International Journal of Statistics and Probability; Vol. 5, No. 1; 2016
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Recursive Deviance Information Criterion for the Hidden Markov Model

Safaa K. Kadhem[1], Paul Hewson[1] & Irene Kaimi[1]

[1]School of Computing, Electronics and Mathematics, Plymouth University, Plymouth, UK

Correspondence: Safaa K. Kadhem, School of Computing, Electronics and Mathematics (Faculty of Science and Engineering), Plymouth University, Plymouth, UK. E-mail: safaa.kadhem@plymouth.ac.uk

Received: October 8, 2015 Accepted: December 6, 2015 Online Published: December 23, 2015

doi:10.5539/ijsp.v5n1p61 URL: http://dx.doi.org/10.5539/ijsp.v5n1p61

Abstract

In Bayesian model selection, the deviance information criterion (DIC) has become a widely used criterion. It is however not defined for the hidden Markov models (HMMs). In particular, the main challenge of applying the DIC for HMMs is that the observed likelihood function of such models is not available in closed form. A closed form for the observed likelihood function can be obtained either by summing all possible hidden states of the complete likelihood using the so-called the forward recursion, or via integrating out the hidden states in the conditional likelihood. Hence, we propose two versions of the DIC to the model choice problem in HMMs context, namely, the recursive deviance-based DIC and the conditional likelihood-based DIC. In this paper, we compare several normal HMMs after they are estimated by Bayesian MCMC method. We conduct a simulation study based on synthetic data generated under two assumptions, namely diversity in the heterogeneity level and also the number of states. We show that the recursive deviance-based DIC performs well in selecting the correct model compared with the conditional likelihood-based DIC that prefers the more complicated models. A real application involving the waiting time of Old Faithful Geyser data was also used to check those criteria. All the simulations were conducted in Python v.2.7.10, available from first author on request.

Keywords: Hidden Markov Models, DIC, recursive likelihood, conditional likelihood, Markov Chain Monte Carlo, the waiting time of Old Faithful geyser data

1. Introduction

Hidden Markov Models (HMMs) have been used to model various types of data: discrete, continuous, univariate, multivariate, mixed and mixture data (MacDonald and Zucchini, 2009). Consequently, they have been widely applied in many fields, such as econometrics (Hamilton, 1989; Billio et al., 1999); finance (Bhar and Hamori, 2004); speech recognition (Rabiner, 1989; Derrode, 2006); and psychology (Visser et al., 2002).

One important issue that is often discussed after fitting models is the model choice issue. In particular, we discuss this issue in the HMM context. When fitting several HMMs to a data set, we seek to determine the number of hidden states of a model that adequately fits those data or more formally the best model among those competitive models. Frequentest criteria such Akaike Information Criterion (AIC) (Akaike, 1973) and the Bayesian information criterion (BIC) (Schwarz, 1978) have been applied to the HMM model choice problem. In the HMM context, they have been used to determine the best HMM by MacDonald and Zucchini (2009). However, the AIC and BIC do not take into account uncertainty about the parameter values and model form. In addition, these criteria might suffer from irregular behavior of the likelihood function that may lead to under-fitting or over-fitting, where the number of hidden states that are analyzed is smaller or greater than the true number of states, respectively (MacDonald and Zucchini, 2009). Alternatively, the Bayesian theory that takes into account the uncertainty about the parameters of the model and also the model selection is used. The deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002) can be viewed as an extended Bayesian version of the AIC the BIC. Similar to the AIC and BIC, the DIC trades off a measure of model adequacy against a measure of complexity. The popularity of the criterion is due to the ease of computing the posterior distribution of the log-likelihood or the deviance. Hence, it has been applied to a wide range of statistical models. Despite its popularity, a number of difficulties can be inherent in this criterion, especially in more complicated models, e.g. mixture and hidden Markov models, where latent (hidden) variables and model parameters are non-identifiable from data (Celeux et al., 2006). In this setting, Celeux et al. (2006) introduced eight formulae for DICs which have been classified according to the nature of the likelihood used; observed, complete and conditional. Celeux et al. (2006) pointed out also that the observed data likelihood of mixture models is sometimes available in closed form, but for the HMMs this is not always the case. Hence, the availability of a closed form for the likelihood function of the HMMs forms a main challenge. The Data Augmentation approach (Tanner and Wong, 1987) is often used with MCMC methods in the Bayesian inference of complicated models such as HMMs, hence, a closed form of the likelihood function of the HMMs can be available and the parameter estimation becomes more

straightforward. Using the Data Augmentation principle, we can obtain a closed form to the likelihood function via two sources exclusively, either via the complete data likelihood or the conditional likelihood. The first source is obtained via summing all possible hidden states of the complete data likelihood. The computation of the observed likelihood function via the complete data likelihood is problematic because of the computational complexity caused by the high dimensions to this function. To overcome this problem, an efficient method called the forward-backward (F-B) (Rabiner, 1989) algorithm that is used for estimating the sequence of underlying hidden states of HMMs and evaluating the likelihood function (Scott, 2002) can be introduced. The observed likelihood function is recursively evaluated using only the forward part of the F-B algorithm to obtain the so-called recursive likelihood or log likelihood. Hence, we propose a DIC based on recursive deviance, which we denote as $\mathrm{DIC}_{\mathrm{rec}}$. The advantage of recursive calculation is that the computational complexity will be reduced from $O(K^T T)$ into $O(K^2 T)$, where T and K are the length of observations and number of hidden states respectively. The second way to obtain a closed form of the likelihood function is to integrate out the hidden states. Hence, another DIC based on a conditional deviance denoted by $\mathrm{DIC}_{\mathrm{con}}$ is proposed.

We contribute to this line of research by developing a new methodology aiming at finding closed forms for the observed likelihood of HMMs. A closed form of the observed likelihood can be obtained either via the complete-data likelihood or conditional. Hence, two of proposed criteria in this paper are introduced. The first criterion, called the recursive deviance-based DIC, is obtained via the evaluation of the observed likelihood by summing all hidden states of the complete likelihood using the forward recursion approach. The second, called the conditional deviance-based DIC, can be obtained via integrating the conditional likelihood. Since HMMs are appropriately described to accommodate the unobserved heterogeneity in data, we also contribute in this paper by taking into account the behaviour of the proposed criteria with different levels of heterogeneity in the data. We also examine these criteria on models with a different number of states to see whether each criterion has a fixed behaviour or not for different number of states. We show via simulation studies that the conditional likelihood-based DIC tends to favour more complicated models, and its estimates are generally unstable as documented from their large numerical standard errors. On the other hand, the DIC based on the recursive likelihoods is more accurately estimated as it has much smaller numerical standard errors compared to the conditional likelihood-based DIC. In addition, it performs well with respect to choosing the correct model.

The rest of this paper is organized as follows. In section 2, the HMMs are introduced. The Bayesian estimation with data augmentation are also reviewed. In section 3, we develop a recursive calculation for the likelihood function. In section 4, we review the general definition of the DIC, propose criteria for hidden Markov models, namely the recursive deviance-based DIC and the conditional likelihood-based DIC and discuss how to compute those criteria from the MCMC output. Section 5 is devoted for fitting the model and also describing the synthetic and real data used in the study. Section 6 shows the results of comparing of the proposed criteria using artificial and real data. The discussions of this paper and future research are introduced in Section 7.

2. Hidden Markov Models

2.1 The Model

The Hidden Markov Model (HMM) is a statistical model that involves two stochastic processes. The first process ($\mathrm{Z}_t = \mathrm{z}_t; t = 1, 2, ..., T$) is an unobserved or hidden process (state process), satisfying the Markov property. The second process ($\mathrm{Y}_t = \mathrm{y}_t, t = 1, 2, ..., T$) is an observed process (state-dependent process). When Z_t is known, the distribution of Y_t can be determined based only on Z_t (MacDonald and Zucchini, 2009). The dependency between Markovian hidden states and observed state can be illustrated in the directed graph in Figure 1. From Figure 1, we can mathematically summarize the relationship between those two processes under the following assumptions:

1- Markov property: $p(\mathrm{Z}_t = \mathrm{z}_t|\mathrm{Z}_{t-1} = \mathrm{z}_{t-1}, \mathrm{Z}_{t-2} = \mathrm{z}_{t-2}, ..., \mathrm{Z}_1 = \mathrm{z}_1) = p(\mathrm{Z}_t = \mathrm{z}_t|\mathrm{Z}_{t-1} = \mathrm{z}_{t-1})$.

2- Conditional independence: $p(\mathrm{Y}_t = \mathrm{y}_t|\mathrm{Y}_{t-1} = \mathrm{y}_{t-1}, \mathrm{Y}_{t-2} = \mathrm{y}_{t-2}, ..., \mathrm{Y}_1 = \mathrm{y}_1, \mathrm{Z}_t = \mathrm{z}_t) = p(\mathrm{Y}_t = \mathrm{y}_t|\mathrm{Z}_t = \mathrm{z}_t)$.

We refer the reader to (Cappe et al., 2006) and (MacDonald and Zucchini, 2009) for good overviews of HMMs.

In this paper, we focus on parametric discrete-time finite state-space HMMs, such that the observations in the assumption (2) can follow distributions from a parametric family, and each hidden state z at time t in assumption (1) takes a discrete value k and it is defined on a known finite state-space, $\Omega_Z \in K$, where K is a known number of the states. Formally, the HMMs can be expressed as a collection of the parameters $\Theta = (\pi, \mathbf{A}, \theta)$ and the number of hidden states K, that are defined in detail as follows:

1. The number of states K, where each hidden state z_t take a discrete value k that belong to the state space $\{1, 2, ..., K\}$.
2. The initial state distribution $\pi = \{\pi_k\}$, where π_k denotes the probability that the model is in the state k at the time $t = 1$, i.e, $\pi_k = p\{\mathrm{z}_1 = k\}$, where $1 \leq k \leq K$.
3. The probability transition matrix $\mathbf{A} = \left\{a_{jk}\right\}$, where a_{jk} is the probability that the state at time t is k, given that the

state at time $t-1$ is j, i.e, $a_{jk} = p\{z_t = k|z_{t-1} = j\}$ such that a_{jk} satisfy the normal stochastic constraints, i.e. $a_{jk} \geq 0$ and $\sum_{k=1}^{K} a_{jk} = 1$, where $1 \leq j,k \leq K$.

4. The state-depended distribution θ, where θ refers to that observations, given a hidden state, have been generated from some parametric HMM. Consider $\mathbf{y} = (y_1, y_2, ..., y_T)$ is an observation sequence that have the probability density functions $f_{\mathbf{y}}(.;\theta), \theta \in \Theta$. The density $f_{\mathbf{y}}(.;\theta)$ follows either a normal distribution with parameters $\theta = \{\theta_k\} = (\mu_k, \sigma_k^2)$, or a poisson distribution with parameter $\theta = \{\theta_k\} = \lambda_k$; $\lambda_k > 0$ where $k = 1, 2, ..., K$.

