Received: 17 July 2020 | Revised: 11 August 2020 | Accepted: 5 October 2020
DOI: 10.1002/sta4.324

WILEY

ORIGINAL ARTICLE

Bayesian group learning for shot selection of professional basketball players

Guanyu Hu[1] | Hou-Cheng Yang[2] | Yishu Xue[3]

[1]Department of Statistics, University of Missouri-Columbia, MO, USA
[2]Department of Statistics, Florida State University, FL, USA
[3]Department of Statistics, University of Connecticut, CT, USA

Correspondence
Hou-Cheng Yang, Department of Statistics, Florida State University, FL, USA.
Email: hy15e@my.fsu.edu

In this paper, we develop a group learning approach to analyze the underlying heterogeneity structure of shot selection among professional basketball players in the NBA. We propose a mixture of finite mixtures (MFM) model to capture the heterogeneity of shot selection among different players based on the Log Gaussian Cox process (LGCP). Our proposed method can simultaneously estimate the number of groups and group configurations. An efficient Markov Chain Monte Carlo (MCMC) algorithm is developed for our proposed model. Simulation studies have been conducted to demonstrate its performance. Finally, our proposed learning approach is further illustrated in analyzing shot charts of selected players in the NBA's 2017–2018 regular season.

KEYWORDS
basketball shot charts, heterogeneity pursuit, log gaussian cox process, mixture of finite mixtures, nonparameteric bayesian

1 | INTRODUCTION

In basketball data analytics, one primary problem of research interest is to study how players choose the locations to make shots. Shot charts, which are graphical representations of players' shot location selections, provide important summary of information for basketball coaches as well as teams' data analysts, as no good defense strategies can be made without understanding the shot selection habits of players in the rival teams. Shot selection data have been discussed from different statistical perspectives. Reich, Hodges, Carlin, and Reich (2006) developed a spatially varying coefficients model for shot-chart data, where the court is divided into small regions and the probability of making a shot in these zones is modeled using the multinomial logit approach. Recognizing the random nature of shot location selection, Miller et al. (2014) analyzed the underlying spatial structure among professional basketball players based on spatial point processes. Franks, Miller, Bornn, and Goldsberry (2015) combined spatial and spatio-temporal processes, matrix factorization techniques, and hierarchical regression models for characterizing the spatial structure of locations for shot attempts. In spatial point processes, locations for points are assumed random and are regarded as realizations of a process governed by an underlying intensity. Spatial point processes are well discussed in many statistical literatures, such as the Poisson process (Geyer, 1998), the Gibbs process (Goulard, Särkkä, & Grabarnik, 1996), and the Log Gaussian Cox process (LGCP Møller, Syversveen, & Waagepetersen, 1998). In addition, they have been applied to different areas, such as ecological studies (Jiao, Hu, & Yan, 2020; Thurman, Fu, Guan, & Zhu, 2015), environmental sciences (Hu, Huffer, & Chen, 2019; Veen & Schoenberg, 2006), and sports analytics (Jiao et al. 2019; Miller et al. 2014). Most existing literatures concentrate on parametric (Guan, 2008) or nonparametric (Geng, Shi, & Hu, 2019; Guan, 2008) estimation of the underlying intensities for the spatial point process and analysis of second-order properties (Diggle, Gómez-Rubio, Brown, Chetwynd, & Gooding, 2007). There are very limited literatures discussing the grouping pattern of multiple point processes. Knowing the group information of different point processes will lead to discovery of the underlying heterogeneity structure of different players.

Jiao et al. (2019) proposed a joint model approach for basketball shot chart data. After model parameter estimates are obtained for different players, they are grouped via *ad hoc* clustering approaches, such as hierarchical clustering. Chen, Pan, Guan, and Wang (2019) developed a group linked Cox process model for analyzing point of interest (POI) data in Beijing. To determine the number of groups, starting from the most complicated model where each observation is its own group, a loss function is used in a series of hierarchical merging steps to combine the groups. In both methods, the inherent uncertainty in estimation for the number of groups is ignored. In contrast, Bayesian models such as the

https://doi.org/10.1002/sta4.324

Dirichlet process (DP; Ferguson, 1973) offer a natural solution that simultaneously estimates the number of groups and the group configurations. However, Miller and Harrison (2013) showed that the Dirichlet process mixture model (DPMM) tends to create tiny extraneous groups. In other words, DPMM does not produce a consistent estimator of the number of groups. In this paper, we employ the mixture of finite mixture (MFM; Miller & Harrison, 2018) approach for learning the group structure of multiple spatial point processes, which, on the contrary, provides consistent estimation for the group numbers.

The contribution of this paper is two-fold. First, we propose a Bayesian group learning method to simultaneously estimate the number of groups and the group configurations. In particular, we use an LGCP to model the spatial pattern of the shot attempts. Based on similarity matrices of the fitted intensity among different players, an MFM model is incorporated for group learning. Moreover, the MFM model has a Pólya urn scheme similar to the Chinese restaurant process, which is exploited to develop an efficient Markov chain Monte Carlo MCMC algorithm without reversible jump or even allocation samplers. Compared with existing approaches (e.g., Jiao et al. 2019), our proposed method does not require prior information for the number of groups, and grouping is incorporated into the structure of the model and performed directly based on the shot selection intensity instead of via *ad hoc* analysis of regression coefficients. In addition, our proposed Bayesian approach reveals interesting shooting patterns of professional basketball players, and the summaries better characterize player types beyond the traditional position categorization.

The rest of the paper is organized as follows. In Section 2, the shot chart data of different players from the 2017–2018 NBA regular season are introduced. In Section 3, we discuss the LGCP and develop the Bayesian group learning method based on MFM. Details of the Bayesian inference are presented in Section 4, including the MCMC algorithm and post MCMC inference methods. Simulation studies are conducted in Section 5. Applications of the proposed methods to NBA players data are reported in Section 6. Section 7 concludes the paper with a discussion.

2 | MOTIVATING DATA

Our data consist of both made and missed field goal attempt locations from the offensive half court of games in the 2017–2018 National Basketball Association (NBA) regular season. The data are available at https://nbasavant.com/index.php. We focus on players who have made more than 400 field goal attempts (FTA). Also, players what just started their careers in the 2017–2018 season, such as Lonzo Ball and Jayson Tatum, are not considered. A total of 191 plays who meet the two criteria above are included in our analysis.

We model a player's shooting location choices as a spatial point pattern on the offensive half court, a 47 ft by 50 ft rectangle, which is the standard size for NBA. We assume the spatial domain $D \in [0, 47] \times [0, 50]$. Indexing the players with $i \in \{1, \dots, 191\}$, the locations of shots, both made and missed, for player i are denoted as $X_i = \{x_{i,1}, \dots, x_{i,T_i}\}$, $\forall x_{i,T_i} \in D$, where T_i is the total number of attempts made by player i on the offensive half court. We select nine players and visualize their shot charts in Figure 1. It can be seen that Clint Capela makes more shot attempts in the painted area, as most of his field goals are slam dunks. JJ Redick, however, prefers to shoot outside the painted area. Our goal is to find groups of similar shooting location habits among the NBA basketball players.

