Joshua C. Chang e-mail: joshchang@ucla.edu [1]

Selection of multistep Markov chain models for free throw shooting in evaluation of the hot hands effect

Abstract.
The hot hands effect in basketball is a simple and seemingly intuitive phenomenon that has proven challenging to detect. For this reason, academics have considered it a fallacy for decades. A key difficulty in the detection of hot handedness is confounding due to possible changes in shot selection. In the more-limited context of free throw shooting, there are fewer confounding factors. Hence, simply modeling the probability of making a free throw as conditional on the outcome of recent free throws (via multistep Markov chains) is reasonable. The detection of hot-handedness then becomes a model selection problem where the job is to evaluate whether the inclusion of this effect makes for a more-predictive model. In this paper, I address the general problem of model selection for multistep Markov chains via comparison of several model selection criteria, each derived as closed-form mathematical formulae. Using simulations, I evaluate the accuracy of these criteria. These methods, when applied to data from the 2016–2017 NBA season, demonstrate evidence of statistical dependencies in LeBron James' free throw shooting. In particular, a model depending on the previous shot (single-step Markovian) is approximately as predictive as a model with independent outcomes. A hybrid jagged model of two parameters, where James shoots a higher percentage after a missed free throw than otherwise, is more predictive than either model.

1. Introduction

With 10.6 seconds remaining on the shot clock in game 7 of the 2016 NBA finals, the Cleveland Cavaliers were up by three points. LeBron James, shaken by a hard foul, went to the free throw line to shoot two free throws. A single make would seal the outcome of the game. James misses the first free throw. That season he was a 73% free throw shooter, slightly off from his career average of 74%. In that moment, one might ask whether the probability of making the second free throw was 73%, or something different altogether. Ignoring the issues of fatigue and the pressure of the situation, perhaps the fact that he had just missed a free throw was indicative of a hidden mental or physical state that predicted another miss. Or perhaps the miss precipitated a correction process that would lead to a more-likely make. Or perhaps it did not matter at all.

LeBron James would go on to make the second free throw and the Cavaliers would go on to win the championship. The outcome of the game crystalized into history, yet our question remained unanswered. Can the outcome of shots recently taken predict the outcome of future shots? This is the central question behind the hot hands phenomenon.

While controversial in analytical circles, belief in the hot hands phenomenon is certainly widespread in both the general public and in athletes [13, 10, 22]. Empirically, the phenomenon has proven to be elusive [5]. In 1980s, examinations of the phenomenon in basketball based on analysis of shooting streaks yielded negative results [13, 16], failing to reject null hypotheses of statistically independent shot outcomes. Based on these early analyses, some studies have explained the widespread belief in hot-handedness by relating it to the Gambler's fallacy [4]. Follow-up studies have examined the effects of belief in this phenomenon under the supposition that it is a fallacy [9]. However, recent analyses, using multivariate methods that can account for factors such as shot difficulty [7, 2, 19], have supported the phenomenon, finding the original studies to be underpowered, [2, 3, 19] or to suffer from methodological issues regarding the weighting of expectation values [19].

This manuscript re-examines the phenomenon from a different analytical perspective than in prior studies. Presently, a broad class

2. Quantitative Methods

2.1. Probabilistic modeling

The outcomes of free throws for a player in a particular game can be represented as string or trajectory of states (miss or make). For example, using "$+$" to denote makes and "$-$" to denote misses, a trajectory of "$++-+$" corresponds to a game where a player makes the first two free throws, misses the third, and makes the fourth. In evaluation of the hot hands phenomenon, we seek to model the probability of observing such trajectories for any given player. In doing so, we may conveniently assess for statistical dependencies in shot outcome on the outcome of prior shots. A simple probabilistic model for paths, incorporating possible statistical dependencies on history, is the multistep Markov chain.

Multistep Markov chains are factorized probability models for discrete-state trajectories, where the probability of a particular trajectory is the product of conditional transition probabilities between possible states. The conditions pertain to the prior locations that a trajectory has visited, or its recent history. For our problem there are two states, however, let us consider the more general problem of a model with any number M states. Assume that a trajectory ξ consists of steps ξ_l, where each step takes a value x_l taken from the set $\{1,2,\ldots,M\}$. We are interested in representations for the trajectory probability of the form

$$\begin{aligned}
\Pr(\xi) &= \prod \Pr\left(\xi_l = x_l | \text{previous } h \text{ states}\right) \\
&= \prod_{l=1}^{L} \Pr(\xi_l = x_l | \xi_{n-1} = x_{l-1}, \ldots, \xi_{l-h} = x_{l-h}) \\
&= \prod_{l}^{L} p_{x_{l-h}, x_{l-h+1}, \cdots, x_l}, \qquad (1)
\end{aligned}$$

where h, a non-negative counting number, represents the number of states worth of memory needed to predict the next state, with appropriate boundary conditions for the beginning of the trajectory. In the context of the hot hands effect, models with $h \geq 1$ encompass statistical dependencies between shot outcomes. Hot hands would correspond to higher

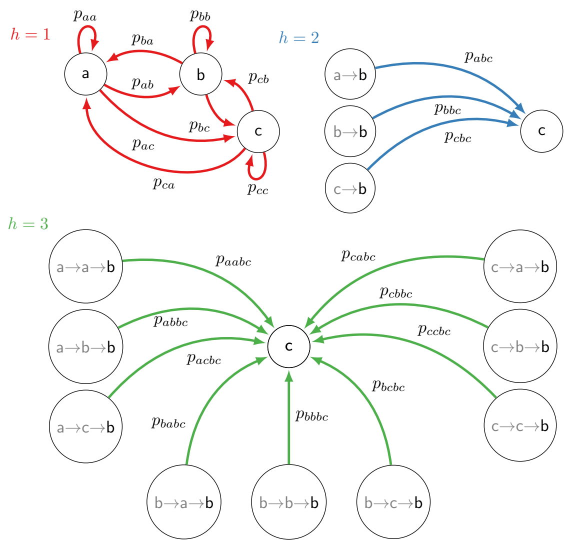

Figure 1. Multi-step finite state Markovian processes parameterized by degree of memory h, demonstrated on a three-state network. For $h = 1$, the statistics of the next state depend solely on the current state and the stochastic system is parameterized by transition probabilities indexed by a tuple. For $h = 2$ and $h = 3$ the statistics depend on the history. Shown are the possible single-step transitions from state **b** to state **c**. For $h = 2$, transition probabilities depend on the current state and the previous state, and all transition probabilities are indexed by 3-tuples. For $h = 3$, all transition probabilities depend on the current state and two previous states and are indexed by 4-tuples.

make probabilities after recent makes and cold hands correspond to lower probabilities after misses.

