Evaluating the hot hand phenomenon using predictive memory selection for multistep Markov Chains: LeBron James' error correcting free throws

Joshua C. Chang*
Clinical Center, National Institutes of Health, Bethesda MD 20892
(Dated: June 26, 2017)

Consider the problem of modeling memory for discrete-state random walks using higher-order Markov chains. This Letter introduces a general Bayesian framework under the principle of minimizing prediction error to select, from data, the number of prior states of recent history upon which a trajectory is statistically dependent. In this framework, I provide closed-form expressions for several alternative model selection criteria that approximate model prediction error for future data. Using simulations, I evaluate the statistical power 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.

This Letter concerns selection of factorized probability models for discrete-state stochastic paths, determining observationally whether and how transitions are dependent on memory. 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

$$\Pr(\xi) = \prod_{l}^{L} p_{x_{l-h}, x_{l-h+1}, \cdots, x_{l-1}, x_l}, \tag{1}$$

where $p_{x_{l-h}, x_{l-h+1}, \cdots, x_l} = \Pr(\xi_l = x_l | \xi_{n-1} = x_{l-1}, \ldots, \xi_{l-h} = x_{l-h})$, and $h \in \mathbb{Z}^+$ 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 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. 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.

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 Letter 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 [7] is chosen. Using these methods, I evaluate the hot hands phenomenon.

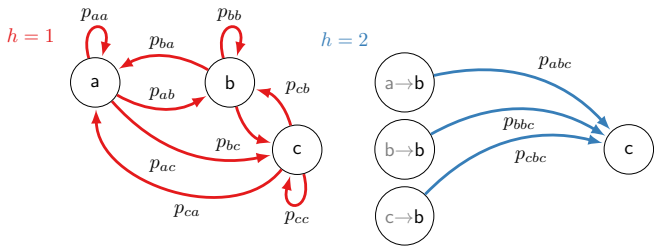

FIG. 1. **Multi-step Markovian processes** by degree of memory h, demonstrated on a three-state network. For $h = 1$, the statistics of the next transition depend solely on the current state, and transition probabilities are indexed by 2-tuple. For $h \geq 2$, statistics depend on history. For $h = 2$, transitions depend on the current state and the previous state, and all probabilities are indexed by 3-tuples. Shown: possible single-step transitions from state **b** to state **c**.

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 $\mathbf{p}_\mathbf{x} = [p_{\mathbf{x},1}, p_{\mathbf{x},2}, \ldots p_{\mathbf{x},M}]$, 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 $\mathbf{p}_\mathbf{x}$ as $\mathbf{p}$.

Generally one has available $J \in \mathbb{N}$ trajectories. Assuming independence, one may write the joint probability, or likelihood, of observing these trajectories as

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

where $N_{\mathbf{x},m}^{(j)}$ is the number of times that the transition

$\mathbf{x} \to m$ occurs in trajectory $\xi^{(j)}$, and $N_{\mathbf{x},m} = \sum_j N^{(j)}_{\mathbf{x},m}$ 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 $\mathbf{N}^{(j)}_{\mathbf{x}}$ 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 $\hat{p}^{\mathrm{MLE}}_{\mathbf{x}} = \mathbf{N}_{\mathbf{x}}/N_{\mathbf{x}}$. Following this approach, the Akaike Information Criterion (AIC), [1, 11, 19] 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 [5]. 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 [10], however, its exact form is problem specific [5]. 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 Letter's approach to rectifying these issues is Bayesian.

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

$$\mathbf{p}_{\mathbf{x}}|\mathbf{N}_{\mathbf{x}} \sim \mathrm{Dirichlet}(\boldsymbol{\alpha}+\mathbf{N}_{\mathbf{x}}) \quad \mathbb{E}_{\mathbf{p}_{\mathbf{x}}|\mathbf{N}_{\mathbf{x}}}[p_{\mathbf{x},m}] = \frac{\alpha + N_{\mathbf{x},m}}{M\alpha + N_{\mathbf{x}}}. \tag{3}$$

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. 3 becomes tightly concentrated about the maximum likelihood estimates.