3 | METHOD

3.1 | Log-Gaussian Cox Process

The shot locations can be denoted as $y = (\boldsymbol{s}_1, \dots, \boldsymbol{s}_\ell)$, with $(\boldsymbol{s}_1, \dots, \boldsymbol{s}_\ell)$ being the locations of points that are observed within a bounded region $\mathcal{B} \subseteq \mathcal{R}^2$. Such a spatial point pattern can be regarded as a realization from a spatial point process Y. Spatial point pattern data are modeled by spatial point processes (Diggle, Besag, & Gleaves, 1976) characterized by a quantity called intensity. Within a region $\mathcal{B}$, the intensity on any location $\boldsymbol{s} \in \mathcal{B}$ can be represented as $\lambda(\boldsymbol{s})$ which is defined as:

$$\lambda(\boldsymbol{s}) = \lim_{|\mathrm{d}\boldsymbol{s} \to 0|} \left(\frac{\mathrm{E}[M(\mathrm{d}\boldsymbol{s})]}{|\mathrm{d}\boldsymbol{s}|} \right),$$

where d$\boldsymbol{s}$ is an infinitesimal region around $\boldsymbol{s}$, $|\mathrm{d}\boldsymbol{s}|$ represents its area, and $M(\mathrm{d}\boldsymbol{s})$ shows the number of events that happened over d$\boldsymbol{s}$. For an area $A \subseteq \mathcal{B}$, we denote by $M_{\boldsymbol{Y}}(A) = \sum_{i=1}^{\ell} 1(s_i \in A)$ the counting process associated with the spatial point process $\boldsymbol{Y}$, which counts the number of points in a realization of $\boldsymbol{Y}$ that fall within an area $A \subseteq \mathcal{B}$. The Poisson distribution has been conventionally used for modeling count data, and correspondingly, the spatial Poisson point process is a popular tool for modeling spatial point pattern data. For a Poisson process Y over $\mathcal{B}$, with its intensity function denoted as $\lambda(\boldsymbol{s})$, the counting process $M_Y(A)$ satisfies:

$$M_Y(A) \sim \text{Poisson}\,(\lambda(A)), \text{ where } \lambda(A) = \int_A \lambda(\boldsymbol{s})\mathrm{d}\boldsymbol{s}.$$

For the Poisson process, it is easy to obtain $E(M_{\boldsymbol{Y}}(A)) = \mathrm{Var}(M_{\boldsymbol{Y}}(A)) = \lambda(A)$. When $\lambda(\boldsymbol{s}) = \lambda$, we have constant intensity over the space $\mathcal{B}$ and, in this special case, $\boldsymbol{Y}$ reduces to a homogeneous Poisson process (HPP). For a more general case, $\lambda(\boldsymbol{s})$ can be spatially varying, which leads to a

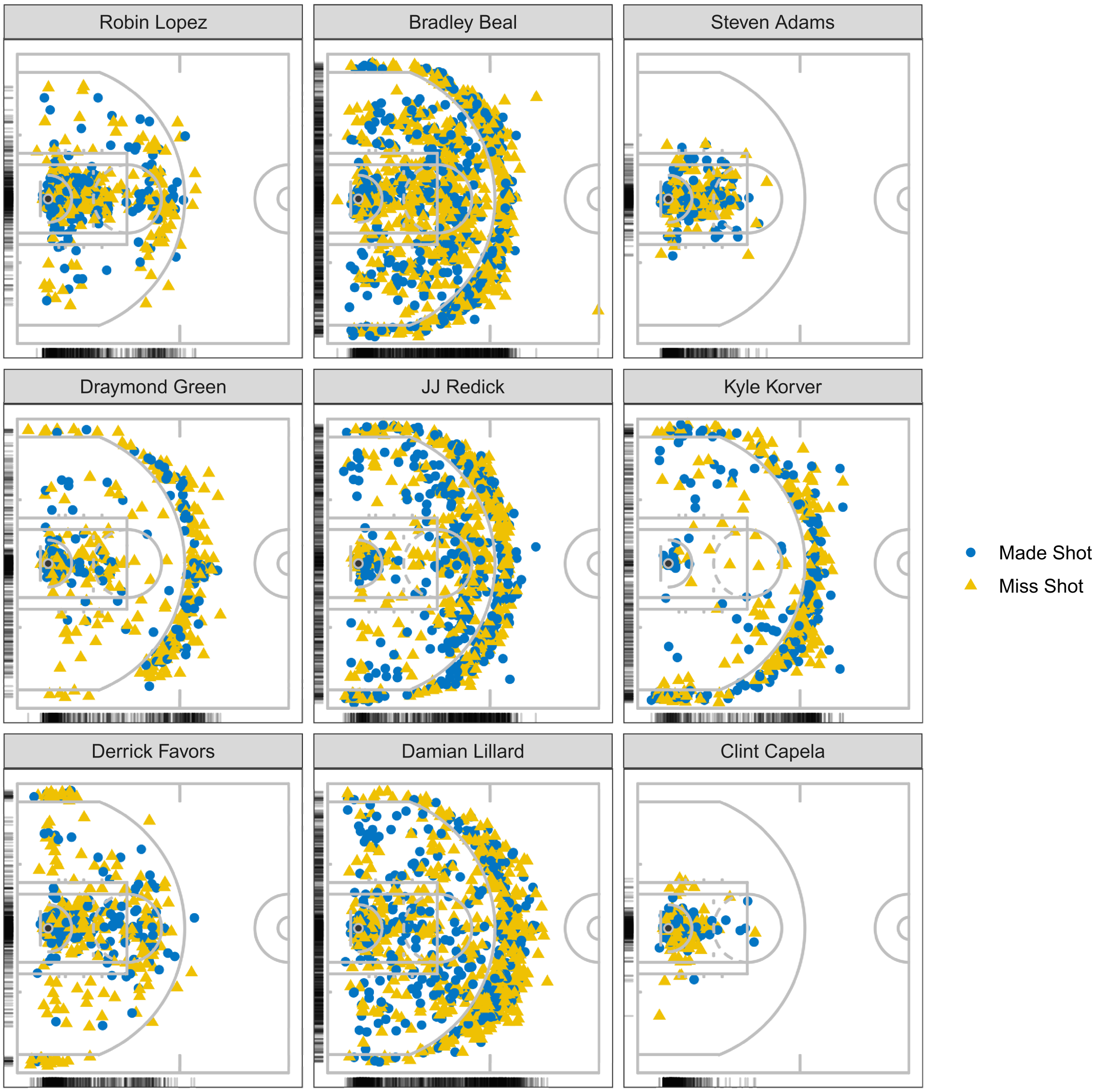


FIGURE 1 Shot charts for selected NBA players

nonhomogeneous Poisson process (NHPP). For the NHPP, the log-likelihood on $\mathcal{B}$ for the observed dataset $\mathbf{y}$ is given by

$$\ell = \sum_{i=1}^{k} \log \lambda(s_i) - \int_{\mathcal{B}} \lambda(s) \mathrm{d}s, \tag{1}$$

where $\lambda(s_i)$ is the intensity function for location s_i. We signify that a set of points $\mathbf{y} = (s_1, s_2, \ldots, s_\ell)$ follows a Poisson process as

$$\mathbf{y} \sim \mathcal{PP}(\lambda(\cdot)). \tag{2}$$

A log-Gaussian Cox process (LGCP) is a doubly-stochastic Poisson process with a spatially varying intensity function modeled as an exponentiated Gaussian process, i.e., Gaussian random field (GRF; Rasmussen & Williams, 2006), which is a spatially continuous random process in which random variables at any location in a space are normally distributed and correlated with random variables at other locations according to a continuous correlation process. The LGCP can be written hierarchically as

$$\begin{aligned} \mathbf{y} &\sim \mathcal{PP}\,(\lambda(\cdot)), \\ \lambda(\cdot) &= \exp\,(Z(\cdot)), \\ Z(\cdot) &\sim \mathcal{GP}\,(0, g(\cdot,\cdot)), \end{aligned} \tag{3}$$

where $g(\cdot,\cdot)$ is the covariance function of the Gaussian process, $Z(\cdot)$. For estimation, the GRF is approximated by the solution to a stochastic partial differential equation (SPDE; see, Lindgren, Rue, & Lindström, 2011, for a review), as SPDEs provide an efficient way of approximating the GRF in a continuous space (Simpson, Illian, Lindgren, Sørbye, & Rue, 2016). Under a purely Bayesian paradigm, model-based Markov chain Monte Carlo (MCMC) can be time-consuming for the LGCP. Therefore, we compute the LGCP using the integrated nested Laplace approximation (INLA; Rue, Martino, & Chopin, 2009), which is an alternative to MCMC for fitting latent Gaussian models, provides a fast and accurate way to fit a potential model, and facilitates computationally efficient inference on point processes. For more details about INLA, we refer the reader to the R-INLA project website at https://www.r&hyphen;inla.org. Details about computation is directed to Section 4.

