Songklanakarin J. Sci. Technol.
43 (3), 824-833, May - Jun. 2021

http://www.sjst.psu.ac.th

Original Article

Widely applicable information criterion for estimating the order in a hidden Markov model

Safaa K. Kadhem[1] and Sadeq A. Kadhim[2*]

[1] *Department of Mathematics and Computer Applications, College of Science, Al-Muthanna University, Samawah, Al Muthanna, Iraq*

[2] *Ministry of Higher Education and Scientific Research, Baghdad, Iraq*

Received: 15 April 2020; Revised: 1 June 2020; Accepted: 10 June 2020

Abstract

This paper considers the determination of the order of hidden Markov models. Recently, a proposed predictive measure, the so-called widely applicable information criterion (WAIC), was derived. This criterion is a convenient alternative to the cross-validation approach due to its less computation processes and quick evaluation. We studied the properties of this criterion applied to hidden Markov models (HMMs) under the Bayesian principle. Such models include serial dependence and overdispersion of observed data. We investigated this criterion via simulation studies and a real data application. It is shown that the introduced criterion performs better with less complicated models, while it tends to over fit some more complicated models.

Keywords: hidden Markov chains models, Markov chain Monte Carlo, integrated posterior predictive density, model selection

1. Introduction

There are many techniques proposed in literature to select the best model for hidden Markov models (HMMs). One of the common methods is the Bayes factors (BF) approach proposed by Kass and Raftery (1995). However, Han and Carlin (2011) mentioned that the approach may be inappropriate for models with high-dimensions. In addition, it can be more sensitive to the prior's specifications (Ando, 2010; Gelman, Hwang & Vehtari, 2014). Different methods have been used for HMMs under the frequentist principle such as the Akaike information criterion (AIC) (Akaike, 1973) and the Bayesian information criterion (BIC) (Schwarz, 1978). However, these criteria can yield problems such as under-fitting or over-fitting due to the irrational behavior of the likelihood function included in these criteria (Johnson, 2007). Furthermore, models evaluation based only on a point estimate for the model parameter using these approaches does not taking into account the full uncertainty in the parameters. Kadhem, Paul and Kaimi (2018) used Bayesian versions for the AIC and BIC that addresses the model selection problem for these models.

The deviance information criterion (DIC), developed by Spiegelhalter, Best, Carlin and van der Linde (2002), has been used for different models including the high-dimensional models. This criterion has the advantage of being easy to calculate in common used software such as WinBUGS. However, this criterion has an unsatisfactory behavior in latent variables models because of the unavailability of closed forms for the likelihood function of these models. (Celeux, Forbes, Robert & Titterington, 2006). In addition, it is based on a concept of so-called focus which may be chosen incorrectly in practice (Spiegelhalter, Best, Carlin, & van der Linde, 2002). Richardson and Green (1997) introduced a method of so-called the reversible jump MCMC to select the best independent mixture. However, it requires some care with respect to mixing performance within each model spaces and amongst competing model spaces. In addition, Fan and Sisson (2011) suggested that this method may have some convergence problems, and also challenges with respect to prior specification on the number of hidden states K.

*Corresponding author
Email address: sadiq2061@gmail.com

This article aims at providing a new prediction-based criterion for HMMs that are based on predictive performance, which is the widely applicable information criterion (WAIC). This criterion was introduced by Watanabe (2009) as an asymptotic version of the leave-one-out cross-validation (LOO-CV). This measure has also received attention from many researchers such as Vehtari and Ojanen, 2012; Gelman, Hwang and Vehtari, 2014. The main benefit of this approach is that it needs to less computational processes than the LOO-CV (Gelman, Hwang & Vehtari, 2014). With the few applications for this criterion, its properties however have not been investigated for the process of determining the numbers of states in HMMs. In this paper we investigated this criterion, in which the existence of the serial dependence and over dispersion in data, believing their influence on the performance of criterion. More specifically, we investigated the sensitivity of the proposed criterion under two scenarios, across the assumption of a fixed number of different complexities. The first scenario included generating data sets with different degrees of serial dependence, while the second scenarios considered the existence of different degrees of over dispersion in the data.

This article is structured in six sections. Section 2 briefly presents the Bayesian definition of the hidden Markov model and also the proposed criterion. Section 3 includes the simulation studies to investigate the new model selection methodology. Section 4 illustrates the results of simulation study. Section 5 addresses the assessing the criterion via a real data application. Conclusions on the paper are presented in Section 6.

