Journal of Data Science 0 (0), 1–15
??? 2025

DOI: 10.6339/25-JDS1196
Data Science in Action

A Bayesian Negative Binomial-Bernoulli Model with Tensor Decomposition: Application to Jointly Analyzing Shot Attempts and Shot Successes in Basketball Games

Kwok-Wah Ho[1]
[1]*Department of Statistics, Chinese University of Hong Kong, Hong Kong*

Abstract

We propose a Bayesian Negative Binomial-Bernoulli model to jointly analyze the patterns behind field goal attempts and the factors influencing shot success. We apply nonnegative CANDECOMP/PARAFAC tensor decomposition to study shot patterns and use logistic regression to predict successful shots. To maintain the conditional conjugacy of the model, we employ a double Pólya-Gamma data augmentation scheme and devise an efficient variational inference algorithm for estimation. The model is applied to shot chart data from the National Basketball Association, focusing on the regular seasons from 2015–16 to 2022–23. We consistently identify three latent features in shot patterns across all seasons and verify a popular claim from recent years about the increasing importance of three-point shots. Additionally, we find that the home court advantage in field goal accuracy disappears in the 2020–21 regular season, which was the only full season under strict COVID-19 crowd control, aside from the short bubble period in 2019–20. This finding contributes to the literature on the influence of crowd effects on home advantage in basketball games.

Keywords *logistic regression; Pólya-Gamma; variational inference*

1 Introduction

Data analytics has proven invaluable to the development of the National Basketball Association (NBA) industry. Player evaluation, team and player match-ups, in-game strategy planning, and other areas have significantly improved. Among these, a primary concern is finding effective methods to understand individual players' field goal attempt patterns and the factors behind shot success. A better understanding can enhance the formulation of attack and defense strategies, as well as the design of appropriate training schemes for players on an individual basis. It is natural to postulate that shot attempt intensity is related to the probability of shot success, as players tend to shoot from locations where they have high confidence. Therefore, we propose a joint model for both shot attempts and shot success to better understand their dependence.

In the literature, there are studies using spatial processes, spatio-temporal processes, matrix factorization methods, and hierarchical regression to model shot attempt patterns (Reich et al., 2006; Miller et al., 2014; Franks et al., 2015; Hu et al., 2021, 2023; Yin et al., 2023). While tensor decomposition methods are popular in the data science literature for analyzing high-dimensional data, they are not commonly applied in basketball data analytics, except for Hu et al. (2022). The authors consider tensor analysis in modeling shot attempt patterns for NBA players in the

[*] Email: remus@cuhk.edu.hk.

Received September 17, 2024; Accepted June 30, 2025

2017–18 regular season. They suggest a Bayesian mixture model of rank-one three-mode tensors to model the players' attempts. Their model requires an MCMC algorithm for estimation and an additional clustering algorithm to be applied to the MCMC output in order to form the players into clusters. In our study, we also postulate that shot attempt patterns are influenced by the "who", "when" and "where" of the shots. Thus, we organize the player identity, game period, and court location of all shots into a 3-mode tensor. Unlike the approach in Hu et al. (2022), we use Negative Binomial instead of Poisson distribution to model shot attempts, and a traditional tensor analysis approach to apply Bayesian nonnegative CANDECOMP/PARAFAC (CP) decomposition (for a concise introduction of CP method, please see Section 3 of Kolda and Bader, 2009). There is no need to employ an additional clustering algorithm in our method, and a computationally efficient variational inference algorithm can be applied. Under a sparsity-promoting prior setting, we consistently obtain a tensor rank of three for all eight regular seasons under study. The evolution of these three latent features, represented by the three rank-one tensors, helps us understand the trends of field goal attempts throughout the study period.

Understanding factors behind achieved shots is another crucial aspect for successful team management. It is natural to expect that shot attempt intensity affects shot success rate because players tend to choose locations and times with confidence to shoot. In this regard, there are several recent studies that analyze shot attempts and shot successes together in a joint model (Jiao et al., 2021; Wong-Toi et al., 2023). In our study, we use logistic regression to predict the outcome of each shot, with a variable reflecting shot attempt intensity estimated from the Negative Binomial part of our model as a predictor. At the same time, we also consider other predictors, including shot distance, remaining time in the period, and whether the shot is taken by the home team.

We have made some interesting observations in our experiments. There is a popular conception that the NBA is undergoing a three-point revolution (Zajac et al., 2023; Freitas, 2021; Rolland et al., 2020) in recent years, and we verify this through a study of the latent features extracted from the tensor decomposition. Additionally, the 2020–21 regular season, which was the only post-COVID season affected by strict crowd restrictions, provides a natural experiment for us to study the effect of home fans. Related studies (Leota et al., 2022; Ehrlich and Potter, 2023; Ganz and Allsop, 2024; Steinfeldt et al., 2022) mainly report results regarding rebound numbers and score differences. In our regression analysis, we find that the home advantage disappears only in the 2020–21 season, verifying the importance of the home fans' effect on shot accuracy as well.

Besides interesting empirical findings, we also contribute to the methodological aspect. Firstly, the joint model involves logit links in both the Negative Binomial part and the logistic regression part. We follow Ma et al. (2024) in using double Pólya-Gamma data augmentation to achieve conditional conjugacy. Secondly, the non-negativity of the latent features obtained from the tensor decomposition is important for part-based interpretation. Thus, we apply truncated normal distributions as priors for the latent features to maintain both non-negativity and conditional conjugacy (Cheng et al., 2020; Hinrichy et al., 2018). Lastly, we adopt a variational inference approach to estimate the model. However, the dispersion parameter r of the Negative Binomial distribution does not allow for conditional conjugacy. Therefore, we construct a Variational Bayes EM algorithm similar to that of Soulat et al. (2021) to solve this problem.

In Section 2, we will first have a discussion about the NBA shot chart data that we use. After that, we shall describe our model and the method of inference in Section 3. The experimental results are summarized in Section 4 and a final discussion will be given in Section 5.