Our main goal is to group players who share similar shot location choices over the court. After the fitted intensity surfaces for n players are obtained over a grid spanning the court (details in Section 4), they are vectorized and denoted as $\hat{\lambda}^{(1)}(\cdot), \ldots, \hat{\lambda}^{(n)}(\cdot)$. An appropriate metric is needed to quantify similarities among the intensities. Let the matrix $\mathbf{C}$ be that $\mathbf{C} \equiv (\hat{\lambda}^{(1)}, \hat{\lambda}^{(2)}, \ldots, \hat{\lambda}^{(n)})$, i.e., a matrix whose column i is the vectorized intensity surface for player i, and denote $\mathbf{C}^{(i)} = \hat{\lambda}^{(i)}$. Then, following the approach in Cervone, D'Amour, Bornn, and Goldsberry (2016), we compute the players' similarity matrix $\mathbf{H}$ as:

$$\mathrm{H}_{i,j} = \exp\left\{ -\left\| \frac{\mathbf{C}^{(i)}}{\sum \mathbf{C}^{(i)}} - \frac{\mathbf{C}^{(j)}}{\sum \mathbf{C}^{(j)}} \right\| \right\}, \tag{4}$$

where $i, j \in \{1, \ldots, n\}$ and $\|\cdot\|$ is L_2 norm. It can be seen that $\mathbf{H}$ is symmetric and $\mathbf{H} \in \mathcal{R}^{n\times n}$.

3.2 | Group Learning via Point Process Intensity

With the similarity matrix $\mathbf{H}$ obtained, we employ nonparametric Bayesian methods to detect grouped patterns in the intensities. Our initial step is to transform the similarity matrix $\mathbf{H}$ so that each entry $H_{i,j}$ is within the range of a Gaussian distribution. Denote the Fisher transformed distance matrix (Fisher, 1915) as $\mathscr{S}$. Its (i, j)th element is calculated as

$$\mathscr{S}_{ij} = \frac{1}{2}\log\left(\frac{1+\mathrm{H}_{i,j}}{1-\mathrm{H}_{i,j}}\right), \tag{5}$$

A larger value of $\mathscr{S}_{ij}$ indicates higher similarity of intensities. We further assume that

$$\begin{gathered} \mathscr{S}_{ij} | \boldsymbol{\mu}, \boldsymbol{\tau}, k \sim \mathrm{N}(\mu_{ij}, \tau_{ij}^{-1}),\ \mu_{ij} = U_{z_i z_j} \\ \tau_{ij} = T_{z_i z_j},\ 1 \le i < j \le n, \end{gathered} \tag{6}$$

where k denotes the number of groups, N() denotes the normal distribution, $z_i \in \{1, \ldots, k\}$ denotes the group membership of player i for $i = 1, \ldots, 191$. The matrices $\boldsymbol{U} = [U_{rs}] \in (-\infty, +\infty)^{k\times k}$ and $\boldsymbol{T} = [T_{rs}] \in (0, +\infty)^{k\times k}$ are both symmetric, with $U_{rs} = U_{sr}$ indicating the mean closeness between any two fitted intensity surfaces in groups r and s, respectively, and $T_{rs} = T_{sr}$ indicating the precision. Note that the diagonal of $\mathscr{S}$ is infinity as the diagonal of $\mathbf{H}$ is always 1. This, however, does not affect our algorithm, as we do not model the diagonal of $\mathbf{J}$ (see Equation (6), $i < j$).

Denote by $\mathcal{Z}_{n,k} = \{(z_1, \ldots, z_n) : z_i \in \{1, \ldots, k\}, 1 \le i \le n\}$ the set of all possible partitions of n players into k groups. With certain $z \in \mathcal{Z}_{n,k}$, denote by $\mathscr{S}_{[rs]}$ the $n_r \times n_s$ sub-matrix of $\mathscr{S}$ consisting of entries $\mathscr{S}_{ij}$ where $z_i = r$ and $z_j = s$. Under model (6), the joint likelihood of $\mathscr{S}$ can be written as

$$P(\mathscr{S} | \boldsymbol{z}, \boldsymbol{U}, \boldsymbol{T}, k) = \prod_{1 \le r \le s \le k} P(\mathscr{S}_{[rs]} | \boldsymbol{z}, \boldsymbol{U}, \boldsymbol{T}), \tag{7}$$

where

$$P(\mathscr{S}_{[rs]} | \boldsymbol{z}, \boldsymbol{U}, \boldsymbol{T}) = \prod_{1 \le i < j \le n : z_i = r, z_j = s} \frac{1}{\sqrt{2\pi T_{rs}^{-1}}} \exp\left\{ -\frac{T_{rs}(\mathscr{S}_{ij} - U_{rs})^2}{2} \right\}.$$

Assuming that the number of groups k is given, independent prior distributions are often assigned to z, U, and T. Such specification can be conveniently incorporated into a finite mixture model. When k is unknown, however, the Dirichlet process mixture prior models (Antoniak, 1974) can be employe as:

$$\mathscr{S}_i \sim F(., \theta_i),\ \theta_i \sim G(.),\ G \sim DP(\eta G_0), \tag{8}$$

with $\mathscr{S}_i = (\mathscr{S}_{i1}, \mathscr{S}_{i2}, \ldots, \mathscr{S}_{in})$, $\theta_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{in})$, and $\theta_{ij} = (\mu_{ij}, \tau_{ij})$. The process G is parameterized by a base measure G_0 and a concentration parameter η. With θ_i for $i = 1, \ldots, n-1$ drawn from G, a conditional prior distribution for a newly drawn θ_n can be obtained via integration (Blackwell & MacQueen, 1973):

$$p(\theta_n | \theta_1, \ldots, \theta_{n-1}) = \frac{1}{n-1+\eta} \sum_{i=1}^{n-1} \delta_{\theta_i}(\theta_n) + \frac{\eta}{n-1+\eta} \mathrm{G}_0(\theta_n), \tag{9}$$

with $\delta_{\theta_i}(\theta_j) = I(\theta_j = \theta_i)$ being the point mass at θ_i. The model can be equivalently obtained with the introduction of group membership z_i's and having K, the number of groups, approach infinity (Neal, 2000):

$$\begin{aligned} \mathscr{S}_i | z_i, \boldsymbol{\theta}^* &\sim F(\boldsymbol{\theta}^*_{z_i}), \\ z_i | \boldsymbol{\pi} &\sim \text{Discrete}(\pi_1, \ldots, \pi_K), \\ \boldsymbol{\theta}^*_c &\sim G_0, \\ \boldsymbol{\pi} &\sim \text{Dirichlet}(\eta/K, \ldots, \eta/K), \end{aligned} \tag{10}$$

where z_i denotes the latent class associated with $\mathscr{S}_i$, $\boldsymbol{\theta}^*_c$ denotes the vector of parameters that determine the distribution of observations from class c, and $\boldsymbol{\theta}^*$ denotes the collection of all such $\boldsymbol{\theta}^*_z$'s. It can be seen that under this construction, the group-specific distribution $F(\cdot|\boldsymbol{\theta}^*_c)$ solely depends on the vector of parameters $\boldsymbol{\theta}^*_c$.

