Joint Probability Distribution Model for High-Order Markov Chains

Jin-Hwa Kim and Byoung-Tak Zhang
Program in Cognitive Science, School of Computer Science and Engineering,
Seoul National University
{jhkim, btzhang}@bi.snu.ac.kr

Abstract

For computational modeling of cognitive behaviors, dealing with sequential information is key to understanding the mechanisms which enable the agent to learn the changing environment and adjust their behaviors. One of the statistical approaches is the high-order Markov chains (MC). However, a naive high-order Markov chain is not suitable to the real world problem which has a large-scale value space for the observations. So, the parsimonious model, using fewer parameters and maintaining the performance of the prediction, is mandatory. For this reason, we propose a parsimonious model using Joint Probability Distribution (JPD) to reasonably reduce the independent parameters. We use the delayed single-paired observations, instead of continuous sequence observations. Even though this does not take into account the direct relationship between previous observations, the JPD models can reasonably fit to the data with fewer parameters than MC. But, JPDs still have too many parameters to defeat Raftery's Mixture Transition Distribution (MTD) model.

1. Introduction

All information which an agent receives from the given environment through its sensory system is sequential. The agent should appropriately process the sequential information to survive or reproduce. The human memory system has a powerful mechanism to store multiple episodes on various time scales [1]. We can recall some of our birthday parties from childhood, and at the same time, recall birthday parties from last year. In addition to these conscious memory mechanisms, our ability to perform skillful movements and behavioral habits seem to be learnt in an unconscious manner [2].

If an agent has a feasible model for this kind of sequential information, at least two factors can be taken advantage by the owner of this model the owner of the model; temporal completion and prediction. If only the partial sequence is available, lost information can be completed using adjacent patterns. This process is called the retrograde completion for the fragmented information, and is used to robustly recall the complete memory or to fill in the interim. While temporal completion is retrograde, prediction is anterograde. A prediction is about what is coming, or something that has not yet happened, and considers the previous pattern to estimate the next part of a sequence. This can be, in the same way, regarded as a temporal completion, however, due to its usefulness for determining next actions, the prediction has a distinct advantage to survive in the dynamic environment.

Therefore, among the various components of computational modeling for cognitive behaviors, modeling of sequential information is critical to understanding the foundation which enables the agent to adapt and learn the changing environment. With regard to the sequential information, one of the well-studied computational models is Markov Chain.

The Markov chain is a probabilistic model to predict the next value using a previously observed sequence. This model is represented by finite states and its transition probabilities. The name of the model was coined by a Russian mathematician, Andrey Andreyevich Markov. It is a simple but powerful method to model sequential information of various research fields, including meteorology, biology, chemistry, physics and cognitive behavior [3, 4].

However, in a situation where the model is required to observe more than one previous value to get a better estimation, the number of independent parameters of high-order Markov chain increases exponentially with the number of observations, which is the order of the Markov chain. Typically, in practice, the number of parameters from the environment is large enough and can be changed dynamically (it makes it even harder to reduce high dimensional data to a lower dimensional one), thus, the naive high-order Markov chain is not suitable to model such real world problem. So, the parsimonious models, using quite fewer parameters and maintaining the performance of the prediction, are necessary [5, 6, 7].

The most notable study of the parsimonious model is of the mixture transition distribution [5]. The most powerful feature of this model is the usage of a very limited number of parameters to estimate, especially using the lambda weighting values as the delaying factors. The range of weighting values includes even the negative part of real values. However, this approach does not allow the simultaneously learning of independent parameters and the lambda weighting values [8]. For review, see [7].

In the present study, we propose a parsimonious model using joint probability distribution to plausibly reduce the independent parameters. We use the delayed single observations, instead of the continuous sequence observations, then predict the next value using the joint probability of the delayed single observations. Even though it does not take into account the relationship between the previous observations directly, it has far fewer parameters than high-order Markov chains with an increase in possible number of states.

2. Markov Chains

We note that a sequence is represented by a discrete-time random variable X_t, which has a value among the finite set $\{1,\cdots,m\}$ for a given time t. To predict the value of X_t, previous observations, of the same variable in the past time, are used.