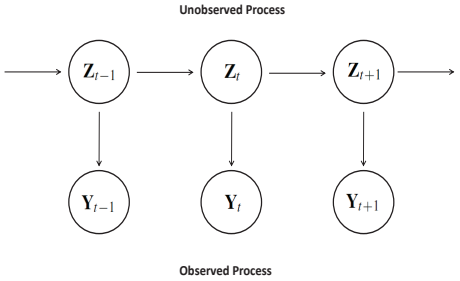

Figure 1. Dependence structure of a HMM

2.2 Bayesian Estimation

The Bayesian model can be written using Bayes Theorem as $p(\Theta|\mathbf{y}) \propto f(\mathbf{y}|\Theta)p(\Theta)$, where $p(\Theta|\mathbf{y})$ denotes the posterior distribution of the model, $f(\mathbf{y}|\Theta)$ and $p(\Theta)$ denote respectively the evidence or the observed data likelihood and prior. The difficulty of applying Bayesian inference for models such as HMMs is attributed to the complexity of evaluating the model evidence or the likelihood $f(\mathbf{y}|\Theta)$. To overcome this problem, techniques such as Monte Carlo Markov Chain (MCMC) are used. The Data Augmentation (Tanner and Wong, 1987) approach is often used to facilitate the estimates process of the parameters of the model. This strategy is commonly used with MCMC methods in Bayesian analysis of HMMs in which the hidden states are introduced as missing data and augmented to the parameter space of the sampler (Robert et al., 1993; Albert and Chib, 1993; Chib, 1996). The posterior distribution can then be written as

$$p(\Theta, \mathbf{z}|\mathbf{y}) \propto f(\mathbf{y}, \mathbf{z}|\Theta)p(\Theta) \propto f(\mathbf{y}|\mathbf{z}, \Theta)p(\mathbf{z}|\Theta)p(\Theta). \tag{1}$$

where the term, $f(\mathbf{y}, \mathbf{z}|\Theta)$, represents the complete data likelihood which can be written according to the Bayes' rules as $f(\mathbf{y}|\mathbf{z}, \Theta)p(\mathbf{z}|\Theta)$ and $p(\Theta)$ represents a prior distribution on Θ. In other words, the complete data likelihood function above can be obtained as follows: Given an initial probability distribution π and transition matrix $\mathbf{A}$, the hidden state sequence $\mathbf{z} = (z_1, z_2, ..., z_T)$, with a known number of states K, can be modelled as $p(\mathbf{z}) = p(z_t, z_{t-1}, ..., z_1; \mathbf{A}, \pi)$. According to Bayes' rules based on the Markov property, we obtain

$$p(\mathbf{z}) = p(z_t|z_{t-1}; \mathbf{A})p(z_{t-1}|z_{t-2}; \mathbf{A}), ...p(z_2|z_1; \mathbf{A})p(z_1|z_0; \pi) = p(z_1|\pi)\prod_{t=2}^{T} p(z_t|z_{t-1}; \mathbf{A}).$$

By conditioning on z_t, the observation y_t can be sampled independently from a chosen parametric distribution,

$$f(y_t|z_t = k, \theta) = \prod_{t=1}^{T} f_k(y_t|\theta_k),$$

where $f_k(.|\theta_k)$ denotes a parametric density function parametrized by θ at the state k. Hence, the complete data likelihood of HMM is then given by

$$f(\mathbf{y}, \mathbf{z}|\pi, \mathbf{A}, \theta) = p(z_1|\pi)\prod_{t=2}^{T} p(z_t|z_{t-1}, \mathbf{A})\prod_{t=1}^{T} f_k(y_t|\theta_k),$$

and the observed data likelihood can be obtained by summing the complete data likelihood above over all possible hidden states

$$f(\mathbf{y}|\Theta) = \sum_{\forall \mathbf{z}}\left[p(z_1|\pi)\prod_{t=2}^{T} p(z_t|z_{t-1}, \mathbf{A})\prod_{t=1}^{T} f_k(y_t|\theta_k)\right]. \tag{2}$$

The calculation of the likelihood for a HMM using traditional methods can be computationally expensive, since it it involves a total of $O(K^T T)$ calculations (Rabiner, 1989; Bishop, 2006). Therefore, it requires more sophisticated methods.

An efficient technique called the forward recursion can be a less expensive way to compute the likelihood function of HMMs (Rabiner, 1989; Chib, 1996; Scott, 2002). We devoted an independent section in this paper to discuss the computation of the likelihood function of the HMMs due to it forms the main part for constructing of the proposed model selection criteria, see Section 3. To complete the Bayesian HMM, we need to specify prior distributions on the parameters $\pi, \mathbf{A}$, and θ. We assume independent Dirichlet priors (Fruhwirth-Schnatter, 2006) on the initial distribution π and each row of the transition matrix $\mathbf{A}$, $\left\{a_{j.}\right\}$, i.e.,

$$\begin{aligned} p(\pi) &= \prod_{k=1}^{K} \pi_k \propto \prod_{k=1}^{K} \pi_k^{\delta_k - 1} = Dir(\delta_1, \delta_2, ..., \delta_K), \\ p(\mathbf{A}) &= \prod_{j=1}^{K} a_{j.} \propto \prod_{j=1}^{K} a_{j.}^{\delta_j - 1} = Dir(\delta_1, \delta_2, ..., \delta_K) \ \ , j, k = 1, 2, \ ..., K, \end{aligned} \tag{3}$$

where δ is a hyperparameter of Dirichlet distribution. Regarding the state-dependent parameter θ, we choose priors on θ, expressed by $p(\theta|\varphi)$, where φ denotes a conjugate hyperparameter. Their priors have the same functional form as the likelihood, and hence the posterior. The complete data posterior of the HMM in equation (1) can then be written as follows

$$\begin{aligned} p(\Theta, \mathbf{z}|\mathbf{y}) &= f(\mathbf{y}, \mathbf{z}|\pi, \mathbf{A}, \theta) p(\mathbf{z}|\pi, \mathbf{A}) p(\theta|\varphi) \\ &= \sum_{\forall \mathbf{z}} \left[p(\mathrm{z}_1|\pi) \prod_{t=2}^{T} p(\mathrm{z}_t|\mathrm{z}_{t-1}, \mathbf{A}) \prod_{t=1}^{T} f_k(\mathrm{y}_t|\theta_k) \right] \times p(\theta|\varphi) Dir(\pi|\delta) \prod_{j=1}^{K} Dir(a_{j.}|\delta). \end{aligned} \tag{4}$$

Analytical computation of the posterior distribution in equation (4) is difficult due to its complex form. A way to address this complexity is to use MCMC algorithms. The MCMC algorithm involves sampling the parameters of the model, Θ, and the hidden state, $\mathbf{z}$, from the desired distribution based on an ergodic Markov chain. We use the Gibbs algorithm (Geman and Geman, 1984), a special case of Metropolis-Hastings algorithm, to treat the high-dimensionality problem in the joint posterior distribution in the equation (4) by partitioning it into blocks (conditional distributions), to make sampling simpler. A two-stage Gibbs sampler is then implemented by alternating between drawing $\mathbf{z}$ from the conditional posterior distribution $p(\mathbf{z}|\Theta, \mathbf{y})$ (data augmentation) and drawing Θ from the conditional posterior distribution $p(\Theta|\mathbf{z}, \mathbf{y})$. A significant advantage of using the Gibbs sampling is that the conditional distributions required are often available for simulation, and hence, the high-dimensionality problem is typically reduced (Casella and Robert, 2004).

3. A Recursive Evaluation of Likelihood Function

The likelihood function forms the main part of many Bayesian model choice criteria. However, with increasing model complexity, Bayesian inference and model selection become more challenging. The reason may be that the likelihood function is analytically unavailable or computationally costly to evaluate. Because of the complicated nature of hidden Markov models, the likelihood function of these models is often unavailable in closed form (Celeux et al., 2006). To overcome this problem, the data augmentation strategy (Tanner and Wong, 1987) is often used. The main idea behind this approach is augmenting the observations by adding hidden states $\mathbf{z}$ introduced as missing data, which leads to obtaining a closed form expression to likelihood function and hence simplifying the MCMC algorithms constructed for computing posterior distributions. As mentioned in Section 2, a closed form of the observed likelihood function in equation (2) is obtained by summing all possible (missing) states $\mathbf{z}$ over the complete data likelihood. However, evaluating of the observed likelihood function over the complete data likelihood is still not a simple task. For this, we rely on an efficient approach called the forward-backward algorithm (Baum et al., 1970; Rabiner, 1989) to achieve this task as shown in the next sub-section.

3.1 Forward-Backward Algorithm

The forward-backward recursion was developed by Baum et al. (1970) to implement an EM algorithm to obtain a maximum likelihood estimate. The observed likelihood function defined in equation (2) in Section 2 is evaluated by summing the complete data likelihood function over all possible hidden states. The evaluation of the observed likelihood function using direct ways requires $O(TK^T)$ steps. Alternatively, we can compute the observed likelihood function recursively using only the forward part of the forward-backward algorithm to obtain the so-called recursive likelihood. This approach requires only $O(TK^2)$ steps (Rabiner, 1989). The recursive computation of the likelihood function is based on defining the forward probabilities which provide the probability of ending up in any particular state, given a partial observation sequence, i.e $p(\mathrm{y}_1, \mathrm{y}_2, ..., \mathrm{y}_t, \mathrm{z}_t|\Theta)$. We define the forward variable $\alpha_t(k)$, that represents the joint probability of the partial observation sequence until time t and the state k at time t, by

$$\alpha_t(k) = p(\mathrm{y}_1, \mathrm{y}_2, ..., \mathrm{y}_t, \mathrm{z}_t = k|\Theta); \ t = 1, 2, ..., T \text{ and } k = 1, 2, ..., K.$$

According to Baum et al. (1970) and Rabiner (1989), the forward probabilities, $\alpha t(k)$, can be calculated recursively by first initializing a forward variable at time $t = 1$,

$$\alpha_1(k) = p(\mathrm{y}_1, \mathrm{z}_1 = k|\Theta) = p(\mathrm{z}_1 = k|\pi) f(\mathrm{y}_1|\mathrm{z}_1 = k, \theta) = \pi_k f(\mathrm{y}_1|\theta_k),$$

and for $t = 2, 3, ..., T$, the updating of the next forward variables over induction steps is done using

$$\begin{aligned}
\alpha_t(k) &= p(\mathrm{y}_1, \mathrm{y}_2, ..., \mathrm{y}_t, \mathrm{z}_t = k|\Theta) \\
&= \sum_{j=1}^{K} p(\mathrm{y}_1, \mathrm{y}_2, ..., \mathrm{y}_t, \mathrm{z}_{t-1} = j, \mathrm{z}_t = k|\Theta) \\
&= \sum_{j=1}^{K} p(\mathrm{z}_t = k, \mathrm{y}_t|\mathrm{z}_{t-1} = j, \mathrm{y}_1, ..., \mathrm{y}_{t-1}, \Theta) p(\mathrm{z}_{t-1} = j, \mathrm{y}_1, ..., \mathrm{y}_{t-1}|\Theta) \\
&= \sum_{j=1}^{K} p(\mathrm{z}_t = k, \mathrm{y}_t|\mathrm{z}_{t-1} = j, \Theta) p(\mathrm{z}_{t-1} = j, \mathrm{y}_1, ..., \mathrm{y}_{t-1}|\Theta) \\
&= \sum_{j=1}^{K} p(\mathrm{y}_t|\mathrm{z}_t = k, \mathrm{z}_{t-1} = j, \Theta) p(\mathrm{z}_t = k|\mathrm{z}_{t-1} = j, \Theta) p(\mathrm{z}_{t-1} = j, \mathrm{y}_1, ..., \mathrm{y}_{t-1}|\Theta) \\
&= \sum_{j=1}^{K} \alpha_{t-1}(j) a_{jk} f(\mathrm{y}_t|\theta_k). \qquad (5)
\end{aligned}$$

Hence, a recursion can be set up that will lead to obtaining $\alpha_t(j)$ for $j = 1, 2, ..., K$ and $t = 1, 2, ..., T$. Note that computing $\alpha_t(k)$ requires the sum of K quantities. Thus a single step in the recursion is $O(K^2)$, and the evaluating of observed likelihood function in equation (2) is $O(TK^2)$, which is far more efficient than the $O(TK^T)$ complexity of the likelihood computed in the direct methods. Given the forward probabilities obtained above, the likelihood can be written as

$$L(\Theta|\mathbf{y}) = f(\mathrm{y}_1, \mathrm{y}_2, ..., \mathrm{y}_T|\Theta) = \sum_{j=1}^{K} f(\mathrm{y}_1, \mathrm{y}_2, ..., \mathrm{y}_T, \mathrm{z}_T = j|\Theta) = \sum_{j=1}^{K} \alpha_T(j).$$

