Zero-inflated Poisson model with clustered regression coefficients: Application to heterogeneity learning of field goal attempts of professional basketball players

Guanyu HU\textsuperscript{1}, Hou-Cheng YANG\textsuperscript{2}, Yishu XUE\textsuperscript{3}*, and Dipak K. DEY\textsuperscript{3}

\textsuperscript{1}Department of Statistics, University of Missouri Columbia, Columbia, MO 65211, U.S.A.\textsuperscript{2}Department of Statistics, Florida State University, Tallahassee, FL 32306, U.S.A.\textsuperscript{3}Department of Statistics, University of Connecticut, Storrs, CT 06269, U.S.A.

Key words and phrases: Bayesian nonparametric; MCMC; mixture of finite mixtures; model-based clustering.

MSC 2020: Primary 62H30; secondary 62P99.

Abstract: Although basketball is a dynamic process sport, played between two sides of five players each, learning some static information is essential for professional players, coaches, and team managers. In order to have a deep understanding of field goal attempts among different players, we propose a zero-inflated Poisson model with clustered regression coefficients to learn the shooting habits of different players over the court and the heterogeneity among them. Specifically, the zero-inflated model captures a large portion of the court with zero field goal attempts, and the mixture of finite mixtures model captures the heterogeneity among different players based on clustered regression coefficients and inflated probabilities. Both theoretical and empirical justification through simulation studies validate our proposed method. We apply our proposed model to data from the National Basketball Association (NBA), for learning players’ shooting habits and heterogeneity among different players over the 2017–2018 regular season. This illustrates our model as a way of providing insights from different aspects.

Résumé: Bien que le basket-ball soit un sport de processus dynamique, joué entre deux équipes de cinq joueurs chacune, l’apprentissage de certaines informations statiques revêt un grand intérêt pour les joueurs professionnels, entraîneurs et chefs d’équipe. Dans le but de bien comprendre les tentatives de tir au panier (field goal) entre les différents joueurs, nous proposons un modèle de régression de Poisson avec surreprésentation de zéros et à coefficients en groupes. Cette approche permet de mieux cerner les habitudes de tir des différents joueurs sur le terrain et l’hétérogénéité parmi eux. Plus précisément, le modèle à inflation de zéro capturera la grande proportion du terrain avec zéro tentative de tir au panier, alors que le modèle de mélanges finis capturera l’hétérogénéité entre les différents joueurs et ce grâce aux coefficients de régression en groupes et aux probabilités gonflées. En plus d’établir quelques résultats théoriques, nous montrons, à travers des études de simulations, que notre approche produit des résultats concluants. Sa mise en application sur des données recueillies au cours de la saison régulière de 2017–2018, nous a permis d’explorer divers aspects des habitudes de tir et d’examiner l’hétérogénéité au sein des joueurs de la National Basketball Association (NBA). La revue canadienne de statistique 51: 157–172; 2023 © 2022 Société statistique du Canada
1. INTRODUCTION

Analyzing players’ “hotspots,” that is, locations from 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, shown in Figure 1, but it is rather rough, as there is no clear-cut way of defining “similarity,” which calls for the need for more rigorous statistical modelling.

Various tools have been proposed to model point patterns. Among them, spatial point processes constitute a family of models that assumes 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 being the Poisson process (Geyer, 1998), the Gibbs process (Goulard, Särkkä & Grabarnik, 1996), and the log-Gaussian Cox process (LGCP; Möller, Syversveen & Waagepetersen, 1998). They also have a wide range of applications, including in ecological studies, environmental sciences (Hu, Huffer & Chen, 2019; Jiao, Hu & Yan, 2021b), and sports analytics. 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 behaviour of individual players. Determining which players are similar to each other remains unanswered by these works.

![Figure 1: Shot charts for selected NBA players.](image-url)
Toward this end, Jiao Hu & Yan (2021a) proposed a marked point process joint modelling approach that takes into account both shot locations and outcomes. The shot locations are modelled with a nonhomogeneous Poisson point process (NHPP), and the intensity is used as a covariate to explain the mark, i.e., the success rate of shots in the joint model. Clustering is performed only on the fitted parameters of the NHPP on an ad hoc basis to identify similarities among players. As such, clustering is heuristic. In pursuit of more statistical rigour, in a study of tree locations, Jiao, Hu & Yan (2021b) proposed a model-based clustering approach that associates each observation with a cluster label and uses the Chinese restaurant process (CRP; Ferguson, 1973) to model the joint distribution of these labels. The number of clusters is readily inferred from the number of unique latent cluster labels. Yin et al. (2020) improved the model of Jiao, Hu & Yan (2021b) when modelling shooting hotspots on the basketball court after partitioning it into many small tiles by using a Markov random fields constraint Dirichlet process for the latent cluster belongings, which effectively encourages local spatial homogeneity. With such a constraint, the identified hotspots are often larger regions instead of individual, scattered tiles, and it is easier for basketball coaches and data analysts to segment the court. This article, however, considers only the similarity of intensities on tiles of a court, but does not account for similarities between players. To determine which players have similar shooting habits, Hu, Yang & Xue (2021) used LGCP 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 a mixture of finite mixtures (MFM; Miller & Harrison, 2018) to perform clustering. Noting that Hu, Yang & Xue (2021) represented the fitted intensity as a vector, which caused loss of information on the spatial domain, Yin, Hu & Shen (2020) proposed a Bayesian nonparametric matrix clustering approach, which preserves spatial contiguity of tiles, to analyze the latent heterogeneity structure of the estimated intensity surfaces. Note that in all five methods, the intensity function plays a certain role, which adds another layer of modelling between the shots and the grouping structure.