2 Data Description

The shot chart data (publicly accessible at http://github.com/swar/nba_api used for our analysis are the regular seasons of NBA league from 2015–16 to 2022–23. We intentionally exclude the so-called "bubble" period (30/7/2020 to 14/8/2020) in the 2019–20 regular season. During the bubble period, only 22 teams (instead of the normal 30 teams) were invited to play, with games being held behind closed doors at the ESPN Wide World of Sports Complex and the teams staying at Disney World hotels. Since the environment and operations differ in many aspects compared to other periods, we decided to exclude the data from these few weeks. However, this data only comprises less than 9% of the data in the 2019–20 regular season.

Our data contain records with various information about every shot in every game. The variables include team name, shot player id, shot made/miss, shot distance, shot region, remaining time to end of the quarter, etc. This abundant information help us to build a joint Bayesian model to analysis the shot attempt patterns and relevant factors behind successful shots.

For each regular season, we only consider shot attempts taken from the frontcourt, as the number of attempts from the backcourt is relatively rare and has a very low success rate. We divide the frontcourt into 18 distinct areas, as outlined in Figure 1 and Table 1. Secondly, we focus our analysis solely on the standard four periods of 12 minutes each, excluding any overtime periods. Lastly, we only include frequent shooters who made at least 400 shot attempts in the entire season. Considering the 2021–22 season as an example, after these filtering steps, we arrive at a 3-mode count tensor of size $223 \times 4 \times 18$, representing the number of shot attempts for each eligible player, game period, and court location. In the next section, we will introduce a Negative Binomial model for this count data and apply a non-negative CP tensor decomposition method to learn the low-rank structure of the tensor and thus identify the common patterns of shot attempts.

Another part of our analysis is to identify the factors behind field goal successes through a logistic regression model. It is natural to expect that, for a given game time, a player will choose a location where they feel more confident and have a better success rate for shooting. Thus, shot attempt intensity is closely related to the shot success rate. This suggests that a variable summarizing the attempt intensities for player × period × location combinations from the Negative Binomial model can serve as a predictor, linking the two parts of our model. The details of this predictor will be discussed in Section 3. Additionally, we include three more predictors in the regression. The first is shot distance, which is understandably related to the likelihood of shot success. The second is a binary variable indicating whether the shooter is from

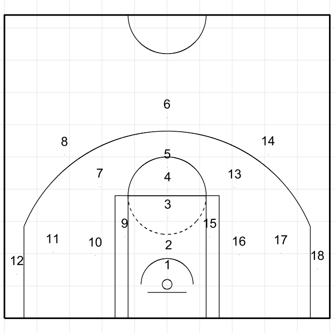

Figure 1: The eighteen zones defined on the front court.

Table 1: Description of the 18 zones. The abbreviations C, LC, L, RC, and R refer to Center, Left Center, Left, Right Center, and Right, respectively.

Zone	Description	Zone	Description
1	C, Restricted Area, <8 feet	2	C, Paint (non-RA), <8 feet
3	C, Paint (non-RA), 8–16 feet	4	C, Mid-Range, 8–16 feet
5	C, Mid-Range, 16–24 feet	6	C, Above Break 3, 24+ feet
7	LC, Mid-Range, 16–24 feet	8	LC, Above Break 3, 24+ feet
9	L, Paint (non-RA), 8–16 feet	10	L, Mid-Range, 8–16 feet
11	L, Mid-Range, 16–24 feet	12	L, Left Corner 3, 24+ feet
13	RC, Mid-Range, 16–24 feet	14	RC, Above Break 3, 24+ feet
15	R, Paint (non-RA), 8–16 feet	16	R, Mid-Range, 8–16 feet
17	R, Mid-Range, 16–24 feet	18	R, Right Corner 3, 24+ feet

the home team, as the home advantage is commonly recognized in sports analysis. The last predictor is the remaining time in the period, as we want to see if time pressure affects shooting performance.

We fit our model for the regular seasons from 2015 to 2023 to investigate the trends regarding shot attempts and successes over time. Additionally, the 2020–21 post-COVID regular season was seriously affected by the strict crowd policies. This provides us with a natural experimental setting to understand the effects of home fans on home-court advantage. We will discuss the details of our empirical findings in Section 4.

3 Bayesian Analysis

3.1 A Negative Binomial-Bernoulli Model

Our objective is to learn the patterns of goal attempts and the factors behind shot successes. The simplest approach is, of course, to model goal attempts and shot successes separately. However, as explained in the previous section, it is natural to expect that they are correlated. Thus, a joint model is necessary to capture this dependence, and a Bayesian approach can facilitate this well. Our model is inspired by the Negative Binomial-Binomial model proposed by Ma et al. (2024). In that paper, the authors apply a Negative Binomial model for the total number of mtDNA chromosomes per cell and use a Binomial model to count how many of those mtDNA copies have a mutation at a given position. The two models are linked by including the logit of the probability parameter in the Negative Binomial model as a predictor for modeling the logit of the probability parameter in the Binomial part.

We consider a similar joint model with two components for goal attempts and shot successes, respectively. In the first component, we model the number of goal attempts using a Negative Binomial model. With reference to Hu et al. (2022), we postulate that field goal choices are mainly affected by three factors: the identity of the player, the game period, and the court location. Therefore, we use a 3-mode tensor to model these effects. Unlike Hu et al. (2022), we apply a more traditional way of tensor decomposition and use the Negative Binomial distribution instead of the Poisson distribution. It is well known that the Negative Binomial distribution has advantages over the Poisson distribution for modeling over-dispersion commonly found in count data. Thus we use a Negative Binomial distribution to model y_i, the number of field goal attempts at the ith player × period × location combination where the subscript i being a triple (i_1, i_2, i_3) that $1 \leqslant i_k \leqslant I_k$ with I_k being the dimension of the k-th mode. We further define $I = I_1 \times I_2 \times I_3$ to be the total number of elements in the tensor. We follow Pillow and Scott (2012) and Neelon (2019) to parametrize the Negative Binomial distribution as

$$p(y_i|r, \psi_i) = \frac{\Gamma(y_i + r)}{\Gamma(r) y_i!}(1 - \psi_i)^r \psi_i^{y_i},$$

where r is a common dispersion parameter and the parameter ψ_i is defined as

$$\psi_i = \frac{\exp(\sum_{d=1}^{D} \prod_{k=1}^{3} a_{i_k,d}^{(k)})}{1 + \exp(\sum_{d=1}^{D} \prod_{k=1}^{3} a_{i_k,d}^{(k)})}, \tag{1}$$

where D is the rank and $\{a_{j,d}^{(k)}\}_{d=1,..,D;k=1,...,3;j=1,...,I_k}$ being the latent variables under a standard CANDECOMP/PARAFAC (CP) decomposition model (Harshman, 1970; Bro, 1997; Sørensen et al., 2012; Rai et al., 2015) for the 3-mode tensor.