Practically, when dealing with a large observation sequence, the calculation of the likelihood function would cause the so-called underflow/overflow problem, that is, the likelihood may converge to zero or diverge to infinity as t increases. To overcome this problem, we use the scaling factors (Rabiner, 1989) to re-scale the $\alpha_t(j)'s$ throughout the recursion by dividing them by $\sum_{j=1}^{K} \alpha_t(j)$.To do this, we define the first scaling coefficient as $c_1 = \dfrac{1}{\sum_{j=1}^{K} \alpha_1(j)}$. Hence, at t=1, the forward variables are re-scaled by multiplying by c_1, that is, $\hat{\alpha}_1(j) = c_1\alpha_1(j)$, for $j = 1, 2, ..., K$, where $\hat{\alpha}$ is used to denote the re-scaled coefficients. Then, applying equation (5) directly to the re-scaled forward variables, we define at $t = 2$,

$$\begin{aligned}
\alpha_2^*(k) &= f(\mathrm{y}_2|\theta_k) \sum_{j=1}^{K} \hat{\alpha}_1(j) a_{jk} \\
&= c_1 f(\mathrm{y}_2|\theta_k) \sum_{j=1}^{K} \alpha_1(j) a_{jk} \\
&= c_1\alpha_2(k).
\end{aligned}$$

Now, let $\hat{\alpha}_2(j) = c_2\alpha_2^*(j)$, where the second scaling coefficient is defined as $c_2 = \dfrac{1}{\sum_{j=1}^{K} \alpha_2^*(j)}$. Then, $\hat{\alpha}_2(j) = c_1c_2\alpha_2(j)$, for $j = 1, 2, ..., K$. For $t > 1$, using the same procedure leads to

$$\begin{aligned}
\alpha_t^*(k) &= f(\mathrm{y}_t|\theta_k) \sum_{j=1}^{K} \hat{\alpha}_{t-1}(j) a_{jk} \\
&= \left(\prod_{r=1}^{t-1} c_r \right) f(\mathrm{y}_t|\theta_k) \sum_{j=1}^{K} \alpha_{t-1}(j) a_{jk} \\
&= \left(\prod_{r=1}^{t-1} c_r \right) \alpha_t(k),
\end{aligned}$$

and allows us to further define

$$\hat{\alpha}_t(k) = \alpha_t^*(k)\frac{1}{\sum_{j=1}^{K}\alpha_t^*(j)} = c_t\alpha_t^*(k) = \left(\prod_{r=1}^{t} c_r\right)\alpha_t(k).$$

As this last expression is valid for $t = T$, it is possible to calculate the observed likelihood function from

$$L(\Theta|\mathbf{y}) = \sum_{j=1}^{K}\alpha_T(j) = \left(\prod_{r=1}^{T} c_r\right)^{-1}\hat{\alpha}_T(j) = \left(\prod_{r=1}^{T} c_r\right)^{-1}, \tag{6}$$

since, $\hat{\alpha}_T(j) = \dfrac{\sum_{j=1}^{K}\alpha_T^*(j)}{\sum_{k=1}^{K}\alpha_T^*(k)} = 1$. The log-likelihood function can be obtained as

$$\ell(\Theta|\mathbf{y}) = \log\left[L(\Theta|\mathbf{y})\right] = -\sum_{t=1}^{T}\log c_t, \tag{7}$$

which depends only on the scaling constants. Since the log-likelihood is computed as a sum of the log scaling coefficients, this will avoid the underflow/overflow problems (Rabiner, 1989). In general, the recursive log-likelihood function computed here forms the main part of the deviance D(Θ), where D(Θ)=-2[log-likelihood], used in the constructing of most likelihood-based criteria. In the next section, we will show how to construct criteria based on the recursive likelihood function.

4. Deviance Information Criterion for HMMs

The deviance information criterion (DIC) was introduced by Spiegelhalter et al. (2002) from a Bayesian perspective. It is used to measure both the goodness of fit of the model and the model complexity. Spiegelhalter et al. (2002) developed this criterion by introducing the theoretical justification to the concept of effective number of parameters as a measure of the complexity of a model. This criterion is based on the concept of the *deviance*, which is defined as $\mathrm{D}(\Theta) = -2\log f(\mathbf{y}|\Theta) + 2\log f(\mathbf{y})$, where $f(\mathbf{y}|\Theta)$ is the likelihood function and $f(\mathbf{y})$ is a function of the data alone which is often set to unity. Thus, the deviance is

$$\mathrm{D}(\Theta) = -2\log f(\mathbf{y}|\Theta) + 2\log(1) = -2\log f(\mathbf{y}|\Theta).$$

The DIC as defined by Spiegelhalter (2002) is then

$$\mathrm{DIC} = \overline{\mathrm{D}(\Theta)} + p_{\mathrm{DIC}},$$

where, $\overline{\mathrm{D}(\Theta)}$, is used as a measure of the goodness of fit and is summarized by the posterior expectation of the deviance,

$$\overline{\mathrm{D}(\Theta)} = \mathrm{E}_{\Theta|\mathbf{y}}\left\{\mathrm{D}(\Theta)\right\} = \mathrm{E}_{\Theta|\mathbf{y}}\left\{-2\log f(\mathbf{y}|\Theta)\right\},$$

and p_{DIC} is the effective number of parameters, is used as a measure for model complexity, which is defined as the difference between the posterior mean of the deviance minus the deviance of posterior means, i.e.

$$p_{\mathrm{DIC}} = \mathrm{E}_{\Theta|\mathbf{y}}\left\{\mathrm{D}(\Theta)\right\} - \mathrm{D}\left\{\mathrm{E}_{\Theta|\mathbf{y}}(\Theta)\right\} = \overline{\mathrm{D}(\Theta)} - \mathrm{D}(\tilde{\Theta}),$$

where $\tilde{\Theta}$ is an estimate of Θ, which is often taken as the posterior mean or mode (Spiegelhalter, 2002; Celeux et al., 2006). Therefore, the DIC is then defined by

$$\begin{aligned}\mathrm{DIC} = \overline{D(\Theta)} + p_{\mathrm{D}} &= \overline{D(\Theta)} + (\overline{D(\Theta)} - \mathrm{D}(\tilde{\Theta})) \\ &= 2\overline{\mathrm{D}(\Theta)} - \mathrm{D}(\tilde{\Theta}) \\ &= 2\left[\mathrm{E}_{\Theta|\mathbf{y}}\left\{-2\log f(\mathbf{y}|\Theta)\right\}\right] - \left[-2\log f(\mathbf{y}|\tilde{\Theta})\right] \\ &= -4\mathrm{E}_{\Theta|\mathbf{y}}\left[\log f(\mathbf{y}|\Theta)\right] + 2\log f(\mathbf{y}|\tilde{\Theta}),\end{aligned} \tag{8}$$

and its the effective number of parameters is

$$p_{\mathrm{DIC}} = -2\mathrm{E}_{\Theta|\mathbf{y}}\left[\log f(\mathbf{y}|\Theta)\right] + 2\log f(\mathbf{y}|\tilde{\Theta}).$$

Given a set of competing models, smaller DIC values indicate a better-fitting model. The DIC criterion is based essentially on the likelihood function of the model since it forms the main part of the deviance used in constructing this criterion. In the context of latent variable models, the likelihood function requires some care because they are often unavailable in closed form. The likelihood function of mixture models is sometimes available in closed form but for HMMs this is not always available (Celeux et al., 2006). In such situation, Celeux et al. (2006) introduced eight formulas to the DIC based on the type of likelihood function of the latent variable model. There are a number of methods by which the likelihood can be defined. Consider a model describing a data set $\mathbf{y}$ with latent variables $\mathbf{z}$ and parameters Θ. The likelihood can take one of the following forms:

1. $f(\mathbf{y}|\Theta)$ is the observed or incomplete likelihood which is computed by marginalizing the complete likelihood shown in the point (2) below.

2. $f(\mathbf{y}, \mathbf{z}|\Theta)$ is the complete likelihood or the joint probability distribution of observations $\mathbf{y}$ and the latent or hidden variables $\mathbf{z}$.

3. $f(\mathbf{y}|\mathbf{z}, \Theta)$ is the conditional likelihood of observation $\mathbf{y}$, given the latent or hidden variable $\mathbf{z}$ and the parameters of the model Θ.

The observed data likelihood $f(\mathrm{y}|\Theta)$ is sometimes called the incomplete data likelihood due the fact that the observed data likelihood cannot be evaluated without involving latent or missing variables. Hence, the analysis is based instead on the so-called complete-data likelihood $f(\mathbf{y}, \mathbf{z}|\Theta)$. In discrete time HMMs with a finite state space, that we adopt in this paper, the incomplete or observed data likelihood can be written as

$$f(\mathbf{y}|\Theta) = \sum_{\mathbf{z}} f(\mathbf{y}|\mathbf{z}, \Theta) f(\mathbf{z}|\Theta) = \sum_{\mathbf{z}} f(\mathbf{y}, \mathbf{z}|\Theta), \tag{9}$$

where, $f(\mathbf{y}|\mathbf{z}, \Theta)$, is the conditional likelihood multiplied by the probability of hidden states $f(\mathbf{z}|\Theta)$, given the parameters of the model, and $f(\mathbf{y}, \mathbf{z}|\Theta)$, is the complete-data likelihood. From the expression above, one can conclude that in order to obtain the observed likelihood function in closed-form, which forms a hard task for HMMs, one can either use the conditional likelihood function $f(\mathbf{y}|\mathbf{z}, \Theta)$, weighted by the density of hidden states $f(\mathbf{z}|\Theta)$, or the complete data function $f(\mathbf{y}, \mathbf{z}|\Theta)$. The approach applied in the expression above (9) is based on the so-called data augmentation strategy (Tanner and Wong, 1987). This strategy is commonly used with MCMC methods in Bayesian analysis of HMMs in which the hidden states are introduced as missing data and added to the parameter space in order to facilitate the estimation process of the parameters of the model and hence implicitly obtain a closed form for the likelihood function (Cappe et al., 2006; Fruhwirth-Schnatter, 2006).

Based on equation (9), we propose two different criteria of the DIC for the HMMs. The first criterion, denoted by $\mathrm{DIC}_{\mathrm{rec}}(\Theta)$, is based on a recursive calculation of the log-likelihood or so-called the *recursive deviance*, see Section 3. To estimate $\mathrm{DIC}_{\mathrm{rec}}(\Theta)$ over MCMC samples, we can obtain a value for the $\ell_{\mathrm{rec}}(\Theta)$ at each iterative step in an MCMC run, of course after evaluate it recursively using the forward algorithm. Hence, a posterior distribution to the $\ell_{\mathrm{rec}}(\Theta)$, this is, $(\ell^{(1)}_{\mathrm{rec}}(\Theta), \ell^{(2)}_{\mathrm{rec}}(\Theta), ..., \ell^{(M)}_{\mathrm{rec}}(\Theta))$, where M is the number of iterations used in an MCMC run, can be obtained along with the posterior distributions of the other parameters of the model. Therefore, the posterior distribution of the $\ell^{(m)}_{\mathrm{rec}}(\Theta)$, $m = 1, 2, ..., M$, can be summarized for obtaining Bayes point estimates, for example, a posterior mean and even a MAP estimate. Algorithm 1 illustrates the computation of the posterior distribution of the recursive log likelihood, adopted from Rabiner (1989), over an MCMC method.

Algorithm 1 : The posterior distribution of recursive log likelihood

Given a sample of parameters $\Theta^{(m)} = (\pi^{(m)}, \mathbf{A}^{(m)}, \theta^{(m)})$ at the m^{th} iteration in an MCMC run:

1. Calculate the scaled forward variables, $\alpha_t(k)$; $t = 1, 2, ..., T$, $k = 1, 2, ..., K$, from Sub-section 3.1.

2. Compute the recursive likelihood function, $L_{\mathrm{rec}}(\Theta|\mathbf{y})$ at the m^{th} iteration, via the scaling factors c_t, $t = 1, 2, ..., T$ as follows:

$$L^{(m)}(\Theta|\mathbf{y}) = \left[\left(\prod_{t=1}^{T} c_t\right)^{-1}\right]^{(m)}.$$

3. Compute the recursive log-likelihood $\ell_{\mathrm{rec}}(\Theta|\mathbf{y})$ at the m^{th} iteration as: $\ell^{(m)}_{\mathrm{rec}}(\Theta|\mathbf{y}) = \log\left[L^{(m)}(\Theta|\mathbf{y})\right] = \left[-\sum_{t=1}^{T} \log c_t\right]^{(m)}$.