2. Materials and Methods

2.1 Bayesian definition of HMMs

The HMMs have pair of random processes, each one with special situation. The first is called the observed process and is denoted by $Y_i = y_i, i = 1,2,\ldots,n$, while the second is called the unobserved process or hidden which is satisfied the Markov properties, and denoted by $Z_i = z_i\,;\ i = 1,2,\ldots,n$. The observable process Y_i can be determined only when the hidden process Z_i is known (Zucchini & MacDonald, 2009). This paper focuses on finite state-space HMMs with a discrete-time, where given a hidden state z at time i the observed process follows a parametric distribution. The HMMs are represented by a set of the parameters symbolized by Θ, where $\Theta = (\pi, A, \theta)$, and the number of unobservable states symbolized by K, which are explained by the following (Rabiner, 1989; Bishop, 2006):

1. The number of unobservable states K which is defined on the discrete state space $\{1,2,\ldots,K\}$.

2. The initial distribution of state represented by the vector $\boldsymbol{\pi} = \{\pi_k\}$, where the element π_k refers to the probability that the system is in the state k at the time $i = 1$, i.e., $\pi_k = \Pr\{z_1 = k\}$, $1 \le k \le K$, where z_1 refers to system's state at the time 1.

3. The transition probabilities of states represented by matrix $A = \{a_{jk}\}$, where the element a_{jk} refers to the probability that the system in the state k at time i, given that the system in the state j at time $i-1$, i.e., $a_{jk} = \Pr\{z_i = k / z_{i-1} = j\}$ such that a_{jk} satisfy the normal stochastic constraints, i.e. $a_{jk} \ge 0$ and $\sum a_{jk} = 1$, where $1 \le j,k \le K$. $_{k=1}$

4. The parameter θ, where θ is a state-based distribution that can take one of parametric distributions. For example, the parameter θ can be expressed to parameter(s) of the normal distribution or Poisson.

By using the Bayes theorem, the Bayesian model is defined as (Frühwirth-Schnatter, 2006):

$$\Pr(\Theta|\boldsymbol{y}) \propto \Pr(\boldsymbol{y}|\Theta)\Pr(\Theta), \tag{1}$$

Where $Pr(\Theta|\boldsymbol{y})$ refers to the posterior distribution, $\Pr(\boldsymbol{y}|\Theta)$ refers to the observed data likelihood and $Pr(\Theta)$ denotes the prior distribution. The use of Bayesian inference for HMMs has challenge due to the complexity of the evaluation of its likelihood $\Pr(\boldsymbol{y}|\Theta)$. Hence, Monte Carlo Markov Chain (MCMC) techniques have been employed to solve such problem. An approach so-called the Data Augmentation (Tanner & Wong, 1987) is often used to ease the estimation process of the model parameters. In context of HMMs, this method is being employed, in which the unobservable states are introduced as "missing data" that are augmented to the parameter space in the sampler (Chib, 1996). The posterior distribution can take the following formula

$$\Pr(\Theta, \boldsymbol{z}|\boldsymbol{y})\, \alpha\, \Pr(\boldsymbol{y}, \boldsymbol{z}|\Theta)\Pr(\Theta)\, \alpha\, \Pr(\boldsymbol{y}|\boldsymbol{z}, \Theta)\Pr(\boldsymbol{z}|\Theta)\Pr(\Theta)\,. \tag{2}$$

Where $Pr(\boldsymbol{y}, \boldsymbol{z}|\Theta) = Pr(\boldsymbol{y}|\Theta, \boldsymbol{z})Pr(\boldsymbol{z}|\Theta)$ refers to the complete data likelihood and $Pr(\Theta)$ refers to the prior distribution on Θ. Briefly, given a transition matrix A and an initial distributionπ, the set of unobservable states $\mathbf{z} = (z_1, z_2, \ldots, z_n)$ is modeled as $Pr(z) = Pr(z_i, z_{i-1}, \ldots, z_1; \boldsymbol{A}; \pi)$. According to the Markovian property, it is expressed as

$$\begin{aligned}\Pr(\boldsymbol{z}) &= \Pr(z_i|z_{i-1};\boldsymbol{A})\Pr(z_{i-1}|z_{i-2};\boldsymbol{A}),\ldots\Pr(z_2|z_1;\boldsymbol{A})\Pr(z_1|z_0;\pi)\\ &= \Pr(z_1|\pi)\prod_{i=2}^{n}\Pr(z_i|z_{i-1};\boldsymbol{A}).\end{aligned} \tag{3}$$

The observation $\boldsymbol{y}_i$, given $\boldsymbol{z}_i$, is sampled independently from a certain parametric distribution,

$$\Pr(y_i|\Theta, z_i = k) = \prod_{i=1}^{n} f_k(y_i|\theta_k), \tag{4}$$

Where $f_k(.\mid\theta_k)$ refers to the density function parameterized by θ_k at kth state. From equations (3) and (4), the complete-data likelihood of HMMs, is then defined as

$$\Pr(\boldsymbol{y},\boldsymbol{z}|\Theta,\pi,\boldsymbol{A}) = \Pr(z_1|\pi)\prod_{i=2}^{n}\Pr(z_i|z_{i-1},\boldsymbol{A})\prod_{i=1}^{n} f_k(y_i|\theta_k). \tag{5}$$