In the case of absolutely no memory ($h = 0$), the path probability is simply the product of the probabilities of being in each of the separate states in a path, $p_{x_1} p_{x_2}, \ldots p_{x_L}$, and there are essentially $M - 1$ free model parameters, where M is the number of states. A model where $h = 0$ implies that the outcomes of previous shot attempts do not inform the outcome of the next attempt. If $h = 1$, the model is single-step Markovian in that only the current state is relevant in determining the next state. These models involve $M(M-1)$ free parameters. Generally, if h states of history are required, then the model is h-step Markovian, and $M^h(M-1)$ parameters are needed (see Fig 1). Hence, the size of the parameter space grows exponentially with memory. Our objective is to determine, based on observational evidence, an appropriate value for h.

For a fixed degree of memory h, we may look at possible history vectors $\mathbf{x} = [x_1, x_2, \ldots, x_h]$ of length h taken from the set $\mathbf{X}_h = \{1, 2, \ldots, M\}^h$. For each $\mathbf{x}$, denote the vector $\mathbf{p_x} = [p_{\mathbf{x},1}, p_{\mathbf{x},2}, \ldots p_{\mathbf{x},M}]$, where $p_{\mathbf{x},m}$ is the probability that a trajectory goes next to state m given that $\mathbf{x}$ represents its most recent history. For convenience, we denote the collection of all $\mathbf{p_x}$ as $\mathbf{p}$ (see below for an example of the notation).

Generally one has available J trajectories. Assuming independence between trajectories, one may write the joint probability, or likelihood, of observing these trajectories as

$$
\begin{aligned}
\Pr(\{\xi^{(j)}\}_{j=1}^{J}|\mathbf{p}) &= \prod_{j=1}^{J} \Pr(\xi^{(j)}|\mathbf{p}) \\
&= \prod_{j=1}^{J} \prod_{\mathbf{x}\in\mathbf{X}_h} \prod_{m=1}^{M} p_{\mathbf{x},m}^{N_{\mathbf{x},m}^{(j)}} = \prod_{\mathbf{x}\in\mathbf{X}_h} \prod_{m=1}^{M} p_{\mathbf{x},m}^{N_{\mathbf{x},m}},
\end{aligned} \tag{2}
$$

where $N_{\mathbf{x},m}^{(j)}$ is the number of times that the transition $\mathbf{x} \to m$ occurs in trajectory $\xi^{(j)}$, and $N_{\mathbf{x},m} = \sum_j N_{\mathbf{x},m}^{(j)}$ is the total number of times the transition is seen.

For convenience, denote $N_{\mathbf{x}} = \sum_m N_{\mathbf{x},m}$, $\mathbf{N}_{\mathbf{x}} = [\mathbf{N}_{\mathbf{x},1}, \mathbf{N}_{\mathbf{x},2}, \ldots, \mathbf{N}_{\mathbf{x},M}]$, and the collection of all $\mathbf{N}_{\mathbf{x}}^{(j)}$ as $\mathbf{N}$. The sufficient statistics of the likelihood are the counts, so we will refer to the likelihood as $\Pr(\mathbf{N} \mid \mathbf{p})$. The maximum likelihood estimator for each parameter vector $\mathbf{p}_{\mathbf{x}}$ is found by maximizing the probability in Eq. 2, and can be written easily as $\hat{p}_{\mathbf{x}}^{\text{MLE}} = \mathbf{N}_{\mathbf{x}}/N_{\mathbf{x}}$.

Example: To make the notation more clear, consider a model for free throw shooting informed using $J = 2$ observed trajectories given: $\{\xi^{(1)} = +-++-++, \xi^{(2)} = +--+-+++++-\}$. Suppose that we set $h = 1$ in this model. This choice implies that we need the counts $N_{++}^{(1)} = 2$, $N_{+-}^{(1)} = 2$, $N_{--}^{(1)} = 0$, $N_{-+}^{(1)} = 2$, $N_{++}^{(2)} = 4$, $N_{+-}^{(2)} = 3$, $N_{--}^{(2)} = 1$, $N_{-+}^{(2)} = 2$ coinciding to the number of makes following makes, misses following makes, misses following misses and makes following misses respectively for each trajectory. In addition, since the first state in each trajectory is stochastic as well, we add two special states representing the outcome of the initial free throw, $N_{\cdot-} = 0$, $N_{\cdot+} = 2$. Aggregating the counts across the trajectories, in the vector notation above, we have $\mathbf{N}_{+} = [N_{--}, N_{++}] = [5,6], \mathbf{N}_{-} = [1,4], \mathbf{N}_{\cdot} = [0,2]$, where the indices are all of length one since our model only considers history vectors of length one.

