Computational Statistics (2025) 40:2195–2222
https://doi.org/10.1007/s00180-024-01560-8

ORIGINAL PAPER

Can the hot hand phenomenon be modelled? A Bayesian hidden Markov approach

Gabriel Calvo[1] · Carmen Armero[1] · Luigi Spezia[2]

Received: 26 October 2023 / Accepted: 16 September 2024 / Published online: 27 September 2024

Abstract
Sports data analytics has been gaining importance over recent years as an essential topic in applied statistics. Specifically, basketball has emerged as one of the iconic sports where the use and immediate collection of data have become widespread. Within this domain, the hot hand phenomenon has sparked a significant scientific controversy, with sceptics claiming its non-existence while other authors provide evidence for it. We propose a Bayesian longitudinal hidden Markov model that examines the hot hand phenomenon in consecutive shots of a basketball team, each of which can be either missed or made. We assume two states (cold or hot) in the hidden Markov chains associated with each math and model the probability of success for each shot with regard the hidden state, the random effects related the match, and the covariates. This model is applied to real data sets of three teams from the USA National Basketball Association: the Miami Heat team and the Toronto Raptors team in the 2005–2006 season, and the Chicago Bulls in the 2022–2023 season. We show that this model is a powerful tool for assessing the overall performance of a team during a game and, in particular, for quantifying the magnitude of team streaks in probabilistic terms.

1 Introduction

Sports data analytics is a relevant topic within applied statistics that has been growing in importance in recent years (Ley and Dominicy 2020). A comprehensive review by Baumer et al. (2023) provides a completed overview of this subject, highlighting recent advancements in the area. One of the most well-known stories is fictionalised by Lewis (2004), who narrates how data analysis influenced decision-making in professional baseball signings. In basketball, the introduction of box

✉ Gabriel Calvo
gabriel.calvo@uv.es

1 Department of Statistics and Operations Research, Universitat de València, Burjassot, Spain

2 Biomathematics and Statistics Scotland, Aberdeen, UK

scores by Chadwick in the 1900s (Terner and Franks 2021) marked a significant milestone in the discipline. Importantly, Manley (1989) developed the first-ever player evaluation metric, known as individual player efficiency. Additionally, Oliver (2004) was a pioneer in the context of player and game analysis in that sport.

The study of success or failure streaks in sport is a vast issue that has its foundation in the literature of psychology, where it is known as “psychological momentum” (Adler 1981). While this subject can be explored in various fields such as politics, finance, or sports, our focus lies solely on the domain of basketball.

The definition of the *hot hand* phenomenon in basketball lacks consensus within the existing literature, although it predominantly revolves around the concept of short-term predictability (Green and Zwiebel 2018). In fact, Avugos et al. (2013) describe the *hot hand* belief in basketball as the conviction that “following a streak of successful baskets, a player is more likely to continue succeeding with the next shot as well." Similar definitions were adopted by many researchers, including Raab et al. (2012); Miller and Sanjurjo (2018); and Terner and Franks (2021). However, from our perspective, the most suitable definition, which we solely consider in this work, is that provided by Gilovich et al. (1985). The concept suggests that a player has a *hot hand* when their performance during a particular period of time is better than expected on the basis of the player’s overall abilities. Belief in this phenomenon is evident both in the media and among fans, as well as in many experts from the world of sport.

Generally, in order to demonstrate the existence or absence of this phenomenon, researchers analysed whether the probability of hitting a shot is truly conditioned by a previous streak of successful baskets. This approach was adopted by many researchers, including Gilovich et al. (1985); Raab et al. (2012); Avugos et al. (2013); and Miller and Sanjurjo (2018). The existence of the *hot hand* phenomenon remains a subject of controversy. Some authors, such as Gilovich et al. (1985), state that while most fans believe in some sort of correlation among hits and misses in basketball, the *hot hand* phenomenon is somewhat of an illusion. On the other hand, other studies contradict Gilovich’s findings. For instance, Miller and Sanjurjo (2018) recently argued that there was a clear bias in the analysis conducted by Gilovich.

This paper proposes a novel approach to model the phenomenon known as the *hot hand*. We draw on the works of authors who have explored this behaviour using latent autoregressive structures (Mews and Ötting 2023) or hidden Markov models (HMM) to analyse its occurrence in basketball (Sun 2004; Wetzels et al. 2016; Sandri et al. 2020) and other sports such as baseball (Albert and Bennett 2001), darts (Ötting et al. 2020) or football (Ötting and Andreas 2022).

HMMs are a class of statistical models whose structure depends on unobservable but temporally correlated states (Rabiner and Juang 1986; MacDonald and Zucchini 1997). According to Zhou et al. (2020), HMMs are particularly well-suited for characterising time series and longitudinal data (Maruotti 2011; Song et al. 2016). Some Bayesian examples of these models can be found in Conesa et al. (2015) and Kang et al. (2019). In this paper we propose a Bayesian longitudinal HMM (BLHMM) that analyses the *hot hand* phenomenon in consecutive basketball shots of a team, each of which can be either missed or made. Two possible latent states (*cold* or *hot*) are assumed in the hidden Markov chains of events, and the probability of success for each shot is modelled by considering the corresponding hidden state, the random effects associated with the match, and the covariates. As a novelty, this model structures each game played by the same team as an observational unit.

The *hot hand* phenomenon has traditionally been analysed from an individual point of view, focusing on assessing the behaviour of a single player. However, our study takes a novel approach by analysing the team. While the majority of the scientific literature on team performance analysis is related to finance (Bosch-Sijtsema et al. 2009; Woolley 2009), there are also examples of similar analyses in sports (Patterson et al. 2005; Thomas et al. 2019). Furthermore, the existence of a phenomenon known as "group identity" in sports has been supported by evidence (Chen et al. 2015; Heere et al. 2011).

We can higlight two important features of our model. First, this model simultaneously assumes dependence between previous shots and the current one through hidden Markov chains. It also assumes that there are moments when the performance of the team is better than others. Secondly, a *hot streak* is defined as a sequence of consecutive moments in which the team is in the *hot* state. Conversely, a *cold streak* is a sequence of consecutive shots in which the team is in the *cold* state.

The structure of the paper is as follows. Section 2 introduces the BLHMM in terms of a sampling model based on two connected processes, one latent (or hidden) and the other one observed, and a prior distribution for all of the unknown elements of the model. Section 3 deals with the posterior distribution for some characteristics of the basketball team performance such as transition probabilities, occupancy and sojourn times, and probabilities of making a basket. Section 4 applies the Bayesian model to three real data sets from different seasons of the National Basketball Association (NBA): the Miami Heat team and the Toronto Raptors team in the 2005–2006 season, and the Chicago Bulls in the 2022–2023 season. This section also includes a specific part devoted to assessing the importance of the latent structure in each of the studies. The paper concludes with a discussion, including potential directions for future research.

2 Bayesian longitudinal hidden Markov modelling

In the Bayesian framework, probabilities are always conditional and will therefore appear as such in their corresponding descriptions. This situation increases the complexity of the notation but clarifies much better the theoretical framework of Bayesian inference.

We propose a BLHMM for analysing the shooting performance of a basketball team in a season or a group of N games as a joint probabilistic model $f(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\psi})$ for the observed process $\boldsymbol{Y}$, the hidden process $\boldsymbol{Z}$, the random effects $\boldsymbol{\psi}$ associated with both processes, and the parameters and hyperparameters $\boldsymbol{\theta}$. This joint distribution can be factorised as

$$f(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\psi}) = \underbrace{f(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\psi})}_{\text{Observable process}} \underbrace{f(\boldsymbol{Z} \mid \boldsymbol{\theta}, \boldsymbol{\psi})}_{\text{Hidden process}} f(\boldsymbol{\psi} | \boldsymbol{\theta})\, \pi(\boldsymbol{\theta}). \tag{1}$$

We assume a general framework of conditional independent and identically distributed random variables among the different games of the season ($i = 1, \dots, N$), so that

$$f(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\psi}) = \left(\prod_{i=1}^{N} f(\boldsymbol{Y}_i \mid \boldsymbol{Z}_i, \boldsymbol{\theta}, \boldsymbol{\psi}) f(\boldsymbol{Z}_i \mid \boldsymbol{\theta}, \boldsymbol{\psi}) \right) f(\boldsymbol{\psi} | \boldsymbol{\theta})\, \pi(\boldsymbol{\theta}). \tag{2}$$

We describe below the different probabilistic elements in (2), with the first two expressed in terms of a generic game i.

The hidden process. This is a hidden Markov chain (HMC) that describes the *hot hand* situation of shots at the basket of the team in game i. The transition probabilities of the HMC are modeled using logistic mixed regression models.

Let $\{Z_{in}, n = 1, \dots, M_i\}$ be a HMC which represents the state, *cold* (C) or *hot* (H), of the team in shooting n at the basket in game i, where M_i is the number of shots in the game i. The transition probability matrix of the HMC is expressed as

$$P_i = \begin{pmatrix} p_i^{(CC)} & p_i^{(CH)} \\ p_i^{(HC)} & p_i^{(HH)} \end{pmatrix}, \tag{3}$$

where $p_i^{(CH)} = P(Z_{i,n+1} = H \mid Z_{in} = C, \boldsymbol{\theta}, \boldsymbol{\psi})$ is the conditional (on $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$) transition probability from the *cold* C to the *hot* state H in the game i, $p_i^{(CC)} = 1 - p_i^{(CH)}$, $p_i^{(HC)} = P(Z_{i,n+1} = C \mid Z_{in} = H, \boldsymbol{\theta}, \boldsymbol{\psi})$, and $p_i^{(HH)} = 1 - p_i^{(HC)}$. Thus, the symbol $\boldsymbol{Z}_i$ of (2) is the random vector $(Z_{i1}, \dots, Z_{iM_i})$.

The complete specification of the Markov chain needs to set a probability distribution for the initial state of the chain $\boldsymbol{\delta}_i = (\delta_i^{(C)}, \delta_i^{(H)}) = (P(Z_{i1} = C \mid \boldsymbol{\theta}, \boldsymbol{\psi}), P(Z_{i1} = H \mid \boldsymbol{\theta}, \boldsymbol{\psi}))$. We know that, from a theoretical standpoint, the values of the discrete latent states are natural numbers and not letters. However, for the sake of readability, we chose to represent the values of the variables in the chain with letters, C and H.