As pertains to Bayesian model selection, Bayes factors are often used [12, 16]. 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 [8].

The Widely Applicable Information Criterion [20, 21] (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 [20]. 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 [15] does not measure prediction fit for new data [8].

Finally Bayesian variants of cross-validation have recently been proposed as alternatives to information criterion [8]. In our problem, k-fold CV, where data is divided into k partitions, can be evaluated in closed form without repeated model fitting. Using $-2 \times$ LPPD as a metric, this Letter 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.

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 Letter recommends the

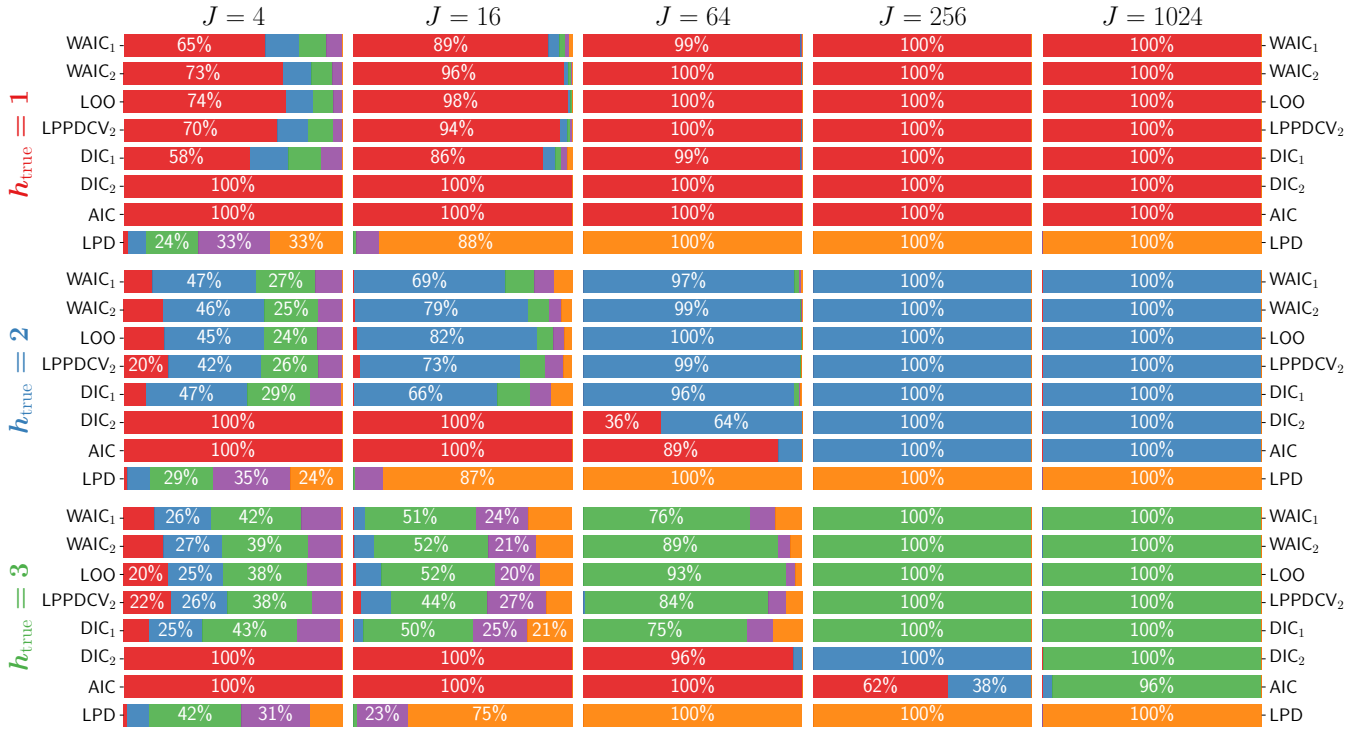

FIG. 2. **Chosen degree of memory** h in simulations for varying true degrees of memory h_{true} and number of observed trajectories J. Rows correspond to model selection under a given degree of memory. Columns correspond to the number of trajectories. Depicted are the percent of simulations in which 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.