One natural way to model the counts directly without employing the intensity surface is Poisson regression. Zhao et al. (2020) proposed a spatial homogeneity pursuit regression model for count value data, where the clustering of locations is done via imposing certain spatial contiguity constraints on MFM. However, the analysis of data of basketball shots poses more challenges. The first challenge comes from the fact that only a 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. Second, existing approaches are subject to different limitations. One approach aims to find areas on the court where the shooting intensities are close and group them together to identify a few large segments on the court that correspond to hotspots and coldspots, but does not consider similarities between players (Yin et al., 2020; Zhao et al., 2020). The other existing approach finds groups of players whose shot intensity surfaces are similar, but these intensity surfaces can be rather “ragged,” and cannot be constructed as a combination of step functions on a few large segments of the court (Hu, Yang & Xue, 2021). Thirdly, under the frequentist paradigm, clustering is often performed as an ad hoc procedure after certain parameter estimation is done. Such clustering is often based on heuristic comparisons, and there is no guarantee of favourable theoretical properties such as consistent estimation for the number of clusters and clustering configurations. This indicates the need for a model-based, coherent approach that incorporates clustering as a built-in component.

To tackle the above three challenges, we propose a Bayesian zero-inflated Poisson (ZIP) regression approach to model field goal attempts of players with different shooting habits. The contributions of this article are threefold. First, the large portion of the court with zero shot attempts is accommodated in the model structure by zero inflation. Next, nonnegative matrix factorization is used to decompose the shooting habits of players into linear combinations of several basis functions, which naturally handles the homogeneity pursuit on the spatial

DOI: 10.1002/cjs.11684
dimension. On the dimension across players, we introduce an MFM prior to the ZIP model to jointly estimate the 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.

The rest of the article is organized as follows. In Section 2, we introduce the motivating data from the 2017 to 2018 regular season. In Section 3, we first review ZIP regression, and then propose our Bayesian clustering method based on MFM. Details of 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 player data are reported in Section 6. Section 7 concludes the article with a discussion.

2. 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 who have made 400 or more field goal attempts. Also, owing to unavailability of data for rookie players, such as Lonzo Ball and Jayson Tatum whose careers started in the 2017–2018 season, a total of 191 players are included in our analysis.

We model each player’s shooting location choices and outcomes as a spatial point pattern on the offensive half court, a 47 ft × 50 ft rectangle, which is the standard size for the NBA. The spatial domain for the basketball court is denoted as \( D \in [0,47] \times [0,50] \). We partition the court into 1 ft × 2 ft blocks, which means that there are in total \( 47 \times 25 = 1175 \) blocks in the basketball court. The shot charts for five selected players are visualized in Figure 1. The numbers of shot attempts in each of the blocks are counted. Hence, these data consist of nonnegative, highly skewed sequence counts with a large proportion of zeros, as most shots are made in the region 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 zeros motivates the use of zero-inflated models for such types of data. Here we define \( y = (y_1,y_2,\ldots,y_n) \), where \( y_i = (y_{i1},y_{i2},\ldots,y_{iJ})^\top \) for \( i = 1,2,\ldots,191 \), and \( J = 1175 \). Each \( y_{ij} \) for \( i = 1,\ldots,191 \) and \( j = 1,\ldots,J \) represents the total number of shots made by the \( i \)th player in the \( j \)th block. For selected players, we counted the number of blocks that have no shot, one shot, two shots, \( \ldots \), and six shots or more as presented in Table 1. Note that Clint Capela has the most number of blocks corresponding to zero shots, which is straightforward as he is a centre and barely shoots from outside the painted area. LeBron James, on the other hand, has the fewest blocks with no shots, which indicates his wide shooting range. Except for Capela, the four other players have a nontrivial, positive number of blocks corresponding to 1, 2, 3, 4, 5, and 6+ shots, indicating that they are comfortable shooting from a larger range. James Harden, in particular, has 45 blocks with 6+ shots, indicating that he has the largest range of “hotspots.”