We assume a logistic mixed regression model for the transition probabilities as follows:

$$\begin{aligned} \operatorname{logit}(p_i^{(CH)} \mid \boldsymbol{\theta}, \boldsymbol{\psi}) &= \boldsymbol{X}_i \boldsymbol{\beta}^{(CH)} + b_i^{(CH)}, \\ \operatorname{logit}(p_i^{(HC)} \mid \boldsymbol{\theta}, \boldsymbol{\psi}) &= \boldsymbol{X}_i \boldsymbol{\beta}^{(HC)} + b_i^{(HC)}, \end{aligned} \tag{4}$$

where $\boldsymbol{X}_i$ is a vector of baseline covariates, and $\boldsymbol{\beta}^{(CH)}$ and $\boldsymbol{\beta}^{(HC)}$ are regression coefficient vectors associated with transitions from C to H and from H to C, respectively. The random effects associated with these transition probabilities within the game are $\boldsymbol{b}_i = (b_i^{(CH)}, b_i^{(HC)})$, which are assumed to follow a conditional multivariate normal distribution as $(\boldsymbol{b}_i | \boldsymbol{\theta}) \sim \mathrm{N}(\boldsymbol{0}, \Sigma_b)$, where Σ_b is a variance-covariance matrix. Note that $\boldsymbol{\beta} = (\boldsymbol{\beta}^{CH)}, \boldsymbol{\beta}^{(HC)})$ and Σ_b are parameters and hyperparameters included in the generic $\boldsymbol{\theta}$ vector. The same applies to the random effects b_i's as elements of $\boldsymbol{\psi}$.

The observed process. A Bernoulli longitudinal model is used to assess the success or failure of shots at the basket in each game i, with probability depending on the state (*cold* or *hot*) of the team. These state-dependent probabilities are modelled through logistic mixed regression models.

Let the random variable Y_{in} be the success (1) or failure (0) of shot n in game i. Thus, the symbol $\boldsymbol{Y}_i$ of (2) is the random vector $(Y_{i1}, \dots, Y_{iM_i})$. Each variable in Y_{in} is a Bernoulli whose probability depends on the latent state Z_{in} as follows,

$$(Y_{in} \mid \boldsymbol{\theta}, \boldsymbol{\psi}, Z_{in} = C) \sim \text{Bern}(\gamma_{in}^{(C)}),$$
$$(Y_{in} \mid \boldsymbol{\theta}, \boldsymbol{\psi}, Z_{in} = H) \sim \text{Bern}(\gamma_{in}^{(H)}),$$

where $\gamma_{in}^{(H)}$ and $\gamma_{in}^{(C)}$ are the probability parameter associated with the *hot* and *cold* states, respectively. The probability vectors $\boldsymbol{\gamma}_i^{(C)} = (\gamma_{i1}^{(C)}, \dots, \gamma_{iM_i}^{(C)})$ and $\boldsymbol{\gamma}_i^{(H)} = (\gamma_{i1}^{(H)}, \dots, \gamma_{iM_i}^{(H)})$ can be modelled through a logistic mixed regression model expressed as

$$\begin{aligned} \text{logit}(\boldsymbol{\gamma}_i^{(C)} \mid \boldsymbol{\theta}, \boldsymbol{\psi}) &= \mathcal{X}_i \boldsymbol{\alpha}^{(C)} + \Xi_i \boldsymbol{a}_i^{(C)}, \\ \text{logit}(\boldsymbol{\gamma}_i^{(H)} \mid \boldsymbol{\theta}, \boldsymbol{\psi}) &= \mathcal{X}_i \boldsymbol{\alpha}^{(H)} + \Xi_i \boldsymbol{a}_i^{(H)}, \end{aligned} \tag{5}$$

where $\mathcal{X}_i$ and Ξ_i are the design matrices for the fixed effects $\boldsymbol{\alpha} = (\boldsymbol{\alpha}^{(C)}, \boldsymbol{\alpha}^{(H)})'$ and the random effects $\boldsymbol{a}_i = (\boldsymbol{a}_i^{(C)}, \boldsymbol{a}_i^{(H)})'$, respectively. Matrix $\mathcal{X}_i$ includes the covariates associated with the success of shooting, for instance the distance to the basket or the type of shot. Random effects $\boldsymbol{a}_i$ are conditionally independent and assumed to follow a conditional multivariate normal distribution $(\boldsymbol{a}_i|\Sigma_a) \sim \text{N}(\boldsymbol{0}, \Sigma_a)$. Note that parameters $\boldsymbol{\alpha}$ and hyperparameters in Σ_a are part of $\boldsymbol{\theta}$ (along with $\boldsymbol{\beta}$, Σ_b, and $\boldsymbol{\delta} = (\boldsymbol{\delta}_1, \dots, \boldsymbol{\delta}_N)$), whereas the $\boldsymbol{a}_i$'s of $\boldsymbol{\psi}$ (along with $\boldsymbol{b} = (\boldsymbol{b}_1, \dots, \boldsymbol{b}_N)$). Therefore, $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \Sigma_a, \Sigma_b, \boldsymbol{\delta})$ and $\boldsymbol{\psi} = (\boldsymbol{a}, \boldsymbol{b})$, where $\boldsymbol{a} = (\boldsymbol{a}_1, \dots, \boldsymbol{a}_N)$. Figure 1 summarises the complete model, and illustrates the relationship between all its elements.

The Bayesian model is completed with the specification of a prior distribution $\pi(\boldsymbol{\theta})$ over the parameters and hyperparameters of the model.

3 Basketball team performance

Bayes' theorem combines the information provided by the prior distribution and the likelihood function to obtain the posterior distribution $\pi(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathcal{D})$, where $\mathcal{D}$ stands for the data and includes the observations $\boldsymbol{y}_i$, (i.e., the realisations of the process $\boldsymbol{Y}_i$ of (1), that is, the baskets made and missed of the shots of the games) and a few covariates. This posterior contains all information of the behaviour of the model and allows direct knowledge of the performance of the team both at a general level (such as the coefficients of the logistic regressions associated with either the hidden transition probabilities or the state dependent probabilities of making a basket) and at a specific level of the different games in the analysed championship. At the specific level, the posterior distribution provides knowledge on the random effects associated with a particular game on transition probabilities or probabilities of making a

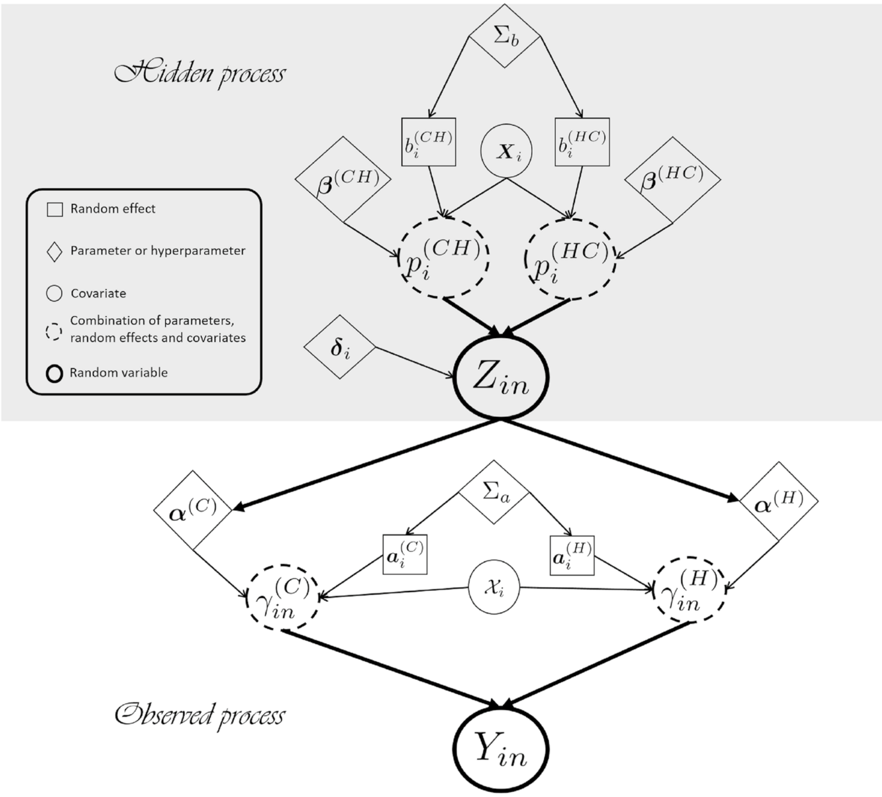

Fig. 1 Graphical representation of the Bayesian longitudinal hidden Markov model. The rectangle with a grey background illustrates the model's hidden process, whereas the area in white depicts the observed process

basket. Beyond the valuable information provided by the posterior distribution of the elements of our model, such as the transition probabilities ($p_i^{(CH)}$ and $p_i^{(HC)}$) or the probabilities of making a basket ($\boldsymbol{\gamma}_i^{(C)}$ and $\boldsymbol{\gamma}_i^{(H)}$), it is intriguing to learn about the behaviour of the hidden process of the model. These include, among others, the t-step transition probabilities, the stationary distributions, and the occupancy and sojourn times. These features will help us to better assist in comprehending the shooting performance of a team during a game. Since the stochastic behaviours of these quantities depends on $(\boldsymbol{\theta}, \boldsymbol{\psi})$, we can apply the Bayesian paradigm and compute the posterior distribution of each of them from the posterior distribution $\pi(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathcal{D})$. We discuss these valuable and interesting characteristics of the model below.

3.1 Transition probabilities

Transition probabilities are conditional probabilities that assess the uncertainty of the chain of moving (or not) from a state to a different one in a single step. In our model, they would be to abandon or remain in the current hot state on the next shot at the basket or to abandon or remain in the current cold state, also on the next shot of the game.