In construction (10), the prior distribution of $(z_1, \ldots, z_n)$, which would allow for automatic inference on the number of groups k, can be obtained by integrating out $\boldsymbol{\pi}$, the mixing proportions. This is also known as the Chinese restaurant process (CRP; Aldous, 1985; Neal, 2000; Pitman, 1995). The conditional distribution for z_i is defined through the metaphor of a Chinese restaurant (Blackwell & MacQueen, 1973):

$$P(z_i = c | z_1, \ldots, z_{i-1}) \propto \begin{cases} |c|, & \text{at an existing table labeled} \, c \\ \eta, & \text{if } c \text{ is a new table} \end{cases}, \tag{11}$$

where $|c|$ denotes the size of group c. Despite its ability to simultaneously estimate the number of groups and the group configuration, the CRP has been shown by Miller and Harrison (2018) to produce redundant tail groups, causing inconsistency in estimation for the number of groups even with the sample size going to infinity. Miller and Harrison (2018) also proposed a modification of the CRP, known as the mixture of finite mixtures (MFM) model, to mitigate this problem. The MFM model can be formulated as:

$$k \sim p(\cdot), \; (\pi_1, \ldots, \pi_k) | k \sim \text{Dirichlet}(\gamma, \ldots, \gamma), \; z_i | k, \boldsymbol{\pi} \sim \sum_{h=1}^{k} \pi_h \delta_h, \; i = 1, \ldots, n, \tag{12}$$

with $p(\cdot)$ being a proper probability mass function on the set of positive integers and δ_h being a point mass at h. Define a coefficient $V_{n-1}(w)$ as

$$V_{n-1}(w) = \sum_{k=1}^{+\infty} \frac{k_{(w)}}{(\gamma k)^{(n-1)}} p(k),$$

where w denotes the number of "existing tables", $k_{(w)} = k(k-1) \ldots (k-w+1)$, and $(\gamma k)^{(n-1)} = \gamma k(\gamma k + 1) \ldots (\gamma k + n - 2)$, $x^{(0)} = 1$, and $x_{(0)} = 1$. Introduction of a new table is slowed down by $V_{n-1}(w+1)/V_{n-1}(w)$, which yields the following conditional prior of θ:

$$P(\theta_n | \theta_1, \ldots, \theta_{n-1}) \propto \sum_{i=1}^{w} (n_i + \gamma) \delta_{\boldsymbol{\theta}^*_i} + \frac{V_{n-1}(w+1)}{V_{n-1}(w)} \gamma G_0(\theta_n), \tag{13}$$

with $\theta^*_1, \ldots, \theta^*_w$ being the unique values taken by $\theta_1, \ldots, \theta_n$. The conditional distribution for the group membership can be expressed analogous to (11) as:

$$P(z_i = c | z_1, \ldots, z_{i-1}) \propto \begin{cases} |c| + \gamma, & \text{at an existing table labeled} \, c \\ V_n(w+1)/V_n(w)\gamma, & \text{if } c \text{ is a new table} \end{cases}. \tag{14}$$

Adapting MFM to our model setting for functional grouping, the model and prior can be expressed hierarchically as:

$$\begin{aligned} & k \sim p(\cdot), \quad \text{where } \; p(\cdot) \text{is a p.m.f on} \{1, 2, \ldots\} \\ & T_{rs} = T_{sr} \overset{\text{ind}}{\sim} \text{Gamma}\,(\alpha, \beta), \; r, s = 1, \ldots, k, \\ & U_{rs} = U_{sr} \overset{\text{ind}}{\sim} \text{N}(\mu_0, k_0^{-1} T_{rs}^{-1}), \; r, s = 1, \ldots, k, \\ & \text{pr}(z_i = j | \boldsymbol{\pi}, k) = \pi_j, \; j = 1, \ldots, k, \, i = 1, \ldots, n, \\ & \boldsymbol{\pi} | k \sim \text{Dirichlet}\,(\gamma, \ldots, \gamma), \\ & \mathscr{S}_{ij} | \boldsymbol{z}, \boldsymbol{U}, \boldsymbol{T}, k \overset{\text{ind}}{\sim} \text{N}(\mu_{ij}, \tau_{ij}^{-1}), \; \mu_{ij} = U_{z_i z_j}, \tau_{ij} = T_{z_i z_j}, 1 \leq i < j \leq n. \end{aligned} \tag{15}$$

We assume $p(\cdot)$ is a Poisson(1) distribution truncated to be positive through the rest of the paper, which has been proved by Miller and Harrison (2018) and Geng, Bhattacharya, and Pati (2019) to guarantee consistency for the mixing distribution and the number of groups. We refer to the hierarchical model above as MFM-PPGrouping.

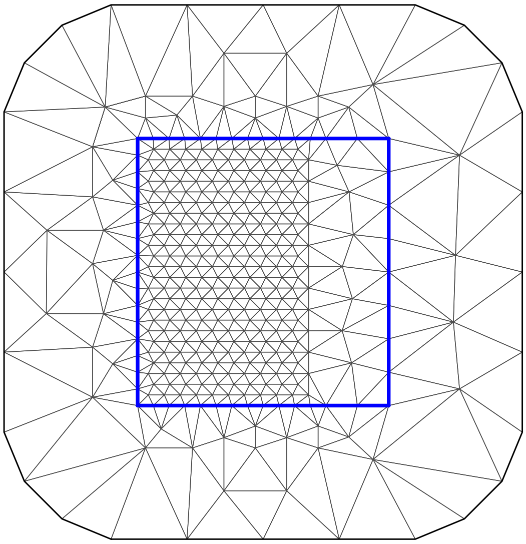


FIGURE 2 Triangulation for the shot data locations over which the "tent" functions are constructed (black line), and the observation locations are inside the (blue) thick-bordered rectangle

4 | BAYESIAN INFERENCE

In this section, we discuss the implementation of INLA estimation for the LGCP, the collapsed sampler algorithm for the proposed MFM-PPGrouping approach, and posterior inference on MCMC samples. INLA tries to partition a region to disjoint triangles (i.e., triangulation) and uses this mesh of discrete sampling locations to estimate a continuous surface in space via interpolation. A set of piecewise linear basis functions, which are typically "tent" or finite element functions, are defined over a triangulation of the domain of interest. The mesh is composed of two regions: the interior mesh, which is where the actions happen; and the exterior mesh, which is designed to alleviate the boundary effects. It is formed by partitioning the region into triangles. The more triangles we have, the more precise our approximation is, at the cost of extended computational time. And the desired mesh would have small triangles where the shot data are dense, and have larger triangles where the shot data are more sparse. Therefore, in our case, the mesh is created to be more dense in the left side of the half court, where most shots are located, as illustrated in Figure 2. A similar mesh can be found in Cervone et al. (2016).

Estimation of the LGCP using INLA is facilitated by the R-package **inlabru** (Bachl, Lindgren, Borchers, & Illian, 2019), which provides an easy access to Bayesian inference for spatial point processes. A benefit of using **inlabru** is that it provides methods for fitting spatial density surfaces, as well as for prediction, while not requiring knowledge of the SPDE theory. With a mesh created as shown in Figure 2, the SPDE can be constructed on the mesh using the function `inla.spde2.pcmatern()`. The "pc" in "pcmatern" is short for "penalized complexity", and it is used to refer to prior distributions over the hyperparameters that are both interpretable and have interesting theoretical properties (see, Simpson, Rue, Riebler, Martins, & Sørbye, 2017, for a discussion).

The LGCP is fitted using the `lgcp()` function in **inlabru** on the mesh. The fitted intensity surface is obtained using the `predict.inla()` function on the pixelized mesh, which generates a 150 × 150 even grid. As we do not consider places off the court, i.e., regions in Figure 2 that are outside the thick-bordered rectangle, this 150 × 150 grid is subsetted to the court only, leaving a 72 × 74 grid. The predicted values obtained on this grid make a 72 × 74 matrix. We represent this matrix using its vectorization, a vector of length 5328, denoted for player i as $\widehat{\lambda}^{(i)}(\cdot)$. With $\widehat{\lambda}^{(1)}(\cdot), \widehat{\lambda}^{(2)}(\cdot), \dots, \widehat{\lambda}^{(n)}(\cdot)$, we obtain $\mathscr{S}$ by (4) and (6). Next, we use MFM-PPGrouping for group learning based on $\mathscr{S}$. The sampler presented in Algorithm 1 (Geng et al. 2019; Hu, Geng, Xue, & Sang, 2020) is used to sample from the posterior distributions for unknown parameters, including k, $z = (z_1, \dots, z_n) \in \{1, \dots, k\}^n$, and $\lambda = (\lambda_1, \dots, \lambda_n)$ in (15). As it marginalizes over the distribution of k, the sampler does not depend on reversible jump and is more efficient than allocation samplers.