2.2. Bayesian modeling

A natural Bayesian formulation of the problem of determining the transition probabilities is to use the Dirichlet conjugate prior on each parameter vector $\mathbf{p}_{\mathbf{x}} \sim \text{Dirichlet}(\alpha)$, hyper-parameterized by α, a vector of size M. This manuscript assumes that $\alpha = \mathbf{1}$, corresponding to a uniform prior. This prior, paired with the likelihood of Eq. 2, yields the posterior distribution on the probabilities,

$$
\mathbf{p}_{\mathbf{x}}|\mathbf{N}_{\mathbf{x}} \sim \text{Dirichlet}(\alpha + \mathbf{N}_{\mathbf{x}}). \tag{3}
$$

According to this distribution, the posterior expectations for the transition probabilities follow

$$
\mathbb{E}_{\mathbf{p}_{\mathbf{x}}|\mathbf{N}_{\mathbf{x}}}[p_{\mathbf{x},m}] = \frac{\alpha + N_{\mathbf{x},m}}{M\alpha + N_{\mathbf{x}}}. \tag{4}
$$

yields a Dirichlet posterior distribution with associated expectation,

$$
\mathbf{p}_{\mathbf{x}}|\mathbf{N}_{\mathbf{x}} \sim \text{Dirichlet}(\alpha + \mathbf{N}_{\mathbf{x}}) \quad \mathbb{E}_{\mathbf{p}_{\mathbf{x}}|\mathbf{N}_{\mathbf{x}}}[p_{\mathbf{x},m}] = \frac{\alpha + N_{\mathbf{x},m}}{M\alpha + N_{\mathbf{x}}}. \tag{5}
$$

In effect, one is assigning a probability of $\alpha/(M\alpha + N_{\mathbf{x}})$ to any unobserved transition, where α can be made small if it is expected that the transition matrix should be sparse.

In the large-sample limit, the choice of α is not important as the posterior distribution of Eq. 5 becomes tightly concentrated about the maximum likelihood estimates.

2.3. Model selection

The parameter h controls the trade-off between complexity and fitting error. From a statistical viewpoint, complexity results in less-precise determination of model parameters, leading to larger prediction errors (overfitting). Conversely, a simple model may not capture the true probability space where paths reside, and fail to catch patterns in the real process (underfit).

This manuscript evaluates several statistical criteria for selecting the number of states h worth of memory to retain in the factorization of Eq. 1, viewing the problem in terms of prediction accuracy. A value h that yields a model that can best predict new unobserved trajectories [11] is chosen. Using these methods, I evaluate the hot hands phenomenon with data from LeBron James' 2016–2017 season.

Prior statistical analyses of the hot hands effect have been rooted in null hypothesis statistical testing. Rather than follow this approach, which requires a subjective choice of a cut off p-value, this manuscript frames this problem as a model selection task. We are interested in whether a model which encompasses effects like hot-handedness is better at predicting free throw outcomes than a model without. There are various existing generalized methods for evaluating how well models predict. To make these methods comparable,

Following this approach, the Akaike Information Criterion (AIC), [1, 25, 15], $\mathrm{AIC} = -2\log\Pr(\mathbf{N}|\hat{p}^{\mathrm{MLE}}) + 2k$, may be used as a metric in order to choose a value of h. Rooted in information theory, the AIC is an asymptotic approximation of the information loss in the representation of data by a model [8]. The model with the smallest AIC, and hence with the smallest approximate loss, is chosen.

A limitation of the AIC is inaccuracy for small datasets. A correction to the AIC known as the AICc exists [14], however, its exact form is problem specific [8]. Fundamentally, the maximum likelihood estimator precludes the existence of unobserved transitions – a property that is problematic if the sample size J is small. It is desirable to regularize the problem by allowing a nonzero probability that transitions that have not yet been observed will occur. This manuscript's approach to rectifying these issues is Bayesian.

As pertains to Bayesian model selection, Bayes factors are often used [17, 21]. If using non-informative model priors, they consist of the likelihood of the data, averaged over the posterior distribution of model parameters. The logarithm of this quantity is known as the log predictive density (LPD). Related to the LPD is the log point-wise predictive density (LPPD), where the same expectation is taken separately for each datapoint and logarithms of these expectations are summed. The LPPD features in alternatives to Bayes factors and the AIC [12].

The Widely Applicable Information Criterion [26, 27] (WAIC) is a Bayesian information criterion with two variants, each featuring the LPPD but differing in how they compute model complexity. The WAIC, unlike the AIC, is applicable to singular statistical models and is asymptotically equivalent to Bayesian leave-one-out cross-validation [26]. The commonly used Deviance Information Criterion (DIC) also resembles the WAIC, consisting of two variants. Both variants use point estimates of the posterior parameters rather than expectations as used in the WAIC. However, the BIC [20] does not measure prediction fit for new data [12].

Finally Bayesian variants of cross-validation have recently been proposed as alternatives to information criterion [12]. In our problem, k-fold CV, where data is divided into k partitions, can be evaluated in closed form without repeated model fitting. Using $-2 \times \text{LPPD}$ as a metric, this manuscript also evaluates two variants of k-fold CV: two-fold cross validation (LPPDCV_2) and leave-one-out cross validation (LOO). Closed-form formulae for computing each model selection criterion are available as SUPPLEMENTAL MATERIAL.