According to (3) and (4), and taking into account that the inverse of the logit function is expressed as a ratio of exponential functions ($\text{logit}(p) = u \;\rightarrow\; p = \exp(u)/(1 + \exp(u))$), we would express p_i^{CH} as

$$p_i^{(CH)} = P(Z_{i\,n+1} = H \mid Z_{in} = C, \boldsymbol{\theta}, \boldsymbol{\psi}) = \frac{\exp\{\boldsymbol{X}_i \boldsymbol{\beta}^{(CH)} + b_i^{(CH)}\}}{1 + \exp\{\boldsymbol{X}_i \boldsymbol{\beta}^{(CH)} + b_i^{(CH)}\}}.$$

Additionally, if we also want to evaluate these probabilities in a general way, without distinguishing between games, we can obtain the marginal transition probability

$$P(Z_{i\,n+1} = H \mid Z_{in} = C, \boldsymbol{\theta}) = \int P(Z_{i\,n+1} = H \mid Z_{in} = C, \boldsymbol{\theta}, \boldsymbol{\psi}) f(\boldsymbol{\psi} \mid \boldsymbol{\theta})\; \mathrm{d}\,\boldsymbol{\psi},$$

which is independent of the game and only depends on $\boldsymbol{\theta}$. Remember that, as indicated above, here the random effects are multivariate normally distributed, $f(\boldsymbol{\psi} \mid \boldsymbol{\theta}) = \mathrm{N}(0, \Sigma_b)$. The same idea can be applied to marginalise out the rest of transition probabilities of the chain.

3.2 Probabilities of making a basket

The probability of making a basket on the n-th shot of game i when the team is in the *cold* or the *hot* state is, as defined in (5), $\gamma_{in}^{(C)}$ and $\gamma_{in}^{(H)}$, respectively. These parameters have been expressed as

$$\gamma_{in}^{(C)} = P(Y_{in} = 1 \mid Z_{in} = C, \boldsymbol{\theta}, \boldsymbol{\psi}),$$
$$\gamma_{in}^{(H)} = P(Y_{in} = 1 \mid Z_{in} = H, \boldsymbol{\theta}, \boldsymbol{\psi}).$$

If interested in a general performance of the team, independently of the game, we can also integrate out the random effects associated with the games, and compute the conditional marginal distributions, analogously to those presented for the transition probabilities:

$$P(Y_{in} = 1 \mid Z_{in} = C, \boldsymbol{\theta}) = \int P(Y_{in} = 1 \mid Z_{in} = C, \boldsymbol{\theta}, \boldsymbol{\psi}) f(\boldsymbol{\psi} \mid \boldsymbol{\theta})\; \mathrm{d}\,\boldsymbol{\psi},$$
$$P(Y_{in} = 1 \mid Z_{in} = H, \boldsymbol{\theta}) = \int P(Y_{in} = 1 \mid Z_{in} = H, \boldsymbol{\theta}, \boldsymbol{\psi}) f(\boldsymbol{\psi} \mid \boldsymbol{\theta})\; \mathrm{d}\,\boldsymbol{\psi},$$

where, in this case, $f(\boldsymbol{\psi} \mid \boldsymbol{\theta}) = f(\boldsymbol{a}_i \mid \Sigma_a) = \mathrm{N}(\boldsymbol{0}, \Sigma_a)$.

In addition, we can compute the probability of a basket on the n-th shot of game i when the state of the team is unknown, i. e. $P(Y_{in} = 1 \mid \boldsymbol{\theta}, \boldsymbol{\psi})$, and it is calculated integrating out the probability associated with the hidden states as

$$P(Y_{in} = 1 \mid \boldsymbol{\theta}, \boldsymbol{\psi}) = \gamma_{in}^{(C)} P(Z_{in} = C \mid \boldsymbol{\theta}, \boldsymbol{\psi}) + \gamma_{in}^{(H)} P(Z_{in} = H \mid \boldsymbol{\theta}, \boldsymbol{\psi}), \tag{6}$$

where

$$\begin{aligned} P(Z_{in} = C \mid \boldsymbol{\theta}, \boldsymbol{\psi}) = {} & P(Z_{in} = C \mid Z_{i1} = C, \boldsymbol{\theta}, \boldsymbol{\psi})\, \delta_i^{(C)} \\ & + P(Z_{in} = C \mid Z_{i1} = H, \boldsymbol{\theta}, \boldsymbol{\psi})\, \delta_i^{(H)}. \end{aligned}$$

Recall that $\delta_i^{(C)} = P(Z_{i1} = C \mid \boldsymbol{\theta}, \boldsymbol{\psi})$ and $\delta_i^{(H)} = P(Z_{i1} = H \mid \boldsymbol{\theta}, \boldsymbol{\psi})$. Furthermore, the calculation of the t-step transition probabilities will be described subsequently.

3.3 *t*-step transition probabilities and stationary distribution

The t-step transition probabilities of the chain for game i are defined as

$$\begin{aligned} p_i^{(CH)}(t) &= P(Z_{i\,n+t} = H \mid Z_{in} = C, \boldsymbol{\theta}, \boldsymbol{\psi}),\ p_i^{(CC)}(t) = 1 - p_i^{(CH)}(t), \\ p_i^{(HC)}(t) &= P(Z_{i\,n+t} = C \mid Z_{in} = H, \boldsymbol{\theta}, \boldsymbol{\psi}),\ p_i^{(HH)}(t) = 1 - p_i^{(CH)}(t). \end{aligned}$$

From the Markov chain theory (e.g., Kulkarni (2016)) we know that the matrix $P_i^{(t)}$ that collects these probabilities can be calculated from the transition probability matrix P_i in (3) as

$$P_i^{(t)} = \frac{1}{p_i^{(CH)} + p_i^{(HC)}} \begin{bmatrix} p_i^{(HC)} & p_i^{(CH)} \\ p_i^{(HC)} & p_i^{(CH)} \end{bmatrix}$$

$$+ \frac{(1 - p_i^{(CH)} - p_i^{(HC)})^t}{p_i^{(CH)} + p_i^{(HC)}} \begin{bmatrix} p_i^{(CH)} & -p_i^{(CH)} \\ -p_i^{(HC)} & p_i^{(HC)} \end{bmatrix}.$$

The Bayesian framework, as explained above, allows us to calculate the posterior distribution of each of the t-step transition probabilities of the chain for each specific game as well as for a generic game.

Since the HMC for game i has a finite state space, and it is irreducible and aperiodic, its stationary distribution will be the limit distribution, $\boldsymbol{\Delta}_i = (\Delta_i^{(C)}, \Delta_i^{(H)})$, which turns out to be

$$\Delta_i^{(C)} = \frac{p_i^{(HC)}}{p_i^{(CH)} + p_i^{(HC)}}, \quad \Delta_i^{(H)} = \frac{p_i^{(CH)}}{p_i^{(CH)} + p_i^{(HC)}}. \tag{7}$$

Note that the vector $\boldsymbol{\Delta}_i$ is also the left eigenvector of the transition probability matrix P_i, associated with the eigenvalue one.

3.4 Occupancy times

The occupancy times in a chain refer to the expected number of visits the process makes to each state of the chain in a given number of transitions (Kulkarni 2016). This is an indicator of the frequency with which the chain visits the different states of the process. In particular, for each game i we represent by $m_i^{(jk)}(t) = \mathrm{E}\,(V_i^{(jk)}(t) \mid \boldsymbol{\theta}, \boldsymbol{\psi})$ the conditional (on $(\boldsymbol{\theta}, \boldsymbol{\psi})$) expected value of the number of visits $V_i^{(jk)}(t)$ to state k from state j in the first t transitions of the chain. In the case of our HMC, these conditional expectations are expressed (Kulkarni 2016) in terms of the transition probabilities as follows

$$\begin{aligned}\begin{bmatrix} m_i^{(CC)}(t) & m_i^{(CH)}(t) \\ m_i^{(HC)}(t) & m_i^{(HH)}(t) \end{bmatrix} &= \frac{t+1}{p_i^{(CH)} + p_i^{(HC)}} \begin{bmatrix} p_i^{(HC)} & p_i^{(CH)} \\ p_i^{(HC)} & p_i^{(CH)} \end{bmatrix} \\ &+ \frac{1-(p_i^{(HC)} + p_i^{(CH)} - 1)^{(t+1)}}{(p_i^{(CH)} + p_i^{(HC)})^2} \begin{bmatrix} p_i^{(CH)} & -p_i^{(CH)} \\ -p_i^{(HC)} & p_i^{(HC)} \end{bmatrix}. \end{aligned} \tag{8}$$

3.5 Sojourn times

The sojourn time of the basketball team in state j, where $j \in \{C, H\}$, during game i, denoted by $\tau_i^{(j)}$, is defined as the number of shots required for the team to transition out of state j for the first time, i.e.

$$\begin{aligned} P(\tau_i^{(j)} = t \mid \boldsymbol{\theta}, \boldsymbol{\psi}) &= P(Z_{i\,n+t} \neq j, Z_{i\,t^*} = j, t^* = n+1, \dots, n+t-1 \mid Z_{in} = j, \boldsymbol{\theta}, \boldsymbol{\psi}) \\ &= [p_i^{(jj)}]^{t-1}\,(1 - p_i^{(jj)}),\ \ t = 0, 1, \dots \end{aligned}$$

Therefore, the conditional distribution of the sojourn time in state C (H) of the game i will be a geometric distribution of parameter $p_i^{(CC)}$ (or $p_i^{(HH)}$).

Consequently, we can compute the posterior distribution for $\tau_i^{(C)}$ and $\tau_i^{(H)}$

$$\begin{aligned} &\pi((\tau_i^{(C)} = t \mid \boldsymbol{\theta}, \boldsymbol{\psi}) \mid \mathcal{D}), \\ &\pi((\tau_i^{(H)} = t \mid \boldsymbol{\theta}, \boldsymbol{\psi}) \mid \mathcal{D}), \end{aligned} \tag{9}$$

or its marginal posterior distribution integrating out the random effects as above.

4 The hot hand model in action: basketball in the NBA

We apply the proposed model to three different NBA basketball teams: a past successful team, a past unsuccessful team, and a current team with an intermediate level. Specifically, we consider the Miami Heat team in the 2005–2006 NBA season, the Toronto Raptors team in the same season, and the Chicago Bulls team in the 2022–2023 season, respectively. We note that the analysis for the first example will be more detailed than in the other two cases to avoid unnecessary repetitions. Moreover, the three analyses were carried out by applying the same BLHMM model, which will be developed in the following section. The full analysis, performed by an R code (R version 4.0.5), and the data are available as supplementary material at https://github.com/gcalvobayarri/hot_hand_model.git.