Notice that under our parametrization of the Negative Binomial model, the expected value of counts is

$$E(y_i) = \frac{r\psi_i}{1 - \psi_i}.$$

Thus a larger value of ψ_i induces a higher number of attempts on average. One of our objectives is to study patterns of field goal attempts related to the 3 modes. To ensure interpretability, we enforce the nonnegativity of the latent variables $\{a_{j,d}^{(k)}\}_{d=1,..,D;k=1,...,3;j=1,...,I_k}$. This parts-based representation allows only addition, not subtraction, leading to a more interpretable model where each extracted latent component represents a distinct part of the data. Specifically, we follow Cheng et al. (2020) to use Truncated Normal distributions as prior distributions for each of the latent variables

$$a_{j,d}^{(k)} \sim TN_{(0,\infty)}\big(0, \lambda_d^{-1}\big), \quad \text{for } d = 1, \dots, D, \ k = 1, \dots, 3, \ j = 1, \dots, I_k,$$

where $TN_{(a,b)}(\mu, \sigma^2)$ denotes a Truncated Normal distribution with the following density function

$$f(x) = \frac{1}{\sigma} \frac{\rho(\frac{x-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}$$

with support on (a, b), and ρ and Φ denote the density function and cumulative distribution function of standard Normal distribution respectively.

The CP rank D is another important quantity to estimate in our model. Exploiting a low-rank tensor factorization structure (i.e., a small value of D) can facilitate model interpretation and noise reduction. In our Bayesian analysis, we apply a sparsity-promoting prior over the λ_d parameters, similar to the automatic relevance determination (ARD) method (Tipping, 2001; Zhao et al., 2015). We set

$$\lambda_d \sim \text{Gamma}(\epsilon, \epsilon), \quad \text{for } d = 1, \dots, D,$$

where Gamma(ϵ, ϵ) represents a Gamma distribution with mean 1 and variance ϵ^{-1}, and the parameter ϵ is set to a small value of 10^{-6} to promote sparsity following the suggestion of Cheng et al. (2020).

The second component of our model is a logistic regression that models the success of each shot. Here I use x_{il} to represent the lth shot in the ith player × period × location combination. We further denote L_i to be the number of shots for the ith combination. So for $i \in \{1, \dots, I\}$ and $l \in \{1, \dots, L_i\}$, the regression model is

$$\begin{aligned} x_{il} &\sim b(1, p_{il}), \\ p_{il} &= \frac{\exp(\xi \phi_i + \boldsymbol{\beta}^T \mathbf{z}_{il})}{1 + \exp(\xi \phi_i + \boldsymbol{\beta}^T \mathbf{z}_{il})}, \\ \phi_i &= \sum_{d=1}^{D} \prod_{k=1}^{3} a^{(k)}_{i_k,d} + \nu_i, \end{aligned} \tag{2}$$

where ξ and $\boldsymbol{\beta}$ are regression coefficients, $\mathbf{z}_{il}$ is a vector of m predictors and ν_i is a mean zero noise term.

Through a logit link, the probability of success, p_{il} is related to a linear combination of $\mathbf{z}_{il}$ and another variable ϕ_i. The idea of including ϕ_i as a predictor parallels to the design of the joint model in Ma et al. (2024). Notice that in equation (1), $logit(\psi_i) = \sum_{d=1}^{D} \prod_{k=1}^{3} a^{(k)}_{i_k,d}$ and hence $E(\phi_i) = logit(\psi_i)$. Recall that a higher value of ψ_i implies a higher value of the average number of goal attempts. Thus including ϕ_i as a predictor is natural to measure the effect of shot attempt intensity of the player × period × location combination on the success of a shot. Here we follow the "explicit noise approach" in Klami (2014) to include a mean zero noise ν_i in constructing ϕ_i. We find that this construction suggests a simpler algorithm and enhances its stability. To complete the model, we specify the prior distributions of the parameters

$$\begin{aligned} \xi \sim N\big(0, \tau_\xi^{-1}\big), &\quad \boldsymbol{\beta} \sim N_m\big(\mathbf{0}, \Omega^{-1}\big), \\ \nu_i \sim N\big(0, \tau_\phi^{-1}\big), &\quad \tau_\phi \sim \text{Gamma}(\delta, \delta), \end{aligned}$$

where the hyperparameters τ_ξ, δ and Ω^{-1} are chosen to be 10^{-6}, 10^{-2} and $10^{-6} I_m$ respectively.

3.2 Pólya Gamma Data Augmentation

When dealing with Bayesian models that have non-conjugate priors, Gibbs sampling or Variational inference methods can become challenging. However, a data augmentation trick can sometimes help overcome this hurdle. The idea is to introduce an additional latent variable $\boldsymbol{\omega}$ and a corresponding conditional distribution $p(\boldsymbol{\omega} \mid \boldsymbol{\beta}, \mathbf{y})$. This transforms the original data model $p(\mathbf{y} \mid \boldsymbol{\beta})$ into a joint probability $p(\mathbf{y}, \boldsymbol{\omega} \mid \boldsymbol{\beta}) = p(\mathbf{y} \mid \boldsymbol{\beta}) p(\boldsymbol{\omega} \mid \boldsymbol{\beta}, \mathbf{y})$. The beauty of this approach is that the desired posterior $p(\boldsymbol{\beta} \mid \mathbf{y})$ can be obtained as the marginal of the joint posterior $p(\boldsymbol{\beta}, \boldsymbol{\omega} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \boldsymbol{\beta}) p(\boldsymbol{\omega} \mid \boldsymbol{\beta}, \mathbf{y}) p(\boldsymbol{\beta})$. If the full conditional distributions $p(\boldsymbol{\beta} \mid \boldsymbol{\omega}, \mathbf{y})$ and $p(\boldsymbol{\omega} \mid \boldsymbol{\beta}, \mathbf{y})$ are tractable, Gibbs sampler or Variation inference can still be applied efficiently.

