Bayesian predictive selection of memory for multistep Markov chains

Joshua C. Chang*
Clinical Center, National Institutes of Health, Bethesda MD 20892
(Dated: June 16, 2017)

Consider the problem of modeling memory for discrete-state random walks using higher-order Markov chains. This Letter introduces a Bayesian framework under the principle of predictive accuracy to select, from data, the number of prior states of recent history upon which a trajectory is statistically dependent. Through simulations, prediction error of several model selection criteria, each provided in closed form, is evaluated.

Our objective is finding factorized probability representations for stochastic paths, determining observationally 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 $\{1, 2, \dots, M\}$. 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 $p_{x_{l-h}, x_{l-h+1}, \cdots, x_l} = \Pr(\xi_l = x_l | \xi_{n-1} = x_{l-1}, \dots, \xi_{l-h} = x_{l-h})$, and $h \in \mathbb{Z}^+$ represents the number of states worth of history 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, $p_{x_1} p_{x_2} \dots p_{x_L}$, and there are essentially M parameters that determine the evolution of the system, where M is the number of states. If $h = 1$, the system is single-step Markovian in that only the current state is relevant in determining the next state. These systems involve M^2 parameters to understand their evolution. In general, if h states of history are required, then the system is h-step Markovian, and M^{h+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.

There is a trade-off between complexity and fitting error that is inherent when varying h. From a statistical viewpoint, complexity results in less-precise determination of network parameters, leading to larger prediction errors. This undesirable consequence of complexity is known as *overfitting*. Conversely, a simple model may not capture the true probability space where paths reside, and fail to catch patterns in the real process.

This Letter evaluates several statistical criteria for selecting the number of states worth of memory to retain in the factorization of Eq. 1, viewing the problem in terms of prediction accuracy. We wish to choose the value of h that yields a model that best predicts new unobserved trajectories [4]. In particular, we seek to test the relationship between finding the best predictive model and uncovering the physical reality.

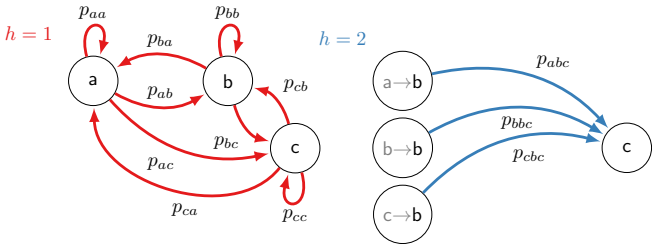

FIG. 1. **Multi-step Markovian processes** by degree of memory h, demonstrated on a three-state network. For $h = 1$, the statistics of the next transition depend solely on the current state, and transition probabilities are indexed by 2-tuple. For $h \geq 2$, the statistics depend on the history. For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

For a fixed degree of memory h, we may look at possible history vectors $\mathbf{x} = [x_1, x_2, \dots, x_h]$ of length h taken from the set $\mathbf{X}_h = \{1, 2, \dots, M\}^h$. For each $\mathbf{x}$, denote the vector $\mathbf{p}_\mathbf{x} = [p_{\mathbf{x},1}, p_{\mathbf{x},2}, \dots p_{\mathbf{x},M}]$, 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 some number of trajectories J. 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 $N_{\mathbf{x},m} = \sum_j N_{\mathbf{x},m}^{(j)}$ is the total number of times the transition is seen across all trajectories.

For convenience, denote $N_\mathbf{x} = \sum_m N_{\mathbf{x},m}$, $\mathbf{N}_\mathbf{x} = [\mathbf{N}_{\mathbf{x},1}, \mathbf{N}_{\mathbf{x},2}, \dots, \mathbf{N}_{\mathbf{x},M}]$, 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 $\mathbf{p_x}$ is found by maximizing the probability in Eq. 2, and can be written easily as $\hat{p}_{\mathbf{x}}^{\mathrm{MLE}} = \mathbf{N_x}/N_{\mathbf{x}}$. Following this approach, the Akaike Information Criterion, [1, 7, 14] which penalizes model complexity, may be used as a metric in order to choose a value of h. Rooted in information theory, the AIC is an approximation of the information loss in the representation of data by a model [3]. The model with the smallest AIC value, and hence with the smallest approximate loss, is chosen.

The aforementioned approach to the problem is simple, however, it has limitations. The AIC, an asymptotic approximation of information loss, is not accurate for for small samples. A modification of the AIC known as the AICc exists [6], however, its exact form is problem specific [3]. More 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. Our 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 $\mathbf{p_x} \sim \mathrm{Dirichlet}(\boldsymbol{\alpha})$, hyper-parameterized by $\boldsymbol{\alpha}$, a vector of size M. This Letter assumes that $\alpha = \mathbf{1}$, corresponding to a uniform prior. This prior, paired with the likelihood of Eq. 2, yields the a posterior Dirichlet distribution with associated expectation,

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

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 asymptotic 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 [8, 11]. 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 pointwise 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 [5].