4.1 Miami Heat shooting in the 2005–2006 NBA season

On 20 June 2006, the Miami Heat team defeated the Dallas Mavericks in the finals and won their first NBA championship. The analysis of the information play-by-play provided by the NBA (2020) about the team is very interesting and gives many clues on Miami successful season. Specifically, we are interested in analysing their shooting performance. We consider all of the *field shots* and *free throws* of all games of the championship, a total of 11 042 shots in 105 games, 5 922 of which were baskets made. For each of these shots the distance in feet from the shooting position to the basket, the game, and the sequential order per game in which each shot was taken were taken into account.

Figure 2 is a shot chart (Zuccolotto and Manisera 2020) of the Miami Heat's *field shots* during the 2005–2006 NBA season. It can be clearly seen that most shots are taken just beyond the *three-point line* or just below the basket. On the other hand, as it might be expected, most of the shots the team took from further away were missed.

4.1.1 Modelling the shooting performance of the Miami Heat team

Now, we apply the Bayesian model presented in Sect. 2. The sampling model is also defined in terms of the two sub-processes, an HMC that accounts for the *hot* and

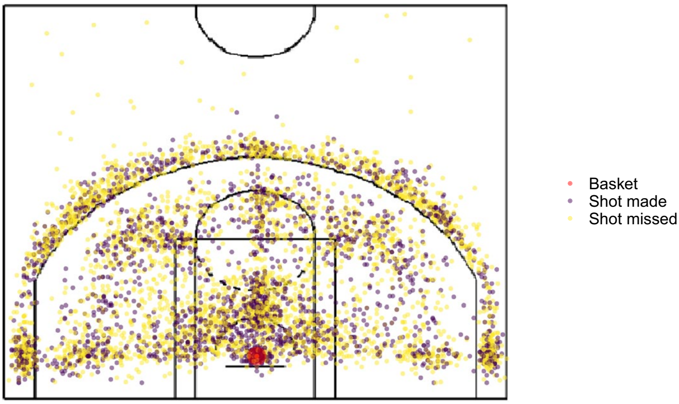

Fig. 2 Shot chart of the Miami Heat during the 2005–2006 season, showing the location of shots on the court, and indicating the shot results and the exact position of the basket

cold states, and an observed Bernoulli longitudinal variable that assesses the success or failure of each shot in relation to the *hot* or *cold* state of the chain.

The transition probabilities of the HMC $\{Z_{in}, n = 1, \ldots, M_i\}$ defined for each game i, $i = 1, \ldots, N = 105$, are described through the logistic mixed regression models

$$\text{logit}(p_i^{(CH)} \mid \boldsymbol{\theta}, \boldsymbol{\psi}) = \beta_{CH} + b_i^{(CH)},$$
$$\text{logit}(p_i^{(HC)} \mid \boldsymbol{\theta}, \boldsymbol{\psi}) = \beta_{HC} + b_i^{(HC)},$$

which express each of the two relevant probabilities, $p_i^{(CH)}$ and $p_i^{(HC)}$, in terms of an intercept (indicating the average logit of the transition probabilities), common to all games and a specific random effect, $b_i^{(CH)}$ and $b_i^{(HC)}$, associated with the game. These random effects are mutually independent and conditionally normally distributed as $(b_i^{(CH)}|\sigma_{CH}) \sim \text{N}(0, \sigma^2_{CH})$ and $(b_i^{(HC)}|\sigma_{HC}) \sim \text{N}(0, \sigma^2_{HC})$. The initial probability vector $\boldsymbol{\delta} = (\delta_C, \delta_H)$ is also considered.

For the observed part of the sampling model, the probabilities $\gamma_{in}^{(H)}$ and $\gamma_{in}^{(C)}$ of making a basket on shot n in game i when the team is respectively in the *hot* or the *cold* state are described as

$$\text{logit}(\gamma_{in}^{(C)}) = \alpha_C + \alpha_d X_{in} + \alpha_{FT} I_{FT}(in) + a_i,$$
$$\text{logit}(\gamma_{in}^{(H)}) = \alpha_H + \alpha_d X_{in} + \alpha_{FT} I_{FT}(in) + a_i,$$

where α_C and α_H are common intercepts for the probability associated with the *cold* and the *hot* state respectively, α_d and α_{FT} are the regression coefficients associated with the covariates X_{in} and $I_{FT}(in)$ that respectively describe the distance from the basket in the n-th shot of the game i and an indicator variable that is 1 when the n-th shot of game i is a *free throw* and zero otherwise. Random effects a_i are assumed normally distributed, i. e. $(a_i|\sigma_a) \sim \text{N}(0, \sigma_a^2)$, and conditional independent given σ_a, for any i.

In order to complete the specification of the BLHMM model we need to elicit a prior distribution for the parameters and hyperpameters of the model. We assume prior independence within a minimally informative prior scenario. In particular, we select a beta distribution for the initial probability $\pi(\delta_C) = \text{Be}(1, 1)$, and uniform distributions for the standard deviation parameters $\pi(\sigma_a) = \pi(\sigma_{CH}) = \pi(\sigma_{HC}) = \text{U}(0, 10)$. The gamma distribution could be an alternative prior distribution, particularly for precision parameters. However, Gelman (2006), and Martínez-Beneito and Botella-Rocamora (2019) argue that the gamma distribution exhibits undesirable behaviour (relative to its percentiles) or converges to an improper distribution, resulting in an improper posterior distribution.

A wide normal distribution is selected for the two regression coefficients, α_d and α_{FT}, associated with the covariates as well as for the two fixed-effect parameters present in the transition probabilities, β_{CH} and β_{HC}. That is, $\pi(\beta_{CH}) = \pi(\beta_{HC}) = \pi(\alpha_d) = \pi(\alpha_{FT}) = \text{N}(0, 10^2)$. This choice is minimally informative although proper, ensuring that the posterior distribution remains proper as well. Finally, to avoid the problem of identifiability because of the label switching issue (see among others McCulloch and Tsay, 1994; Frühwirth-Schnatter, 2001; Spezia, 2009), we include the following restriction in the α_C's and α_H's prior distributions

$$\pi(\alpha_C) = \pi(\alpha_H) = \mathrm{N}(0, 10^2),\ \ \alpha_C \leq \alpha_H.$$

4.1.2 Posterior distribution

The complexity of the BLHMM model makes the posterior distribution analytically intractable. We approximated it by means of Markov chain Monte Carlo (MCMC) sampling methods (Tanner 2012) via the JAGS software (Plummer 2003). Three parallel chains were run for 30 000 iterations each after 30 000 iterations as a burn-in. In addition, based on the estimated autocorrelation in the sample, and in order to reduce it, the chains were also thinned at every 30th iteration.

The posterior distribution provides useful information on the general performance of the Miami Heat team in a game. Table 1 shows the posterior summary for the parameters and hyperparameters included in the model. Furthermore, we computed a standard convergence diagnostic measure, $\hat{R}$ (Gelman et al. 2013), for each parameter and hyperparameter. It is noteworthy that all $\hat{R}$ values were close to 1, which indicates good convergence of the MCMC algorithm.

First, we focus on the elements of the hidden model. The posterior means of the common intercepts included in the transition probabilities β_{CH} and β_{HC} clearly indicate a higher probability of moving from the *hot* state to the *cold* state than from the *cold* state to the *hot* state. In particular, the posterior mean of the probability of switching from C to H in a generic game is 0.38, and from H to C is 0.59. Thus, remaining in the *cold* state in one transition is more likely than switching to the hot, or than remaining in the *hot* state. This can also be observed in Fig. 3, which displays the posterior distribution of the transition probabilities for a generic game.

Table 1 Posterior summaries (mean, standard deviation, 95% credible interval) and diagnostic statistic $\hat{R}$ for the parameters and hyperparameters of the Miami Heat shooting performance *hot hand* model

Subprocess		mean	sd	$q_{0.025}$	$q_{0.975}$	$\hat{R}$
Hidden	β_{CH}	−0.49	0.05	−0.58	−0.39	1.001
	β_{HC}	0.38	0.06	0.27	0.49	1.000
	δ_C	0.55	0.06	0.43	0.68	1.000
	σ_{CH}	0.07	0.05	0.00	0.18	1.011
	σ_{HC}	0.10	0.07	0.00	0.25	1.048
Observed	α_C	−0.15	0.07	−0.29	−0.01	0.999
	α_H	12.59	1.19	10.52	14.97	1.014
	α_d	−0.42	0.05	−0.51	−0.33	1.017
	α_{FT}	6.37	0.69	5.16	7.75	1.017
	σ_a	0.15	0.08	0.00	0.31	1.024

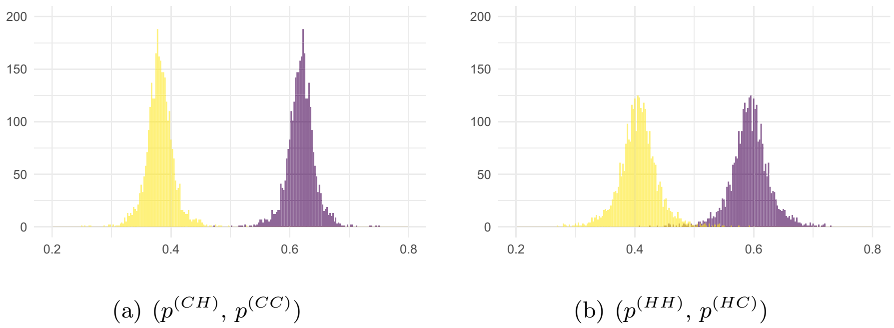

Fig. 3 Posterior distribution histograms (with frequencies) of the transition probabilities (**a**) $p^{(CC)}$ (in purple) and $p^{(CH)}$ (in yellow), and (**b**) $p^{(HC)}$ (in purple) and $p^{(HH)}$ (in yellow) integrating the random effects associated with the game

Moreover, the expected posterior probability of being in the *cold* state when the game starts is around 0.55 according to the posterior mean of δ_C.