After obtaining posterior samples of $\{z_1, z_2, \dots, z_n\}$, posterior inference for the group configurations needs to be carried out so that the values are nominal integers denoting group belongings. This renders the posterior mean unsuitable for our purpose. We adopt Dahl's method (Dahl, 2006). Define a membership matrix $\mathcal{B}^{(\ell)}$ as:

$$\mathcal{B}^{(\ell)} = (\mathcal{B}^{(\ell)}(i,j))_{i,j\in\{1:n\}} = (z_i^{(\ell)} = z_j^{(\ell)})_{n\times n}, \tag{16}$$

where $\ell = 1, \dots, B$ indexes the number of post-burnin MCMC iterations, and $z_i^{(\ell)}$ and $z_j^{(\ell)}$ denote the memberships for players i and j, respectively. An entry $\mathcal{B}^{(\ell)}(i,j)$ equals 1 if $z_i^{(\ell)} = z_j^{(\ell)}$, and 0 otherwise. An element-wise mean of the membership matrices can be obtained as

$$\overline{\mathcal{B}} = \frac{1}{B}\sum_{t=1}^{B}\mathcal{B}^{(t)},$$

where the summation is also element-wise. The posterior iteration with the smallest squared distance to $\overline{\mathcal{B}}$ is obtained by

$$C_{LS} = \text{argmin}_{\, c\in(1:B)} \sum_{i=1}^{n}\sum_{j=1}^{n}(\mathcal{B}^{(c)}(i,j) - \overline{\mathcal{B}}(i,j))^2. \tag{17}$$

The estimated parameters, together with the group assignments $\boldsymbol{z}$, are obtained from the C_{LS}th post burn-in iteration. With the Dahl's method, our Bayesian grouping method is summarized in Algorithm 2.

Algorithm 1 Collapsed sampler for MFM-PPGrouping

```
procedure C-MFM-PPGROUPING
Initialize: let $\boldsymbol{z} = (z_1, \ldots, z_n)$, $\boldsymbol{U} = (U_{rs})$, $\boldsymbol{T} = (T_{rs})$.
    for each iter = 1 to M do
Update $\boldsymbol{T} = (T_{rs})$ conditional on $z$ as
    $p(T_{rs} \mid \mathscr{S}, z) \sim \text{Gamma}\left(\alpha + n_{rs}/2, \beta + (n_{rs}-1)\text{var}(A_{[rs]})/2 + \frac{k_0 n_{rs}(\bar{A}_{[rs]} - \mu_0)^2}{2(k_0 + n_{rs})}\right)$
Update $\boldsymbol{U} = (U_{rs})$ conditional on $z$ as
    $p(U_{rs} \mid \mathscr{S}, T_{rs}, z) \sim \mathrm{N}\left(\frac{k_0\mu_0 + n_{rs}\bar{A}_{[rs]}}{k_0 + n_{rs}}, ((k_0 + n_{rs})T_{rs})^{-1}\right)$
    where $A_{[rs]} = (\mathscr{S}_{ij}; z_i = r, z_j = s, i \neq j)$, $\bar{A}_{[rs]} = \left(\sum_{z_i=r,z_j=s,i\neq j}\mathscr{S}_{ij}\right)/n_{rs}$ and $n_{rs} = \sum_{i\neq j} I(z_i = r, z_j = s), r = 1, \ldots, k; s = 1, \ldots, k$. Note that $k$ denotes the number of groups yielded by the current $z$.
Update $\boldsymbol{z} = (z_1, \ldots, z_n)$ conditional on $\boldsymbol{U} = (U_{rs})$ and $\boldsymbol{T} = (T_{rs})$. For each $i$ in $(1, \ldots, n)$, $P(z_i = c \mid z_{-i}, \mathscr{S}, U, T)$ can be obtained in closed form as
    $\propto \begin{cases} [|c| + \gamma]\left[\prod_{j>i} \frac{1}{\sqrt{2\pi T_{cz_j}^{-1}}} e^{-\frac{T_{cz_j}(\mathscr{S}_{ij} - U_{cz_j})^2}{2}}\right]\left[\prod_{k<i} \frac{1}{\sqrt{2\pi T_{z_k c}^{-1}}} e^{-\frac{T_{z_k c}(\mathscr{S}_{ki} - U_{z_k c})^2}{2}}\right] & \text{at an existing table } c \\ \frac{V_n(|\mathcal{C}_{-i}|+1)}{V_n(|\mathcal{C}_{-i}|}\gamma m(\mathscr{S}_i) & \text{if } c \text{ is a new table} \end{cases},$
    with $\mathcal{C}_{-i}$ being the partition obtained by removing $z_i$ and
    $m(\mathscr{S}_i) = \prod_{t=1}^{|\mathcal{C}_{-i}|} \frac{\Gamma(\alpha_n)}{\Gamma(\alpha)} \frac{\beta^{\alpha}}{\beta_n^{\alpha_n}} \left(\frac{k_0}{k_0 + n_t}\right)^{\frac{1}{2}} (2\pi)^{\frac{-n_t}{2}},$
    where $\alpha_n = \alpha + n_t/2, n_t = \sum_{i\neq j} I(z_j = t), \beta_n = \beta + (n_t - 1)\text{var}(A_{[t]})/2 + \frac{k_0 n_t(\bar{A}_{[t]} - \mu_0)^2}{2(k_0+n_t)}, A_{[t]} = (\mathscr{S}_{ij}; z_j = t, i \neq j)$ and $\bar{A}_{[t]} = \left(\sum_{z_j=t,i\neq j}\mathscr{S}_{ij}\right)/n_t$.
    end for
end procedure
```

Algorithm 2 Bayesian group learning procedure for basketball players

```
Fit LGCPs for $n$ different players $\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}, \ldots, \boldsymbol{y}^{(n)}$ via inlabru and get $n$ underlying intensity surface $\hat{\lambda}^{(1)}(\cdot), \hat{\lambda}^{(2)}(\cdot), \ldots, \hat{\lambda}^{(n)}(\cdot)$,
Use (4) and (5) to construct matrix $\boldsymbol{S}$ and matrix $\mathscr{S}$ and based on $\hat{\lambda}^{(1)}(\cdot), \hat{\lambda}^{(2)}(\cdot), \ldots, \hat{\lambda}^{(n)}(\cdot)$,
Get $B$ posterior samples of $\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}, \ldots, \boldsymbol{z}^{(B)}$ from $\mathscr{S}$ via Algorithm 1,
Summary posterior samples by Dahl's method.
```

5 | SIMULATION

5.1 | Simulation Setup

A total of three groups are designed, each of which has its own base intensity as shown in Figure 3. The first group corresponds to players most of whose shots are in the painted area; the second group corresponds to players whose shot locations are widely distributed in every location from the painted area to the three-point line. A player in the third group has more shots at the three-point line and inside the painted area. To create some variation between players within the same group so that their shot charts are not generated from exactly the same intensity surface, a noise term has been added to the base surfaces, so that for player i in group k,

$$\lambda_{k,i}(\cdot) = \lambda_k(\cdot) + |\epsilon_{k,i}|, \; k = 1, 2, 3, \; i = 1, \; \ldots \;, 25, \tag{18}$$

where $\lambda_k(\cdot)$ denote the three base intensity surfaces in Figure 3, $\epsilon_{k,i}$ is generated from a multivariate normal distribution with mean $\mathbf{0}$ and variance $0.5\mathbf{I}$, and an absolute value step is taken to ensure that the summation produces a positive and valid intensity surface.