By summing over all possible hidden states in the complete data-based likelihood, in equation (5), we obtain the observed data likelihood

$$\Pr(\boldsymbol{y}|\Theta) = \sum_{z}[\Pr(z_1|\pi)]\prod_{i=2}^{n}\Pr(z_i|z_{i-1},\boldsymbol{A})\prod_{i=1}^{n} f_k(y_i|\theta_k),]. \tag{6}$$

The calculation of $Pr(\boldsymbol{y}|\Theta)$ for HMMs in equation (6) requires $O(K^n T)$ of computational processes (Rabiner, 1989; Bishop, 2006). To ease its computation, we use the forward-backward recursion method proposed by (Rabiner, 1989). In order to complete the definition of Bayesian HMMs, we have to specify prior distributions on the model parameters. We assume independent Dirichlet priors (Frühwirth-Schnatter, 2006) on the initial distribution π and each row $\{a_{j.}\}$ in the transition matrix **A**, thus,

$$\Pr(\pi) = \prod_{k=1}^{K}\pi_k\,\alpha\prod_{k=1}^{K}\pi_k^{\delta_k-1} = Dir\,(\delta_1,\delta_2,\ldots,\delta_K)\ , \tag{7}$$

$$\Pr(\boldsymbol{A}) = \prod_{j=1}^{K}a_j\ \alpha\prod_{j=1}^{K}a_{j.}^{\delta_j-1} = Dir\,(\delta_1,\delta_2,\ldots,\delta_K)\ , j,k = 1,2,\ldots,K \tag{8}$$

The symbol δ is a hyper-parameter of Dirichlet distribution and "Dir" is shortcut to Dirichlet distribution. With respect to the parameter θ, we specify priors on θ, expressed by $\Pr(\theta|\varphi)$, where φ refers to a conjugate hyper-parameter. The posterior distribution of HMMs in equation (2) is then given as

$$\begin{aligned}&\Pr(\boldsymbol{\Theta},\boldsymbol{z},\boldsymbol{y})\\ &= \Pr(\boldsymbol{y},\boldsymbol{z}|\theta,\pi,\boldsymbol{A})\Pr(\mathbf{z}|\pi,\mathbf{A})\Pr(\theta|\varphi)\\ &= \sum_{z}[\Pr(z_1|\pi)]\prod_{i=2}^{n}\Pr(z_i|z_{i-1},\boldsymbol{A})\prod_{i=1}^{n} f_k(y_i|\theta_k)]\ \times\Pr(\theta|\varphi)Dir(\pi|\delta)\prod_{j=1}^{K}Dir\,(a_{j.}|\delta).\end{aligned} \tag{8}$$

The form of posterior distribution in equation (8) is not feasible as it is a sum over K^n elements for single chain, and this complexity in model increases as the number of unobservable states K and the data length n are increasing. For this purpose, we use the MCMC approach, called the Gibbs sampler (Geman & Geman, 1984), to sample the model parameters, Θ. With respect to the hidden states, z, they are sampled using so-called the forward-backward Gibbs (FBG) algorithm (Chib, 1996; Scott, 2002). This method is based on simulating the entire sequence of hidden states from the posterior distribution, conditional on the model parameters, through a backward recursion.

2.2 Basic definition of the WAIC

Let assume $\tilde{y} = (\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_n)$ is a sequence of out-of-sample predictions. The out-of-sample logarithm predictive density for a single predicted observation, can be defined as (Gelman, Hwang & Vehtari, 2014):

$$\log Pr_{post}(\tilde{y}_i) = \log\{E_{post}[\Pr(\tilde{y}_i|\theta)]\} = \log\left\{\int \Pr(\tilde{y}_i|\theta)\,\Pr_{post}(\theta|\boldsymbol{y})\; d\theta\right\} \tag{9}$$

The term $\Pr_{post}(\theta|\mathbf{y})$ refers to the posterior distribution and $\log \Pr_{post}(\tilde{y}_i)$ refers to the logarithm predictive density of the predicted observation $\tilde{y}_i, i = 1,2,\ldots,n$ outputted by the posterior distribution $\Pr_{post}(\theta\,|\mathbf{y})$. Hence, we can find expected values of all future points $\tilde{y}_i$. The expectation of logarithm predictive density (elpd) can be given as:

$$elpd = E_f\{\log Pr_{post}(\tilde{y}_i)\} = \int \{\log Pr_{post}(\tilde{y}_i)\} f(\tilde{y}_i)\, d\boldsymbol{y}, \tag{10}$$

where $f(.)$ refers to some data-based distribution. Gelman, Hwang and Vehtari (2014) suggested that the posterior distribution, $Pr_{post}(.)$, is known, but the distribution based on real data $f(.)$ is unknown. Therefore, they proposed to use the within-sample data with a bias correction term. Hence, the log-pointwise predictive density (lppd_y), based on available within-sample data y, can be defined as follows:

$$\begin{aligned}\widehat{lppd}_y &= \log\prod_{i=1}^{n} Pr_{post}(y_i) = \sum_{i=1}^{n}\log E_\theta[\Pr(y_i|\theta)],\\ &= \sum_{i=1}^{n}\log\int \Pr(y_i|\theta)\Pr(\theta|\boldsymbol{y})\, d\theta.\end{aligned} \tag{11}$$