Given a posterior distribution of the recursive log-likelihood, $\ell^{(m)}_{\mathrm{rec}}(\Theta)$, $m = 1, 2, ..., M$, obtained using Algorithm 1, we can obtain Bayes point estimates from this posterior which will be used later in constructing the proposed $\mathrm{DIC}_{\mathrm{rec}}(\Theta)$. Note that the criterion proposed in this work, $\mathrm{DIC}_{\mathrm{rec}}(\Theta)$, which is based on a complete likelihood, is inspired from the criterion

DIC_5 based on the complete-likelihood proposed by Celeux et al. (2006). From the DIC's formula in equation (8), we propose two versions to this criterion with different parameterization methods. Under the first parameterization method, we can define the first version as

$$\begin{aligned} \mathrm{DIC}_{\mathrm{rec}_1} &= -4\mathrm{E}_{\Theta}\left[\log L_{\mathrm{rec}}(\Theta|\mathbf{y})\right] + 2\log L_{\mathrm{rec}}(\hat{\Theta}|\mathbf{y}), \\ &= -4\mathrm{E}_{\Theta}\left[\ell_{\mathrm{rec}}(\Theta|\mathbf{y})\right] + 2\ell_{\mathrm{rec}}(\hat{\Theta}|\mathbf{y}), \\ &= \frac{-4}{M}\sum_{m=1}^{M}\ell_{\mathrm{rec}}^{(m)}(\Theta|\mathbf{y}) + 2\ell_{\mathrm{rec}}(\hat{\Theta}|\mathbf{y}), \end{aligned} \tag{10}$$

where $\hat{\Theta} = \bar{\Theta} = (\bar{\pi}, \bar{\mathbf{A}}, \bar{\theta})$ in the second term of the equation above are the posterior means of the parameters of the HMM obtaining after ending up the MCMC run, i.e

$$\ell_{\mathrm{rec}}(\bar{\Theta}) = \log L_{\mathrm{rec}}(\bar{\Theta}|\mathbf{y}) = \log\left(\sum_{\mathbf{z}} f(\mathbf{y},\mathbf{z}|\bar{\Theta})\right) = \log\left\{\sum_{\mathbf{z}}\left(p(\mathrm{z}_1|\bar{\pi})\left[\prod_{t=2}^{T} p(\mathrm{z}_t|\mathrm{z}_{t-1},\bar{\mathbf{A}})\right]\prod_{t=1}^{T} f(\mathrm{y}_t|\mathrm{z}_t,\bar{\theta})\right)\right\}.$$

Note that an approximation for the $\ell_{\mathrm{rec}}(\bar{\Theta})$ is obtained using the Algorithm 1 via just one iteration step, given parameter estimates $\bar{\Theta} = (\bar{\pi}, \bar{\mathbf{A}}, \bar{\boldsymbol{\theta}})$ summarized from the MCMC output. The effective number of parameters, denoted by $p_{\mathrm{DIC}_{\mathrm{rec}_1}}$, is then

$$\begin{aligned} p_{\mathrm{DIC}_{\mathrm{rec}_1}} &= -2\mathrm{E}_{\Theta}\left[\ell_{\mathrm{rec}}(\Theta)\right] + 2\ell_{\mathrm{rec}}(\bar{\Theta}), \\ &= \frac{-2}{M}\sum_{m=1}^{M}\ell_{\mathrm{rec}}^{(m)}(\Theta|\mathbf{y}) + 2\ell_{\mathrm{rec}}(\hat{\Theta}|\mathbf{y}). \end{aligned} \tag{11}$$

In the second parameterization method, we propose that the second term in the $\mathrm{DIC}_{\mathrm{rec}_1}$ can take another form. We can set a maximum a posteriori (MAP) estimate which is derived from the posterior density of the $\ell_{\mathrm{rec}}(\Theta)$ itself. In other words, we propose

$$\begin{aligned} \hat{\ell}_{\mathrm{rec(MAP)}}(\Theta) &= \log \hat{f}_{\mathrm{MAP}}(\mathbf{y},\mathbf{z}|\Theta) \\ &= \underset{f}{\mathrm{argmax}}\left\{\log f(\mathbf{y},\mathbf{z}|\Theta)\right\} \\ &= \underset{f}{\mathrm{argmax}}\ \log\left[\sum_{\mathbf{z}}\left(p(\mathrm{z}_1|\pi^{(m)})\left[\prod_{t=2}^{T} p(\mathrm{z}_t|\mathrm{z}_{t-1},\mathbf{A}^{(m)})\right]\prod_{t=1}^{T} f(\mathrm{y}_t|\mathrm{z}_t,\theta^{(m)})\right)\right] \\ &= \underset{\ell_{\mathrm{rec}}^{(m)}(\Theta)}{\mathrm{argmax}}\left\{\ell_{\mathrm{rec}}^{(m)}(\Theta)\right\}. \end{aligned}$$

Therefore, another version of the $\mathrm{DIC}_{\mathrm{rec}}$ can be proposed as follows

$$\mathrm{DIC}_{\mathrm{rec}_2} = -4\mathrm{E}_{\Theta}\left[\ell_{\mathrm{rec}}(\Theta)\right] + 2\hat{\ell}_{\mathrm{rec(MAP)}}(\Theta), \tag{12}$$

and

$$p_{\mathrm{DIC}_{\mathrm{rec}_2}} = -2\mathrm{E}_{\Theta}\left[\ell_{\mathrm{rec}}(\Theta)\right] + 2\hat{\ell}_{\mathrm{rec(MAP)}}(\Theta). \tag{13}$$

Note that the hidden states $\mathbf{z}$ in both $\mathrm{DIC}_{\mathrm{rec}_1}$ and $\mathrm{DIC}_{\mathrm{rec}_2}$ are dealt with as missing data. Also note that the first terms of the $\mathrm{DIC}_{\mathrm{rec}_2}$ and $p_{\mathrm{DIC}_{\mathrm{rec}_2}}$ in equations (12) and (13) respectively are similar to those used with the $\mathrm{DIC}_{\mathrm{rec}_1}$.

The second criterion proposed in this paper is the DIC based on the conditional recursive, denoted by $\mathrm{DIC}_{\mathrm{con}}$, which is inspired from the DIC_7 based on conditional likelihood in Celeux (2006). From the relationship in equation (13), there is still another possibility for obtaining a closed form for the observed likelihood function for the HMM via the conditional likelihood function $f(\mathbf{y}|\mathbf{z},\Theta)$. In this setting, obtaining a closed form of $f(\mathbf{y}|\Theta)$ requires integrating out the hidden states $\mathbf{z}$. It means, the hidden states $\mathbf{z}$ will be treated here as added parameters which will be estimated along with the parameters of the model Θ. Thus, an approximation to the observed log-likelihood based on a conditional likelihood function, $\ell_{\mathrm{con}}(\Theta)$, is defined as:

$$\hat{\ell}_{\mathrm{con}}(\Theta|\mathbf{y}) = \mathrm{E}_{\Theta,\mathbf{z}}\left[\log f(\mathbf{y}|\mathbf{z},\Theta)\right] = \int \log f(\mathbf{y}|\mathbf{z},\Theta) f(\mathbf{z}|\Theta) d\mathbf{z} \approx \frac{1}{M}\sum_{m=1}^{M} f(\mathbf{y}|\mathbf{z}^{(m)},\Theta^{(m)}).$$

Given a posterior sample of each parameter, $\Theta^{(m)} = (\pi^{(m)}, \mathbf{A}^{(m)}, \theta^{(m)})$ and hidden states $\mathbf{z}^{(m)}$ induced by an MCMC run, the proposed $\mathrm{DIC}_{\mathrm{con}}$ is given by

$$\begin{aligned}\mathrm{DIC}_{\mathrm{con}} &= -4\mathrm{E}_{\mathbf{z},\Theta}\left[\log f(\mathbf{y}|\mathbf{z},\Theta)\right] + 2\left[\log f(\mathbf{y}|\hat{\mathbf{z}},\hat{\Theta})\right] \\ &= \frac{-4}{M}\sum_{m=1}^{M}\left[\log f(\mathbf{y}|\mathbf{z}^{(m)},\Theta^{(m)})\right] + 2\left[\underset{\hat{f}}{\operatorname{argmax}}\left\{\log \hat{f}(\mathbf{y}|\mathbf{z},\Theta)\right\}\right] \\ &= \frac{-4}{M}\sum_{m=1}^{M}\left[\ell_{\mathrm{con}}^{(m)}(\Theta|\mathbf{y})\right] + 2\left[\underset{\ell_{\mathrm{con}}^{(m)}(\Theta|\mathbf{y})}{\operatorname{argmax}}\left\{\ell_{\mathrm{con}}^{(m)}(\Theta|\mathbf{y})\right\}\right],\end{aligned} \tag{14}$$

where the second term in equation (14) represents a MAP estimate of the conditional log-likelihood itself, and its effective number of parameters is

$$p_{\mathrm{DIC}_{\mathrm{con}}} = \frac{-2}{M}\sum_{m=1}^{M}\left[\log f(\mathbf{y}|\mathbf{z}^{(m)},\Theta^{(m)})\right] + 2\left[\underset{\ell_{\mathrm{con}}^{(m)}(\Theta)}{\operatorname{argmax}}\left\{\ell_{\mathrm{con}}^{(m)}(\Theta)\right\}\right]. \tag{15}$$

5. Methods

In this section, we describe the estimation process of parameters of the model and provide a simulation study involving synthetic and real data used in this study. We developed codes using Python (Python: v.2.7.10, see also Rossum (1995)) to generate the synthetic data, perform the estimation process and also model selection using the proposed criteria.

5.1 Fitting Algorithm

Before evaluating the proposed criteria, we have to estimate the parameters of the models adopted in this study which are the hidden states $\mathbf{z}$ and all parameters, $\Theta = (\boldsymbol{\pi}, \mathbf{A}, \mu, \sigma^2)$. We use the Gibbs sampler in the estimation process. We need to specify priors to the parameters. The priors with respect to the initial state distributions π and each row $\left\{a_{j.}\right\}$ in the transition matrix $\mathbf{A}$ are independently given the Dirichlet distribution with hyperparameter $\delta_k = 1$, $\forall\, j,k \in K$, where $K = 2, 3$, Fruhwirth-Schnatter (2006),

$$\left\{a_{j.}\right\} \text{ and } \boldsymbol{\pi} \overset{ind.}{\sim} Dir(1, 1, ..., 1_k).$$

For the parameters of the state-depended distributions μ and σ^2, we assumed a normal distribution as a conjugate prior on the mean parameter with hyperparameters $\lambda = 0$ and $\zeta = 0.01$, i.e., $\mu_k \sim \mathbb{N}(0, 100)$, $\forall\ k = 1, ..., K$, and $K = 2, 3$, and InvGamma with hyper-parameters $a = 0.1$ and $b = 0.1$ as a conjugate prior on the variance parameter, i.e., $\sigma_k^2 \sim InvGamma(0.1, 0.1)$, $\forall\ k = 1, ..., K$, and $K = 2, 3$. The full conditional posterior distributions of the parameters of the models, see Cappe et al. (2005), are given by

$$\begin{aligned}p(\pi|\mathbf{y},\mathbf{z},\mu,\sigma^2) &\propto \prod_{k=1}^{K}\pi^{N_k+\delta-1} = Dir(N_k+\delta_k), \\ p(a_{j.}|\mathbf{y},\mathbf{z},\mu,\sigma^2) &\propto \prod_{j=1}^{K}a_{j.}^{N_{j.}+\delta_j-1} = Dir(N_{j.}+\delta_j),\end{aligned}$$

where $N_k = \sum_{t=1}^{T}\mathbb{I}(\mathrm{z}_t = j)$ and $N_{jk} = \sum_{t=1}^{T}\mathbb{I}(\mathrm{z}_{t+1} = k, \mathrm{z}_t = j)$, for $j,\ k = 1, 2, ..., K$,

$$p(\mu_k|\mathbf{y},\mathbf{z},\sigma^2,\mathbf{A}) = \mathbb{N}\left\{\frac{\lambda\zeta + \sigma_k\sum_{t:\mathrm{z}_t=k}\mathrm{y}_t}{n_k\sigma_k + \lambda}, \frac{1}{n_k\sigma_k+\lambda}\right\},$$

where $n_j = \sum_{t=1}^{T}\mathbb{I}(\mathrm{z}_t = j)$ denotes the number of observations generated from the state j. The full conditional distribution of the variance parameter σ^2 is given by

$$p(\sigma_k^2|\mathbf{y},\mathbf{z},\mu,\mathbf{A}) = InvGamma\left\{a + \frac{n_k}{2}, b + \frac{\sum_{t:\mathrm{z}_t=k}(\mathrm{y}_t-\mu_k)^2}{2}\right\}.$$

We used the artificial identifiability constraints (Diebolt and Robert, 1994) on the mean parameter, such that, $\mu_k < \mu_{k+1}$, to avoid the label switching problem . We also adopted Gelman and Rubin statistics R (Gelman and Rubin, 1992), which based on multiple chains and the use of the between-sequence and within sequence variances, to check the convergence.