On the other hand, for the observed Bernoulli sub-process the posterior distribution of the common intercepts α_C and α_H shows a large difference of the magnitude of α_H with respect to α_C as seen in their means $\text{E}(\alpha_C \mid \mathcal{D}) = -0.15$ and $\text{E}(\alpha_H \mid \mathcal{D}) = 12.59$. Following the same procedure as above for the transition probabilities, we have a posterior expected probability 0.46 of making a shot from a distance of 0 feet to the basket when the team is in the *cold* state, and almost 1 in the case of the *hot* state when the distance to the basket is 0.25 feet.

The two covariates considered in the modelling are relevant. The coefficient parameter associated with the distance to the basket, α_d, is completely negative, $\text{E}(\alpha_d \mid \mathcal{D}) = -\,0.42$, with a small posterior standard deviation, $\text{SD}(\alpha_d \mid \mathcal{D})=0.03$, which means that the probability of successful shooting decreases when the distance is further away. In addition, the credible interval of the coefficient associated with the *free throws* has a large positive expected value, $\text{E}(\alpha_{FT} \mid \mathcal{D}) = 6.37$. Therefore, as it might be expected, making a *free throw* is easier than making a *field goal* from that distance, likely owing to the absence of opposition. Finally, the random effects associated with the games included in the success probability are also important since their associated standard deviation σ_a has a posterior mean 0.15. It can be stated that a relevant portion of the variability can be attributed to the variations in the scoring efficiency of the team in the different games.

All information we extract from the posterior distribution provides an immense number of possibilities to assess different aspects of the performance of the team. We can discuss some of them right now.

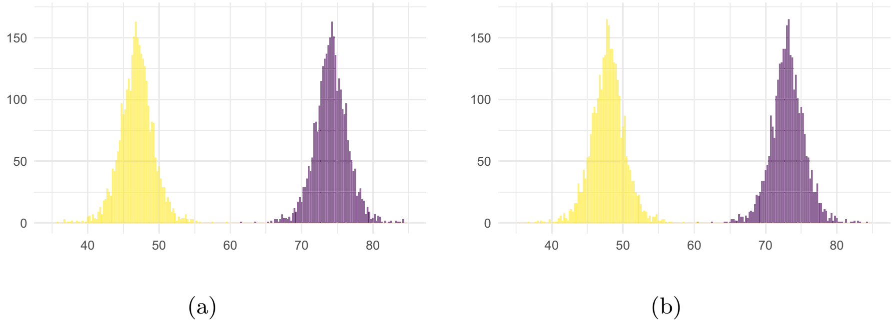

Fig. 4 Posterior distribution histograms (with frequencies) of the occupancy times in the states *hot* (in yellow) and *cold* (in purple) in a game with $N = 120$ shots that starts (**a**) in the *cold* and (**b**) in the *hot* state

4.1.3 Occupancy times

Occupancy times admit many different posterior outputs. Here, for illustration we only focus on the posterior distribution of occupancy times in the *cold* and the *hot* state in a generic game with $N = 120$ shots. Figure 4 shows the posterior distribution of the occupancy times in the *cold* and the *hot* state and the two possibilities of initial state, *cold* and *hot*.

It is interesting to note that the team spends more time in the *cold* state than in the *hot* (with a posterior mean of 74.05 and 46.85, respectively), and that the state in which the team starts playing is practically irrelevant because there is hardly any difference between the two figures.

4.1.4 Stationary distributions

Although we could consider a stationary distribution of the chain associated with each game of the season, we focus on the posterior distribution $\pi((\Delta^{(C)}, \Delta^{(H)} \mid \boldsymbol{\theta}) \mid \mathcal{D})$ of the stationary distribution of the chain corresponding to a generic game. Table 2 shows a summary of the posterior distribution of the stationary distribution for the *cold* and *hot* state.

In summary, the probability of being in the *cold* state for the Miami Heat team in the 2005–2006 season was around 0.61 (thus, the probability of being in the *hot* state was 0.39). Both distributions have very little variability.

Table 2 Posterior summaries (mean, standard deviation and 95% credible interval) for the stationary distribution of the hidden Markov chain associated with a generic game

	Mean	sd	$q_{0.025}$	$q_{0.975}$
$\Delta^{(C)}$	0.61	0.02	0.57	0.65
$\Delta^{(H)}$	0.39	0.02	0.35	0.43

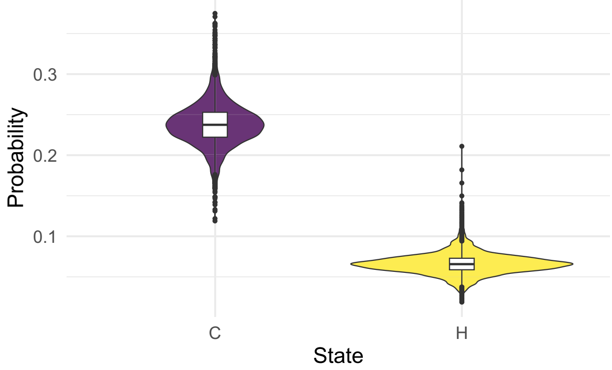

Fig. 5 Violin and box plot of the posterior probability distribution for a *cold* and a *hot* streak

4.1.5 Sojourn times

We consider a *cold* or a *hot* streak when the team stays in the same state for more than three shots. This number of shots is an arbitrary choice that can only be justified in order to illustrate the potential of our modelling and improve the understanding of the behaviour of the team. Figure 5 shows a violin plot of the posterior distribution of the probability for a *cold* and for a *hot* streak in a generic game of the Miami Heat. There, we observe the probability of a *cold* streak (around 0.25) is nearly three times higher than the probability of a *hot* streak (less than 0.1). The much smaller amplitude of the *hot* distribution compared to the *cold* one is also evident.

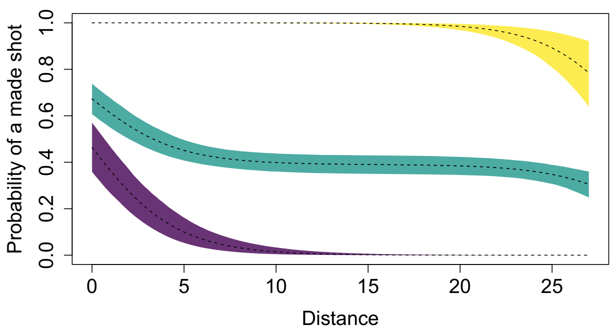

Fig. 6 Posterior means and 95% credible interval for the probabilities of making a basket depending on the distance, in feet, in three different scenarios. Purple curve represents the case when the team is in the *cold* state, the yellow plot is for the *hot* state, and the green plot is the general case when the current state of the team is unknown

4.1.6 Probability of making a basket

The probability of making a basket depends on the state the team is currently visiting, the distance to the basket at which the shot is made, and whether the shot is a *free throw* or not. Figure 6 shows the posterior mean and the 95% credible interval for the probability of making a basket depending on the distance at which the shot (non-*free throw*) is made when the team is in the *cold* state, *hot* state or in the case where the state of the team is unknown.

When the team is in the *cold* state, one can see that for easy shots (i.e. shots close to the basket) there is a probability of around 0.5 of making a basket. However, from a distance of 10 feet, it is almost impossible to succeed. On the other hand, when the team is in the *hot* state, it is very likely to make a successful shot up to 15 feet. Then, from this distance, the probability starts to decrease slowly. Further, for a shot, in which the current state is unknown, the probability of success is also negatively related to the distance to the basket. Shots closer to the basket have a probability of success around 0.7, whereas for intermediate shots this probability stabilises at around 0.5. Finally, the probability of making a three-point shot drops to 0.4. It is important to mention that when we consider that the state of the team, hot or cold, is unknown ($P(Y_{in} = 1 \mid \boldsymbol{\theta}, \boldsymbol{\psi})$), it does not mean that we are considering the existence of a 'third' state but that we are expressing uncertainty to the current state, hot or cold, of the team. As we recall, to calculate the probability of making a basket when we do not know what exactly is the current state in which the team is, we integrate out the hidden states, as shown in the expression (6).

It can be interesting to visualise the team's performance over the different games of the season. In that sense, Figure 7 shows the mean and the 95th percentile credible interval of the probability of hitting a shot from 25 feet, in which the team's state is unknown, in each of the games of the season. We observe a very marked regularity in all of the games of the season, very little variability among games and a very stable behaviour in all phases of the season.

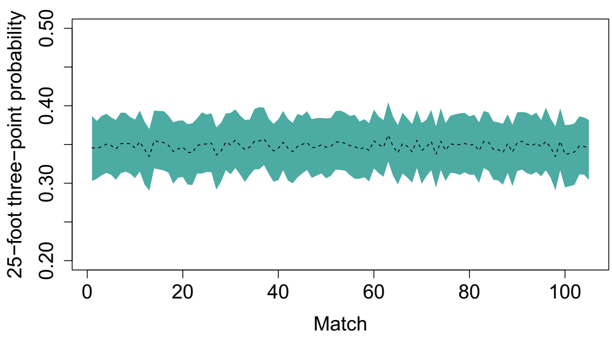

Fig. 7 Posterior mean and credible interval for the probability of making a basket from 25 feet distance in the general case when the state of the team is unknown

4.1.7 Evidence in favour of the BLHMM

Now, we compare the previous model, denoted as $\mathcal{M}_1$, with a simpler model, $\mathcal{M}_2$, without hidden structure and joint distribution factorised as:

$$f(\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\psi}) = \left(\prod_{i=1}^{N} \prod_{n=1}^{M_i} f(Y_{in} \mid \boldsymbol{\theta}, \boldsymbol{\psi}) \right) f(\boldsymbol{\psi} | \boldsymbol{\theta}) \, \pi(\boldsymbol{\theta}).$$

Here, each variable Y_{in} follows a Bernoulli distribution with a probability parameter γ_{in}:

$$(Y_{in} | \boldsymbol{\theta}, \boldsymbol{\psi}) \sim \text{Bern}(\gamma_{in}).$$

Each success probability parameter γ_{in} is associated with the covariates and the random effects through the logit function, as follows:

$$\text{logit}(\gamma_{in}) = \alpha + \alpha_d X_{in} + \alpha_{FT} I_{FT}(in) + a_i.$$

In this expression, all elements of the model have similar meaning as we indicated for model $\mathcal{M}_1$. In addition, we selected the same prior distribution for all shared parameters and hyperparameters across $\mathcal{M}_1$ and $\mathcal{M}_2$.