An approximate to the integral in equation (11) is then obtained by integrating out the posterior samples, $\theta^{(m)}, m = 1,2,\ldots,M$ in an MCMC run. Given $lppd_y$, Gelman, Hwang and Vehtari (2014) introduced two definitions for the effective number of parameters p_{WAIC}:

$$p_{WAIC_1} = 2\sum_{i=1}^{n}\{\log[E_\theta \Pr(y_i|\theta)] - E_\theta \log \Pr(y_i|\theta)\}, \tag{12}$$

and the second form can be given by

$$p_{WAIC_2} = \sum_{i=1}^{n} V_\theta[\log \Pr(y_i|\theta)], \tag{13}$$

The term V_θ refers to the variance of individual terms in the logarithm of predictive density summed over the n observed points. The second version p_{WAIC_2} is more stable as suggested by Gelman, Hwang and Vehtari (2014). The WAIC can be then given by

$$\begin{aligned}WAIC &= -2\,\widehat{lppd}_y + 2p_{WAIC_j},\\ &= -2\sum_{i=1}^{n}\log E_\theta[\Pr(y_i|\theta)] + 2p_{WAIC_j};\; j\\ &= 1, 2,.\end{aligned} \tag{14}$$

2.3 The WAIC for HMMs

Let assume a set of observations $\boldsymbol{y}$ induced from a HMM and a set of latent variables $\boldsymbol{z}$ such that each z_i is specified for each corresponding observation y_i. Then, it can model $\boldsymbol{y}$ and $\boldsymbol{z}$ for some distribution parameterized by $\Theta = (\pi, A, \theta)$. By following Gelman, Hwang and Vehtari (2014), we can define so-called the integrated pointwise predictive density (ilppd) as follows

$$\widehat{ilppd}_y = \log \prod_{i=1}^{n} Pr_{post}(y_i) = \sum_{i=1}^{n} \log E_{\boldsymbol{z},\boldsymbol{\theta}} \left[\Pr(y_i|\boldsymbol{\theta},\boldsymbol{z})\,|\boldsymbol{y}\right],$$

$$= \sum_{i=1}^{n} \log \left\{ \int_{\Theta} \int_{z} \Pr(y_i|\boldsymbol{\theta}, z_i)\ Pr(\boldsymbol{\theta},\boldsymbol{z}|\boldsymbol{y})\ d\boldsymbol{\theta} d\boldsymbol{z} \right\}, \tag{15}$$

which can be approximated over integrating out the model parameter, $\boldsymbol{\theta}$, and latent variables, $\boldsymbol{z}$. The term $Pr(y_i|\boldsymbol{\theta},\boldsymbol{z})$ in above equation refers to the pointwise predictive density of point data, weighted by the joint posterior distribution, $Pr(\boldsymbol{\theta},\boldsymbol{z}|\boldsymbol{y})$, of the model parameters. Thus, by integrating individually over each latent variable z_i and the parameter θ, we can obtain the integrated pointwise predictive density of each data point, y_i. The integrated logarithm pointwise predictive density in equation (15) is approximated by the posterior draws of the model parameters over MCMC sweeps. To define the effective number of parameter p_{WAIC}, we use the second version as it is more stable (Gelman, Hwang & Vehtari, 2014):

$$p_{WAIC_2} = \sum_{i=1}^{n} V_{z,\theta} \left[\log \Pr(y_i|\boldsymbol{z},\theta)\right], \tag{16}$$

where $V_{z,\theta}$ is the variance of individual terms in the ilppd summed over the n observations. Hence, the WAIC for the HMMs can represented by

$$WAIC = -2\,\widehat{ilppd}_y + 2p_{WAIC_2},$$

$$= -2 \sum_{i=1}^{n} \log E_{z,\theta} \left[\Pr(y_i|\boldsymbol{z},\theta)\right] + 2p_{WAIC_2}. \tag{17}$$

In the appendix of this paper, approximated WAIC, ilppd and p_{WAIC} are given.

3. Results and Discussion

3.1 Simulation study

This section contains two scenarios designed as a simulation study to include assessing the performance of the proposed criterion. We take into account a HMM in which the parameter space follows the Poisson distribution (Zucchini, & MacDonald, 2009). The two scenarios designed under this study are given in Table (1) which their details are more explained in Table 2. We assume three true models with different complexities, K_0=2, K_0=3 and K_0=4, where K_0 refers to the order of assumed true model. From this experiment, we assume there is over-dispersion problem in data (i.e. variance > mean) as in Table 2 and also assume degrees of serial dependence in data as shown in Figure 3. The aim is to know to how our proposed criterion is sensitive to these assumptions. In the first scenario, we do the follows. For the first true model with 2-state, K_0 = 2, we assume a weak serial dependence through the structure S1, while we expect strong correlation between the data through the structure S2. The same way, we repeat the same above scenario for the models K_0 = 3: S3 for a simple serial dependence and S4 for strong serial dependence, and with respect to the model K_0 = 4: S5 for a simple serial Dependence and S6 for strong serial dependence. Overall, given n=500 observations, we report the percentage of times out of 100 replications that the model out of K = 2, …, 7 competition models, fitted to data, will select the correct order K_0 as shown in Table (3). Table (4) presents the frequencies of the values of the ilppd, WAIC and p_{WAIC} for all true models, each one with two structures with respect to the degree of the serial dependence in data.