5.2 Study Design

We designed a Monte Carlo simulation study to evaluate the performance of our proposed criteria using synthetic data generated from normal HMMs and also a real application involving the waiting time of Old faithful geyser data.

5.2.1 Synthetic Data

In this sub-section, we generate synthetic data from a normal HMM, in which we observe normal variables $y_t \sim N(\mu_k, \sigma_k^2), k = 1, 2, ..., K$, whose parameters (μ_k, σ_k^2) depend on a hidden state z_t such that $z_1, z_2, ..., z_T$ is a Markov chain. We consider two assumptions. The first assumption involves data from models with a different number of states to examine the influence of diversity in number of states on behaviour of proposed criteria. In this setting, we assume that data have been generated from two and three state normal HMM respectively. In the second assumption, we assume different levels of the heterogeneity in the data by assuming different values of the variance parameter σ^2 for the normal HMMs adopted in the first assumption, to examine the effect of the heterogeneity in the data on the behaviour of proposed criteria. Under the assumptions above, we generate three synthetic databases, each one of length $T = 150$, with a different level of heterogeneity, from the following two and three state normal HMM:

$$\begin{pmatrix}\pi_1\\ \pi_2\end{pmatrix} = \begin{pmatrix}0.6\\ 0.4\end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix}0.9 & 0.1\\ 0.3 & 0.7\end{pmatrix}, \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix} = \begin{pmatrix}10\\ 20\end{pmatrix},$$

with three cases for the variance parameter $\sigma_k^2, k = 1, 2$, which are: $\sigma^2_{\text{case}_1} = (0.5, 0.2)$, $\sigma^2_{\text{case}_2} = (0.5, 1.2)$ and $\sigma^2_{\text{case}_3} = (0.5, 4)$, and,

$$\begin{pmatrix}\pi_1\\ \pi_2\\ \pi_3\end{pmatrix} = \begin{pmatrix}0.3\\ 0.3\\ 0.4\end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix}0.9 & 0.1 & 0.0\\ 0.1 & 0.8 & 0.1\\ 0.1 & 0.2 & 0.7\end{pmatrix}, \begin{pmatrix}\mu 1\\ \mu_2\\ \mu_3\end{pmatrix} = \begin{pmatrix}2\\ 6\\ 12\end{pmatrix},$$

with also three cases of σ_k^2, $k = 1, 2, 3$, which are: $\sigma^2_{\text{case}_1} = (0.5, 1, 1.2)$, $\sigma^2_{\text{case}_2} = (0.5, 1, 4)$ and $\sigma^2_{\text{case}_3} = (0.5, 2, 4)$. The density plots of these data are shown in the Figures (2-7).

5.2.2 Real Data

This study also includes a real application involving the waiting time of Old Faithful geyser data. The Old Faithful geyser data consist of 299 observations which represent the waiting time, in minutes, between two successive eruptions. Old Faithful is a geyser in the Yellowstone National Park in Wyoming, USA (Hardle, 1991). These data have been used by Robert and Titterington (1998); MacDonald and Zucchini (2009) for modelling normal HMMs with a different number of states. A histogram of these 299 observations is shown in Figure 8 which also shows plots of six fitted densities to be discussed later.

6. Results

We ran the DG sampler for 12000 iterations and adopted the last 10000 iterations, after discarding 2000 iterations as a burn-in period, to summarize the posterior chain of each conditional. Tables 1 and 2 show the results of the parameter estimates for the two and three state normal HMMs respectively.

Table 1. Parameter estimates of 2-state normal HMMs fitted to three synthetic data sets each one with a different level in the parameter of variance σ^2.

Case	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{a}_{11}$	$\hat{a}_{22}$	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$
$\sigma^2_{\text{case}_1}$	0.723	0.277	0.878	0.683	10.015	19.947	0.561	0.822
$\sigma^2_{\text{case}_2}$	0.791	0.209	0.914	0.686	10.003	19.739	0.568	1.633
$\sigma^2_{\text{case}_3}$	0.707	0.293	0.873	0.681	10.038	19.798	0.535	3.967

Table 2. Parameter estimates of 3-state normal HMMs fitted to three synthetic data sets each one with a different level in the parameter of variance σ^2.

Case	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{a}_{11}$	$\hat{a}_{12}$	$\hat{a}_{22}$	$\hat{a}_{23}$	$\hat{a}_{32}$	$\hat{a}_{33}$	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\sigma}_3^2$
$\sigma^2_{\text{case}_1}$	0.428	0.394	0.178	0.853	0.131	0.806	0.112	0.135	0.715	2.023	6.026	11.779	0.480	0.973	1.226
$\sigma^2_{\text{case}_2}$	0.578	0.324	0.101	0.918	0.061	0.794	0.096	0.332	0.534	2.074	6.115	12.681	0.497	1.021	2.735
$\sigma^2_{\text{case}_3}$	0.508	0.353	0.139	0.826	0.107	0.629	0.199	0.515	0.297	1.990	5.726	11.619	0.678	1.208	3.592

Regarding the presence of the label switching problem, we saw no evidence of label switching in the MCMC run for both adopted models. Perhaps the reason is that the $K!$ modes are well separated in the high-dimensional parameter space, see

Figures (2-7). For checking convergence, the convergence statistic $\sqrt{\hat{R}}$ of most parameters was below 1.2. Convergence is also suggested by the parameter estimates in the Table 1 and 2 compared with true parameters. Given MCMC samples to the hidden states $\mathbf{z}$ and all parameters $\Theta = (\boldsymbol{\pi}, \mathbf{A}, \theta)$, we can obtain the posterior distribution of the recursive log-likelihood using Algorithm 1 illustrated in Section 4. In the next step, we will evaluate the performance of the proposed criteria. We need to know whether the proposed criteria are able to select the correct model among competitive models. For this purpose, we assume six test models, as competing models, with a different number of states $K_{\text{test}} = 2, 3, ..., 7$ of each normal HMM. We fit these test models to the data that have been generated from a normal HMM with 2 states, and also to those generated from normal HMM with 3 states (assuming that 2 and 3 normal HMM are the true models), respectively.

6.1 Two-State Normal HMM

For 2-state normal HMM, we fit six test models with a different number of states $K_{\text{test}} = 2, 3, ..., 7$ using the Gibbs sampler. Figures 2-4 show the density plots of the six test models and the density of the original model (2-state normal HMM). Table 3 shows the results of all criteria and their the effective numbers of the parameters p_{DIC}, applied to the normal HMMs with 2 state. The values of the criteria are random quantities as they are based on the posterior distributions. Hence, their values might be unstable. For this purpose, we ran 10 independent MCMC runs, and calculated the averages and standard deviations. From Table 3, both recursive versions of the DIC, $\text{DIC}_{\text{rec}_1}$ and $\text{DIC}_{\text{rec}_2}$, behave well for selecting the correct model in all heterogeneity cases. On the other hand, the conditional-based DIC, DIC_{con}, began first in selecting the correct model at the first and second levels of heterogeneity and then tended to select a more complicated model with 7 states in the third level of heterogeneity. We can conclude that the DIC_{con} is not stable for non homogeneous data. From Figure 2 with the first heterogeneity level, $\sigma^2_{\text{case}_1} = (0.5, 0.2)$ and Figure 3 with the second heterogeneity level, $\sigma^2_{\text{case}_2} = (0.5, 1.2)$, it appears there is no a substantial improvement in the fit after the 2-state model.

Table 3. Estimates of the proposed criteria to the 2-state model with three levels of variance. The numbers in brackets indicate standard deviations.

Heter.level	K_{test}	$\text{DIC}_{\text{rec}_1}$	$p_{\text{DIC}_{\text{rec}_1}}$	$\text{DIC}_{\text{rec}_2}$	$p_{\text{DIC}_{\text{rec}_2}}$	DIC_{con}	$p_{\text{DIC}_{\text{con}}}$
	2	**387.901** (0.158)	12.306 (0.136)	**380.805** (0.022)	5.209 (0.018)	**210.994** (0.073)	10.839 (0.020)
	3	390.245 (0.304)	13.459 (0.005)	382.552 (0.356)	5.766 (0.005)	220.519 (5.958)	20.385 (9.594)
	4	391.327 (0.222)	14.084 (0.020)	383.927 (0.146)	6.684 (0.003)	216.222 (1.353)	16.874 (0.448)
Case_1	5	420.315 (1.946)	26.097 (1.204)	401.506 (0.003)	7.59 (0.001)	230.704 (11.685)	23.06 (5.249)
	6	428.591 (1.922)	31.348 (1.498)	405.297 (1.510)	8.054 (0.219)	249.075 (3.388)	32.353 (2.036)
	7	426.611 (0.047)	28.93 (0.408)	406.506 (1.104)	8.825 (0.397)	249.834 (1.706)	34.137 (4.674)
	2	**527.285** (0.002)	3.965 (0.001)	**527.171** (0.002)	3.852 (0.004)	**384.167** (0.140)	2.101 (0.101)
	3	530.53 (0.096)	5.782 (0.011)	529.619 (0.217)	4.872 (0.067)	399.436 (1.639)	17.281 (1.033)
	4	534.682 (0.424)	7.707 (0.036)	533.441 (0.591)	6.465 (0.095)	408.767 (2.912)	25.691 (4.328)
Case_2	5	538.211 (0.796)	9.283 (0.065)	536.831 (0.967)	7.903 (0.586)	404.749 (7.394)	20.751 (3.587)
	6	541.861 (0.287)	11.458 (0.352)	538.948 (0.054)	8.546 (0.084)	404.051 (5.361)	31.903 (3.044)
	7	549.227 (0.133)	15.327 (0.006)	543.262 (0.007)	9.363 (0.281)	400.004 (8.714)	39.081 (10.797)
	2	**697.541** (0.011)	4.456 (0.001)	**697.223** (0.026)	4.138 (0.002)	518.935(0.271)	3.255 (0.067)
	3	701.747 (0.136)	6.734 (0.398)	700.344 (0.181)	5.332 (0.026)	536.769 (4.459)	9.474 (6.283)
	4	704.813 (0.810)	10.737 (0.069)	701.865 (0.006)	8.594 (0.008)	595.930 (6.954)	49.743 (7.855)
Case_3	5	706.422 (2.435)	12.003 (1.322)	703.184 (0.420)	8.765 (0.056)	472.645 (21.804)	23.688 (17.753)
	6	710.066 (0.9129)	13.239 (0.693)	705.943 (0.398)	9.116 (0.259)	478.303 (7.554)	28.074 (7.804)
	7	721.466 (1.330)	19.973 (0.061)	714.277 (0.053)	11.281 (1.739)	**454.265** (24.039)	43.866 (16.544)

