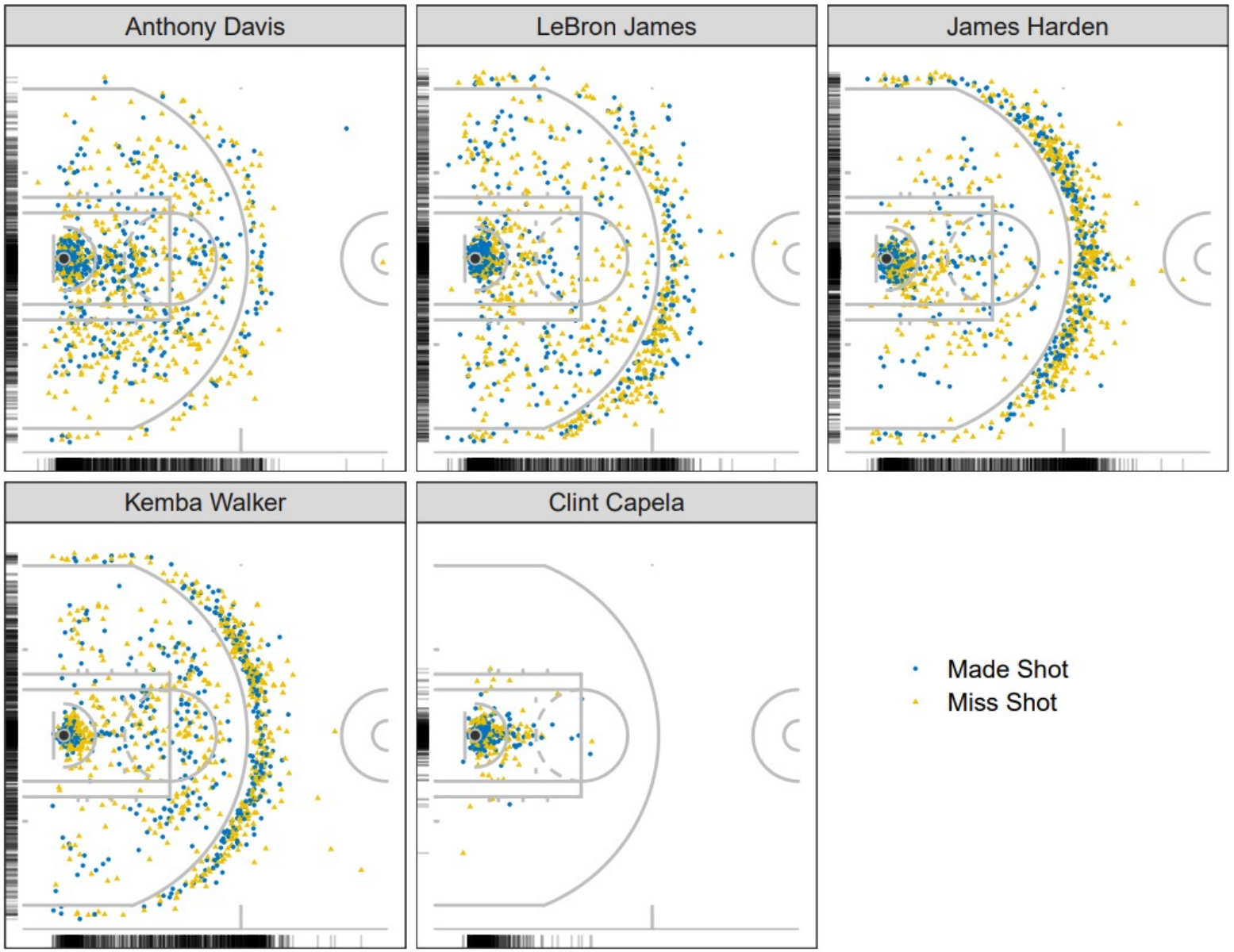

Figure 1: Shot charts for selected NBA players.

Descriptive Analysis

- On the offensive half court, a 47 ft by 50 ft rectangle, which is the standard size for NBA.
- The spatial domain for the basketball court is denoted as D 2 [0; 47] x [0; 50]. We partition the court to 1 ft _ 2 ft blocks, which means that there are in total 47 x 25 = 1175 blocks in the basketball court.
- The shot charts for selected players are visualized in Figure 1.
- The numbers of shot attempts in each of the blocks are counted. Hence, this data consists of non-negative, highly skewed sequence counts with a large proportion of zeros, as most shots are made in the range from the painted area to the three-point line, and many of the blocks between the three-point line and mid-court line have no corresponding positive values. This abundance of 0's motivates the usage of zero-inflated models for such type of data.