The first-order Markov chain uses the Markov assumption, which means that a random variable X_t depends only on the previous observation X_{t-1}. Therefore we can get

$$\begin{aligned} &P(X_t = i_0 \mid X_0 = i_t, X_1 = i_{t-1}, \cdots, X_{t-1} = i_1) \\ &= P(X_t = i_0 \mid X_{t-1} = i_1) \\ &= q_{i_1 i_0}(t) \end{aligned}$$

where $i_t, \cdots, i_0 \in \{1,\cdots,m\}$. In this notation, $q_{i_1 i_0}(t)$ is the probability of the observation of i_0, immediately after the observation of i_1, for a given time t. For the prediction to estimate the next value, we can assume that the transition probabilities are time-invariant.

However, in general, the random variable X_t depends not only on the previous observation X_{t-1}, but on the last l observations. As mentioned before, the time-invariant assumption enables to estimate $q_{i_l \cdots i_0}$ as

$$\hat{q}_{i_l\cdots i_0} = \frac{n_{i_l\cdots i_0}}{\sum_{i_0=1}^{m} n_{i_l\cdots i_0}}$$

where $n_{i_l\cdots i_0}$ is the number of observations, and the log-likelihood of this model can be obtained by

$$LL = \sum_{i_l\cdots i_0=1}^{m} n_{i_l\cdots i_0} \log(\hat{q}_{i_l\cdots i_0}).$$

For the review of Markov chains and the reduced form, please refer to well-organized previous studies [3, 4, 5].

3. Joint Probability Distribution Model

In many cases, the effective number of parameters from the environment is more likely to be large enough to make an efficient model. In Markov chains, a larger number of successive observations are used for the better prediction, which means the more states should be considered for the prediction. The number of independent parameters of the model grows exponentially with the number of observation l. Thus, in this study, the main purpose of the proposed model is to reduce the number of the independent parameters of the model while maintaining the reasonable performance.

First, we notice that many biological mechanisms use the coincidence of neural signals for processing temporal information. For example, Barn owls can localize a sound source by the coincidence detectors in their auditory neural circuits. The coincidence detectors catch the temporal differences of the signals from left and right ears, which are made by the synaptic circuit length or the delayed amounts from it [9, 10, 11]. Furthermore, even in the synaptic formation of binocular systems, the coincidental competition from the inputs of left and right eyes are a critical factor to rearranging the synaptic networks in the visual cortex [12]. If the temporal correlation of the inputs from two eyes are broken, in the case of strabismus, a condition in which the two eyes cannot be perfectly aligned to focus, the binocular vision is not available, even though each visual inputs from the eyes are normal.

The proposed coincidence-based model use the independent 1 to l delayed observations with the present observation. Therefore, we need the l transition matrices with the only m rows. For this model, the $m(m-1)\cdot l$ independent parameters are needed. See the difference with $m^l(m-1)$ of the lth-order Markov chain. The transition probability $\hat{q}_{i_l\cdots i_0}$ is rewritten as

$$\widehat{q}_{i_l\cdots i_0} = \prod_{g=1}^{l} \hat{q}_{i_g i_0}^{(g)} \Big/ \sum_{i_0=1}^{m}\prod_{g=1}^{l} \hat{q}_{i_g i_0}^{(g)}$$

for a given time t,

$$X_{t-g} = i_g, \quad X_t = i_0.$$

Second, we can improve this approach through the numerical maximization method. After defining the coincidence-based model, fit the model to the observed sequence using the gradient descent of the log-likelihood function. Because it reflects the whole sequence while iterating the adjustment procedure gradually, it helps to avoid the locally optimized estimation [8]. The log-likelihood function of this is defined as,

$$LL = \sum_{i_l\cdots i_0=1}^{m} n_{i_l\cdots i_0}\cdot\log\left(\widehat{q}_{i_l\cdots i_0}\right) = \sum_{i_l\cdots i_0=1}^{m} n_{i_l\cdots i_0}\cdot\log\left(\frac{\prod_{g=1}^{l}\hat{q}_{i_g i_0}^{(g)}}{\sum_{i_0=1}^{m}\prod_{g=1}^{l}\hat{q}_{i_g i_0}^{(g)}}\right). \quad (1)$$