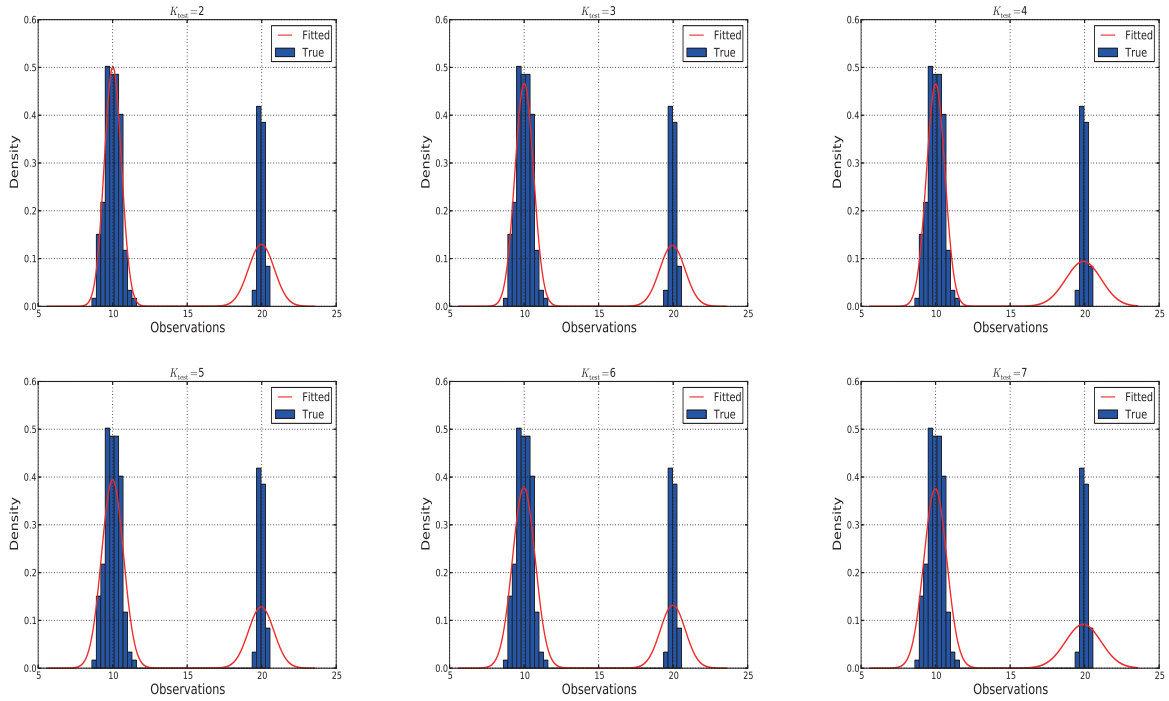

Figure 2. Fitting six test models to data generated from two state normal HMM with variance $\sigma^2_{\text{case}_1} = (0.5, 0.2)$.

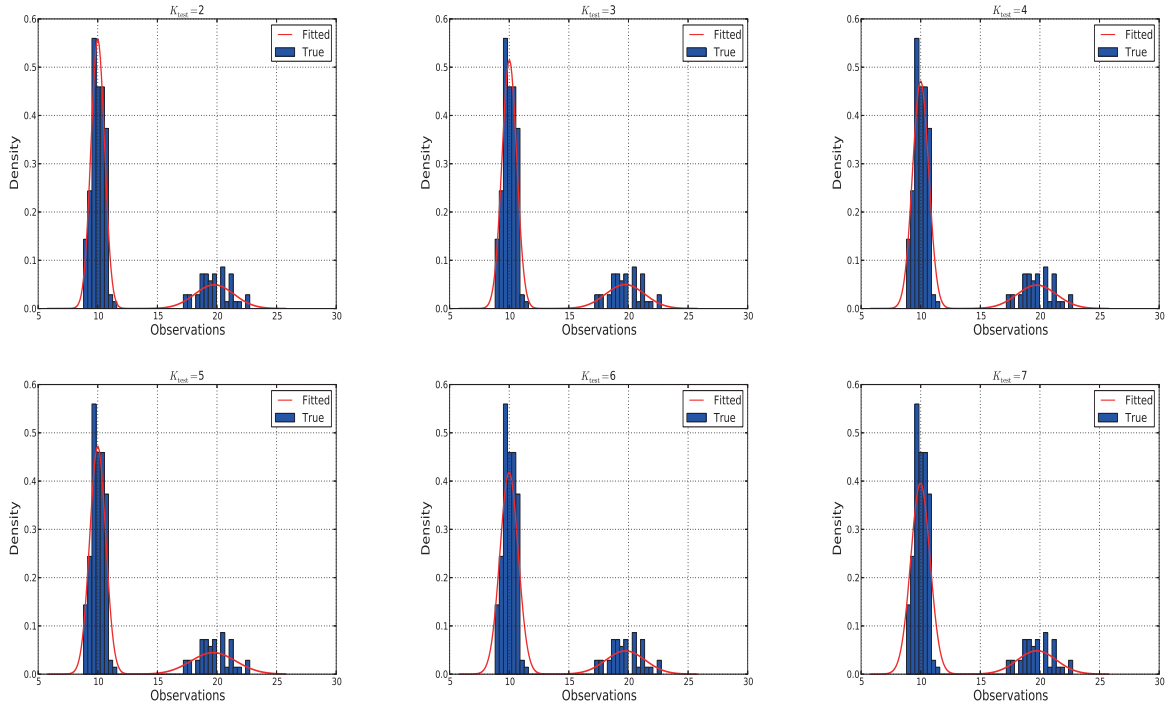

Figure 3. Fitting six test models to data generated from two state normal HMM with variance $\sigma^2_{\text{case}_2} = (0.5, 1.2)$.

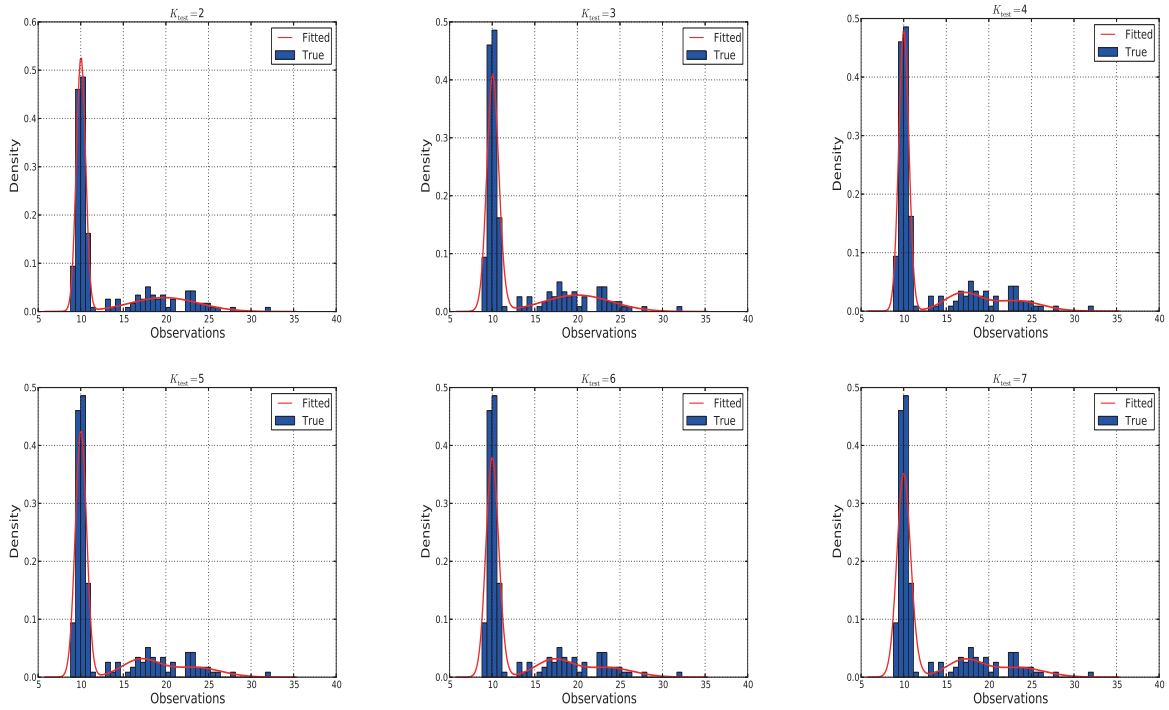

Figure 4. Fitting six test models to the data generated from two state normal HMM with variance $\sigma^2_{\text{case}_3} = (0.5, 4)$.

In contrast, in Figure 4 with the third heterogeneity level, $\sigma^2_{\text{case}_3} = (0.5, 4)$, there appears to be a fixed improvement after $K_{\text{test}} = 3$ which continues until the model with $K_{\text{test}} = 7$ which was selected by the DIC_{con}.

6.2 Three State Normal HMM

Regarding the model with 3 states, we also fit six test models with a different number of states $K_{\text{test}} = 2, 3, ..., 7$ to the normal HMM with three states using the Gibbs sampler.

Table 4. Estimates of the proposed criteria to the 3-state model with three levels of variance. The numbers in brackets indicates standard deviations.

Heter. level	K_{test}	$\text{DIC}_{\text{rec}_1}$	$p_{\text{DIC}_{\text{rec}_1}}$	$\text{DIC}_{\text{rec}_2}$	$p_{\text{DIC}_{\text{rec}_2}}$	DIC_{con}	$p_{\text{DIC}_{\text{con}}}$
	2	820.423 (0.002)	19.073 (0.001)	805.3 (0.001)	3.949 (0.002)	728.499 (0.141)	31.081 (0.035)
	3	**649.153** (0.009)	12.608 (0.003)	**645.090** (0.064)	8.546 (0.009)	**450.866** (0.059)	14.074 (0.016)
	4	659.290 (1.771)	17.881 (1.587)	650.311 (0.486)	8.902 (0.126)	481.773 (3.255)	46.334 (13.12)
Case_1	5	661.236 (0.354)	18.353 (0.373)	652.864 (0.001)	9.980 (0.001)	486.658 (1.881)	56.897 (1.874)
	6	672.013 (5.217)	23.906 (0.937)	658.443 (3.788)	10.336 (0.212)	490.696 (12.499)	62.727 (10.866)
	7	678.13 (0.005)	25.488 (0.004)	663.621 (0.001)	10.979 (0.001)	496.985 (11.585)	71.688 (11.297)
	2	698.171 (0.004)	9.210 (0.051)	692.991 (0.152)	4.030 (0.032)	622.498 (0.36)	20.557 (0.173)
	3	**616.395** (0.034)	12.268 (0.039)	**612.602** (0.096)	8.475 (0.022)	**451.234** (0.272)	13.397 (0.143)
	4	628.073 (0.111)	18.068 (0.19)	618.572 (0.086)	8.567 (0.007)	470.501 (3.687)	29.1 (3.874)
Case_2	5	637.424 (0.532)	21.302 (0.323)	624.637 (0.029)	8.516 (0.179)	479.162 (1.865)	35.248 (1.567)
	6	641.049 (0.468)	23.318 (0.625)	627.707 (0.077)	9.977 (0.081)	489.266 (1.408)	56.376 (1.364)
	7	648.605 (1.121)	25.224 (1.559)	633.625 (1.074)	10.244 (0.488)	482.601 (4.274)	52.642 (4.165)
	2	906.724 (0.019)	34.807 (0.992)	874.919 (0.548)	3.002 (0.001)	778.038 (0.937)	32.4 (0.114)
	3	**831.099** (0.186)	21.826 (0.121)	**817.135** (0.003)	7.863 (0.005)	604.663 (0.414)	23.084 (0.025)
	4	853.851 (0.665)	29.939 (0.593)	831.805 (0.027)	6.393 (0.006)	643.596 (2.653)	56.963 (2.679)
Case_3	**5**	836.487 (2.131)	24.322 (2.547)	823.284 (0.044)	11.118 (0.005)	**596.312** (4.923)	57.652 (4.454)
	6	852.345 (0.096)	30.447 (0.043)	832.107 (0.046)	10.209 (0.012)	609.018 (6.280)	67.645 (6.340)
	7	846.199 (1.951)	33.579 (2.985)	825.228 (0.022)	12.608 (0.230)	607.114 (1.203)	86.546 (2.535)