3. METHODOLOGY

3.1. Zero-Inflated Poisson Regression

In this section, we briefly discuss the ZIP distribution. There are some models that are capable of dealing with excess zero counts (Mullahy, 1986; Lambert, 1992), and zero-inflated models are one of them. Zero-inflated models are two-component mixture models that combine a count component and a point mass at zero with a count distribution such as Poisson, geometric, or negative binomial (see Cameron & Trivedi, 2005 for a discussion). We denote the observed count values of the \( i \)th player as \( y_i = (y_{i1},y_{i2},\ldots,y_{i1175})^\top \). Hence, the probability distribution of

The Canadian Journal of Statistics / La revue canadienne de statistique

DOI: 10.1002/cjs.11684
Table 1: Number of blocks corresponding to the number of shots for selected players.

| Observed value | Davis | James | Harden | Walker | Capela |
|----------------|-------|-------|--------|--------|--------|
| 0              | 836   | 788   | 853    | 837    | 1123   |
| 1              | 161   | 212   | 149    | 159    | 25     |
| 2              | 88    | 91    | 69     | 82     | 8      |
| 3              | 38    | 35    | 31     | 35     | 2      |
| 4              | 18    | 16    | 23     | 17     | 4      |
| 5              | 9     | 8     | 5      | 9      | 0      |
| 6+             | 25    | 25    | 45     | 36     | 13     |

the ZIP random variable \( y_{ij} \) can be written, for a nonnegative integer \( y \), as

\[
pr(y_{ij} = y) = \begin{cases} 
\rho_i + (1 - \rho_i) \exp\{-\mu_{ij}\} & \text{if } y = 0 \\
(1 - \rho_i)^{\frac{y}{\mu_{ij}}} \exp\{-\mu_{ij}\} & \text{if } y > 0 
\end{cases},
\]

where \( \rho_i \) is the probability of extra zeros, and \( \mu_{ij} \) is the mean parameter of the Poisson distribution. For the rest of the article, we denote this distribution as \( \text{ZIP}(\mu_{ij}, \rho_i) \). It can be seen from (1) that ZIP reduces to the standard Poisson model when \( \rho_i = 0 \). Also, we know that \( pr(y_{ij} = 0) > \exp\{-\mu_{ij}\} \) indicates zero inflation. The mean parameter \( \mu_{ij} \) is linked to the explanatory variables through log links as

\[
\log(\mu_{ij}) = x_j^\top \beta_i,
\]

where \( x_j \) is a vector of covariates \( x_j = (1, x_{1,j}, \ldots, x_{p,j})^\top \), and \( \beta_i = (\beta_{0i}, \beta_{1i}, \ldots, \beta_{pi})^\top \) are the corresponding regression coefficients including the intercept \( \beta_{0i} \).

3.2. Nonnegative Matrix Factorization for Spatial Basis

To capture shot styles of individual players, following Jiao, Hu & Yan (2021a), we construct spatial basis functions using historical data. Shot data for a total of 359 players who took 100 shots or more in the 2016–2017 regular season is used as input. First, kernel density estimation is used to estimate the shooting frequency matrix \( \lambda = (\lambda_1, \ldots, \lambda_{1175}) \) for each individual player. Similar checking of empirical correlation between the kernel density values on blocks on the court is performed as in Miller et al. (2014), and the existence of long-range correlations in nonstationary patterns motivates the usage of a basis construction method that captures such long-range correlation via global spatial patterns. This motivates the use of nonnegative matrix factorization (NMF; Sra & Dhillon, 2005) in our modelling effort.

NMF is a dimensionality reduction technique which assumes that a matrix \( \Lambda \) can be approximated by the product of two low-rank matrices

\[
\Lambda \approx WB,
\]

where the matrix \( \Lambda \in \mathbb{R}^{N \times V}_+ \) is composed of \( N \) data points of length \( V \), the basis matrix \( B \in \mathbb{R}^{M \times V}_+ \) is composed of \( M \) basis vectors, and the weight matrix \( W \in \mathbb{R}^{N \times M}_+ \) is composed.
of the $N$ nonnegative weight vectors that scale and linearly combine the basis vectors to reconstruct $\Lambda$. The matrices $W$ and $B$ are obtained by minimizing certain divergence criteria (e.g., Kullback–Leibler divergence or Euclidean distance), with the constraint that all matrix elements remain nonnegative. With the nonnegativity restriction for both the weight vectors and basis vectors, NMF eliminates redundant cases where negative bases “cancel out” positive bases. The remaining basis is often more sparse and focuses on partial representations, which can be combined to represent the whole story (Lee & Seung, 1999).

In our application, with data for the 359 players as a $359 \times 1175$ matrix, we use the R package NMF (Gaujoux & Seoighe, 2010) to calculate the bases. From the practical perspective, we would like each base to be representative of a specific type of shooting habits. At the same time, from the statistical perspective, with the ZIP regression model presented later, where the bases are used as covariates, we would want as little multicollinearity as possible. Combining these two factors together, and after examining the results obtained for $M \in \{3, 4, 5, 6, 7\}$ bases, we chose $M = 5$ for both its practical interpretability and representativeness. The five bases each correspond to a certain shot type, as illustrated in Figure 2. These five bases correspond to, respectively, top of key threes, long two-pointers, restricted area, wing and corner threes, and perimeter shots. A player’s shooting habit can be approximated by a weighted combination of these bases. Nevertheless, looking at the scales, one can see that the maximum value is over 400, and the values are highly skewed. The natural logarithm of the basis values is taken to reduce extreme values, and a normalization step is subsequently performed so that values in each basis vector have mean 0 and standard deviation 1. When such normalized basis functions are used as covariates in modelling an individual player’s shots, their corresponding coefficients, or weight vectors, can be regarded as a characterization for the shooting style of a player. The other four sets of bases obtained for $M \in \{3, 4, 6, 7\}$ are briefly discussed in the Supplementary Material.