LOO criterion:

$$\text{LOO} = -2 \sum_j \sum_{\mathbf{x}} \log \left(\frac{B(\mathbf{N}_{\mathbf{x}} + \boldsymbol{\alpha})}{B(\mathbf{N}_{\mathbf{x}} - \mathbf{N}_{\mathbf{x}}^{(j)} + \boldsymbol{\alpha})} \right), \qquad (4)$$

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. 4 decomposes completely into a sum of logarithms of Gamma functions, and is hence easy to implement in standard scientific software packages.

The hot-hand phenomenon: A real-world application of these methods is the evaluation of "hot-hand phenomenon" (or fallacy) in the context of basketball free-throw shooting. This controversial phenomenon pertains to the belief that recent success (or failure) predicts immediate future success (or failure). In 1985, the first systematic examination of the phenomenon in basketball based on analysis of shooting streaks yielded negative results [9]. Follow-up studies have examined the effects of belief in this phenomenon under the supposition that it is a fallacy [6]. However, recent analyses, using multivariate methods that can account for factors such as shot difficulty [4, 14], have supported the phenomenon, finding the original study to be underpowered [2, 14].

The methodology of this Letter can be used to evaluate this phenomenon 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 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.

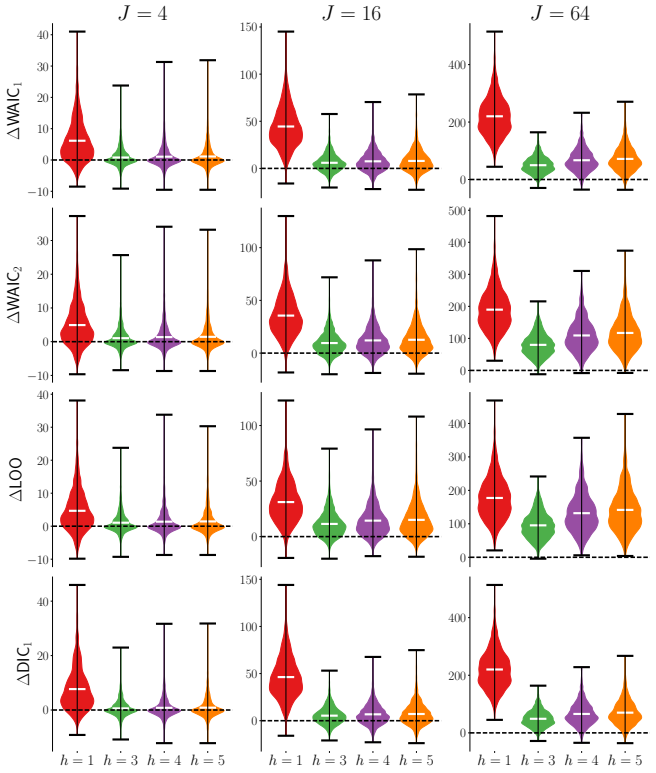

FIG. 3. **Distributions of computed selection criteria relative to a true model ($h_{\text{true}} = 2$),** $\Delta\text{Criterion}(h) = \text{Criterion}(h) - \text{Criterion}(h_{\text{true}})$. Density plots with minimum, maximum, and mean of the selection criteria for each model relative to that of 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).