The main difficulties of our model come from the likelihood for both parts of our model because of the logit links. There are no conjugate priors for the parameters of both distributions. Fortunately, the Pólya Gamma data augmentation suggested in Polson et al. (2013) helps to build conditional conjugacy that leads to a efficient variational inference algorithm to be introduced in the next section. Details of our augmentation scheme is contained in Supplementary Material (S.1). This method is widely used in many Bayesian applications involving Binomial or Negative Binomial likelihoods. In particular, Ma et al. (2024) consider a Binomial-Negative Binomial spatial model that requires double augmentation similar to our case. However, they do not consider tensor decomposition and focus on Gibbs sampling that are different from us.

3.3 Variational Inference

While MCMC algorithms like Gibbs sampling are promising in estimating the posterior distribution of our model, the computational burden for simulation is daunting because of the large number of parameters and auxiliary variables. Therefore we follow the Bayesian CP tensor decomposition literature (Cheng et al., 2020; Hinrichy et al., 2018; Soulat et al., 2021; Takayama et al., 2022; Yang et al., 2025; Zhao et al., 2015) to approximate the posterior distribution using Variational Bayes (VB) inference methods. VB algorithms generally converge much faster than MCMC algorithms and are thus extensively used in machine learning and data science problems to handle large datasets and complicated models. In particular, we make use of mean-field approximation as the full conditional distributions are known distributions. For completeness, we briefly describe the VB method with mean-field approximation below.

Consider a general statistical problem with data $\mathscr{D}$ and parameters Θ, the target of VB method is to find a distribution $q(\Theta)$ (so-called the variational distribution) to approximate the true posterior distribution $p(\Theta|\mathscr{D})$ by minimizing the Kullback–Leibler (KL) divergence, that is

$$\begin{aligned}\mathrm{KL}\left(\frac{q(\Theta)}{p(\Theta|\mathscr{D})}\right) &= \int q(\Theta)\ln\left(\frac{q(\Theta)}{p(\Theta|\mathscr{D})}\right)d\Theta \\ &= \ln p(\mathscr{D}) - \int q(\Theta)\ln\left(\frac{p(\mathscr{D},\Theta)}{q(\Theta)}\right)d\Theta \\ &= \ln p(\mathscr{D}) - \mathcal{L}(q),\end{aligned}$$

where $\ln p(\mathscr{D})$ represents the model evidence, and its lower bound (so-called the ELBO) is defined by $\mathcal{L}(q) = \int q(\Theta)\ln(\frac{p(\mathscr{D},\Theta)}{q(\Theta)})d\Theta$. Since the model evidence is a constant, the maximum of the lower bound occurs when the KL divergence vanishes, which implies that $q(\Theta) = p(\Theta|\mathscr{D})$.

The idea of mean-field approximation is to make an assumption that the variational distribution can be factorized with respect to each parameter Θ_j such that

$$q(\Theta) = \prod_j q_j(\Theta_j).$$

It is important to note that this factorized form is the only assumption made about the variational distribution. The specific functional forms of the individual factors $q_j(\Theta_j)$ can be explicitly derived one by one. The optimised form of the jth factor based on the maximization of $\mathcal{L}(q)$ is given by

$$\ln q_j(\Theta_j) = \mathbb{E}_{-j}\big[\ln p(\mathscr{D},\Theta)\big] + \text{const.}, \tag{3}$$

where $\mathbb{E}_{-j}[\cdot]$ denotes an expectation w.r.t. the q distributions over all variables except Θ_j.

As we mentioned before, we treat the dispersion parameter r as an unknown constant to be estimated. In other words, we shall construct a Variational EM algorithm for estimation. Given an estimate of r, we update the variational distributions of other parameters using mean-field approximation. After that we update r by maximizing the ELBO. These two steps will be repeated until the ELBO values stablize at a maximum. Soulat et al. (2021) also consider a Variational EM algorithm with similar structure for a Negative Binomial count tensor model.

For the variational distribution on the parameters $\Theta = \{\{a_{ld}^{(k)}\}, \{\lambda_d\}, \{\phi_i\}, \{w_i\}, \{v_{il}\}, \tau_\phi, \xi, \boldsymbol{\beta}\}$, we adopt the factorization

$$q(\Theta) = \left(\prod_{d=1}^{D}\prod_{k=1}^{3}\prod_{l=1}^{I_k} q\big(a_{ld}^{(k)}\big)\right)\left(\prod_{d=1}^{D} q(\lambda_d)\right)\left(\prod_{i=1}^{I} q(\phi_i)\right)\left(\prod_{i=1}^{I} q(w_i)\right)\left(\prod_{i=1}^{I}\prod_{l=1}^{J_i} q(v_{il})\right)q(\tau_\phi)q(\xi)q(\boldsymbol{\beta}).$$

For each iteration of the algorithm, given an estimate of r, the variational E-step is to update each component $q_j(\Theta_j)$ one by one based on equation (3). Because of conditional conjugacy, all components of the variational distribution are well-known distributions. For example $q_j(\beta)$ is Multivariate Normal distributions and $q_j(\phi_i)$ is Normal distribution. Details of the formulae for updating the variational distributions are reported in the Supplementary Material (S.2).

After updating the variational distributions, we search r that maximize the ELBO. The maximization step is detailed in the Supplementary Material (S.2). The above variational-E step and maximization step will be iterated until the ELBO value stablize at its maximum value.

4 Application to NBA Data

In this section, we summarize our experimental results for the NBA data. In the first part, we will discuss the findings from the Negative Binomial component of our model. For all the regular seasons analyzed, our algorithm consistently chooses a rank of 3 for the CP tensor decompositions. Thus, there are three latent features behind the distributions of field goal attempts. We will compare the similarities and differences of these three factors across the NBA seasons studied. In the second part, we will focus on the logistic regression component of our model and report on the significance of various predictors of shot successes. To provide a clear and concise description, we will only report part of the graphs and numerical results to highlight our findings. More graphical results can be found in the Supplementary Material (S.3).