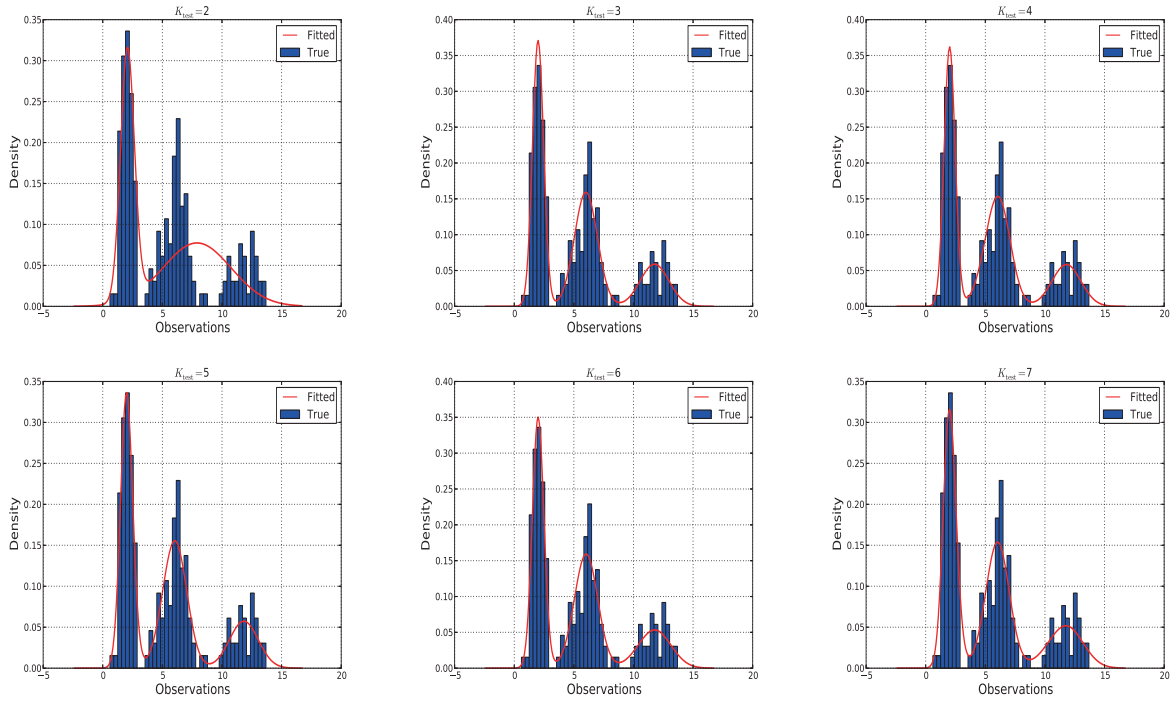

Figure 5. Fitting six test models to data generated from three state normal HMM with variance $\sigma^2_{\text{case}_1} = (0.5, 1, 1.2)$.

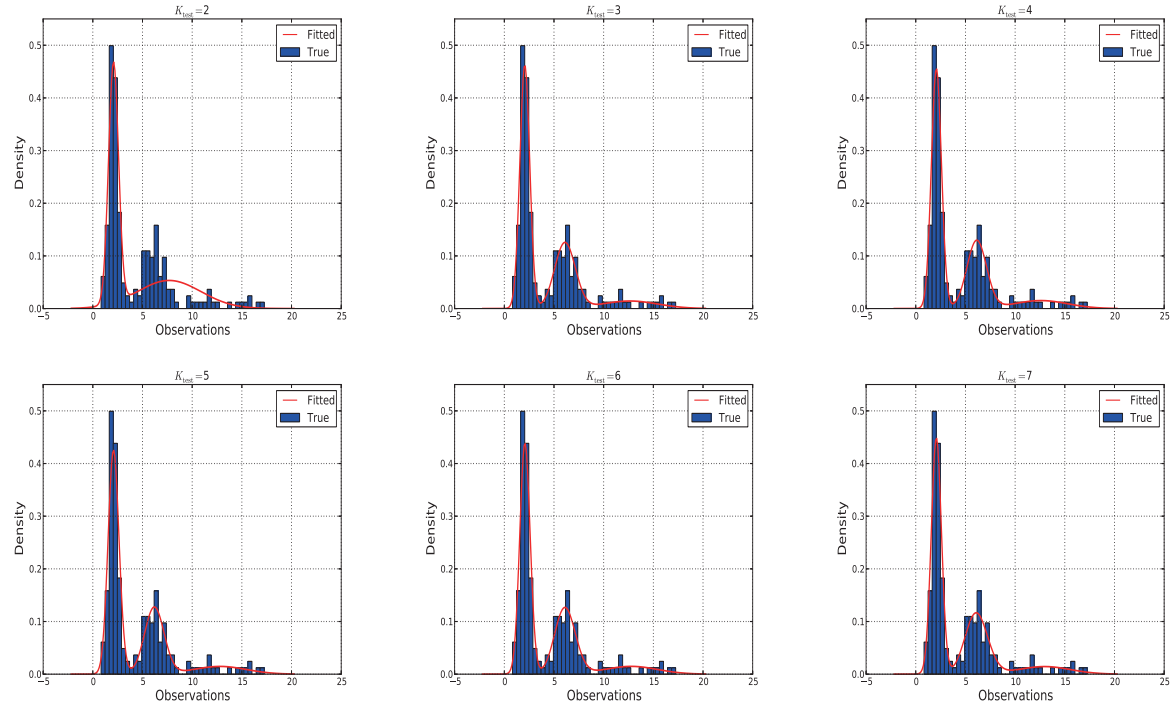

Figure 6. Fitting six test models to the data generated from three state normal HMM with variance $\sigma^2_{\text{case}_2} = (0.5, 1, 4)$.

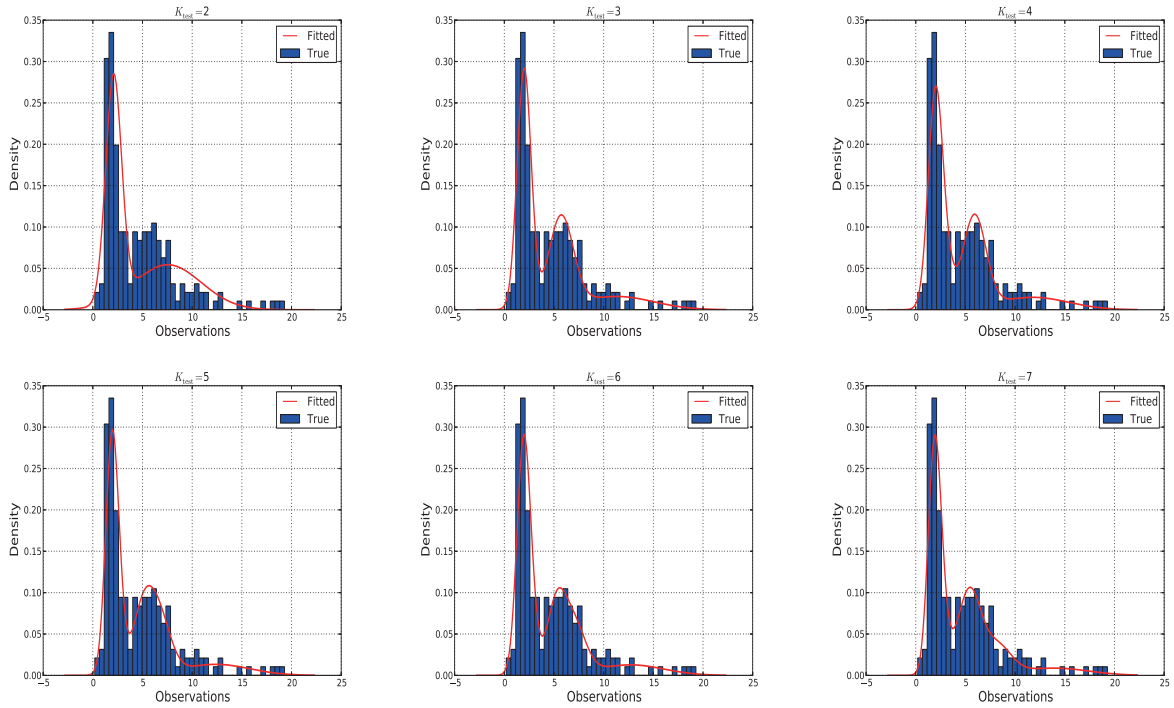

Figure 7. Fitting six test models to the data generated from three state normal HMM with variance $\sigma^2_{\text{case}_3} = (0.5, 2, 4)$.

We also used 10 MCMC runs to compute the estimates and the standard deviations of the proposed criteria . We can see from Table 4 that both versions of the DICs_{rec} are also selecting the correct model, i.e. $K_{\text{test}} = K_{\text{true}} = 3$, for all the cases of heterogeneity. This suggests, both proposed criteria, $\text{DIC}_{\text{rec}_1}$ and $\text{DIC}_{\text{rec}_2}$, were not affected by any level of the assumed heterogeneity and followed a fixed behavior although of the diversity in the number of the states. Conversely, the DIC_{con} behaves in the same way as for the normal HMM with 2 states. It also selects the correct model in the first and second cases of the heterogeneity and then tends to select the more complicated model, $K_{\text{test}} = 5$, in the third heterogeneity level. Figures 5-7 illustrate the densities of six test models fitting to the synthetic data generated from a 3-state normal HMM. We can clearly see from the Figures 5 and 6 with the heterogeneity levels, $\sigma^2_{\text{case}_1} = (0.5, 1, 1.2)$ and $\sigma^2_{\text{case}_2} = (0.5, 1, 4)$, respectively that there is no improvement in the fitting after the model with $K_{\text{test}} = 3$. In contrast, a slight improvement is observed in the fitting in Figure 7, $K_{\text{test}} = 7$, with the third heterogeneity level, $\sigma^2_{\text{case}_3} = (0.5, 2, 4)$. Nevertheless, the $K_{\text{test}} = 5$ was more favourable state as it corresponds to the smallest value for the criterion DIC_{con}.

6.3 Real Application

In this sub-subsection, we look at the performance of the proposed criteria, DICs_{rec} and DIC_{con}, on models fitted to the waiting time of Old Faithful geyser data that have been described in the Section 5. We follow the same scenario used to examine the proposed criteria with synthetic data. We also fit six normal HMMs as test models with $K_{\text{test}} = 2, 3, ..., 7$, to the waiting time of Old Faithful geyser data. We used the Gibbs sampler to estimate these test models. Figure 8 shows the plots of the densities of the six test models fitted to the waiting time of Old Faithful geyser data.

Using the proposed criteria, we try to decide which model adequately fits the data. Each one of the six test models has been estimated using the DG sampler with the same information used to estimate models generated from the simulation data (sub-section 5.1), which are the priors, the burn-in period, the number of adopted iterations for analysis and the number of MCMC runs (10 runs) used for checking the randomness of the criteria values. Table 5 shows the averages of the proposed criteria $\text{DIC}_{\text{rec}_1}$, $\text{DIC}_{\text{rec}_2}$ and DIC_{con}, their effective numbers of parameters and the standard deviations appearing in parentheses. We can see from Table 5 that estimates of both DICs_{rec} remain stable over all test states K_{test} compared with estimates of the DIC_{con} which have high standard deviations starting after state $K_{\text{test}} = 5$. We can also see from Table 5 that both versions of the DIC_{rec} select the normal HMM with $K_{\text{test}} = 5$, while the DIC_{con} prefers a more complicated model, $K_{\text{test}} = 6$. Note that the model with $K_{\text{test}} = 6$, as shown in the Figure 8, can provide a more adequate fitting to the data. However, it is not often favoured in applications since a large number of parameters can lead to high variance in parameter estimates and overfitting.