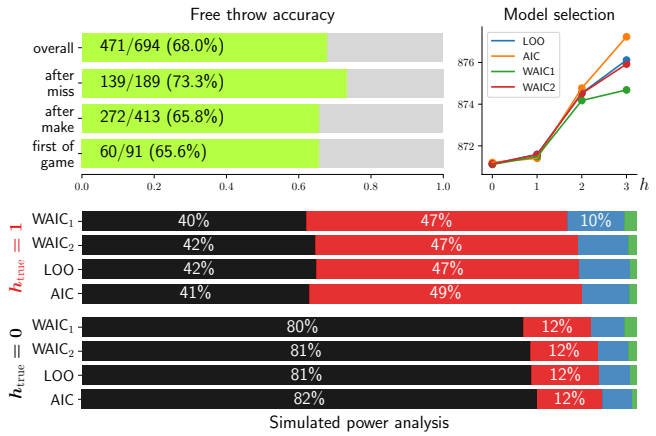

FIG. 4. **LeBron James' free throw accuracy for the 2016-2017 season** and evaluation of the hot hands phenomenon. *Model selection criteria* for degree $h \in \{0, 1, 2, 3\}$ based on four criteria compared. Lower is better and $h = 0$ is slightly favored over $h = 1$ using all criteria. *Simulated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

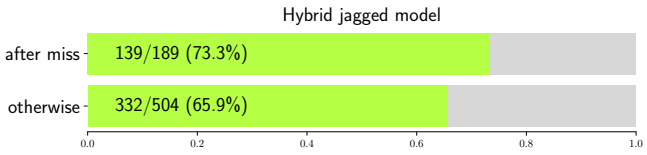

FIG. 5. **Hybrid jagged memory model** for free throw outcome where shots are independent except immediately after a miss. AIC: 869.40, WAIC_1: 869.45, WAIC_2: 869.52, LOO: 869.52. For reference, all selection criteria for the fully independent ($h = 0$) model are approximately 871 (Fig. 4).

This Letter 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 Letter 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 [7]. 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 [17, 18]. 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 [13], human digital trails [17], DNA sequences, protein folding [22], and eye movements [3]. 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.

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.

* joshchang@ucla.edu

[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] 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.

[4] 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.

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

[6] 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.

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

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

[9] 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.

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

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

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

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

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

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

Supplemental Material: Computation of alternate model validation criteria

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

$$\mathrm{AIC} = -2\sum_{\mathbf{x}}\sum_{m=1}^{M} N_{\mathbf{x},m} \log\left(\frac{N_{\mathbf{x},m}}{N_{\mathbf{x}}}\right) + 2M^{q+1}, \quad (5)$$

where 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

$$\mathrm{DIC} = -2\sum_{\mathbf{x}}\sum_{m=1}^{M} N_{\mathbf{x},m} \log\left(\frac{N_{\mathbf{x},m}+\alpha}{N_{\mathbf{x}}+M\alpha}\right) + 2k_{\mathrm{DIC}}, \quad (6)$$

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

$$\begin{aligned}
k_{\mathrm{DIC1}} &= -2\Bigg\{ \sum_{\mathbf{x}}\sum_{m=1}^{M} N_{\mathbf{x},m} \log\left(\frac{N_{\mathbf{x},m}+\alpha}{N_{\mathbf{x}}+M\alpha}\right) \\
&\quad - \sum_{j}\sum_{\mathbf{x}} \mathbb{E}_{\mathbf{p}_{\mathbf{x}}|\mathbf{N}} \log \mathbf{p}_{\mathbf{x}}^{\mathbf{N}_{\mathbf{x}}^{(j)}} \Bigg\} \\
&= 2\Bigg\{ \sum_{\mathbf{x}}\sum_{m=1}^{M} N_{\mathbf{x},m} \log\left(\frac{N_{\mathbf{x},m}+\alpha}{N_{\mathbf{x}}+M\alpha}\right) \\
&\quad - \sum_{j}\sum_{\mathbf{x}}\sum_{m=1}^{M} \mathbf{N}_{\mathbf{x},m}^{(j)} \left[\psi(\alpha+N_{\mathbf{x},m}) - \psi(M\alpha+N_{\mathbf{x}})\right] \Bigg\} \\
&= 2\Bigg\{ \sum_{\mathbf{x}}\sum_{m=1}^{M} N_{\mathbf{x},m} \log\left(\frac{N_{\mathbf{x},m}+\alpha}{N_{\mathbf{x}}+M\alpha}\right) \\
&\quad - \sum_{\mathbf{x}}\sum_{m=1}^{M} N_{\mathbf{x},m} \left[\psi(\alpha+N_{\mathbf{x},m}) - \psi(M\alpha+N_{\mathbf{x}})\right] \Bigg\},
\end{aligned} \quad (7)$$