4.1 Results for Shot Attempts

In each run of our algorithm, we start with an initial CP tensor rank of 15, and after estimation, we consistently find that only 3 dominant latent features remain. Each feature is a rank-1 3-mode tensor describing the patterns related to the player, period, and location of the feature. For easier discussion, the analysis below is based on the mean of the variational distributions of the latent feature elements. We first study the 3 features based on the location mode. We report the relative importance of the 18 locations for the 3 features for alternate years from 2015–16 to 2021–22 in Figure 2. This choice allows us to observe the trend over time while keeping the number of graphs manageable.

For each panel in Figure 2, zones with darker colors and larger areas represent those with larger means of the variational distributions and are therefore more important. The first to fourth rows correspond to the seasons 2015–16, 2017–18, 2019–20 and 2021–22, respectively. We first notice that the left column can be interpreted as a factor mainly for 3-point attempts (with some from the restricted area), while the right column pertains to attempts concentrated in the restricted area and near the basket. The pattern in the middle column is more diffuse, but we can still observe that attempts are primarily from mid-range and the paint area.

How about the change in patterns over time, i.e., along the columns? The changes in the left and right columns are not obvious. However, if we focus on the middle column, there is a clear trend that we can observe. For the 2015–16 season, we do not observe any significance in 3-point attempts. However, in later seasons, the circles for 3-point attempts from the center and the wings grow larger and darker. This observation aligns with the trend of the growing importance of 3-point shooting observed in recent years.

Next, we would like to study the three latent features with respect to the game period mode. As the patterns are similar for all seasons, we only report here the result of season 2015–16. Figure 3 shows the mean values of the variational distributions for the four periods (normalized

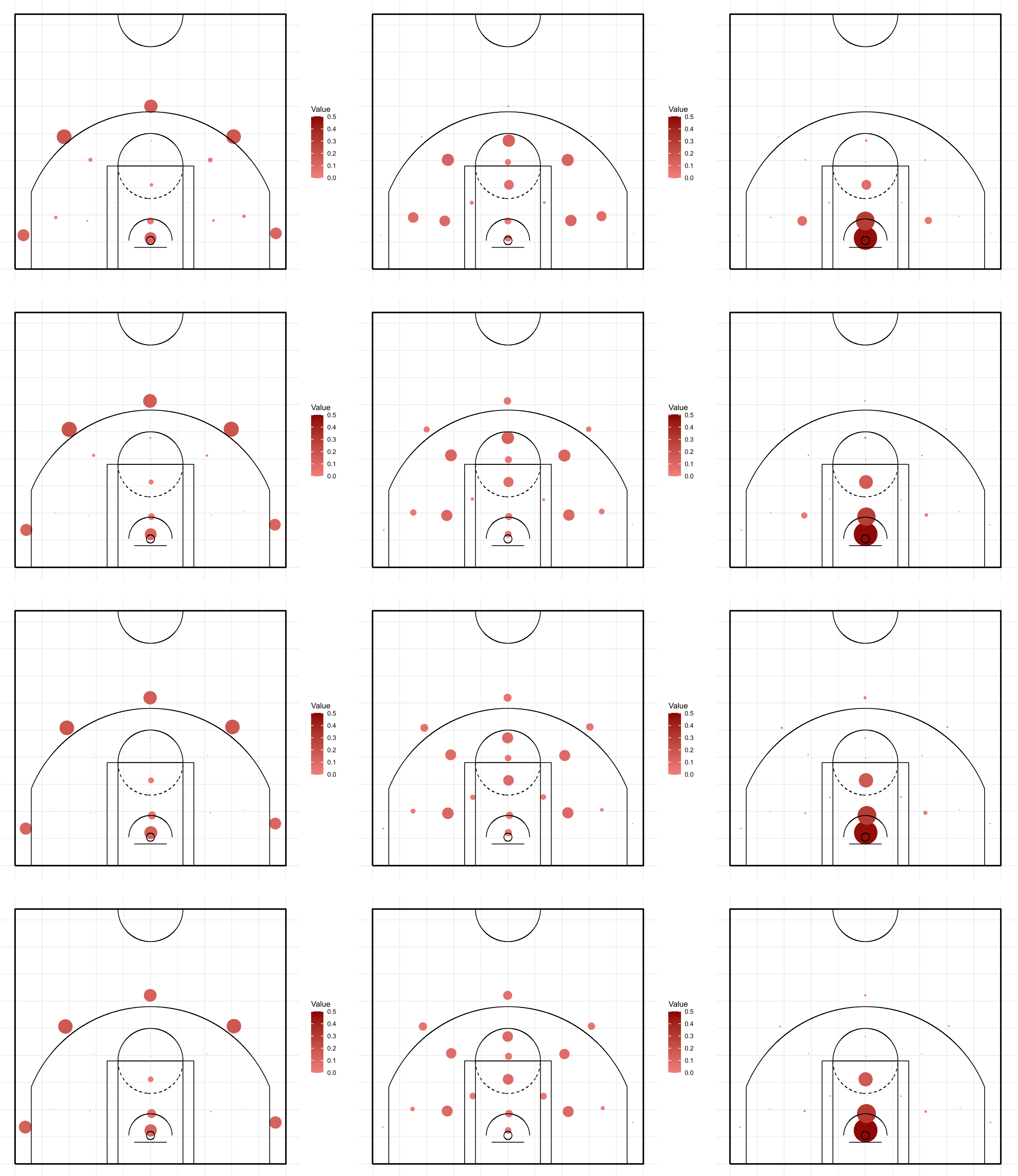

Figure 2: The first, second, third and fourth rows are for seasons 2015–16, 2017–18, 2019–20 and 2021–22 respectively. The columns correspond to the three latent factors of the location mode. Zones with darker colors and larger areas are with larger posterior means for the latent factor.

so that they add up to 1 for easier comparison) for each latent feature. We notice that for both the first and third latent features, the distribution among the four periods is quite even. However, the second latent feature, which corresponds to mid-range and paint area attempts for