For the adjustment of parameters, the partial derivative of the log-likelihood function with respect to the k delayed parameter is

$$\begin{aligned}\frac{\partial LL}{\partial q_{i_k i_0}^{(k)}} &= \frac{n_{i_k i_0}}{q_{i_k i_0}^{(k)}} - \sum_{\substack{i_l\cdots i_{k-1}\cdots\\ i_{k+1}\cdots i_1=1}}^{m} n_{i_l\cdots i_0}\cdot\frac{\prod_{g=1}^{l} q^{(g)}{}_{i_g i_0}}{\sum_{i_0=1}^{m}\prod_{g=1}^{l} q^{(g)}{}_{i_g i_0}}\cdot\frac{1}{q_{i_k i_0}^{(k)}}\\ &= \frac{1}{q_{i_k i_0}^{(k)}}\cdot\left(n_{i_k i_0} - \sum_{\substack{i_l\cdots i_{k-1}\cdots\\ i_{k+1}\cdots i_1=1}}^{m} n_{i_l\cdots i_0}\cdot\frac{\prod_{g=1}^{l} q^{(g)}{}_{i_g i_0}}{\sum_{i_0=1}^{m}\prod_{g=1}^{l} q^{(g)}{}_{i_g i_0}}\right)\\ &= \frac{1}{q_{i_k i_0}^{(k)}}\cdot\left(\left\langle n_{i_k i_0}\right\rangle_{data} - \left\langle n_{i_k i_0}\right\rangle_{model}\right).\end{aligned} \quad (2)$$

And the learning rule is

$$q_{i_k i_0}^{(k)} \leftarrow q_{i_k i_0}^{(k)} + \eta\cdot\frac{\partial LL}{\partial q_{i_k i_0}^{(k)}} \quad (3)$$

where η is a learning rate. For the stability of the learning convergence, we can choose the inverse of the number of iterations plus one [13]. However, the inverse of the prior parameter $q_{i_k i_0}^{(k)}$ in the equation (2) has the similar effect of annealing the learning fluctuations. Therefore, we can set the learning rate η as one.

4. Results

To evaluate how well the models fit to the observations and predict to the unknown upcoming observations considering to the number of independent parameters for the models, we measured the Bayesian information criterion (BIC) for each model with LL [14, 15]. The BIC for the model is defined by

$$BIC = -2LL + p\log(n),$$

where p is the number of independent parameters of the model and n is the number of observations which are used by training. The doubled LL negatively contributes to BIC and the number of independent parameters positively contributes to BIC, thus the lower is the better.

For the experiment, we use the historically observed data of wind directions at Koeberg, South Africa. The data has 744 hourly wind directions for the month of May 1985. Each one of the wind directions can be represented as a number 1 to 16, which means North for 1, NNE for 2, NE for 3, ENE for 4 and so on; clockwise numbering is used. For the terminology of directions, refer to the convention of the cardinal directions. The same data is available in MacDonald & Zucchini (1997).

Table 1 shows the results of the experiments. For the name of models, MC # means the #-order Markov chain, MTD # means the mixture transition distribution with # delays, JPD # means the proposed model, the joint probability distribution model with # delays. The column for the # of parameters describes the number of independent parameters used in the corresponding model. LL and BIC is described in the Joint Probability Distribution Model section and Results section, respectively.

The LL of JPD 2 is higher than MTD 2's. However, for the numbers of parameters for each models are different, the BIC of JPDs have disadvantages to the others. For the LL of JTD 3, although it has more

Table 1: The abbreviation of the models, MC means the given-order Markov chain, MTD means the mixture transition distribution with the given delays, JPD means the proposed model with the given delays. The second column describes the number of independent parameters used in the model. LL and BIC is described in the Joint Probability Distribution Model section and Results section. The MTD results are from Berchtold & Raftery (2002).

Model	# of parameters	LL	BIC
MC 1	11	-413.3	899.1
MC 2	27	-374.9	927.9
MC 3	39	-346.2	949.7
MTD 2	**11**	**-393.4**	**859.3**
MTD 3	12	-393.2	865.6
JPD 2	24	-389.6	937.4
JPD 3	36	-407.4	1052.2

parameters for learning, its LL is worse than JTD 2's. Notice that this is unlikely for MC or MTD cases.

Figure 1 shows that the LL of the JPD 2 is asymptotically converged to around -389.6. JPDs can reasonably fit to the data with fewer parameters than MC. It also empirically proves that the learning rule (3) works for the given observations.