Different Modeling Strategies

- Marked point process joint modeling approach which takes into account both shot locations and outcomes. The fitted model parameters are grouped using ad hoc approaches to identify similarities among players.
- Model-based clustering approach that incorporates the Chinese restaurant process to account for the latent grouped structure. The number of clusters is readily inferred from the number of unique latent cluster labels.
- This was further improved using Markov random fields constraint Dirichlet process for the latent cluster belongings, which effectively encourages local spatial homogeneity

Methodology

- One natural way to model the counts directly without employing the intensity surface is the Poisson regression by proposing a spatial homogeneity pursuit regression model for count value data, where clustering of locations is done via imposing certain spatial contiguity constraints on MFM (Mixture of Finite Mixture).
- Data of basketball shots, however, poses more challenges. The first challenge comes from the fact that only few shots are made by players in the region near the half court line, which means there is a large portion of the court that corresponds to no attempts. Secondly, existing approaches either only perform clustering on the spatial domain. Thirdly, to demonstrate its superiority over heuristic comparison and grouping, a model-based approach should have favorable theoretical properties such as consistent estimation for the number of clusters.

Methodology

- Marked point process joint modeling approach that takes into account both shot locations and outcomes. The fitted model parameters are grouped using ad hoc approaches to identify similarities among players.
- Proposed a model-based clustering approach that incorporates the Chinese restaurant process to account for the latent grouped structure. The number of clusters is readily inferred from the number of unique latent cluster labels.
- We propose a Bayesian zero-inflated Poisson (ZIP) regression approach to model field goal attempts of players with different shooting habits.

Three-Fold Contributions

- First, the large proportion of the court with zero shot attempts is accommodated in the model structure by zero inflation.
- Next, non-negative matrix factorization is utilized to decompose the shooting habits of players into linear combinations of several basis functions, which naturally handles the homogeneity pursuit on the spatial dimension. On the dimension across players, we for the first time introduce a MFM prior in ZIP model to jointly estimate regression coefficients and zero inflated probability and their clustering information.
- Finally, we provide both theoretical and empirical justification through simulations for the model's performance in terms of both estimation and clustering.

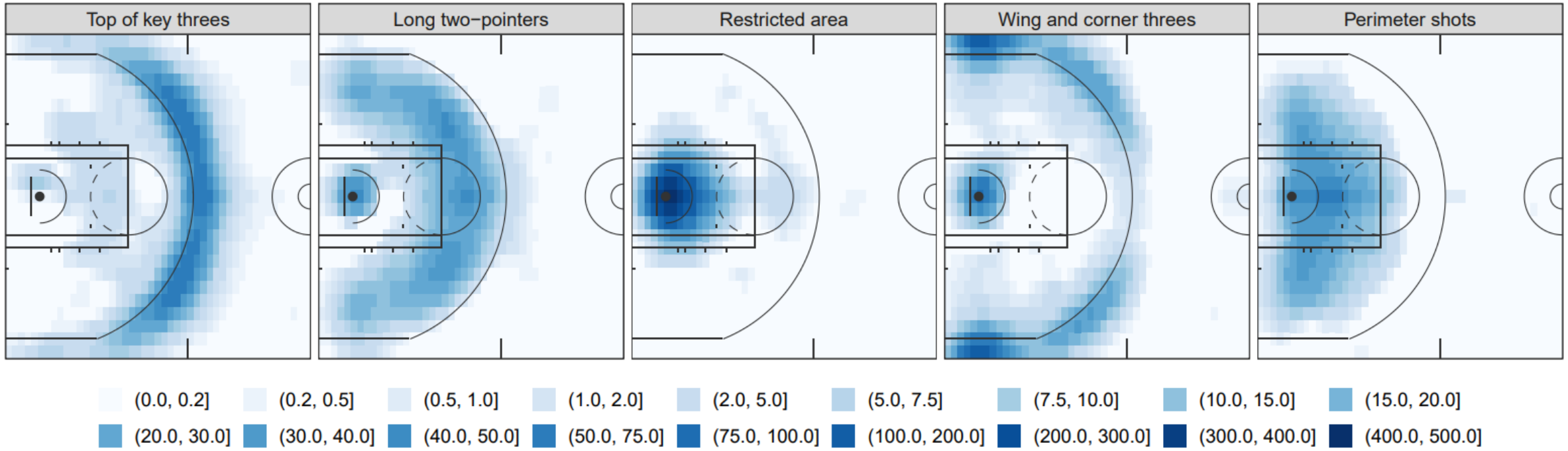

Figure 2: Visualization of basis functions obtained by NMF for K = 5. Each basis function represents the intensity function of a particular shot type.