3.3. ZIP with Clustered Regression Coefficients

Consider a ZIP regression model with varying coefficients as follows:

$$y_{ij} \sim ZIP \left( \exp \left( x_j^\top \beta_i \right), \rho_i \right), \quad i = 1, \ldots, n, \quad j = 1, \ldots, J,$$

(2)

where $\beta_i$ is a $p + 1$-dimensional regression coefficient. From Gelfand et al. (2003), a Gaussian process prior can be assigned on regression coefficients to obtain varying patterns. Assuming the existence of clusters in $\beta$ and $\rho$, and denoting the true cluster label for player $i$ as $z_i$, the

![Figure 2: Visualization of basis functions obtained by NMF for $M = 5$. Each basis function represents the intensity function of a particular shot type.](image-url)
parameters for cluster $i$ as $\beta_{zi}$ and $\rho_{zi}$, the model in (2) can be rewritten as

$$y_{ij} \sim \text{ZIP}\left(\exp\left(x_j^T \beta_{zi}\right), \rho_{zi}\right), \quad i = 1, \ldots, n, \quad j = 1, \ldots, J,$$

where $z_i \in \{1, \ldots, k\}$, with $k$ being the total number of clusters.

One popular way to model the joint distribution of $z_1, \ldots, z_k$ is the Chinese restaurant process (CRP; Blackwell & MacQueen, 1973). Despite its favourable property as a method to simultaneously estimate the number of clusters and the clustering configuration, it has been proven to produce extraneous clusters in the posterior, even when the sample size goes to infinity, which renders the estimation for the number of clusters inconsistent (Miller & Harrison, 2018). A slowed-down version of the CRP in terms of producing new tables, mixture of finite mixtures (MFM; Miller & Harrison, 2018) is proposed to mitigate this problem:

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

where $p(\cdot)$ is a proper probability mass function on $\{1, 2, \ldots\}$, $(\pi_1, \ldots, \pi_k)$ is the vector of mixture weights such that $\sum_{i=1}^{k} \pi_i = 1$, the vector $\pi = (\pi_1, \ldots, \pi_k)$, and $\delta_h$ is a point mass at $h$. Compared to the CRP, the introduction of new tables is slowed down by the factor $V_n(t+1)/V_n(t)$, which allows for a model-based pruning of the tiny extraneous clusters. The coefficient $V_n(t)$ is precomputed as

$$V_n(t) = \frac{\sum_{n=1}^{+\infty} k_{(n)}(t)}{(\gamma k)^{(n)}},$$

where $k_{(n)} = (k-1) \cdots (k-t+1)$, and $(\gamma k)^{(n)} = \gamma^k (\gamma k + 1) \cdots (\gamma k + n - 1)$. The conditional distributions of $z_i, i = 2, \ldots, n$ under (4) can be defined in a Pólya urn scheme:

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

with $|c|$ being the number of elements in cluster $c$, and $t$ the number of existing clusters. Adapting MFM to our model setting for clustering, the model and prior can be expressed hierarchically as

$$y_{ij} \sim \text{ZIP}\left(\exp\left(x_j^T \beta_{zi}\right), \rho_{zi}\right), \quad i = 1, \ldots, n, \quad j = 1, \ldots, J,$$

$$\beta_h \sim \text{N}\left(0, \Sigma_0\right), \quad h = 1, \ldots, k,$$

$$\rho_h \sim \text{U}(0, 1), \quad h = 1, \ldots, k,$$

$$z_i|k, \pi \sim \sum_{h=1}^{k} \pi_h \delta_h,$$

$$(\pi_1, \ldots, \pi_k)|k \sim \text{Dir}(\gamma, \ldots, \gamma),$$

$$k \sim p(\cdot),$$

DOI: 10.1002/cjs.11684 The Canadian Journal of Statistics / La revue canadienne de statistique
where \( p(\cdot) \) is a Poisson(\( \psi \)) distribution, truncated to be positive through the rest of the article, which has been proven by Miller & Harrison (2018) and Geng, Bhattacharya & Pati (2019) to guarantee consistency for the mixing distribution and the number of groups, \( \gamma = 1 \), \( \delta_h \) denotes the Dirac measure, and \( \Sigma_0 \) is a hyperparameter for the base distribution of \( \beta \)'s. We refer to the hierarchical model above as MFM-ZIP.

3.4. Theoretical Property

In this section, we study the theoretical property of MFM-ZIP. We assume that the parameter space \( \Theta^\ast \) is the compact parameter space for all the model parameters (i.e., mixture weights, regression coefficients, and zero inflated probability) given a fixed number of clusters. The mixing measure \( G = \sum_{i=1}^{k} \pi_i \delta_{\theta_i} \), where \( \delta \) is the point mass measure, and \( \theta_h = \{ \beta_h, \rho_h \} \) is the collection of regression coefficients and zero-inflation probability in cluster \( h \) for \( h = 1, \ldots, k \).