Table 1. Different structures regarding the levels of the dependency and over-dispersion for 2 state PHMM

K_0	Structure	Parameters		
		π	A	λ
2	S1	[0.50 0.50]	$\begin{bmatrix}0.2 & 0.8\\0.8 & 0.2\end{bmatrix}$	[3 9]
	S2	[0.50 0.50]	$\begin{bmatrix}0.2 & 0.8\\0.8 & 0.2\end{bmatrix}$	[3 9]
3	S3	[0.33 0.33 0.34]	$\begin{bmatrix}0.70 & 0.20 & 0.10\\0.20 & 0.60 & 0.20\\0.10 & 0.10 & 0.80\end{bmatrix}$	[3 9 20]
	S4	[0.33 0.33 0.34]	$\begin{bmatrix}0.10 & 0.20 & 0.70\\0.20 & 0.60 & 0.20\\0.80 & 0.10 & 0.10\end{bmatrix}$	[3 9 20]
4	S5	[0.25 0.25 0.25 0.25]	$\begin{bmatrix}0.70 & 0.00 & 0.10 & 0.20\\0.00 & 0.80 & 0.20 & 0.00\\0.10 & 0.00 & 0.70 & 0.20\\0.10 & 0.10 & 0.00 & 0.80\end{bmatrix}$	[3 9 20 30]
	S6	[0.25 0.25 0.25 0.25]	$\begin{bmatrix}0.10 & 0.10 & 0.40 & 0.40\\0.30 & 0.10 & 0.50 & 0.10\\0.10 & 0.60 & 0.10 & 0.20\\0.40 & 0.10 & 0.20 & 0.30\end{bmatrix}$	[3 9 20 30]

Table 2. Four scenarios on the generating model with 2-state represented by four structures obtained from Table 1.

Structure	Assumptions
S1	low dependence as shown in figure (1), and high over-dispersion as the variance so large than the mean (s^2=15.193 and $\bar{y}$ = 5.776)
S2	strong dependence as shown in figure (1), and high over-dispersion as the variance so large than the mean (s^2=15.193 and $\bar{y}$ = 5.776)
S3	strong dependence as shown in figure (1), and high over-dispersion as the variance so large than the mean (s^2=65.869 and $\bar{y}$=11.288)
S4	strong dependence as shown in figure (1), and high over-dispersion as the variance so large than the mean (s^2=65.869 and $\bar{y}$ =11.288)
S5	strong dependence as shown in figure (1), and high over-dispersion as the variance so large than the mean (s^2=65.869 and $\bar{y}$=11.288)
S6	strong dependence as shown in figure (1), and high over-dispersion as the variance so large than the mean (s^2=65.869 and $\bar{y}$=11.288)

Table 3. The number of times (percentage) that the competition models from K=2 to K=7 components are chosen by the criterion for PHMMs with complexities K_0=2, 3 and 4, each one two structures.

K	K_0=2		K_0=3		K_0=4	
	S1	S2	S3	S4	S5	S6
2	**97.00%**	**93.50%**	0.00%	3.00%	0.00%	0.00%
3	3.00%	7.00%	**88.00%**	**77.00%**	12.00%	22.00%
4	0.00%	0.00%	12.00%	11.00%	**63.00%**	**44.00%**
5	0.00%	0.00%	0.00%	0.00%	**25.00%**	**34.00%**
6	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
7	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 4. Results of model selection for the competing models fitted to earthquake data.

k	2	3	4	5	6	7
ilppd	-294.945	-294.736	-294.356	-293.736	-293.242	-293.051
$p_{WAICvar}$	24.091	24.187	25.771	27.973	32.145	34.112
$WAIC_{var}$	638.072	**637.846**	640.254	643.418	650.774	654.326

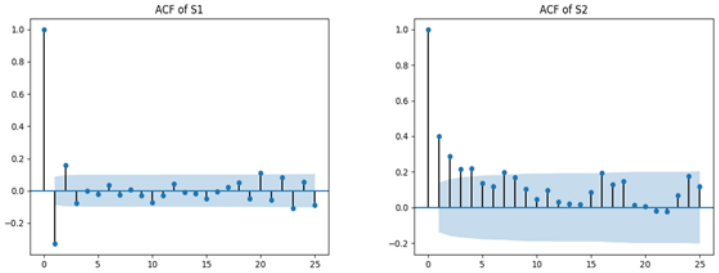

Figure 1. Theoretical ACF of S1 (left) and S2 (right) for data generated from a 2-state PHMM