The Widely Applicable Information Criterion [15, 16] (WAIC) is a Bayesian information criterion that consists of two variants, WAIC_1, and WAIC_2, 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 cross-validation [15]. The commonly used Deviance Information Criterion (DIC) also resembles the WAIC, consisting of two variants DIC_1 and DIC_2. Both variants of the DIC use point estimates of the posterior parameters rather than expectations as used in the WAIC. However, unlike the WAIC and cross-valdation, Bayes factors, DIC, and other methods using the BIC [10] do not have as their objective the maximization of predictive model fit onto new data [5].

Finally Bayesian variants of cross-validation have recently been proposed as alternatives to information criterion [5]. It is of note that the WAIC provides an asymptotic approximation of cross-validation (CV). In our problem, k-fold CV, where data is divided into k partitions, can be evaluated in closed form without repeated model fitting. Using $-2 \times \mathrm{LPPD}$ 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).

To compare these methods, we evaluated them using simulations. Closed-form formulae for computing each model selection criterion are available as SUPPLEMENTAL MATERIAL. Our test system is composed of $M = 8$ states, with designated start and absorbing states. For each given value of h, we generated for each $\mathbf{x} \in \mathbf{X}_h$ a single set of true transition probabilities drawn from $\mathrm{Dirichlet}(\mathbf{1})$ distributions. For each of these random networks of a fixed h, we randomly sampled trajectories of a given sample size J 10^4 times, determining from each sample of J the degree of h to use as 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_{true} and the columns represent increasing sample sizes when viewed from left to right. Generally, it is seen that as the number of samples increases, all selection criteria except for the LPD (Bayes factors) improve in their ability to select the true model. In general, the AIC does well if the true memory is small, but requires more data than many of the competing methods in order to resolve larger degrees of memory.

The two variants of the WAIC, LOO, 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 the masses of these quantities from zero. By $J = 64$, for instance, 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.