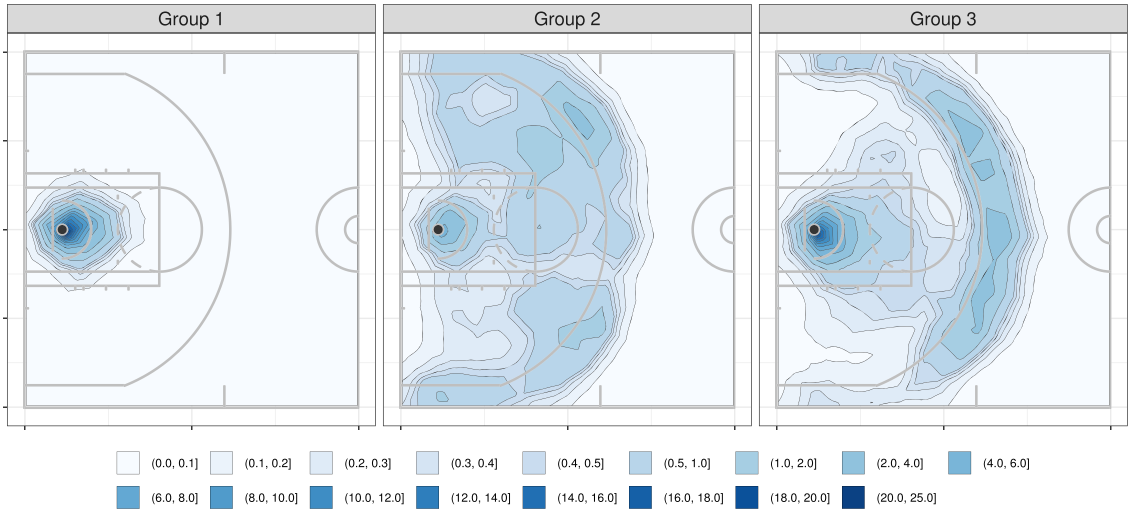
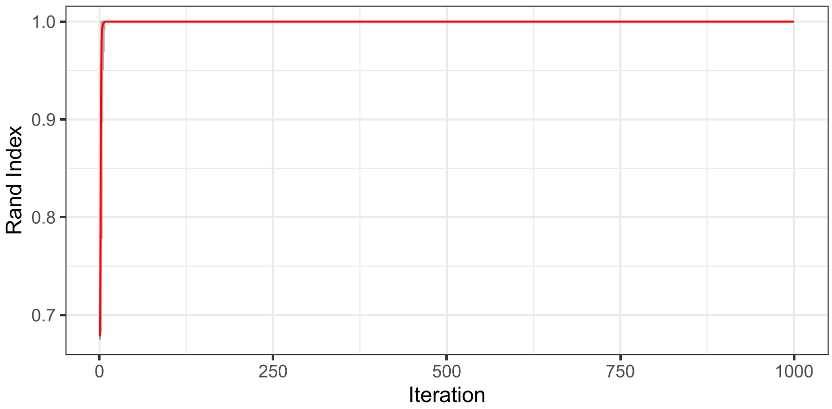


FIGURE 3 Visualization of three bases for the three groups in simulation design

FIGURE 4 Rand Index trace plot for single replicate simulated data. Dark grey line indicates for each random seed. Red line is the average rand index for 50 random seeds



With 75 valid intensity surfaces having both between group variation and within group variation, shot locations for the 75 players are generated following the Poisson process as in Equation (2). Algorithm 2 is implemented on these 75 players, and we examine both the estimation for the number of groups as well as congruence of group belongings with the true setting in terms of modulo labeling by the Rand index (RI; Rand, 1971), the computation of which is facilitated by the R-package **fossil** (Vavrek, 2011). The RI ranges from 0 to 1 with a higher value indicating better agreement between a grouping scheme and the true setting. In particular, a value of 1 indicates perfect agreement.

5.2 | Simulation Results

We run our algorithm with 1,000 MCMC iterations, with the first 500 iterations as burn-in for each replicate data. We examine it is sufficient for the chain to converge and stabilize. The numbers are chosen to be sufficiently large for the chain to converge and stabilize. To verify this, with a single replicate of data, 50 separate MCMC chains are run with different random seeds and hence initial values, and 50 final grouping schemes are obtained. The RI is calculated for these 50 chains at each iteration, giving 50 traces, which are visualized in Figure 4. It can be observed that covergence is attained after a small number of iterations, and the band of the 50 traces is rather tight after convergence.

Proceeding to 50 separate replicates of data, our proposed algorithm was run, and 50 RI values were obtained by comparing with the true setting. They average to 0.9988, which indicates rather accurate grouping ability of the proposed approach. In addition, performance comparisons of our proposed method with three competing methods are made. We compare our method to the K-means algorithm, Density-based spatial grouping of applications with noise (DBSCAN), and mean shift grouping. Grouping recovery performances of all the four methods are measured using the RI. The 50 final RI's obtained for the three competitors average to 0.9005, 0.7642, and 0.7380, respectively, indicating the superior performance of our proposed approach. We also show the number of clusters covered by our algorithm; see Figure 5. We see that there are fourty two replicates having three clusters.

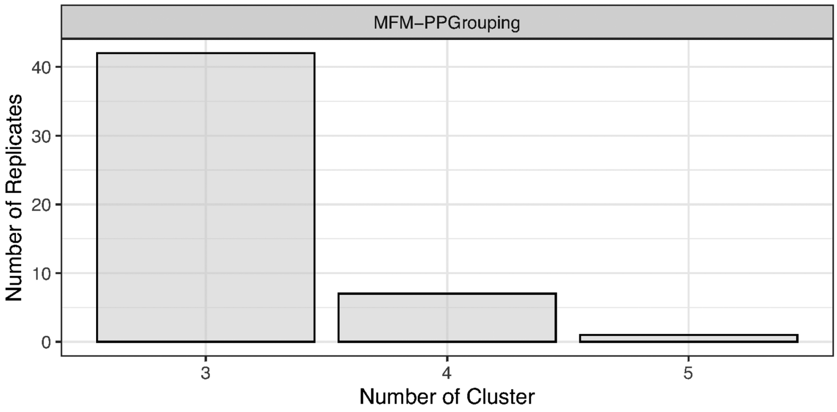


FIGURE 5 Histogram for the number of clusters produced by MFM-PPGrouping in 50 simulation replicates