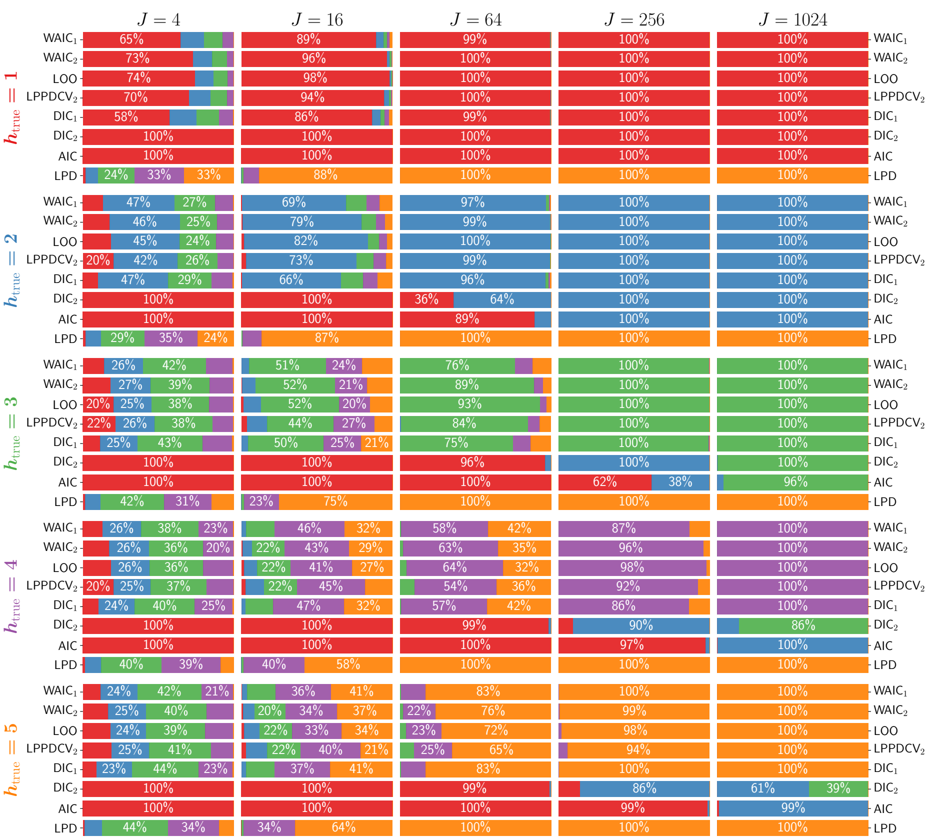

Figure 2. Chosen degree of memory h in simulations for varying true degrees of memory h_{true} and number of observed trajectories J. Rows correspond to model selection under a given degree of memory. Columns correspond to the number of trajectories. Depicted are the percent of simulations in which each degree of memory is selected using the different model evaluation criteria (percents of at least 20 are labeled). Colors coded based on degree of memory: (1: red, 2: blue, 3: green, 4: purple, 5: orange). *Example:* For $h_{\text{true}} = 1$ and $J = 4$, the WAIC_1 criteria selected $h = 1$ approximately 65% of the time.

The Akaike information criterion (AIC) is defined through the formula $\text{AIC} = -2\sum_{\mathbf{x}} \log \Pr(\mathbf{N_x}|\hat{\mathbf{p}}_{\text{MLE}}) + 2k$ and can be computed exactly as

$$\mathrm{AIC} = -2\sum_{\mathbf{x}}\sum_{m=1}^{M} N_{\mathbf{x},m}\log\left(\frac{N_{\mathbf{x},m}}{N_{\mathbf{x}}}\right) + 2M^{q+1}, \tag{6}$$

where in this context we define $0 \times \log(0) = 0$.

The deviance information criterion (DIC) is similar to the AIC and defined as $\mathrm{DIC} = -2\sum_{\mathbf{x}}\log p\left(\mathbf{N}_{\mathbf{x}} \mid \mathbf{p}_{\mathbf{x}} = \mathbb{E}_{\mathbf{p}_{\mathbf{x}}|\mathbf{N}_{\mathbf{x}}}\mathbf{p}_{\mathbf{x}}\right) + 2k_{\mathrm{DIC}}$. It may also be computed by evaluating the closed form expression

$$\mathrm{DIC} = -2\sum_{\mathbf{x}}\sum_{m=1}^{M} N_{\mathbf{x},m}\log\left(\frac{N_{\mathbf{x},m}+\alpha}{N_{\mathbf{x}}+M\alpha}\right) + 2k_{\mathrm{DIC}}, \tag{7}$$

where we are assuming that one uses the posterior mean as the point estimate of the model parameters, and also where the effective model complexity k_{DIC} has two variants

$$\begin{aligned}
k_{\mathrm{DIC1}} &= -2\left\{\sum_{\mathbf{x}}\sum_{m=1}^{M} N_{\mathbf{x},m}\log\left(\frac{N_{\mathbf{x},m}+\alpha}{N_{\mathbf{x}}+M\alpha}\right)\right.\\
&\left.-\sum_{j}\sum_{\mathbf{x}}\mathbb{E}_{\mathbf{p}_{\mathbf{x}}|\mathbf{N}}\log\mathbf{p}_{\mathbf{x}}^{\mathbf{N}_{\mathbf{x}}^{(j)}}\right\}\\
&= 2\left\{\sum_{\mathbf{x}}\sum_{m=1}^{M} N_{\mathbf{x},m}\log\left(\frac{N_{\mathbf{x},m}+\alpha}{N_{\mathbf{x}}+M\alpha}\right)\right.\\
&\left.-\sum_{j}\sum_{\mathbf{x}}\sum_{m=1}^{M}\mathbf{N}_{\mathbf{x},m}^{(j)}\left[\psi(\alpha+N_{\mathbf{x},m})-\psi(M\alpha+N_{\mathbf{x}})\right]\right\}\\
&= 2\left\{\sum_{\mathbf{x}}\sum_{m=1}^{M} N_{\mathbf{x},m}\log\left(\frac{N_{\mathbf{x},m}+\alpha}{N_{\mathbf{x}}+M\alpha}\right)\right.\\
&\left.-\sum_{\mathbf{x}}\sum_{m=1}^{M} N_{\mathbf{x},m}\left[\psi(\alpha+N_{\mathbf{x},m})-\psi(M\alpha+N_{\mathbf{x}})\right]\right\},
\end{aligned} \tag{8}$$

and $k_{\mathrm{DIC2}} = 2\mathrm{var}_{\mathbf{p}|\mathbf{N}}\left[\log\Pr\left(\mathbf{N} \mid \mathbf{p}\right)\right]$, which may be computed

$$
\begin{aligned}
k_{\mathrm{DIC2}} &= 2\mathrm{var}_{\mathbf{p_x}}\left[\sum_{\mathbf{x}}\sum_{m} N_{\mathbf{x},m}\log p_{\mathbf{x},m}\right] \\
&= 2\sum_{\mathbf{x}}\mathrm{var}_{\mathbf{p_x}}\left(\sum_{m} N_{\mathbf{x},m}\log p_{\mathbf{x},m}\right) \\
&= 2\sum_{\mathbf{x}}\sum_{m}\sum_{n} N_{x,m}N_{x,n}\mathrm{cov}(\log p_{\mathbf{x},m},\log p_{\mathbf{x},n}) \\
&= 2\sum_{\mathbf{x}}\sum_{m}\sum_{n} N_{x,m}N_{x,n} \\
&\qquad\times\left[\psi'(\alpha+N_{\mathbf{x},m})\delta_{mn}-\psi'(M\alpha+N_{\mathbf{x}})\right] \\
&= 2\sum_{\mathbf{x}}\left(\sum_{m} N_{\mathbf{x},m}^{2}\psi'(\alpha+N_{\mathbf{x},m})-(N_{\mathbf{x}})^{2}\psi'(M\alpha+N_{\mathbf{x}})\right)
\end{aligned}
\tag{9}
$$