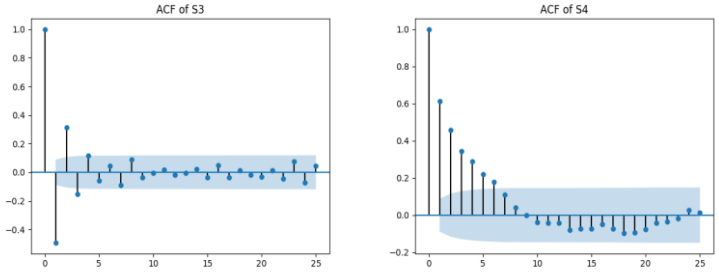

Figure 2. Theoretical ACF of S3 (left) and S4 (right) for data generated from a 3-state PHMM

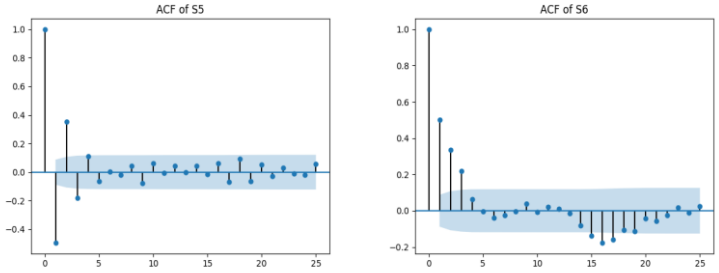

Figure 3. Theoretical ACF of S5 (left) and S6 (right) for data generated from a 4-state PHMM

3.2. Results

Under the assumptions assumed in the study, we report the proposed criterion's performance as shown in Table 3. For PHMM with K_0=2, with the structure S1, it can see that our criterion selects the higher ratio to select the right model (97%), with a slight ratio for over-fitting (3%). For the same model, but with the structure S2 in which exist a strong serial dependence, the correct model is also selected with high ratio (93.5%) which it suggests that either existence or not of the serial dependence in data has no effect. For the model with complexity K_0=3, the criterion also performs well with respect to the structure S3 but with ratio 88% and over-fitting is 12%. On other hand, with same model, the criterion selects the true order with ratio 76% and appears an over-fitting is 11% and slight under-fitting which is 3%. When the complexity of model K_0=4 increases, the criterion tends to choose the correct model with less ratio is 63% in structure S5, comprising with other models, with over-fitting and under-fitting are 25% and 12%, respectively, while the ratio of selecting the true model decreases to 44% under structure S6 with high quantity of over-fitting with is 34 and under-fitting is 22%. Overall, the criterion has high efficiency with models that have a lower order, either exists or not the serial dependency in data, but it appears sensitive with respect to exist the correlation in data when the complexity of the model increases.

3.3. Real data application

We consider here a real data application involving count data of earthquake data to evaluate the proposed criterion. These data consist of a series of length 107 of major earthquakes (magnitude 7 or greater) which occurred in the world between 1900 and 2006. The average and variance of the data are: $\bar{y}$ = 19:364 and s^2 = 51:091, respectively, which suggests that these data have the over-dispersion problem. In addition, the ACF appears a high serial correlation in the data as shown in Figure (5). Thus, the PHMM can be applied here to accommodate the over-dispersion and serial correlation to this kind of data. Zucchini and MacDonald (2009) fitted several Poisson hidden Markov models to these data with a different number of states using classical estimates-based AIC and BIC. They concluded that the model with K=3 is the best to adequately represent these data. We assume a number of competing models in order to fit to these data, with an upper bound K_{max} = 7. By the same way followed in simulation study, we adopted 15,000