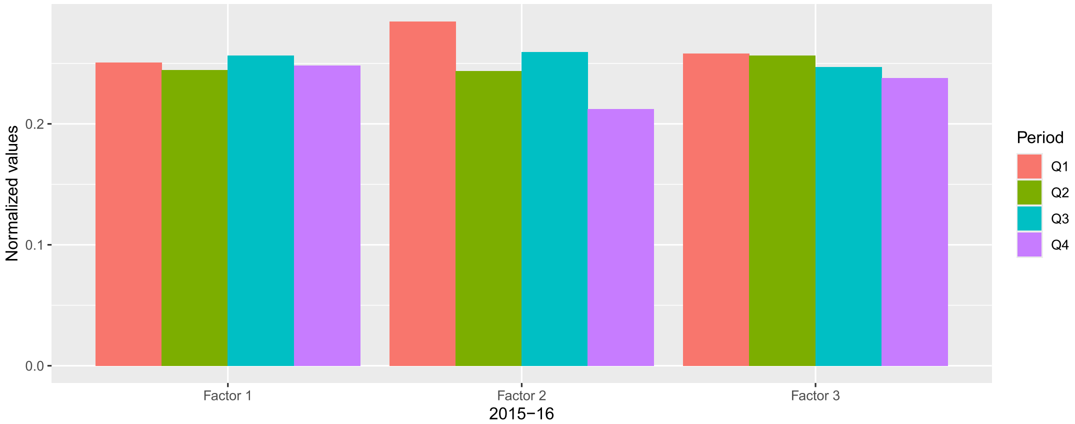

Figure 3: The three bar graphs are correspond to the three latent factors of the period mode of the regular season 2015–16. Each bar graph shows the normalized posterior means of the factor values for the four quarters.

the location mode, is less even. Notably, more attempts are made in the first quarter compared to the fourth quarter.

Lastly, we study the distribution of players across the three latent factors. Here, we will not investigate individual players. Instead, we aim to examine the differences among the five player positions: center (C), power forward (PF), small forward (SF), point guard (PG), and shooting guard (SG) (position data downloaded from https://www.basketball-reference.com). To study the difference across the five position groups, we pick the 2015–16 season for demonstration. Figure 4 reports, for each latent feature and player position, a frequency histogram of the player mode normalized posterior means (multiplied by 100 for better illustration). Again, a larger value on the x-axis represents higher importance.

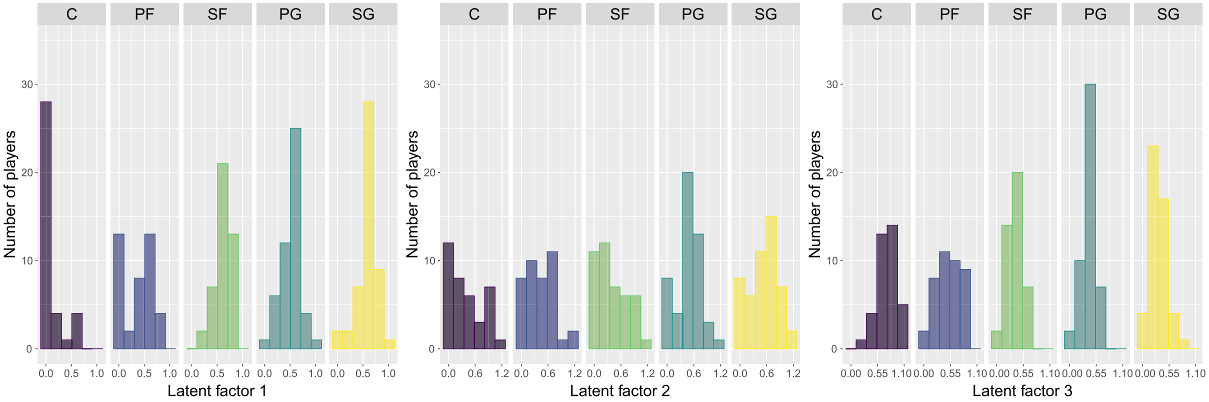

Figure 4: Each panel demonstrate histograms for the five positions for a latent factor of the player mode in the 2015–16 regular season. Each histogram reveals the frequency distribution of the normalized posterior means of factor values for players in a certain position. C, PF, SF, PG and SG correspond to the positions of Center, Power Forward, Small Forward, Point Guard and Shooting Guard respectively.

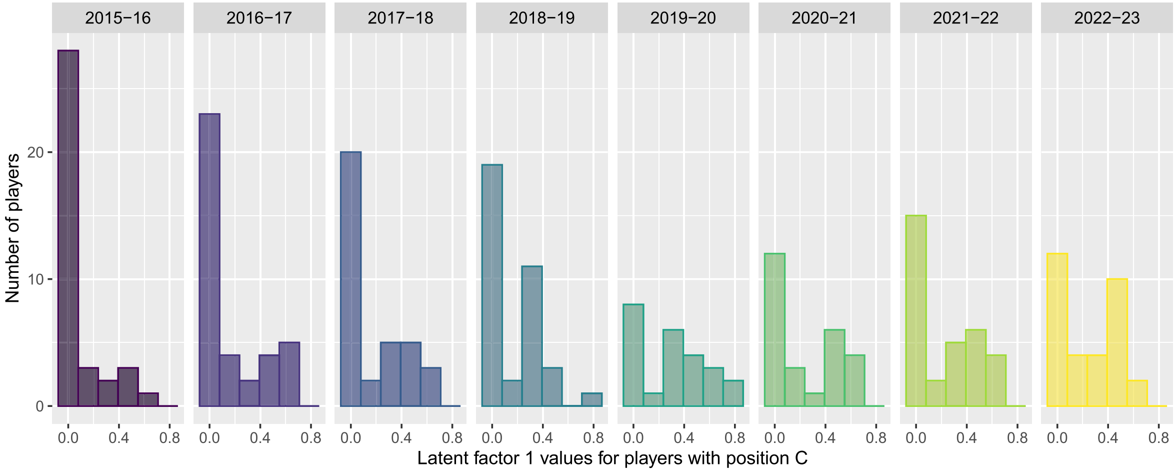

Figure 5: The eight histograms show the trend from regular season 2015–16 to 2022–23. Each histogram reveals the frequency distribution of the normalized posterior means of the latent factor 1 values for players in the Center position.