Bayes factors are ratios of the probability of the dataset given two models and their corresponding posterior parameter distributions. In the case of this application, the likelihood completely factorizes into a product of transition probabilities and each model's corresponding term in a Bayes factor is the exponential of its log predictive density (LPD). The LPD can be computed exactly

$$
\begin{aligned}
\mathrm{LPD} &= \log\mathbb{E}_{\mathbf{p}|\mathbf{N}}\left[\Pr\left(\mathbf{N}\mid\mathbf{p}\right)\right] \\
&= \log\mathbb{E}_{\mathbf{p}|\mathbf{N}}\left(\prod_{\mathbf{x}}\prod_{m=1}^{M} p_{\mathbf{x},m}^{N_{\mathbf{x},m}}\right) \\
&= \sum_{\mathbf{x}}\log\left(\frac{B(2\mathbf{N_x}+\alpha)}{B(\mathbf{N_x}+\alpha)}\right).
\end{aligned}
\tag{10}
$$

Related to the LPD is the log pointwise predictive density (LPPD), where the expectation in the LPD is broken down "point-wise." For our application, we will consider trajectories to be points and write the LPPD as

$$
\begin{aligned}
\mathrm{LPPD} &= \sum_{j}\sum_{\mathbf{x}}\log\mathbb{E}_{\mathbf{p_x}|\mathbf{N_x}}\left[\Pr\left(\mathbf{N}_{\mathbf{x}}^{(j)}\mid\mathbf{p_x}\right)\right] \\
&= \sum_{j}\sum_{\mathbf{x}}\log\mathbb{E}_{\mathbf{p_x}|\mathbf{N_x}}\left(\prod_{m=1}^{M} p_{\mathbf{x},m}^{N_{\mathbf{x},m}^{(j)}}\right) \\
&= \sum_{j}\sum_{\mathbf{x}}\log\left(\frac{B(\mathbf{N_x}+\mathbf{N}_{\mathbf{x}}^{(j)}+\alpha)}{B(\mathbf{N_x}+\alpha)}\right).
\end{aligned}
\tag{11}
$$

The WAIC is defined as $\mathrm{WAIC} = -2\mathrm{LPPD} + 2k_{\mathrm{WAIC}}$, where the effective model sizes are computed exactly as

$$
\begin{aligned}
k_{\mathrm{WAIC1}} &= 2\mathrm{LPPD} - 2\sum_j \sum_{\mathbf{x}} \mathbb{E}_{\mathbf{p_x}|\mathbf{N}} \log \mathbf{p}_{\mathbf{x}}^{\mathbf{N}_{\mathbf{x}}^{(j)}} \\
&= 2\mathrm{LPPD} - \sum_j \sum_{\mathbf{x}} \sum_{m=1}^{M} N_{\mathbf{x},m}^{(j)} \mathbb{E}_{\mathbf{p_x}|\mathbf{N_x}} \left(\log p_{\mathbf{x},m}\right) \\
&= 2\mathrm{LPPD} \\
&\quad - 2\sum_j \sum_{\mathbf{x}} \sum_{m=1}^{M} N_{\mathbf{x},m}^{(j)} \left[\psi(N_{\mathbf{x},m}+\alpha) - \psi(N_{\mathbf{x}}+M\alpha)\right] \\
&= 2\mathrm{LPPD} \\
&\quad - 2\sum_{\mathbf{x}} \sum_{m=1}^{M} N_{\mathbf{x},m} \left[\psi(N_{\mathbf{x},m}+\alpha) - \psi(N_{\mathbf{x}}+M\alpha)\right],
\end{aligned}
\tag{12}
$$

and

$$
\begin{aligned}
k_{\mathrm{WAIC2}} &= \sum_j \sum_{\mathbf{x}} \mathrm{var}_{\mathbf{p_x}} \left[\log \Pr\left(\mathbf{N}_{\mathbf{x}}^{(j)} \mid \mathbf{p_x}\right)\right] \\
&= \sum_j \sum_{\mathbf{x}} \mathrm{var}_{\mathbf{p_x}} \left\{ \log \left(\prod_{m=1}^{M} p_{\mathbf{x},m}^{N_{\mathbf{x},m}^{(j)}} \right) \right\} \\
&= \sum_j \sum_{\mathbf{x}} \mathrm{var}_{\mathbf{p_x}} \left[\sum_{m=1}^{M} N_{\mathbf{x},m}^{(j)} \log p_{\mathbf{x},m} \right] \\
&= \sum_j \sum_{\mathbf{x}} \sum_{m=1}^{M} \sum_{n=1}^{M} N_{\mathbf{x},m}^{(j)} N_{\mathbf{x},n}^{(j)} \mathrm{cov}\left(\log p_{\mathbf{x},m}, \log p_{\mathbf{x},n}\right) \\
&= \sum_j \sum_{\mathbf{x}} \sum_{m=1}^{M} \sum_{n=1}^{M} N_{\mathbf{x},m}^{(j)} N_{\mathbf{x},n}^{(j)} \Big[\psi'\left(\alpha + N_{\mathbf{x},n}\right) \delta_{nm} \\
&\qquad\qquad - \psi'\left(M\alpha + N_{\mathbf{x}}\right) \Big] \\
&= \sum_j \sum_{\mathbf{x}} \Big[\sum_{m=1}^{M} [N_{\mathbf{x},m}^{(j)}]^2 \psi'(\alpha + N_{\mathbf{x},m}) \\
&\qquad\qquad - [N_{\mathbf{x}}^{(j)}]^2 \psi'(M\alpha + N_{\mathbf{x}}) \Big].
\end{aligned}
\tag{13}
$$