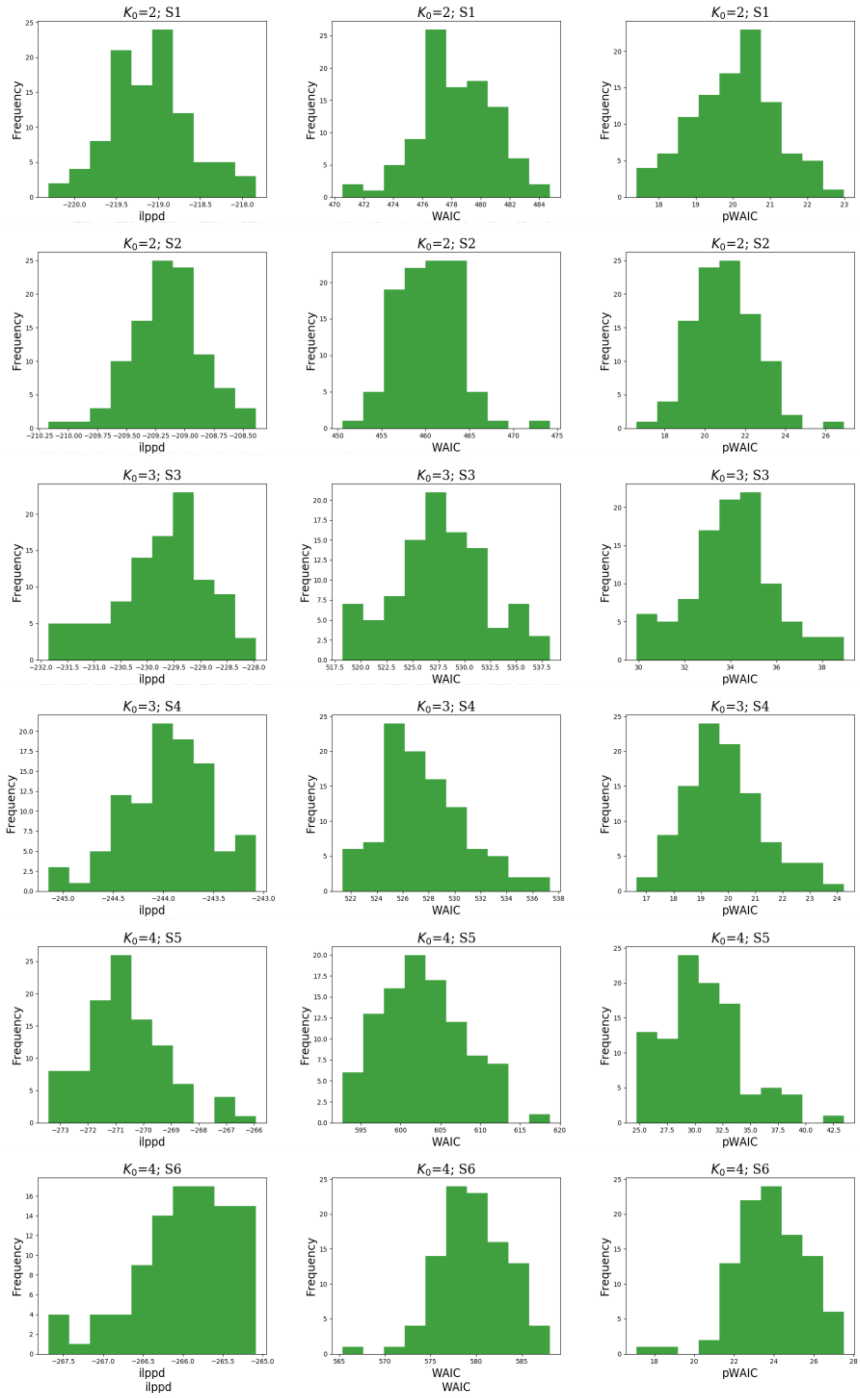

Figure 4. Histograms of the frequencies of values of the ilppd, WAIC and pW AIC for the models K0 = 2, K0 = 3 and K0 = 4 with scenarios S1-S6

Table 5. Parameter estimates of a PHMM with 3 state fitted to the earthquakes data.

$\hat{\pi}$	$\hat{A}$	$\hat{\lambda}$
$\begin{pmatrix} 0.3103 \\ 0.4771 \\ 0.2126 \end{pmatrix}$	$\begin{pmatrix} 0.8114 & 0.1302 & 0.0584 \\ 0.0866 & 0.8203 & 0.0931 \\ 0.0654 & 0.2154 & 0.7192 \end{pmatrix}$	$\begin{pmatrix} 12.7129 \\ 19.3490 \\ 29.2053 \end{pmatrix}$

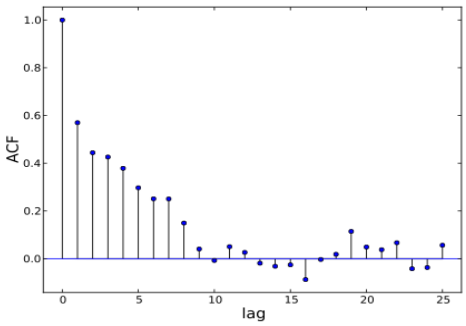

Figure 5. Sample ACF for the earthquake count data

samples for inference after burning the first 5,000 samples. We put artificially identifiable constraints on the mean parameter, λ, to avert the issue of identifiability, i.e.: $\lambda_1 < \lambda_2 < ... < \lambda_k$. From table (4), it can see that the WAIC chooses the model with K=3. The results of selection the best model fitted to those data are also supported by the fitting results in Figure (6). It can be noted from the Figure 6 that the competition model with K=3 shows an appropriate goodness of fit and there are not more states will be needed to the model. Table 5 shows the result of model estimation for the selected model which are very close to those estimated by Zucchini and MacDonald (2009).