To conduct a model comparison between $\mathcal{M}_1$ and $\mathcal{M}_2$, we adopt two approaches. Firstly, we compute the conditional predictive ordinate (*CPO*) for each observation from the cross-validated predictive density, as outlined in Gelfand and Dey (1994). Secondly, we compute the Watanabe-Akaike information criterion, developed by Watanabe (2010).

4.1.8 Cross-validated conditional predictive ordinate

The fundamental idea underlying this approach is based on the assumption that if the estimated model is correct, each observation can be considered as a random variable drawn from the cross-validated predictive density. The *CPO*s for models $\mathcal{M}_1$ and $\mathcal{M}_2$ are defined as

$$CPO_{in}^{\mathcal{M}_1} = f(Y_{in} \mid \mathcal{D}^{-(in)}, \mathcal{M}_1), \tag{10}$$

$$CPO_{in}^{\mathcal{M}_2} = f(Y_{in} \mid \mathcal{D}^{-(in)}, \mathcal{M}_2). \tag{11}$$

Here, $\mathcal{D}^{-(in)}$ denotes all of the data in $\mathcal{D}$ except for the n-th shot in game i. Higher *CPO* values from a particular model compared to other provide support for that model. We compute these values following the approach in Ntzoufras (2011). This procedure applies the self-normalized importance sampling technique to approximate *CPO* values obtained from samples of the posterior distribution computed through the complete data $\mathcal{D}$.

Specifically, we consider the difference $CPO_{d_{in}} = CPO_{in}^{\mathcal{M}_1} - CPO_{in}^{\mathcal{M}_2}$ as a normally distributed random variable, i. e.,

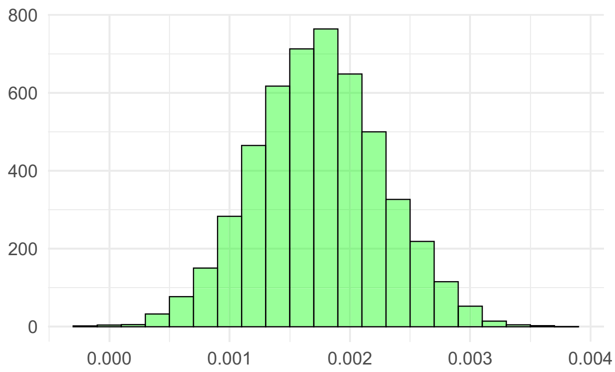

Fig. 8 Posterior distribution histogram of the mean parameter μ_d

$$(CPO_{d_{in}} \mid \mu_d, \sigma_d) \sim \mathrm{N}(\mu_d, \sigma_d^2). \tag{12}$$

In this context, bolstered by the central limit theorem, we chose a broad normal distribution for the marginal prior of the mean parameter, i.e., $\pi(\mu_d) = \mathrm{N}(0, 10^2)$. In addition, we opted for a uniform distribution for the standard deviation, $\pi(\sigma_d) = \mathrm{U}(0, 10)$.

Figure 8 shows the approximate posterior distribution of μ_d. It is evident that the marginal posterior is almost entirely positive, providing substantial evidence in favour of $\mathcal{M}_1$ over $\mathcal{M}_2$. In other words, the BLHMM provides a better fit for this data than a model lacking a hidden structure.

4.1.9 Watanabe–Akaike information criterion

According to Gelman et al. (2014), the Watanabe-Akaike information criterion (*WAIC*) provides a fast and computationally convenient alternative to the cross-validation approach. Its goal is to estimate the predictive accuracy of a model while penalising its complexity. The respective calculations for the two models, as implemented in the R package `rjags` (Plummer 2023), are presented in the following expressions:

Table 3 Estimation of *WAIC* for models $\mathcal{M}_1$ and $\mathcal{M}_2$ analysing the shooting performance of the Miami Heat team

	$\mathcal{M}_1$	$\mathcal{M}_2$
WAIC	14352.93	14371.59

$$\begin{aligned} WAIC^{\mathcal{M}_1} = -2\Bigg[\sum_{i=1}^{N}\sum_{n=1}^{M_i} \log\left(\frac{1}{S}\sum f(Y_{in} \mid \boldsymbol{Z}^{(s)}, \boldsymbol{\theta}^{(s)}, \boldsymbol{\psi}^{(s)}, \mathcal{M}_1)\right) \\ - \sum_{i=1}^{N}\sum_{n=1}^{M_i} V_{s=1}^{S}\left(\log f(Y_{in} \mid \boldsymbol{Z}^{(s)}, \boldsymbol{\theta}^{(s)}, \boldsymbol{\psi}^{(s)}, \mathcal{M}_1)\right)\Bigg], \end{aligned} \tag{13}$$

$$\begin{aligned} WAIC^{\mathcal{M}_2} = -2\Bigg[\sum_{i=1}^{N}\sum_{n=1}^{M_i} \log\left(\frac{1}{S}\sum f(Y_{in} \mid \boldsymbol{\theta}^{(s)}, \boldsymbol{\psi}^{(s)}, \mathcal{M}_2)\right) \\ - \sum_{i=1}^{N}\sum_{n=1}^{M_i} V_{s=1}^{S}\left(\log f(Y_{in} \mid \boldsymbol{\theta}^{(s)}, \boldsymbol{\psi}^{(s)}, \mathcal{M}_2)\right)\Bigg]. \end{aligned} \tag{14}$$

Here, S denotes the number of draws simulated from the posterior distribution, and $\boldsymbol{Z}^{(s)}$, $\boldsymbol{\theta}^{(s)}$, and $\boldsymbol{\psi}^{(s)}$ represent the posterior draws for the hidden states, the parameters, and the random effects, respectively. Furthermore, $V_{s=1}^{S}(\cdot)$ denotes the sample variance derived from the posterior outputs.

Table 3 presents the computed *WAIC* for the two models. According to our results, although the estimate for model $\mathcal{M}_1$ is slightly lower compared with that for $\mathcal{M}_2$, there is no enough evidence favouring the more complex model $\mathcal{M}_1$, as the values calculated are very close. This contrasts with the results from the *CPO* approach. Further analysis will show that, in other cases, both approaches align more closely.

4.2 Toronto Raptors shooting in the 2005–2006 NBA season

We review the 2005–2006 season of the Toronto Raptors, in which they struggled and failed to qualify for the *Playoff* phase. We made use of the same information that we had for the Miami Heat, which included, for each shot: the type of shot, the distance to the basket in feet, the specific game, and the sequential order of the shot

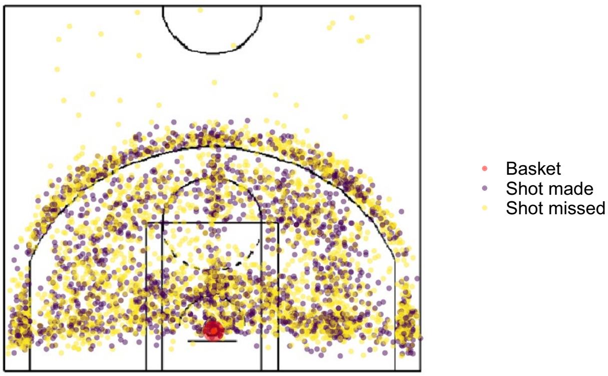

Fig. 9 Shot chart of the Toronto Raptors during the 2005–2006 season, showing the location of shots on the court, and indicating the shot results and the exact position of the basket

Table 4 Posterior summaries (mean, standard deviation, 95% credible interval) and diagnostic statistic $\hat{R}$ for the parameters and hyperparameters of the Toronto Raptors shooting performance *hot hand* model

Subprocess		Mean	Sd	$q_{0.025}$	$q_{0.975}$	$\hat{R}$
Hidden	β_{CH}	−0.38	0.05	−0.47	−0.28	1.000
	β_{HC}	0.37	0.06	0.25	0.49	0.999
	δ_C	0.60	0.06	0.47	0.72	1.002
	σ_{CH}	0.07	0.05	0.00	0.20	1.011
	σ_{HC}	0.15	0.08	0.01	0.32	1.007
Observed	α_C	−0.86	0.12	−1.10	−0.64	1.001
	α_H	11.57	1.78	8.72	15.75	1.010
	α_d	−0.34	0.06	−0.47	−0.24	1.011
	α_{FT}	6.56	0.91	5.06	8.58	1.011
	σ_a	0.25	0.10	0.04	0.44	1.003

for each game. The Toronto Raptors team attempted 8728 shots in 82 games, and 2089 of them were *free throws*. Figure 9 illustrates the precise locations and outcomes of each shot for the team.