We first notice that players in the center (C) position are much more dominant in the right panel compared to the left and middle panels. This is reasonable because centers are usually the largest players controlling the restricted area near the basket. Another observation is that shooting guards (SG) are more prominent in the left panel compared to the middle and right panels. This makes sense, as shooting guards are typically the best shooters on the team and generally take more 3-point attempts.

Besides comparing across position groups, it is also interesting to observe any patterns of change for a group over time. Among the five groups, the most notable change comes from the center (C) group. Figure 5 shows the change in the distributions of the latent factor 1 values for the center group over time. Recall that latent factor 1 corresponds to 3-point attempts in the location mode. We can see that the fraction of center players with high importance in this factor grows over time. Again, this observation provides further evidence for the growing importance of 3-point throws in recent years.

4.2 Results for Shot Successes

We now turn to a discussion about the logistic regression estimation results for shot successes. We have four predictors under consideration and Table 2 summarizes the means and standard deviations of the variational distributions of the corresponding parameters. The predictor "Attempt Intensity" in Table 2 refers to the variable ϕ_i defined in equation (2) of Section 3.1. "Minutes left", "Distance" and "Home" refers respectively, to the minutes left to end of the period, distance to basket and whether the shooter is from the home team.

We notice that "Attempt Intensity", "Minutes left" and "Distance" are all significant in all seasons with their signs reasonable. "Attempt Intensity" is positively correlated to shot success as we postulate in Section 3.1. It is also natural that the success rate is higher when either "Minutes left" is larger because of less pressure or "Distance" is shorter for better accuracy. The most interesting finding comes from the "Home" variable. This predictor is significant (or at least marginally significant) for all seasons except the 2020–21 season. Recall that the unique feature

Table 2: Results of the Logistic regression part for the shot successes. Reported values are posterior means with posterior standard deviation in parentheses.

	Intercept	Attempt intensity	Minutes left	Distance	Home
2015–16	−0.1589(0.0129)	0.1334(0.0016)	0.0057(0.0015)	−0.0302(0.0005)	0.0188(0.0100)
2016–17	−0.1319(0.0128)	0.1347(0.0016)	0.0066(0.0015)	−0.0309(0.0005)	0.0421(0.0099)
2017–18	−0.1811(0.0129)	0.1618(0.0016)	0.0063(0.0015)	−0.0300(0.0005)	0.0302(0.0099)
2018–19	−0.1747(0.0125)	0.1524(0.0015)	0.0074(0.0014)	−0.0317(0.0005)	0.0312(0.0096)
2019–20	−0.1466(0.0151)	0.1504(0.0019)	0.0066(0.0017)	−0.0309(0.0006)	0.0209(0.0116)
2020–21	−0.0763(0.0144)	0.1415(0.0017)	0.0079(0.0016)	−0.0318(0.0005)	**0.0075**(**0.0109**)
2021–22	−0.0566(0.0131)	0.1502(0.0016)	0.0046(0.0015)	−0.0360(0.0005)	0.0216(0.0100)
2022–23	0.0034(0.0129)	0.1369(0.0015)	0.0062(0.0014)	−0.0368(0.0005)	0.0356(0.0098)

Remark: For the 2019–20 regular season, only the pre-bubble period data are used.

of this season is the strict post-COVID crowd restriction policies that were applied. Therefore, the effect of home fans is almost eliminated, while other home court advantages, such as travel and familiarity with the court, are still present. Thus, the result we observe can be interpreted as evidence that the home fans' effect is a decisive component of home court advantage regarding shot accuracy.

5 Discussion

Our contribution lies in both the methodological and application sides. On the methodological side, we promote the use of tensor analysis in sports data analytics, which is not common. In addition, the model and algorithm we propose incorporate data augmentation, tensor decomposition, and variational inference. On the application side, our results are consistent, reasonable, and inspiring. We consistently obtain three interpretable latent features to understand the patterns with respect to "who", "when" and "where" behind shot attempt behavior. The well-known phenomenon of the growing importance of three-point shots is also observed in our results. More interestingly, we find that a major driver of home advantage on shot successes is the presence of fans in arena.

There are several directions for extending this work. On the application side, the joint model we proposed can be applied to studies where we want to understand the pattern behind some exposures and at the same time we want to learn the effect of exposures on outcomes. For example, we may want to understand the features behind the number of infectious cases regarding location, time and disease type. At the same time, we may postulate that the exposure value may affect the chance of death and hence a joint model is appropriate.

On the technical side, possible extensions include considering other tensor decomposition methods like Tucker decomposition (Tucker, 1966) instead of CP decomposition to allow for more flexibility, and using a Multiplicative Gamma Process (MGP) prior (Takayama et al., 2022) instead of an ARD-type prior to exert more control over sparsity and the number of ranks in the decomposition. Certainly, MCMC algorithms can be developed to replace the variational EM algorithm to obtain more exact posterior inference, but the computational time will be exceptionally high for our large datasets.

Supplementary Material

We have included a supplementary section about the details of the Pólya-Gamma augmentation, the variational EM algorithm outlined in Section 3, and more graphs for the empirical analysis. The codes for downloading the shot chart data and for generating the major results are included in https://github.com/kwho1/NBA_JDS.

Acknowledgement

The authors thank the editor, associate editor and referees for their constructive comments which has led to significant improvement of this paper.

References

Bro R (1997). PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*, 38: 149–171. https://doi.org/10.1016/S0169-7439(97)00032-4

Cheng L, Tong X, Wang S, Wu YC, Poor HV (2020). Learning nonnegative factors from tensor data: Probabilistic modeling and inference algorithm. *IEEE Transactions on Signal Processing*, 68: 1792–1806. https://doi.org/10.1109/TSP.2020.2975353

Ehrlich J, Potter J (2023). Estimating the effect of attendance on home advantage in the National Basketball Association. *Applied Economics Letters*, 30(11): 1471–1482. https://doi.org/10.1080/13504851.2022.2061898

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: 94–121.