5. Discussion

Although JPD 2 is better than MTD 2 at LL, due to its redundancy in parameters, the BICs of JPDs have inferior values to the others. The critical reason for these results is that there are quite a number of independent parameters, which are not reduced for the not-observed prior series. In the cases of MC 2, there are many zero probabilities in the transition matrix because of its absence in the series. Therefore, eventually the model has fewer number of parameters than 48. Moreover, for the learning, the wind direction data have a relatively small value space to compare with MC.

Additionally, the LL of JTD 3, which has more parameters for fitting, is worse than JTD 2's. This indicates that there is a limit to appropriately model for the given data using the longer delayed cues. The possible interpretation of the phenomena is that the elements of the transition matrix Q of g is fluctuating during the iteration of learning when the amount of delay g is lengthy. It can be the weakest point of this proposed model.

In summary, the parsimonious model, which uses the joint probability distribution, is proposed to make a model using fewer parameters while maintaining the reasonable performance. We show that the likelihood of the model is stably converged to describe the given data. However, it still has many parameters to defeat Raftery's MTD because the MTD method reduced the number of independent parameters using the weighting parameters even though it needs the EM learning with much more risk of fluctuations. Further research on other types of data is possible, and we look forward to applying this model on the physically constrained situations, which only allows a simple Hebbian learning rule.

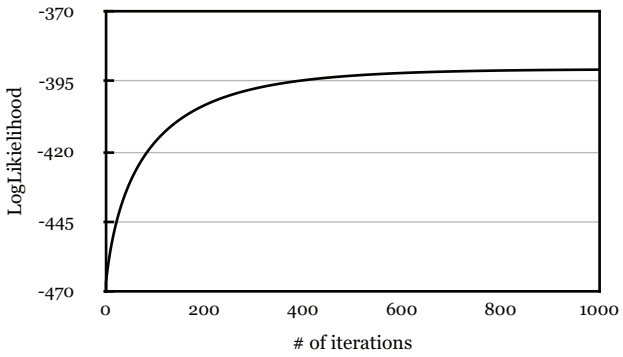

Figure 1: Using the learning rule of the equation (3), the iteration of learning is run up to 1000 times. The log-likelihood of the JPD 2 is asymptotically converged to the around of -389.6. JPDs can reasonably fit to the data with the fewer parameters than MC.

6. Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2010-0017734-Videome), supported in part by KEIT grant funded by the Korea government (MKE) (KEIT-10035348-mLife, KEIT-10044009).

7. References

[1] Fortin NJ, Agster KL, Eichenbaum HB, Critical role of the hippocampus in memory for sequences of events. Nat Neurosci., 2002.

[2] Henke K, A model for memory systems based on processing modes rather than consciousness. Nat Rev Neurosci 11:523–532, 2010.

[3] Kemeny, JG, & Snell, JL, Finite markov chains (Vol. 210). New York: Springer-Verlag, 1976.

[4] Brémaud P, Markov chains: Gibbs fields, Monte Carlo simulation, and queues (Vol. 31). springer, 1999.

[5] Pegram G, An autoregressive model for multilag Markov chains. Journal of Applied Probability:350–362, 1980.

[6] Raftery AE, A model for high-order Markov chains. Journal of the Royal Statistical Society Series B (Methodological):528–539, 1985.

[7] Berchtold A, Raftery AE, The mixture transition distribution model for high-order Markov chains and non-Gaussian time series. Statistical Science:328–356, 2002.

[8] Dempster AP, Laird NM, Rubin DB, Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B (Methodological):1–38, 1977.

[9] Jeffress, L. A., A place theory of sound localisation. J. Comp. Physiol. Psychol., 41:35–39, 1948.

[10] Knudsen, E. I., Blasdel, G. G., and Konishi, M., Sound localization by the barn owl (tyto alba) measured with the search coil technique. J. Comp. Physiol., 133:1–11, 1979.

[11] Carr, C. E. and Konishi, M., A circuit for detection of interaural time differences in the brain stem of the barn owl. J. Neurosci., 10:3227–3246, 1990.

[12] Sireteanu R,Thiel A, Fikus S, Iftime A, Distortions in two-dimensional visual space perception in strabismic observers, Vision Research, 33, 677–690, 1993.

[13] Bertsekas DP, Tsitsiklis JN, Neuro-Dynamic Programming. Athena Scientific, 1996.

[14] Katz RW, On some criteria for estimating the order of a Markov chain. Technometrics 23:243–249, 1981.

[15] MacDonald WZAIL, Hidden Markov Models for Time Series. :89-92;167-180, 2009.