4.2.1 Posterior results

Table 4 presents the approximate posterior summary for the parameters and hyperparameters of the Toronto Raptors' model. Two remarkable aspects can be observed from this table. Firstly, according to the posterior mean of β_{CH} and β_{HC}, as it was the case with Miami Heat, the probability of transitioning from the *cold* state to the *hot* state is lower than that of transitioning from the *hot* state to the *cold* state. This suggests that remaining in the *cold* state is more likely than remaining in the *hot* state. Secondly, observing the intercept parameters of the observed process, α_C and α_H, and comparing them with those from the Miami Heat analysis (Table 1), we see that both approximate posterior distributions from the Toronto Raptors model are lower

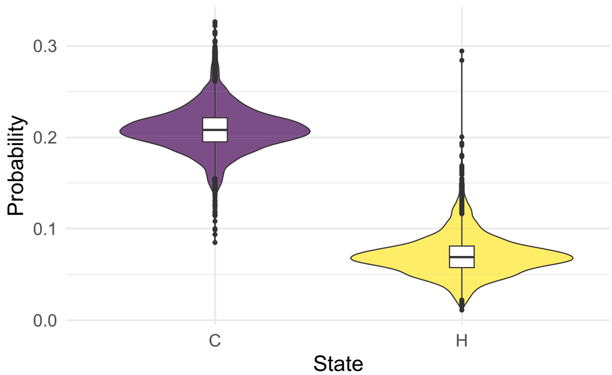

Fig. 10 Violin and box plot of the posterior probability distribution for a *cold* and a *hot* streak, for the Toronto Raptors' model

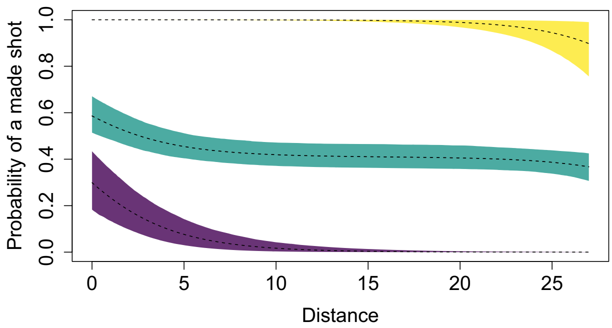

Fig. 11 Posterior means and 95% credible interval for the probabilities of making a basket depending on the distance, in feet, for the Toronto Raptors' case, in three different scenarios. Purple curve represents the case when the team is in the *cold* state, the yellow plot is for the *hot* state, and the green plot is the general case when the state of the team is unknown

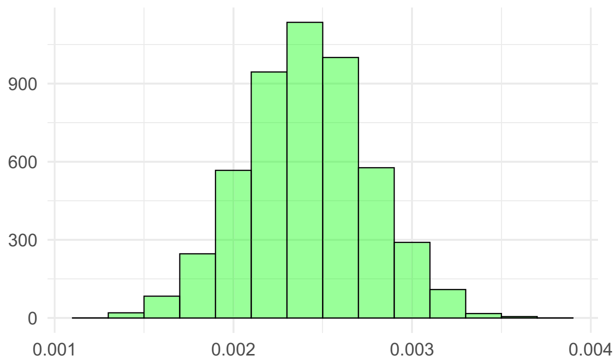

Fig. 12 Posterior distribution histogram of the mean parameter μ_d, for the Toronto Raptors' case

Table 5 Estimation of *WAIC* for models $\mathcal{M}_1$ and $\mathcal{M}_2$ analysing the shooting performance of the Toronto Raptors team

	$\mathcal{M}_1$	$\mathcal{M}_2$
WAIC	4620.95	11159.66

than those from the Miami Heat model. This implies that, in general, the two states of the Miami Heat team exhibit more efficiency.

For the sake of brevity, we have decided to show only two results of the team's output only but they provide a clear idea of the team performance: one associated with the latent process and the other with the observed process. Regarding the hidden part of the model, Fig. 10 illustrates the approximate probability of remaining three shots or more in the same state, namely *cold* and *hot*, respectively. We can observe that these probabilities do not differ substantially from those obtained by the Miami Heat team (Fig. 5).

However, when examining the approximate probability of success regarding the shot distance in Fig. 11, we realise that, with the exception of the scenario in which the team is in the *hot* state, the probabilities of making a shot are lower compared to the Miami Heat case (Fig. 6).

4.2.2 Model comparison

Following the same approaches as outlined in Sect. 4.1.3, we compare the BLHMM with a simpler model that does not incorporate a latent structure by utilising cross-validated *CPO* and computing their respective *WAIC*s. Figure 12 provides a visual representation of the approximate posterior distribution of the mean parameter μ_d of the *CPO* differences, as defined in expression (12). This figure suggests that the BLHMM is favourable. Furthermore, observing the *WAIC* estimations in both models presented in Table 5, there is clear evidence in favour of $\mathcal{M}_1$. In conclusion, both approaches support the more complex model over the simpler one.

4.3 Chicago Bulls shooting in the 2022–2023 NBA season

We now consider the 2022–2023 season of the Chicago Bulls, which also did not result in their qualification for the NBA *Playoffs* following a defeat in the *Play-In* against the Miami Heat. The Chicago Bulls attempted a total of 9 109 shots in 84 games, and 1 825 of them were *free throws*.

Examining the short chart of the Chicago Bulls team (Fig. 13), we observe that the spatial distribution of shot locations considerably differs from those in the previous two cases. Notably, both the Toronto Raptors and the Miami Heat during the 2005–2006 season (Figs. 9 and 2) tended to avoid shooting directly from the painted area in front of the basket. In contrast, the shot chart for the Chicago Bulls does

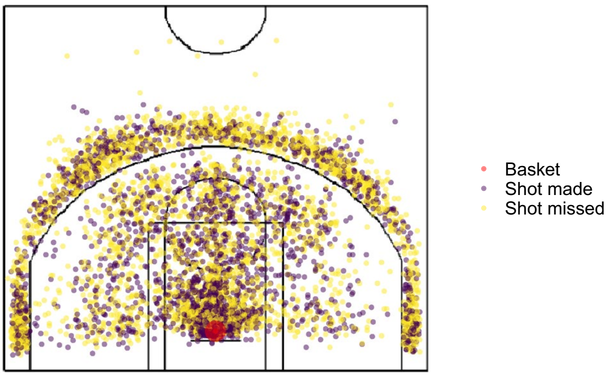

Fig. 13 Shot chart of the Chicago Bulls during the 2022–2023 season, showing the location of shots on the court, and indicating the shot results and the exact position of the basket

Table 6 Posterior summaries (mean, standard deviation, 95% credible interval) and diagnostic statistic $\hat{R}$ for the parameters and hyperparameters of the Chicago Bulls shooting performance *hot hand* model

Subprocess		Mean	Sd	$q_{0.025}$	$q_{0.975}$	$\hat{R}$
Hidden	β_{CH}	−0.35	0.05	−0.45	−0.24	0.999
	β_{HC}	0.23	0.06	0.11	0.36	1.001
	δ_C	0.55	0.06	0.43	0.67	1.002
	σ_{CH}	0.07	0.06	0.00	0.20	1.014
	σ_{HC}	0.13	0.09	0.00	0.31	1.021
Observed	α_C	0.22	0.11	0.00	0.44	1.007
	α_H	11.34	1.22	9.10	13.91	1.008
	α_d	−0.37	0.04	−0.46	−0.29	1.008
	α_{FT}	5.96	0.59	4.85	7.23	1.006
	σ_a	0.16	0.09	0.01	0.34	1.013

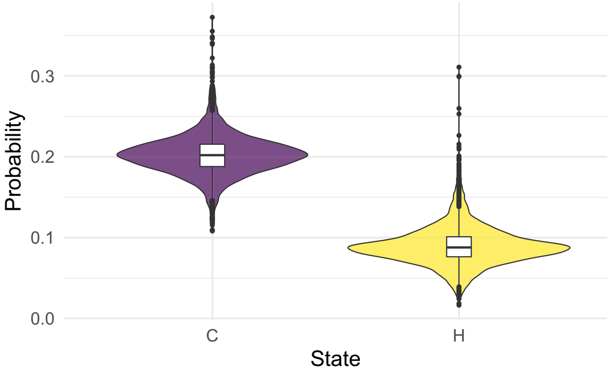

Fig. 14 Violin and box plot of the posterior probability distribution for a *cold* and a *hot* streak, for the Chicago Bulls' model

not show such a gap. Moreover, the frequency of three-point attempts by this team seems considerably higher.

4.3.1 Posterior results

The posterior results obtained from the recent analysis of the Chicago Bulls performance in the 2022–2023 season, as summarised in Table 6, reveal some similarities and differences with the results of the two previous teams. On the one hand, according to the approximate posterior distribution of β_{CH} and β_{HC}, we observe that

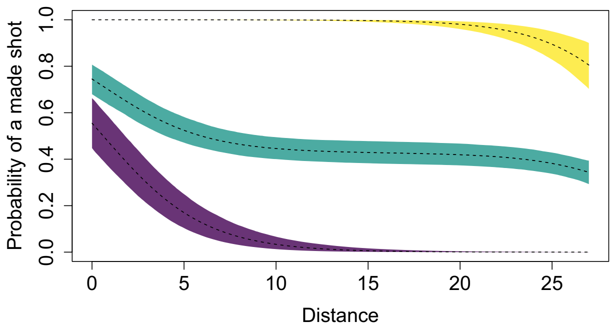

Fig. 15 Posterior means and 95% credible interval for the probabilities of making a basket depending on the distance, in feet, for the Chicago Bulls' case, in three different scenarios. Purple curve represents the case when the team is in the *cold* state, the yellow plot is for the *hot* state, and the green plot is the general case when the state of the team is unknown

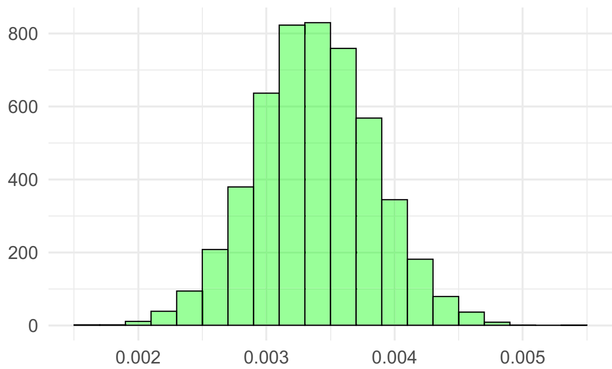

Fig. 16 Posterior distribution histogram of the mean parameter μ_d, for the Chicago Bulls' case

Table 7 Estimation of *WAIC* for models $\mathcal{M}_1$ and $\mathcal{M}_2$ analysing the shooting performance of the Chicago Bulls team

	$\mathcal{M}_1$	$\mathcal{M}_2$
WAIC	5634.04	11447.94

the probability of transitioning from the *cold* to the *hot* state is lower than the probability of moving in the opposite direction. This fact has been consistently observed in all three analyses conducted using our approach. On the other hand, the posterior distribution of α_C is predominantly positive. In other words, the probability of shooting success when the Chicago Bulls were in the cold state is higher than that observed for the Miami Heat or the Toronto Raptors.