Finally, as an alternative to information criterion, we may use cross-validation. In particular, the log posterior predictive density under leave-one-out cross validation (LOO) has a particularly simple form,

$$
\mathrm{LOO} = -2\sum_j \sum_{\mathbf{x}} \log \left(\frac{B(\mathbf{N_x}+\alpha)}{B(\mathbf{N_x} - \mathbf{N}_{\mathbf{x}}^{(j)} + \alpha)} \right). \tag{14}
$$

The leave-one-out version of cross validation is a specific case of k-fold cross validation, where k is precisely the number of data points. At the other extreme of this type of cross validation is 2-fold cross validation, which can be computed exactly as

$$\begin{aligned}\mathrm{LPPDCV}_2 = &-2\sum_{j=1}^{J/2}\sum_{\mathbf{x}}\log\left(\frac{B(\mathbf{N}_{\mathbf{x}}^{+}+\mathbf{N}_{\mathbf{x}}^{(j)}+\alpha)}{B(\mathbf{N}_{\mathbf{x}}^{+}+\alpha)}\right)\\ &-2\sum_{j=J/2}^{J}\sum_{\mathbf{x}}\log\left(\frac{B(\mathbf{N}_{\mathbf{x}}^{-}+\mathbf{N}_{\mathbf{x}}^{(j)}+\alpha)}{B(\mathbf{N}_{\mathbf{x}}^{-}+\alpha)}\right),\end{aligned} \tag{15}$$

where $\mathbf{N}_{\mathbf{x}}^{\pm}$ constitute the transition counts of the last $J/2$ trajectories or the first $J/2$ trajectories respectively, so that $\mathbf{N}_{\mathbf{x}}^{-}+\mathbf{N}_{\mathbf{x}}^{+}=\mathbf{N}_{\mathbf{x}}$.

3. Results

3.1. Evaluation of selection criteria

Simulations provided tests of these methods. The test system is composed of $M = 8$ states, with designated start and absorbing states. For each given value of h, I generated for each $\mathbf{x}\in\mathbf{X}_h$ a single set of true transition probabilities drawn from Dirichlet($\mathbf{1}$) distributions. For each of these random networks of a fixed h, I randomly sampled trajectories of a given sample size J 10^4 times, determining from each sample of J trajectories the degree of h chosen by each of the methods.

Fig. 2 provides the frequency that each of five models ($h = 1,\ldots,5$) was chosen based on the selection criteria compared. Each row corresponds to a given true degree of memory $h_{\mathrm{true}}\in\{1,2,3\}$ and sample sizes increase along columns when viewed from left to right. Generally, as the number of samples increases, all selection criteria except for the LPD (Bayes factors) improve in their ability to select the true model. The LPD consistently selects a more-complex (higher-h) model. The AIC does well if h_{true} is small, but requires more data than many of the competing methods in order to resolve larger degrees of memory.

LOO, the two variants of the WAIC, and DIC_1 perform roughly on par. Since each criterion selects the model with the lowest value, it is desirable that $\Delta\mathrm{Criterion}(h) = \mathrm{Criterion}(h)-\mathrm{Criterion}(h_{\mathrm{true}})>0$, for $h\neq h_{\mathrm{true}}$. Fig. 3 explores the distributions of these quantities in the case where $h_{\mathrm{true}}=2$. As sample size J increases, there is clearer separation of these quantities from zero. By $J = 64$, no models where $h = 1$ are selected using any of the criteria. The WAIC_2 and LOO criteria perform about the same whereas the WAIC_1 criteria and the DIC_1 criteria lag behind in separating themselves from zero.

Informed by these tests, this manuscript recommends the LOO criterion:

$$\mathrm{LOO} = -2\sum_{j}\sum_{\mathbf{x}}\log\left(\frac{B(\mathbf{N}_{\mathbf{x}}+\alpha)}{B(\mathbf{N}_{\mathbf{x}}-\mathbf{N}_{\mathbf{x}}^{(j)}+\alpha)}\right), \tag{16}$$

where B is the multivariate Beta function. LOO performed slightly better than WAIC_2 in the included tests, while being somewhat simpler to compute. Eq. 16 decomposes completely into a sum of logarithms of Gamma functions, and is hence easy to implement in standard scientific software packages.