Let \( K_0, G_0, \) and \( P_0 \) be the true number of clusters, the true mixing measure, and the corresponding probability measure, respectively. Then the following proposition establishes the posterior consistency and contraction rate for the cluster number \( K \) and mixing measure \( G \). The proof is based on the general results for Bayesian mixture models in Guha, Ho & Nguyen (2021).

Proposition 1. Let \( \Pi_n(\cdot|y_1, \ldots, y_n) \) be the posterior distribution obtained from a given random sample \( y_1, \ldots, y_n \). Assume that the parameters of interest are restricted to a compact space \( \Theta^\ast \). Then we have

\[
\Pi_n(K = K_0|y_1, \ldots, y_n) \to 1, \quad \Pi_n(W(G, G_0) \lesssim (\log n/n)^{-1/2}|y_1, \ldots, y_n) \to 1,
\]

almost surely under \( P_0 \) as \( n \to \infty \), where \( W \) is the Wasserstein distance.

Proposition 1 shows that our proposed Bayesian method is able to correctly identify the unknown number of clusters and the latent clustering structure with posterior probability tending to 1 as the number of observations increases.

In order to prove Proposition 1, we need to verify that the conditions (P.1)–(P.4) in Guha, Ho & Nguyen (2021) hold. Condition (P.1) is satisfied since we restrict our parameters of interest to a compact space \( \Theta^\ast \) and since the uniform distribution and multivariate normal distribution are first-order identifiable. Condition (P.2) also holds since we assign a nonzero continuous prior on \( \beta \)'s and \( \rho \)'s on the parameters within a bounded support. The uniform distribution and multivariate normal distribution are sufficient for Condition (P.3). Condition (P.4) holds since we choose a truncated Poisson distribution on \( q(\cdot) \). The proof can be finished by using the results in Theorem 3.1 of Guha, Ho & Nguyen (2021).

4. BAYESIAN INFERENCE

For the hierarchical ZIP model with MFM introduced in (6), the set of parameters is denoted as \( \Theta = \{ (\beta_i, \rho_i, z_i, \pi, k) : i = 1, \ldots, n \} \). If we choose \( (k-1) \sim \text{Poisson}(\psi) \) and \( \gamma = 1 \) in (4), the mixture weights \( \pi_1, \ldots, \pi_k \) can be constructed following the stick-breaking (Sethuraman, 1994) approximation:

- **Step 1.** Generate \( \eta_1, \eta_2, \ldots \sim \text{i.i.d.} \ Exp(\psi) \),
- **Step 2.** \( k = \min\{j : \sum_{k=1}^{j} \eta_k \geq 1\} \),
- **Step 3.** \( \pi_h = \eta_h \), for \( h = 1, \ldots, k - 1 \),
- **Step 4.** \( \pi_k = 1 - \sum_{h=1}^{k-1} \pi_h \).
The prior for the hyperparameter $\psi$ is Gamma$(1,1)$. With the prior distributions specified above, the posterior distribution of these parameters based on the data $D = \{y_i, x_j : i = 1, \ldots, n, j = 1, \ldots, p\}$ is given by

$$
\pi(\Theta|D) \propto L(\Theta|D)\pi(\Theta) = \prod_{i=1}^{n} f(y_i, x_1, \ldots, x_j|\beta_{i,j}, \rho_{i,j}, z_i)\pi(\Theta),
$$

where $\pi(\Theta)$ is the joint prior for all the parameters. Because of the unavailability of the analytical form for the posterior distribution of $\Theta$, we employ the MCMC sampling algorithm to sample from the posterior distribution, and then obtain the posterior estimates of the unknown parameters. Computation is facilitated by the nimble (de Valpine et al., 2017) package in R (R Core Team, 2013). The implementation code is given in the Supplementary Material, available online at https://ys-xue.github.io/MFM-ZIP-Basketball-Supplemental/. Another important task is to do the posterior inference for clustering labels. We carry out posterior inference on the clustering labels based on Dahl’s method (Dahl, 2006), which proceeds as follows:

1. **Step 1.** Define membership matrices $A(t) = (A(t)(i,j))_{i,j=1,\ldots,n} = \left(1(z_{i,j} = z_{t,j})\right)_{n \times n}$, where $t = 1, \ldots, T$ is the index for the retained MCMC draws after burn-in, and $1(\cdot)$ is the indicator function.

2. **Step 2.** Calculate the element-wise mean of the membership matrices over MCMC draws $\overline{A} = \frac{1}{T} \sum_{t=1}^{T} A(t)$.

3. **Step 3.** Identify the most representative $\overline{A}$ draw based on minimizing the element-wise Euclidean distance $\frac{1}{T} \sum_{t=1}^{T} (A(t)(i,j) - \overline{A}(i,j))^2$ among the retained $t = 1, \ldots, T$ posterior draws.