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

$$\begin{aligned}
k_{\mathrm{DIC2}} &= 2\mathrm{var}_{\mathbf{p}_{\mathbf{x}}}\left[\sum_{\mathbf{x}}\sum_{m} N_{\mathbf{x},m}\log p_{\mathbf{x},m}\right] \\
&= 2\sum_{\mathbf{x}} \mathrm{var}_{\mathbf{p}_{\mathbf{x}}}\left(\sum_{m} N_{\mathbf{x},m}\log p_{\mathbf{x},m}\right) \\
&= 2\sum_{\mathbf{x}}\sum_{m}\sum_{n} N_{x,m}N_{x,n}\mathrm{cov}(\log p_{\mathbf{x},m}, \log p_{\mathbf{x},n}) \\
&= 2\sum_{\mathbf{x}}\sum_{m}\sum_{n} N_{x,m}N_{x,n} \\
&\qquad \times \left[\psi'(\alpha+N_{\mathbf{x},m})\delta_{mn} - \psi'(M\alpha+N_{\mathbf{x}})\right] \\
&= 2\sum_{\mathbf{x}}\left(\sum_{m} N_{\mathbf{x},m}^2\psi'(\alpha+N_{\mathbf{x},m}) - (N_{\mathbf{x}})^2\psi'(M\alpha+N_{\mathbf{x}})\right)
\end{aligned} \quad (8)$$

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

$$\begin{aligned}
\mathrm{LPD} &= \log \mathbb{E}_{\mathbf{p}|\mathbf{N}}\left[\Pr\left(\mathbf{N} \mid \mathbf{p}\right)\right] \\
&= \log \mathbb{E}_{\mathbf{p}|\mathbf{N}}\left(\prod_{\mathbf{x}}\prod_{m=1}^{M} p_{\mathbf{x},m}^{N_{\mathbf{x},m}}\right) \\
&= \sum_{\mathbf{x}} \log\left(\frac{B(2\mathbf{N}_{\mathbf{x}}+\boldsymbol{\alpha})}{B(\mathbf{N}_{\mathbf{x}}+\boldsymbol{\alpha})}\right).
\end{aligned} \quad (9)$$

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

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

The WAIC is defined as $\mathrm{WAIC} = -2\mathrm{LPPD} + 2k_{\mathrm{WAIC}}$,

where the effective model sizes are computed exactly as

$$\begin{aligned}
k_{\mathrm{WAIC1}} &= 2\mathrm{LPPD} - 2\sum_j \sum_{\mathbf{x}} \mathbb{E}_{\mathbf{p_x}|\mathbf{N}} \log \mathbf{p}_{\mathbf{x}}^{\mathbf{N}_{\mathbf{x}}^{(j)}} \\
&= 2\mathrm{LPPD} - \sum_j \sum_{\mathbf{x}} \sum_{m=1}^{M} N_{\mathbf{x},m}^{(j)} \mathbb{E}_{\mathbf{p_x}|\mathbf{N_x}} \left(\log p_{\mathbf{x},m}\right) \\
&= 2\mathrm{LPPD} \\
&\quad - 2\sum_j \sum_{\mathbf{x}} \sum_{m=1}^{M} N_{\mathbf{x},m}^{(j)} \left[\psi(N_{\mathbf{x},m} + \alpha) - \psi(N_{\mathbf{x}} + M\alpha)\right] \\
&= 2\mathrm{LPPD} \\
&\quad - 2\sum_{\mathbf{x}} \sum_{m=1}^{M} N_{\mathbf{x},m} \left[\psi(N_{\mathbf{x},m} + \alpha) - \psi(N_{\mathbf{x}} + M\alpha)\right],
\end{aligned} \tag{11}$$