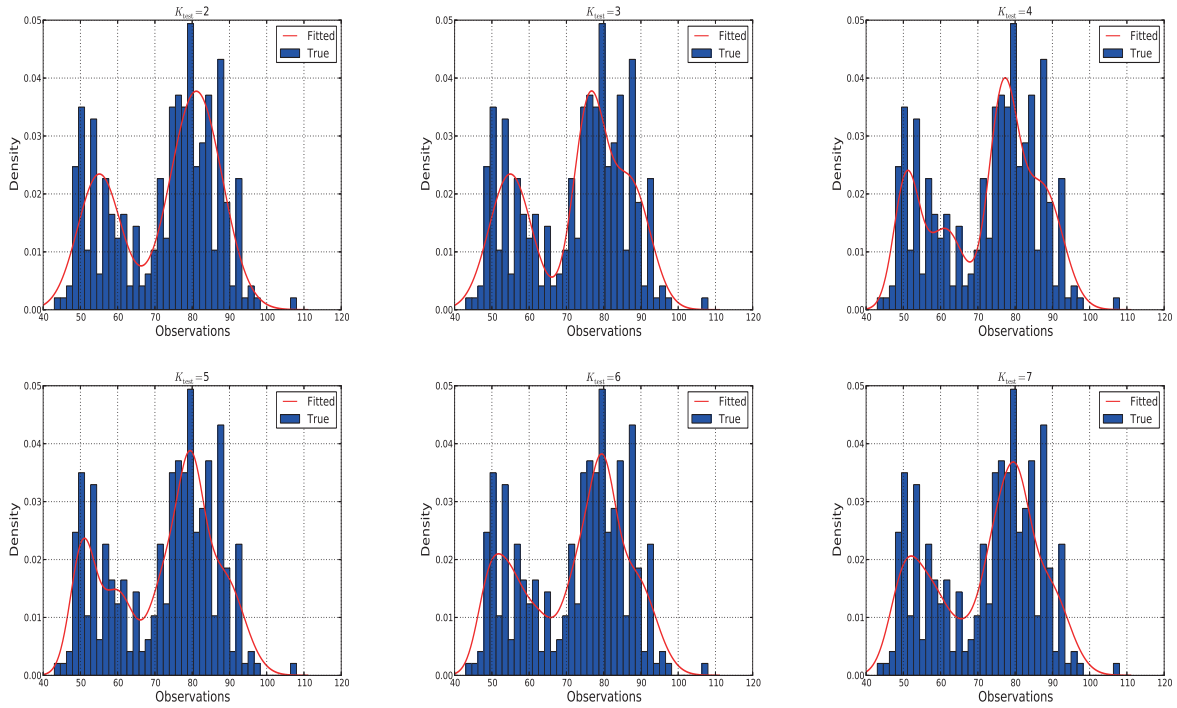

Figure 8. Histograms of the densities fitted of six tested models to a normal HMM fitted to the waiting time of faithful geyser data. The number of adopted iterations is 10000 (main) and discarded is 2000 iterations (burn-in).

Table 5. Results the estimates of the proposed criteria obtained from fitting a normal HMM with a different number of states, to the waiting time of Old Faithful geyser data.

K_{test}	$\text{DIC}_{\text{rec}_1}$	$p_{\text{DIC}_{\text{rec}_1}}$	$\text{DIC}_{\text{rec}_2}$	$p_{\text{DIC}_{\text{rec}_2}}$	DIC_{con}	$p_{\text{DIC}_{\text{con}}}$
2	2223.342 (0.107)	8.708 (0.112)	2217.312 (0.032)	2.717 (0.002)	1956.482 (0.035)	3.618 (0.037)
3	2162.555 (0.001)	18.885 (0.032)	2149.124 (0.052)	5.064 (0.023)	1767.888 (0.565)	17.558 (0.378)
4	2151.579 (0.213)	24.786 (0.016)	2132.962 (0.077)	6.368 (0.112)	1651.496 (0.858)	32.069 (0.702)
5	**2139.382** (0.034)	23.988 (0.654)	**2122.987** (0.532)	7.060 (0.269)	1603.155 (1.055)	63.670 (1.762)
6	2159.143 (0.788)	31.144 (0.612)	2133.861 (0.099)	6.674 (0.087)	**1596.483** (1.705)	79.707 (2.654)
7	2177.822 (1.010)	42.302 (1.719)	2141.474 (0.132)	4.933 (0.204)	1628.212 (3.022)	109.711 (4.112)

7. Discussion and Future Research

We have illustrated two main ways to obtain a closed form for the observed likelihood of HMMs. Hence, two criteria are proposed. The first proposed criterion, the recursive deviance-based DIC, is based on the observed likelihood obtained by summing all possible states of the complete data likelihood using forward recursion. We also have developed two different parameterization strategies to that criterion. The second proposed criterion, the conditional deviance-based DIC, is based on the observed likelihood obtained by integrating out the hidden states of the conditional likelihood. In order to examine the proposed criteria, we have performed a simulation study involving the assumptions of diversity in the number of states and the heterogeneity in the data, with the goal of understanding the effect of these assumptions on the behavior of the proposed criteria. The simulation study has demonstrated that both versions of the recursive deviance-based DIC are able to select the correct model and they are robust to sensitivity tests conducted in this study. On the other hand, the conditional deviance-based DIC agrees first with the recursive deviance-based DICs in selecting the correct model for the 2 and 3 state models, especially at moderate levels of heterogeneity. It then tends to select a more complex model when the data have a high level of heterogeneity. We can conclude that the high heterogeneity in the data leads to producing new modes and this may explain why more complicated models are selected by some criteria such as the DIC_{con}. In the real application involving the waiting time of the Old Faithful geyser data, we have found that the conditional deviance-based DIC tends to select more complicated models, with 6 states, while, the DICs based on recursive deviances select

a less complexity model, with 5 states, and have more stable estimates. Some literature has pointed out different results with respect to select an HMM fitted for the waiting time of the Old Faithful geyser data. For instance, MacDonald and Zucchini (2009) have obtained 4 states using the AIC for those data. They only investigated 2 to 4 states and did not consider a larger number to test the their model fitting to those data. Their work was based on an MLE estimator which in itself did not account for uncertainty in parameter estimation. Graphically, Robert and Titterington (1998) proposed that 3 states are an adequate for fitting those data. Although they used the Bayesian approach to estimate the model, they did not use a particular criterion to support their suggestion.

We can illustrate some remarks of interesting with respect to the proposed criteria. Firstly, we have noted that the estimates obtained with the conditional deviance-based DIC were smaller than those obtained by both versions of recursive deviance-based DICs. In addition, they were more variable as shown from their standard deviations in the Tables 3 and 4. In contrast, the estimated values obtained from both DICs_{rec} are more stable. The unstable estimates of the DIC_{con} may be attributed the problem of the unbounded likelihood, which is not discussed in this paper. Briefly, when one puts a mean of a state-dependent distribution equal to one of the observations and allows the corresponding variance to tend to zero, the likelihood becomes arbitrarily large, and hence the DIC becomes small. This problem often occurs when densities rather than probabilities are used in the likelihood. For more details about this problem, see Robert and Titterington (1998), MacDonald and Zucchini (2009, p.50) and Fruhwirth-Schnatter (2006, p.334). This case is avoided in the recursive likelihood as it is bounded by 0 and 1. More clearly, in the recursive calculation of the observed likelihood function using the forward recursion, we use the scaling procedure (Rabiner, 1989) for normalizing the forward variables to prevent the underflow problem. Hence, the densities (i.e. the state-dependent densities involved in the definition of the likelihood) are automatically converted into probabilities which sum up to 1. Secondly, we have noted that both versions of the recursive deviance-based DIC, $\text{DIC}_{\text{rec}_1}$ and $\text{DIC}_{\text{rec}_2}$, have somewhat similar values and stable standard deviations as shown in Tables 3 and 4. This stable behaviour of both versions is not surprising as the first term of both criteria is the same and the second term of both the $\text{DIC}_{\text{rec}_1}$ and the $\text{DIC}_{\text{rec}_2}$, is essentially derived from the same posterior distribution of the recursive likelihood $(\ell_{\text{rec}}^{(1)}(\Theta), \ell_{\text{rec}}^{(2)}(\Theta), ..., \ell_{\text{rec}}^{(M)}(\Theta))$, (see the definitions of the $\text{DIC}_{\text{rec}_1}$ and the $\text{DIC}_{\text{rec}_2}$ in Section 4). Third, the recursive likelihood used in the DICs_{rec}, is evaluated using the forward recursion with a low computational cost, $O(TK^2)$, compared with the traditional methods that require $O(TK^T)$ calculations. Nevertheless, this recursive calculation is expensive from a computational viewpoint compared with the conditional likelihood, DIC_{con}, which requires only $O(T)$ of calculations.

In conclusion, we suggest that both versions of the recursive deviance-based DICs, as they have somewhat similar values, may be a useful guide to the model choice problem for more general state-space models where the likelihood is often unavailable in closed form. We can also propose introducing a practical study to examine the computational time of both criteria, the recursive deviance-based DICs and the conditional deviance-based DICs. We leave this proposal for future research.

References

Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle* (In B. N. Petrov and F. Csaki (Eds.), Second international symposium on information theory (pp. 267-281) ed.). Budapest: Academiai Kiado.

Albert, J. H. , & Chib S. (1993). Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *Journal of Business and Economic Statistics, 11*, 1-15. http://dx.doi.org/10.2307/1391303

Baum, L. E., Petrie T. , Soules G. , & Weiss N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics, 41*(1), 164-171. http://dx.doi.org/10.1214/aoms/1177697196

Bhar, R. , & Hamori S. (2004). *Hidden Markov Models: Applications to Financial Economics*. Advanced Studies in Theoretical and Applied Econometrics. Dordrecht, The Netherlands: Kluwer Academic Pubilshers.

Billio, M., Monfort A., & Robert C. (1999). Bayesian estimation of switching ARMA models. *Journal of Econometrics, 93*, 229-255. http://dx.doi.org/10.1016/S0304-4076(99)00010-X

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, Heidelberg.

Cappe, O., Moulines E. , & Ryden T. (2005). *Inference in Hidden Markov Models* (2nd ed.). Springer-Verlag, Berlin, Heidelberg, NewYork.

Casella, G., & Robert C. P. (2004). *Monte Carlo Statistical Methods* (2nd ed.). New York: Springer-Verlag.

Celeux, G., Forbes F. , Robert C. P. , & Titterington D. M. (2006). Deviance information criteria for missing data models.

Bayesian Analysis(1), 651-674. http://dx.doi.org/10.1214/06-BA122

Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics, 75*, 79-97. http://dx.doi.org/10.1016/0304-4076(95)01770-4

Derrode, S., & Pieczynski L. (2006). Contextual estimation of hidden Markov chains with application to image segmentation. *In ICASSP*(2), 689-692. http://dx.doi.org/10.1109/icassp.2006.1660436

Diebolt, J. & Robert, C.P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the American Statistical Association, 96*, 194-209.

Fruhwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. Springer, New York.

Gelman, A. & Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457-511. http://dx.doi.org/10.1214/ss/1177011136

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. http://dx.doi.org/10.1109/TPAMI.1984.4767596

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica, 57*, 357-384. http://dx.doi.org/10.2307/1912559

Hardle, W. (1991). *Smoothing techniques*. Springer Series in Statistics, New York. http://dx.doi.org/10.1007/978-1-4612-4432-5

MacDonald, I. L. , & Zucchini W. (2009). *Hidden Markov Models for Time Series: An Introduction Using R*. Chapman and Hall, London.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceeding of the IEEE, 77*(2), 257-286. http://dx.doi.org/10.1109/5.18626

Robert, C. , & Titterington D. (1998). Reparameterization strategies for hidden Markov models and Bayesian approaches to maximum-likelihood estimation. *Statist. & Computing, 8*, 145-158. http://dx.doi.org/10.1023/A:1008938201645

Robert, C. P., Celeux G. , & Diebolt J. (1993). Bayesian estimation of hidden Markov models: A stochastic implementation. *Statistics and Probability Letters, 16*(1), 77-83. http://dx.doi.org/10.1016/0167-7152(93)90127-5

Rossum, G. V. (May 1995). *Python tutorial*. Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam. (see also: http://python.org/)

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics, 6*, 461-464. http://dx.doi.org/10.1214/aos/1176344136

Scott, S. L. (2002). Bayesian mehtods for Hidden Markov Models: Recursive computing in the 21th century. *Journal of the American Statistical Association, 97*, 337-351. http://dx.doi.org/10.1198/016214502753479464

Spiegelhalter, D. J., Best N. G. , Carlin B. P. , & van der Linde A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of Royal Statistical Society B, 64*, 583-639. http://dx.doi.org/10.1111/1467-9868.00353

Tanner, M. Y., & Wong W. H. (1987). The Calculation of Posterior Distribution by Data Augmentation. *Journal of the American Statistical Association, 82*, 528-550. http://dx.doi.org/10.1080/01621459.1987.10478458

Visser, I., Maartje E. , & Peter C. (2002). Fitting Hidden Markov Models to psychological data. *Scientific Programming, 10*, 185-199. http://dx.doi.org/10.1155/2002/874560