The posterior estimates of cluster memberships $z_1, \ldots, z_n$ and other model parameters $\beta$’s and $\rho$’s can also be obtained using Dahl’s method.

5. SIMULATION STUDY

5.1. Simulation Setup

We have two scenarios in our simulation: balanced type and imbalanced type. A total of 75 players are separated into three different groups for each type. Under the balanced design, each group contains an equal number of players. Under the imbalanced design, the group sizes are 10, 35, and 30, respectively. The spatial domain is the same as for the motivating data in Section 2. We generate data $\{y_{ij}; \forall i = 1, \ldots, 75; j = 1, \ldots, 1175\}$ from a ZIP model with different mean parameter and probability parameter of extra zeros. In our simulation setting, our covariates contain an intercept term and five basis function terms (see Section 3.2). Different values of the coefficient $\beta$ are used: $(-1,1.2,0.95,1.1,0.8)^\top$, $(-0.4,0.6,0.7,0.5,0.8,0.3)^\top$, and $(-0.9,0.2,0.1,0.3,0.2,0.4)^\top$, corresponding to each cluster, respectively. We set the true probability parameter of extra zeros for each cluster to be $(0.1,0.3,0.4)$.

We examine both the estimation for number of groups and congruence of estimated membership with their true membership in terms of modulo labelling 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 7000 MCMC iterations, with the first 2000 iterations removed as burn-in for each replicate. The chain length has been examined to ensure convergence and
stabilization. Proceeding to 100 separate replicates of data, our proposed algorithm was run and 100 RI values were obtained by comparing the inferred clusters with the true setting. For each scenario, we also calculate the coverage rate, which equals the percentage of replicates that our proposed algorithm accurately recovers the true number of clusters. The coverage rate for each scenario is 98% and 93%, respectively. We also compare our method to three competitor methods. The first one is the commonly seen $K$-means algorithm, implemented using the \texttt{kmeans()} function. The number of clusters needs to be specified for $K$-means, and for an apples-to-apples comparison we supplied the number of clusters specified by our method to $K$-means in each simulation replicate. The second one is the high-dimensional supervised classification and clustering (HDC; Bergé, Bouveyron & Girard, 2012), implemented in R-package \texttt{HDclassif}, function \texttt{hddc()}. The method is based on a new Gaussian mixture model that combines dimension reduction and model constraints on covariance matrices. With such constraints, the number of parameters in the model is much smaller than other model-based methods, which makes it stable and efficient for high-dimensional applications, and appropriate as a competitor method for clustering 1175-dimensional vectors. As the \texttt{hddc()} function provides the Bayesian information criterion (BIC) as a measure for the goodness of fit, we first tried setting the number of clusters to values in \{1, 2, …, 10\}, and chose the value with the smallest BIC, i.e., the best fit, as the final result for comparison. The last method is the mean shift grouping (Lisic, 2018), implemented in R-package \texttt{meanShiftR}, function \texttt{meanShiftR()}. It is a steepest ascent classification algorithm originating from Fukunaga & Hostetler (1975), where classification is performed by fixed-point iteration to a local maximum of a kernel density estimate. Grouping recovery performances of all four methods are measured using the RI. As $K$-means and the mean shift algorithm cannot infer the number of clusters, such values need to be prespecified, and we supply them with the number of clusters inferred by our method in each replicate. The clustering performances as measured by the RI are shown in Figure 3. It can be seen that for the first three methods, there are many replicates where the methods yield perfect performance, as indicated by the thin line at RI = 1 in the plot. The proposed method, however, has the highest average RI. The mean shift algorithm performs quite poorly in both cases. It is not capable of discerning the grouped pattern in the data, and always clustered all players into one cluster.

![Figure 3: Boxplot of Rand indices for the proposed method and three other competing approaches over 100 simulation replicates. Average RIs are shown below each box.](image-url)
We provide parameter estimation in Table 2. For each of the three $\beta$’s, the average parameter estimate denoted by $\hat{\beta}_{\ell,m}$ ($\ell = 1, \ldots, 75; m = 1, \ldots, 6$) in 100 simulations is calculated as

$$\hat{\beta}_{\ell,m} = \frac{1}{100} \sum_{r=1}^{100} \hat{\beta}_{\ell,m,r},$$

where $\hat{\beta}_{\ell,m,r}$ denotes the posterior estimate for the $m$th coefficient of player $\ell$ in the $r$th replicate. We use different metrics to evaluate the posterior performance. Those metrics include the mean absolute bias (MAB), the mean standard deviation (MSD), the mean of mean squared error (MMSE), and the mean coverage rate (MCR) of the 95% highest posterior density (HPD) intervals in the following ways:

$$\text{MAB} = \frac{1}{75} \sum_{\ell=1}^{75} \frac{1}{100} \sum_{r=1}^{100} |\hat{\beta}_{\ell,m,r} - \beta_{\ell,m}|,$$

$$\text{MSD} = \frac{1}{75} \sum_{\ell=1}^{75} \sqrt{\frac{1}{99} \sum_{r=1}^{100} \left(\hat{\beta}_{\ell,m,r} - \hat{\beta}_{\ell,m}\right)^2},$$