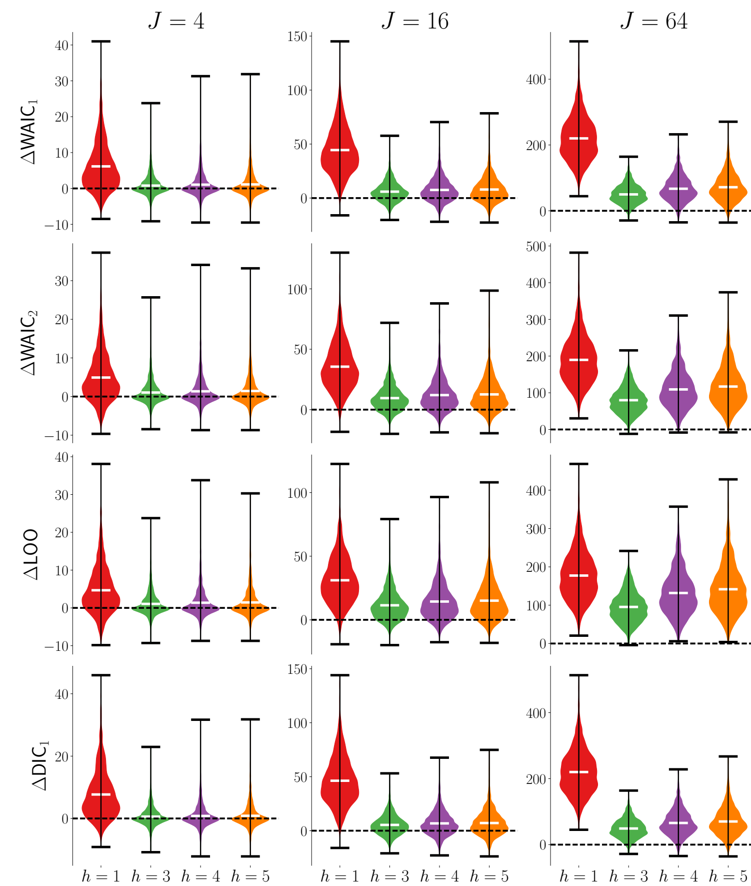

Figure 3. Distributions of computed selection criteria relative to a true model ($h_{\mathbf{true}} = 2$), $\Delta\text{Criterion}(h) = \text{Criterion}(h) - \text{Criterion}(h_{\text{true}})$. Density plots with minimum, maximum, and mean of the selection criteria for each model relative to that of the true model are shown at various sample sizes J. Values above zero mean that the true model is favored over a particular model. Ideally, mass should be above zero for accurate selection of the true model (zero drawn as dashed line).

3.2. The hot-hand phenomenon

The methodology of this manuscript can be used to evaluate this hot-hands effect in the controlled context of free throws. During the 2016-2017 season, in 91 games, LeBron James attempted at least a single free throw, hitting 471 of 693 overall (Fig. 4). Conditioning the hit probabilities by the outcome of the preceding free throw in the same game, James shot a slightly better percentage after missing a free throw than otherwise. However, the $h = 0$ model is favored slightly over $h = 1$ as it appears that the dataset is underpowered for the selection of $h = 1$. In simulations of free throw trajectories, where the number of free throws per game was drawn from a Poisson distribution that approximated the number distribution in the dataset, and outcomes were drawn for the fitted $h = 1$ model, $h = 1$ was chosen slightly under half the time (Fig. 4).

However, examining the model parameters in the case of $h = 1$, one sees that the hitting probabilities are similar in all cases except after a miss (Fig. 4). This observation

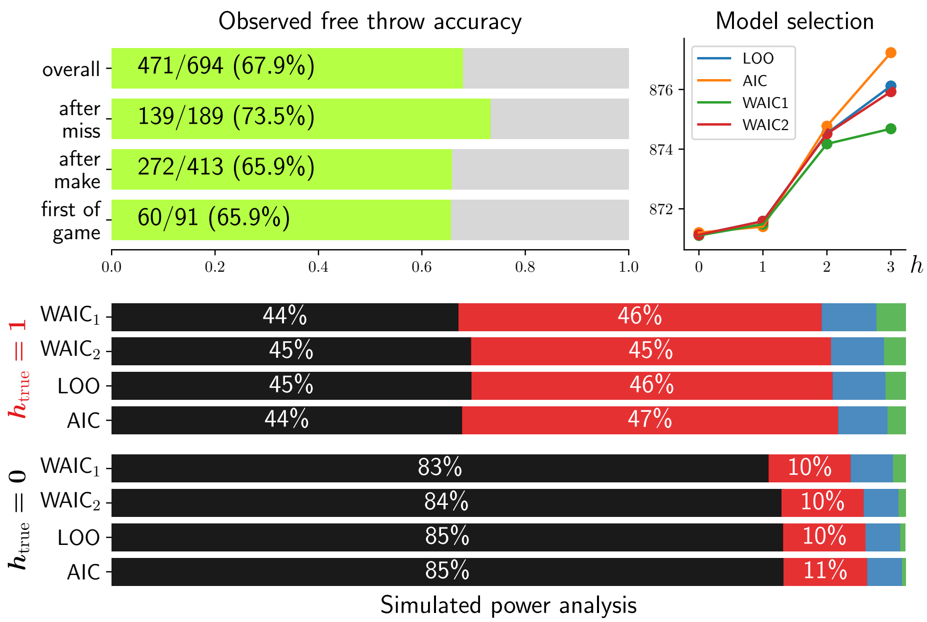

Figure 4. LeBron James' free throw accuracy for the 2016-2017 season and evaluation of the hot hands phenomenon. *Model selection criteria* for degree $h \in \{0,1,2,3\}$ based on four criteria compared. Lower is better and $h = 0$ is slightly favored over $h = 1$ using all criteria. *Simulated power analysis* showing the frequency that each value h is chosen for simulated sets of free throw trajectories.

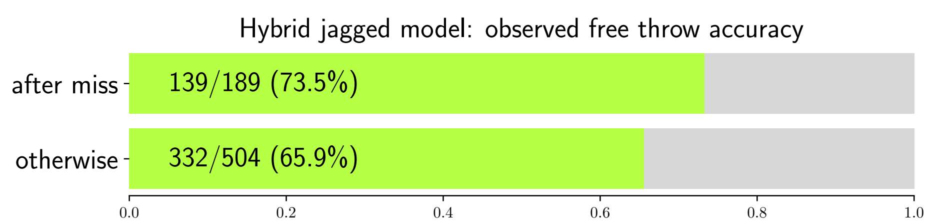

Figure 5. Hybrid jagged memory model for free throw outcome where shots are independent except immediately after a miss. AIC: 869.40, WAIC$_1$: 869.45, WAIC$_2$: 869.52, LOO: 869.52. For reference, all selection criteria for the fully independent ($h = 0$) model are approximately 871 (Fig. 4).

suggests a model with jagged memory: independence of outcome except after a miss. Having one fewer parameter than the full $h = 1$ model, this jagged model is favorable to both the $h = 0$ and $h = 1$ models (Fig. 5). Hence, at least for this season, the most predictive model of James' free throw shooting tells a story of error correction rather than a story of hot hands.

4. Discussion

This manuscript has shown that LOO and its approximation, the WAIC, can learn from data the physical reality of the degree of memory in a system. Regardless of the selection criterion used, the determination of h is not truly certain except asymptotically when an unlimited amount of data are available. However, one may use a simulation procedure like the one used in this manuscript in order to estimate the degree of uncertainty.