According to Fig. 14, the Chicago Bulls' posterior probability of a *cold* and a *hot* streak is remarkably similar to those of the Toronto Raptors team, at approximately

0.2 for the *cold* streak and close to 0.1 for the *hot* one. Notably, the approximate posterior probability of successfully making a shot with regard to the distance to the basquet, as shown in Fig. 14, indicates a high success rate in the *cold* state for easier shots. Thus, the Chicago Bulls exhibit the best performance in the *cold* state compared to the other teams analysed previously (Fig. 15).

4.3.2 Model comparison

In the case of the Chicago Bulls, both model comparison procedures (*CPO* and *WAIC*) point to the same direction. The comparison between the BLHMM and the model without a hidden structure through the cross-validation approach shows a posterior mean for the difference between both CPOs of about 0.0035 (see Fig. 16), the highest value observed so far. Additionally, observing the estimates of the *WAIC* in Table 7, we can see that there is strong evidence in favour of the BLHMM model $\mathcal{M}_1$. Thus, in the case of the Chicago Bulls team, it is clear that the model with a latent structure offers advantages over the simpler model.

5 Conclusions

We observed a common situation in the performance of the three basketball teams we studied: the *cold hand* was more likely than the *hot hand*. In fact, in the same match, *cold streaks* seem to occur more frequently than *hot streaks*. Moreover, according to the posterior distribution of the transition probabilities, the two states are clearly differentiated; the team's performance is very different depending on which state it is in.

We believe our BLHMM can be employed as a valuable tool for analysing the performance of a team during a match. It could also be used to compare the performances of different teams, providing information that is not possible to obtain through the traditional modelling techniques. For instance, our model allows analysing differences between transition probabilities and determining the frequency of occupancy of each state. It can also assess the effectiveness of the team by quantifying the probability of making a basket through relevant covariates as well as provide an important set of dynamic team performance outcomes. In addition, the probability associated with both states, cold or hot, at the beginning of each game could be investigated, i.e. how δ_C varies for each team. Additionally, it has the potential to be applied in the analysis of an opponent's team performance during a game. It is worth noting that, with the exception of the first example of the Miami Heat team in the 2005–2006 NBA season when comparing the two models using WAIC, we provide evidence that supports our approach over a simpler model. Furthermore, we would like to highlight that the histograms of the mean difference of the CPO values, as well as the values obtained from WAIC, indicate the same result. In fact, in the first example, we observe the smallest CPOs compared to the other two examples, while the WAIC values are also observed to be the closest to each other.

Finally, for future research, as the choice of the number of states in the hidden Markov model is a crucial and complex issue that extends beyond the scope of our current paper, the inclusion of a third intermediate state that is neither *hot* nor *cold* is an interesting possibility. Furthermore, studying the possibility of a higher number of latent states could be worth exploring. It could also be highly valuable to incorporate the match time as a covariate in the transition matrix. By doing so, we would be able to study how the team behaves at each specific moment of the match. In this sense, exploring continuous-time Markov chains could be a worthwhile approach to better model this phenomenon.

Acknowledgements Gabriel Calvo's research was partially funded by the ONCE Foundation, the Universia Foundation, and the Spanish Ministry of Education and Professional Training, grant FPU18/03101. This paper is part of the project PID2022-136455NB-I00, funded by Ministerio de Ciencia, Innovación y Universidades of Spain (MCIN/AEI/10.13039/501100011033/FEDER, UE) and the European Regional Development Fund. Luigi Spezia's research was funded by the Scottish Government's Rural and Environment Science and Analytical Services Division. Comments from Fergus Chadwick and two anonymous referees improved the quality of the final paper.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

References

Adler P (1981) Momentum: a theory of social action (sociological observations). Sage Publications

Albert J, Bennett J (2001) Streakiness (or, the hot hand). Curve ball. Springer, Berlin, pp 111–144

Avugos S, Köppen J, Czienskowski U, Raab M, Bar-Eli M (2013) The â€œhot handâ€ reconsidered: A meta-analytic approach. Psychol Sport Exerc 14(1):21–27

Baumer BS, Matthews GJ, Nguyen Q (2023) Big ideas in sports analytics and statistical tools for their investigation. Computational Statistics, Wiley Interdisciplinary Reviews, p e1612

Bosch-Sijtsema PM, Ruohomäki V, Vartiainen M (2009) Knowledge work productivity in distributed teams. J Knowl Manage 13(6):533–546

Chen Z, Ong D, Sheremeta RM (2015) Competition between and within universities: Theoretical and experimental investigation of group identity and the desire to win. J Econ Psychol 93:102551

Conesa D, Martínez-Beneito M, Amorós R, López-Quílez A (2015) Bayesian hierarchical Poisson models with a hidden Markov structure for the detection of influenza epidemic outbreaks. Stat Methods Med Res 24(2):206–223

Frühwirth-Schnatter S (2001) Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. J Am Stat Assoc 96(453):194–209

Gelfand AE, Dey DK (1994) Bayesian model choice: asymptotics and exact calculations. J R Stat Soc Ser B (Methodol) 56(3):501–514

Gelman A (2006) Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2013) Bayesian data analysis. CRC Press, Boca Raton
Gelman A, Hwang J, Vehtari A (2014) Understanding predictive information criteria for Bayesian models. Stat Comput 24:997–1016
Gilovich T, Vallone R, Tversky A (1985) The hot hand in basketball: on the misperception of random sequences. Cogn Psychol 17(3):295–314
Green B, Zwiebel J (2018) The hot-hand fallacy: cognitive mistakes or equilibrium adjustments? Evidence from major league baseball. Manage Sci 64(11):5315–5348
Heere B, James J, Yoshida M, Scremin G (2011) The effect of associated group identities on team identity. J Sport Manage 25(6):606–621
Kang K, Cai J, Song X, Zhu H (2019) Bayesian hidden Markov models for delineating the pathology of Alzheimerâ€™s disease. Stat Methods Med Res 28(7):2112–2124
Kulkarni VG (2016) Modeling and analysis of stochastic systems. Chapman and Hall/CRC, Boca Raton
Lewis M (2004) Moneyball: the art of winning an unfair game. WW Norton & Company, New York
Ley C, Dominicy Y (2020) Science meets sports: when statistics are more than numbers. Cambridge Scholars Publishing, Newcastle upon Tyne
MacDonald IL, Zucchini W (1997) Hidden Markov and other models for discrete-valued time series, vol 110. CRC Press, Boca Raton
Manley M (1989) Basketball heaven. Doubleday Books, New York
Martínez-Beneito MA, Botella-Rocamora P (2019) Disease mapping: from foundations to multidimensional modeling. CRC Press, Boca Raton
Maruotti A (2011) Mixed hidden Markov models for longitudinal data: an overview. Int Stat Rev 79(3):427–454
McCulloch RE, Tsay RS (1994) Statistical analysis of economic time series via Markov switching models. J Time Ser Anal 15(5):523–539
Mews S, Ötting M (2023) Continuous-time state-space modelling of the hot hand in basketball. AStA Adv Stat Anal 107(1–2):313–326
Miller JB, Sanjurjo A (2018) Surprised by the hot hand fallacy? A truth in the law of small numbers. Econometrica 86(6):2019–2047
NBA (2020) NBAstuffer. https://www.nbastuffer.com/analytics101/playbyplay-data/. Accessed: 2022-05-03
Ntzoufras I (2011) Bayesian modeling using WinBUGS. John Wiley & Sons, New Jersey
Oliver D (2004) Basketball on paper: rules and tools for performance analysis. Potomac Books, Inc
Ötting M, Andreas G (2022) A regularized hidden Markov model for analyzing the 'hot shoe' in football. Stat Modell 22(6):546–565
Ötting M, Langrock R, Deutscher C, Leos-Barajas V (2020) The hot hand in professional darts. J R Stat Soc Ser A Stat Soc 183(2):565–580
Patterson MM, Carron AV, Loughead TM (2005) The influence of team norms on the cohesion-self-reported performance relationship: a multi-level analysis. Psychol Sport Exerc 6(4):479–493
Plummer M (2003) Jags: a program for analysis of Bayesian graphical models using Gibbs sampling. In: Proceedings of the 3rd International Workshop on Distributed Statistical Computing, volume 124. Vienna, Austria
Plummer M (2023) rjags: Bayesian Graphical Models using MCMC. R package version 4-15
Raab M, Gula B, Gigerenzer G (2012) The hot hand exists in volleyball and is used for allocation decisions. J Exp Psychol Appl 18(1):81
Rabiner L, Juang B (1986) An introduction to hidden Markov models. IEEE ASSP Mag 3(1):4–16
Sandri M, Zuccolotto P, Manisera M (2020) Markov switching modelling of shooting performance variability and teammate interactions in basketball. J R Stat Soc Ser C 69(5):1337–1356
Song X, Xia Y, Zhu H (2016) Hidden Markov latent variable models with multivariate longitudinal data. Biometrics 73(1):313–323
Spezia L (2009) Reversible jump and the label switching problem in hidden Markov models. J Stat Plan Inference 139(7):2305–2315
Sun Y (2004) Detecting the hot hand: an alternative model. In: Proceedings of the annual meeting of the cognitive science society, vol 26
Tanner MA (2012) Tools for statistical inference. Springer, Berlin
Terner Z, Franks A (2021) Modeling player and team performance in basketball. Annu Rev Stat Appl 8:1–23

Thomas WE, Brown R, Easterbrook MJ, Vignoles VL, Manzi C, D'Angelo C, Holt JJ (2019) Team-level identification predicts perceived and actual team performance: longitudinal multilevel analyses with sports teams. Br J Soc Psychol 58(2):473–492

Watanabe S (2010) Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J Mach Learn Res 11:3571–3594

Wetzels R, Tutschkow D, Dolan C, van der Sluis S, Dutilh G, Wagenmakers E-J (2016) A Bayesian test for the hot hand phenomenon. J Math Psychol 72:200–209

Woolley AW (2009) Means vs ends: Implications of process and outcome focus for team adaptation and performance. Organ Sci 20(3):500–515

Zhou J, Song X, Sun L (2020) Continuous time hidden markov model for longitudinal data. J Multivar Anal 179:104646

Zuccolotto P, Manisera M (2020) Basketball data science: with applications in R. CRC Press, Boca Raton