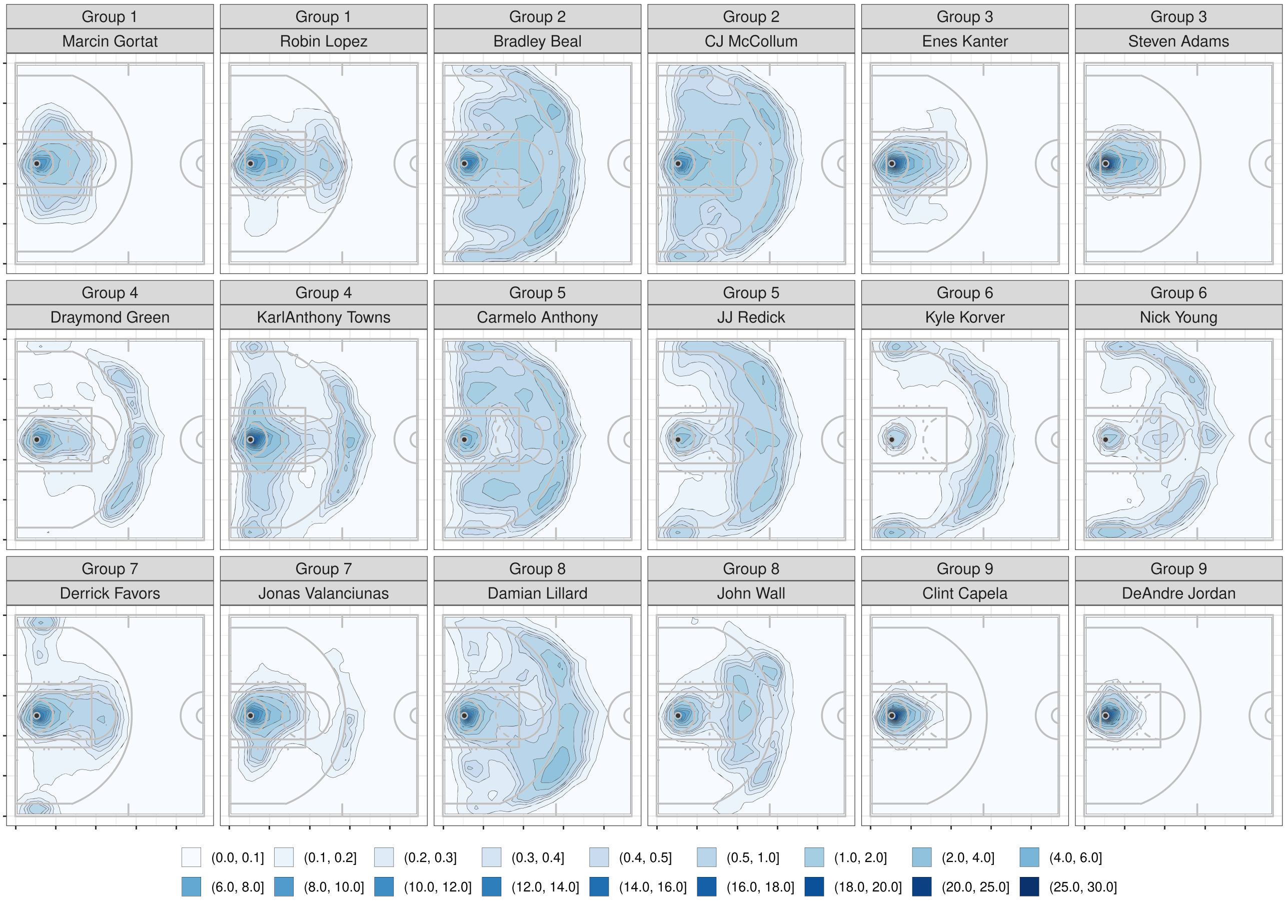


FIGURE 6 Fitted intensities with contour lines for two selected players in each of the nine identified groups

6 | ANALYSIS OF NBA PLAYERS

In this section, we apply the proposed method to the analysis of players' shot data in the 2017-2018 NBA regular season. Only the locations of shots are considered regardless of the players' positions on the court (e.g., point guard, power forward, etc.). As a starting point, a predictive intensity matrix is obtained for each player using **inlabru**. Algorithms 1 and 2 are subsequently used to identify the groups. We run 1,000 MCMC iterations and the first 500 iterations as the burn-in period. The result from the MFM model suggests that the 191 players are classified into nine groups. The sizes of the nine groups are 10, 59, 8, 19, 24, 8, 10, 51, and 2 respectively. Visualizations of intensity matrices with contour for two selected players from each group are presented in Figure 6.

Several interesting observations can be made from the visualization results. First, we discuss groups 1, 3, and 9. The contours for group 1 and group 3 are wider than the contours for group 9, where most shots are located near the hoop. Clint Capela and DeAndre Jordan (in group 9), for example, are both good at making alley-oops and slam dunks. Only very few shots are made by these players outside the painted area. Despite

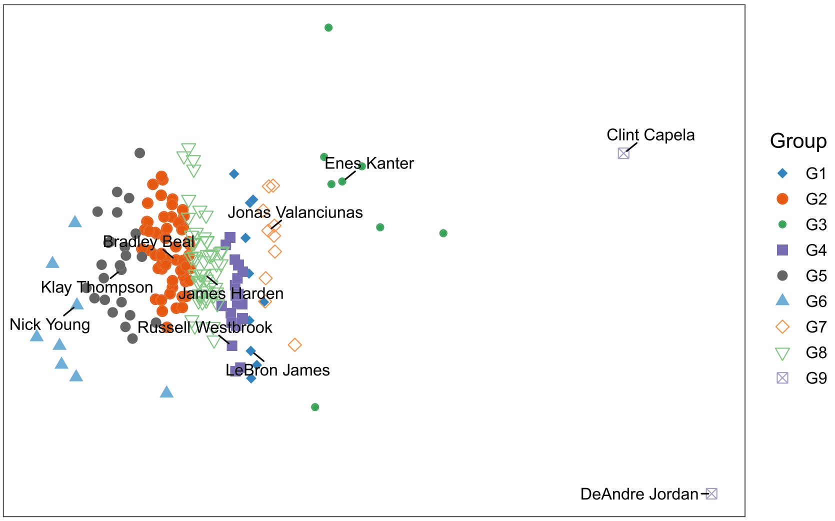


FIGURE 7 Visualization of the night identified groups of players, with selected players' names annotated

the similarity between groups 1 and 3, it can be seen that for group 3, most of the shots are made within the painted area, while there are quite a number of shots outside the painted area for group 1, indicating wider shooting ranges for the corresponding players.

Groups 4 and 7 share some characteristic in common as most of the shot locations are around the hoop. Players in group 4, however, are also able to shoot frequently beyond the three-point line at a wider range of angles, while players in group 7 also shoot beyond the three-point line at very limited angles.

Groups 2, 5, and 8 also bear some resemblance with each other. A first look at the fitted intensity contours indicates that players in these groups are able to make all types of shots, including three-pointers, perimeter shots, and also shots over the painted area. Group 2 is differentiated from the other two as most of the shots are made around the hoop, and the intensities for perimeter shots and three-pointers are rather similar. For group 5, however, the intensities for shots around the hoop is much lower than that for group 2. For group 8, players make most shots around the hoop and also make some perimeter shots as well as three-pointers. Compared with the other two groups, they have a more narrow range of angles to make perimeter shots and three-pointers. Most shots are located between 45 degrees up and down angle from the horizontal line across the hoop.

Group 6 has little similarity with any other groups. As can be seen, most shots are located either near the hoop, or beyond the three-point line. There are very few perimeter shots. Kyle Korver and Nick Young, both of whom are well-known catch-and-release shooters, fall in this group.

As further verification, we use multidimensional scaling to lower the dimension of the fitted intensity matrices for players to 2 so that similarities in their shooting habits can be visualized. See Figure 7. Separation of the nine groups is quite clear. Group 9, for example, with its unique strong preference for alley-oops and slam dunks, stands far from others.

Finally, to make sure the group configuration presented here is not a random occurrence but reflects the true pattern demonstrated by the data, we run 50 separate MCMC chains with different random seeds and initial values and obtained 50 final grouping schemes. The RI between each scheme and the present grouping scheme is calculated, and they average to 0.948, indicating high concordance of conclusion regardless of the random seeds.

7 | CONCLUSION

In this paper, we proposed using MFM to capture heterogeneity of different NBA players based on the LGCP. Our group learning method provides a quantitative summary of different players shot habits other than traditional position categorization. Our simulation results indicated that our proposed methods achieve good grouping accuracy.

The real data application gives us the information about player's shooting habit locations. Players can understand their own shooting habits, and they can also strengthen their weaker shooting locations. On the other hand, the professional coach can formulate a defensive strategy to reduce the opponent's score with these information. Our grouping results will provide a good guidance for team managers trading the players with similar shot patterns.

A few topics beyond the scope of this paper are worth further investigation. In this paper, a two-stage group learning method is proposed. A unified approach is an interesting alternative in future work. In addition, incorporating auxiliary information such as player position, or historical information could also be taken into account for grouping in our future work. Jointly modeling spatial field goal percentage and shot selection will provide more detail instructions for professional coaches.

ORCID

Guanyu Hu https://orcid.org/0000-0003-1410-1665
Hou-Cheng Yang https://orcid.org/0000-0002-8679-4280
Yishu Xue https://orcid.org/0000-0002-9660-6087

REFERENCES

Aldous, D. J. (1985). Exchangeability and related topics, *École d'été de Probabilités de Saint-Flour XIII-1983*. Berlin, Heidelberg: Springer, pp. 1–198.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, *2*(6), 1152–1174.

Bachl, F. E., Lindgren, F., Borchers, D. L., & Illian, J. B. (2019). inlabru: An R package for Bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution*, *10*(6), 760–766.

Blackwell, D., & MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, *1*(2), 353–355.

Cervone, D., D'Amour, A., Bornn, L., & Goldsberry, K. (2016). A multiresolution stochastic process model for predicting basketball possession outcomes. *Journal of the American Statistical Association*, *111*(514), 585–599.