Freitas L (2021). Shot distribution in the NBA: Did we see when 3-point shots became popular? *German Journal of Exercise and Sport Research*, 51: 237–240. https://doi.org/10.1007/s12662-020-00690-7

Ganz SC, Allsop K (2024). A mere fan effect on home-court advantage. *Journal of Sports Economics*, 25(1): 30–53. https://doi.org/10.1177/15270025231200890

Harshman RA (1970). Foundations of the PARAFAC procedure: Model and conditions for an "explanatory" multimodal factor analysis. *UCLA Working Papers Phonetics*, 16: 1–84.

Hinrichy JL, Nielseny SFV, Madseny KH, Mørup M (2018). Variational bayesian partially observed non-negative tensor factorization. In: *IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6.

Hu G, Xue Y, Shen W (2022). Multidimensional heterogeneity learning for count value tensor data with applications to field goal attempt analysis of NBA players. arXiv preprint: https://arxiv.org/abs/2205.09918.

Hu G, Yang HC, Xue Y (2021). Bayesian group learning for shot selection of professional basketball players. *Stat*, 10: e4324.

Hu G, Yang HC, Xue Y, Dey DK (2023). Zero-inflated Poisson model with clustered regression coefficients: Application to heterogeneity learning of field goal attempts of professional basketball players. *The Canadian Journal of Statistics*, 51(1): 157–172. https://doi.org/10.1002/cjs.11684

Jiao J, Hu G, Yan J (2021). A bayesian marked spatial point processes model for basketball shot chart. *Journal of Quantitative Analysis in Sports*, 17(2): 77–90. https://doi.org/10.1515/jqas-2019-0106

Klami A (2014). Polya-Gamma augmentations for factor models. *JMLR: Workshop and Conference Proceedings*, 39: 112–128.

Kolda T, Bader B (2009). Tensor decompositions and applications. *Siam Review*, 51(3): 455–500. https://doi.org/10.1137/07070111X

Leota J, Hoffman D, Mascaro L, Czeisler M, Nash K, Drummond S, et al. (2022). Home is where the hustle is: the influence of crowds on effort and home advantage in the National Basketball Association. *Journal of Sports Sciences*, 40(20): 2343–2352. https://doi.org/10.1080/02640414.2022.2154933

Ma X, Brynjarsdóttir J, LaFramboise T (2024). A double Pólya-Gamma data augmentation scheme for a hierarchical negative binomial – binomial data model. *Computational Statistics and Data Analysis*, 199: 108009. https://doi.org/10.1016/j.csda.2024.108009

Miller AC, 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 (ICML)*, Xing E, Jebara T (Eds.), volume 32(1), 235–243.

Neelon B (2019). Bayesian zero-inflated negative binomial regression based on Pólya-Gamma mixtures. *Bayesian Analysis*, 14(3): 829–855. https://doi.org/10.1214/18-BA1132

Pillow J, Scott J (2012). Fully bayesian inference for neural models with negative-binomial spiking. *Advances in Neural Information Processing Systems*, 25: 1907–1915.

Polson NG, Scott JG, Windle J (2013). Bayesian inference for logistic models using Pólya-Gamma latent variables. *Journal of the American Statistical Association*, 108(504): 1339–2349. https://doi.org/10.1080/01621459.2013.829001

Rai P, Hu C, Harding M, Carin L (2015). Scalable probabilistic tensor factorization for binary and count data. In: *IJCAI'15: Proceedings of the 24th International Conference on Artificial Intelligence*, 3770–3776.

Reich B, Hodges J, Carlin B, Reich A (2006). A spatial analysis of basketball shot chart data. *The American Statistician*, 60(1): 3–12. https://doi.org/10.1198/000313006X90305

Rolland G, Vuillemot R, Bos W, Rivière N (2020). Characterization of space and time-dependence of 3-point shots in basketball. In: *MIT Sloan Sports Analytics Conference*.

Sørensen M, De Lathauwer L, Comon P, Icart S, Deneire L (2012). Canonical polyadic decomposition with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 33: 1190–1213. https://doi.org/10.1137/110830034

Soulat H, Keshavarzi S, Margrie T, Sahani M (2021). Probabilistic tensor decomposition of neural population spiking activity. In: *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Wortman Vaughan J (Eds.), volume 34, 15969–15980. 2021.

Steinfeldt H, Dallmeyer S, Breuer C (2022). The silence of the fans: The impact of restricted crowds on the margin of victory in the NBA. *International Journal of Sport Finance*, 17: 165–177. https://doi.org/10.32731/ijsf/173.082022.04

Takayama H, Zhao Q, Hontani H, Yokota T (2022). Bayesian tensor completion and decomposition with automatic CP rank determination using MGP shrinkage prior. *SN Computer Science*, 3: 225. https://doi.org/10.1007/s42979-022-01119-8

Tipping ME (2001). Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1: 211–244.

Tucker L (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3): 279–311. https://doi.org/10.1007/BF02289464

Wong-Toi W, Yang H, Shen W, Hu G (2023). A joint analysis for field goal attempts and percentages of professional basketball players: Bayesian nonparametric resource. *Journal of Data Science*, 21(1): 68–86. https://doi.org/10.6339/22-JDS1062

Yang Z, Yang T, Wang H, Zhao H, Liu D (2025). Bayesian nonnegative tensor completion with automatic rank determination. *IEEE Transactions on Image Processing*, 34: 2036–2051. https://doi.org/10.1109/TIP.2024.3459647

Yin F, Hu G, Shen W (2023). Analysis of professional basketball field goal attempts via a bayesian matrix clustering approach. *Journal of Computational and Graphical Statistics*, 32(1): 49–60. https://doi.org/10.1080/10618600.2022.2085727

Zajac T, Mikolajec K, Chmura P, Konefal M, Krzysztofik M, Makar P (2023). Long-term trends in shooting performance in the NBA: An analysis of two- and three-point shooting across 40 consecutive seasons. *International Journal of Environmental Research and Public Health*, 20(3): 1924.

Zhao Q, Zhang L, Cichocki A (2015). Bayesian CP factorization of incomplete tensors with automatic rank determination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9): 1751–1763. https://doi.org/10.1109/TPAMI.2015.2392756