As a result of these tests, this Letter recommends the

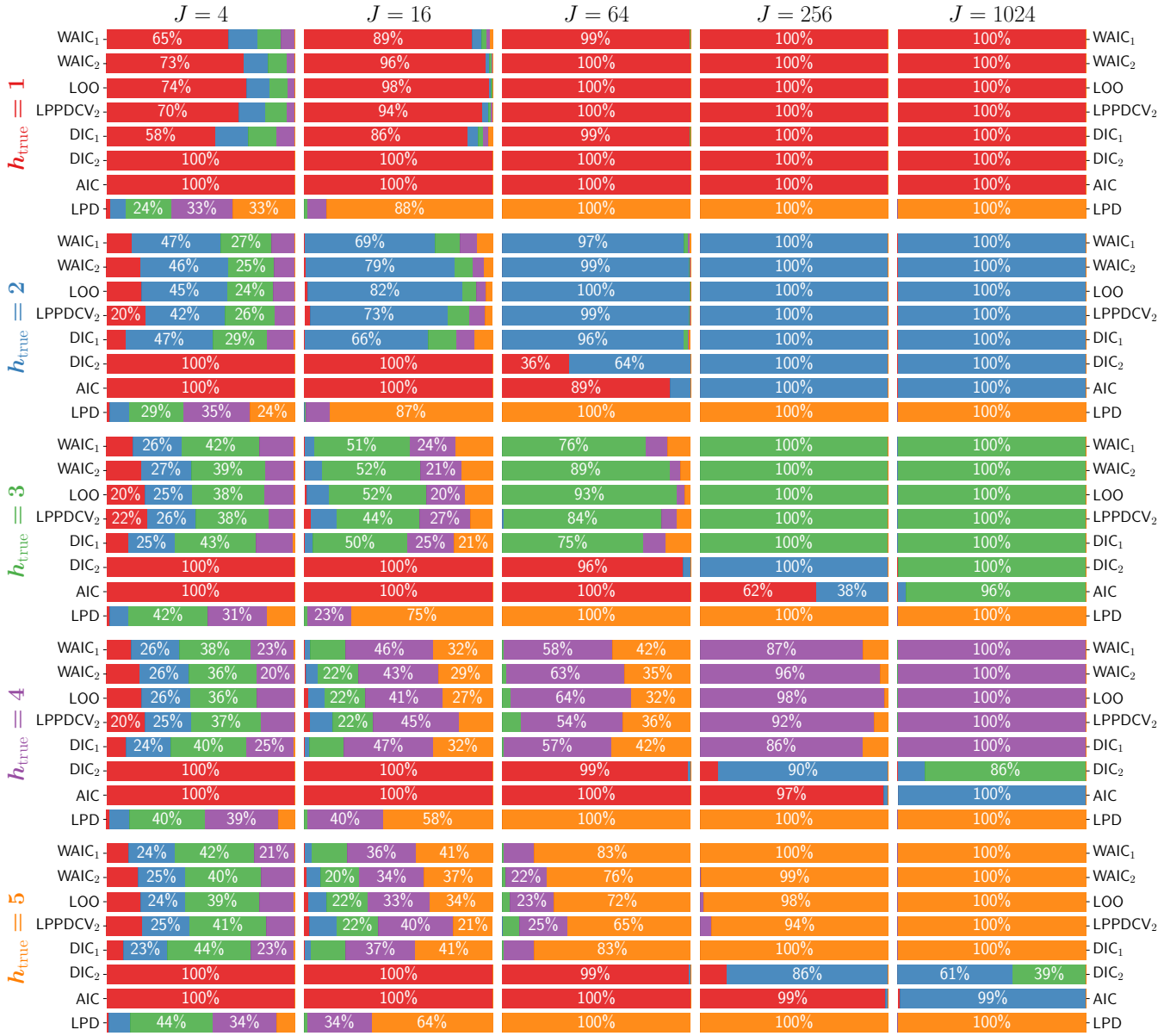

FIG. 2. **Chosen degree of memory** h in simulations for varying true degrees of memory h_{true} and number of observed trajectories J. Rows correspond to model selection under a given degree of memory. Columns correspond to the number of trajectories. Depicted are the percent of simulations in which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

leave-one-out cross validation 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.

This Letter has shown that LOO and WAIC_2 can learn from data the physical reality of the degree of memory in a system. It is important to comment on the uncertainty in such determinations. Regardless of the selection criterion used, the determination of h is not truly certain except in an asymptotic sense where one has an unlimited amount of data 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)

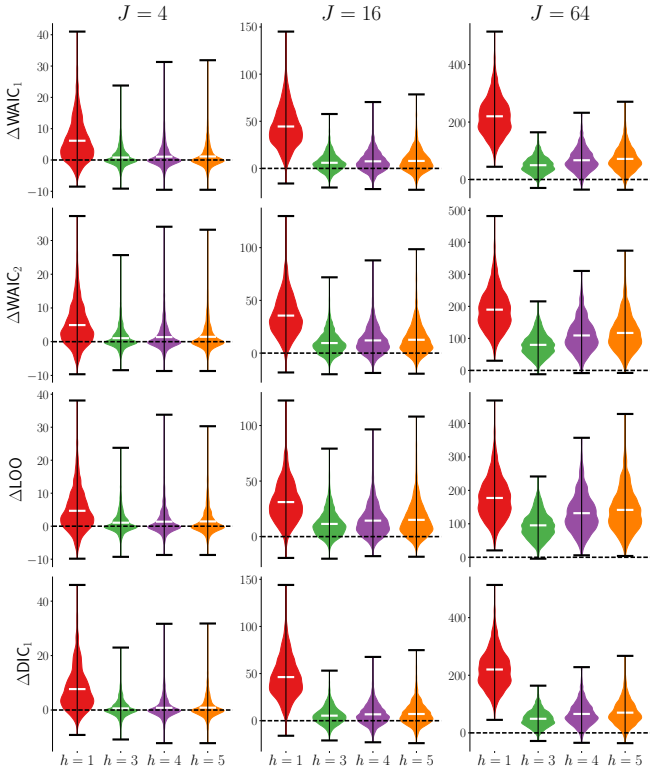

FIG. 3. **Distributions of computed selection criteria relative to a true model ($h_{\mathbf{true}} = 2$),** $\Delta\text{Criterion}(h) = \text{Criterion}(h) - \text{Criterion}(h_{\text{true}})$. Density plots with minimum, maximum, and mean of the selection criteria for each model relative to that of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

are biased, in opposite directions. 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 [4]. Bayes factors with flat model priors as investigated here, on the other hand, are known to automatically penalize complexity indirectly through increased posterior entropy. Yet, when enough data are present, posterior entropy is low while many model parameters may be highly concentrated about zero and the Bayes factor is happy to select the more-complex model even if it is inconsistent with physical truth. 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 [12, 13]. 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 less memory.

Models of structure similar to Eq. 1 have appeared in limitless contexts such as in text analysis [9], analysis of human digital trails [12], DNA sequence analysis, protein folding [17], and biology [2]. As we have seen, many methods tend to asymptotically select the correct model. However, studies are seldom in the asymptotic regime and the use of the methods mentioned in this Letter to reanalyze data from prior studies may prove fruitful in uncovering previously overlooked physics. The general method mentioned in this Letter can also be extended to model averaging in order to generate jagged models with no single fixed degree of memory.

Acknowledgements The author thanks the United States Social Security Administration and the Intramural Research Program at the NIH Clinical Center for funding this research. Additionally, the author is thankful for helpful comments from 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.

* 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] 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.

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

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

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

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

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

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

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

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