Chen, Y., Pan, R., Guan, R., & Wang, H. (2019). A case study for Beijing point of interest data using group linked Cox process. *Statistics and Its Interface*, *12*(2), 331–344.

Dahl, D. B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. In Kim-Anh Do, M. V. (Ed.), *Bayesian inference for gene expression and proteomics*, Vol. *4*. Cambridge, United Kingdom: Cambridge University Press, pp. 201–218.

DeGroot, M. H. (2005). *Optimal statistical decisions*, Vol. *82*: John Wiley & Sons.

Diggle, P. J., Besag, J., & Gleaves, J. T. (1976). Statistical analysis of spatial point patterns by means of distance methods. *Biometrics*, *3*(32), 659–667.

Diggle, P. J., Gómez-Rubio, V., Brown, P. E., Chetwynd, A. G., & Gooding, S. (2007). Second-order analysis of inhomogeneous spatial point processes using case–control data. *Biometrics*, *63*(2), 550–557.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, *1*(2), 209–230.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, *10*(4), 507–521.

Franks, A., Miller, A., Bornn, L., & Goldsberry, K. (2015). Characterizing the spatial structure of defensive skill in professional basketball. *The Annals of Applied Statistics*, *9*(1), 94–121.

Geng, J., Bhattacharya, A., & Pati, D. (2019). Probabilistic community detection with unknown number of communities. *Journal of the American Statistical Association*, *114*(526), 893–905.

Geng, J., Shi, W., & Hu, G. (2019). Bayesian nonparametric nonhomogeneous Poisson process with applications to USGS earthquake data. arXiv preprint arXiv:1907.03186.

Geyer, C. (1998). Likelihood inference for spatial point processes. In Kendall, W. S. (Ed.), *Stochastic Geometry: Likelihood and Computation*, Vol. *80*. London, UK: CRC Press, pp. 79.

Goulard, M., Särkkä, A., & Grabarnik, P. (1996). Parameter estimation for marked Gibbs point processes through the maximum pseudo-likelihood method. *Scandinavian Journal of Statistics*, *23*(3), 365–379.

Guan, Y. (2008). On consistent nonparametric intensity estimation for inhomogeneous spatial point processes. *Journal of the American Statistical Association*, *103*(483), 1238–1247.

Hu, G., Geng, J., Xue, Y., & Sang, H. (2020). Bayesian spatial homogeneity pursuit of functional data: An application to the us income distribution. arXiv preprint arXiv:2002.06663.

Hu, G., Huffer, F., & Chen, M.-H. (2019). New development of Bayesian variable selection criteria for spatial point process with applications. arXiv preprint arXiv:1910.06870.

Jiao, J., Hu, G., & Yan, J. (2019). A Bayesian joint model for spatial point processes with application to basketball shot chart. arXiv preprint arXiv:1908.05745.

Jiao, J., Hu, G., & Yan, J. (2020). Heterogeneity pursuit for spatial point pattern with application to tree locations: A Bayesian semiparametric recourse. arXiv preprint arXiv:2003.10043.

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(4), 423–498.

Miller, A., Bornn, L., Adams, R., & Goldsberry, K. (2014). Factorized point process intensities: A spatial analysis of professional basketball. In *Proceedings of the 31st International Conference on Machine Learning* (Xing, E. P., & Jebara, T., Eds.), Proceedings of Machine Learning Research, *32*, PMLR, Bejing, China, pp. 235–243.

Miller, J. W., & Harrison, M. T. (2013). A simple example of Dirichlet process mixture inconsistency for the number of components. In *Advances in Neural Information Processing Systems 26* (Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., & Weinberger, K. Q., Eds.), Red Hook, USA: Curran Associates, Inc., pp. 199–206.

Miller, J. W., & Harrison, M. T. (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, *113*(521), 340–356.

Møller, J., Syversveen, A. R., & Waagepetersen, R. P. (1998). Log Gaussian cox processes. *Scandinavian Journal of Statistics*, *25*(3), 451–482.

Murphy, K. P. (2007). Conjugate bayesian analysis of the gaussian distribution. def 1(2σ2), 16.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, *9*(2), 249–265.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, *102*(2), 145–158.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*(336), 846–850.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press.

Reich, B. J., Hodges, J. S., Carlin, B. P., & Reich, A. M. (2006). A spatial analysis of basketball shot chart data. *The American Statistician*, *60*(1), 3–12.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(2), 319–392.

Simpson, D., Illian, J. B., Lindgren, F., Sørbye, S. H., & Rue, H. (2016). Going off grid: Computationally efficient inference for log-Gaussian Cox processes. *Biometrika*, *103*(1), 49–70.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., & Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, *32*(1), 1–28.

Thurman, A. L., Fu, R., Guan, Y., & Zhu, J. (2015). Regularized estimating equations for model selection of clustered spatial point processes. *Statistica Sinica*, *25*(1), 173–188.

Vavrek, M. J. (2011). Fossil: Palaeoecological and palaeogeographical analysis tools. *Palaeontologia Electronica*, *14*(1), 1T. R package version 0.4.0.

Veen, A., & Schoenberg, F. P. (2006). Assessing spatial point process models using weighted K-functions: Analysis of California earthquakes. In Baddeley, A., Gregori, P., Mateu, J., Stoica, R., & Stoyan, D. (Eds.), *Case Studies in Spatial Point Process Modeling*. New York, NY: Springer New York, pp. 293–306.

How to cite this article: Hu G, Yang H-C, Xue Y. Bayesian group learning for shot selection of professional basketball players. *Stat*. 2021;10:e324. https://doi.org/10.1002/sta4.324

APPENDIX A: DERIVATION OF $m(\mathscr{S}_i)$

We start with

$$P\left(z_i = c | z_{-i}, \mathscr{S}, U, T\right) \propto \frac{V_n(|\mathcal{C}_{-i}|+1)}{V_n(|\mathcal{C}_{-i}|}\gamma m(\mathscr{S}_i),$$

where $m(\mathscr{S}_i)$ is the marginal likelihood of terms in matrix. $\mathscr{S}$ is related to i. Since $\mathscr{S}$ is symmetric, it is sufficient enough to consider $\mathscr{S}_{i1} \dots \mathscr{S}_{in}$

$$\begin{aligned} m(\mathscr{S}_i) &= \prod_{t=1}^{|\mathcal{C}_{-i}|} m(\mathscr{S}_{i[t]}), \\ \mathscr{S}_{i[t]} &= \{\mathscr{S}_{ij}; z_j = t,\ i \neq j\} \\ m(\mathscr{S}_{i[t]}) &= \int \prod_{j \in [t]} N(\mathscr{S}_{ij} | U_{z_i t}, T_{z_i t}) N(U_{z_i t} | \mu_0, k_0^{-1}, T_{z_i t}^{-1}) \mathrm{Gamma}(T_{z_i t} | \alpha, \beta) dU_{z_i t} dT_{z_i t} \\ &= \frac{\Gamma(\alpha_n)}{\Gamma(\alpha)} \frac{\beta^\alpha}{\beta_n^{\alpha_n}} \left(\frac{k_0}{k_0 + n_t}\right)^{\frac{1}{2}} (2\pi)^{\frac{-n_t}{2}} \end{aligned} \tag{A1}$$

where $\alpha_n = \alpha + n_t/2$, $n_t = \sum_{i \neq j} I(z_j = t)$, $\beta_n = \beta + (n_t - 1)\mathrm{var}(A_{[t]})/2 + \frac{k_0 n_t (\bar{A}_{[t]} - \mu_0)^2}{2(k_0 + n_t)}$, $A_{[t]} = (\mathscr{S}_{ij}; z_j = t, i \neq j)$ and $\bar{A}_{[t]} = \left(\sum_{z_j = t, i \neq j} \mathscr{S}_{ij}\right)/n_t$. The integral is a straightforward deviation from Normal data with Normal-Gamma prior (See, for example, DeGroot, 2005; Murphy, 2007).