4. Conclusions

We introduced more recent model selection criterion, named the widely applicable information criterion (WAIC), for hidden Markov models. We examined our criterion via simulated databases, given scenarios that included the existence the serial dependency and over-dispersion in data, as well as data real application. It was found that the new methodology performs well in simulation for HMMs with less complicated models. Overall, the criterion has high efficiency with models that have a lower order, either the serial dependency in data exists or not. The criterion had sensitivity for existing correlation in the data when the complexity of the model increases. We expect that the model work well with mixture models where the dependency assumption is not included in those models and this, thus, could be a future study of interesting. In addition, we propose to use new Bayesian methods to estimate the model such as Hamiltonian Monte Carlo method. The latter method may give more reliable estimations for the model parameters. It can be investigated in future studies.

Acknowledgements

The authors thank the editor and reviewers for their notes and constructive suggestions that contributed to the improvement of this article.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov & F. Csaki (Eds.), *Proceedings of the Second International Symposium on Information* (pp. 267–281), Budapest.

Ando, T. (2010). *Bayesian model selection and statistical modeling*. Boca Raton, FL: Chapman and Hall.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Berlin, Germany: Springer.

Celeux, G., Forbes, F., Robert, C. P., & Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, *1*(4), 651–673.

Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics*, *75*(1), 79–97.

Fan, Y., & Sisson, S. A. (2011). Reversible jump MCMC. In S. Brooks, A. Gelman, G. Jones, & X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo, Chapter 3* (pp. 67–91). Boca Raton, FL: CRC press.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. New York, NY: Springer.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criterion for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*(6), 721–741.

Han, C., & Carlin, B. P. (2001). Markov chain Monte Carlo methods for computing Bayes factors: A comparative review. *Journal of the American Statistical Association*, *96*(455), 1122-1132.

Johnson, M. (2007). Why doesn't EM find good HMM POS-taggers?. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (pp. 296–305). Stroudsburg, PA: Association for Computational Linguistics.

Kadhem, S. K., Hewson, P., & Kaimi, I. (2018). Using hidden Markov models to model spatial dependence in a network. *Australian and New Zealand Journal of Statistics*, *60*(4), 423–446.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceeding of the IEEE*, *77*(2), 257–286.

Richardson, S., & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *59*(4), 731–792.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461-464.

Scott, S. L. (2002). Bayesian methods for Hidden Markov models: Recursive computing in the 21th century. *Journal of the American Statistical Association*, *97*(457), 337–351.

Spiegelhalter, D., Best, N., Carlin, B., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639.

Tanner, M. Y., & Wong, H. (1987). The Calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*(398), 528–540.

Vehtari, A., & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, *6*, 142–228.

Watanabe, S. (2009). *Algebraic geometry and statistical learning theory*. Cambridge, MA: Cambridge University Press.

Zucchini, W., & MacDonald, I. L. (2009). *Hidden Markov models for time series: An introduction using R.* London, England: Chapman and Hall.

Appendix

A. Computation of the WAIC, ilppd and p_{WAIC} for Normal HMM

Consider a normal HMM with k of components, the ilppd, p_{WAIC} and WAIC are approximated as:

$$\widehat{\text{ilppd}}_y = \sum_{i=1}^{n} \log \mathrm{E}_{\{\mu,\sigma^2,z\}} \{\phi\,(y_i|\mu,\sigma^2,\mathbf{z})|\mathbf{y}\},$$

$$\approx \sum_{i=1}^{n} \log\left(\frac{1}{M}\sum_{m=1}^{M} \phi(y_i\,|\mu_{z_i^{(m)}}^{(m)}, \sigma_{z_i^{(m)}}^{2\,(m)})\right); \tag{22}$$

Where $\phi(y_i\,|\mu,\sigma^2,z)|\boldsymbol{y}$ is the component-specific density of normal HMM. The penalty term of WAIC is given by:

$$p_{WAIC_{var}} = \sum_{i=1}^{n} V_{\{\mu,\sigma^2,z\}}\,[\log \phi(y_i\,|\mu,\sigma^2,z)],$$

$$\approx \sum_{i=1}^{n} V_{m=1}^{M} \log \phi\left(y_i \middle| \mu_{z_i^{(m)}}^{(m)}, \sigma^{2\,(m)}_{z_i^{(m)}}\right), \tag{23}$$

Where $V_{m=1}^{M}$ denotes the samples variance through full MCMC run. Finally, the WAIC then is approximated as follows:

$$WAIC = -2\{\textstyle\sum_{i=1}^{n} \log E_{z,\theta}\,[\phi\;(y_i|z,\mu,\sigma^2)|\mathbf{y}]\} + 2\,p_{WAIC_{var}},$$

$$\approx -2\left\{\sum_{i=1}^{n} \log\left(\frac{1}{M}\sum_{m=1}^{M} \phi\;(y_i \middle|\; \mu_{z_i^{(m)}}^{(m)}, \sigma^{2\,(m)}_{z_i^{(m)}})\right)\right\} + 2p_{WAIC_{var}}\;, \tag{24}$$