Importantly, both the AIC and LPD (Bayes factors) are biased in opposite situations, in opposite directions. For small datasets, the AIC tends to sparsity, which runs counter to the typical situation in linear regression problems where the AIC can favor complexity with too few data, a situation ameliorated by the more-stringent AICc [11]. Bayes factors with flat model priors as investigated here, on the other hand, consistently select a higher value of h given more data. Notably, alternative Bayes factors methods for selecting the degree of memory also include model-level priors that behave like the penalty term in the AIC [24, 23]. Since the upper bound of the LPD is the logarithm of the likelihood found from the MLE procedure, this selection method is more stringent in the low sample-size regime than the pure AIC and hence will suffer from the same bias towards selecting models with less memory.

Models of structure similar to Eq. 1 have appeared in limitless contexts such as analysis of text [18], human digital trails [23], DNA sequences, protein folding [28], and eye movements [6]. As we have seen, many methods tend to asymptotically select the correct model. However, studies are seldom in the asymptotic regime and using these methods to reanalyze data from prior studies may prove fruitful in uncovering previously overlooked memory effects, particularly in systems of a large number of states.

5. Acknowledgements:

I thank the United States Social Security Administration and the Intramural Research Program at the NIH Clinical Center for funding. Additionally, I thank members of the Biostatistics and Rehabilitation Section in the Rehabilitation Medicine Department at NIH, John P. Collins in particular, and also Carson Chow at NIDDK for helpful discussions.

References

[1] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[2] Jeremy Arkes. Revisiting the hot hand theory with free throw data in a multivariate framework. *Journal of Quantitative Analysis in Sports*, 6(1):2, 2010.

[3] Jeremy Arkes. Misses in “hot hand” research. *Journal of Sports Economics*, 14(4):401–410, 2013.

[4] Peter Ayton and Ilan Fischer. The hot hand fallacy and the gambler?s fallacy: Two faces of subjective randomness? *Memory & cognition*, 32(8):1369–1378, 2004.

[5] Michael Bar-Eli, Simcha Avugos, and Markus Raab. Twenty years of ?hot hand? research: Review and critique. *Psychology of Sport and Exercise*, 7(6):525–553, 2006.

[6] Mario Bettenbühl, Marco Rusconi, Ralf Engbert, and Matthias Holschneider. Bayesian selection of markov models for symbol sequences: Application to microsaccadic eye movements. *PloS one*, 7(9):e43388, 2012.

[7] Andrew Bocskocsky, John Ezekowitz, and Carolyn Stein. The hot hand: A new approach to an old 'fallacy'. In *8th Annual Mit Sloan Sports Analytics Conference*, 2014.

[8] Kenneth P Burnham and David R Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media, 2003.

[9] Bruce D Burns. The hot hand in basketball: Fallacy or adaptive thinking. In *Proceedings of the twenty-third annual meeting of the cognitive science society*, pages 152–157, 2001.

[10] Colin F Camerer. Does the basketball market believe in thehot hand,'? *The American Economic Review*, 79(5):1257–1261, 1989.

[11] Gerda Claeskens, Nils Lid Hjort, et al. *Model selection and model averaging*, volume 330. Cambridge University Press Cambridge, 2008.

[12] Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014.

[13] Thomas Gilovich, Robert Vallone, and Amos Tversky. The hot hand in basketball: On the misperception of random sequences. *Cognitive psychology*, 17(3):295–314, 1985.

[14] Clifford M Hurvich and Chih-Ling Tsai. Regression and time series model selection in small samples. *Biometrika*, pages 297–307, 1989.

[15] Richard W Katz. On some criteria for estimating the order of a markov chain. *Technometrics*, 23(3):243–249, 1981.

[16] Jonathan J Koehler and Caryn A Conley. The ?hot hand? myth in professional basketball. *Journal of Sport and Exercise Psychology*, 25(2):253–259, 2003.

[17] Michael Lavine and Mark J Schervish. Bayes factors: what they are and what they are not. *The American Statistician*, 53(2):119–122, 1999.

[18] SS Melnyk, OV Usatenko, and VA Yampol'skii. Memory functions of the additive markov chains: applications to complex dynamic systems. *Physica A: Statistical Mechanics and its Applications*, 361(2):405–415, 2006.

[19] Joshua Benjamin Miller and Adam Sanjurjo. Surprised by the gambler's and hot hand fallacies? a truth in the law of small numbers. 2016.

[20] Leelavati Narlikar, Nidhi Mehta, Sanjeev Galande, and Mihir Arjunwadkar. One size does not fit all: On how markov model order dictates performance of genomic sequence analyses. *Nucleic acids research*, 41(3):1416–1424, 2013.

[21] David Posada and Thomas R Buckley. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic biology*, 53(5):793–808, 2004.

[22] Justin M Rao. Experts? perceptions of autocorrelation: The hot hand fallacy among professional basketball players. *Unpublished technical manuscript. California. San Diego. Downloaded from http://www. justinmrao. com/playersbeliefs. pdf.(July 11th 2012)*, 2009.

[23] Philipp Singer, Denis Helic, Behnam Taraghi, and Markus Strohmaier. Detecting memory and structure in human navigation patterns using markov chain models of varying order. *PloS one*, 9(7):e102070, 2014.

[24] Christopher C Strelioff, James P Crutchfield, and Alfred W Hübler. Inferring markov chains: Bayesian estimation, model comparison, entropy rate, and out-of-class modeling. *Physical Review E*, 76(1):011106, 2007.

[25] Howell Tong. Determination of the order of a markov chain by akaike's information criterion. *Journal of Applied Probability*, 12(03):488–497, 1975.

[26] Sumio Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec):3571–3594, 2010.

[27] Sumio Watanabe. A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, 14(Mar):867–897, 2013.

[28] Zheng Yuan. Prediction of protein subcellular locations using markov chain models. *FEBS letters*, 451(1):23–26, 1999.