$$\text{MMSE} = \frac{1}{75} \sum_{\ell=1}^{75} \frac{1}{100} \sum_{r=1}^{100} \left(\hat{\beta}_{\ell,m,r} - \beta_{\ell,m}\right)^2,$$

$$\text{MCR} = \frac{1}{75} \sum_{\ell=1}^{75} \frac{1}{100} \sum_{r=1}^{100} \left\{\hat{\beta}_{\ell,m,r} \in \text{95% HPD interval}\right\},$$

where $1_{\{\cdot\}}$ denotes the indicator function. The four metrics for each $\beta$ under the balanced and imbalanced designs are presented in Table 2. With high clustering accuracy as indicated by the RI, the estimated $\beta$ for each cluster is close to its corresponding true value, which can be seen.
by the small numerical values in the MAB column. The estimation performance is stable in the sense that the MSD values are also small. The MCR under the balanced design fluctuates around its nominal value of 0.95, and under the imbalanced design, the values are overall lower owing to the influence of mis-clustered players, but still remain close to or greater than 0.9. As suggested by a referee, it is also interesting to examine the performance of the proposed model when ZIP

FIGURE 4: Visualization of shooting patterns for four selected players from each group.
is not the true data-generating model. We have done additional simulation studies under this scenario, and the details are included in the Supplementary Material.

6. REAL DATA APPLICATION

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). We run 15,000 MCMC iterations and discard the first 5000 iterations as burn-in. The result from the MFM-ZIP model suggests that the 191 players are to be classified into four groups. The sizes of the four groups are 29, 110, 48, and 4, respectively. We show the shot attempt counts made by the four selected players on blocks of the court in Figure 4. The players for each group are shown in Section 3 of the Supplementary Material.

Several interesting observations can be made from Figure 4. It can be seen that each group of players have their own favourite shooting locations. Players in Group 1 make the most shots near the hoop, which is confirmed by the regression coefficients in Table 3, as their coefficients for the third and fifth basis functions are the largest. Clint Capela and DeAndre Jordan, e.g., are both good at making alley-oops and slam dunks. Andre Drummond and Dwight Howard are also centres who rarely leave the painted area.

Players in Group 2 make the most shots beyond the three-point line, as they have the largest parameter estimates for the first and fourth basis functions when compared to other groups. As shown in Figure 4, JJ Redick and Stephen Curry are both well-known shooters.

A first look at the plots for Group 3 indicates that players in this group are able to make all types of shots, including three-pointers, perimeter shots, and shots over the painted area. We find the players in this group are often the leaders in their teams, and usually have the most possession. The parameter estimates also confirm the observation. Their $\hat{\beta}_0$ values are the largest among all groups, indicating an overall higher probability for making shots. Compared with those of the players of Group 2, their shots are more evenly distributed, which can be reflected by the larger parameter estimate for the basis functions corresponding to areas within the three-point line.

For Group 4, we find that most of their shots are close to the hoop and around the perimeter, and they have fewer shots around or beyond the three-point line. From the estimation result, the coefficients for the second and third basis functions are larger than other basis functions, but similar in value. In addition, their $\hat{\beta}_1$ values are the second smallest among all four groups while their $\hat{\beta}_4$ values are the smallest, which again indicates their disfavour toward shooting beyond the three-point line. Note that the presented analysis is based on the 2017–2018 regular season, which was before Giannis Antetokoumpo increased his three-point shots in the 2019–2020 season.

| Group  | $\rho$ | $\hat{\beta}_0$ | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ | $\hat{\beta}_4$ | $\hat{\beta}_5$ |
|--------|-------|----------------|----------------|----------------|----------------|----------------|----------------|
| Group 1 | 0.583 | -1.857 | 0.256 | 0.573 | 1.374 | 0.531 | 1.061 |
| Group 2 | 0.618 | -0.409 | 0.731 | 0.200 | 0.359 | 0.581 | 0.392 |
| Group 3 | 0.515 | -0.147 | 0.511 | 0.237 | 0.505 | 0.493 | 0.474 |
| Group 4 | 0.322 | -1.205 | 0.258 | 1.101 | 1.056 | 0.426 | 0.766 |

Table 3: Performance of parameter estimates under the real data.
7. DISCUSSION

Based on theoretical justifications and empirical studies, our proposed methods successfully solved the three challenges raised in Section 1. Based on the results shown in Section 5, our proposed methods could accurately estimate the parameters in the ZIP model and recover the number of clusters and clustering configurations with different proportions of zero counts. Compared with several benchmark clustering methods such as $K$-means, high-dimensional supervised classification and clustering, and mean shift grouping, our methods have higher clustering concordance without any tuning steps.

In the analysis of the NBA shot chart data, four field goal attempt patterns, their corresponding zero-inflation probabilities, and regression coefficients of each group were identified. The results provide valuable insights to players, coaches, and managers. The players can obtain more descriptive analysis of their current offence patterns, and thereby develop customized training plans with pertinence; the coaches can organize their offence and defence strategies more efficiently for different opponents; the managers will make better data-informed decisions on player recruiting and trading during the off-season. Our proposed method is not only for field goal attempts but also can be used in other applications such as processing and passing data (Arbues-Sanguesa et al., 2020; Hurault, Ballester & Haro, 2020) in soccer games.