and

$$\begin{aligned}
k_{\mathrm{WAIC2}} &= \sum_j \sum_{\mathbf{x}} \mathrm{var}_{\mathbf{p_x}} \left[\log \Pr\left(\mathbf{N}_{\mathbf{x}}^{(j)} \mid \mathbf{p_x}\right)\right] \\
&= \sum_j \sum_{\mathbf{x}} \mathrm{var}_{\mathbf{p_x}} \left\{ \log \left(\prod_{m=1}^{M} p_{\mathbf{x},m}^{N_{\mathbf{x},m}^{(j)}} \right) \right\} \\
&= \sum_j \sum_{\mathbf{x}} \mathrm{var}_{\mathbf{p_x}} \left[\sum_{m=1}^{M} N_{\mathbf{x},m}^{(j)} \log p_{\mathbf{x},m} \right] \\
&= \sum_j \sum_{\mathbf{x}} \sum_{m=1}^{M} \sum_{n=1}^{M} N_{\mathbf{x},m}^{(j)} N_{\mathbf{x},n}^{(j)} \mathrm{cov}\left(\log p_{\mathbf{x},m}, \log p_{\mathbf{x},n}\right) \\
&= \sum_j \sum_{\mathbf{x}} \sum_{m=1}^{M} \sum_{n=1}^{M} N_{\mathbf{x},m}^{(j)} N_{\mathbf{x},n}^{(j)} \Big[\psi'\left(\alpha + N_{\mathbf{x},n}\right) \delta_{nm} \\
&\qquad\qquad - \psi'\left(M\alpha + N_{\mathbf{x}}\right) \Big] \\
&= \sum_j \sum_{\mathbf{x}} \Bigg[\sum_{m=1}^{M} [N_{\mathbf{x},m}^{(j)}]^2 \psi'(\alpha + N_{\mathbf{x},m}) \\
&\qquad\qquad - [N_{\mathbf{x}}^{(j)}]^2 \psi'(M\alpha + N_{\mathbf{x}}) \Bigg].
\end{aligned} \tag{12}$$

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,

$$\mathrm{LOO} = -2\sum_j \sum_{\mathbf{x}} \log \left(\frac{B(\mathbf{N_x} + \boldsymbol{\alpha})}{B(\mathbf{N_x} - \mathbf{N}_{\mathbf{x}}^{(j)} + \boldsymbol{\alpha})} \right). \tag{13}$$

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

$$\begin{aligned}
\mathrm{LPPDCV}_2 = &-2\sum_{j=1}^{J/2} \sum_{\mathbf{x}} \log \left(\frac{B(\mathbf{N}_{\mathbf{x}}^{+} + \mathbf{N}_{\mathbf{x}}^{(j)} + \boldsymbol{\alpha})}{B(\mathbf{N}_{\mathbf{x}}^{+} + \boldsymbol{\alpha})} \right) \\
&- 2\sum_{j=J/2}^{J} \sum_{\mathbf{x}} \log \left(\frac{B(\mathbf{N}_{\mathbf{x}}^{-} + \mathbf{N}_{\mathbf{x}}^{(j)} + \boldsymbol{\alpha})}{B(\mathbf{N}_{\mathbf{x}}^{-} + \boldsymbol{\alpha})} \right),
\end{aligned} \tag{14}$$

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 $\mathbf{N}_{\mathbf{x}}^{-} + \mathbf{N}_{\mathbf{x}}^{+} = \mathbf{N}_{\mathbf{x}}$.