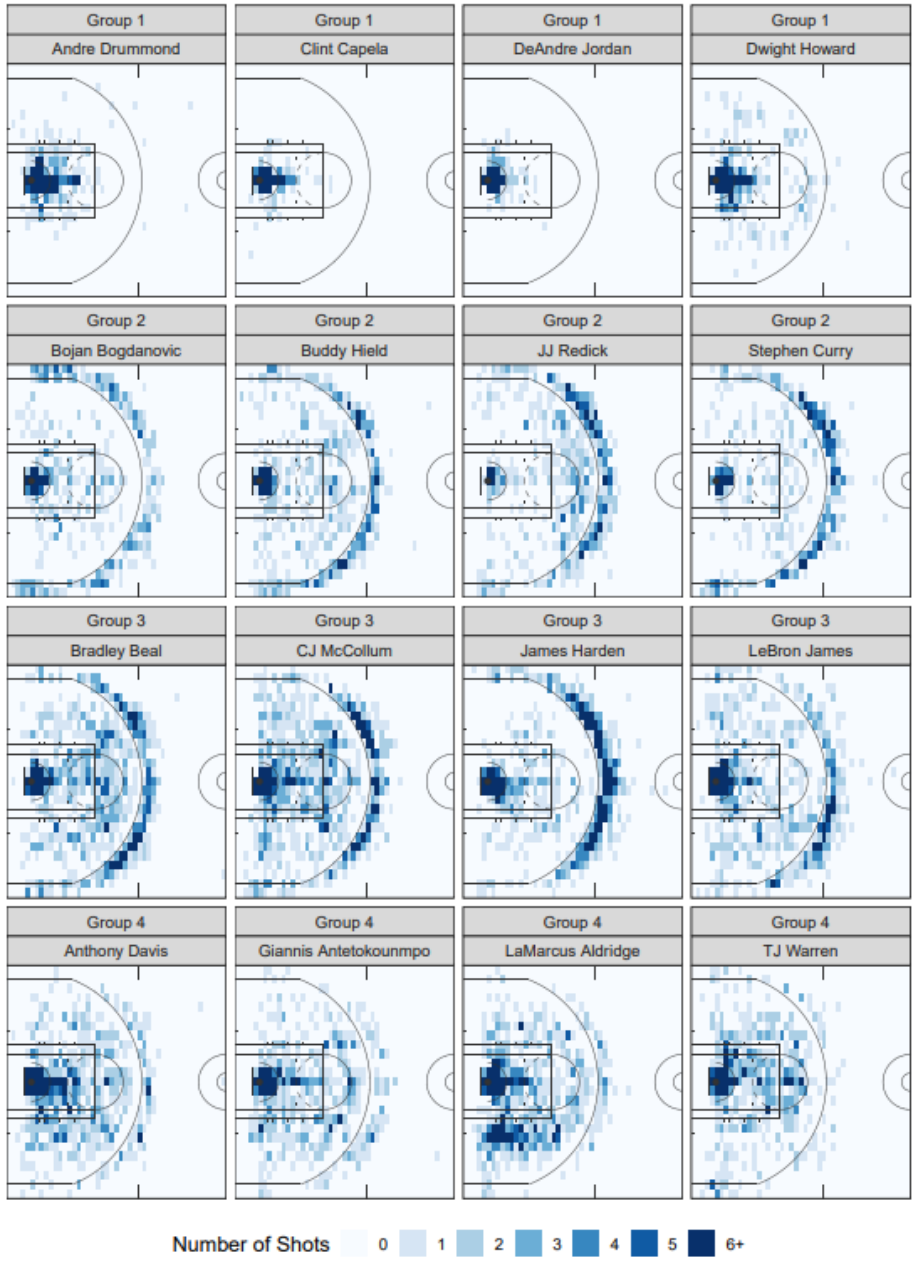

Figure 3: Visualization of shooting patterns for four selected players from each group.

Main References

1. “Zero Inflated Poisson Model with Clustered Regression Coefficients: an Application to Heterogeneity Learning of Field Goal Attempts of Professional Basketball Players.” *Canadian Journal of Statistics*, March 2023, Vol. 51, Issue 1, Pages 157-172. G. Hu, H-C Yang, Y. Xue and D. K. Dey.

2. “Spatial Clustering Regression of Count Value Data via Bayesian Mixture of Finite Mixtures”. P Zhao, HC Yang, DK Dey, G Hu - Proceedings of the 29th ACM SIGKDD conference. Pages 504-512. KDD ’23, August 6ś10, 2023, Long Beach, CA, USA

- ACM ISBN 979-8-4007-0103-0/23/08.
- https://doi.org/10.1145/3580305.3599509

UCONN Sports Analytics Symposium (UCSAS)

- While there are many well established sports analytics conferences, they are often not accessible to students due to technical level, cost, or space limitations. UConn, recognized nationally for its teams in sports such as basketball, baseball, and hockey, among others, hosts the UConn Sports Analytics Symposium (UCSAS), which will focus specifically on undergraduate and graduate students who are interested in sports analytics or more broadly, data science. UCSAS, started in 2019, aims to:
- Showcase sports analytics to students at an accessible level;
- Train students in data analytics with application to sports data; and
- Foster collaboration between academic programs and the sports industry.
- https://statds.org/events/ucsas2024/photos/ucsas2024.pdf

Journals on Sports Statistics

Journal of Quantitative Analysis in Sports

Journal of Sports Analytics

International Journal of Computer Science in Sport

Journal of Sports Science and Medicine

International Journal of Sports Science and Engineering

Catch the future star players