There are several possible directions for further investigation. Spatial correlation over the court is accounted for nonparametrically by the basis functions. Considering either stationary or nonstationary model-based spatial correlation in our proposed modelling framework is a natural extension. In this article, our posterior sampling was based on stick-breaking representations. Developments of more scalable inference algorithms (e.g., variational inference) are critical for large-scale data. In addition, proposing a matrix-on-scale for other covariates such as age and salary will help us to extend applications of the proposed method. In this article, we adopted the ZIP framework to deal with the excessive number of zeros. There is, however, the possibility that the assumptions of the Poisson distribution are violated. An interesting future project would be to decompose the basketball court into fewer regions that could be of either regular or irregular shape, and consider both the excess zero issue and under/over dispersion for each subregion. The zero-inflated Conway–Maxwell Poisson distribution (Barriga & Louzada, 2014) would be a potential model to use. Finally, building a heterogeneity learning model with auxiliary information from different players, such as age and position on the court, merits future research from both methodological and applied perspectives.

REFERENCES

Arbues-Sanguesa, A., Martin, A., Fernández, J., Ballester, C., & Haro, G. (2020). Using player’s body-orientation to model pass feasibility in soccer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 886–887.

Barriga, G. D. & Louzada, F. (2014). The zero-inflated Conway–Maxwell–Poisson distribution: Bayesian inference, regression modeling and influence diagnostic. Statistical Methodology, 21, 23–34.

Bergé, L., Bouveyron, C., & Girard, S. (2012). HDclassif: An R package for model-based clustering and discriminant analysis of high-dimensional data. Journal of Statistical Software, 46, 1–29.

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

Cameron, A. C. & Trivedi, P. K. (2005). Microeconometrics: Methods and Applications. Cambridge University Press, Cambridge, UK.

Dahl, D. B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. In Do, K.-A. & Peter Müller, M. V. (Eds.), Bayesian Inference for Gene Expression and Proteomics, Vol. 4, Cambridge University Press, Cambridge, UK, 201–218.

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., & Bodik, R. (2017). Programming with models: Writing statistical algorithms for general model structures with NIMBLE. Journal of Computational and Graphical Statistics, 26, 403–413.
Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1, 209–230.

Fukunaga, K. & Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21, 32–40.

Gaujoux, R. & Seoighe, C. (2010). A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics*, 11, 367.

Gelfand, A. E., Kim, H.-J., Sirmans, C., & Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98, 387–396.

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

Geyer, C. (1998). Likelihood inference for spatial point processes. In Kendall, W. S. (Ed.), *Stochastic Geometry: Likelihood and Computation*, Vol. 80, CRC Press, Boca Raton, FL, 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, 365–379.

Guha, A., Ho, N., & Nguyen, X. (2021). On posterior contraction of parameters and interpretability in Bayesian mixture modeling. *Bernoulli*, 27, 2159–2188.

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*.

Hu, G., Yang, H.-C., & Xue, Y. (2021). Bayesian group learning for shot selection of professional basketball players. *Stat*, 10, e324.

Hurault, S., Ballester, C., & Haro, G. (2020). Self-supervised small soccer player detection and tracking. In *Proceedings of the 3rd International Workshop on Multimedia Content Analysis in Sports*, 9–18.

Jiao, J., Hu, G., & Yan, J. (2021a). A Bayesian joint model for spatial point processes with application to basketball shot chart. *Journal of Quantitative Analysis in Sports*, 17, 77–90.

Jiao, J., Hu, G., & Yan, J. (2021b). Heterogeneity pursuit for spatial point pattern with application to tree locations: A Bayesian semiparametric recourse. *Environmetrics*, 32, e2694.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Nature*, 401, 788–791.

Lee, D. D. & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401, 788–791.

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

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

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33, 341–365.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

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, 3–12.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4, 639–650.

Sra, S. & Dhillon, I. (2005). Generalized nonnegative matrix approximations with Bregman divergences. *Advances in Neural Information Processing Systems*, 18, 283–290.

Vavrek, M. J. (2011). fossil: Palaeoecological and palaeogeographical analysis tools. *Palaeontologia Electronica*, 14, 16 1T, R package version 0.4.0.

Yin, F., Hu, G., & Shen, W. (2020). Analysis of professional basketball field goal attempts via a Bayesian matrix clustering approach. *arXiv preprint arXiv:2010.08495*.
Yin, F., Jiao, J., Hu, G., & Yan, J. (2020). Bayesian nonparametric estimation for point processes with spatial homogeneity: A spatial analysis of NBA shot locations. *arXiv preprint arXiv:2011.11178.*

Zhao, P., Yang, H.-C., Dey, D. K., & Hu, G. (2020). Bayesian spatial homogeneity pursuit regression for count value data. *arXiv preprint arXiv:2002.06678.*

Received 7 April 2021
Accepted 7 November 2021